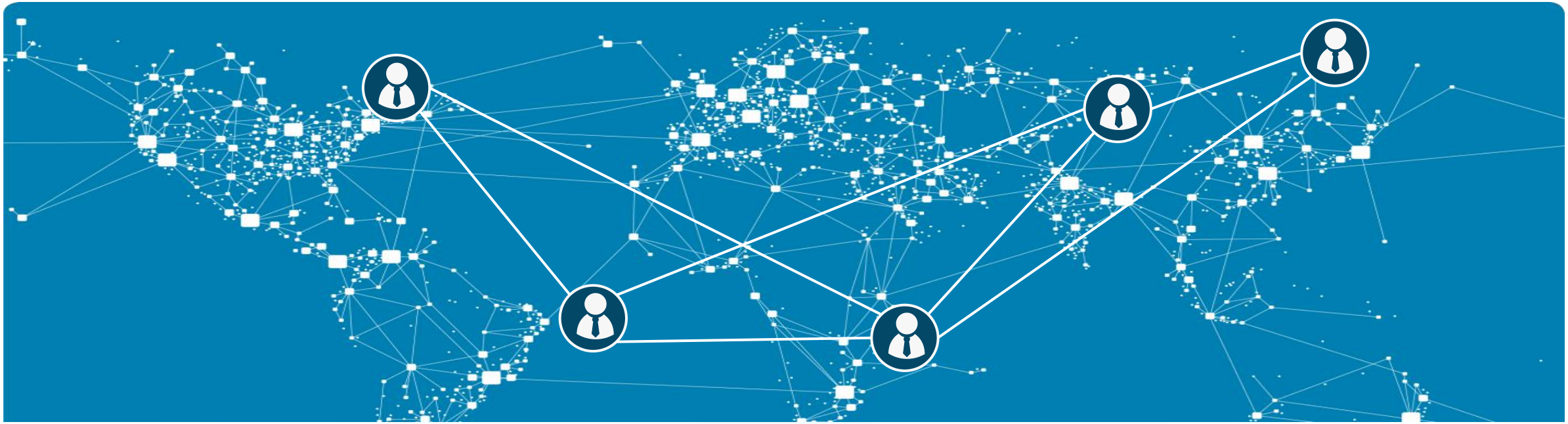
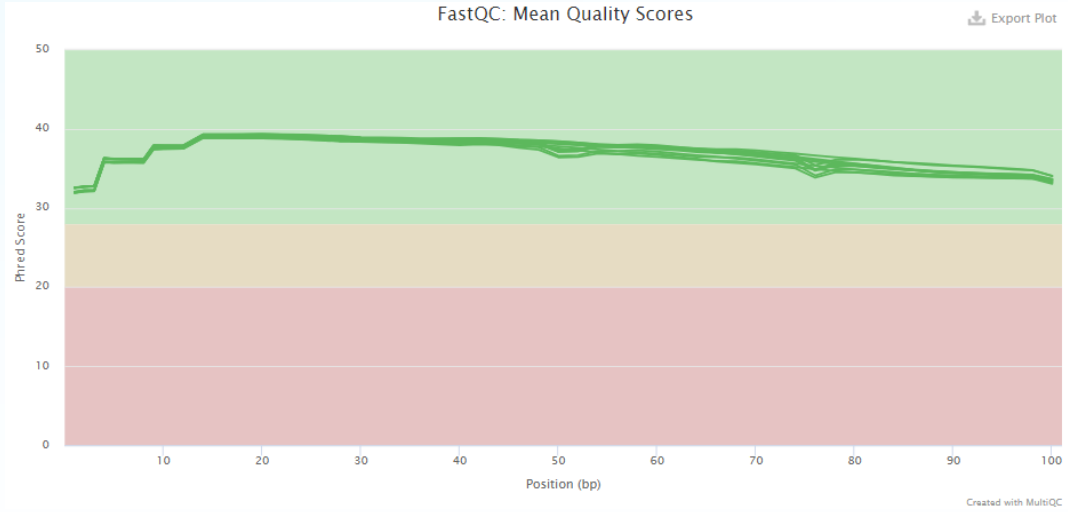
微生物组—宏基因组分析专题研讨会第18期



22质控和去宿主

易生信

2023年4月8日



数据分析的基本思想——三步走

大数据 → 大表 → 小表 → 图

```
@HISEQ:549:HLYNYBCXY:1:1101:1267:2220 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGTTGGGCG
+
DDDDDIHHHIIIIIIIIHIIHIIIIIIIIIIHIHIHIIIIIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:1887:2204 1:N:0:CACTCAAT
TACGAGTATGAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTCTA
+
DDDD@H<GHIIIIIIIIIIIIIIIIIIIIHIIIHIIIIIIIIIIGIIIIIIIIFH
@HISEQ:549:HLYNYBCXY:1:1101:2196:2168 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCCATAAACGATGACAA
+
DDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:2025:2183 1:N:0:CACTCAAT
ATATCGCGAGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGCG
+
DDDD@E@HIGHIIIHHFHIIHIIIIFHHIIIHHGIHIIHIIICHDEHHIIIHGH
@HISEQ:549:HLYNYBCXY:1:1101:2052:2198 1:N:0:CACTCAAT
CACGAGACAGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGGGTA
+
D@DD@H=?CCHHIIIIIIIIIIIIIIIIIIIIIIIIIIIGI0CHIIIIIHIHIIH@
```

序列: 10^6 ~ 10^9

ID	WT6	WT3	OE4	WT2	OE3	WT1
OTU_265	18	18	6	11	20	15
OTU_36	63	77	57	194	155	163
OTU_102	20	44	18	77	18	43
OTU_49	106	92	25	137	76	65
OTU_270	9	5	22	5	22	5
OTU_1865		0	3	0	0	2
OTU_58	77	75	28	84	53	64
OTU_1110		6	3	3	2	2
OTU_30	100	142	78	111	124	145
OTU_51	87	79	21	38	42	102
OTU_1353		0	1	2	0	1
OTU_1137		0	1	0	3	0
OTU_18	166	150	126	318	130	265
OTU_4	498	343	189	804	224	626
OTU_3	459	690	340	1039	568	580
OTU_704	3	14	12	8	9	4
OTU_14	176	283	110	314	169	232

特征表: $10^{1\sim3}$ X $10^{3\sim5}$

Sample	berger_parker	buzas_gibson		chao1
WT6	0.042	0.0381	1388.9	0.992	0.817
WT3	0.0453	0.0425	1474.9	0.992	0.828
OE4	0.0359	0.0414	1476.4	0.993	0.828
WT2	0.0642	0.0244	1203.0	0.985	0.773
OE3	0.0426	0.0396	1716.9	0.991	0.807
WT1	0.0586	0.0293	1317.0	0.988	0.788
WT4	0.0518	0.0359	1353.2	0.991	0.813
OE5	0.0361	0.0441	1622.8	0.993	0.824
OE2	0.0466	0.0472	1733.3	0.992	0.827
OE6	0.0432	0.0523	1759.5	0.994	0.840
WT5	0.0435	0.0252	1181.6	0.987	0.776
OE1	0.0374	0.0524	1591.2	0.994	0.852
KO4	0.0558	0.0325	1474.1	0.990	0.796
KO1	0.0552	0.0409	1651.6	0.990	0.813
KO5	0.0732	0.025	1306.2	0.986	0.772
KO2	0.0509	0.0445	1675.3	0.992	0.825
KO3	0.0571	0.0329	1489.8	0.990	0.800
KO6	0.0518	0.0334	1215.9	0.991	0.813

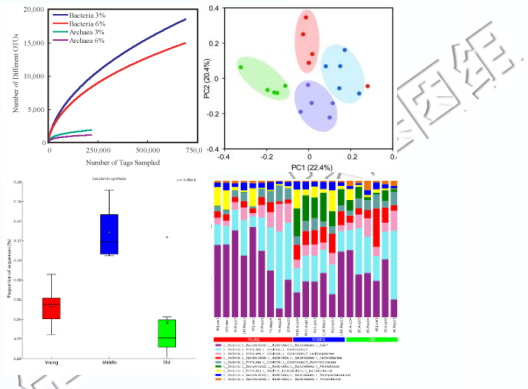


图: $10^{1\sim3}$个点和统计信息

宏基因组有参分析基本思路

16S rRNA基因扩增子

宏基因组

u/vsearch

QIIME 2

MetaPhlAn2

Kraken

HUMAnN2/bowtie2/diamond

物种组成

	Sample 1	Sample 2	Sample 3
OTU _1	4	0	2
OTU _2	1	0	0
OTU _3	2	4	2

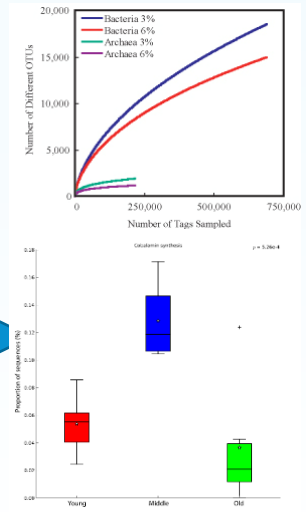
PICRUSt

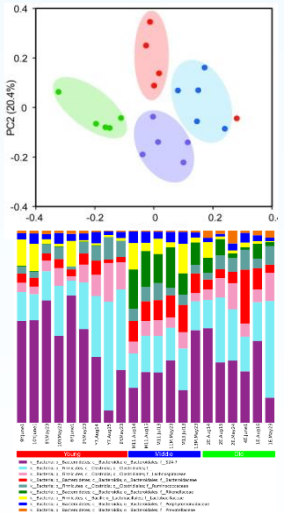
Tax4Fun

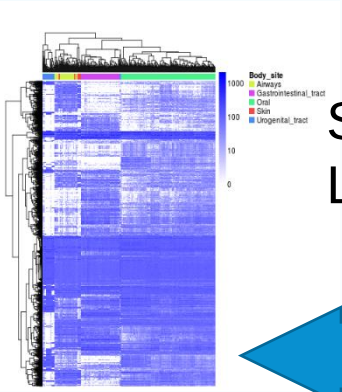
	Sample 1	Sample 2	Sample 3
K00001	20	15	18
K00002	1	2	0
K00003	4	5	4

功能组成

STAMP / LEfSe / R





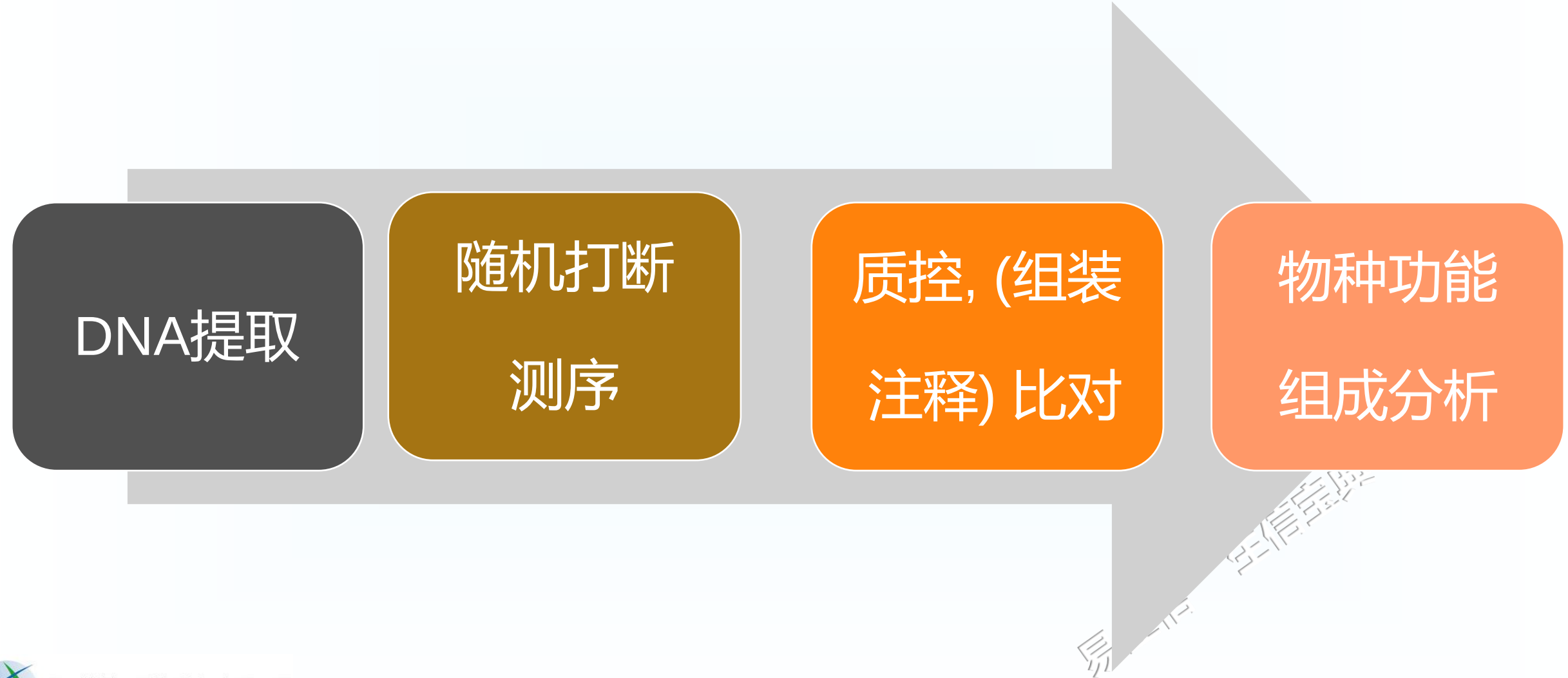


STAMP / LEfSe / R

Protein Cell：扩增子和宏基因组数据分析实用指南

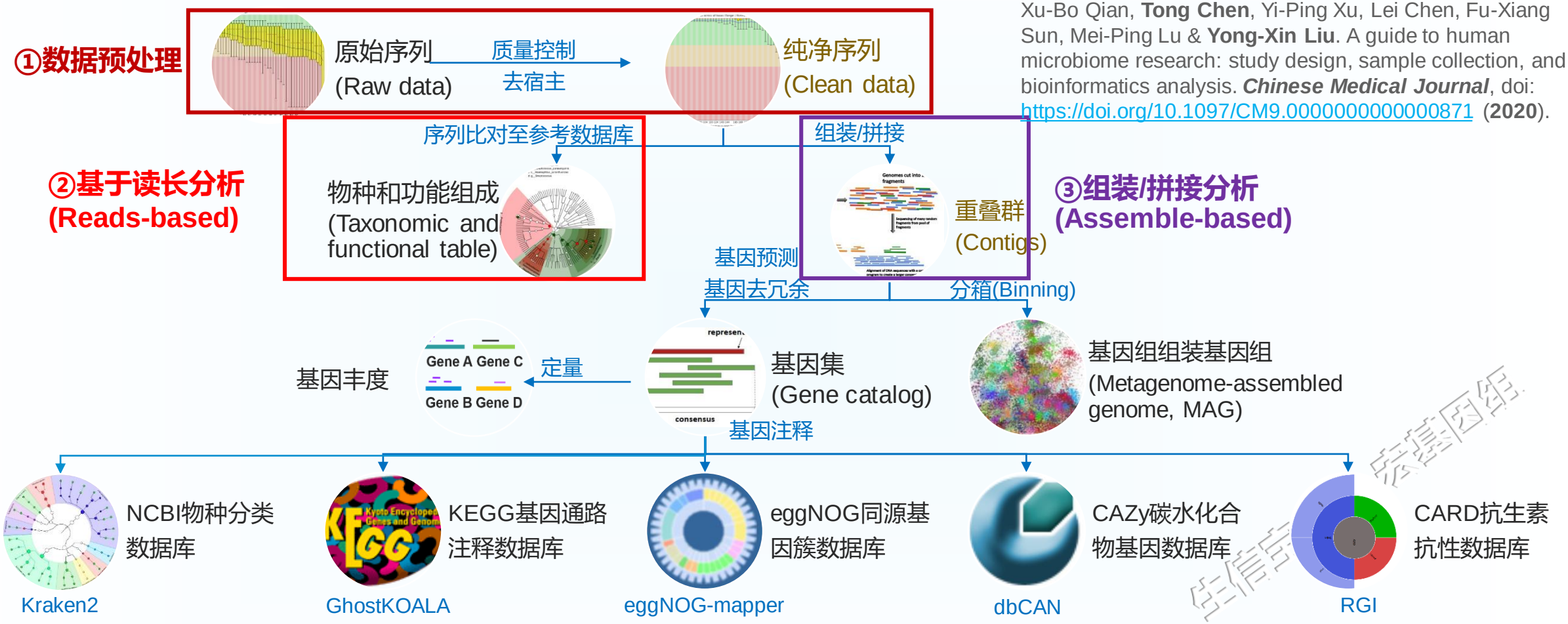
熟记此图，胸中有丘壑

宏基因组实验分析流程



宏基因组分析流程

Xu-Bo Qian, **Tong Chen**, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu & **Yong-Xin Liu**. A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. ***Chinese Medical Journal***, doi: https://doi.org/10.1097/CM9.0000000000000871 (**2020**).



常用物种和功能基因注释数据库(图标右)和对应的软件(图标下)

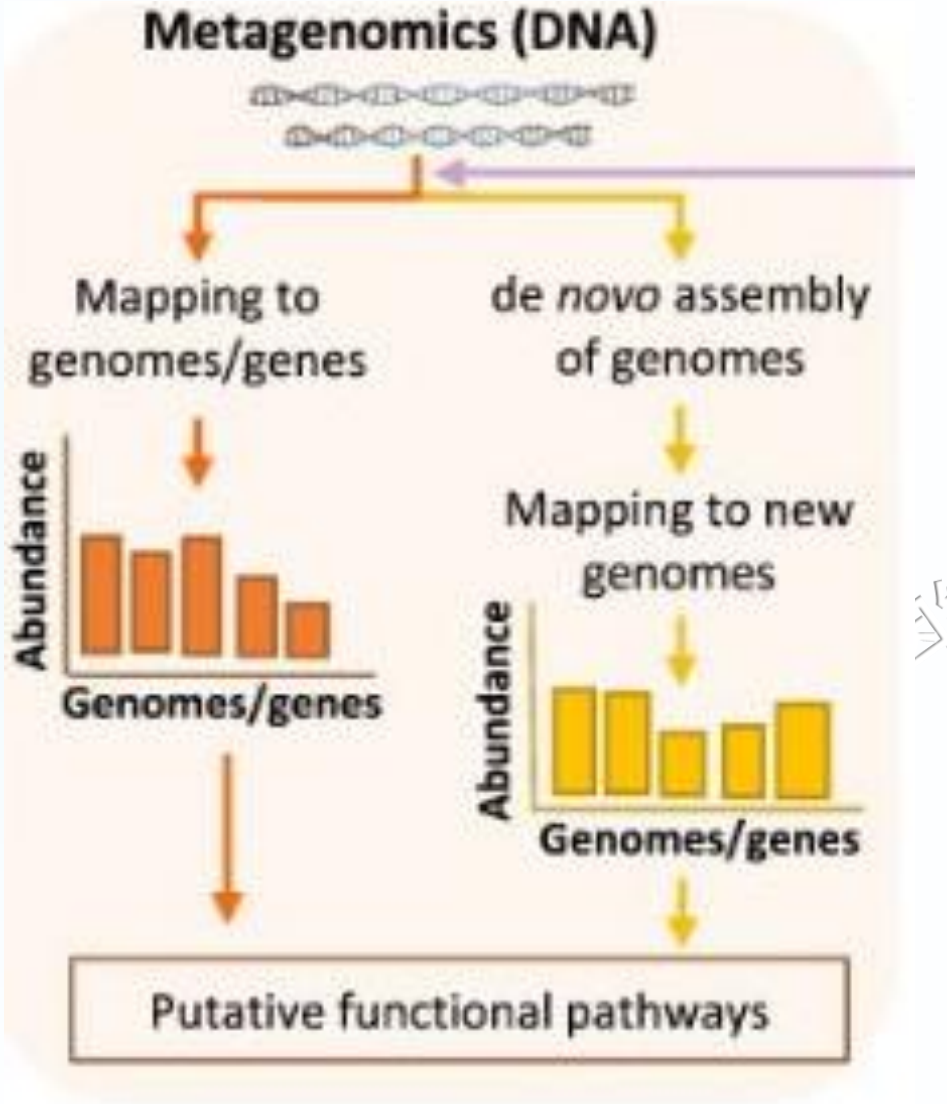
宏基因组测序技术可以回答的科学问题

回答3个科学问题：

1. 样品中有什么？
物种组成(包括宿主、细菌、真菌、病毒、原生动物等)

2. 样品中有哪些功能基因？
功能基因组成——潜在的功能

3. 组间物种和功能差异？
分组有关的物种分类(界/门/纲/目/科/属/种/株)和功能(通路/模块/同源簇/基因)



宏基因组基于读长(Reads-based)的分析流程

一. **软件安装和数据库部署**

二. KneadData质控

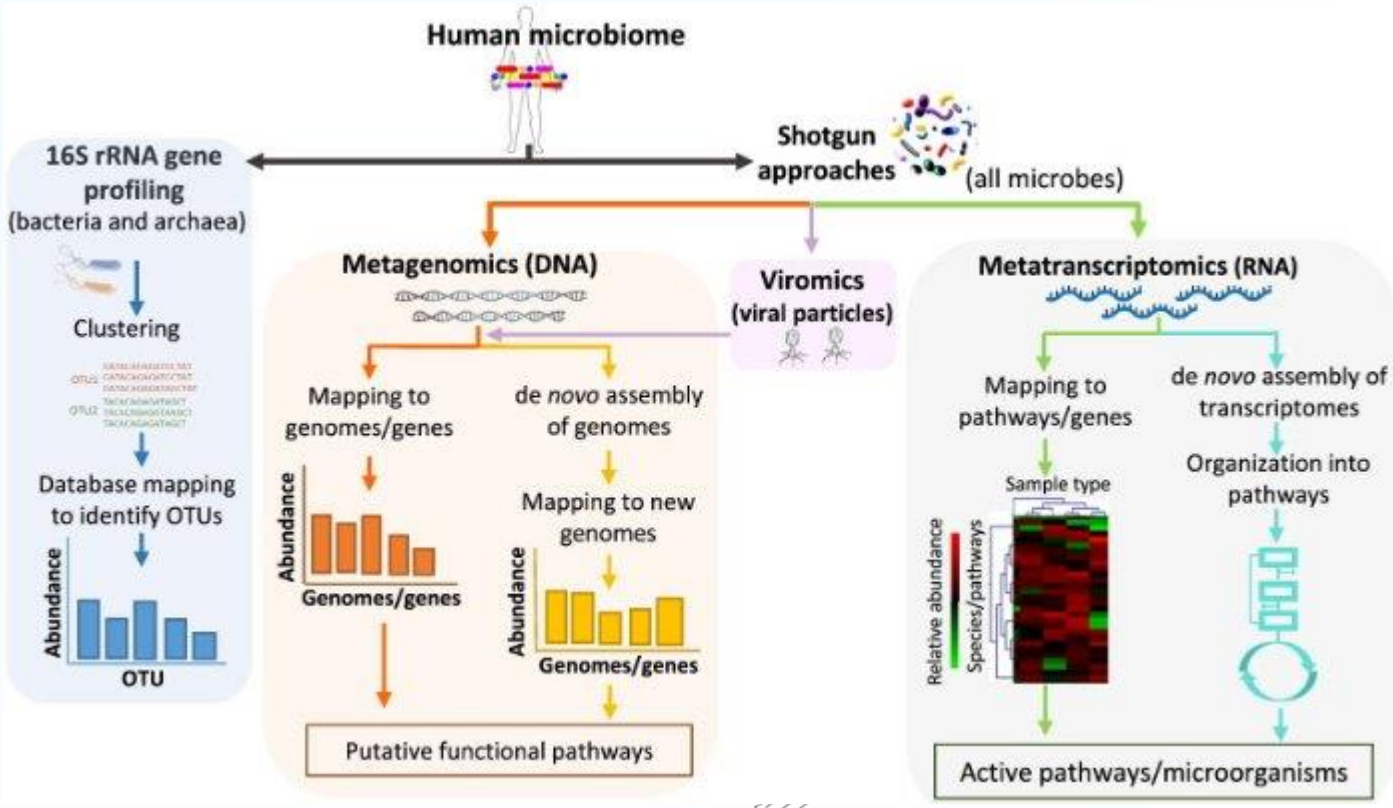
三. MetaPhlAn2物种组成

四. HUMAnN2功能组成

五. GraPhlAn可视化物种

六. LEfSe分析物种差异

七. STAMP功能组成分析



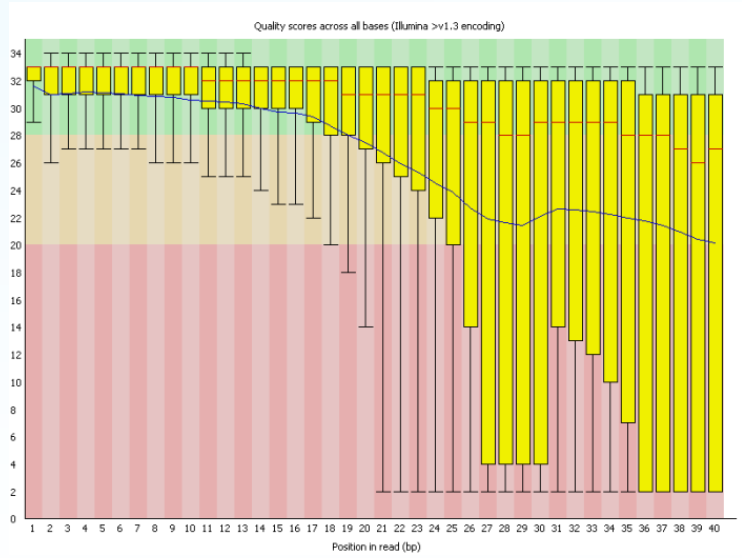
本节内容大纲

一. 软件安装和数据库部署

- Conda简介与安装
- 软件安装
- 数据库部署

二. KneadData质控

- FastQC评估和MultiQC汇总结果
- KneadData质控和去宿主
- FastQC再评估和MultiQC汇总



- Conda是(Python，R，Java，C等)软件包和环境管理系统，用于安装多个版本的软件包及其依赖关系，并在它们之间轻松切换。
- 开源软件，支持Windows、MacOS和**Linux(软件最多)**三大主流系统
- 容易安装、升级软件及依赖包；
- 方便创建、保存、加载和切换不同的环境变量(如Python2/3)
- Conda由本地软件(Anaconda/**Miniconda**)和远程软件仓库组成
- 推荐安装Miniconda
- 生物软件安装必添加Bioconda频道

https://conda.io/docs/

推荐Miniconda3

- 最流行的Python数据科学管理平台
- https://conda.io/miniconda.html 推荐下载Linux python3 64位版本

```
# 下载软件，可根据官网下载最新版本

wget -c https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh

# 安装，如管理员推荐安装目录设为conda，普通用户根据个人喜好设定或使用默认值~/miniconda3，其它选项全yes

bash Miniconda3-latest-Linux-x86_64.sh -b -f
```

详细教程见：Nature Method：Bioconda解决生物软件安装的烦恼

- **Bioconda是conda系统的生物信息软件专用频道，包括4部分：**
- 可用软件清单 http://bioconda.github.io/conda-package_index.html
- 软件布署系统，方便用户定制软件及依赖关系
- 8527个生物信息软件/包及多版本，如收录fastqc就有29个版本
- 超千人添加、修改、升级和维护软件清单
- 2017年发布于bioRxiv；2018年以通讯发表于***Nature Methods***，以后可以优雅的引用它(吃水不忘挖井人)，三年内被引600+次
- 添加频道：conda config --add channels bioconda

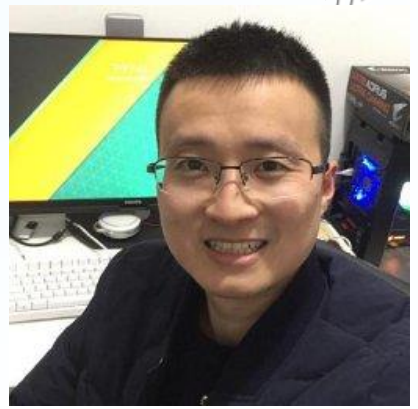
Nature Method：Bioconda解决生物软件安装的烦恼 **https://bioconda.github.io/**
Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475-476, doi:10.1038/s41592-018-0046-7 (2018).

(可选)清华/北外维护的Anaconda镜像站加速下载

```
# 添加北外镜像加速下载
site= https://mirrors.bfsu.edu.cn/anaconda/
conda config --add channels ${site}/pkgs/free/
conda config --add channels ${site}/pkgs/main/
conda config --add channels ${site}/pkgs/r/
conda config --add channels ${site}/cloud/conda-forge/
conda config --add channels ${site}/cloud/bioconda/
# 如果不可用，请手动在conda配置文件 ~/.condarc 中手动删除
```

常用生物信息、宏基因组小工具推荐

- 陈实富GitHub主页 https://github.com/OpenGene
  - fastp 0.23.2: Fastq序列质控
  - MutScan v1.14.1: 突变位置检测和可视化
  - repaq v0.3.0: Fastq序列高压缩比快速解压
  - Fastv 0.8.1：微生物检测，如SARS-CoV-2
- 沈伟GitHub主页 https://github.com/shenwei356
  - 通用工具支持Windows / Linux / MacOS的32/64位系统，支持下载或conda安装
  - seqkit 2.4：序列处理
  - csvtk v0.25.0：表格处理
  - taxonkit v0.14.1: NCBI物种信息查询和整理
  - rush v0.5.0: 任务并行管理软件



fastp: fastq数据质量评估和质控

- 主页：https://github.com/OpenGene/fastp
- 安装 conda install fastp -c bioconda
- 下载 wget http://opengene.org/fastp/fastp 添加权限 chmod a+x ./fastp
- 示例：适合单独质控或无需去宿主的环境样本，分析速度极快

```
mkdir -p temp/qc
i=C1
fastp -i seq/${i}_1.fq.gz -o temp/qc/${i}_1.fastq -I seq/${i}_2.fq.gz -O
temp/qc/${i}_2.fastq
```

- 质控前后报告见 fastp.html

seqkit: fastq数据基本统计和操作

- seqkit：序列梳理神器-统计、格式转换、长度筛选、质量值转换、翻译、反向互补、抽样、去重、滑窗、拆分等30项全能
- 安装 conda install seqkit -c bioconda
- 可选在 https://github.com/shenwei356/seqkit/releases 发布页下载
- 样本批量统计 seqkit stat seq/*.fq.gz

```
(base) yongxin@yongxin:/mnt/c/meta$ seqkit stat seq/*.fq.gz
file             format  type  num_seqs     sum_len  min_len  avg_len  max_len
seq/C1_1.fq.gz   FASTQ   DNA     75,000   7,575,000      101      101      101
seq/C1_2.fq.gz   FASTQ   DNA     75,000   7,575,000      101      101      101
seq/C2_1.fq.gz   FASTQ   DNA     75,000   7,575,000      101      101      101
seq/C2_2.fq.gz   FASTQ   DNA     75,000   7,575,000      101      101      101
```

质控软件安装

- # 质量评估软件fastqc

  conda install fastqc
  fastqc -v # FastQC v0.12.1

- # 多样品评估报告汇总multiqc

  conda install multiqc
  multiqc --version # multiqc, version 1.14

- # 质量控制流程kneaddata，安装最新/指定版解决ID问题

  conda install kneaddata
  kneaddata --version # 0.12.0
  # 如有问题，可用=指定版本
  # conda install kneaddata=0.12.0

注意记录安装软件版本！

默认安装工作环境兼容的最新版，保证可运行且功能最全

有问题时安装指定版本，确保分析结果正确；

质控相关数据库安装——人类基因组

- # 查看可用数据库

  kneaddata_database

- # 包括人类基因组human_genome bowtie2/bmtagger、转录组 、小鼠基因组、核糖体SILVA128数据库

- # 如下载人类基因组bowtie2索引至指定数据目录

  mkdir -p ~/db/kneaddata/human_genome
  kneaddata_database --download human_genome bowtie2 ~/db/kneaddata/human_genome

- 其它物种可自行下载并使用bowtie2建索引，可参考代码或下方链接教程

•宏基因组有参流程Metagenomics Tutorial (HUMAnN2)

自定义基因组构建bowtie2索引-Kneaddata去宿主

- 大多数基因组可在ensembl genome下载。此处以拟南芥为例，访问http://plants.ensembl.org/index.html，选择Arabidopsis thaliana —— Download DNA sequence (FASTA)，选择toplevel右键复制链接

```
# 新建目录、进入并下载链接
mkdir -p ${db}/kneaddata/ath && cd  ${db}/kneaddata/ath
wget -c http://ftp.ensemblgenomes.org/pub/plants/release-51/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz
gunzip Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz
# 简化文件名
mv Arabidopsis_thaliana.TAIR10.dna.toplevel.fa tair10.fa
# bowtiew建索引，输入文件，输出文件前缀，9线程2分
time bowtie2-build -f tair10.fa tair10 --threads 9 --seed 1
```

宏基因组基于读长(Reads-based)的分析流程

一. 软件安装和数据库部署

二. **KneadData质控**

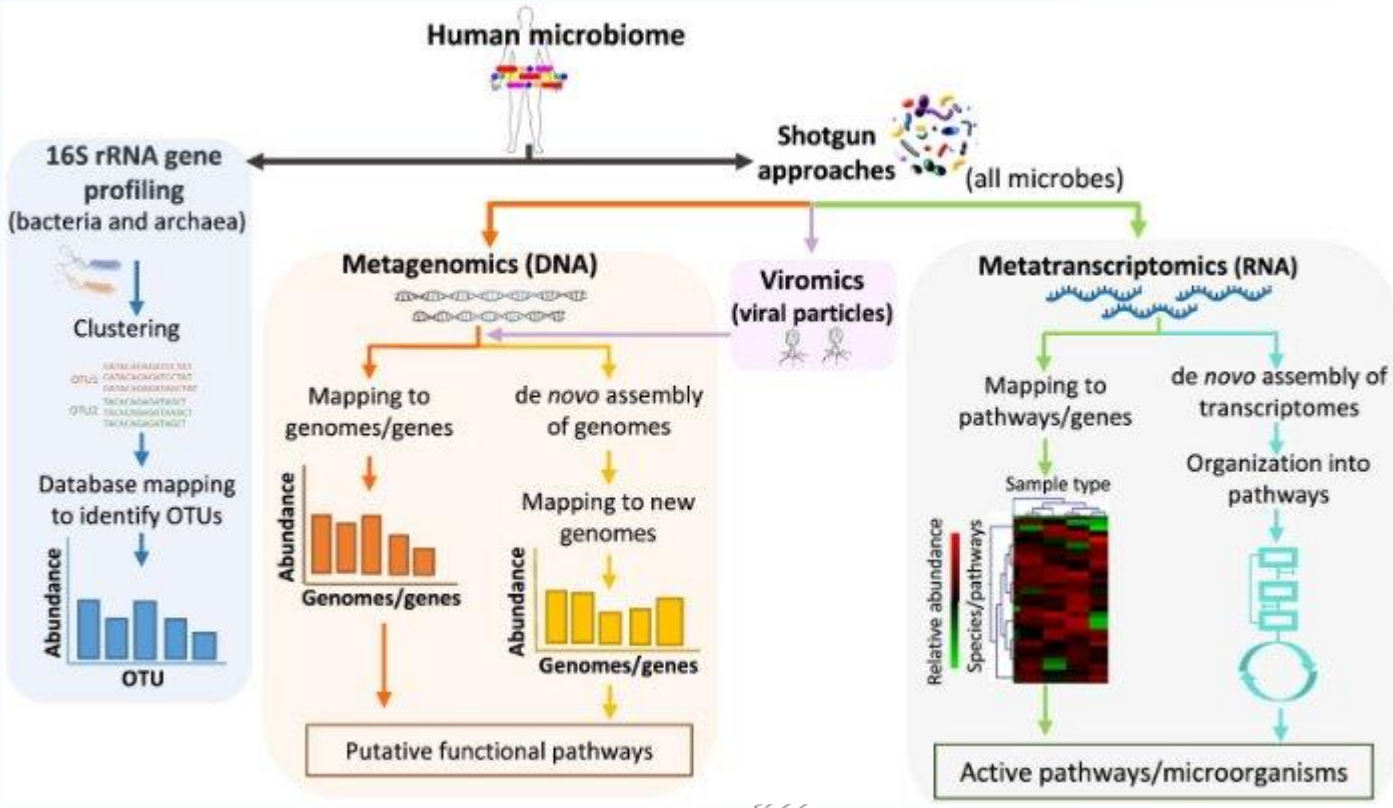
三. MetaPhlAn2物种组成

四. HUMAnN2功能组成

五. GraPhlAn可视化物种

六. LEfSe分析物种差异

七. STAMP功能组成分析



分析开始前必须设置环境变量

- # 公共数据库database位置，如db公用可能为/db，而自己下载可能为~/db
- **db=~/db**
- # Conda软件software安装目录，如db公用可能为/conda，而自己下载可能为~/miniconda3
- **soft=~/miniconda3**
- # wd为项目工作目录work directory，如meta
- **wd=~/meta**

了解宏基因组分析起始文件(上传到服务器)

- 测序数据：成对测序文件seq/*.fq.gz，通常为压缩的gz格式

C1_1.fq.gz C2_1.fq.gz
C1_2.fq.gz C2_2.fq.gz

```
@SRR3586062.883556
CTTGGGGCTGCTGAGCTTCATGCTCCCCTCCTGCCTCAAGGACAATAAGGAGATCTTCGACAAGCCTGCAGCAGCTCGCATCG
+
CCCFFFFFHHHHHIJJJJJJJIJIJJJJGIJDGIJEIIJIJJJJJJJJIJJJJIJJIJJJJJHHHFFFFECEEEDDDDD?BDD
@SRR3586062.3376311
GACGGTGTCCTCAGGACCCTTCAGTGCCTTCATGATCTGCTCAGAGGTGATGGAGTCACGGACGAGATTCGTCGTGTCAGCAC
+
@@@DDDDAFF?DF;EH+ACHIIICHDEHGIGBFE@GCGDGG?D?G@BGHG@FHCGC;CC:;8ABH>BECCBCB>;8ABCCC@A
```

- 实验设计：样本名和分组 result/metadata.txt

SampleID	Group	Replicate	Sex	Individual	GSA	CRR
C1	Cancer	1	Male	p136	CRA002355	CRR117732
C2	Cancer	2	Male	p143	CRA002355	CRR117733

序列格式详解fasta&fastq，样品命名 注意事项 实例

宏基因组数据质量

- 常用Illumina NovaSeq6000 PE150，或BGI-Seq500 PE100
- 数据质量评估——FastQC

Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010). Cited by 12235
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- 去除引物、接头和低质量序列——Trimmomatic

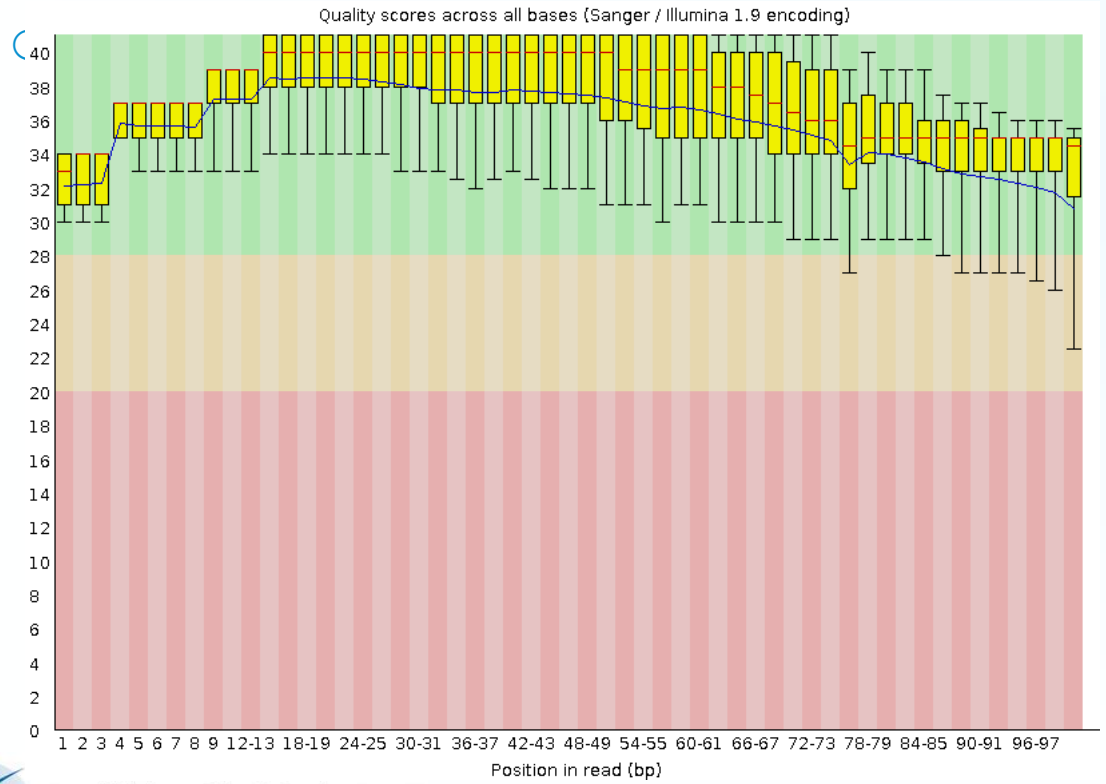
Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014). Cited by 39668

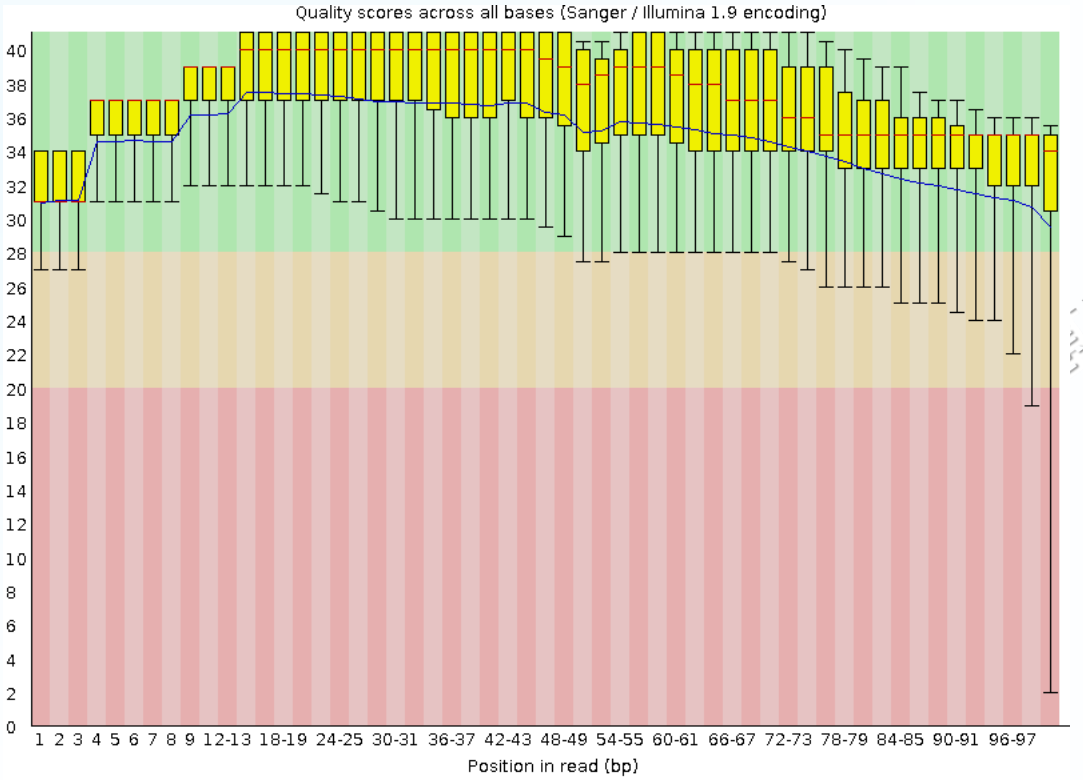
- 去除宿主——Bowtie 2比对宿主基因组；筛选非宿主序列

Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357, doi:10.1038/nmeth.1923 (2012). Cited by 38410

FastQC质量评估

- fastqc seq/*.gz -t 1 # fastqc批量，12个双端样本24个文件，设置1线程即仅允许1个文件同时处理，可根据服务器性能合理选择

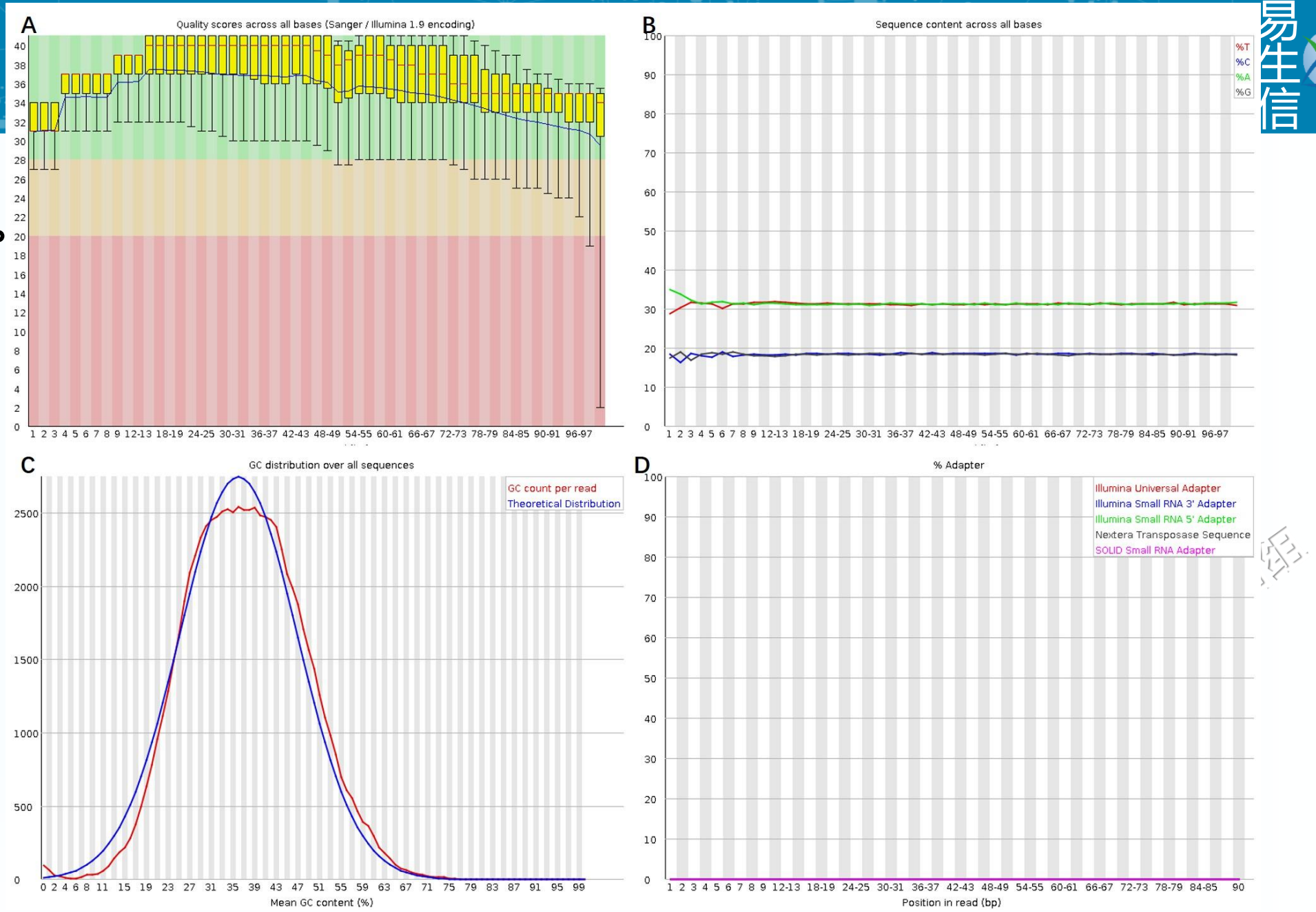




数据的质量控制软件——fastQC

FastQC质量评估报告中的重要结果。

A. 每个碱基的质量(Per base sequence quality)。
B. 序列中每个位置上碱基的含量(Per base sequence content)。
C. 所有序列的GC含量(Per sequence GC content)分布与理论值分布曲线。
D. 接头含量(Adapter Content)。
以样本C2右端序列为列，详见seq/C2_2_fastqc.html



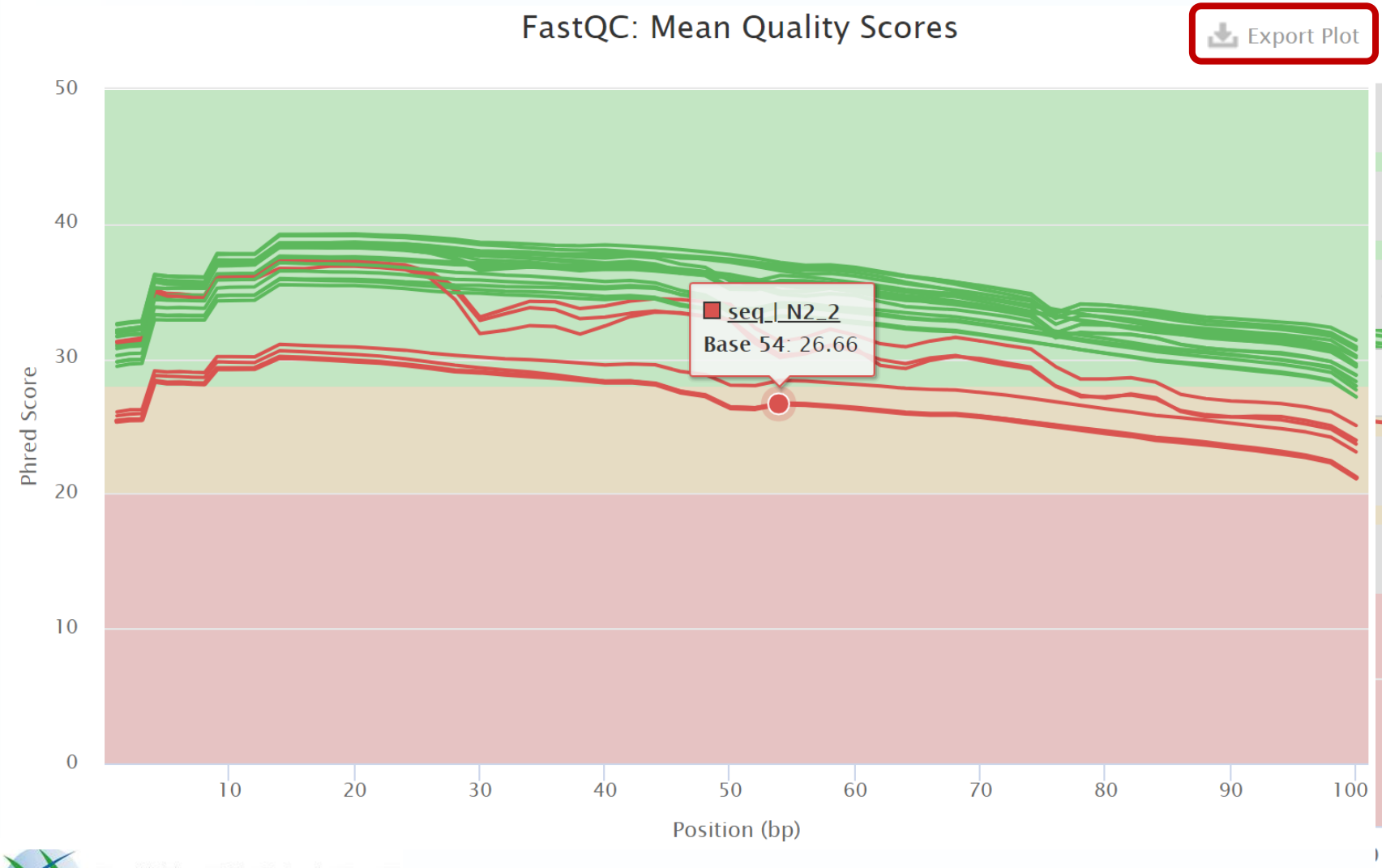
MultiQC多样本汇总比较

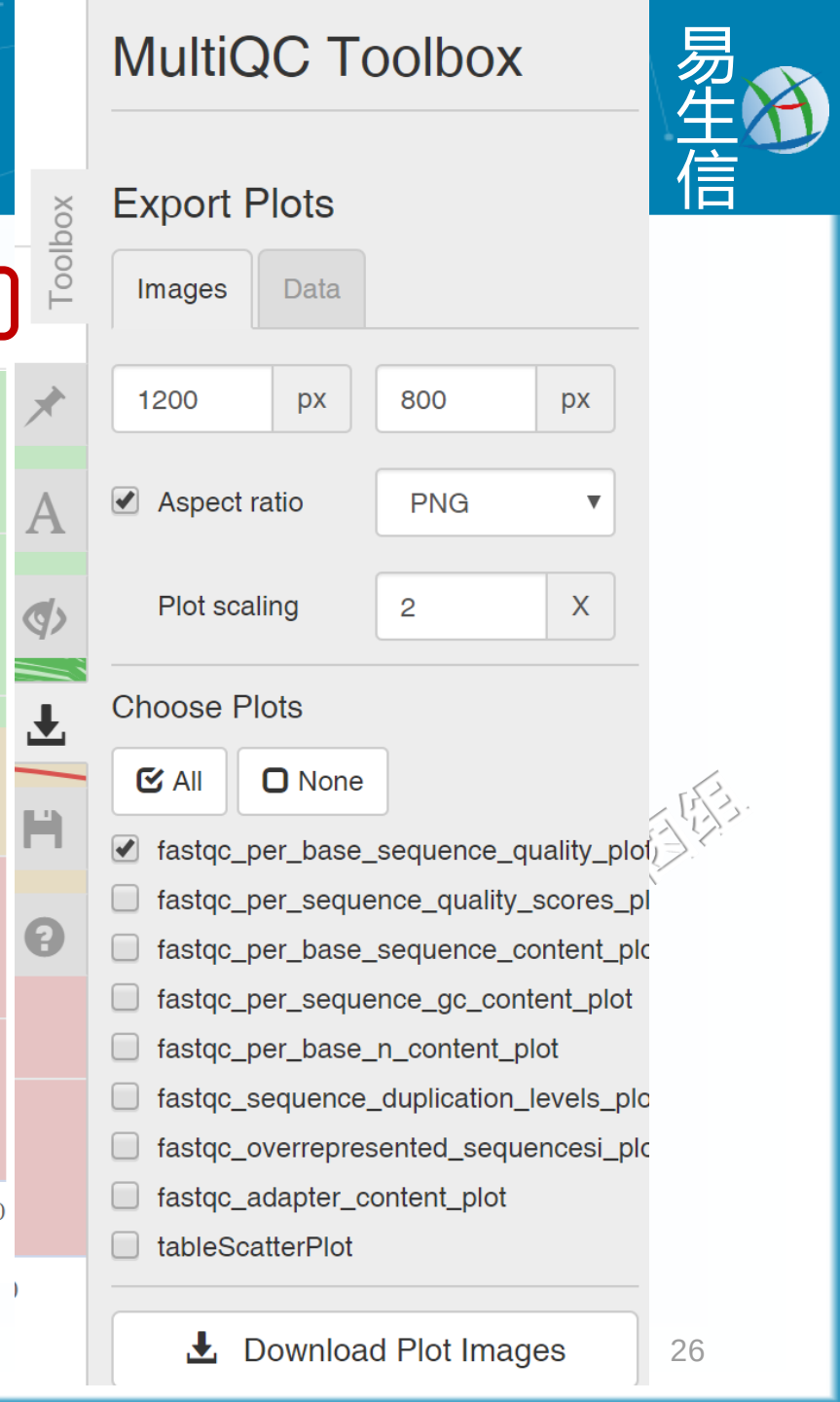
- # 生成多样品报告比较
- multiqc -d seq/ -o result/qc
- # 查看右侧result/qc目录中multiqc_report.html，可交互式报告

Sample Name ▲	% Dups	% GC	M Seqs
seq | C1_1	0.1%	37%	0.1
seq | N1_1	1.6%	40%	0.1
seq | C1_2	0.2%	37%	0.1
seq | N1_2	3.4%	40%	0.1

Philip Ewels, Måns Magnusson, Sverker Lundin & Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. ***Bioinformatics*** 32, 3047-3048, doi:10.1093/bioinformatics/btw354 (**2016**). Cited by 689

图片和数据导出





MultiQC质量评估汇总报告中的重要结果。

A. 综合统计(General Statistics)。

B. 平均质量值(Mean Quality Scores)。

C. 每个碱基的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。

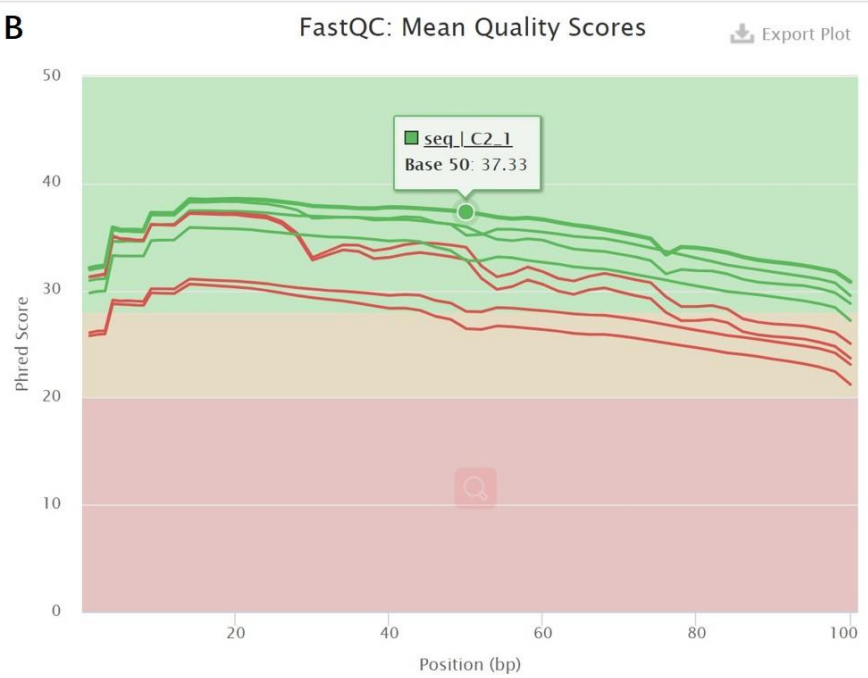
A

General Statistics

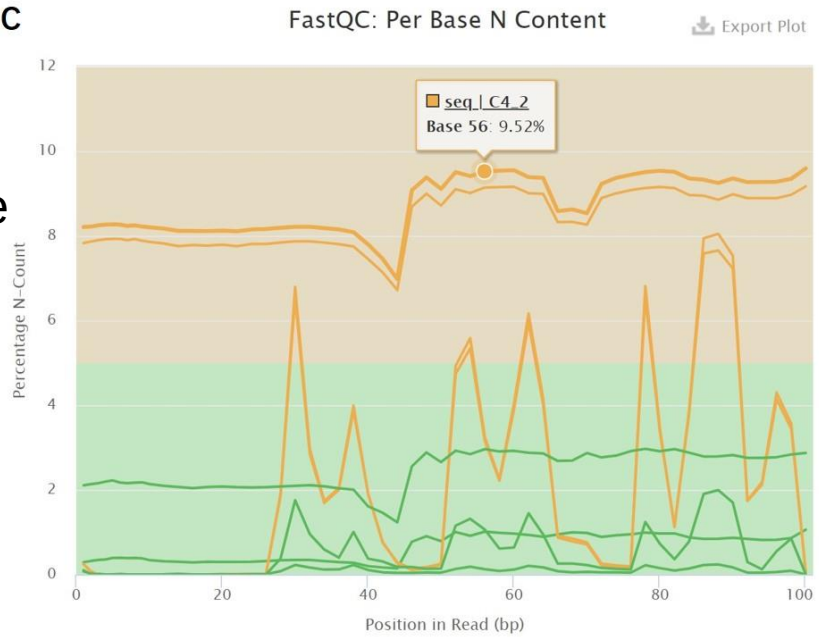
Copy table | Configure Columns | Plot | Showing 8/8 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
seq | C2_1	0.1%	37%	0.1
seq | C2_2	0.2%	37%	0.1
seq | C3_1	10.3%	46%	0.1
seq | C3_2	4.6%	45%	0.1
seq | C4_1	6.1%	38%	0.1
seq | C4_2	12.8%	38%	0.1
seq | C5_1	0.6%	43%	0.1
seq | C5_2	1.7%	43%	0.1

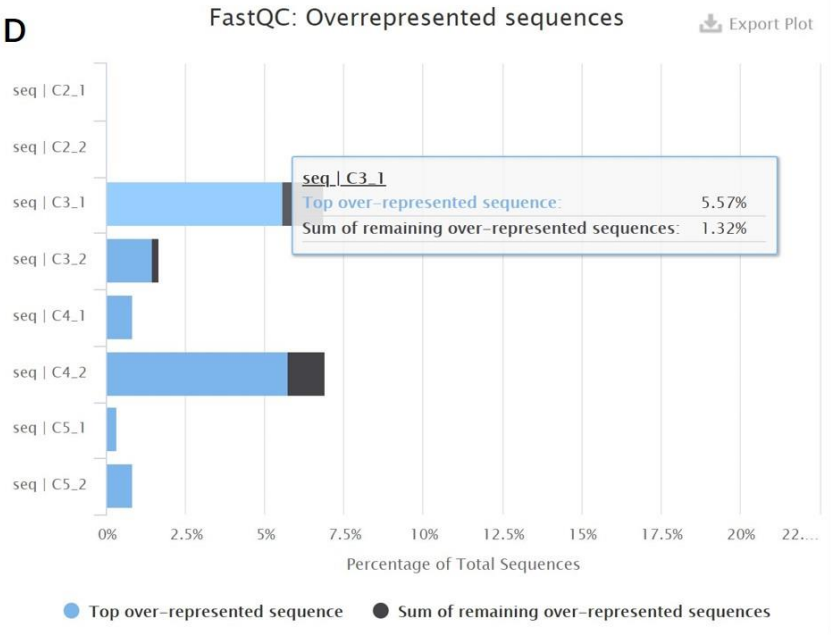
B



C



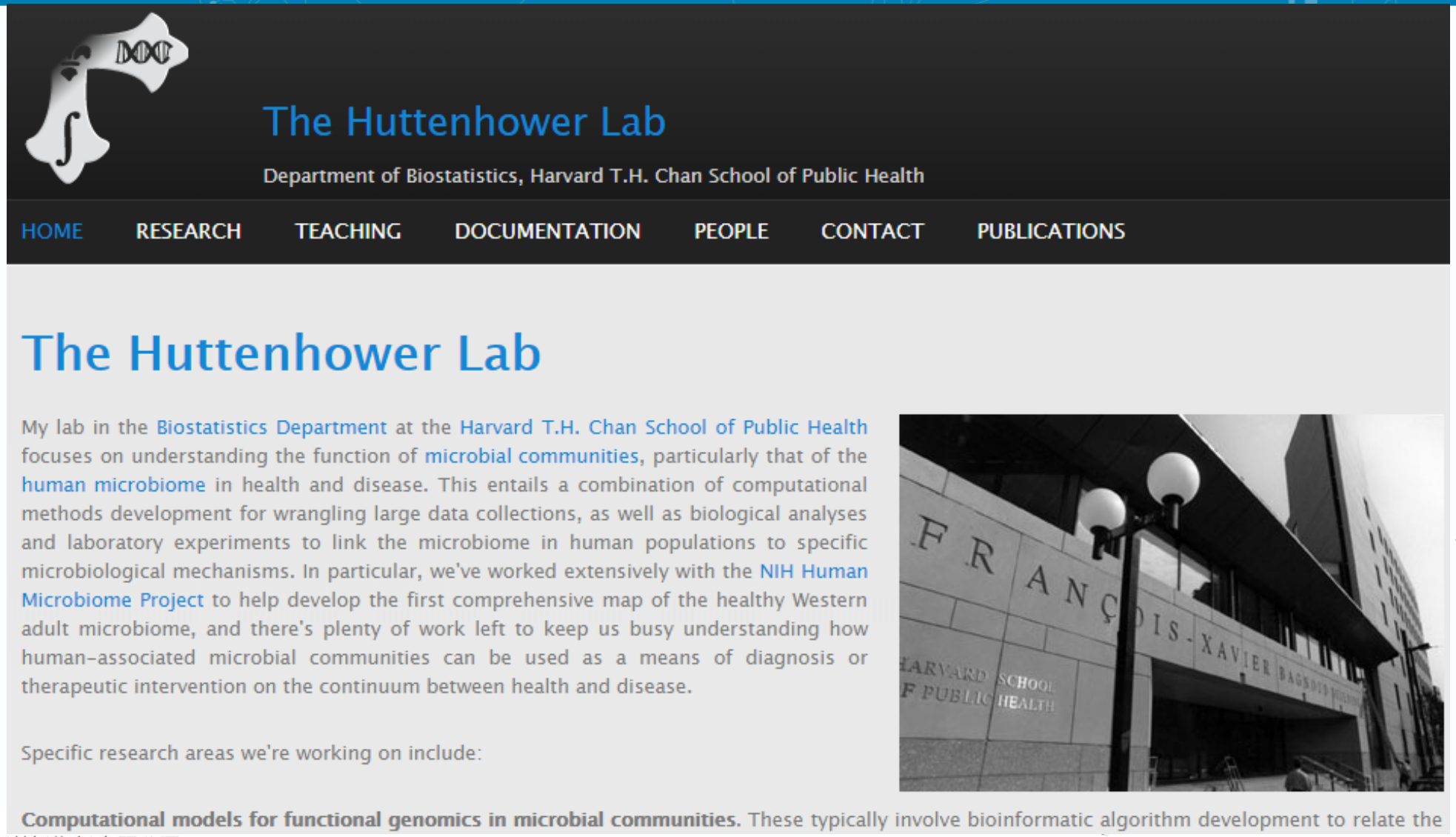
D



KneadData — 宏基因组质控和去宿主流程

- 质控包括去除低质量和接头、比对宿主基因组、去除宿主序列三步
- 依赖Trimmomatic、Bowtie 2、Samtools、Python等
- 由Huttenhower实验室提供了此步的解决方案：KneadData
  http://huttenhower.sph.harvard.edu/kneaddata
- 文章还在投稿中 (TBD)，已经被引用近500次
- 流程采用Python编写，支持pip和Conda安装
- 预构建了人类、小鼠数据库；可自定义Bowtie 2索引数据库

Curtis Huttenhower lab



实验室主页：http://huttenhower.sph.harvard.edu/
Google学术主页: https://scholar.google.co.jp/citations?hl=zh-CN&user=yFncM6AAAAAJ

Curtis Huttenhower Google学术主页

Curtis Huttenhower

关注

Department of Biostatistics, Harvard School of Public Health
在 hsph.harvard.edu 的电子邮件经过验证

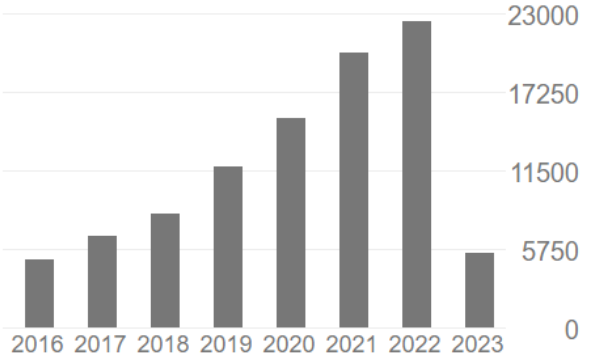
computational metagenomics　human microbiome　biological data mining

10篇Nature专题报导人类微生物组计划2(iHMP)成果及展望
Nature：iHMP之“微生物组与炎症性肠病”

标题	引用次数	年份
Metagenomic biomarker discovery and explanation N Segata, J Izard, L Waldron, D Gevers, L Miropolsky, WS Garrett, ... Genome biology 12, 1-18	9674	2011
Structure, function and diversity of the healthy human microbiome nature 486 (7402), 207-214	8947	2012
Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 E Bolyen, JR Rideout, MR Dillon, NA Bokulich, CC Abnet, GA Al-Ghalith, ... Nature biotechnology 37 (8), 852-857	8640	2019
Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences MGI Langille, J Zaneveld, JG Caporaso, D McDonald, D Knights, ... Nature biotechnology 31 (9), 814-821	7427	2013
The treatment-naive microbiome in new-onset Crohn's disease D Gevers, S Kugathasan, LA Denson, Y Vázquez-Baeza, W Van Treuren, ... Cell host & microbe 15 (3), 382-392	2792	2014

引用次数　查看全部

	总计	2018 年至今
引用	106293	83979
h 指数	123	110
i10 指数	242	223



开放获取的出版物数量　查看全部

10 篇文章　226 篇文章

无法查看的文章　可查看的文章

根据资助方的强制性开放获取政策

KneadData——宏基因组质控流程依赖关系

- http://huttenhower.sph.harvard.edu/kneaddata
- **Trimmomatic** (version >= 0.33) (automatically installed)
- **Bowtie2** (version >= 2.2) (automatically installed)
- Python (version >= 2.7)
- Java Runtime Environment
- TRF (optional)
- FastQC (optional)
- SAMTools (only required if input file is in BAM format)

以C1单样品质控为例(正对照确保软件可用)

- **-i输入文件，-o输出目录**，-v输出计算过程，**-t线程数，** **--trimmomatic位置和参数，** --bowtie2-options 参数， **-db 宿主基因组索引位置**

```
time kneaddata -i seq/C1_1.fq.gz -i seq/C1_2.fq.gz \
  -o temp/qc -v -t 3 --remove-intermediate-output \
  --trimmomatic ~/miniconda3/envs/kneaddata/share/trimmomatic/ --trimmomatic-options
'ILLUMINACLIP:~/miniconda3/envs/kneaddata/share/trimmomatic/adapters/TruSeq2-PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \
  --bowtie2-options '--very-sensitive --dovetail' -db
~/db/kneaddata/human_genome/hg37dec_v0.1
```

多个样品如何批量分析，并管理好资源分配呢?

并行管理软件 rush / parallel

- 现实中是有一大堆样品，for可以单个或全部提交任务效率都很低，如何让服务器性能允许下并行加速分析，并有序管理队伍呢?
- 国人开发了跨平台的并行管理工具rush，官网下载 或 conda安装

  conda install rush

  官网：https://github.com/shenwei356/rush
- (可选)Parallel是Perl语言编写，可提供并行任务数量管理的功能，保证任务高效有这序完成，作者要求引用，如不想引用也可付10000欧元购买。可以直接在Ubuntu仓库中安装或conda安装

  sudo apt install parallel

  conda install parallel

Tange, O. *Gnu parallel 2018*. (Lulu. com, 2018).

方法1. rush并行管理质量控制(质控)实例

- 样本名列表从命令行管道传入，-j 2控制2个任务并行，红色为需要修改的部分

```
tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \
  "kneaddata -i seq/{1}_1.fq.gz -i seq/{1}_2.fq.gz \
  -o temp/qc -v -t 3 --remove-intermediate-output \
  --trimmomatic ~/miniconda3/envs/kneaddata/share/trimmomatic/ \
  --trimmomatic-options 'ILLUMINACLIP:
~/miniconda3/envs/kneaddata/share/trimmomatic/adapters/TruSeq2-
PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \
  --reorder --bowtie2-options '--very-sensitive --dovetail' \
  -db ~/db/kneaddata/human_genome/hg37dec_v0.1"
```

(备选)方法2. parallel并行

- 示例：对所有样品进行质控，同时保持最多3个样本在运行。
- -j为任务数，--xapply是对两个参数按顺序使用而非组合方式

```
parallel -j 3 --xapply \
  "kneaddata -i seq/{1}_1.fq.gz -i seq/{1}_2.fq.gz\
  -o temp/qc -v -t 3 --remove-intermediate-output \
  --trimmomatic ~/miniconda3/envs/kneaddata/share/trimmomatic/ --trimmomatic-options
'ILLUMINACLIP ~/miniconda3/envs/kneaddata/share/trimmomatic/adapters/TruSeq2-
PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \
  --bowtie2-options '--very-sensitive --dovetail' -db ~db/kneaddata/human_genome/hg37dec_v0.1"
\
  ::: `tail -n+2 result/metadata.txt|cut -f1`
```

质控去宿主 结果文件简化统一(与质控一致)

- awk的system命令批处理系统命令，mv实现移动即改名
- # 移动实现简化名，如C1_1_kneaddata_paired_1.fastq为C1_1.fastq

```
awk '{system("mv `pwd`/temp/qc/"$1"_1_kneaddata_paired_1.fastq
temp/qc/"$1"_1.fastq")}' <(tail -n+2 result/metadata.txt)
# 右端链接为新简化名
awk '{system("mv `pwd`/temp/qc/"$1"_1_kneaddata_paired_2.fastq
temp/qc/"$1"_2.fastq")}' <(tail -n+2 result/metadata.txt)
# 检查结果
ls -l temp/qc/
```

- for循环仅能使用单个变量，awk命令适合基于metadata多个列变量的文件下载、重命名、链接等批量操作

质控结果汇总表

```
# 合并所有样本统计结果为表
kneaddata_read_count_table --input temp/qc -output temp/kneaddata.txt
# 筛选重要的列，并查看结果
cut -f 1,2,4,12,13 temp/kneaddata.txt | sed 's/_1_kneaddata//' > result/qc/sum.txt
cat result/qc/sum.txt
```

Sample	raw pair1	raw pair2	trimmed p	trimmed p	decontam	decontam	final pair1	final pair2	final orpha	final orpha
C1	75000	75000	65243	65243	64809	64809	64809	64809	670	6042
C2	75000	75000	47971	47971	30944	30944	30944	30944	1210	7632
C3	75000	75000	49504	49504	28643	28643	28643	28643	950	5469
C4	75000	75000	60685	60685	57149	57149	57149	57149	848	6631
C5	75000	75000	62110	62110	61928	61928	61928	61928	977	8705
C6	75000	75000	65249	65249	65211	65211	65211	65211	727	6349
N1	75000	75000	60059	60059	53662	53662	53662	53662	753	6276
N2	75000	75000	46195	46195	31873	31873	31873	31873	1050	7098
N3	75000	75000								
N4	75000	75000								
N5	75000	75000								
N6	75000	75000								

```
Sample    raw pair1    trimmed pair1    final pair1    final pair2
------    ---------    -------------    -----------    -----------
C1        75000.0      65316.0          64276.0        64276.0
C2        75000.0      48082.0          29293.0        29293.0
```

csvtk -t pretty result/qc/sum.txt

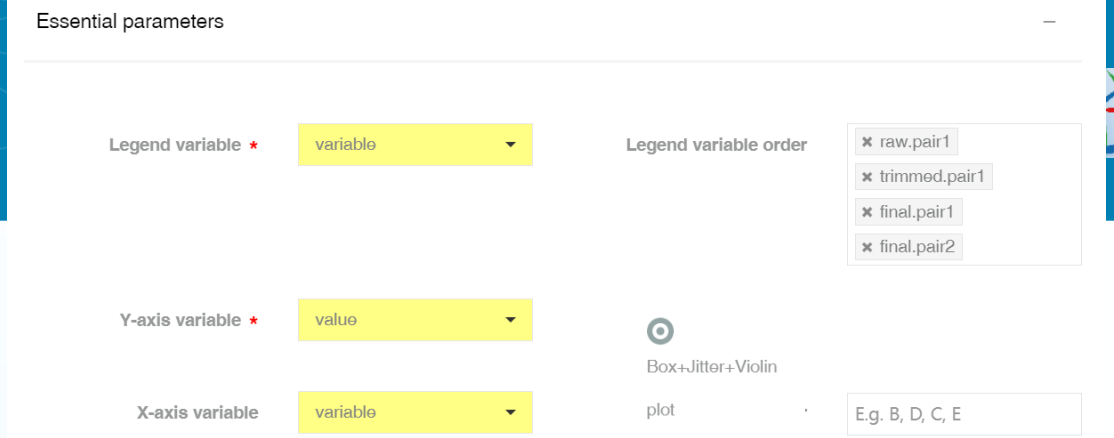
质控结果统计和可视化

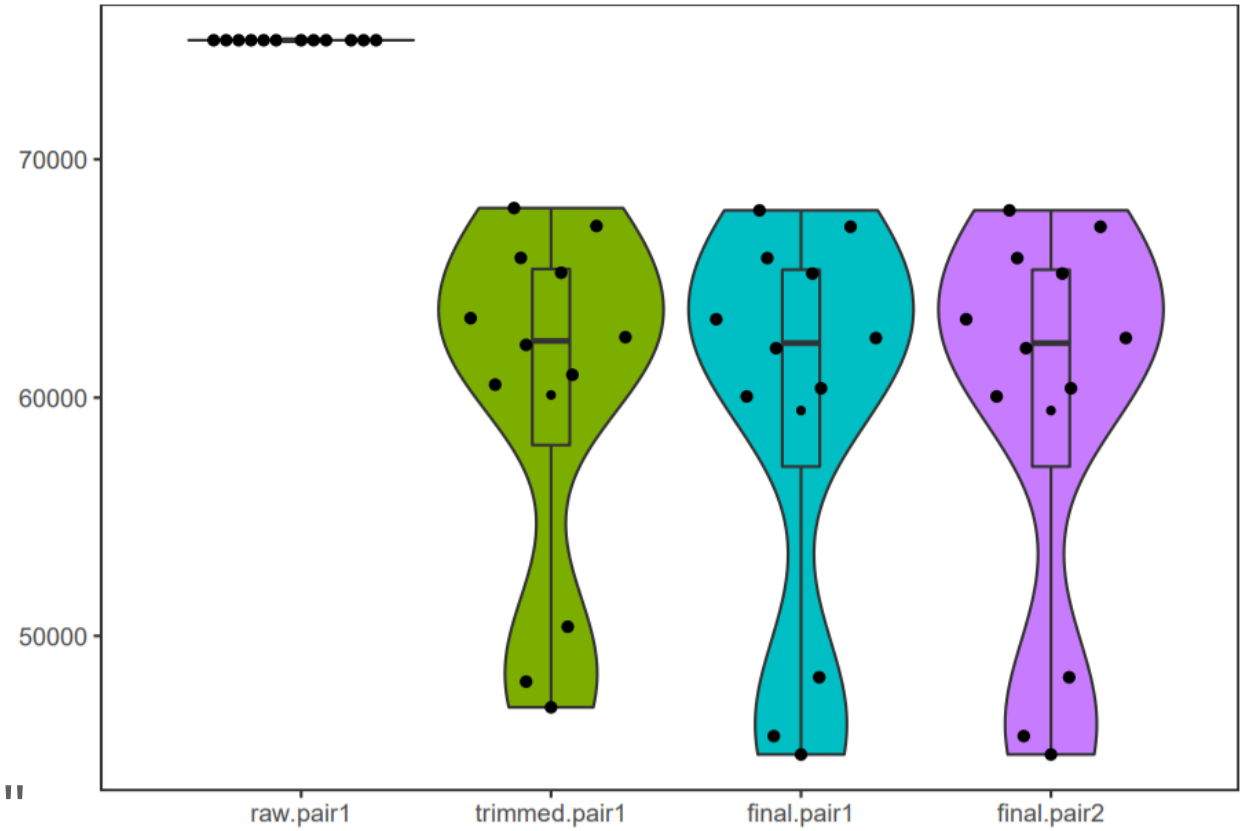
- # 用R代码统计下质控结果

```
Rscript -e
"data=read.table('result/qc/sum.txt',
header=T, row.names=1, sep='\t');
summary(data)"
```

- # R转换宽表格为长表格

```
Rscript -e "library(reshape2);
data=read.table('result/qc/sum.txt',
header=T,row.names=1, sep='\t');
write.table(melt(data),
file='result/qc/sum_long.txt',sep='\t',
quote=F, col.names=T, row.names=F)"
```





https://www.bic.ac.cn/ImageGP/index.php/Home/Index/Boxplot.html

1.4 质控后质量再评估
trimmomatic + bowtie2 + fastqc 三个软件报告汇总

易生信

```
fastqc temp/qc/*_1_kneaddata_paired_* -t 6
multiqc -d temp/qc/ -o result/qc/ # 结果为multiqc_report_1.html
```

Sample Name	% Aligned	% Dropped	% Dups	% GC	M Seqs
decompressed_dclWWc_N2_1		25.7%			
decompressed_tx4J0T_C3_1		22.3%			
decompressed_uwtl85_C2_1		21.6%			

质控步骤，扔掉低质量的比例

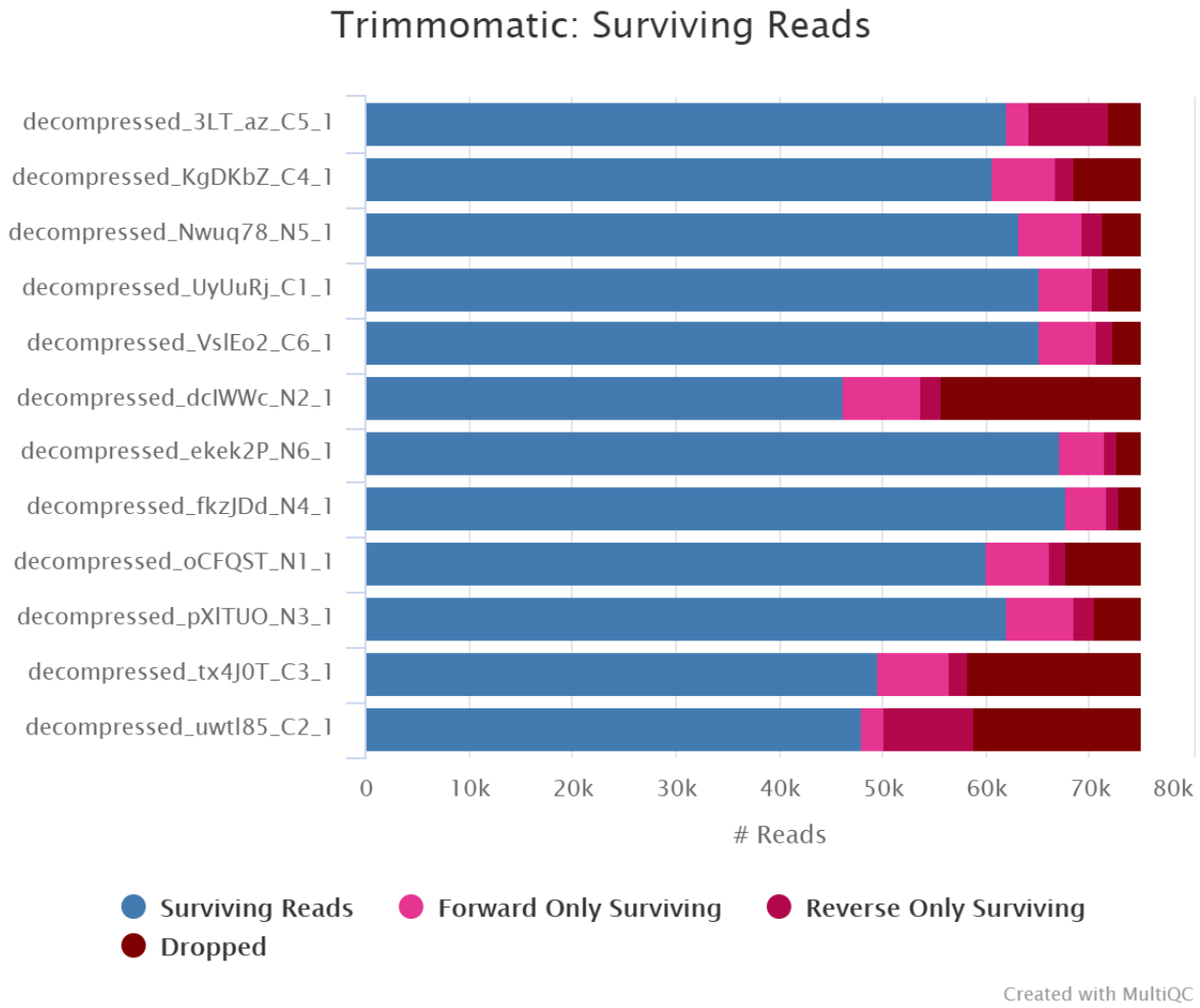
C2_1_kneaddata.trimmed.single.1	4.5%				
C3_1_kneaddata.trimmed.single.1	4.2%				
N2_1_kneaddata.trimmed.single.1	3.7%				

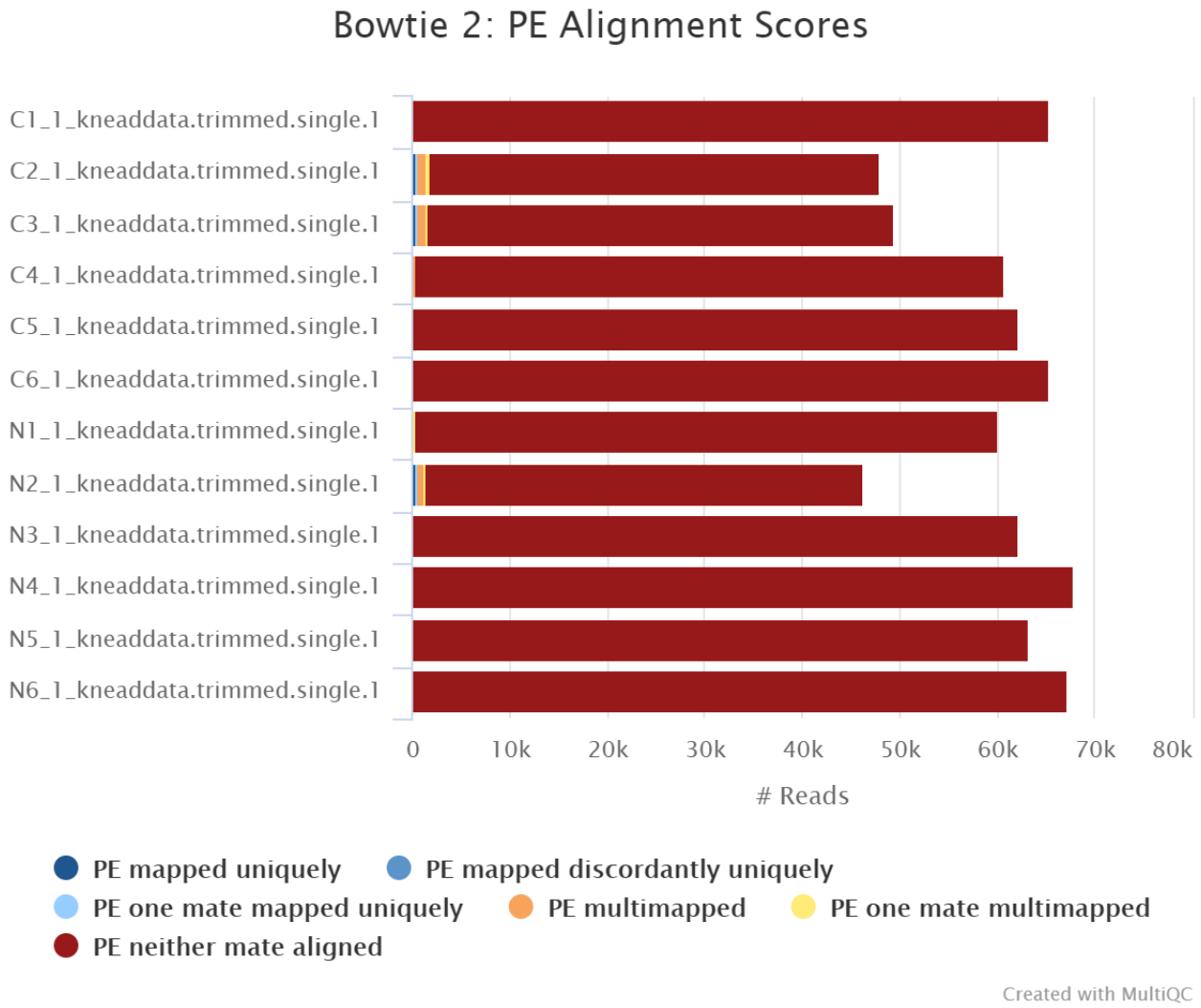
去宿主步骤，宿主含量

temp | qc | N5_1_kneaddata_paired_1			0.1%	45%	0.1
temp | qc | C2_1_kneaddata_paired_1			3.9%	44%	0.0
temp | qc | C2_1_kneaddata_paired_2			3.9%	44%	0.0

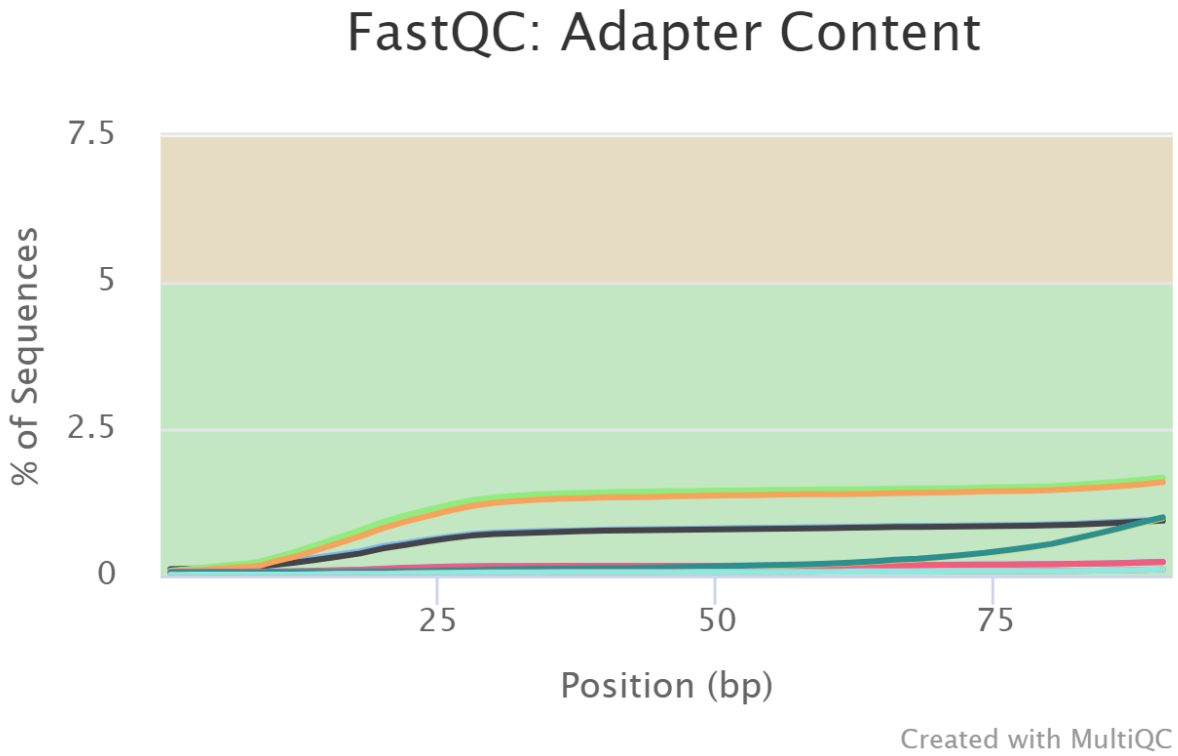
质量评估步骤，基本信息

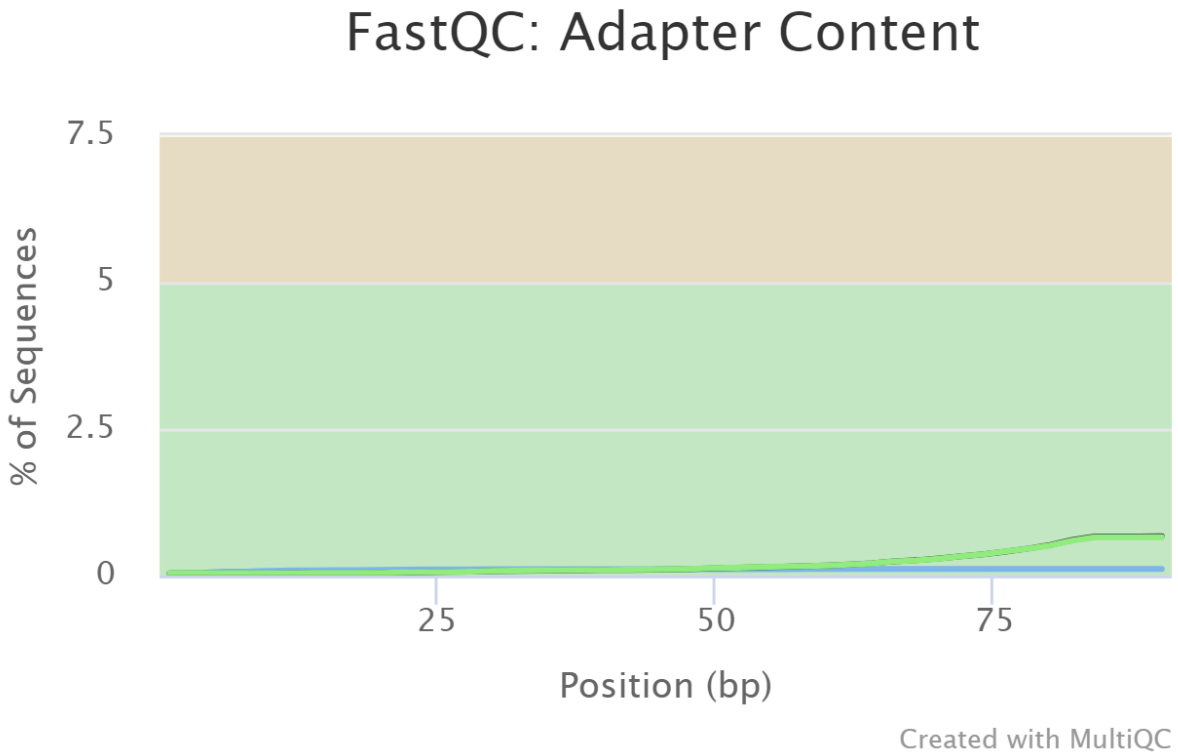
Trimmomatic质控+Bowtie2比对宿主柱状图展示





接头含量 FastQC: Adapter Content





质控前 vs 质控后

总结

- Conda是软件安装和管理神器，Bioconda频道是生物学家的福音，8千多个生信软件及数十万个版本满足你各种需求，记得引用它；
- 很多软件还依赖数据库需要手动下载，如人类基因组用于去宿主；
- 哈佛大学Huttenhover组编写的质控流程KneadData，整合质控需要的Trimmomatic，Bowtie 2等软件和宿主基因组数据库，解决软件数据库选择和安装、流程脚本、参数选择等众多烦恼；
- MultiQC用于质控前后的评估和汇总，包括FastQC、Trimmomatic和Bowtie 2的汇总、可视化，方便阅读、比较和图表导出；
- 多任务管理专家 rush，备选parallel (Perl编写、依赖包容易报错)

参考资源

- 宏基因组公众号文章目录　　生信宝典公众号文章目录
- 科学出版社《微生物组数据分析》——50+篇
- Bio-protocol《微生物组实验手册》——153篇
- Protein Cell：扩增子和宏基因组数据分析实用指南
- CMJ：人类微生物组研究设计、样本采集和生物信息分析指南
- 加拿大生信网 https://bioinformatics.ca/ 宏基因组课程中文版
- 美国高通量开源课程 https://github.com/ngs-docs
- Curtis Huttenhower http://huttenhower.sph.harvard.edu/
- Nicola Segata http://segatalab.cibio.unitn.it/

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识