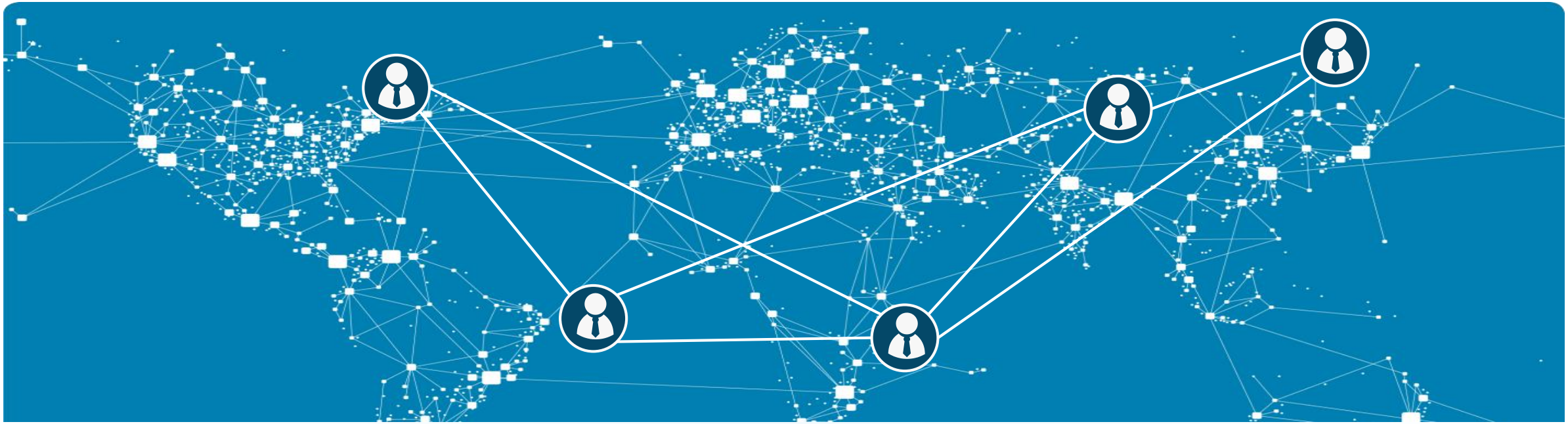
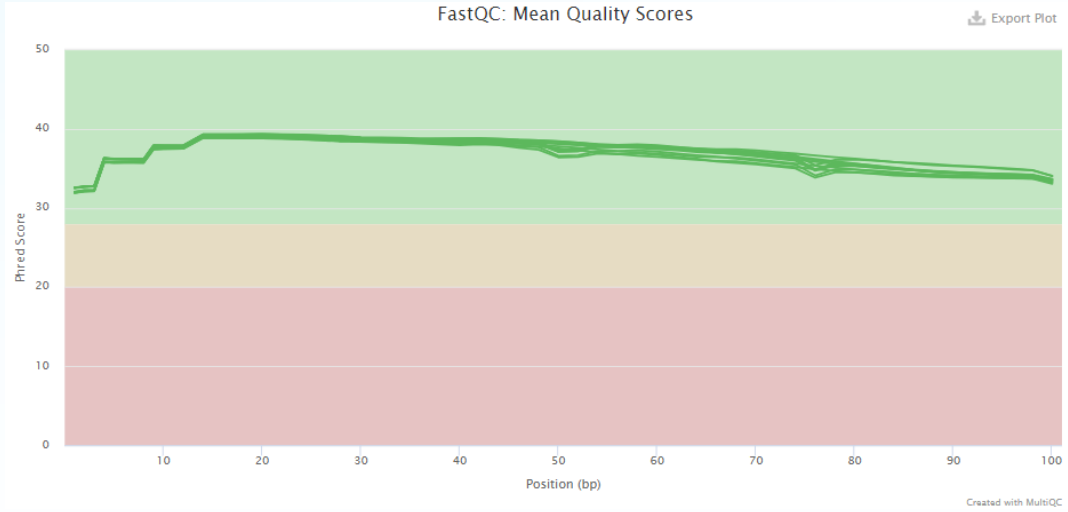


22质控和去宿主

易生信

2019年6月22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
+
DDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:2025:2183 1:N:0:CACTCAAT
ATATCGCGAGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGCG
+
DDDD@E@HIGHIIIHHFHIIHIIIIFHHIIIHHGIHIIHIIICHDEHHIIIIHGH
@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

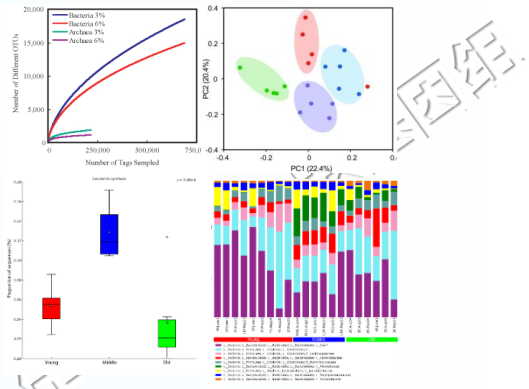


图: $10^{1\sim3}$个点和统计信息

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

MetaPhlAn2

HUMAnN2

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

	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

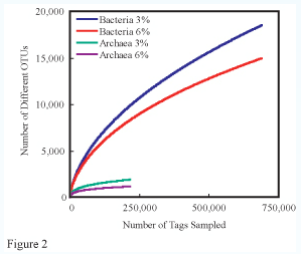
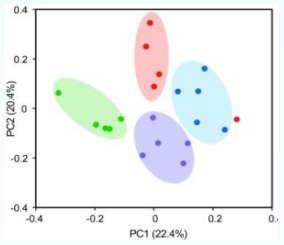
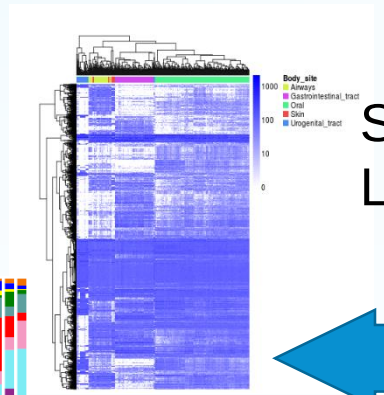
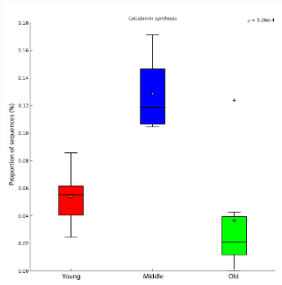


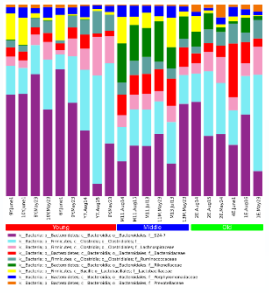
Figure 2





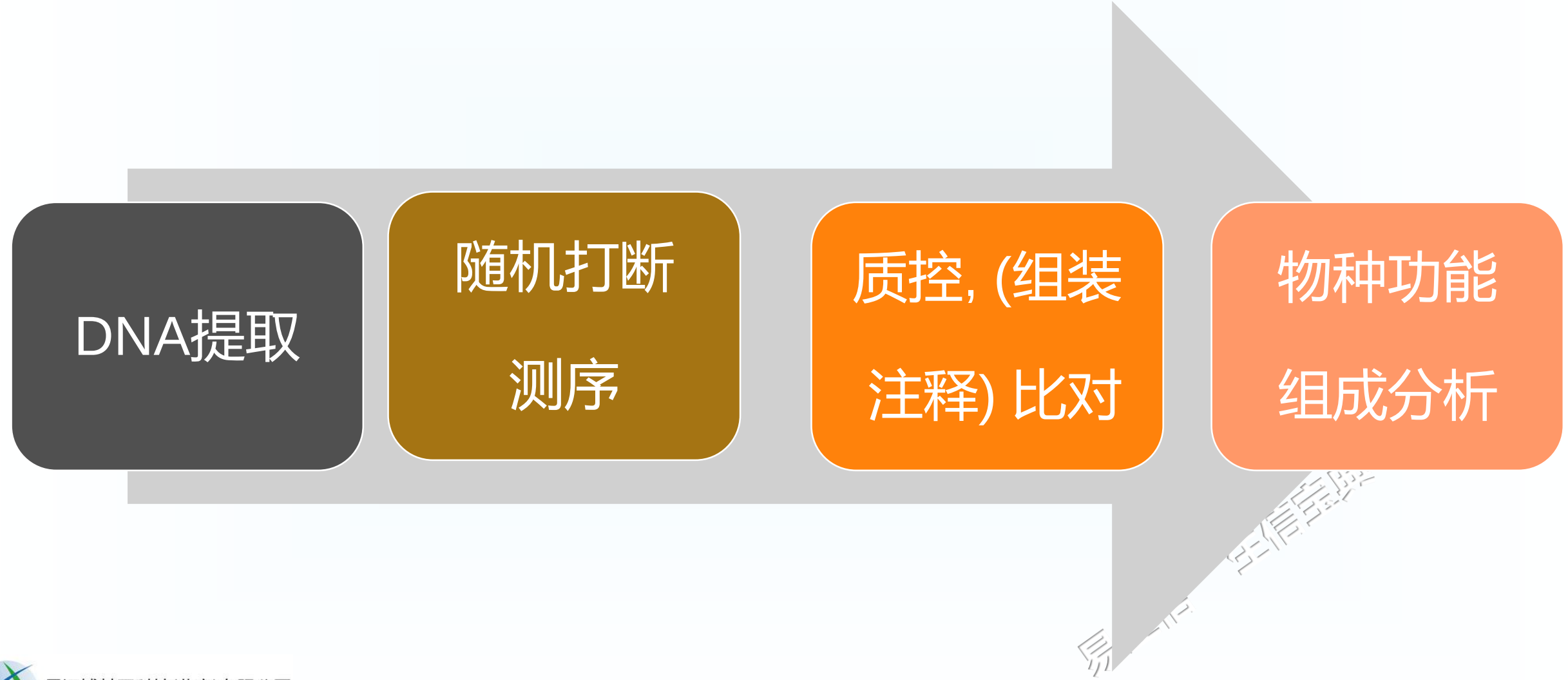
STAMP / LEfSe / R





熟记此图，胸中有丘壑

宏基因组实验分析流程



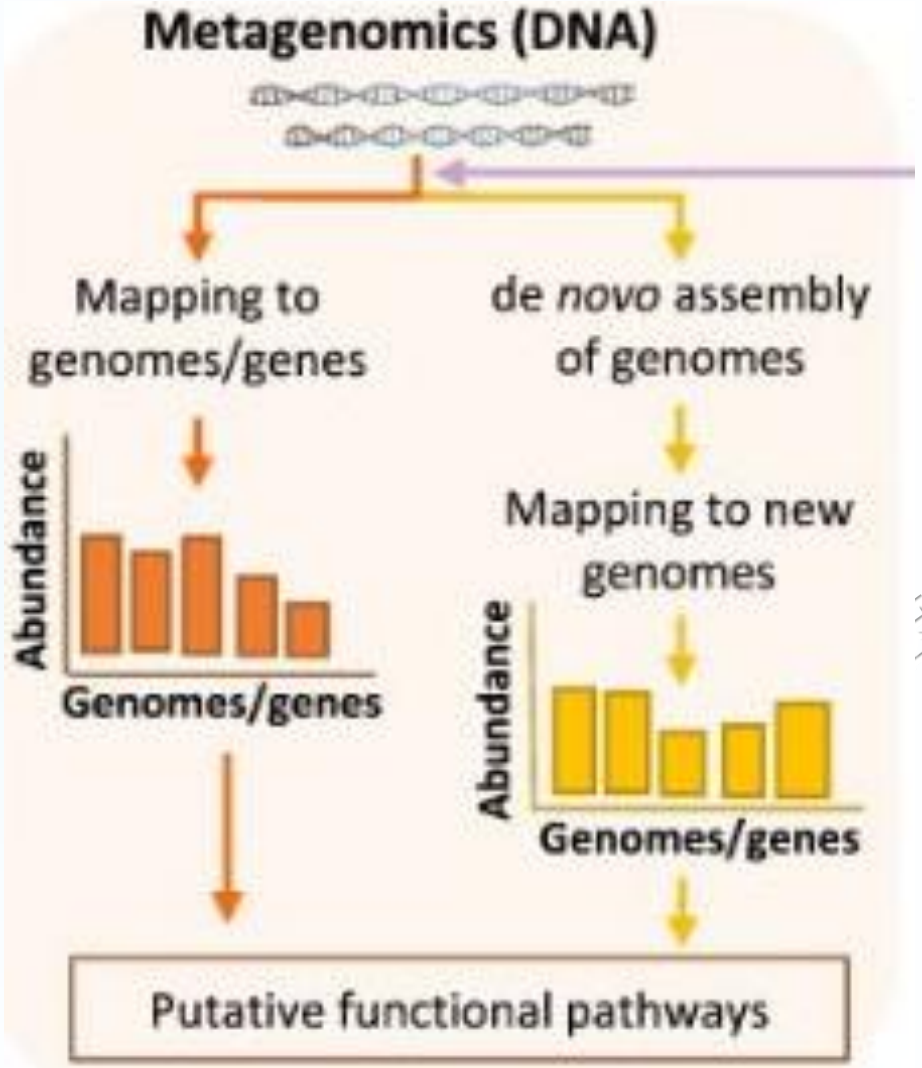
宏基因组技术可以回答的科学问题

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)



宏基因组有参(Reference-based)流程

一. **软件安装和数据库部署**

二. KneadData质控

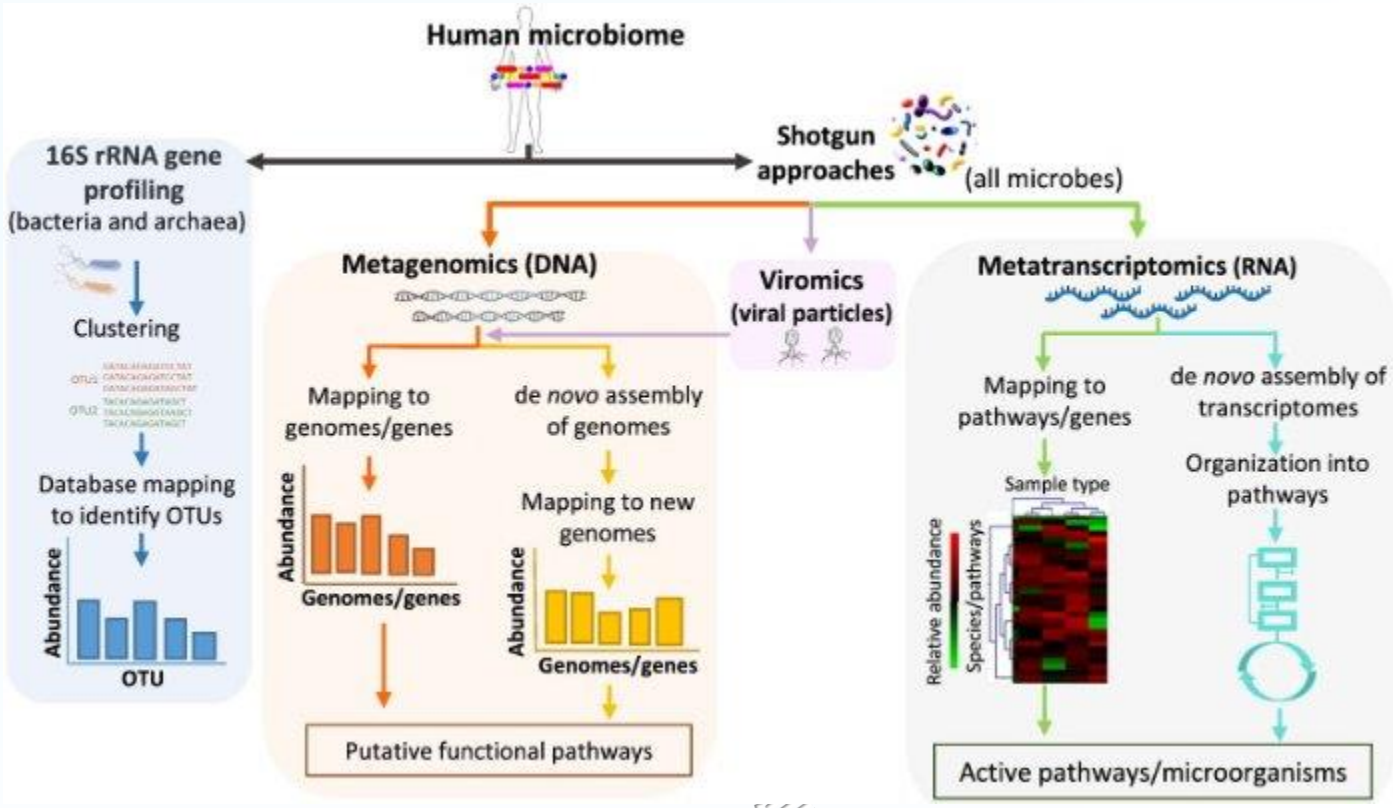
三. MetaPhlAn2物种组成

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



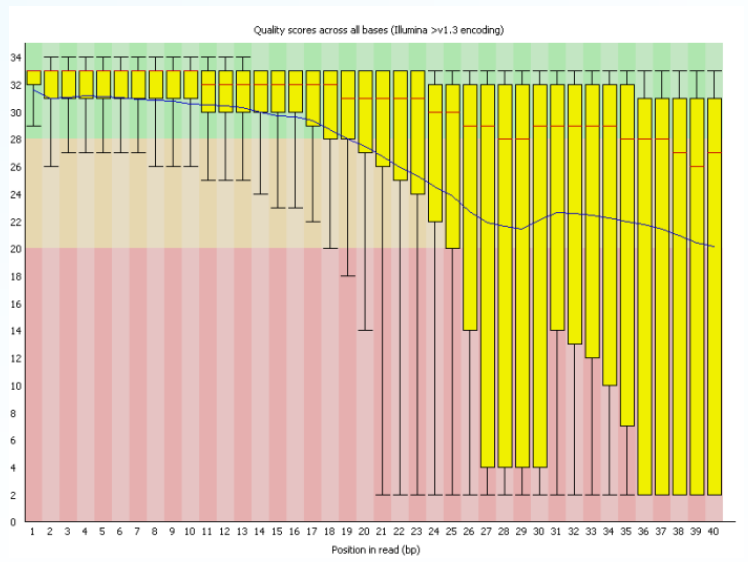
本节内容大纲

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

易生信

- 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，空间充足网速快推荐Anaconda(大小2G)
- 生物软件安装必添加Bioconda频道

https://conda.io/docs/

Miniconda2

- 最流行的Python数据科学管理平台
- https://conda.io/miniconda.html 推荐下载Linux python2.7 64位版本

```
# 下载软件，可根据官网下载最新版本

wget -c https://repo.continuum.io/miniconda/Miniconda2-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~/miniconda2，其它选项全yes

bash Miniconda2-latest-Linux-x86_64.sh
```

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

- 最流行的Python数据科学管理平台之一
- Anaconda较Miniconda默认多安装几百个Python包(此方法可选)
- https://www.anaconda.com 推荐下载Linux Python2.7 64-Bit (x86)版

```
# 下载软件，可根据官网下载最新版本

wget -c https://repo.anaconda.com/archive/Anaconda2-2019.03-Linux-x86_64.sh

# 管理员推荐安装目录设为conda2，普通用户建议默认，其它选项全yes

bash Anaconda2-2019.03-Linux-x86_64.sh
```

- **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 https://bioconda.github.io/conda-recipe_index.html
- 软件布署系统，方便用户定制软件及依赖关系；
- 12609个生物信息软件及常多版本，如收录fastqc常用的20个版本；
- 超250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了(吃水不忘挖井人)。
- 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/**

Bioconda: a sustainable and comprehensive software distribution for the life sciences, 2018, Nature Methods

质控软件安装

- # 质量评估软件fastqc

  conda install fastqc

  fastqc -v # FastQC v0.11.8

- # 多样品评估报告汇总multiqc

  conda install multiqc

  multiqc --version # multiqc, version 1.5

- # 质量控制流程kneaddata，安装指定版本，0.72不支持我们的数据

  conda install kneaddata=0.6.1

注意记录安装软件版本！

默认安装最新版，保证错误最少且功能最全

有问题时安装指定版本，保证可成功运行；

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human_transcriptome 、小鼠基因组mouse_C57BL 、ribosomal_RNA SILVA128数据库

- # 如下载人类基因组bowtie2索引至指定数据目录

  kneaddata_database --download human_genome bowtie2 ~/kneaddata/human_genome

- 其它物种可自行下载并使用bowtie2建索引，可参考下方链接教程

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

有参分析流程MetaPhlAn2、HUMAnN2

- # 安装MetaPhlAn2、HUMAnN2和所有依赖关系

  conda install humann2
  humann2_databases # 显示可用数据库

- # 微生物泛基因组数据库5.37G

  humann2_databases --download chocophlan full ~/db/humann2

- # UniRef90功能基因diamond索引 10.3G

  humann2_databases --download uniref uniref90_diamond ~/db/humann2

设置humann2默认参数：数据库位置、线程数

- # 显示参数

  humann2_config --print

- # 如修改线程数

  humann2_config --update run_modes threads 8

  humann2_config --update database_folders protein ~/db/humann2/uniref

  humann2_config --update database_folders nucleotide ~/db/humann2/chocophlan

- # metaphlan2数据库默认会自动下载，位于程序所在目录的db_v20和databases下各一份。系统安装软件需管理员运行一次下载数据

宏基因组有参流程——实战分析大纲

一. 软件安装和数据库部署

二. **KneadData质控**

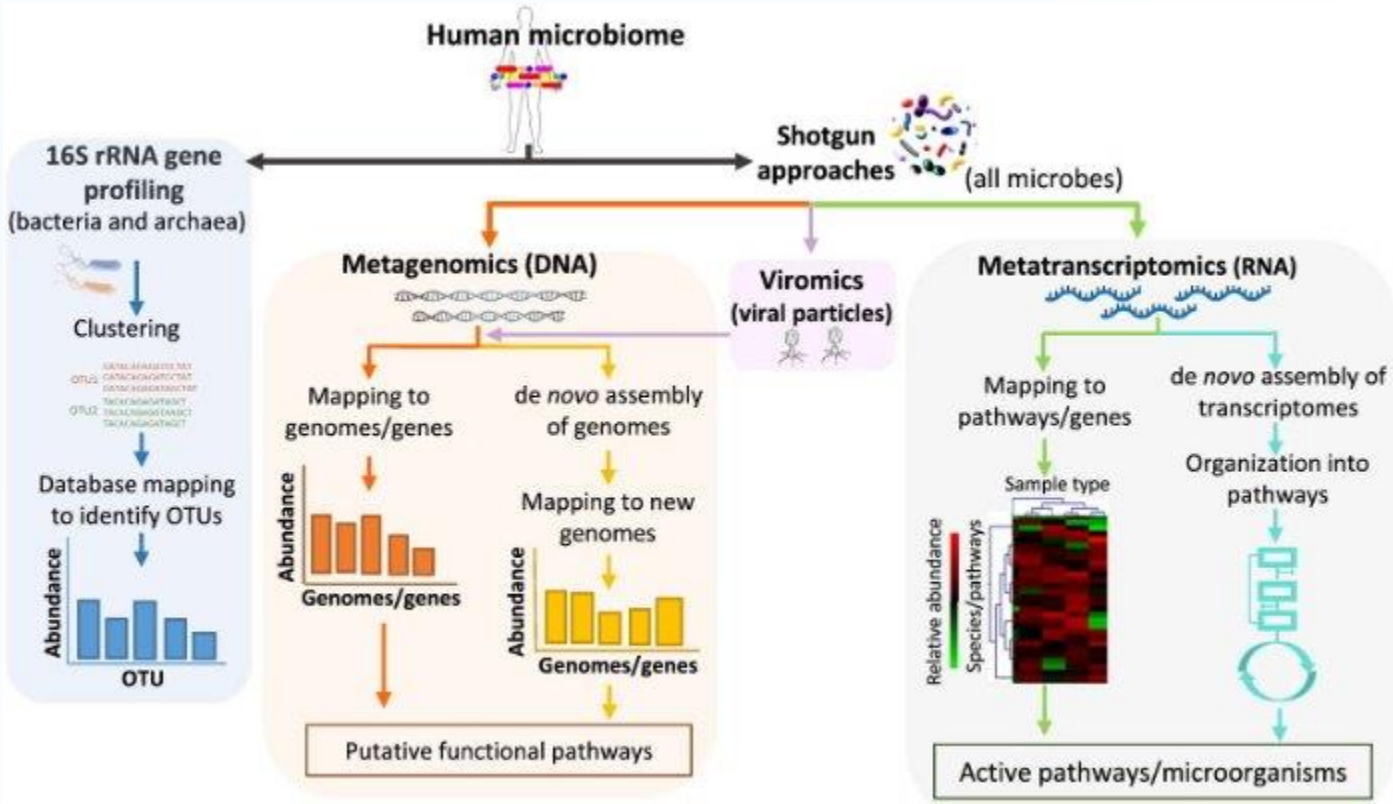
三. MetaPhlAn2物种组成

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



分析开始前需要设置的环境变量

- # 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~/miniconda2
- **soft=/conda**
- # 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

```
p136C_1.fq.gz  p136N_1.fq.gz  p144C_1.fq.gz  p144N_1.fq.gz  p153C_1.fq.gz  p153N_1.fq.gz
p136C_2.fq.gz  p136N_2.fq.gz  p144C_2.fq.gz  p144N_2.fq.gz  p153C_2.fq.gz  p153N_2.fq.gz
```

```
@SRR3586062.883556
CTTGGGGCTGCTGAGCTTCATGCTCCCCTCCTGCCTCAAGGACAATAAGGAGATCTTCGACAAGCCTGCAGCAGCTCGCATCGACGCCCTCATCGCTGAGG
+
CCCFFFFFHHHHHIJJJJJJJJIJIJJJJGIJDGIJEIIJIJJJJJJJJJIJJJJIJJIJJJJJHHHFFFFECEEEDDDDD?BDDDDDDBDDDDDDDDBBBDD
@SRR3586062.3376311
GACGGTGTCCTCAGGACCCTTCAGTGCCTTCATGATCTGCTCAGAGGTGATGGAGTCACGGACGAGATTCGTCGTGTCAGCACGTAGGATGCGGTCGCCTG
+
@@@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	Sex	Individual
p136C	Cancer	M	p136
p136N	Normal	M	p136

•学习全基因组测序数据分析 2fasta&fastq
•样品命名 注意事项 实例

宏基因组数据质量

- 常用Illumina PE150 (双端 150 bp)，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 2552

- 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 9887

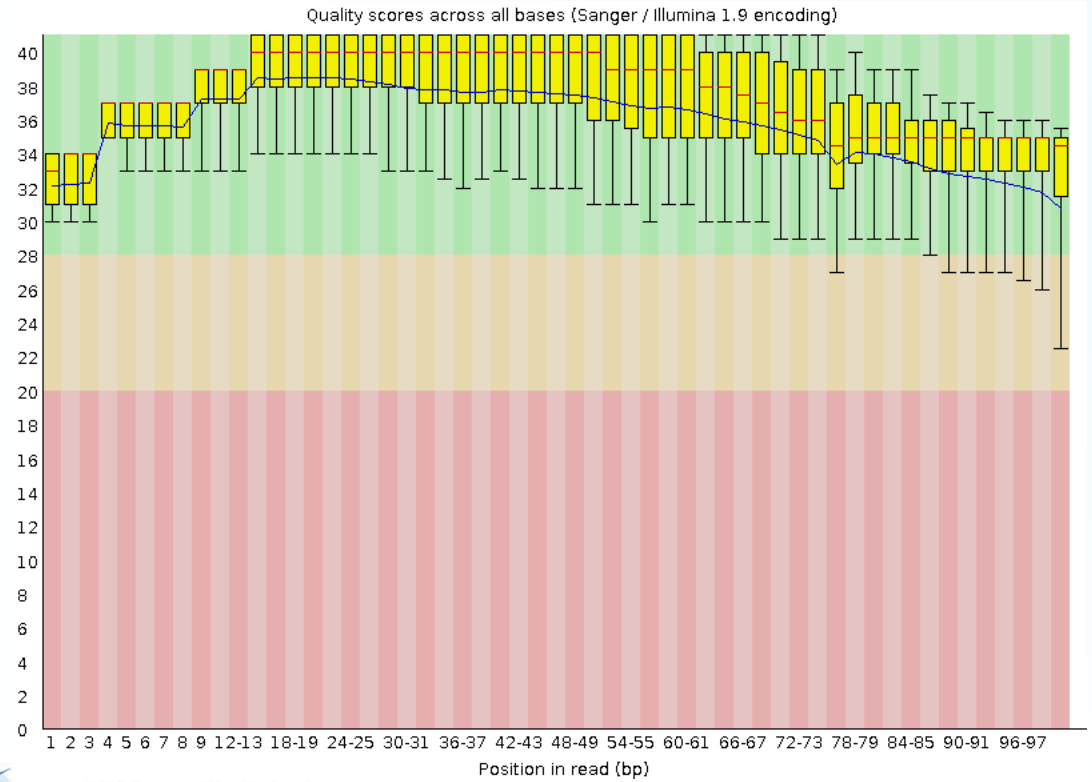
- 去除宿主——bowtie2比对宿主基因组；筛选非宿主序列

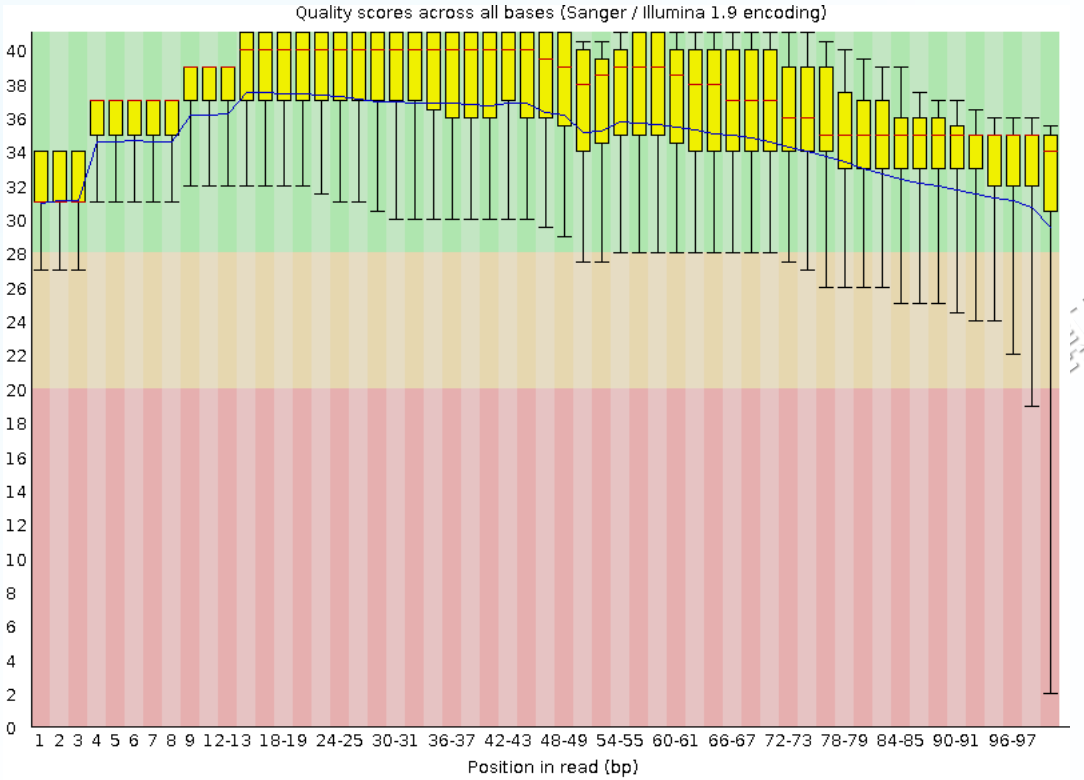
Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357, doi:10.1038/nmeth.1923 (2012). Cited by 14346

- 有时还要去PhiX

FastQC质量评估

- # fastqc每个文件一个线程，6个双端样本12个文件，设置6线程
- time fastqc seq/*.gz -t 6 # 9s

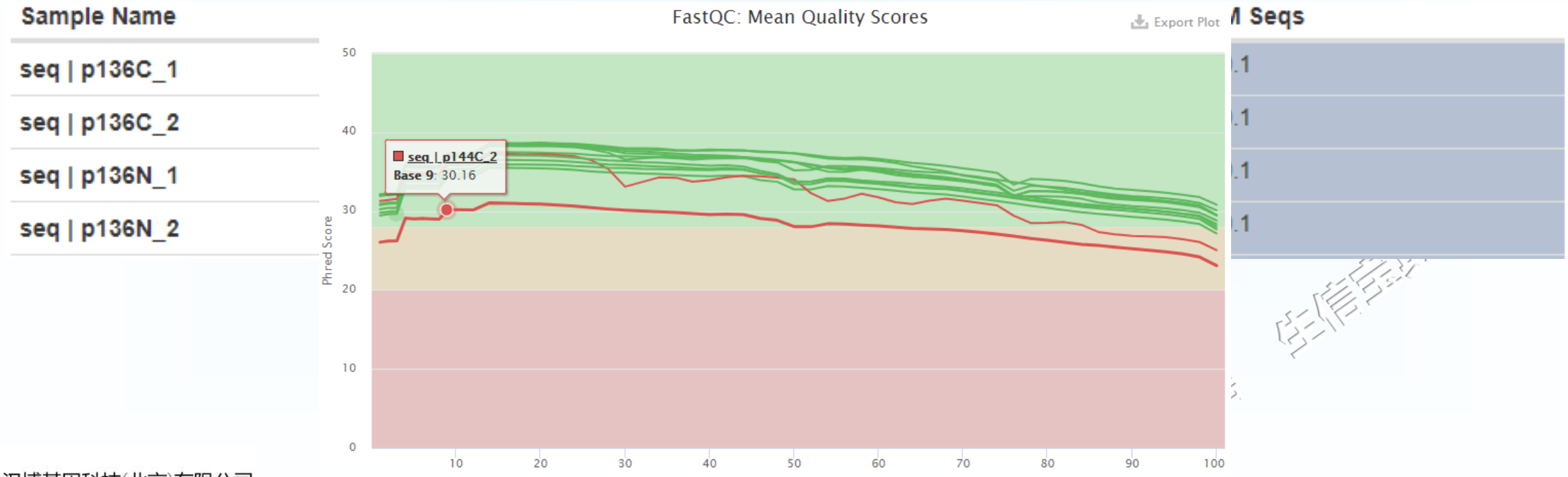




•数据的质量控制软件——fastQC

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

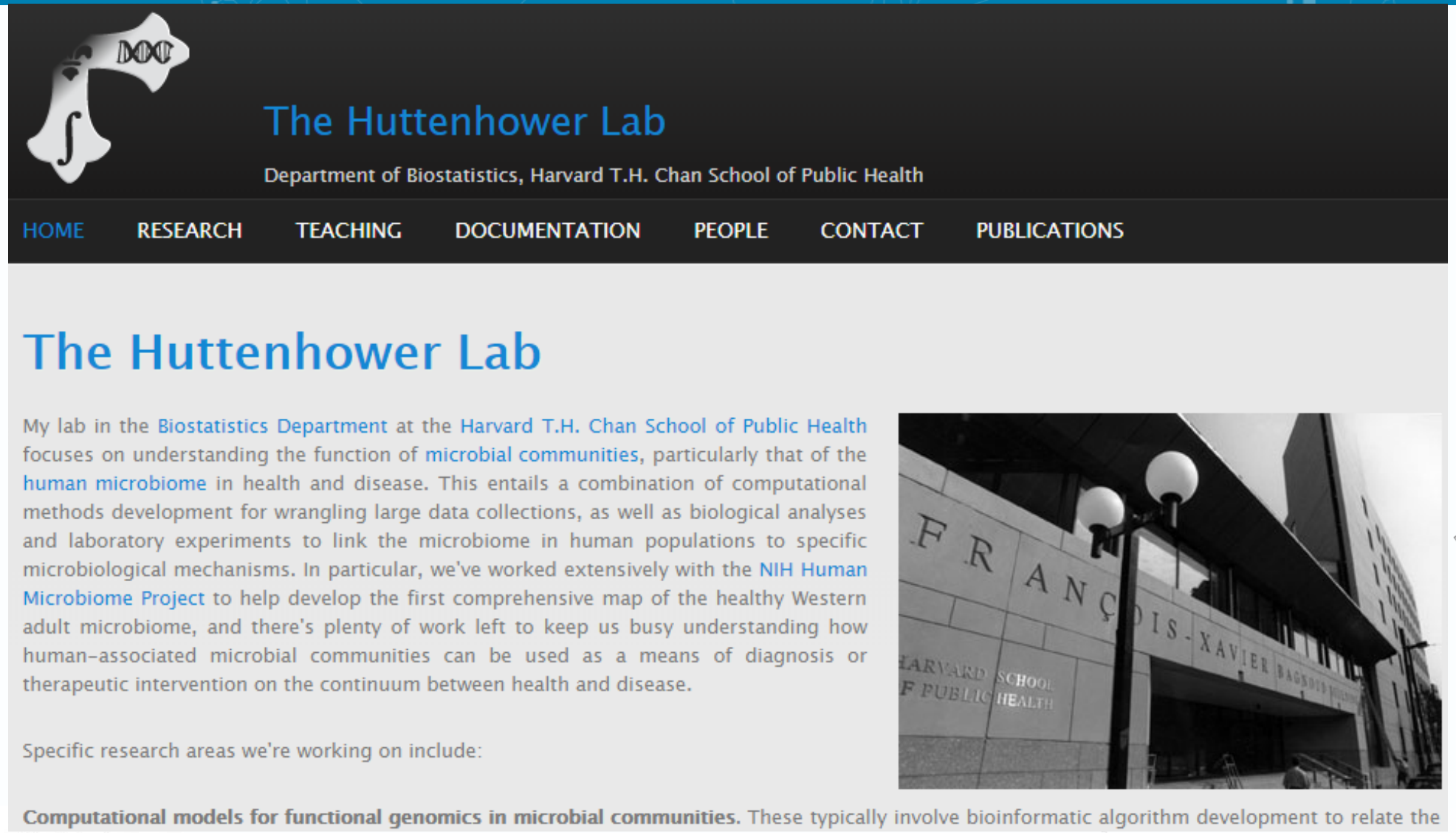


•2数据质控fastqc, Trimmomatic, MultiQC, khmer

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3部分；
- 依赖Trimmomatics、Bowtie2、Python脚本等，如何快速完成此分析呢?
- 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)
- 支持Conda安装
- 预构建了人类、小鼠数据库

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

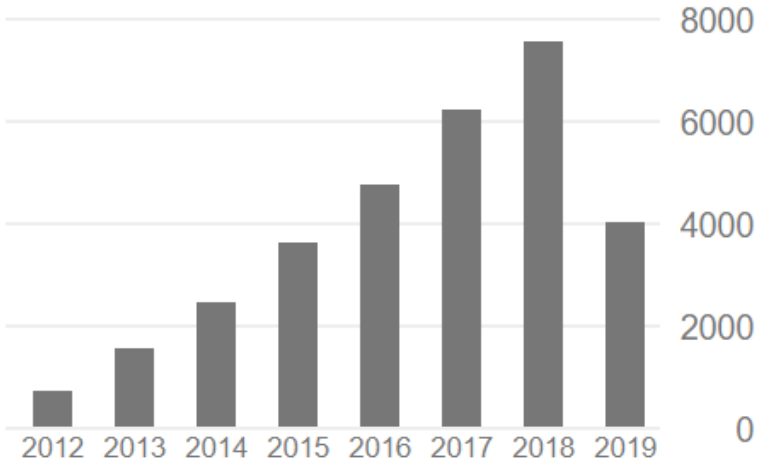
Department of Biostatistics, Harvard School of Public Health
在 hsph.harvard.edu 的电子邮件经过验证

computational metagenomics　human microbiome
biological data mining

标题	引用次数	年份
Structure, function and diversity of the healthy human microbiome C Huttenhower, D Gevers, R Knight, S Abubucker, JH Badger, ... nature 486 (7402), 207	4251	2012
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	2614	2013

引用次数　查看全部

	总计	2014 年至今
引用	32472	28768
h 指数	75	71
i10 指数	162	156



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p136C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/p144C_1.fq.gz -i seq/p144C_2.fq.gz \
  -o temp/qc -v -t 3 --remove-intermediate-output \
  --trimmomatic ${soft}/share/trimmomatic/ --trimmomatic-options 'SLIDINGWINDOW:4:20 MINLEN:50' \
  --bowtie2-options '--very-sensitive --dovetail' -db ${db}/kneaddata/human_genome/Homo_sapiens
```

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?
- Parallel是GUN收录的官方程序
- 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
- 可以直接在Ubuntu仓库中安装

  sudo apt install parallel
- 作者要求引用，如不想引用也可付10000欧元购买

Tange, O. *Gnu parallel 2018*. (Lulu. com, 2018). Cited by 15

并行质量控制(质控)实例

- 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

```
time parallel -j 3 --xapply \
  "kneaddata -i {1} -i {2} \
  -o temp/qc -v -t 3 --remove-intermediate-output \
  --trimmomatic ${soft}/share/trimmomatic/ --trimmomatic-options
'SLIDINGWINDOW:4:20 MINLEN:50' \
  --bowtie2-options '--very-sensitive --dovetail' -db
${db}/kneaddata/human_genome/Homo_sapiens" \
  ::: seq/*_1.fq.gz ::: seq/*_2.fq.gz
```

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 seq/kneaddata_read_counts.txt

查看结果

cat seq/kneaddata_read_counts.txt

Sample	raw pair1	raw pair2	trimmed pair1	trimmed pair2	trimmed orphan1	trimmed orphan2	decontaminated	Homo_sapiens pair1
p136C_1_kneaddata	75000	75000	65316	65316	5061	1635	65278	65278	673	6023	65278	65278	673	6023
p136N_1_kneaddata	75000	75000	60548	60548	6048	1721	60022	60022	839	6855	60022	60022	839	6855
p143C_1_kneaddata	75000	75000	48082	48082	2116	8597	45648	45648	1466	9098	45648	45648	1466	9098
p143N_1_kneaddata	75000	75000	47003	47003	7520	2135	44901	44901	1373	8607	44901	44901	1373	8607
p144C_1_kneaddata	75000	75000	50387	50387	6974	1832	48102	48102	1308	7326	48102	48102	1308	7326
p144N_1_kneaddata	75000	75000	62217	62217	6477	1848	62062	62062	851	7483	62062	62062	851	7483
p146C_1_kneaddata	75000	75000	60959	60959	6028	1850	60362	60362	898	6977	60362	60362	898	6977
p146N_1_kneaddata	75000	75000	67958	67958	3807	1236	67856	67856	499	4532	67856	67856	499	4532
p153C_1_kneaddata	75000	75000	62540	62540	1567	7752	62502	62502	942	8367	62502	62502	942	8367
p153N_1_kneaddata	75000	75000	63336	63336	6043	1901	63284	63284	778	7152	63284	63284	778	7152
p156C_1_kneaddata	75000	75000	65866	65866	4811	1693	65856	65856	662	5836	65856	65856	662	5836
p156N_1_kneaddata	75000	75000	67204	67204	4242	1378	67173	67173	532	5085	67173	67173	532	5085

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

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

General Statistics

Copy table | Configure Columns | Plot | Showing 24/24 rows and 5/7 columns.

Sample Name	% Aligned	% Dropped	% Dups	% GC	M Seqs
decompressed_1CWW75_p144C_1		21.1%			
decompressed_RSipJU_p153C_1		4.2%			
decompressed_X9vWji_p136C_1		4.0%			
decompressed_gZOHJl_p144N_1		5.9%			
decompressed_i5cn2c_p153N_1		5.0%			
decompressed_yfS4Dp_p136N_1		8.9%			
p136C_1_kneaddata.trimmed.single.1	0.1%				
p136N_1_kneaddata.trimmed.single.1	0.8%				
p144C_1_kneaddata.trimmed.single.1	4.2%				
p144N_1_kneaddata.trimmed.single.1	0.2%				
p153C_1_kneaddata.trimmed.single.1	0.1%				
p153N_1_kneaddata.trimmed.single.1	0.1%				
temp | qc | p136C_1_kneaddata_paired_1			0.1%	36%	0.1
temp | qc | p136C_1_kneaddata_paired_2			0.1%	36%	0.1
temp | qc | p136N_1_kneaddata_paired_1			1.0%	39%	0.1
temp | qc | p136N_1_kneaddata_paired_2			1.0%	39%	0.1

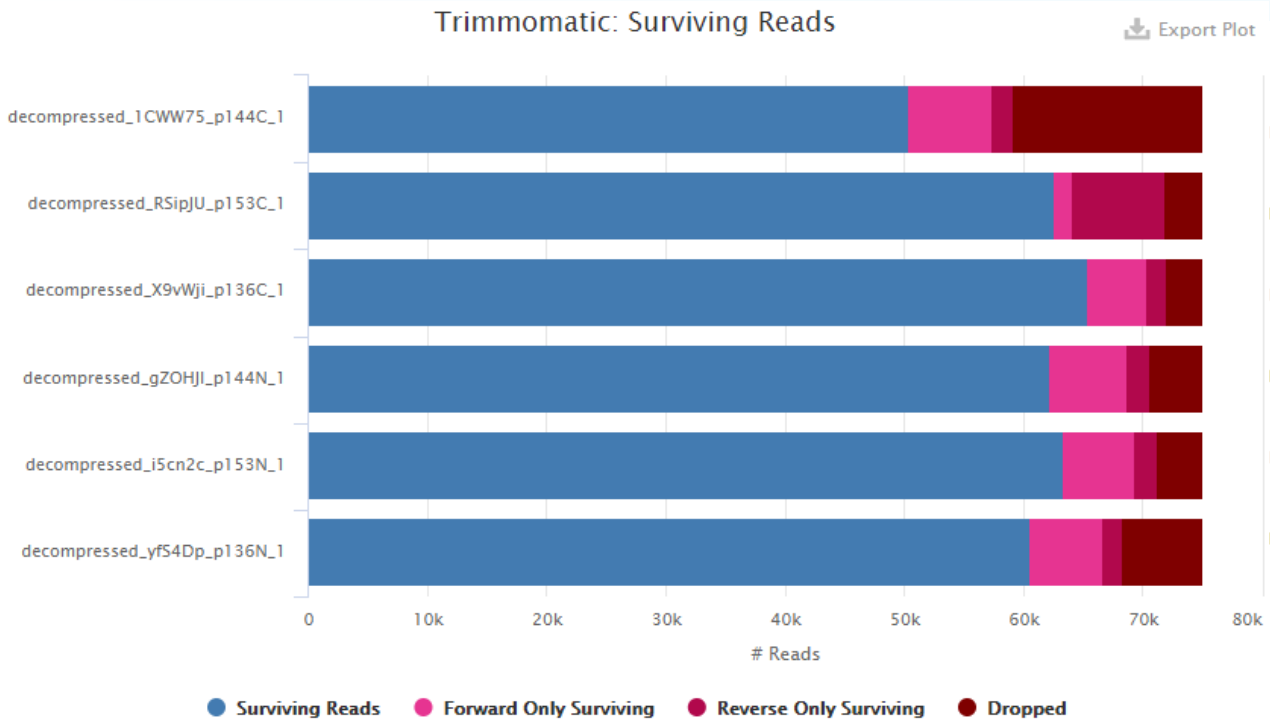
质控步骤，扔掉低质量的比例

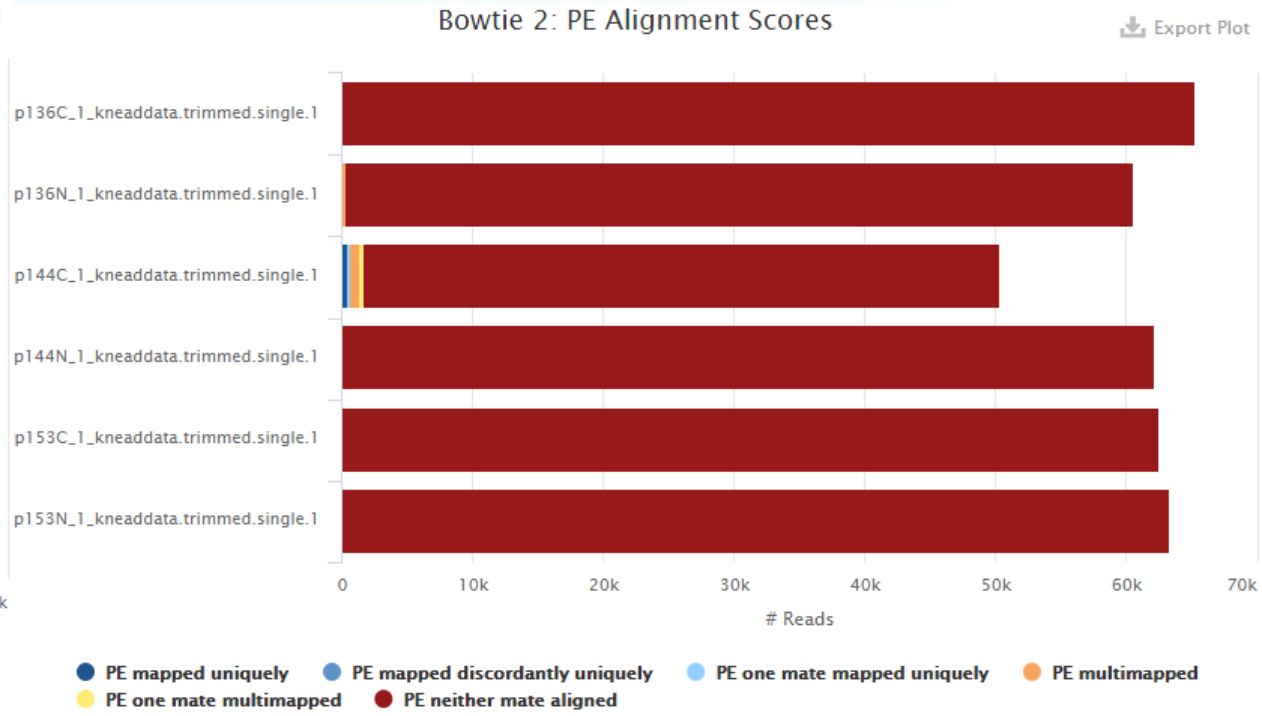
去宿主步骤，宿主含量

质量评估步骤，基本信息

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

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





总结

- Conda是软件安装和管理神器，Bioconda频道是生物学家的福音，超1.2万个生信软件及数万版本满足你各种需求，记得引用NM的文章；
- 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；
- 质控需要trimmomatics，bowtie2和宿主基因组等多软件和数据库，哈佛大学Huttenhover组编写的质控流程KneadData一站解决软件数据库，以及分析方法，参数选择等众多烦恼；
- MultiQC用于质控前后的评估和结果汇总，包括fastqc、trimmomatic和bowtie2结果的汇总、可视化方便阅读、比较和图表导出；
- parallel 多样本批量队列管理工具，多任务、多线程管理专家 。

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
- 加拿大生信网 https://bioinformatics.ca/
- 加拿大生信网宏基因组课程中文版——挖掘微生物组生物标记
- 美国高通量开源课程 https://github.com/ngs-docs
- The Huttenhower Lab http://huttenhower.sph.harvard.edu/
- LangilleLab https://github.com/LangilleLab/microbiome_helper
- Susan Holmes http://statweb.stanford.edu/~susan/

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