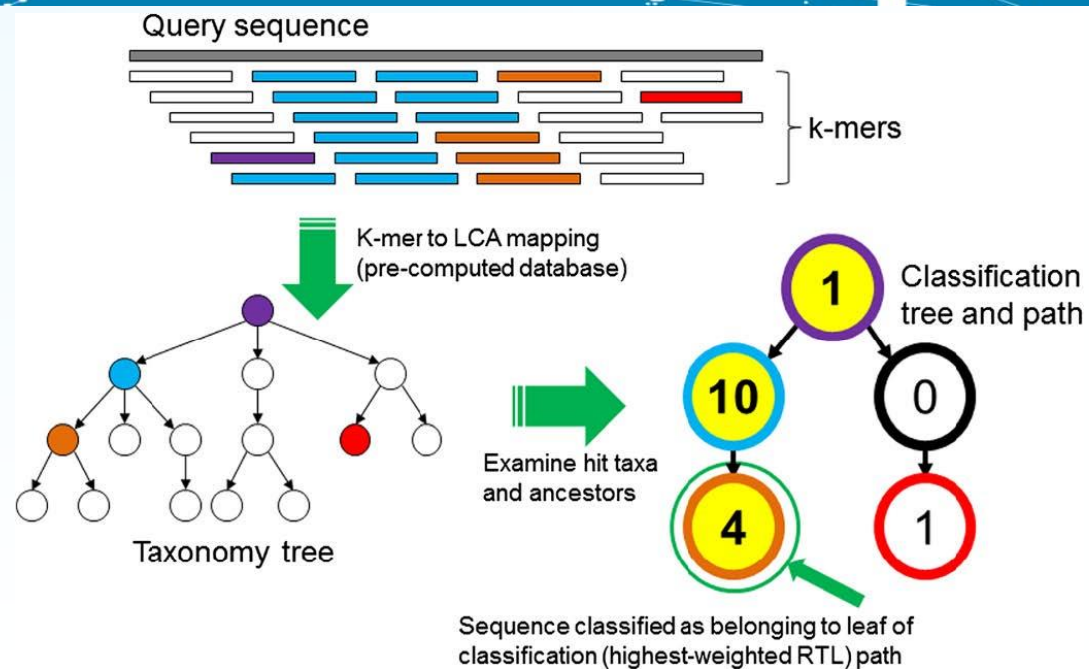




25 Kraken2物种注释

易生信
2023年7月22日



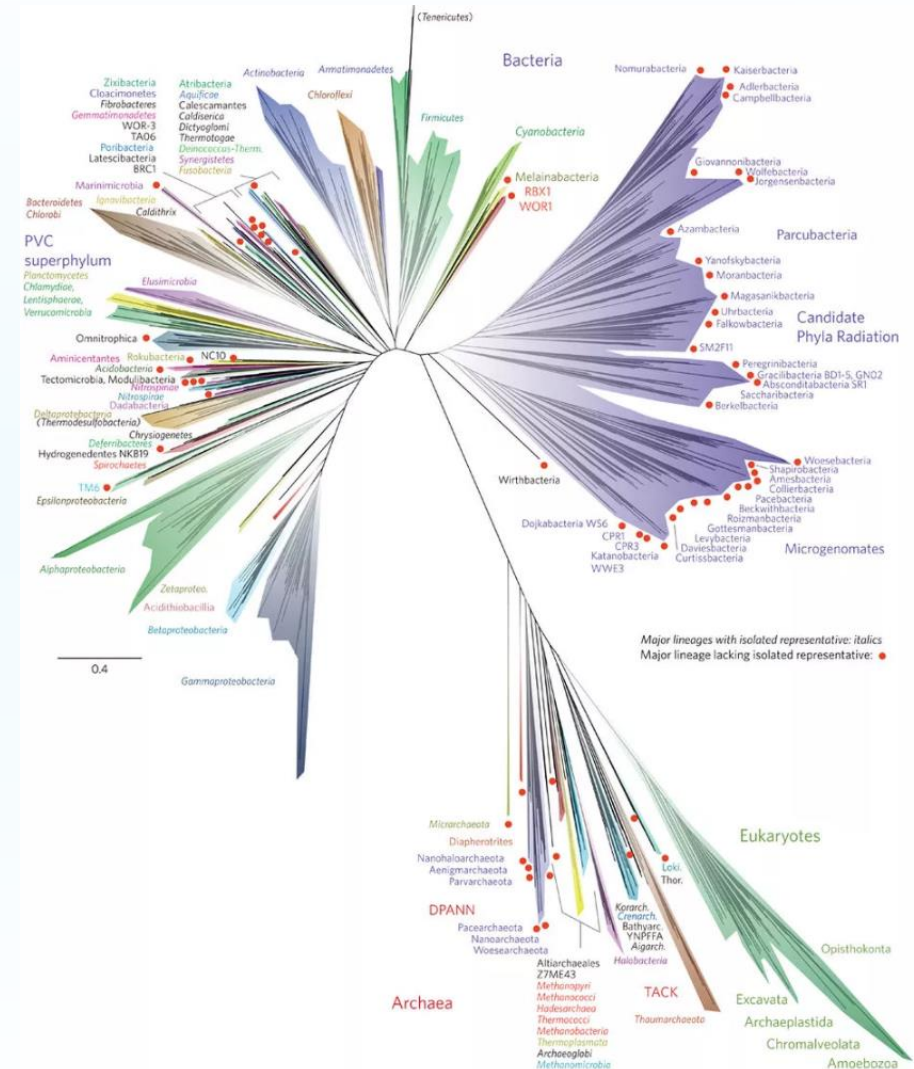
宏基因组分析

主要步骤常用软件和数据库



物种分类学注释

- 分类学(taxonomy): 是一门研究生物类群间的异同以及异同程度, 阐明生物间的亲缘关系、进化过程和发展规律的科学。
- 主要分为细菌、古菌和真核生物三大类;
- 常用七级分类法: 界(Kingdom)、门(Phylum)、纲(Class)、目(Order)、科 (Family)、属 (Genus)、种 (Species)



Laura A. Hug, ..., Jillian F. Banfield. 2016. A new view of the tree of life. **Nature Microbiology** 1: 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>



物种注释——相当于地址

- 界(Kingdom)、门(Phylum)、纲(Class)、目(Order)、科(Family)、属(Genus)、种(Species)
- 动物界、脊索动物门、哺乳纲、食肉目、熊科、大熊猫属、大熊猫
- 动物界、脊索动物门、哺乳纲、灵长目、人科、人属、智人种
- 国、省、市、县、镇、村、屯
- 中国、黑龙江省、哈尔滨市、五常县、冲河镇、三家子村、大排地屯
- 微生物进化快，属种不能保证与功能一致，常用株(Strain)关联功能
- 扩增子只测序部分16S序列，信息有限，仅能确定属水平

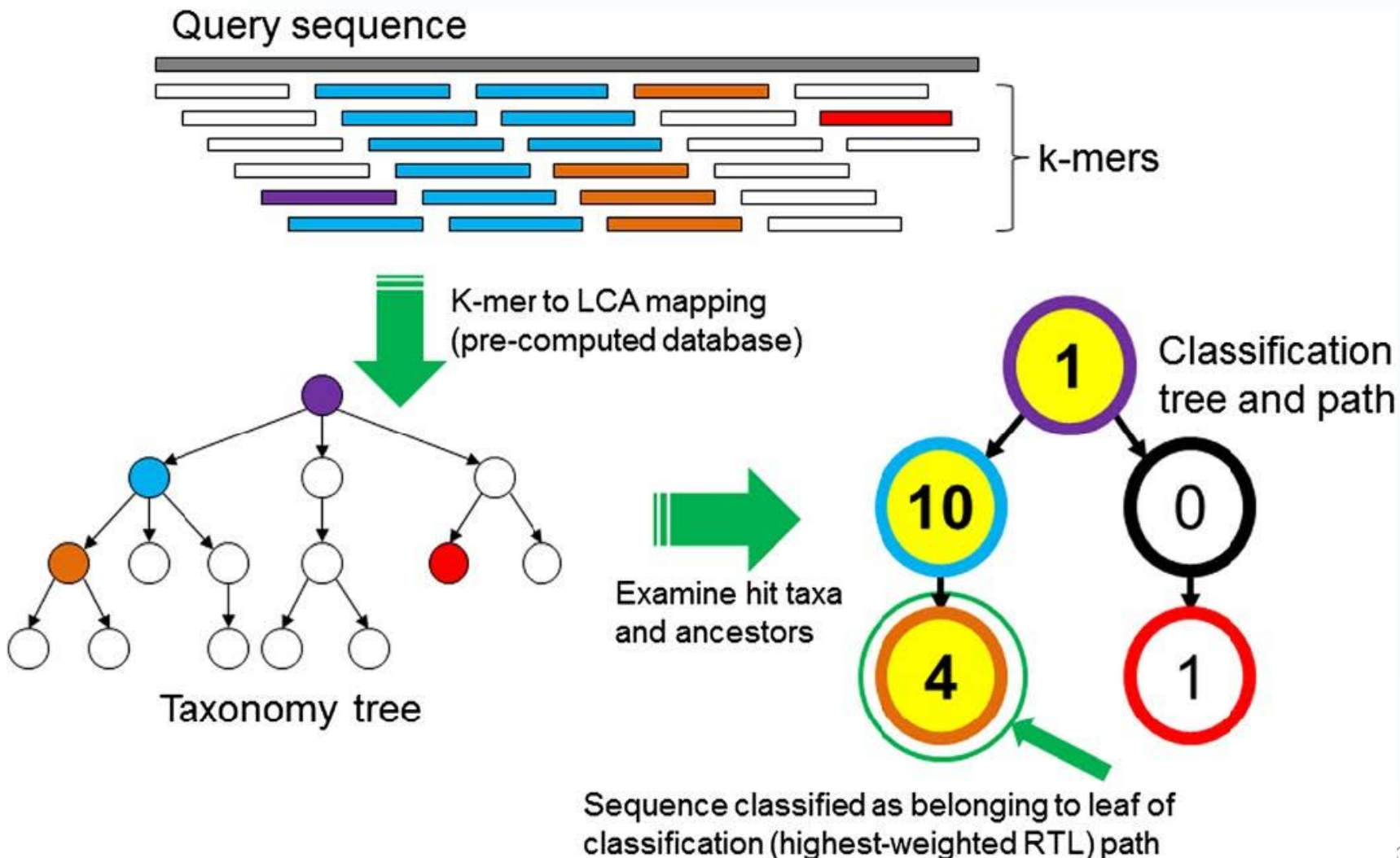


- NCBI——NR非冗余序列，NCBI发布的序列包含物种Taxonomy ID
- MetaPhlAn2——整理已发表基因组Marker基因数据库
- GTDB——基因组细菌120/古菌122单拷贝基因
- GreenGenes/RDP——原核生物核糖体(16S)数据库
- SILVA——原核、真核核糖体(16/18S)数据库



- 比对方法：与有物种注释的序列数据库比对，通过相似度进行物种注释；这种方法受限于数据库，且比对结果不准确。常用blast、diamond等。
- LCA(Lower Common Ancestor最低共同祖先)：此类方法常基于K-mer进行分类注释；目前认为方法较准确，但是注释到的物种信息很少，常用软件有Kraken系列、RDP classifier、Sintax等。

Kraken序列分类算法: LCA



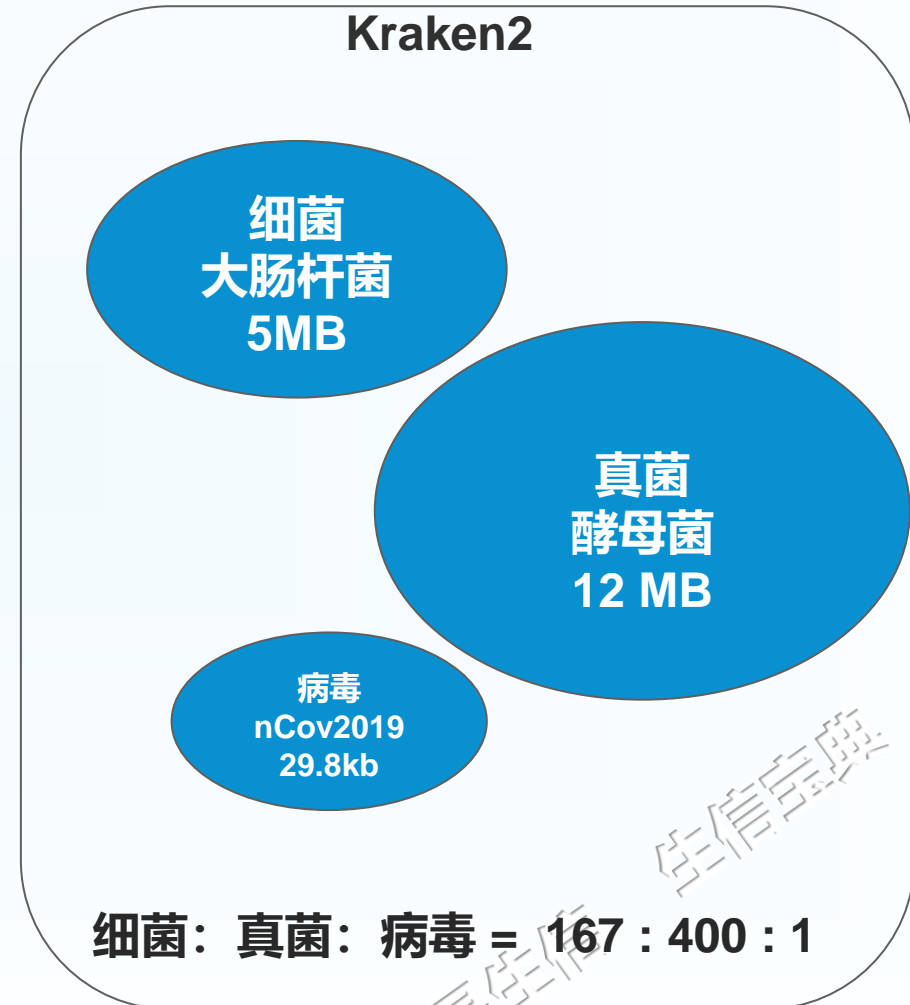
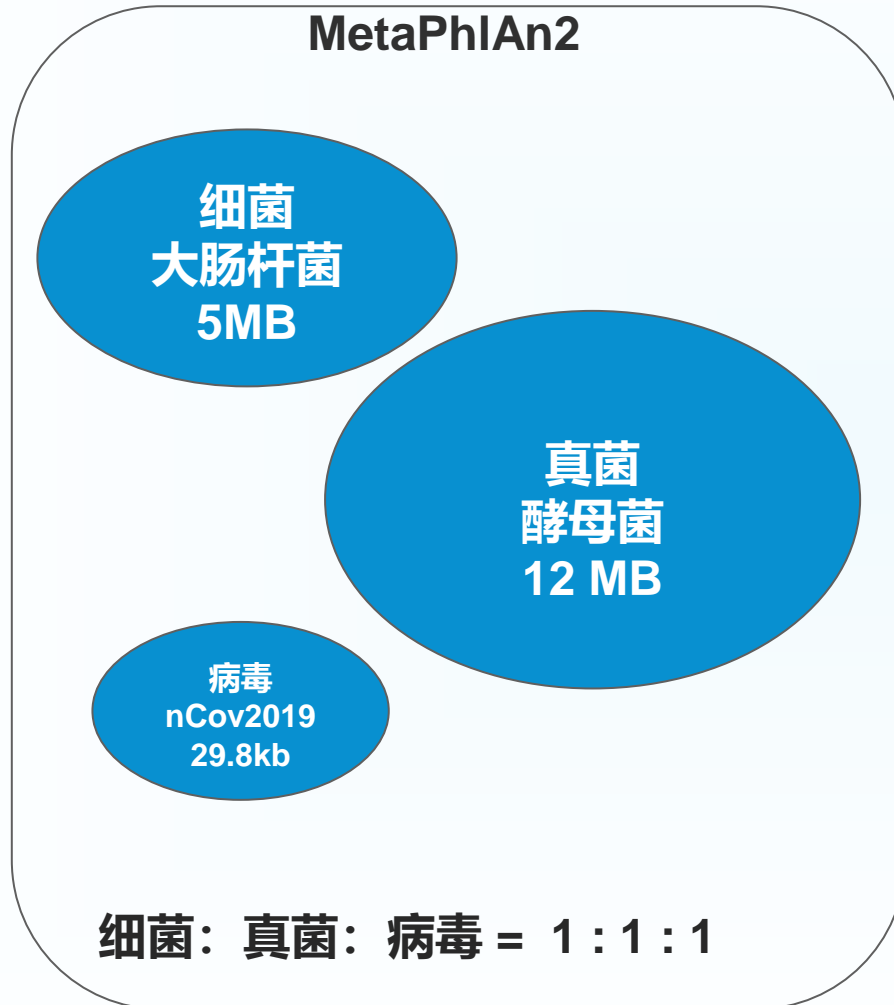
Lower Common Ancestor
最低共同祖先

软件默认阈值为0,
存在过分类问题

[Kraken: 使用精确比对的超快速宏基因组序列分类软件](#)

相对丰度：分类(taxonomic) vs 序列(sequence)

Nature子刊：刘洋或、Rob Knight等评测不同宏基因组物种定量方法及其对结果的影响



Zheng Sun, Shi Huang, Meng Zhang, Qiyun Zhu, Niina Haiminen, Anna Paola Carrièri, Yoshiki Vázquez-Baeza, Laxmi Parida, Ho-Cheol Kim, Rob Knight & Yang-Yu Liu. (2021). Challenges in benchmarking metagenomic profilers. *Nature Methods*, doi: <https://doi.org/10.1038/s41592-021-01141-3>

- Kraken有安装数据库过大，结果可读性差，需要二次转换等缺点。
- kraken2横空出世 <https://github.com/DerrickWood/kraken2>

DerrickWood / **kraken2** Public

<> Code Issues 327 Pull requests 16 Actions Proj

master 5 branches 7 tags Go to file

BenLangmead Merge pull request #697 from ch4rr0/pyth... df20a8f

data	Add small viral testing set
docs	Prep for 2.1.2
scripts	Merge pull request #697 from ch4rr0/python
src	Merge pull request #675 from ch4rr0/maskin

Scholar

Kraken

About 34,300 results (0.06 sec)

[HTML] **Kraken**: ultrafast metagenomic sequence classification using exact alignments

DE Wood, SL Salzberg - Genome biology, 2014 - genomebiology.biomedcentral.com

... **Kraken**, along with a reduced memory version of **Kraken** (MiniKraken), quick execution versions of both (**Kraken**-Q and MiniKraken-Q), and **Kraken** ... genomes from GenBank (**Kraken**-GB).

☆ Save Cite Cited by 3531 Related articles All 20 versions

[HTML] Improved metagenomic analysis with **Kraken 2**

DE Wood, J Lu, B Langmead - Genome biology, 2019 - Springer

Although **Kraken**'s k-mer-based approach provides a fast taxonomic classification of metagenomic sequence data, its large memory requirements can be limiting for some applications. ...

☆ Save Cite Cited by 2078 Related articles All 14 versions

Derrick E. Wood, Jennifer Lu, Ben Langmead. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20: 257. <https://doi.org/10.1186/s13059-019-1891-0>



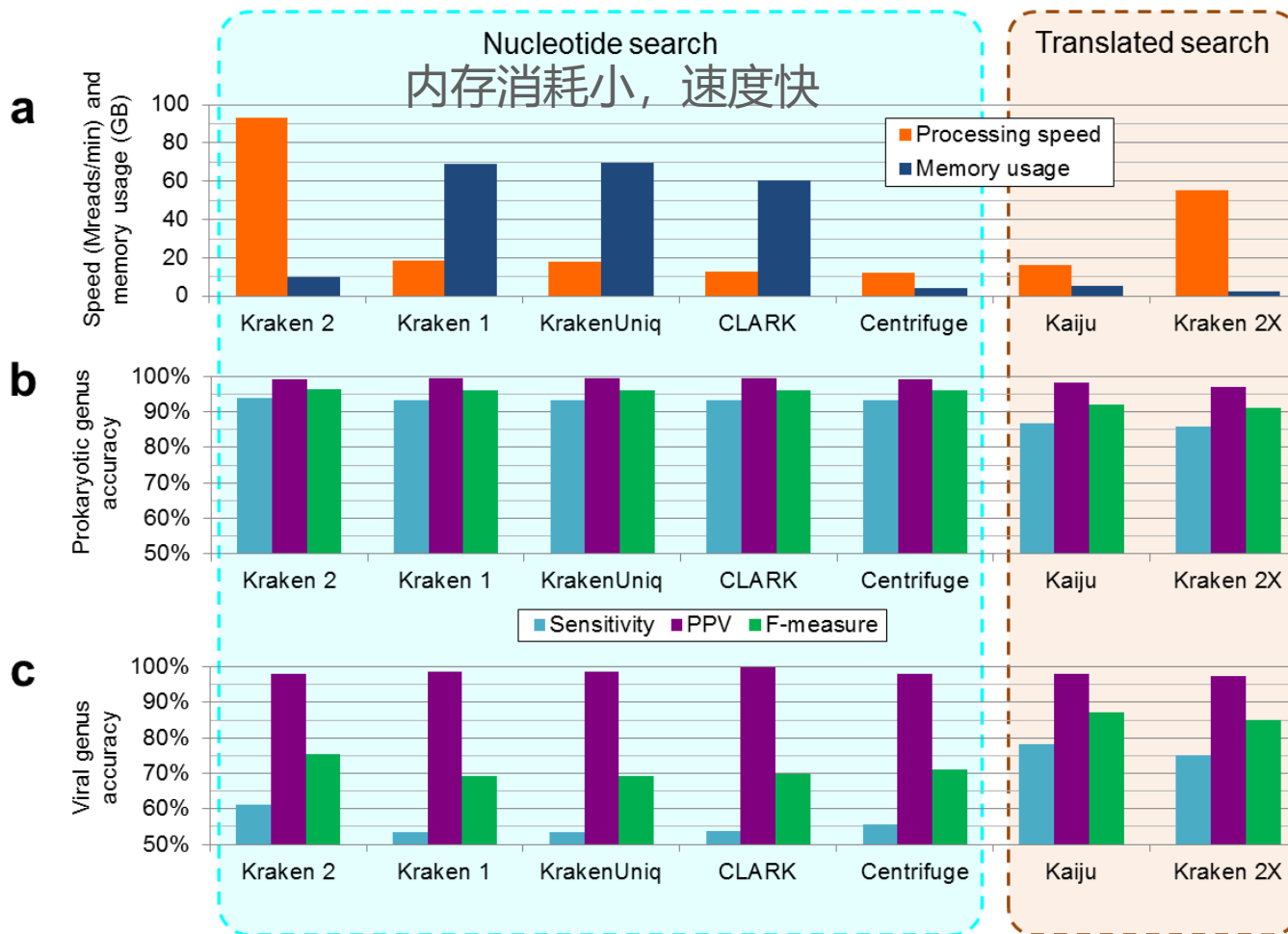
Kraken2与其它工具比较

图1. Kraken 2与其他工具之间的比较。

(a) 显示了每个分类器的处理速度 (M) 和内存使用情况 (GB)，对16线程下的5000万对配对模拟读长进行评估的。显示了 (b) 40个原核基因组和 (c) 10个病毒基因组的准确性结果。此处显示的是灵敏度 Sensitive，阳性预测值 (PPV) 和F值的结果。“Kraken 2X”是针对蛋白质数据库进行翻译搜索的Kraken 2

Simon H. Ye, Katherine J. Siddle, Daniel J. Park, Pardis C. Sabeti. 2019.
Benchmarking metagenomics tools for taxonomic classification. *Cell* 178: 779-794.

<https://doi.org/10.1016/j.cell.2019.07.010>



Kraken2安装和数据库配置

- 安装基于LCA算法的物种注释软件Kraken2
conda install kraken2
- 下载数据库: <https://benlangmead.github.io/aws-indexes/k2>
- 小内存/演示使用迷你库(PlusPFP-8/16) , 包括标准+原生动物+真菌+植物及Bracken2索引, 仅8G或16G, 可选完整版144G
- **不是数据库时间、体积大小版对注释比例影响非常大**
- 自定义数据库, 标准模式只下载5种数据库: 古菌archaea、细菌bacteria、人类human、载体UniVec_Core、病毒viral
kraken2-build --standard --threads 24 --db ~db/kraken2

基于NCBI数据库的Kraken2物种注释

多样本并行物种注释，推荐1个任务，最多3个，使用3倍内存

```
mkdir -p temp/kraken2
```

```
tail -n+2 result/metadata.txt|cut -f1|rush -j 1 \
```

```
'kraken2 --db ${db}/kraken2/plusfp8g --paired temp/qc/{1}_?.fastq \
```

```
--threads 3 --use-names --report-zero-counts \
```

```
--report temp/kraken2/{1}.report \
```

```
--output temp/kraken2/{1}.output'
```

屏幕会输出各样品注释比例，和运行时间 10 - 20 min



Kraken2实现kraken2结果的格式转换和筛选

○ 安装

```
conda install kraken2 -c bioconda
```

○ 批量转换kraken2的report结果为mpa格式(metaphlan格式, 可直接进行LEfSe分析)

```
for i in `tail -n+2 result/metadata.txt|cut -f1`;do  
    kreport2mpa.py -r temp/kraken2/${i}.report \  
        --display-header \  
        -o temp/kraken2/${i}.mpa  
done
```

易生信
毕生缘
宏基因组



Kraken2基于NCBI数据库注释reads层面

汇总样品物种组成表

```
mkdir -p result/kraken2
```

```
tail -n+2 result/metadata.txt|cut -f1|rush -j 1 \
```

```
'tail -n+2 temp/kraken2/{1}.mpa | sort | cut -f 2 | sed "1 s/^/{1}\n/" >  
temp/kraken2/{1}_count '
```

```
header=`tail -n 1 result/metadata.txt | cut -f 1`
```

```
tail -n+2 temp/kraken2/${header}.mpa | sort | cut -f 1 | sed "1 s/^/Taxonomy\n/" >  
temp/kraken2/0header_count
```

```
paste temp/kraken2/*count > result/kraken2/tax_count.mpa
```



物种组成表

Taxonomy	C1	C2	C3	C4	C5	C6	N1	N2	N3	N4	N5	N6	
d__Bacteria	44672	21242	19694	42120	45493	48646	33573	22230	45868	50203	52489	45116	
d__Bacteria p__Actinobacteria			304	880	383	2330	107	561	1402	6151	5284	2732	139
d__Bacteria p__Actinobacteria c__Actinobacteria				54	871	380	2319	103	416	1118	6081	5279	272
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces gaoshouyii				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces hongkongensis				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces howellii				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces israelii				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces oris				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces pacaensis				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces radicidentis				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces slackii				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces sp. Chiba101				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces sp. dk561				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces sp. oral taxon 414				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces sp. oral taxon 897				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces sp. Z16				12	394	61	46	28	63	571	1393	457	
d__Bacteria p__Actinobacteria c__Actinobacteria o__Actinomycetales f__Actinomycetaceae g__Actinomyces s__Actinomyces viscosus				12	394	61	46	28	63	571	1393	457	

- 本地/在线使用LEfSe差异比较，GraPhlAn或microbiomeViz可视化
- R语言统计分析alpha, beta和物种组成和可视化
- 直接使用STAMP差异比较和可视化



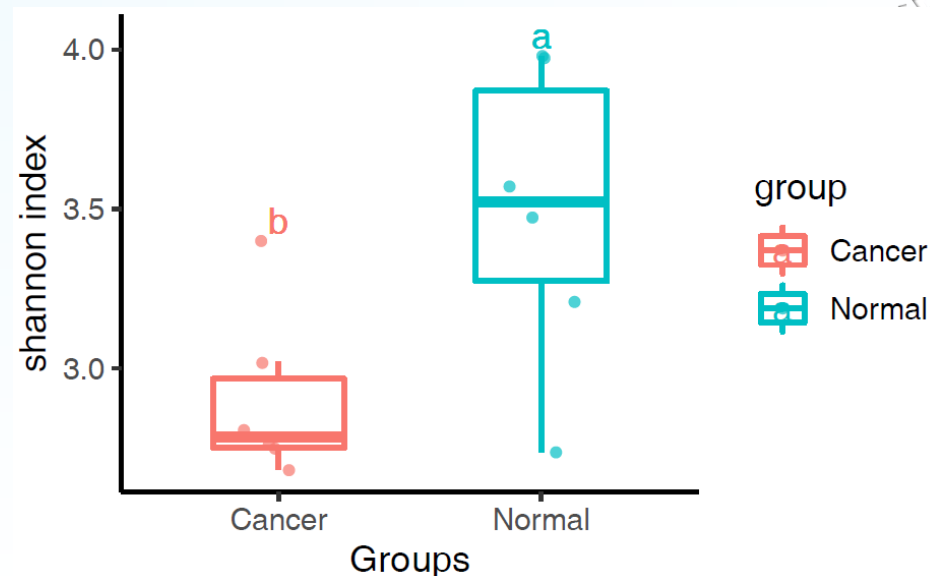
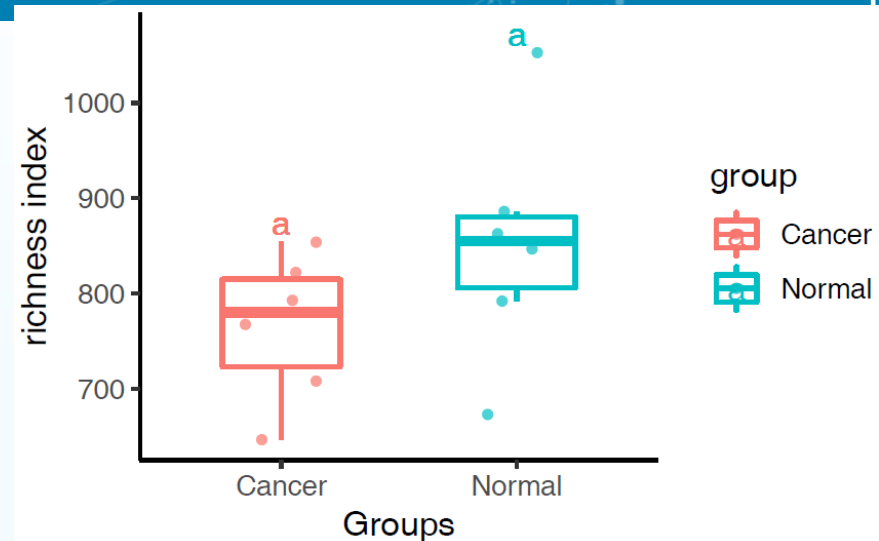
3StatPlot.sh - 物种Kraken2 - Alpha多样性

提取种级别、抽平、计算6种alpha多样性指数

```
Rscript $sd/kraken2alpha.R \  
--input result/kraken2/tax_count.mpa \  
--depth 0 \  
--species result/kraken2/tax_count.txt \  
--normalize result/kraken2/tax_count.norm \  
--output result/kraken2/tax_count.alpha
```

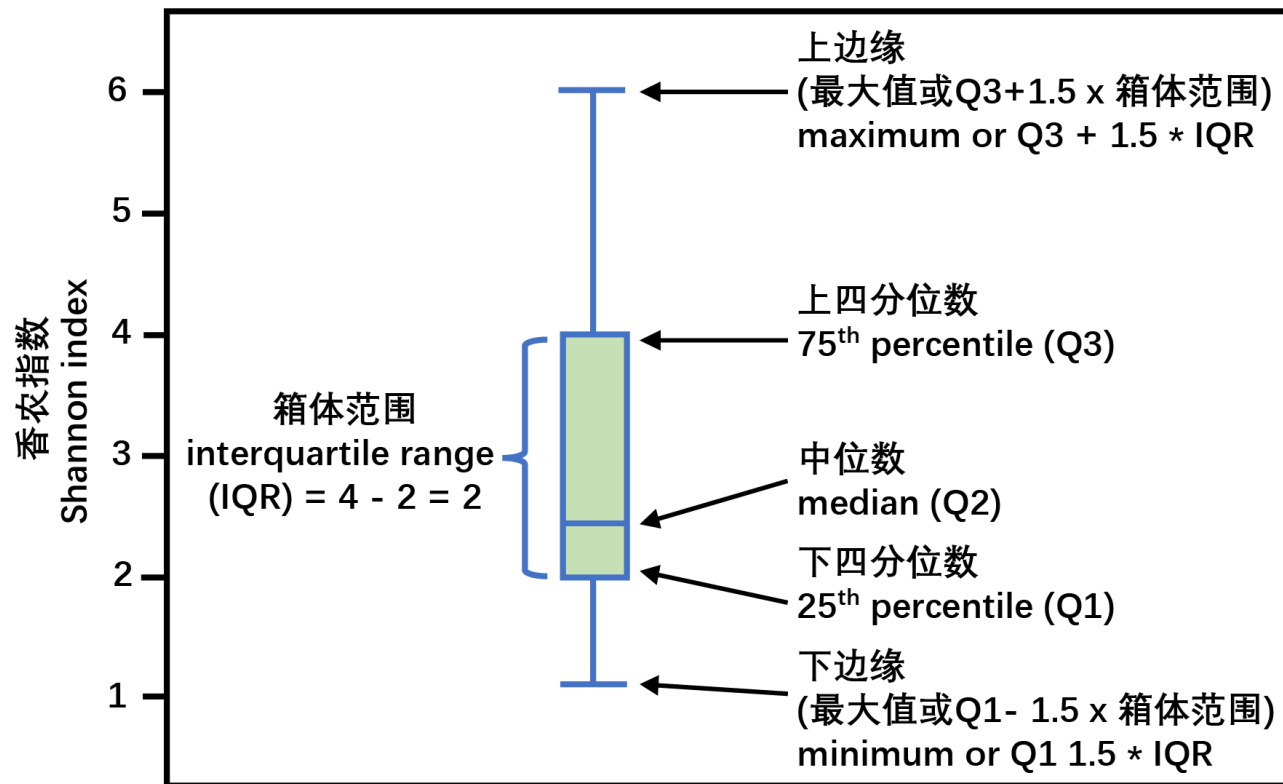
绘制箱线图,可选richness/chao1/shannon...

```
Rscript $sd/alpha_boxplot.R \  
-i result/kraken2/tax_count.alpha \  
-a shannon \  
-d result/metadata.txt \  
-n Group \  
-o result/kraken2/ \  
-w 89 -e 59
```

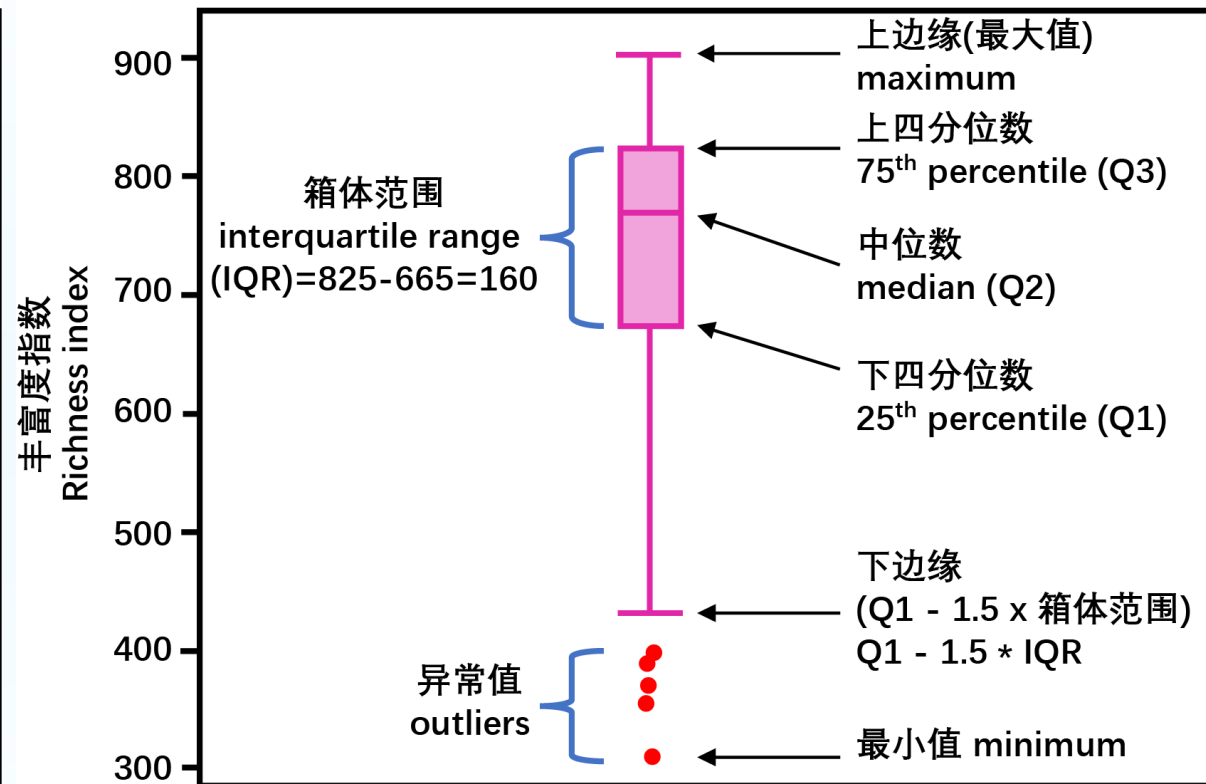


箱线图的基本知识

Alpha多样性香农指数箱线图(Boxplot of Shannon index)



Alpha多样性丰富度指数箱线图(Boxplot of richness index)



中文：百度百科箱形图：<https://baike.baidu.com/item/箱形图>

英文：斯坦福大学统计公开课<https://lagunita.stanford.edu/courses/Medicine/MedStats-SP/SelfPaced/about>

图注写法：[NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用*](#)

3StatPlot.sh - 物种Kraken2 - 热图

调整输入文件为spf文件，即物种丰度表格

可选分类级Kingdom / Phylum / Class / Order / Family / Genus / Species、分类显示数量

Rscript

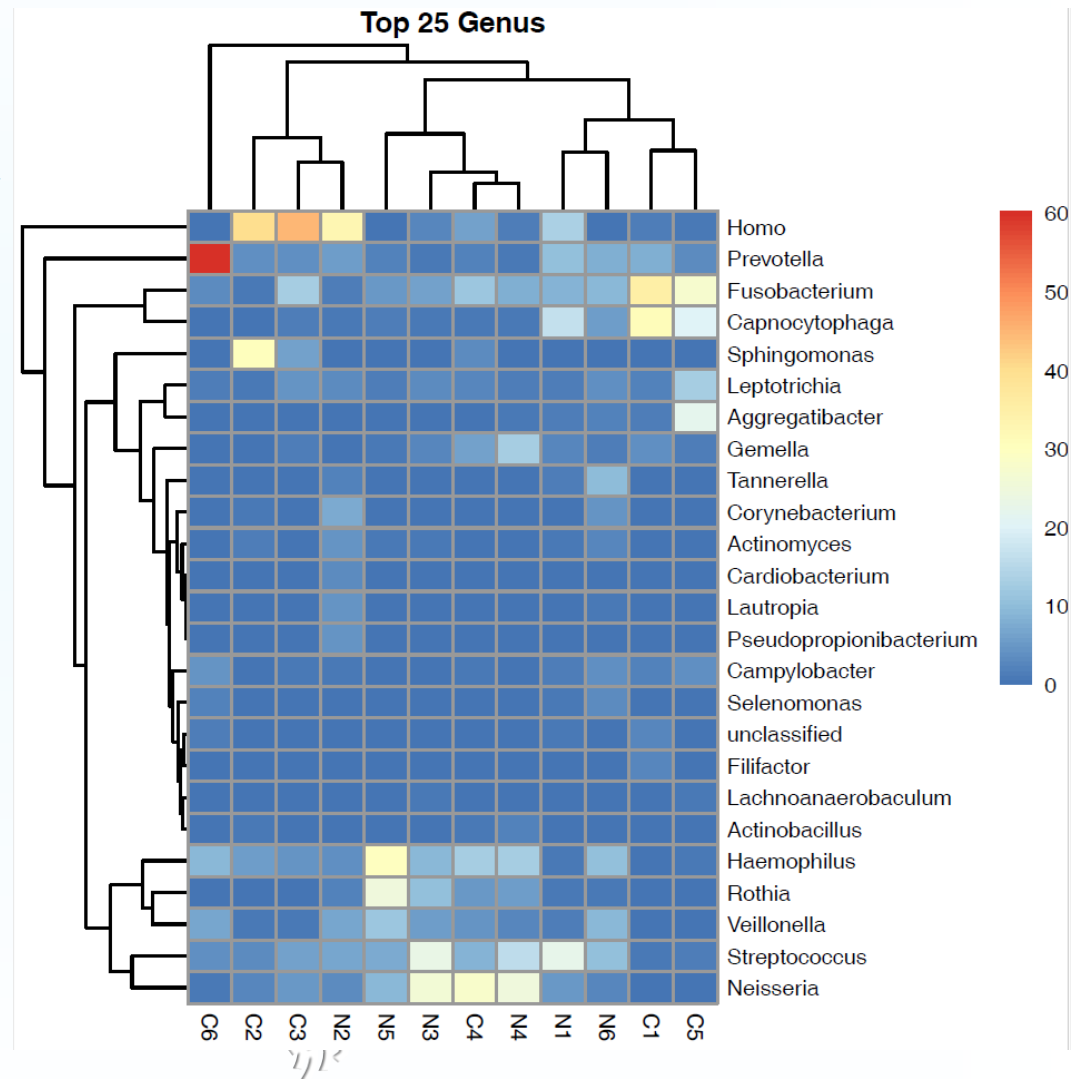
db/script/metaphlan_hclust_heatmap.R \

-i result/kraken2/tax_count.spf \

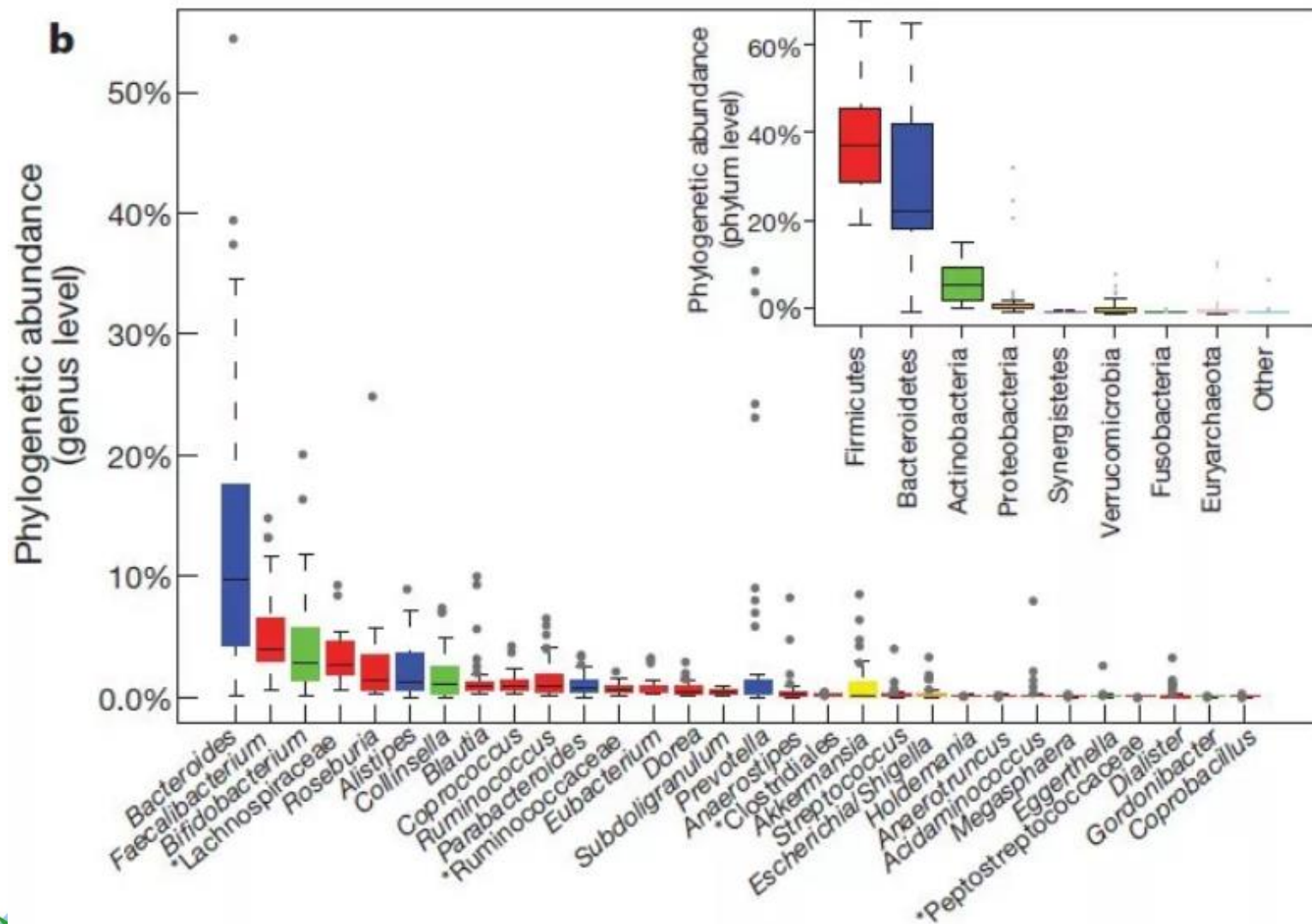
-t **Genus** \

-n **25** \

-o result/kraken2/heatmap_Genus



箱线图展示最高丰度的30个属和8个门

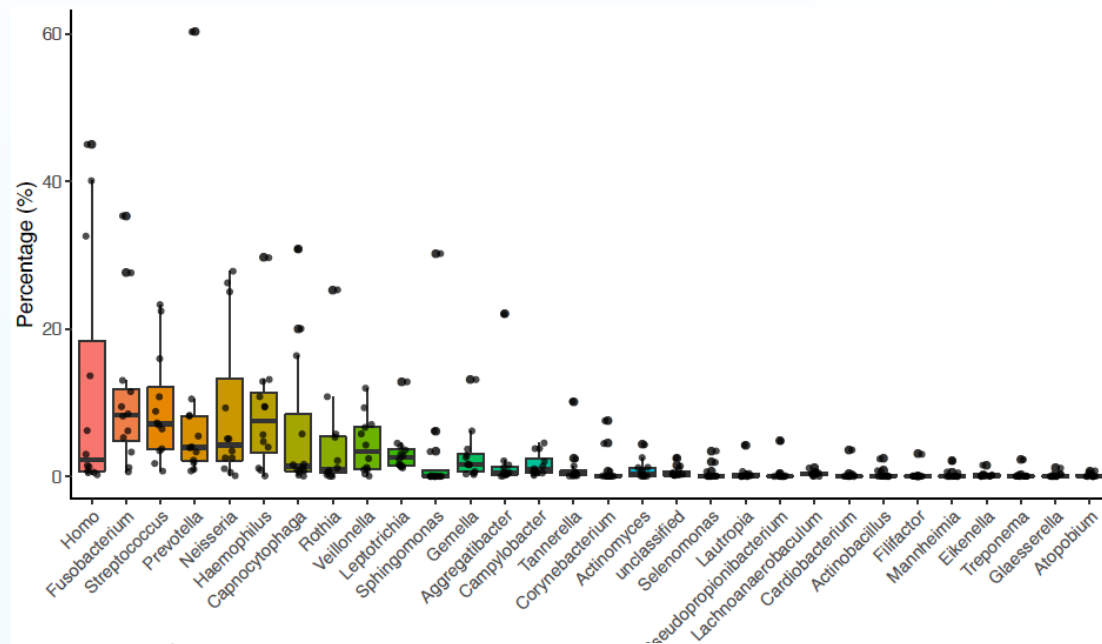


箱线图展示最高丰度的30个属。按门着色。同时角上有门水平箱线图。属和门水平丰度计算采用有参比对，85%相似度，65%覆盖度的阈值。未分类的属显示更高水平标注了星号。

3StatPlot.sh - 物种Kraken2 - 箱线图

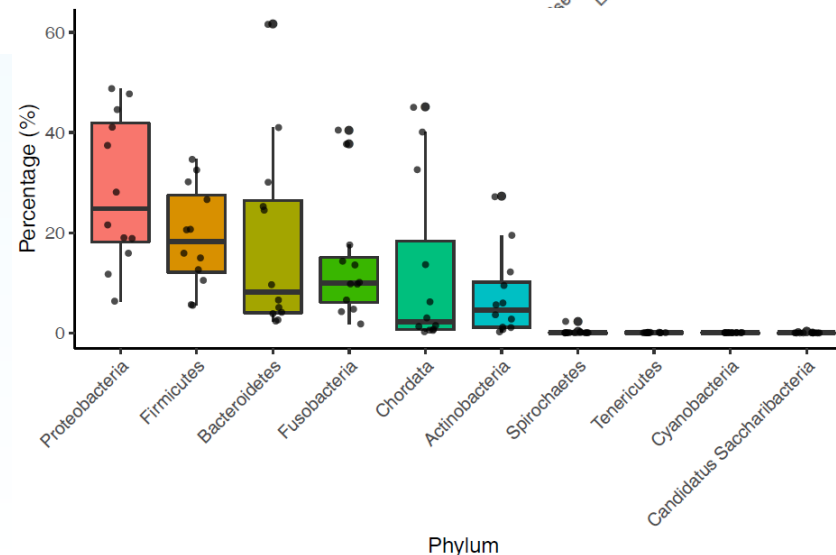
绘制属水平Top30箱线图

```
Rscript ${db}/script/metaphlan_boxplot.R \  
-i result/kraken2/tax_count.spf \  
-t Genus \  
-n 30 \  
-o result/kraken2/boxplot_Genus
```



绘制门水平Top10箱线图

```
Rscript ${db}/script/metaphlan_boxplot.R \  
-i result/kraken2/tax_count.spf \  
-t Phylum \  
-n 10 -w 4 -e 2.5 \  
-o result/kraken2/boxplot_Phylum
```



Bracken估计Kraken2结果丰度

- -d为数据库与kraken2一致，-i为kraken2报告文件
- r是读长，此处默认为100，通常为150
- l为分类级，本次种级别(S)丰度估计，可选域、门、纲、目、科、属、种：**D,P,C,O,F,G,S**，t是阈值，默认0，越大越可靠，但可用数据越少

tax=P

```
for i in `tail -n+2 result/metadata.txt|cut -f1`;do
```

```
bracken -d ${db}/kraken2/mini \
```

```
    -i temp/kraken2/${i}.report \
```

```
    -r 100 -l ${tax} -t 0 \
```

```
    -o temp/bracken/${i}
```

```
done
```



Bracken结果描述

- 结果描述：共7列，分别为物种名、ID、分类级、读长计数、补充读长计数、**总数、百分比**

name	Taxonomy id	Taxonomy lvl	Kraken assigned reads	Added reads	New est reads	Fraction Total reads
Phixviricota	2732412	P	0	0	0	0
Microsporidia	6029	P	0	0	0	0
Hofneiviricota	2732410	P	0	0	0	0
Proteobacteria	1224	P	1869	14	1883	0.05368
Peploviricota	2731361	P	0	0	0	0
Spirochaetes	203691	P	9	0	9	0.00026
Cercozoa	136419	P	0	0	0	0

Bracken结果整合和筛选

- 样本整合为表，同Kraken2类似
- Microbiome Helper中filter_feature_table.R按出现频率筛选，如1%至少筛选掉全为0的行，默认为20%

```
Rscript ~/db/EasyMicrobiome/script/filter_feature_table.R \  
-i result/kraken2/bracken.${tax}.txt \  
-p 0.01 \  
-o result/kraken2/bracken.${tax}.0.01
```

- # 种水平去除人类P:Chordata,S:Homo sapiens

```
grep -v 'Homo sapiens' result/kraken2/bracken.S.0.01 \  
> result/kraken2/bracken.S.0.01-H
```



3StatPlot.sh - 物种Kraken2 – Alpha/Beta多样性

- 多样性计算需要抽平并计算alpha多样性, -d指定最小样本量, 默认0为最小值, 抽平文件bracken.S.norm, alpha多样性bracken.S.alpha
tax=S

```
Rscript $sd/otutab_rare.R \  
--input result/kraken2/bracken.${tax}.txt \  
--depth 0 --seed 1 \  
--normalize result/kraken2/bracken.${tax}.norm \  
--output result/kraken2/bracken.${tax}.alpha
```

- Beta多样性距离矩阵计算

```
usearch -beta_div result/kraken2/bracken.${tax}.norm \  
-filename_prefix result/kraken2/beta/
```



3StatPlot.sh - 物种Kraken2 –Beta多样性

- Bracken的Reads更多， Alpha多样性丰富度大于Kraken2的结果
- Beta多样性可选距离有 bray_curtis, euclidean, jaccard, manhattan

dis=bray_curtis

```
Rscript $sd/beta_pcoa.R \  
--input result/kraken2/beta/${dis}.txt \  
--design result/metadata.txt \  
--group Group \  
--width 89 --height 59 \  
--output result/kraken2/pcoa.${dis}.pdf
```

统计结果文件:

beta_pcoa_stat.txt

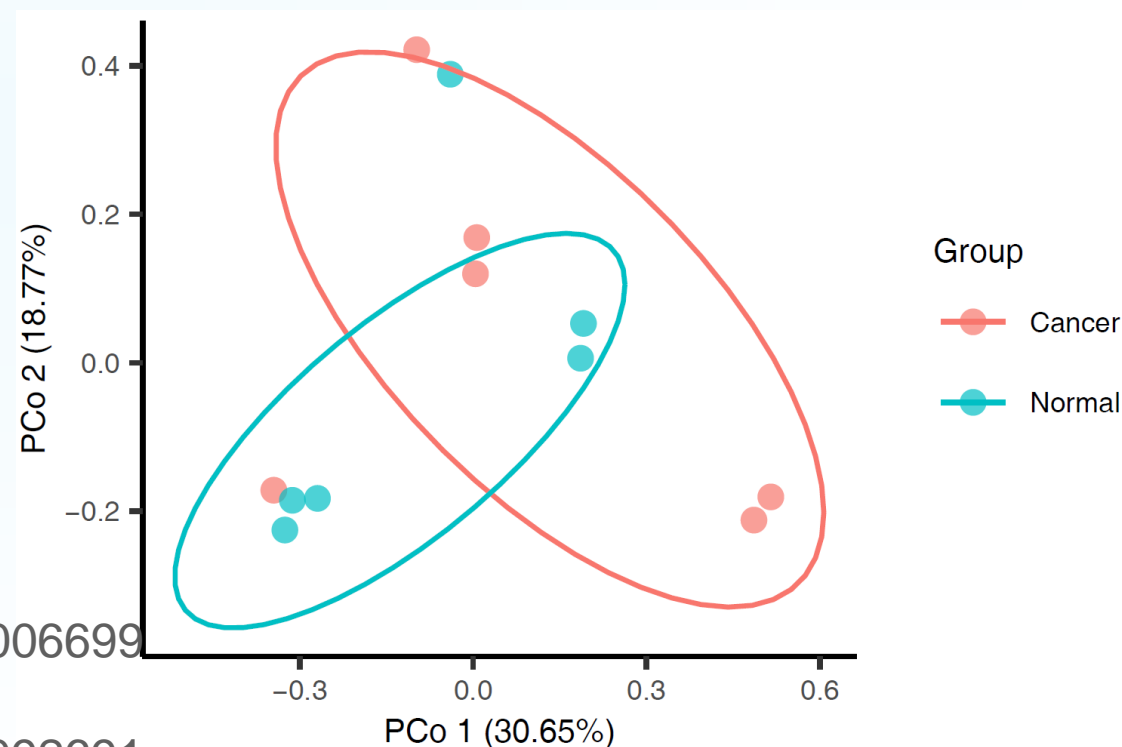
P值有波动但比较稳定

Sun Jan 03 16:19:07 2021

Cancer Normal 0.300669933006699

Sun Jan 03 17:55:04 2021

Cancer Normal 0.309269073092691



3StatPlot.sh - 物种Kraken2 – 堆叠柱状图

- 以门(P)/种(S)水平为例，结果包括output.sample/group.pdf两个文件

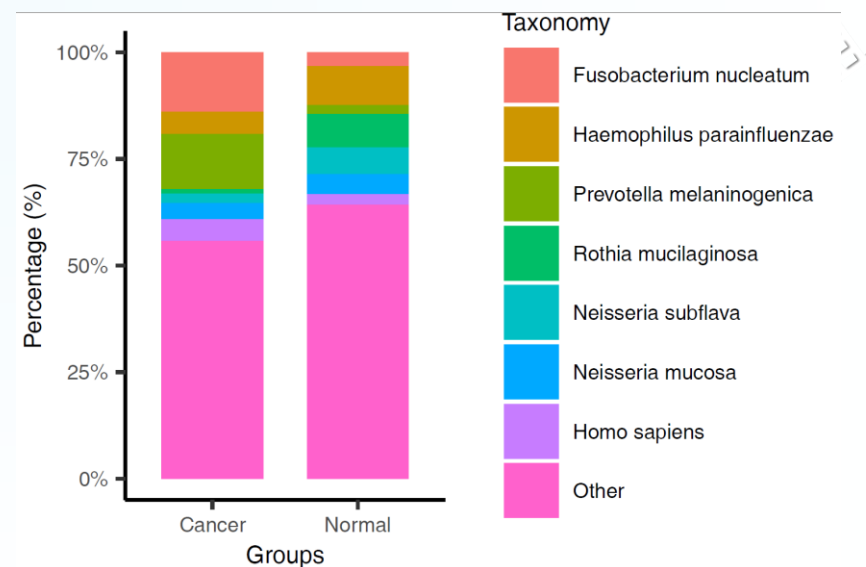
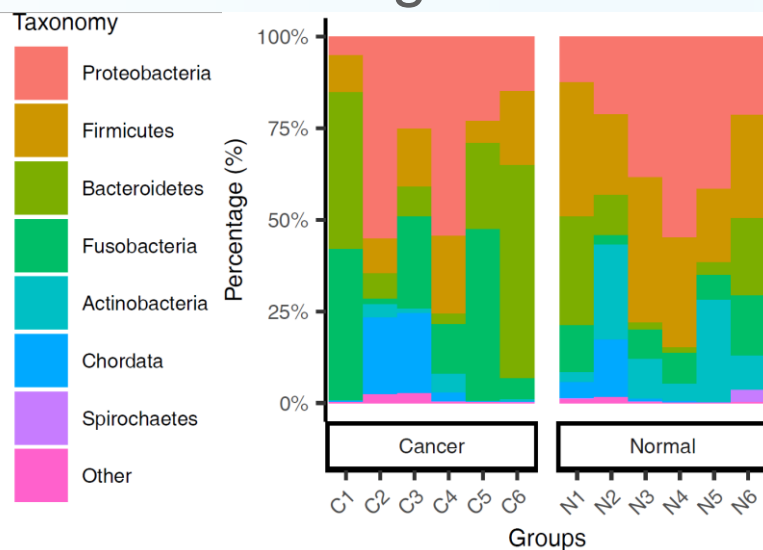
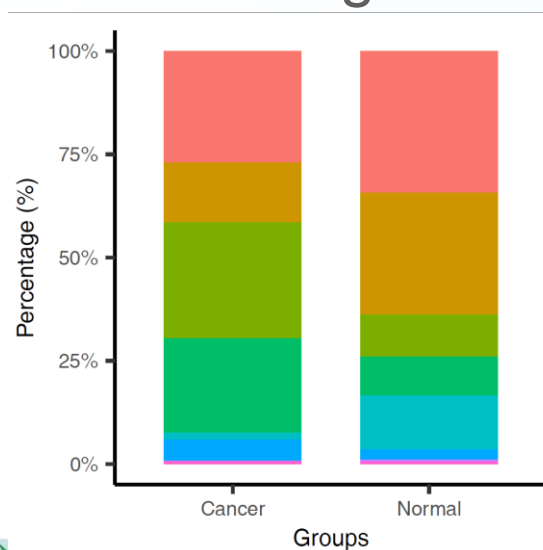
tax=S

```
Rscript ${sd}//tax_stackplot.R \
```

```
--input result/kraken2/bracken.${tax}.txt --design result/metadata.txt \
```

```
--group Group --output result/kraken2/bracken.${tax}.stackplot \
```

```
--legend 8 --width 89 --height 59
```



- 物种注释(界门纲目科属种)类似于地址，表明物种间关系远近，不同分类数据库结果差别较大，分类方法常用比对Best hit和精确匹配LCA；
- Kraken2运行速度快、数据库可大可小、结果为计数型counts格式，可抽平开展多样性分析，也可绘制各级热图和箱线图进行整体描述；
- 下游有Bracken2丰度重估计，KrakenTools格式转换和筛选；
- 物种组成表下游STAMP/LEfSe和扩增子课程R语言多样性分析；
- 常用的物种可视化工具有GranPhlAn(公认最美，使用复杂、输入文件准备复杂)、microbiomeViz(R中重复LEfSe结果)、Metacoder(非常有特色)和Krona(跨平台、交互式网页结果)等多种风格可选



- [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

附. 去宿主-取非植物33090和动物(人)33208序列

```
parallel -j 3 \  
  "extract_kraken_reads.py \  
  -k temp/kraken2/{1}.output \  
  -r temp/kraken2/{1}.report \  
  -1 temp/qc/{1}_1_kneaddata_paired_1.fastq \  
  -2 temp/qc/{1}_1_kneaddata_paired_2.fastq \  
  -t 33090 33208 --include-children --exclude \  
  --max 20000000 --fastq-output \  
  -o temp/kraken2_qc/{1}_1.fq \  
  -o2 temp/kraken2_qc/{1}_2.fq" \  
 ::: `tail -n+2 result/metadata.txt|cut -f1`
```

