宏基因组分析

主要步骤常用软件和数据库

软件安装
Conda

分箱
MetaWRAP

质控去宿主
KneadData

KEGG
COG CAZy

CentOS
Ubuntu

组装
Megahit

功能注释
HUMAnN2

基因预测
Prokka

物种注释
Kraken2

非冗余基因集
Cd-hit
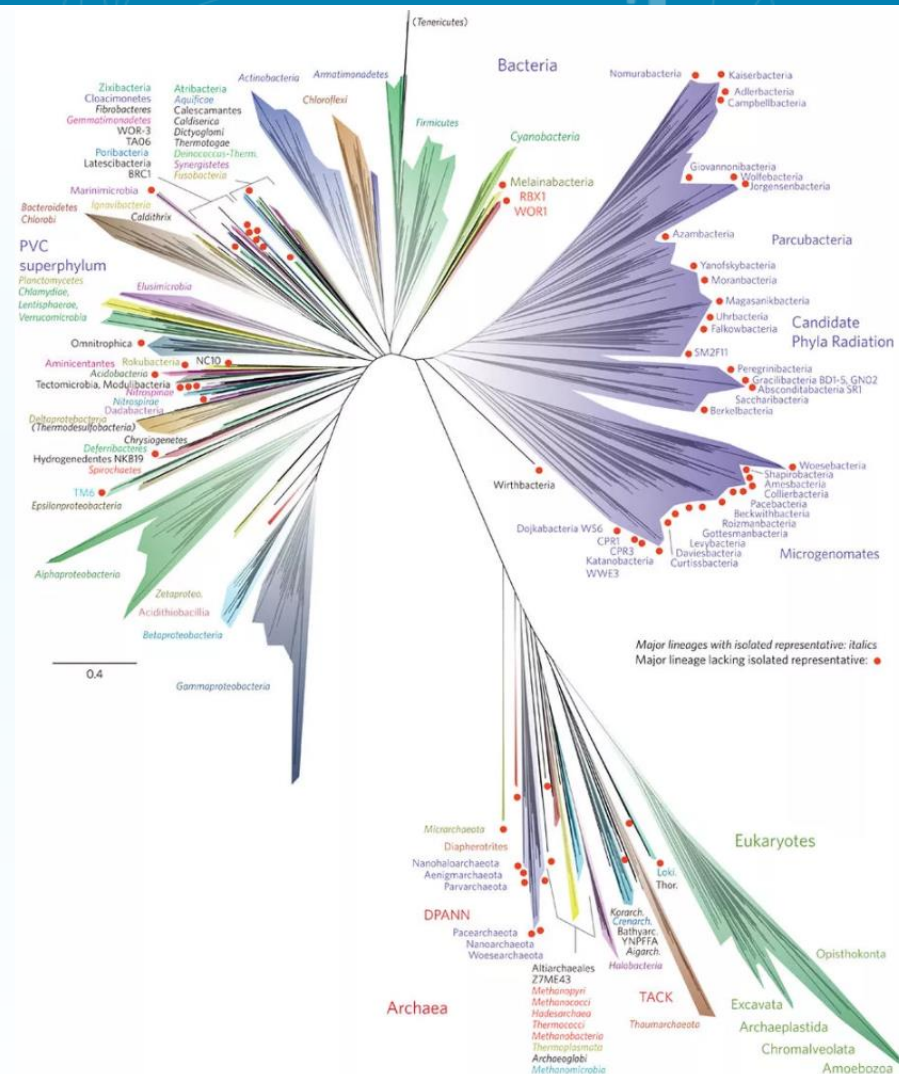
- 分类学(taxonomy)：是一门研究生物类群间的异同以及异同程度，阐明生物间的亲缘关系、进化过程和发展规律的科学。

- 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)

Hug, L. A. et al. A new view of the tree of life. **Nature Microbiology** 1, 16048, doi:10.1038/nmicrobiol.2016.48 (2016).

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

3

- 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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

https://baike.baidu.com/item/%E7%95%8C%E9%97%A8%E7%BA%B2%E7%9B%AE%E7%A7%91%E5%B1%9E%E7%A7%8D

- 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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

o 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。

o 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。

易汉博基因科技(北京)有限公司
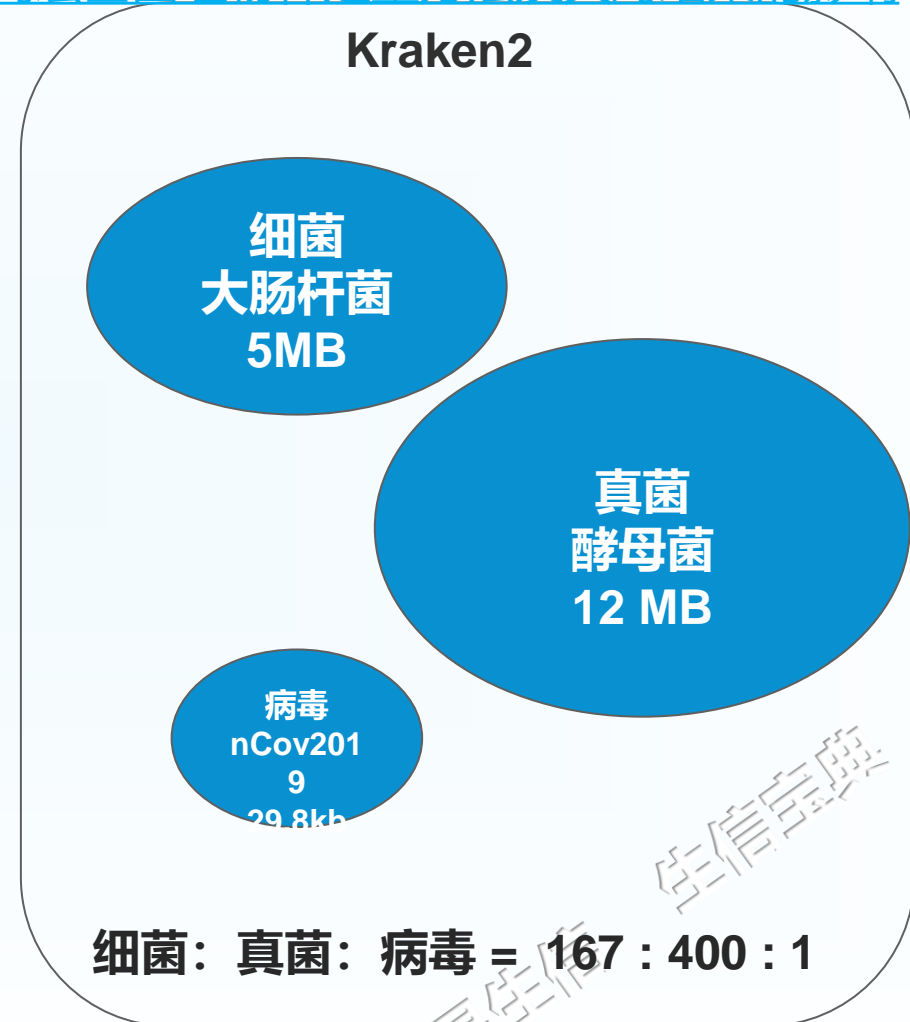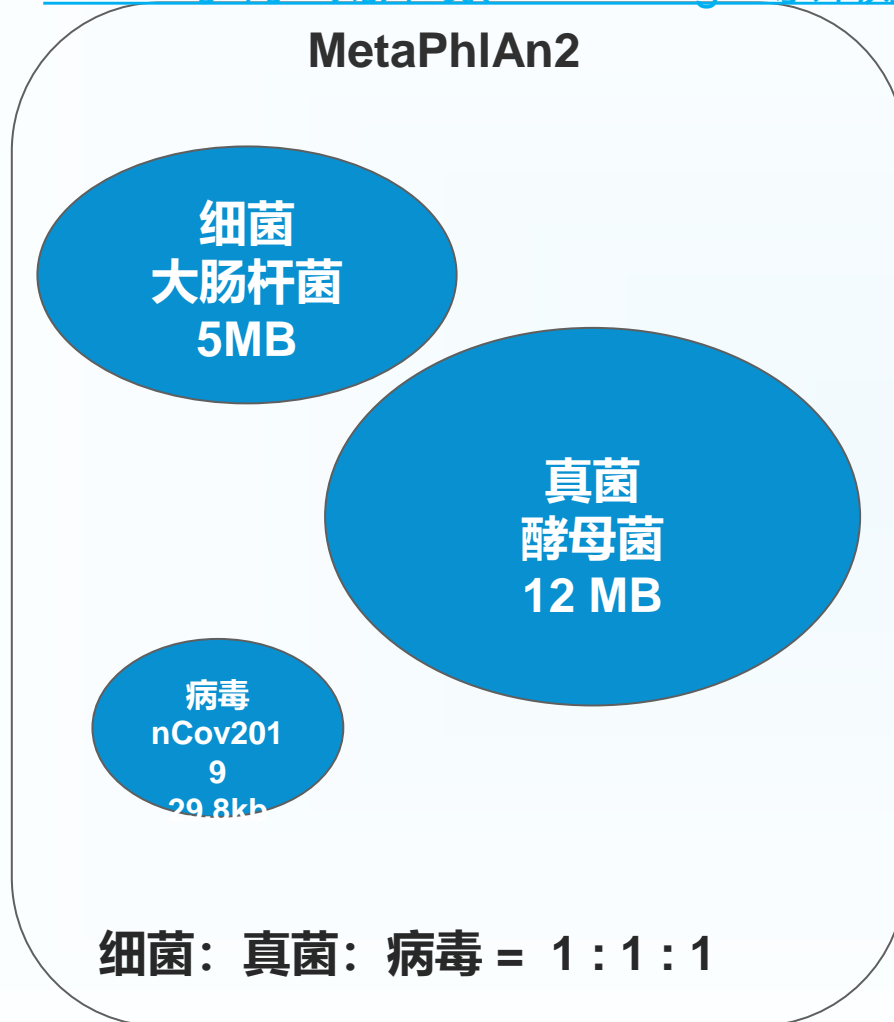EHBIO Gene Technology (Beijing) co., LTD
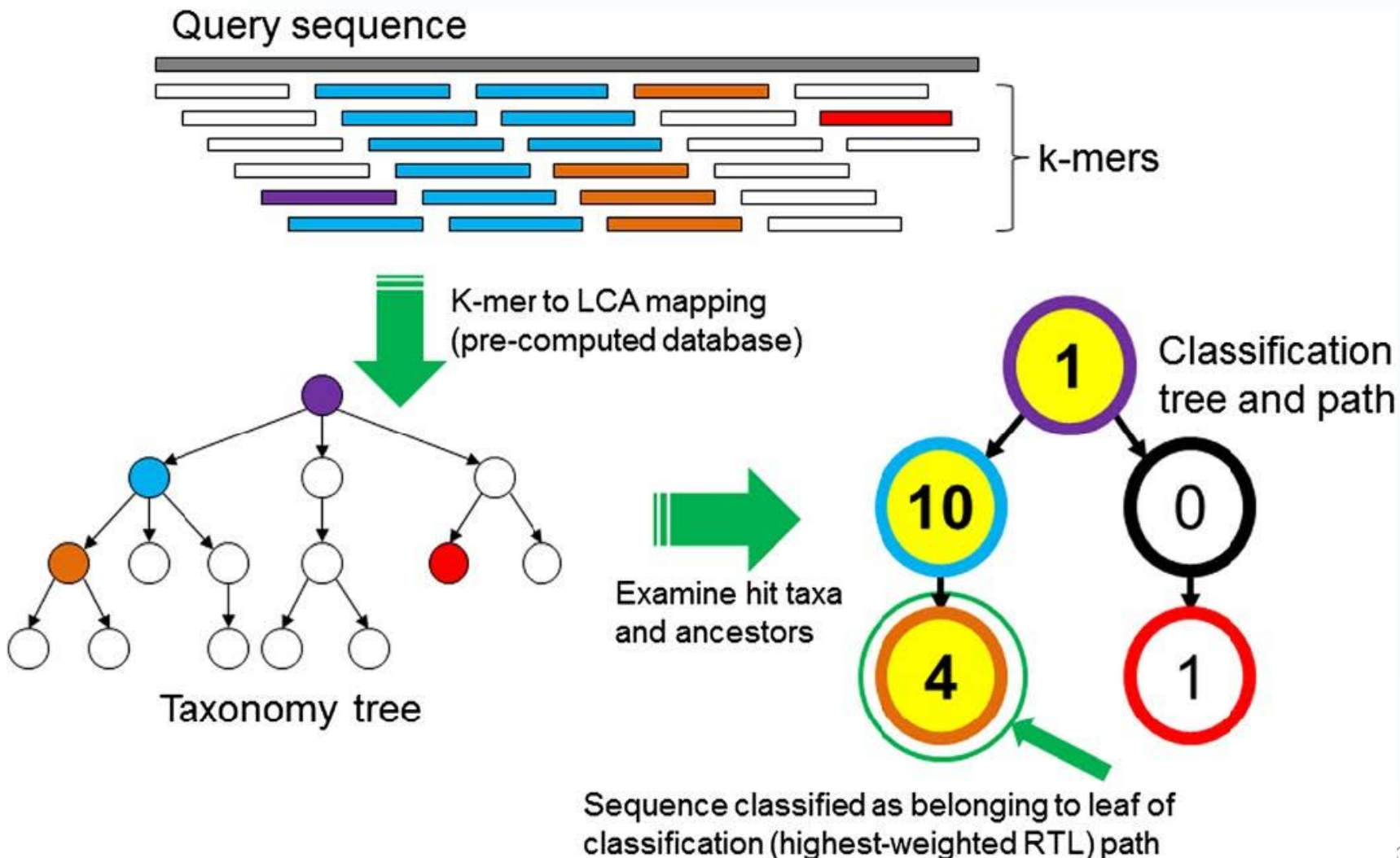
Lower Common Ancestor
最低共同祖先

软件默认认阈值为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](#)

Nature子刊：刘洋彧、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

MetaPhlAn2

细菌
大肠杆菌
5MB

真菌
酵母菌
12 MB

病毒
nCov2019
29.8kb

细菌：真菌：病毒 = 1：1：1

Kraken2

细菌
大肠杆菌
5MB

真菌
酵母菌
12 MB

病毒
nCov2019
29.8kb

细菌：真菌：病毒 = 167：400：1

Zheng Sun, Shi Huang, Meng Zhang, Qiyun Zhu, Niina Haiminen, Anna Paola Carrieri, 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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

8

Karken——序列物种分类系统

Kraken
Taxonomic Sequence Classification System

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

CCB » Software » Kraken

ABOUT KRAKEN

Kraken is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies. Previous attempts by other bioinformatics software to accomplish this task have often used sequence alignment or machine learning techniques that were quite slow, leading to the development of less sensitive but much faster abundance estimation programs. Kraken aims to achieve high sensitivity and high speed by utilizing exact alignments of k-mers and a novel classification algorithm.

In its fastest mode of operation, for a simulated metagenome of 100 bp reads, Kraken processed over 4 million reads per minute on a single core, over 900 times faster than Megab estimation program MetaPhlAn. Kraken's accuracy is comparab high precision.

Kraken is written in C++ and Perl, and is designed for use with compiled and run it under the Mac OS.

[HTML] Kraken: ultrafast metagenomic sequence classification using exact alignments
DE Wood, SL Salzberg - Genome biology, 2014 - genomebiology.biomedcentral.com
Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences. Previous programs designed for this task have been relatively slow and computationally expensive, forcing researchers to use faster abundance estimation programs, which only classify small subsets of metagenomic data. Using exact alignment of k-mers, Kraken achieves classification accuracy comparable to the fastest BLAST program. In its fastest mode, Kraken classifies 100 base pair reads at a rate of over ...
☆ 〃 Cited by 2193   Related articles   All 17 versions 》

DOWNLOADS AND DOCUMENTS

- Kraken 2 source code release

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

Wood D E, Salzberg S L. Kraken: ultrafast metagenomic sequence classification using exact alignments[J]. Genome biology, 2014, 15(3): R46. http://ccb.jhu.edu/software/kraken

易生信

o https://github.com/fbreitwieser/krakenuniq

KrakenUniq: confident and fast metagenomics classification using unique k-mer counts

False-positive identifications are a significant problem in metagenomics classification. KrakenUniq (formerly KrakenHLL) is a novel metagenomics classifier that combines the fast k-mer-based classification of Kraken with an efficient algorithm for assessing the coverage of unique k-mers found in each species in a dataset. On various test datasets, KrakenUniq gives better recall and precision than other methods and effectively classifies and distinguishes pathogens with low abundance from false positives in infectious disease samples. By using the probabilistic cardinality estimator HyperLogLog, KrakenUniq runs as fast as Kraken and requires little additional memory.

Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biology* 19, 198, doi:10.1186/s13059-018-1568-0 (2018).

Kraken2

○ 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

Derrick E. Wood, Jennifer Lu & Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257, doi:10.1186/s13059-019-1891-0 (**2019**).

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

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

Cell：20种宏基因组学物种分类工具大比拼

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD
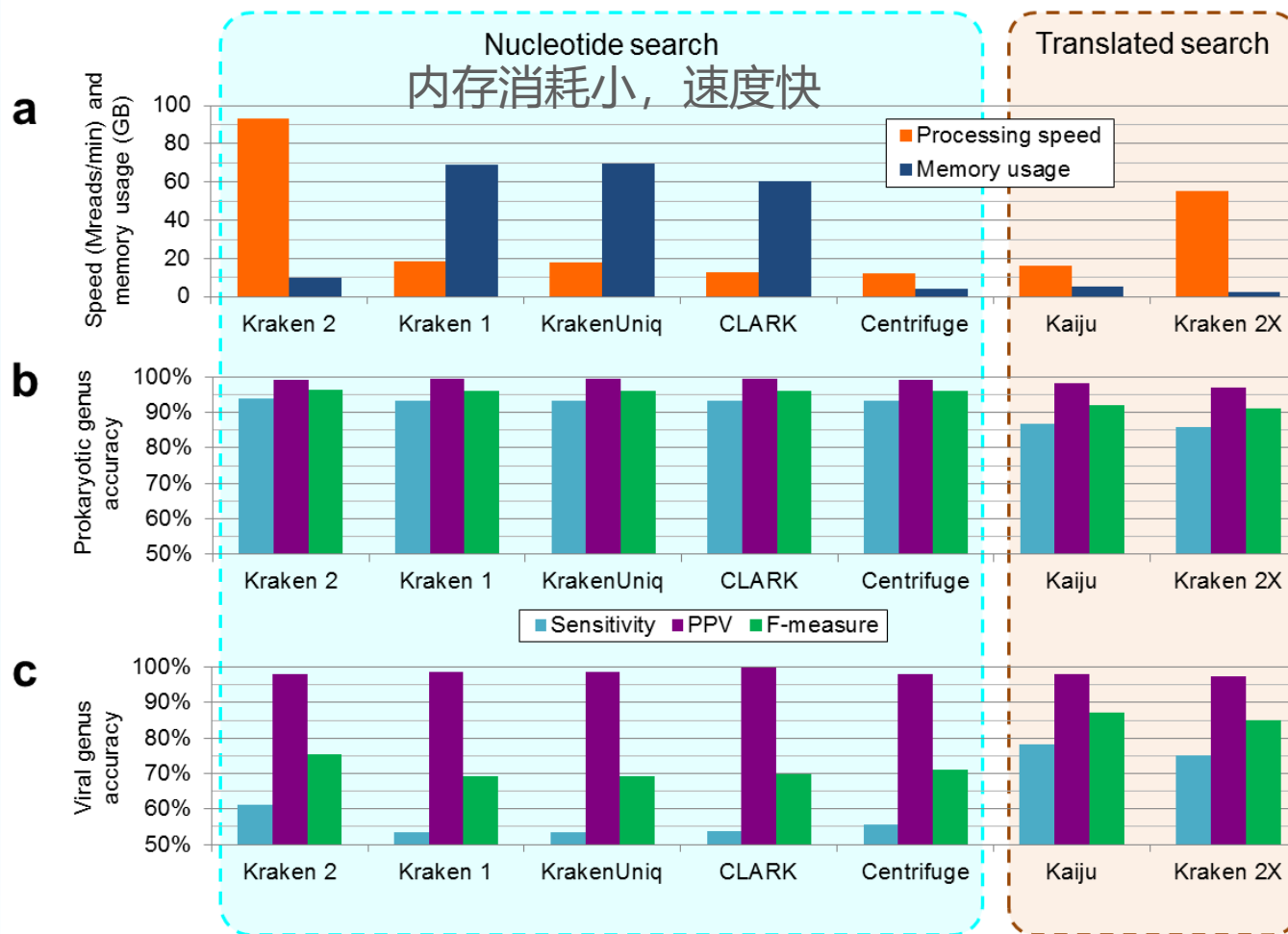
○ 安装基于LCA算法的物种注释软件Kraken2

conda install kraken2

○ 下载数据库，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

○ 小内存/演示使用迷你库(PlusPFP-8)，包括标准+原生动物+真菌+植物及Bracken2索引，仅8G，详见：https://benlangmead.github.io/aws-indexes/k2

○ 新版收录更高的序列多样性，注释比例更高。如以测序数据C1样本的65K条双端序列为例，8G数据库仅从2020/12更新为2021/05，序列注释率从39%上升为52%，版本2.0.9和2.1.1对结果无影响

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

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/mini --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
```

○ 安装

conda install krakentools -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
```

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	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces						9	388	57	25	27	49	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces gaoshouyii									1	0	0	0
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces hongkongensis									0	0	2	7
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces howellii									0	2	0	0
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces israelii									0	0	0	0
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces oris									0	319	29	2
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces pacaensis									8	0	3	13
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces radicidentis									0	2	2	0
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces slackii									0	3	1	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces sp. Chiba101									0	1	1	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces sp. dk561									0	0	0	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces sp. oral taxon 414									0	4		
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces sp. oral taxon 897									0	0		
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces sp. Z16									0	3	0	
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actinomyces|s__Actinomyces viscosus									0	37	13	0

- 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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

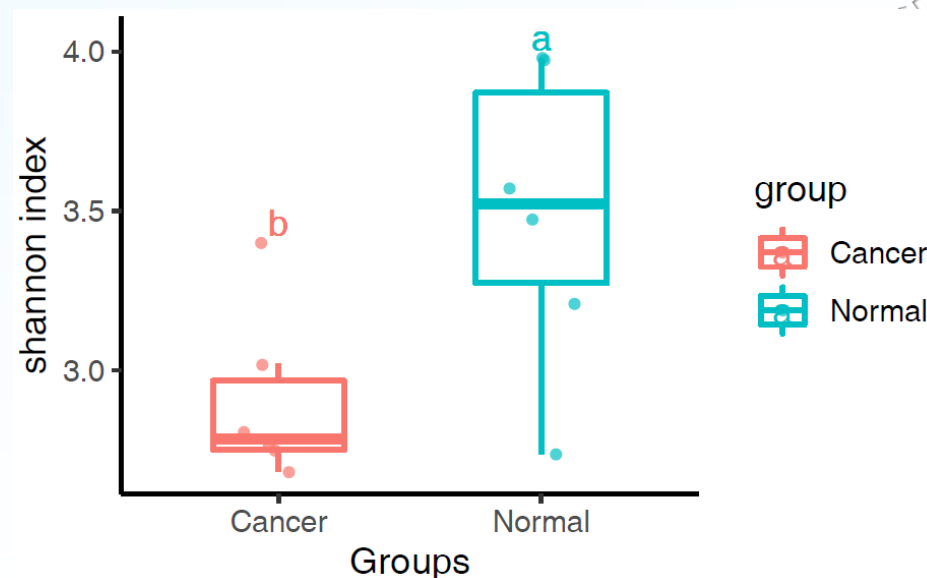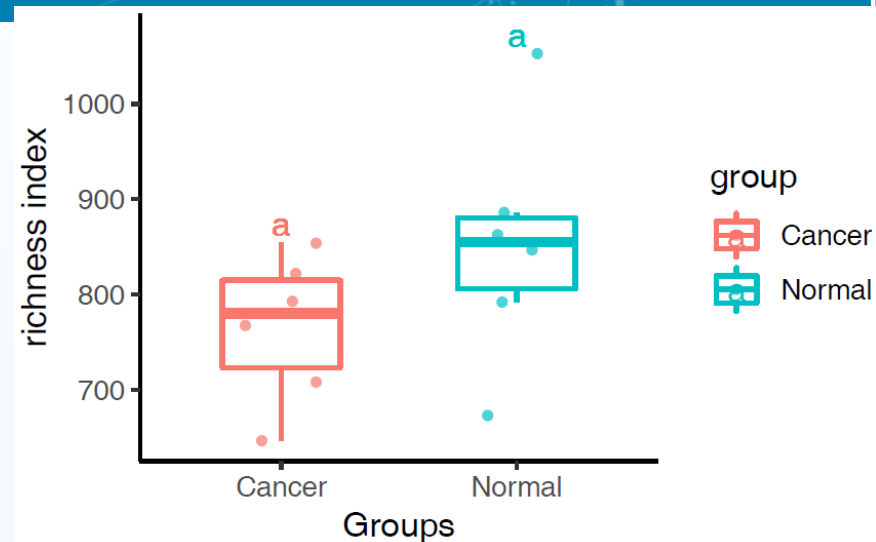microbiomeViz：绘制lefse结果中Cladogram

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
```
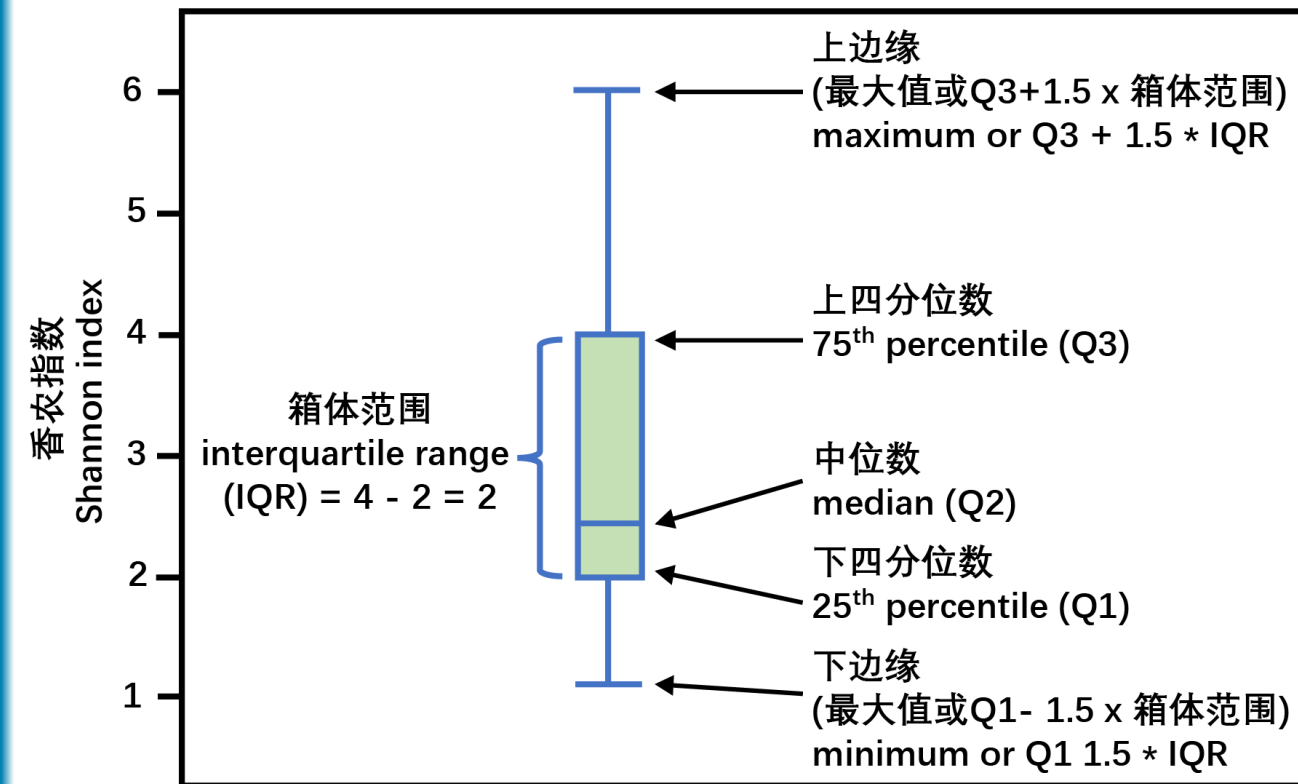
绘制箱线图,可选richness/chao1/shannon…

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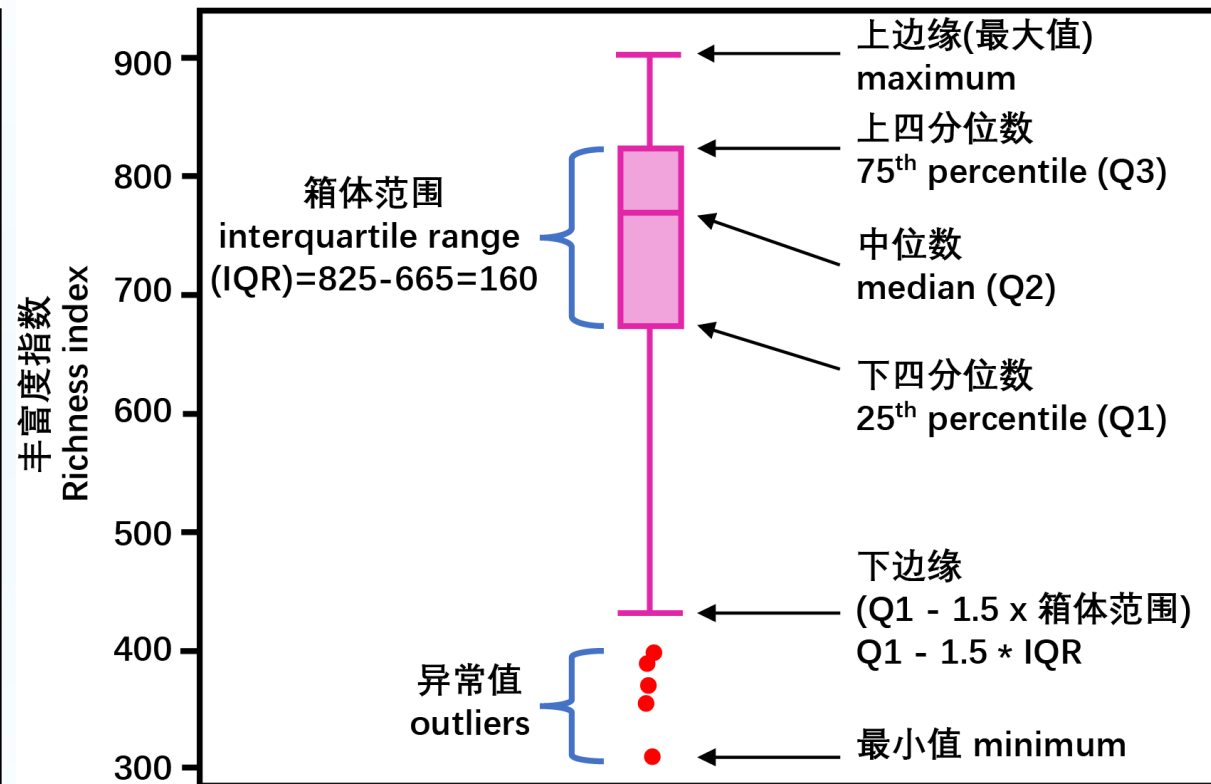
Alpha多样性香农指数箱线图(Boxplot of Shannon index)

上边缘
(最大值或Q3+1.5 x 箱体范围)
maximum or Q3 + 1.5 * IQR

上四分位数
75th percentile (Q3)

箱体范围
interquartile range
(IQR) = 4 - 2 = 2

中位数
median (Q2)

下四分位数
25th percentile (Q1)

下边缘
(最大值或Q1- 1.5 x 箱体范围)
minimum or Q1 1.5 * IQR

香农指数
Shannon index

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

上边缘(最大值)
maximum

上四分位数
75th percentile (Q3)

箱体范围
interquartile range
(IQR)=825-665=160

中位数
median (Q2)

下四分位数
25th percentile (Q1)

下边缘
(Q1 - 1.5 x 箱体范围)
Q1 - 1.5 * IQR

异常值
outliers

最小值 minimum

丰富度指数
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*

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

《微生物组数据分析与可视化实战》

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
```
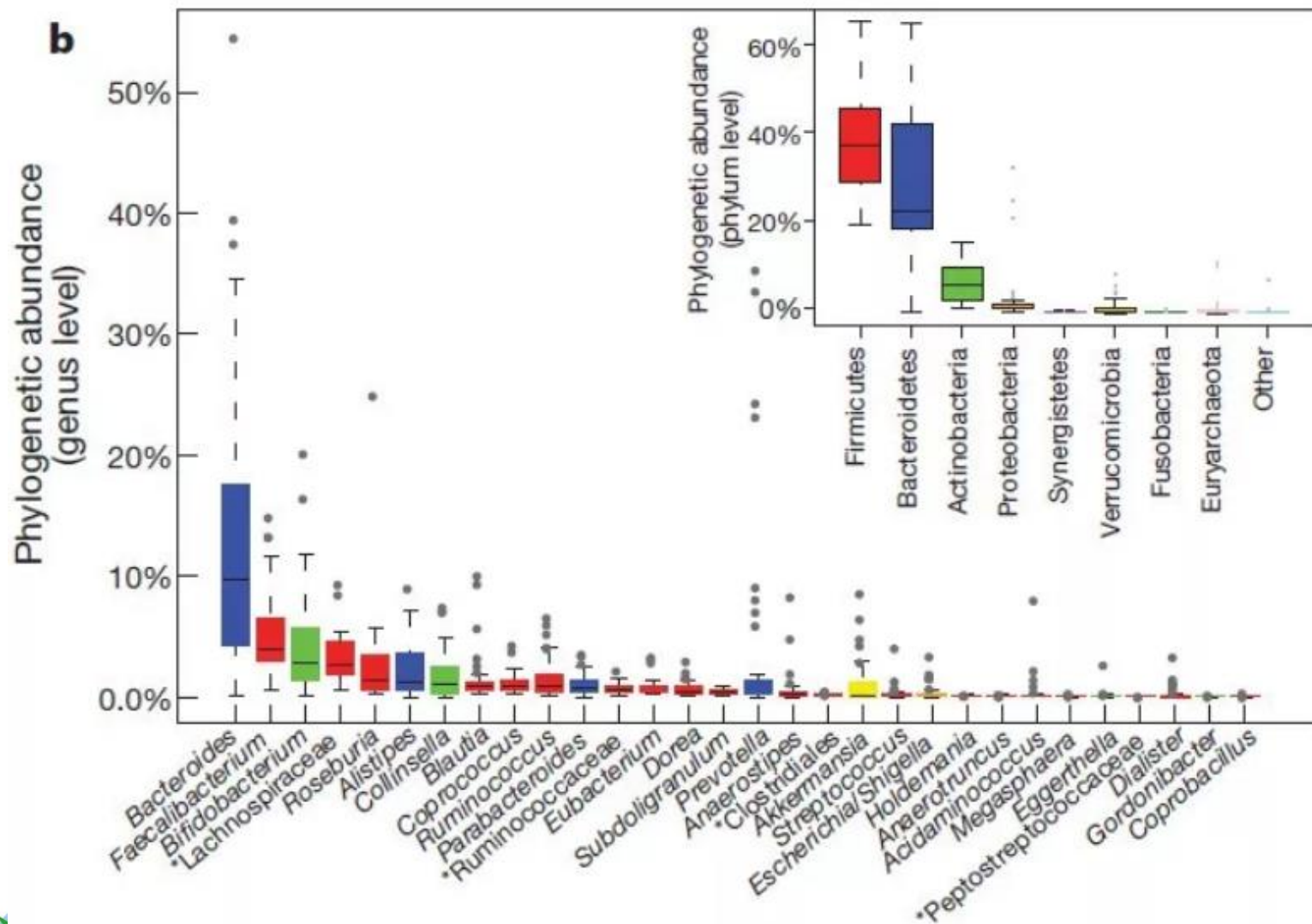

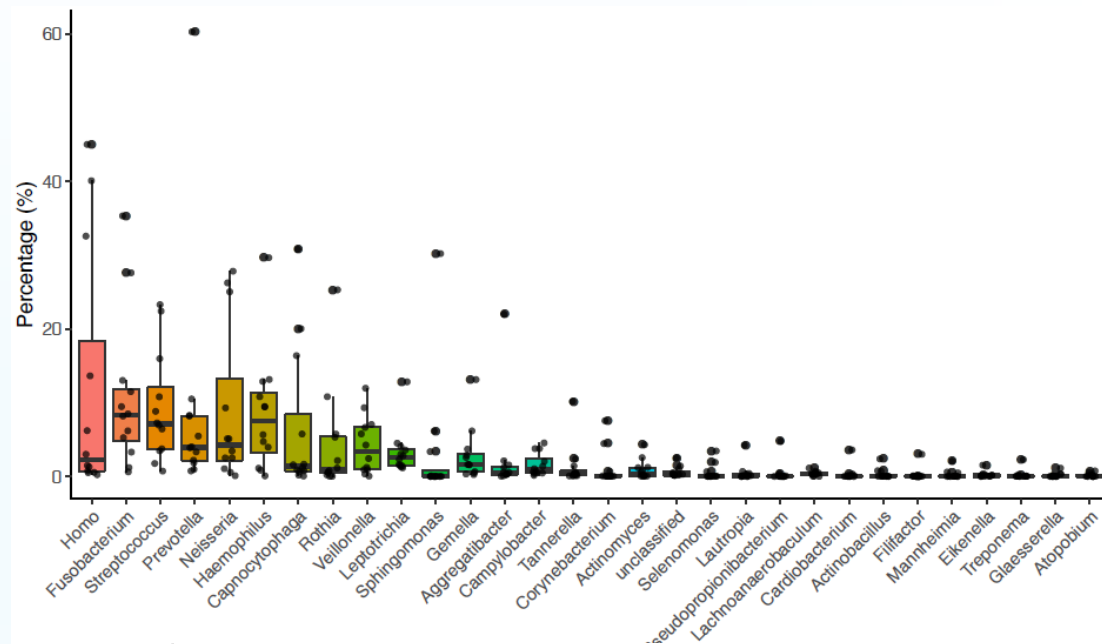Top 25 Genus
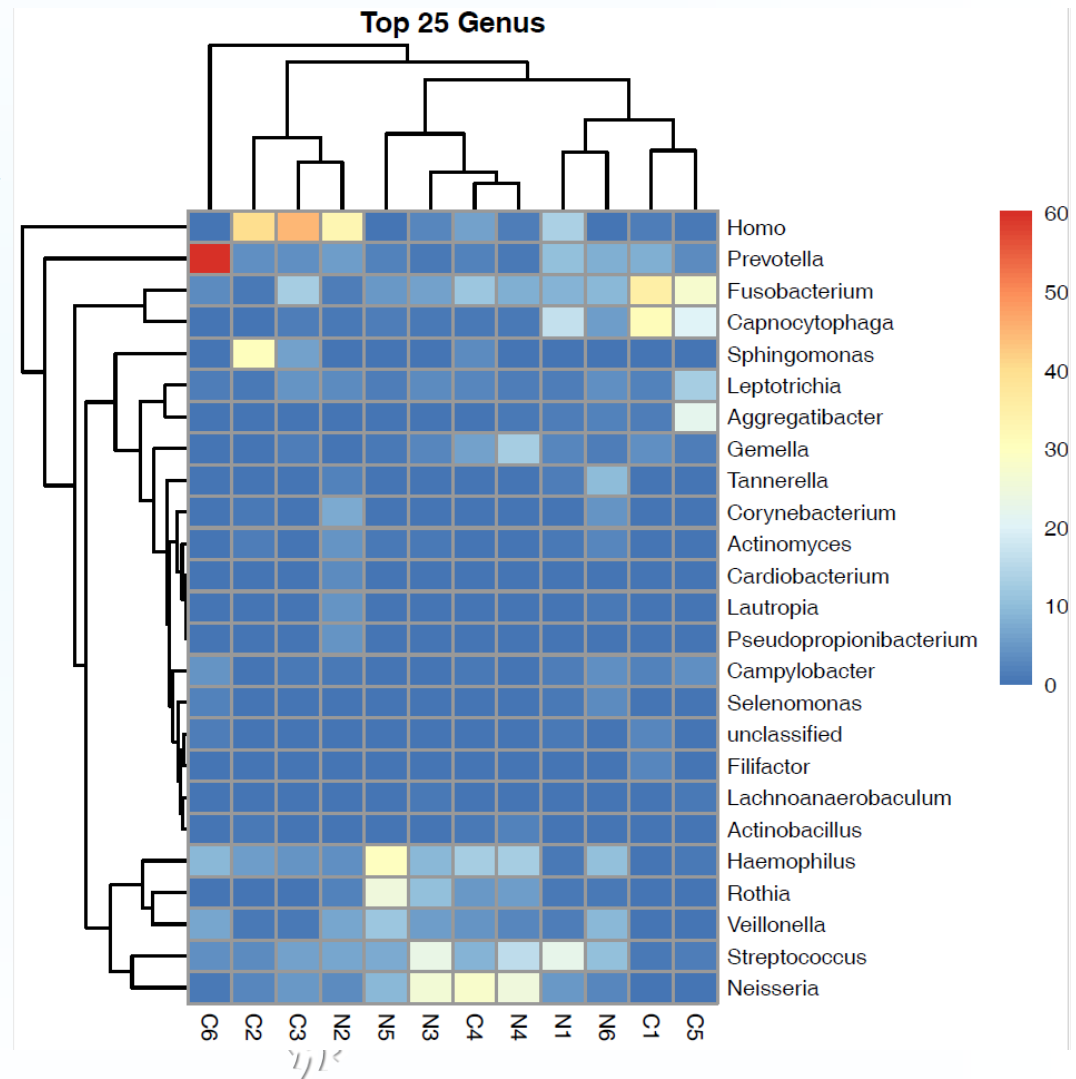
箱线图展示最高丰度的30个属。按门着色。同时角上有门水平箱线图。属和门水平丰度计算采用有参比对，85%相似度，65%覆盖度的阈值。未分类的属显示更高水平标注了星号。

○ # 绘制属水平Top30箱线图

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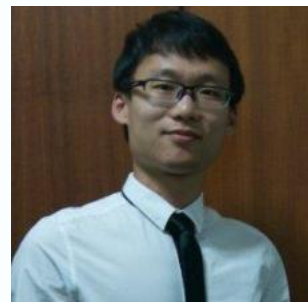 -t Genus \
  -n 30 \
  -o result/kraken2/boxplot_Genus

○ # 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
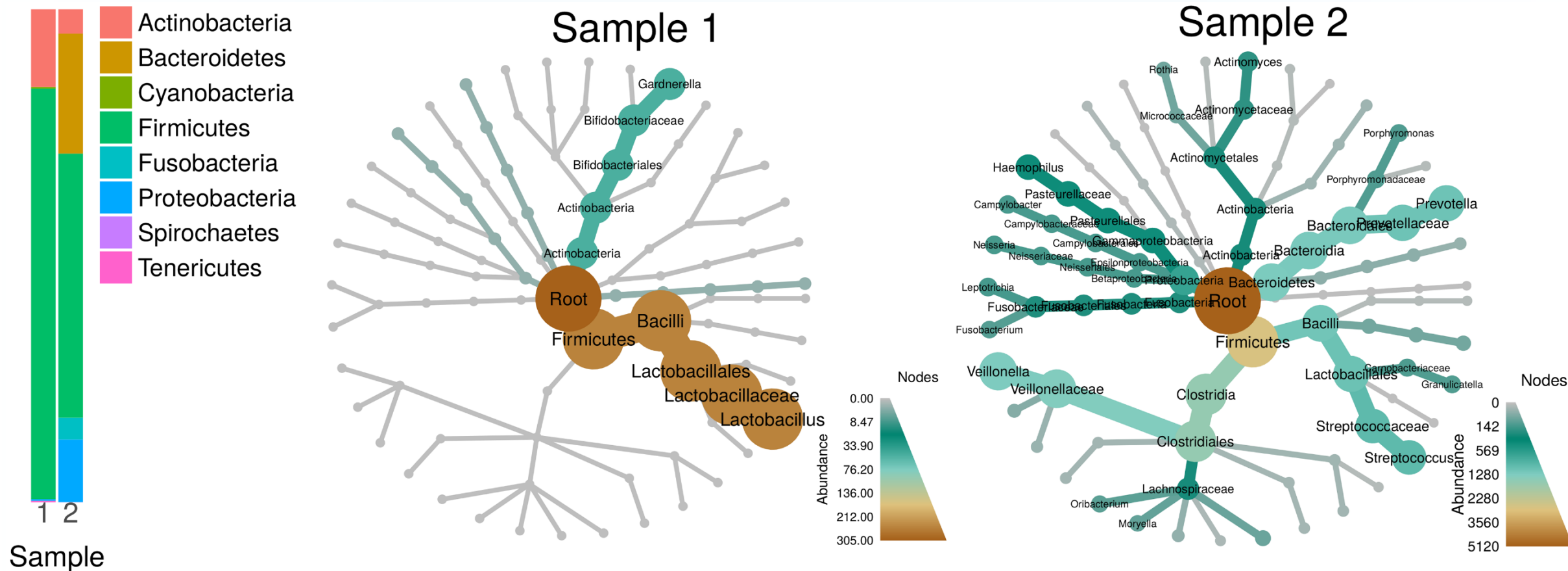

易汉博基因科技（北京）有限公司
EHBIO Gene Technology (Beijing) co., LTD

易生信，毕生缘；培训版权所有。

22
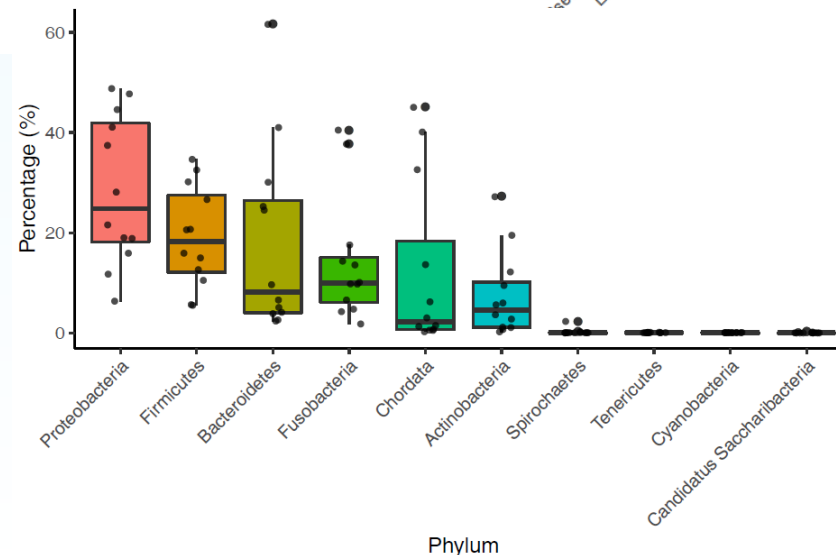
microbiomeViz：绘制lefse结果中Cladogram

- 作者：李陈浩，新加坡基因组所，博士在读

- 博客主页 https://lchblogs.netlify.com/

- GitHub https://github.com/lch14forever/microbiomeViz

- 两步完成LEfSe的Cladogram
  按平均相对丰度绘制树枝结点大小
  按差异类别着背景色并添加标签

- 详细教程见右下角文章或官网
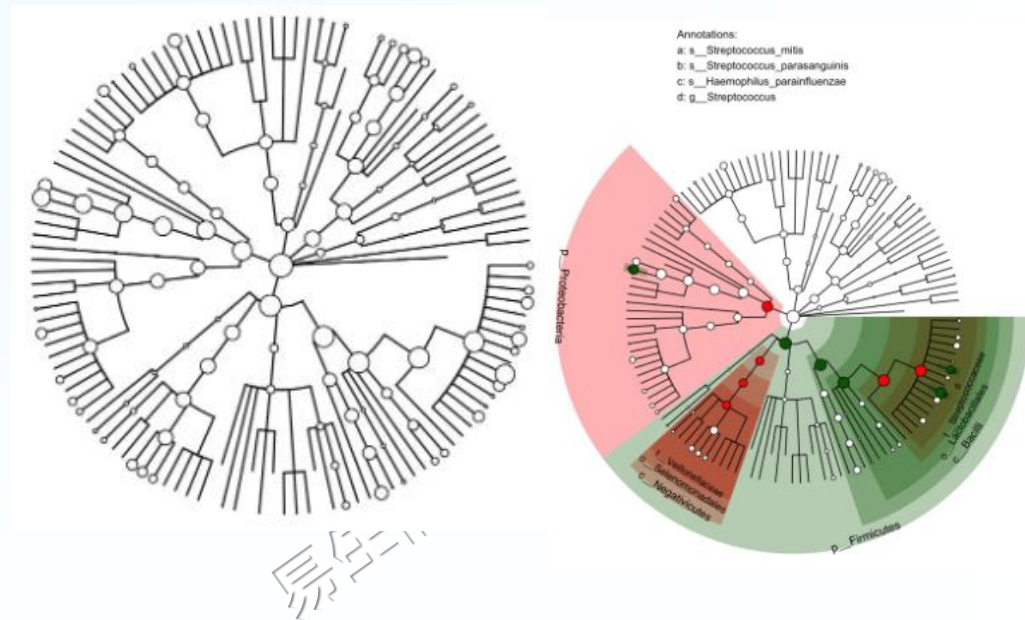
Annotations:
a: s__Streptococcus_mitis
b: s__Streptococcus_parasanguinis
c: s__Haemophilus_parainfluenzae
d: g__Streptococcus

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

microbiomeViz：绘制lefse结果中Cladogram

Foster ZSL, Sharpton TJ, Grünwald NJ (2017) Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol 13(2): e1005404. https://doi.org/10.1371/journal.pcbi.1005404
主页：https://grunwaldlab.github.io/metacoder_documentation/index.html
应用：Gut：人体最初的微生物起源与生殖健康

Krona绘制物种或功能组成圈图

- ○ Krona采用多图层饼图，交互式探索层级数据。软件同时支持Linux命令行脚本和图片界面下Excel插件+模板，方便所有用户使用。

- ○ 官方主页：https://github.com/marbl/Krona/wiki

Krona绘制物种或功能组成圈图
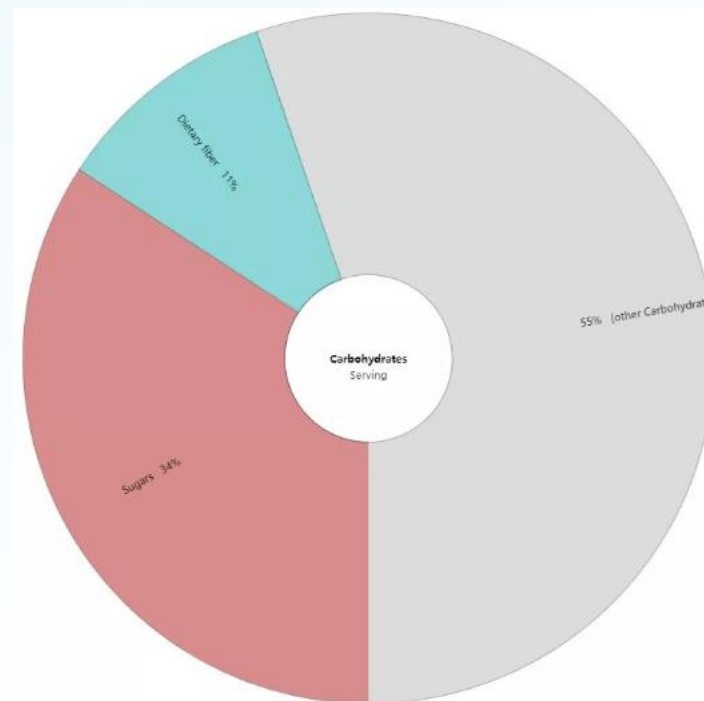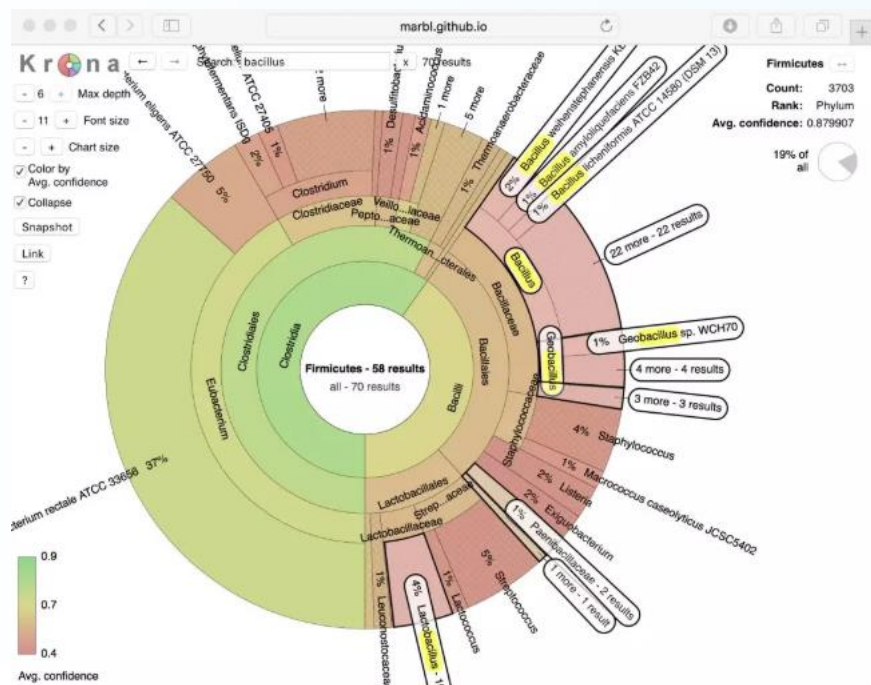
- -d为数据库与kraken2一致，-i为kraken2报告文件

- 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 -I ${tax} -t 0 \
    -o temp/bracken/${i}
done
```

○ 样本整合为表，同Kraken2类似

○ Microbiome Helpler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/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
```

○ 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/
```

易生信，毕生缘；培训版权所有。

○ 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
```
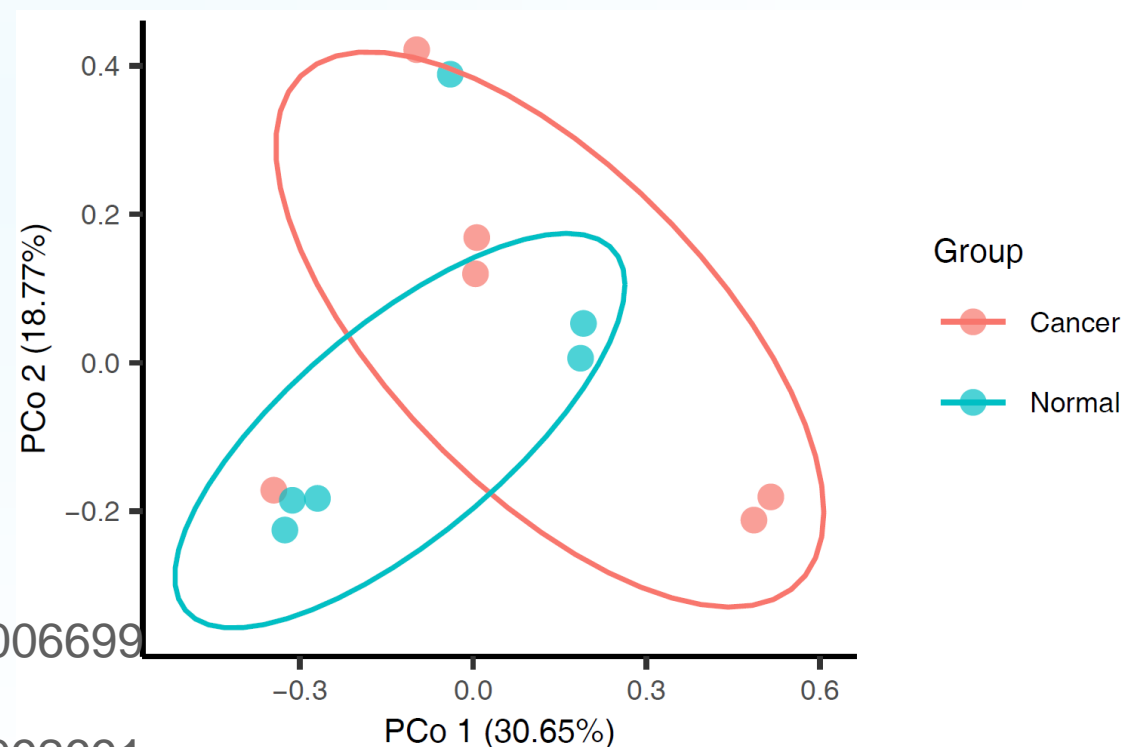

统计结果文件：   Sun Jan 03 16:19:07 2021
beta_pcoa_stat.txt   Cancer   Normal   0.300669933006699
P值有波动但比较稳定   Sun Jan 03 17:55:04 2021
                        Cancer   Normal   0.309269073092691

易生信，毕生缘；培训版权所有。

○ 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
```
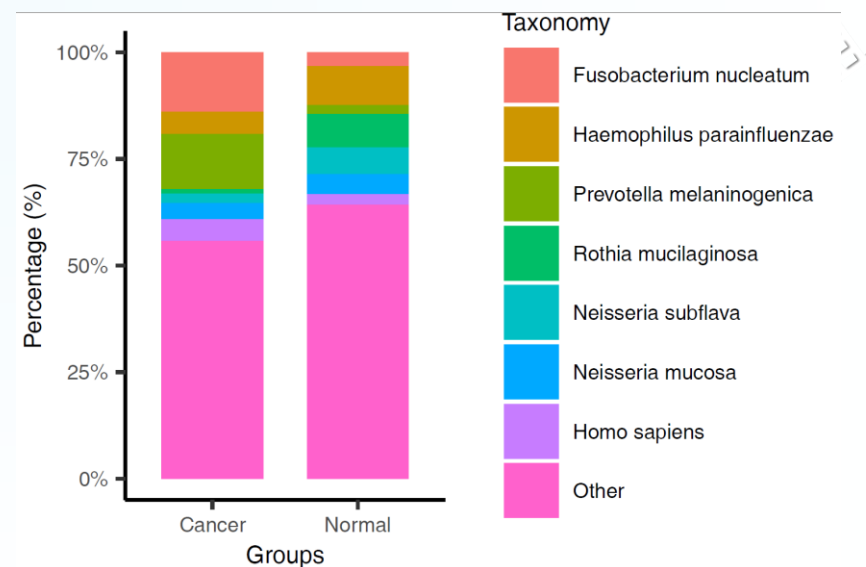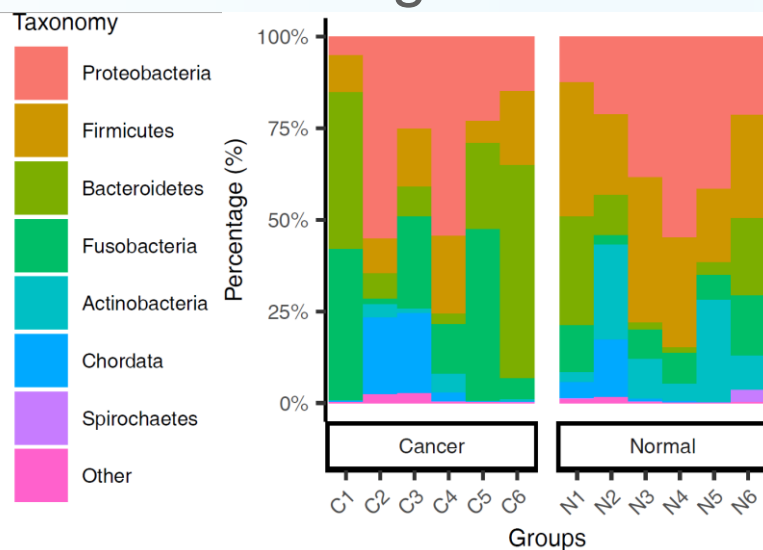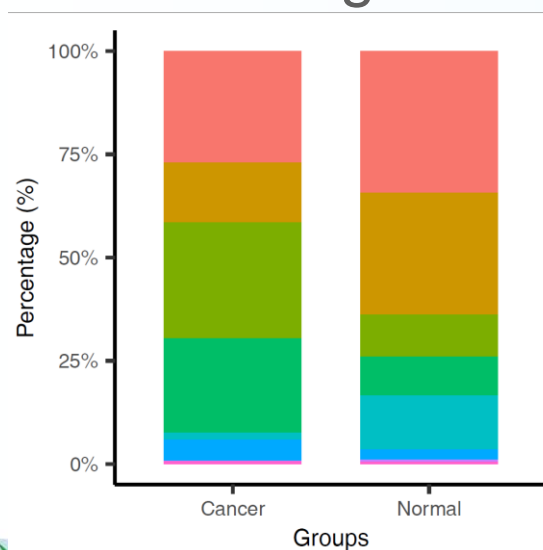
易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

○ 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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

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/**

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

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

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

```
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`
```