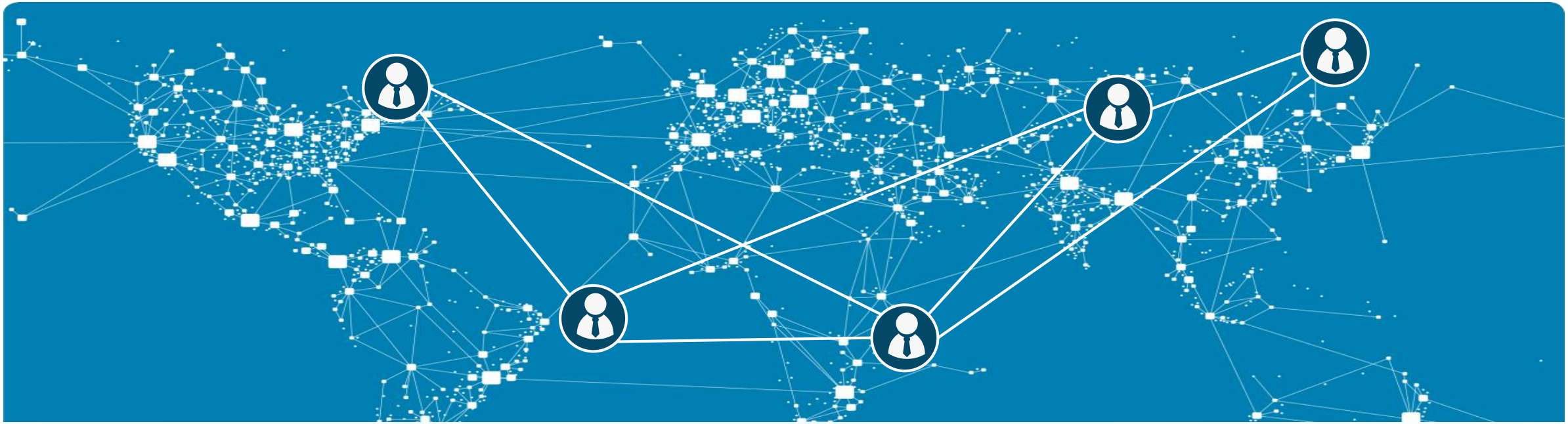


31物种注释Kraken2和可视化

易生信

2019年6月23日

宏基因组无参(*De novo*)分析

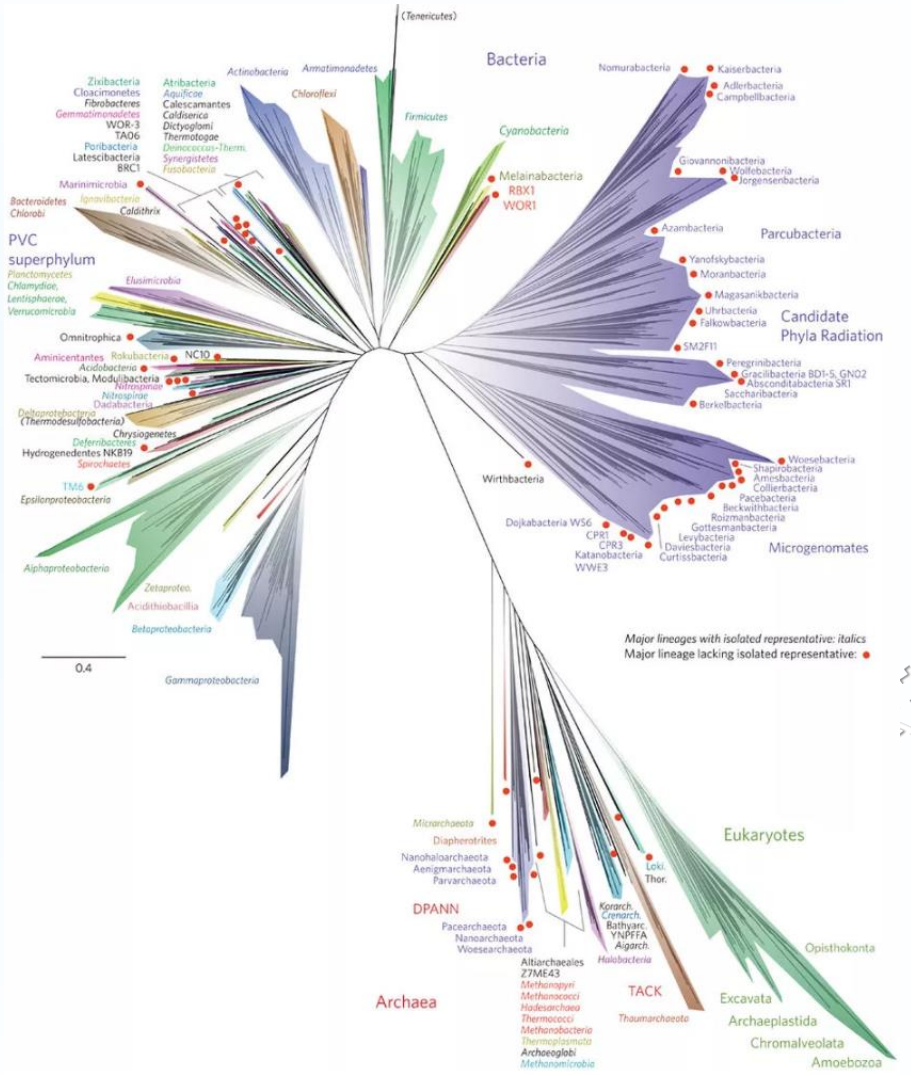
主要步骤常用软件和数据库

目录

目录

物种分类学注释

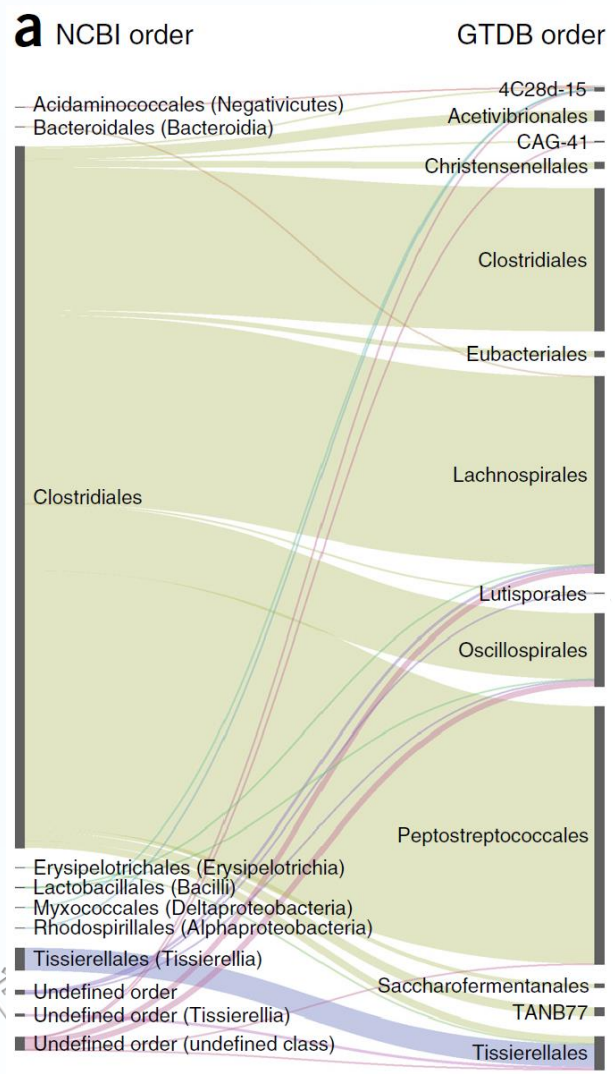
- 分类学(taxonomy)：是一门研究生物类群间的异同以及异同程度，阐明生物间的亲缘关系、进化过程和发展规律的科学。
- 主要分为细菌、古菌和真核生物三大类；
- 常用七级分类法：界(Kingdom)、门(Phylum)、纲(Class)、目(Order)、科（Family）、属（Genus）、种(Species)



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

Nature Biotechnology：根据基因组系统发育进行细菌标准化分类大幅修正生命之树

- 以细菌中普遍存在的120个单拷贝蛋白质为基础，得到基因组分类数据库(GTDB)；
- 涵盖了94759个细菌基因组，在属、种分辨率水平上描述了99个门，其中不可培养细菌占14.4%；
- **58%在NCBI分类系统中已收录基因组的分类地位有变动，例如新系统中变形菌门重新划为6个不重叠的新类群；**
- 一些难以确定分类地位的物种(如不可培养微生物)也被系统的整合了进来。



A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life, 2018, Nature Biotechnology https://www.mr-gut.cn/papers/read/1087194889

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

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

物种注释数据库

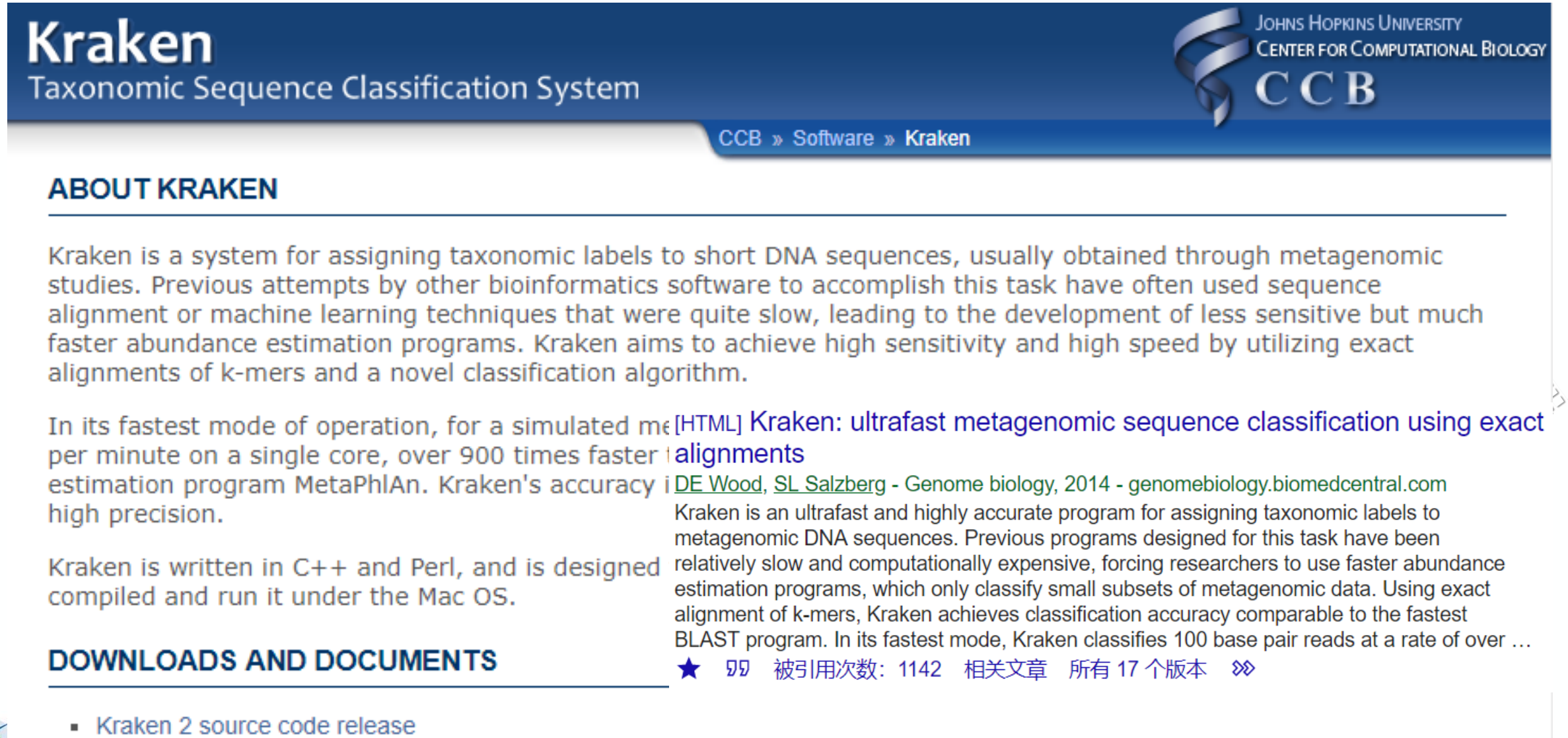
- NCBI——NR非冗余序列，NCBI发布的序列包含物种Taxonomy ID
- MetaPhlAn2——整理已发表基因组Marker基因数据库
- GreenGene——细菌16S物种数据库
- RDP——细菌核糖体数据库
- SILVA——原核、真核核糖体数据库

物种注释方法

- 比对方法：与有物种注释的序列数据库进行比对，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、RDP classifier等。

- LCA(Lower Common Ancestor最低共同祖先)：一般这类方法是基于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、KrakenUniq、Kraken2、Sintax等。

Karken——序列物种分类系统



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

KrakenUniq: 基于唯一K-mer获得特异宏基因组分类

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

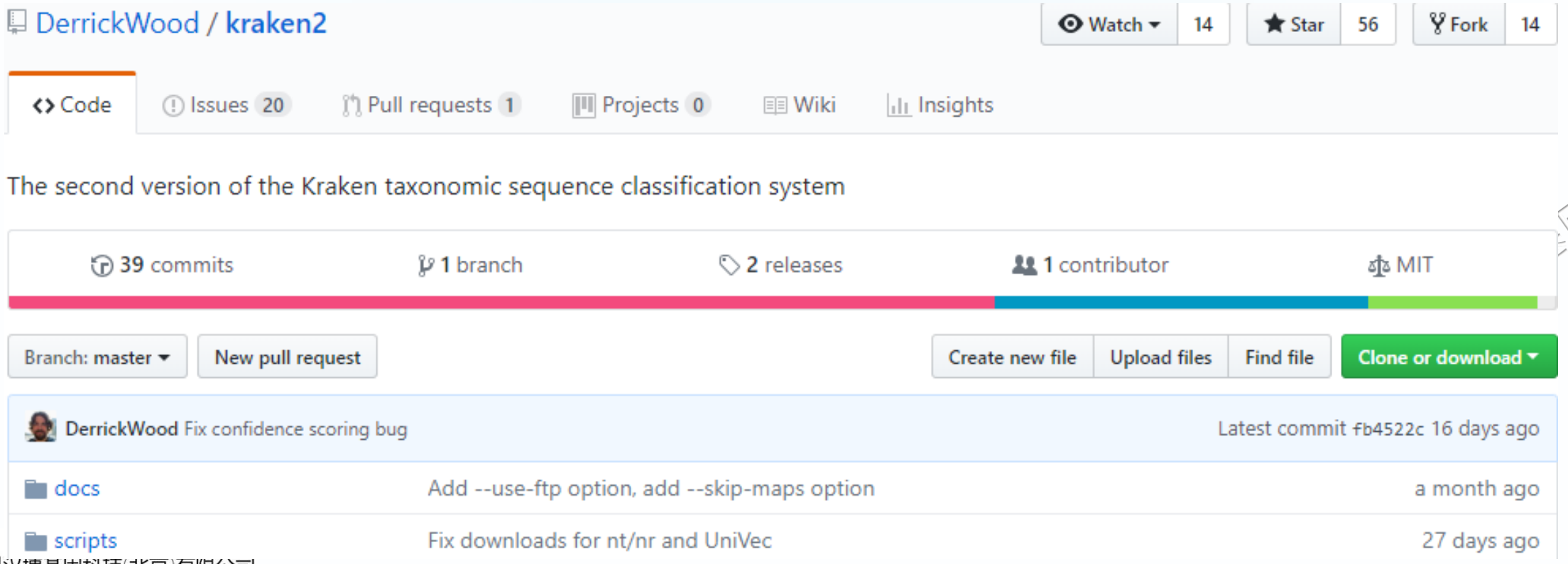
KrakenUniq: confident and fast metagenomics classification using unique k-mer counts

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



Kraken2安装和数据库配置

- # 基于LCA算法的物种注释 kraken2 https://ccb.jhu.edu/software/kraken/
- conda install kraken2
- # 下载数据库
- kraken2-build --standard --threads 24 --db /db/kraken2
- # 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
- # 此步下载数据 > 64 GB，下载时间由网速决定，单建索引时间4.5小时间，24线程35分完成

2.7 Kraken2基于NCBI数据库物种注释reads

2.7.1 多样本并行物种注释

```
mkdir -p temp/kraken2

time parallel -j 3 \

  'kraken2 --db ${db}/kraken2 --paired temp/qc/{1}_1_kneaddata_paired*.fastq \

  --threads 3 --use-names --use-mpa-style --report-zero-counts \

  --report temp/kraken2/{1}_report \

  --output temp/kraken2/{1}_output' \

  ::: `tail -n+2 result/metadata.txt | cut -f 1`

# 屏幕会输出各样品注释比例，和运行时间 1 - 10 min
```

2.7 Kraken2基于NCBI数据库注释reads层面

2.7.2 汇总样品物种组成表

```
mkdir -p result/kraken2
parallel -j 6 \
  sort temp/kraken2/{1}_report | cut -f 2 | sed "1 s/^/{1}\n/" > temp/kraken2/{1}_count' \
  ::: `tail -n+2 result/design.txt | cut -f 1`
header=`tail -n 1 result/design.txt | cut -f 1`
sort temp/kraken2/${header}_report | cut -f 1 | sed "1 s/^/Taxonomy\n/" > temp/kraken2/0header_count
paste temp/kraken2/*count > result/kraken2/taxonomy_count.txt
```

物种组成表

Taxonomy	p136C	p136N	p144C	p144N	p153C	p153N
d__Bacteria	42779	31219	18501	42228	43114	48946
d__Bacteria|p__Actinobacteria	18075	10898	6720	15026	18509	13707
d__Bacteria|p__Actinobacteria|c__Actinobacteria	13687	9452	2540	12985	18509	13687
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales				13687	8521	2444	11698	18509	12812
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae					13678	8453	2431	11459	13034
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Rothia					13499	8445	2315	11428
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Rothia|s__Rothia mucilaginosa									8367
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Rothia|s__Rothia dentocariosa									1703
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter						1267	766		80
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter alpinus									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. PGP41									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter crystallopo									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. PAMC 25									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. DCT5									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. FB24									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. Rue61a									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. QXT-31									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. YC-RL1									
d__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Micrococcales|f__Micrococcaceae|g__Arthrobacter|s__Arthrobacter sp. LS16									

- 本地/在线使用LEfSe差异比较，GraPhLan/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

2.7.3 物种多样性分析

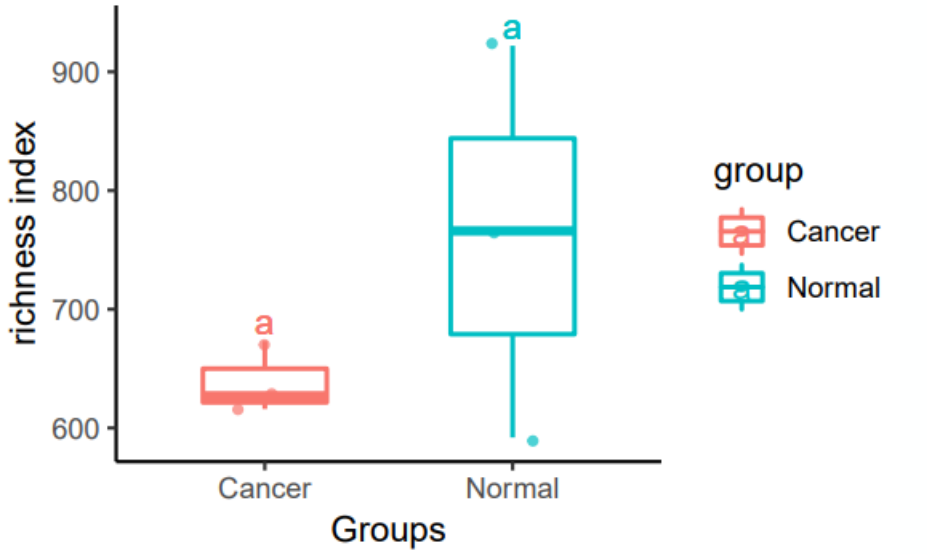
提取种级别注释并抽平至最小测序量，计算6种alpha多样性指数

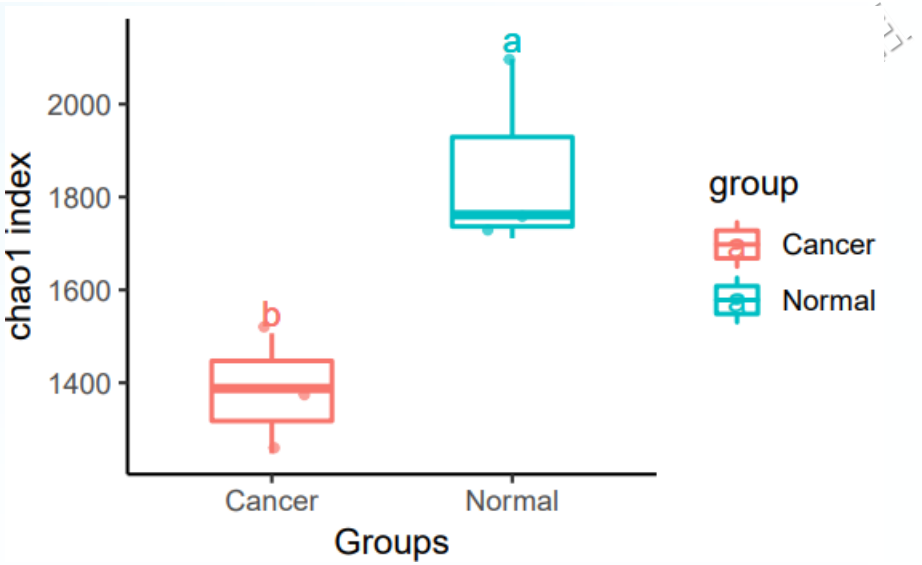
Rscript ./script/kraken2alpha.R -i result/kraken2/taxonomy_count.txt

绘制Alpha多样性指数，结果为输入文件+类型 richness/chao1/ACE/shannon/simpson

Rscript ./script/alpha_boxplot.R -i result/kraken2/taxonomy_count.alpha.txt -t shannon \

-d result/metadata.txt -n group -w 4 -e 2.5





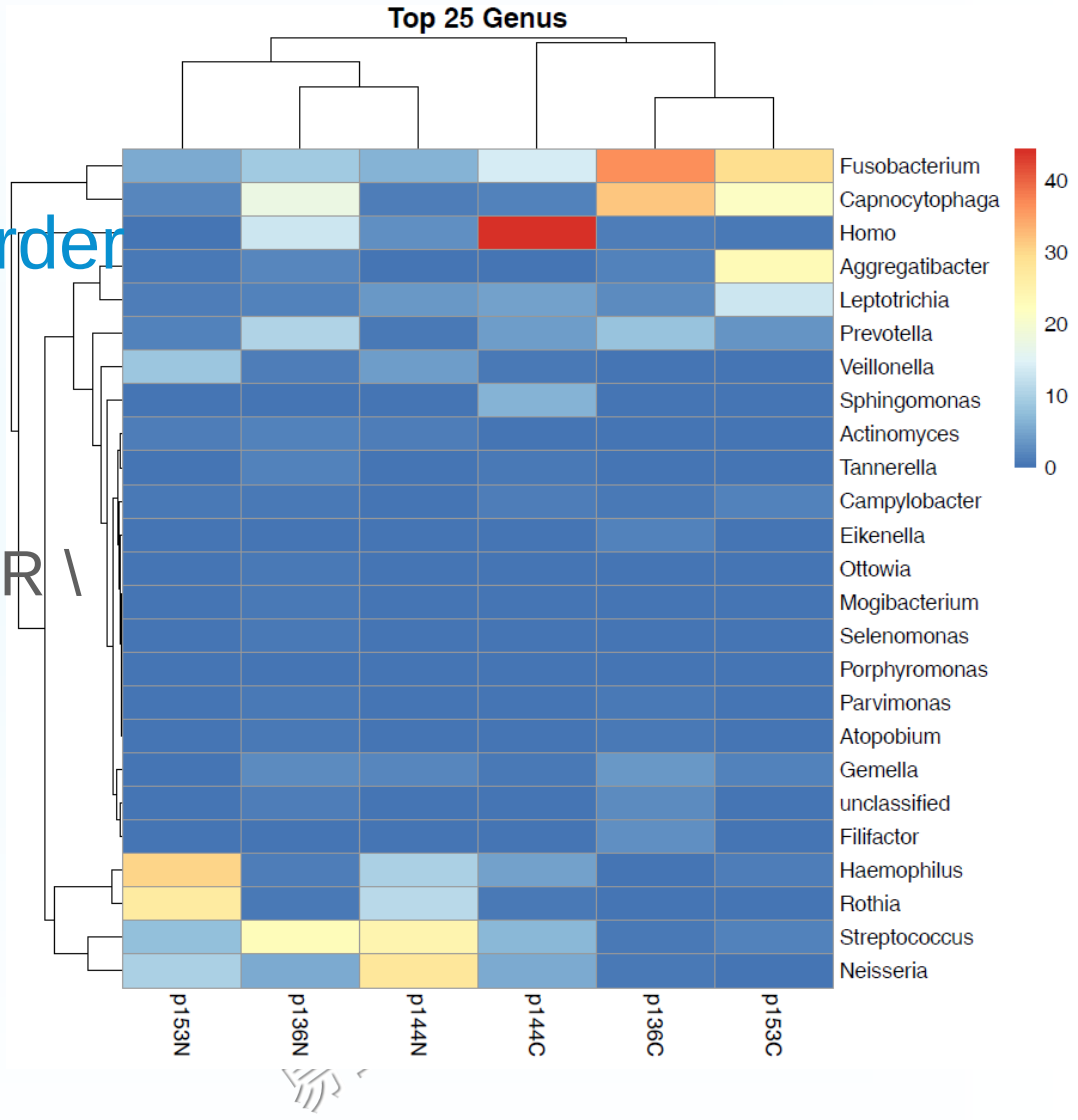
2.7.4 物种组成——热图

输入spf文件，即物种和丰度表格

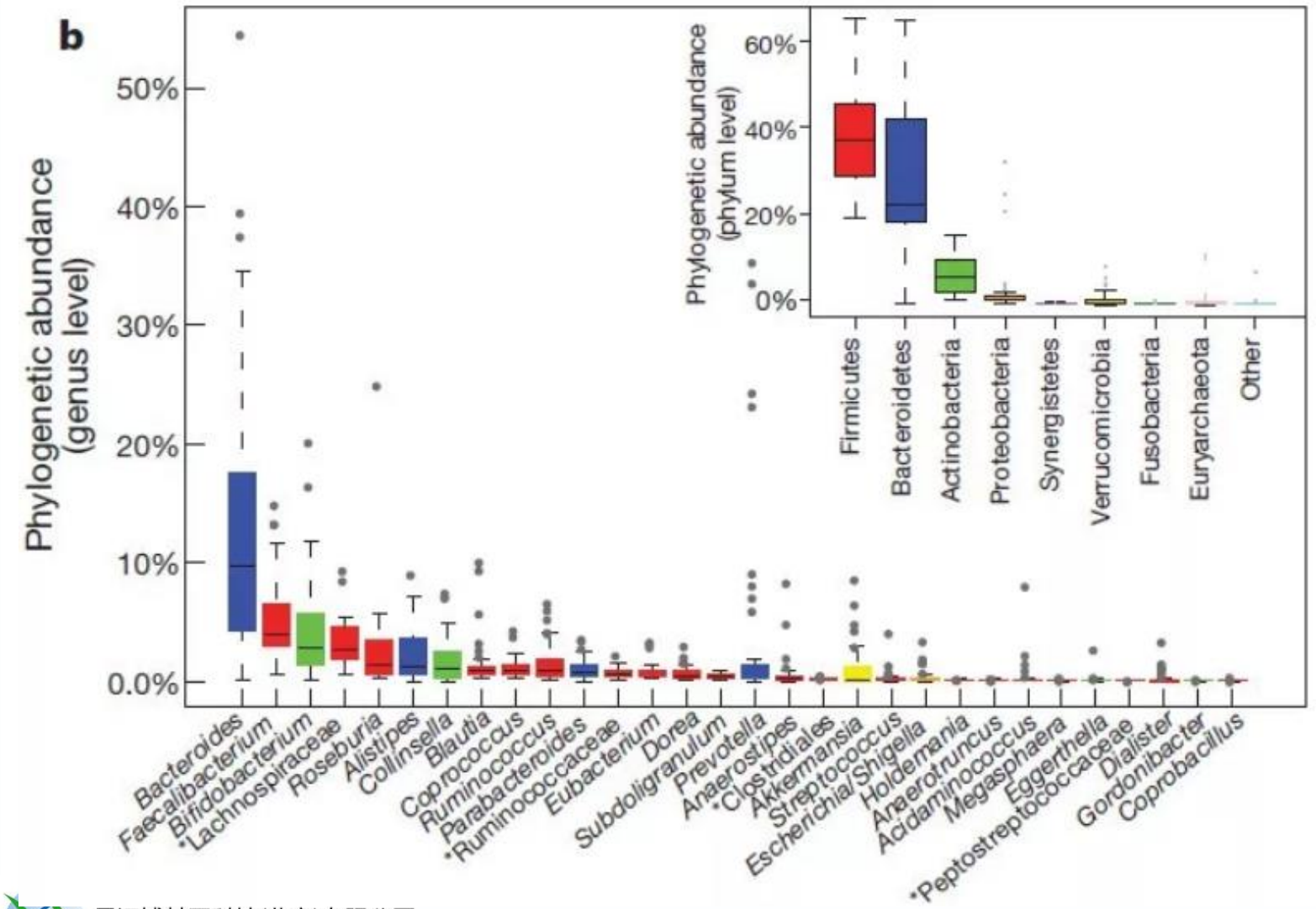
可选分类级Kingdom / Phylum / Class / Order / Family / Genus / Species、分类显示数量

输出文件名

```
Rscript db/script/metaphlan_hclust_heatmap.R \
  -i result/kraken2/taxonomy_count.norm.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



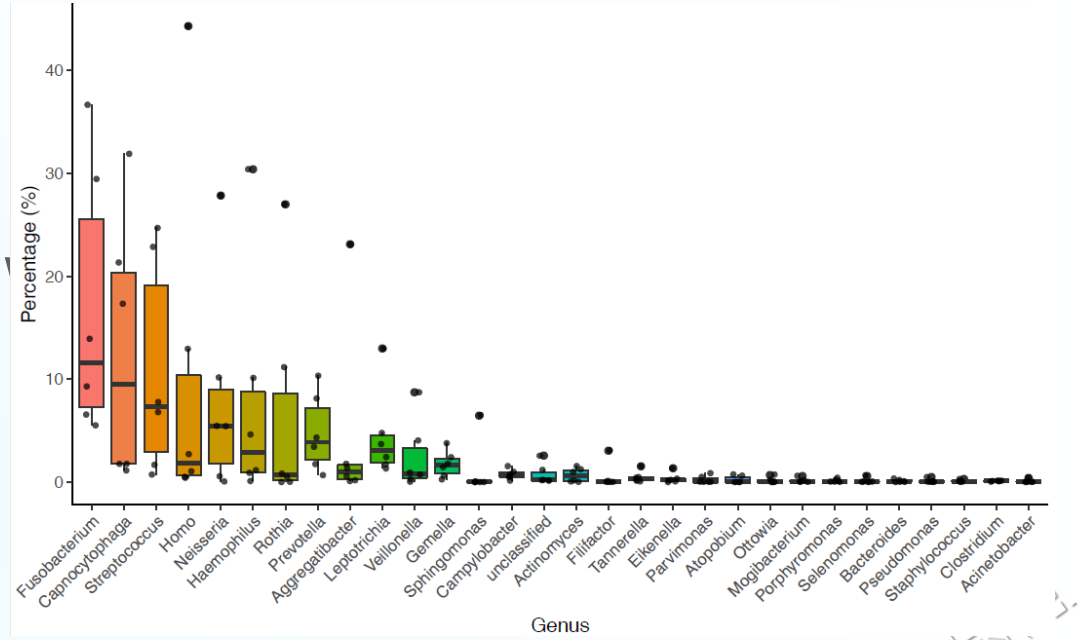
箱线图展示最高丰度的30个属。按门着色。同时角上有门水平箱线图。属和门水平丰度计算采用有参比对，85%相似度，65%覆盖度的阈值。未分类的属显示更高水平标注了星号。

•Nature：人类肠道微生物组的肠型

2.7.4 物种组成——热图

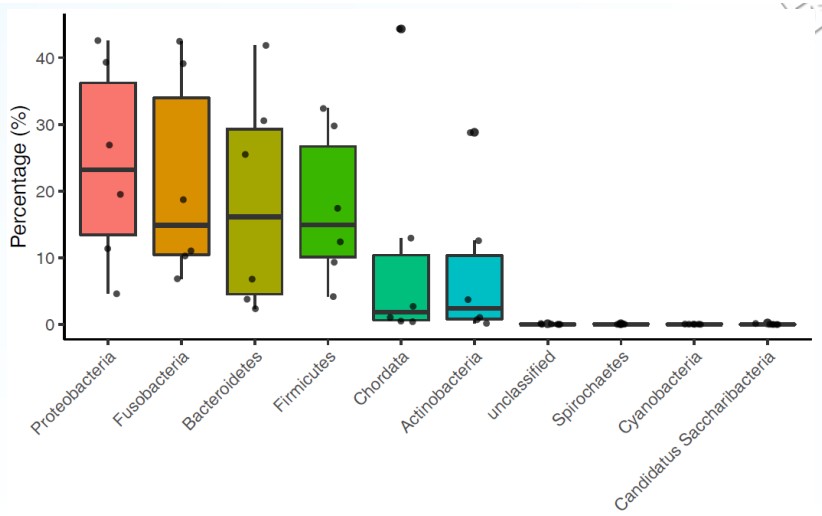
- # 绘制属水平Top30箱线图

```
Rscript db/script/metaphlan_boxplot.R \
  -i result/kraken2/taxonomy_count.norm.spf \
  -t Genus \
  -n 30 \
  -o result/kraken2/boxplot_Genus
```

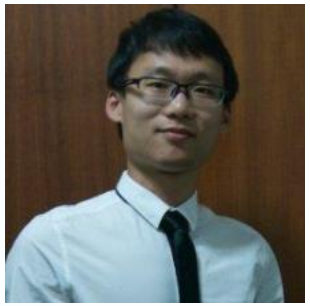


- # 绘制门水平Top10箱线图

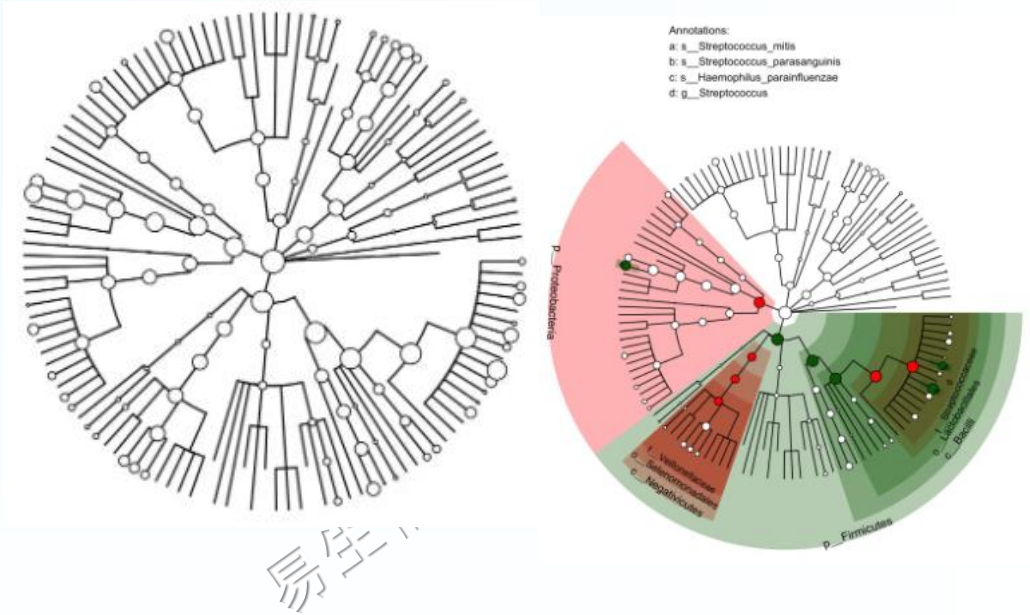
```
Rscript db/script/metaphlan_boxplot.R \
  -i result/kraken2/taxonomy_count.norm.spf \
  -t Phylum \
  -n 10 -w 4 -e 2.5 \
  -o result/kraken2/boxplot_Phylum
```



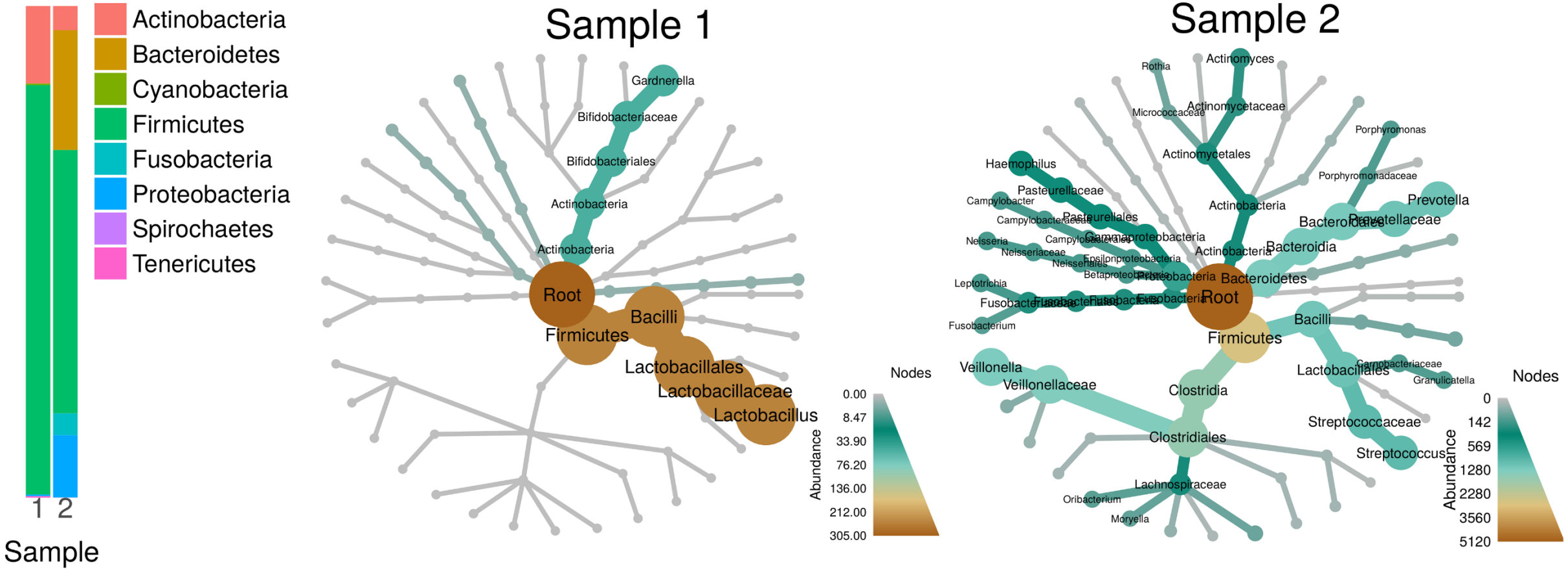
microbiomeViz：绘制lefse结果中Cladogram



- 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



Metacoder: Tools for Parsing, Manipulating, and Graphing Taxonomic Abundance Data



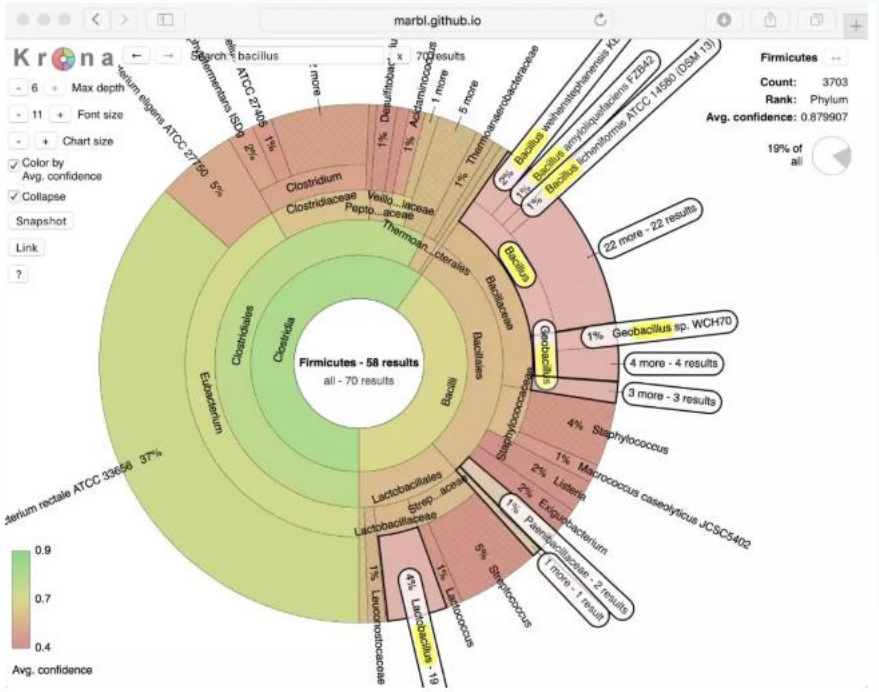
Foster ZSL, Sharpton TJ, Grünwald NJ (2017) Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Comput Biol 13(2): e1005404. https://doi.org/10.1371/journal.pcbi.1005404

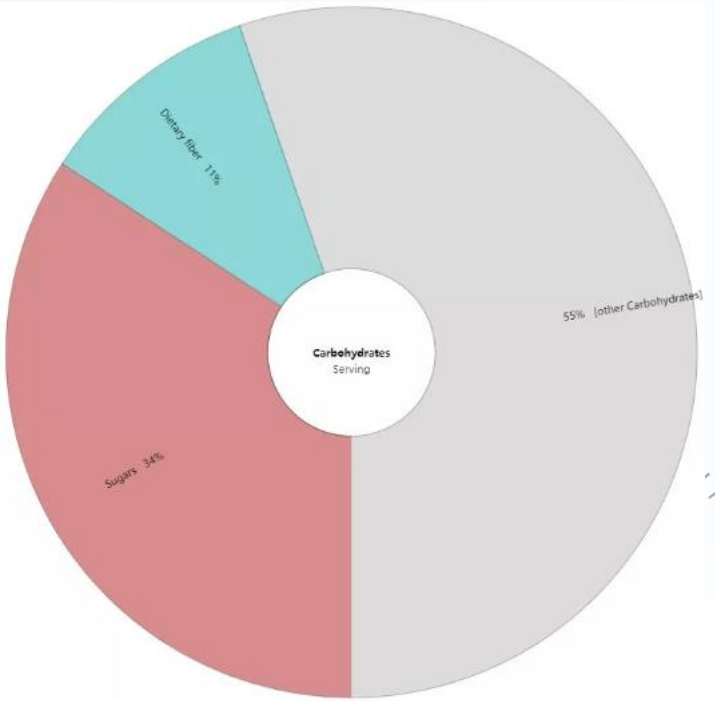
主页： https://grunwaldlab.github.io/metacoder_documentation/index.html

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





总结

- 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，多种分类方法间差异较大，理解原理仅供参考；
- Kraken2运行更快、数据库选择更灵活、结果更可读；
- Kraken2结果为reads counts格式，可计算alpha多样生richness和chao1，还可以绘制各级别热图和箱线图整体描述；
- 物种组成表下游可接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)

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/
- 微生物组助手 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