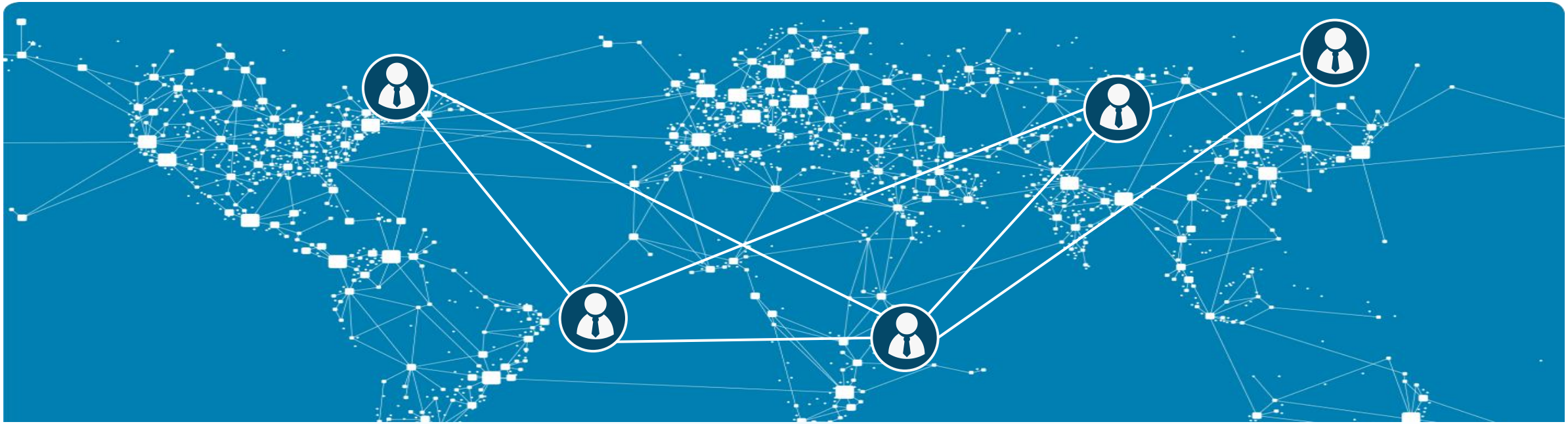
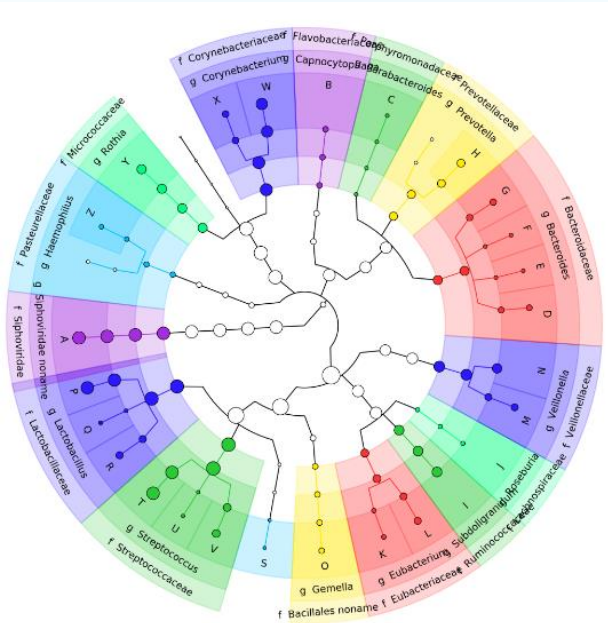
微生物组—宏基因组分析专题研讨会第20期

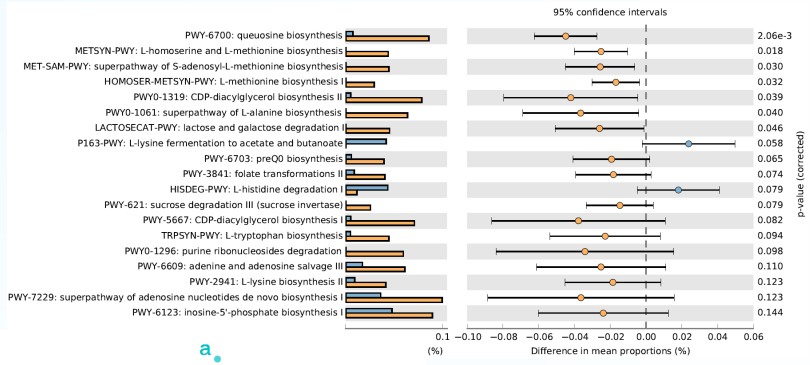


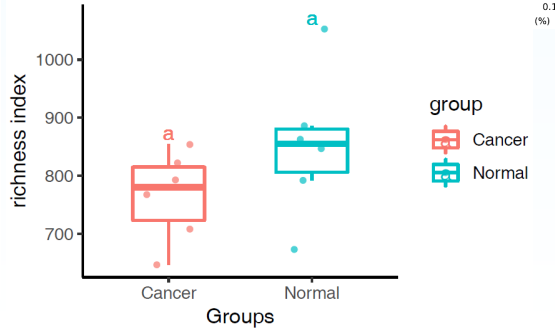
30 总结

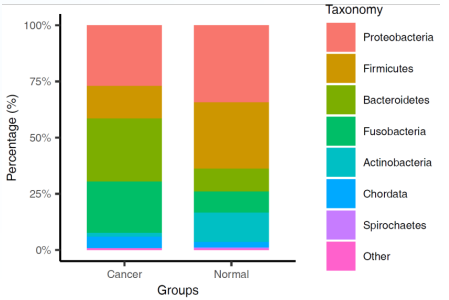
易生信

2023年11月26日

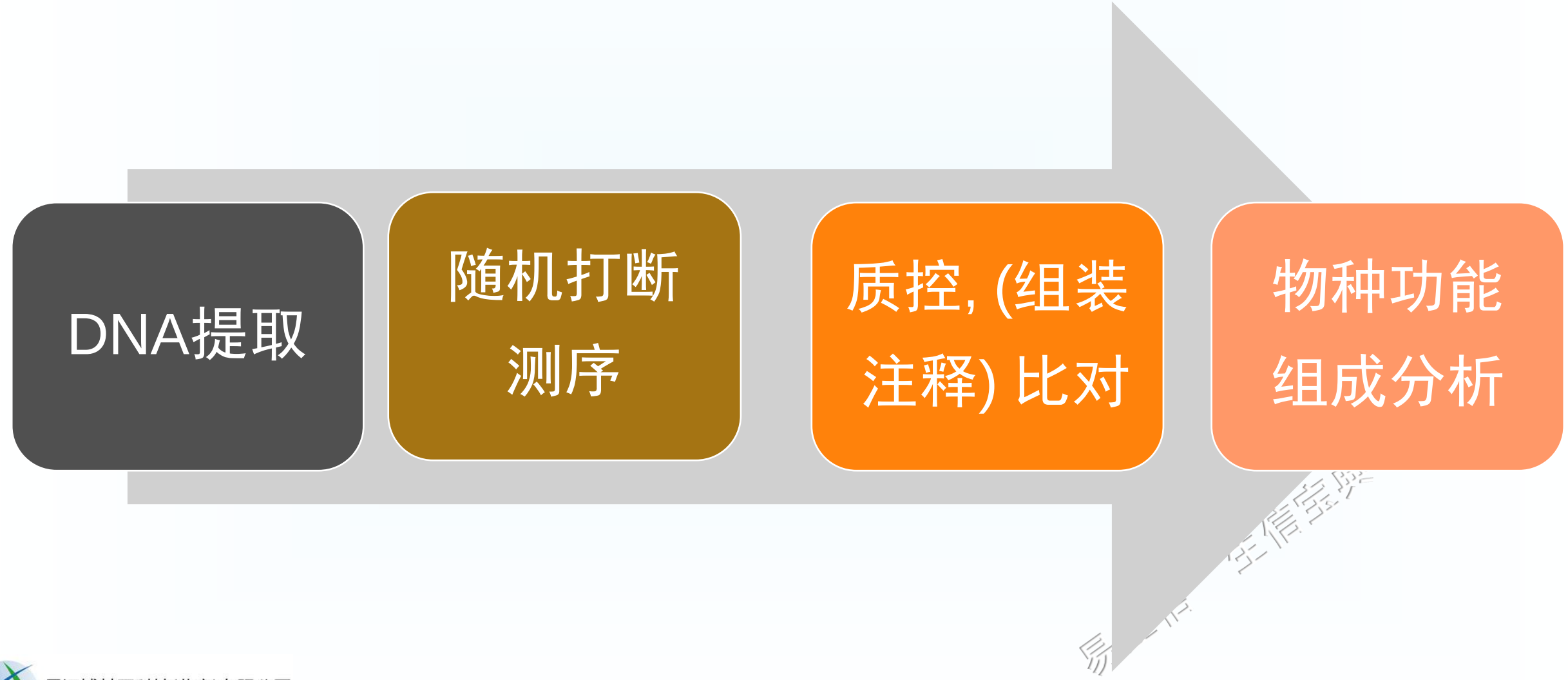








宏基因组实验分析流程



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

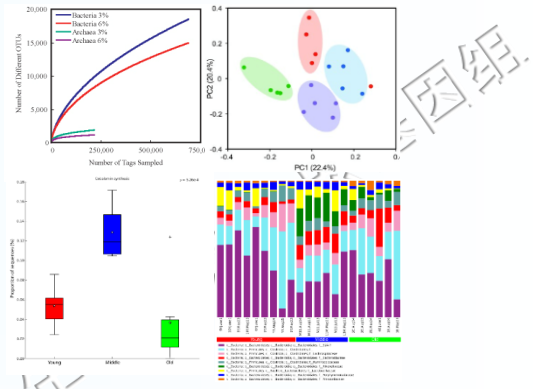
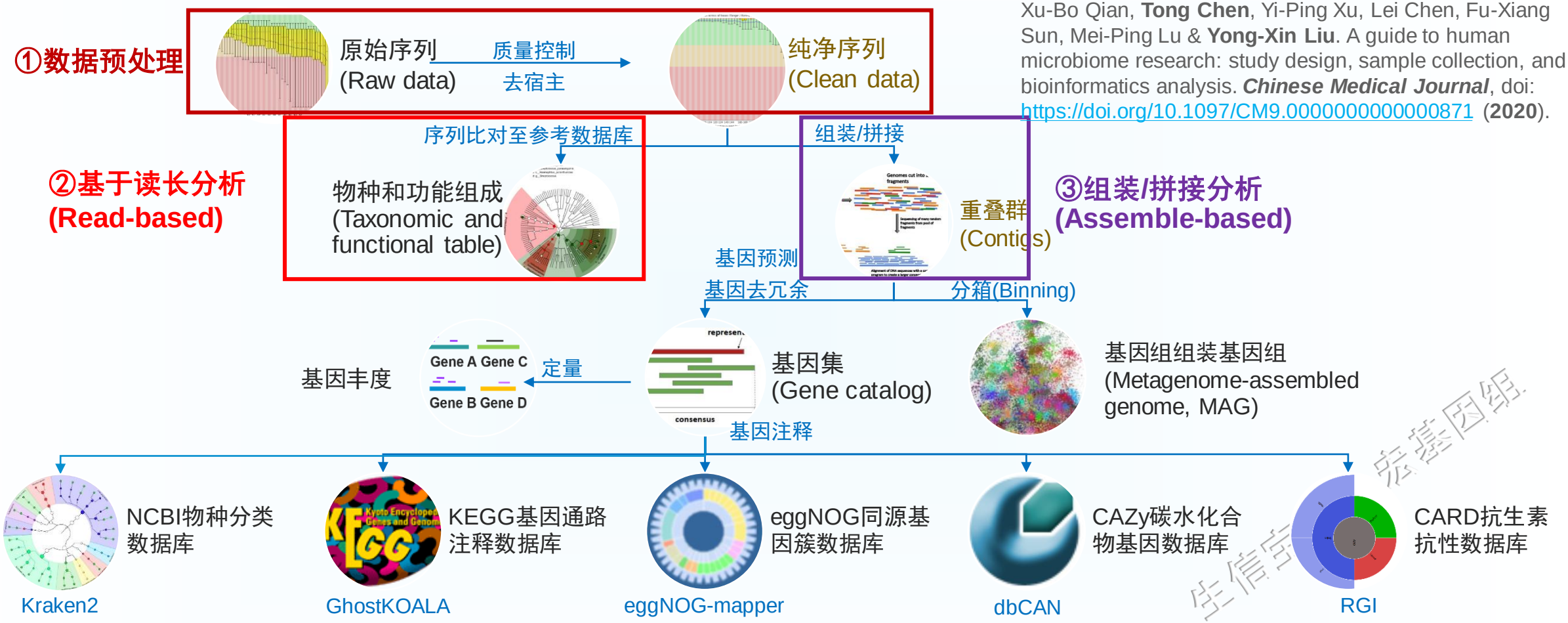


图: $10^{1\sim3}$个点和统计信息

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)

- **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
- 8627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).

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；

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 C3_1.fq.gz C5_1.fq.gz N1_1.fq.gz N3_1.fq.gz N5_1.fq.gz
C1_2.fq.gz C3_2.fq.gz C5_2.fq.gz N1_2.fq.gz N3_2.fq.gz N5_2.fq.gz

```
@SRR3586062.883556
CTTGGGGCTGCTGAGCTTCATGCTCCCCTCCTGCCTCAAGGACAATAAGGAGATCTTCGACAAGCCTGCAGCAGCTCGCATCGACGCCCTCATCGCTGAGG
+
CCCFFFFFHHHHHIJJJJJJJIJIJJJJGIJDGIJEIIJIJJJJJJJJIJJJJIJJIJJJJJHHHFFFFECEEEDDDDD?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	Replicate	Sex	Individual	GSA	CRR
C1	Cancer	1	Male	p136	CRA002355	CRR117732
C2	Cancer	2	Male	p143	CRA002355	CRR117733
N6	Normal	6	Female	p156	CRA002355	CRR117743

序列格式详解fasta&fastq，样品命名 注意事项 实例

fastp批量数据质量评估和质控

```
# -j 2: 表示同时处理2个样本
time tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \
  "fastp -i seq/{1}_1.fq.gz -I seq/{1}_2.fq.gz \
    -j temp/qc/{1}_fastp.json -h temp/qc/{1}_fastp.html \
    -o temp/qc/{1}_1.fastq  -O temp/qc/{1}_2.fastq \
    > temp/qc/{1}.log 2>&1 "
# 质控后结果汇总
echo -e "SampleID\tRaw\tClean" > temp/fastp
for i in `tail -n+2 result/metadata.txt|cut -f1`;do
    echo -e -n "$i\t" >> temp/fastp
    grep 'total reads' temp/qc/${i}.log|uniq|cut -f2 -d ':'|tr '\n' '\t' >> temp/fastp
    echo "" >> temp/fastp
done
sed -i 's/ //g;s/\t$//' temp/fastp
```

rush并行Kneaddata去宿主

- **-i输入文件，-o输出目录，-t线程数， -db 宿主基因组索引位置**

```
time tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \
  "sed '1~4 s/ 1:/.1:/;1~4 s/$/\/1/' temp/qc/{}_1.fastq > /tmp/{}_1.fastq; \
  sed '1~4 s/ 2:/.1:/;1~4 s/$/\/2/' temp/qc/{}_2.fastq > /tmp/{}_2.fastq; \
  kneaddata -i1 /tmp/{1}_1.fastq -i2 /tmp/{1}_2.fastq \
  -o temp/hr --output-prefix {1} --bypass-trim --bypass-trf --reorder \
  --bowtie2-options '--very-sensitive --dovetail' \
  -db ${db}/kneaddata/human/hg37dec_v0.1 --remove-intermediate-output -v -t 3; \
  rm /tmp/{}_1.fastq /tmp/{}_2.fastq"
```

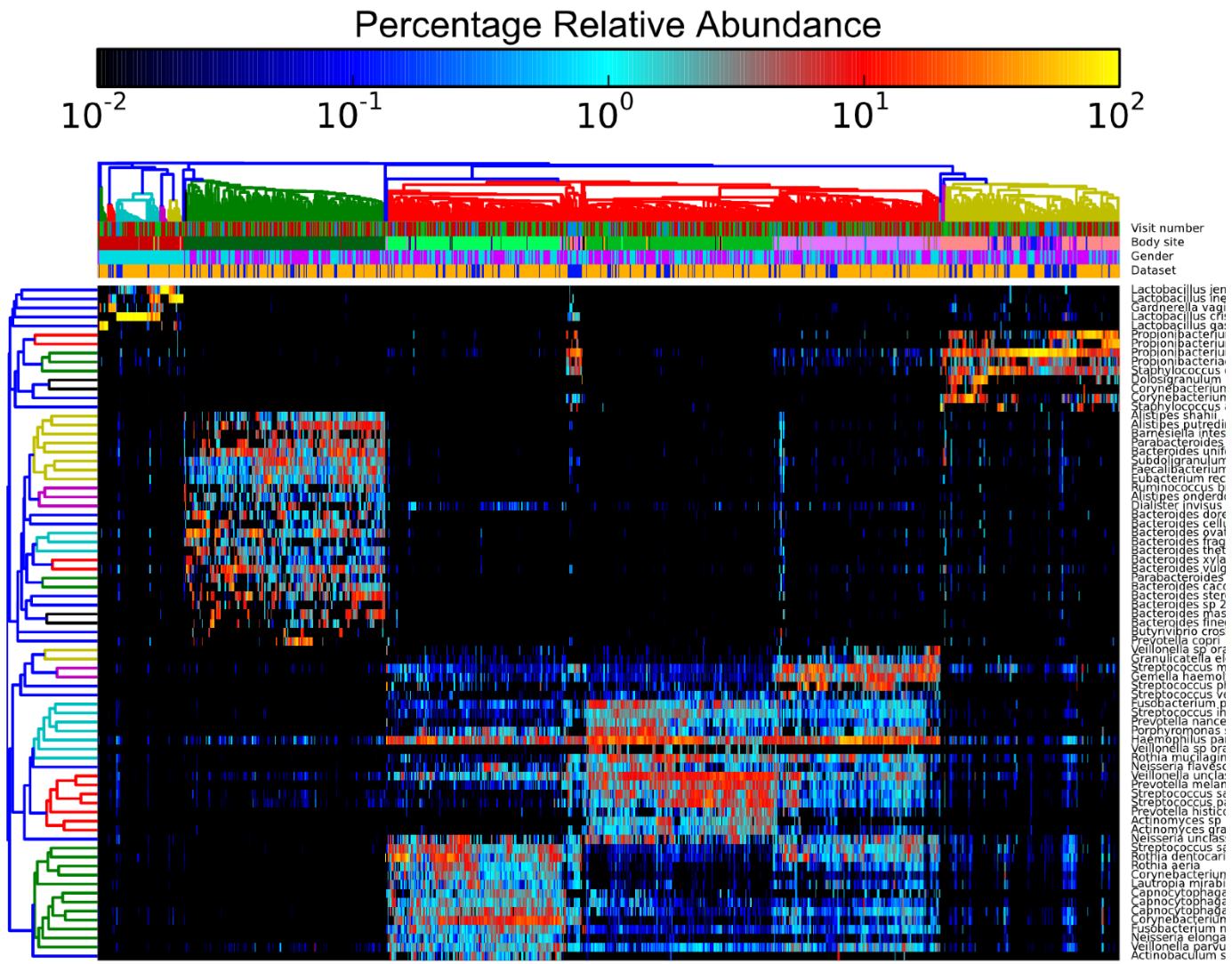
2.2 HUMAnN3计算物种和功能组成

```
mkdir -p temp/humann3
```

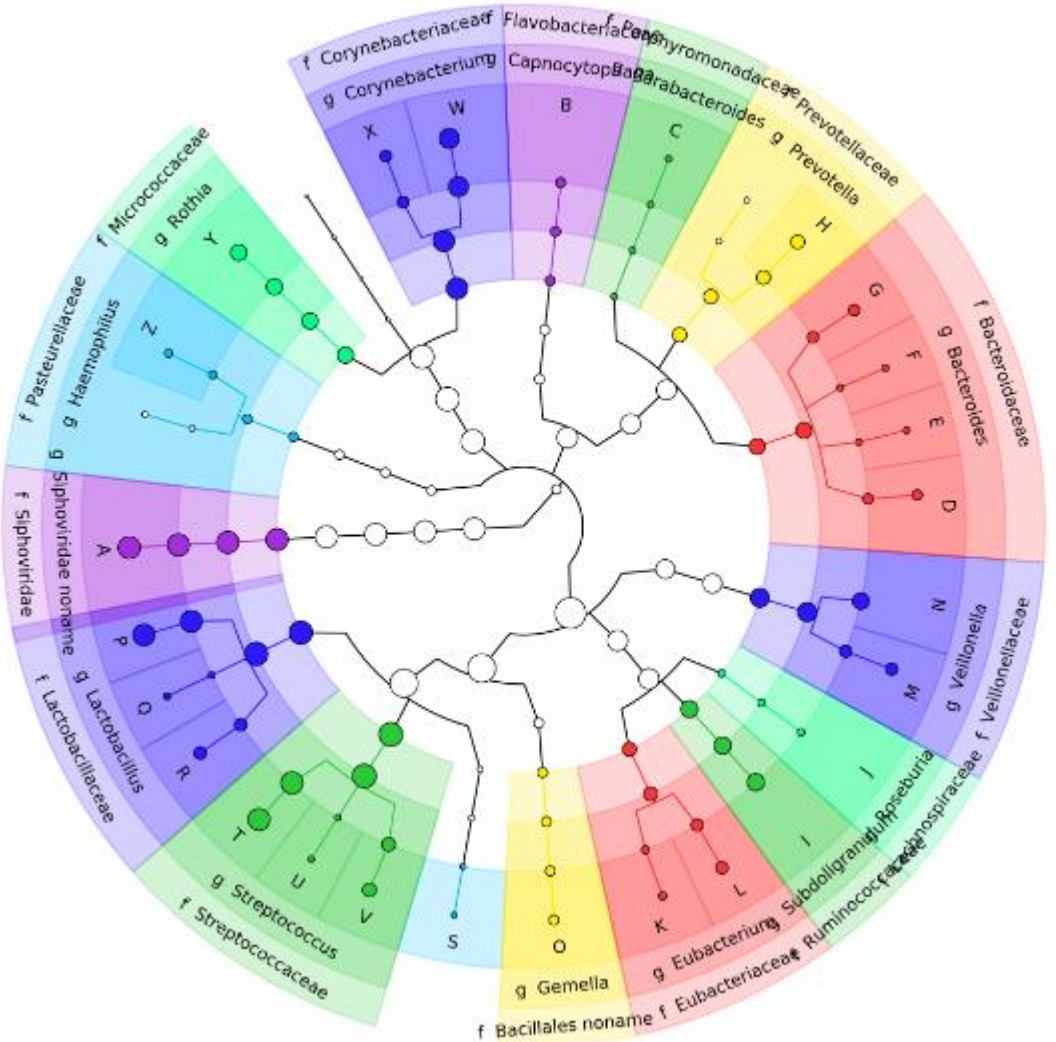
- 如果数据库位置正确，只需输入文件和输出目录，经rush管理批量任务队列

```
DEFAULT_DB_FOLDER=~/db/metaphlan4
tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \
  'humann2 --input temp/concat/{1}.fq  \
  --output temp/humann3/'
# 核心步骤，测序数据2X8=16线程，用时1h，真实数据可能要几小时至几天
```

MetaPhlAn4结果常用展示方式



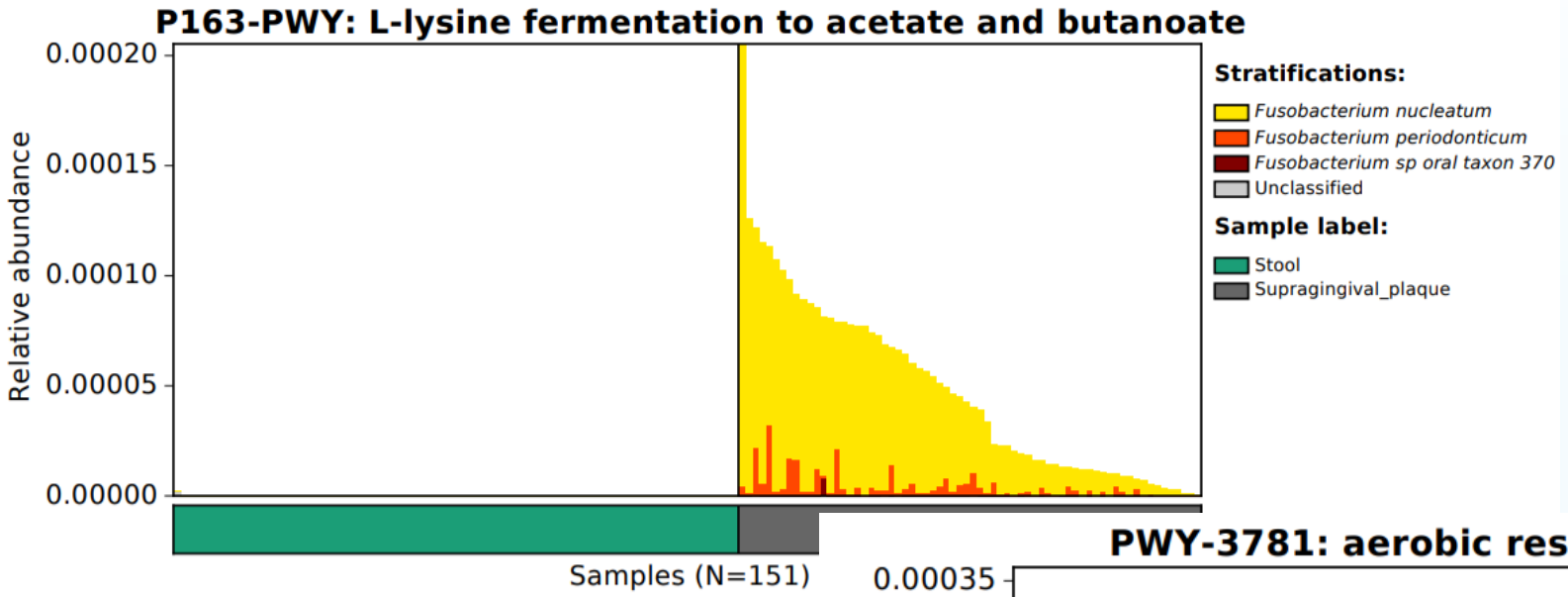
metaphlan_hclust_heatmap.py (huamnn2中)

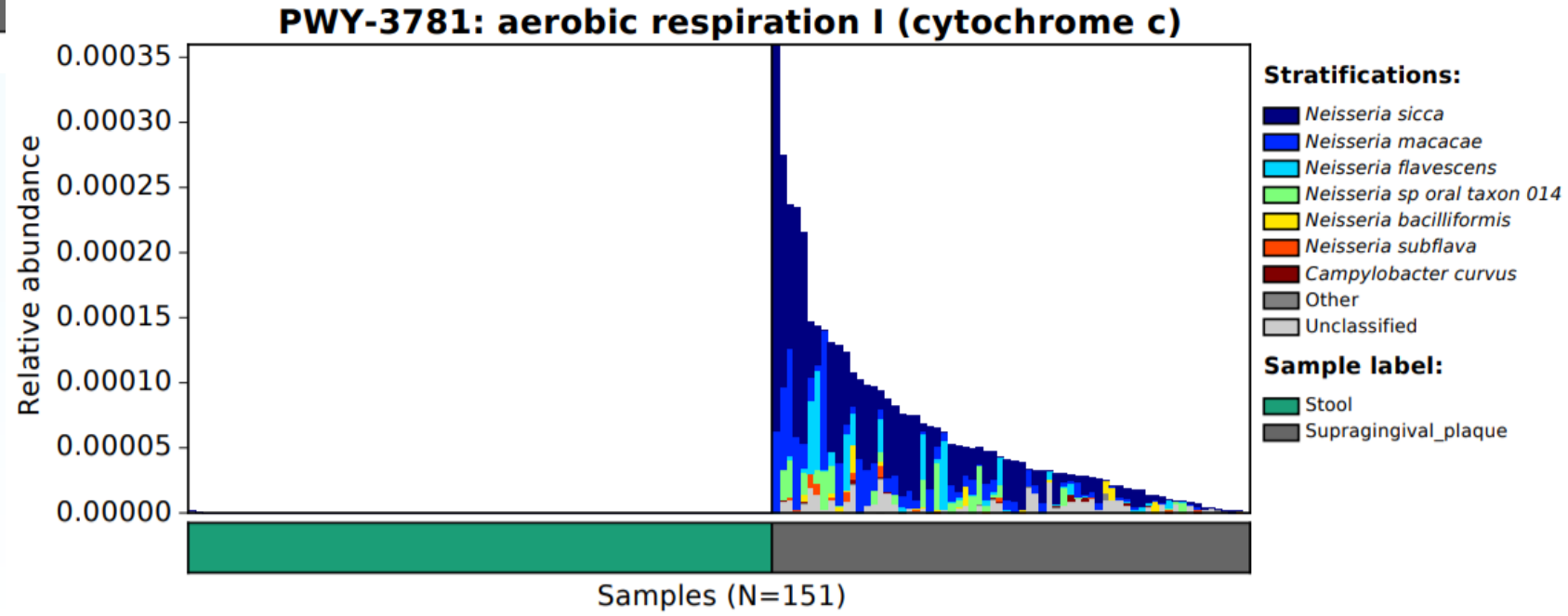


export2graphlan.py

MetaPhlAn4一条命令获得宏基因组物种组成

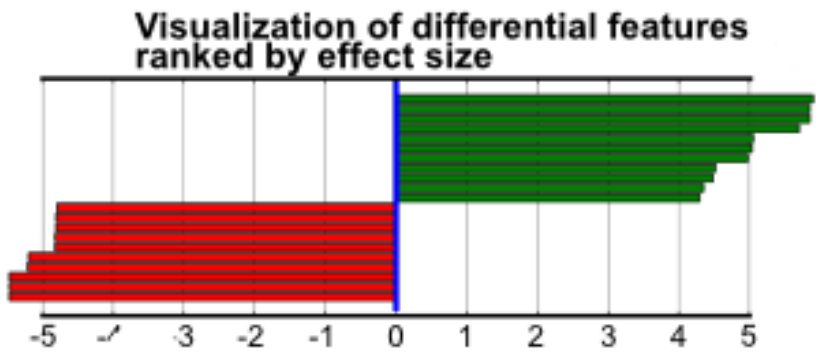
humann_barplot绘制功能的物种组成

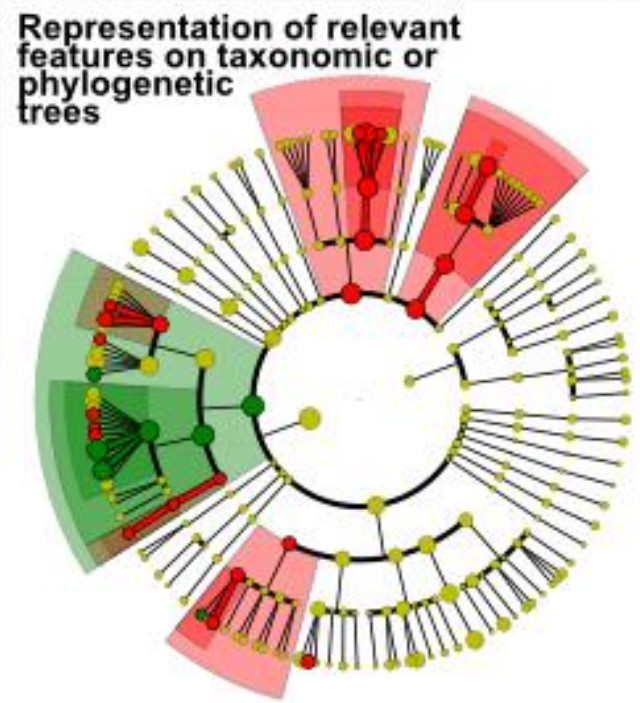


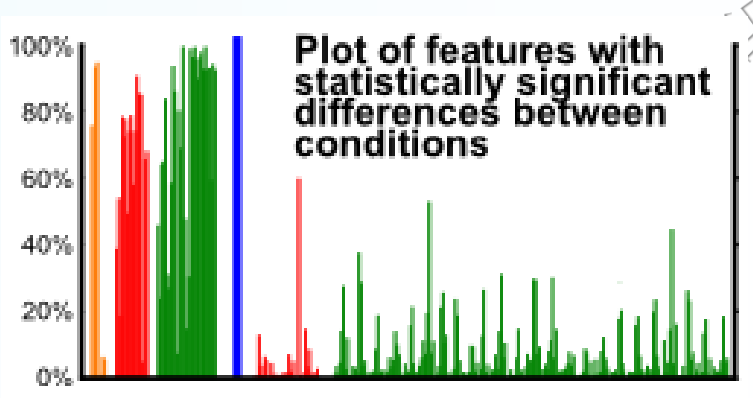


Linear discriminant analysis Effect Size (LEfSe)

- LEfSe分析即LDA Effect Size分析，是一种用于发现和解释**高维度数据生物标志**(基因、通路和分类单元等)的分析工具，可以进行两个或多个分组的比较，它强调**统计意义**和**生物相关性**，能够在组与组之间寻找具有统计学差异的生物标志（Biomarker）。

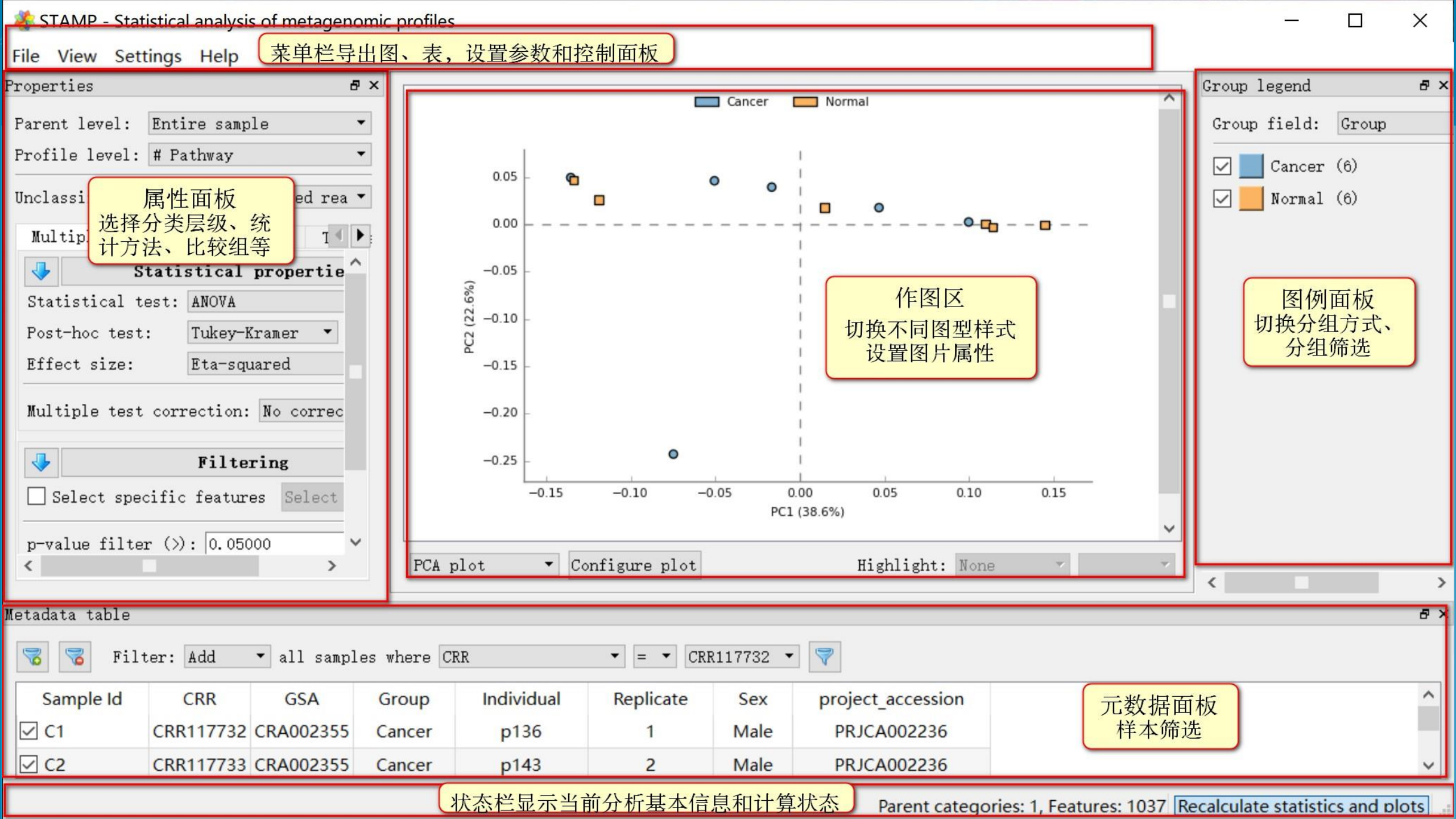




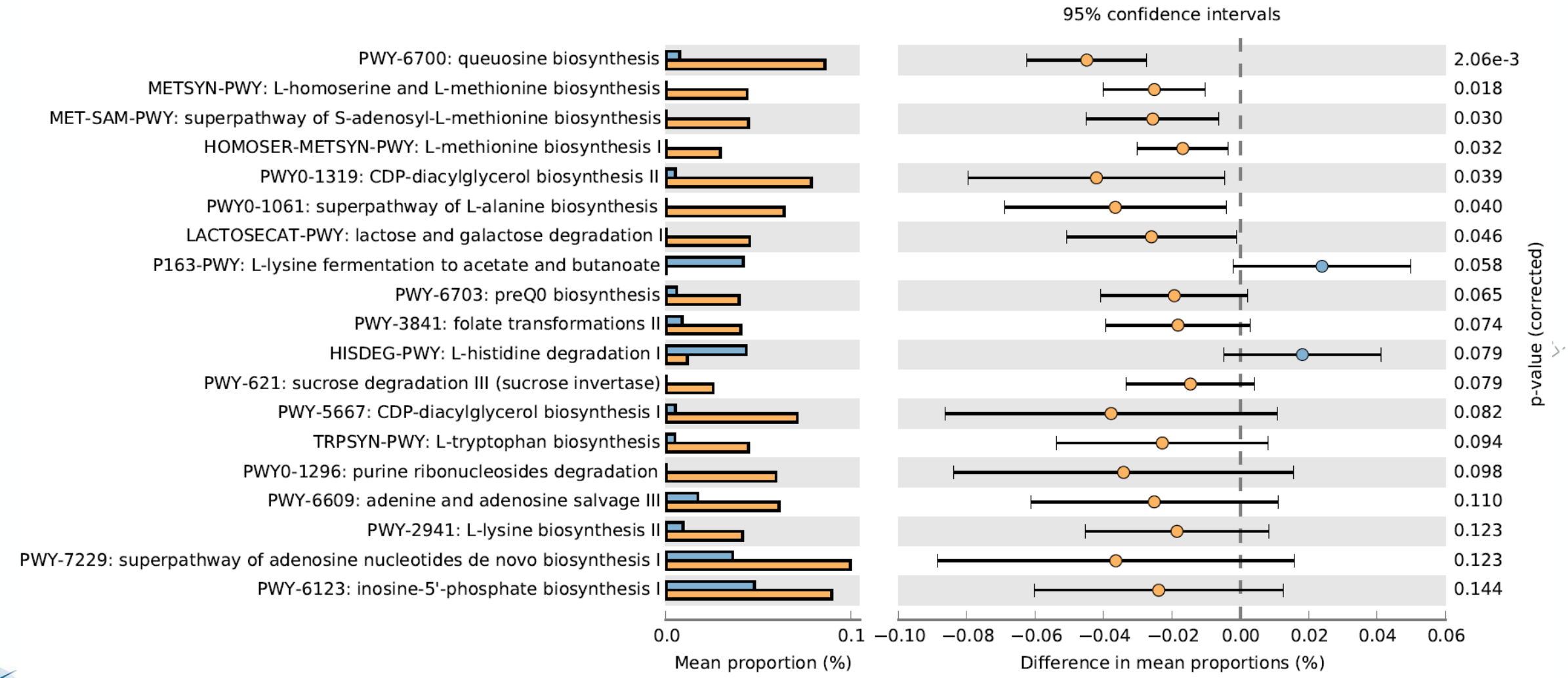


Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, *12*(6), R60.

LDA分析、作图及添加置信-ggord


STAMP - Statistical analysis of metagenomic profiles
File View Settings Help
菜单栏导出图、表，设置参数和控制面板
Properties
Parent level: Entire sample
Profile level: # Pathway
属性面板
选择分类层级、统计方法、比较组等
Statistical propertie
Statistical test: ANOVA
Post-hoc test: Tukey-Kramer
Effect size: Eta-squared
Multiple test correction: No correc
Filtering
Select specific features
Select
p-value filter (): 0.05000
Cancer
Normal
PC2 (22.6%)
PC1 (38.6%)
作图区
切换不同图型样式
设置图片属性
PCA plot
Configure plot
Highlight: None
Group legend
Group field: Group
Cancer (6)
Normal (6)
图例面板
切换分组方式、分组筛选
Metadata table
Filter: Add all samples where CRR = CRR117732
Sample Id CRR GSA Group Individual Replicate Sex project_accession
C1 CRR117732 CRA002355 Cancer p136 1 Male PRJCA002236
C2 CRR117733 CRA002355 Cancer p143 2 Male PRJCA002236
元数据面板
样本筛选
状态栏显示当前分析基本信息和计算状态
Parent categories: 1, Features: 1037
Recalculate statistics and plots

STAMP结果组间差异功能扩展柱状图



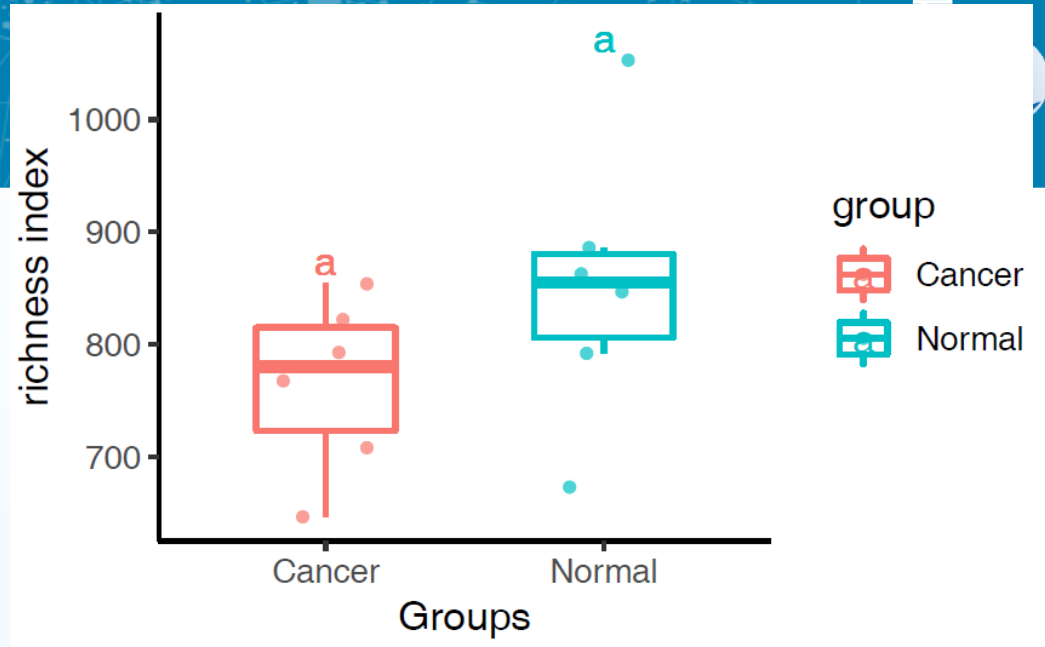
均值　　差异

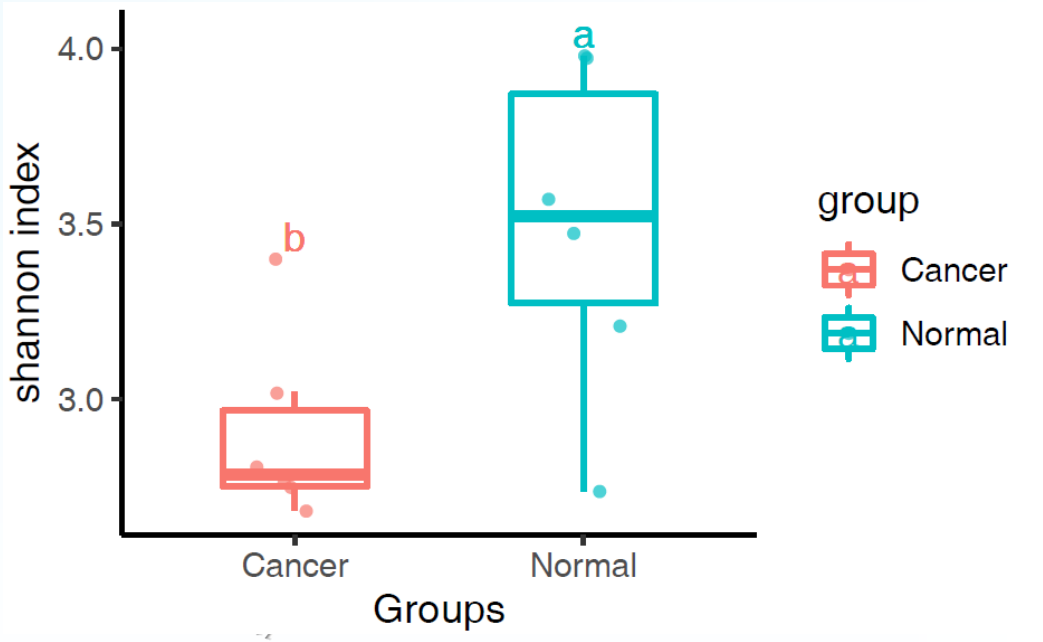
基于NCBI数据库的Kraken2物种注释

```
### 多样本并行物种注释

mkdir -p temp/kraken2

tail -n+2 result/metadata.txt|cut -f1|rush -j 3 \

  'kraken2 --db ~/db/kraken2/pluspfp16g --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物种多样性分析

```
# 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
```

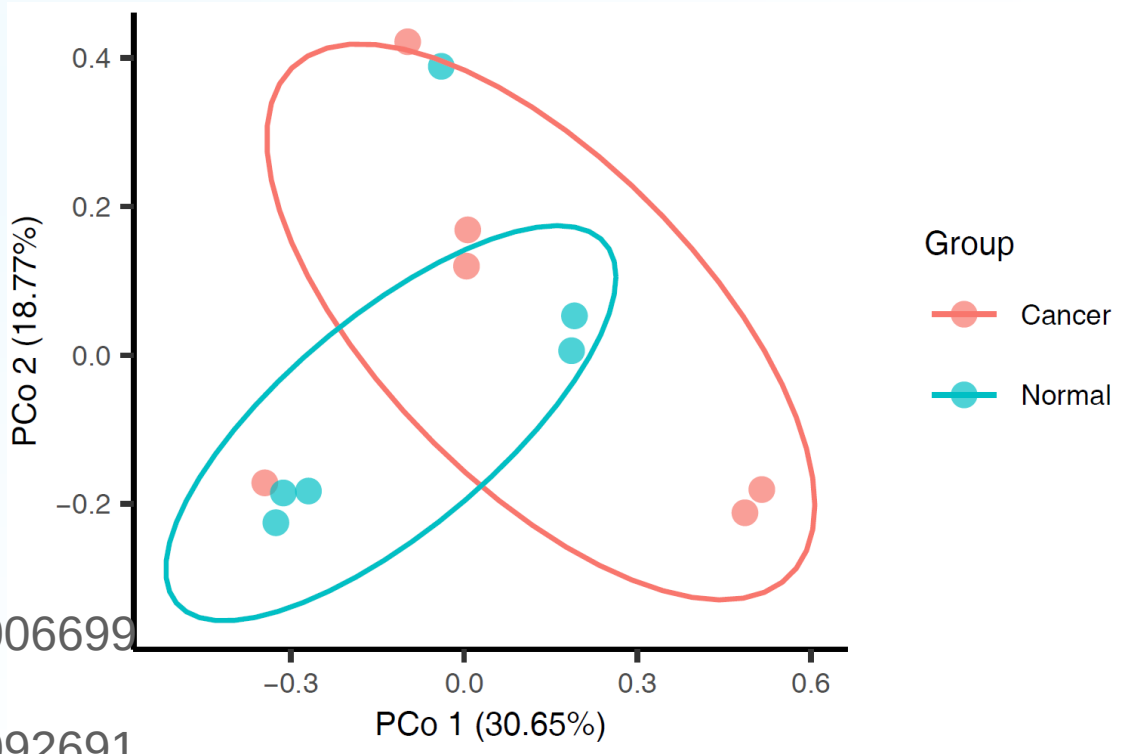




多样性可视化

- 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
```

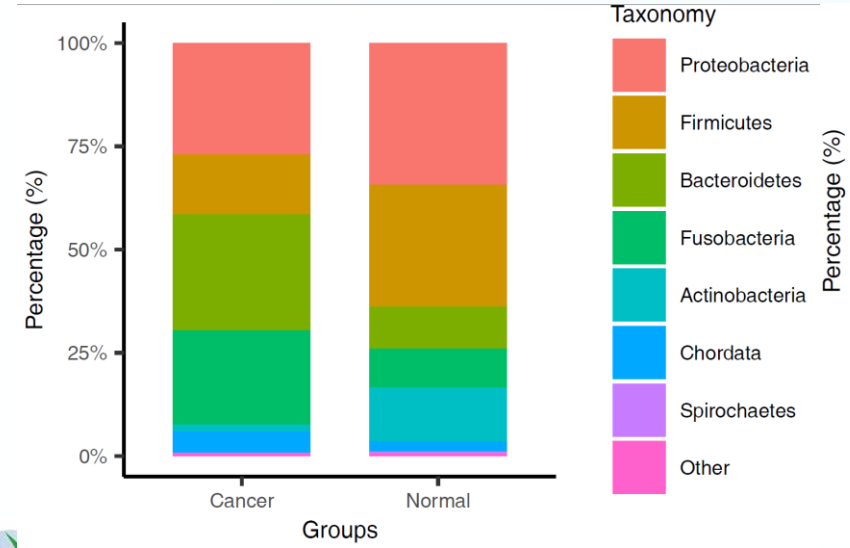


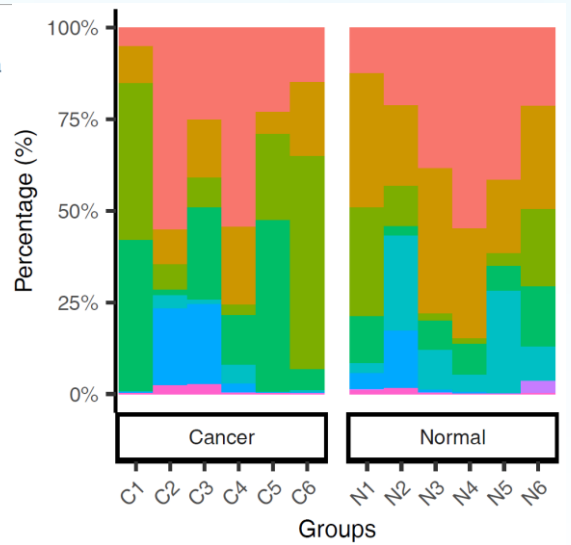

```
统计结果文件：          Sun Jan 03 16:19:07 2021
beta_pcoa_stat.txt      Cancer  Normal  0.300669933006699
P值有波动但比较稳定     Sun Jan 03 17:55:04 2021
                        Cancer  Normal  0.309269073092691
```

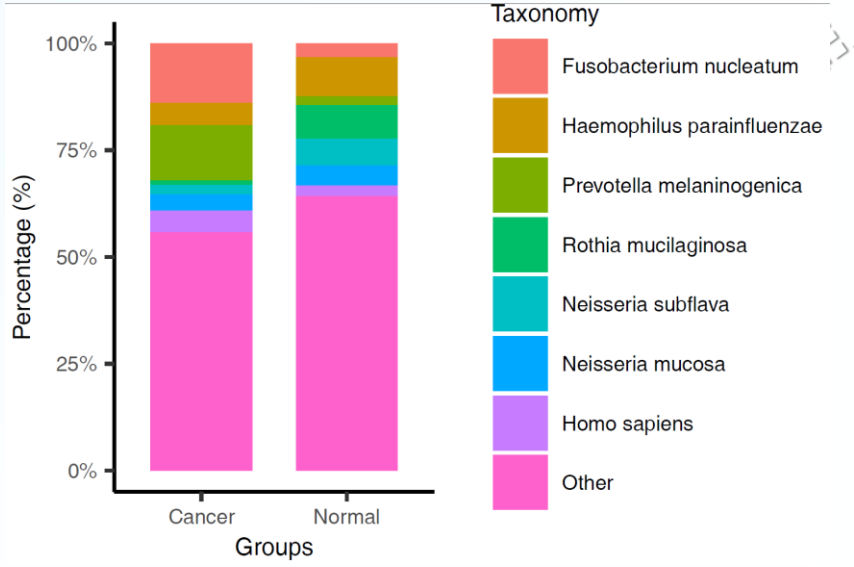
物种组成

- 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
```





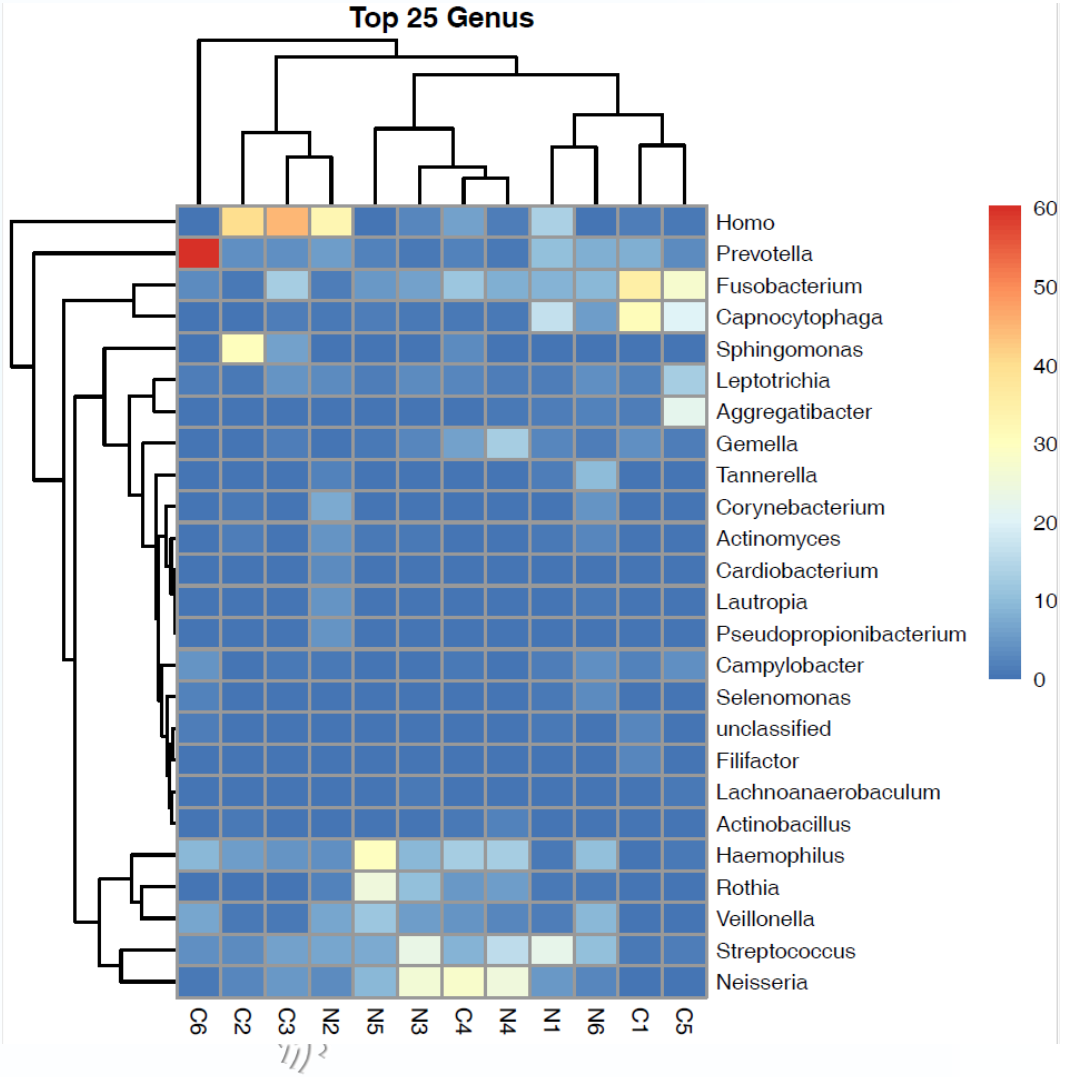


物种组成——热图

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
```



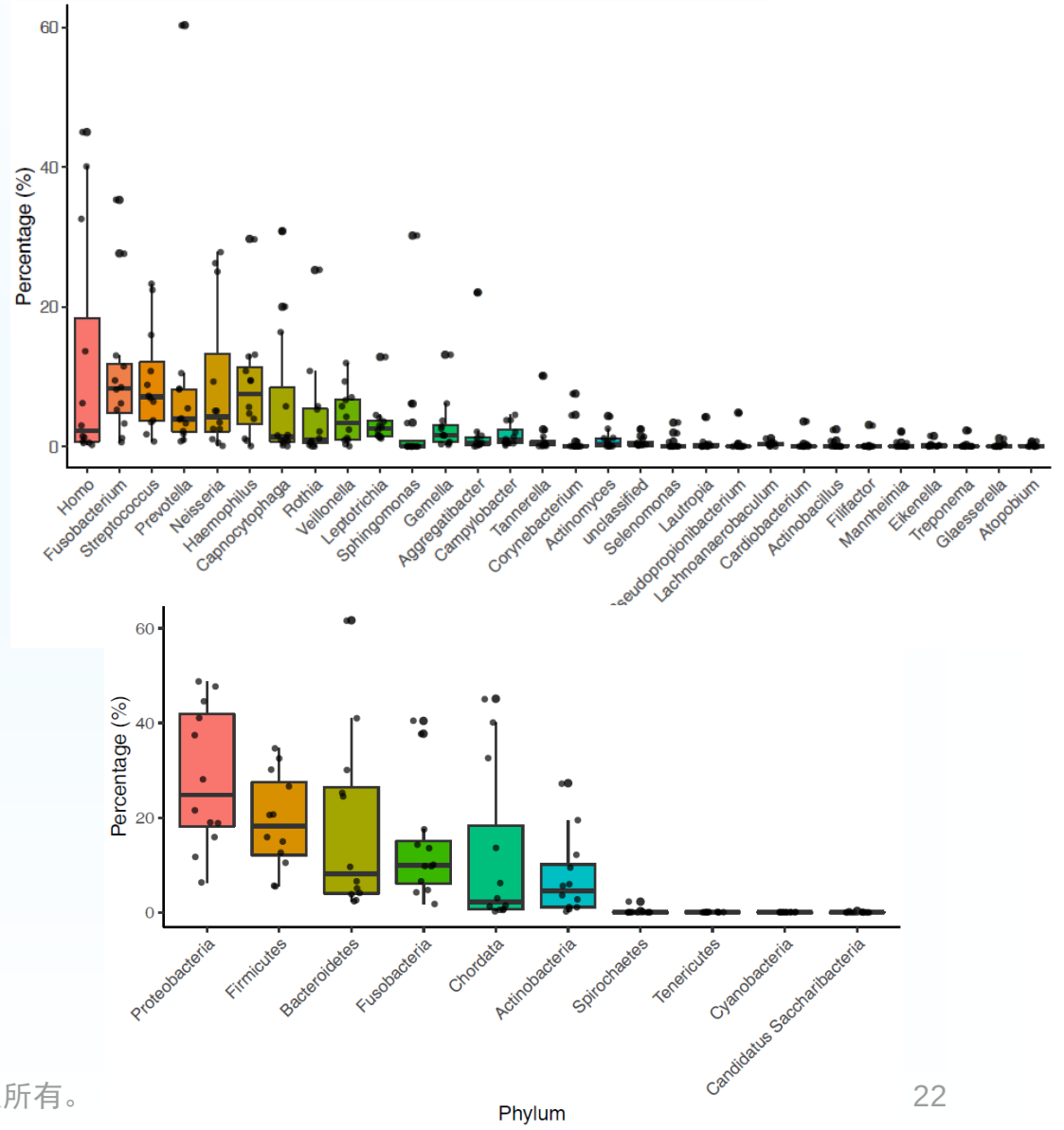
物种组成——箱线图

- # 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
```

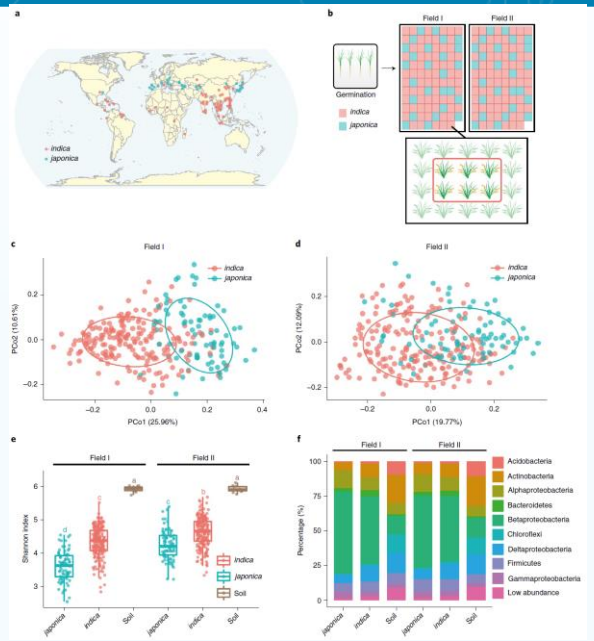
- # 绘制门水平Top10箱线图

```
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
```



高水平文章发表前三部曲

一. 图片拼图美化



二. 原始数据上传存档

三. 整理图表对应数据和分析代码

Yong-Xin Liu, Yuan Qin, **Tong Chen**, Meiping Lu, Xubo Qian, Xiaoxuan Guo & Yang Bai. A practical guide to amplicon and metagenomic analysis of microbiome data. ***Protein & Cell*** **2021**,12:315, https://doi.org/10.1007/s13238-020-00724-8

Protein Cell：扩增子和宏基因组数据分析实用指南

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