

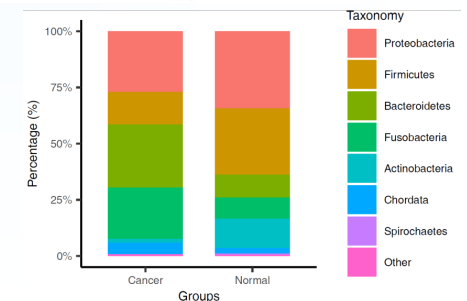
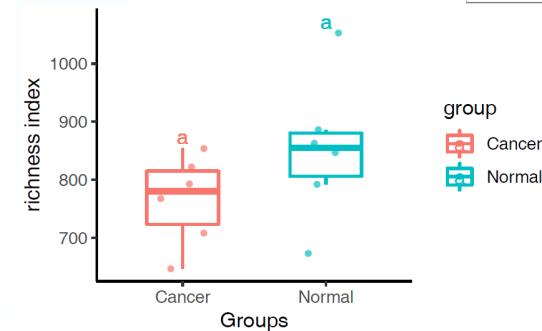
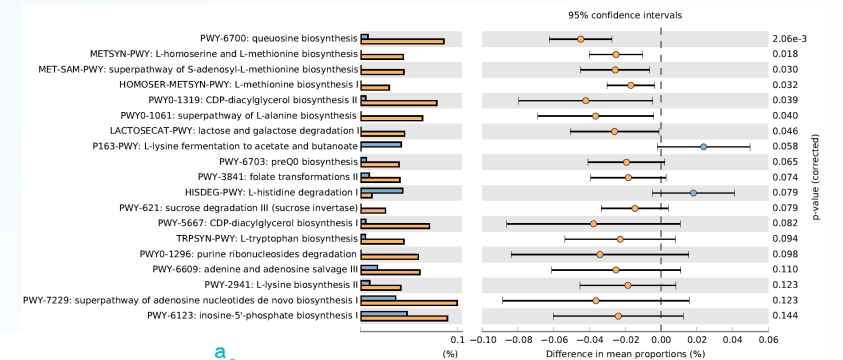
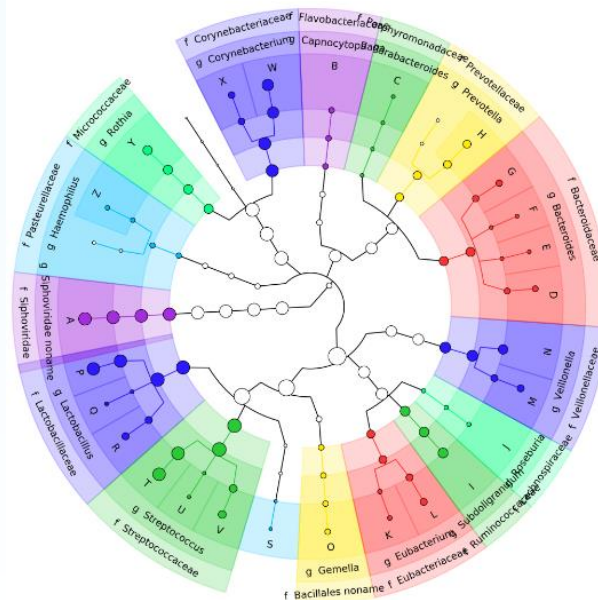
微生物组—宏基因组分析专题研讨会第18期



30 总结

易生信

2023年4月9日



宏基因组实验分析流程

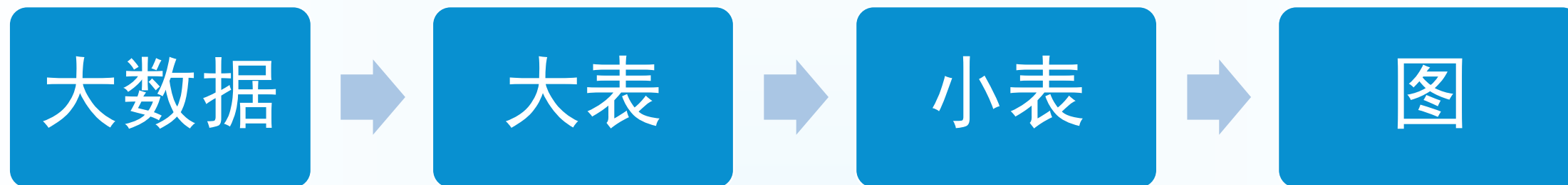
DNA提取

随机打断
测序

质控, (组装
注释) 比对

物种功能
组成分析

数据分析的基本思想——三步走



```

@HISEQ:549:HLNYBCXY:1:1101:1267:2220 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCTGGTAGTCCACGCTGTAACGTTGGGCG
+
DDDDDIHHIIIIIIIIHHIIIIIIIIHHIIIIIIIIHHIIIIIIIIHHIIIIIIII
@HISEQ:549:HLNYBCXY:1:1101:1887:2204 1:N:0:CACTCAAT
TACGAGTATGAACAGGATTAGATACCTGGTAGTCCACGCCCTAAACGATGTCTA
+
DDDD@H~GHIIIIIIIIIIIIIIIIIIIHHIIIIIIIIIIIIIIIIIIIIIIII
@HISEQ:549:HLNYBCXY:1:1101:2196:2168 1:N:0:CACTCAAT
TCGTCGCTCGAACAGGATTAGATACCTGGTAGTCCACGCCCTAAACGATGACAA
+
DDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@HISEQ:549:HLNYBCXY:1:1101:2025:2183 1:N:0:CACTCAAT
ATATCGCGAGAACAGGATTAGATACCTGGTAGTCCACGCCCTAAACGATGAGCG
+
DDDD@E@HIGHIIHHFHHIIIIIFHHIIHHGIHHIIIIICHDEHHIIIIHGH
@HISEQ:549:HLNYBCXY:1:1101:2052:2198 1:N:0:CACTCAAT
CAGGAGACAGAACAGGATTAGATACCTGGTAGTCCACGCTGTAACGATGGGTA
+
D@DD@H=7CCHIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  
```

序列: $10^6 \sim 10^9$

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	0	2
OTU_58	77	75	28	84	53	64
OTU_1110	6	3	3	2	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	1
OTU_1137	0	1	0	3	0	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: $10^{1-3} \times 10^{3-5}$

Sample	berger_parker	buzas_gibson	chaol
WT6	0.042	0.0381	1388.9
WT3	0.0453	0.0425	1474.9
OE4	0.0359	0.0414	1476.4
WT2	0.0642	0.0244	1203.0
OE3	0.0426	0.0396	1716.9
WT1	0.0586	0.0293	1317.0
WT4	0.0518	0.0359	1353.2
OE5	0.0361	0.0441	1622.8
OE2	0.0466	0.0472	1733.3
OE6	0.0432	0.0523	1759.5
WT5	0.0435	0.0252	1181.6
OE1	0.0374	0.0524	1591.2
K04	0.0558	0.0325	1474.1
K01	0.0552	0.0409	1651.6
K05	0.0732	0.025	1306.2
K02	0.0509	0.0445	1675.3
K03	0.0571	0.0329	1489.8
K06	0.0518	0.0334	1215.9

统计表: $1 \sim N \times 10^{1-3}$

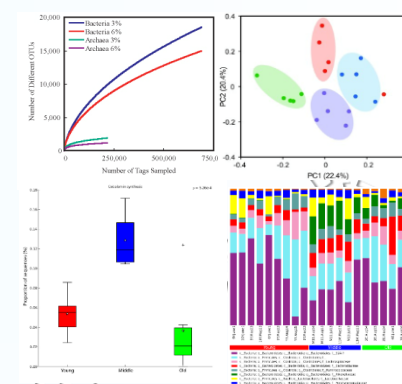


图: 10^{1-3} 个点和统计信息

宏基因组分析

主要步骤常用软件和数据库

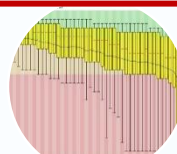


[Protein Cell: 扩增子和宏基因组数据分析实用指南](#)

宏基因组分析流程

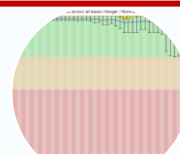
Xu-Bo Qian, **Tong Chen**, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu & **Yong-Xin Liu**. A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chinese Medical Journal*, doi: <https://doi.org/10.1097/CM9.0000000000000871> (2020).

①数据预处理



原始序列
(Raw data)

质量控制
去宿主

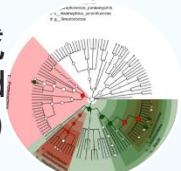


纯净序列
(Clean data)

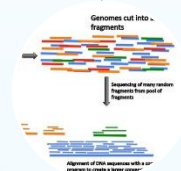
②基于读长分析 (Read-based)

序列比对至参考数据库

物种和功能组成
(Taxonomic and
functional table)



组装/拼接



③组装/拼接分析 (Assemble-based)

重叠群
(Contigs)

基因预测

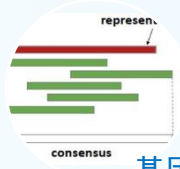
基因去冗余

分箱(Binning)

基因丰度



定量

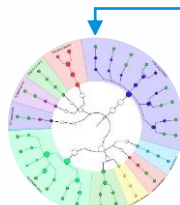


基因集
(Gene catalog)

基因注释



基因组组装基因组
(Metagenome-assembled
genome, MAG)



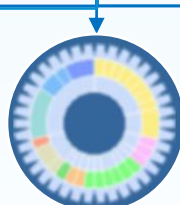
NCBI物种分类
数据库

Kraken2



KEGG基因通路
注释数据库

GhostKOALA



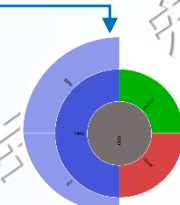
eggNOG同源基
因簇数据库

eggNOG-mapper



CAZy碳水化合
物基因数据库

dbCAN



CARD抗生素
抗性数据库

RGI

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
- 8527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**

易生信
生信宝典
宏基因组



- 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
+
CCCCFFFFHHHHHIJJJJJJJIJIIJJJJGIJDGIJEIJJJJJJJJJJJJJJJJJJJJHHEEEEDDDDD?BDDDDDBDDDDDDDDBBBD
@SRR3586062.3376311
GACGGTGTCCTCAGGACCCTTCAGTGCCTTCATGATCTGCTCAGAGGTGATGGAGTCACGGACGAGATTGTCGTGTCAGCACGTAGGATGCGGTCGCCTG
+
@@@DDDDAFF?DF;EH+ACHIICHDEHGIGBFE@GCGDGG?D?G@BGHG@FHC GC;CC::8ABH>BECCBCB>;8ABCCC@A@#####

- | 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 |

并行质量控制(质控)实例

- 样本名列表从命令行管道传入，-j 2控制2个任务并行，红色为需要修改的部分

```
tail -n+2 result/metadata.txt|cut -f1|rush -j 2 \  
"kneaddata -i seq/{1}_1.fq.gz -i seq/{1}_2.fq.gz \  
-o temp/qc -v -t 3 --remove-intermediate-output \  
--trimmomatic ~/miniconda3/envs/kneaddata/share/trimmomatic/ \  
--trimmomatic-options 'ILLUMINACLIP:  
~/miniconda3/envs/kneaddata/share/trimmomatic/adapters/TruSeq2-  
PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \  
--reorder --bowtie2-options '--very-sensitive --dovetail' \  
-db ~/db/kneaddata/human_genome/hg37dec_v0.1"
```



2.2 HUMAnN2计算物种和功能组成

```
mkdir -p temp/humann2
```

- 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

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

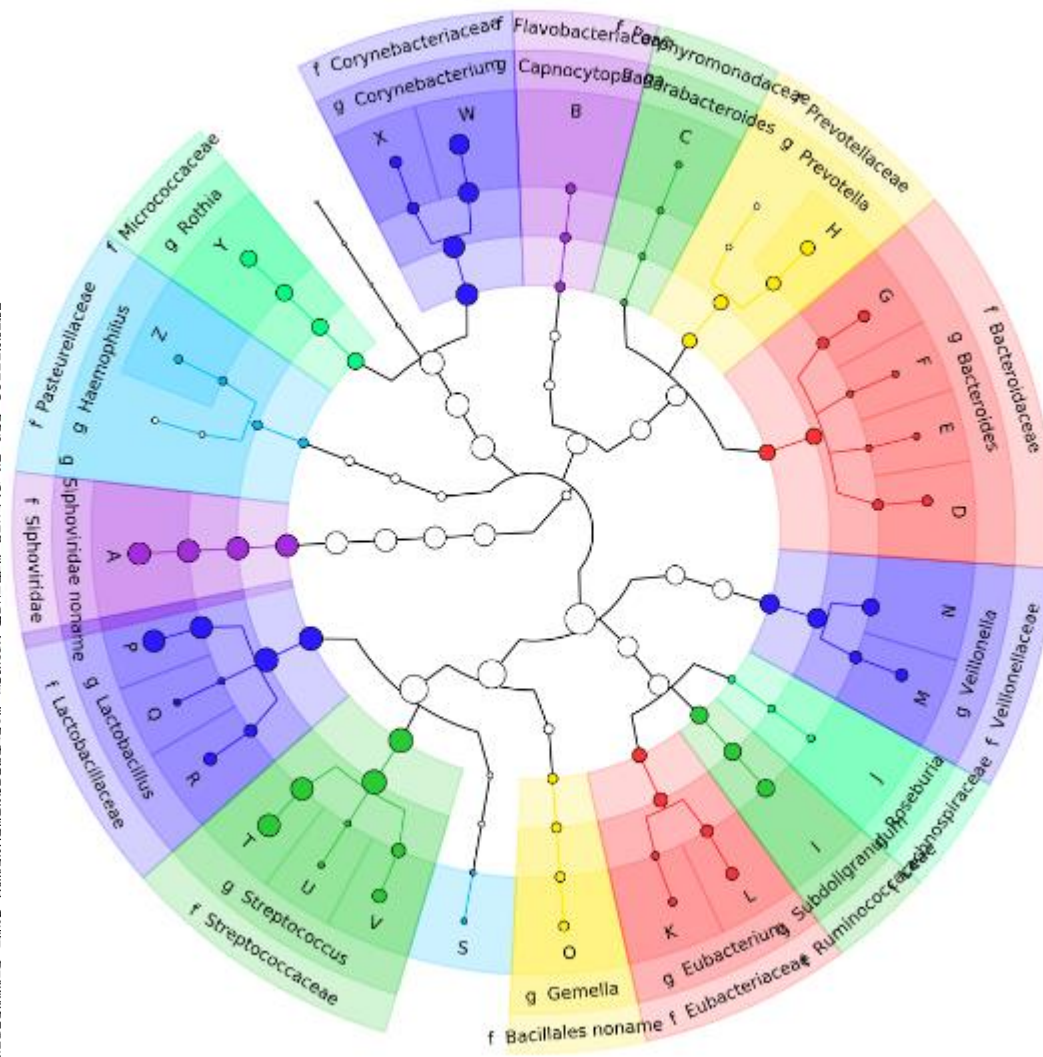
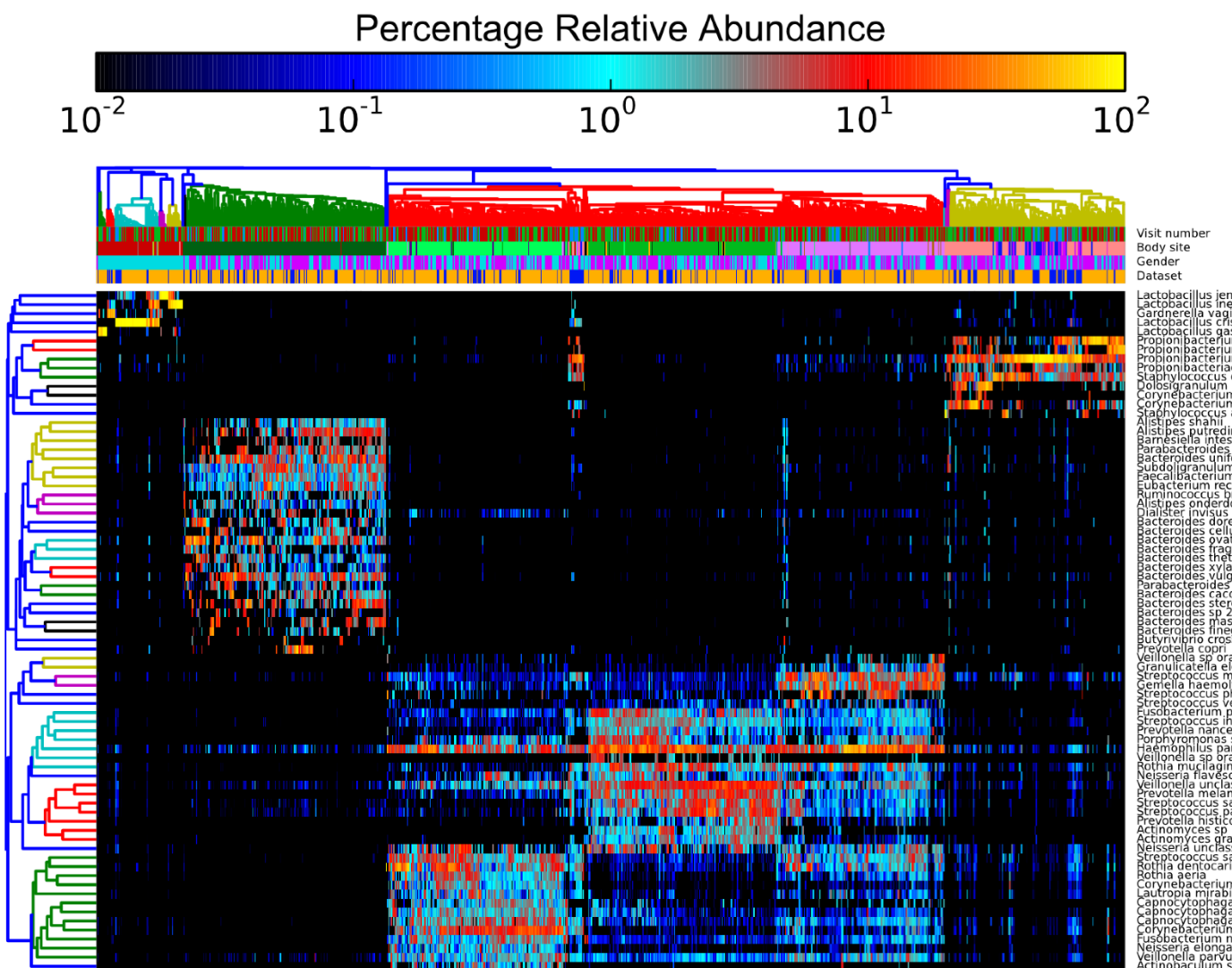
```
'humann2 --input temp/concat/{1}.fq \
```

```
--output temp/humann2/'
```

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



MetaPhlAn2结果常用展示方式

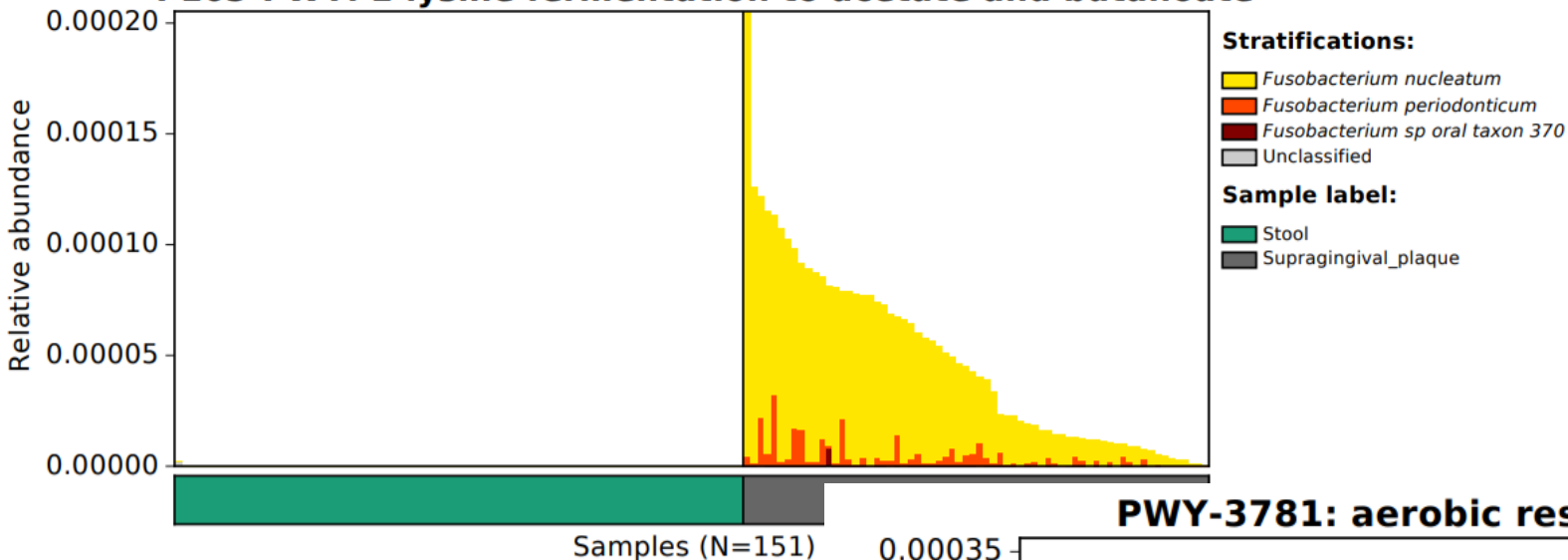


```
''' export2graphlan.py
```

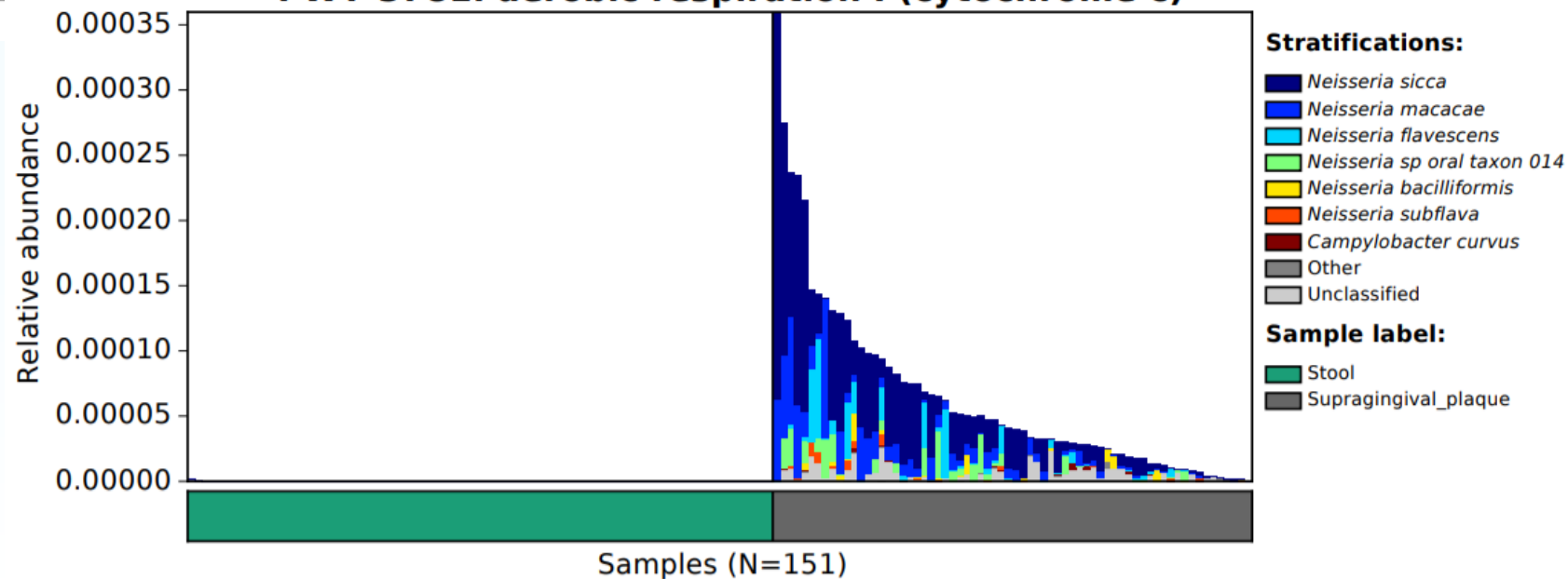
MetaPhlAn2一条命令获得宏基因组物种组成¹²

humann2_barplot绘制功能的物种组成

P163-PWY: L-lysine fermentation to acetate and butanoate

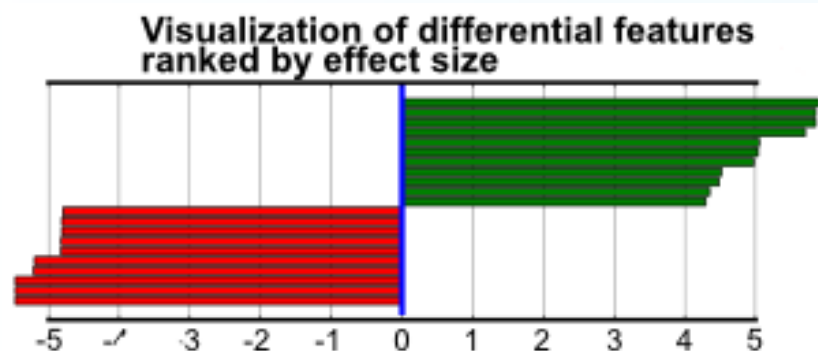


PWY-3781: aerobic respiration I (cytochrome c)

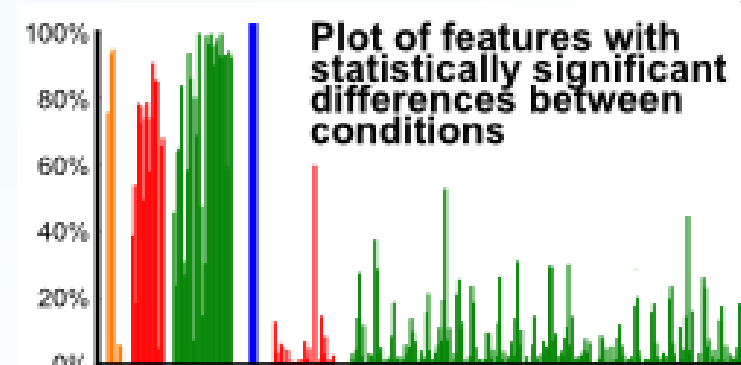
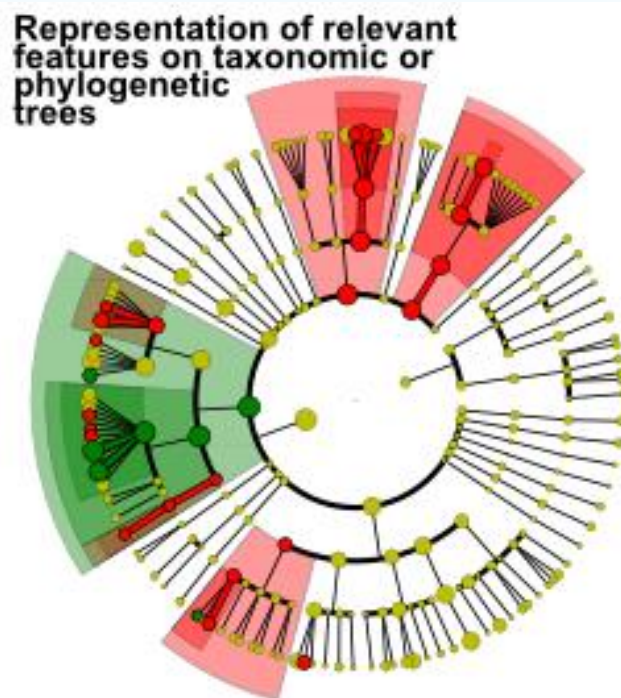


Linear discriminant analysis **Effect Size** (LEfSe)

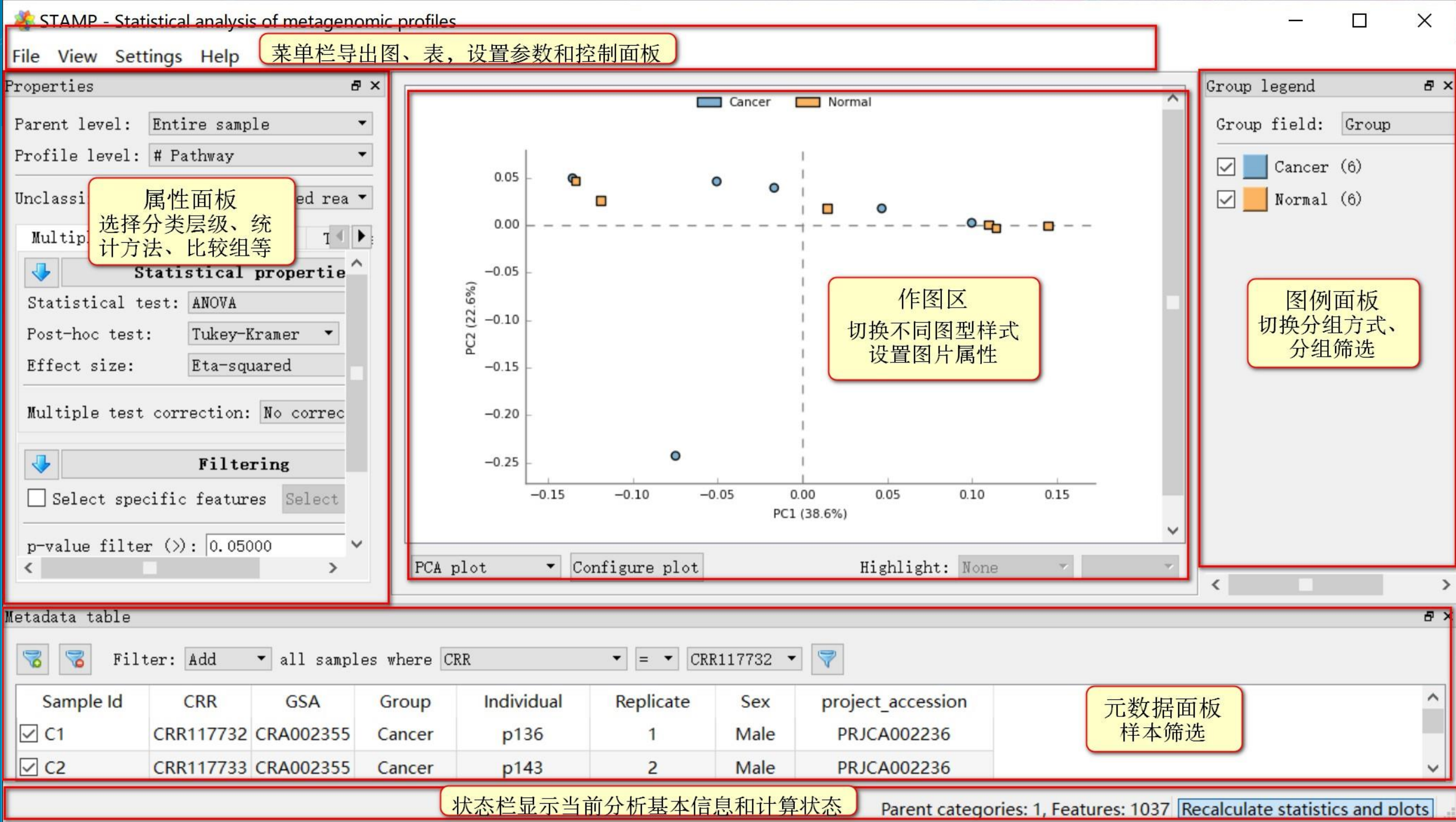
- LEfSe分析即LDA Effect Size分析，是一种用于发现和解释高维度数据生物标志(基因、通路和分类单元等)的分析工具，可以进行两个或多个分组的比较，它强调统计意义和生物相关性，能够在组与组之间寻找具有统计学差异的生物标志（Biomarker）。



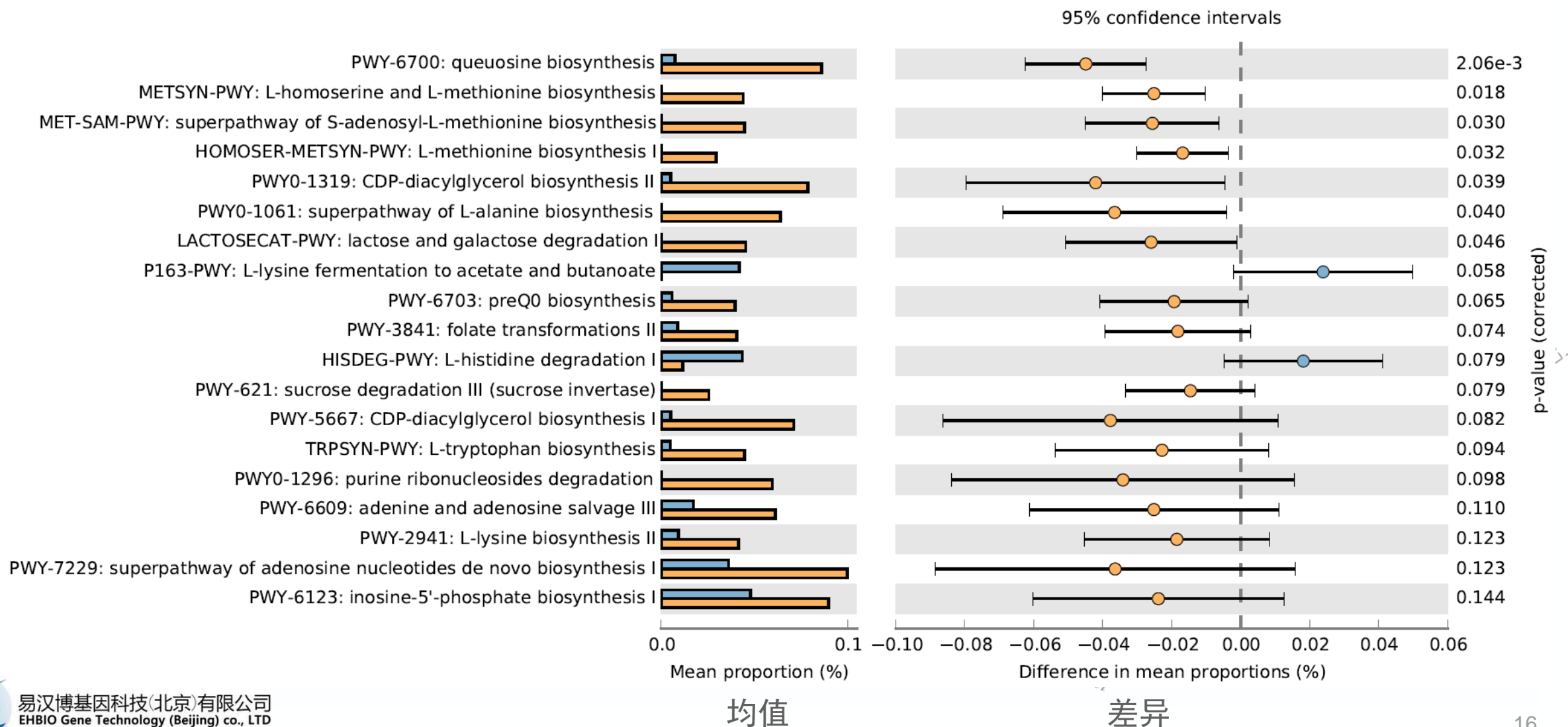
Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6), R60.



LDA分析、作图及添加置信-ggord



STAMP结果组间差异功能扩展柱状图



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/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

易生信
生信宝典
宏基因组



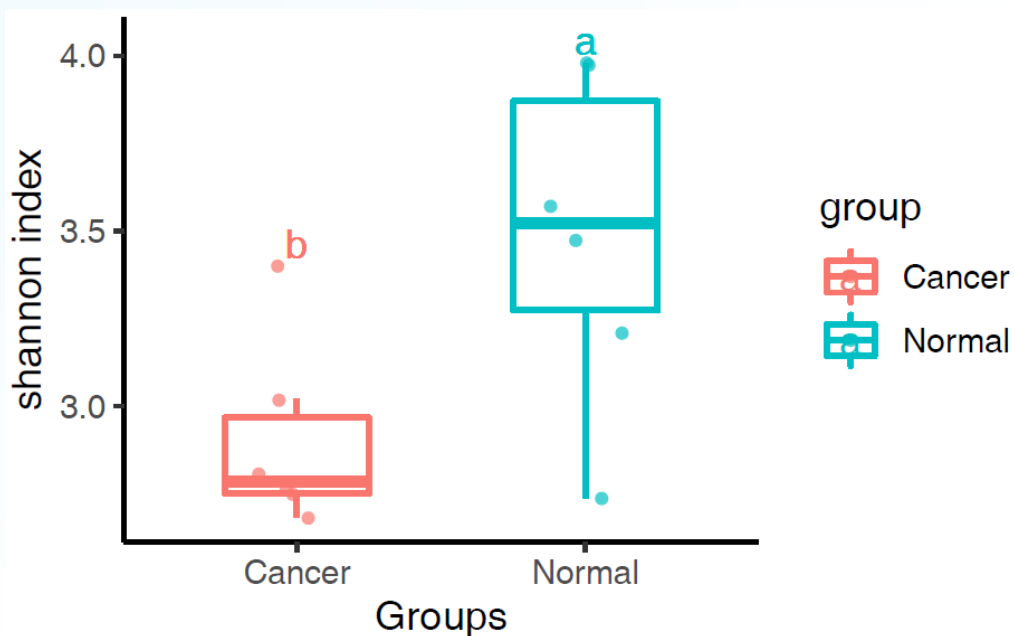
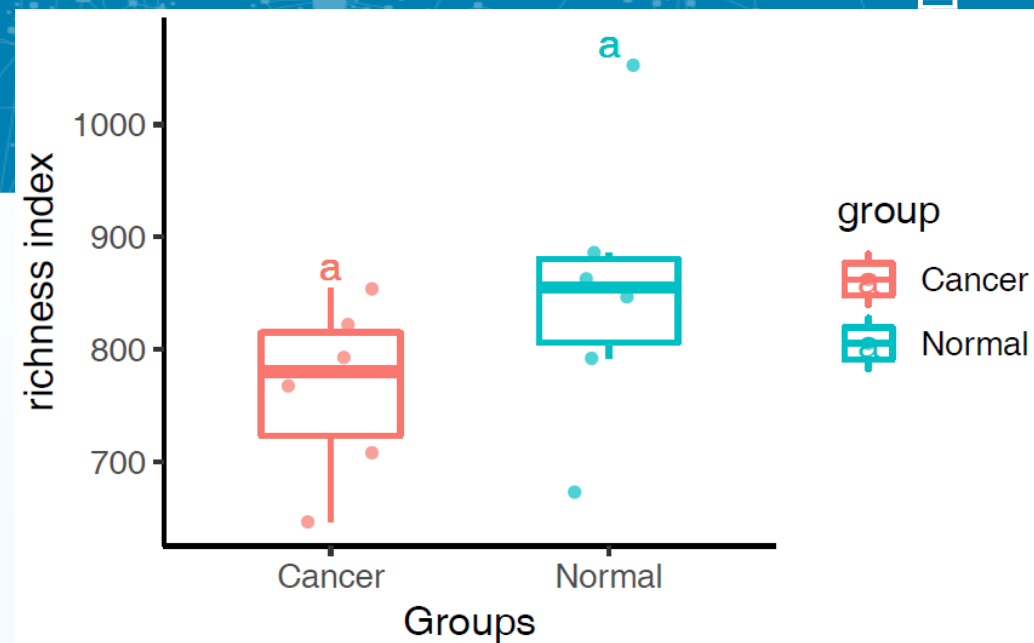
Kraken2物种多样性分析

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
```



- 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

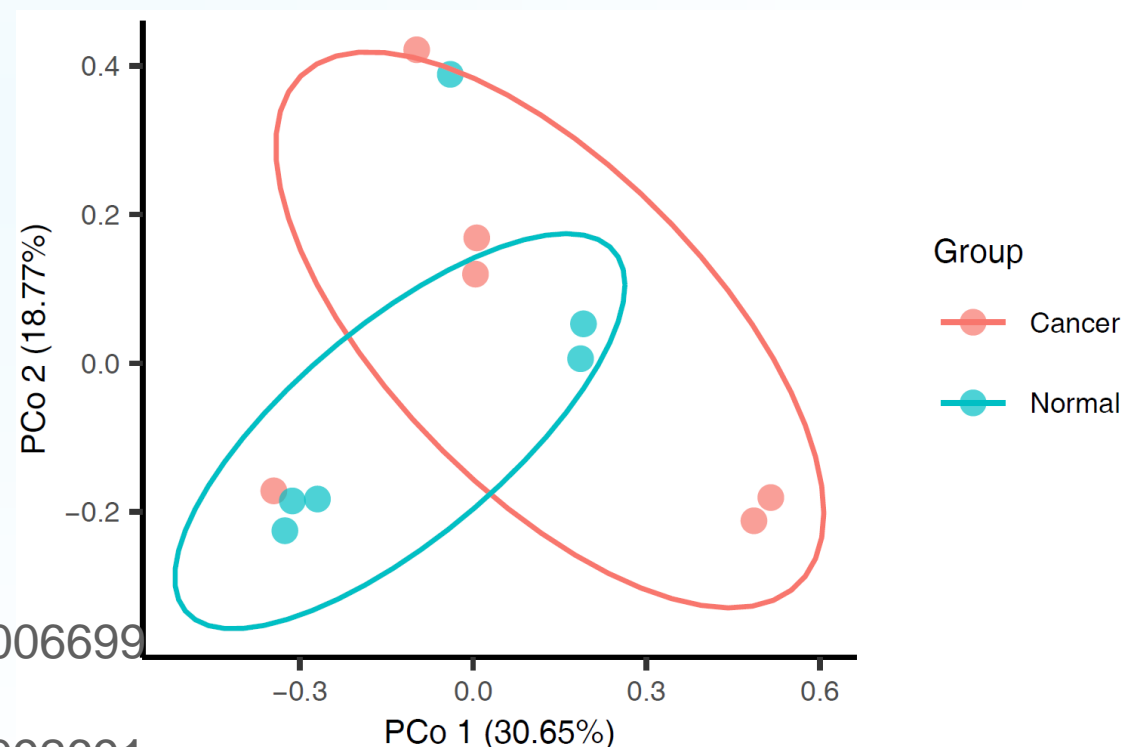
P值有波动但比较稳定

Sun Jan 03 16:19:07 2021

Cancer Normal 0.300669933006699

Sun Jan 03 17:55:04 2021

Cancer Normal 0.309269073092691



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

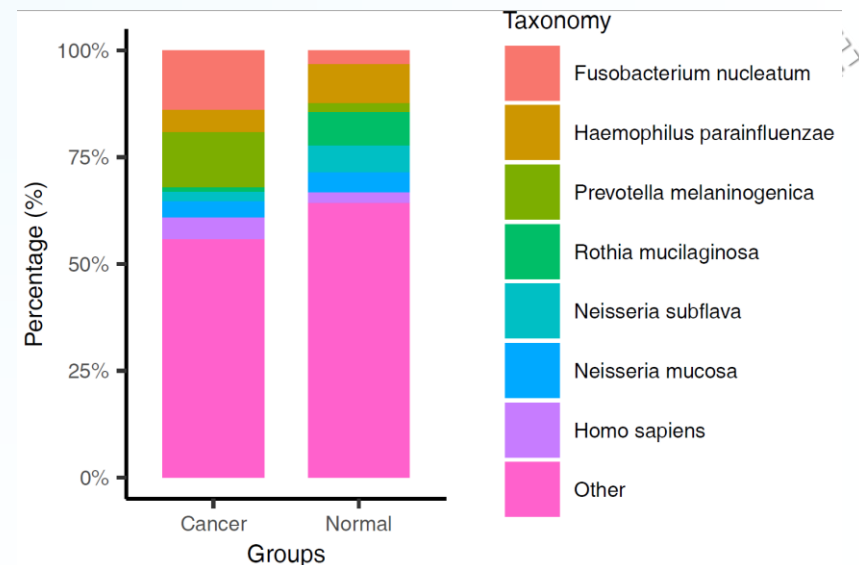
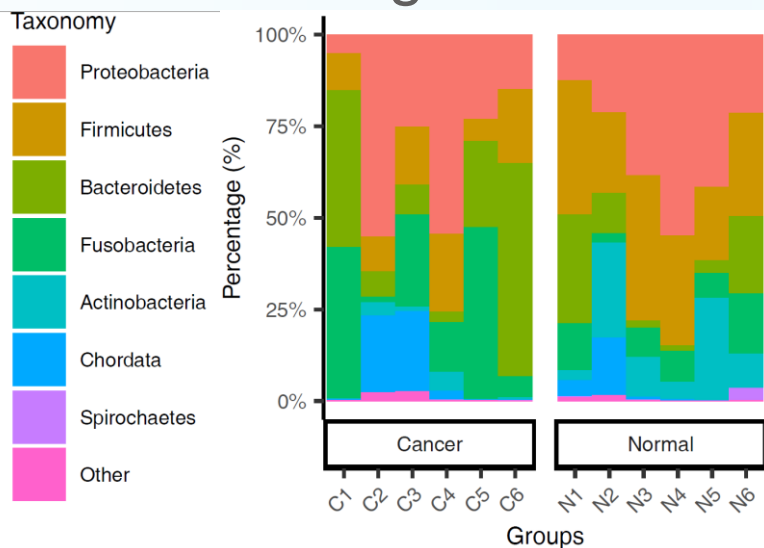
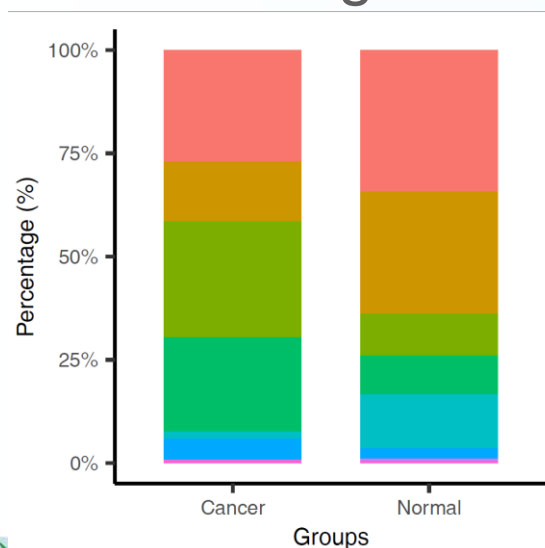
tax=S

```
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

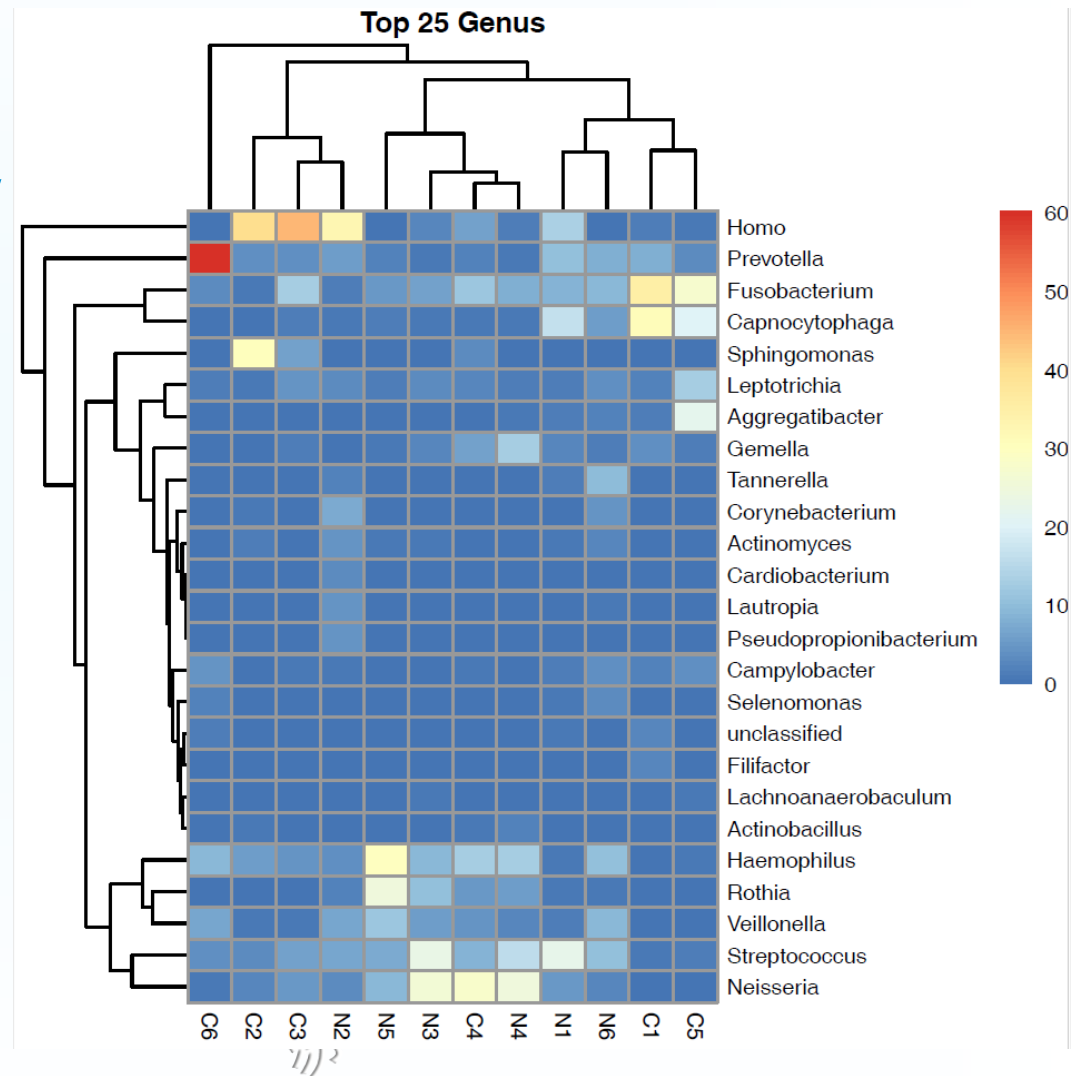
db/script/metaphlan_hclust_heatmap.R \

-i result/kraken2/tax_count.spf \

-t **Genus** \

-n **25** \

-o result/kraken2/heatmap_Genus



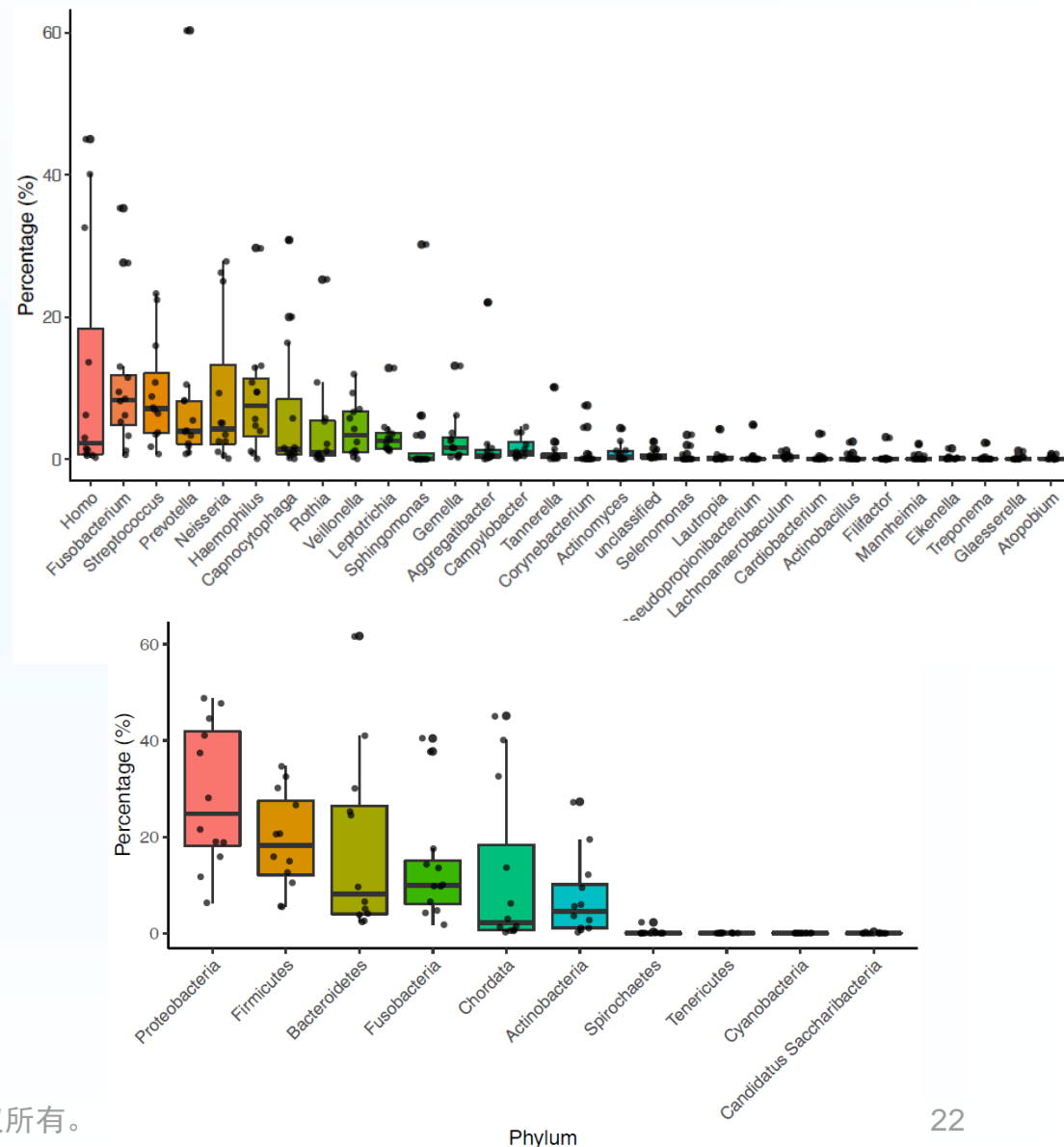
物种组成——箱线图

○ # 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
```

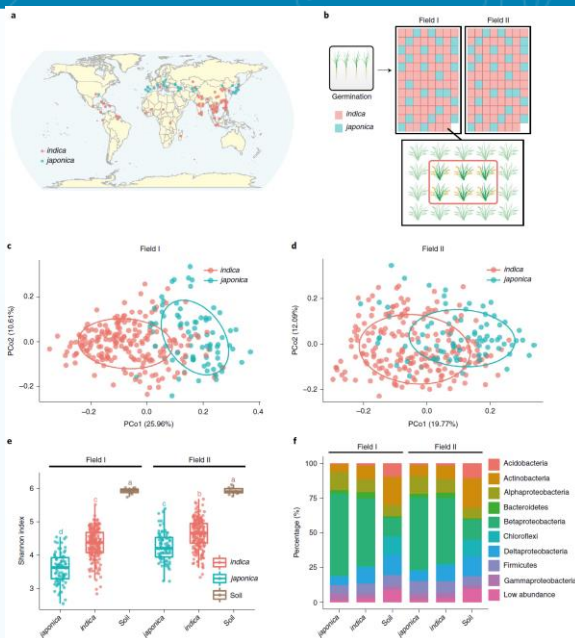


高水平文章发表前三部曲

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

