中国土壤学会土壤生物和生化专业委员会暨
土壤生物多样性与生物化学过程研讨会

微生物组数据分析和结果解读

刘永鑫
中科院遗传发育所 高工 青促会会员
宏基因组公众号 创始人
2021年6月23日

中国科学院遗传与发育生物学研究所
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences

自我介绍

学士 微生物学
硕士 遗传学

博士 生物信息
博后 遗传学
高工 微生物组

- 以第一/通讯作者(共同)在***Nature Biotechnology***、***Nature Protocols***、***Current Opinion in Microbiology***等发表论文12篇，在***Science***、***Cell Host & Microbe***等杂志合作发表论文14篇，累计被引用3986次(Google学术，截止2021/6/22)。主持国自然、中科院项目2项，参与3项；共同主编在编专著2部；获得软件著作权1项；参与申请国内专利3项；兼职mSystems、Bioinformatics、JGG、BMC系列等10余个期刊审稿人。2017年创办宏基因组公众号，分享原创文章2千余篇，关注人数近12万，阅读量累计2100万+。

近年合作发表的文章

Science Contents News Careers Journals

SHARE RESEARCH ARTICLE

A specialized metabolic network selectively modulates *Arabidopsis* root microbiota

https://doi.org/10.1126/science.aau6389

nature biotechnology ARTICLES

https://doi.org/10.1038/s41587-019-0104-4

CORRESPONDENCE

https://doi.org/10.1038/s41587-019-0209-9

nature biotechnology

nature protocols PROTOCOL

https://doi.org/10.1038/s41596-020-00444-7

Cell Host & Microbe

https://doi.org/10.1016/j.chom.2020.03.004

SCIENCE CHINA Life Sciences

https://doi.org/10.1007/s11427-019-9521-2

Protein & Cell

Presenting Novel Discoveries in Biological Sciences

https://doi.org/10.1007/s13238-020-00724-8

Current Opinion in Microbiology

https://doi.org/10.1016/j.mib.2019.10.010

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb

www.sciencedirect.com

https://doi.org/10.1016/j.gpb.2014.04.003

Chinese Medical Journal 中华医学杂志英文版

https://doi.org/10.1097/cm9.0000000000000871

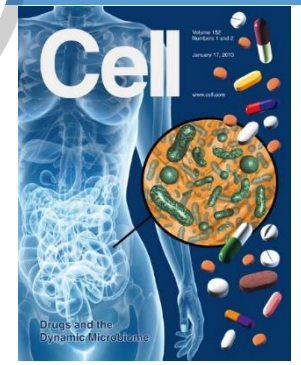
遗传 Hereditas (Beijing)

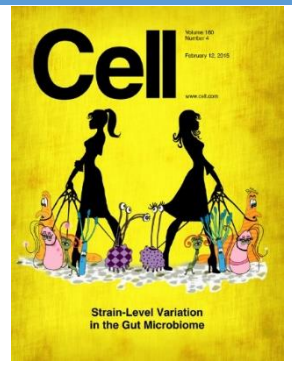
https://doi.org/10.16288/j.yczz.19-222

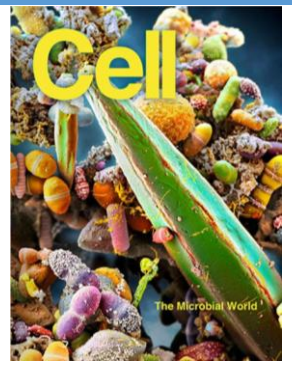
CEPAMS

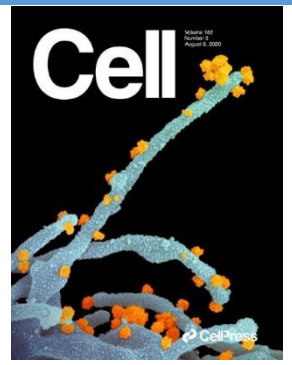
刘永鑫：想学菌群生物信息分析-21分钟带你入门

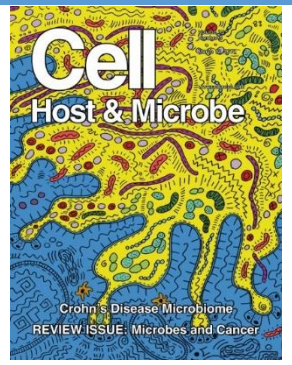
为什么微生物组这么热?

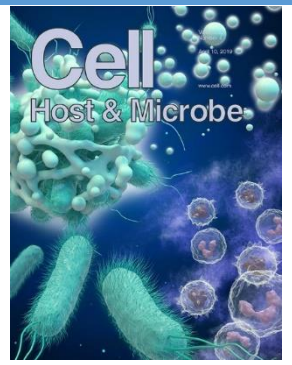


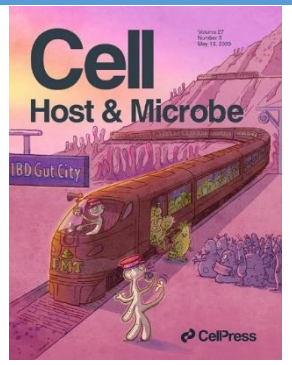


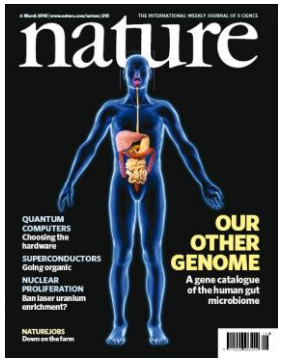


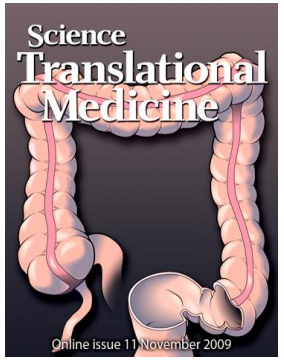






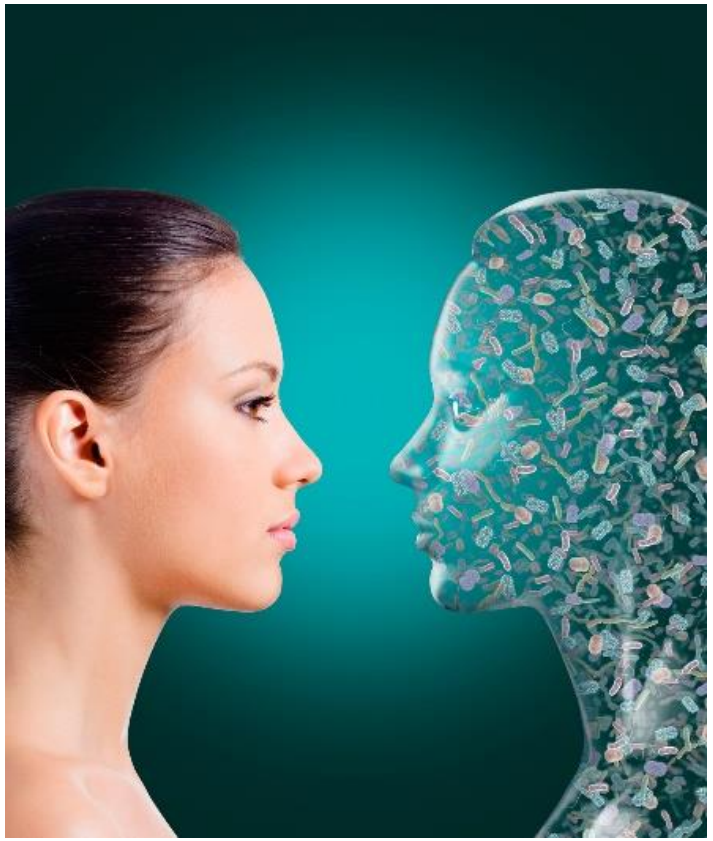


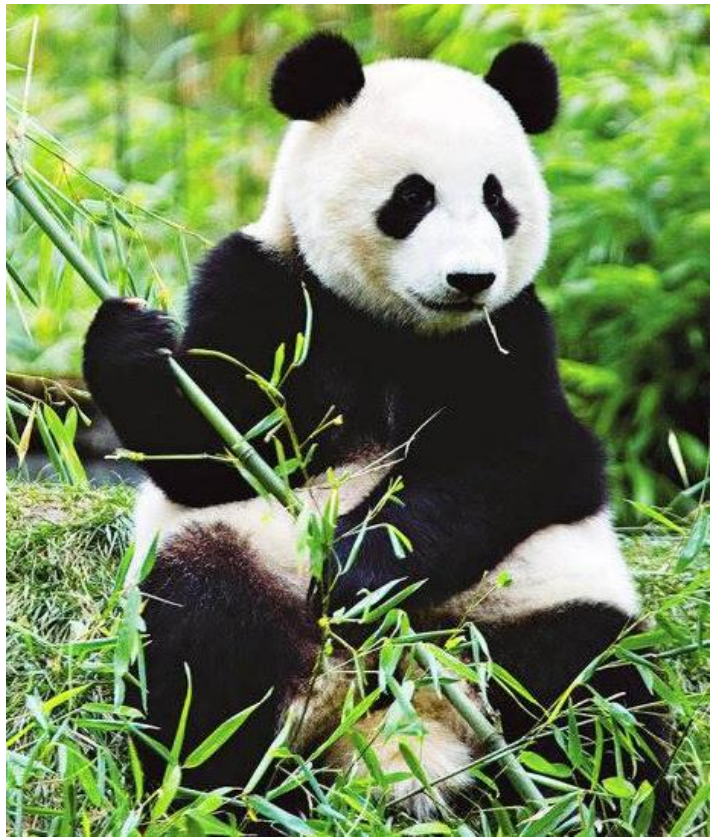




CEPAMS

动植物生活环境中充满微生物





脱离了微生物的生物学研究是不完整的

报告提纲

- 读懂文章图表
- 扩增子分析流程
- 宏基因组分析流程
- 多样性分析和可视化
- 高分文章套路

报告提纲

- **读懂文章图表**
- 扩增子分析流程
- 宏基因组分析流程
- 多样性分析和可视化
- 高分文章套路

CEPAMS

读懂文章图表

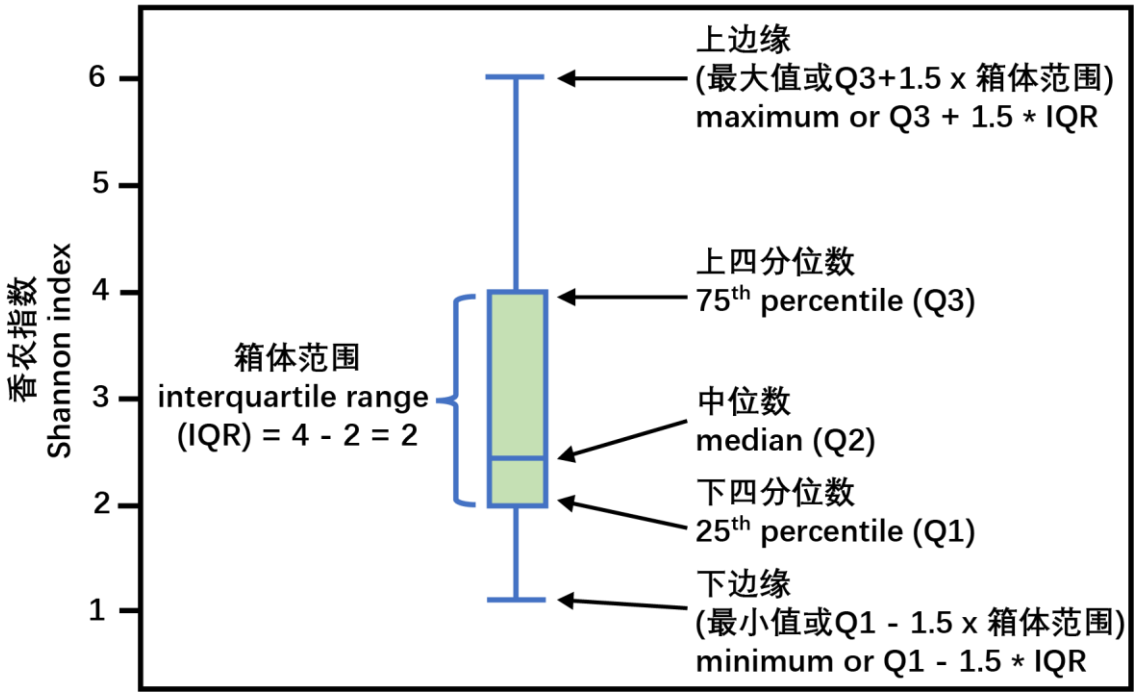
- Alpha多样性——箱线图、折线图、维恩图
- Beta多样性——热图、散点图、箱线图
- 物种组成——堆叠柱状图、冲击图、圈图
- 差异特征——热图、火山图、曼哈顿图、三元图
- 多组比较——维恩图、集合图、桑基图
- 共存相关——corrplot、相关热图、网络图

CEPAMS

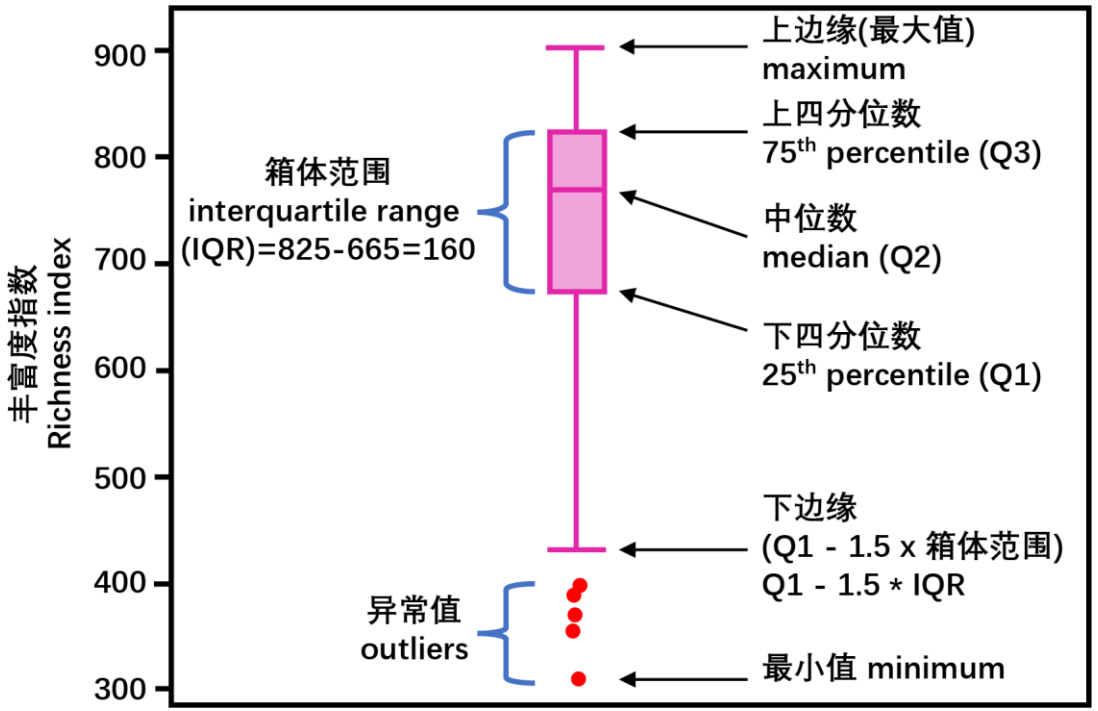
•扩增子图表解读-理解文章思路

箱线图的基本知识

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



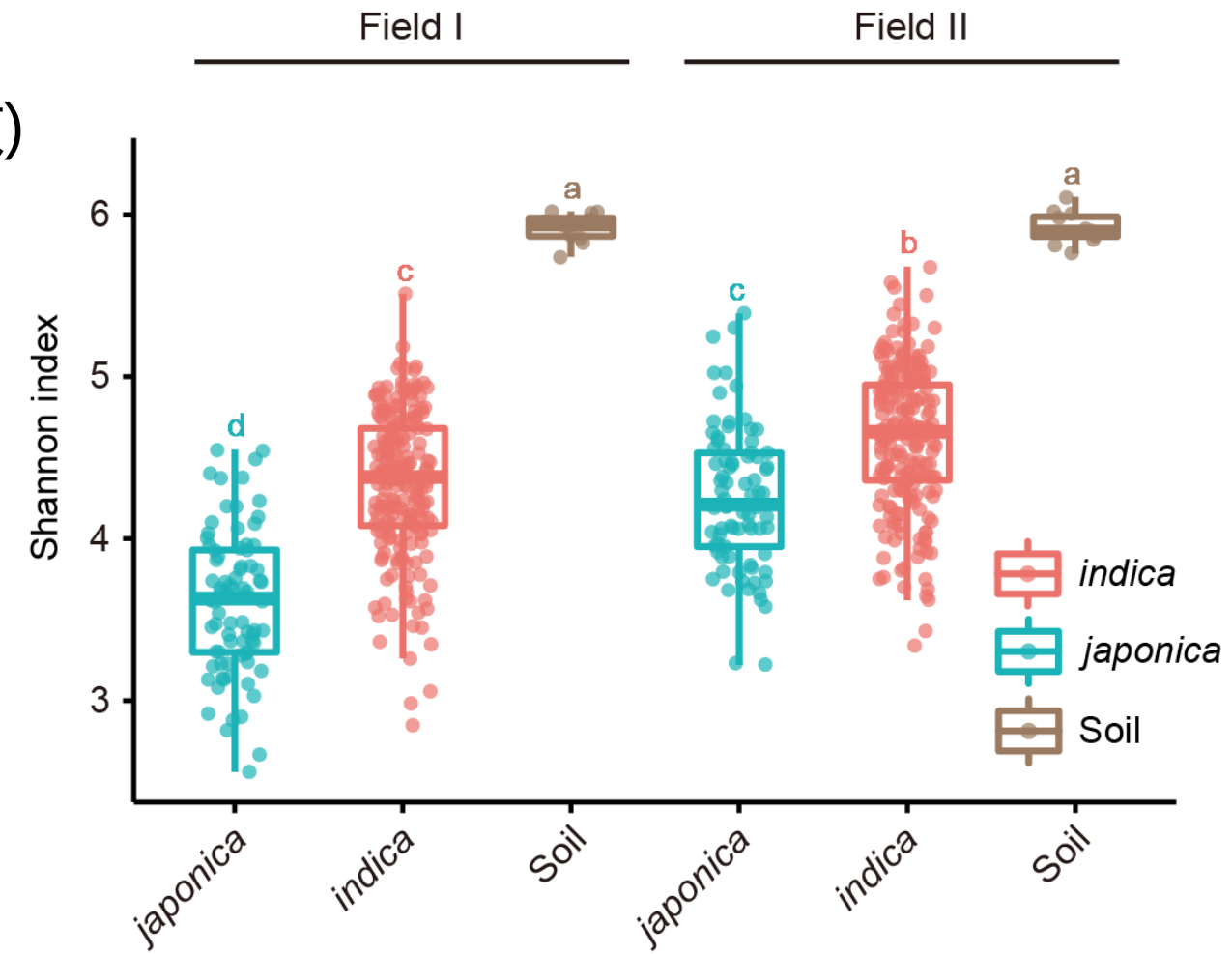
Alpha多样性丰富度指数箱线图(Boxplot of richness index)



[Alpha多样性箱线图(样章，11图2视频)]

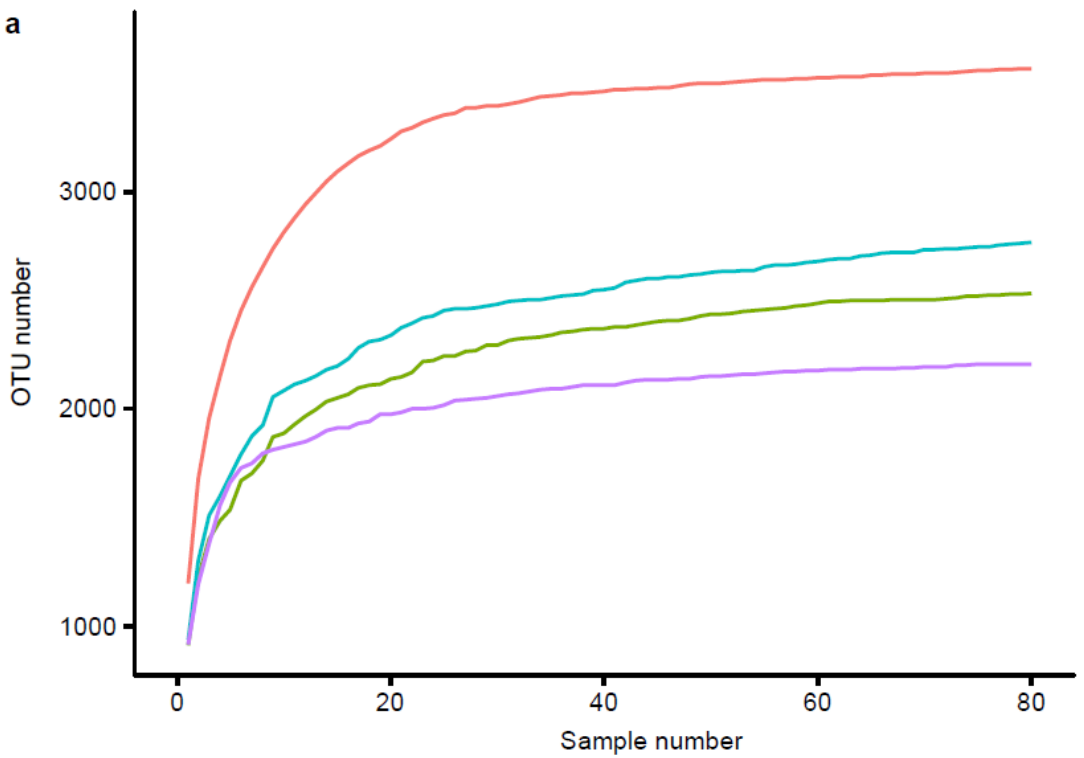
亚种和土壤在两地的香农指数

箱体上中下线分别为75、50(中位数)和25分位数，轴须线最长不超过1.5x箱体范围。
字母用于区分组间是否存在显著区别，不同字母表示组间存在显示差异($P < 0.05$，ANOVA, Tukey-HSD test)。
图中的样本量如下：地块1：籼稻 ($n = 201$), 粳稻 ($n = 80$), 土壤 ($n = 12$); 地块2, 籼稻 ($n = 201$), 粳稻 ($n = 81$), 土壤 ($n = 12$)

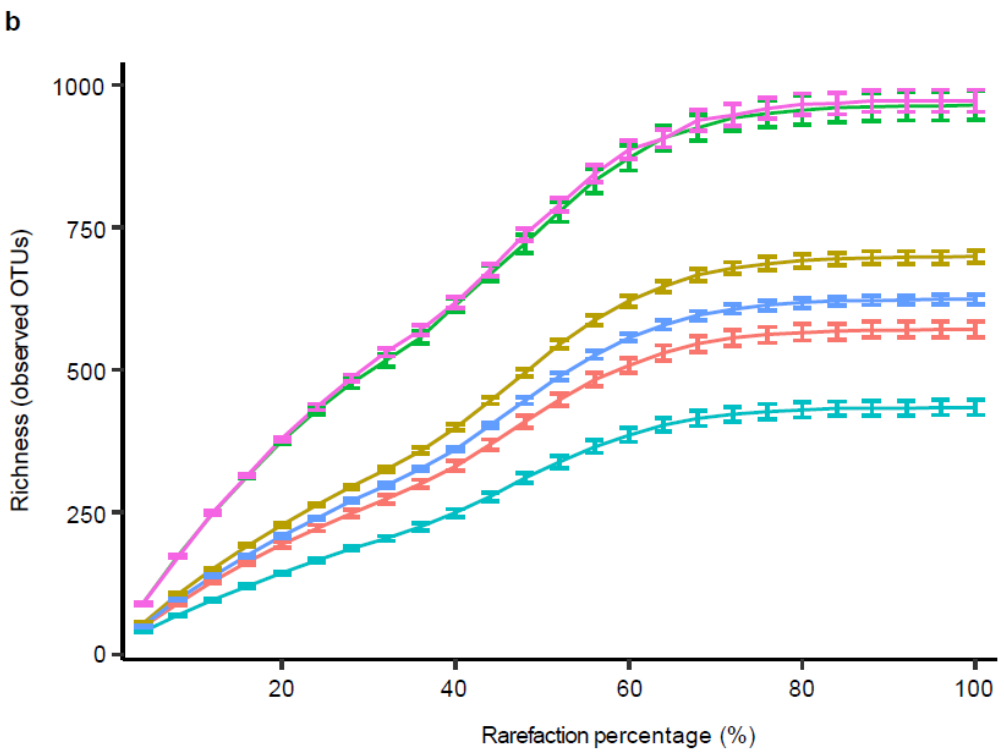


- NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用
- Alpha多样性箱线图(样章，11图2视频)

稀释曲线展示OTUs/样本饱和度



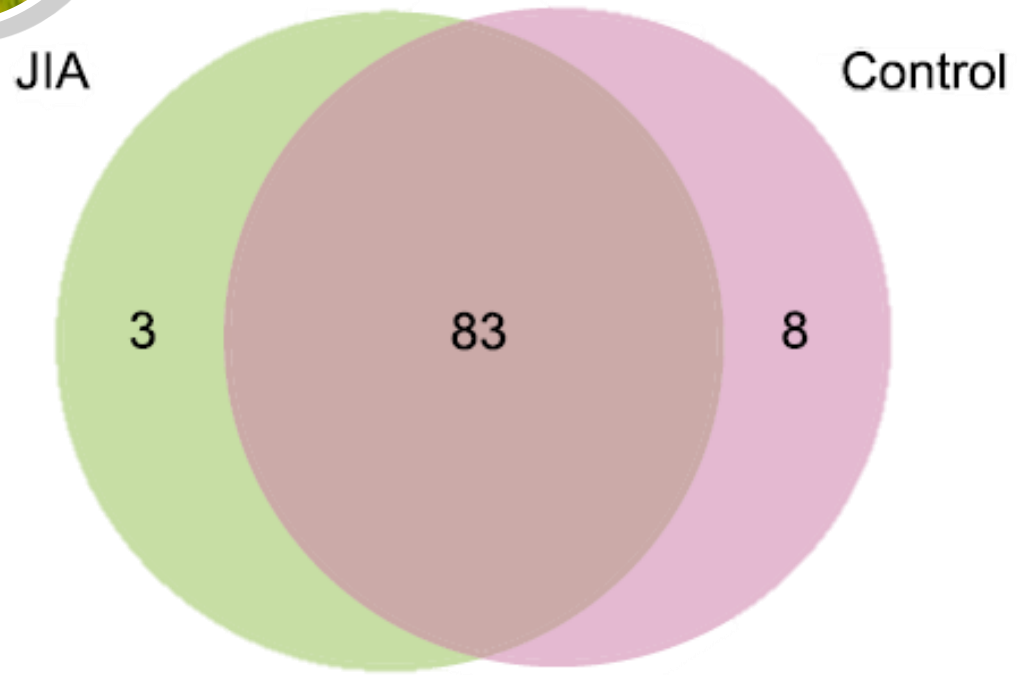
样本 vs OTUs稀释曲线-显示样本量充足



测序量百分比 vs OTUs稀释曲线+标准误
显示测序量充足和组间差异

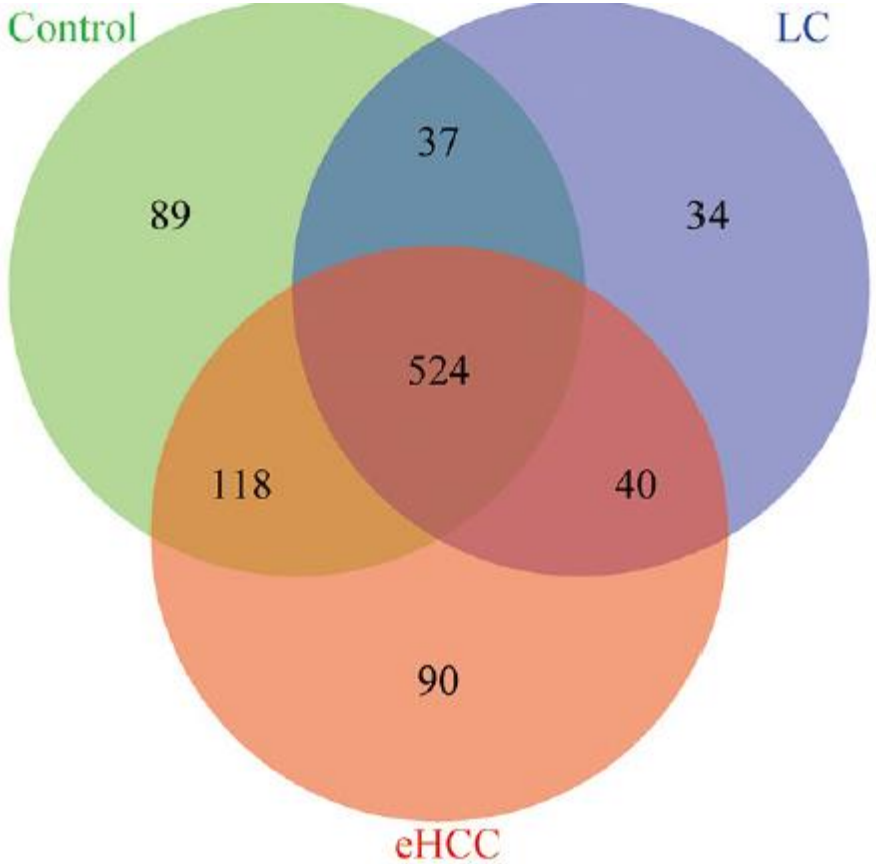
•212.样本量和测序深度的Alpha多样性稀释曲线

维恩图展示组间共有/特有



维恩图显示了两组有83个属是相同的，但是JIA组有3个属是独有的，对照组有8个属是独有的

BMC：幼儿关节炎患儿肠道菌群的特征



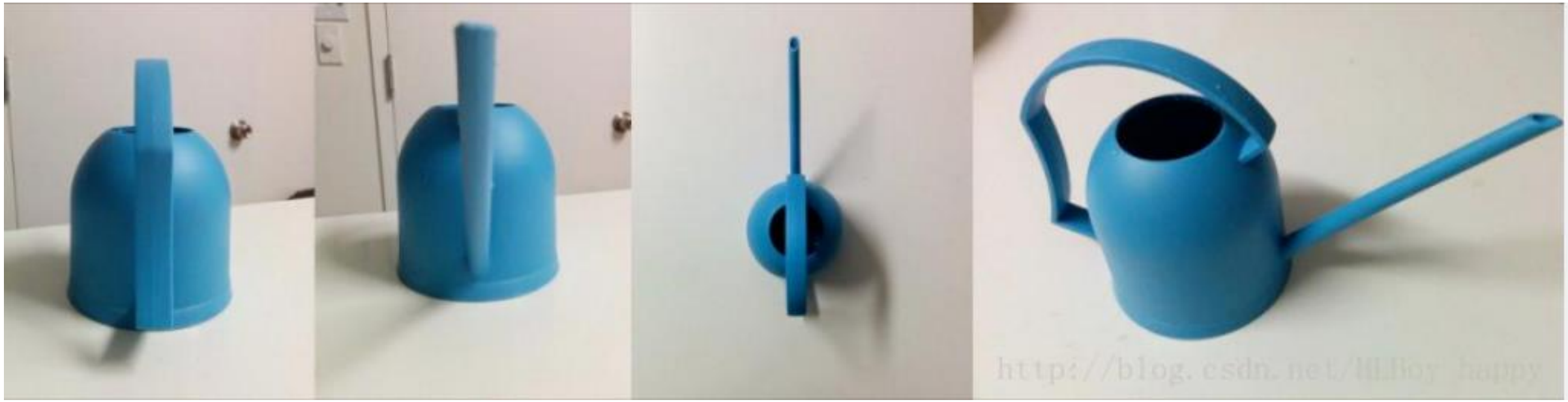
维恩图展示了Control健康人、LC肝硬化、eHCC早期肝癌三组间独有和共有的OTU

Gut：早期肝癌肠道生物标志物鉴定

•213.维恩图和集合图(Venn&Upset)

排序分析基本思想-降维(举个栗子)

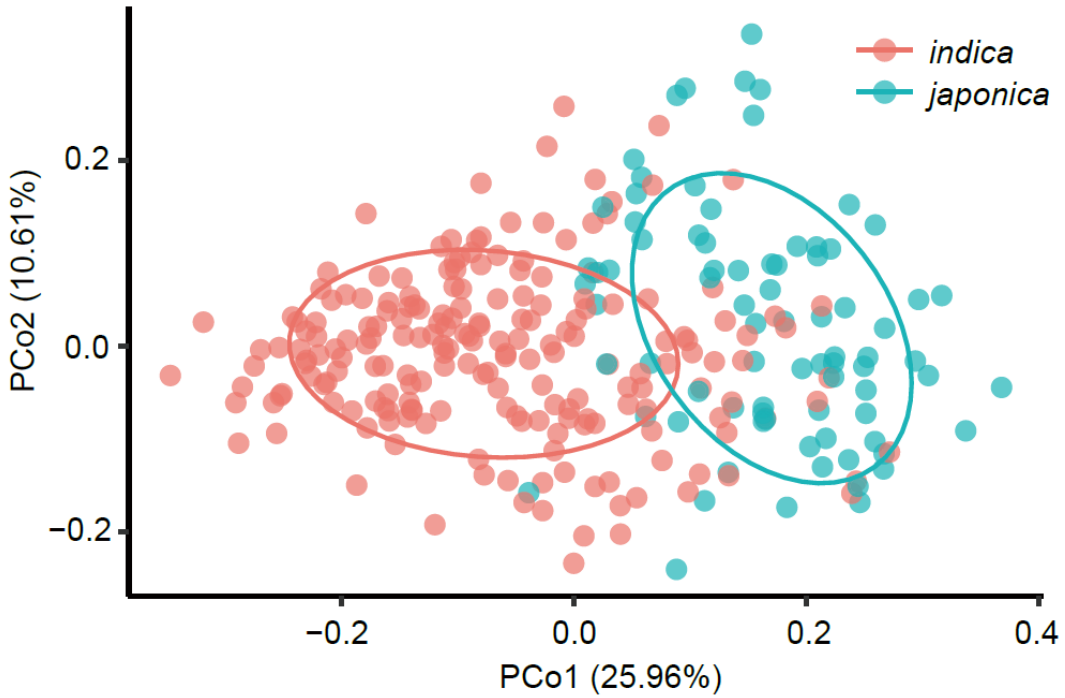
假如你是一本养花工具宣传册的摄影师，你正在拍摄一个水壶。水壶是三维的，但是照片是二维的，为了更全面的把水壶展示给客户，你需要从不同角度拍几张图片。下图是你从水壶背面、正面、正上方、斜上方的照片：



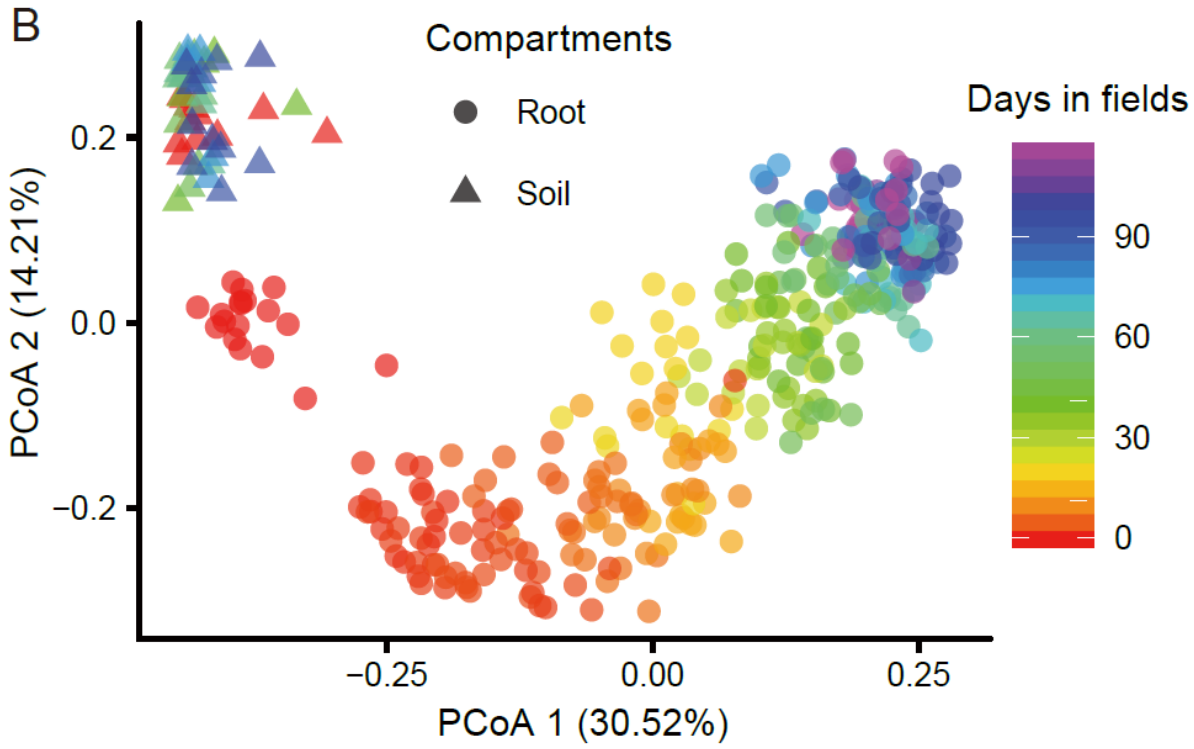
主成分 / 主坐标分析的基本思想：用2-3维展示高维数据的主体

•221.Beta多样性PCoA和NMDS排序

主坐标分析(PCoA)展示样本主要差异



两组间明显的微生物组差异

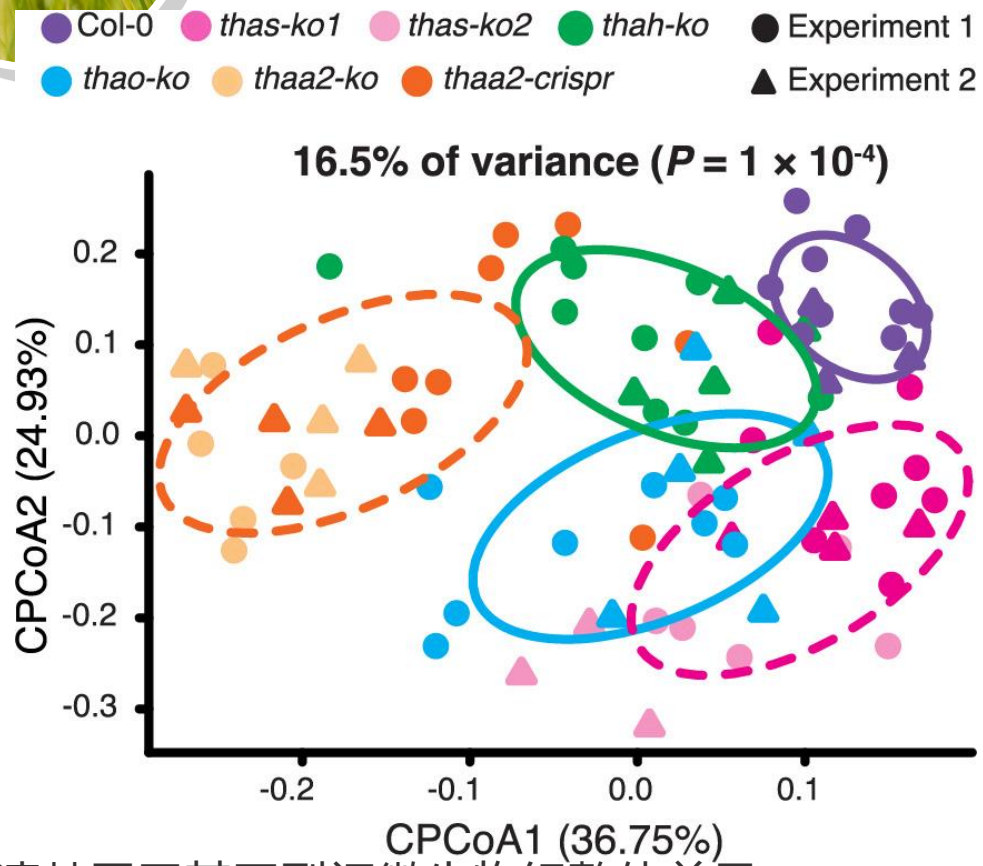


微生物组结构在时间上动态变化

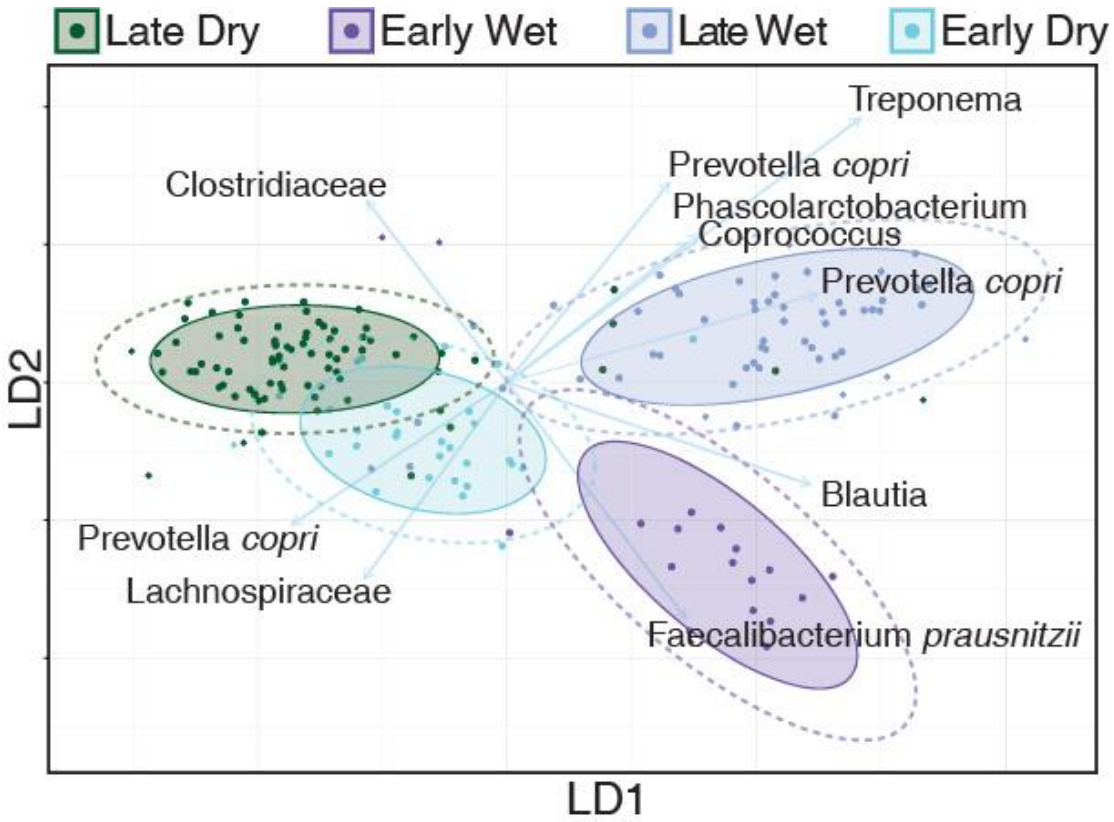
CEPAMS

•221.Beta多样性PCoA和NMDS排序

限制性排序CPCoA / LDA



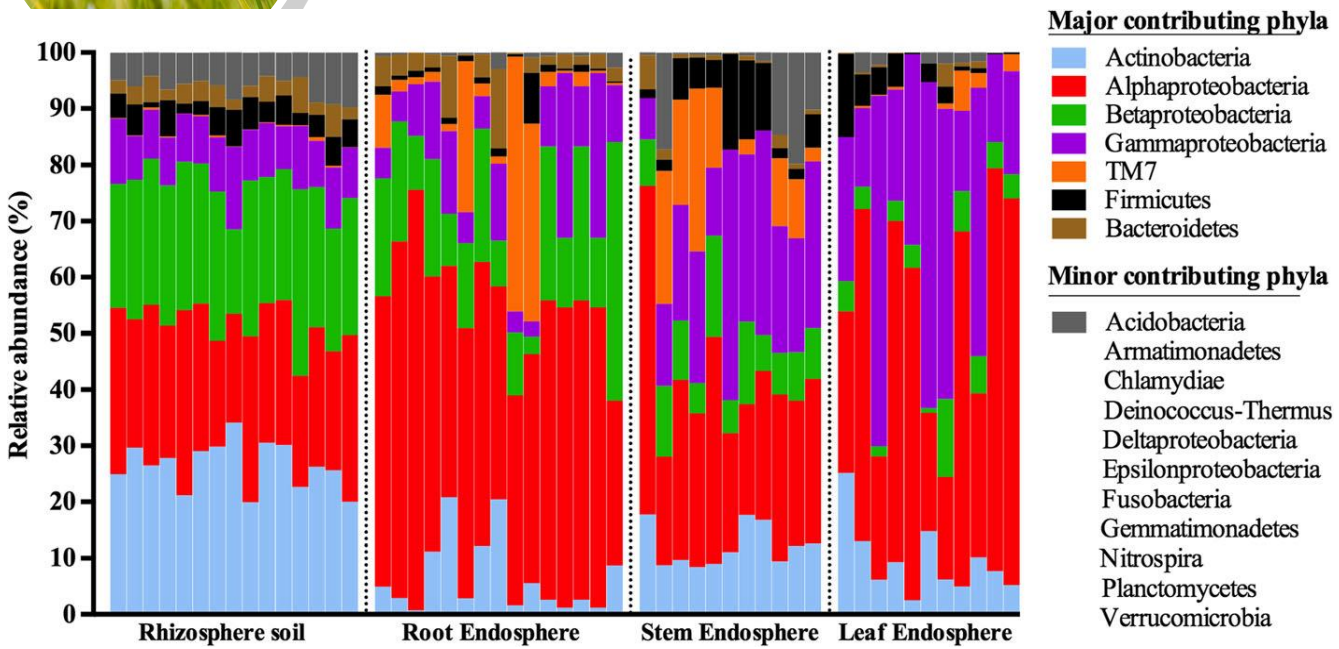
更清楚展示基因型间微生物组整体差异(Science, 2019)



展示季节间微生物组整体差异(Science, 2017)

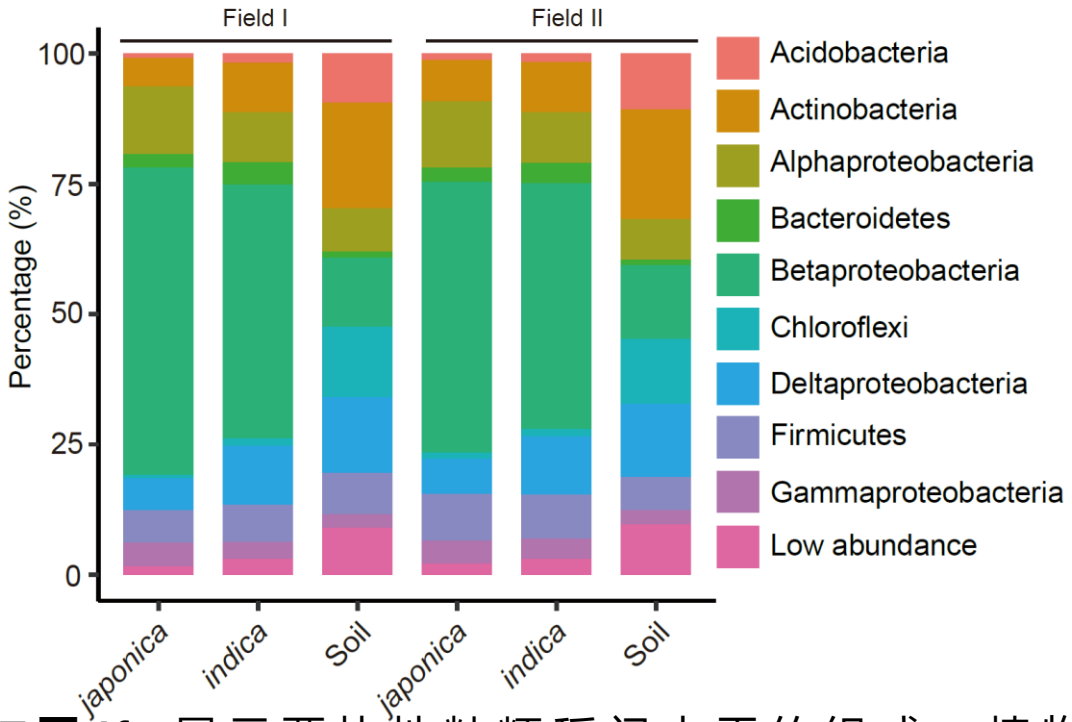
•222.Beta多样性限制性排序CPCoA/CCA/RDA/LDA

物种组成堆叠柱状图



Microbiome图4. 各样品OTU在门水平分类的相对丰度柱状图
图中展示了每个Compartment的每个样品门水平的相对丰度；其中Proteobacteria由于组比较大，也将其分成了apha，beta，gamma，delta，epislon五类展示；高丰度的前7类用彩色显示，其它低丰度的门用灰色显示。

Microbiome: 简单套路发高分文章--杨树微生物组



NBT 图 1f. 展示两块地籼粳稻门水平的组成。植物中Proteobacteria较大，进一步拆分为Alpha/Beta/Delta/Gemma纲。图中的样本量如下：地块1：籼稻 (*n* = 201), 粳稻 (*n* = 80), 土壤 (*n* = 12); 地块2, 籼稻 (*n* = 201), 粳稻 (*n* = 81), 土壤 (*n* = 12)。

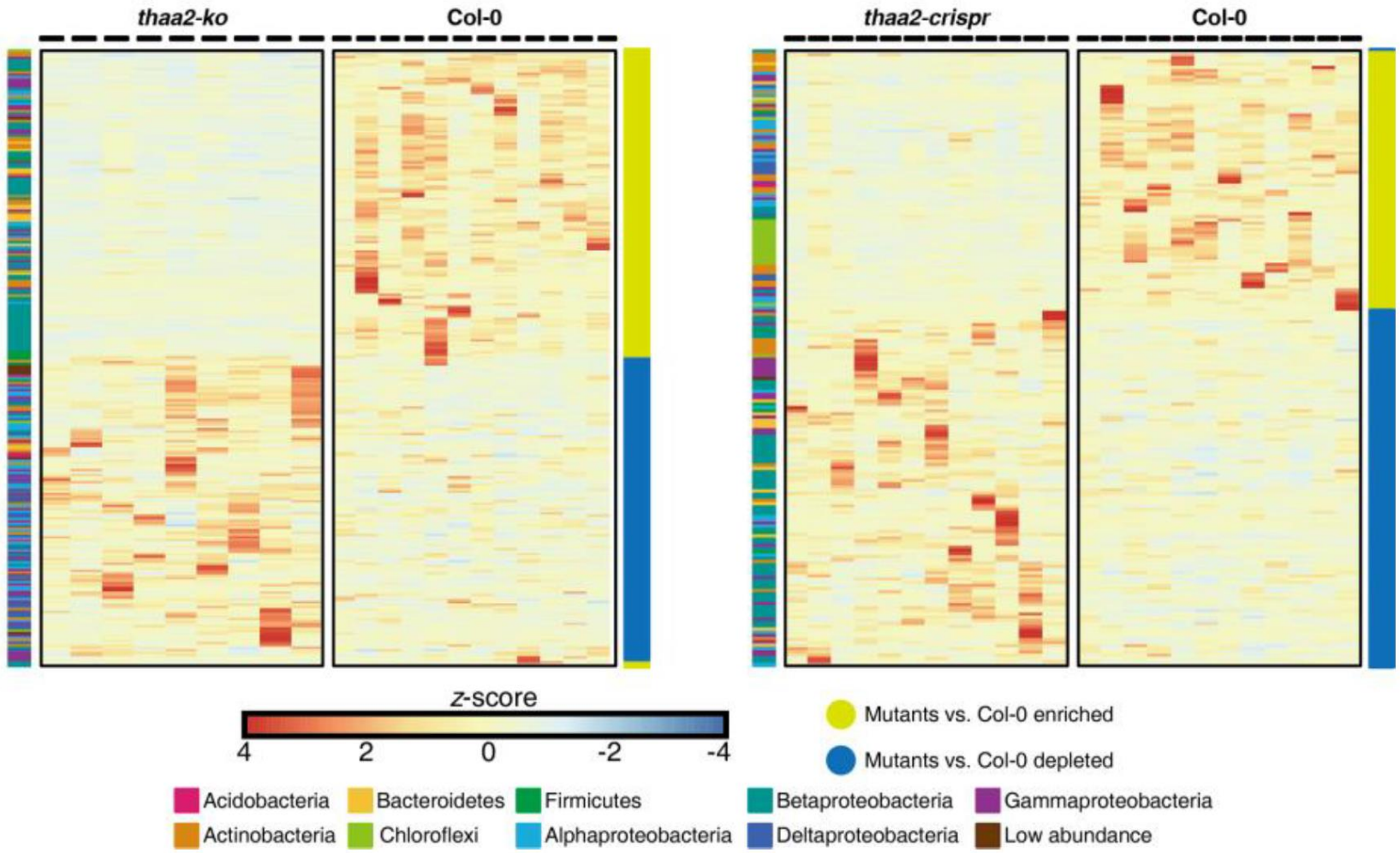
NBT：水稻基因调控根系微生物组参与氮利用

CEPAMS

•231.菌群物种组成堆叠柱状图、弦图、词云

热图展示组间差异

pheatmap包绘制
呈现每个样品的细节，整体的一致性，并配合左、右侧注释物种分类和变化类型
z-score水平标准化
(x-μ)/σ(减均值除方差)，呈现组间差异



Science：拟南芥三萜化合物特异调控根系微生物组

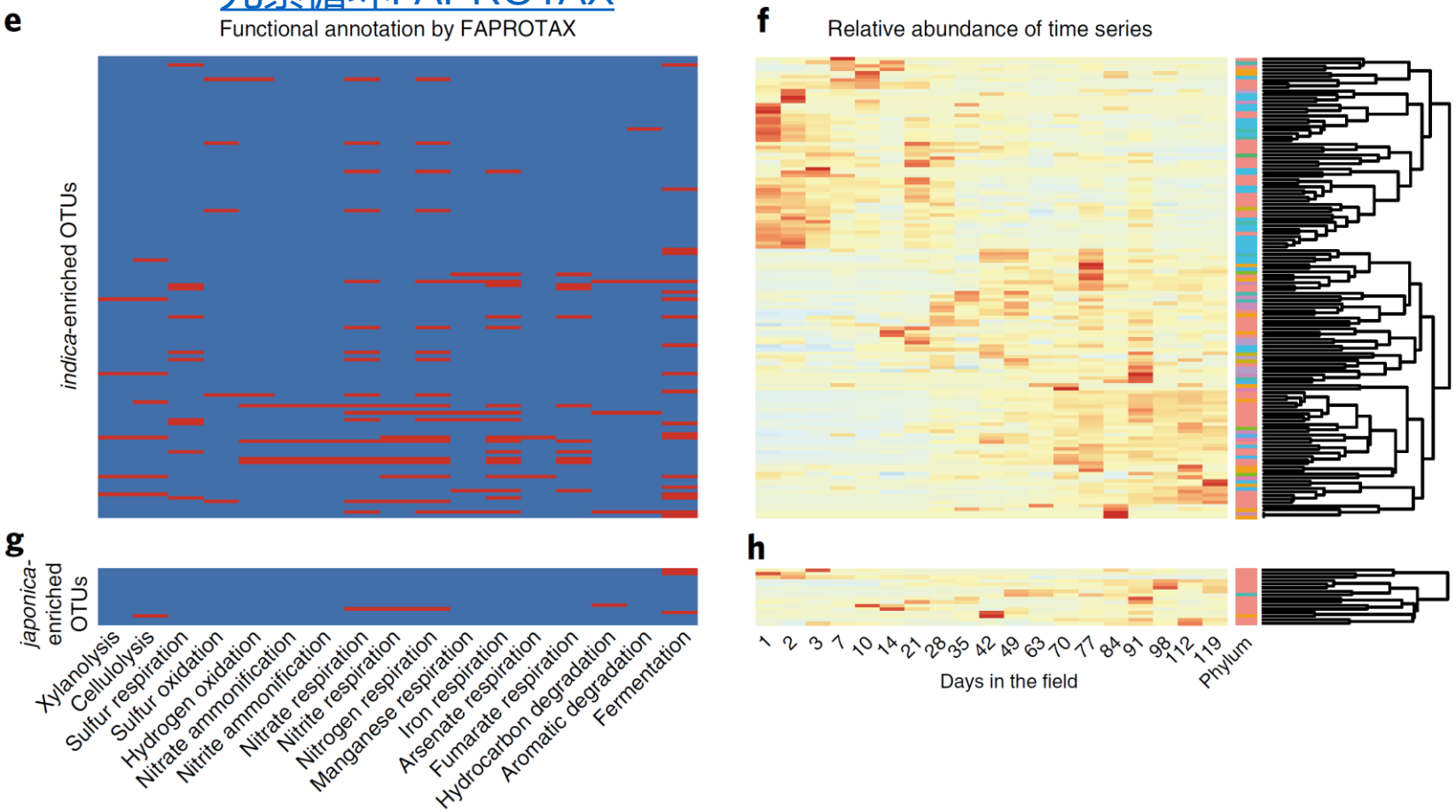
CEPAMS

热图展示功能有无和时间序列

pheatmap包绘制

双色热图展示每个OTUs中是否存在某种功能

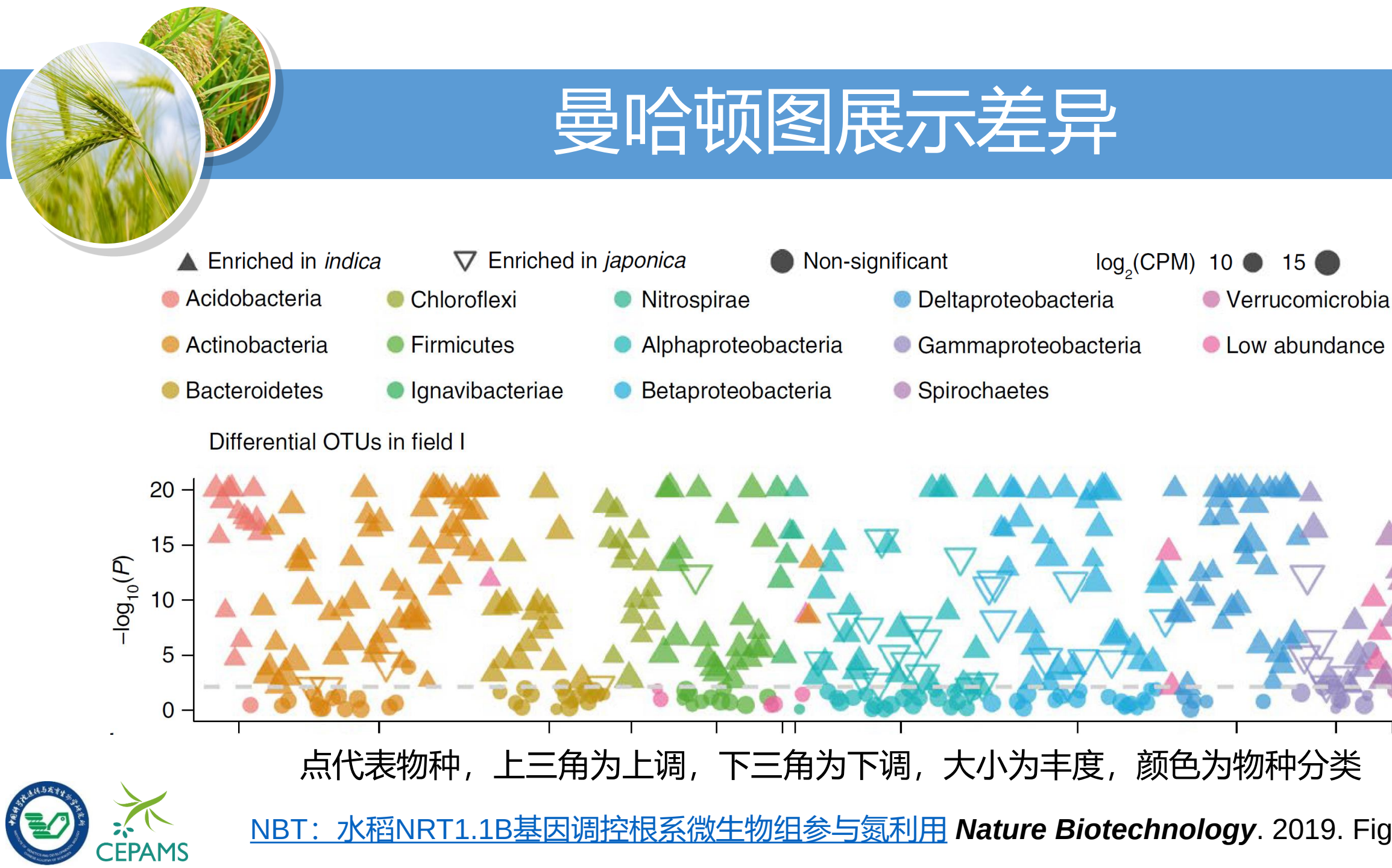
元素循环FAPROTAX



热图Z-Score展示每个OTUs在时间序列中的动态变化，方便观察丰度富集的时期

NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用

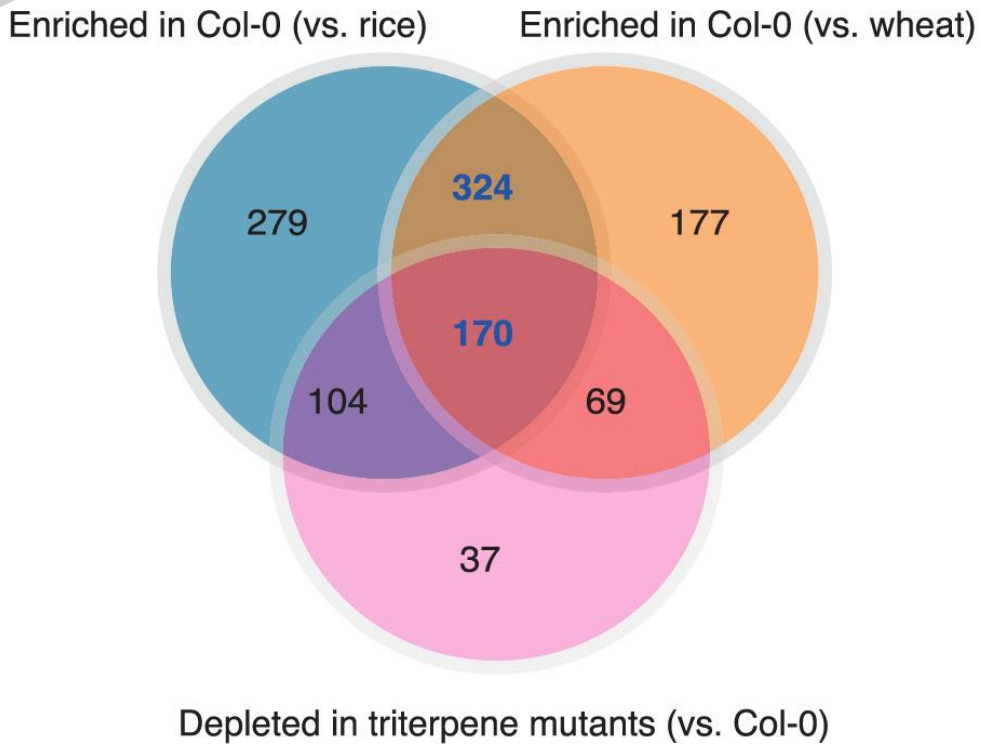
曼哈顿图展示差异

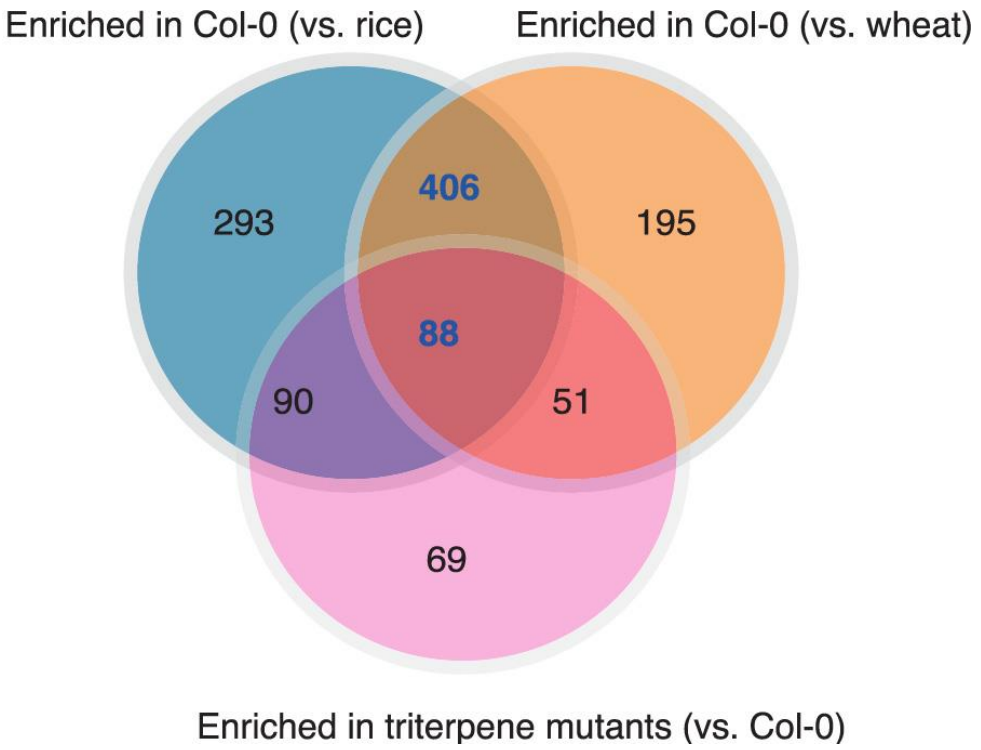


点代表物种，上三角为上调，下三角为下调，大小为丰度，颜色为物种分类

NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用 ***Nature Biotechnology***. 2019. Fig 3a/b

三组上下调OTUs分别比较

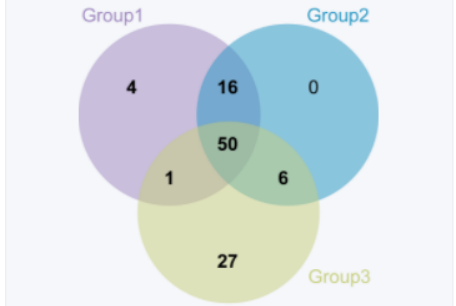




三萜通路的突变体特异调控的根系细菌类群。 维恩图展示了的拟南芥三萜突变体中下调(左)或富集(右)的OTUs，与水稻和小麦与拟南芥野生型相比变化的OTUs大量重叠。

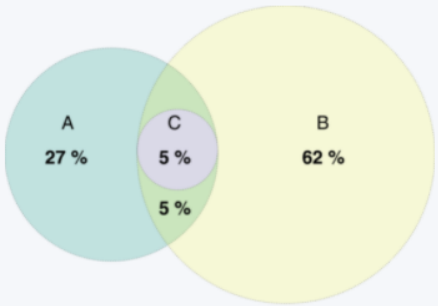
[Science：拟南芥三萜化合物特异调控根系微生物组* 专家点评]

E Venn维恩图网站

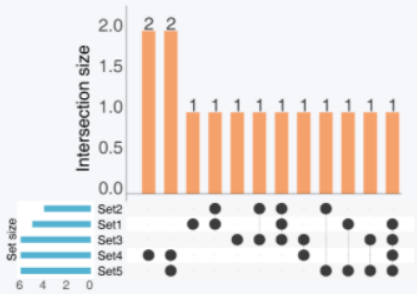


Interactive venn diagram

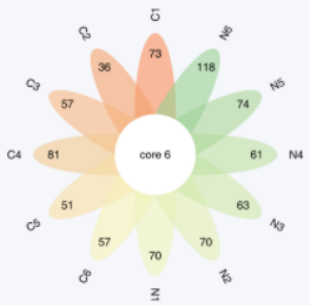
Interactve edwards layout



Euler diagram

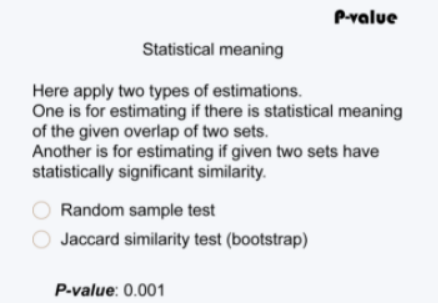


UpSet plot

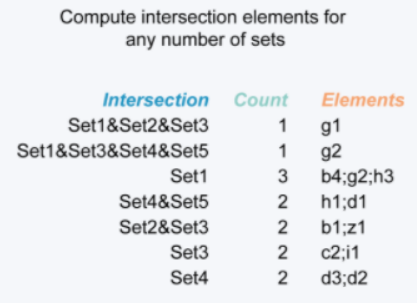


Flower plot

Venn network



Estimate set similarity

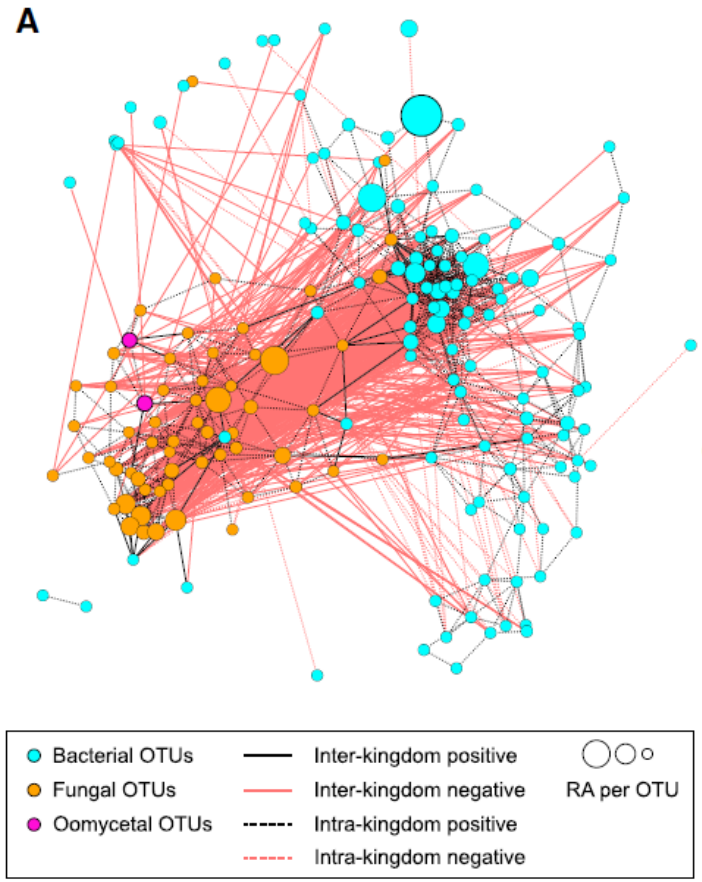


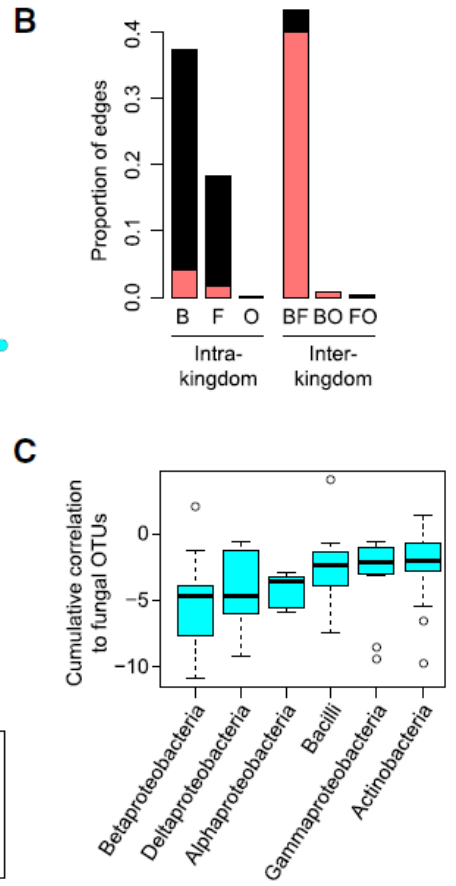
Venn calculator

http://www.ehbio.com/test/venn/#/

网络-跨界相关研究

细菌、真菌、卵菌相关网络分析
A. OTU相关性网络。每个节点代表一个OTU,每个节点之间的边代表正相关（黑色）或者负相关（红色）（$p < 0.05$, correlation values < -0.6 or > 0.6）。属于不同微生物界的OTU具有不同的颜色，并且节点大小反映它们在根内表中的相对丰度。用虚线表示界内部相关性，用实线表示界间相关性。
B. 根内表网络中呈现正相关（黑色）和负相关（红色）相关性的边的比例。B，细菌；F真菌；O：卵菌。
C. 在细菌和真菌OTU之间的微生物网络中测量的累积相关性得分。
细菌（左）和真菌（右）OTU按纲水平分组（每个纲多于5个OTU），并根据它们与真菌和细菌OTU的累积相关性分数进行排序。

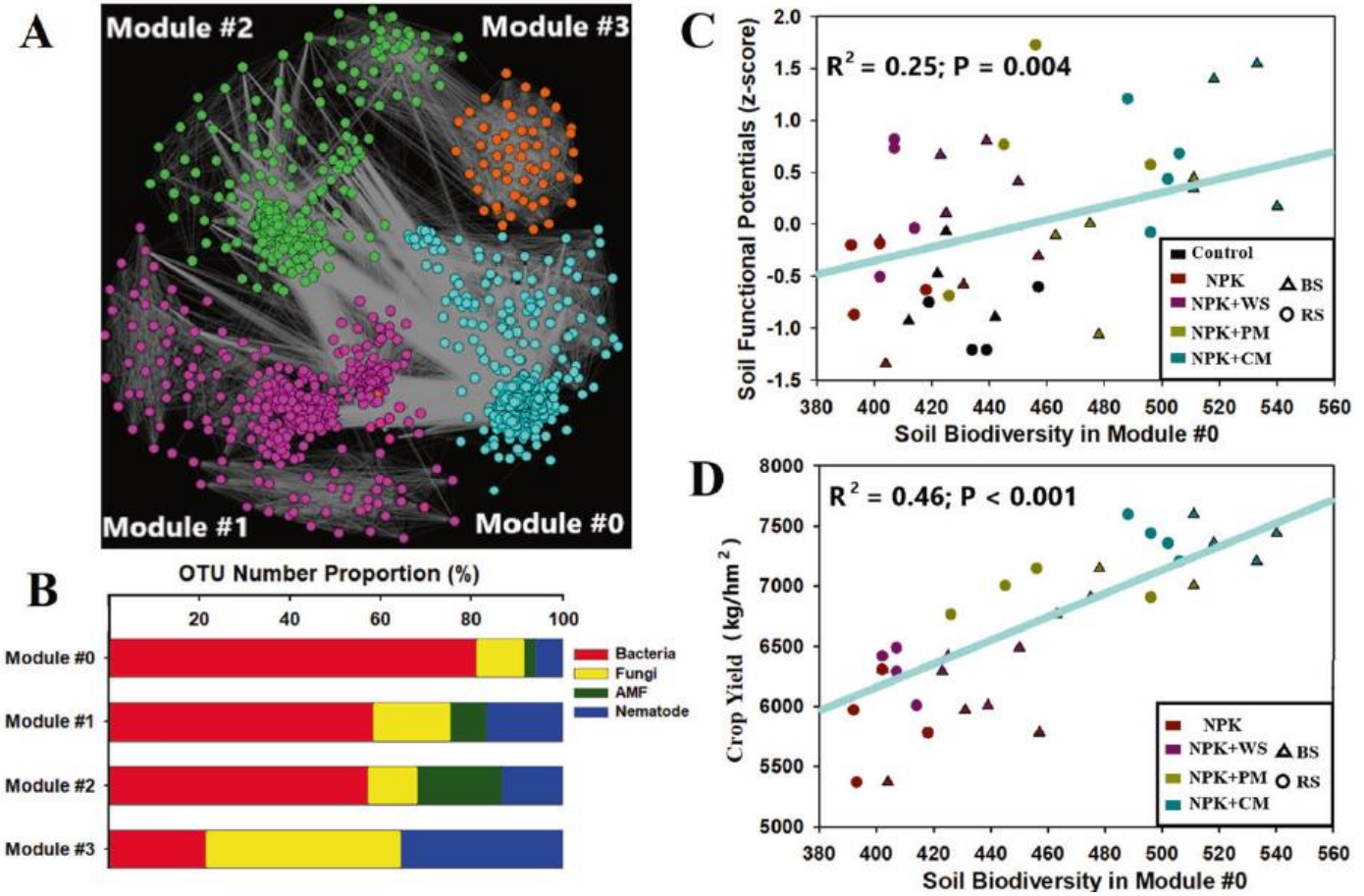




•Cell： 根部微生物跨界的互作促进拟南芥生存

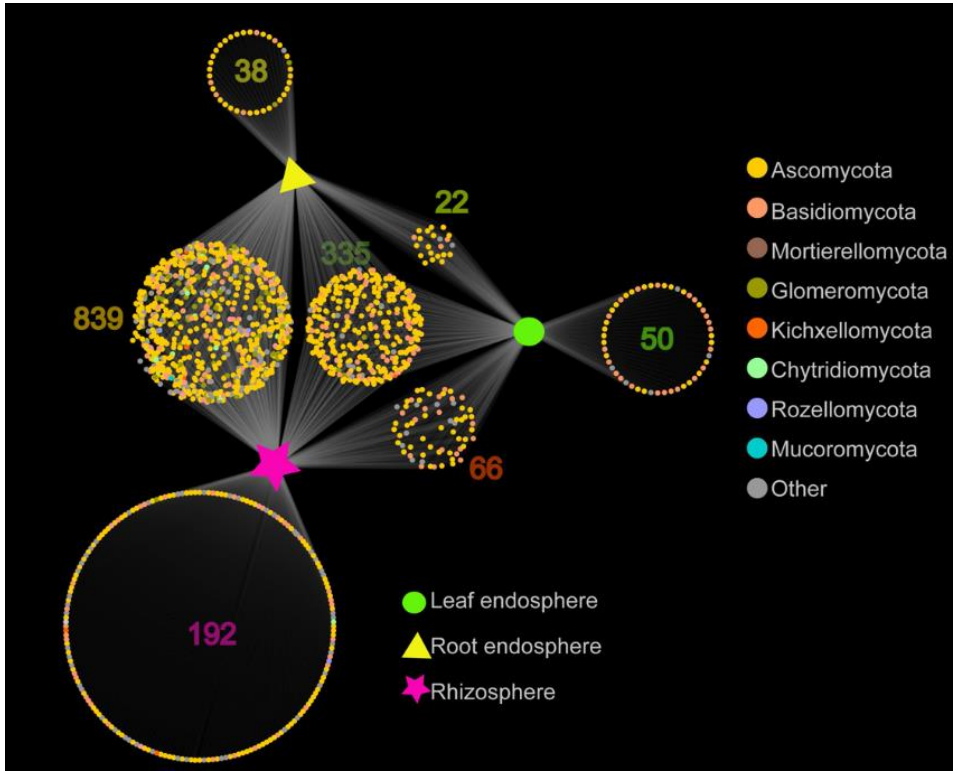
网络-模块化研究

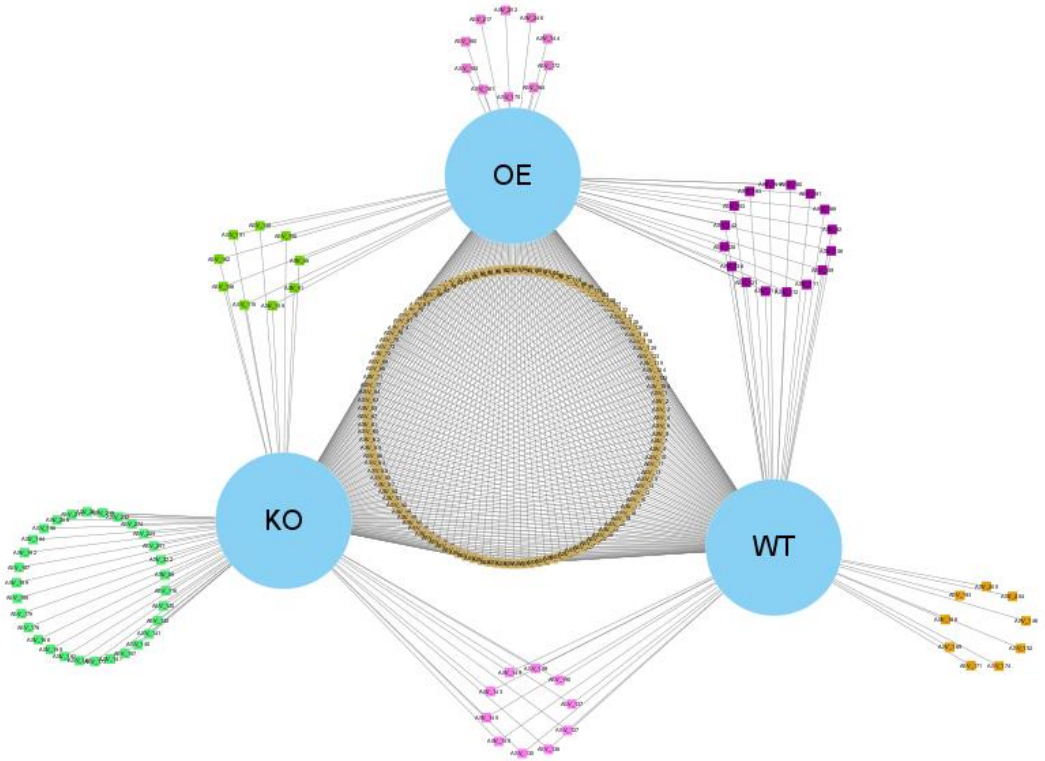
基于多营养网络的生态集群。
(A) 四个主要微生物生态集群的网络图(Module #0-3); (B)不同微生物生态集群中优势菌群的OTU比例; (C) &(D)线性回归分析解析关键微生物菌群生物多样性与土壤功能潜力、小麦作物产量之间的关系。



•ISME: 中科院南京土壤所褚海燕组揭示关键菌群的生物多样性决定作物产量

网络图-不同组共有或特有OTU





使用网络图展示Venn图集合及Cytoscape操作演示

http://www.ehbio.com/test/venn/#/

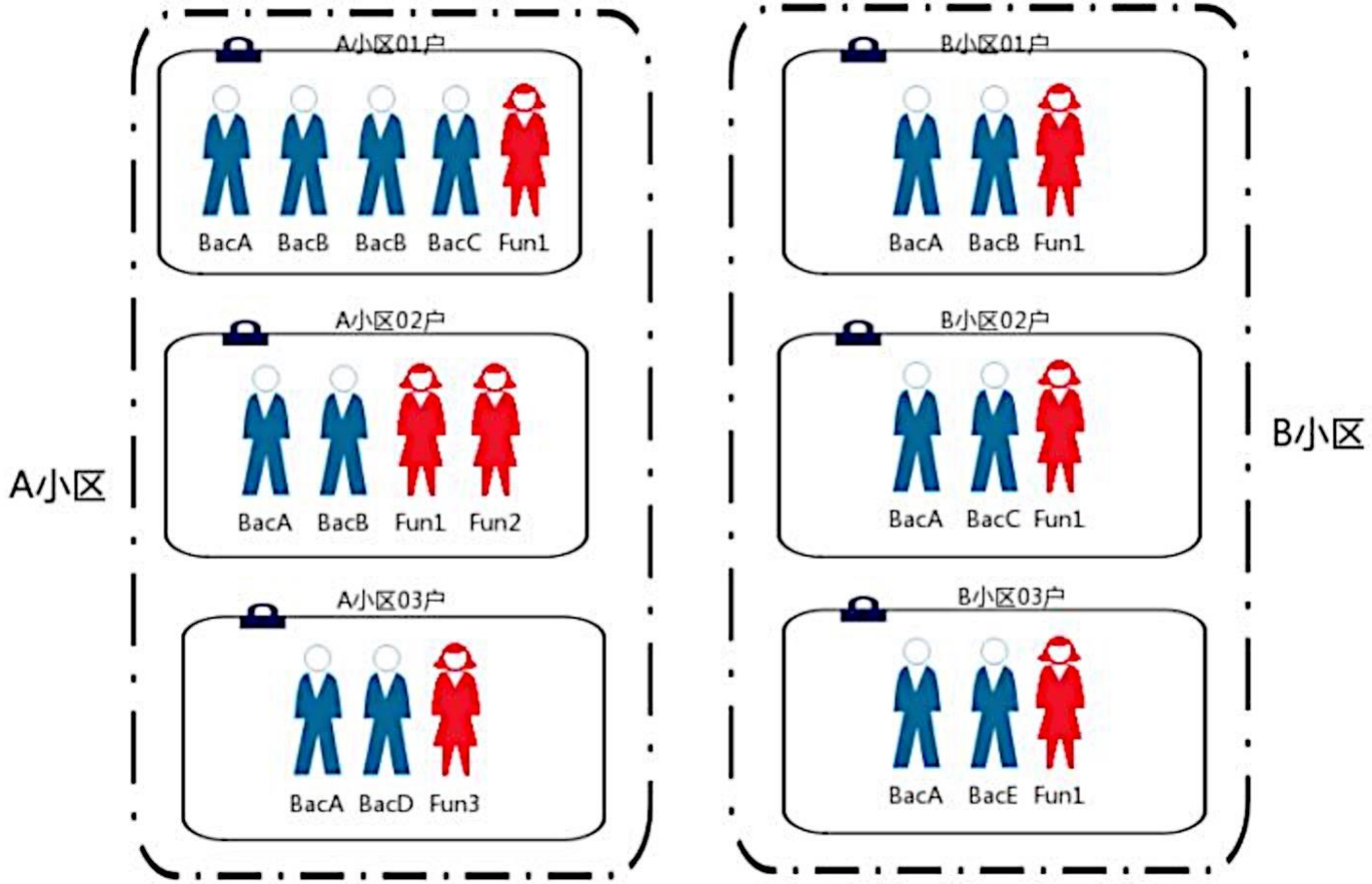
CEPAMS

报告提纲

- 读懂文章图表
- **扩增子分析流程**
- 宏基因组分析流程
- 多样性分析和可视化
- 高分文章套路

CEPAMS

扩增子分析类似人口普查



- **小区：实验组**
- **家户：样品**
- **男生：样品中的细菌**
  - **BacA：北京人**
  - **BacB：山东人**
  - **BacB：山东人**
  - **BacC：东北人**
  - **“省份”这一规则进行分类**
- **女生：样品中的真菌**

扩增子(标记基因，Marker genes)

DNA提取

扩增

测序

质控定量

聚类/去噪

多样性

分析

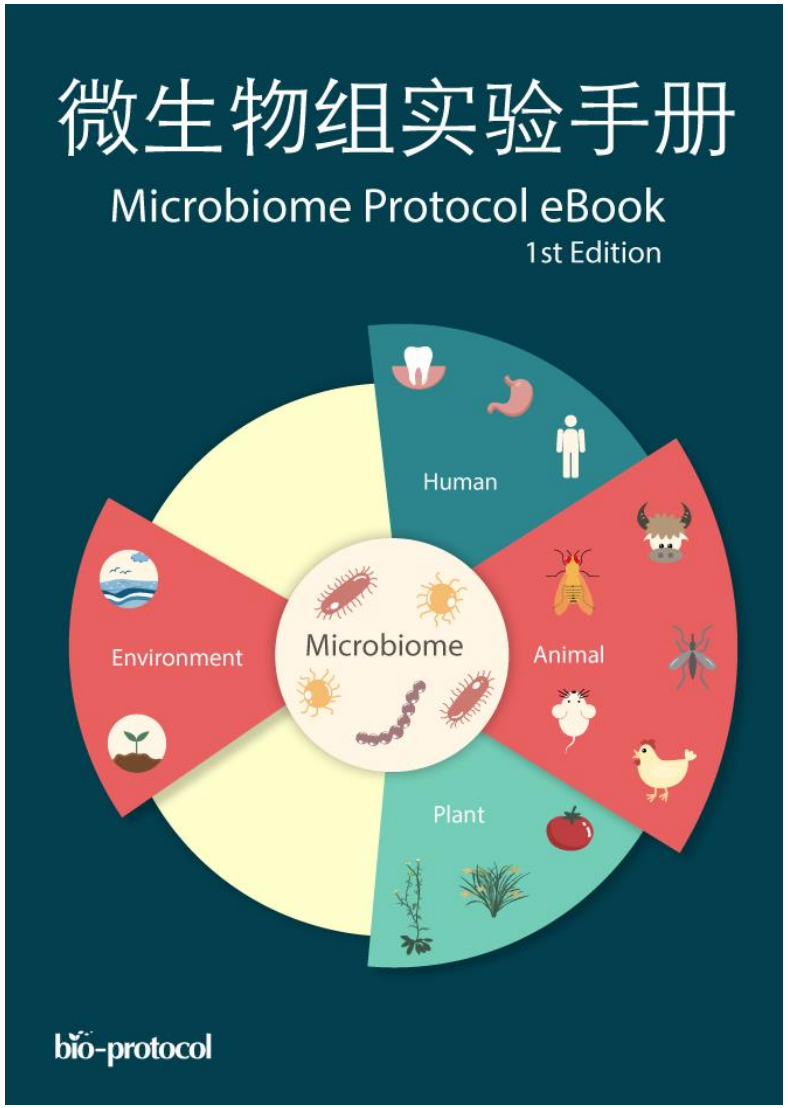
Yong-Xin Liu, Yuan Qin, **Tong Chen**, Meiping Lu, Xubo Qian, Xiaoxuan Guo & Yang Bai. A practical guide to amplicon and metagenomic analysis of microbiome data. ***Protein Cell*** 41, 1-16, doi: https://doi.org/10.1007/s13238-020-00724-8 (2020).

Protein Cell：扩增子和宏基因组数据分析实用指南

《微生物组实验手册》

101家单位，357人参与，发表147篇实验方法，80余万字

涵盖样品制备、培养组、扩增子、宏基因组、宏转录组、代谢组、单菌基因组、相关分子生物学和微生物学实验等



主页：https://bio-protocol.org/bio101/mpb

科学顾问

刘双江

中国科学院微生物研究所

朱宝利

中国科学院微生物研究所

朱永官

中国科学院城市环境研究所

特邀主编

刘永鑫

中国科学院遗传与发育生物学研究所

褚海燕

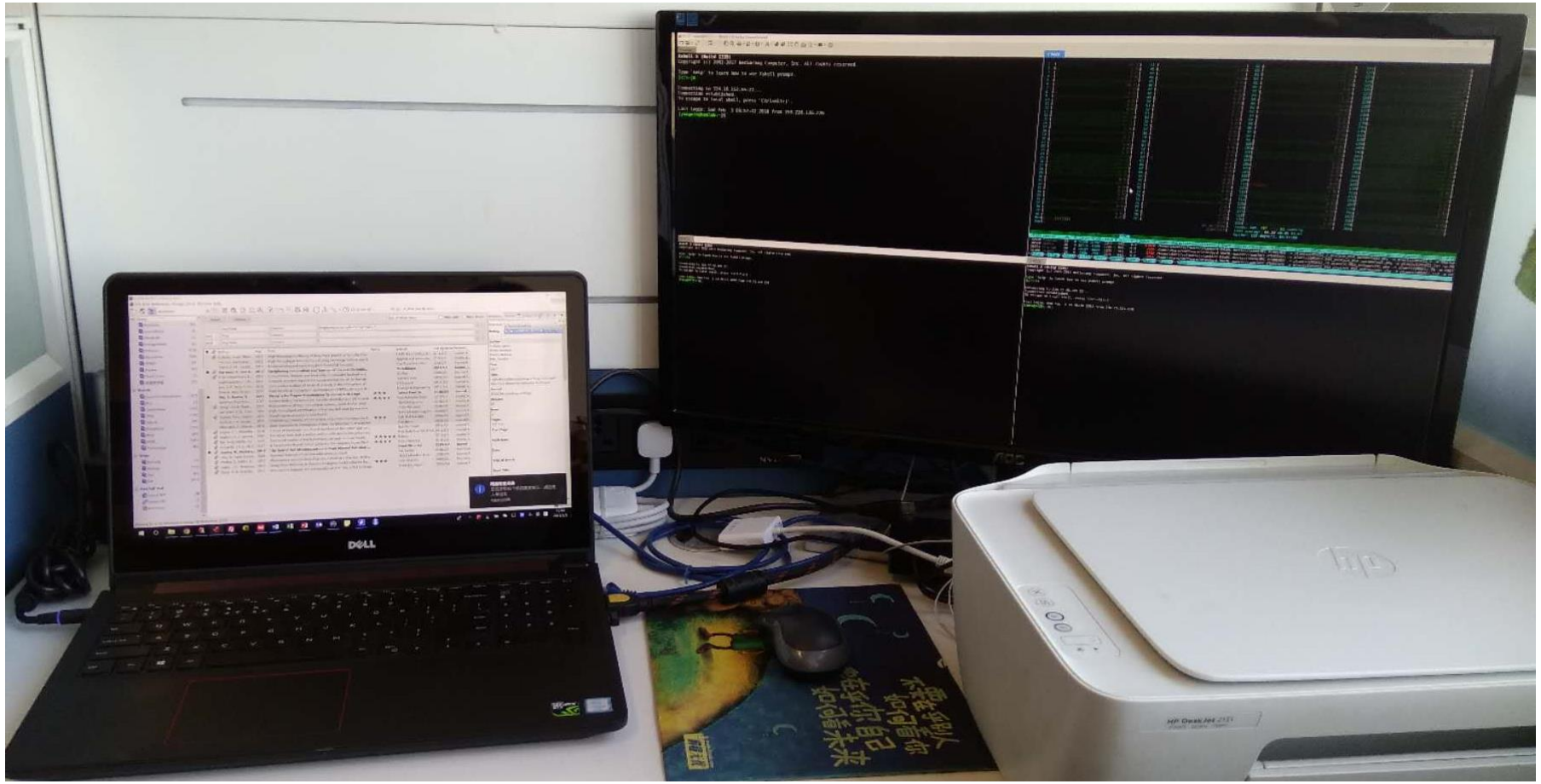
中国科学院南京土壤研究所

白洋

中国科学院遗传与发育生物学研究所

长期项目，投稿可以联系我，微信 meta-genomics

硬件——笔记本+显示器



双显示器方便多任务管理、阅读文献和多图比较

常用分析软件

- 数据分析环境Shell + R + IDE：GitForWindows、R +R包、RStudio

- 扩增子分析流程：USEARCH & VSEARCH，Win子系统+QIIME 2
- 辅助工具：序列工具seqkit、表格工具csvtk、并行管理rush
- 差异分析和可视化：STAMP
- 网络分析及可视化：Cytoscape、Gephi
- 图片排版：Adobe Illustrator
- 登录服务器：XShell 或PuTTY；上传下载文件：Filezilla 或 WinSCP

我常用的脚本和数据库

- 下载github仓库： https://github.com/yongxinliu/db
  - 方法1. 网页中点“Code” —— Download ZIP，下载后解压
  - 方法2. 命令行中 git clone git@github.com:YongxinLiu/db.git
- 如何使用脚本？
  - 将下载的db目录复制到windows的c盘(/c)，或Linux/Mac家目录(~)
  - 添加可执行程序至环境变量(以Windows中RStudio中Terminal为例)
  - export PATH=$PATH:/c/db/win/
  - 使用前设置目录变量，方便以后多次使用
  - sd=/c/db/script
  - R语言绝对路径使用R脚本
  - Rscript ${sd}/alpha_boxplot.R -h

CEPAMS

扩增子分析软件和数据库

Quantitative Insights Into Microbial Ecology

qiime

USEARCH

Ultra-fast sequence analysis

mothur

16S、18S和ITS分析流程

扩增子分析神器USEARCH 简介

GREENGENES

The 16S rRNA Gene Database and Tools

rdp

unite

community

silva

high quality ribosomal RNA databases

16S、18S、ITS数据库

微生物扩增子数据库大全

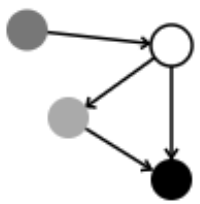
NAR：UNITE真菌鉴定ITS数据库

CEPAMS

QIIME1/2——最流行

QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible,

Learn more



Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Interactively explore your d beautiful visualizations that pro perspectives.

QIIME

About 40,800 results (**0.05** sec)

QIIME allows analysis of high-throughput community sequencing data

JG Caporaso, J Kuczynski, J Stombaugh, K Bittinger... - Nature ..., 2010 - nature.com

To the Editor: High-throughput sequencing is revolutionizing microbial ecology studies. Efforts like the Human Microbiome Projects 1 and the US National Ecological Observatory Network 2 are helping us to understand the role of microbial diversity in habitats within our ...

☆ 99 Cited by 24628 Related articles All 17 versions

Reproducible, interactive, scalable and extensible microbiome data science using **QIIME** 2

E Bolyen, JR Rideout, MR Dillon, NA Bokulich... - Nature ..., 2019 - nature.com

Conclusions Because of a clear trend toward more engagement and transparency with research participants, we should expect more research participants to exercise their HIPAA access right in coming years. The committee's recommendations ensure that researchers ...

☆ 99 Cited by 2296 Related articles All 30 versions

- NBT：QIIME 2可重复、交互式的微生物组分析平台
- 1简介和安装Introduction&Install

CEPAMS

USEARCH——最好用

http://www.drive5.com/usearch/

- 由于USEARCH即好用，但收费，出现了模仿者VSEARCH，方便大家免费使用。
- 有多平台版本，轻松分析扩增子
- 想免费分析大数据的有福了

[HTML] **VSEARCH**: a versatile open source tool for metagenomics
T Rognes, T Flouri, B Nichols, C Quince, F Mahé - PeerJ, 2016 - peerj.com
Background **VSEARCH** is an open source and free of charge multithreaded 64-bit tool for processing and preparing metagenomics, genomics and population genomics nucleotide sequence data. It is designed as an alternative to the widely used USEARCH tool for which
☆ 99 Cited by 3115 Related articles All 22 versions »

CEPAMS

易扩增子(EasyAmplicon)

- QIIME、USEARCH和Mothur，但仍分别存在依赖关系过多导致的安装困难、大数据收费和使用界面不友好等问题
- 易扩增子实现简单易用、可重复和跨平台地开展扩增子分析
- 核心采用体积小、安装方便、计算速度快且跨平台的软件USEARCH，同时整合VSEARCH以突破USEARCH免费版限制
- 选用RStudio的图形界面对流程代码文档管理和运行，实现命令行和/或鼠标点击操作方式开展扩增子可重复分析
- 提供数10个脚本，实现特征表过滤、重采样、分组均值等常用计算，并为STAMP、LEfSe、PICRUSt1/2等提供标准的输入文件

CEPAMS

项目：https://github.com/YongxinLiu/EasyAmplicon
教程：MPB：易扩增子：易用、可重复和跨平台的扩增子分析流程

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
+
DDDDDIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIII
@HISEQ:549:HLYNYBCXY:1:1101:2025:2183 1:N:0:CACTCAAT
ATATCGCGAGAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGCG
+
DDDD@E@HIGHIIIHHFHIIHIIIIFHHIIIHHGIHIIHIIICHDEHHIIIIHGH
@HISEQ:549:HLYNYBCXY:1:1101:2052:2198 1:N:0:CACTCAAT
CACGAGACAGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGATGGGTA
+
D@DD@H=?CCHHIIIIIIIIIIIIIIIIIIIIIIIIIIGI0CHIIIIIHIHIIH@
```

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

统计表: 1~N X $10^{1\sim3}$

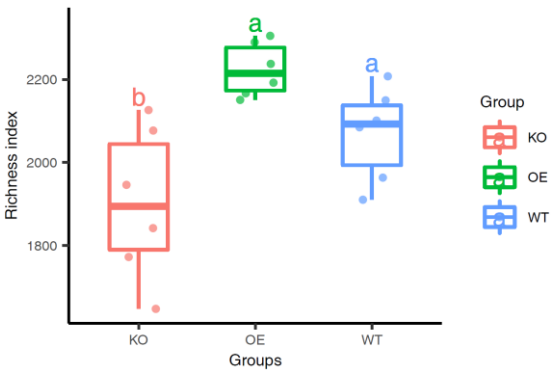


图: $10^{1\sim3}$个点和统计信息

Protein Cell：扩增子和宏基因组数据分析实用指南

刘永鑫：想学菌群生物信息分析-21分钟带你入门

CEPAMS

1. 认识文件格式

- 测序原始数据：seq/*.fq.gz

```
@HISEQ:549:HLYNYBCXY:1:1101:2135:2154 1:N:0:CAGGCGAT
ACGCTCGACAAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGTGTGCTGGGCGTCGGGGGGCTTGCCCCT
+
@DDDDHIIIIIIIHHIIIIGHIHICGHIIIIIIH<FHF?CHHIHHCGHHHHIFHCHE@G@EF?HHHHCHID/EEHCEHHI
```

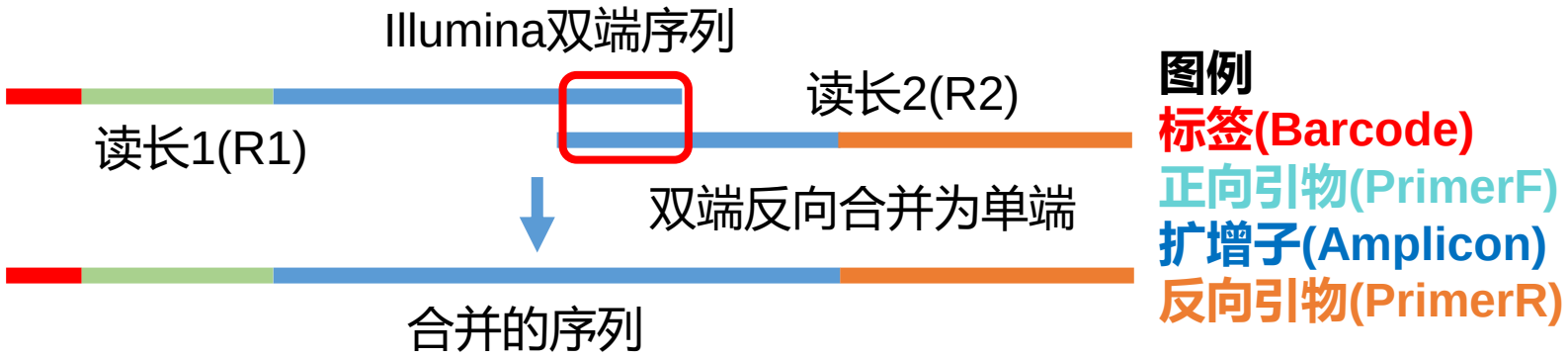
- 实验设计/样品信息：metadata.txt (制表符分隔文本文本，可用Excel或纯文本编辑器编写，如Editplus)

SampleID	Group	Date	Site	CRA	CRR	BarcodeSequence	LinkerPrimerSequence	ReversePrimer
KO1	KO	2017/6/30	Chaoyang	CRA002352	CRR117575	ACGCTCGACA	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
KO2	KO	2017/6/30	Chaoyang	CRA002352	CRR117576	ATCAGACACG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
OE1	OE	2017/6/30	Chaoyang	CRA002352	CRR117581	TCTCTATGCG	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC
OE2	OE	2017/6/30	Chaoyang	CRA002352	CRR117582	TACTGAGCTA	AACMGGATTAGATACCCKG	ACGTCATCCCCACCTTCC

- 注意事项：有行列标题，**行为样品名(字母开头+数字组合)**，列为**分组信息(至少1列，可多列)**、**地点和时间(提交数据必须)**、及其它属性。

序列格式详解fasta&fastq，样品命名 注意事项 实例

2. 双端序列合并



R1

TCGTCGCTCGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGTTGGGCGCTAGGTGTGGGGGACATTCACGTTCTCCG
TGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGTTGAAACTCAAAGGAATTGACGGGGACCCGCGCA
AGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGTCTTGACATCCATGGAACCCTGCAGAGATGC

R2

ACGTCATCCCCACCTTCCTCCGGTTTGTCACCGGCGGTCTCCTTAGAGTTCCCAACTAAATGATGGCAACTAAGGACAAGGGTT
GCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCAGCACCTGTCTCATGGTTCCTTACGGC
ACCCCCGCATCTCTGCAGGGTTCCATGGATGTCAAGACCAGGTAAGGATCTTCGCGTGGCATCGAAGTAAAACACAGGCACC

R2_RC
反向互补

GGTGCCTGTGTTTTACTTCGATGCCACGCGAAGATCCTTACCTGGTCTTGACATCCATGGAACCCTGCAGAGATGCGGGGGTGC
CGTAAGGAACCATGAGACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAA
CCCTTGTCCTTAGTTGCCATCATTTAGTTGGGAACTCTAAGGAGACCGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGT

双端序列合并的实现

- 一条命令实现双端序列合并

```
vsearch -fastq_mergepairs seq/WT1_1.fq.gz -reverse seq/WT1_2.fq.gz \
        -fastqout temp/WT1.merged.fq -relabel WT1.
```

- 解释：扩增子分析 -序列合并 序列1 -反向序列 序列2 -输出 合并结果
- 小技巧，使用变量替换文件名可变部分，方便修改

```
i=WT1 # 如果你的文件名为human_skin_180910_beijing_1.fq
vsearch -fastq_mergepairs seq/${i}_1.fq -reverse seq/${i}_2.fq \
        -fastqout temp/${i}_merge.fq -relabel ${i}.
```

CEPAMS

理解命令和命令行参数

盖个房子?

瓦匠 把砖 盖成房子

1. 谁能干：找人

瓦匠

2. 对谁干：材料

村东头砖厂

3. 结果：盖好的房子

你家马路对面的新房子

把双端测序文件按末端互相合并?

vsearch -fastq_mergepairs seq/WT1_1.fq -reverse seq/WT1_2.fq -fastqout temp/WT1_merge.fq

1. 谁能干：具体程序

vsearch -fastq_mergepairs

2. 对谁干：输入文件

seq/WT1_1.fq -reverse seq/WT1_2.fq

3. 结果：输出文件

-fastqout temp/WT1_merge.fq

CEPAMS

易扩增子(EasyAmplicon)

- - pipeline.sh # 流程脚本
- - pipeline_mac.sh # Mac版
- - result/ # 示例结果
- - result/Diversity-tutorial.Rmd
- 多样性分析交互脚本
- 使用有道云笔记Markdown阅读
- 每季度更新

项目代码：https://github.com/YongxinLiu/EasyAmplicon
中文教程：MPB：易扩增子：易用的扩增子分析流程

CEPAMS

QIIME 2进一步学习

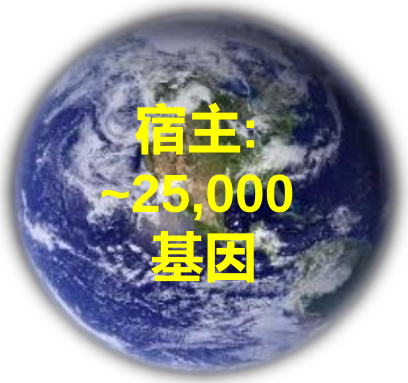
- 简明教程(5千字，1天入门)
  - MPB：使用QIIME 2分析微生物组16S rRNA基因扩增子测序数据(视频)
- 官方教程中文版(10万字，32节，半个月系统学习)
  - NBT：QIIME 2可重复、交互式的微生物组分析平台
  - 1简介和安装Introduction&Install
  - 2插件工作流程概述Workflow
  - 3老司机上路指南Experienced
- 英文原版(最新版见官网)
  - https://docs.qiime2.org/

CEPAMS

报告提纲

- 读懂文章图表
- 扩增子分析流程
- **宏基因组分析流程**
- 多样性分析和可视化
- 高分文章套路

CEPAMS



Nature：人类肠道微生物组参考基因集
NBT：人类微生物组千万基因的参考基因集
NBT：20万个基因组的肠道微生物参考基因组集
Cell：人体肠道细菌与自身细胞的比例1：1

CEPAMS

宏基因组

DNA 提取

随机打断测序

质控，(组装注释) 比对

物种组成功能分析

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*** 41, 1-16, doi: https://doi.org/10.1007/s13238-020-00724-8 (2020).

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

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)

③组装/拼接分析 (Assemble-based)

组装/拼接 → 重叠群 (Contigs)

基因预测 → 基因去冗余 → 基因集 (Gene catalog)

分箱(Binning) → 基因组草图 (Draft genome)

基因丰度 ← 定量 ← 基因集 (Gene catalog)

基因注释

- 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

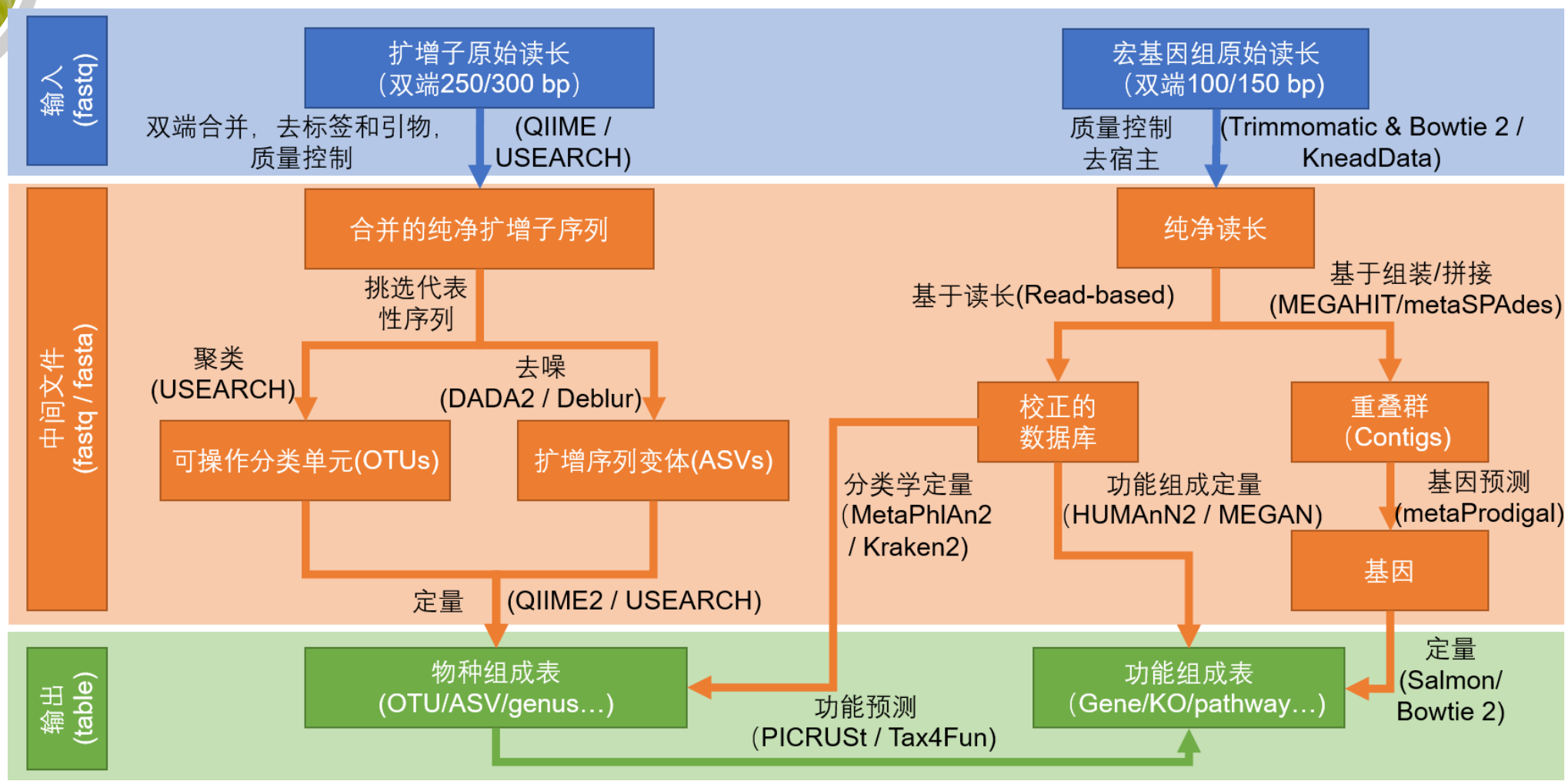
常用物种和功能基因注释数据库(图标右)和对应的软件(图标下)

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**).

CMJ：人类微生物组研究设计、样本采集和生物信息分析指南

CEPAMS

扩增子和宏基因组分析流程



Protein Cell：扩增子和宏基因组数据分析实用指南

硬件——服务器/集群

服务器

集群

CONDA BIOCONDA

- Conda是(Python, R, Java, C等)软件包和环境管理系统，用于安装多个版本的软件包及其依赖关系，并在它们之间轻松切换
  - wget -c https://repo.continuum.io/miniconda/Miniconda2-latest-Linux-x86_64.sh
  - bash Miniconda2-latest-Linux-x86_64.sh
- **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
- 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300+次
- 添加频道：conda config --add channels bioconda

CEPAMS

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).

国家微生物科学数据中心

•微生物组常用数据库国内备份站

微生物组数据库国内下载

国际数据下载

- 核酸及蛋白质序列数据库 (11)
- 基因组数据库 (6)
- 蛋白质结构及功能数据库 (15)
- 文献数据库 (1)
- 物种及元数据库 (4)
- 宏基因组数据库 (1)
- Blast数据库 (9)

工具资源下载

- 微生物组软件包 (2)
- 扩增子数据库 (4)
- 宏基因组数据库 (9)

http://nmdc.cn/datadownload

宏基因组数据库

HUMAnN物种和功能注释-HUMAnN2+MetaPhlAn2数据库(主流)

主页： http://www.huttenhower.org/humann2

描述： 宏基因组数据有参快速物种和功能通路定量软件。 下载，解压，具体路径通过humann2_config设置database_folders中utility_mapping、protein和nucleotide的值。

序号	版本	大小	更新时间	下载链接	描述
1	full_mapping_1_1	593M	2020-09-23	tar.gz	功能描述(utility_mapping)
2	full_chocophlan_plus_...	5.4G	2020-09-23	tar.gz	微生物泛基因组(nucleotide)，建立功能与物种组成的联系
3	uniref90_annotated_1...	5.9G	2020-09-23	tar.gz	UniRef蛋白(protein)序列diamond索引

宏基因组数据库

HUMAnN物种和功能注释-HUMAnN3+MetaPhlAn3数据库(测试版)

主页： http://www.huttenhower.org/humann3

描述： 宏基因组数据有参快速物种和功能通路定量软件。比HUMAnN2的数据库增大3倍。 下载，解压，具体路径通过humann_config设置database_folders中utility_mapping、protein和nucleotide的值。

微生物组常用数据库国内备份站

CEPAMS

并行管理软件parallel / rush

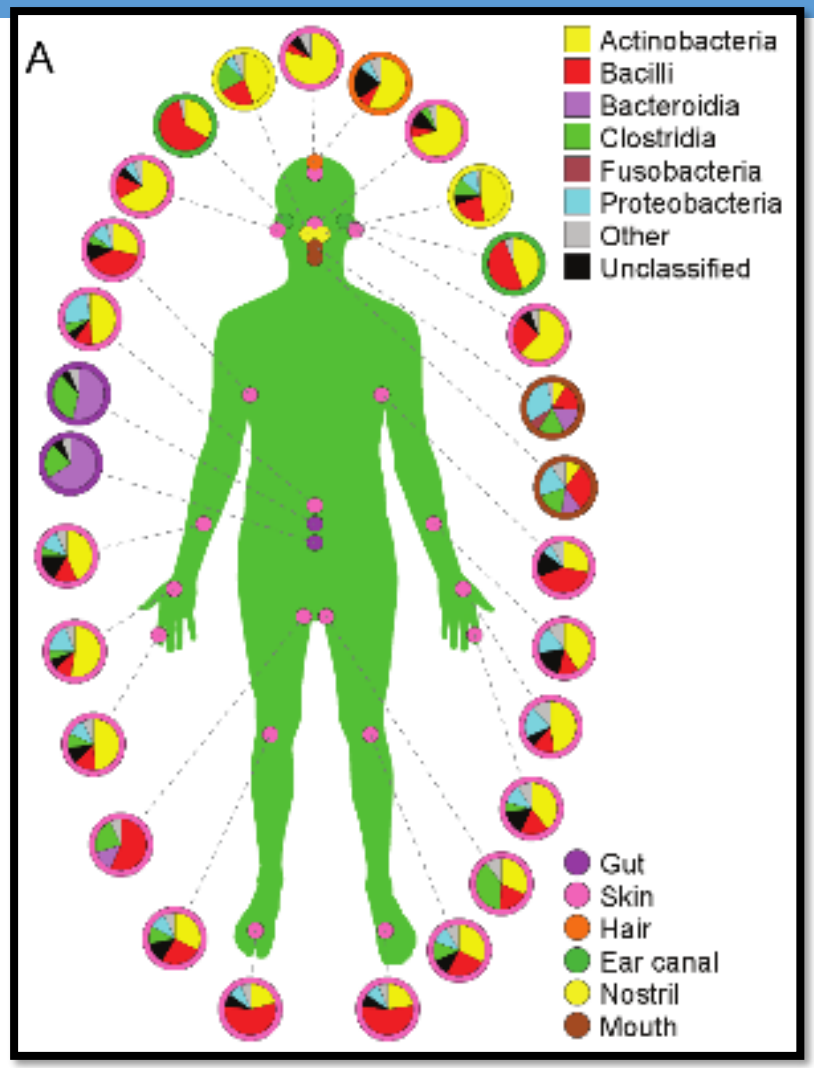
- 现实中是有一大堆样品，for可以单个或全部提交任务效率都很低，如何让服务器性能允许下并行加速分析，并有序管理队伍呢?
- Parallel是Perl语言编写，可提供并行任务数量管理的功能，保证任务高效有这序完成，作者要求引用，如不想引用也可付10000欧元购买。可以直接在Ubuntu仓库中安装或conda安装
  - sudo apt install parallel
  - conda install parallel
- 国人开发了跨平台的并行管理工具rush，官网下载 或 conda安装
  - conda install rush
  - 官网：https://github.com/shenwei356/rush

Tange, O. *Gnu parallel 2018*. (Lulu. com, 2018).

CEPAMS

宏基因组分析流程

一．简介—定义、方法和数据库
二．数据质量控制与并行计算
三．HUMAnN2定量物种和功能
四．差异统计和可视化方法
五．Kraken2物种注释
六．组装、基因注释和定量
七．常用功能注释数据库
八．分箱单菌基因组



另一套参考教材：挖掘微生物组生物标记

并行质量控制(质控)实例

- 示例：对所有样品进行质控，同时保持最多3个样本在运行。
- -j为任务数，--xapply是对两个参数按顺序使用而非组合方式
  - **time parallel -j 3 --xapply **
  - "kneaddata -i seq/{1}_1.fq.gz -i seq/{1}_2.fq.gz \
  - -o temp/qc -v -t 3 --remove-intermediate-output \
  - --trimmomatic /conda2/envs/metagenome_env/share/trimmomatic/ --trimmomatic-options 'ILLUMINACLIP:/conda2/envs/metagenome_env/share/trimmomatic/adapters/TruSeq2-PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \
  - --bowtie2-options '--very-sensitive --dovetail' -db ${db}/kneaddata/human_genome/Homo_sapiens" \
  - **::: `tail -n+2 result/metadata.txt|cut -f1`**

注意修改软件、数据库位置。

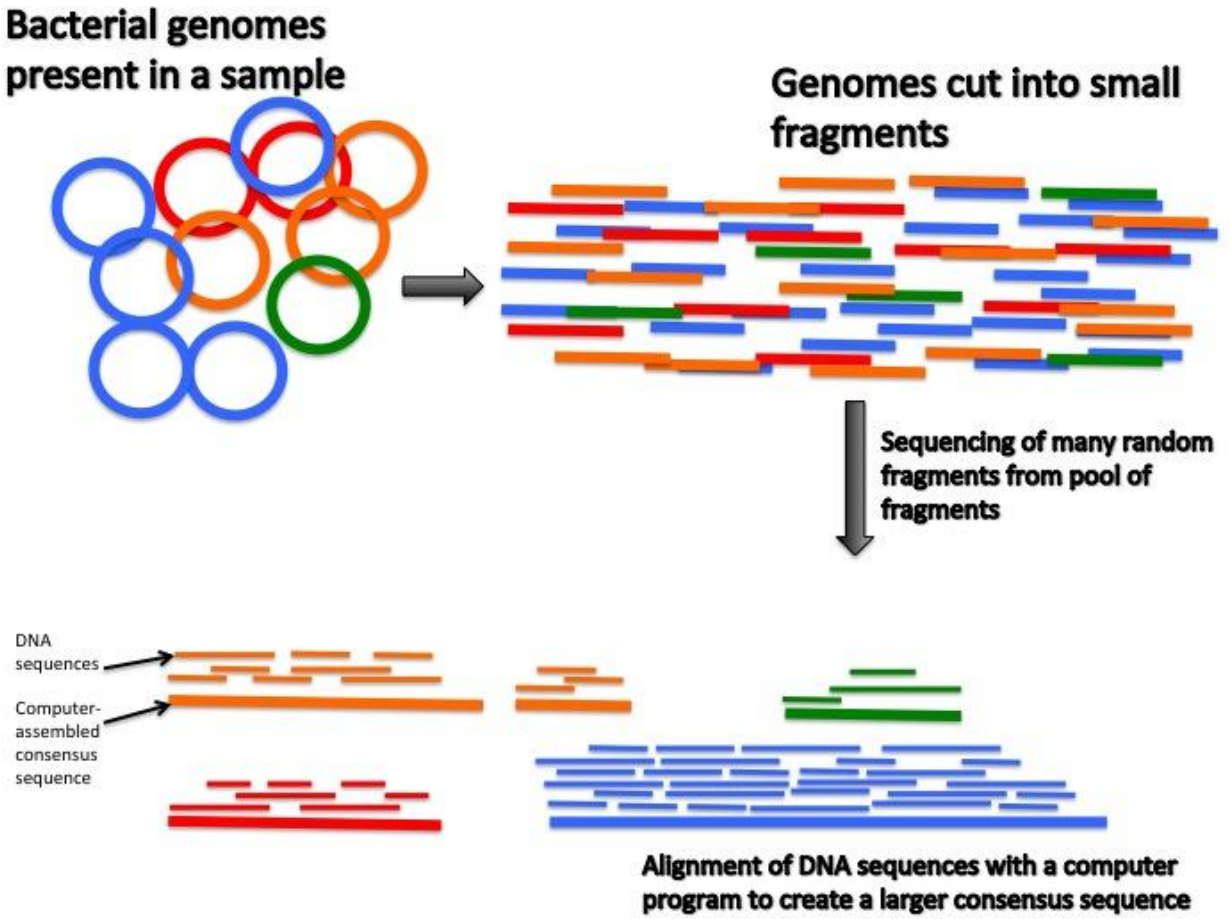
CEPAMS

物种丰度：分类(taxonomic) vs 序列(sequence)

- 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

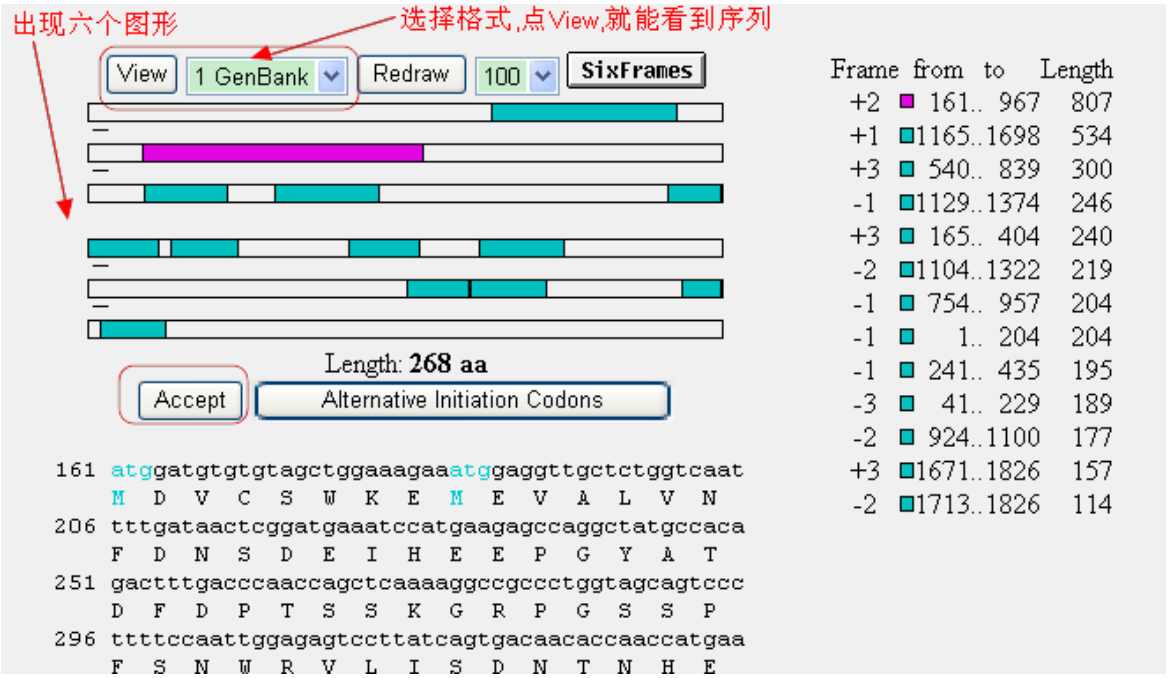
Zheng Sun, Shi Huang, Meng Zhang, Qiyun Zhu, Niina Haiminen, Anna Paola Carrieri, Yoshiki Vázquez-Baeza, Laxmi Parida, Ho-Cheol Kim, Rob Knight & Yang-Yu Liu. (**2021**). Challenges in benchmarking metagenomic profilers. ***Nature Methods***, doi: https://doi.org/10.1038/s41592-021-01141-3

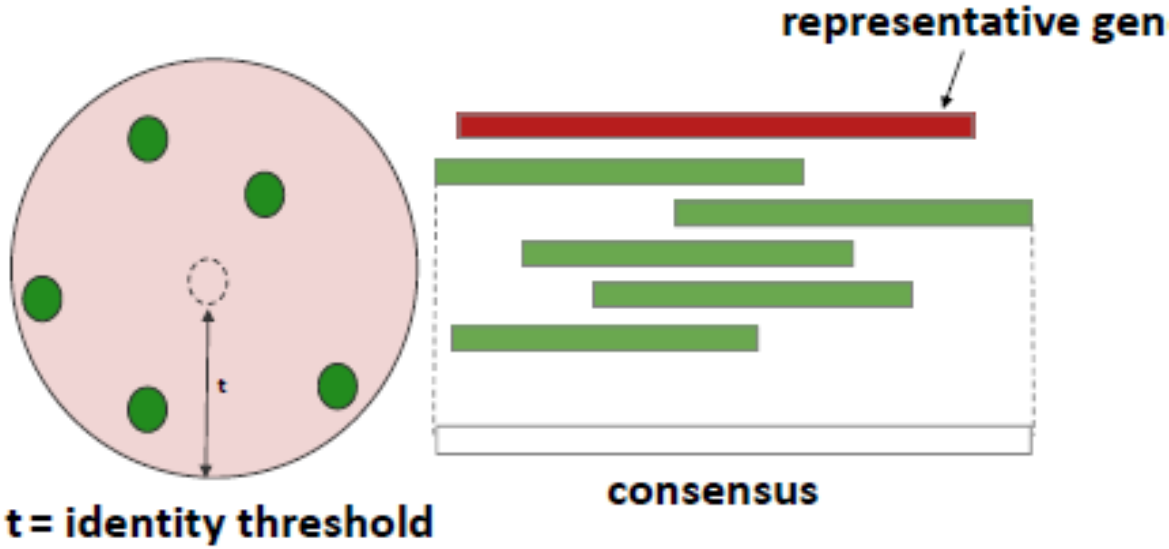
组装/拼接 (Assemble)的基本原理



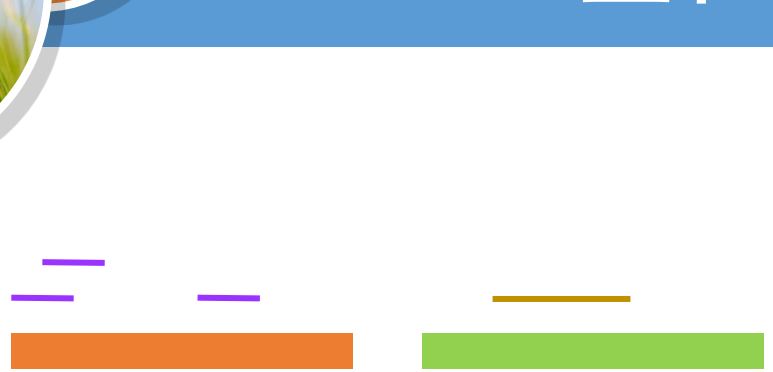
https://teachthemicrobiome.weebly.com/sequencing-the-microbiome.html

基因预测和聚类


出现六个图形
选择格式,点View,就能看到序列
View
1 GenBank
Redraw
100
SixFrames
Length: 268 aa
Accept
Alternative Initiation Codons
Frame from to Length
+2 161.. 967 807
+1 1165..1698 534
+3 540.. 839 300
-1 1129..1374 246
+3 165.. 404 240
-2 1104..1322 219
-1 754.. 957 204
-1 1.. 204 204
-1 241.. 435 195
-3 41.. 229 189
-2 924..1100 177
+3 1671..1826 157
-2 1713..1826 114
161 atggatgtgtgtagctggaaagaaatggaggttgctctggtcaat
M D V C S W K E M E V A L V N
206 tttgataactcggatgaaatccatgaagagccaggctatgccaca
F D N S D E I H E E P G Y A T
251 gactttgacccaaccagctcaaaaggccgccctggtagcagtccc
D F D P T S S K G R P G S S P
296 ttttccaattggagagtccttatcagtgacaacaccaaccatgaa
F S N W R V L I S D N T N H E


representative gen
t
consensus
t = identity threshold

基因定量和功能注释

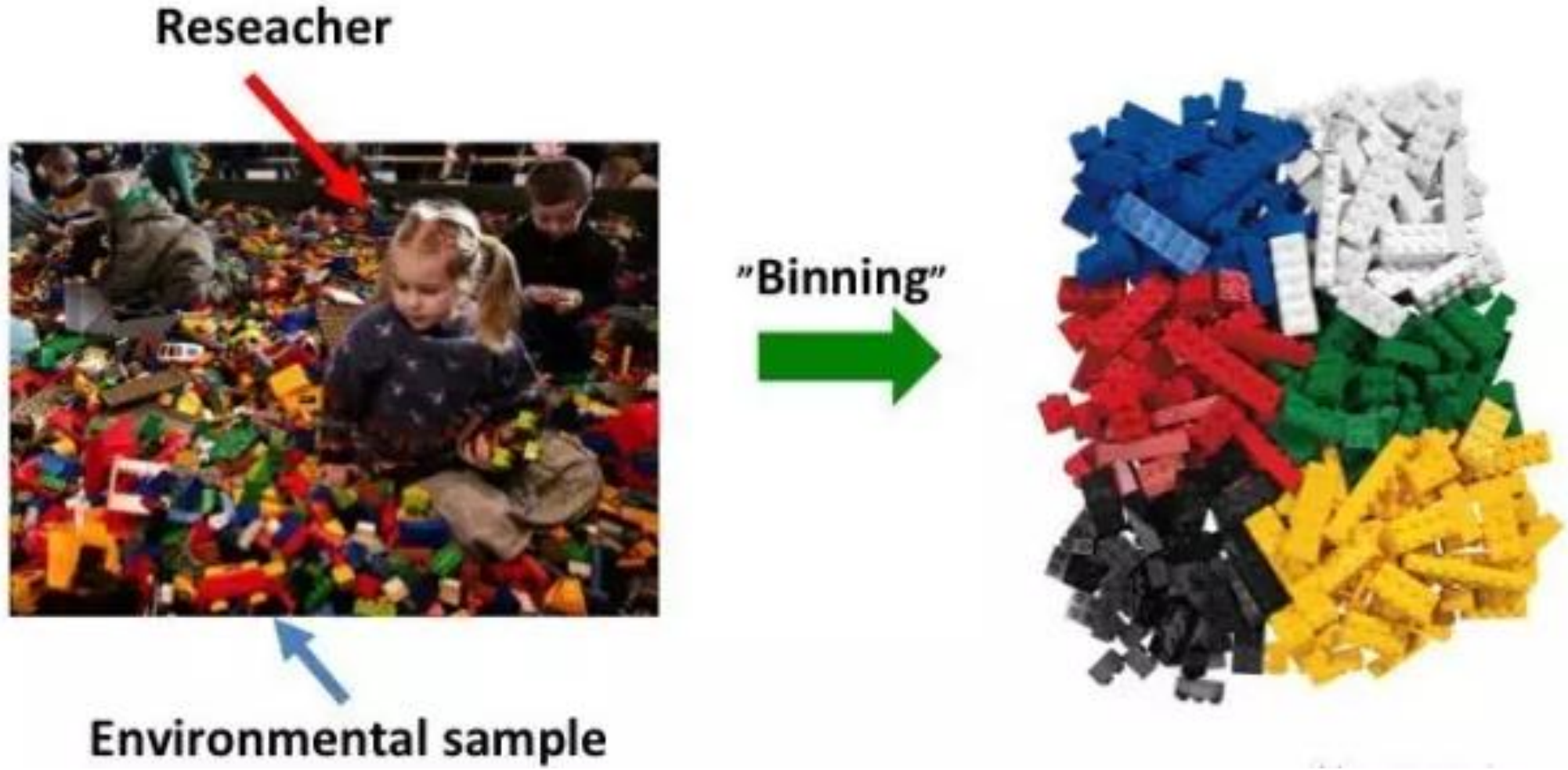


EggNOG 5.0.0

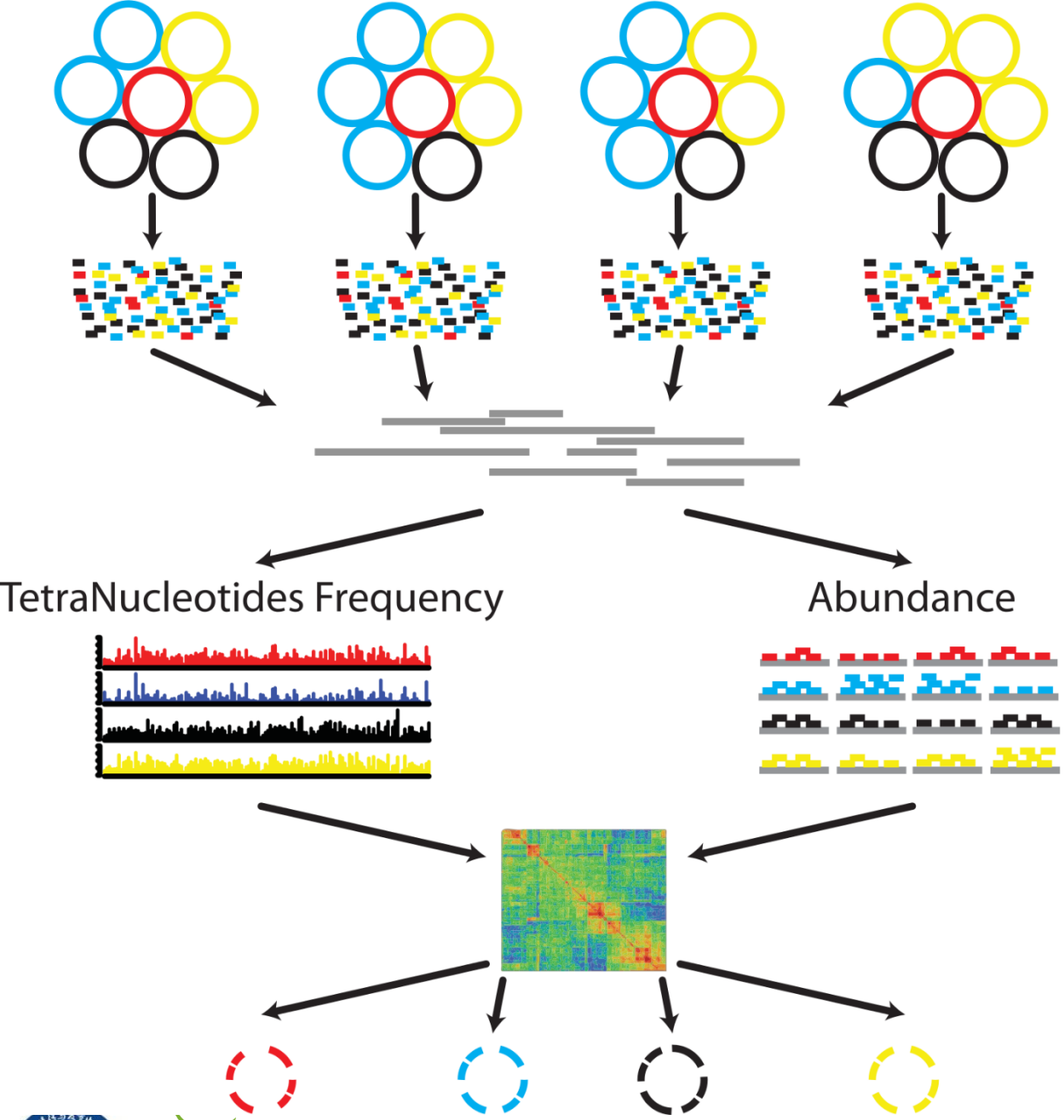
KEGG Kyoto Encyclopedia of Genes and Genomes

CAZy Carbohydrate-Active enZYmes

分箱的原理



《一文读懂宏基因组binning》



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

1. 不同时间地点的样品
2. 宏基因组测序
3. 组装为重叠群
4. 计算重叠群4核苷酸频率
5. 计算样本中重叠群丰度
6. 计算成对距离矩阵
7. 迭代获得分箱

Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT. ***PeerJ* 3**, e1165, doi:10.7717/peerj.1165 (2015).

CEPAMS

Nature子刊：宏基因组中挖掘原核基因组的分析流程

易宏基因组(EasyMetagenome)

- 1SoftDb.sh：软件和数据库安装
- 2Pipeline.sh：分析流程
  - 数据质量控制和去宿主
  - 有参分析流程HUMAnN2
  - 组装拼接流程
  - 分箱和基因组分析
- 3StatPlot.sh：统计和可视化代码
- 使用有道云笔记Markdown阅读
- 每季度更新

项目：https://github.com/YongxinLiu/EasyMetagenome

CEPAMS

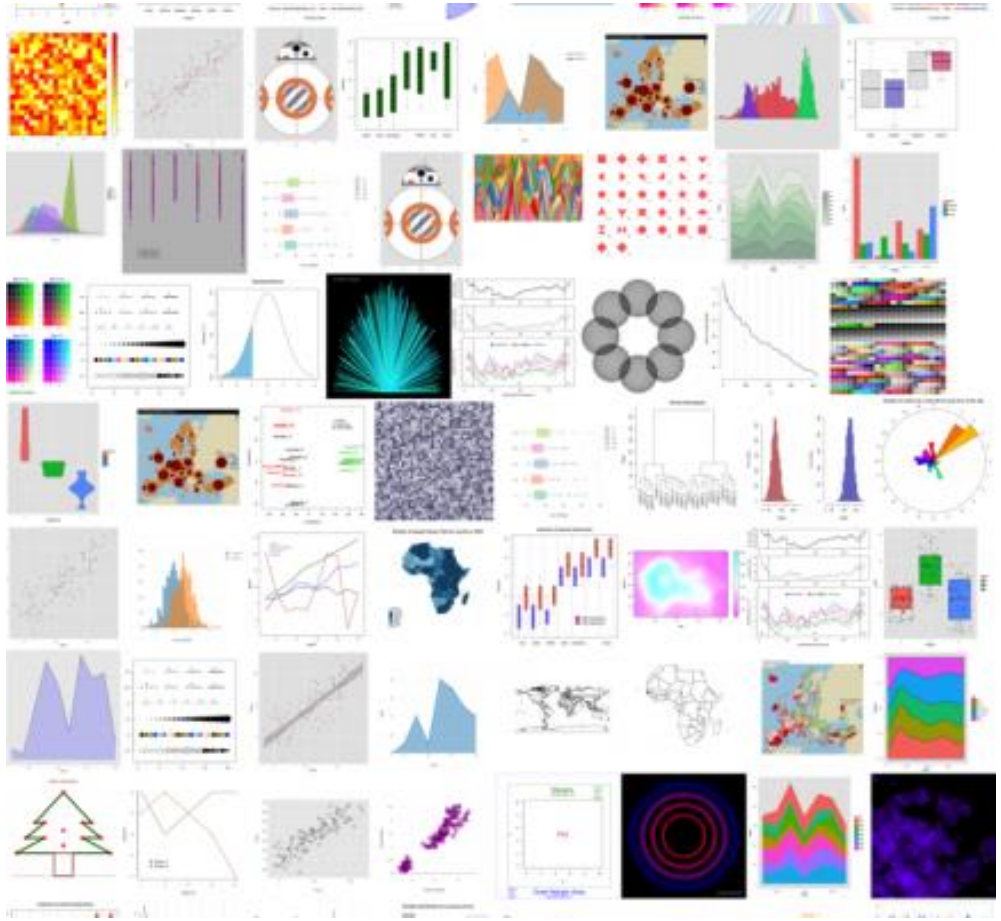
报告提纲

- 读懂文章图表
- 扩增子分析流程
- 宏基因组分析流程
- **多样性分析和可视化**
- 高分文章套路

CEPAMS

常用的可视化工具

Excel / GraphPad / SigmaPlot
图形界面、入门简单、商业收费、不可重复



R / Python / Java
代码界面，入门略难，开源免费、可重复

CEPAMS

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | 拖拽上传 | Project: (None)

R_basic.r

Source on Save | Run | Source

```
84
85  # 系统默认是灰主题，改为经典主题
86  p1 + theme_classic()
87
88  # 主题包括修改标题、坐标轴标签内容，字体等
89  p1 + theme_classic() +
90    labs(x = "X axis", y = "Y axis", title = "Title") +
91    theme(axis.title = element_text(size = 7),
92          axis.text =  element_text(size = 7),
93          plot.title = element_text(hjust = 0.5, size = 9))
94
95
```

94:1 (Top Level) R Script

代码编辑区

Environment | History | Connections | Presentation

Import Dataset | List

Global Environment

Data

mtcars	32 obs. of 11 variables
p	List of 9
p1	List of 9

环境变量/历史

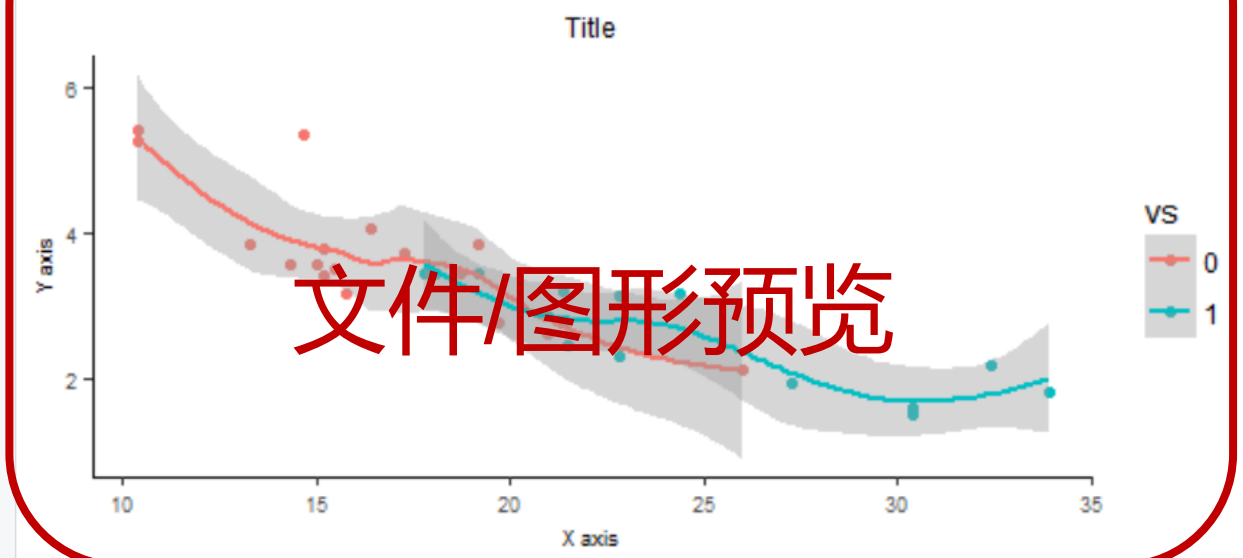
Console ~/

```
> p + geom_point() + stat_smooth() # 点图+拟合平滑曲线
`geom_smooth()` using method = 'loess'
> # 设置主题
> # 散点和拟合确定使用，可进一步保存
> p1 = p + geom_point() + stat_smooth()
> # 系统默认是灰主题，改为经典主题
> p1 + theme_classic()
`geom_smooth()` using method = 'loess'
> # 主题包括修改标题、坐标轴标签内容，字体等
> p1 + theme_classic() +
+   labs(x = "X axis", y = "Y axis", title = "Title") +
+   theme(axis.title = element_text(size = 7),
+         axis.text =  element_text(size = 7),
+         plot.title = element_text(hjust = 0.5, size = 9))
`geom_smooth()` using method = 'loess'
>
```

代码执行区

Files | Plots | Packages | Help | Viewer

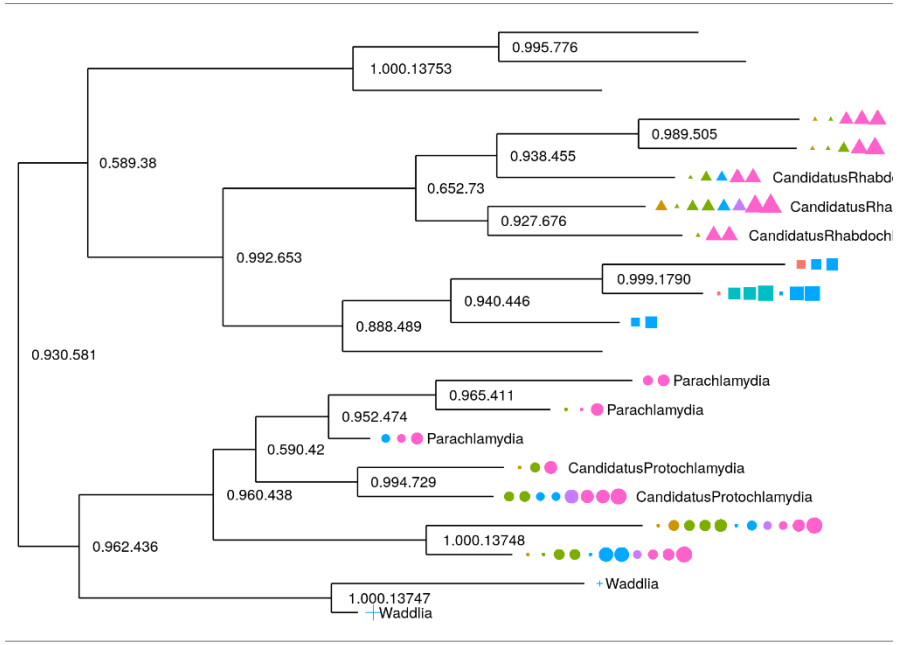
Zoom | Export | Publish



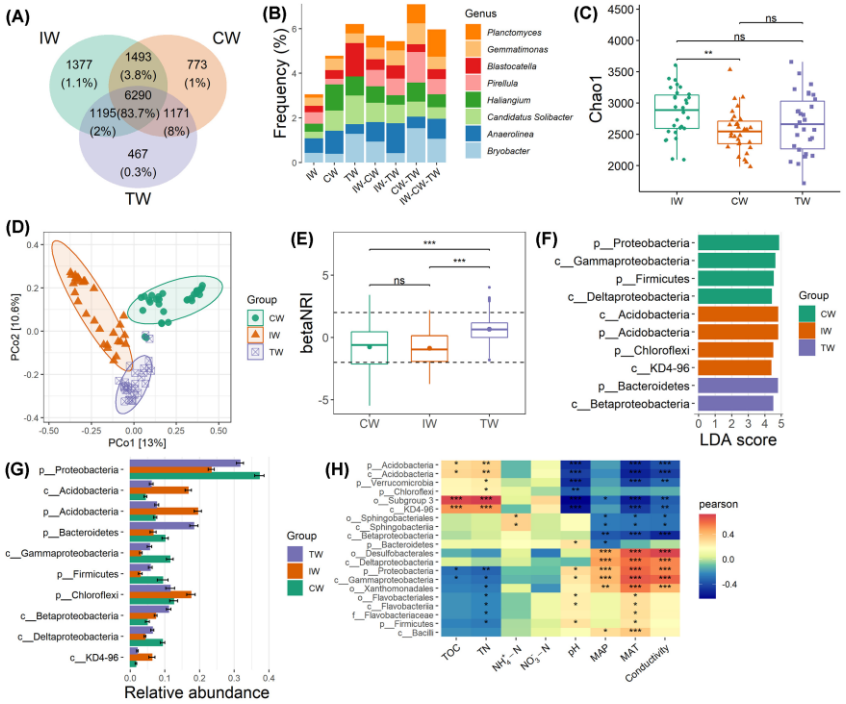
文件/图形预览

0:06 2018/3/1

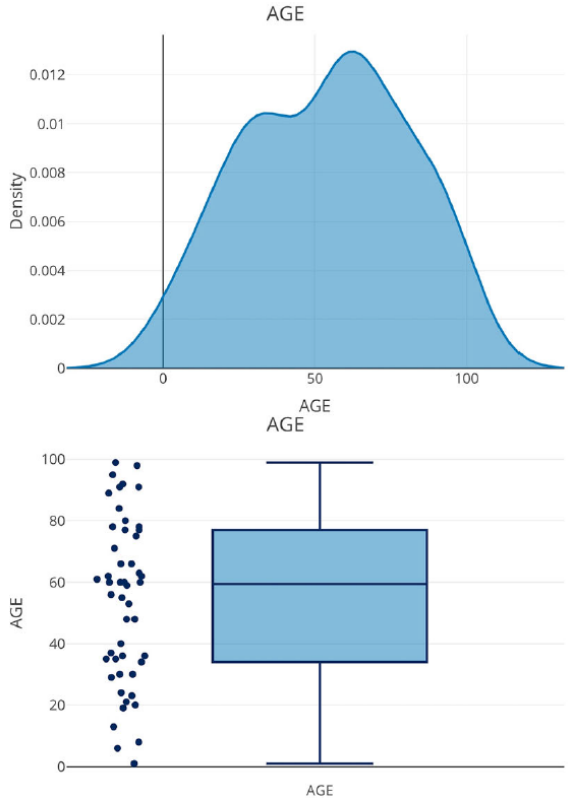
可视化的常用R包



phyloseq



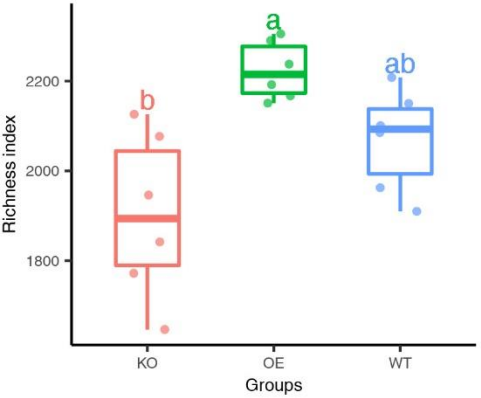
microeco

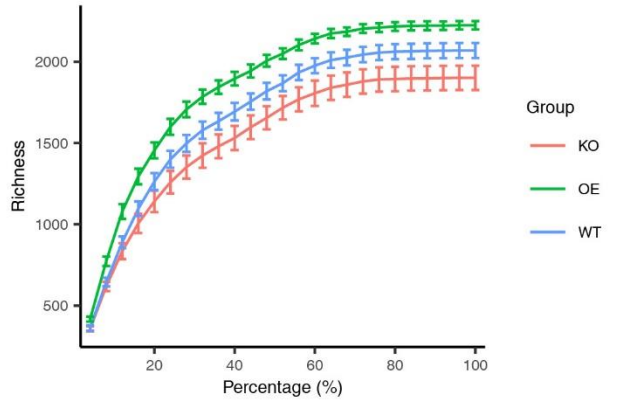


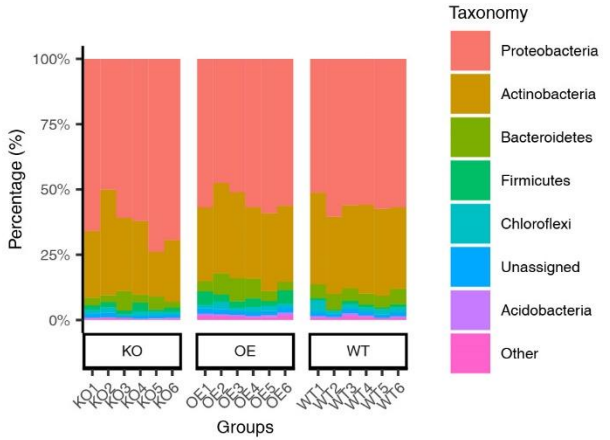
animalcules

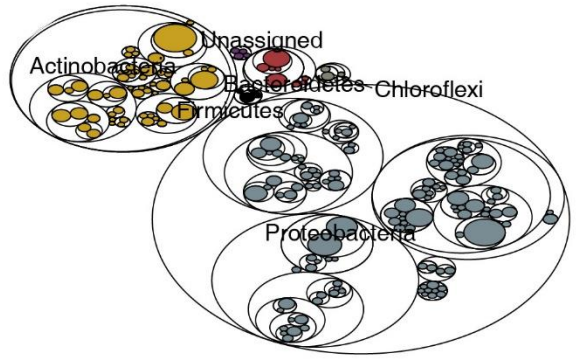
- [微生物组统计和可视化——phyloseq入门]
- [FEMS：微生物群落生态学数据挖掘的R包microeco]
- [R包animalcules-一键式交互探索微生物组数据] [Microbiome：animalcules-文章解读]

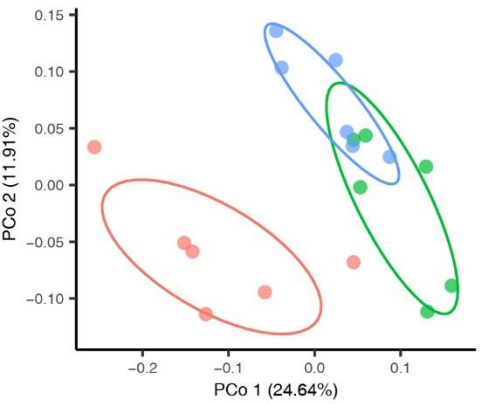
统计可视化amplicon包

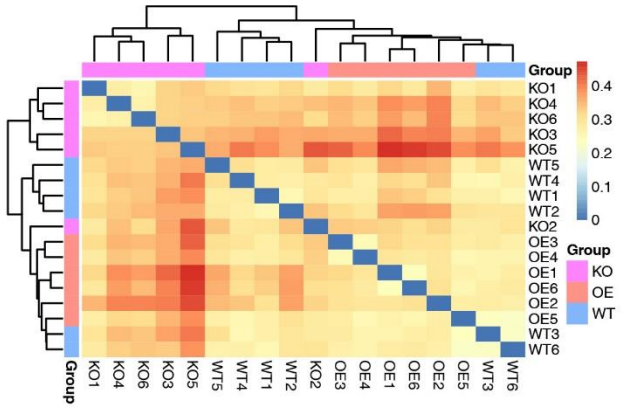


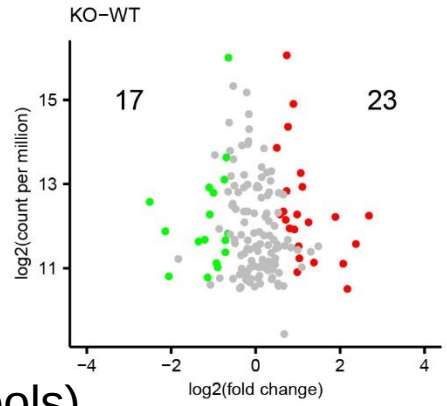


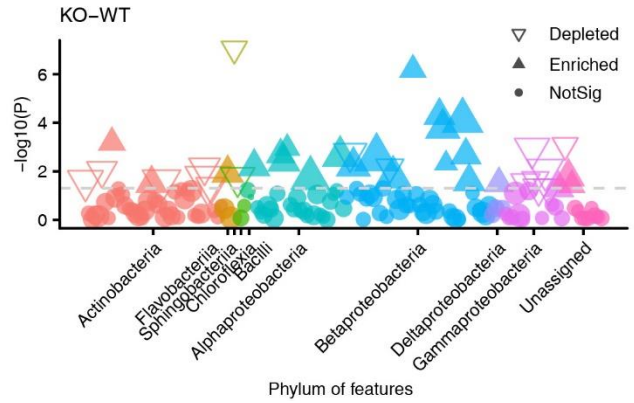













```
library(devtools)
install_github("microbiota/amplicon")
```

[微生物组可视化包：https://github.com/microbiota/amplicon](https://github.com/microbiota/amplicon)

一键生成发表级别图样式，欢迎大家为此项目贡献你的代码和函数

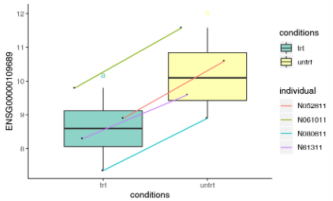
CEPAMS

ImageGP——在线绘制20种图/分析

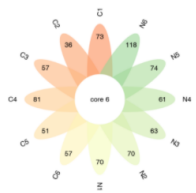
http://www.ehbio.com/ImageGP/

ImageGP 2图+代码可重复分析

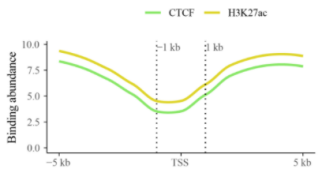
Boxplot



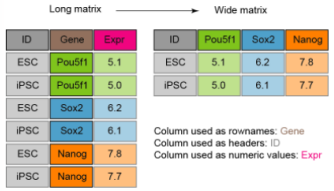
Flower plot



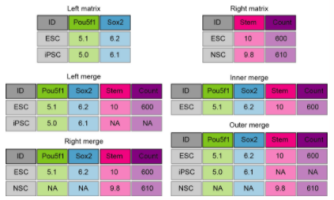
Line plot



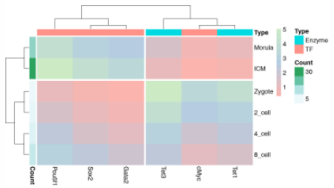
Long to wide matrix



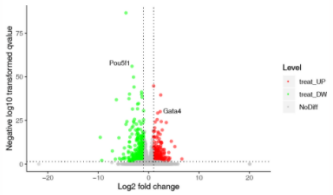
Merge matrixes



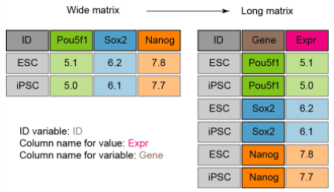
Pretty heatmap



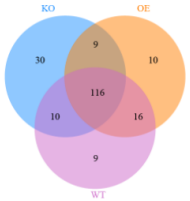
Volcano plot



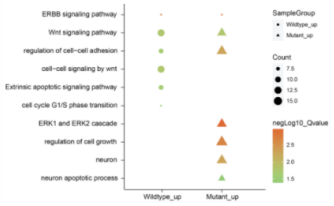
Wide to long matrix



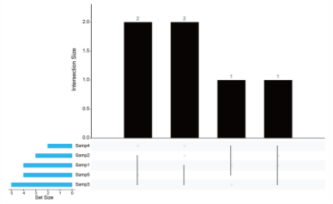
VennDiagram plot



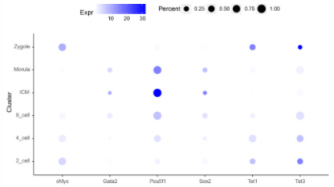
Enrichment plot



Upsetview plot



Scatter plot



http://www.ehbio.com/Cloud_Platform/front

报告提纲

- 读懂文章图表
- 扩增子分析流程
- 宏基因组分析流程
- 多样性分析和可视化
- **高分文章套路**

CEPAMS

高分文章套路

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

四．方法、摘要和封面可视化

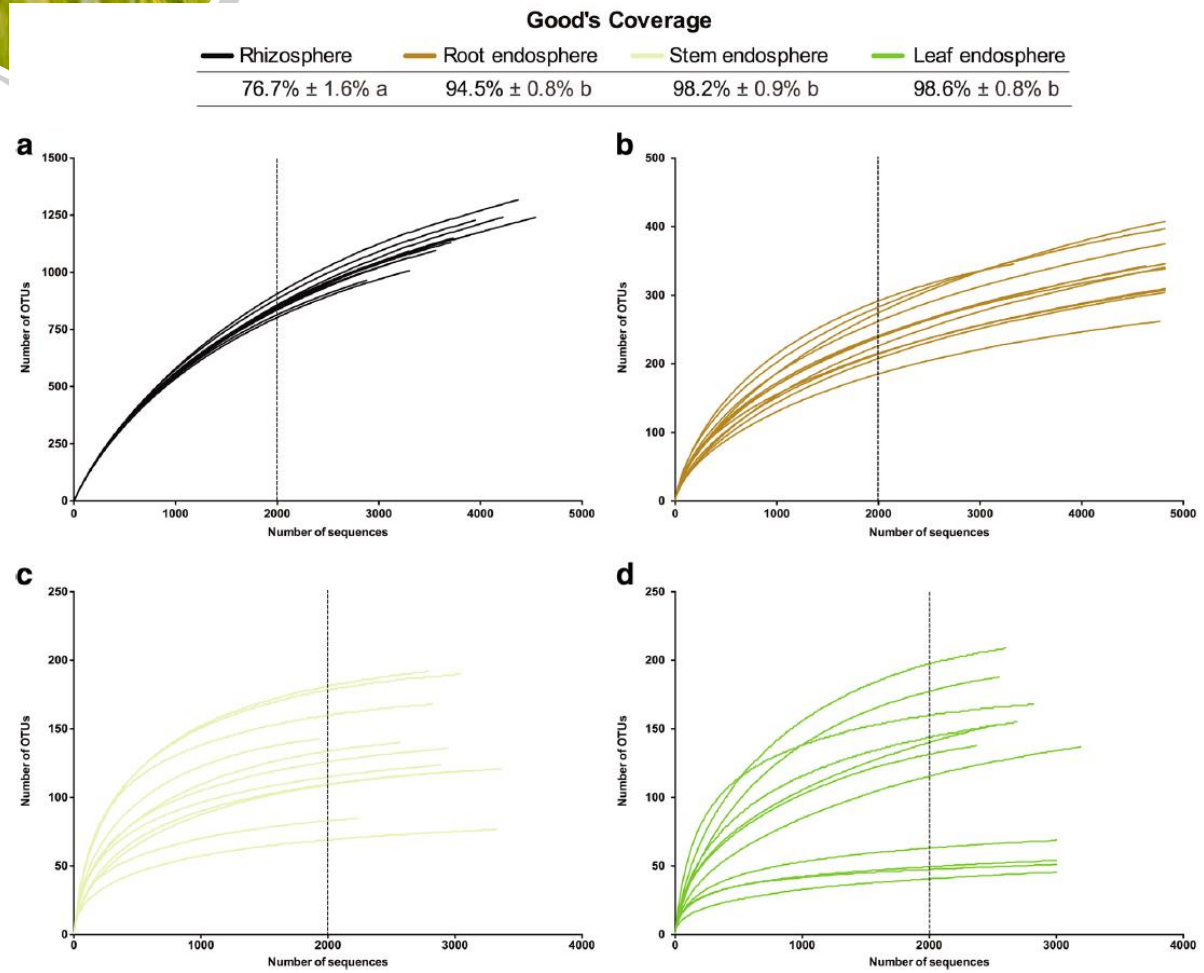
五．投稿经验和杂志选择

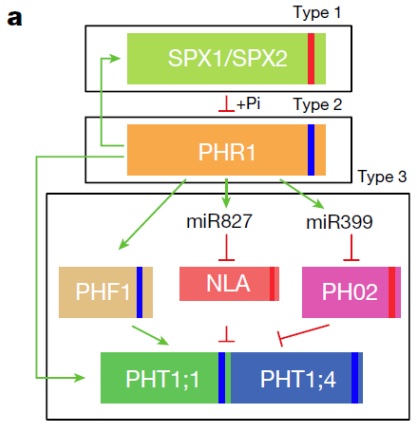
Yong-Xin Liu, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo & Yang Bai. (**2021**). A practical guide to amplicon and metagenomic analysis of microbiome data. ***Protein & Cell*** 12, 315-330, doi: https://doi.org/10.1007/s13238-020-00724-8

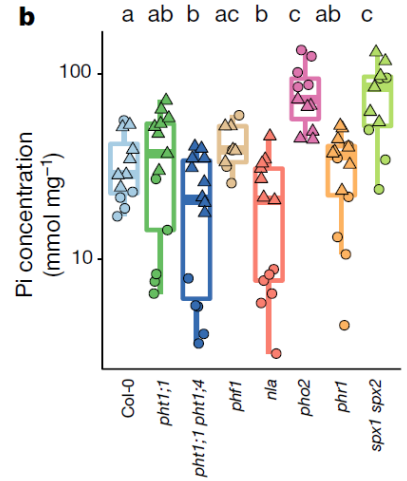
Protein Cell：扩增子和宏基因组数据分析实用指南

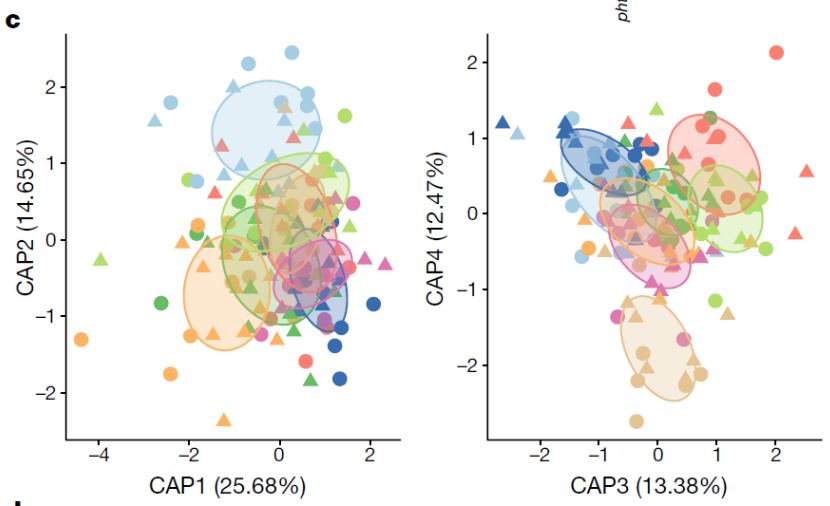
CEPAMS

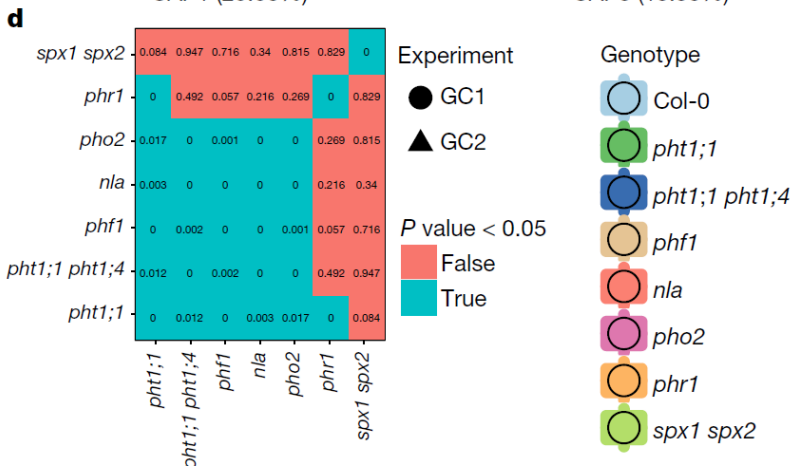
文章图的差距











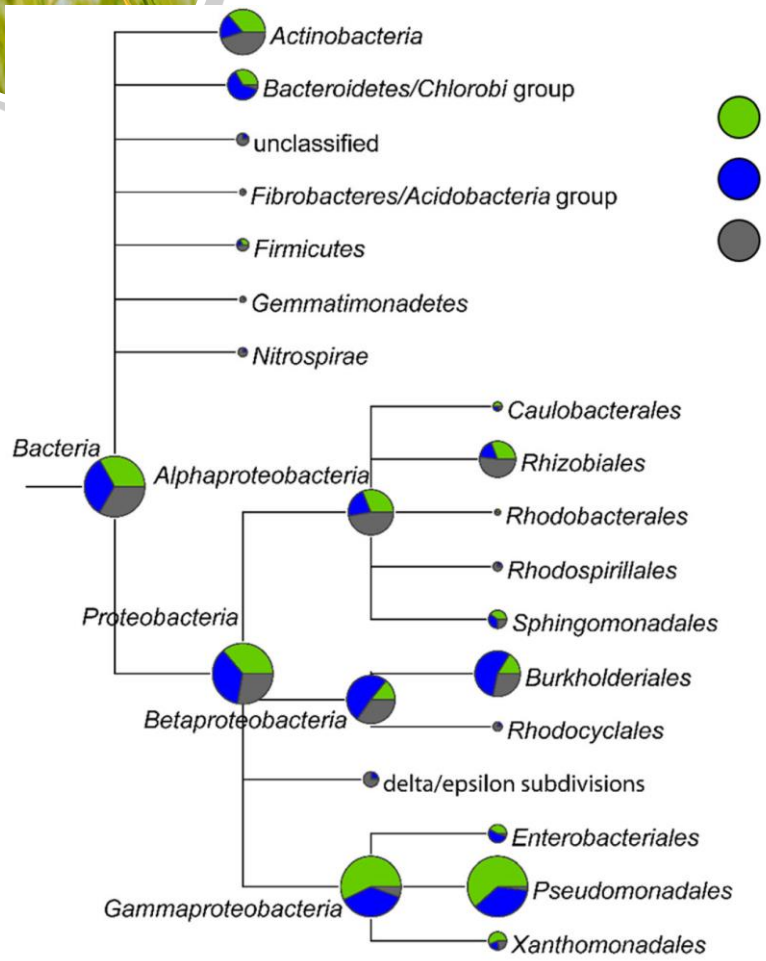
Nature：根系菌群参与磷胁迫和免疫的平衡

内容和表达方式丰富，排版更细致矢量图清晰

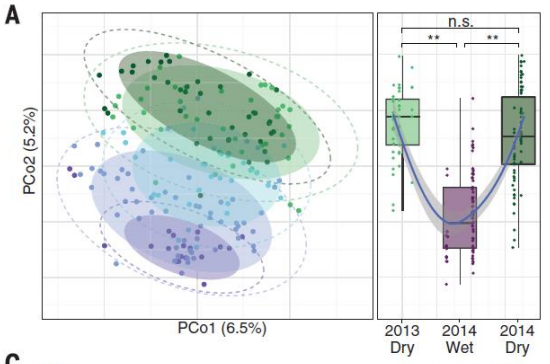
Microbiome: 简单套路发高分文章--杨树微生物组

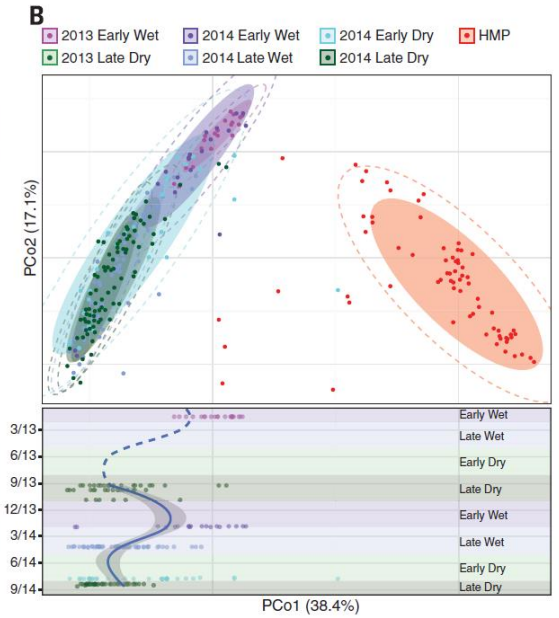
CEPAMS

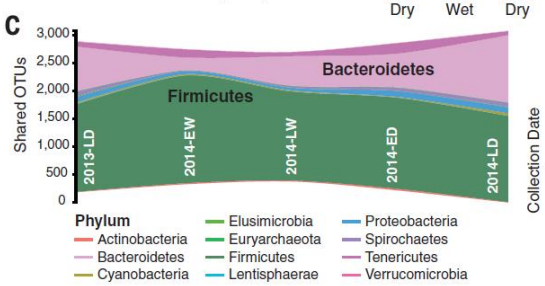
文章图的差距

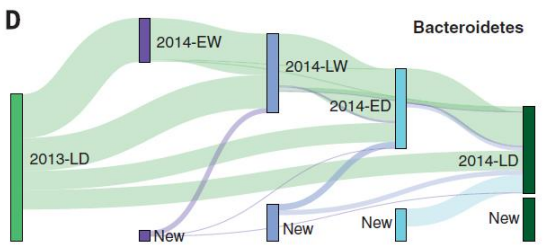


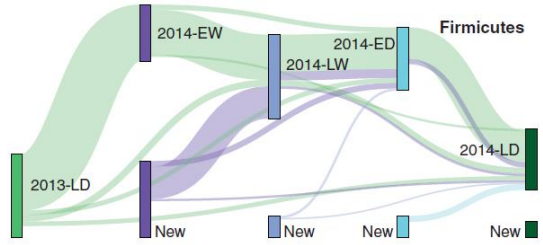
Microbiome：芝麻菜中肠杆菌科主导核心微生物组

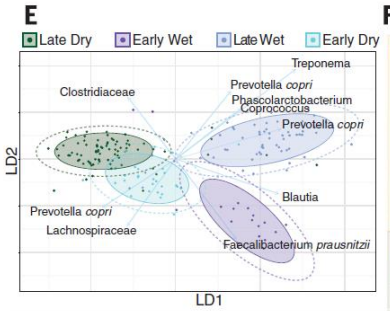


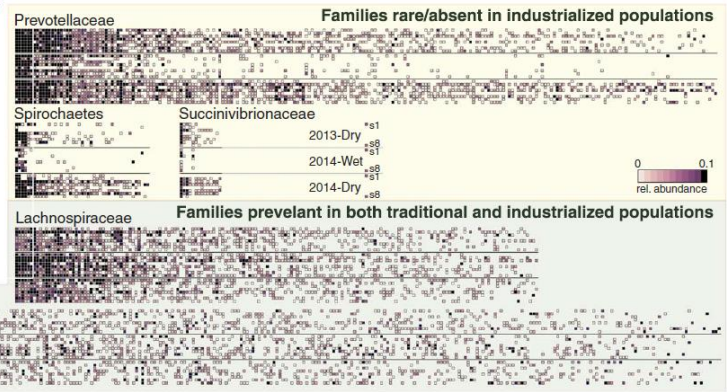












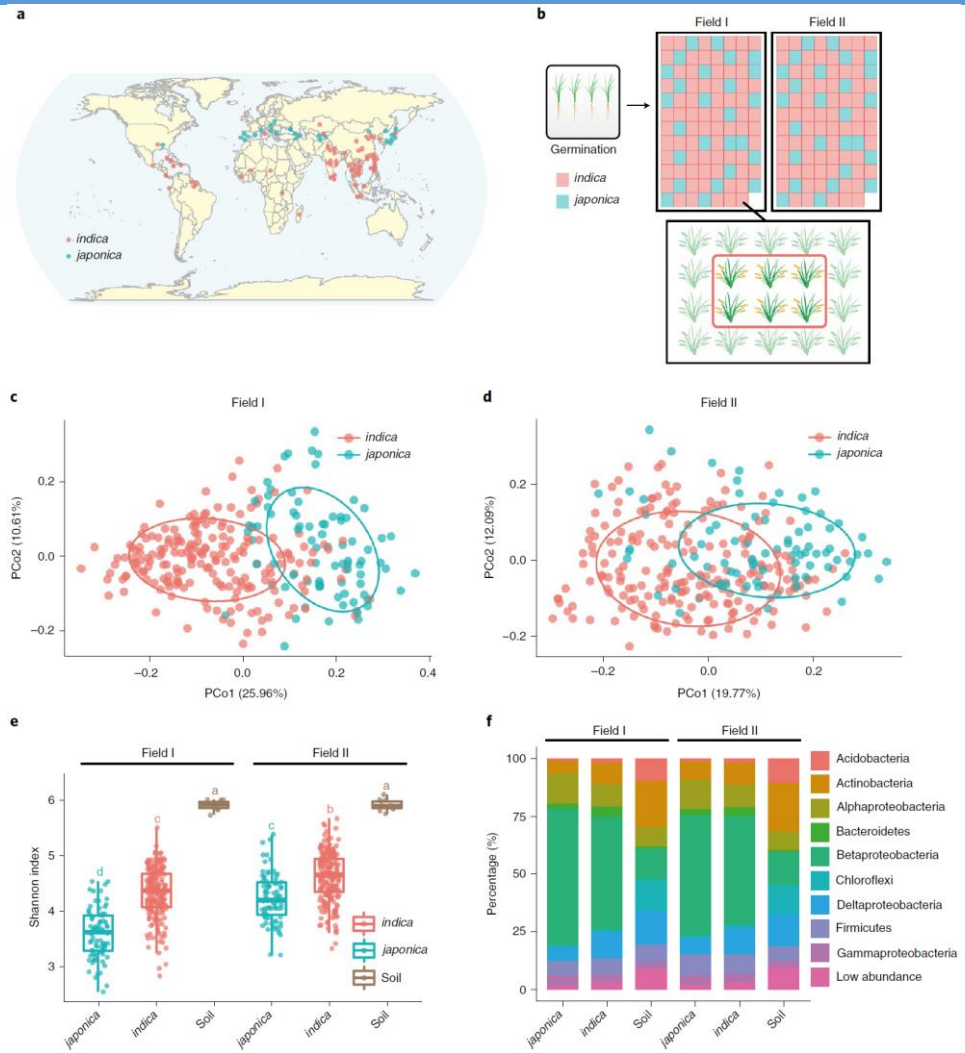
Science：哈扎人的肠道菌群

图形种类多样

PCoA
箱线图
拟合曲线
桑基图
流图
热图

套路总结

- 矢量图清晰美观、4-6个子图
- 多角度描述、多图形样式
- 开篇图1——总体描述
  1. 材料来源地图/材料关系图(a)
  2. 实验设计图+取样方式(b)
  3. 样本组间整体差异PCA/PCoA(c)+多年/多点重复(d)
  4. 其它角度整体描述，如多样性/均匀度(e)+物种/功能组成(f)——图中可进一步分面展示重复实验突出规律一致性



[NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用]

Jingying Zhang#, **Yong-Xin Liu**#, Na Zhang#, et. al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. ***Nature Biotechnology***. 2019. doi:10.1038/s41587-019-0104-4

CEPAMS

套路总结

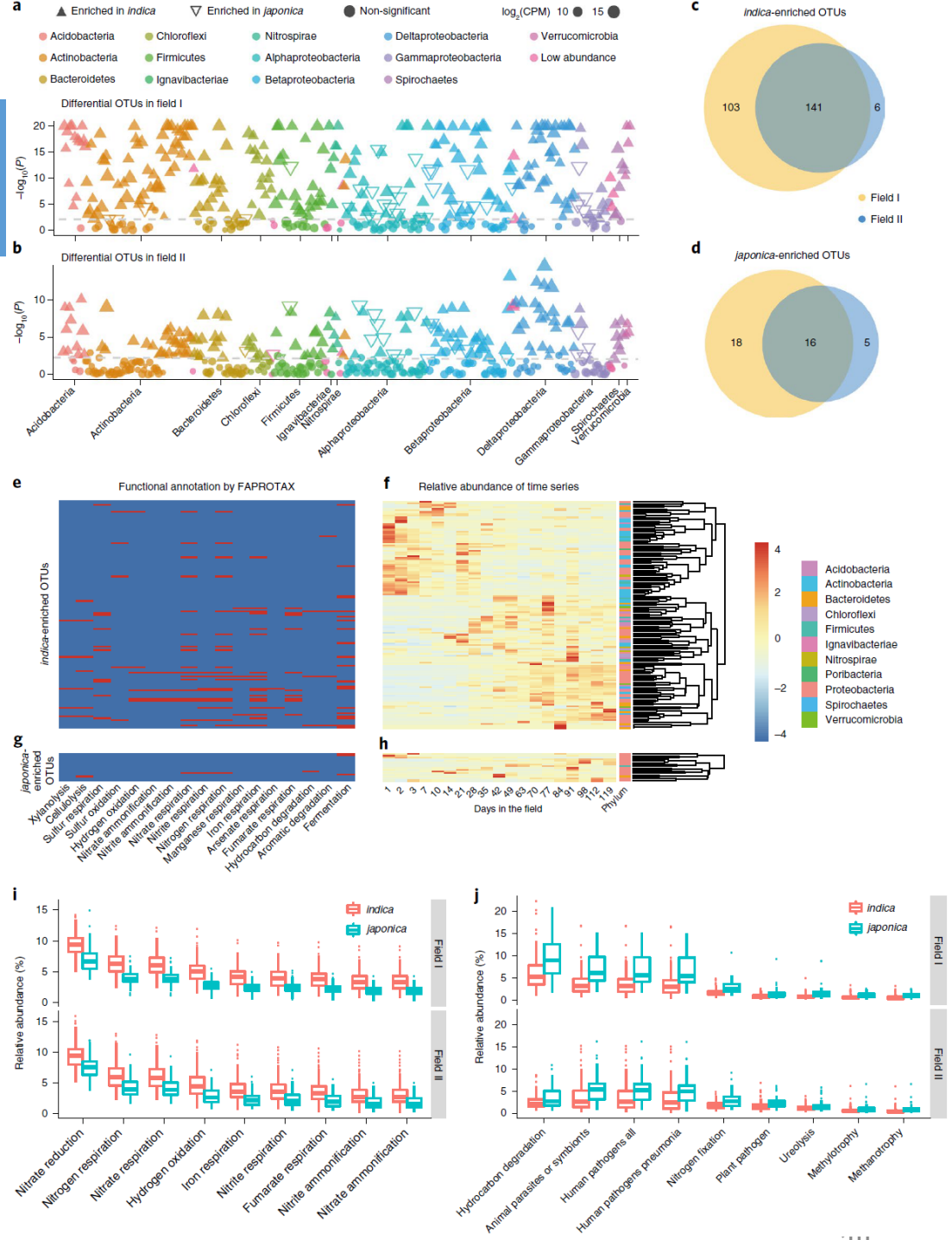
图2-4：细节

1.组间物种差异：热图/火山图/曼哈顿图(a/b)展示上、下调物种分类

2.重复比较：多年、多点或多组间可重复差异数量(c/d)；跨时间、地点间可检测才具有普适性(高分文章得分点)

3.差异物种的功能描述：功能注释/KEGG通路(e/g)+时间序列的变化规律(f/h)

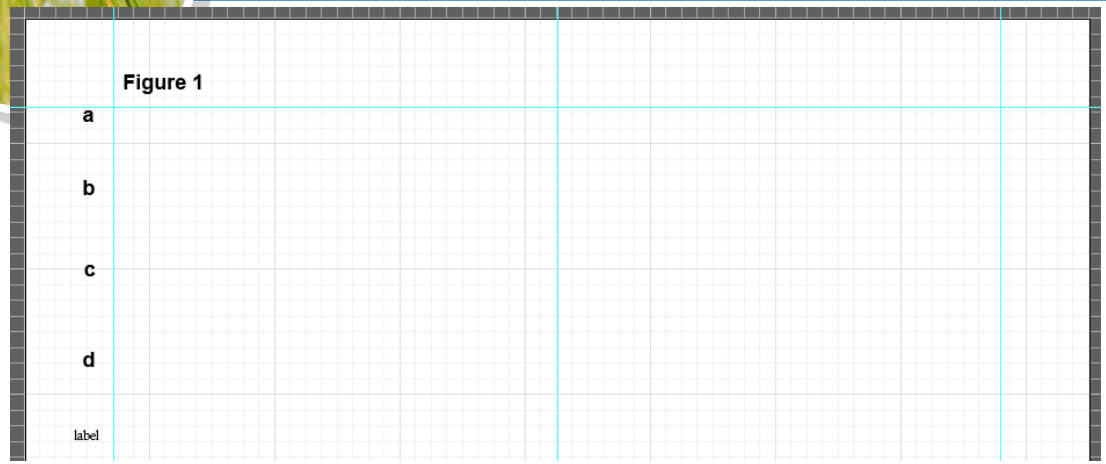
4.差异功能：箱线图、条形图、扩展柱状图展示差异功能(i/j)的细节，差异可分上/下调两类，重复可分面如上下图展示多地点重复



- [NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用]

CEPAMS

最常用的半/全版格式(Nature)



半版图

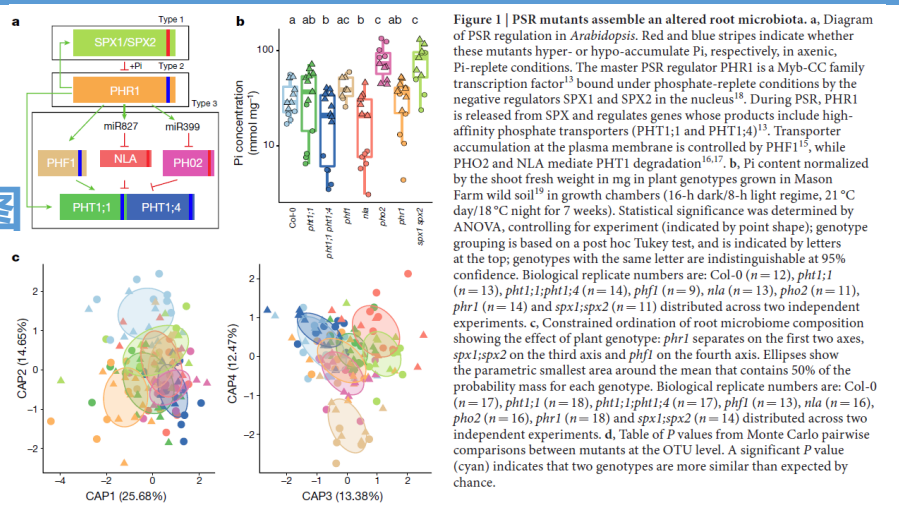


Figure 1 | PSR mutants assemble an altered root microbiota. a, Diagram of PSR regulation in *Arabidopsis*. Red and blue stripes indicate whether these mutants hyper- or hypo-accumulate Pi, respectively, in axenic, Pi-replete conditions. The master PSR regulator PHR1 is a Myb-CC family transcription factor[13] bound under phosphate-replete conditions by the negative regulators SPX1 and SPX2 in the nucleus[18]. During PSR, PHR1 is released from SPX and regulates genes whose products include high-affinity phosphate transporters (PHT1;1 and PHT1;4)[13]. Transporter accumulation at the plasma membrane is controlled by PHF1[15], while PHO2 and NLA mediate PHT1 degradation[16,17]. **b**, Pi content normalized by the shoot fresh weight in mg in plant genotypes grown in Mason Farm wild soil[19] in growth chambers (16-h dark/8-h light regime, 21 °C day/18 °C night for 7 weeks). Statistical significance was determined by ANOVA, controlling for experiment (indicated by point shape); genotype grouping is based on a post hoc Tukey test, and is indicated by letters at the top; genotypes with the same letter are indistinguishable at 95% confidence. Biological replicate numbers are: Col-0 ($n = 12$), *pht1;1* ($n = 13$), *pht1;1;pht1;4* ($n = 14$), *phf1* ($n = 9$), *nla* ($n = 13$), *pho2* ($n = 11$), *phr1* ($n = 14$) and *spx1;spx2* ($n = 11$) distributed across two independent experiments. **c**, Constrained ordination of root microbiome composition showing the effect of plant genotype: *phr1* separates on the first two axes, *spx1;spx2* on the third axis and *phf1* on the fourth axis. Ellipses show the parametric smallest area around the mean that contains 50% of the probability mass for each genotype. Biological replicate numbers are: Col-0 ($n = 17$), *pht1;1* ($n = 18$), *pht1;1;pht1;4* ($n = 17$), *phf1* ($n = 13$), *nla* ($n = 16$), *pho2* ($n = 16$), *phr1* ($n = 18$) and *spx1;spx2* ($n = 14$) distributed across two independent experiments. **d**, Table of *P* values from Monte Carlo pairwise comparisons between mutants at the OTU level. A significant *P* value (cyan) indicates that two genotypes are more similar than expected by chance.

全版图

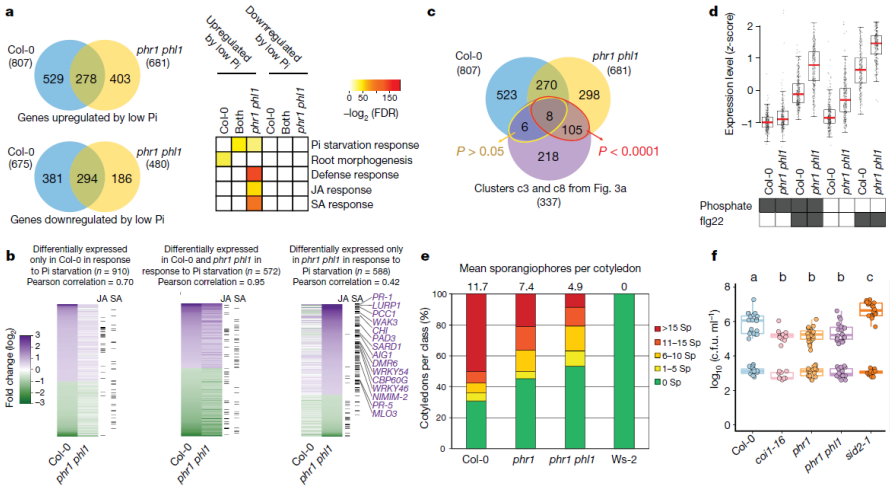


Figure 4 | Loss of PHR1 activity results in enhanced activation of plant immunity. a, Venn diagram (left) showing the overlap between genes upregulated and downregulated in Col-0 and *phr1;phl1* in response to marker genes differentially expressed following chronic exposure to flg22 (Extended Data Fig. 9). Averaged from six biological replicates. **e**, *phr1* exhibits enhanced disease resistance to the biotrophic oomycete pathogen

排版模板：*template/AI_Nature.ai*，颜色模式RGB
页面纸张为Letter，图片宽度为单栏89mm，双栏183mm，
图片最长大小为247mm

字体使用**Arial(英文黑体)**；氨基酸AA或核酸AGCT使用Courier系列等宽字体

子图编号为小写字母8 pt加粗，图注字体为5 - 7 pt,
推荐投稿使用12 pt加粗(子图分明) 和8 pt (文字可读性更好)

https://github.com/yongxinliu/db

CEPAMS

AI学习三部曲和扩展阅读

- 1基础入门和基本图形绘制
- 2模式图
- 3排版
- 在线浏览器，在线PS，在线AI，在线编程
- 论文Figures，你不能不知道的秘密
- 图表色彩运用原理的全面解析
- Graphpad经典绘图工具初学
- GraphPad Prism — 简单又好用的生物数据统计绘图软件

CEPAMS

中国高通量测序数据中心——GSA

推荐上传中科院基因组所 GSA

全中文界面 gsa.big.ac.cn

中文表格模板、中文帮助教程

邮箱 、QQ群技术支持 548170081

教育网、科技网传输速度极快

无人值守，24小时全自助提交

杂志广泛接受

http://gsa.big.ac.cn/

CEPAMS

上传数据的基本步骤

1. http://gsa.big.ac.cn/ 注册账号(首次)并登陆(“文档”按扭中有中文图文教程)
2. 提交 – BioProject提交入口 – 新BioProject(按提示填写项目基本简介)，提交成功后记录项目编号
3. FileZilla上传数据至 ftp://submit.big.ac.cn，帐号和密码同注册账号
4. 提交 – 新建GSA – 填写提交者和基本信息 – 选择样本类型：如Metagenome – 下载 Metagenome_or_environmental.cn.xlsx 模板填写样本名称、时间、地点、物种等信息并上传，可进行格式检查 – 下一步
5. 元数据表格GSA_Template.cn.xlsx下载，填写样本测序信息、文件名和md5值(md5sum *.fq.gz)等，并上传，可进行格式检查 – 文件位置选择FTP – 再次确认填写内容并提交

- 提示1：通常归档需要1-2天，数据越多越慢，归档后可随时释放
- 提示2：按中文模板填写，可参考我的模板，网站经常更新，尽量以官网最新下载模板为准

CEPAMS

中国核酸数据库GSA数据提交指南 扩增子16S实例 宏基因组实例

提交扩增子、宏基因组数据参考模板

1	2	3	4	5	6	7	8	9	10
nple_n	*public_description	project_accessio	mple_	*organism	host	isolation_source	collection_dat	eographic_locati	latitude_longitude
K01	Knock-out replicate 1	PRJCA002236	K01	Microbiota	Arabidopsi	Arabidopsis root	2017/6/30	China: Beijing	40.00 N 116.22 E
K02	Knock-out replicate 2	PRJCA002236	K02	Microbiota	Arabidopsi	Arabidopsis root	2017/6/30	China: Beijing	40.00 N 116.22 E
K03	Knock-out replicate 3	PRJCA002236	K03	Microbiota	Arabidopsi	Arabidopsis root	2017/7/2	China: Beijing	40.00 N 116.22 E
K04	Knock-out replicate 4	PRJCA002236	K04	Microbiota	Arabidopsi	Arabidopsis root	2017/7/2	China: Beijing	40.00 N 116.22 E

- 样本实验和数据信息： GSA_Template.cn_amplicon.xlsx

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
*ID	*Experiment title	Project access	ample	nple acc	*Platform	*Library Construction / Experimental Design	brary nan	*Strategy	*Source	electi	*Layout	gth for r	gth for m	rt size
E1	16S rDNA amplicon of kr	PRJCA002236	K01		Illumina HiSeq 2500	DNA for each sample was extracted with FastDNA SPIN Kit (MP		AMPLICON	METAGENOMIC	PCR	PAIRED	250	250	441
E2	16S rDNA amplicon of kr	PRJCA002236	K02		Illumina HiSeq 2500	DNA for each sample was extracted with FastDNA SPIN Kit (MP		AMPLICON	METAGENOMIC	PCR	PAIRED	250	250	441
E3	16S rDNA amplicon of kr	PRJCA002236	K03		Illumina HiSeq 2500	DNA for each sample was extracted with FastDNA SPIN Kit (MP		AMPLICON	METAGENOMIC	PCR	PAIRED	250	250	441

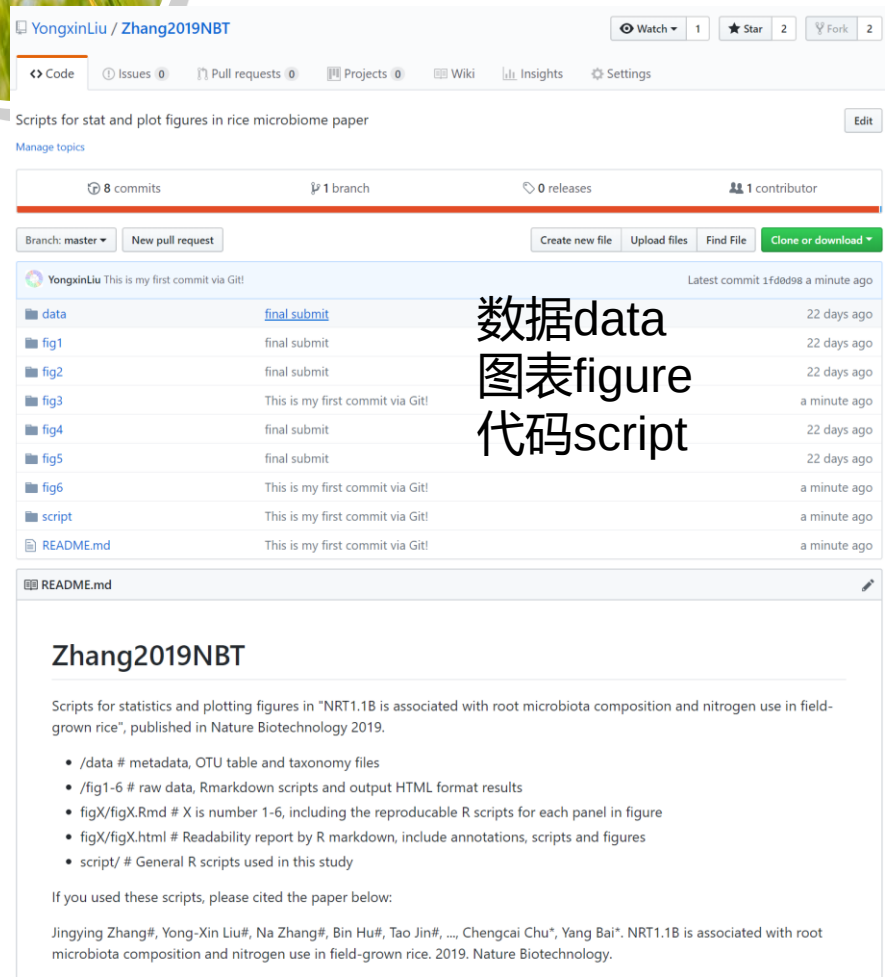
1	2	3	4	5	6	7	8	9
*ID	*Run title	Project access	perimen	data fil	*File name 1	*MD5 checksum 1	File name 2	MD5 checksum 2
R1	16S rDNA amplicon of knock-out replicate 1	PRJCA002236	E1	fastq	KO1_1.fq.gz	180d95da80536083bab9f5059e9d300c	KO1_2.fq.gz	33756be503f150603ef2aa4808aa3016
R2	16S rDNA amplicon of knock-out replicate 2	PRJCA002236	E2	fastq	KO2_1.fq.gz	c8bd3361ce2d3e192bc235540fffa995	KO2_2.fq.gz	1edb49a24f10decc6101feb0c5bf9ac1
R3	16S rDNA amplicon of knock-out replicate 3	PRJCA002236	E3	fastq	KO3_1.fq.gz	7617114a1f4a64c04bae83208e07d21c	KO3_2.fq.gz	7c3d4571c13a1d0347722e69ed899628

GSA_Template.cn_metagenome.xlsx

CEPAMS

https://github.com/yongxinliu/db 中 template 目录

数据代码图表整理示例



数据data
图表figure
代码script

简介、文件描述和引文

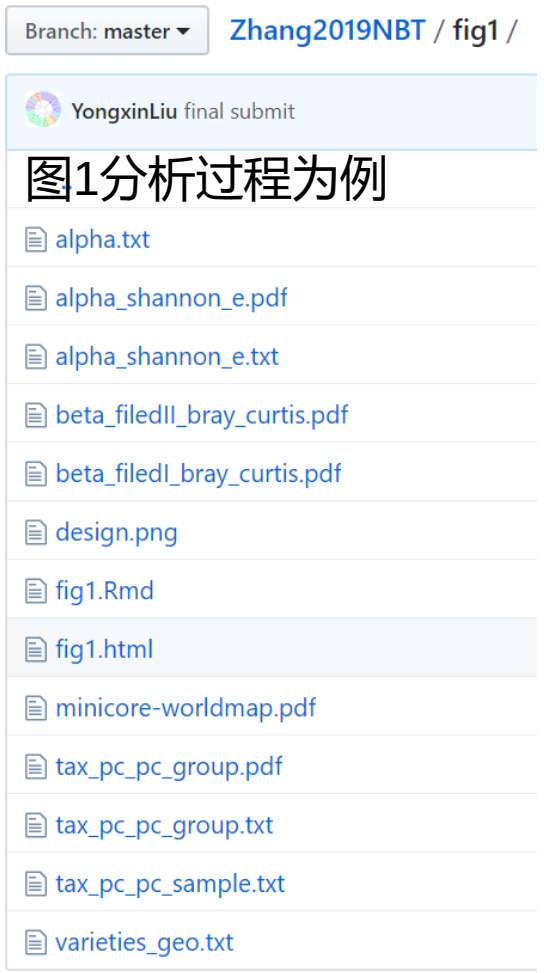


fig1.Rmd 分析过程代码及注释

fig1.html 分析过程代码、注释和图表混排网页，方便阅读

*.txt 分析原始数据或统计结果表格

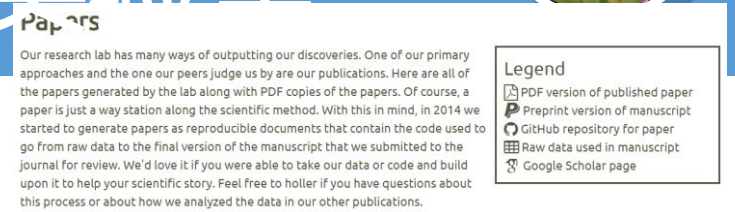
*.png 位图，如照片或绘画

*.pdf 矢量图，R语言绘制图片常用保存格式，方便查看和编辑

https://github.com/YongxinLiu/Zhang2019NBT

CEPAMS

提供开源可重复分析的实验室



- **密歇根大学Pat Schloss http://www.schlosslab.org**
- 期坦福大学Susan Holmes http://statweb.stanford.edu/~susan/
- 德国马普Paul Schulze-Lefert https://github.com/garridoo
- 北卡教堂山Jeffery L. Dangl https://github.com/isaisg / surh
- EMBL-EBI Robert D. Finn https://github.com/Finn-Lab/MGS-gut
- 比利时鲁汶大学 Jeroen Raes https://github.com/raeslab
- 贝勒医学院 Christopher J. Stewart https://github.com/StewartLab
- 遗传发育所/CEPAMS白洋 https://github.com/microbiota

CEPAMS

文件附件整理

Science
AAAS

Supplementary Materials for

A specialized metabolic network selectively modulates *Arabidopsis* root microbiota

Ancheng C. Huang,[1#] Ting Jiang,[2,3,4#] Yong-Xin Liu,[2,3] Yue-Chen Bai,[5,6] James Reed,[1] Baoyuan Qu,[2,3] Alain Goossens,[5,6] Hans-Wilhelm Nützmann,[7] Yang Bai,[2,3,4*] Anne Osbourn[1*]
Correspondence to: anne.osbourn@jic.ac.uk or ybai@genetics.ac.cn

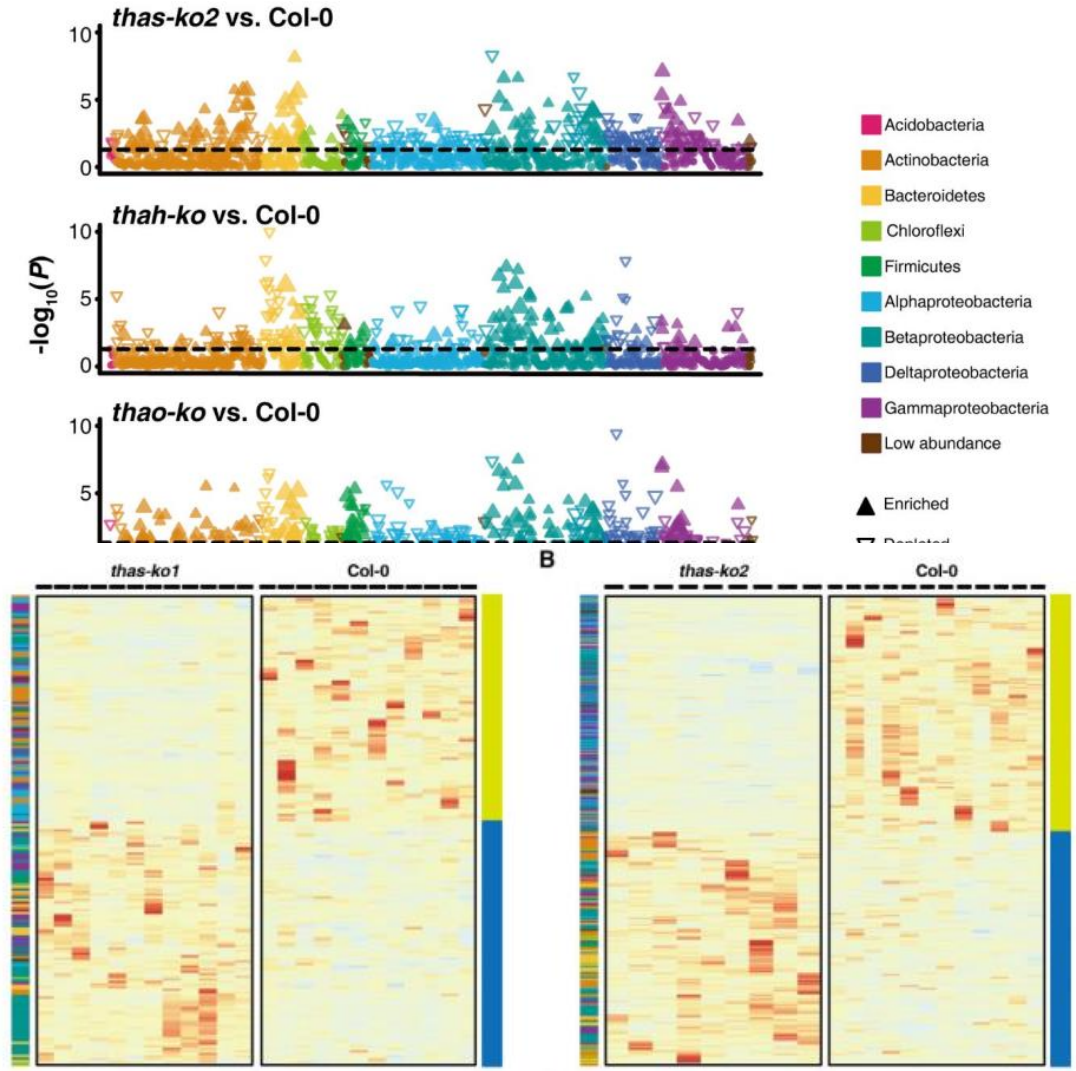
This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S64
Tables S1 to S17
References (46-60)

Other Supplementary Materials for this manuscript include the following:
Tables S18-65 are in a separate excel file (captions listed below).

Table S18: Metadata.
Table S19: OTU table.
Table S20: Representative sequences.
Table S21: OTU taxonomy.
Table S22: PERMANOVA by Adonis Using Bray-Curtis Distance Matrices, *P* value corrected by FDR.

2 / 113



CEPAMS

https://science.sciencemag.org/content/364/6440/eaau6389

文件附件整理：附表

- 样本元数据
- OTU、ASV表
- 代表序列和物种注
- 所有图对应的表

Metadata

SampleID
ThasKO2r1
ThasKO2r2
ThasKO2r3
ThasKO2r4
ThasKO2r9
ThasKO2r10
ThasKO2r11
ThasKO2r16
ThasKO1r1
ThasKO1r2
ThasKO1r3
ThasKO1r7
ThasKO1r8
ThasKO1r9
ThasKO1r10
ThasKO1r13
ThasKO1r14
ThasKO1r15
ThasKO1r16
ThasKO1r17
Thaa2KOr7
Thaa2KOr8

OTU table

OTU	Th...
OTU_1	
OTU_10	
OTU_100	
OTU_1000	
OTU_10001	
OTU_10002	
OTU_10003	
OTU_10006	
OTU_10007	
OTU_1001	
OTU_10013	
OTU_10018	
OTU_10019	
OTU_1002	
OTU_10020	
OTU_10022	
OTU_10024	
OTU_10025	
OTU_10028	
OTU_10031	
OTU_10033	
OTU_10035	
OTU_10036	

Table

OTU taxonomy

OTU	Kingdom	Phylum	Class	Order
OTU_1	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales
OTU_10	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales
OTU_100	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales
OTU_1000	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales
OTU_10001	Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcales
OTU_10002	Bacteria	Nitrospirae	Nitrospira	Nitrospirales
OTU_10003	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales
OTU_10006	Bacteria	Verrucomicrobia	Subdivision3	Xanthomonadale
OTU_10007	Bacteria	Proteobacteria	Deltaproteobacteria	Desulfuromonad
OTU_1001	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales
OTU_10013	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales
OTU_10018	Bacteria	Proteobacteria	Deltaproteobacteria	Rhizobiales
OTU_10019	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales
OTU_1002	Bacteria	Proteobacteria	Deltaproteobacteria	Myxococcales
OTU_10020	Bacteria	Firmicutes	Bacilli	Bacillales
OTU_10022	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales
OTU_10024	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadal
OTU_10025	Bacteria	Acidobacteria	Acidobacteria_Gp5	Unassigned
OTU_10028	Bacteria	Firmicutes	Clostridia	Clostridiales
OTU_10031	Bacteria	Firmicutes	Bacilli	Bacillales
OTU_10033	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadal
OTU_10035	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillales
OTU_10036	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales

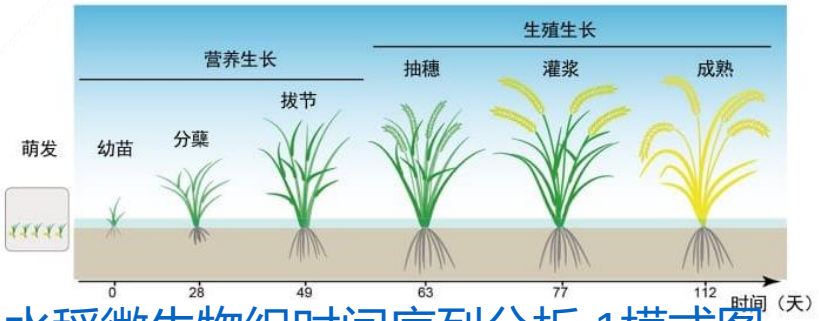
Table S18 | Table S19 | Table S20 | Table S21 | Table S22 | Table S23 | Table S24

CEPAMS

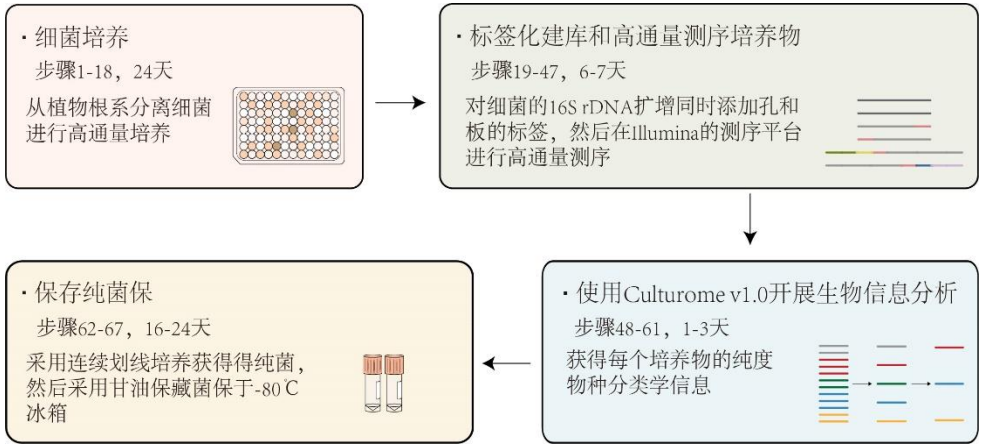
https://science.sciencemag.org/content/suppl/2019/05/08/364.6440.eaau6389.DC1

材料和方法可视化

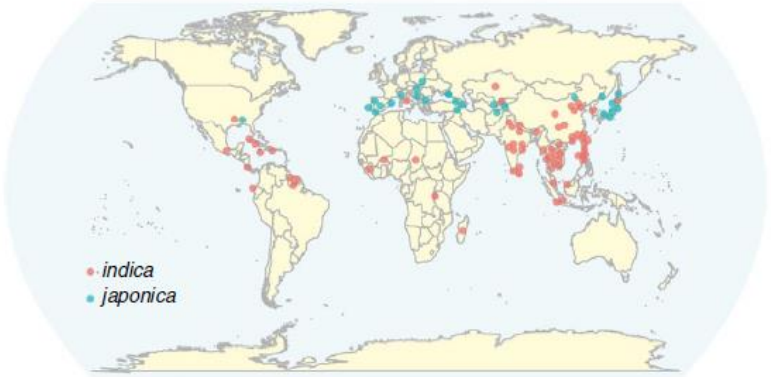
水稻田间全生育期示意图

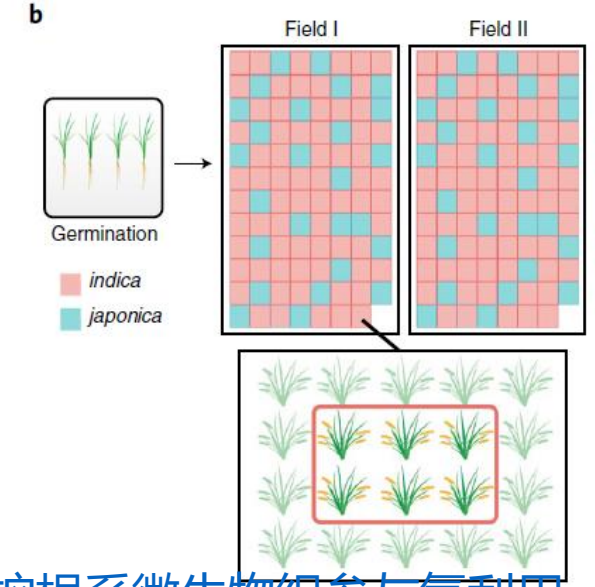


SCLS：水稻微生物组时间序列分析 1模式图



Nature子刊：高通量分离培养和鉴定根系细菌的方法





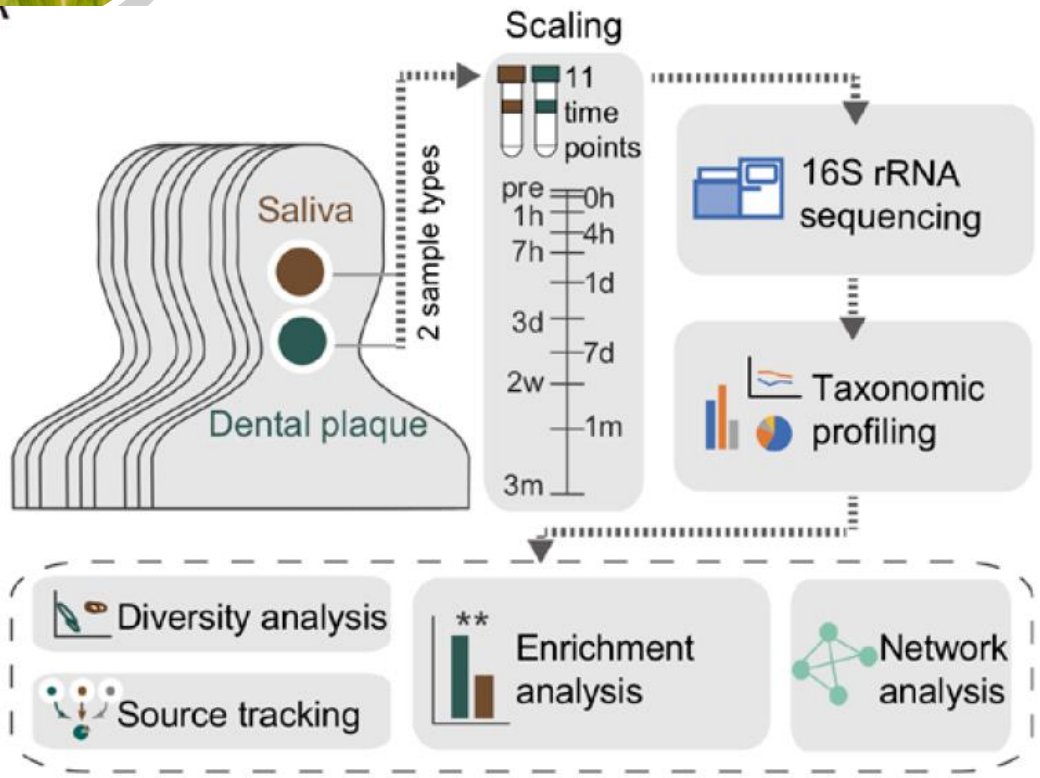
NBT封面：水稻NRT1.1B基因调控根系微生物组参与氮利用

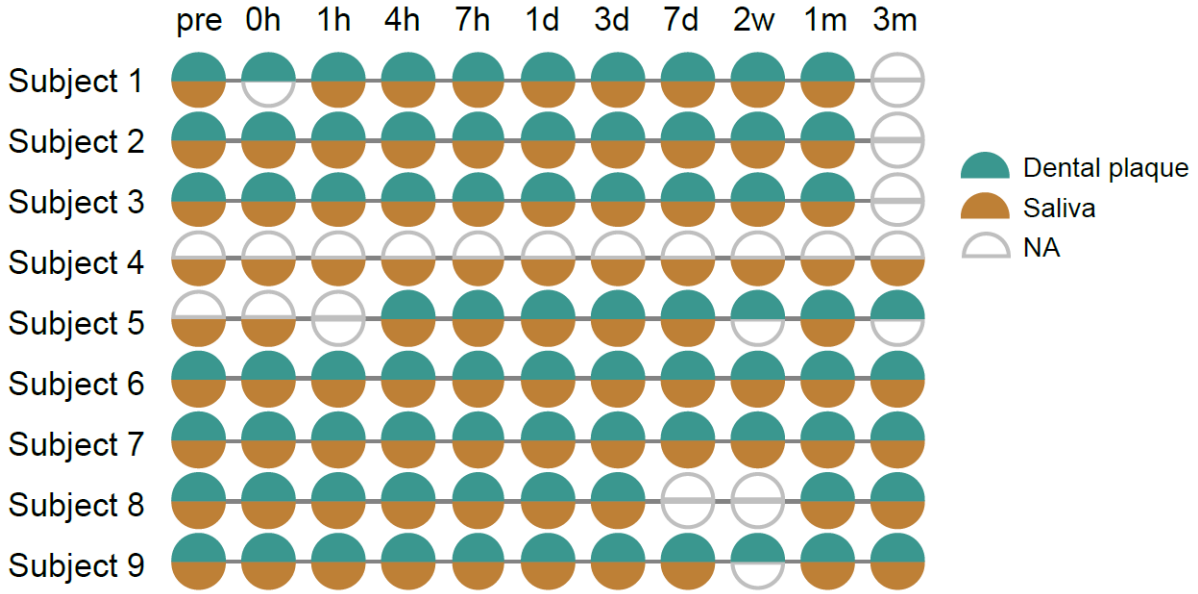
a. 地图，展示材料的全球来源；通常用于展示取样来源

b. 田间实验和取样设计。两组用不同颜色体现随机种植分布，两地块体实验存在大规模重复，单点取样放大图展示取样方案，小区存在保护行防品种混杂。

CEPAMS

材料和分析方法可视化示例



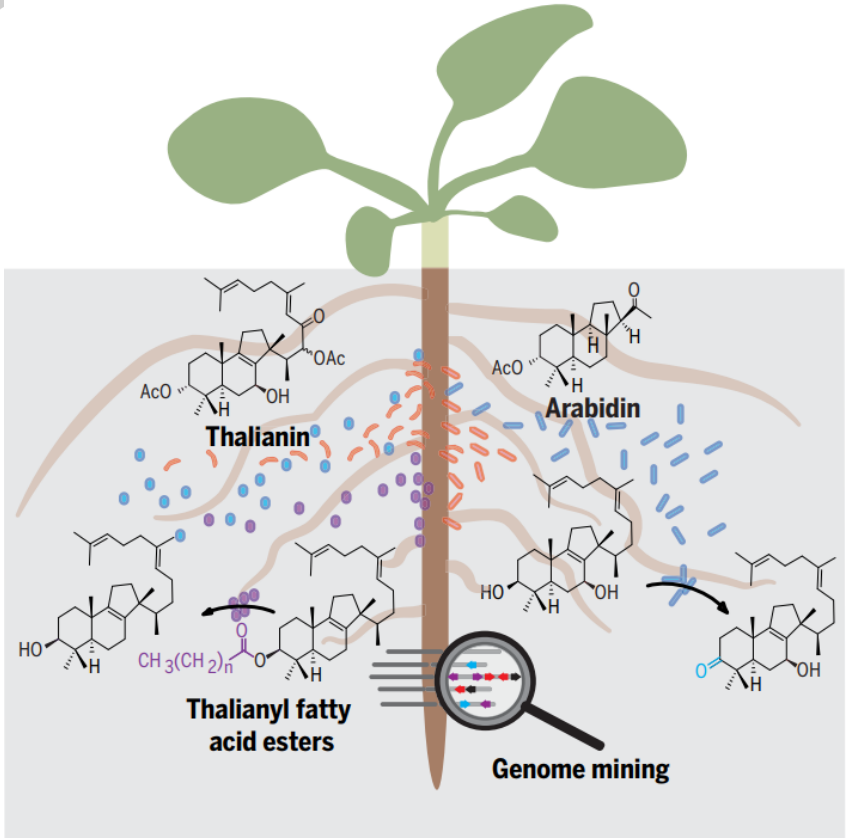


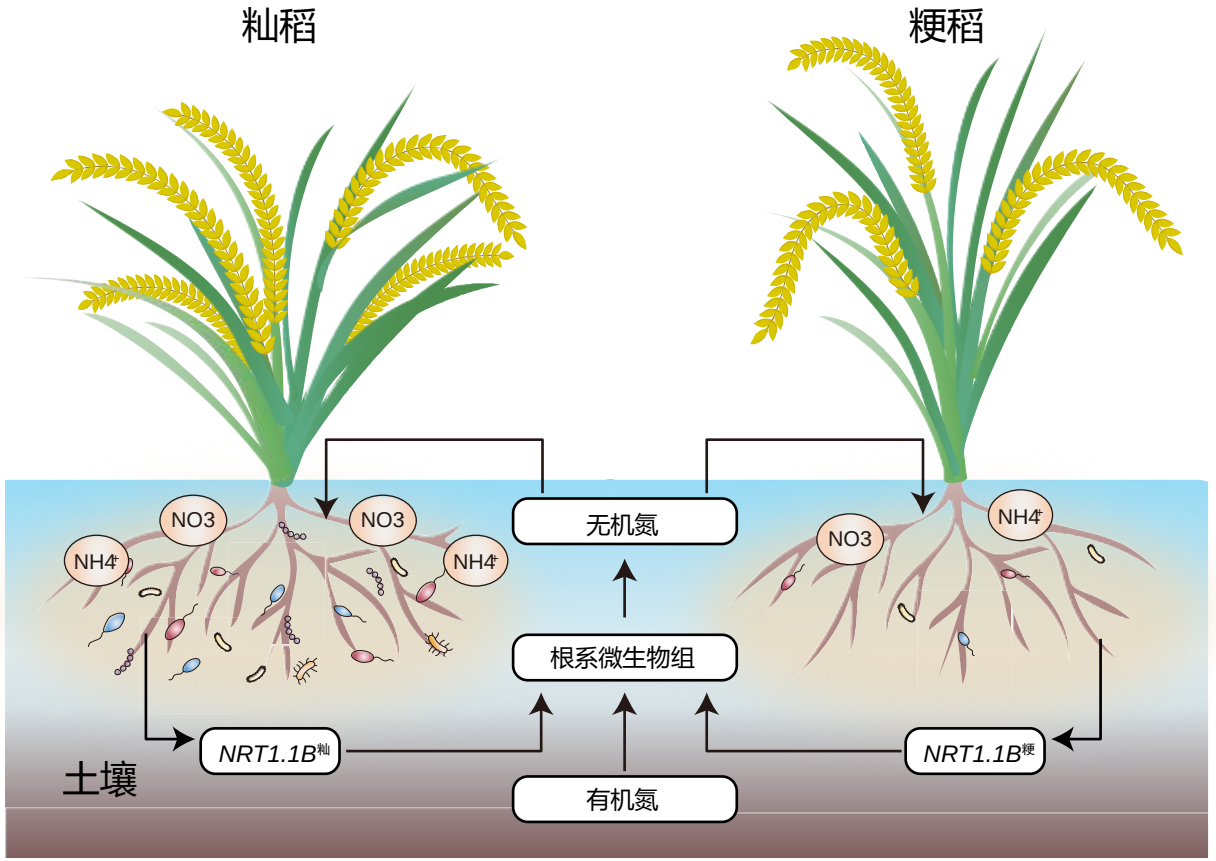
个体在时间序列上的取样分布图

研究对象、取样部分、取样时间、测序方法、和分析方法(物种组成、多样性、溯源分析、差异比较、网络分析)的可视化

Gut：人体口腔菌群的稳定性和动态变化规律

摘要可视化

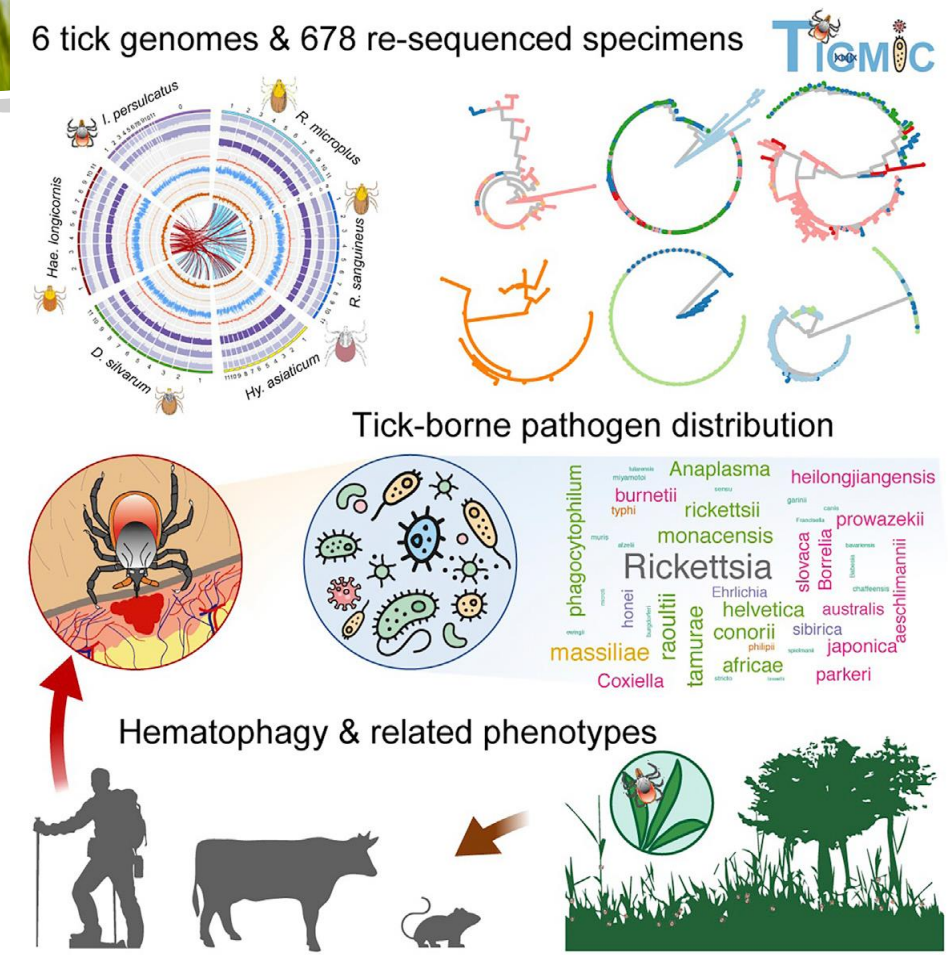




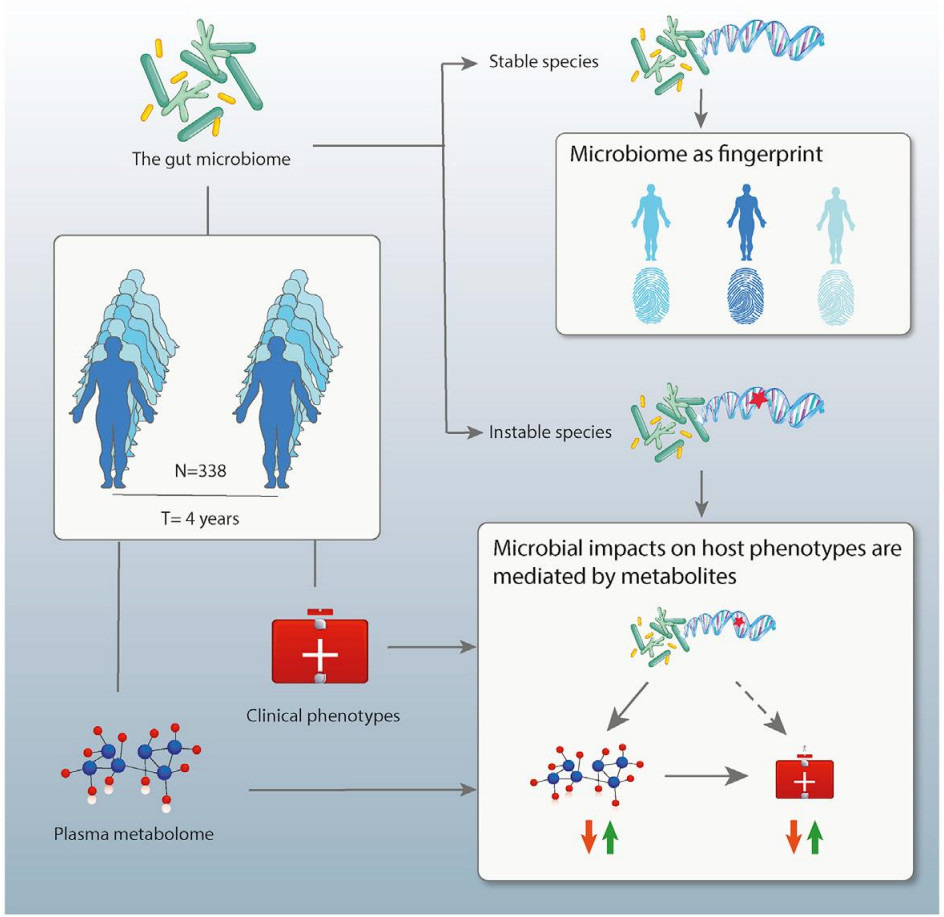
Science：拟南芥三萜化合物特异调控根系微生物组

NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用

Cell系列期刊的图形摘要



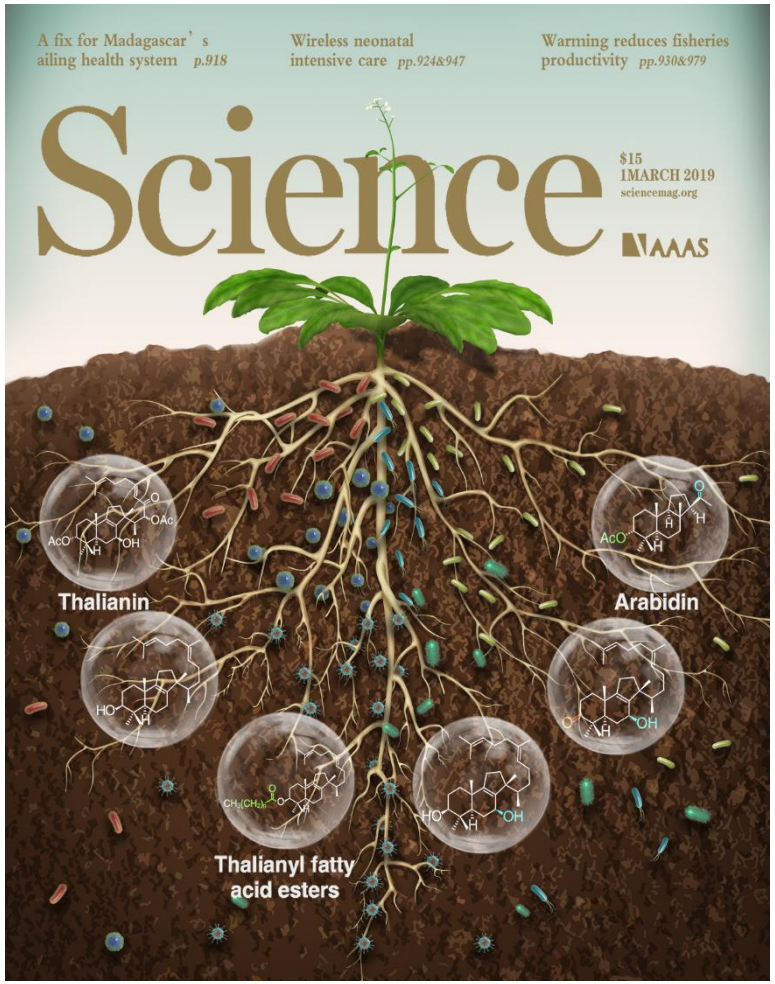
[Cell：蜱虫基因组和宏基因组数据]

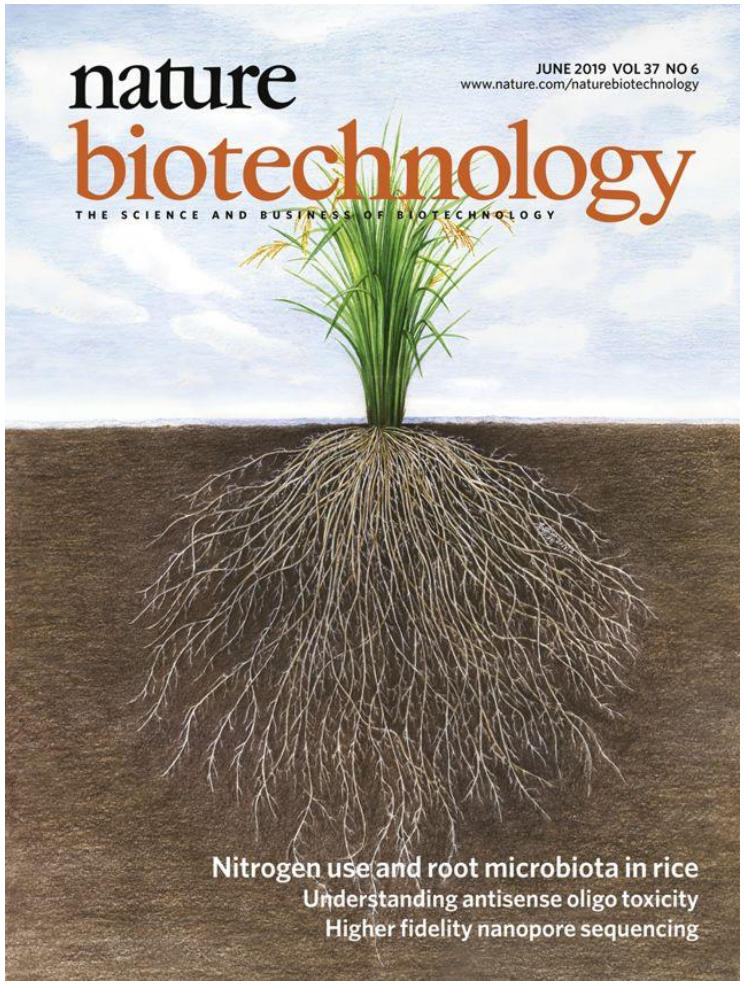


[Cell：人体肠道菌群的长期遗传稳定性和个体特异性]

CEPAMS

封面可视化



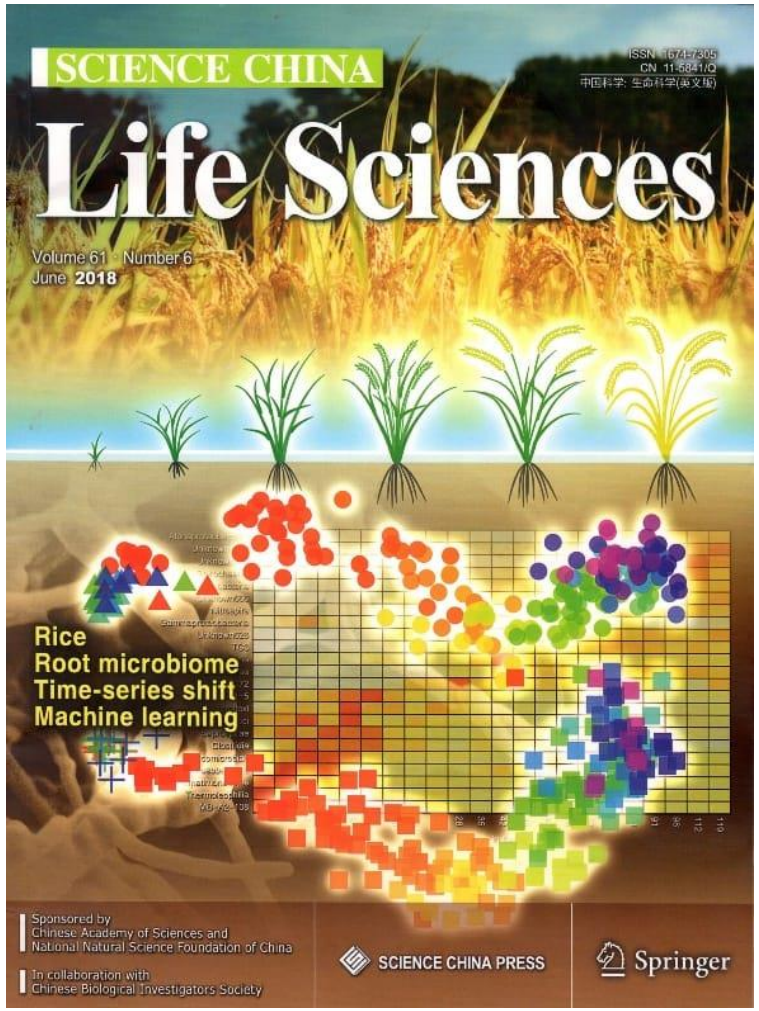


Science：拟南芥三萜化合物特异调控根系微生物组　NBT：水稻NRT1.1B基因调控根系微生物组参与氮利用

CEPAMS

封面可视化

Nature子刊：高通量分离培养细菌方法



SCLS：水稻微生物组时间序列分析

杂志	投稿	送审	接收
Nature系列	15	7	3
Science	3	2	1
Cell系列	4	4	1

Manuscript(s)

- Submitted (1):
- Accepted (3):
- Published:
- Declined (11):

咨询是否适合　　补实验/分析再投(可能多次)　　终投

(预投稿) → 投稿 → 拒稿/大修 → 拒稿/小修 → 编辑修改 → 接收排版 → 发表

拒稿/**送审**　　编辑部拒稿/非正式接收　　校稿

Science

拒稿/**送审**　　编辑部拒稿/正式接收确定发表时间　　校稿

投稿 → 同行预审 → 拒稿/大修 → 拒稿/小修 → 接收 → 编辑排版 → 发表

专业编辑和科学家共同初审　　补实验/分析再投(可能多次)　　可选提前在线发表

CEPAMS

本领域主要杂志和影响因子

nature biotechnology 36.558

nature | methods 30.822

nature microbiology 15.540

Cell Host & Microbe 15. 925

Gut 19.819

Microbiome 11.607

The ISME Journal Multidisciplinary Journal of Microbial Ecology 9.180

WATER RESEARCH 9. 130

Global Change Biology 8.555

Environmental Science & Technology 7.864

environment INTERNATIONAL 7.577

gut microbes 7.74

mBio 6.784

mSystems 6.633

AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology 4.016

frontiers in Microbiology 4.235

nature REVIEWS MICROBIOLOGY 34.209

AMERICAN SOCIETY FOR MICROBIOLOGY | Clinical Microbiology Reviews 22.556

FEMS MICROBIOLOGY REVIEWS 13.920

Trends in Microbiology 13.546

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology and Molecular Biology Reviews 12.568

ANNUAL REVIEW OF MICROBIOLOGY 11.000

Current Opinion in Microbiology 8.134

Cell Research 20.16

Fungal Diversity 15.386

Molecular Plant 12.084

Protein & Cell 10.164

Science Bulletin 9.511

GPB 7.051

(GIGA)n SCIENCE 5.993

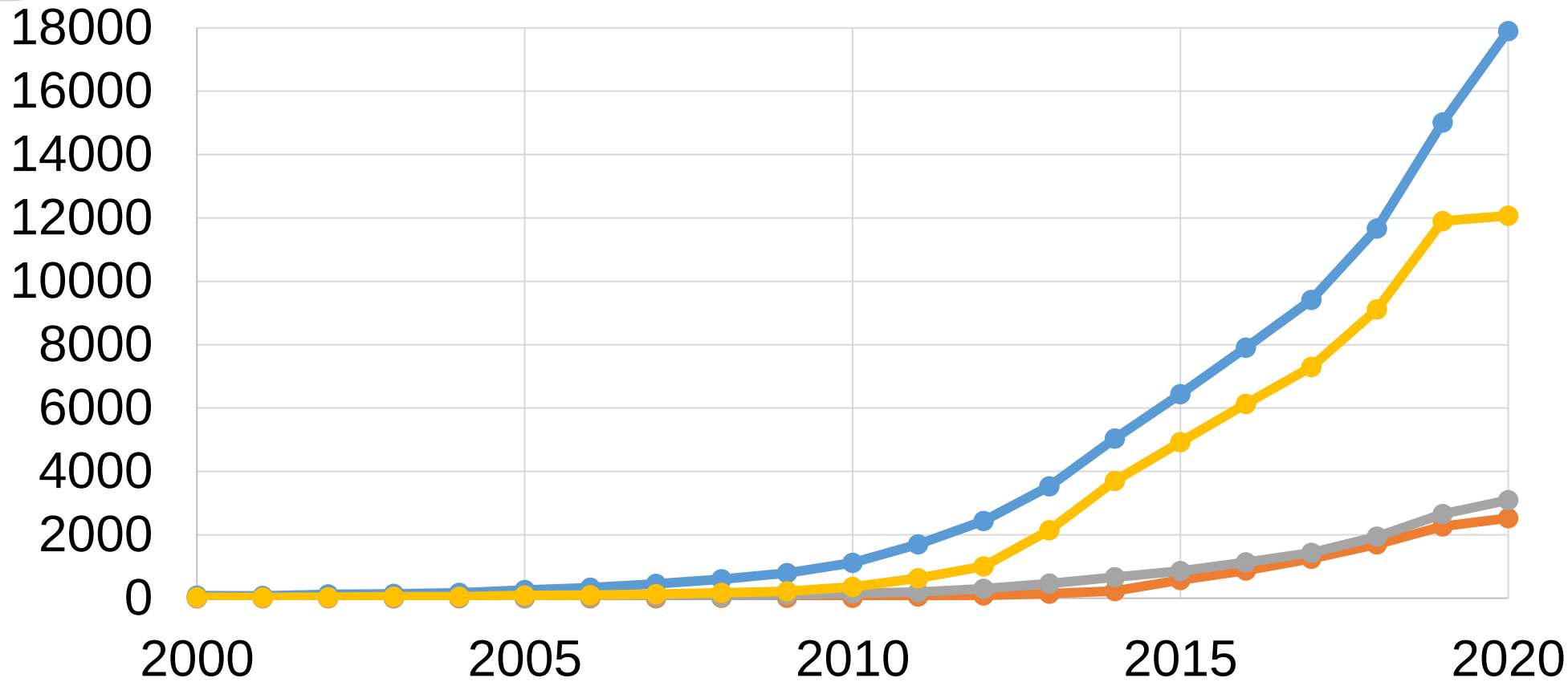
JGG Journal of Genetics and Genomics 5.065

SCIENCE CHINA Life Sciences 4.611

CEPAMS

- 宏基因组领域杂志简介及最新影响因子

微生物组领域的发展—环境开始内卷

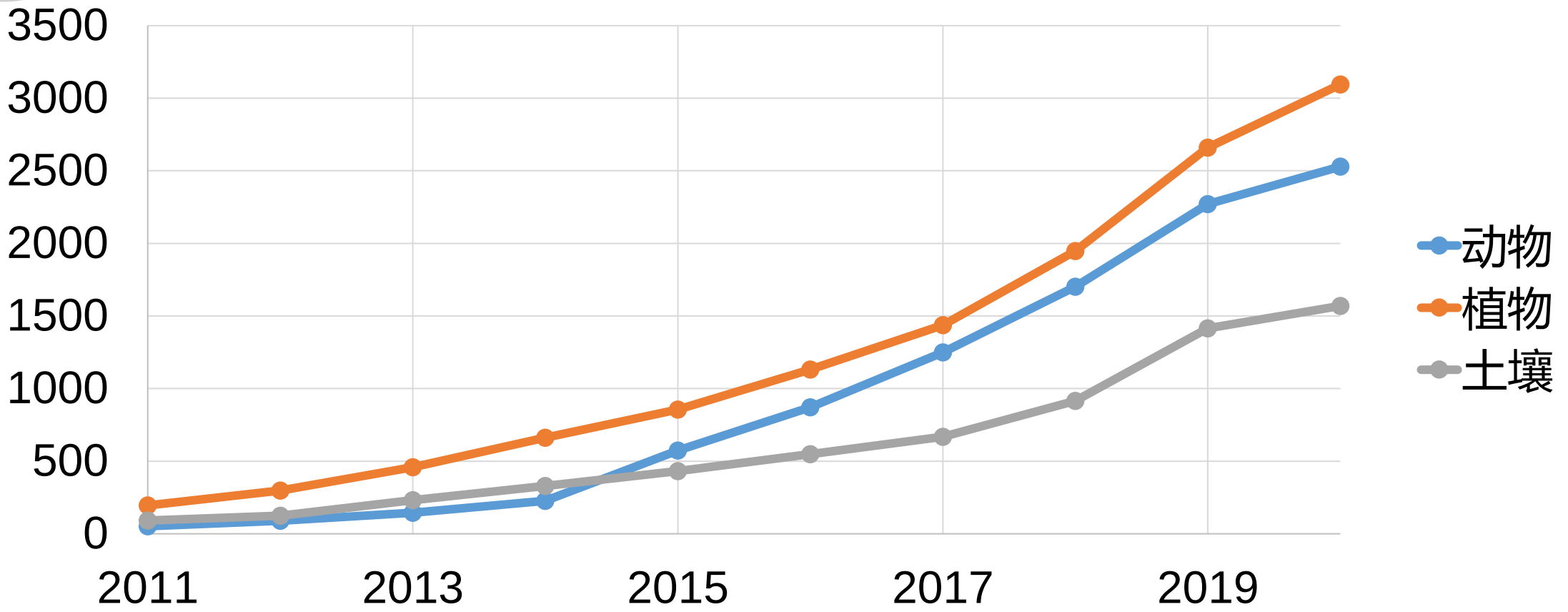


https://pubmed.ncbi.nlm.nih.gov/

Microbiome + human/animal/plant/environment

CEPAMS

经历了10年快速发展——2021全面内卷



https://pubmed.ncbi.nlm.nih.gov/

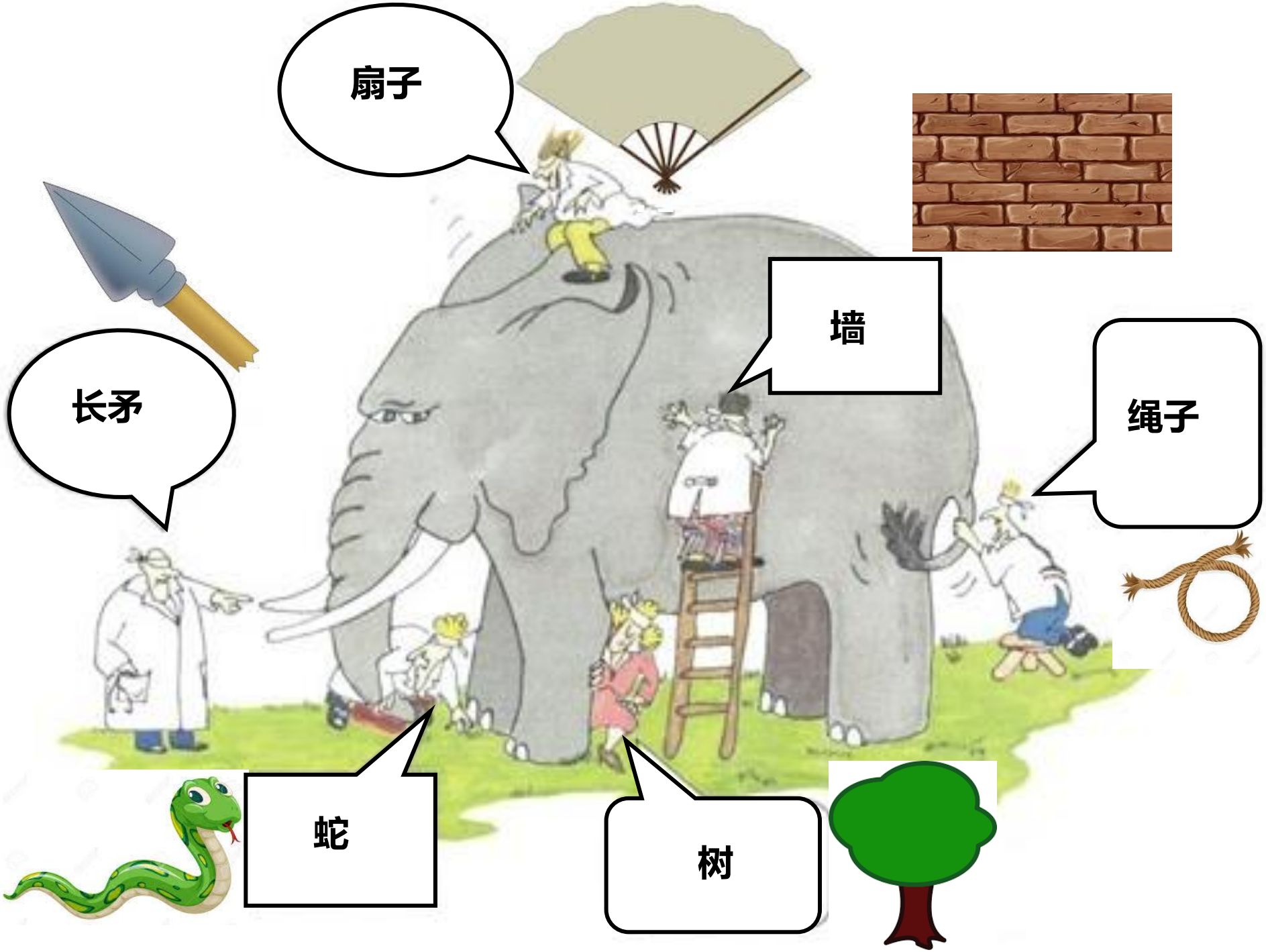
Microbiome + animal/plant/soil

CEPAMS

合作：实验+分析=1+1>2

	实验生物学家	计算生物学家	合作
丰富的实验材料	✓		✓
精湛的实验操作	✓		✓
强大的背景知识	✓		✓
数据分析的思路		✓	✓
方法的测试与比较		✓	✓
分析结果描述		✓	✓
生物学意义解读	✓	✓	✓+✓

实验生物学家与计算生物学家的知识体系优势互补


扇子
墙
绳子
长矛
蛇
树

总结

- 微生物组为什么这么热？脱离了微生物的研究是不完整的
- 数据分析基本思想：降维——降维——可视化
- 扩增子分析：主流有QIIME、mothur、USEARCH，可选我整合的EasyAmplicon全分析流程或最新的QIIME 2(有中文教程)
- 宏基因组分析：安装用Conda和Bioconda通道，数据库有微生物所备份、流程参考EasyMetagenome
- 可视化方案：常用样式有Alpha多样性、Beta多样性、物种组成、差异比较、相关分析、网络分析、机器学习、进化分析
- 可视化R包：有microeco和animalcules等，以及我维护的amplicon
- 高分文章特点：意义重大，读者广泛，逻辑性强，多重证据组图，数据共享，可重复分析，方法摘要可视化，多人合作

CEPAMS

宏基因组(公众号、个人微信)

扫码关注宏基因组，获取专业学习资料

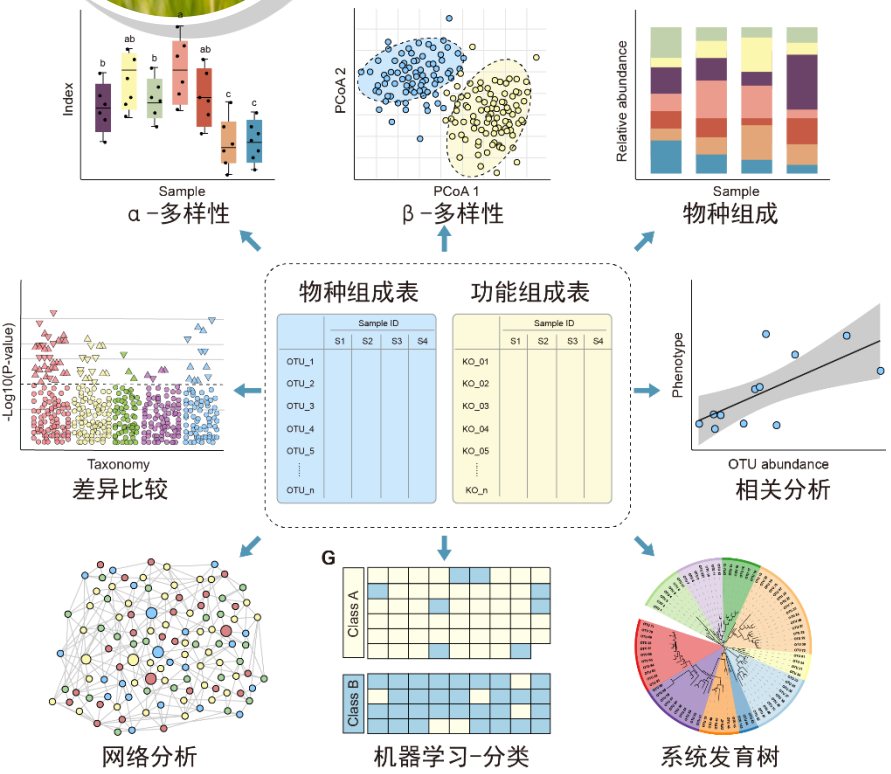
每天坚持学习进步一点点

$(1 + 0.01)^{356} = 37.8$

加我进入同行交流群

备注：姓名-单位-职称/年级-研究方向

《微生物组数据分析》—科学出版社



众筹编写《微生物组数据分析与可视化实战》——成为宏基因组学百科全书的创始人(目录)

编者序：初衷、计划、要求、优势、目标和展望

内容格式要求：背景知识、实例解读、实战代码；帮助同行看得懂文章，会归纳结果和写文章，实现自主分析；

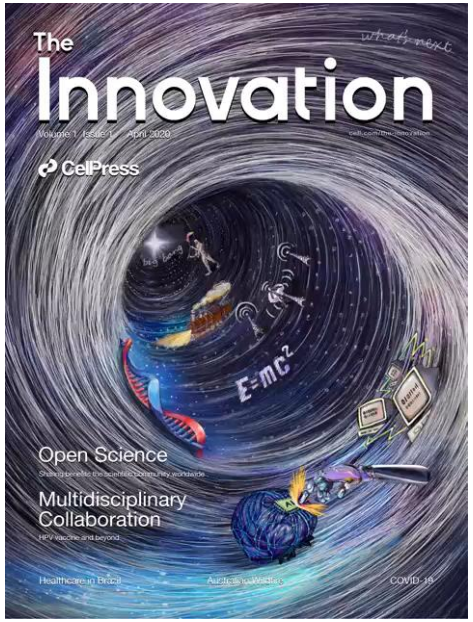
计划：1年发行中文专著和建立在线百科全书平台，优秀稿件推荐发表中、英文方法期刊；2年建立我国自主、有影响力的微生物组分析平台；3年走出去，发表国际有影响力的方法文章和出版外文图书

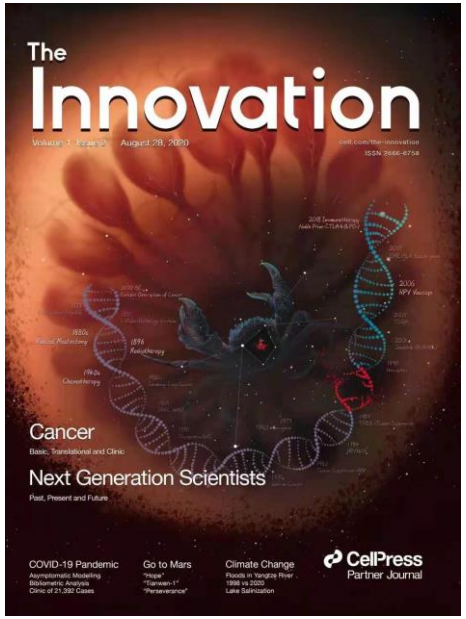
编号	姓名	单位	研究方向	职称/年	创作/参与章节
1	刘永鑫	中科院遗传发育所	微生物组数据分析	工程师	扩增子/宏基因组分析流程,Bet
2	文涛	南京农业大学	微生物组数据分析	博4	可视化章节正文，R包amplicon
3	钱旭波	浙江大学	儿童风湿与肠道菌	主任医师	基本概念、统计、英文版发行
4	吴一磊	中科院微生物所	环境微生物16s 数	科研助理	ggplot2可视化？R统计分析基础
5	徐俊	北京大学	肠道微生物与肠道	助理研究	真菌组18S/ITS
6	蔡伟伟	北京交通大学	环境微生物	讲师	微生物网络分析
7	查富蓉	华中农业大学/凌恩	微生物组数据分析	硕士	宏基因组抗性基因/整合子分析；
8	常帆	陕西省微生物研究所	生物废弃物综合利用	助理研究	数据库和数据上传
9	陈亮	中科院微生物研究所	三代测序与肠道微生	助理研究	network分析与真菌组
10	邓子祺	西班牙植物生物技术	系统发育工具开发；	博1	利用ETE构建、绘制和分析系统发
11	付先恒	中国科学院大学	微生物生态与土壤养	博3	ggplot2可视化及ASV表后续统计分
12	葛德燕	中国科学院动物研究	动物进化与系统学	副研究员	野生动物系统发育重建与胃肠道宏
13	何茂章	安徽医科大学	肠道宏基因组与代谢	讲师	普氏分析，相关分析的可视化，随
14	李鸿毅	浙江大学/某测序公	土壤微生物生态/扩	博士	lefse结合原代码进行各参数解读
15	李辉	中山大学中山医学院	医学微生态学	博2	mothur; Community type
16	李杰	常熟理工	三代测序和冠状病毒	讲师	三代测序Nanopore分析冠状病毒
17	李金优	浙江大学	肠道衰老与肠道微生	实验员	图片排版和美化，R统计和绘图
18	李瑞琳	中国科学院计算机所	高性能计算与生物信	助理研究	宏基因组数据挖掘算法与分析流程
19	李苏梅	深圳市人民医院	临床药理病理	副研究员	功能注释数据库
20	李文譞	新加坡国立大学	环境微生物抗药性	博后	文章套路总结-抗生素抗性
21	李延明	堪萨斯大学医学中心	机器学习，大数据分	助理教授	机器学习的常用算法
22	李雨泽	西北农林科技大学	农田微生物生态	硕2	co-occurrence网络分析
23	林禾雨	墨尔本大学	环境微生物多组学	博3	分箱专题
24	刘华	中国科学技术大学	生命医学/微生物学	博3	非限制性排序，ggplot2扩展包
25	刘云	吉林大学	宏基因组binning算	讲师	分箱专题
26	吕波	湖南师范大学	胁迫对动物肠道菌群	硕1	R可视化、宏转录组
27	毛杰	南京农业大学	WES数据分析	硕士	R统计与绘图，ggplot2可视化
28	孙江伟	瑞典卡罗林斯卡医	宏基因组学分析	博3	统计基础
29	谭乔文	同济大学	饮用水微生物组	硕3	分类树构建/机器学习的常用算法/
30	王敬敬	中国科学院天津工业	微生物肥料，根际微	副研究员	Gephi绘制网络图，CANNOCO进行R

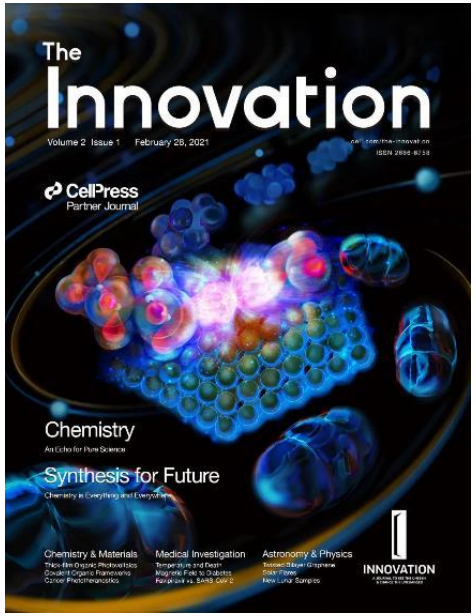
全部文章点击链接《微生物组数据分析》，**持续更新，欢迎参与**

The Innovation: A Journal to See the Unseen and Change the Unchanged

CellPress







The Innovation 是一本由中科院青年科学家与Cell Press出版社于2020年共同创办的**综合性英文学术期刊**，向科学界展示鼓舞人心的跨学科发现，鼓励研究人员专注于科学的本质和自由探索的初心，领域覆盖**化学、材料科学、纳米技术、医学、物理学、生物学、地球科学和工程学等全部自然科学**。The Innovation已被DOAJ，ADS，Scopus等数据库收录。2020年，从投稿到发表的周期是56天。

欢迎赐稿

Website 1: https://www.cell.com/the-innovation/home
Website 2: http://www.the-innovation.org/