

宏基因组  
meta-genome



# 高水平组学文章的分析 和可视化套路

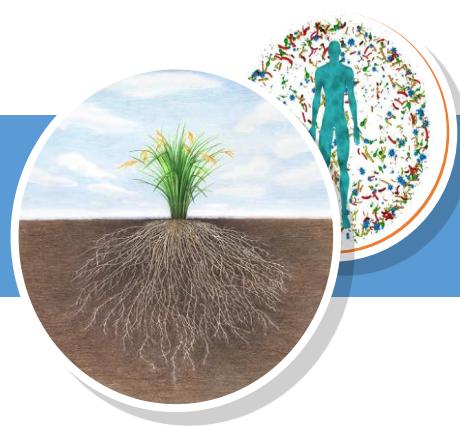
刘永鑫

中国科学院遗传与发育生物学研究所  
2019年6月30日

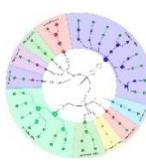
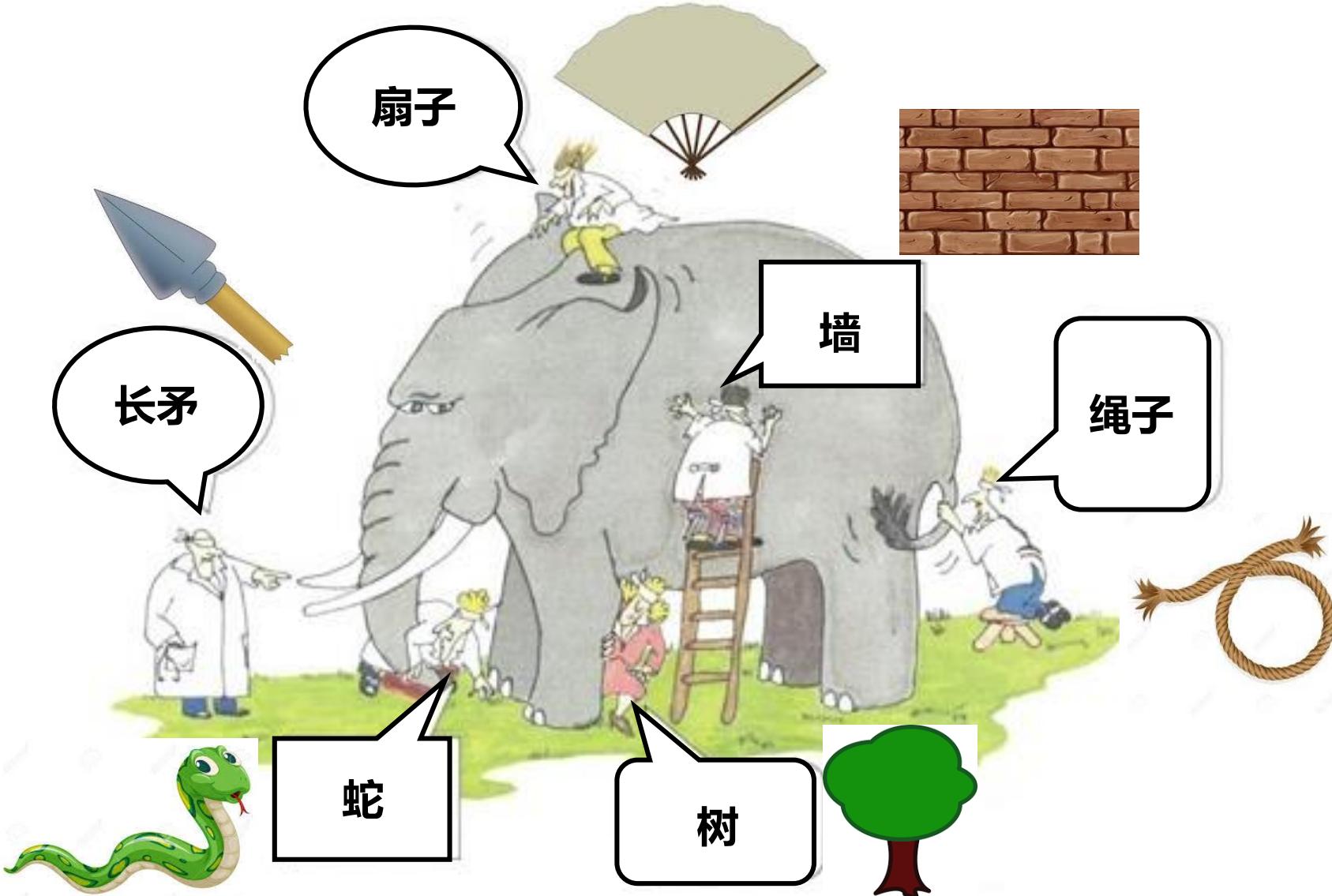


# 什么是组学？

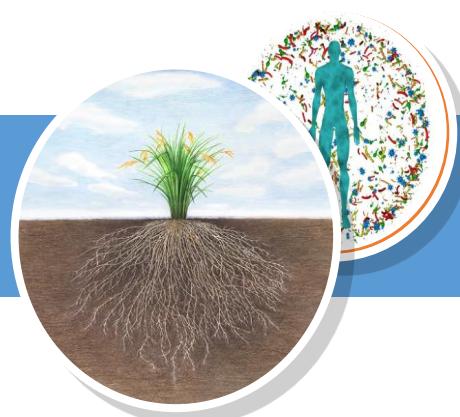
- 组学(Omics)主要包括**基因组学（Genomics）**，**转录组学（transcriptomics）**，**蛋白组学（Proteinomics）**，**代谢组学（Metabolomics）**，脂类组学(lipidomics)，免疫组学(Immunomics)，糖组学(glycomics)，RNA组学(RNomics)学，影像组学(Radiomics)，超声组学(Ultrasomics)等。
- Omics是组学的英文称谓，它的词根'-ome'英译是一些种类个体的系统集合，例如**Genome（基因组）是构成生物体所有基因的组合**，基因组学( Genomics ) 这门学科就是研究这些基因以及这些基因间的关系。



# 组学的优势——更全面地看问题



宏基  
因组

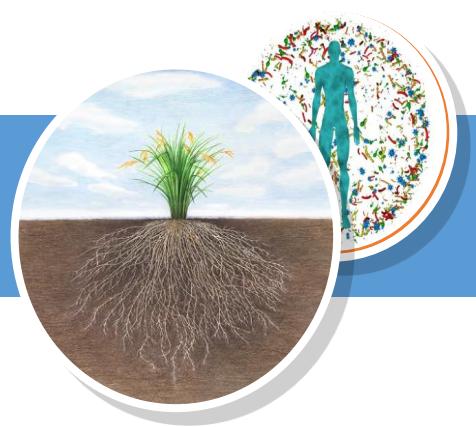


# 我的组学研究历程

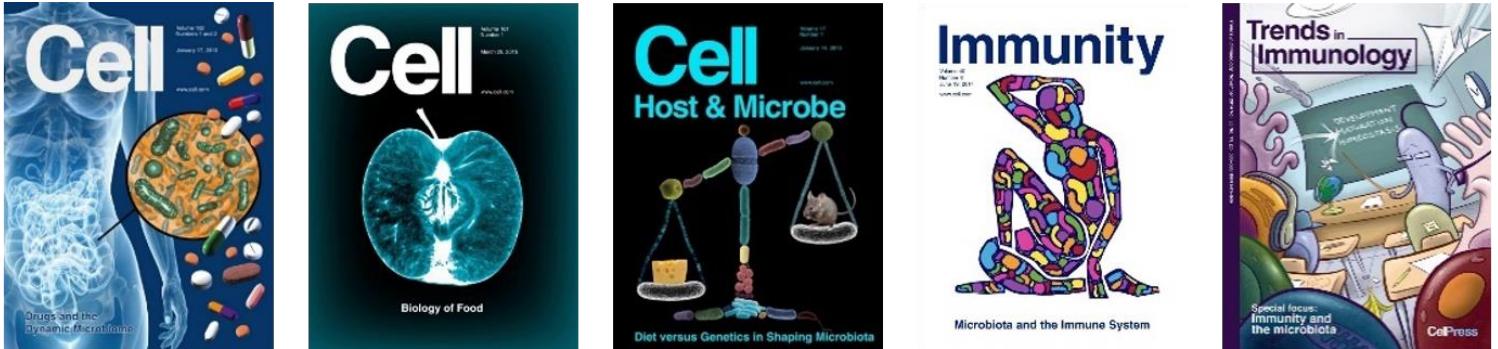


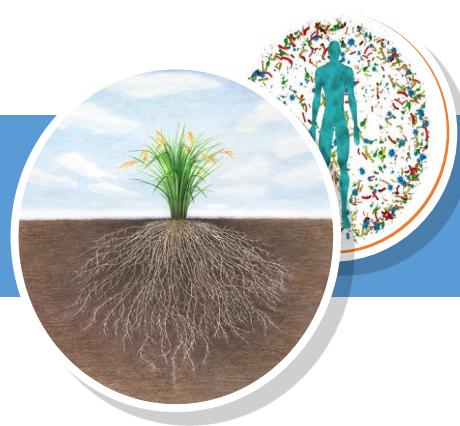
2004	东北农业大学	学士	微生物学	食用真菌
2008	东北农业大学	硕士	作物遗传育种	miRNA组
2011	中科院遗传发育所	博士	生物信息	小RNA组
2014	中科院遗传发育所	博士后	细胞生物学	转录组、表观组
2016	中科院遗传发育所	工程师	微生物组	扩增子、宏基因组



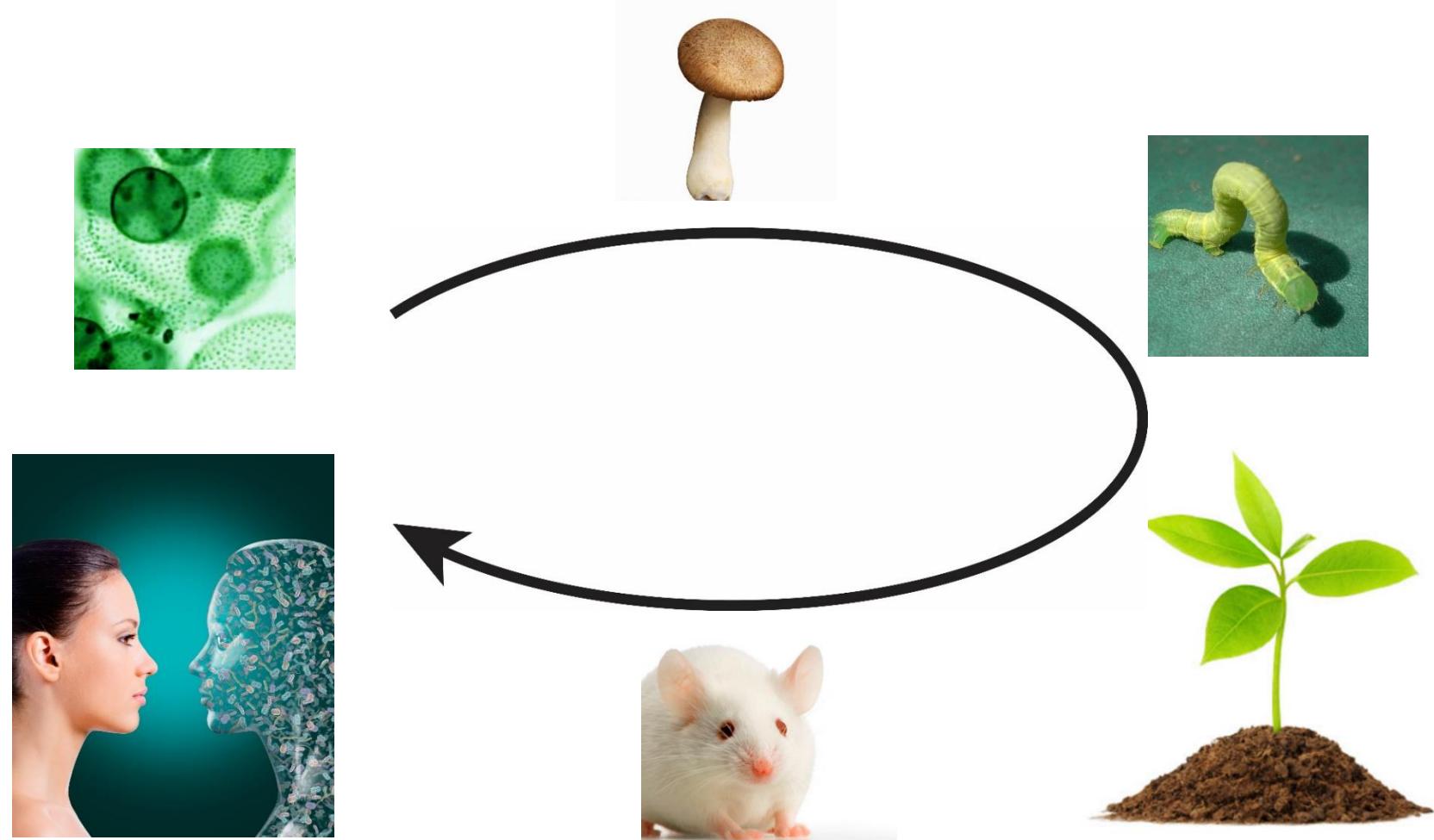


# 微生物组领域为什么这么热？





# 动植物生活环境充满微生物



脱离了微生物的生物学研究是不完整的



# 近年我参与发表的微生物组文章



nature  
biotechnology

ARTICLES

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

RESEARCH ARTICLE

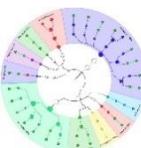
**NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice.**

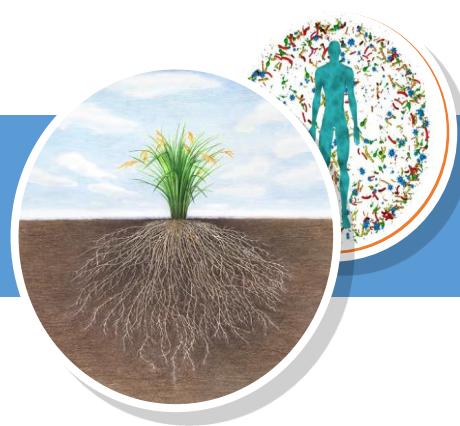
PLANT SCIENCE

Jingying Zhang<sup>1,2,10</sup>, Yong-Xin Liu<sup>1,2,10</sup>, Na Zhang<sup>1,2,3</sup>, Yuan Qin<sup>1,2,3</sup>, Pengxu Yan<sup>4,5,6</sup>, Xiaoning Zhang<sup>1,2,3</sup>, Xin Wang<sup>1,2</sup>, Chao Wang<sup>1,2</sup>, Hui Wang<sup>1,2,3</sup>, Bac Ruben Garrido-Otero<sup>8,9</sup>, Chengcai Chu<sup>1,3\*</sup> and Ancheng C. Huang<sup>1\*</sup>, Ting Jiang<sup>2,3,4\*</sup>, Yong-Xin Liu<sup>2,3</sup>, Yue-Chen Bai<sup>5,6</sup>, James Reed<sup>1</sup>, Baoyuan Qu<sup>2,3</sup>, Alain Goossens<sup>5,6</sup>, Hans-Wilhelm Nützmann<sup>1†</sup>, Yang Bai<sup>2,3,4‡</sup>, Anne Osbourn<sup>1‡</sup>

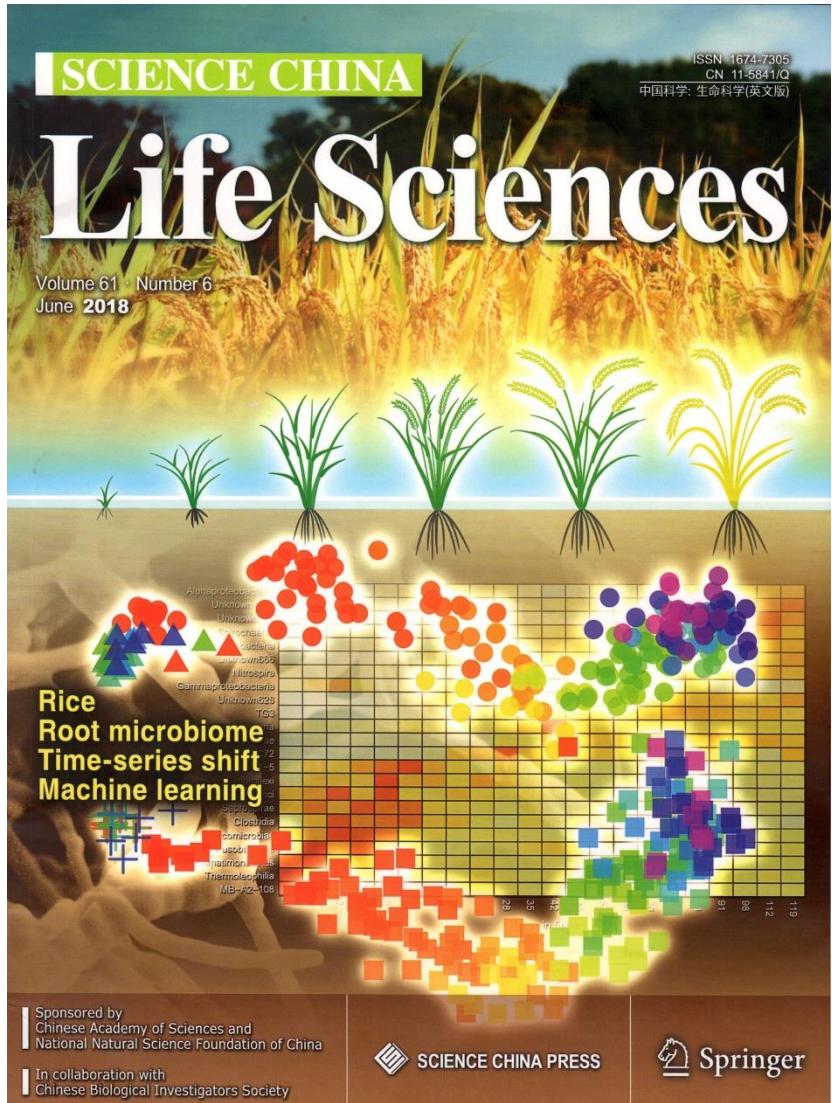
Jingying Zhang, Yong-Xin Liu, Na Zhang, et. al. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. *Nature Biotechnology*. 2019. doi:10.1038/s41587-019-0104-4

Ancheng C. Huang, Ting Jiang, Yong-Xin Liu, et. al. A specialized metabolic network selectively modulates Arabidopsis root microbiota. *Science*. 2019, 364: eaau6389. doi:10.1126/science.aau6389



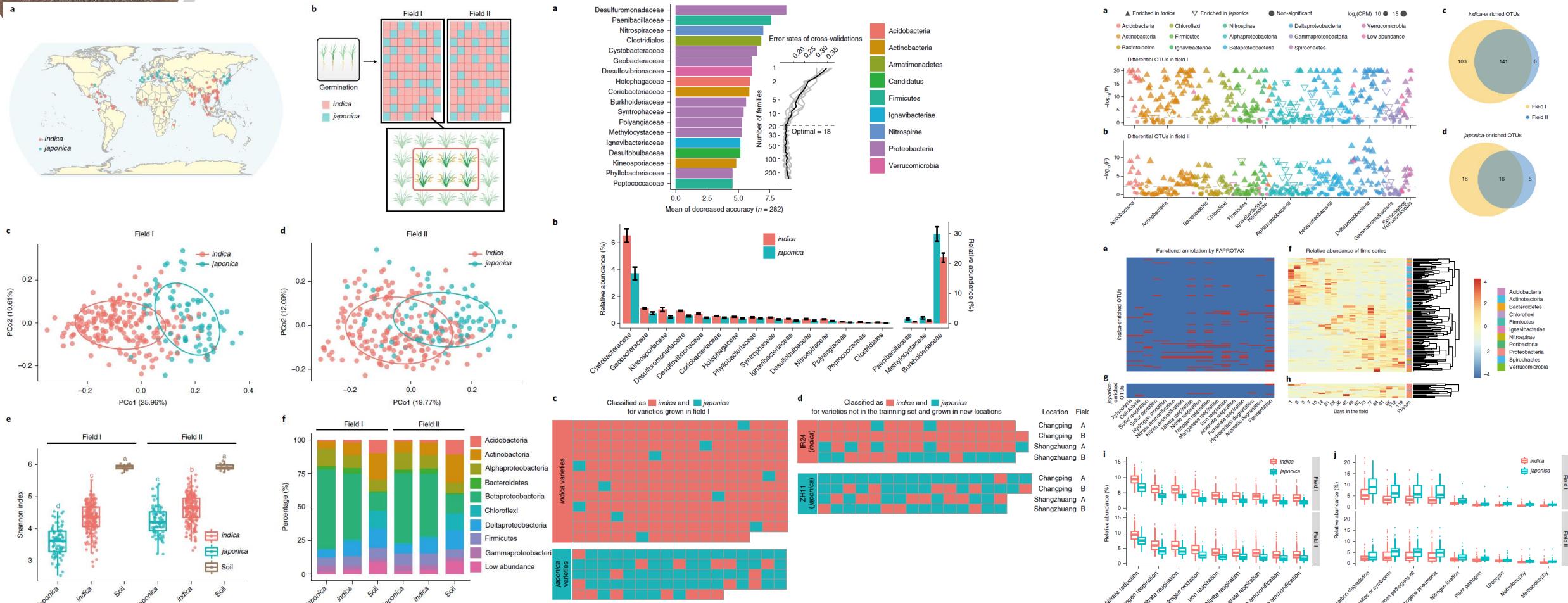


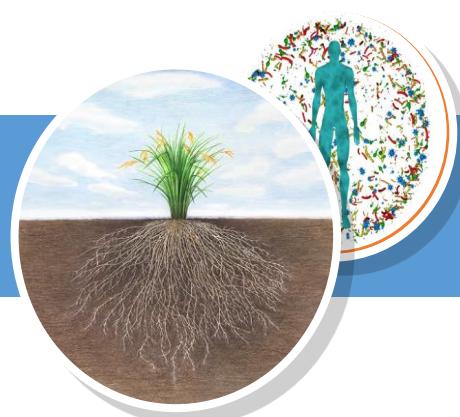
# 其中一部分也被入为封面文章





# 组学文章可视化结果常用图表

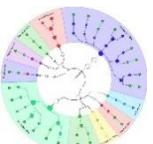




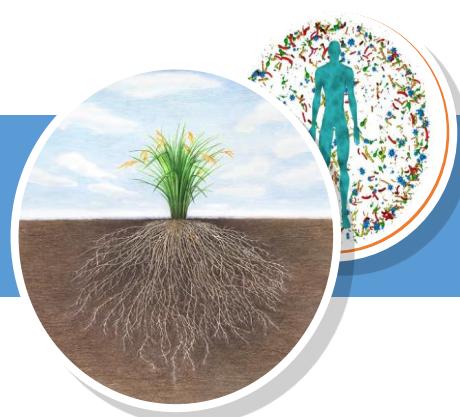
# 目录



- 整体概述
  - 材料地图和/或实验设计
  - 样本概述:  $\alpha$ 多样性箱线图、 $\beta$ 多样性PCoA散点图、物种组成柱状图
- 细节展示
  - 物种组间差异: 曼哈顿图、韦恩图
  - 功能注释或差异: 热图有/无或时间序列、箱线图
- 应用场景——机器学习挖掘生物标记
  - 随机森林分类——区分不同组, 如疾病诊断、来源或品种鉴定
  - 随机森林回归——时间序列, 如年龄、死亡时间预测



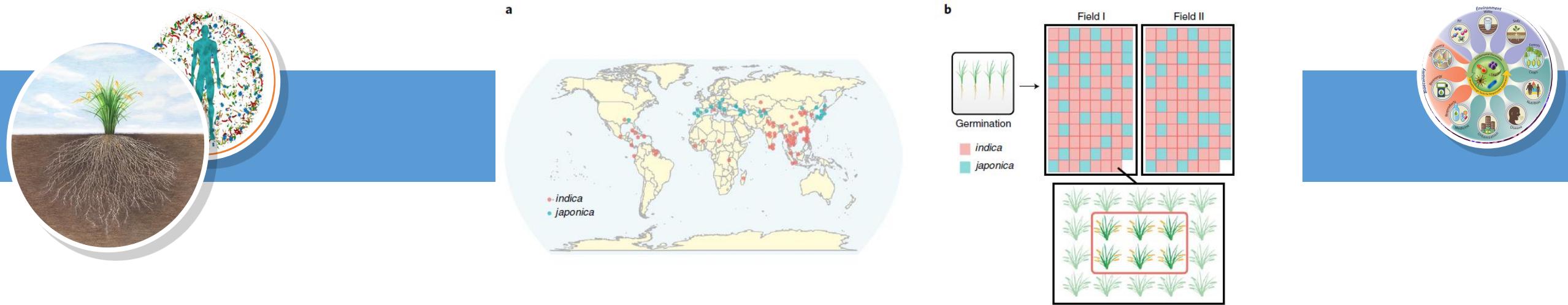
宏基  
因组



# 目录



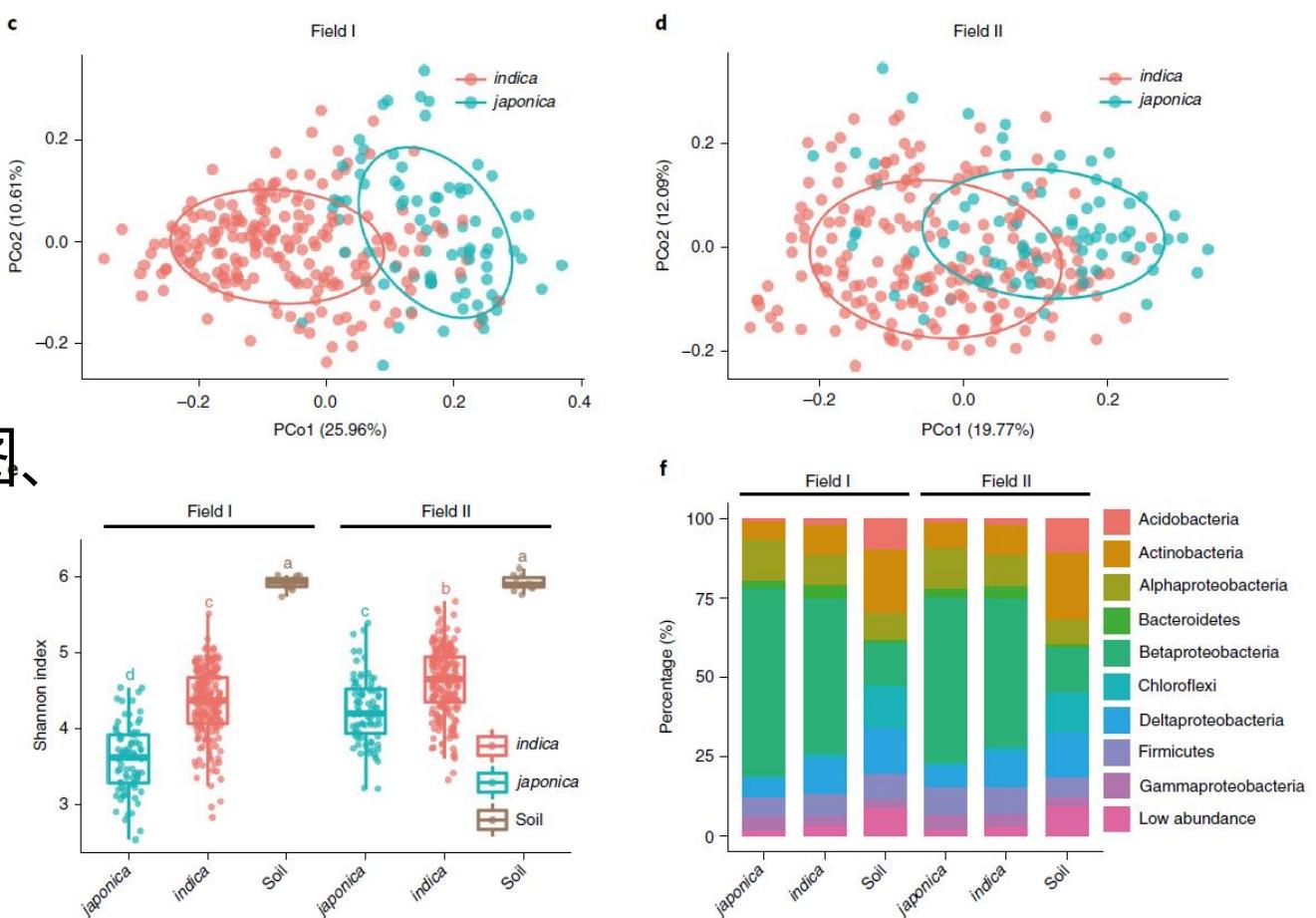
- 整体概述
  - 材料地图和/或实验设计
  - 样本概述:  $\alpha$ 多样性箱线图、 $\beta$ 多样性PCoA散点图、物种组成柱状图
- 细节展示
  - 物种组间差异: 曼哈顿图、韦恩图
  - 功能注释或差异: 热图有/无或时间序列、箱线图
- 应用场景——机器学习挖掘生物标记
  - 随机森林分类——区分不同组, 如疾病诊断、来源或品种鉴定
  - 随机森林回归——时间序列, 如年龄、死亡时间预测

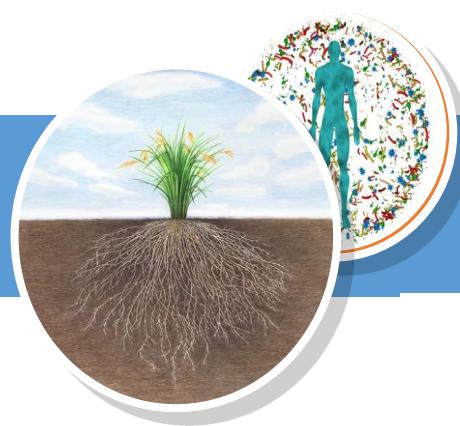


# 整体概述

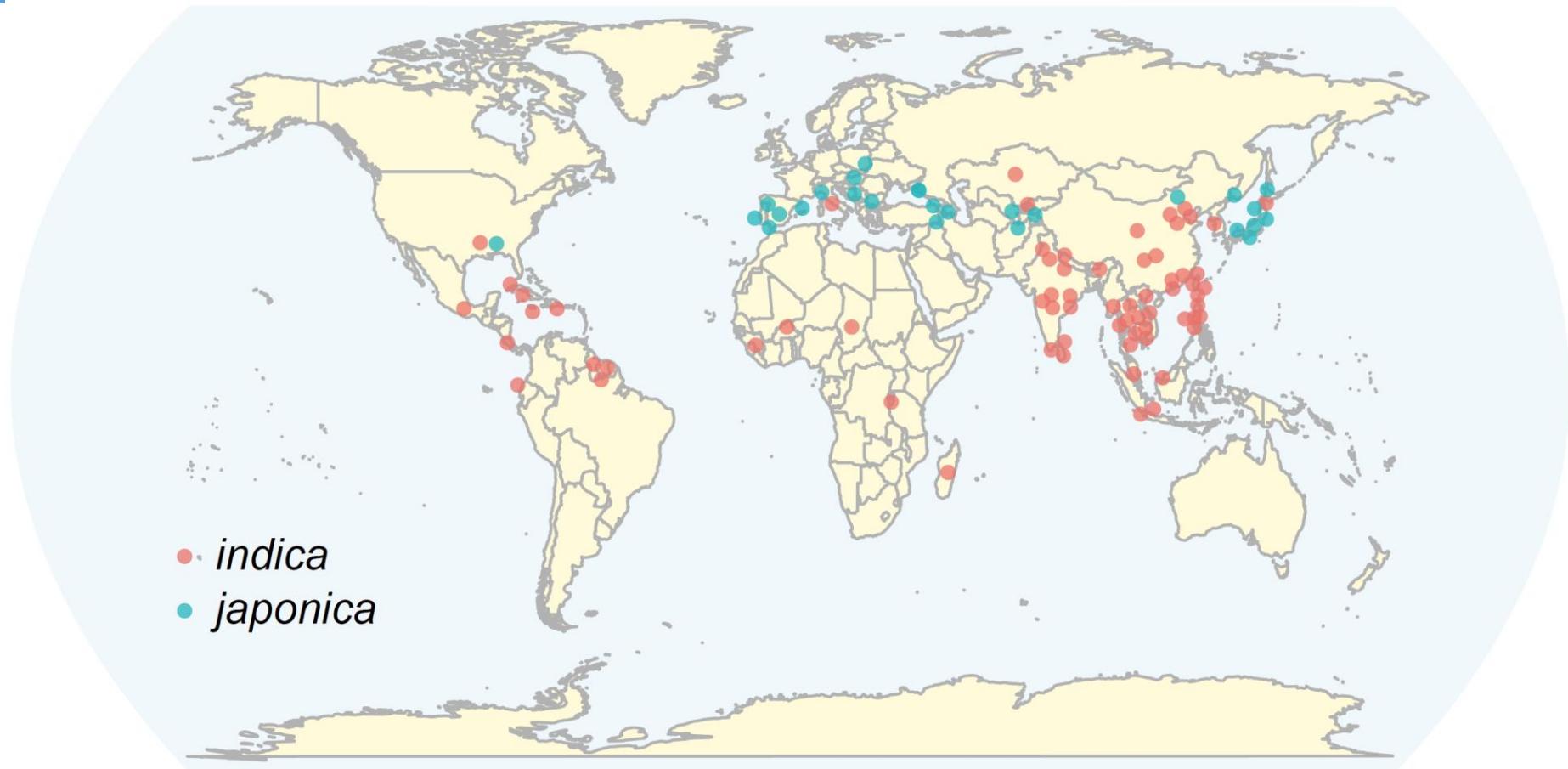
- 材料地图
- 实验设计
- 样本概述

- α多样性箱线图、
- β多样性PCoA散点图、
- 物种组成柱状图



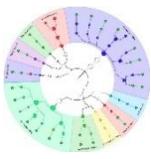


# 材料地图



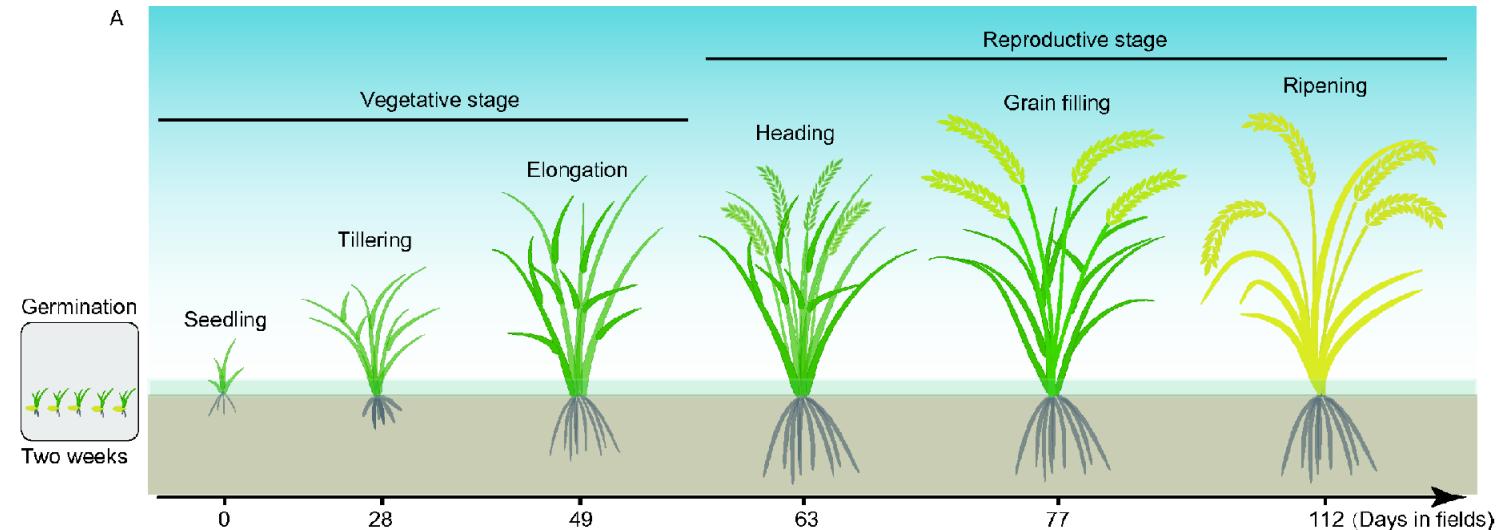
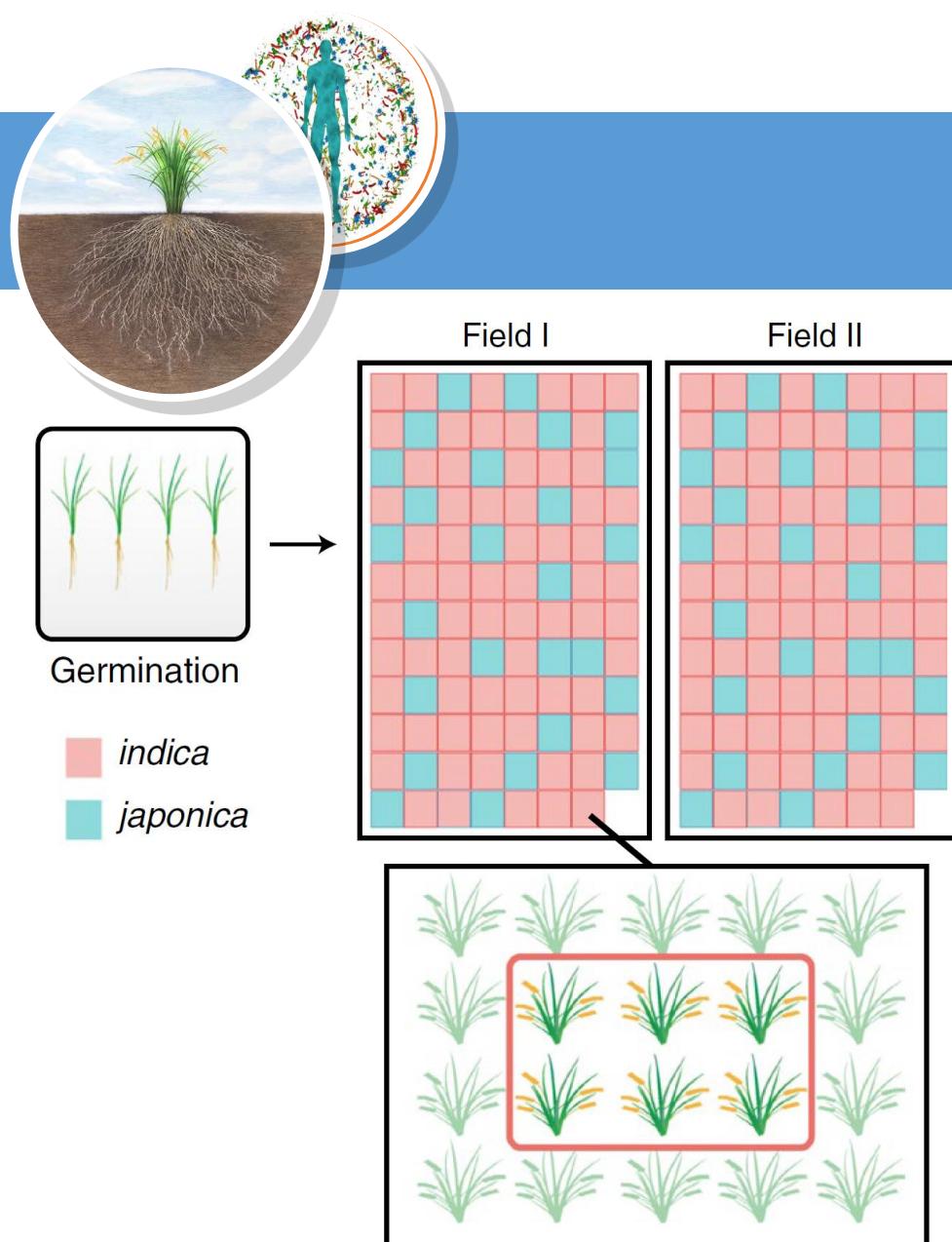
R语言maps结合ggplot2包，基于样本来源经纬度绘制  
展示来源44个国家的95个品种的分布

Jingying Zhang, Yong-Xin Liu, et. al. *Nature Biotechnology*. 2019. Fig 1a





# 实验设计

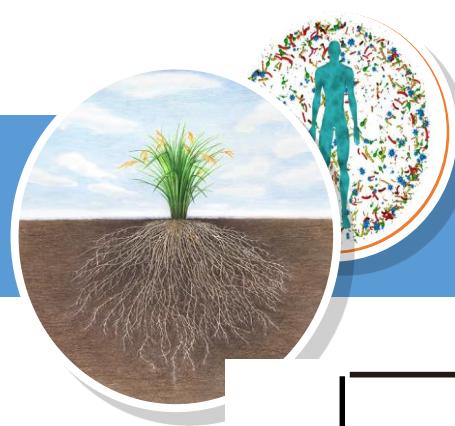


Root microbiota shift in rice correlates with resident time in the field and developmental stage

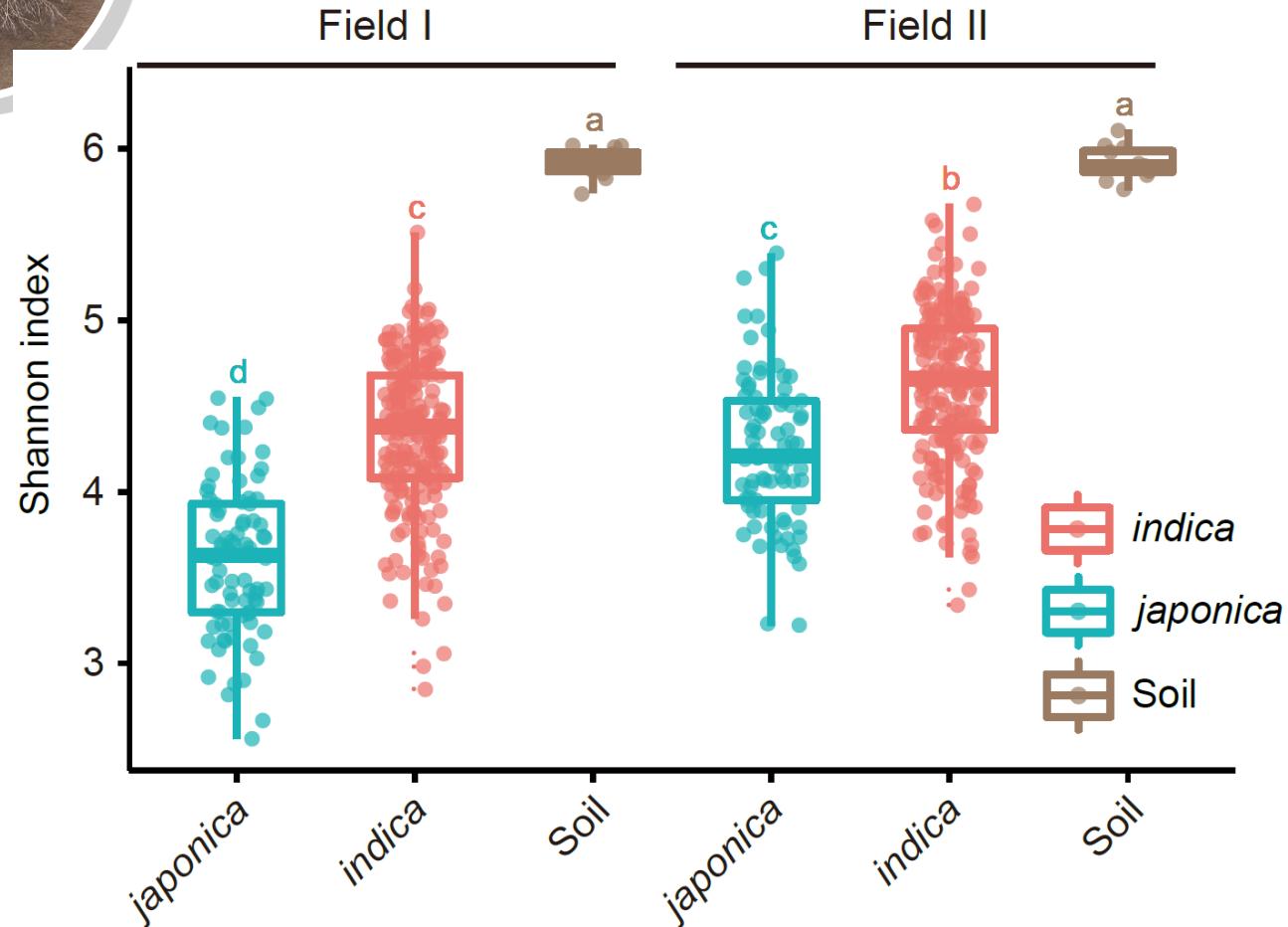
J Zhang, N Zhang, YX Liu, X Zhang, B Hu, Y Qin... - *Science China Life ...*, 2018 - Springer

Land plants in natural soil form intimate relationships with the diverse root bacterial microbiota. A growing body of evidence shows that these microbes are important for plant growth and health. Root microbiota composition has been widely studied in several model plants and crops; however, little is known about how root microbiota vary throughout the plant's life cycle under field conditions. We performed longitudinal dense sampling in field trials to track the time-series shift of the root microbiota from two representative rice cultivars .

☆ 99 Cited by 10 Related articles All 8 versions



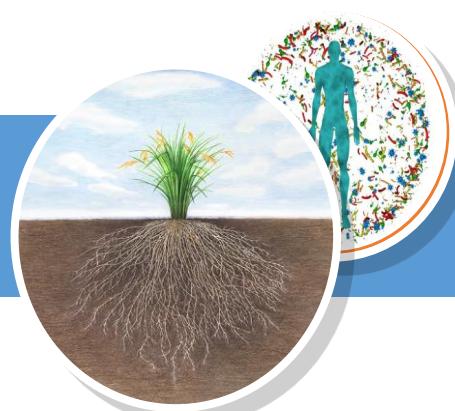
# α多样性箱线图



agricolae或multcompView包  
如LSD.test 统计组间显著性  
ggplot2包

geom\_boxplot 绘制箱线图  
geom\_jitter 添加抖动样本点  
geom\_text 添加显著性分组

结果表明不同地点组间存在物种丰富度和均匀度的一致显著差异



# β多样性PCoA散点图

vegan包

vegdist 计算样本间距离

capscale 排序分析

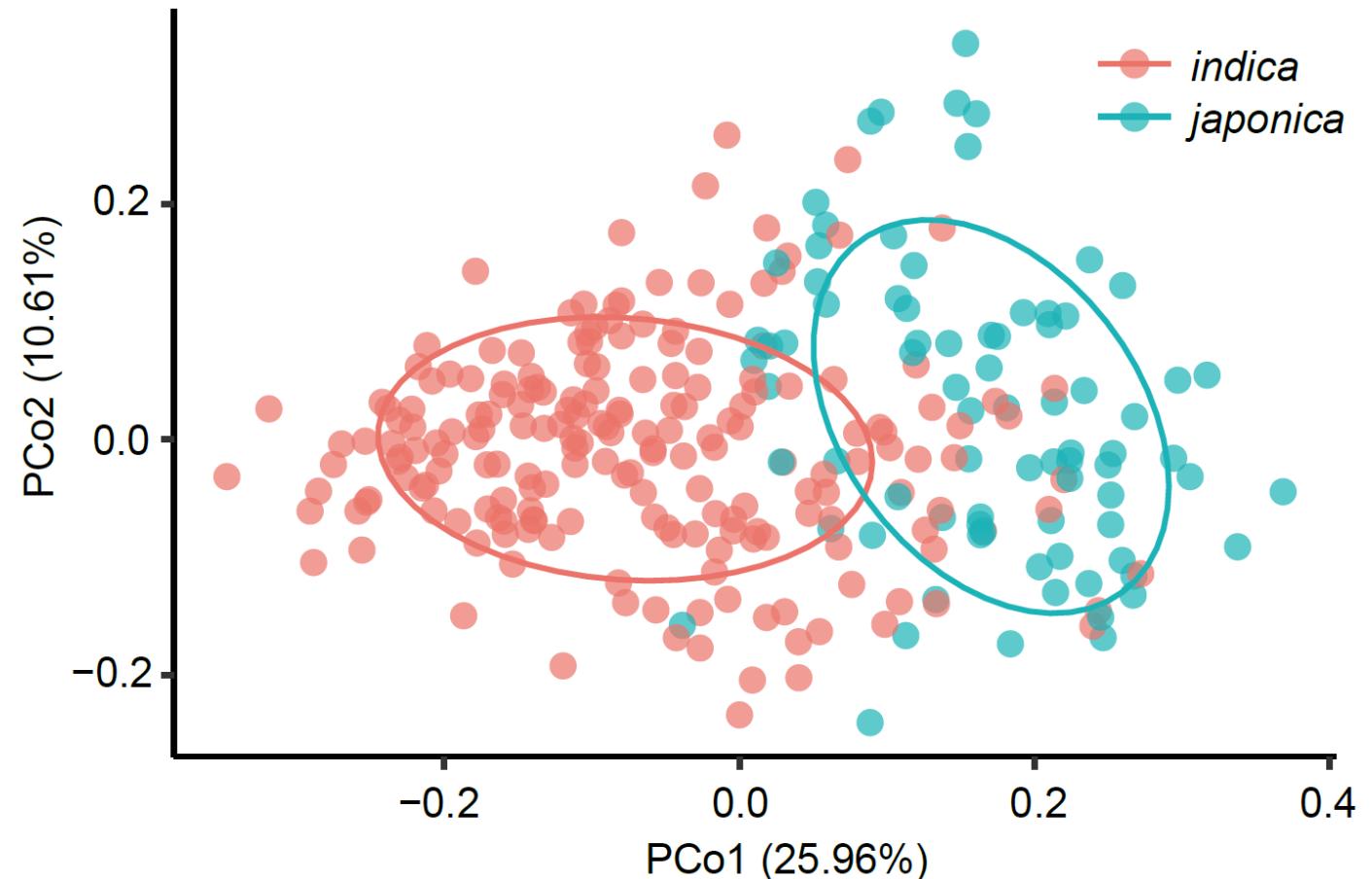
ggplot2包

geom\_point 绘制散点图

stat\_ellipse 添加统计椭圆

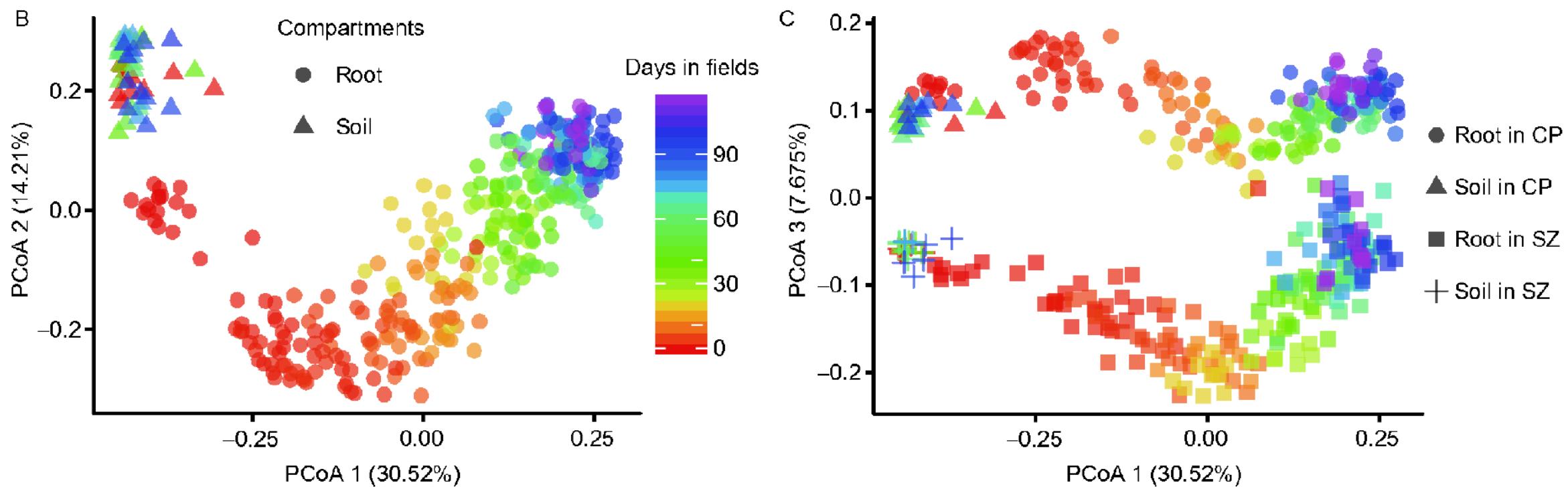
labs 添加差异解释率

结果表明亚种间微生物组群落  
结构在第一主轴明显分开  
第一主轴可解释25.96%的差异

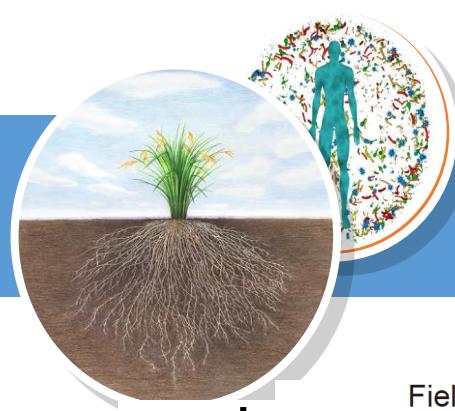




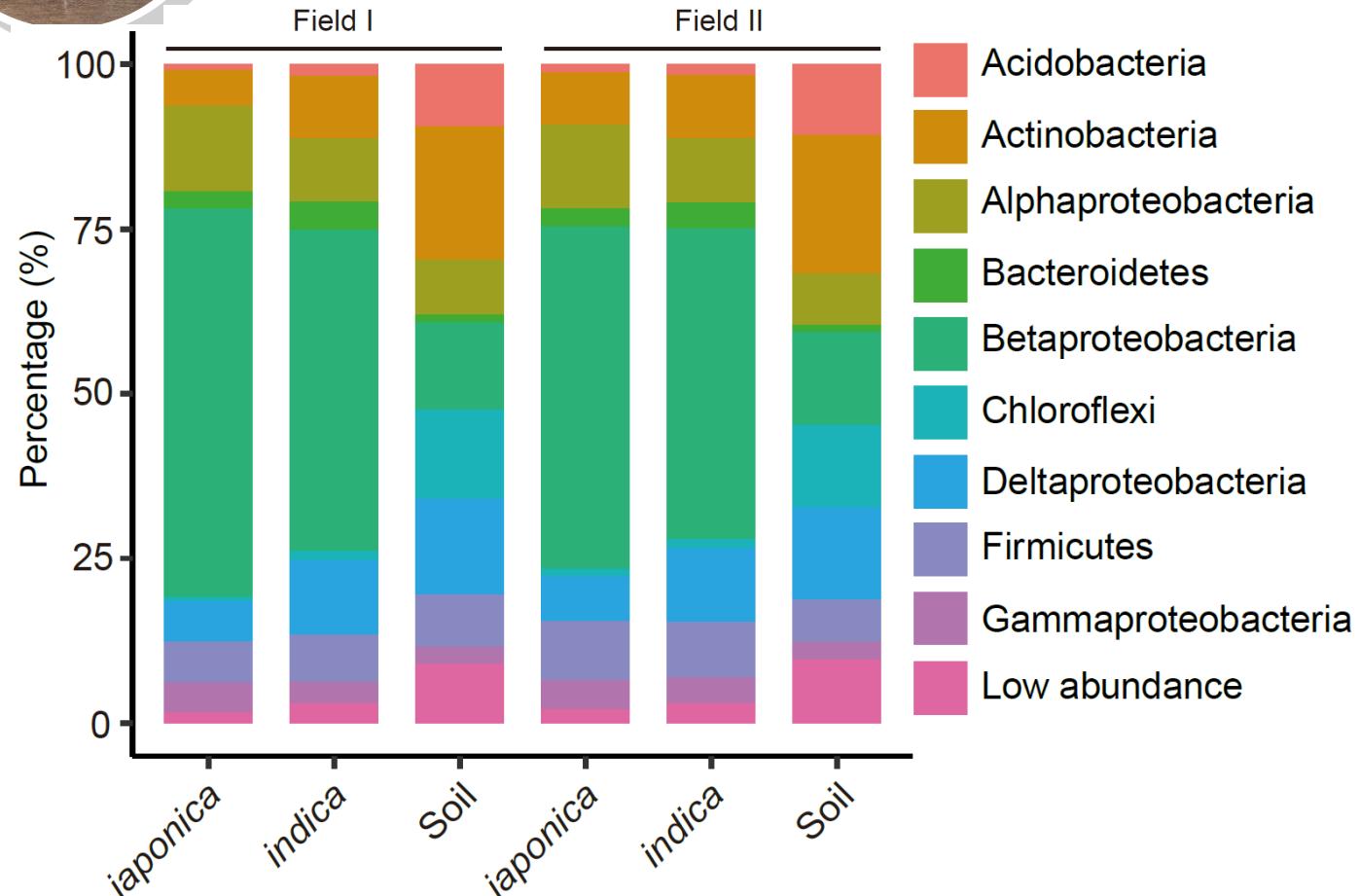
# 散点图-连续型着色展示时间梯度上变化



`p + scale_color_gradientn(colours = rainbow(7))`  
<https://github.com/microbiota/Zhang2018SCLS>



# 物种组成堆叠柱状图



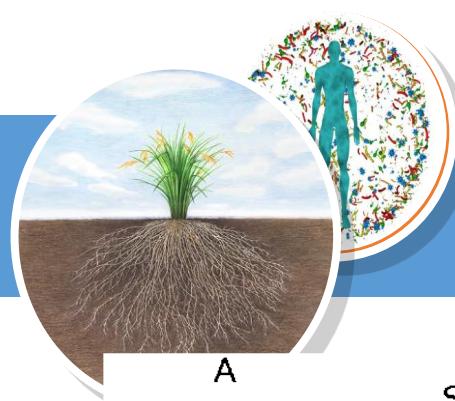
## 数据分析

样本标准化为百分比  
按组求均值

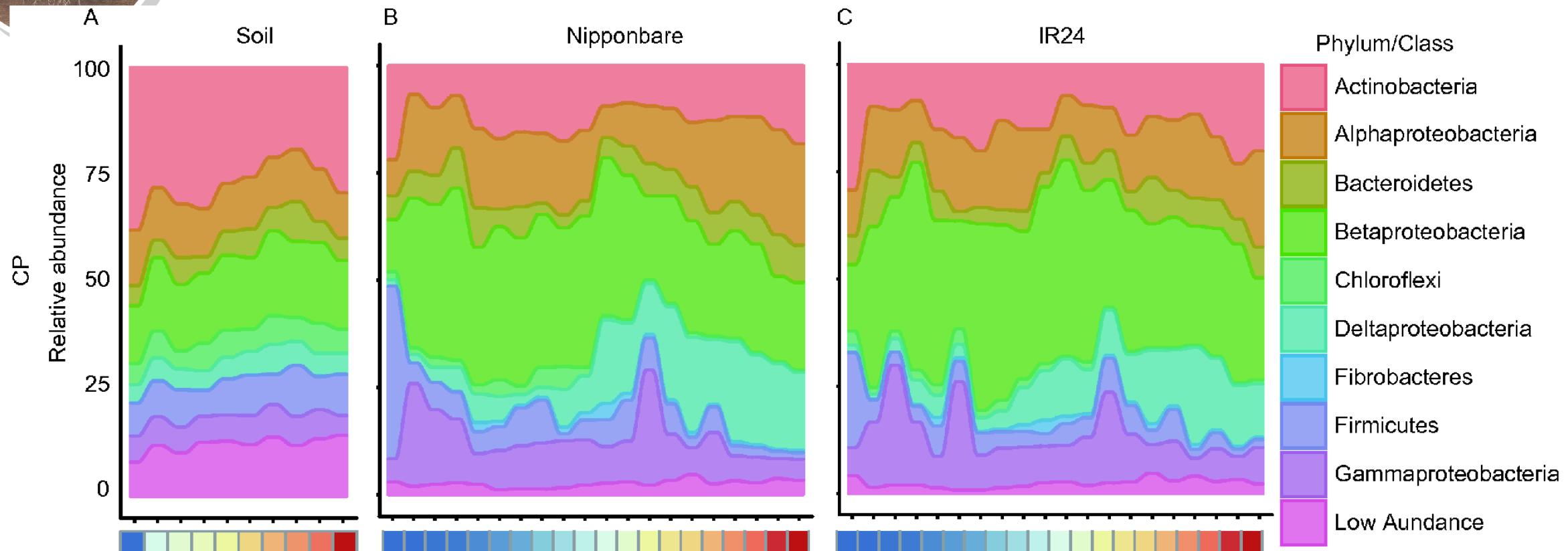
## ggplot2包

`geom_bar(stat = "identity")` 绘制堆叠柱状图

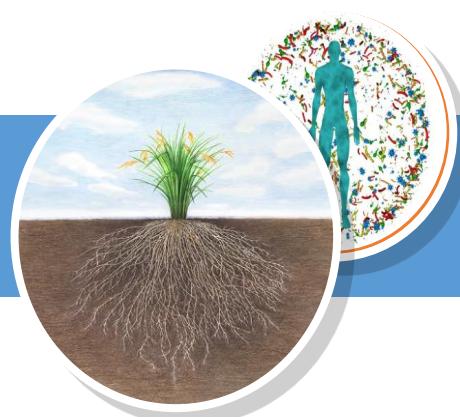
结果表明亚种间存在的物种组成在门水平即存在差异且不同地点可重复



# 冲击图展示物种组成动态变化



```
library(ggalluvial)
p + geom_alluvium(aes(fill = phylumpro, colour = phylumpro), alpha = .75)
```

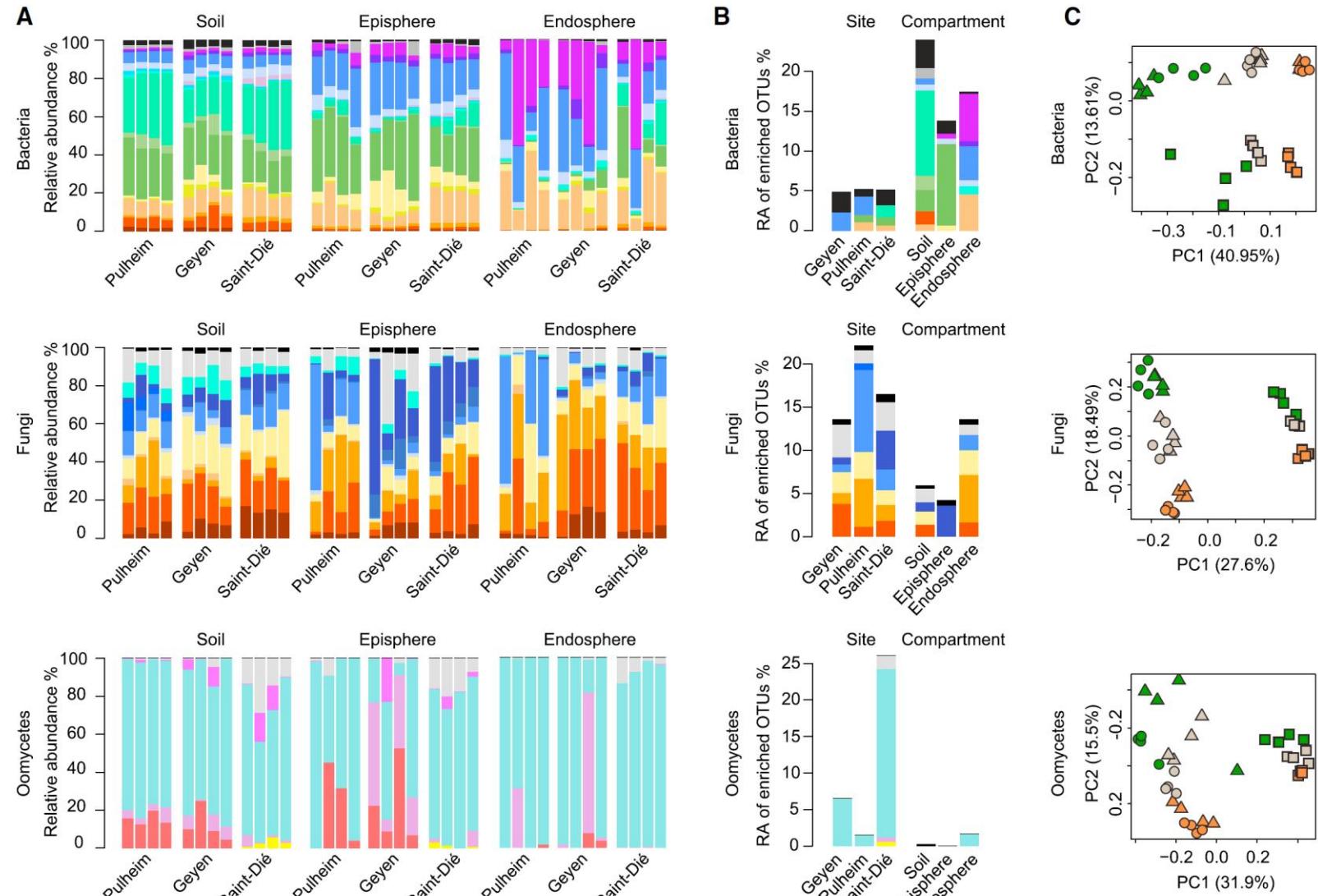
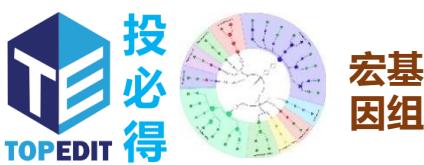


# 关于重复

1. 多组学  
细菌、真菌、卵菌
2. 多部位  
土壤、根表、根内
3. 多地点  
德国Geyen、Pulheim、法国

样本量 = 4个生物学重复  $\times$  3种组学  $\times$  3个部位  $\times$  3个地点 = 108个

Cell : 根部微生物跨界的互作  
促进拟南芥生存





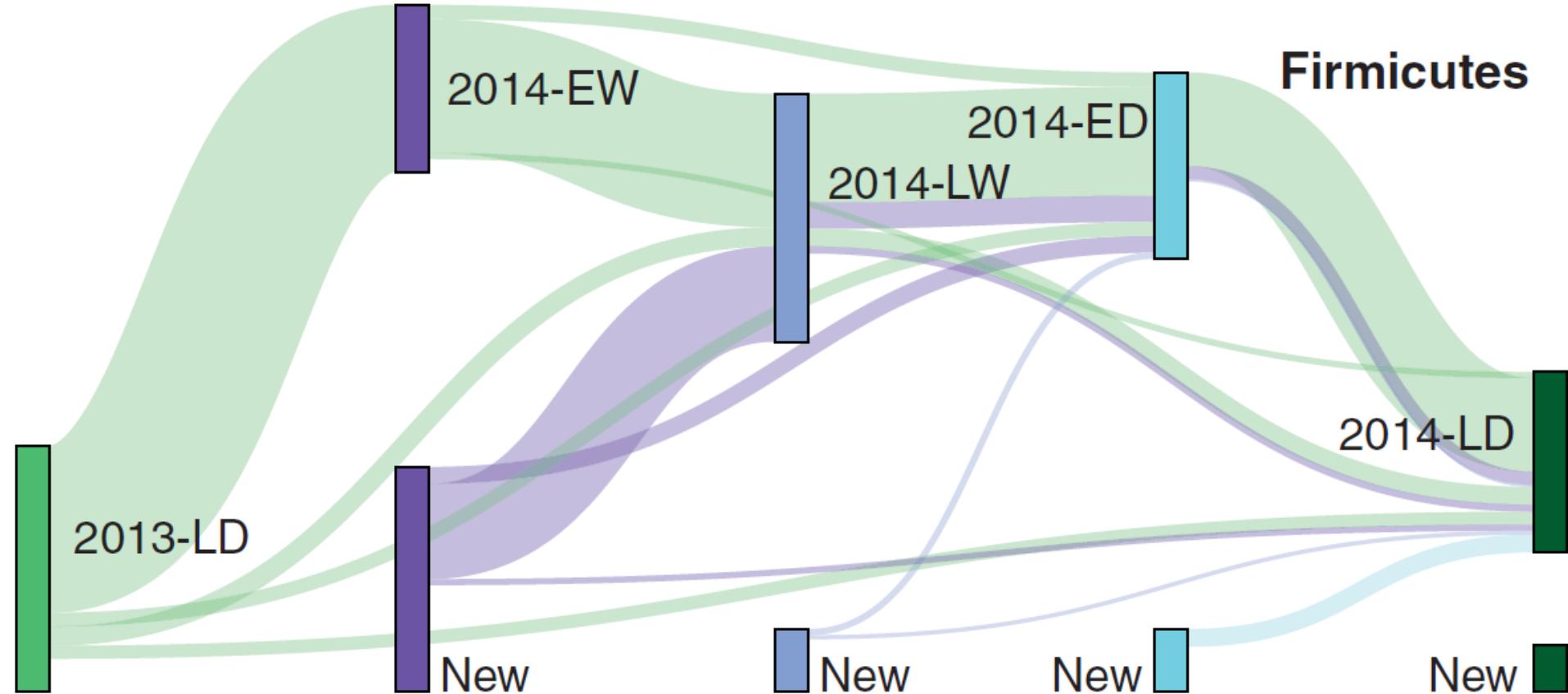
# 重复之时间点

一年内连续采样

三分法：旱季、雨季、旱季(冬-夏-冬)

五分法：冬、春、夏、秋、冬

样本量 = 41 + 19 + 58 + 30 + 40 = 188

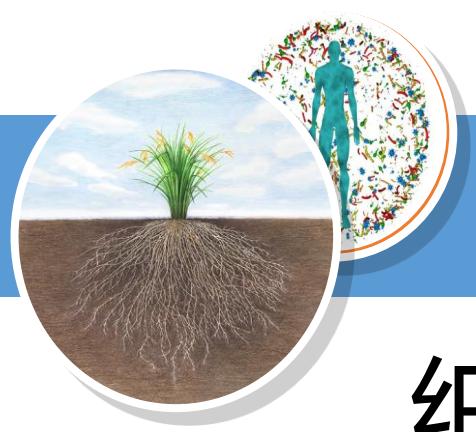




# 目录

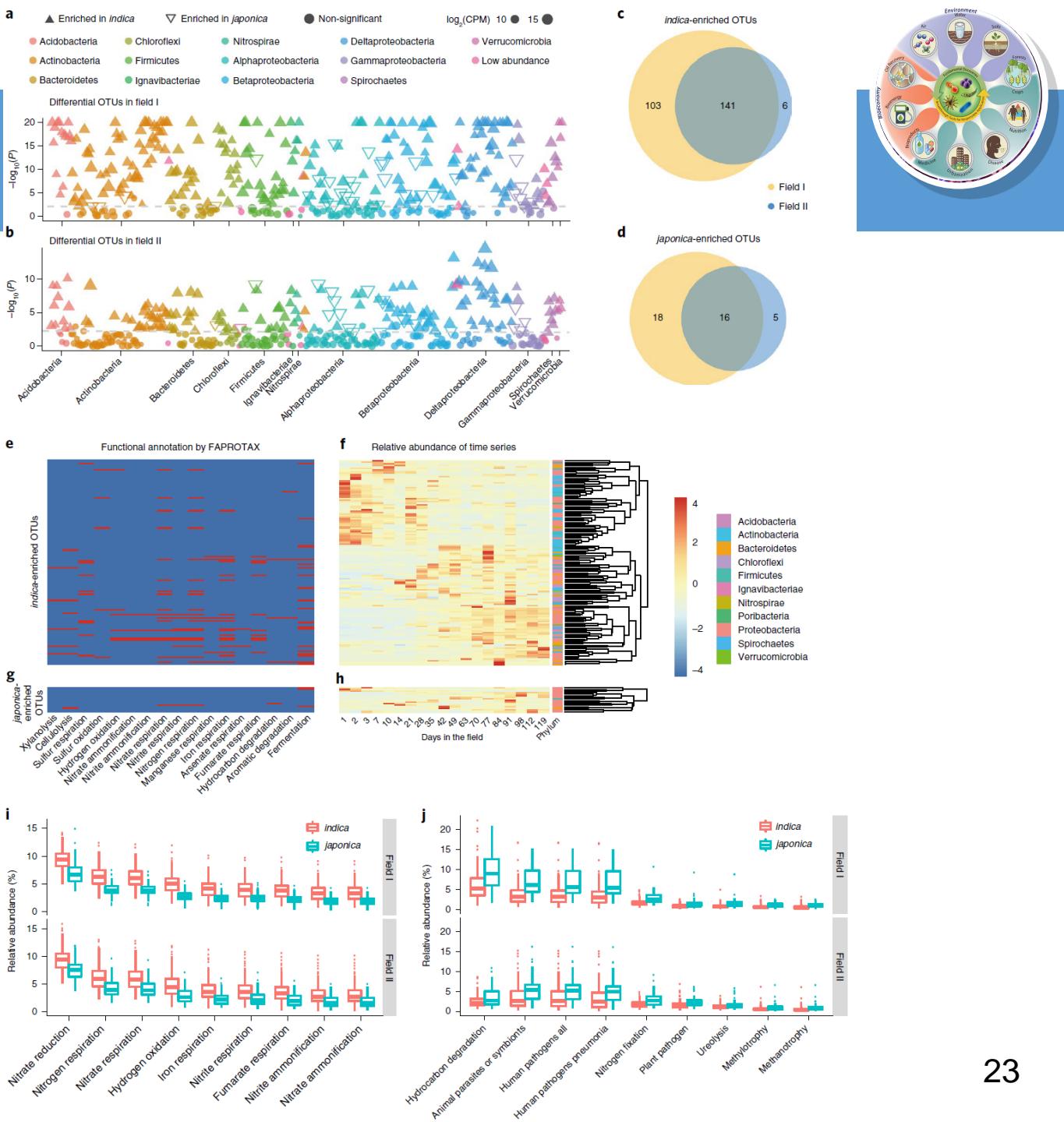


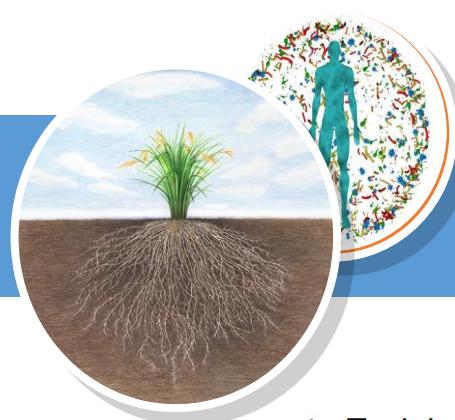
- 整体概述
  - 材料地图和/或实验设计
  - 样本概述:  $\alpha$ 多样性箱线图、 $\beta$ 多样性PCoA散点图、物种组成柱状图
- 细节展示
  - 物种组间差异: 曼哈顿图、韦恩图
  - 功能注释或差异: 热图有/无或时间序列、箱线图
- 应用场景——机器学习挖掘生物标记
  - 随机森林分类——区分不同组, 如疾病诊断、来源或品种鉴定
  - 随机森林回归——时间序列, 如年龄、死亡时间预测



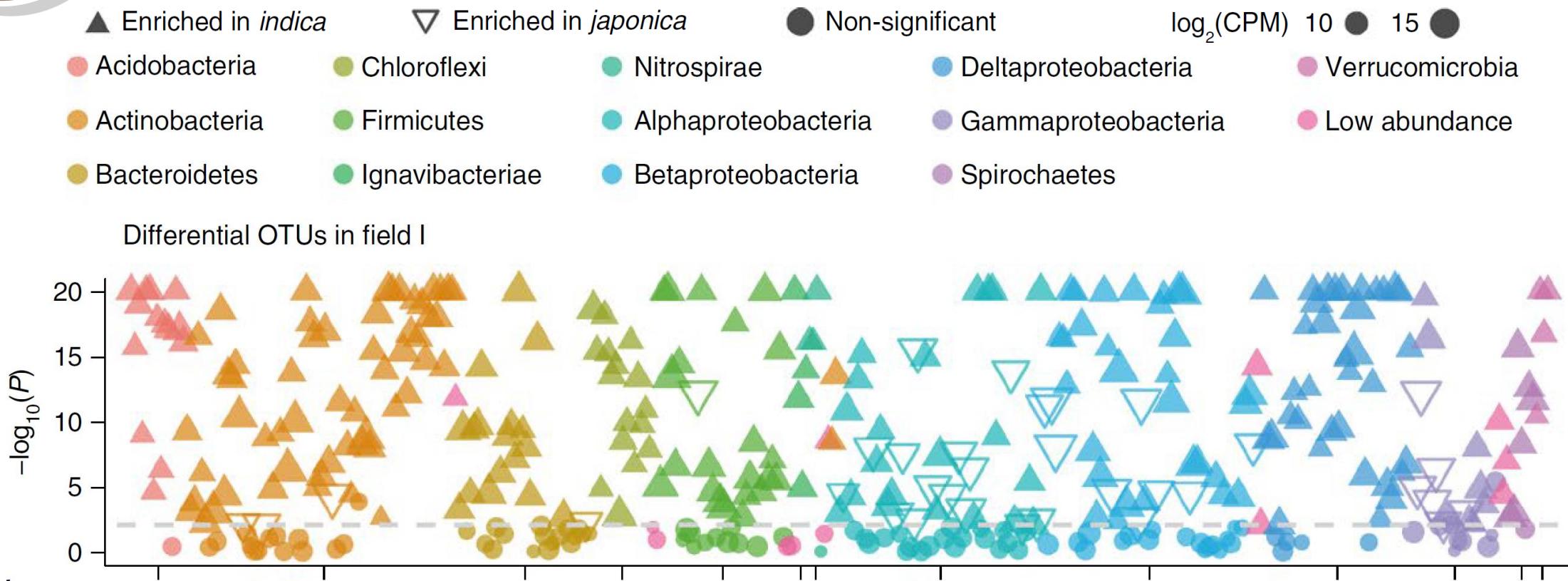
# 细节展示

1. 物种组间差异:
  - 曼哈顿图
  - 韦恩图
2. 功能注释或差异:
  - 热图有/无
  - 热图时间序列
  - 箱线图



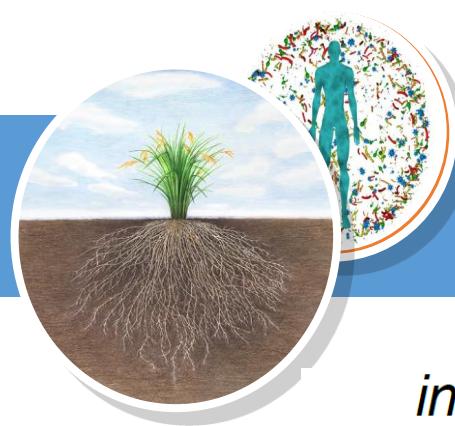


# 曼哈顿图展示差异

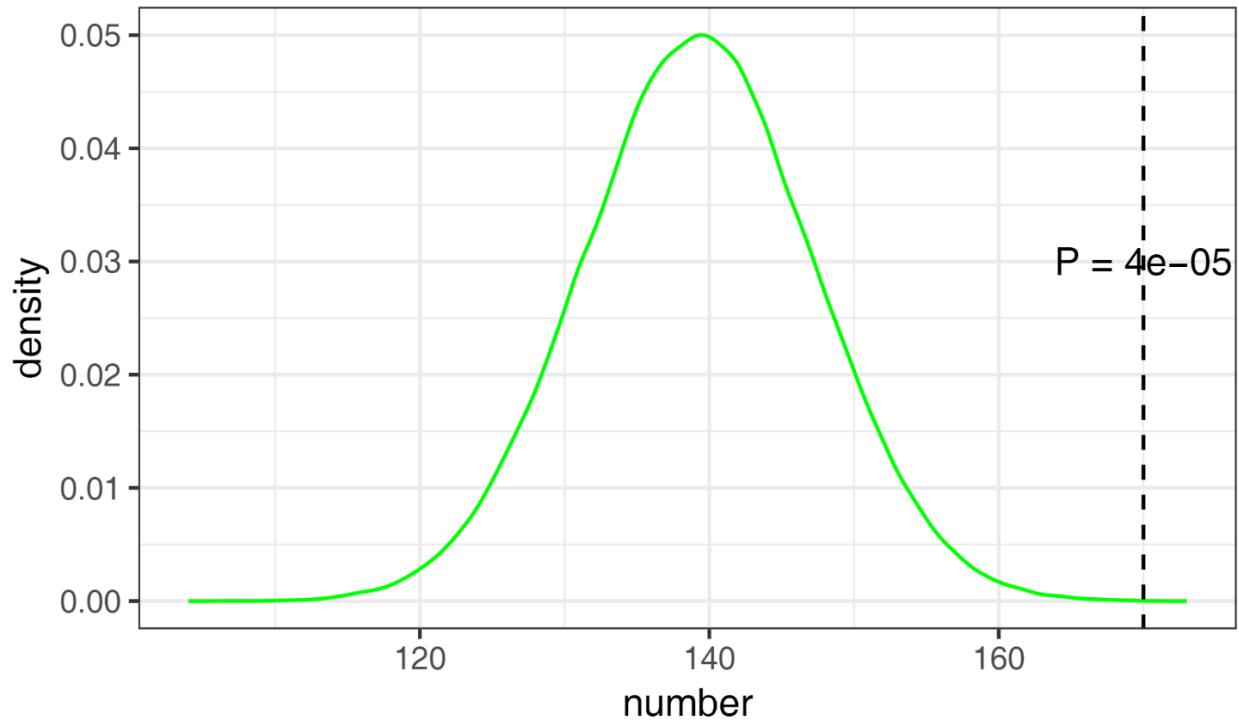
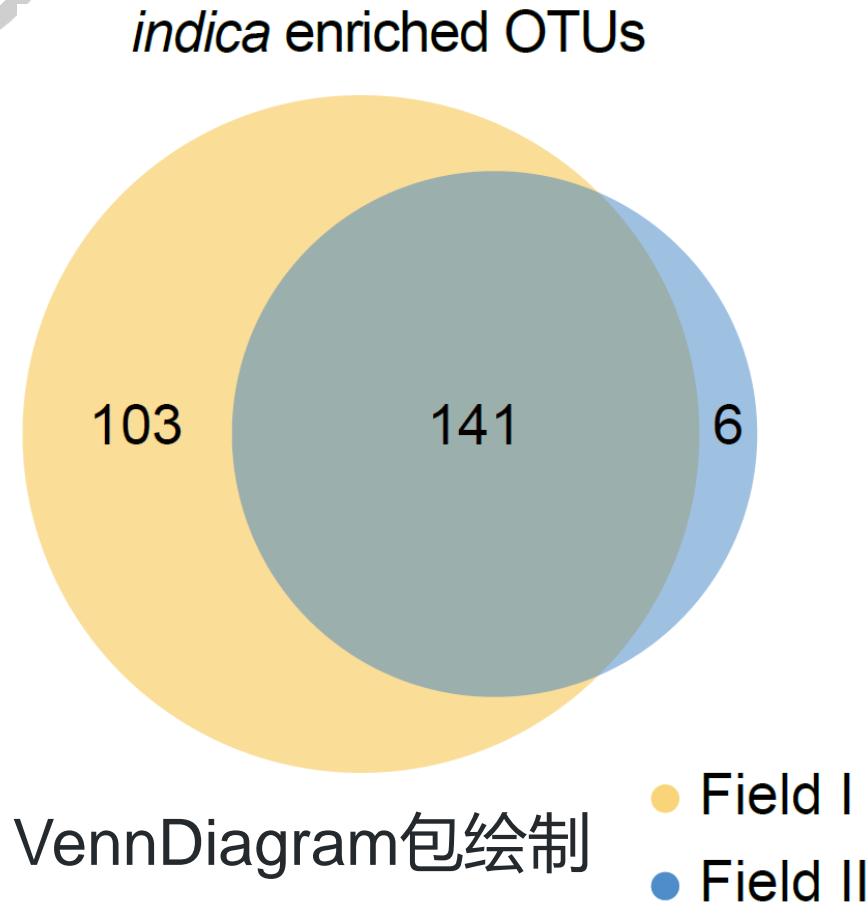


点代表物种，上三角为上调，下三角为下调，大小为丰度，颜色为物种分类  
 ggplot + geom\_point + geom\_hline + scale\_shape\_manual + scale\_size

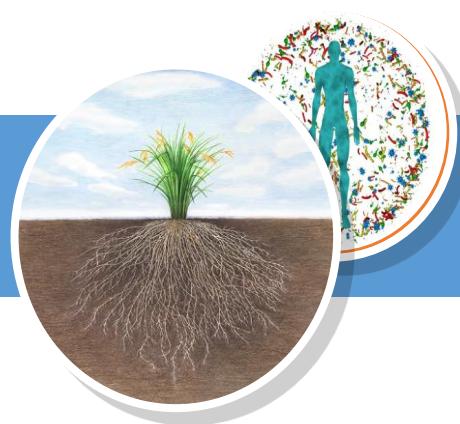
Jingying Zhang, Yong-Xin Liu, et. al. *Nature Biotechnology*. 2019. Fig 3a/b



# 韦恩图展示一致性



<https://github.com/microbiota/Huang2019SCIENCE>  
参考 <https://github.com/genomicsclass/labs>



pheatmap包  
绘制

双色热图展  
示每个OTUs  
中是否存在  
某种功能

# 热图展示有无或梯度



e

Functional annotation by FAPROTAX

f

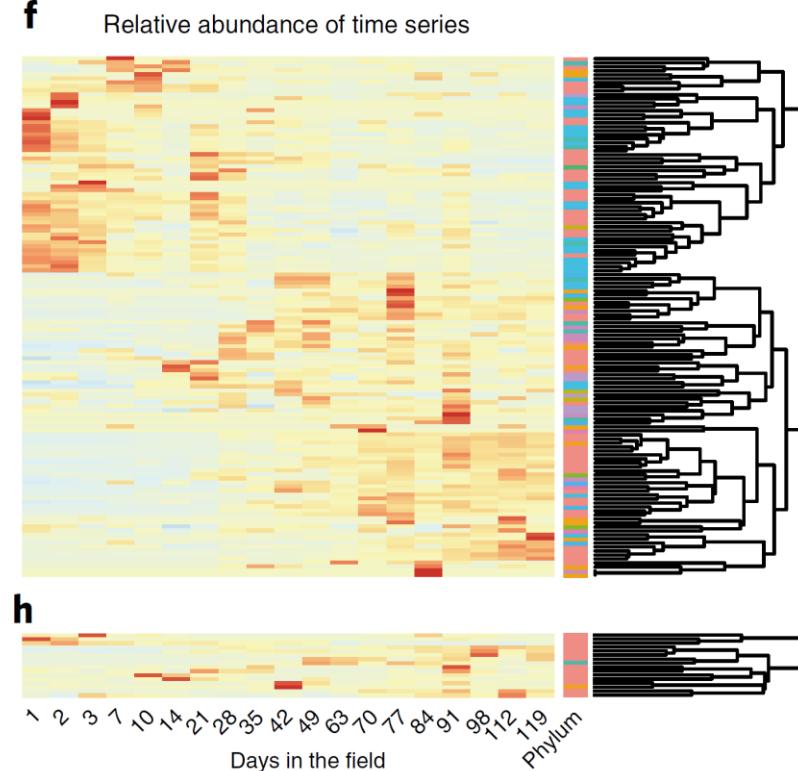
Relative abundance of time series

g

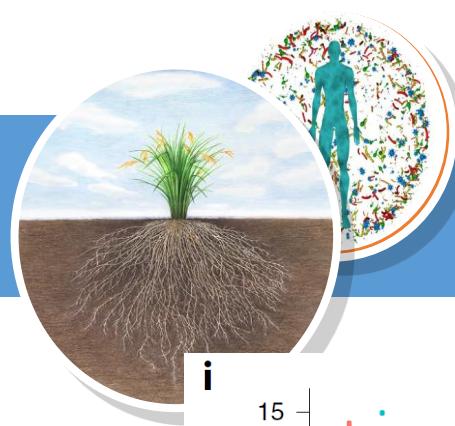
japonica-enriched OTUs

- Xylanolysis
- Cellulolysis
- Sulfur respiration
- Sulfur oxidation
- Hydrogen oxidation
- Nitrate ammonification
- Nitrite respiration
- Nitrogen respiration
- Manganese respiration
- Iron respiration
- Arsenate respiration
- Fumarate respiration
- Hydrocarbon degradation
- Aromatic degradation
- Fermentation

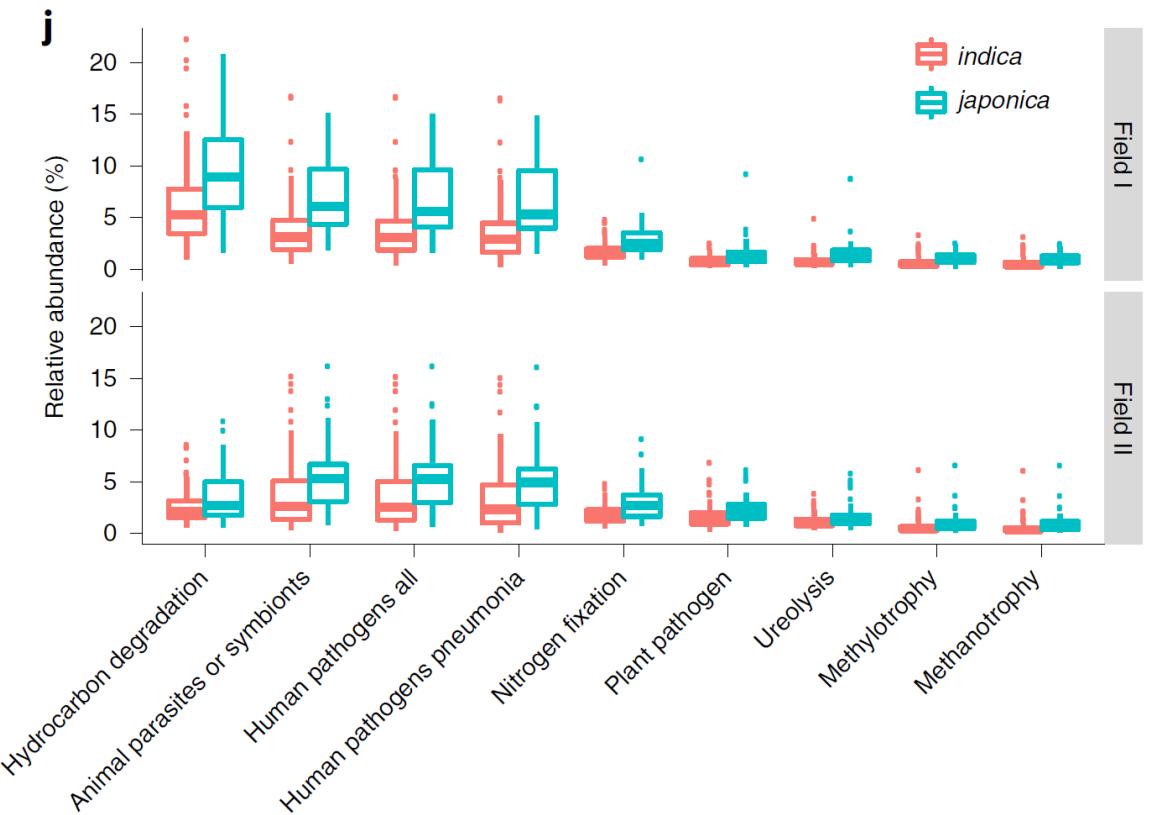
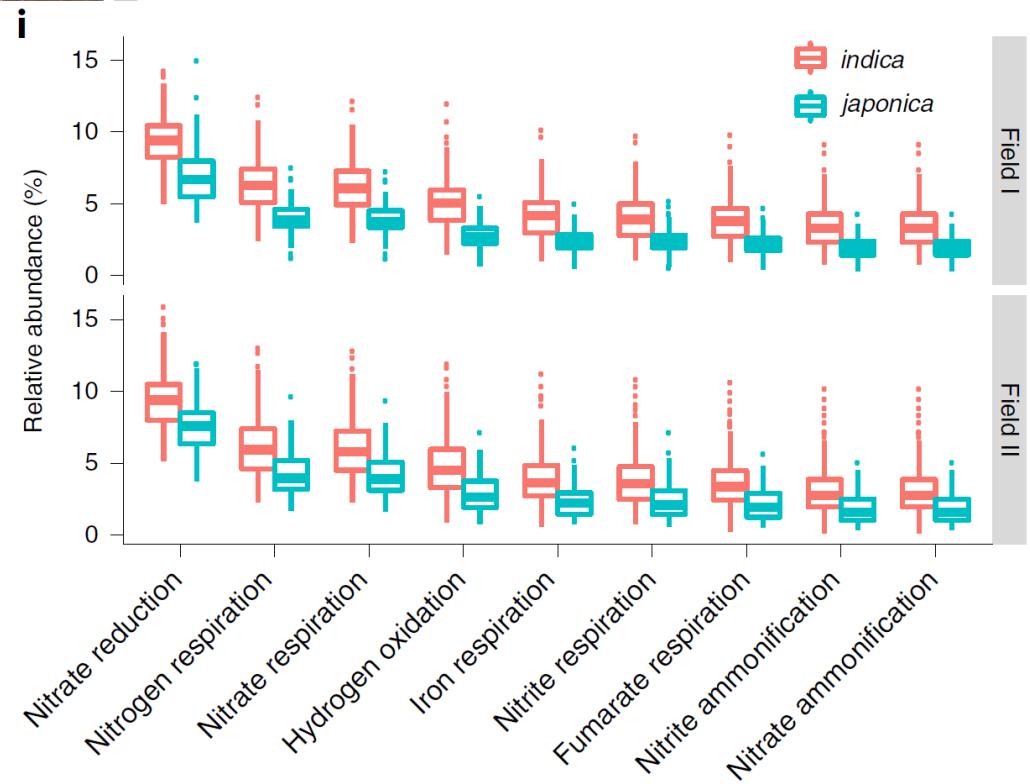
h



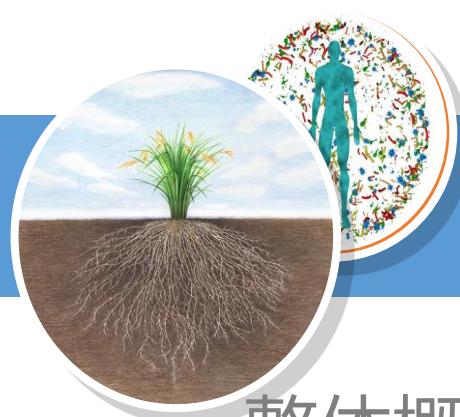
热图Z-Score  
展示每个  
OTUs在时间  
序列中的动  
态变化，方  
便观察丰度  
富集的时期



# 箱线图展示功能通路组间差异



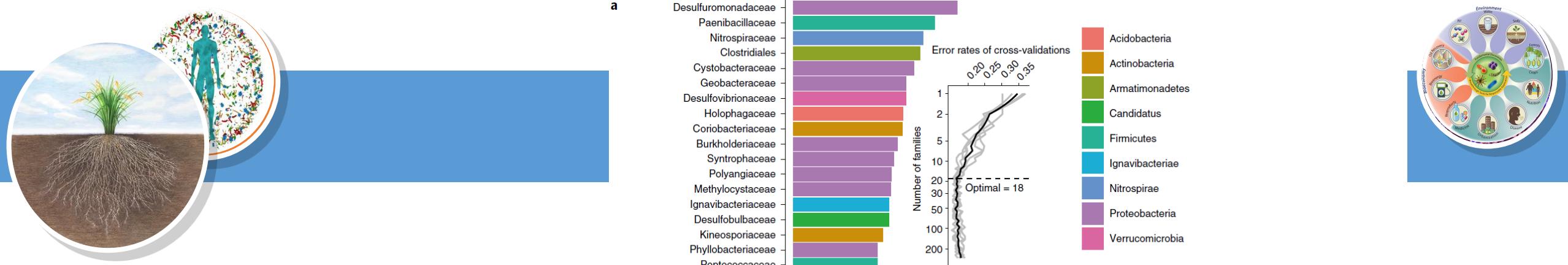
ggplot2绘制默认格式箱线图，左右分别为两组特异富含的功能通路的相对丰度，上下为两块地的重复facet\_grid(soiltype ~ .)



# 目录

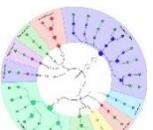
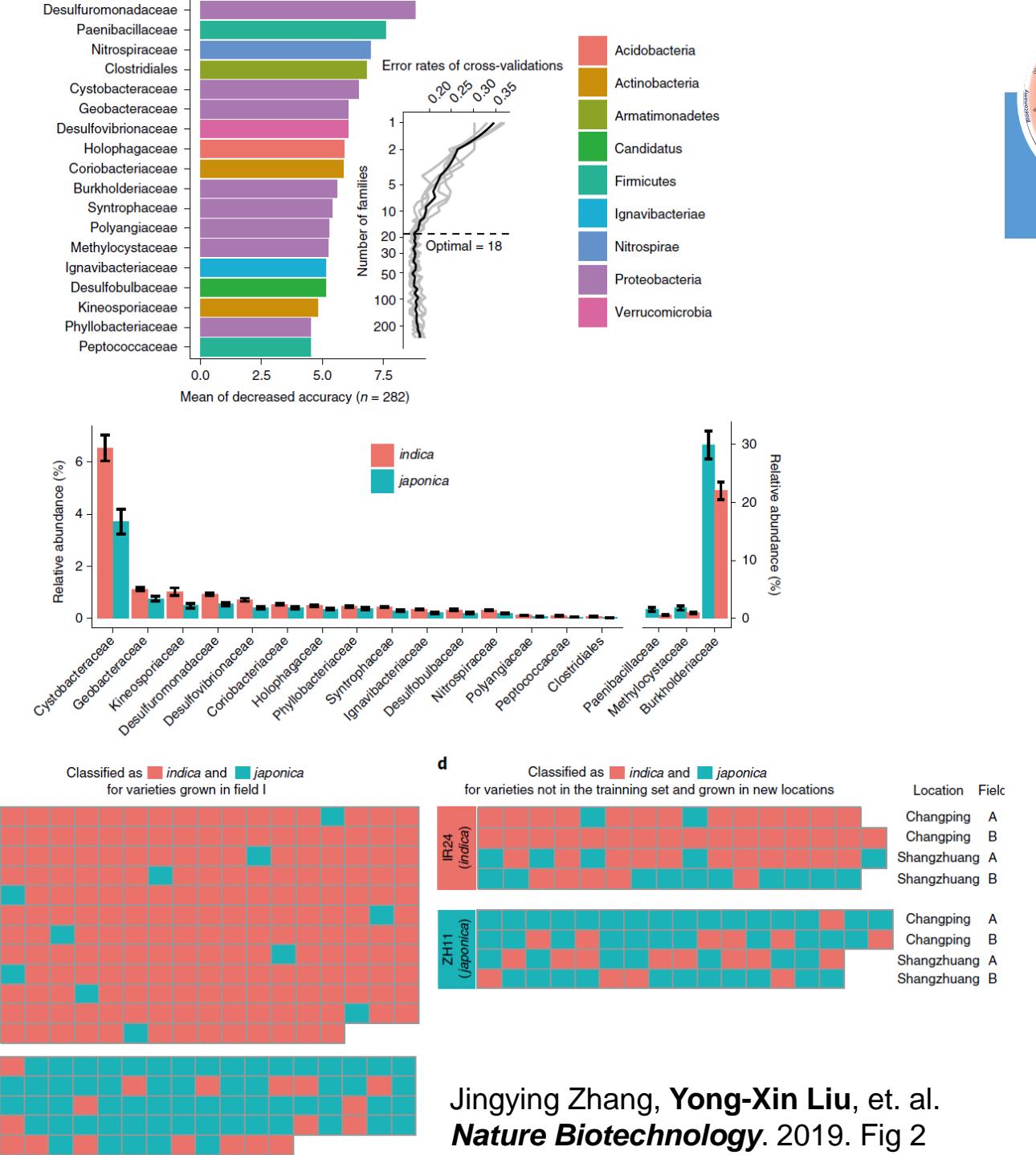


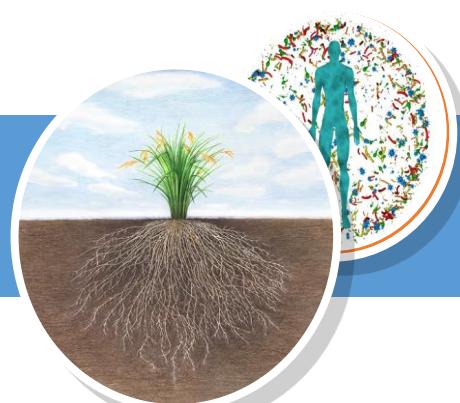
- 整体概述
  - 材料地图和/或实验设计
  - 样本概述:  $\alpha$ 多样性箱线图、 $\beta$ 多样性PCoA散点图、物种组成柱状图
- 细节展示
  - 物种组间差异: 曼哈顿图、韦恩图
  - 功能注释或差异: 热图有/无或时间序列、箱线图
- 应用场景——机器学习挖掘生物标记
  - 随机森林分类——区分不同组, 如疾病诊断、来源或品种鉴定
  - 随机森林回归——时间序列, 如年龄、死亡时间预测



## 随机森林分类

1. 鉴定特征的重要性
2. 交叉验证选择合适的特征组合
3. 特征在组间相对丰度
4. 同地和异地的验证

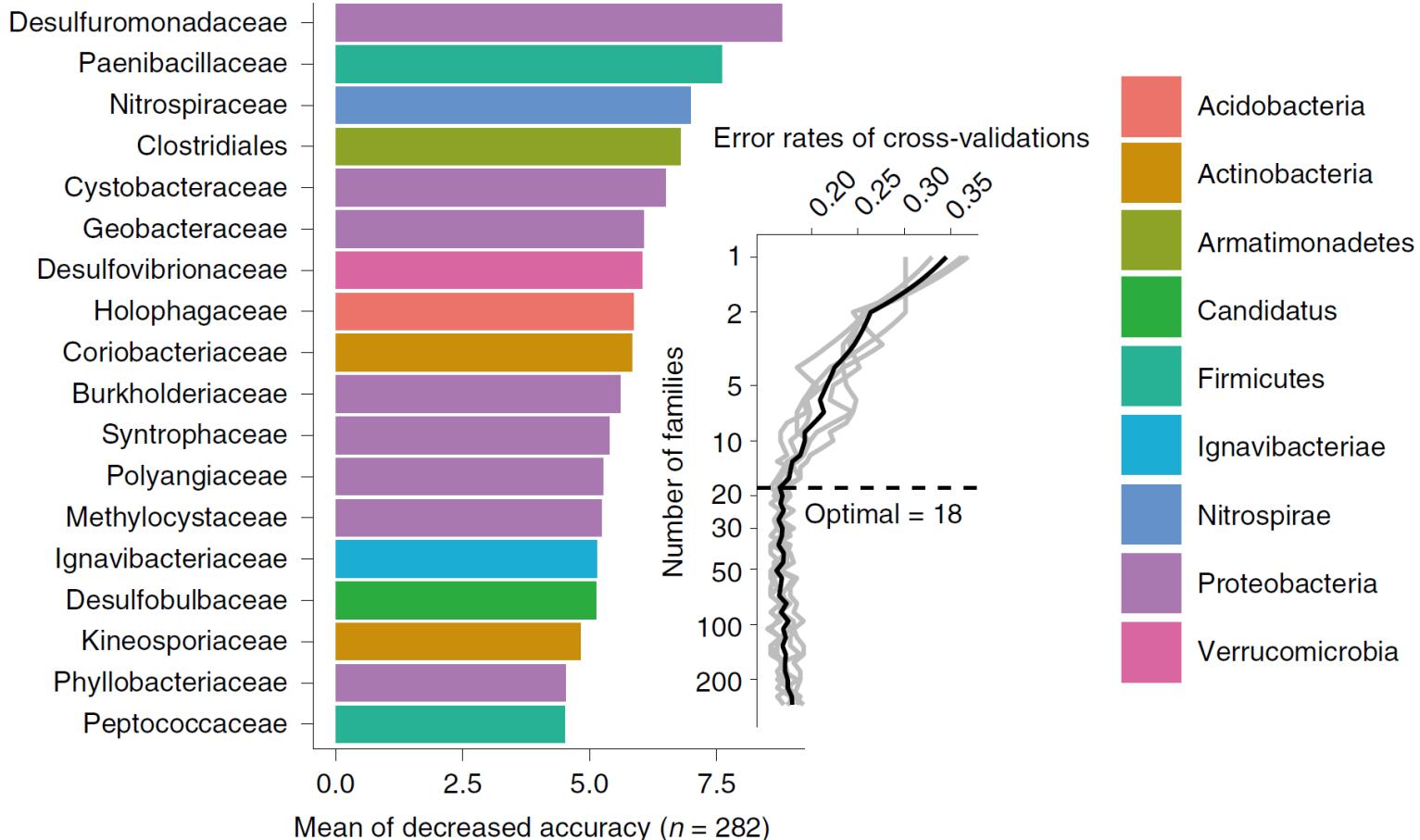


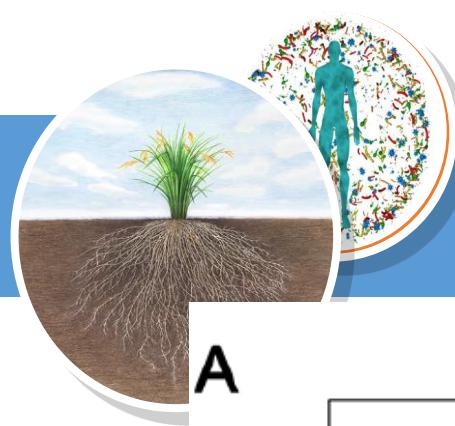


# 随机森林挖掘重要分类特征

randomForest包对样本离散型进行分类，连续型进行回归

1. randomForest建模和导出特征贡献度importance, ggplot2柱状图可视化
2. 多次交叉验证，人为选择合适的特征组合，ggplot2折线图可视化

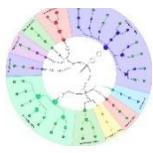
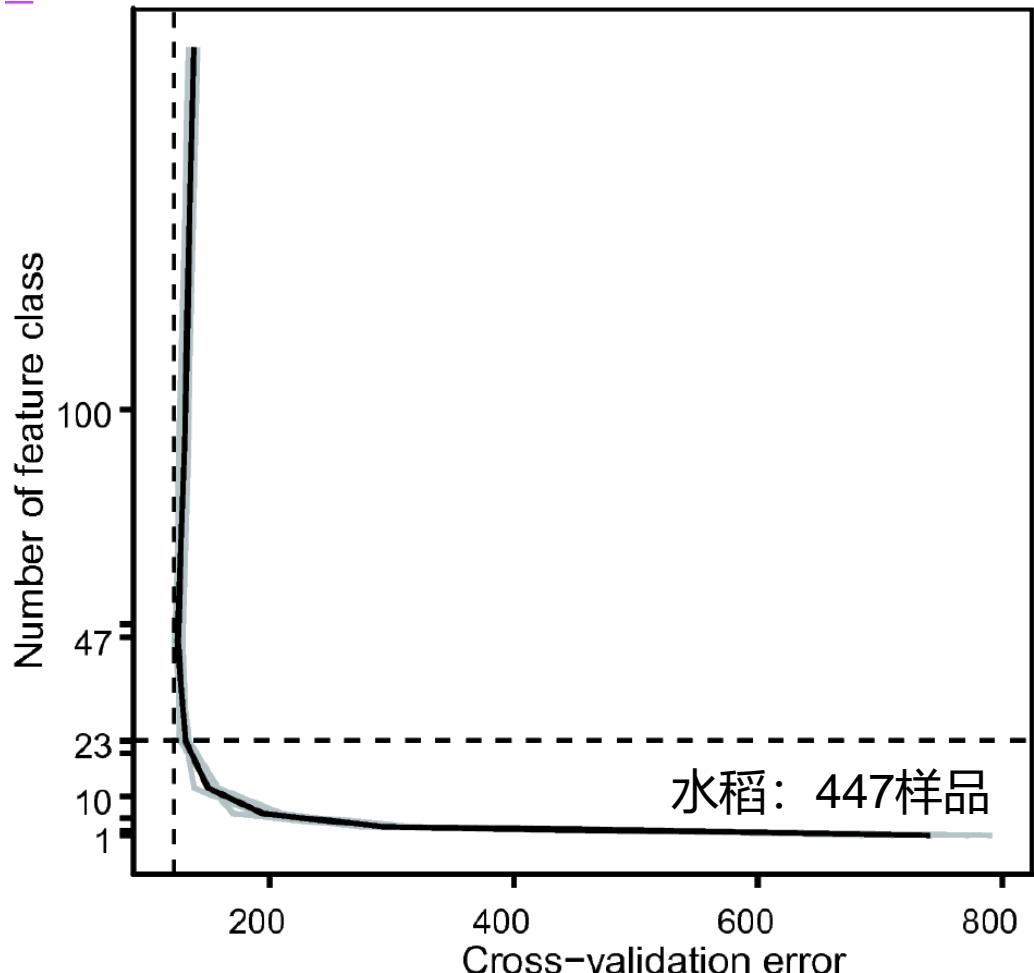
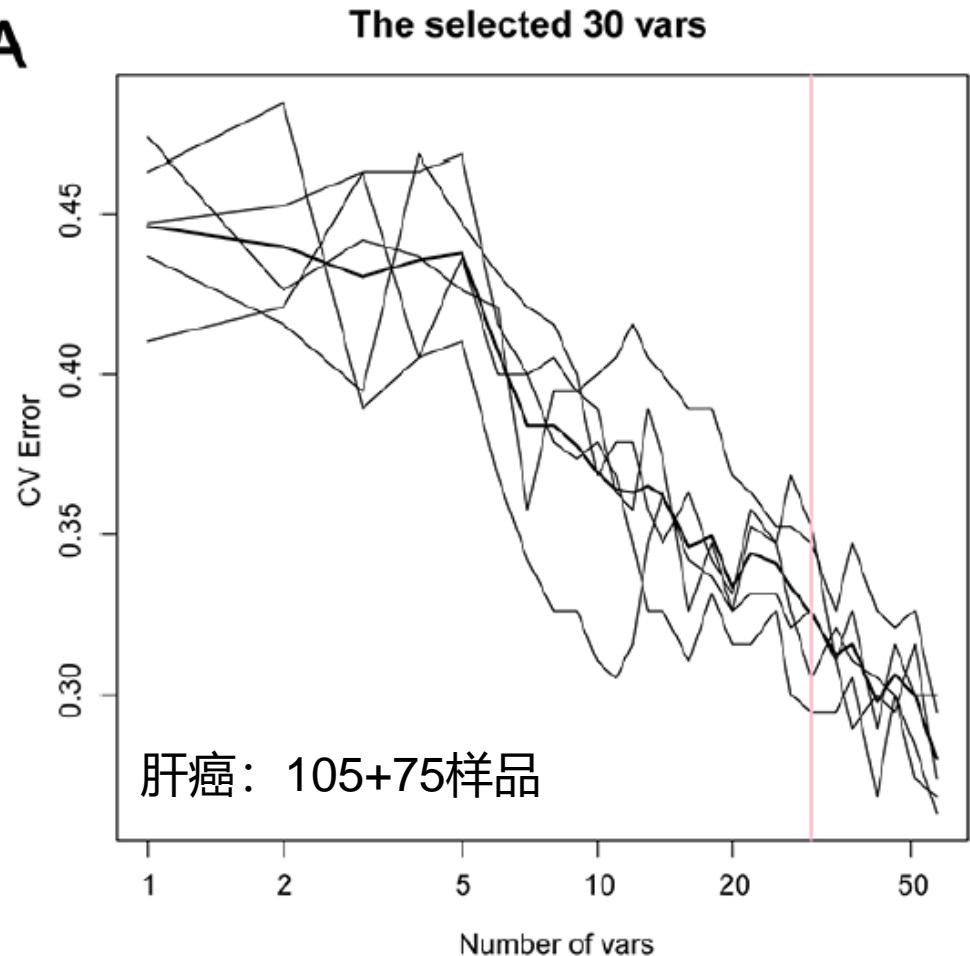


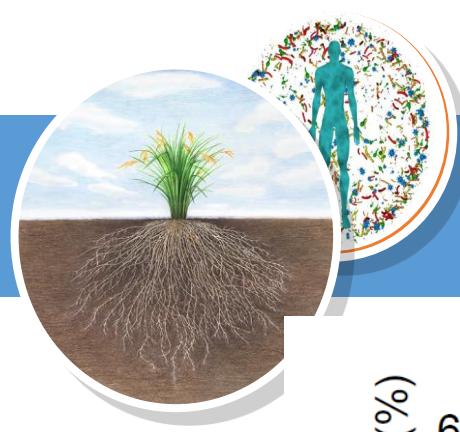


# 寻找生物标志物：交叉验证确定较优数量

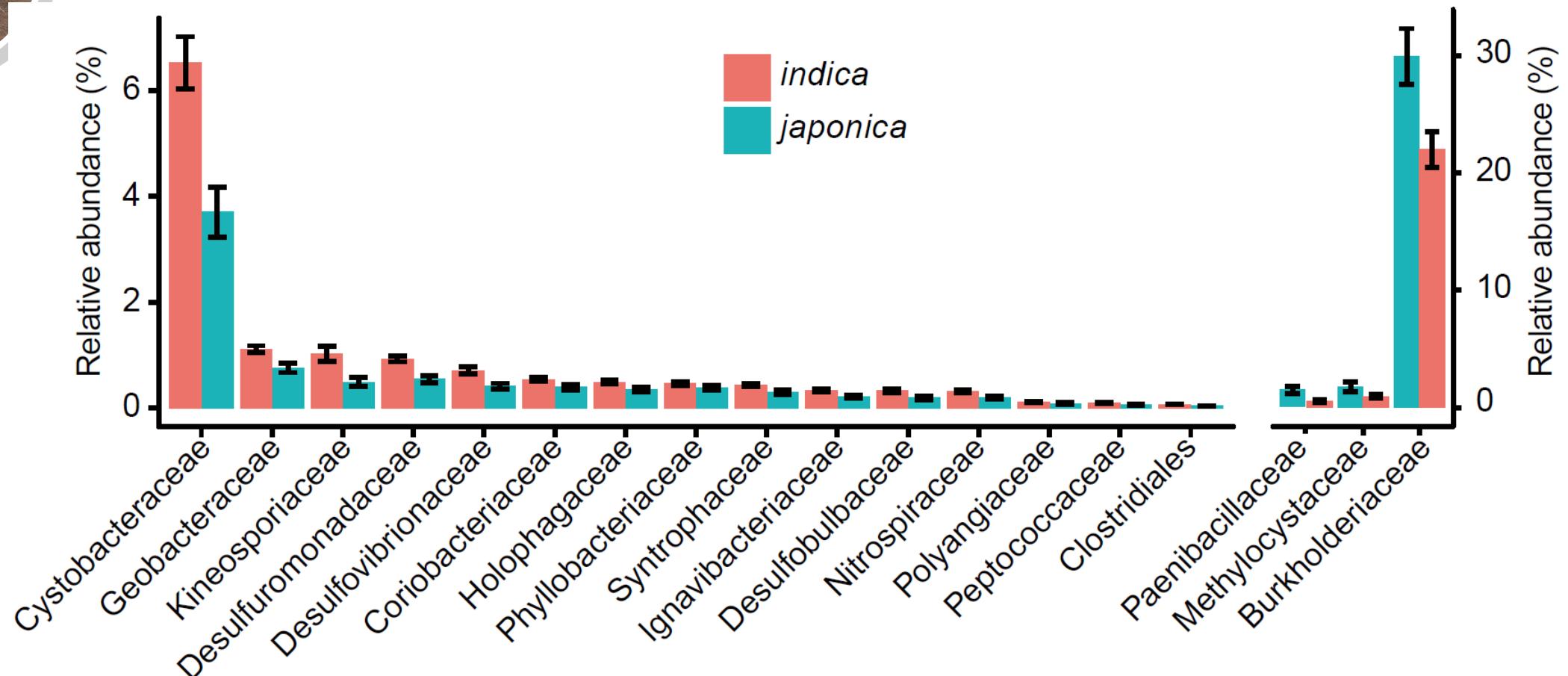


A

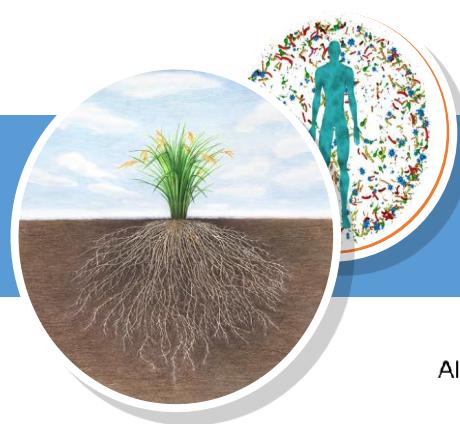




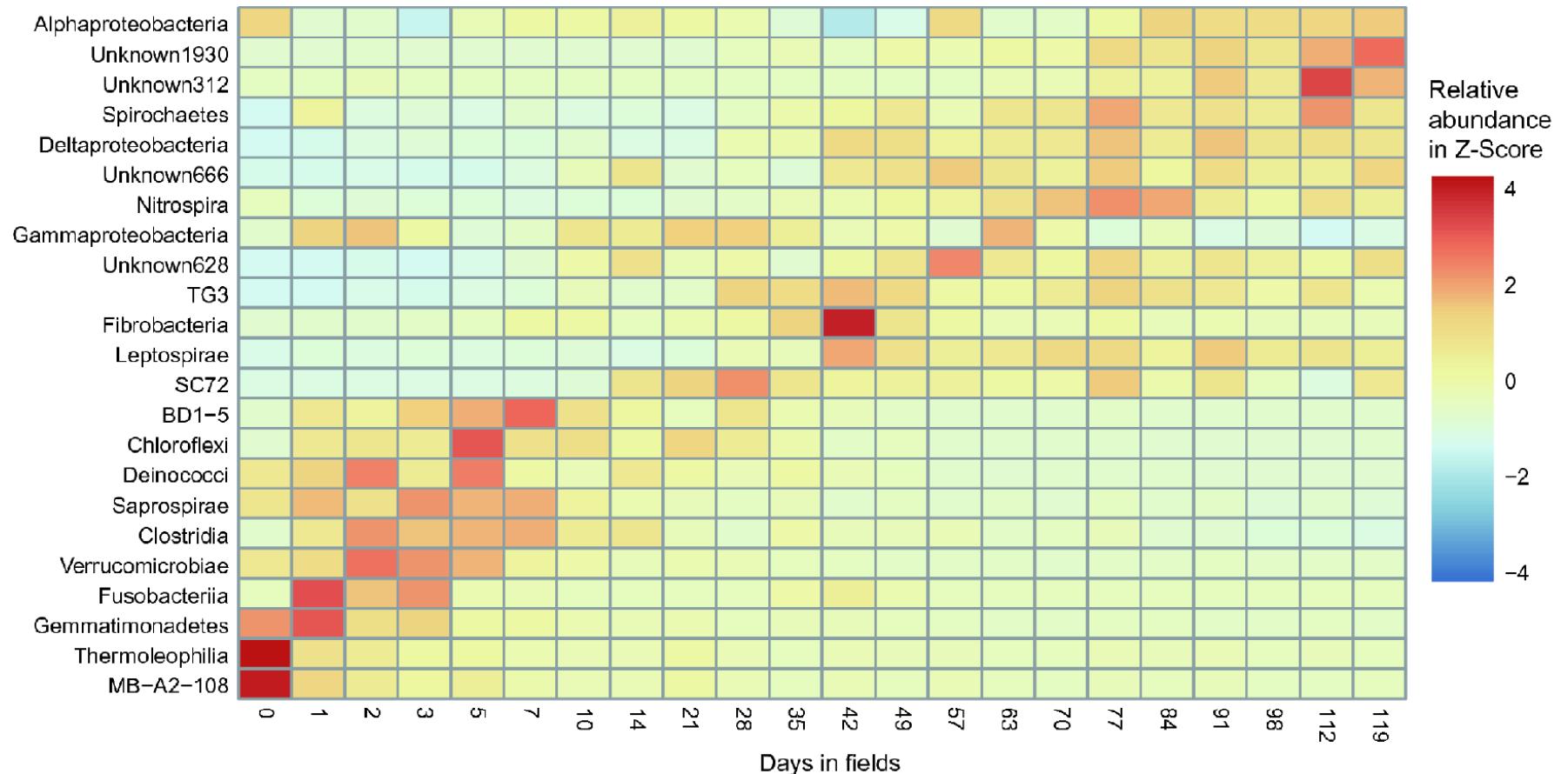
# 重要分类特征在组间差异



柱状图+标准误展示特征丰度的组间差异，分类再排序

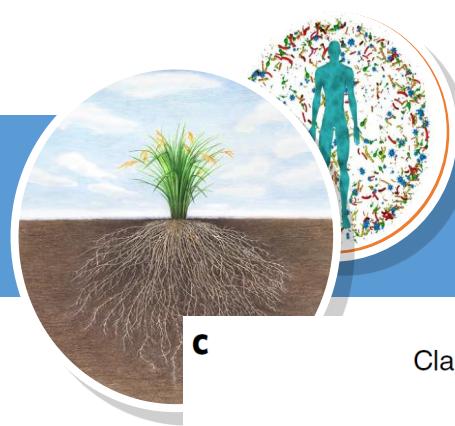


# 重要分类特征在时间上变化



heatmap绘制，数据行按最高丰度出现的时间排序，列按时间序列排序

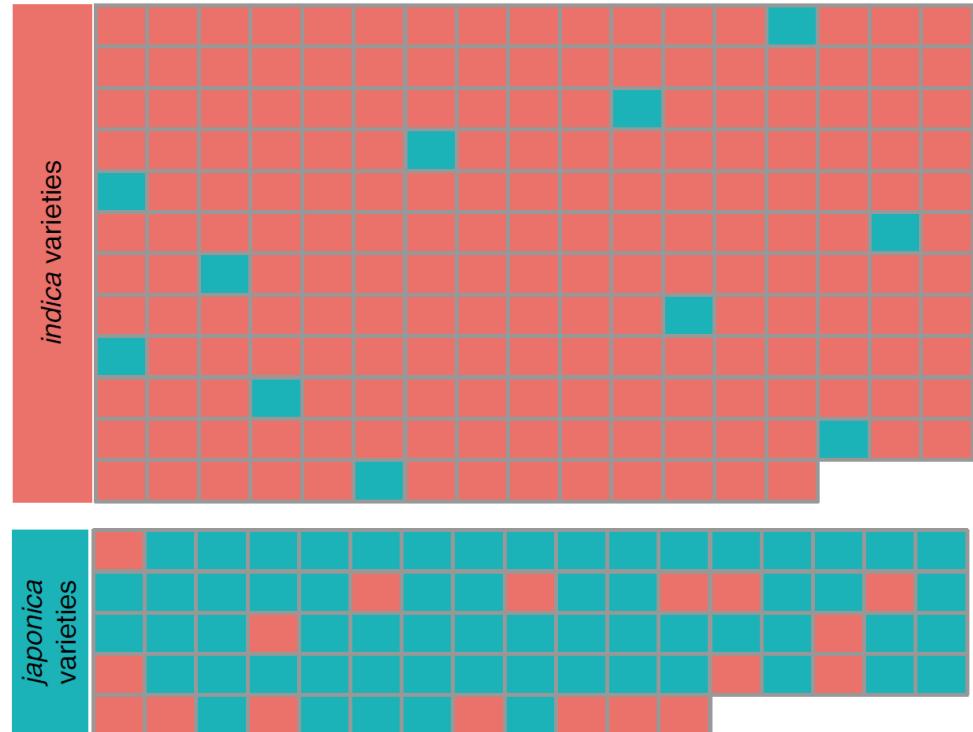
Jingying Zhang, Na Zhang, Yong-Xin Liu, et. al. Root microbiota shift in rice correlates with resident time in the field and developmental stage. *Science China Life Sciences*. 2018. doi:10.1007/s11427-018-9284-4 Fig 4B



# 模型普适性评估——预测样本分类

c

Classified as ■ *indica* and ■ *japonica*  
for varieties grown in field I



d

Classified as ■ *indica* and ■ *japonica*  
for varieties not in the training set and grown in new locations



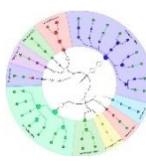
Location Field

Changping A  
Changping B  
Shangzhuang A  
Shangzhuang B

Changping A  
Changping B  
Shangzhuang A  
Shangzhuang B

heatmap 展示预测结果，需  
要考虑布局换行和末行补齐

模型对另一块地相同品种(c)和不同地点不同品种(d)预测

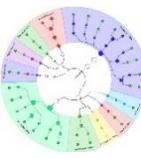


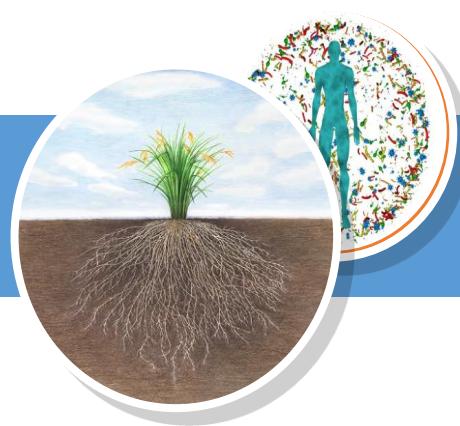


# 总结



- 文章的逻辑结构：总——分——应用，3 – 4个组图，最多6个
- 图的逻辑：3-6个子图构成一个组图，说明一个主题
- 1总：地图和实验设计、 $\alpha$ 多样性箱线图、 $\beta$ 多样性PCoA散点图、物种组成柱状图
- 2分：差异比较曼哈顿/火山图、韦恩图进一步比较重复性、热图结合有/无或丰度注释、箱线图展示差异功能
- 3应用：建模(特征数量和贡献度)、特征在组间差异、模型验证(预测不同来源数据测试准确性和普适性)
- 文章的逻辑以吸引人和可读性为主，而非实际实验或分析的顺序

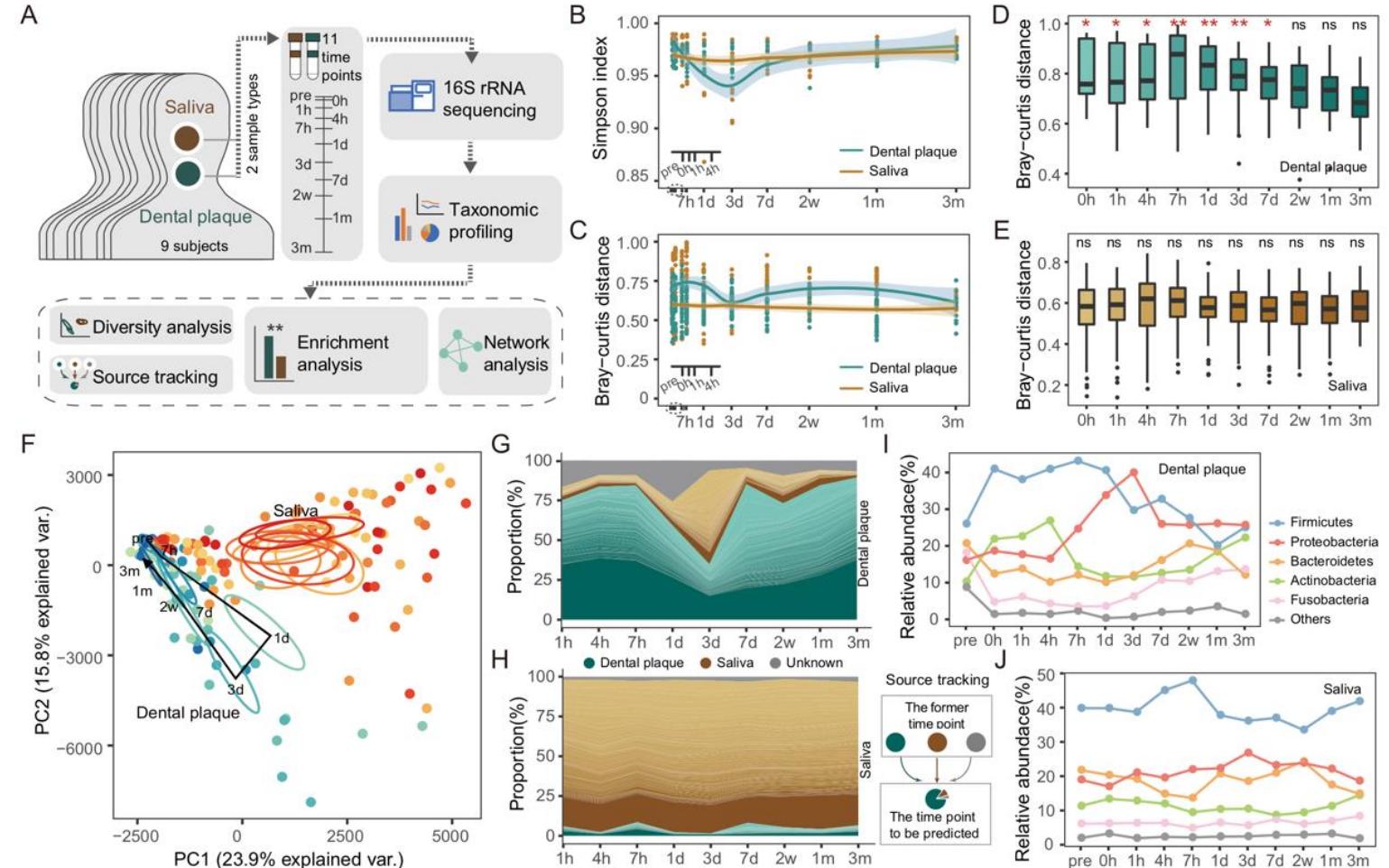


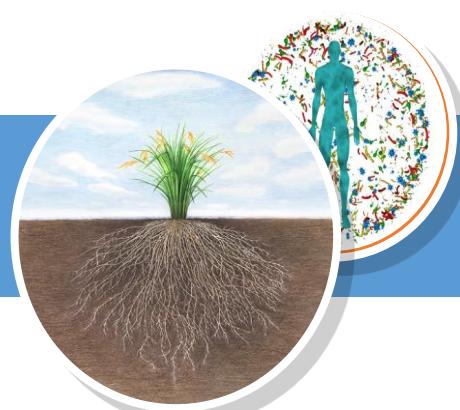


# 规律的验证1. 整体描述



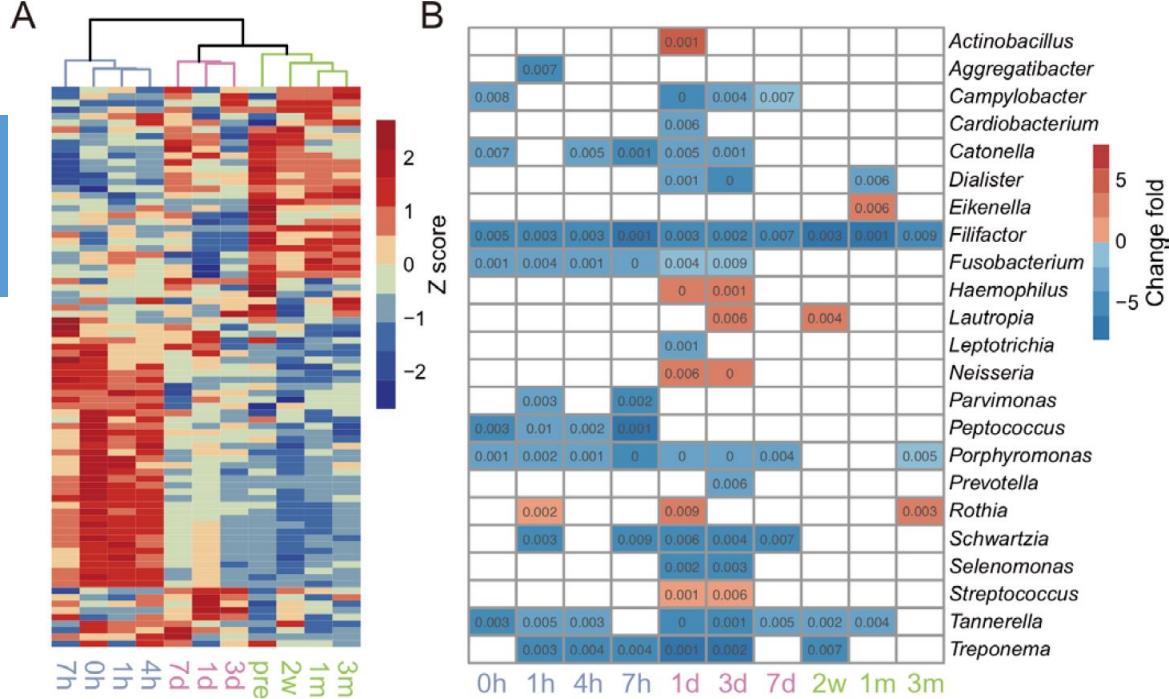
- A. 实验设计
- B. Alpha多样性
- C. Beta多样性拟合
- D. 牙菌斑Beta箱线
- E. 唾液Beta箱线
- F. PCA分析+置信椭圆
- G. 牙菌斑来源
- H. 唾液来源
- I. 牙菌斑物种组成
- J. 唾液物种组成



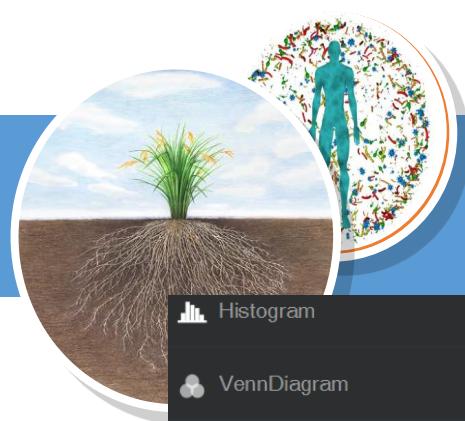


## 2. 细节展示

- A. 属水平热图组成+聚类
- B. 属水平显著差异+P值+差异倍数
- C. 网络分析
  - a. 线型实/虚代表正负相关
  - b. 线颜色代表时间短、中、长(与A对应)
  - c. 线粗细代表相关系数绝对值
  - d. 点大小代表相对丰度
  - e. 点数字代表节点度数
  - f. 点标签为属名
  - g. 有时还有点颜色，点边框色属性可添加



Wang, J., Jia, Z., Zhang, B., Peng, L. & Zhao, F. Tracing the accumulation of in vivo human oral microbiota elucidates microbial community dynamics at the gateway to the GI tract. *Gut*, gutjnl-2019-318977, doi:10.1136/gutjnl-2019-318977 (2019).



# ImageGP——在线绘制20种图

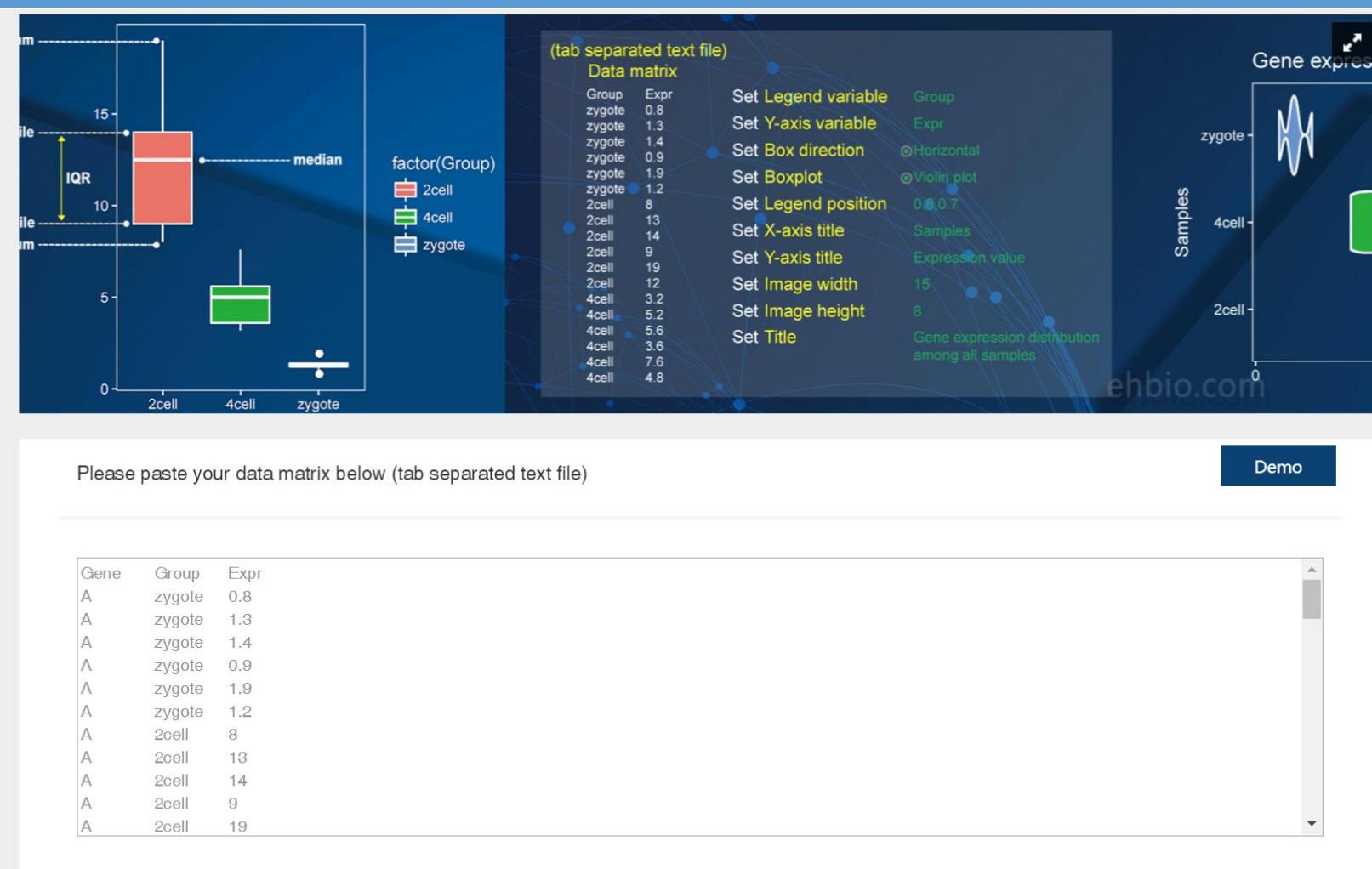
The screenshot displays the ImageGP web application interface. On the left, a sidebar lists various visualization tools: Histogram, VennDiagram, UpsetView, Density plot, PCA plot, PCoA plot, CPCoA plot, Sankey, PICRUSt, LEFSe, FAPROTAX, BugBase, Problem Feedback, and Manuals. Below these are links for FAQ and Code wall, with a 'NEW!' badge. A purple '反馈建议' (Feedback) button is also present. The main area features a 'Image previews' section with nine thumbnail images of different types of plots: stacked bar charts, density plots, PCA plots, Venn diagrams, scatter plots, contour plots, and heatmaps. To the right of the thumbnails is a large search bar containing the URL 'http://www.ehbio.com/ImageGP/'. Below the search bar is a 'Google Scholar' logo and a 'Articles' section with a blue diamond icon. A large text box on the right contains the URL 'www.ehbio.com/ImageGP' and the text 'About 30 results (0.08 sec)'.



# 粘贴数据 - 选择参数 - 一键出图



- [Home](#)
- [Line plot](#)
- [GO Enrichment plot](#)
- [Pretty Heatmap](#)
- [Boxplot](#)
- [Scatter plot](#)
- [Bar plot](#)
- [Volcano plot](#)
- [Manhattan plot](#)
- [Histogram](#)
- [VennDiagram](#)
- [UpsetView](#)
- [Density plot](#)
- [PCA plot](#)
- [PCoA plot](#)
- [CPCoA plot](#)





# 提供开源可重复分析的实验室

- 密歇根大学Pat Schloss <http://www.schlosslab.org>
- 斯坦福大学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/>
- 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>



# 数据代码图表整理示例



<https://github.com/YongxinLiu/Zhang2019NBT>

YongxinLiu / Zhang2019NBT

Code Issues Pull requests Projects Wiki Insights Settings

Scripts for stat and plot figures in rice microbiome paper

Manage topics

8 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find File Clone or download

YongxinLiu This is my first commit via Git!

data	final_submit	22 days ago
fig1	final submit	22 days ago
fig2	final submit	22 days ago
fig3	This is my first commit via Git!	a minute ago
fig4	final submit	22 days ago
fig5	final submit	22 days ago
fig6	This is my first commit via Git!	a minute ago
script	This is my first commit via Git!	a minute ago
README.md	This is my first commit via Git!	a minute ago

README.md

Zhang2019NBT 简介、文件描述和引文

Scripts for statistics and plotting figures in "NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice", published in Nature Biotechnology 2019.

- ./data # metadata, OTU table and taxonomy files
- ./fig1-6 # raw data, Rmarkdown scripts and output HTML format results
- figX/figX.Rmd # X is number 1-6, including the reproducible R scripts for each panel in figure
- figX/figX.html # Readability report by R markdown, include annotations, scripts and figures
- script/ # General R scripts used in this study

If you used these scripts, please cited the paper below:

Jingying Zhang#, Yong-Xin Liu#, Na Zhang#, Bin Hu#, Tao Jin#, ..., Chengcai Chu\*, Yang Bai\*. NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice. 2019. Nature Biotechnology.

Branch: master Zhang2019NBT / fig1 /

YongxinLiu final submit

..

- alpha.txt
- alpha\_shannon\_e.pdf
- alpha\_shannon\_e.txt
- beta\_filedl\_bray\_curtis.pdf
- beta\_filedl\_bray\_curtis.pdf
- design.png
- fig1.Rmd
- fig1.html
- minicore-worldmap.pdf
- tax\_pc\_pc\_group.pdf
- tax\_pc\_pc\_group.txt
- tax\_pc\_pc\_sample.txt
- varieties\_geo.txt

fig1.Rmd 分析代码及注释

fig1.html 分析代码、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

\*.pdf 矢量图，常用保存格式，方便查看和编辑

图1分析目录文件

<https://github.com/YongxinLiu/Zhang2019NBT>



# 分析代码样式和网页预览效果



Branch: master ▾ Zhang2019NBT / fig1 / fig1.Rmd

Find file Copy path

YongXinLiu final submit 94b8505 22 days ago

1 contributor

322 lines (264 sloc) | 12.2 KB

Raw Blame History

```

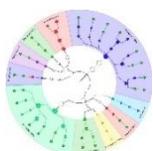
1 ---
2 title: "Figure 1. Root microbiota of indica and japonica."
3 author: "Yong-xin Liu"
4 date: "2019/2/20"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 # Clean workspace
11 rm(list=ls())
12 # Load setting and functions
13 source("../script/stat_plot_functions.R")
14 # Set output directory
15 output_dir<-"./"
16 ```
17
18 ## a. World map
19
20 (a) Diagram of original collection sites (44 countries) of indica (red) and japonica (blue) rice.
21
22 ```{r geomap, echo=TRUE}
23 library(dplyr)
24 library(mapproj)
25 library(ggplot2)
26 library(maps)
27
28 geatable = read.table("varieties_geo.txt", header = T, sep = "\t")
29 worldmap = map_data("world")
30
31 fig1 = ggplot(geatable, aes(longitude, Latitude, color = Subspecies)) +
32   geom_polygon(data = worldmap, aes(x = long, y = lat, group = group, fill = NA), color = "grey70", size = 0.25) +
33   geom_point(size = 2.5, alpha = 0.5)+ scale_colour_brewer(palette = "Set1")+
34   scale_fill_brewer(palette = "Set1")+
35   coord_cartesian()+
36   scale_y_continuous(breaks = (-3:3)*30)+
37   scale_x_continuous(breaks = (-6:6)*30) +
38   labs(x="Longitude", y="Latitude", colour = "Subspecies" ) +
39   theme_tufte()
40 ggsave(paste0(output_dir, "minicore-worldmap.pdf", sep=""), fig1, width = 9, height = 5)
41 fig1
42
43
44

```

标题、作者、日期

设置通用参数和加载  
通用依赖关系

图1a, 绘制材料来  
源地图的代码



宏基  
因组

fig1.Rmd 代码  
在线或Rstudio查看

Figure 1. Root microbiota of indica and japonica.

Yong-Xin Liu  
2019/2/20

a. World map

a. Diagram of original collection sites (44 countries) of indica (red) and japonica (blue) rice.

```

library(dplyr)
library(mapproj)

## Warning: package 'mapproj' was built under R version 3.4.4

## Loading required package: sp

## Checking rgeos availability: FALSE
## Note: when rgeos is not available, polygon geometry computations in mapproj depend on gplib,
## which has a restricted licence. It is disabled by default;
## to enable gplib, type gplibPermit()

library(ggplot2)
library(maps)

geatable = read.table("varieties_geo.txt", header = T, sep = "\t")
worldmap = map_data("world")

fig1 = ggplot(geatable, aes(longitude, Latitude, color = Subspecies)) +
  geom_polygon(data = worldmap, aes(x = long, y = lat, group = group, fill = NA), color = "grey70", size = 0.25) +
  geom_point(size = 2.5, alpha = 0.5)+ scale_colour_brewer(palette = "Set1")+
  scale_fill_brewer(palette = "Set1")+
  coord_cartesian()+
  scale_y_continuous(breaks = (-3:3)*30)+
  scale_x_continuous(breaks = (-6:6)*30) +
  labs(x="Longitude", y="Latitude", colour = "Subspecies" ) +
  theme_tufte()
ggsave(paste0(output_dir, "minicore-worldmap.pdf", sep=""), fig1, width = 9, height = 5)
fig1

```

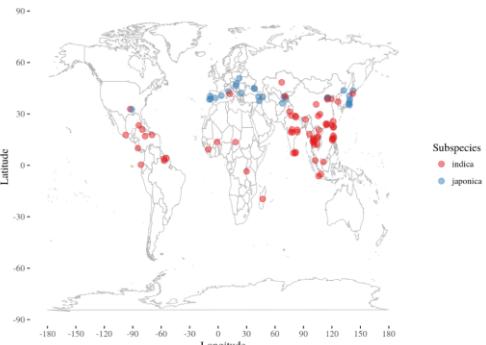


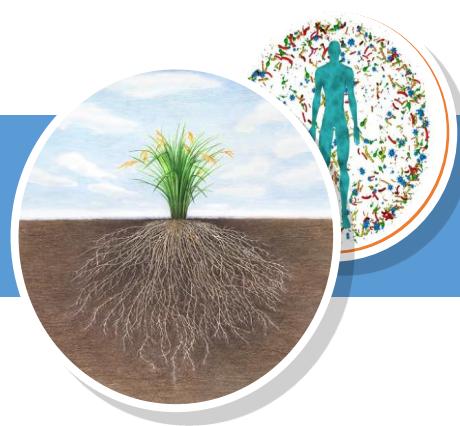
fig1.html可由Rmd生成  
下载用浏览器查看

标题、作者和时间  
带格式

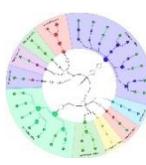
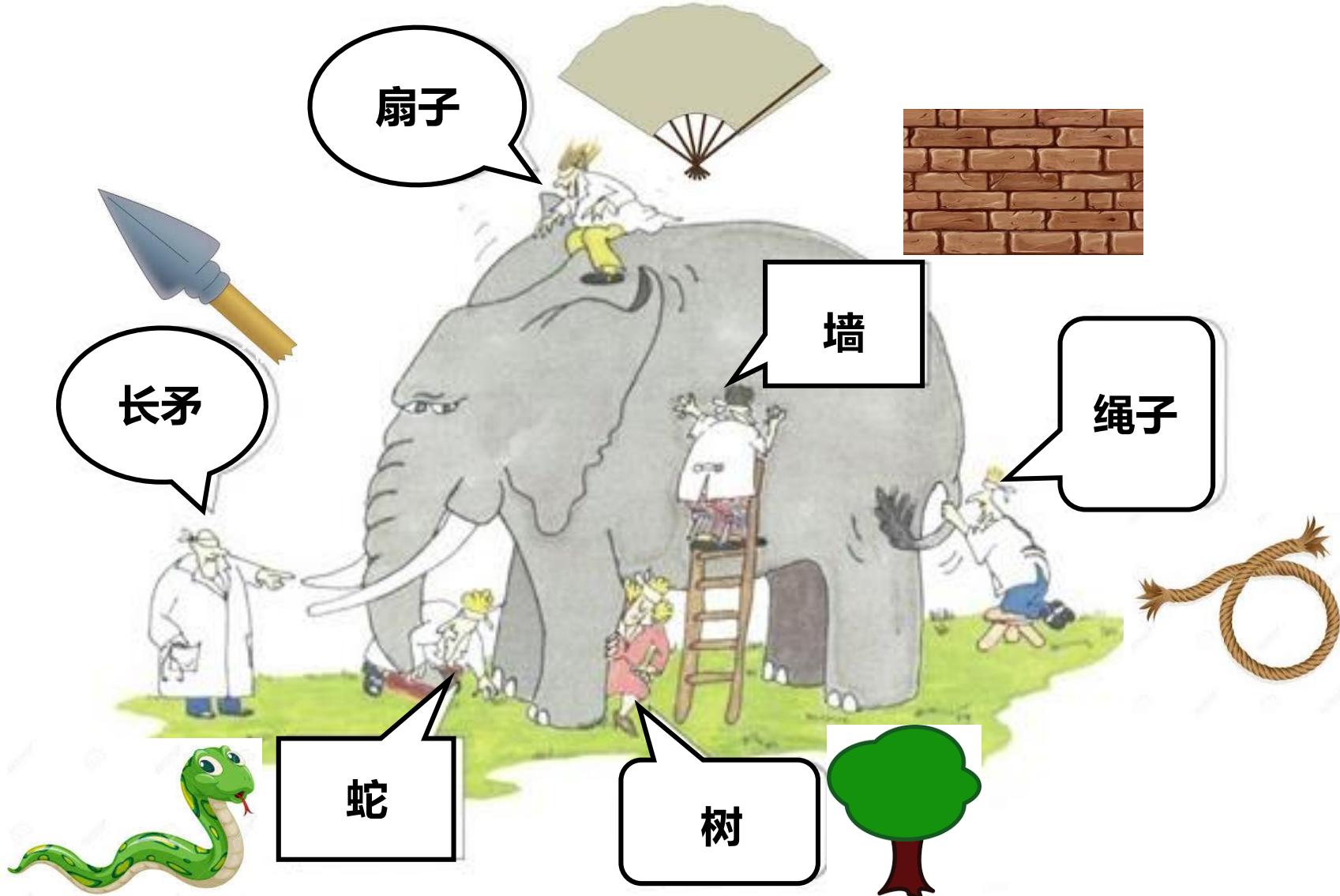
运行中的提示信息

分析代码

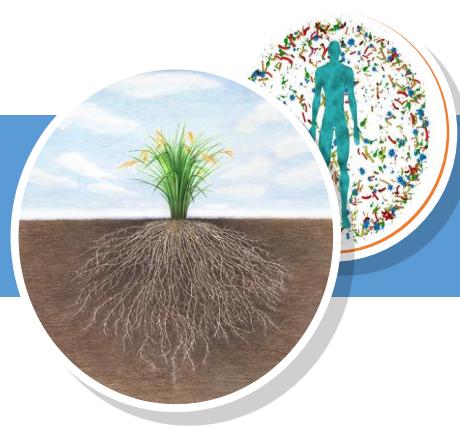
绘图结果



# 科研的过程——盲人摸象



宏基  
因组



# 投必得



投必得 - 您的论文专家  
[www.TopEditSCI.com](http://www.TopEditSCI.com)

母语编辑 学术团队  
论文润色, 编辑, 翻译  
写作指导 学术讲座



# 宏基因组



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

