24差异分析2-STAMP

易生信
2023年6月10日

目录

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

特征表统计分析和可视化的常用方法比较

类型	优点	缺点
STAMP	1. 跨平台，图形界面，学习成本低； 2. 实时计算，方便调整； 3. 勾选样本或分组可筛选样品； 4. 多种比较类型、统计方法可选； 5. 统计同时自动绘图； 6. 支持多达6种图形样式； 7. 本地分析，无需联网。	1. Python环境不稳定，容易报错； 2. Linux/Mac安装困难； 3. 不允许层级注释存在非严格层级； 4. 不支持中文路径； 5. 关闭后再打开才能分析新项目； 6. 功能扩展性差； 7. 图片可调参数少。
R	1. 跨平台，代码方式修改灵活； 2. 完全可重复计算； 3. 统计方法、可视化方法种类多。	1. 安装包依赖关系多； 2. 包数量过多，方法分散； 3. 学习成本高；
LEfSe	1. 整合的统计模型更适合宏基因组； 2. 流程化操作，无需统计方法选择； 3. 多层级混合分析，多组比较； 4. 出图漂亮。	1. 安装复杂，依赖关系多； 2. 需要Linux服务器或云平台； 3. 无其它方法可选； 4. 公用服务器难用。

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

http://kiwi.cs.dal.ca/Software/STAMP

STAMP

STAMP is a software package for analyzing taxonomic or metabolic profiles that promotes 'best practices' in choosing appropriate statistical techniques and reporting results. Statistical hypothesis tests for pairs of samples or groups of samples is support along with a wide range of exploratory plots. STAMP encourages the use of effect sizes and confidence intervals in assessing biological importance. A user friendly, graphical interface permits easy exploration of statistical results and generation of publication quality plots for inferring the biological relevance of features in a metagenomic profile. STAMP is open source, extensible via a plugin framework, and available for all major platforms.

Announcements

- June 26, 2015: STAMP v2.1.3 released. Minor bug fix to scatter plot to properly handle profiles contained a single feature.
- June 15, 2015: STAMP v2.1.2 released. Minor enhancements to extended error bar, heatmap, and profile bar plots.
- June 7, 2015: STAMP v2.1.1 released. Resolves issue with v2.1.0 installation.
- June 4, 2015: STAMP v2.1.0 released. Resolves the _hierarchy_wrap issue. Requires numpy >= 1.9.1, scipy >= 0.15.1, matplotlib >= 1.4.2.
- Previous announcements

Documentation

- Quick installation instructions (Microsoft Windows, Linux, Apple's Mac OS X)
- User's Guide
- Google Group
- FAQs
- Version history

Downloads

Please uninstall previous versions of STAMP before installing a new release.

- STAMP v2.1.3 setup package for Microsoft Windows (~42MB)
- Linux and OS X users can follow the instructions to install from source
- STAMP GitHub Repository
- Previous versions

Identifying biologically relevant differences between metagenomic communities

DH Parks, RG Beiko - Bioinformatics, 2010 - academic.oup.com

Motivation: Metagenomics is the study of genetic material recovered directly from environmental samples. Taxonomic and functional differences between metagenomic samples can highlight the influence of ecological factors on patterns of microbial life in a wide range of habitats. Statistical hypothesis tests can help us distinguish ecological influences from sampling artifacts, but knowledge of only the P-value from a statistical hypothesis test is insufficient to make inferences about biological relevance. Current …

☆ Save   ")) Cite   Cited by 908   Related articles   All 12 versions

Parks D H, Beiko R G. Identifying biologically relevant differences between metagenomic communities[J]. Bioinformatics, 2010, 26(6): 715-721.

STAMP: statistical analysis of taxonomic and functional profiles

DH Parks, GW Tyson, P Hugenholtz, RG Beiko - Bioinformatics, 2014 - academic.oup.com

STAMP is a graphical software package that provides statistical hypothesis tests and exploratory plots for analysing taxonomic and functional profiles. It supports tests for comparing pairs of samples or samples organized into two or more treatment groups. Effect sizes and confidence intervals are provided to allow critical assessment of the biological relevancy of test results. A user-friendly graphical interface permits easy exploration of statistical results and generation of publication-quality plots. Availability and implementation .

☆ Save   ")) Cite   Cited by 2771   Related articles   All 11 versions

Parks D H, Tyson G W, Hugenholtz P, et al. STAMP: statistical analysis of taxonomic and functional profiles[J]. Bioinformatics, 2014, 30(21): 3123-3124.

Donovan Parks

Bioinformatic Consultant
Verified email at uq.edu.au

bioinformatics   metagenomics   biogeography   machine learning
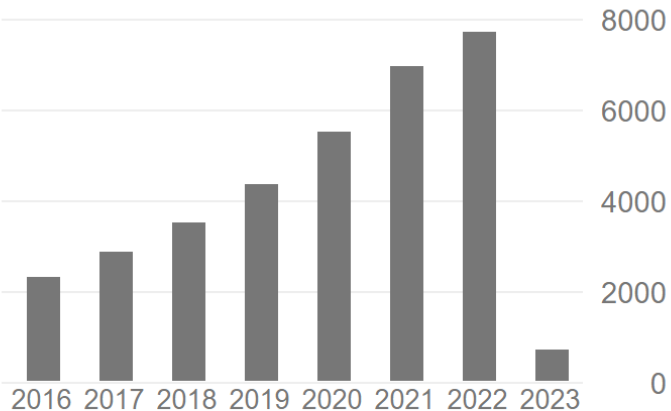information visualization

https://scholar.google.com/citations?user=A87iE7wAAAAJ

TITLE	CITED BY	YEAR
Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities PD Schloss, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, ... Applied and environmental microbiology 75 (23), 7537-7541	18461	2009
CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes DH Parks, M Imelfort, CT Skennerton, P Hugenholtz, GW Tyson Genome research 25 (7), 1043-1055	5397	2015
STAMP: statistical analysis of taxonomic and functional profiles DH Parks, GW Tyson, P Hugenholtz, RG Beiko Bioinformatics 30 (21), 3123-3124	2771	2014
A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life GTDB-细菌基因组分类数据库 DH Parks, M Chuvochina, DW Waite, C Rinke, A Skarshewski, ... Nature biotechnology 36 (10), 996-1004	2028	2018
GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database PA Chaumeil, AJ Mussig, P Hugenholtz, DH Parks Bioinformatics	1714	2019

Cited by   VIEW ALL

	All	Since 2018
Citations	40012	28873
h-index	41	37
i10-index	56	52

2016 2017 2018 2019 2020 2021 2022 2023

Public access   VIEW ALL

1 article   37 articles

not available   available

Based on funding mandates

EnBio Gene Technology (Beijing) co., LTD

5

散点(PCA，scatter)、箱线、(扩展)柱状、热图(丑且易报错)

输入文件：特征表 + 实验设计(分组信息)

○ 特征表可以是门、纲、目、科、属、OTU/ASV表，以及多层级组合(16S物种名层级不严格，不可用)，包括目录中tax_*.txt文件。

Class	KO1	KO2	KO3	KO4	KO5	KO6
k__Archaea|p__Crenarchaeota|c__Thaumarchaeota|c	0.00319	0.00911	0	0	0	0
k__Bacteria|p__[Thermi]|c__Deinococci|o__Deinoc	0	0	0	0	0	0
k__Bacteria|p__[Thermi]|c__Deinococci|o__Deinoc	0	0.00304	0	0	0	0
k__Bacteria|p__|c__|o__|f__	0	0	0	0	0	0
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0	0	0	0	0	0
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0	0	0	0	0.00269	0
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0	0	0	0	0	0
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0	0.00304	0	0	0.00539	0.00282
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0.00319	0.00304	0.00264	0.00281	0	0.00564
k__Bacteria|p__Acidobacteria|c__[Chloracidobact	0.00956	0.01215	0.00528	0	0.00539	0.02256
k__Bacteria|p__Acidobacteria|c__|o__|f__	0.00319	0	0	0	0	0.00282
k__Bacteria|p__Acidobacteria|c__Acidobacteria-5	0.01594	0.01215	0.00528	0.00844	0.01616	0.03102
k__Bacteria|p__Acidobacteria|c__Acidobacteria-6	0	0	0	0	0	0
k__Bacteria|p__Acidobacteria|c__Acidobacteria-6	0	0.00304	0	0	0	0.00282
k__Bacteria|p__Acidobacteria|c__Acidobacteria-6	0.02231	0.05771	0.02638	0.01407	0.02154	0.0282

特征表

SampleID	group	genotype	site
KO1	A	KO	Beijing
KO2	A	KO	Beijing
KO3	A	KO	Sanya
KO4	A	KO	Sanya
KO5	A	KO	Harbin
KO6	A	KO	Harbin
OE1	B	OE	Beijing
OE2	B	OE	Beijing
OE3	B	OE	Sanya
OE4	B	OE	Sanya
OE5	B	OE	Harbin
OE6	B	OE	Harbin

分组信息/元数据Metadata
(metadata.txt)

○ 特征表(OTU/ASV)：result/otutab.txt

○ 门水平汇总表：result/tax/sum_p.txt

○ ……

○ 属水平汇总表：result/tax/sum_g.txt

○ 存在问题：特征表没有筛选时特征数量较大，低丰度不准确结果容易出现假阳性、高丰度差异结果经多重比较校正造成假阴性。分类级汇总表缺少上级分类学信息，无法描述其归属。

STAMP助手输入文件准备助手 format2stamp.R 一键生成STAMP 要求输入文件

- 准备特征表(otutab.txt)、物种注释(taxonomy.txt)

- 2.1 命令行生成输入文件

- 注意输出输出文件位置

  ```
   mkdir -p result/stamp
      Rscript ${db}/script/format2stamp.R --input result/otutab.txt \
       --taxonomy result/taxonomy.txt --threshold 0.01 \
       --output result/stamp/tax
  ```

- 输出结果位于result/stamp目录中，有tax_1-8的结果，1-7为界、门、纲、目、科、属和种水平分类汇总，8为筛选OTU表，如0.01代表筛选丰度均值>万分之一。

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

○ result目录中准备otutab.txt和taxonomy.txt文件；

○ RStudio中打开format2stamp.Rmd，结果默认为result/stamp目录中

○ 检查输入输出文件名和位置、OTU/ASV相对丰度筛选阈值

○ 点击Knit按钮运行程序

R format2stamp.Rmd

📄 otutab.txt

📄 taxonomy.txt

输入文件

输出文件

1-7为界、门、纲、目、科、属和种水平分类汇总，

8为筛选OTU表，如0.01代表筛选丰度均值>万分之一

📄 format2stamp.html

📄 tax_1Kingdom.txt

📄 tax_2Phylum.txt

📄 tax_3Class.txt

📄 tax_4Order.txt

📄 tax_5Family.txt

📄 tax_6Genus.txt

📄 tax_7Species.txt

📄 tax_8OTU0.01.txt

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

1. STAMP安装和运行

○ Windows用户安装程序在 public\win\STAMP2.1.3 目录中；

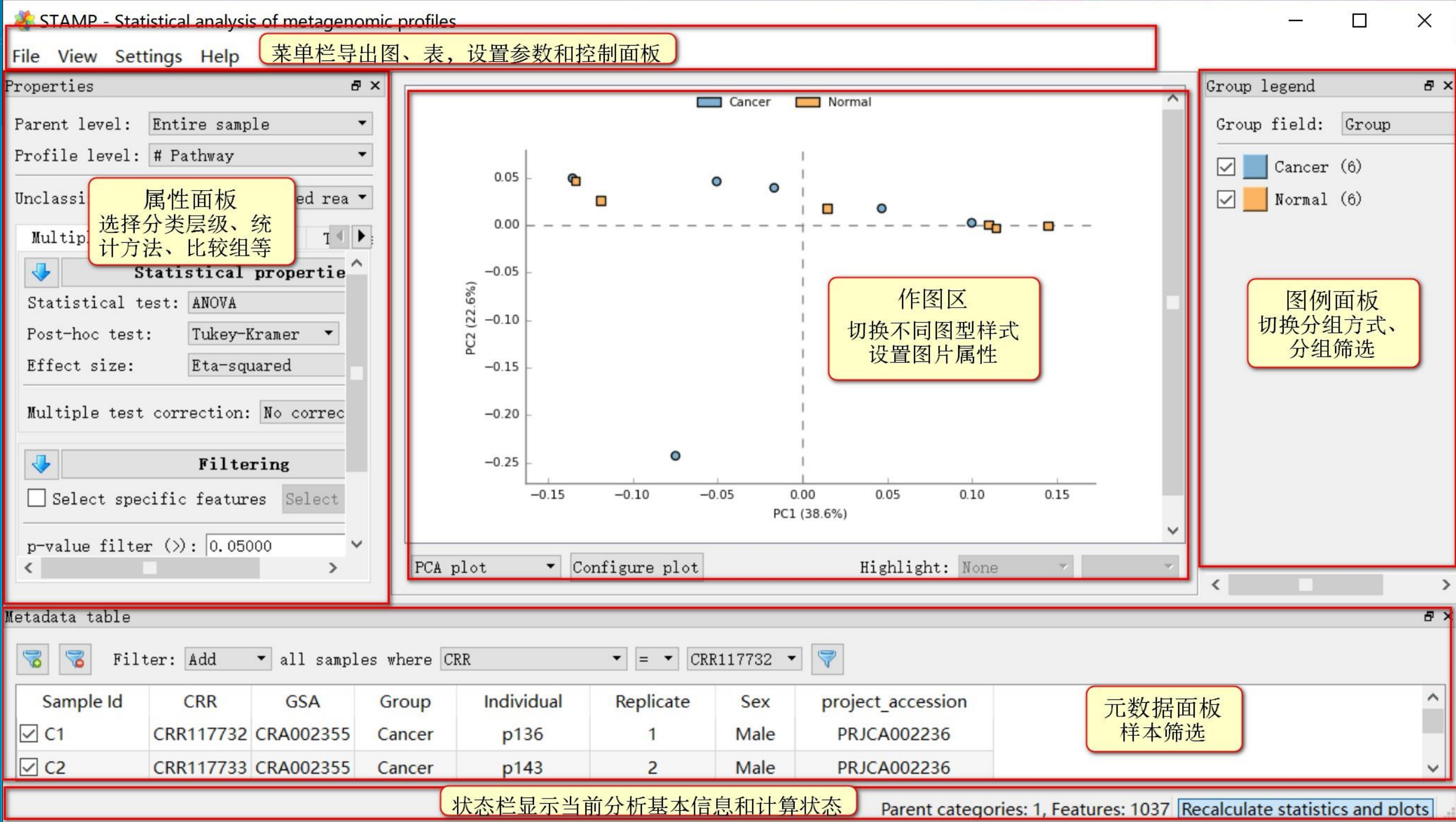○ 按要求一路安装即可(最好不要修改程序默认安装位置)；

○ 安装后程序名为STAMP的花瓣图标(有时图标为白纸)，双击运行；

易生信

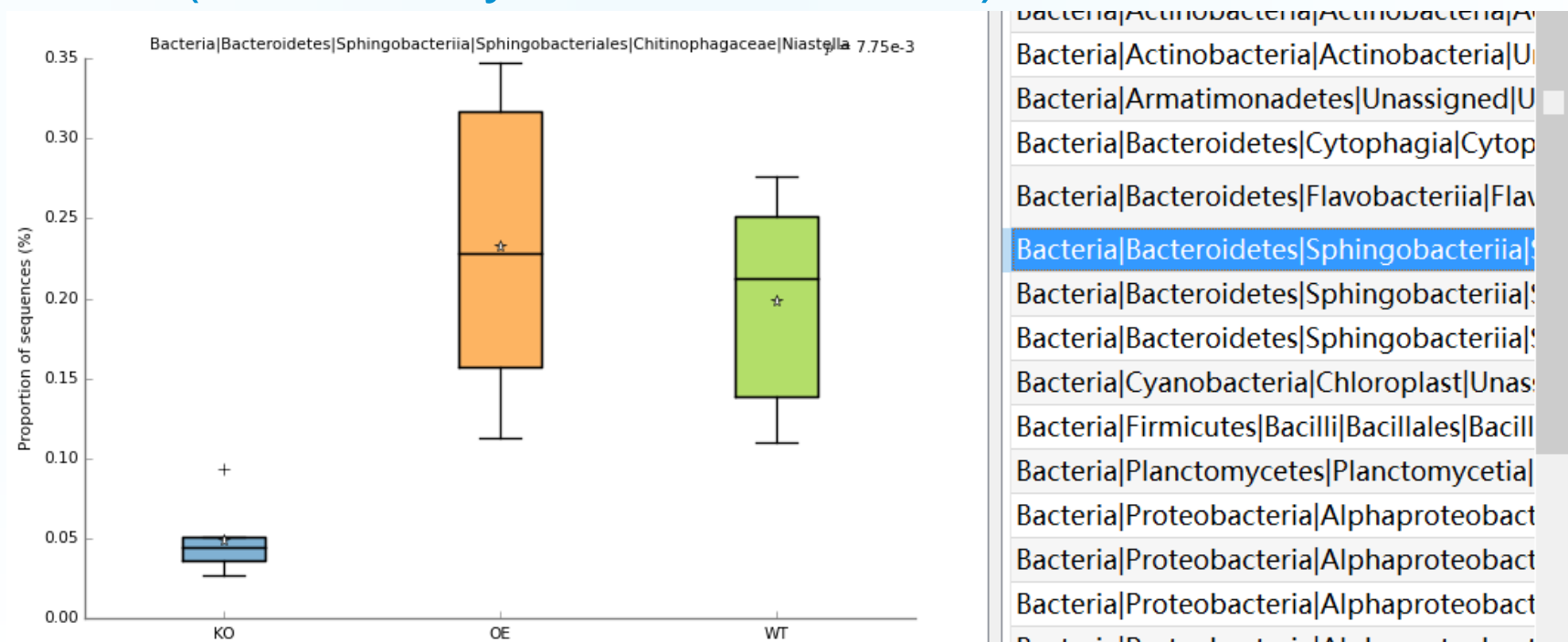- File – Load data ——以属genus为例

- 选择 丰度组成 和 元数据 文件, OK

- 开始随机选组进行PCA分析

- 右侧Group field中切换分组

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD
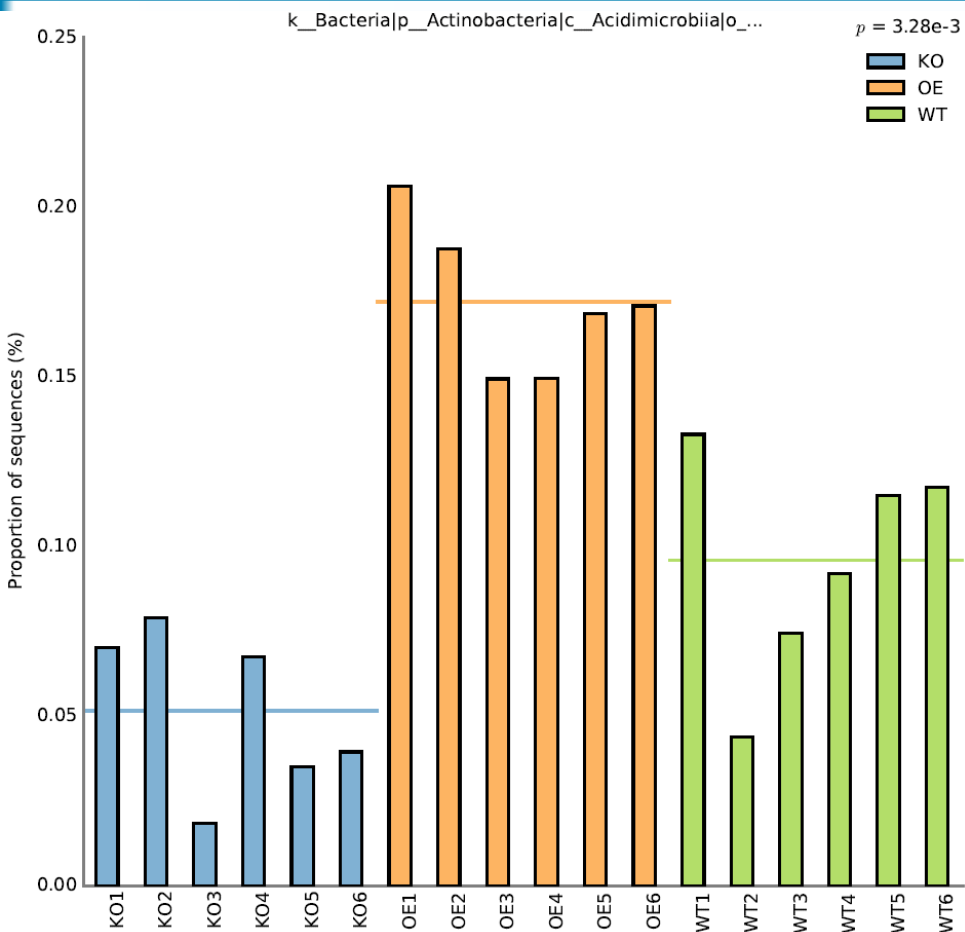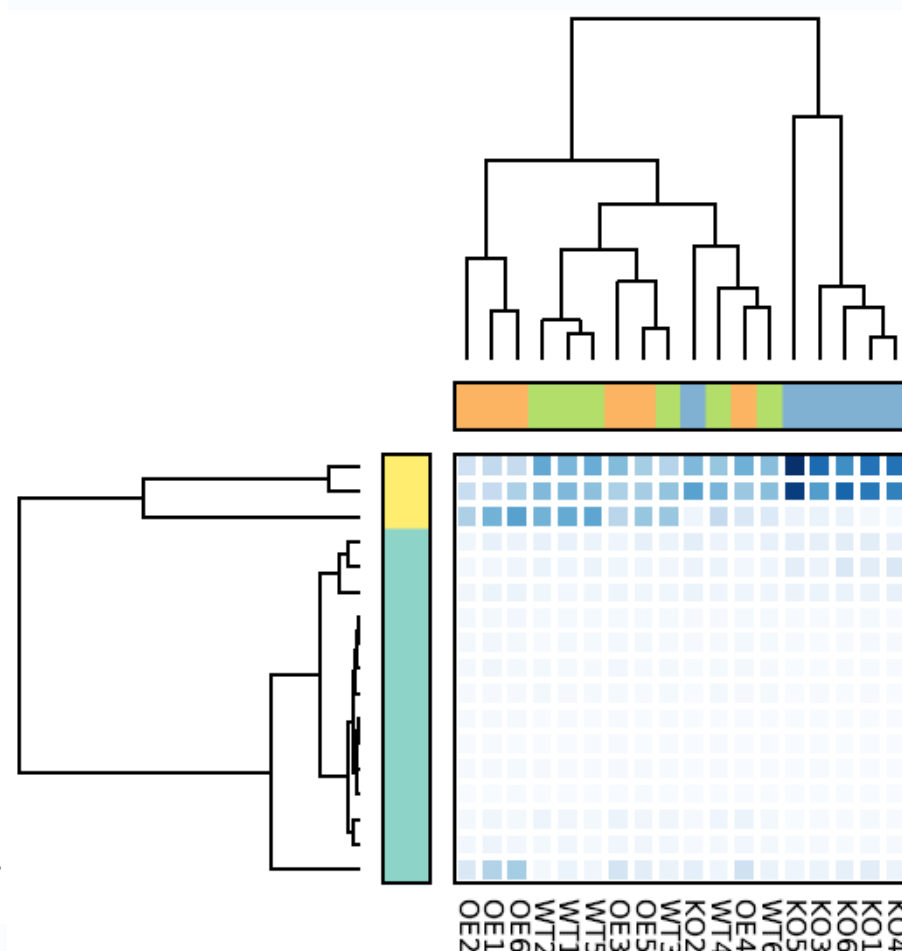
- 左侧多组比较(Multiple groups) – 校正多组检验(Multiple test correction) – None修改为Storey FDR(结果太多改为BH FDR)

- 图片下方类型的主坐标轴分析(PCA plot)改为 箱线图(Box plot)，右侧勾选只显示差异(Show only active features)，选择具体条目展示如下

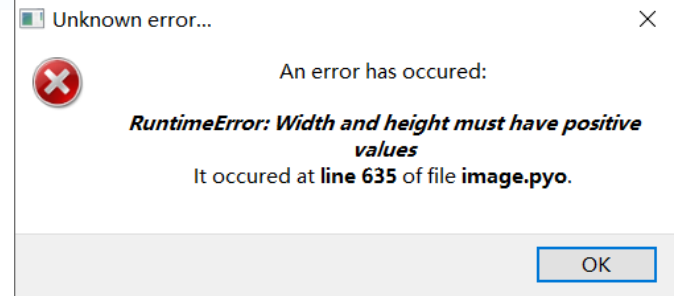柱状图(Bar plot)

热图(Heatmap)，不好看，容易报错，不推荐使用
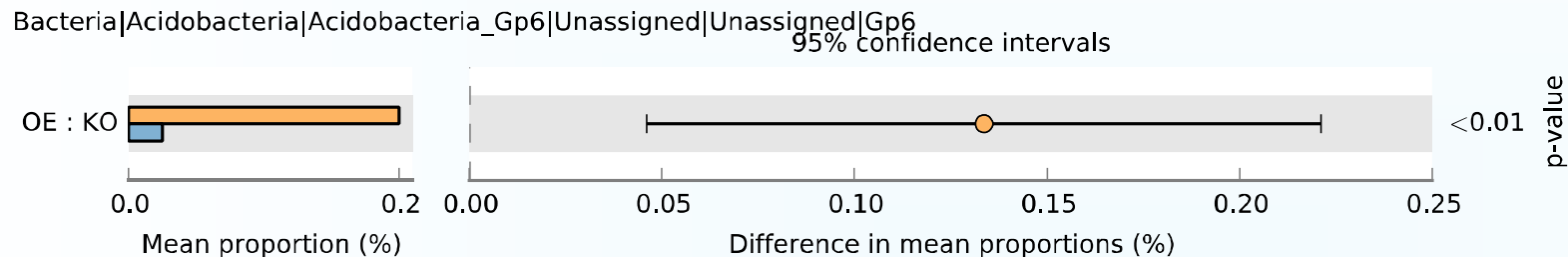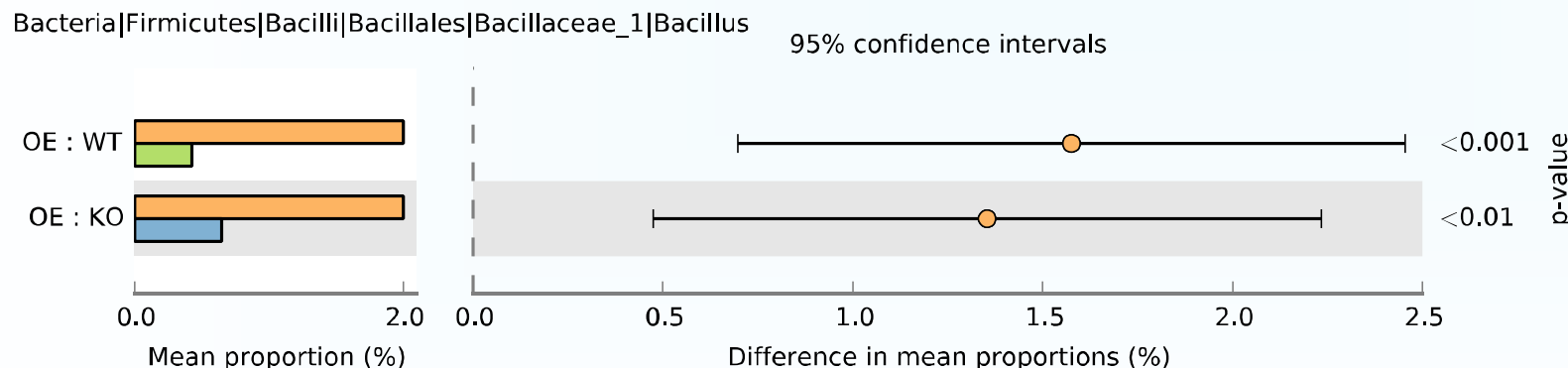(报错点OK，关闭程序后打开重新分析)

○ 三组中仅有一对显著差异

Bacteria|Acidobacteria|Acidobacteria_Gp6|Unassigned|Unassigned|Gp6

○ 三组中有两对显著差异

Bacteria|Firmicutes|Bacilli|Bacillales|Bacillaceae_1|Bacillus

○ 三组中两两均显著差异-本数据中没有

操作：选择Two groups – Group1选择KO, Group2选择OE – 统计方法默认 t-test，推荐选择Welch's t-test，各方法优缺点如下：

| 统计假设方法	描述
t检验	T检验，亦称学生t检验(Student's t test)，假设两组有相同的方差，当假设成立时，它比Welch's检验更强，**主要用于样本量较小(例如 $n < 30$)，总体标准差σ未知的正态分布**。当$n > 3$时有较高的**准确度**和精确度。
Welch's t-test	t-test的一种变形，用于当两组**无法满足方差相同**的假设时使用。
White's无参t-test	无参数的检验，由White为临床宏基因组数据分析提出。此方法使用排序过程移除标准t-test的正态假设。此外，它使用启发式鉴定松散的特征，可采用Fisher精确检验和pooling的策略，**适合组样本一致，或小于8个样品**。大数据集计算**极耗时**。

置信区间方法	描述
DP: t-test inverted	只有当方差相等的t检验可用。
Scheffe	考虑所有可能的比较，而Tukey-Kramer只考虑成对均值。此种方法较保守
DP: Welch's inverted	为Welch's t检验提供置信区间。
DP: bootstrap	适合White's 无参t-test

多种检验校正方法	描述
Benjamini-Hochberg FDR	控制假阳性率FDR
Bonferroni	控制整体错误率的经典方法，被批评太保守
Sidak	在整体错误率控制中使用不多，但均匀分布数据上比Bonferroni更强，但需要假设个体检验是独立的
Storey's FDR	控制FDR的新方法，比BH更强，需要更多计算资源。

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

强度顺序：Storey's FDR < Benjamini-Hochberg FDR < Bonferroni
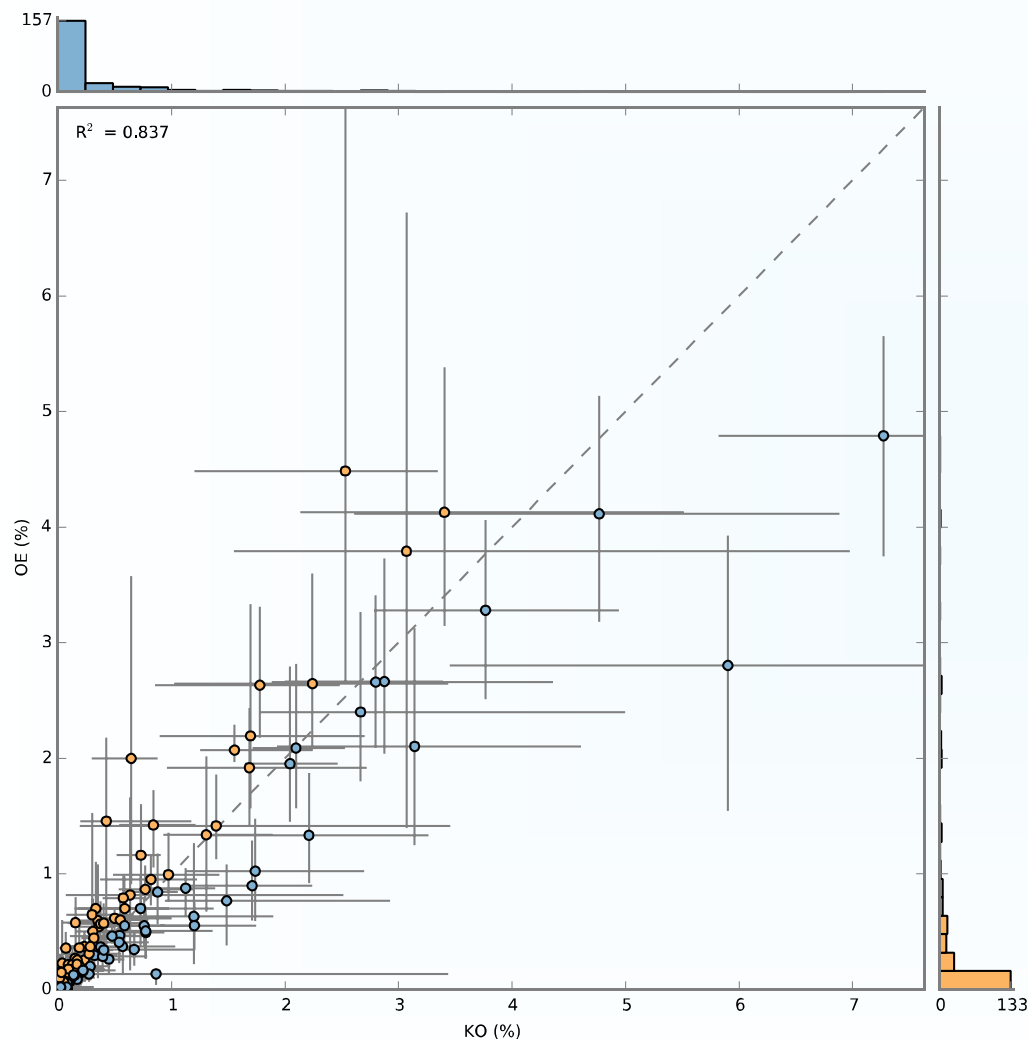
散点图(Scatter plot)——相关分析
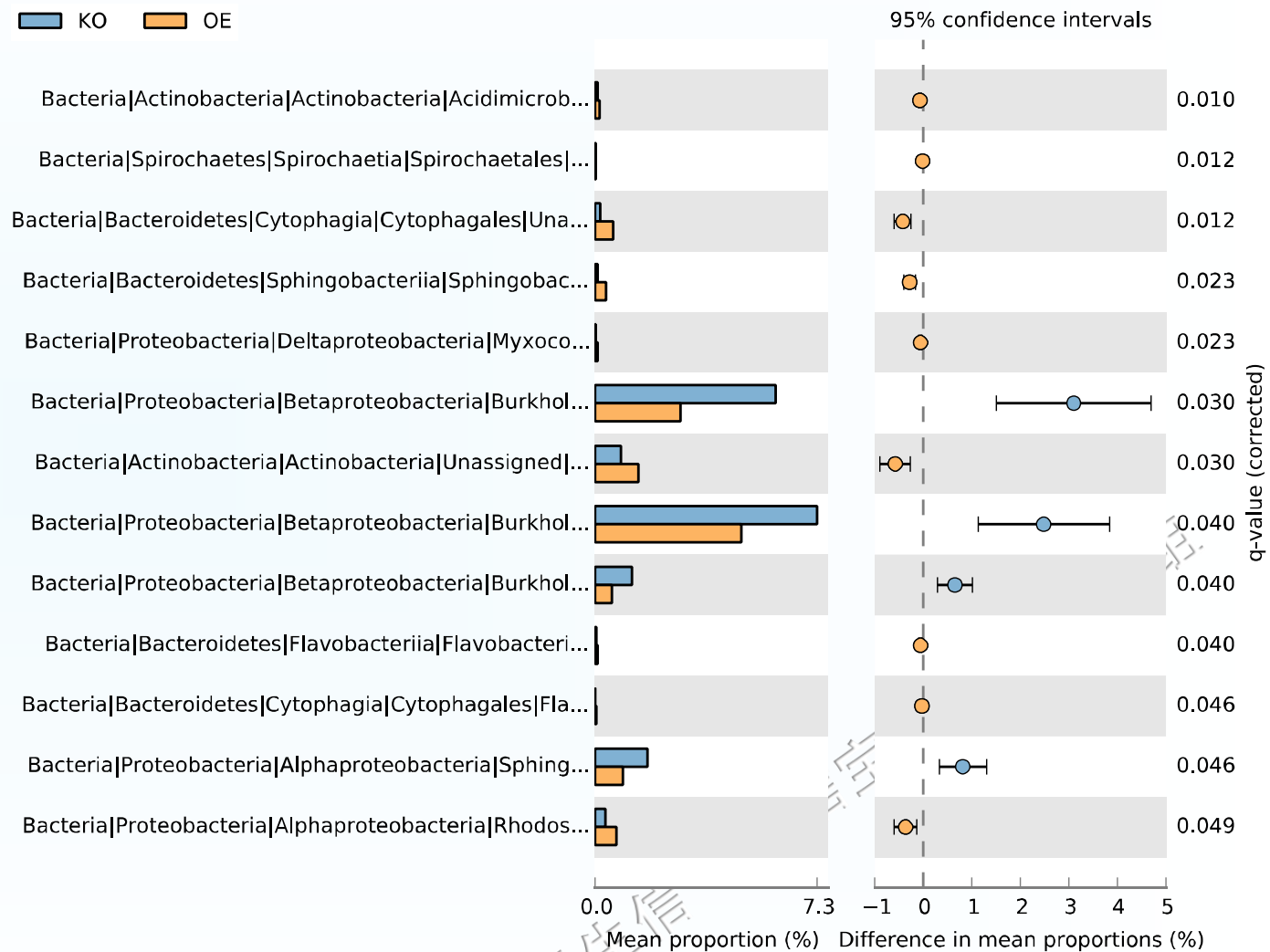
扩展柱状图(Extended error bar plot)——所有差异

```
compare="KO-WT"
Rscript ${db}/script/compare_stamp.R \
  --input result/otutab.txt --metadata result/metadata.txt \
  --group Group --compare ${compare} --threshold 0.1 \
  --method "t.test" --pvalue 0.05 --fdr "none" \
  --width 189 --height 159 \
  --output result/stamp/${compare}
```
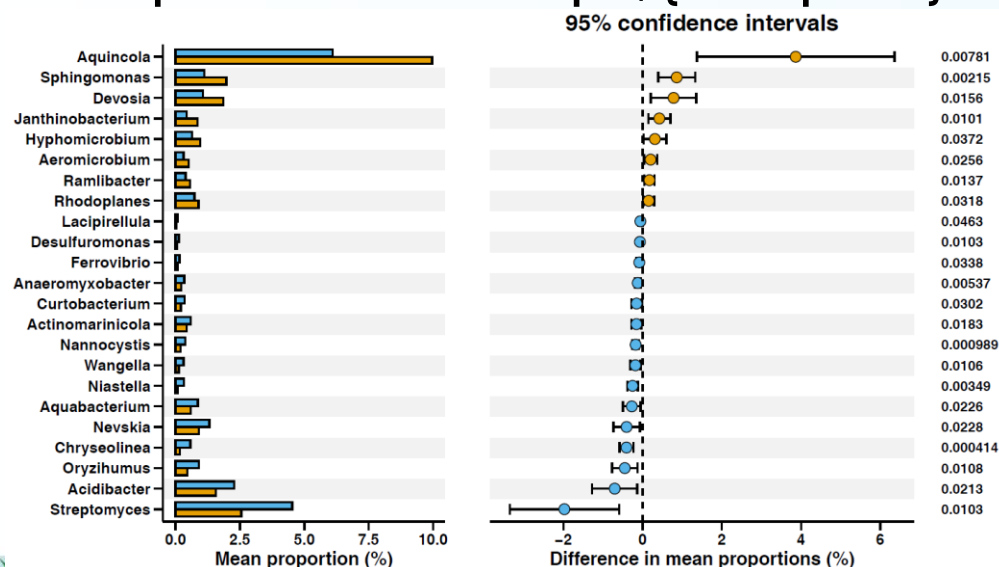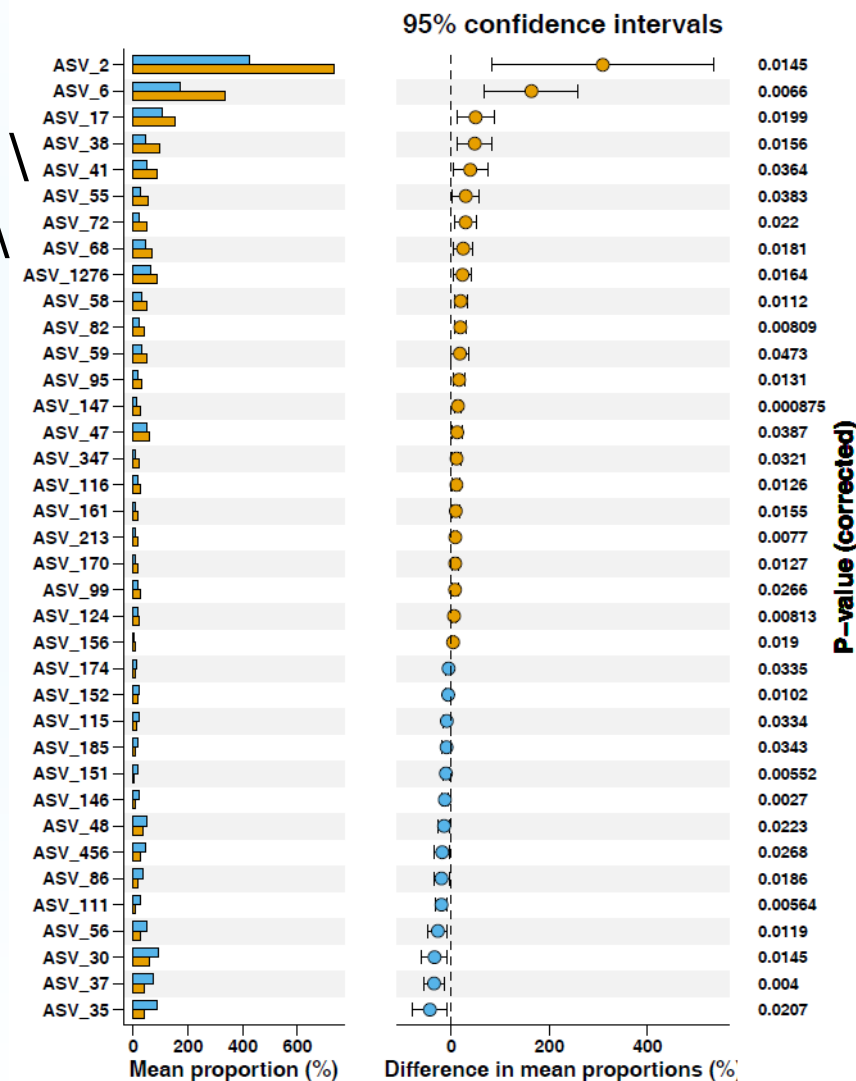

STAMP矢量图PDF文字仅为点线，不可修改字体，编辑需要重打；R语言输出图在AI中方便编辑文字。

•R语言完美重现STAMP结果图

- 图片在下方Configure plot中可设置图片大小、列宽、图例位置等；菜单Setting – Preferences中调置标签显示长度

- 注意bar / box / heatmap 中Field to plot选择百分比或原始数据

- 保存图片：File – Save plot (Ctrl + S)，保存类型默认png，推荐修改为pdf(矢量图更清晰、细节可修改)

- 保存两组比较统计表：View – Two group statistics table (Ctrl + G)，可勾选Show only active features只保存显著结果，格式有tsv/txt

- 此外还有多组、两样品等统计，结果保存同两组类似

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

- STAMP安装要用默认目录，不要修改到含有中文的目录

- 输入文件为丰度矩阵表+分组信息，路径不要有中文

- 多组比较推荐使用PCA，ANOVA +FDR+ 单Feature boxplot

- 两组比较推荐使用Welch's t-test + Extened errorbar / heatmap

- 导出图片为PDF格式，即高清，又方便AI修改编辑；表格为tsv格式，方便Excel打开编辑

- 最终的统计方法由具体项目和实际结果决定

- 统计是辅助你的帮手，不是指挥你的领导；注意FDR校正的原理

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

进一步阅读

- **宏基因组公众号文章目录 生信宝典公众号文章目录**

- **科学出版社《微生物组数据分析与可视化实战》——30+篇**

- **Bio-protocol《微生物组实验手册》计划——200+篇**

- **Protein Cell：扩增子和宏基因组数据分析实用指南**

- **CMJ：人类微生物组研究设计、样本采集和生物信息分析指南**

- **扩增子图表解读 分析流程 统计绘图**

- **QIIME2中文教程-把握分析趋势**

- **扩增子16S分析专题研讨论会——背景介绍**

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD

扫码关注生信宝典，学习更多生信知识

扫码关注宏基因组，获取专业学习资料

易生信，没有难学的生信知识

易汉博基因科技(北京)有限公司
EHBIO Gene Technology (Beijing) co., LTD