



31功能预测PICRUST1/2

易生信
2023年6月11日



- PICRUST简介
- 原理与应用
- 本地+在线操作
- 进一步学习+PICRUST2

易生信 生信宝典 宏基因组



- **PICRUST简介**
- 原理与应用
- 本地+在线操作
- 进一步学习+PICRUST2

易生信 生信宝典 宏基因组



如何将物种的“身份”和“功能”对应起来？

- 16S扩增子测序较高的性价比揭示菌群的具体物种组成，从而解答“群落中有谁？”的基本问题。然而很多时候，我们更希望知道菌群行使的具体功能，也就是解释“它们在干什么？”
- 2013年一款名为PICRUSt的菌群代谢功能预测利器应运而生，相关论文发表在顶级期刊《*Nature Biotechnology*》上。

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences

[MGI Langille](#), [J Zaneveld](#), [JG Caporaso](#)... - *Nature* ..., 2013 - [nature.com](#)

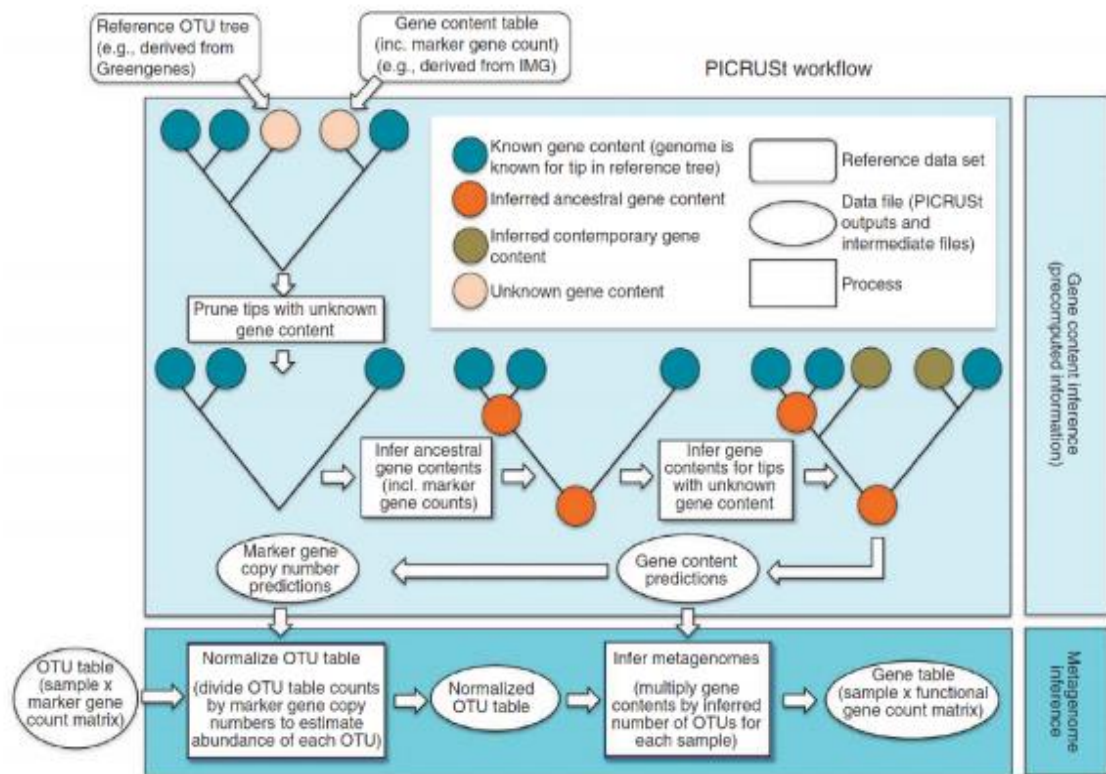
Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. Here we describe PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict ...

☆ Save ㉞ Cite Cited by 7254 Related articles All 23 versions



► PICRUST:

Phylogenetic Investigation of Communities by Reconstruction of Unobserved States



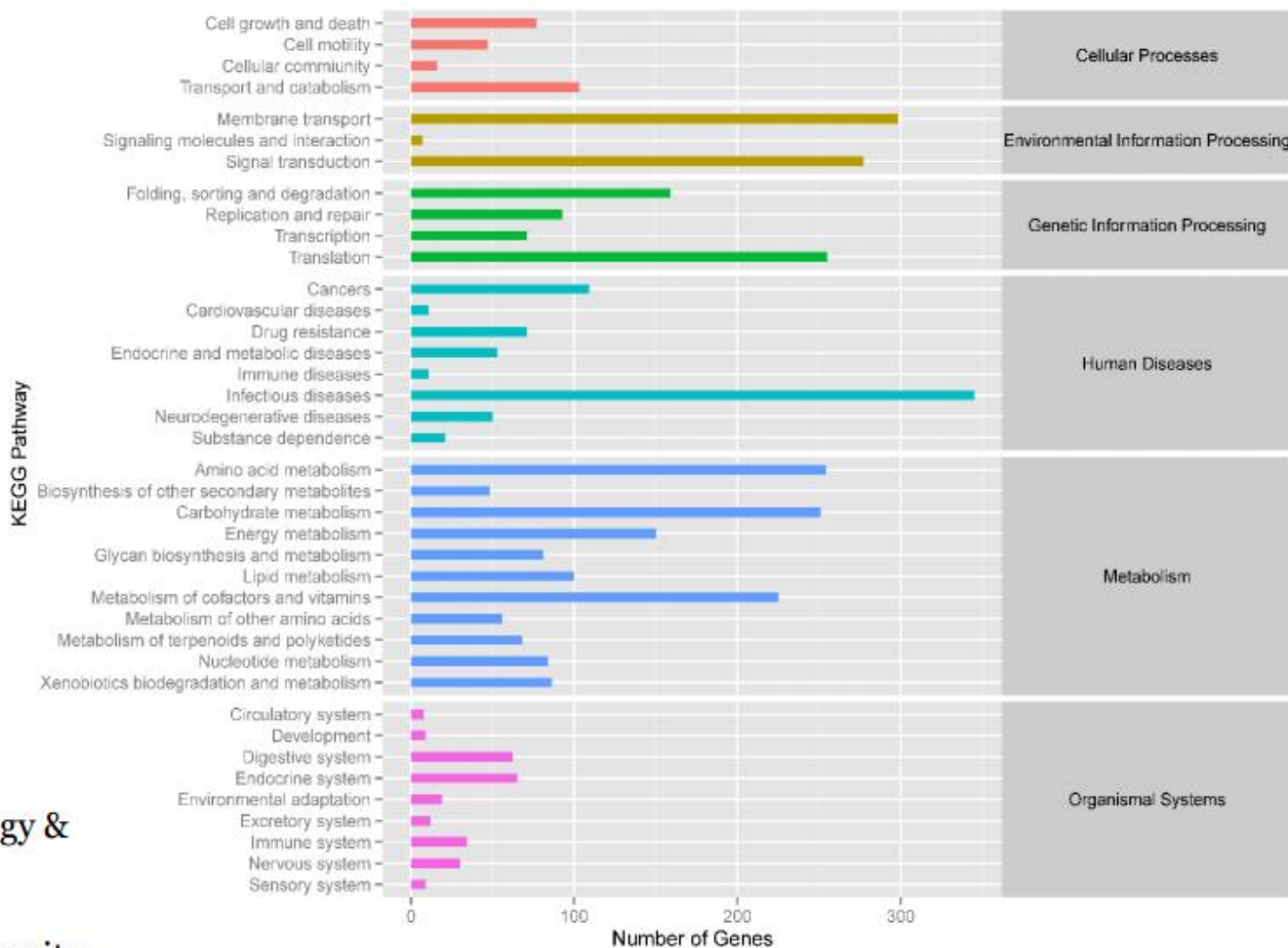
(Langille *et al.*, 2013, Nature Biotechnology)



Curtis Huttenhower

Associate Prof. of Computational Biology & Bioinformatics,
Department of Biostatistics,
School of Public Health, Harvard University

http://huttenhower.sph.harvard.edu/galaxy/tool_runner?tool_id=PICRUST_normalize



- PICRUST简介
- 原理与应用
- 本地+在线操作
- 进一步学习

易生信 生信宝典 宏基因组



PICRUSt的总体思路

- 先根据已测微生物基因组的16S rRNA基因全长序列，推断它们的共同祖先的基因功能谱；
- 对GreenGenes数据库中其它未测物种的基因功能谱进行推断，构建古菌和细菌域全谱系的基因功能预测谱；
- 最后，将测序得到的菌群组成“映射”到数据库中，对菌群代谢功能进行预测。

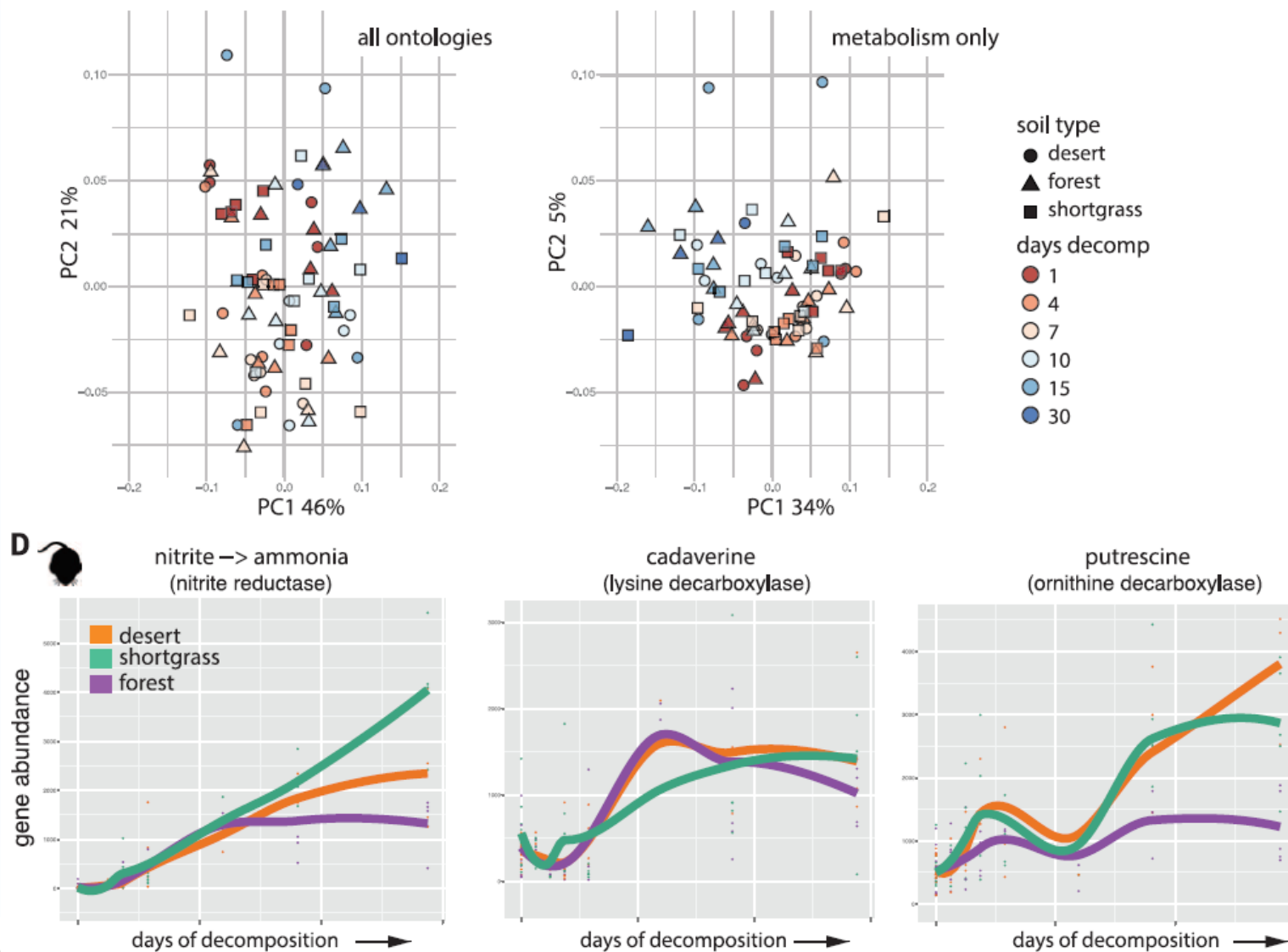


具体步骤如何实现

- 对测序获得的16S rRNA基因序列，进行“封闭式”参考OTU划分 (**Closed-reference OTU picking**)，通过与Greengenes数据库比对，寻找每一条测序序列的“参考序列最近邻居”，并归为参考OTU；
- 根据“参考序列最近邻居”的rRNA基因拷贝数，对获得的OTU丰度矩阵进行校正；
- 根据“参考序列最近邻居”对应的KEGG/EggNOG等基因功能谱数据，换算预测菌群的整体代谢功能。



案例1——尸体分解者代谢基因随时间变化

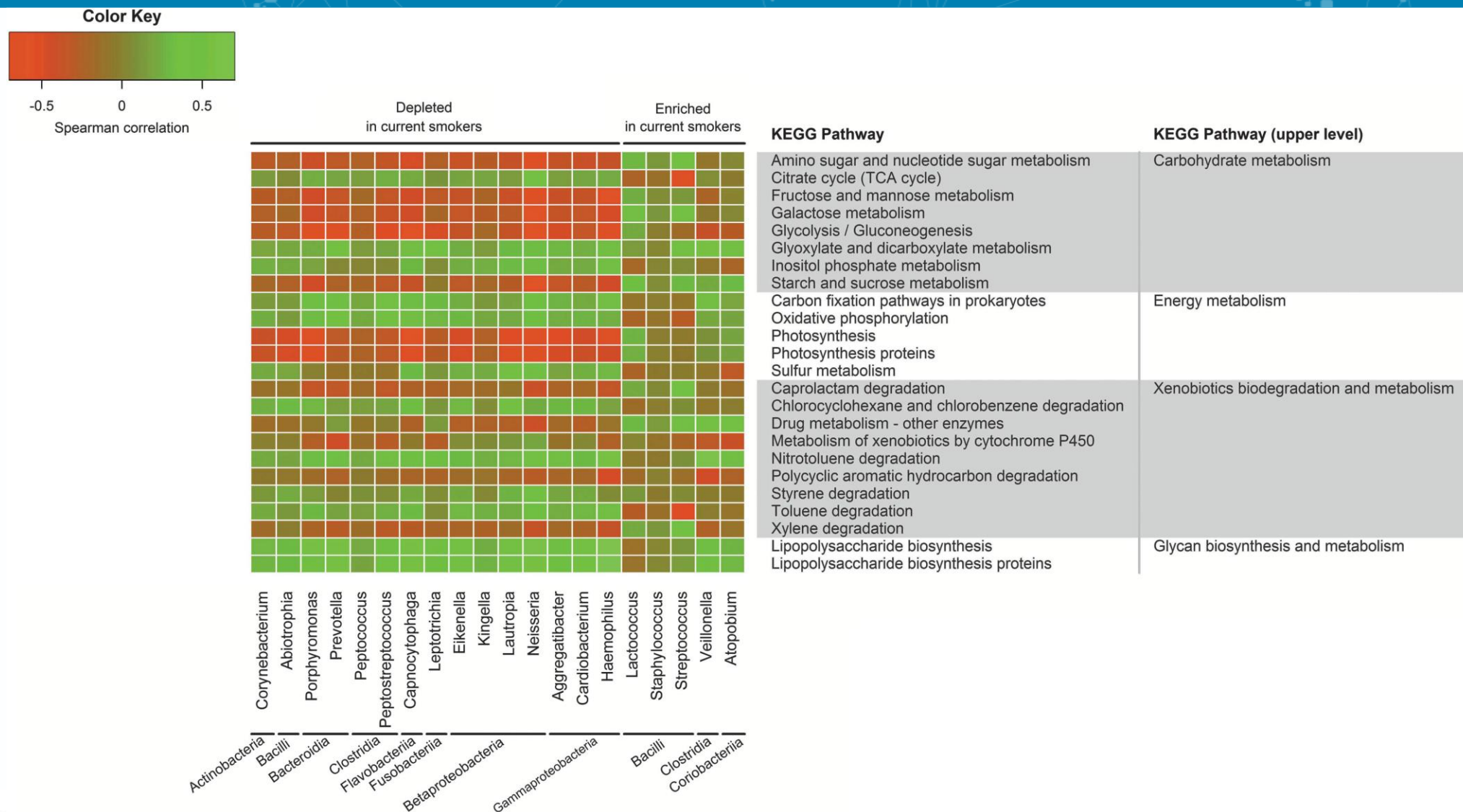


宏基因组

信宝興



案例2——吸烟者口腔差异菌群与关联相关功能



宏基因组



- PICRUST简介
- 原理与应用
- 本地+在线操作
- 进一步学习

易生信 生信宝典 宏基因组



文件准备：比对greengene数据库生成OTU表

- 代码参考：pipeline.sh —— ## 9. 有参比对
- Greengene数据库是最旧、最准，支持最广泛的数据库，基于它的OTUs表可进行功能预测(PICRUSt)和形态学预测(Bugbase)
- 生成OTU表 Create OTUs table

```
usearch -otutab temp/filtered.fa -otus ${db}/gg/97_otus.fasta \
-otutabout result/gg/otutab.txt -threads 4
```

#OTU	ID	K01	K02	K03	K04	K05	K06	OE1	OE2	OE3	OE4	OE5	OE6	WT1	WT2	WT3	WT4	WT5	WT6
1000161	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
100054	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	1	2	0	1
1000757	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1001013	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100104	7	0	0	1	0	2	1	2	3	2	2	2	8	1	5	0	1	0	1
1001967	0	1	1	0	0	1	9	2	0	0	0	2	2	3	2	2	0	0	0
100208	1	0	0	0	0	0	0	2	1	0	0	2	4	0	1	2	0	6	3
100307	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1003344	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1004427	2	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0
1004452	0	3	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1

方法1. ImageGP —— PICRUST一键分析

Input data

* Input data type *i*

☒ Paste data ☐ Select uploaded file

* Paste data to text area *i*

```
634665 0 0 0 0 0 0 0 0 0 0
614253 0 0 0 0 0 0 0 0 0 0
831178 0 0 0 0 0 0 0 0 0 0
749133 0 0 0 0 0 0 0 0 0 0
787185 0 0 0 0 0 0 0 0 0 0
800713 0 0 0 0 0 0 0 0 0 0
```

注意修改 “#OTU ID” 为 “OTUID”

Check Data *i*

Check Data

Essential parameters

Type of functional predictions *i*

☒ KEGG Orthologs ☐ COG ☐ RFAM

Output weighted NSTI *i*

☐ Yes ☒ No

Pathway Hierarchy Level *i*

1 2 3

Result file format *i*

☐ TXT ☐ BIOM ☒ TXT & BIOM

Submit



1 PICRUSt Result

1.1 OTU Biom Table

1.2 NSTI Sample Summary Result

1.3 PICRUSt Predict Function Result Based on KEGG

1.4 PICRUSt Predict Function Result in Function Level 1

1.4.1 PICRUSt Predict Function Result in Function Level 1 (ko)

1.5 PICRUSt Predict Function Result in Function Level 2

1.5.1 PICRUSt Predict Function Result in Function Level 2 (ko)

1.6 PICRUSt Predict Function Result in Function Level 3

Result Table Based on KEGG (stamp)

Result Table Based on KEGG (ko)

Result Table Based on KEGG (ko.L3.stamp)

Result Table Based on KEGG (ko.L3)

Result Table Based on KEGG (ko.L2)

Result Table Based on KEGG (ko.L1)

Result Table Based on KEGG (ko.L1.stamp)

Result Table Based on KEGG (ko.L2.stamp)

共4个级别，每个级别有原始
和stamp格式结果

方法2. 本地操作(限Linux系统, 安装picrust环境)

- 需conda install picrust 安装软件环境
- # 转换为OTU表通用格式, 方便下游分析和统计

```
biom convert -i otutab.txt -o otutab.biom --table-type="OTU table" --to-json
```
- # 校正拷贝数

```
normalize_by_copy_number.py -i otutab.biom -o otutab_norm.biom -c /db/picrust/16S_13_5_precalculated.tab.gz
```
- # 预测宏基因组KO表, biom方便下游归类, txt方便查看分析

```
predict_metagenomes.py -i otutab_norm.biom -o ko.biom -c /db/picrust/ko_13_5_precalculated.tab.gz
```

```
predict_metagenomes.py -f -i otutab_norm.biom -o ko.txt -c /db/picrust/ko_13_5_precalculated.tab.gz
```

[1KO通路PICRUST](#)



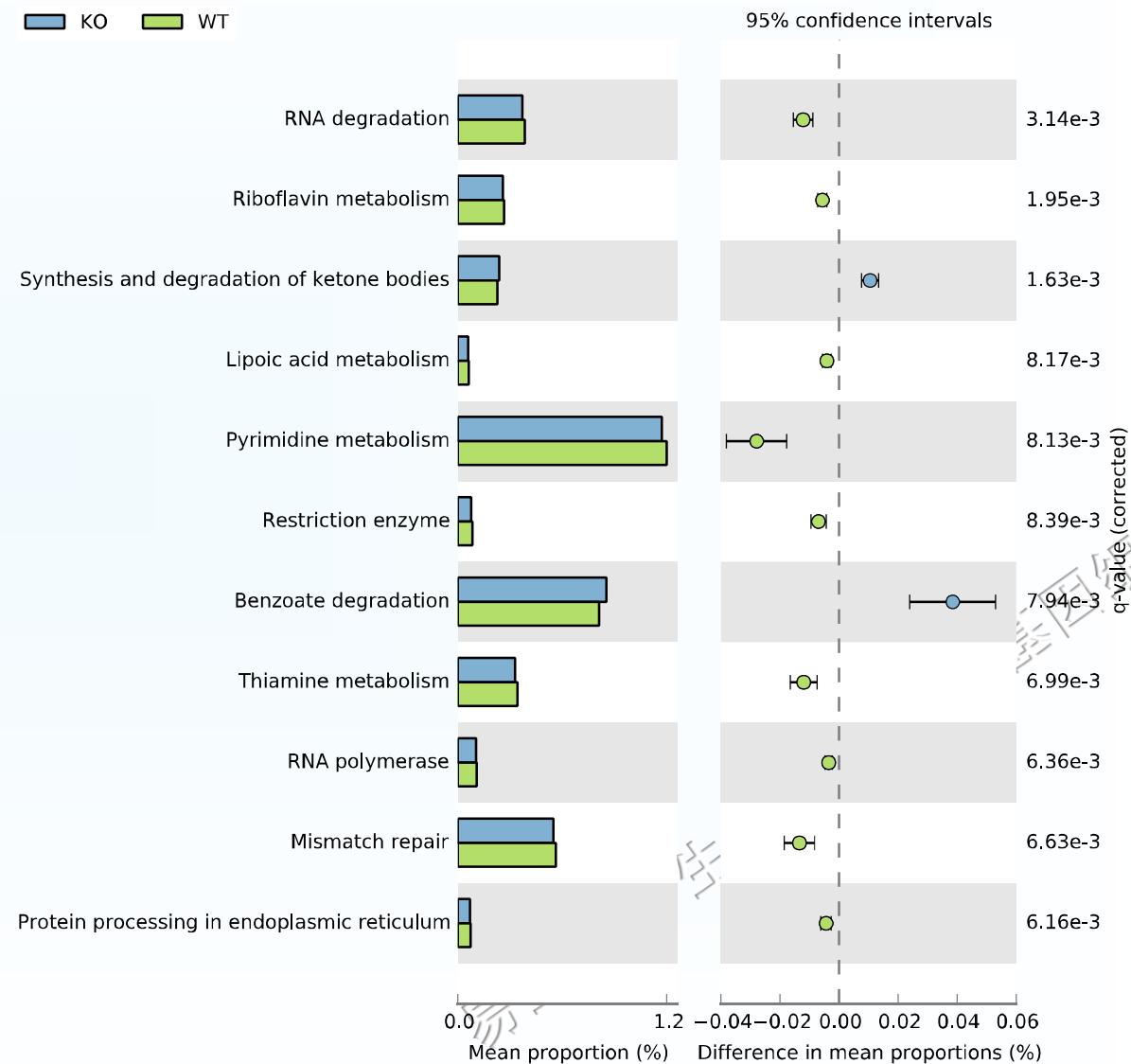
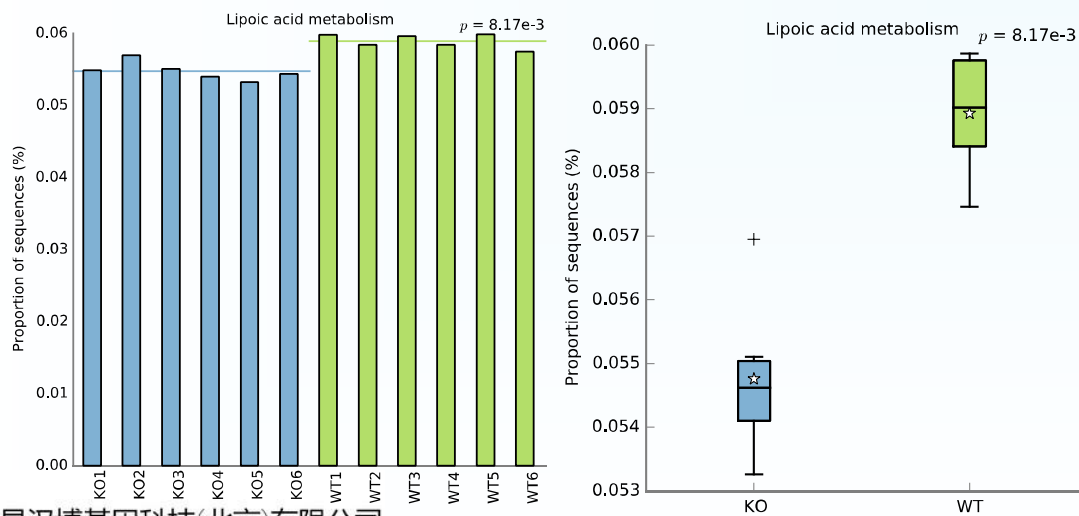
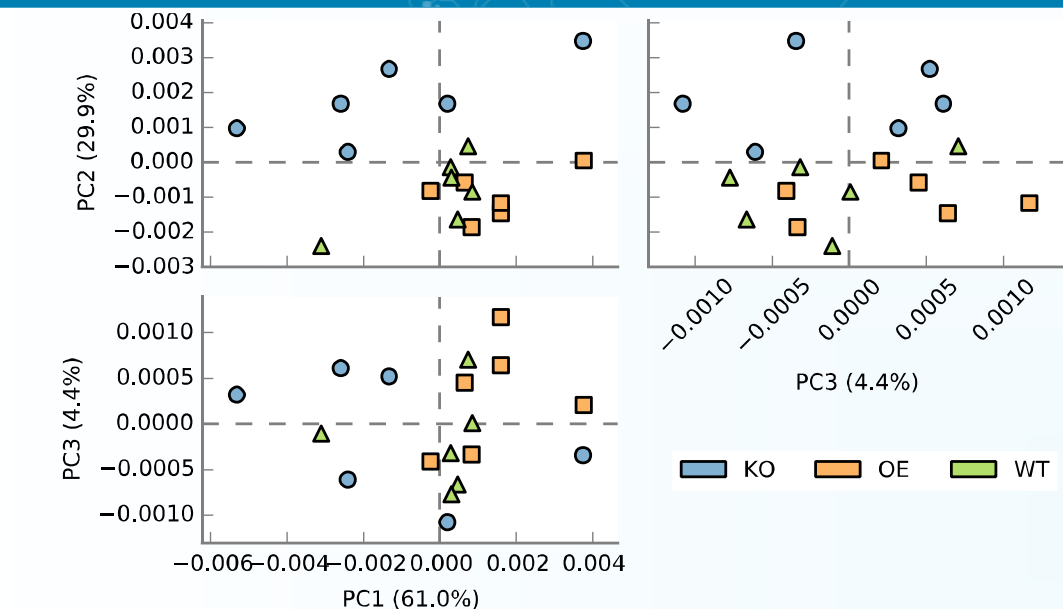
KEEG层级汇总，生成STAMP输入文件

- # 按功能级别分类汇总，-c输出KEGG_Pathways，分1-3级

```
sed -i '/# Constr/d;s/#OTU //' ko.txt
num=`tail -n1 ko.txt|wc -w`
for i in 1 2 3;do
    categorize_by_function.py -f -i gg/ko.biom -c KEGG_Pathways -l ${i} -o
    gg/ko${i}.txt
    sed -i '/# Constr/d;s/#OTU //' gg/ko${i}.txt
    paste <(cut -f $num gg/ko${i}.txt) <(cut -f 1-[$num-1] gg/ko${i}.txt) >
    gg/ko${i}.spf
done
```

- # ko1只有8类, ko2有41个类，推荐ko3级别328类别统计

STAMP统计分析*L3.stamp.txt



·R语言完美重现STAMP结果图

KEGG层级功能柱状图

差异比较pathway

compare="KO-WT"

Rscript \${db}/script/compare.R \

--input result/picrust/\${l}.mat.txt --design result/metada

--group Group --compare \${compare} --threshold 0 \

--method wilcox --pvalue 0.05 --fdr 0.2 \

--output result/picrust/

整体可视化

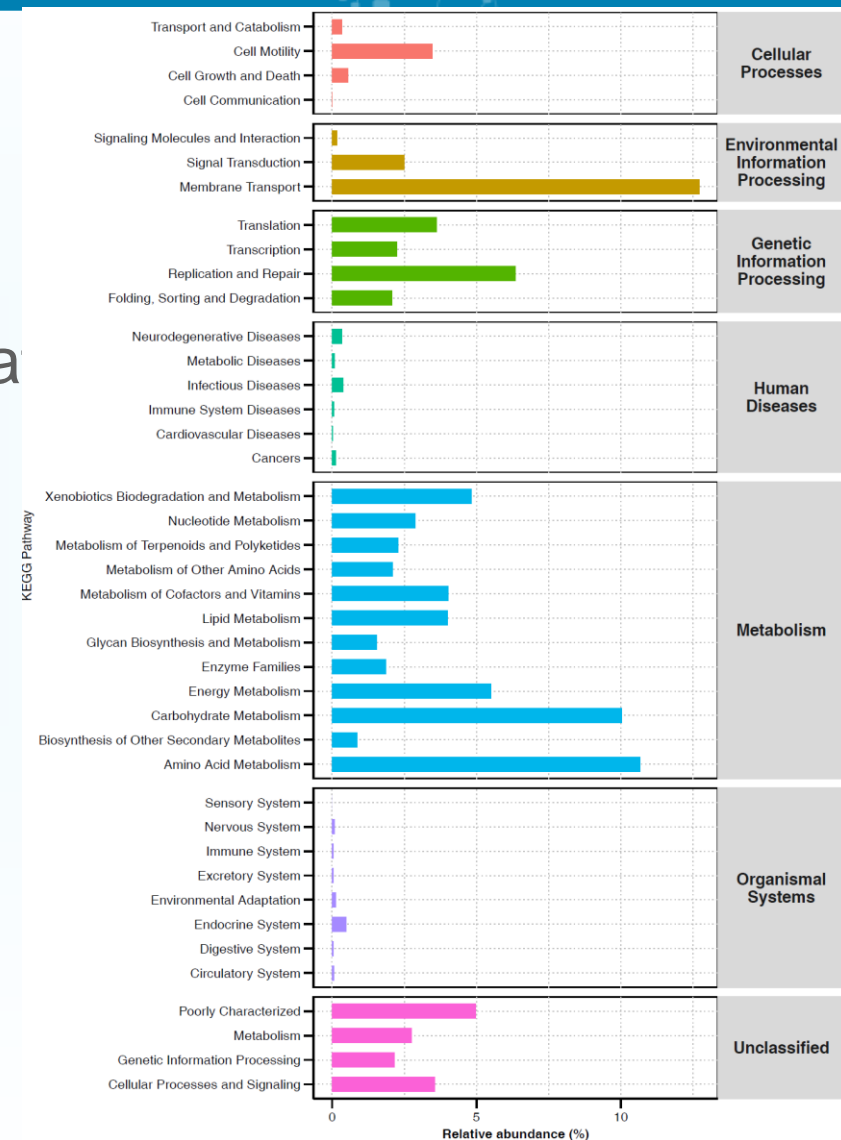
Rscript \${db}/script/compare_hierarchy_facet.R \

--input result/picrust/\${compare}.txt \

--data MeanB \

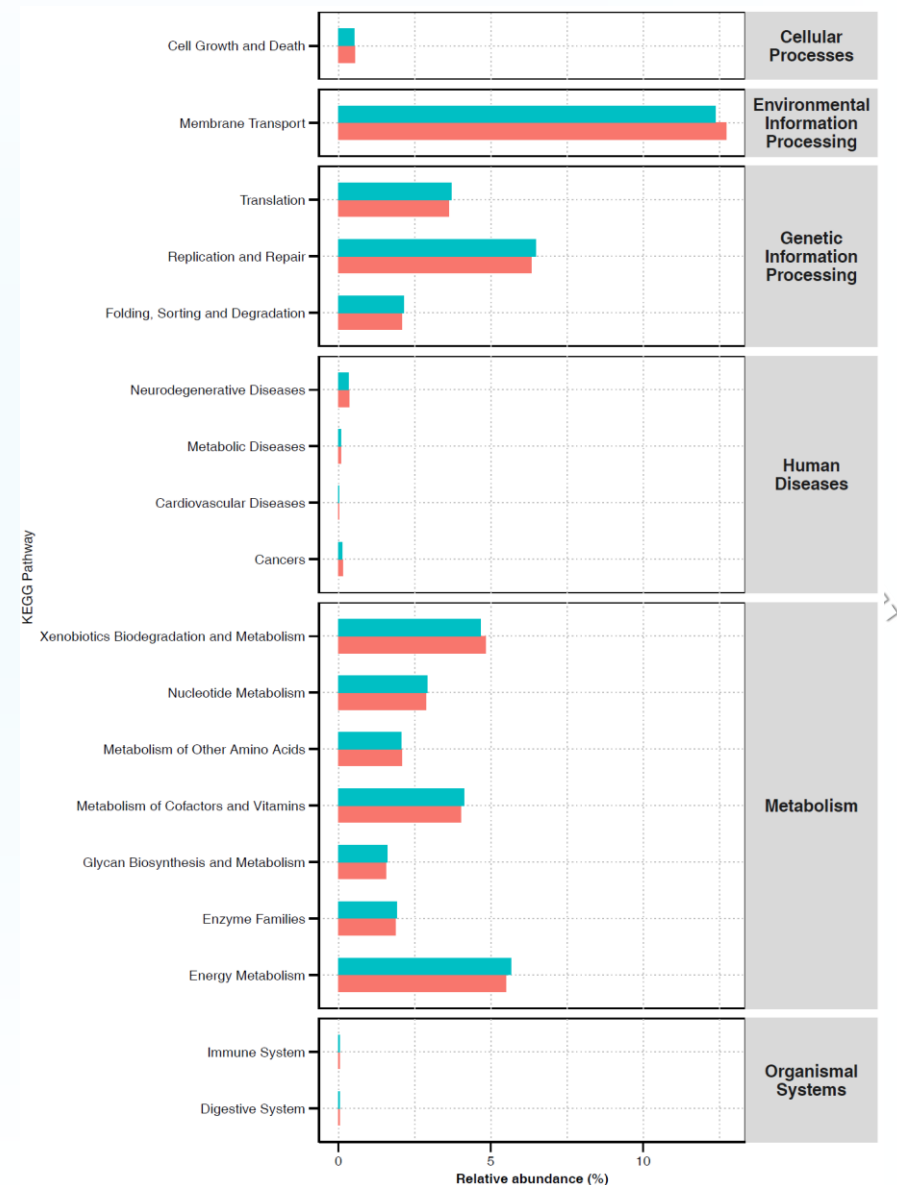
--annotation result/picrust/\${l}.anno.txt \

--output result/picrust/\${compare}.MeanB.bar.pdf



KEGG层级功能柱状图-差异比较

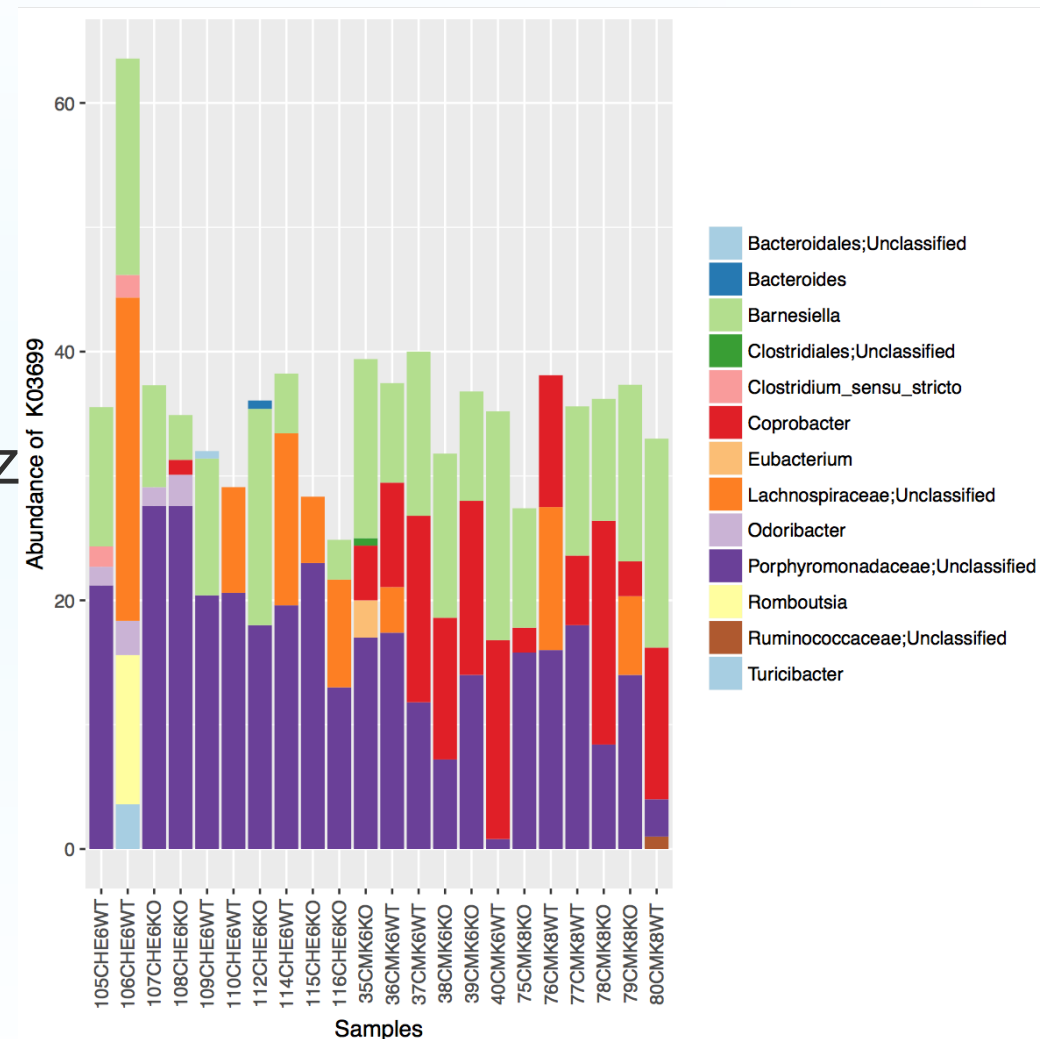
```
# 绘制A/B两组显著差异柱状图，按高分分类级分面
Rscript ${db}/script/compare_hierarchy_facet2.R \
--input result/picrust/${compare}.txt \
--pvalue 0.05 --fdr 0.1 \
--annotation result/picrust/${l}.anno.txt \
--output result/picrust/${compare}.bar.pdf
```



差异功能的物种组成贡献(microbiome_helper)

```
# 计算某个KO在所有菌中的丰度组成
metagenome_contributions.py \
  -i otutab_norm.biom \
  -l K01362,K01104 \
  -o metagenome_contributions.txt \
  -c /db/picrust/ko_13_5_precalculated.tab.gz
```

```
# 展示功能对应的菌组成
plot_metagenome_contributions.R \
  --input metagenome_contributions.txt \
  --output K01362_contrib.pdf \
  --function_id K01362 --rank Genus
```



方法3.——1. 文件上传(选学/自学)

- 访问PICRUSt在线服务器 <http://huttenhower.sph.harvard.edu/galaxy>
- 数据和上传：Greengenes参考OTU表result/gg97_otutab.txt，或用官方测试数据；先从左侧下部 Get Data -- Upload File -- Choose local file 选择本地文件；再单击Start，完成后点Close。

Galaxy / Hutlab

Tools

search tools

Text Manipulation

PICRUSt

GrPhIA

MetaPhlAn

MetaPhlAn2

Send Data

LEfSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Extract Features

Phenotype Association

NGS TOOLBOX BETA

NGS: QC and manipulation

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

MaAsLin

Get Data

Upload File from your computer

Thanks for visiting our lab's tools number of resources for metagenomics. left for an overview of available tools.

Our lab's research interests include microbiome systems biology, and look at our additional research and publications.

The tools are available here without charge. You can create an account and/or login to save your work.

If you have any comments, questions, or suggestions, please contact us.

Download from web or upload from disk

Regular Composite

Please wait...1 out of 1 remaining.

Name	Size	Type	Genome	Settings	Status
gg97_otutab.txt	270 KB	picrust	unspecified (?)		Adding to history...

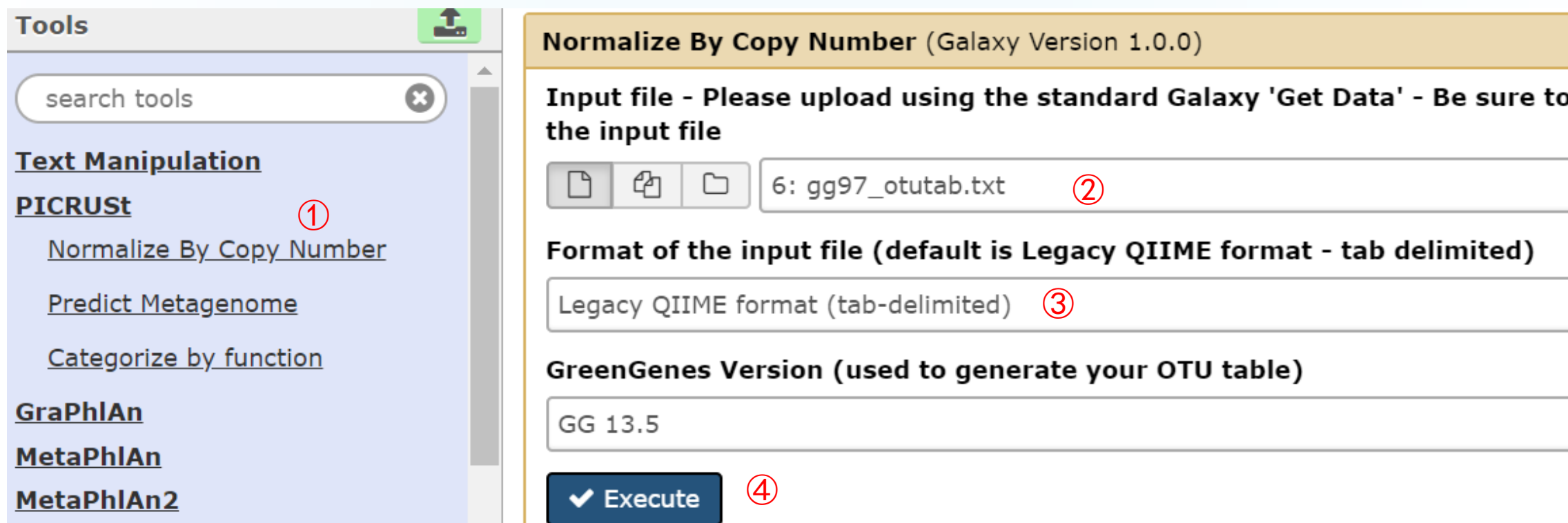
Type (set all): Auto-detect

Genome (set all): unspecified (?)

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

2. 拷贝数标准化 Normalize By Copy Number

- 左侧工具栏 [PICRUSt--NormalizeByCopyNumber](#)，Format 选择 tab-delimited，点击 Execute。



Tools

search tools

Text Manipulation

PICRUSt ①

- [Normalize By Copy Number](#)
- [Predict Metagenome](#)
- [Categorize by function](#)

GraPhlAn

MetaPhlAn

MetaPhlAn2

Normalize By Copy Number (Galaxy Version 1.0.0)

Input file - Please upload using the standard Galaxy 'Get Data' - Be sure to the input file

6: gg97_otutab.txt ②

Format of the input file (default is Legacy QIIME format - tab delimited)

Legacy QIIME format (tab-delimited) ③

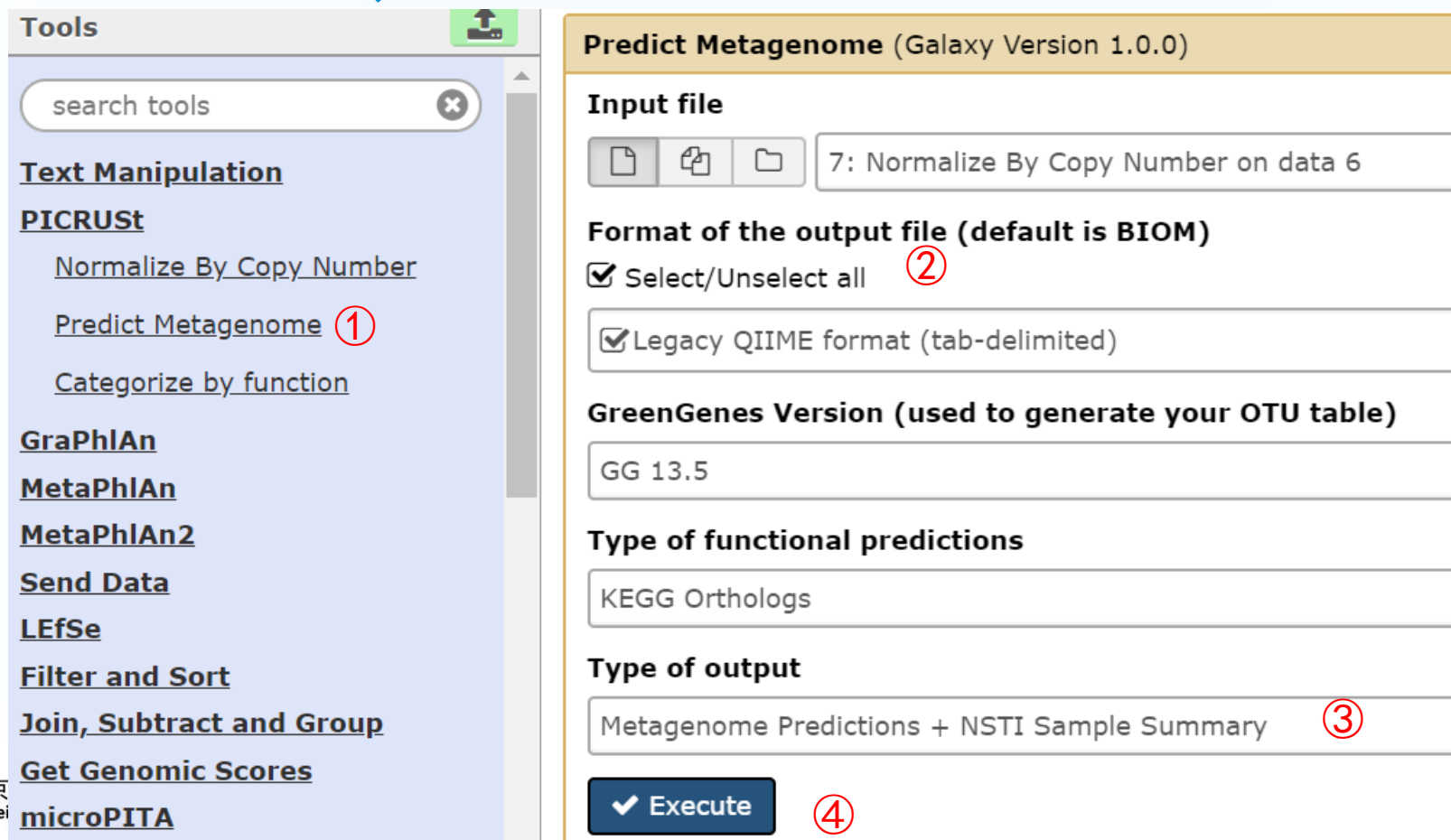
GreenGenes Version (used to generate your OTU table)

GG 13.5

✓ Execute ④

3. 预测宏基因组 Predict Metagenome

- 左侧工具栏 PICRUST -- Predict Metagenome, 选择Select, 再选第二项+NSTI输出格式, 点击execute即可。



Tools

search tools

Text Manipulation

PICRUST

[Normalize By Copy Number](#)

[Predict Metagenome](#) ①

[Categorize by function](#)

GraPhlAn

MetaPhlAn

MetaPhlAn2

Send Data

LefSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Predict Metagenome (Galaxy Version 1.0.0)

Input file

7: Normalize By Copy Number on data 6

Format of the output file (default is BIOM)

☒ Select/Unselect all ②

☒ Legacy QIIME format (tab-delimited)

GreenGenes Version (used to generate your OTU table)

GG 13.5

Type of functional predictions

KEGG Orthologs

Type of output

Metagenome Predictions + NSTI Sample Summary ③

④




宏基因组

信宝典









结果文件：KO表和评估值




- 点击Predict展开预览
- 磁盘图标下载

9: Predict Metagenome on data 7










6,909 lines
 format: **picrustp**, database: ?

# Constructed from biom file			
#OTU ID	K01	K02	K03
K01365	0.0	0.0	0.0
K01364	0.0	0.0	0.0
K01361	34.0	28.0	30.0

10: Predict Metagenome on data 7




18 lines
 format: **picrustp**, database: ?

#Sample	Metric	Value
K01	Weighted NSTI	0.07913
K02	Weighted NSTI	0.08605
K03	Weighted NSTI	0.07537
K04	Weighted NSTI	0.07485

易生信 生信宝典 宏基因组

KO表(OTU表转换)

- 同OTU表，多一行注释
- 最后列为KEGG描述

# Constructed from biom file																			KEGG_Description
#OTU_ID	K01	K02	K03	K04	K05	K06	OE1	OE2	OE3	OE4	OE5	OE6	WT1	WT2	WT3	WT4	WT5	WT6	
K01365	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	cathepsin-L [EC:3.4.22.15]
K01364	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	2.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	streptopain [EC:3.4.22.10]
K01361	34.0	28.0	30.0	21.0	17.0	25.0	93.0	41.0	56.0	56.0	30.0	99.0	16.0	10.0	21.0	17.0	12.0	23.0	lactocepin [EC:3.4.21.96]
K01360	2.0	1.0	2.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	proprotein convertase subtilisin/kexin-t
K01362	43674.0	38959.0	46565.0	46523.0	56257.0	46638.0	40868.0	36192.0	40302.0	40073.0	43432.0	40490.0	48129.0	56124.0	46664.0	46578.0	51170.0	44897.0	None
K02249	4.0	5.0	3.0	3.0	4.0	3.0	33.0	7.0	7.0	9.0	5.0	16.0	6.0	2.0	4.0	3.0	2.0	4.0	competence protein ComGG
K05841	476.0	859.0	569.0	708.0	618.0	502.0	1147.0	997.0	952.0	693.0	1076.0	1267.0	1515.0	1247.0	1205.0	965.0	1359.0	774.0	sterol-3beta-glucosyltransferase [EC:2.4
K05844	2596.0	2818.0	2969.0	2514.0	2515.0	3011.0	3377.0	3490.0	3103.0	3076.0	3040.0	3233.0	2803.0	2886.0	3272.0	2846.0	2732.0	3299.0	ribosomal protein S6 modification protei
K05845	4832.0	6960.0	6492.0	6292.0	5445.0	5698.0	5587.0	6003.0	5406.0	5111.0	5835.0	5893.0	6602.0	5996.0	6265.0	6761.0	7439.0	6053.0	osmoprotectant transport system substrat
K05846	10932.0	14827.0	13457.0	13417.0	11774.0	12902.0	12079.0	12645.0	11650.0	11016.0	12354.0	12551.0	13567.0	12611.0	13444.0	13678.0	14772.0	13104.0	osmoprotectant transport system permease
K05847	5974.0	8032.0	7509.0	7278.0	6163.0	6977.0	7080.0	7301.0	6678.0	6151.0	7123.0	7357.0	7644.0	7019.0	7622.0	7727.0	8005.0	7331.0	osmoprotectant transport system ATP-bind
K00508	3695.0	4866.0	5101.0	4155.0	3588.0	4618.0	3659.0	4783.0	4679.0	3907.0	4166.0	3878.0	4344.0	4350.0	4593.0	4527.0	4679.0	4449.0	linoleoyl-CoA desaturase [EC:1.14.19.3]
K00500	6298.0	4454.0	6610.0	6620.0	9234.0	6573.0	4578.0	4514.0	5436.0	5665.0	5521.0	4631.0	6197.0	7995.0	5977.0	6311.0	6770.0	6086.0	phenylalanine-4-hydroxylase [EC:1.14.16.
K00507	11904.0	11414.0	12139.0	12691.0	13825.0	13262.0	12671.0	10934.0	11995.0	11578.0	13068.0	12724.0	14706.0	16256.0	14327.0	14232.0	15822.0	13145.0	stearoyl-CoA desaturase (delta-9-desatur
K00504	63.0	65.0	54.0	44.0	37.0	55.0	104.0	67.0	65.0	58.0	59.0	79.0	60.0	49.0	86.0	84.0	45.0	70.0	peptidylglycine monooxygenase [EC:1.14.1
K00505	242.0	265.0	321.0	226.0	116.0	190.0	215.0	315.0	644.0	444.0	236.0	328.0	363.0	297.0	278.0	477.0	457.0	298.0	tyrosinase [EC:1.14.18.1]
K12521	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	minor pilin subunit PapF
K01101	495.0	573.0	642.0	643.0	889.0	661.0	588.0	689.0	433.0	526.0	484.0	563.0	486.0	423.0	473.0	456.0	509.0	543.0	4-nitrophenyl phosphatase [EC:3.1.3.41]
K01103	4284.0	2692.0	3952.0	4147.0	5946.0	4059.0	2658.0	2296.0	3289.0	3623.0	3230.0	2480.0	4114.0	5589.0	3669.0	3856.0	4277.0	3647.0	6-phosphofructo-2-kinase / fructose-2,6-
K12520	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	minor pilin subunit PapE
K01104	19629.0	20204.0	21658.0	20554.0	22590.0	20489.0	19426.0	19570.0	20173.0	19280.0	20151.0	19324.0	21909.0	23740.0	22263.0	22119.0	23133.0	21558.0	protein-tyrosine phosphatase [EC:3.1.3.4
K05592	6983.0	8068.0	8549.0	7814.0	7327.0	8318.0	8630.0	8484.0	7662.0	7212.0	8313.0	8575.0	8083.0	8286.0	8837.0	8063.0	8423.0	8425.0	ATP-dependent RNA helicase DeaD [EC:3.6.
K05593	76.0	70.0	59.0	51.0	37.0	81.0	201.0	93.0	128.0	124.0	86.0	134.0	80.0	88.0	86.0	86.0	70.0	70.0	aminoglycoside-6-adenylyltransferase [EC
K05590	0.0	0.0	1.0	1.0	0.0	0.0	39.0	1.0	0.0	27.0	0.0	25.0	0.0	0.0	0.0	0.0	0.0	2.0	ATP-dependent RNA helicase SrmB [EC:2.7.
K05591	5602.0	3605.0	5153.0	5733.0	7938.0	5797.0	3622.0	3289.0	4513.0	4772.0	4514.0	3795.0	5354.0	7108.0	4894.0	5281.0	5745.0	5007.0	ATP-independent RNA helicase DbpA [EC:3.
K05596	8690.0	7368.0	9082.0	9522.0	11918.0	9375.0	5669.0	5816.0	7536.0	7453.0	7396.0	6062.0	8747.0	10522.0	8021.0	8528.0	9691.0	7905.0	LysR family transcriptional regulator, c
K05597	3866.0	2282.0	3617.0	3794.0	5733.0	3480.0	2294.0	2069.0	3003.0	3330.0	2952.0	2232.0	3790.0	5273.0	3273.0	3515.0	3975.0	3295.0	glutamin-(asparagin)-ase [EC:3.5.1.38]

预测结果准确度NSTI(Nearest Sequenced Taxon Index)

- NSTI表征某样品中所有微生物OTU与其亲缘关系最近的已测序基因组间系统进化距离的平均值，该值越小表示预测结果越可信。
- 人肠道微生物样品的预测结果最好($\text{NSTI}=0.03 \pm 0.02\text{s.d.}$)
- 其次是土壤样品($\text{NSTI}=0.17 \pm 0.02\text{s.d.}$)
- 其他哺乳动物肠道样品波动较大($\text{NSTI}=0.14 \pm 0.06\text{s.d.}$)
- 究较少的高盐微生物席样品的预测准确度最低($\text{NSTI}=0.23 \pm 0.07\text{s.d.}$)

3.2 预测宏基因组 – 生成biom格式可进一步分类

- 左侧工具栏 PICRUSt -- Predict Metagenome, 文件确认是Copy Number文件, 其它全用默认, 结果可继续分析

Tools

search tools

Text Manipulation

PICRUSt

[Normalize By Copy Number](#)

[Predict Metagenome](#) ①

[Categorize by function](#)

GraPhlAn

MetaPhlAn

MetaPhlAn2

Send Data

LEfSe

Filter and Sort

Join, Subtract and Group

Get Genomic Scores

microPITA

Extract Features

Predict Metagenome (Galaxy Version 1.0.0) Options

Input file

7: Normalize By Copy Number on data 6

Format of the output file (default is BIOM)

☐ Select/Unselect all

☐ Legacy QIIME format (tab-delimited) ②

GreenGenes Version (used to generate your OTU table)

GG 13.5

Type of functional predictions

KEGG Orthologs

Type of output

Metagenome Predictions ③

Execute ④

4. 按功能类别分类汇总

- 工具栏 PICRUSt -- Categorize by function, 输入选择Predict Metagenome on data xx, 级别使用1/2/3可选, 输出格式改为tab, 点击Execute。结果就是转换为KO的表(功能分类), 和OTU类似。

Tools

search tools

Text Manipulation

PICRUSt

[Normalize By Copy Number](#)

[Predict Metagenome](#)

[Categorize by function](#) ①

GraPhlAn

MetaPhlAn

MetaPhlAn2

Categorize by function (Galaxy Version 1.0.0)

Input file

9: Predict Metagenome on data 7 ②

此步的输入必须是biom格式结果, 即3.2步默认参数生成结果

KEGG Pathway

3 ③

Type of output

Legacy QIIME format (tab-delimited) ④

Execute

KO功能分类结果L1-L3

- 以我们的结果为例
- 宏基因组KO表中，有6908个功能条目
- L3有327条，L2有40条，L1有7大类

pathway1.txt	9
pathway2.txt	42
pathway3.txt	329
KO.txt	6910

#OTU ID	#OTU ID	#OTU ID	K01	K02
1,1,1-Trichloro	Amino Acid Meta	Cellular Processes	2181052	1874721
ABC transporter	Biosynthesis of	Environmental Information Processing	7510620	7274295
Adherens juncti	Cancers	Genetic Information Processing	7064020	6892976
Adipocytokine s	Carbohydrate Me	Human Diseases	499032	460804
African trypano	Cardiovascular	Metabolism	24639759	24888984
Alanine, aspart	Cell Communicat	None	101891	100488
Aldosterone-reg	Cell Growth and	Organismal Systems	434121	436613
Alzheimer's dis	Cell Motility	Unclassified	6667902	6257232

1/12

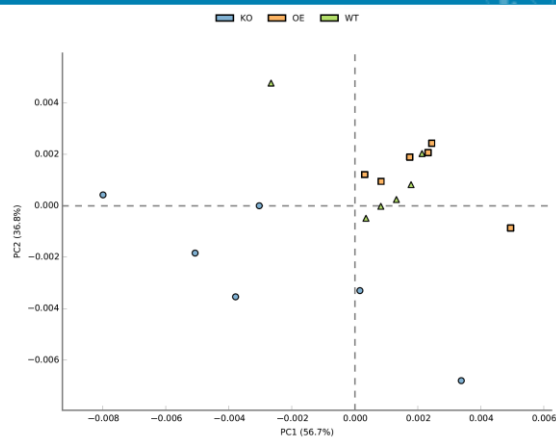


结果的KO表可以当做一个OTUs表进行统计分析

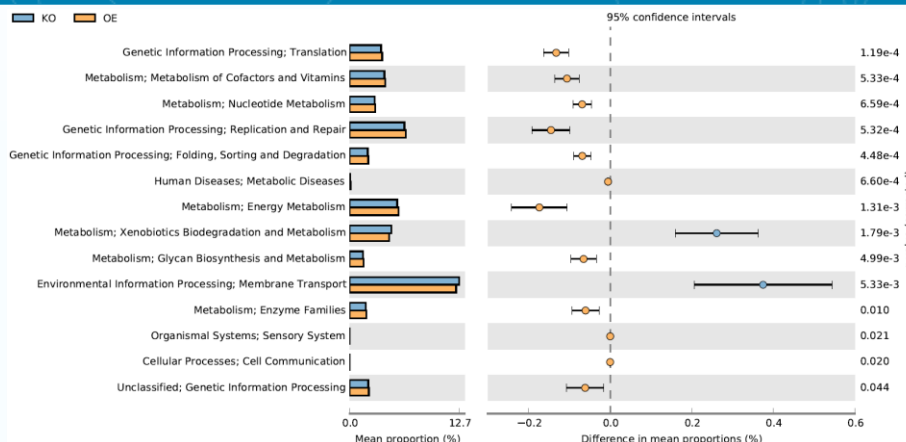
- STAMP 统计绘图：多组、两组比较，PCA、barplot、boxplot、heatmap、extended barplot
- R语言统计绘图：Alpha多样性——boxplot；Beta多样性 —— PCoA；组间差异——火山图、热图
- LEfSe差异分析：组间差异功能标记

易生信 生信宝典 宏基因组

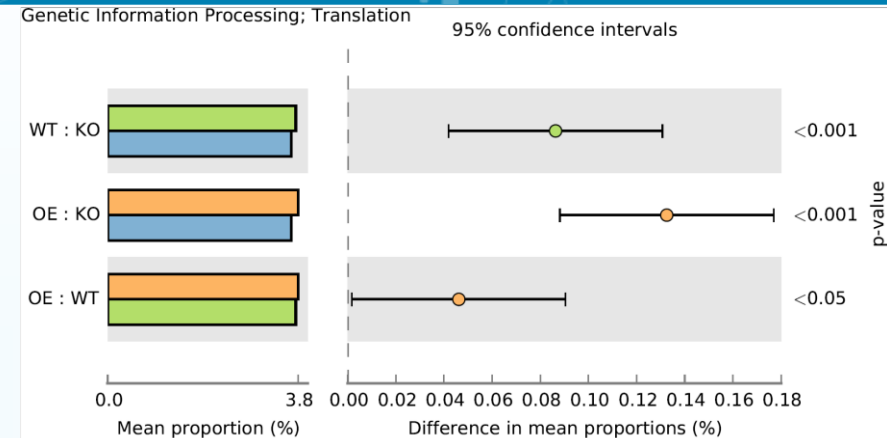
STAMP在各层级整合和细节分析结果展示



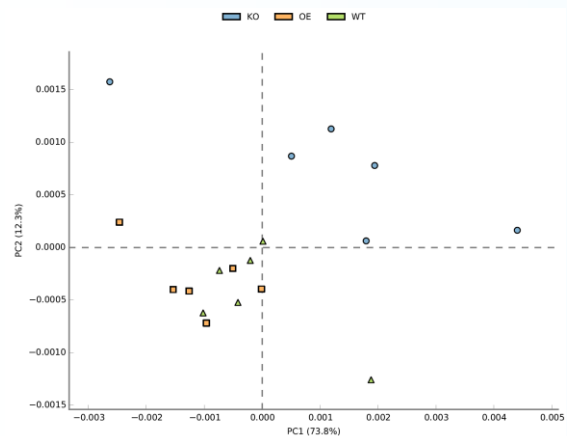
Pathway 2整体PCA



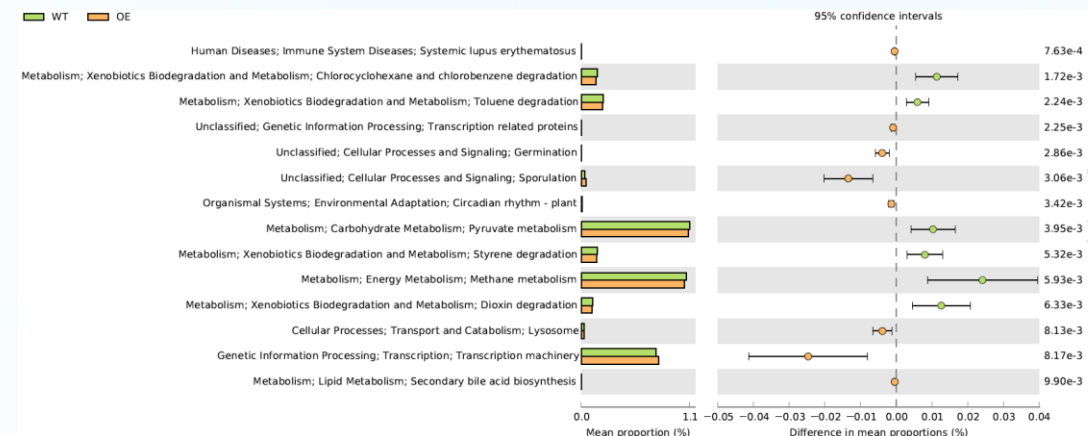
Pathway 2细节二组差异扩展柱状图(含L1)



Pathway 2细节多组比较事后检验(post-hoc)



Pathway 3整体PCA



Pathway 3细节二组差异扩展柱状图(含L2)

pathway1.txt
pathway2.txt
pathway3.txt
KO.txt

9
42
329

通常在通路2、3级分析，
数量不多也不少。太少没
啥说的，太多看不过来

- PICRUST简介
- 原理与应用
- 本地+在线操作
- 进一步学习

易生信 生信宝典 宏基因组



PICRUSt2: 任意16S序列预测宏基因组


- PICRUSt最新版1.1.4 <https://picrust.github.io/picrust/>
- PICRUSt2于2020年6月正式发表于Nature Biotechnology:

PICRUSt2 for prediction of metagenome functions

[GM Douglas](#), [VJ Maffei](#), [JR Zaneveld](#), [SN Yurgel...](#) - Nature ..., 2020 - nature.com

To the Editor—One limitation of microbial community marker-gene sequencing is that it does not provide information about the functional composition of sampled communities.

PICRUSt1 was developed in 2013 to predict the functional potential of a bacterial community on the basis of marker gene sequencing profiles, and now we present PICRUSt2 (<https://github.com/picrust/picrust2>), which improves on the original method. Specifically, PICRUSt2 contains an updated and larger database of gene families and reference ...

☆ Save  Cite Cited by 1515 Related articles All 6 versions

宏基因组



PICRUSt2的优缺点

- 主页: <https://github.com/picrust/picrust2>
- 第二版的主要优点:
 1. 不再只依赖Greengenes, 任何OTU或ASV均可直接预测;
 2. 预测基因组数据扩大了10倍以上;
 3. 输出 MetaCyc 和 MinPath 结果更准确且与宏基因组分析软件 HUMAnN2结果保持一致
- 缺点: 只支持Linux/Mac, 至少16G以上内存
- 详见: 软件使用教程



PICRUSt2安装和使用

- 安装指定版2.3.0_b，不加参数安装最新版，参数可能有变化

```
conda create -n picrust2 -c bioconda -c conda-forge picrust2=2.3.0_b
```

- 启动picrust2工作环境

```
conda activate picrust2
```

- 基于特征序列和表开展分析

```
picrust2_pipeline.py -s otus.fa -i otutab.txt \  
-o picrust2 -p 8
```

- 基于特征序列和表开展分析

```
add_descriptions.py -i KO_metagenome_out/pred_metagenome_unstrat.tsv.gz  
-m KO -o KO_metagenome_out/pred_metagenome_unstrat_descrip.tsv.gz
```



PICRUSt2主要结果

○ E.C. : [pred_metagenome_unstrat_descrip.tsv.gz](#)

function	description	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2
EC:1.1.1.1	Alcohol dehydrogenase	82528.86	92478.69	98803.72	97071.53	97450.66	93277.27	81081.31	82891.35
EC:1.1.1.100	3-oxoacyl-[acyl-carrier-protein]	189254.5	220111.1	220337.7	215184.3	211407.6	210296.7	188899.4	190916
EC:1.1.1.102	3-dehydrosphinganine reductase	156.33	159.5	318.33	347.17	124.83	137.67	176.5	167.67

○ KO: [pred_metagenome_unstrat_descrip.tsv](#)

function	description	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2
K00001	E1.1.1.1, adh; alcohol dehydrogenase	27382.03	30201.88	33393.11	32112.69	35466.87	31491.3	24797.15	26854.63
K00002	AKR1A1, adh; alcohol dehydrogenase	0	1	0	1	0	0	18	0
K00003	E1.1.1.3; homoserine dehydrogenase	29262.57	31605.39	33731.4	32935.71	34413.49	33166.09	27664.69	27693.45

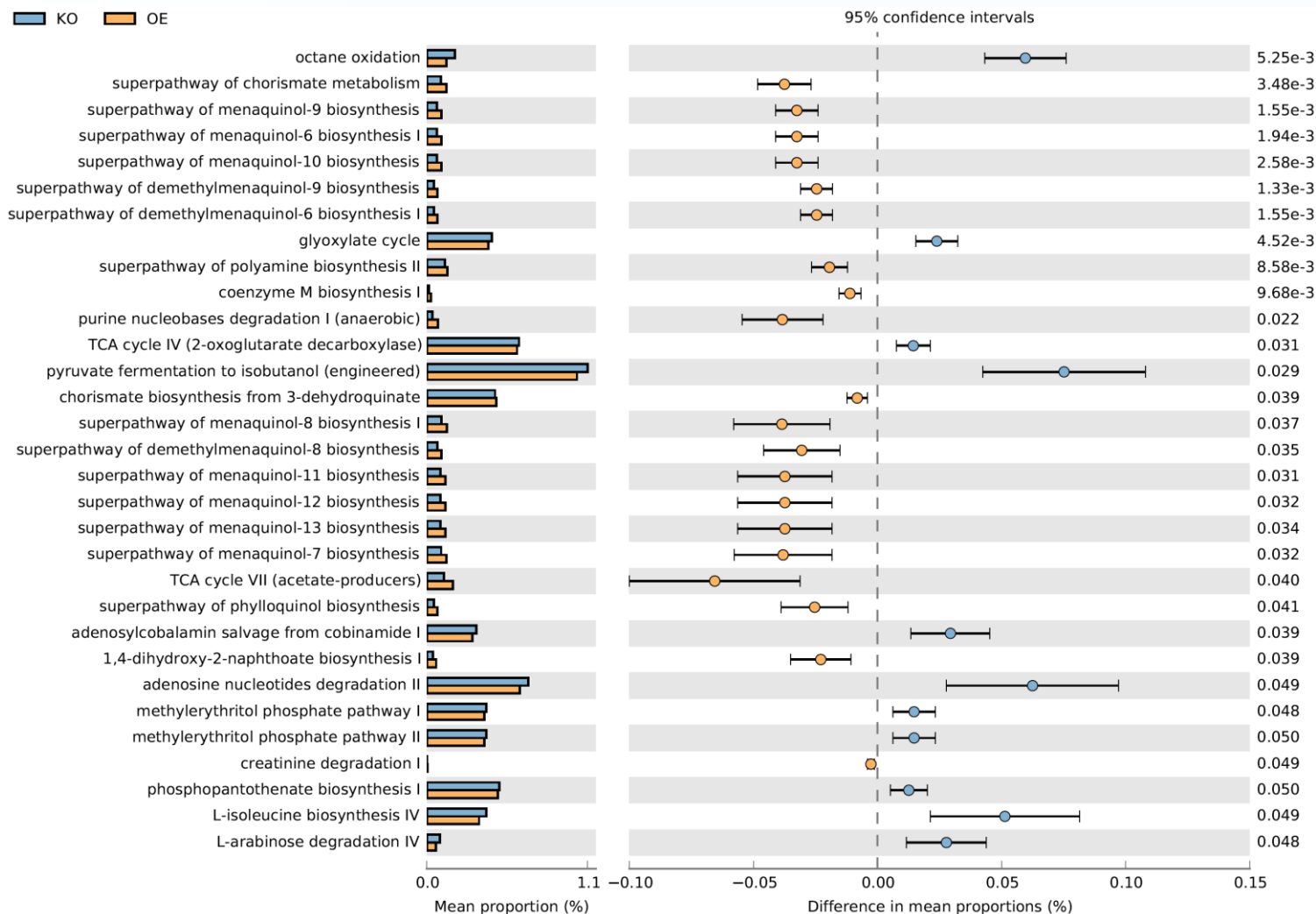
○ Pathway : [path_abun_unstrat_descrip.tsv.gz](#)

pathway	description	KO1	KO2	KO3	KO4	KO5	KO6	OE1	OE2
1CMET2-PWY	N10-formyl-tetrahydrofolate biosynthesis	32547.02	34157.25	38745.35	36515.75	39943.82	36825.05	30588.08	31455.05
3-HYDROXYPHENYLACE	4-hydroxyphenylacetate degradation	5185.698	5607.09	6628.57	6215.858	5553.78	5236.02	5724.779	5125.748
AEROBACTINSYN-PWY	aerobactin biosynthesis	116.8037	111.7943	106.9449	272.4736	146.0153	114.1764	204.104	96.059

可直接用STAMP差异比较，可用R、LEfSe等进行差异比较和生物标记鉴定



PICRUSt2结果使用STAMP统计Pathway中KOs vs OE



pathway	description	KO: mean	Difference
1CMET2-PWY	N10-formyl-tetrahydrof	0.489577	0.008429
3-HYDROXYPHEN	4-hydroxyphenylacetate	0.077115	-0.00607
AEROBACTINSYN	aerobactin biosynthesis	0.001938	-8.33E-05
ALL-CHORISMAT	superpathway of chorism	0.101197	-0.03748
ANAEROFRUCAT	homolactic fermentation	0.503431	-0.00204
ANAGLYCOLYSIS	glycolysis III (from glucos	0.512474	-0.01121
ARG+POLYAMIN	superpathway of arginin	0.302554	0.016227
ARGDEG-PWY	superpathway of L-argin	2.00E-05	-0.00187
ARGORNPROST-	arginine, ornithine and p	0.135871	-0.01985
ARGSYN-PWY	L-arginine biosynthesis I	0.411625	-0.01916
ARGSYNBSUB-PV	L-arginine biosynthesis I	0.511917	0.010086
ARO-PWY	chorismate biosynthesis	0.509542	-0.00552
ASPASN-PWY	superpathway of L-aspa	0.123856	-0.0224
AST-PWY	L-arginine degradation I	0.035229	0.006791
BIOTIN-BIOSYNT	biotin biosynthesis I	0.421802	0.024768
BRANCHED-CHA	superpathway of branch	0.728606	0.001515
CALVIN-PWY	Calvin-Benson-Bassham	0.548704	0.007312
CATECHOL-ORT	catechol degradation to	0.096211	0.011449
CENTFERM-PWY	pyruvate fermentation to	0.009986	-0.01368
CHLOROPHYLL-S	chlorophyllide a biosynt	0.034249	-0.00053
COA-PWY	coenzyme A biosynthesis	0.46577	-0.00309
COBALSYN-PWY	adenosylcobalamin salva	0.352834	0.029407
CODH-PWY	reductive acetyl coenzym	0.00641	-0.00315

C:\amplicon\result\picrust2\pathways_out\pathway_KO_OE.pdf

pathway_KO_OE.tsv

整合KEGG的KO至Pathway及更高分类级

○ # 解压KO表

```
zcat KO_metagenome_out/pred_metagenome_unstrat.tsv.gz > KEGG.KO.txt
```

○ # 按KEGG的Level 1/2/3合并

```
python3 ${db}/script/summarizeAbundance.py \
-i KEGG.KO.txt \
-m ${db}/kegg/KO1-4.txt \
-c 2,3,4 -s ',+,+', -n raw \
-o KEGG
```

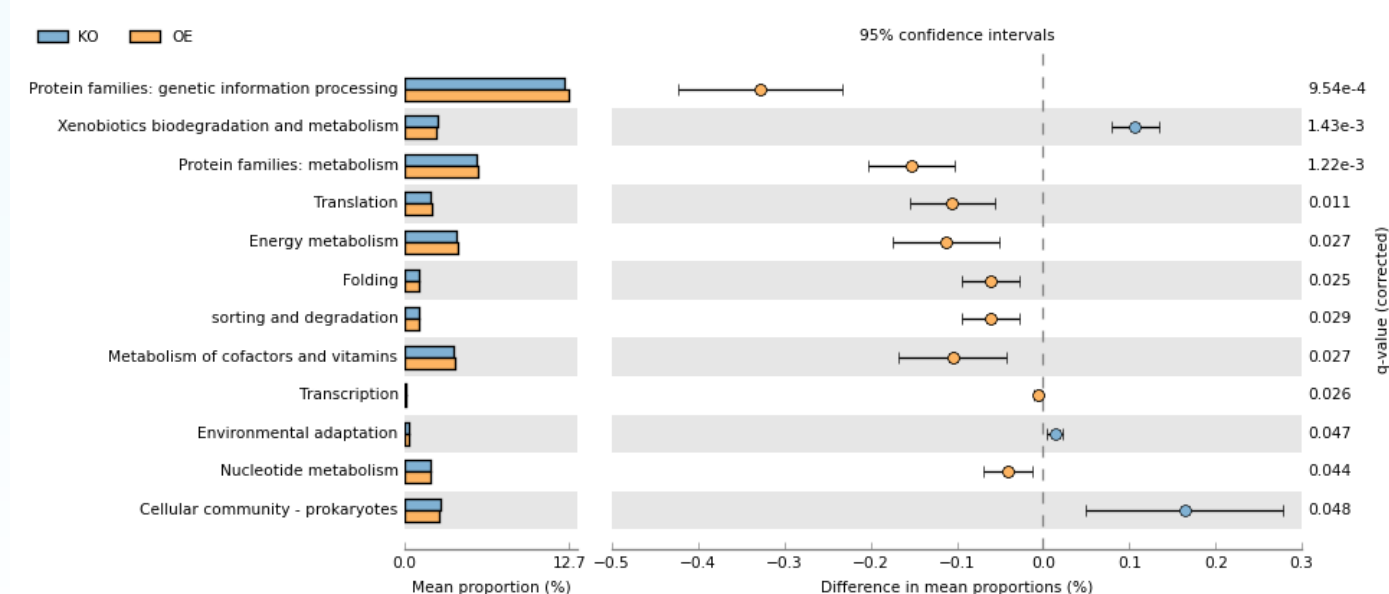
各级别功能条目数量

6707 KEGG.KO.txt

474 KEGG.Pathway.raw.txt

9 KEGG.PathwayL1.raw.txt

55 KEGG.PathwayL2.raw.txt



STAMP展示PathwayL2中KOvsOE的差异功能通路

易生信，毕生缘；培训版权所有。



- PICRUST有着目前最广泛的应用，本地、服务器均可方便使用，自带KEGG层级汇总，但局限于13年版的GreenGenes数据库。
- PICRUST 2 需要16GB以上内存的Linux系统下运行，支持KEGG、EC和MetaCyc三类功能注释。KEGG的层级汇总需手动制作，我们提供了summarizeAbundance.py脚本一键汇总。
- 结果可用STAMP快速进行整体(PCA)、局部多组比较+事后检验(post-hoc)、两组比较(extended bar plot)。通常在Pathway的Level 2或3级展开分析，结果数量可读性强。

进一步学习

- 0概述：根据16S预测微生物群落功能最全攻略，对以下4种常用的功能预测工具进行简介、实例说明和点评，必读；
- 1KO通路PICRUST：本地和在线分析，及统计分析可视化指导；
- 2元素循环FAPROTAX：元素循环相关菌代谢预测，只有本地版python脚本；识别菌种名称；
- 3表型bugbase：表型预测、革兰氏、氧气利用等；有在线和本地版，基于Greengene数据库；
- 4KO通路Tax4Fun：R包基于作者整理的Sliva123数据，准确率不可知，输出结果为百分比；最新版Tax4Fun2
- PICRUST2：OTU/ASV等16S序列随意预测宏基因组，参考数据库增大10倍





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

