

# 诺禾致源宏基因组项目高级分析

## 目 录

目 录	1
1. 环境因子分析	3
1.1 环境因子相关性分析	3
1.1.1 分析方法	3
1.1.2 结果展示	3
1.1.3 交付结果	4
1.1.4 参考文献	5
1.2 CCA/RDA 分析	5
1.2.1 分析方法	5
1.2.2 结果展示	5
1.2.3 交付结果	6
1.2.4 参考文献	7
1.3 VPA 分析	7
1.3.1 分析方法	7
1.3.2 结果展示	7
1.3.3 交付结果	7
1.3.4 参考文献	8
2. 病原与宿主互作数据库 ( PHI ) 注释	8
2.1 分析方法	8
2.2 结果展示	8
2.3 交付结果	8
2.4 参考文献	8
3. 分泌蛋白预测	9
3.1 分析方法	9
3.2 结果展示	9
3.3 交付结果	9
3.4 参考文献	9
4. III 型分泌系统效应蛋白预测	10
4.1 分析方法	10
4.2 结果展示	10
4.3 交付结果	10
4.4 参考文献	10
5. 分泌系统相关蛋白预测	11

5.1	分析方法.....	11
5.2	结果展示.....	11
5.3	交付结果.....	11
6.	细菌致病菌毒力因子 ( VFDB ) 注释.....	12
6.1	分析方法.....	12
6.2	结果展示.....	12
6.3	交付结果.....	12
6.4	参考文献.....	12
7.	CAG/MLG 分析 .....	13
7.1	分析方法.....	13
7.2	结果展示.....	13
7.3	参考文献.....	15
8.	CNV 拷贝数变异分析.....	15
8.1	分析方法.....	15
8.2	结果展示.....	16
8.3	参考文献.....	17
9.	肠型分析.....	17
9.1	分析方法.....	17
9.2	结果展示.....	17
9.3	参考文献.....	19
10.	噬菌体分析.....	19
10.1	分析方法.....	19
10.2	结果展示.....	20
10.3	参考文献.....	21

# 1、环境因子分析

## 1.1 环境因子相关性分析

相关分析（correlation analysis），是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度，是研究随机变量之间的相关关系的一种统计方法。

研究环境因子与物种之间，或环境因子与功能基因之间的相关性，目前常用的有Spearman相关性分析和Mantel test分析，来检验两两矩阵之间的相关性，并得到显著性的p值。

### 1.1.1 分析方法

#### (1) Spearman 相关性分析

Spearman相关性分析以Spearman相关系数作为量度，Spearman相关系数又称秩相关系数，是利用两变量的秩次大小作线性相关分析，对原始变量的分布不作要求，属于非参数统计方法，适用范围较广泛。

用Spearman秩相关来研究环境因子与微生物种丰度数据（或基因丰度）之间的相互变化关系，得到两两之间的相关性和显著性P值。

使用软件：R vegan包，输入文件：物种文件（或基因文件）、环境因子文件。

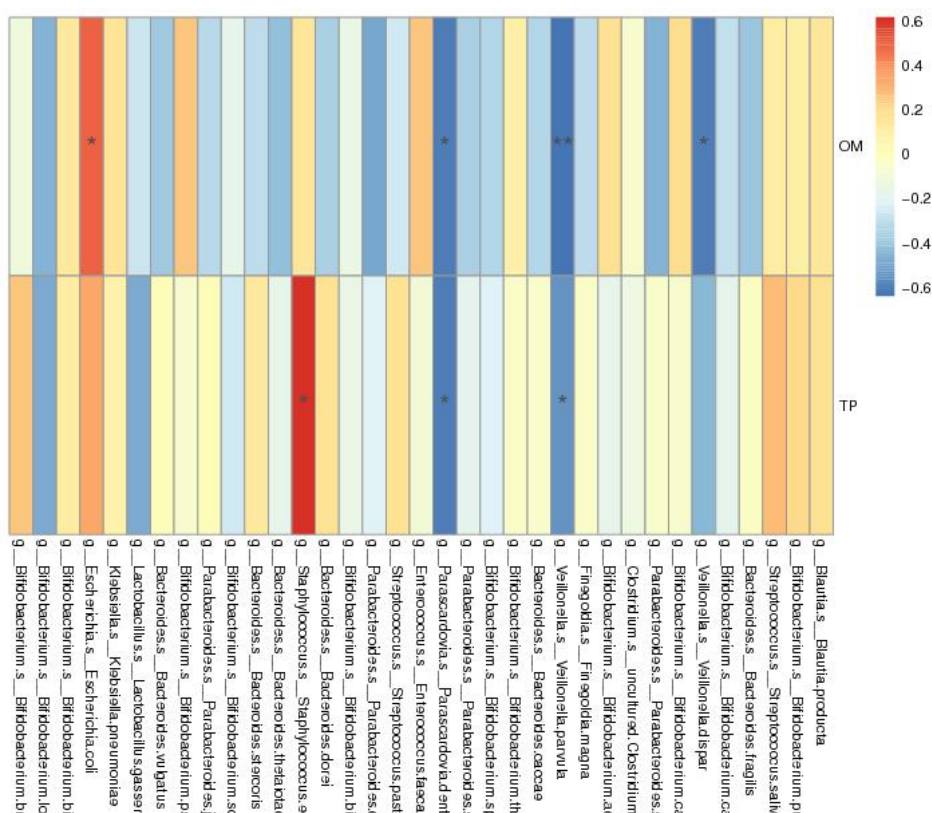
#### (2) Mantel test分析

Mantel test是对两个矩阵相关关系的检验，可用于计算环境因子和微生物群落数据或者环境因子和功能注释数据的相关性。

使用软件：R vegan包，输入文件：物种注释矩阵文件或功能注释矩阵文件、环境因子文件。

### 1.1.2 结果展示

#### (1) Spearman 相关性分析结果



spearman 相关性分析热图

说明：纵向为环境因子信息，横向为物种信息，中间热图对应的值为spearman相关系数r，介于-1，1之间，r<0为负相关，r>0为正相关，标\*表示显著性检验p值 <0.05。


## (2) Mantel test分析结果


Variable	r	P
DOC+AN+AP+AK+Ca+Mg+Urease+ACPase	0.4865	0.01
DOC+AN+AP+AK+Ca+Mg	0.5202	0.005
DOC+AN+AP+AK+Urease+ACPase	0.3696	0.01
AN+AP+AK+Ca+Mg+Urease+ACPase	0.3949	0.015
DOC+Ca+Mg+Urease+ACPase	0.6118	0.008
AN+AP+AK+Urease+ACPase	0.3467	0.024
AN+AP+AK	0.2636	0.089
DOC+Ca+Mg	0.7614	0.001
Ca+Mg	0.6051	0.002

说明：Variable是环境因子信息，r为相关系数，P为显著性检验p值。


## 1.1.3 交付结果

结果目录：spearman

 correlation.pdf

 correlation.png

 correlation.xls


 pvalue.xls

correlation.pdf/correlation.png : spearman 热图

correlation.xls : spearman 分析结果

pvalue.xls: 显著性p值

结果目录: mantel\_test

 mantel\_test\_table.xls

mantel\_test.txt : Mantel test分析结果

#### 1. 1. 4 参考文献

【1】 Algina, J., & Keselman, H. J. (1999). Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test significance. *Psychological Methods*, 4(1), 76-83.

【2】 Yang S L, Zhang J, Xu X J. Influence of the Three Gorges Dam on downstream delivery of sediment and its environmental implications, Yangtze River[J]. *Geophysical research letters*, 2007, 34(10).

### 1. 2 CCA/RDA 分析

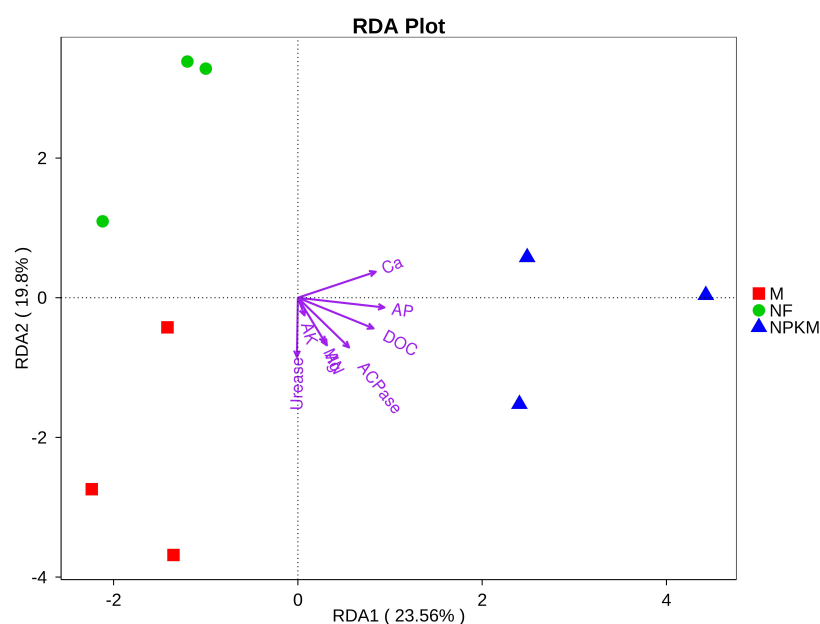
主要用来反映菌群与环境因子之间的关系，可以检测环境因子、样品、菌群三者之间的关系或者两两之间关系，可得到影响样品分布的重要环境驱动因子。

#### 1. 2. 1 分析方法

CCA/RDA 分析是基于对应分析发展的一种排序方法，将对应分析与多元回归分析相结合，每一步计算均与环境因子进行回归，又称多元直接梯度分析。RDA 是基于线性模型，CCA 是基于单峰模型。具体方法，首先对物种数据矩阵或功能数据作除趋势对应分析，即DCA(Detrended correspondence analysis)分析，根据梯度值确定线性模型（RDA）和单峰模型(CCA)哪个最合适（DCA 分析结果中Axislength 的前4 个轴中最大的值如果大于4.0，应该选CCA，如果3.0~4.0 之间，选RDA 和CCA 均可，如果小于3.0，RDA 的结果要好于CCA），交付结果中RDA 和CCA 的结果均会给出。

使用软件：R 软件vegan 包，输入文件：物种矩阵文件或功能矩阵文件、环境因子文件。（注：需要提供环境因子的数据，比如pH 值，温度值，临床因子等。）

### 1.2.2 结果展示



说明：在RDA 排序图内，环境因子一般用箭头表示，箭头连线的长度代表某个环境因子与群落分布和种类分布间相关程度的大小，箭头越长，说明相关性越大，反之越小。箭头连线和排序轴的夹角代表某个环境因子与排序轴的相关性大小，夹角越小，相关性越高；反之越低。环境因子之间的夹角为锐角时表示两个环境因子之间呈正相关关系，钝角时呈负相关关系。

### 1.2.3 交付结果

结果目录：CCA/RDA

dca.csv  
rda.env.csv cca.env.csv  
rda.sample.csv cca.sample.csv  
rda.sp.csv cca.sp.csv  
rdaenvfit.csv ccaenvfit.csv

CCA/RDA 图：

CCA\_sample\_env.pdf RDA\_sample\_env.pdf  
CCA\_sample\_env.png RDA\_sample\_env.png  
CCA\_sample\_env2.pdf RDA\_sample\_env2.pdf  
CCA\_sample\_env2.png RDA\_sample\_env2.png

dca.csv: DCA 结果, 判断用CCA 分析还是RDA 分析的文件;

\*.sample.csv: 各样本在各排序轴上的计算值;

\*.sp.csv: 各物种在各排序轴上的计算值;

\*.env.csv: 各环境因子在各排序轴上的计算值;

\*envfit.csv: 各环境因子对排序结果的相关性系数及显著性检验值;

#### 1.2.4 参考文献

【1】Sheik CS, Mitchell TW, Rizvi FZ, Rehman Y, Faisal M, et al. (2012) Exposure of Soil Microbial Communities to Chromium and Arsenic Alters Their Diversity and Structure. PLoS ONE 7(6): e40059. doi:10.1371/journal.pone.0040059.

### 1.3 VPA 分析

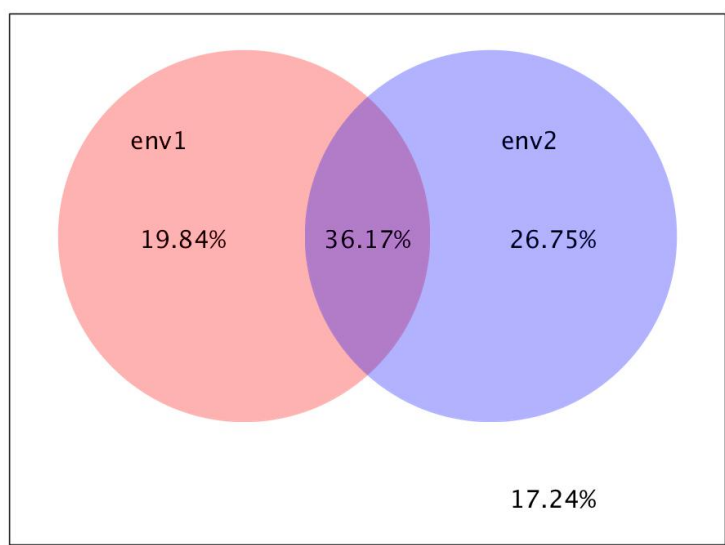
VPA 分析 (Variance partitioning canonical correspondence analysis (CCA)), 配合 CCA/RDA使用, 重点研究各环境因子对微生物群落分布的解释率, 可得到影响样品分布的环境驱动因子的影响大小。

#### 1.3.1 分析方法

约束排序的主要类型有RDA和CCA, 在约束排序里, 只展示能被环境因子所解释的物种分布变化量, 如果想分析某几个环境因子对物种分布的解释量或者单个环境因子分别的解释部分就要用到VPA分析。

使用软件: R软件vegan包, 输入文件: 物种矩阵文件或功能矩阵文件、环境因子文件。(注: 需要提供环境因子的数据, 比如 pH值, 温度值, 临床因子等, 以及想比较的环境因子分组。)

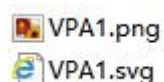
#### 1.3.2 结果展示



说明：圆圈交叉部分为两组环境因子所共有的解释量，圆圈外17.24%为不能的解释量

### 1.3.3 交付结果

结果目录：VPA



### 1.3.4 参考文献

【1】Yang S L, Zhang J, Xu X J. Influence of the Three Gorges Dam on downstream delivery of sediment and its environmental implications, Yangtze River[J]. Geophysical research letters, 2007, 34(10).

## 2. 病原与宿主互作数据库 ( PHI ) 注释

### 2.1 分析方法

PHI 全称为 Pathogen Host Interactions Database，病原与宿主互作数据库，其内容经过实验验证，主要来源于真菌、卵菌和细菌病原，感染的宿主包括动物、植物、真菌以及昆虫。

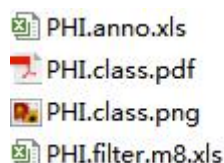
使用 BLAST 软件，把目标物种的氨基酸序列，与 PHI 数据库进行比对，把目标物种的基因和其相对应的功能注释信息结合起来，得到注释结果。由于每一条序列比对结果可能超过一条，为保证其生物意义，注释时保留一条最优比对结果作为该基因的注释。最后提供的 BLAST 结果为M8格式。

### 2.2 结果展示



Gene_id	Identity	E_value	PHI_ACCESS	Gene_Name	Protein	#NCBI_TAX_Species	Phenotype	taxonomy		
nana_2176	43.67	#####	PHI:2469	AcrB	Q7WTQ9	552	Erwinia amylovora	Reduced virulence	k_Bacteria;p_Proteobacteria	
nana_2178	47.29	2.00E-35	PHI:2356	MoHYR1	G4W178	148305	Magnaporthe oryzae	Reduced virulence	k_Bacteria;p_Actinobacteria;c	
nana_2179	43.92	1.00E-50	PHI:1566	Gz0E006	I1RC95	5518	Gibberella zeae (Lethal		k_Bacteria;p_Proteobacteria;c	
nana_2180	46.13	1.00E-80	PHI:624	ssaN	AA068935	216597	Salmonella enteri	Reduced virulence	k_Bacteria;p_Proteobacteria;c	
nana_2181	40.72	5.00E-41	PHI:2445	ColR	Q5H3K9	64187	Xanthomonas oryzae	Reduced virulence	k_Bacteria;p_Actinobacteria;c	
nana_2186	40.4	1.00E-16	PHI:2293	cycA	B0XTA5	746128	Aspergillus fumig	Reduced virulence	k_Bacteria;p_Proteobacteria;c	
nana_2186	52.31	1.00E-20	PHI:2971	CspR	Q82ZX2	1351	Enterococcus faec	Reduced virulence	k_Bacteria;p_Proteobacteria;c	
nana_2186	53.36	2.00E-77	PHI:877	MGC_00385	RDK03005	318829	Magnaporthe oryzae	Reduced virulence	k_Bacteria;p_Actinobacteria;c	

## 2.3 交付结果



PHI.class.pdf: PHI 数据库注释分类条形图  
 PHI.class.png: PHI 数据库注释分类条形图  
 PHI.filter.m8: PHI 数据库进行BLAST 比对结果  
 PHI.anno.xls: PHI 数据库注释的结果文件（含物种注释信息）

## 2.4 参考文献

- 【1】 Winnenbunrg R, Baldwin T K, Urban M, et al. PHI-base: a new database for pathogen-host interactions[J]. Nucleic acids research, 2006, 34(suppl 1): D459-D464.
- 【2】 Baldwin T K, Winnenbunrg R, Urban M, et al. The pathogen-host interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity[J]. Molecular plant-microbe interactions, 2006, 19(12): 1451-1462.
- 【3】 Winnenbunrg R, Urban M, Beacham A, et al. PHI-base update: additions to the pathogen-host interaction database[J]. Nucleic acids research, 2008, 36(suppl 1): D572-D576.

## 3. 分泌蛋白预测

### 3.1 分析方法

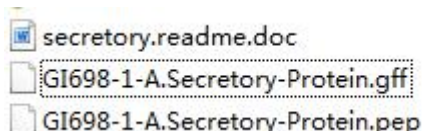
分泌蛋白是指在细胞内合成后，分泌到细胞外起作用的蛋白质。分泌蛋白的 N 端有一般由 15~30 个氨基酸组成的信号肽。使用信号肽预测工具 SignalP 注释蛋白序列是否是分泌蛋白。

### 3.2 结果展示

```
##gff-version 2
##sequence-name source feature start end score N/A ?
## -----
gene11 SignalP-4.1 SIGNAL 1 24 0.817 . . YES
gene22 SignalP-4.1 SIGNAL 1 24 0.784 . . YES
gene56 SignalP-4.1 SIGNAL 1 20 0.720 . . YES
gene85 SignalP-4.1 SIGNAL 1 24 0.679 . . YES
gene207 SignalP-4.1 SIGNAL 1 17 0.600 . . YES
gene218 SignalP-4.1 SIGNAL 1 23 0.721 . . YES
gene238 SignalP-4.1 SIGNAL 1 19 0.810 . . YES
gene336 SignalP-4.1 SIGNAL 1 23 0.776 . . YES
gene373 SignalP-4.1 SIGNAL 1 22 0.659 . . YES
```

### 3.3 交付结果

结果目录：SignalP



secretory.readme.doc: 交付结果说明

\*.Secretory-Protein.pep: 鉴定为分泌蛋白的氨基酸序列文件

\*.Secretory-Protein.gff: 鉴定为分泌蛋白的基因列表

### 3.4 参考文献

【1】 Petersen T N, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions[J]. Nature methods, 2011, 8(10): 785- 786.

## 4. III 型分泌系统效应蛋白预测

### 4.1 分析方法

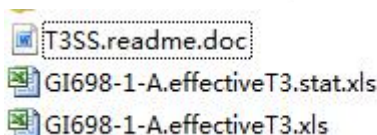
III 型分泌系统（Type III secretion system, T3SS）主要是革兰氏阴性菌的分泌蛋白分泌到细胞外的运输途径，因此 III 型分泌系统效应蛋白（Type III secretion system Effector protein）与革兰氏阴性致病菌致病机理有关。

使用软件 EffectiveT3 对输入的氨基酸序列进行预测，通过其内部特定的计算模型对每条氨基酸序列进行评分，分值越高，可信度越高，选出评分高于阈值的序列，认为这些序列为 III 型分泌系统效应蛋白。

### 4.2 结果展示

	A	B	C	D
1	Gene_id	Description	Score	is secreted
2	artGM000001	locus=Scaffold1:1007:1627:+	-1	FALSE
3	artGM000002	locus=Scaffold1:1590:2651:-	-1	FALSE
4	artGM000003	locus=Scaffold1:2755:4917:+	-1	FALSE
5	artGM000004	locus=Scaffold1:5037:5654:+	-1	FALSE
6	artGM000005	locus=Scaffold1:5659:6687:+	-1	FALSE
7	artGM000006	locus=Scaffold1:6761:7507:-	-1	FALSE
8	artGM000007	locus=Scaffold1:7534:8313:-	-1	FALSE
9	artGM000008	locus=Scaffold1:8325:9380:-	-1	FALSE

#### 4.3 交付结果



T3SS.readme.doc: 交付结果说明

\*.effectiveT3.xls: EffectiveT3 软件预测结果文件

\*.effectiveT3.stat.xls: EffectiveT3 软件预测结果统计

#### 4.4 参考文献

【1】 Arnold R, Brandmaier S, Kleine F, et al. Sequence-based prediction of type III secreted proteins[J]. PLoS pathogens, 2009, 5(4): e1000376.

### 5. 分泌系统相关蛋白预测

#### 5.1 分析方法

基于 nr 等 11 个数据库整合的序列库比对的方法，提取属于分泌系统相关蛋白的结果。

#### 5.2 结果展示

##### TnSS

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Type	Gene_id	[nr]	{KEGG}	{COG}	{CAZy}	{SwissProt}	{TrEMBL}	{PHI}	{IPR}	{GO}	{Secretory_Protein}	{T3SS}
2	type I	PccS1GM002253	{type I sec}	{K02022 AB}	{COG0845 M}	{NA}	{NA}	{CGDAZ8 PECC}	{NA}	{IPR003997}	{GO:0009306}	{NA}	{false}
3	type I	PccS1GM002254	{type I sec}	{K06148 AB}	{COG2274 A}	{NA}	{NA}	{CGDAZ9 PECC}	{NA}	{IPR001140}	{GO:0000166}	{NA}	{false}
4	type I	PccS1GM002255	{type I sec}	{NA}	{COG1538 O}	{NA}	{NA}	{CGDE00 PECC}	{NA}	{IPR003423}	{GO:0005215}	{YES}	{false}
5	type II	PccS1GM000329	{twitching m}	{K02669 pi}	{COG2805 T}	{NA}	{Uncharacteri}	{CGDE29 PECC}	{NA}	{IPR001482}	{GO:0000166}	{NA}	{false}
6	type II	PccS1GM000456	{putative m}	{K02682 pp}	{COG4969 T}	{NA}	{Prepilin p}	{CGDE36 PECC}	{NA}	{IPR000983}	{GO:0008565}	{NA}	{false}
7	type III	PccS1GM000416	{Hrp depend}	{NA}	{COG3395 U}	{NA}	{NA}	{CGDEB6 PECC}	{NA}	{IPR010737}	{NA}	{NA}	{false}
8	type III	PccS1GM000661	{putative t}	{NA}	{NA}	{NA}	{NA}	{K4PH65 PECC}	{NA}	{IPR019110}	{NA}	{YES}	{false}

##### SwissProt


Gene_id	Identity	E_value	Subject_i	Subject_taxonomy
nana_2174	77.5	#####	O86509	AGUA_STRCK_Bacteria;p_Actinobacteri
nana_2174	47.5	1.80E-37	Q2CEE2	AT_OCEGH_k_Bacteria;p_Actinobacteri
nana_2174	51.3	1.50E-57	P9WN70	G6PD1_MYCK_Bacteria;p_Actinobacteri
nana_2174	58.6	#####	P9WP47	CSTA_MYCK_k_Bacteria;p_Actinobacteri


##### TrEMBL

Gene_id	Identity	E_value	Subject_i	Subject_ctaxonomy						
nana_2174	75.7	4.80E-59	B6AM58	B6AM58_9Ek_Bacteria;p_Nitrospirae;c_Nitrospira;o_Nitrospirales;f_Nitrospiraceae;g_Nitrospira						
nana_2174	60	1.60E-62	X0WAL3	X0WAL3_9Zk_Bacteria;p_Bacteroidetes;c_Cytophagia;o_Cytophagales						
nana_2174	75.5	2.60E-171	F5XNG1	F5XNG1_Mlk_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Propionibacteriaceae;g_Propionibacterium						
nana_2174	85.4	2.30E-173	H0BIL2	H0BIL2_9Ak_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Streptomyces						
nana_2174	46.9	6.30E-80	G6FSE8	G6FSE8_9Ck_Bacteria;p_Proteobacteria;c_Deltaproteobacteria;o_Myxococcaceae;g_Myxococcus						

## 5.3 交付结果

### TnSS


 TnSS.stat.xls

 TnSS.xls

\*.TnSS.xls：分泌系统相关蛋白结果汇总

\*.TnSS.stat.xls：分泌系统相关蛋白结果统计表

### SwissProt

 SwissProt.anno.xls

 SwissProt.filter.m8.xls

SwissProt.anno.xls：SwissProt 数据库注释的结果文件（含物种注释信息）

SwissProt.filter.m8.xls：SwissProt 数据库进行 BLAST 比对结果

### TrEMBL

 TrEMBL.anno.xls

 TrEMBL.filter.m8.xls

TrEMBL.anno.xls：TrEMBL 数据库注释的结果文件（含物种注释信息）

TrEMBL.filter.m8.xls：TrEMBL 数据库进行 BLAST 比对结果

## 6. 细菌致病毒力因子 ( VFDB ) 注释

### 6.1 分析方法

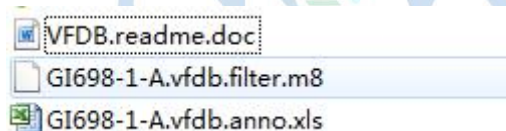
VFDB 数据库全称为 Virulence Factors of Pathogenic Bacteria，是致病细菌、衣原体和支原体的毒力因子数据库，除收录毒力基因的物种信息、基本特征描述外，还提供毒力基因功能和致病机制的详细描述。其包含26 个种，共459 个致病因子，24 个致病岛，2,505 个与毒力因子相关的基因。

使用BLAST 软件，把目标物种的氨基酸序列，与 VFDB 数据库进行比对，把目标物种的基因和其相对应的功能注释信息结合起来，得到注释结果。由于每一条序列比对结果可能超过一条，为保证其生物意义，注释时保留一条最优比对结果作为该基因的注释。最后提供的BLAST 结果为M8 格式。

### 6.2 结果展示

Gene_id	Identity	E_value	VFDB_inte	VF_id	VF_name	Related_g	Character	Structure	Functions	Mechanism	Taxonomy
nana_2175	40.21	3.00E-15	VFG0869	VF0215	Dispersin	aatC - A2	Encoded	typical	Promotes	-	k_Bacteria;p_Actinobacteria;c
nana_2175	52.11	6.00E-64	VFG1214	VF0082	Type IV	pilR - tw	PilA, B,	-	Attaches	The C-ter	k_Bacteria
nana_2180	49.11	#####	VFG2144	VF0344	TTSS	yscN - ty	Type III	-	Translocates	Chlamydia	k_Bacteria;p_Proteobacteria;c
nana_2180	40.19	4.00E-40	VFG1301	VF0003	Capsule	cap8E - c	Produced	-	Prevent	r-	k_Bacteria;p_Actinobacteria;c
nana_2180	50.83	6.00E-46	VFG1222	VF0082	Type IV	pilM - ty	PilA, B,	-	Attaches	The C-ter	k_Bacteria;p_Proteobacteria;c
nana_2182	42.33	2.00E-37	VFG1206	VF0272	FbpAEC	fbpC - ir-	-	-	Encodes	s-	k_Bacteria;p_Proteobacteria;c
nana_2187	45.83	4.00E-93	VFG1249	VF0273	Flagella	flaR - tw-	-	-	Swimming	-	k_Bacteria;p_Planctomycetes;c
nana_2189	49.15	4.00E-87	VFG0141	VF0085	LPS	waaA - li	Two disti-	-	Mediates	Binding	k_Bacteria;p_Proteobacteria;c
nana_2189	48.67	5.00E-33	VFG1824	VF0317	DevRS	devR - hyA	two-com-	-	Controls	-	k_Bacteria;p_Chloroflexi;c_U
nana_2189	47.67	1.00E-44	VFG1412	VF0006	CSH	csH -	distal	-	Translocates	-	k_Bacteria;p_Actinobacteria;c

### 6.3 交付结果



VFDB.readme.doc: 交付结果说明

\*.vfdb.filter.m8: VFDB 数据库进行 BLAST 比对结果

\*.vfdb.anno.xls: VFDB 数据库注释的结果文件（含物种注释信息）

### 6.4 参考文献

【1】Chen L, Xiong Z, Sun L, et al. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors[J].

Nucleic acids research, 2011: gkr989.

【2】 YangJ,ChenL,SunL,etal.VFDB2008release:anenhancedweb-based resourcefor comparativepathogenomics[J].Nucleicacidsresearch,2008, 36(suppl 1): D539-D542.

【3】 Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors[J]. Nucleic acids research, 2005, 33(suppl 1): D325-D328.

## 7. CAG/MLG 分析

### 7.1 分析方法

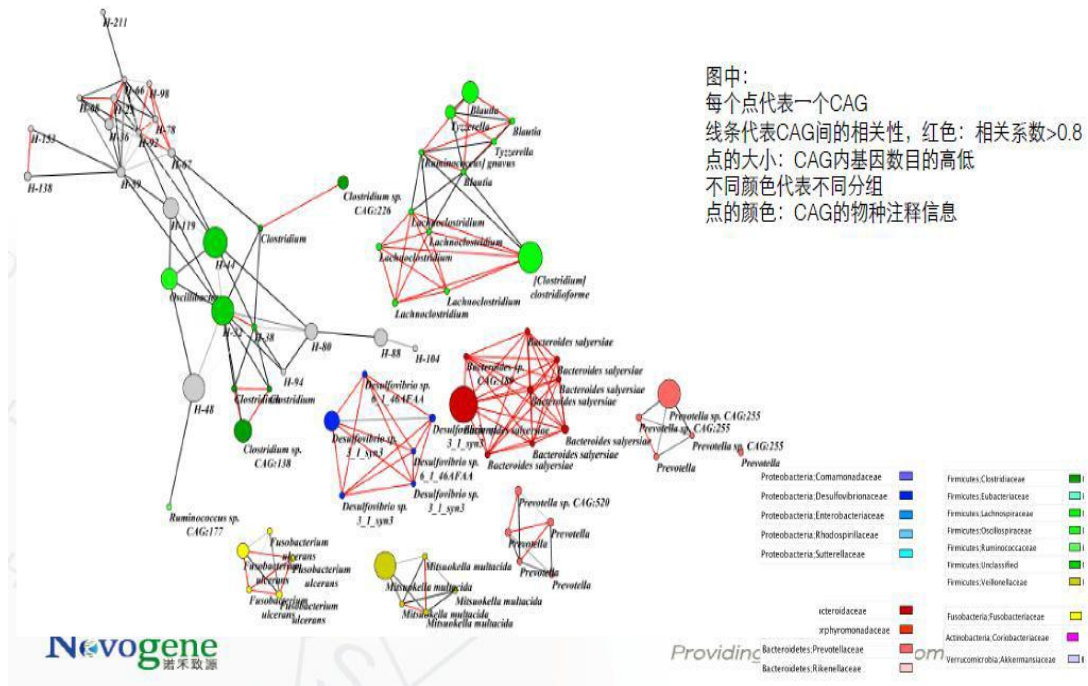
对在样品中存在的基因，根据丰度一致性进行聚类。一般认为聚为一类的基因簇是某一物种的全部或部分基因，根据 cluster 中基因数目的多少，可将 cluster 分为 CAG、MLG、MGS 等。

- 1、从基因丰度表出发，挑选显著变化的基因。
- 2、根据第一步得到的基因在各个样品间的丰度信息，计算基因与基因之间的相关性，将相关性高的基因进行聚类，最终得到cluster。
- 3、从cluster出发，挑选出CAG和MLG进行后续分析。CAG：cluster中基因数目在50及以上；MLG：cluster中基因数目在700及以上。
- 4、将CAG中的基因和reference genome及NR库进行blast比对，从比对结果出发，对CAG进行物种注释。
- 5、从CAG在各样品中的丰度信息出发，进行组间差异CAG分析，富集CAG分析。
- 6、挑选MLG，进行后续组装单菌分析。

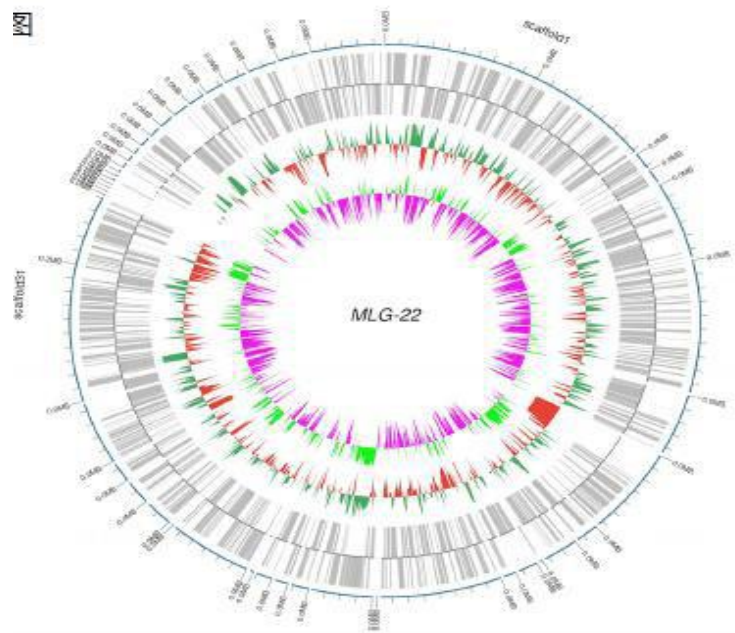
### 7.2 结果展示

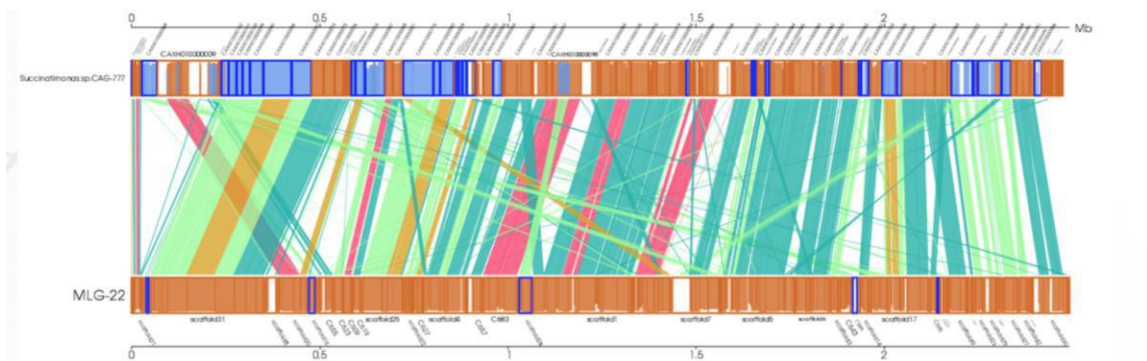
- 1、CAG 网络图





## 2、MLG 组装单菌结果展示 （基因组圈图以及共线性分析）





### 7.3 参考文献

【1】Nielsen H B, Almeida M, Juncker A S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes[J]. Nature Biotechnology, 2014, 32(8):822-828.

【2】Feng Q, Liang S, Jia H, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence.[J]. Nature Communications, 2015, 6.

【3】Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes.[J]. Nature, 2012, 490(7418):55-60.

【4】Qin N, Yang F, Li A, et al. Alterations of the human gut microbiome in liver cirrhosis.[J]. Nature, 2014, 513(7516):59-64.

## 8. CNV 拷贝数变异分析

### 8.1 分析方法

人类消化道有大量微生物，被统称为肠道微生物组。肠道微生物组在人类代谢食物、抵御感染和应答药物等过程中都发挥重要的作用。许多人类疾病如肝硬化、代谢综合症、糖尿病、动脉硬化和神经系统疾病等都与微生物组失衡有关。肠道菌群组成因人而异，最近 Sharon Greenblum 等人的研究发现，在 100 多人微生物组中，有几十个菌种五千多个基因存在着拷贝数的显著差异。这些基因虽然出现在每个目标菌种中，但接近四分之一基因都存在拷贝数变异。

#### 1、数据预处理

采用 Illumina HiSeq 测序平台测序获得的原始数据(Raw Data)存在一定比例低质量数据，为了保证后续分析的结果准确可靠，需要对原始的测序数据进行预处理，获取用于后续分析的有效数据(Clean Data)。

#### 2、数据比对和KO注释



采用BWA软件将得到的Clean Data与参考基因组聚类簇（genome cluster）进行比对，计算参基因组聚类簇的单碱基位点的测序覆盖度。采用DIAMOND软件将参考基因组聚类簇的对应的蛋白序列与KEGG数据库（<http://www.kegg.jp>）进行比对，得到参考基因组聚类簇上基因的KO注释信息。

### 3、计算拷贝数

根据参基因组聚类簇的单碱基位点的测序覆盖度和基因的KO注释信息计算KO的覆盖度，然后根据13个marker KO的平均覆盖度（看做参基因组聚类簇的菌株个数）计算特定样品中特定参基因组聚类簇上特定KO的拷贝数（Vkcs, Copy number Variation in each KO k, in each cluster c, and in each sample s）。

### 4、计算 Highly Variable KCs 和 Set-specific VariableKCs

对于至少出现在10个样品中的可检测到的KCs，计算每个KCs在所有样品中Vkcs的MAD值（绝对中位差），所有KCs的MAD值形成一个偏正态分布，计算得到分布的平均值和标准差。

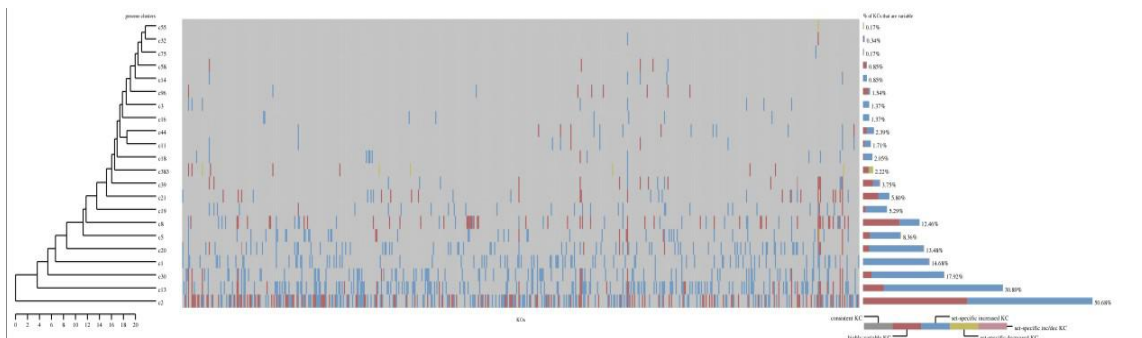
以偏正态分布的平均值加上2倍的标准差为阈值（T1），当KCs的MAD值高于此阈值（T1）时，则定义该KCs为Highly Variable KCs；对于特定KCs，以所有样品Vkcs的中值加上T1为阈值（T2），当Vkcs超过此阈值（T2）的样品数超过所有的样品的10%，则定义改KCs为Set-specific increased Variable KCs；对于特定KCs，以所有样品Vkcs的中值减去T1为阈值（T3），当Vkcs低于此阈值（T3）的样品数超过所有的样品的10%，则定义改KCs为Set-specific decreased Variable KCs。

### 5、计算 Host-state-associatedKCs

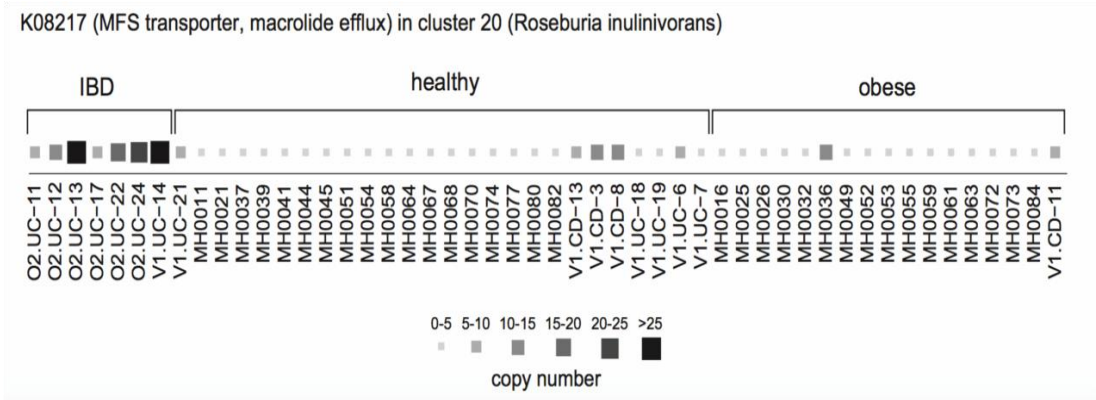
对于不同组间样品的Vkcs进行wilcox检验，计算p值和FDR校正后的p值（q值），从而筛选出组间显著性差异的KCs，即Host-state-associated KCs。

## 8.2 结果展示

### 1、 Heatmap of Variable KCs



2、 Host-state-associated KCs 分析结果展示



8.3 参考文献

【1】 Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. Cell 2015;160:583-94.

【2】 Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature 2013;493:45-50.

【3】 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 2014;42:D199-205.

9. 肠型分析

9.1 分析方法

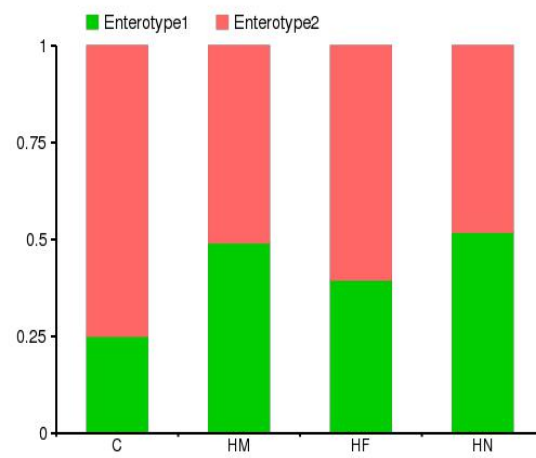
肠型是一种宿主-微生物共生稳态时的微生物群落结构，能够体现共生微生物协助宿主适应环境的能力；肠型的形成主要受到宿主长期形成的生活习性尤其是饮食习惯的影响，能够在一定程度上体现宿主可能存在的健康或疾病风险，因此明确肠型对于准确发掘与疾病相关的肠道功能菌以及疾病风险的预测和个体化治疗具有重要作用。

分析流程主要含有以下分析点：

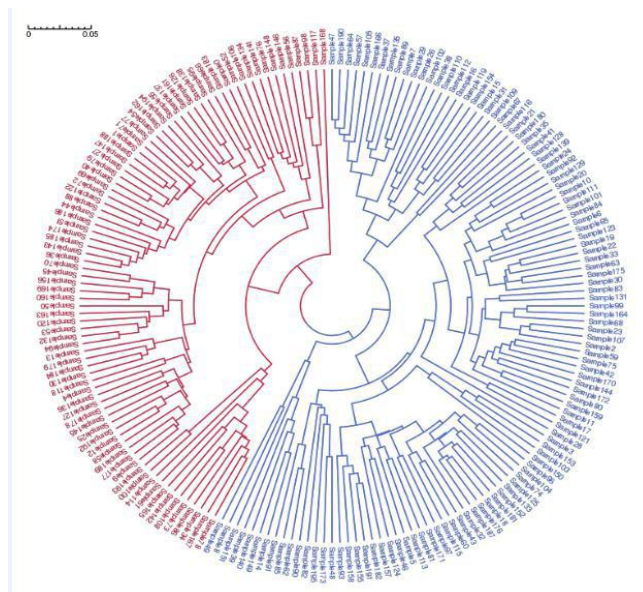
- 1、肠型的确定及肠型中各个表型的分布情况统计
- 2、基于肠型的降维分析——PCoA/PCA 分析
- 3、基于肠型的显著性差异物种分析

9.2 结果展示

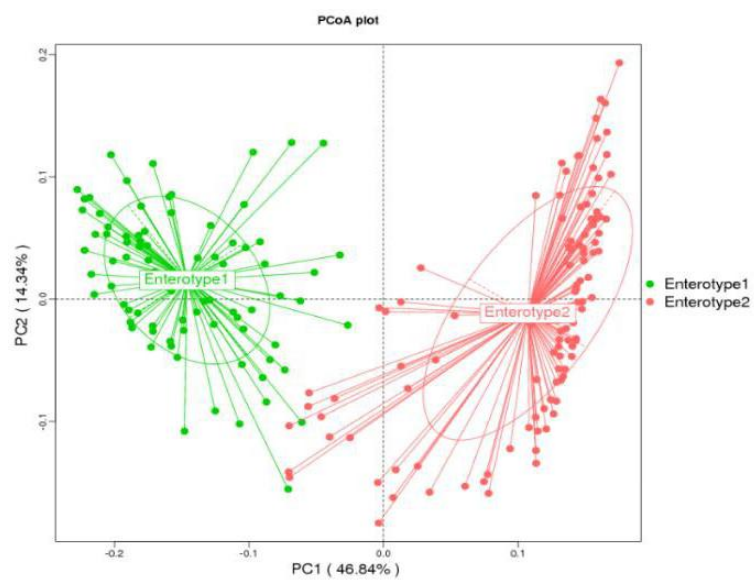
1、 肠型中各个表型的分布图



2、样品聚类图



3、PCoA 图



### 9.3 参考文献

- 【1】 Manimozhiyan A, Jeroen R, Eric P, et al. Enterotypes of the human gut microbiome.[J]. Nature, 2011, 473(7346):174-180.
- 【2】 Wu G D, Christian H, Kyle B, et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes[J]. Science, 2011, 334(6052):105-108.
- 【3】 Moeller A H, Degnan P H, Pusey A E, et al. Chimpanzees and humans harbour compositionally similar gut enterotypes.[J]. Nature Communications, 2012, 3(6):542-555.
- 【4】 Bresson F, Maury L, Bonniec P L, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice[J]. Neuropsychologia, 2014, 111(26):E2703-E2710.

## 10. 噬菌体分析

### 10.1 分析方法

病毒是肠道微生物的重要组成部分，而噬菌体是病毒中种类最多的一部分，但是至今人们对肠道微生物中的病毒和噬菌体研究甚少。噬菌体在肠道微生态中起着非常重要的作用，它们通过侵染细菌等肠道微生物，进而影响菌群的繁殖以及功能，从而改变着肠道微生物的组成和功能。

1、基于非冗余基因集的噬菌体注释；

- 1) 基于宏基因组分析的标准流程得到非冗余的基因集；
- 2) 采用PSIBLAST将非冗余的基因集比对到POGs(噬菌体同源蛋白数据库)中, 进行噬菌体的注释；
- 3) 最终结合基因的丰度计算每种噬菌体在各个样品中的丰度, 并进行PCA、heatmap、cluster、Metastats等统计学分析, 挖掘噬菌体在样品间及组间差异信息。

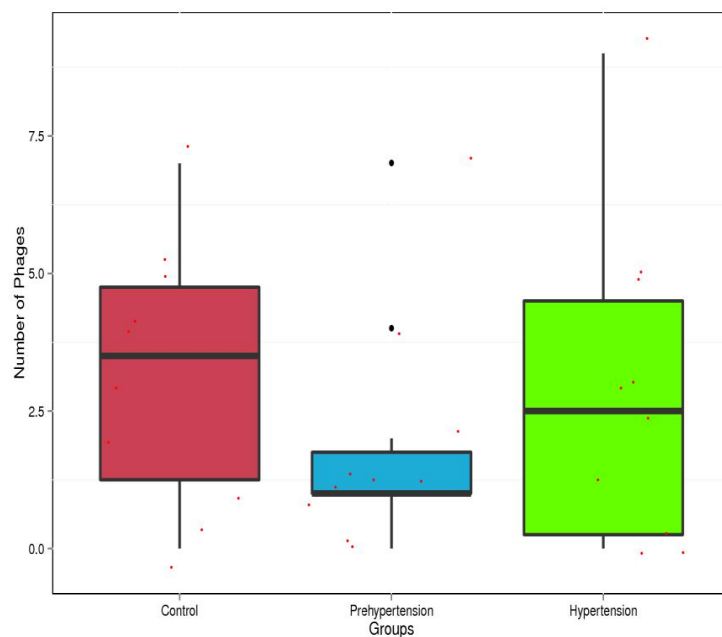
2、基于scaffigs的前噬菌体预测及分析：

- 1) 基于组装后得到的每个样品的scaffigs信息, 结合每条scaffigs上的基因预测结果, 采用Phage\_Finder2.1进行前噬菌体区域的预测；
- 2) 将预测结果中噬菌体区域的基因序列与POGs数据库的注释确定其噬菌体的信息；
- 3) 噬菌体区域上下游的序列与NT库进行blast比对后, 采用LCA算法注释确定其宿主的信息；
- 4) 将预测得到的噬菌体区域的基因序列与ARDB、VFDB数据库进行比对, 寻找噬菌体区域的抗生素抗性基因与毒力因子；

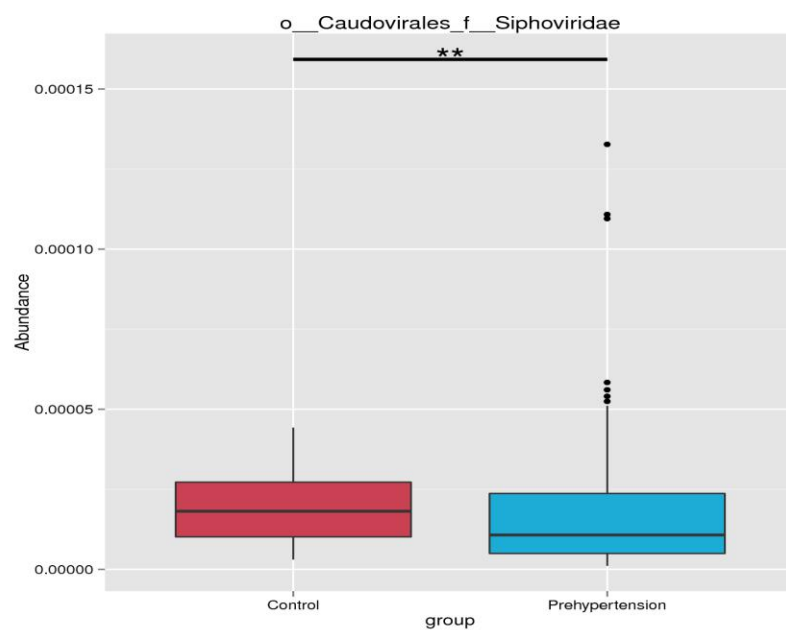
5)根据噬菌体的丰度和宿主的丰度计算PtoH值,并针对噬菌体的丰度以及 PtoH值 进行箱图、热图等统计学分析。

## 10.2 结果展示

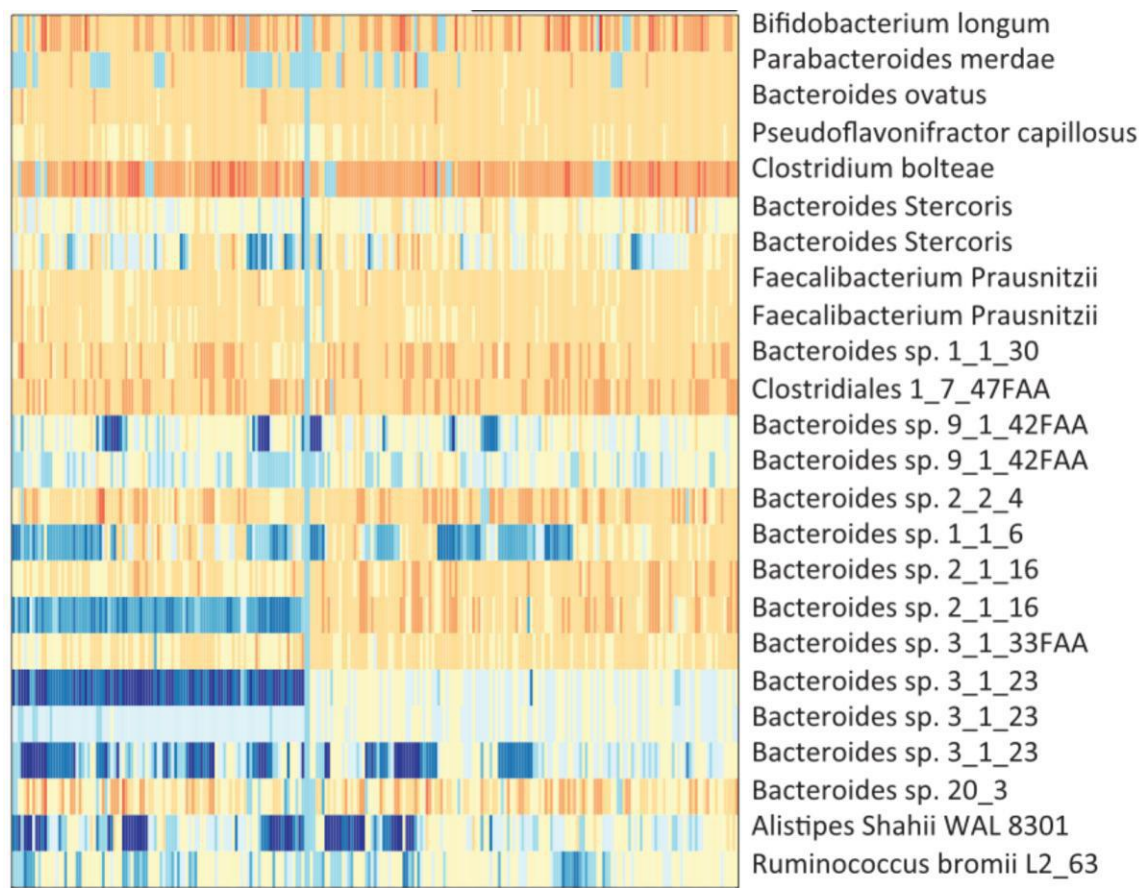
### 1、组间噬菌体数目箱图



### 2、显著性差异噬菌体分析



3、基于PtoH值的热图分析



10.3 参考文献

【1】 Vijay-Kumar M, Aitken JD, Carvalho FA, Cullender TC, Mwangi S, Srinivasan S, et al. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science 2010;328:228-31.

【2】 Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med 2011;3:14.

【3】 Iida N, Dzutsev A, Stewart CA, Smith L, Bouladoux N, Weingarten RA, et al. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. Science 2013;342:967-70.

【4】 Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 2012;490:55-60.