微生物基因组测序结果解读及方案设计

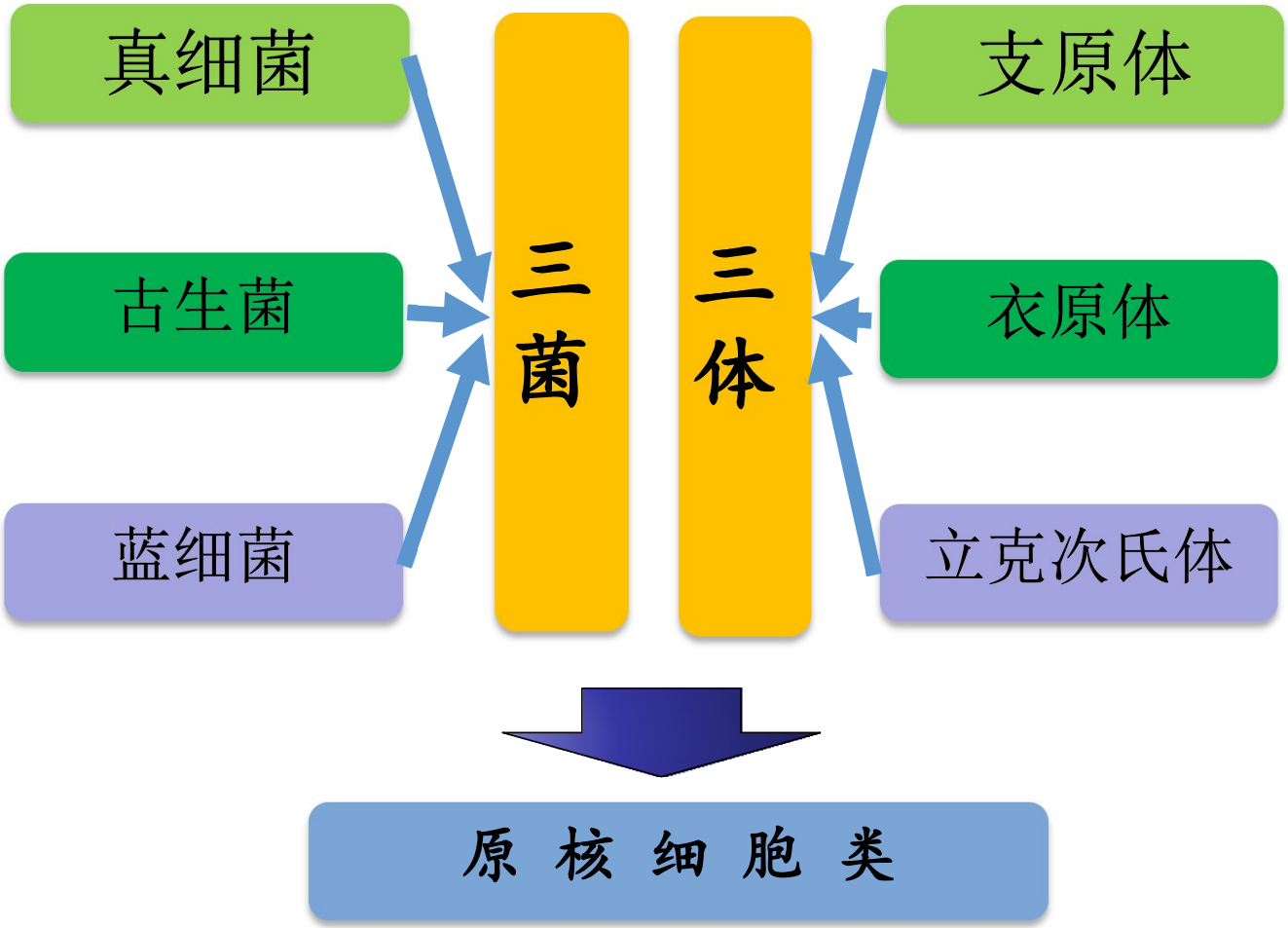
王梅 中级产品工程师
微生物基因组产品部
mdna@majorbio.com
021-31050579

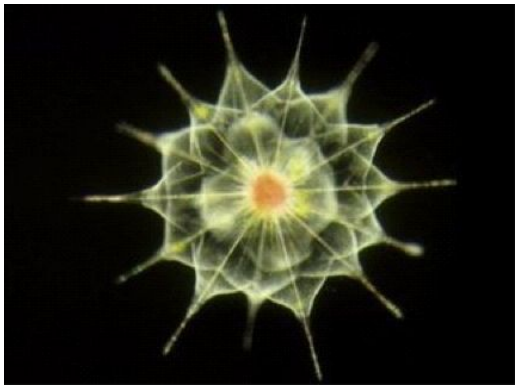
内容

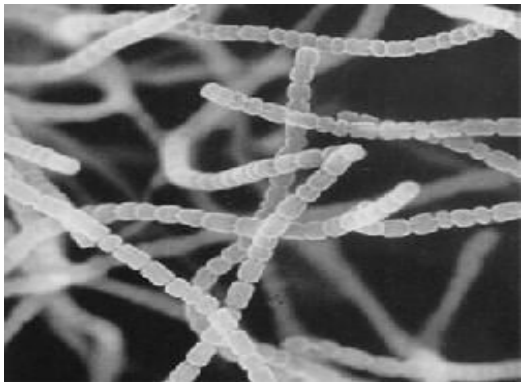
1. 组学层面微生物学研究的必要性
2. 微生物基因组测序技术路线与流程
3. 微生物全基因组测序结果解读
4. 微生物基因组研究方案设计

1 组学层面微生物学研究的必要性

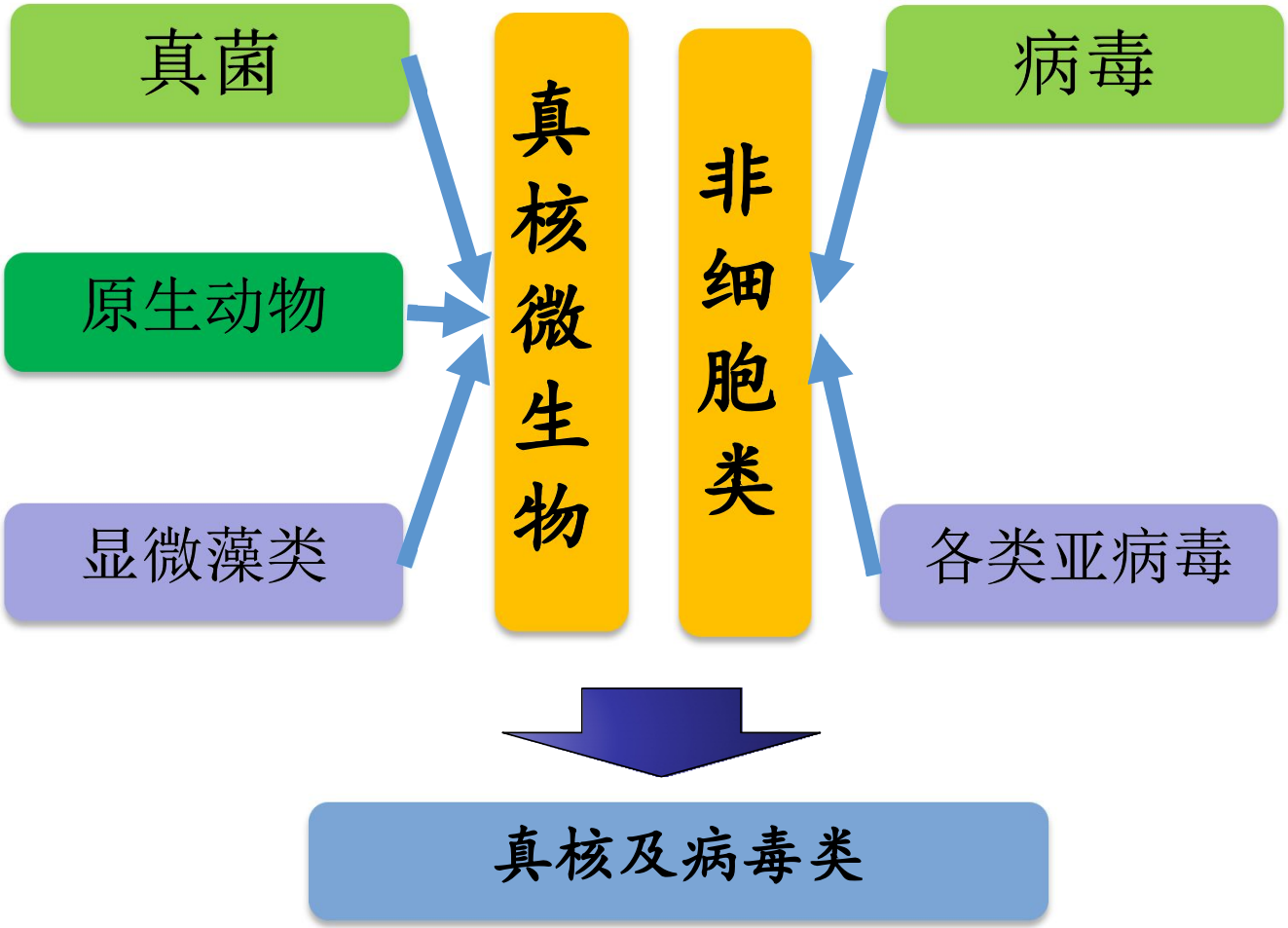
无处不在的微生物：分类

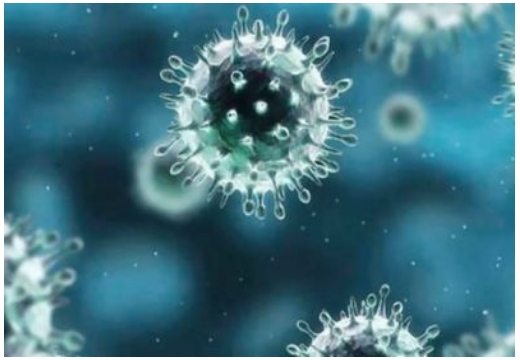




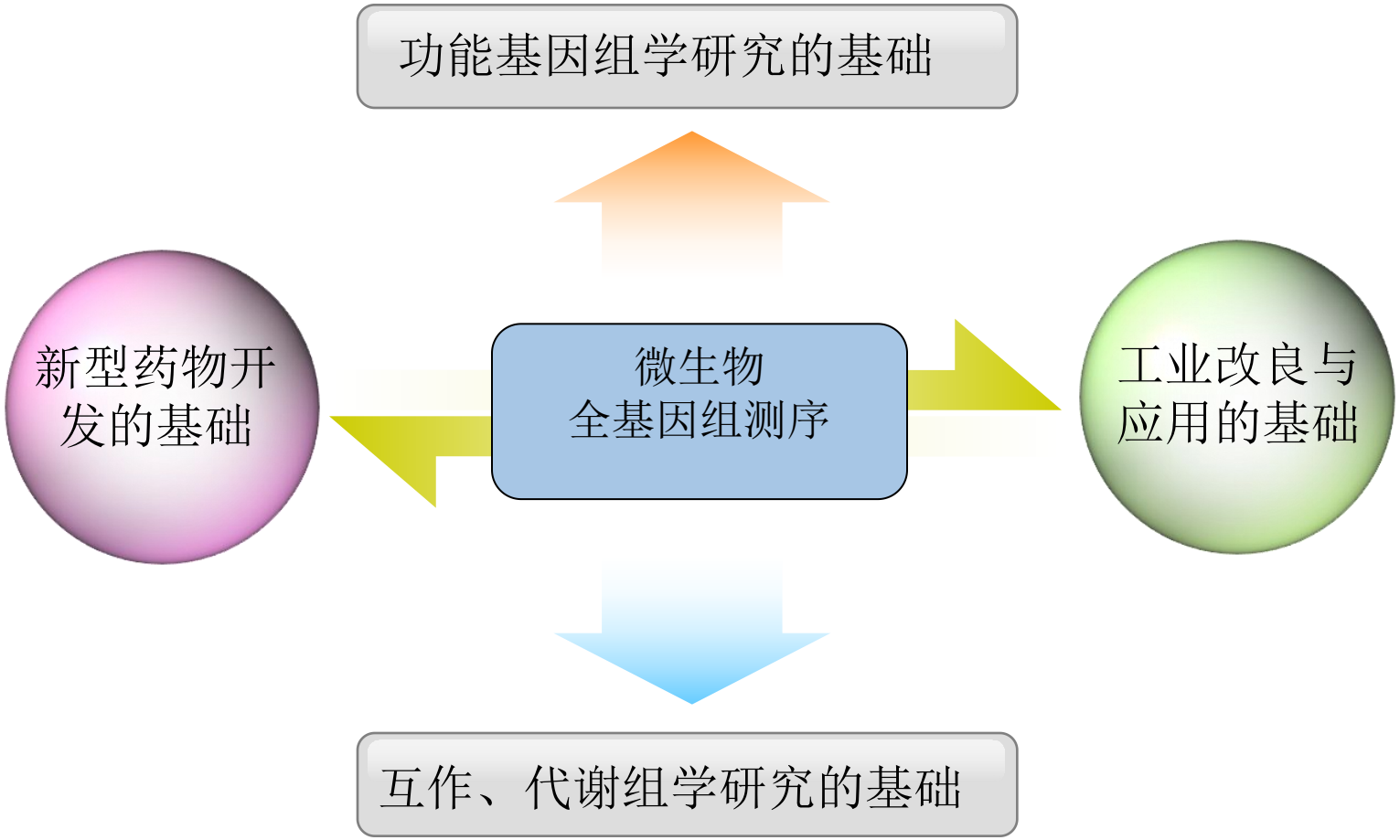


无处不在的微生物：分类





微生物全基因组测序已成基石研究



国际和国家层面上的重大测序项目

1990年
人类基因组计划启动
(human genome project, HGP)

1994年
美国能源部：微生物基因组计划
(mocrobial genome project, MGP)

2008年
美国国立卫生研究院：
人类微生物组计划
（human microbiome project, HMP）

欧盟：人类肠道宏基因组计划
（metagenomics of the human intestinal tract

2011年
多国：地球微生物计划启动
(earth microbiome project, EMP)

2013年
多国：十万食源性病原微生物基因组计划

2016年
美国国家微生物组计划
（national microbiome initiative, NMI）

2017年
中国国家微生物基因组计划

美吉生物
Majorbio

2 微生物基因组测序技术路线与流程

三个基础概念：

细菌基因组扫描图

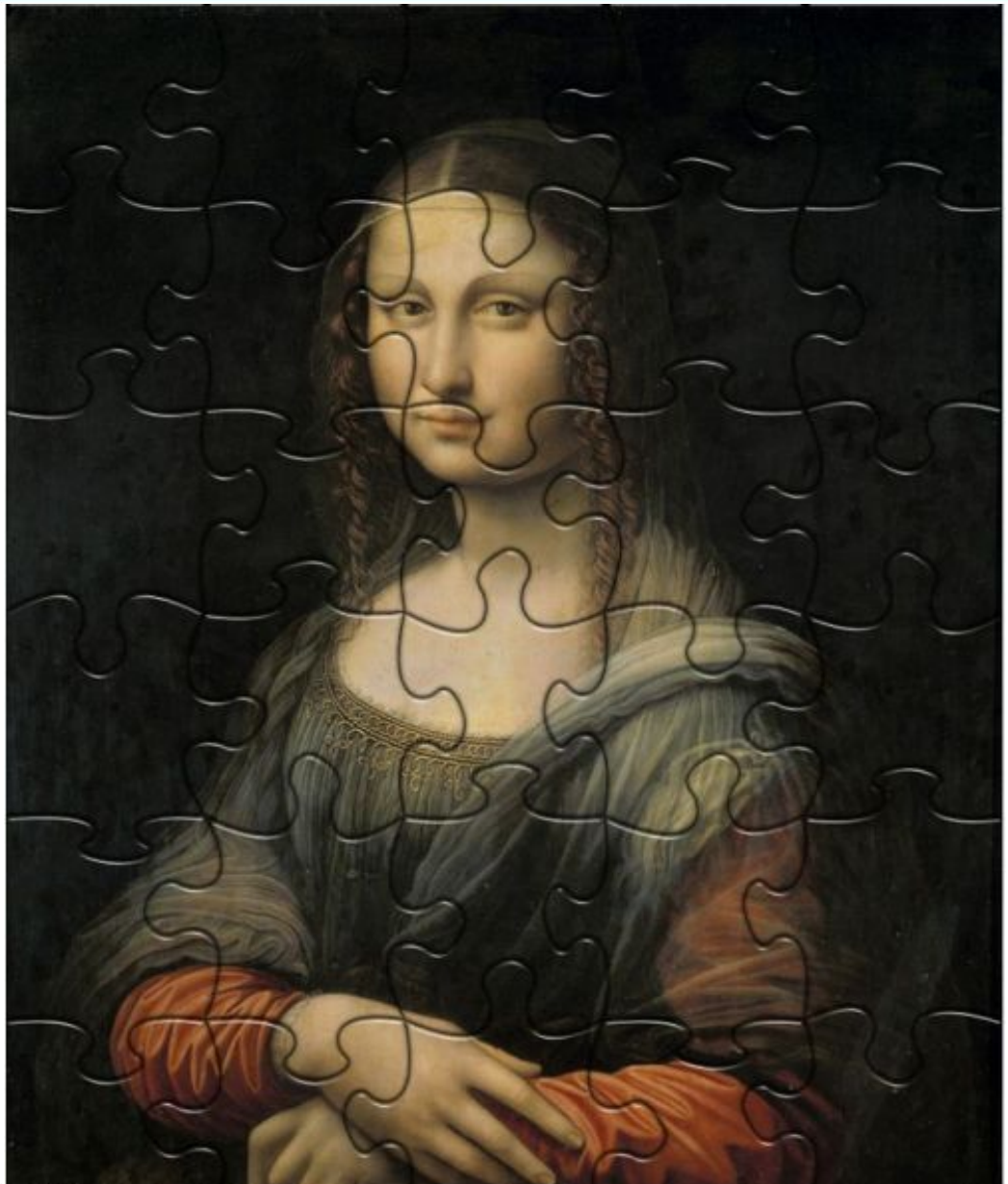
- 组装得到基因组片段（**scaffold**）；
- scaffold之间顺序不定；
- 存在**Gap**；

细菌基因组精细图

- 组装得到**scaffold**；
- scaffold更长，数目更少；
- scaffold顺序不确定
- 存在**Gap**；

说明：细菌基因组精细图概念是在测序早起提出的，由于当时技术限制，扫描图的结果很差，完成图很难达到，所以就出现了中间产品—精细图，在当前技术水平下，扫描图水平已经很高，且完成图的获得相对容易，很少提及精细图

细菌基因组完成图



- 0 Gap；
- 获得全长序列和全部基因序列；
- 单碱基错误率小于十万分之一。

测序技术路线与流程

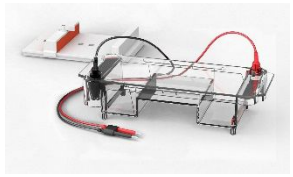
A 质控检测　B 片段化和建库　C 上机测序　D 数据分析

纯度检测

NanoDrop分析检测核酸浓度，分析OD260/OD280（1.8-2.0）以及OD260/OD230的吸光值比例，判定纯度

完整性检测

利用琼脂糖凝胶电泳检测DNA样品的片段大小和范围

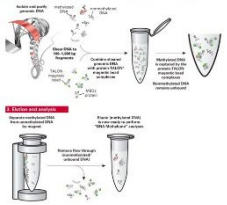


总量检测

利用Qubit/TBS-380精准检测双链DNA的含量

富集检测

检测磁珠对特定长度片段的吸附能力



测序技术路线与流程

A 质控检测 | B 片段化和建库 | C 上机测序 | D 数据分析

基因组片段化

根据基因组完整性情况和测序需要，利用 G-tubes 方法将基因组DNA 处理成 8-20 kb 的片段应用于PacBio测序；利用Covaris M220将基因组DNA处理成400 bp左右的短片段用于Illumina测序。

末端修复

PacBio测序末端补平，片段两端分别连接环状单链得到一个套马环结构；Illumina测序末端加A补平。

片段筛选

根据测序需要，筛选特定长度的片段用于上机测序。

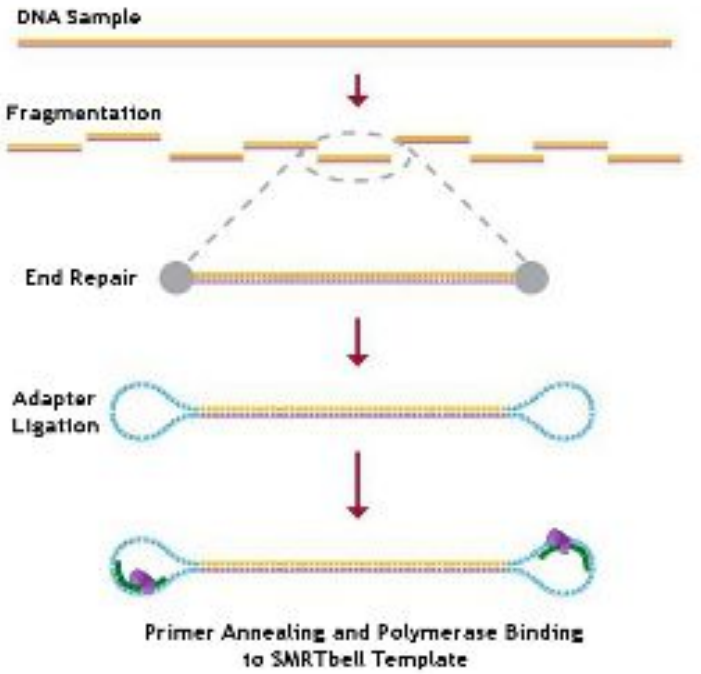
测序技术路线与流程

A 质控检测

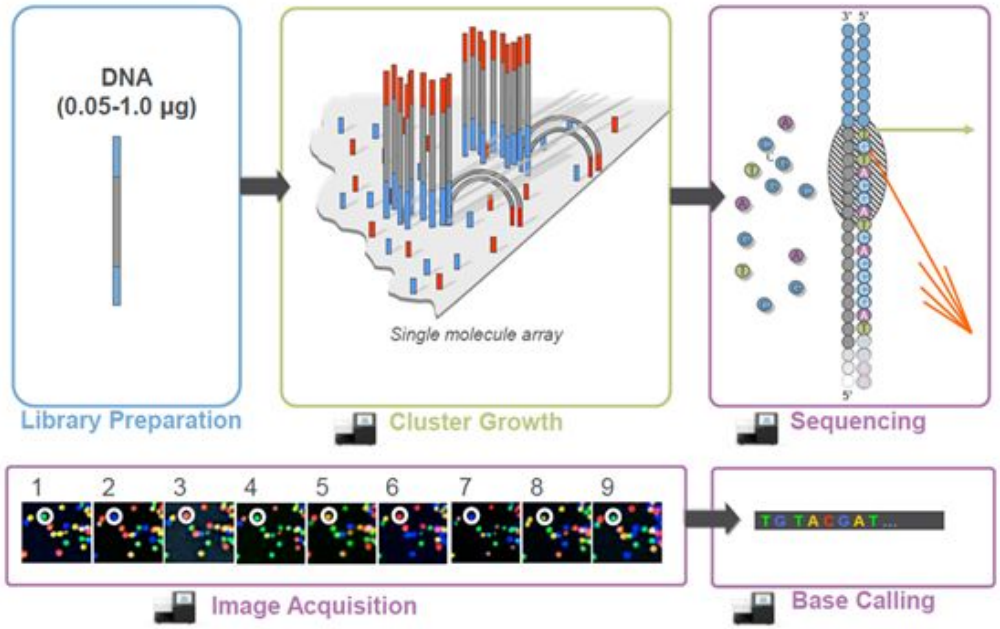
B 片段化和建库

C 上机测序

D 数据分析



PacBio三代测序



Illumina二代测序

测序技术路线与流程

A 质控检测

B 片段化和建库

C 上机测序

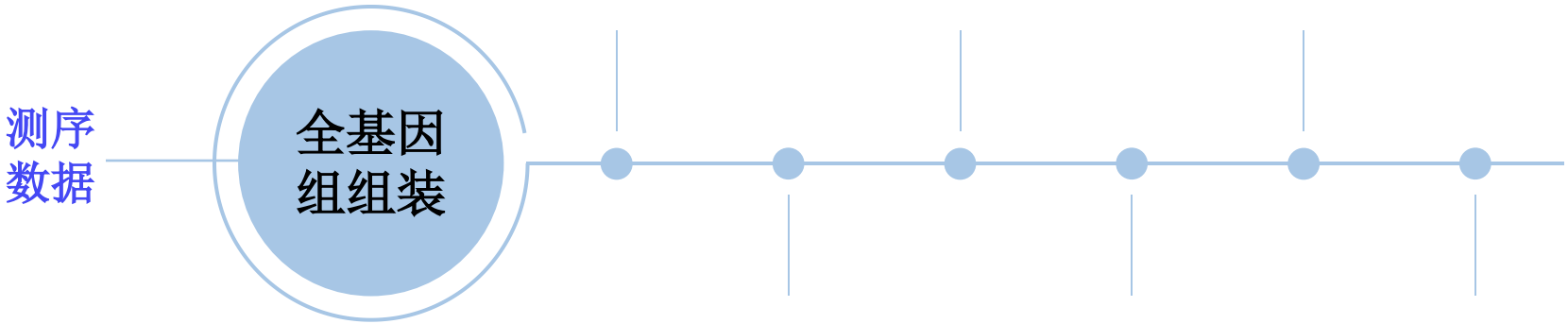
D 数据分析

标准分析	高级分析I	高级分析II	个性化分析I	个性化分析II	个性化分析III
原始数据统计与处理	GI预测分析	同源基因分析	耐药基因注释	SNP、InDel检测及注释	其他个性定制
基因组组装	CRISPR分析	基于全基因组的进化树分析	毒力基因注释	SV检测和注释	
基因预测	前噬菌体序列分析	共线性分析	碳水化合物相关酶数据库注释		
重复序列分析	圈图绘制	泛基因组分析	转座单元预测与分析		
tRNA&rRNA预测	假基因预测	ANI分析	分泌蛋白预测分析(信号肽和跨膜螺旋)		
Nr/Swiss-prot注释			基因簇比较作图分析		
COG/GO/KEGG注释			数据上传NCBI数据库		

美吉生物
Majorbio

3 微生物基因组测序结果解读

原核基因组分析流程



基因组注释

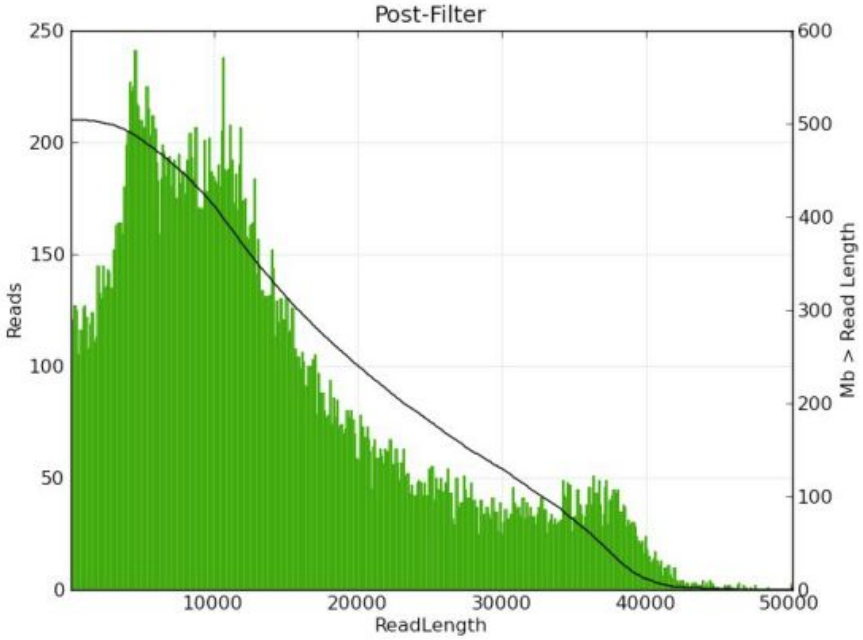
① nr库注释
② KEGG注释
③ COG注释
④ GO注释
⑤ ncRNA注释
⑥ 重复序列分析
⑦ 插入序列分析

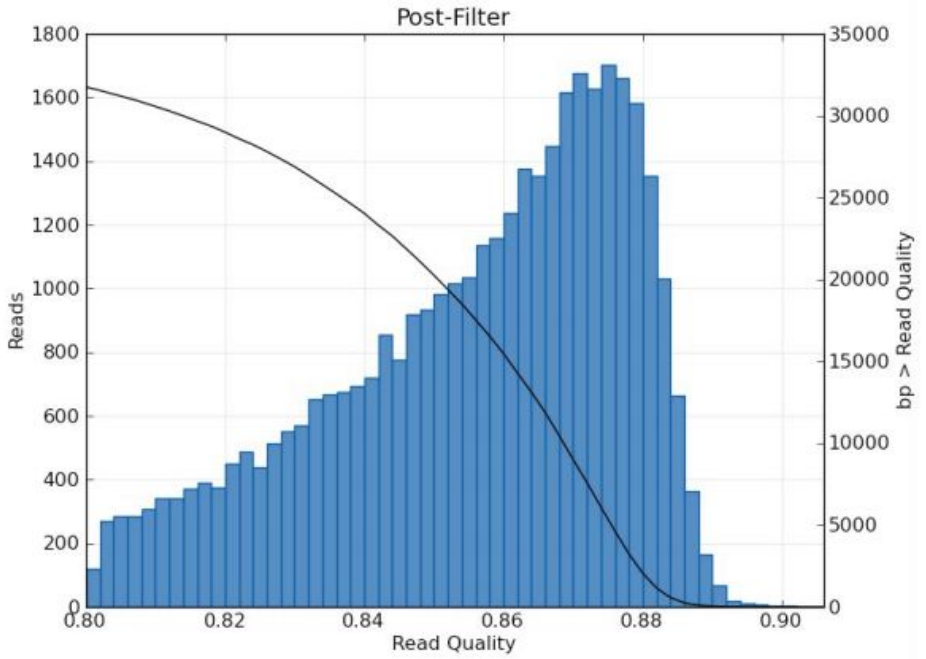
病原系统分析

① 基因岛分析
② 抗性基因分析
③ 毒力基因分析
④ 互作基因分析
⑤ 病原体生物攻防因子分析
⑥ 双组分分析
⑦ 群体感应分析

比较基因组分析

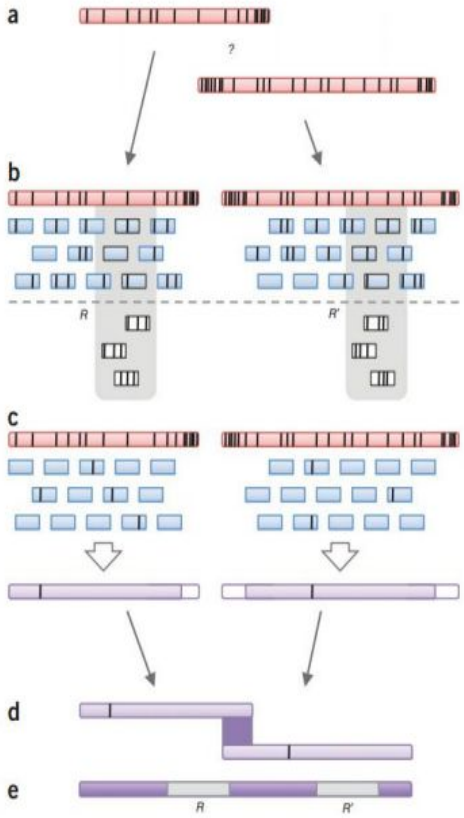
① 共线性分析
② 进化树分析
③ 泛基因组分析
④ GWAS分析
⑤ ANI、TNI分析
⑥ SNP、InDel分析
⑦ 群体进化分析

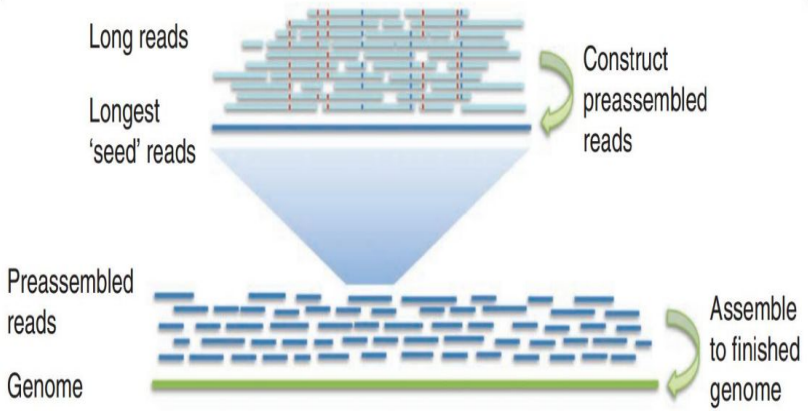




- 单分子测序原始数据在组装前，每条 read 都经过质量过滤。有两种校正方式，一是利用PacBio自身的数据进行校正，二是利用Illumina二代测序结果进行校正。一般情况下，测序深度达到70×以上时，PacBio数据可以较好地完成自我校正。
- 测序reads 的长度是评估三代测序优劣的重要指标，上图为单分子测序 Clean 数据 reads 的长度和质量分布统计图。

利用HGAP、soapdenovo等多种组装软件及算法进行组装，调整参数，选取最优组装结果。





（方法参考：Chin C S, Alexander D H, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data[J]. Nature methods, 2013, 10(6): 563-569.）

基因组组装是整个流程中最重要的部分，是后续分析的保障，也是最难的部分。对于每一个项目，我们会进行多次组装，并对结果进行人工校正，

No.2--基因组组装

- 一般情况下，单独利用PacBio三代测序数据就可以组装出一个完整的细菌基因组。有些细菌基因组较为复杂，特别是含有多个内源质粒的基因组，使用Illumina + PacBio相组合的方式可以获得更优的组装结果。

Category	Property
No. of all scaffolds	1
Bases in all scaffolds	4,793,207 bp
G+C content	64.03%
N rate	0
No. of all contigs	1
Bases in all contigs	4,793,207 bp
No. of large contigs(> 1000 bp)	1

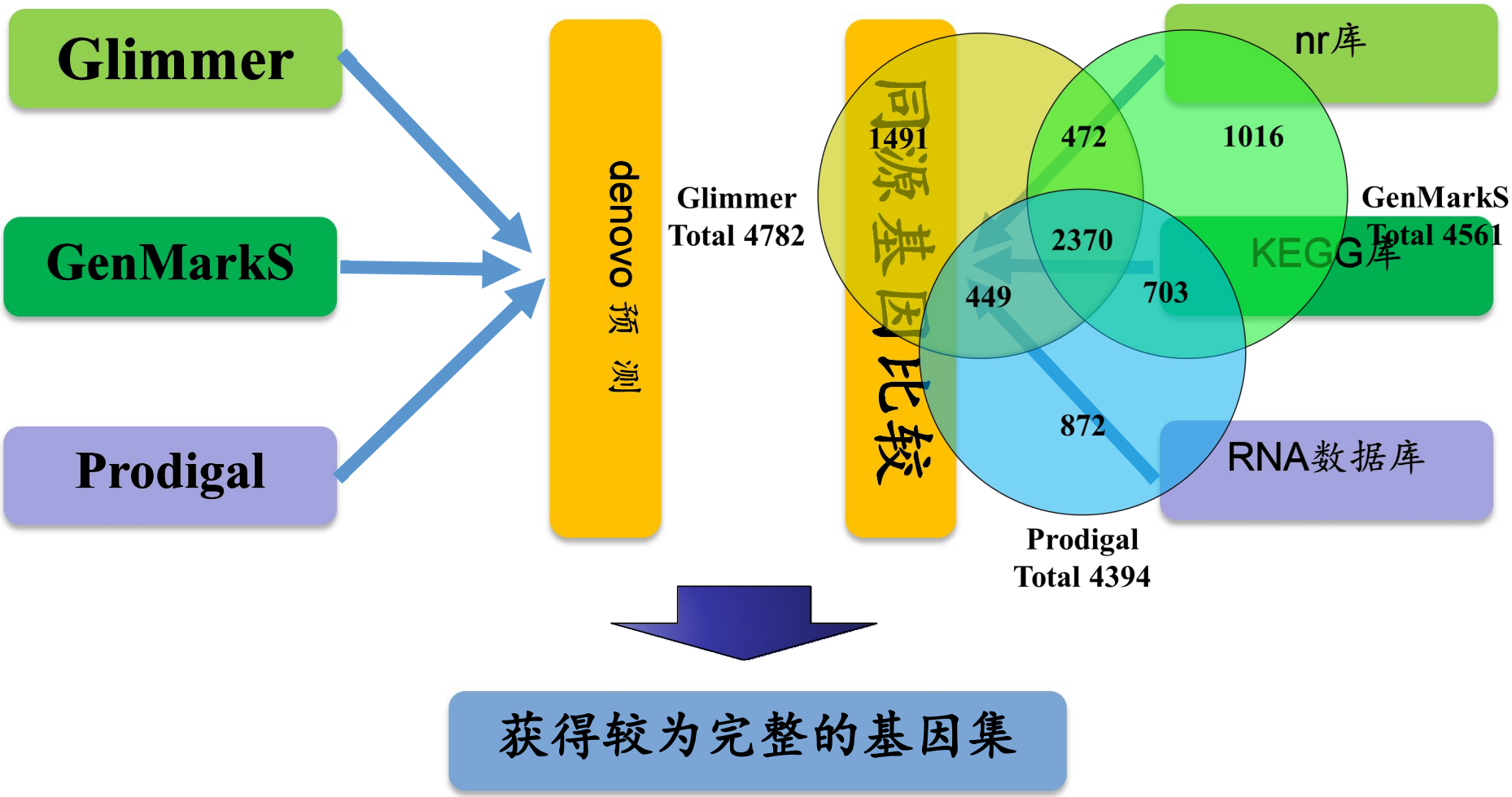
No.2--基因组组装

组装结果：

- 位置：assembly文件夹，.fna序列文件
- 标准的FASTA格式文件
- 完成图组装得到1条scaffold，扫描图组装得到多条scaffold
- 建议使用notepad++软件打开

```
>scaffold 1
tatggccttcaacaacatttttgggattttaggcgttttgccgaagcgccgagagcagct
ctctgcagcttaattcggtaaaatgtcaggatttaaaaggagtgacgtctgtggacagcc
atacctctgaactatggcagcaaattctatccatcatacaaaccaagctgagtaagccga
gttacgacacttggtttaaggctaccaaggcagcgaaactaaatgaccactccattgtga
tttctgcaccgacaacttttgccgtggaatggcttgaaagccgctataccaagctagtcg
gagcaacggtttatgaaatttgggcaaacaactagaggtcaagttcgttattgaagaga
acaagcccgctgaggtcgaccttcagcaacaacctcagcagcagccggtcgttcatgaag
aagctgtgtcccatatgctgaatcccaaatatacattcgatacattcgtcatcggatcgg
ggaaccgttttgcccatgcggcatcgctggccgtcgccgaggcgccggcaaaagcttaca
atccgctgtttttgtacggtggtgtggggctgggaaaaacgcatctgatgcacgctatcg
gacactatattttggagcacaatccgaccagcaaggtcgtttatttatcgtcggagaagt
ttacgaatgaattcattaatgccatccgggacaaccgcggggaaagtttccggaataaat
atcgcaacattgatattttgctcattgatgatattcaattcattgcgggcaaggaatcga
cgcaggaggaatttttccacacgttcaatgcgcttcatgaggaacgcaagcagattataa
tctcaagcgatcggccgcctaaagaaattccaacgctggaagaacggctgcgctctcgct
tcgagtggggacttattacggatattcaaccgccagatctggagacgagaattgctattc
ttcggaaaaaggcgcgggcggaaaacctggatattcctaatgaggccatgatgtatatcg
ctaatcaaattgatacaaacatccgtgagctggaaggggcgcttattcgcgttgtcgctt
attcctccttaaccaatcaggatgtcacaagtcatctcgcagctgaggcgttgaaagata
ttatcccgtccagtcgtccaaaaatgatcacgattcaggatatacagcatcaagtcgggg
aattttacaatctacggttggaggattttaaagcgcgtaagcggacaaaggctgtagctt
ttccacgacagattgccatgtacctgtctcgtgagctaaccgactattctttgccaaaaa
tcggcgaagctttcggtgggcgtgatcatacgactgtcattcatgcgcacgaaaaaatct
ccaaatccattcaagtggatcaggatctgtttaaagttatcaacagcctaatagaaaaaa
tcaaaaatccaacctgaataagttcagagcctatacacaatttatacacatgtggatagg
cttgattttattagggttacagtgacttatccacatattcagtgcgcctattactattac
taatatattttaaagatatacatcacctaaataaggcccgcgtcaacgcgtgcttgaagg
attaaaacccgtaggaggaacttatgaagatcagcattctgaaaaacgttttgaacgagg
```

No.3--基因预测


Glimmer
GenMarkS
Prodigal
denovo 预测
同源基因比较
Glimmer
Total 4782
1491
472
1016
2370
449
703
872
nr库
GenMarkS
Total 4561
KEGG库
RNA数据库
Prodigal
Total 4394
获得较为完整的基因集

No.3--基因预测

预测结果：

预测基因的位点信息

gene	start	end	strand
orf_0001	111	1457	+
orf_0002	1644	2786	+
orf_0003	2834	3052	+
orf_0004	3107	4222	+

预测基因的核酸序列

```
>orf00001 50 1447 scaffold1 50 1447
ATGCTTTGGACAGACTGCTTAACTCGCTTGCGACAAGAGCTCTCT
>orf00002 1545 2693 scaffold1 1545 2693
GTGCGTTTGAAAATCGCTAAAGAAAGTTTACTCAATGTTTTATCG
>orf00003 2708 3790 scaffold1 2708 3790
ATGCATCTTACGCGCTTAAATATTGAACGTGTGCGTAATTTAAAA
>orf00004 3843 6311 scaffold1 3843 6311
ATGAGTTCAGAGTCTCAATCAGCCTCTCAAACAGAACAAACCAAT
```

预测基因的氨基酸序列

```
>orf00001_1 50 1447 scaffold1 50 1447
MLWTDCLTRLRQELSDNVFAMWIRPLVAEEVEGILRLYAPNPYWTRYIQEN
EQLSEGRVRQVEILVDSRPGSILSSSEQPATTTAALQTAPIPQPTKVKREP
SKSSKKKLLNPQFTFSLFVEGRSNQMAAETCRKVLTQLGASQHNPLFLYGP
QAVGNALLQAKPNARVMYMTSESFVQDFVSSLQKGKVEEFKKNCRSLDLLL
KEASLVEFFYTFNALLDESKQIILTSDRYPKELTELDPRLVSRFSWGLSVG
IEILLKKAENSGVDLPRNCALFIAQQVVANVRELEGALNKVVAISRFKGAP
LKDVLAIRARTISVENIQRVVSEYFRIPLKELVGPKRTRIYARPRQLAMGL
```

- 使用BLAST 2.2.28+软件对预测基因的蛋白序列分别进行 nr数据库功能注释、KEGG数据库功能注释、Swiss_Prot数据库功能注释、 GO 数据库功能注释， 汇总上述各个数据库的注释信息，导出在 annotation.table.xls文件中。

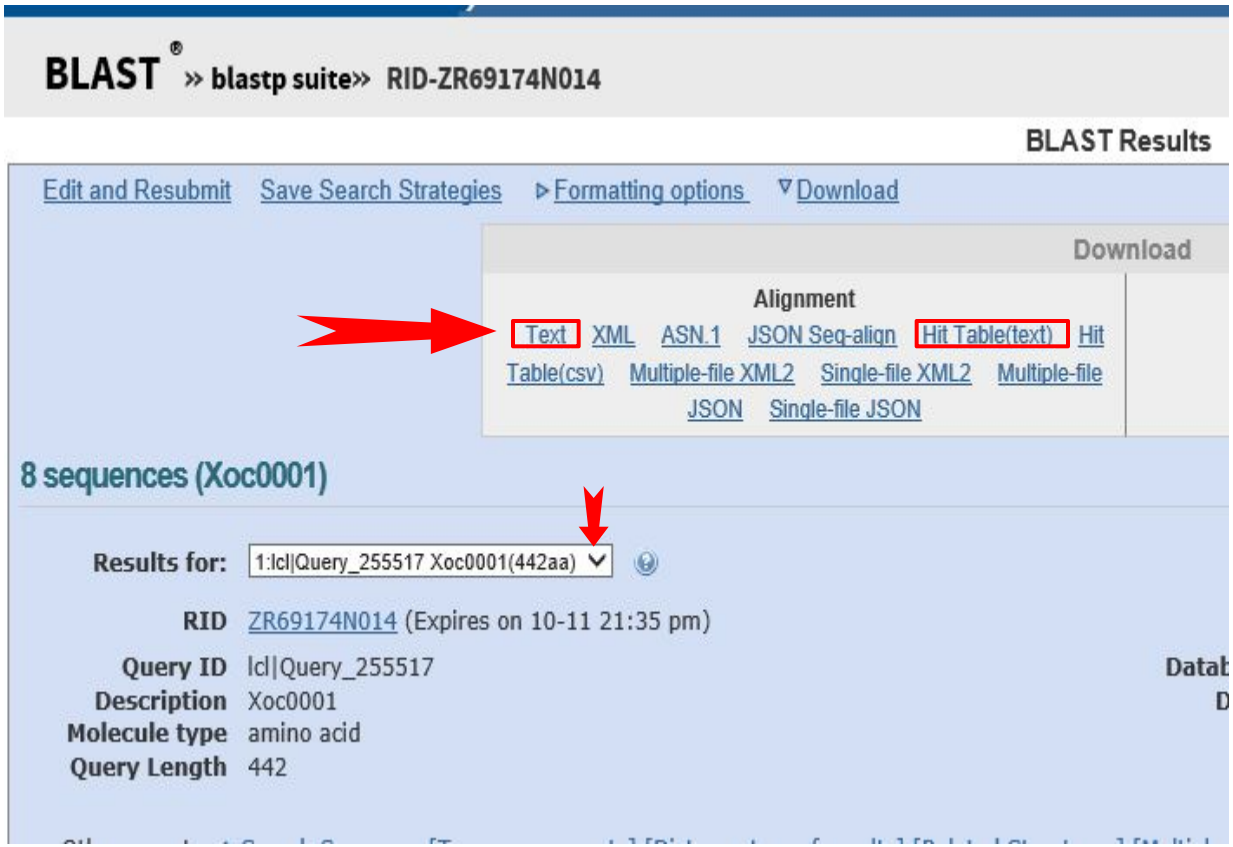
annotation.table.xls 文件各列含义说明

Query name	Query length	NR-tophit description	NR-tophit similarity	Swissprot description	Swissprot similarity	GO	COG	KO/gene	KEGG-gene-name
orf1	432	Chromosomal Replication initial	80%	RecName：dnaA	78%	GO：000016	COG0593	KO2313	dnaA
orf2	367	……	……	……	……	……	……	……	……
orf3	805	……	……	……	……	……	……	……	……
orf4	225	……	……	……	……	……	……	……	……

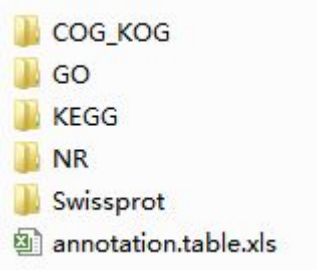
No.4--基因注释

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA注释
- 重复序列分析

https://blast.ncbi.nlm.nih.gov/Blast.cgi



基因注释结果文件：



Nr注释结果

score	HSP-length	%-Identity	Query-name	Q-length	Q-begin	Q-end	Hit-name	description
2346	465	100%	orf1	465	1	465	gi|446887	Chromosomal Replication initial
orf2	……	……	……	……	……	……	……	
orf3	……	……	……	……	……	……	……	
orf4	……	……	……	……	……	……	……	

No.4--基因注释-KEGG

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA注释
- 重复序列分析

http://www.kegg.jp/blastkoala/

BlastKOALA

Result

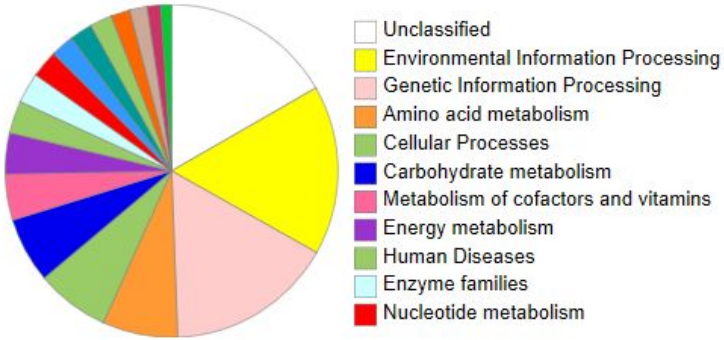
Your BlastKOALA job Help

Query dataset: 4953 entries
Taxonomy group: (Taxonomy ID: xanthomonas)
KEGG database searched: family_eukaryotes.pep,genus_prokaryotes.pep
Job submitted: Fri May 6 17:27:36 JST 2016
Job completed: Fri May 6 21:37:30 JST 2016

Annotation data View | Download

Summary 2184 entries (44.1%) annotated

Functional category

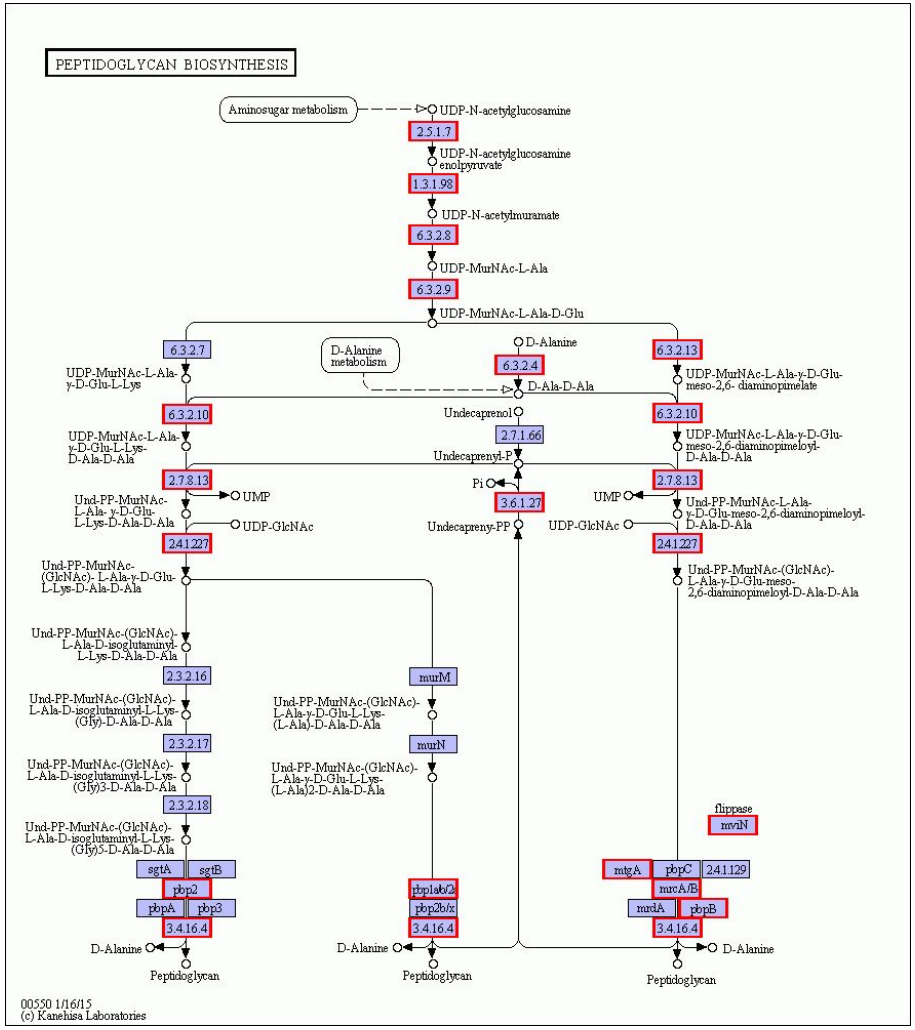


See color codes

KEGG Mapper Reconstruct Pathway
Reconstruct Brite
Reconstruct Module

KEGG pathway分类

- KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因和基因组百科全书）是目前最完善的代谢通路分析数据库。
- 基因产物并不是孤立存在而各自发挥作用的，不同基因产物之间通过有序的相互协调来一起行使具体的生物学功能。
- KEGG 数据库中丰富的通路信息将有助于我们从系统水平去了解基因的生物学功能，例如代谢途径、遗传信息传递以及细胞学过程等一些复杂的生物过程。



KEGG注释结果：

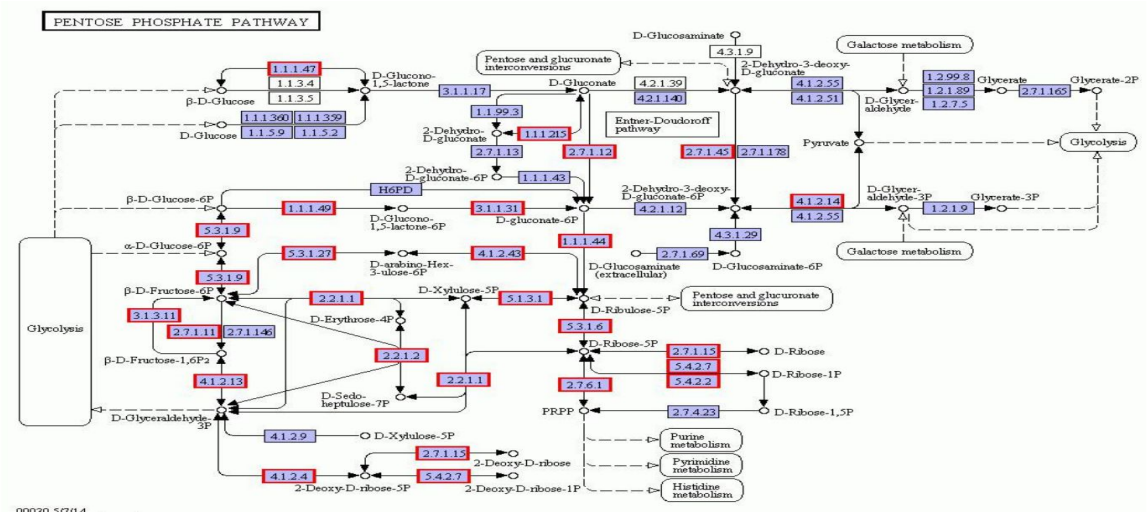
每个orf的注释结果

#Query	Ko id(Gene id)	Ko name(Gene name)	hyperlink	Paths
orf00001_1	K02313	dnaA	http://www	path:ko02020;path:ko04112
orf00002_1	K02338	DPO3B, dnaN	http://www	path:ko00230;path:ko00240;]
orf00003_1	K03629	recF	http://www	path:ko03440

每个通路（Ko）中注释到的基因

PathWay	Pathway_definition	number_of_seqs	seqs_kos/genes_list	pathway_imagename
path:ko00830	Retinol metabolism	2	orf01837_1(K00121);c	ko00830.png
path:ko00650	Butanoate metabolis:	35	orf00100_1(K01692);c	ko00650.png
path:ko03430	Mismatch repair	15	orf00002_1(K02338);c	ko03430.png
path:ko00760	Nicotinate and nico	17	orf00046_1(K00767);c	ko00760.png

每个基因注释到的代谢通路图



红色表示本次测序所研究基因能够注释到这些基因产物上

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA注释
- 重复序列分析
- 插入序列分析

ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd

cdd.info	334 B	2016/6/28 下午1:57:00
cdd.tar.gz	3.6 GB	2016/6/28 下午2:07:00
cdd.versions	39.4 MB	2016/6/28 下午3:49:00
cddannot.dat.gz	446 kB	2016/6/28 下午3:48:00
cddannot_generic.dat.gz	252 kB	2016/6/28 下午3:48:00
cddid.tbl.gz	4.9 MB	2016/6/28 下午3:49:00
cddid_all.tbl.gz	5.4 MB	2016/6/28 下午3:49:00
cddmasters.fa.gz	12.0 MB	2016/6/28 下午3:49:00
cdtrack.txt	995 kB	2016/6/28 下午3:49:00
family_superfamily_links	1.2 MB	2016/6/28 下午3:49:00
fasta.tar.gz	332 MB	2016/6/28 下午3:50:00
little_endian/		2014/12/3 上午12:00:00
rpsbproc/		2015/11/6 上午12:00:00
sparcle/		2016/9/30 下午7:55:00
temp/		2016/9/15 下午1:38:00

使用更加灵敏的**rpsblast**工具进行比对分析，获取**COG**编号

Functional categories

J A K L B D Y V T M N Z W U O X C G E F H I P Q R S

All COGs

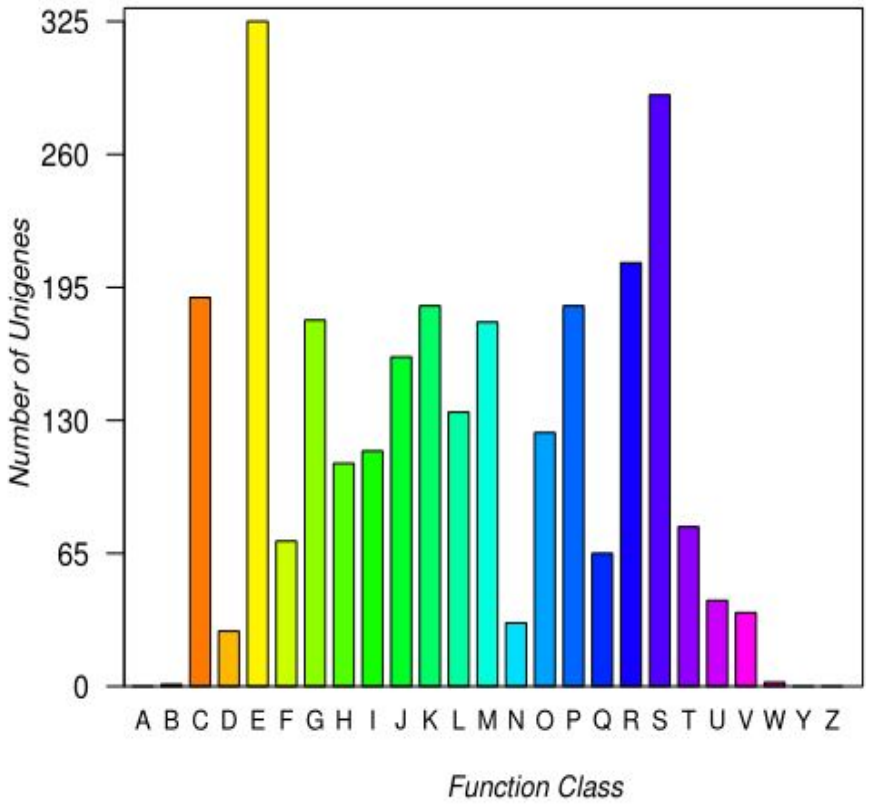
ARCHAEA	
CRENARCHAEOTA [21]	
Acihos	Acidianus hospitalis W1
Acisac	Acidilobus saccharovorans 345-15
Aerper	Aeropyrum pernix K1
Callag	Caldisphaera lagunensis DSM 15908
Calmag	Caldivirga maquilingensis IC-167
Deskam	Desulfurococcus kamchatkensis 1221n
Ferfon	Fervidicoccus fontis Kam940
Hypbut	Hyperthermus butylicus DSM 5456

COG功能注释

- COG 是 Clusters 是在对已完成基因组测 的，一个 COG 对应于 以对预测蛋白进行功能

```
INFORMATION STORAGE AND PROCESSING
 [J] Translation, ribosomal structure and biogenesis
 [A] RNA processing and modification
 [K] Transcription
 [L] Replication, recombination and repair
 [B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING
 [D] Cell cycle control, cell division, chromosome partitioning
```

- 基因归类。



A : RNA processing and modification
B : Chromatin structure and dynamics
C : Energy production and conversion
D : Cell cycle control, cell division, chromosome partitioning
E : Amino acid transport and metabolism
F : Nucleotide transport and metabolism
G : Carbohydrate transport and metabolism
H : Coenzyme transport and metabolism
I : Lipid transport and metabolism
J : Translation, ribosomal structure and biogenesis
K : Transcription
L : Replication, recombination and repair
M : Cell wall/membrane/envelope biogenesis
N : Cell motility
O : Posttranslational modification, protein turnover, chaperones
P : Inorganic ion transport and metabolism
Q : Secondary metabolites biosynthesis, transport and catabolism
R : General function prediction only
S : Function unknown
T : Signal transduction mechanisms
U : Intracellular trafficking, secretion, and vesicular transport
V : Defense mechanisms
W : Extracellular structures
Y : Nuclear structure
Z : Cytoskeleton

COG（ Clusters of Orthologous Groups of proteins ）注释结果：

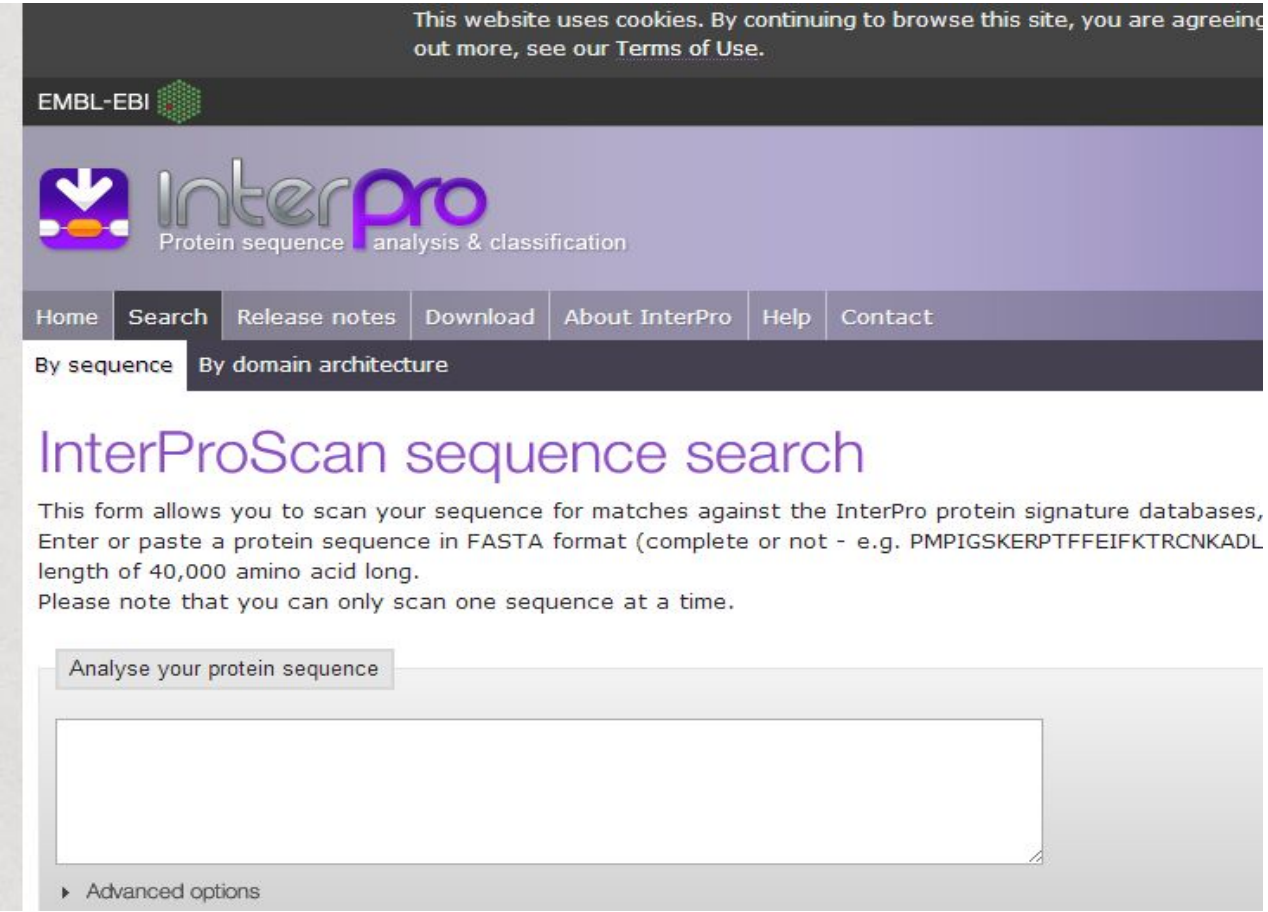
每个orf的cog注释结果详情，比对结果，cog编号，cog分类

Query name	Query length	Start-of-query	end-of-query	Hit description	COG/KOG/NOG_group	group_description	group_categeries	coverage	identity
orf1	462	1	462	DNA replication initiator	COG0593	it binds …	[L]	100%	95.20%
orf2	……	……	……	……	……	……	……	……	……
……	……	……	……	……	……	……	……	……	……

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA注释
- 重复序列分析

下载**InterProScan** 单机版软件和**panther-data-9.0**数据包

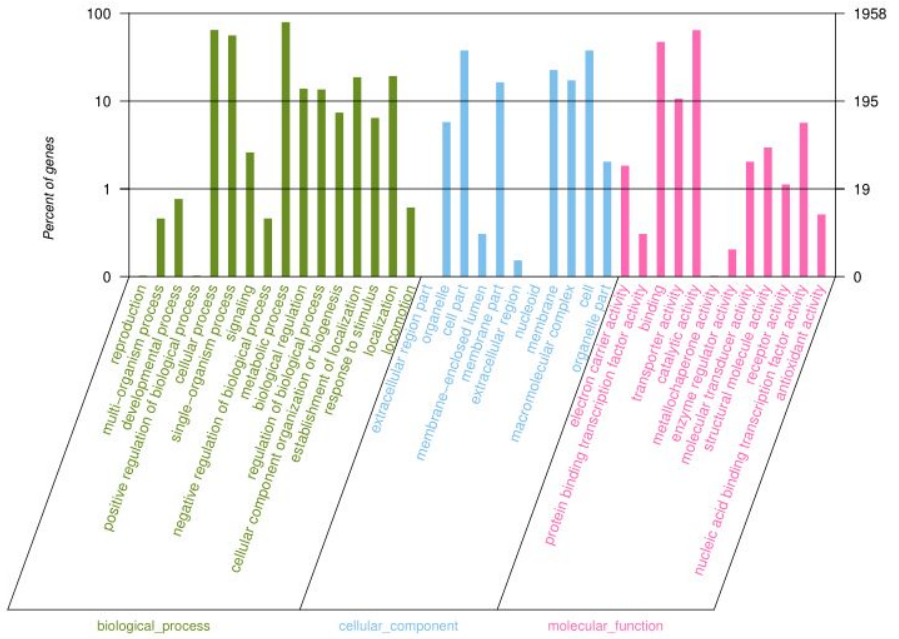
http://www.ebi.ac.uk/interpro/search/sequence-search



GO功能注释

- GO 是基因本体论 Gene Ontology 的缩写。由于不同物种、不同数据库中的关于基因和基因产物等生物学术语的描述存在差异，当查询某个研究领域的相关信息时，生物学家需要花费大量的时间和精力去分析生物学术语之间的联系，而 Gene Ontology 项目的目的就是为了标准化这些生物学术语，方便交流。

- GO 注释划分为 3 个层面的内容：
  - Cellular component 细胞组分
  - Molecular function 分子功能
  - Biological process 生物学过程



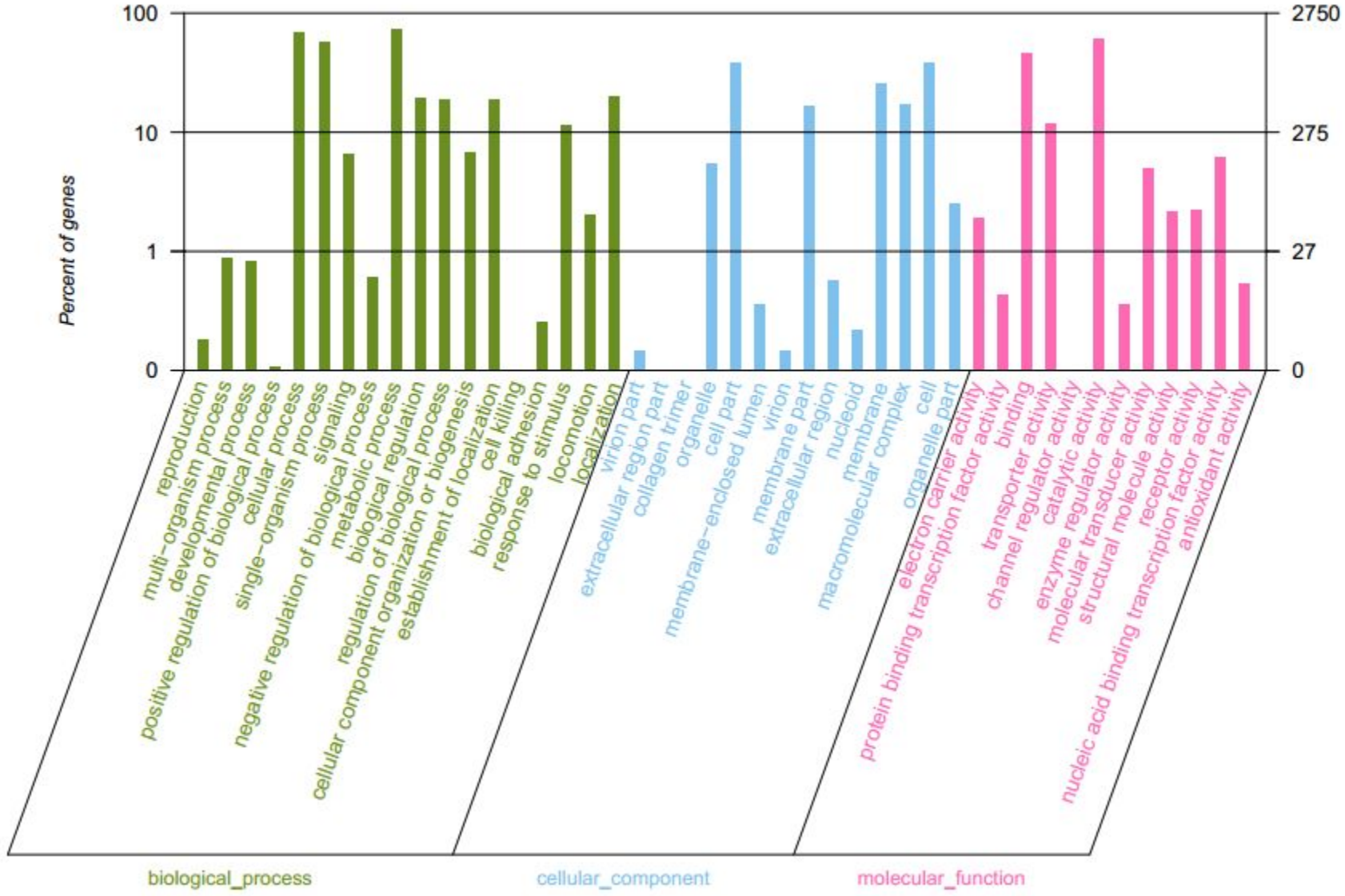
GO（ Gene Ontology）注释结果：

每个基因的go号

1	orf00001_1	GO:0017111	chromosomal replication initiator protein
2	orf00001_1	GO:0005524	
3	orf00001_1	GO:0003688	
4	orf00001_1	GO:0006270	
5	orf00001_1	GO:0006275	
6	orf00001_1	GO:0005737	
7	orf00001_1	GO:0046809	
8	orf00001_1	EC:3.6.1.15	
9	orf00002_1	GO:0005737	dna polymerase beta subunit
10	orf00002_1	GO:0006261	
11	orf00002_1	GO:0003887	
12	orf00002_1	GO:0090305	
13	orf00002_1	GO:0003677	
14	orf00002_1	GO:0008408	
15	orf00002_1	GO:0009360	
16	orf00002_1	EC:2.7.7.7	

三种分类层次中每个子分类中包含的基因

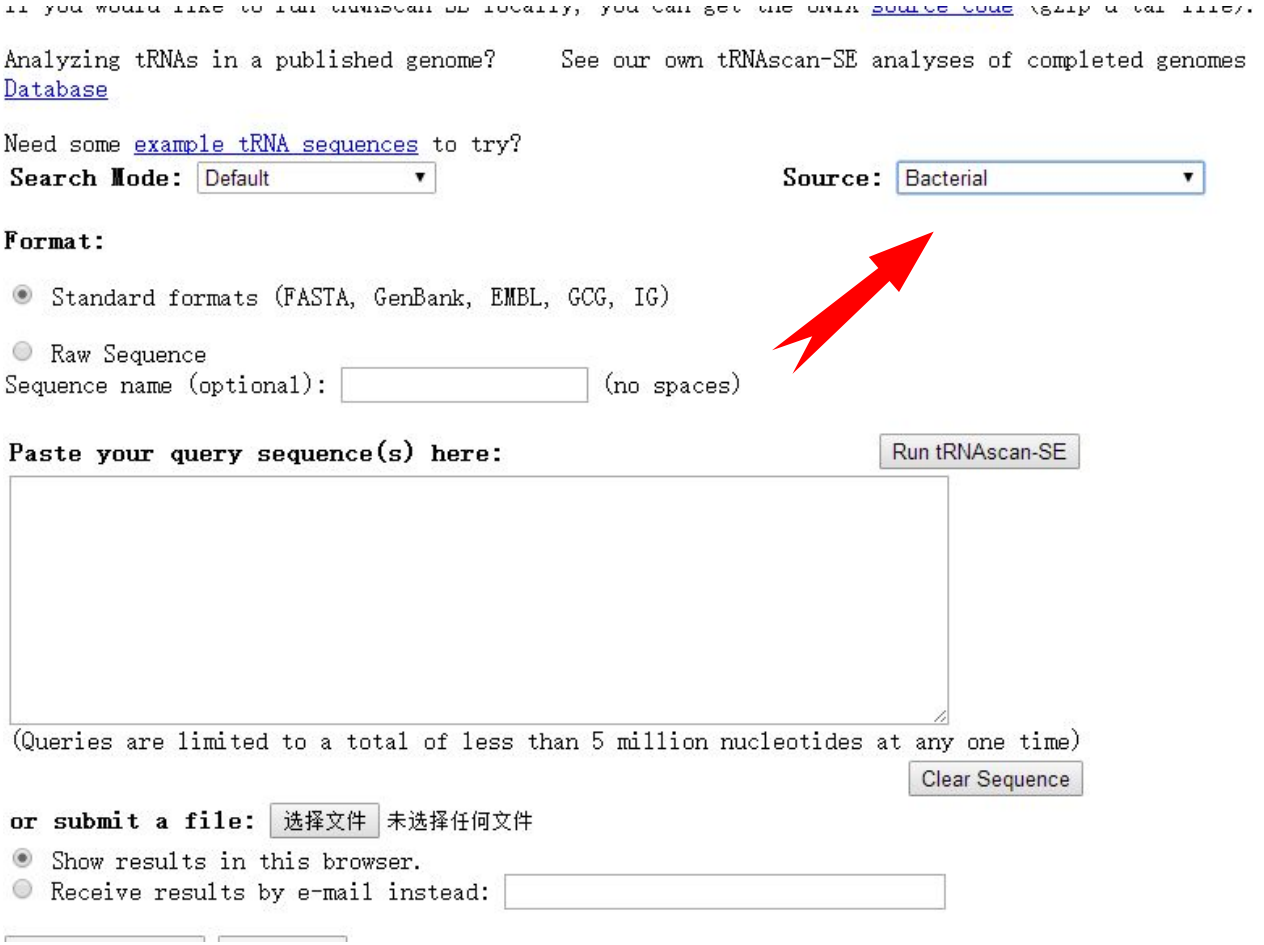
1	molecular_function	electron carrier activity	GO:0009055		
2	molecular_function	protein binding transcription factor act			
3	molecular_function	binding	GO:0005488	1023	orf03517_1(G
4	molecular_function	transporter activity	GO:0005215	256	
5	molecular_function	catalytic activity	GO:0003824	1354	
6	molecular_function	metallochaperone activity	GO:0016530		
7	molecular_function	enzyme regulator activity	GO:0030234		
8	molecular_function	molecular transducer activity	GO:00600		
9	molecular_function	structural molecule activity	GO:00051		
10	molecular_function	receptor activity	GO:0004872	32	orf0
11	molecular_function	nucleic acid binding transcription facto			
12	molecular_function	antioxidant activity	GO:0016209	15	
13	cellular_component	virion part	GO:0044423	4	orf02220_1(G
14	cellular_component	extracellular matrix	GO:0031012	2	
15	cellular_component	extracellular region part	GO:0044421		
16	cellular_component	organelle	GO:0043226	105	orf00426_1(G


Percent of genes
100
10
1
0
2750
275
27
0
reproduction
multi-organism process
developmental process
positive regulation of biological process
cellular process
single-organism process
signaling
negative regulation of biological process
metabolic process
biological regulation
regulation of biological process
cellular component organization or biogenesis
establishment of localization
cell killing
biological adhesion
response to stimulus
locomotion
localization
virion part
extracellular region part
collagen trimer
organelle
cell part
membrane-enclosed lumen
virion
membrane part
extracellular region
nucleoid
membrane
macromolecular complex
cell
organelle part
electron carrier activity
protein binding transcription factor activity
binding
transporter activity
channel regulator activity
catalytic activity
enzyme regulator activity
molecular transducer activity
structural molecule activity
receptor activity
nucleic acid binding transcription factor activity
antioxidant activity
biological_process
cellular_component
molecular_function

GO注释结果

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA/rRNA注释
- 重复序列分析

http://selab.janelia.org/tRNAscan-SE/



- 分别利用不同的预测软件对基因组中包含的 rRNA 和 tRNA 进行预测。

基因组 rRNA 预测统计表

Sequence name	Begin	End	+/-	Attribute
chromosome1	1599857	1601338	-	16S_rRNA
chromosome1	1795464	1795573	-	5S_rRNA
chromosome1	1795791	1798693	-	23S_rRNA
chromosome1	1799466	1800947	-	16S_rRNA

tRNA注释结果：

所有预测tRNA在基因组上的位点信息及类型

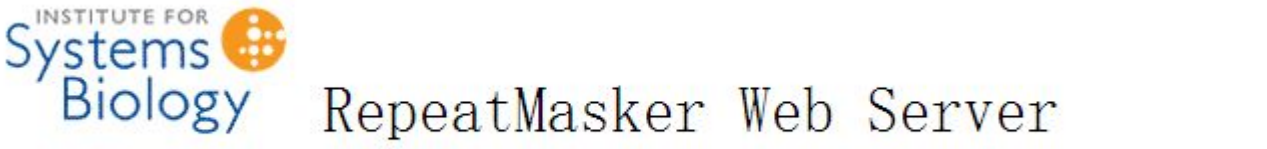
Sequence Name	tRNA#	tRNA Bounds Begin	End	tRNA Type	Anti Codon	Intron Bounds Begin	End	Cove Score
scaffold1	1	20010	20086	Ile	GAT	0	0	98.53
scaffold1	2	20142	20217	Ala	TGC	0	0	89.91
scaffold1	3	180647	180723	Ile	GAT	0	0	98.53
scaffold1	4	180779	180854	Ala	TGC	0	0	89.91
scaffold1	5	285811	285886	Thr	TGT	0	0	94.17
scaffold1	6	285943	286026	Tyr	GTA	0	0	62.89
scaffold1	7	286064	286139	Gly	TCC	0	0	91.76
scaffold1	8	286153	286227	Thr	GGT	0	0	83.04
scaffold1	9	287641	287716	Trp	CCA	0	0	81.08
scaffold1	10	343663	343747	Leu	GAG	0	0	59.44

tRNA的二级结构

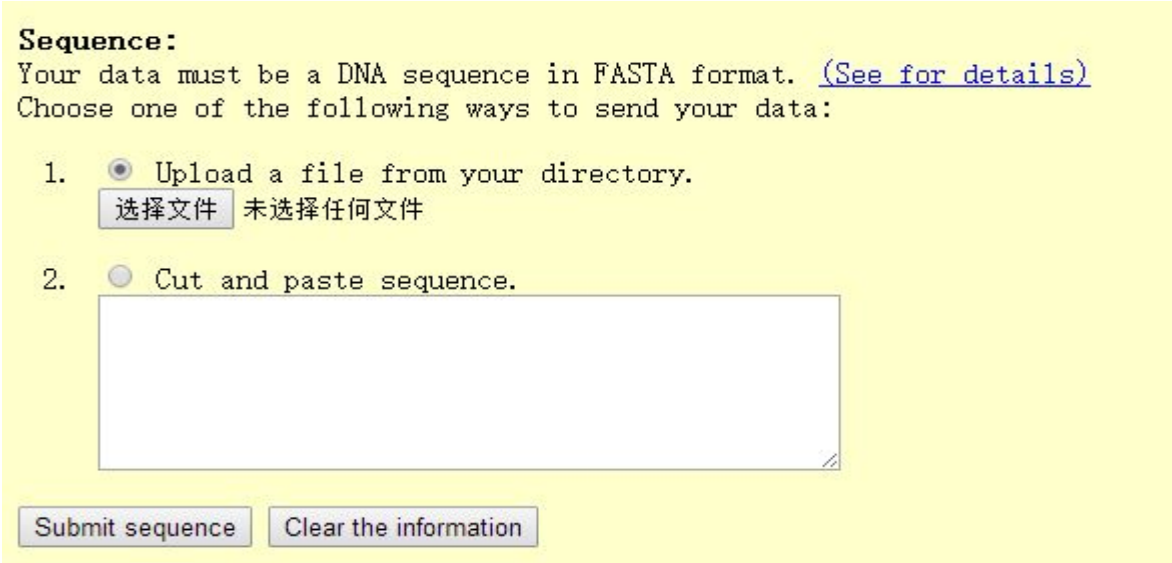
```
scaffold1.trna1 (11747-11823)   Length: 77 bp
Type: Ile   Anticodon: GAT at 35-37 (11781-11783)   Score: 88.31
         *    |    *    |    *    |    *    |    *    |    *    |    *    |    *
Seq: GGGCCTATAGCTCAGGCGGTtAGAGCGCTTCGCTGATAACGAAGAGGtCGGAGGTTCGAGTCCTCCTAGGCCCACGA
Str: >>>>>>>..>>>>.........<<<<.>>>>>.......<<<<<.....>>>>>.......<<<<<<<<<<<....

scaffold1.trna2 (11948-12023)   Length: 76 bp
Type: Ala   Anticodon: TGC at 34-36 (11981-11983)   Score: 86.37
         *    |    *    |    *    |    *    |    *    |    *    |    *    |    *
Seq: GGGGCCTTAGCTCAGTTGGTAGAGCGCTGCCTTTGCAAGGCAGATGtCAGGAGTTCGAATCTCCTAGGCTCCACAA
Str: >>>>>>>..>>>>........<<<<.>>>>>.......<<<<<.....>>>>>.......<<<<<<<<<<<....
```

- nr库注释
- KEGG库注释
- COG库注释
- GO库注释
- tRNA注释
- 重复序列分析

http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker



http://tandem.bu.edu/trf/trf.html



重复序列分析结果： 采用不同的软件分别对不同类型的重复序列进行预测

SSR 预测结果

ID	SSR nr.	SSR type	SSR size	start	end
scaffold1	1	p3	(GCA)5	15	250365	250379
scaffold1	2	p3	(GCT)5	15	291016	291030
scaffold1	3	p1	(T)10	10	1164046	1164055
scaffold1	4	p6	(AGGTTC)8	48	2274148	2274195
scaffold1	5	p1	(A)10	10	2886284	2886293
scaffold1	6	p1	(A)10	10	3481624	3481633
scaffold1	7	p6	(GAAGGT)5	30	3568094	3568123
scaffold1	8	p1	(A)10	10	3736597	3736606
scaffold1	9	p6	(TGATGG)11	66	3836161	3836226

串联重复序列预测结果

```
##gff-version 3
scaffold1	TRF	TandemRepeat	17513	17827	630	+	.	ID=TR01;PeriodSize=9;CopyNumber=35.0;PercentMatches=100
scaffold1	TRF	TandemRepeat	346619	346667	89	+	.	ID=TR02;PeriodSize=24;CopyNumber=2.0;PercentMatches=96;
scaffold1	TRF	TandemRepeat	585197	585263	91	+	.	ID=TR03;PeriodSize=24;CopyNumber=2.8;PercentMatches=86;
scaffold1	TRF	TandemRepeat	586446	586876	691	+	.	ID=TR04;PeriodSize=30;CopyNumber=14.4;PercentMatches=92
scaffold1	TRF	TandemRepeat	660190	660353	86	+	.	ID=TR05;PeriodSize=21;CopyNumber=8.6;PercentMatches=73;
scaffold1	TRF	TandemRepeat	660204	660349	179	+	.	ID=TR06;PeriodSize=42;CopyNumber=3.7;PercentMatches=76;
scaffold1	TRF	TandemRepeat	660237	660358	163	+	.	ID=TR07;PeriodSize=54;CopyNumber=2.3;PercentMatches=87;
scaffold1	TRF	TandemRepeat	660192	660349	248	+	.	ID=TR08;PeriodSize=75;CopyNumber=2.1;PercentMatches=90;
```

◆耐药基因注释（CARD）

ORF_ID	CONTIG	START	STOP	ORIENTATION	ARO	ARO_name	ARO_category
orf03436_1	scaffold1	3149516	3147771	-	ARO:3000776	adeC	efflux pump conferring antibiotic resistance, tetracycline resistance gene
orf00809_1	scaffold1	752431	751754	-	ARO:3000535	macB	efflux pump conferring antibiotic resistance, macrolide resistance gene
orf03955_1	scaffold1	3659451	3658462	-	ARO:3003577	PmrE	gene altering cell wall charge conferring antibiotic resistance, polymyxin resi
orf02216_1	scaffold1	2042968	2041106	-	ARO:3000617,	mecC, mec	antibiotic resistance gene cluster, cassette, or operon, antibiotic target repl

◆毒力基因注释（VFDB）

Score	HSP-Len	%-Simil	Query-Name	Q-Len	Q-Begin	Q-End	Hit-Name	Description
148	139	52%	orf00015_1	440	38	176	VFG038255(gi:46982	(pbpG) D-alanyl-D-alanine endo
148	180	47%	orf00057_1	457	7	175	VFG011688(gi:12161	(CJJ81176_1422) capsular biosy
311	492	46%	orf00074_1	556	88	547	VFG016546(gi:42561	(tnp) IS1634CB transposase [Ca

美吉生物 Majorbio

◆碳水化合物数据库注释（CAZy）

Class	Genes_Count	Genes_List	Class_Definition
GH	32	orf00020_1,	Glycoside Hydrolases
GT	26	orf00404_1,	Glycosyl Transferases
PL	0	-	Polysaccharide Lyases
CE	20	orf00271_1,	Carbohydrate Esterases
AA	4	orf00410_1,	Auxiliary Activities
CBM	14	orf00180_1,	Carbohydrate-Binding Modules

Family	Genes_Count	Genes_List	Class	Class_Definition
AA3	1	orf00410_1	AA	Auxiliary Activities
AA4	1	orf01556_1	AA	Auxiliary Activities
AA6	2	832_1, orf00	AA	Auxiliary Activities
CBM50	12	688_1, orf02	CBM	Carbohydrate-Binding Modules
CBM63	1	orf01402_1	CBM	Carbohydrate-Binding Modules
CBM65	1	orf01739_1	CBM	Carbohydrate-Binding Modules
CE1	6	531_1, orf01	CE	Carbohydrate Esterases
CE10	2	564_1, orf02	CE	Carbohydrate Esterases
CE12	1	orf00641_1	CE	Carbohydrate Esterases
CE14	2	269_1, orf04	CE	Carbohydrate Esterases
CE3	3	, orf02367_1,	CE	Carbohydrate Esterases
CE4	4	640_1, orf00	CE	Carbohydrate Esterases
CE7	1	orf00688_1	CE	Carbohydrate Esterases
CE9	1	orf03758_1	CE	Carbohydrate Esterases
GH1	4	503_1, orf03	GH	Glycoside Hydrolases
GH109	5	, orf01230_1,	GH	Glycoside Hydrolases

CAZy数据库包含：

糖苷水解酶类（GHs）

糖苷转移酶类（GTs）

多糖裂解酶（PLs）

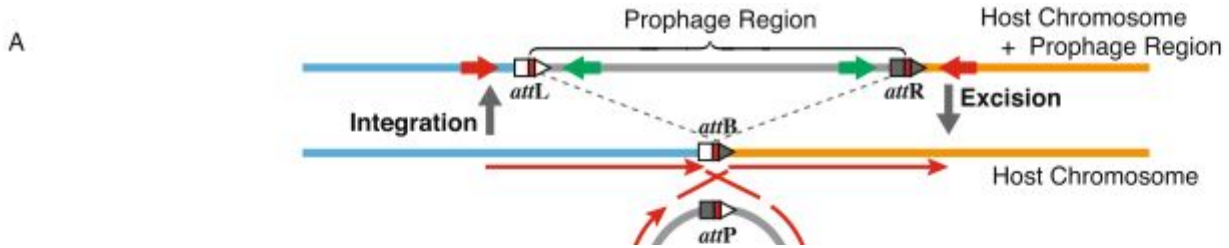
糖水化合物酯酶类（CEs）

辅助模块酶类（AAs）

六大类家族

◆ 前噬菌体分析

整合在宿主基因组上的温和噬菌体，带有前噬菌体的菌称为溶源菌。条件适宜时可能会打破前噬菌体状态，变为可增殖型而进行自主增殖，并使细胞裂解。



ID	Length	Completeness	Position	Possible phage	GC%
1	22648bp	incomplete	chromosome1[1961485-1984132]	PHAGE_Acanth_mimivirus_NC_014649(4), ...	58.15

注 :ID:前噬菌体编号 ;Length :预测出的前噬菌体长度 ;Completeness :完整度 ;Position :预测到的前噬菌体位置 ;Possible phage：数据库中与该区域相似的序列 ID 号；GC%：GC 含量。

ΦLH-1 attL (959,267-959,328)
AAGCGAATTGTCGTGAGTTCGAACCTCACTCGCTTCACTTTTCACAAGCCGTTATATACCTG
ΦLH-1 attR (998,034-998,095)
GGGATTTTTCGCACGTGTTCAACTGTCACTCGCTTCATTTTTTTACAAGCTGTCTATATATC
ΦLH-1 attB
AAGCGAATTGTCGTGAGTTCGAACCTCACTCGCTTCATTTTTTTACAAGCTGTCTATATATC
ΦLH-1 attP
GGGATTTTTCGCACGTGTTCAACTGTCACTCGCTTCACTTTTCACAAGCCGTTATATACCTG

ΦLH-2 attL (1,400,311-1,400,382)
AAAAGTAGAGTTTAGATCTTATATTACTTAGAAAAATAAAAACGCGTTACATTAATATGCTGTTAAATCAAC
ΦLH-2 attR (1,437,638-1,437,709)
TGTGACTCCAGATAATTCCCAAATCACTTAGAAAAATAAAAACGCGTCAACTCAAGAGCTGAATTGTCCATA
ΦLH-2 attB
AAAAGTAGAGTTTAGATCTTATATTACTTAGAAAAATAAAAACGCGTCAACTCAAGAGCTGAATTGTCCATA
ΦLH-2 attP
TGTGACTCCAGATAATTCCCAAATCACTTAGAAAAATAAAAACGCGTTACATTAATATGCTGTTAAATCAAC

◆ 前噬菌体分析 http://phast.wishartlab.com/index.html

前噬菌体的序列信息

```
>1.864985_882644
TACTTATGAAATATACGTGCTTTATCACGTTG
>2.1240470_1255087
TGTTAGATGTCTGTGATGGGCAACCAATAGTA
>3.1688974_1722093
GCTCTAAATTGAGCGCTTTTTATTAAATTCAA
>4.2241487_2263990
CTTTTAAACGAATAGCAATAATCTTGTCGGAT
```

每个前噬菌体序列的注释信息

REGION	REGION_LENGTH	COMPLETENESS(score)	SPECIFIC_KEYWORD	REGION_POSITION	MOST_COMMON_PHAGE_NAME(hit_genes_count)
1	17.6Kb	intact(100)	terminase,tail,po	864985-882644	PHAGE_Psychr_Psymv2_NC_023734(4),PHAGE_Bur
2	14.6Kb	incomplete(60)	transposase,termi	1240470-1255087	PHAGE_Psychr_pOW20_A_NC_020841(6),PHAGE_Ac
3	33.1Kb	intact(120)	head,capsid,recom	1688974-1722093	PHAGE_Acinet_vB_AbaS_TRS1_NC_031098(16),PH
4	22.5Kb	incomplete(60)	capsid,head,porta	2241487-2263990	PHAGE_Psychr_pOW20_A_NC_020841(6),PHAGE_Ps

◆ 基因岛**GI**预测分析

由于环境的选择，不同的基因组之间往往会出现基因的水平转移，这些外来基因往往会以基因岛的形式存在于基因组中。

基因岛分析结果表

Island start	Island end	Length	Method	Locus	Gene start	Gene end	Strand	Product
258539	275983	17444	Predicted by at least one method	orf00255	258881	259927	1	integrase
258539	275983	17444	Predicted by at least one method	orf00256	260438	260656	1	DNA-binding protein
258539	275983	17444	Predicted by at least one method	orf00257	261306	261485	-1	hypothetical protein

◆ 基因岛分析

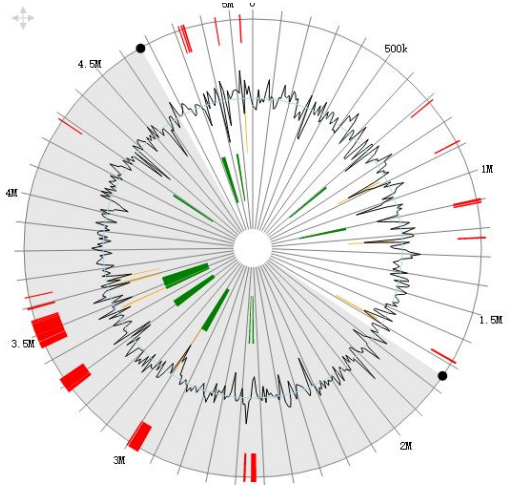
每个基因岛中包括的基因信息

Island start	Island end	Length	Method	Gene name	Gene ID Locus	Gene start	Gene end	Strand	Product External
17506	24809	7303	Predicted by at least one method		orf00015	17506	17880	-1	hypothetical protein
17506	24809	7303	Predicted by at least one method		orf00016	22931	23299	-1	MULTISPECIES: hydrola
17506	24809	7303	Predicted by at least one method		orf00017	23881	24210	1	hypothetical protein
17506	24809	7303	Predicted by at least one method		orf00018	24267	24809	-1	flavoprotein [Acinetol
61516	66167	4651	Predicted by at least one method		orf00055	61516	62811	1	hypothetical protein
61516	66167	4651	Predicted by at least one method		orf00056	62835	63614	1	glycosyltransferase g
61516	66167	4651	Predicted by at least one method		orf00057	63691	64671	-1	hypothetical protein
61516	66167	4651	Predicted by at least one method		orf00058	64770	66167	1	hypothetical protein

基因岛的序列信息

```
>17506_24809 7303
TAATATTTCTTTGGATTCTTTGGATTCTTTGGATTCTTTGGATTCTTTGGATTCTTTGGATTCTTTGGATTCTTTGGA
>61516_66167 4651
TGAACTTAATTCATAAATTAAAAAATGATCCTAGAATTAGAAATTCCCTATGGATGCTTATTGAAAAGGCTCTATCAT
>219864_225778 5914
TGAGTCAACCTGAACAATGTTGGTATATCAGAACTCCCATTCAGCAAGGGGAAATATTTTCATCTTGGTTGATCCGTT
>752886_762163 9277
TGAGTAAATTTATAAAGTCTTTAGCAATTTTTGCAGCTTCTGCTCCAATCATTAATAATAAAGCCACTGCTGCTATTG
>955887_960835 4948
TATCTTTGCTTATTAGATTGTGAAAGAAGTGATGCCGCTAGTGATTTTGTAGTATCACTATATTTGTCACTTTGTAAT
>1019913_1024589 4676
TGTTTAATCTAATTGCTAGATTAAATGATTTTATTGAGCATTTAAGCTATTACACAAGCCAAGTGTCTCCAGATCAGC
```

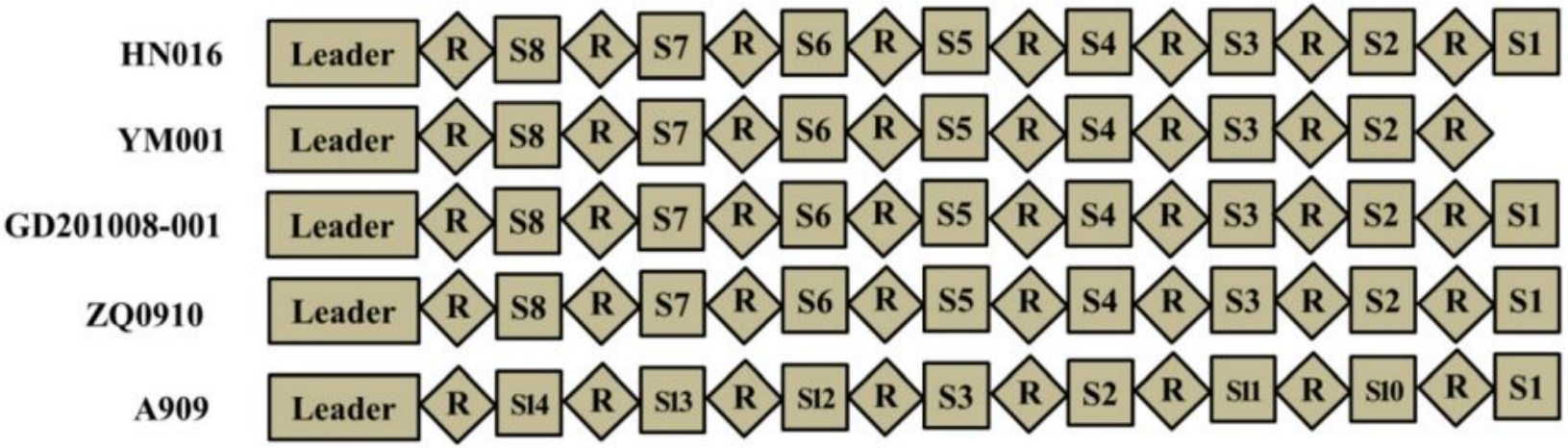
基因岛在基因组上的位置分布



◆ CRISPR分析

CRISPR 是一串包含多个短而重复的序列的碱基序列，重复序列之间是一些约 30bp 的"spacer DNA"。在原核生物中，CRISPR 起到免疫系统的作用，对外来的质粒和噬菌体序列具有抵抗作用。CRISPR 能识别并使入侵的功能元件沉默。

◆ CRISPR分析



参考文献：Comparative genome analysis identifies two large deletions in the genome of highly-passaged attenuated Streptococcus agalactiae strain YM001 compared to the parental pathogenic strain HN016

◆ 分泌蛋白预测分析

信号肽
预测结果

```
##Gene ID
##Length(aa)
##SignalP
##Description: D value is the score that is used to discriminate signal peptides from non-signal peptides;if D > D-cutoff,S
```

Gene ID	Length(aa)	SignalP	Description
orf00001_1	465	NO	D=0.121 D-cutoff=0.570 Networks=SignalP-noTM
orf00002_1	382	NO	D=0.213 D-cutoff=0.570 Networks=SignalP-noTM
orf00003_1	360	NO	D=0.186 D-cutoff=0.570 Networks=SignalP-noTM
orf00004_1	822	NO	D=0.102 D-cutoff=0.570 Networks=SignalP-noTM
orf00005_1	130	YES	Cleavage site between pos. 24 and 25: AFA-AD D=0.912 D-cutoff=0.570 Networks=SignalP-noTM

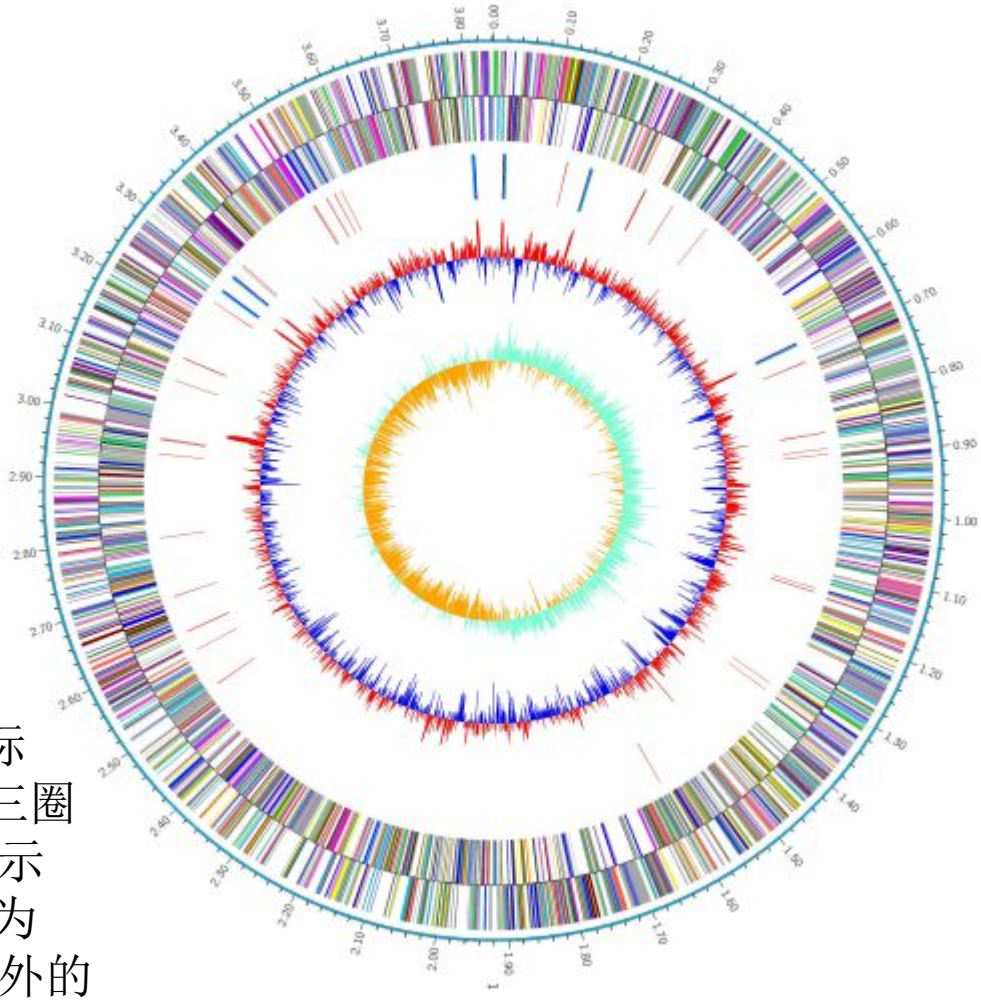
跨膜结构域预测

```
##Gene ID
##Length(aa)
##Number of predicted TMHs: The number of predicted transmembrane helices.
##Exp number of AAs in TMHs: The expected number of amino acids intransmembrane helices. If this number is larg
##Exp number, first 60 AAs: The expected number of amino acids in transmembrane helices in the first 60 amino a
##Total prob of N-in: The total probability that the N-term is on the cytoplasmic side of the membrane.
##POSSIBLE N-term signal sequence: a warning that is produced when "Exp number, first 60 AAs" is larger than 10
##Topology: eg. outside:1-316:表示从1-316个氨基酸在膜外
```

Gene ID	Length(aa	Number of	Exp numbe	Exp number, first 60 AAs	Total prob of N-in	POSSIBLE	Topology
orf00001_	465	0	0.00373	0.00079	0.00324	NO	outside:1-465;
orf00002_	382	0	0.0019	0.00046	0.01462	NO	outside:1-382;
orf00003_	360	0	0.06113	0.05451	0.02155	NO	outside:1-360;
orf00004_	822	0	0.01247	0	0.00058	NO	outside:1-822;
orf00005_	130	0	3.49197	3.49197	0.39433	NO	outside:1-130;
orf00006_	185	3	74.09714	23.81436	0.97066	YES	inside:1-20;TMhelix:2
orf00007_	643	0	0.07283	0.07214	0.00359	NO	outside:1-643;
orf00008_	334	0	8.29315	5.96176	0.38004	NO	outside:1-334;
orf00009_	335	0	4.81193	4.78049	0.2327	NO	outside:1-335;

◆ 细菌基因组圈图

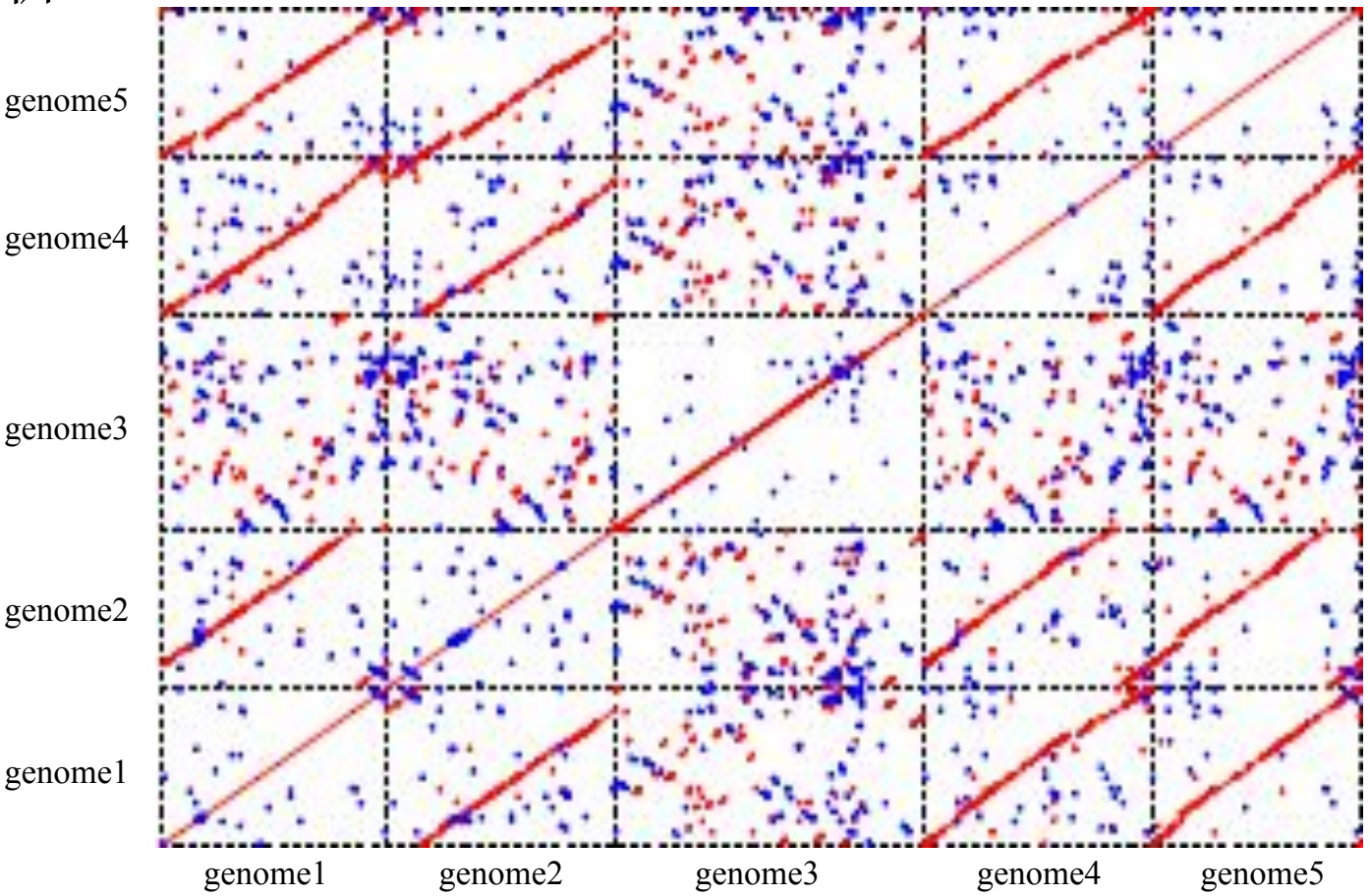
基因组圈图可以全面展示基因组的特征，如基因在正、反义链上的分布情况、基因的 COG 功能分类情况、GC 含量、基因组岛、同源基因等。



注：圈图的最外面一圈为基因组大小的标识，每一个刻度为 0.1Mb；第二圈和第三圈为正链、负链上的 CDS，不同的颜色表示 CDS 不同的 COG 的功能分类；第四圈为 rRNA 和 tRNA；第五圈为 GC 含量，向外的红色部分表示该区域 GC 含量高于全基因组平均 GC 含量，峰值越高表示与平均 GC 含量差值越大，向内的蓝色部分表示该区域

- 共线性分析

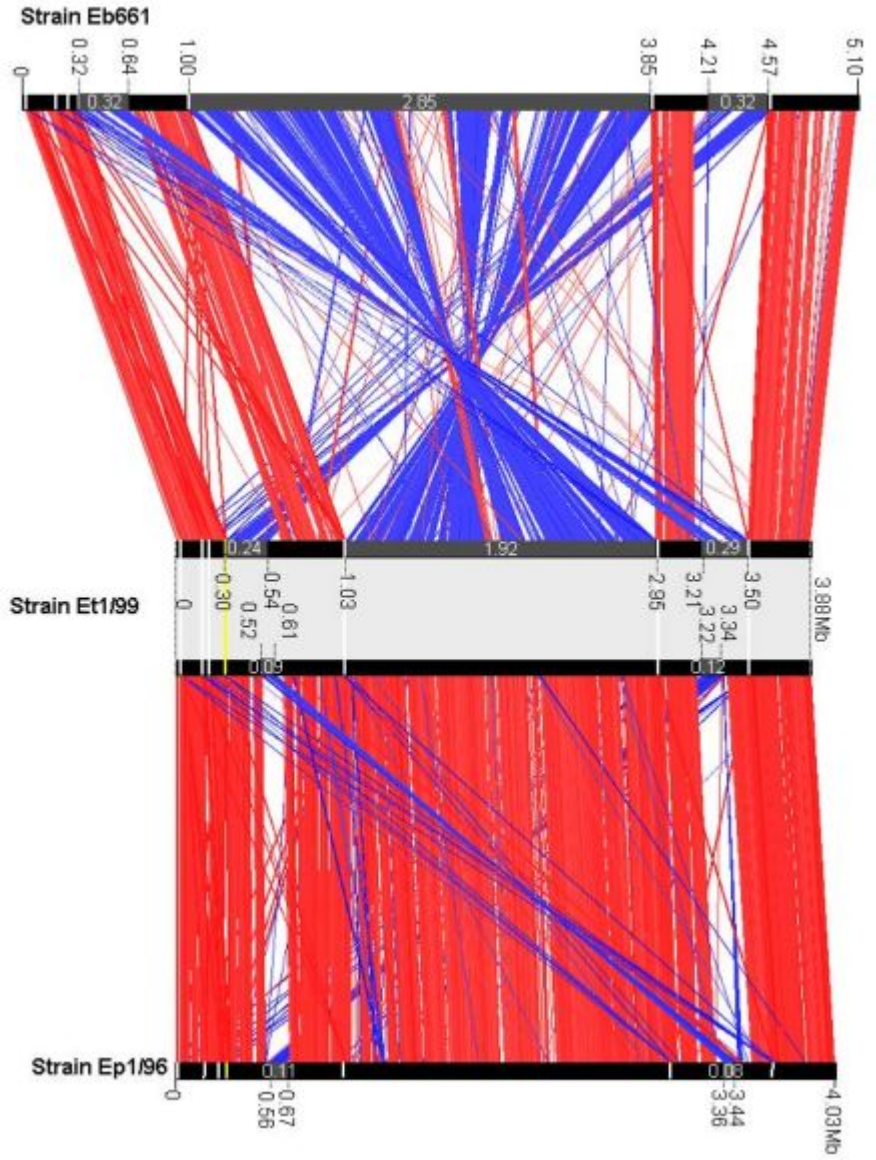
三种常用软件：
Mummer
ACT
Mauve



Mummer

- 共线性分析

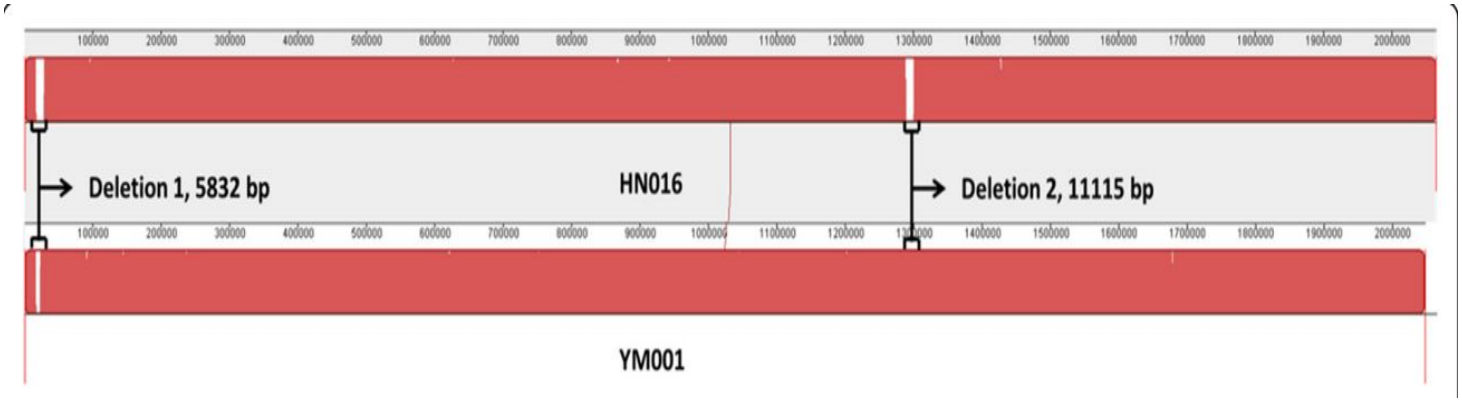
三种常用软件：
Mummer
ACT
Mauve



ACT

- 共线性分析

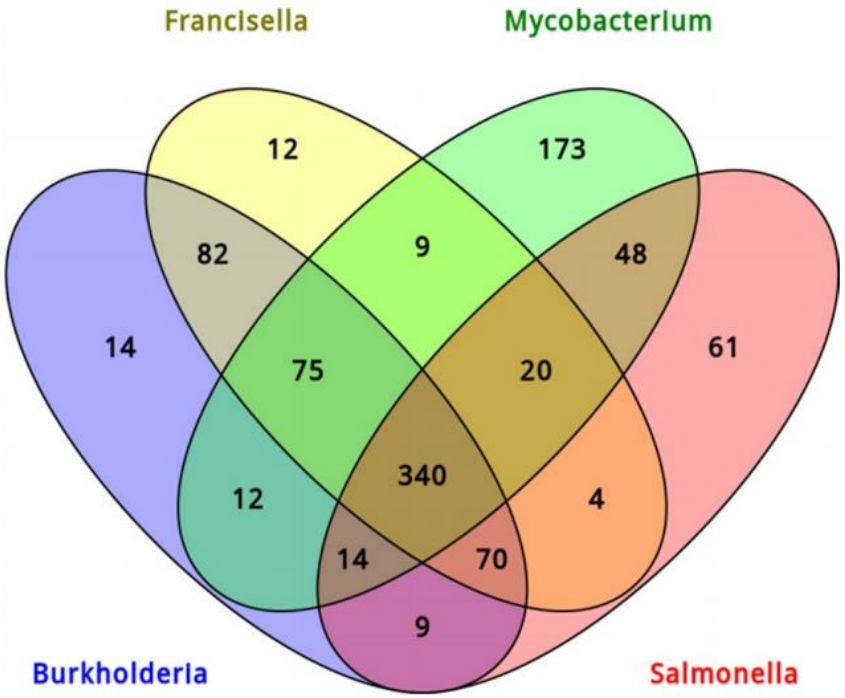
三种常用软件：
Mummer
ACT
Mauve

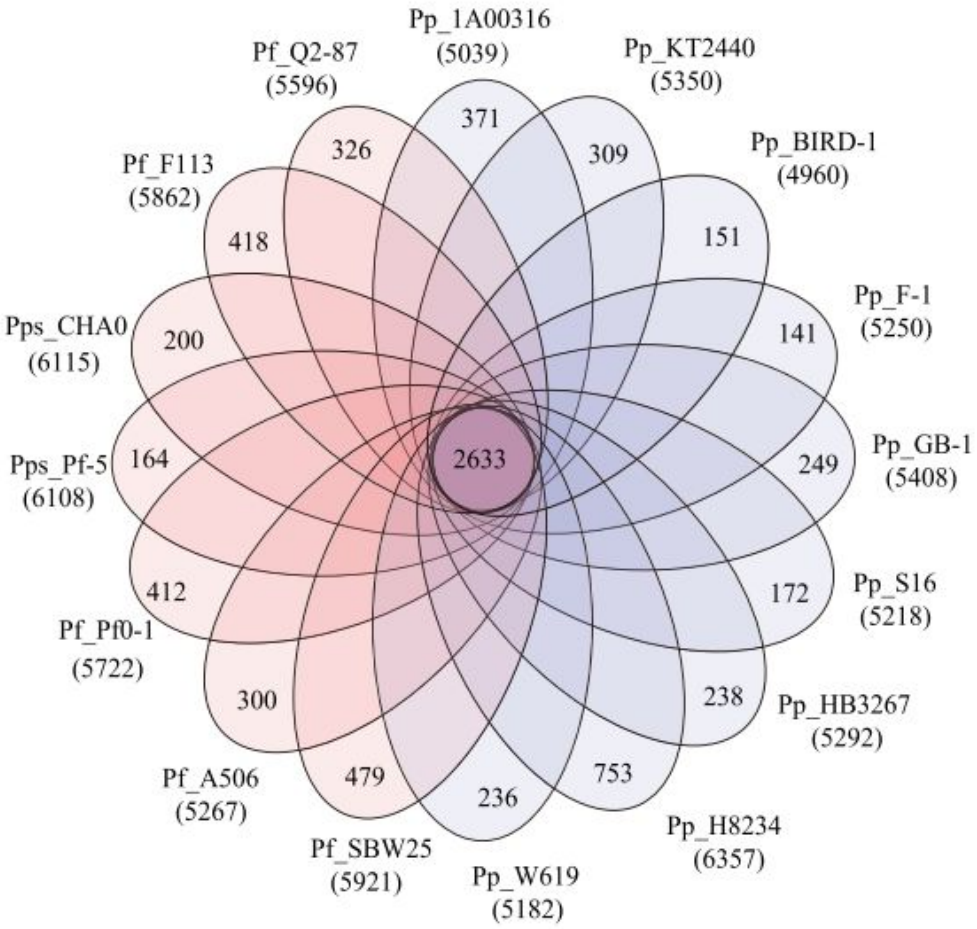


Mauve

参考文献：Comparative genome analysis identifies two large deletions in the genome of highly-passaged attenuated Streptococcus agalactiae strain YM001

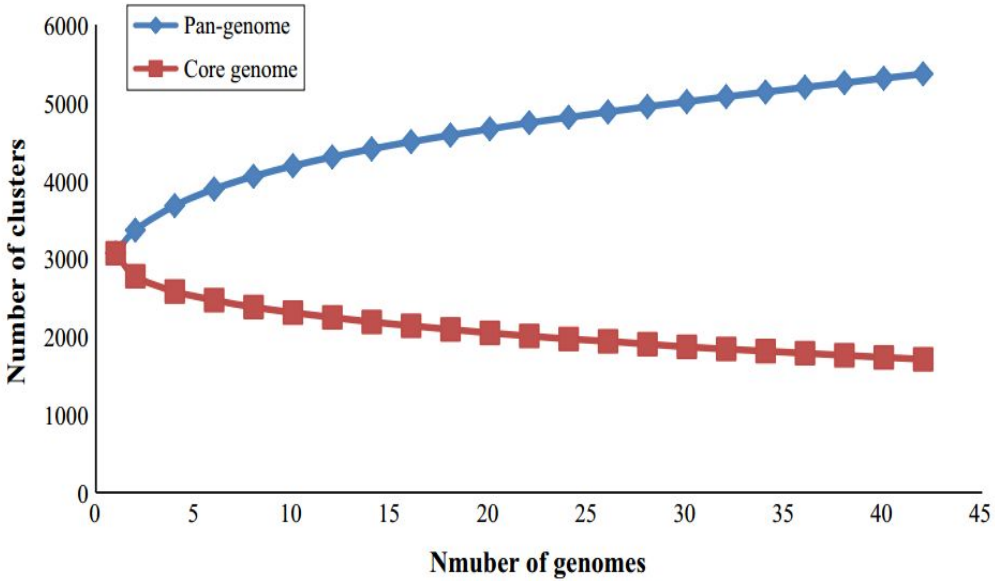
- 同源基因分析





参考文献：Analysis of pan-genome to identify the core genes and essential genes of Brucella spp.

- 泛基因组分析



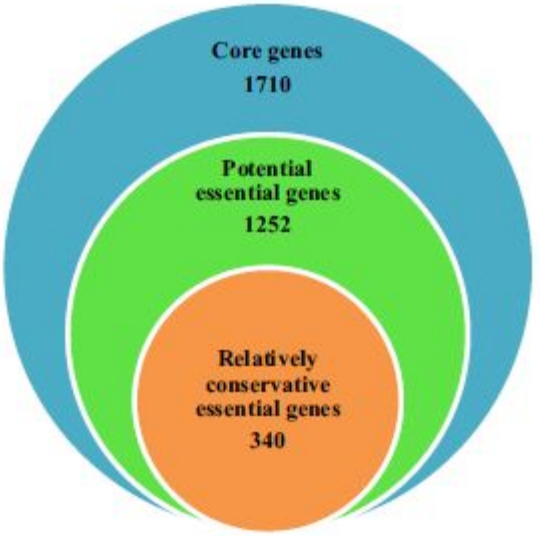
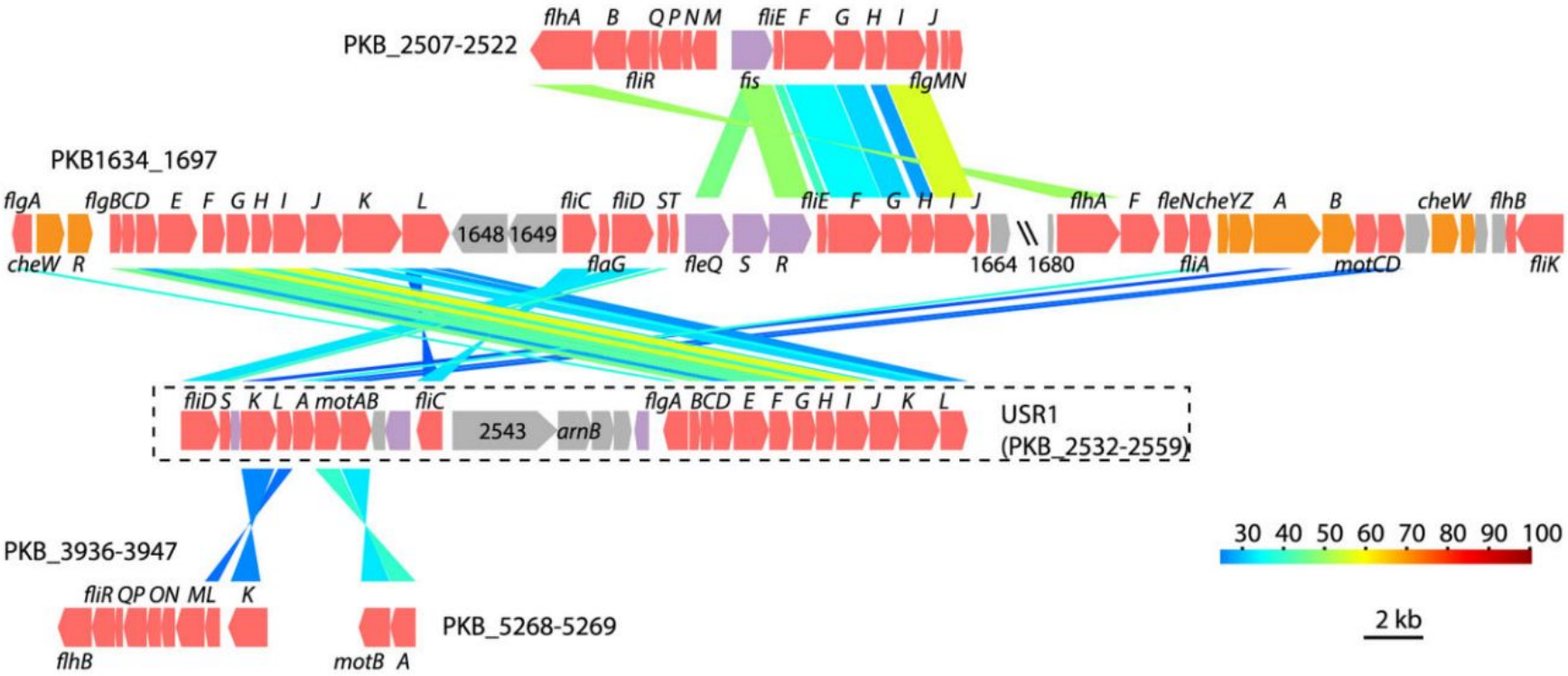


Fig. 6 Relationship between core genes and essential genes. The core genome comprised 1710 clusters, and we selected one gene as representative of each cluster. The number of homologs found in the DEG was 1252. The number of persistent nonessential genes that belong to the core genome but are not essential genes was 458 (1710–1252)

参考文献：Analysis of pan-genome to identify the core genes and essential genes of Brucella spp.

- 基因簇分析



参考文献：Comparative genome analysis of Pseudomonas knackmussii B13, the first bacterium known to degrade chloroaromatic compounds

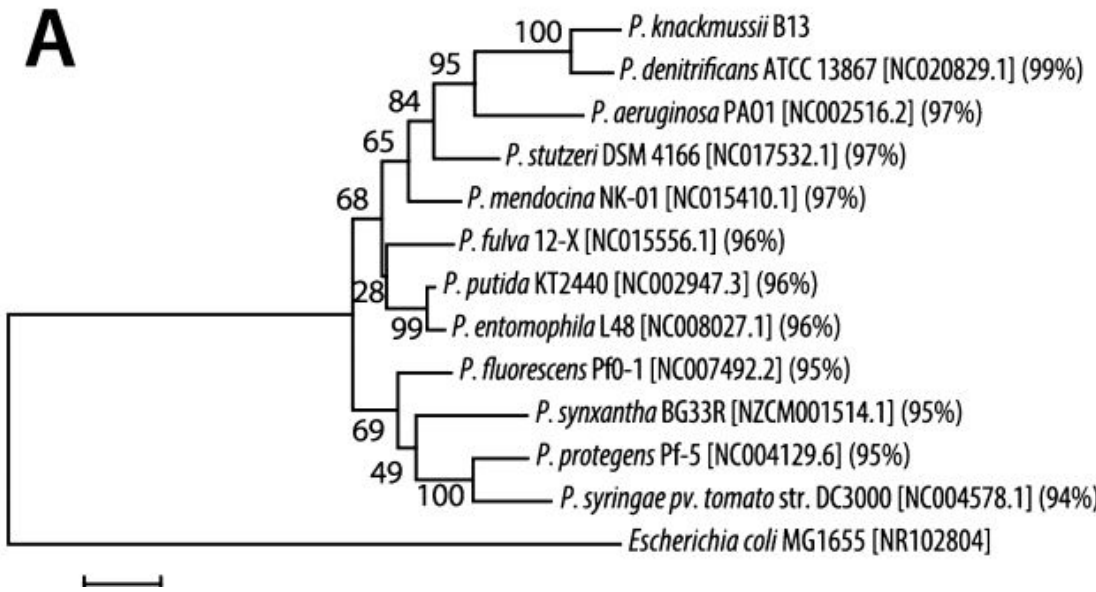
- 进化树分析

三种方式：

基于单基因或多基因建树

基于全基因组中单拷贝基因建树

基于全基因组SNP建树



参考文献：Comparative genome analysis of Pseudomonas knackmussii B13, the first bacterium known to degrade
chloroaromatic compounds

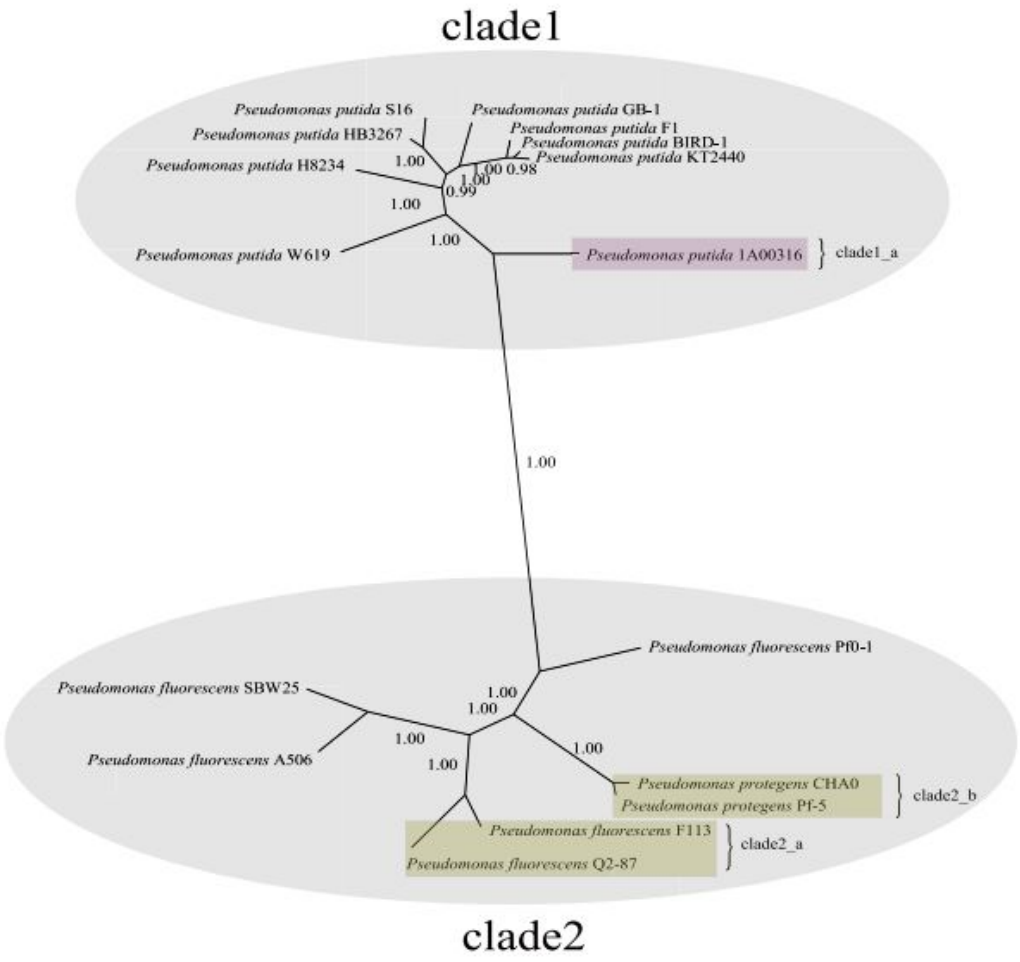
- 进化树分析

三种方式：

基于单基因或多基因建树

基于全基因组中单拷贝基因建树

基于全基因组SNP建树



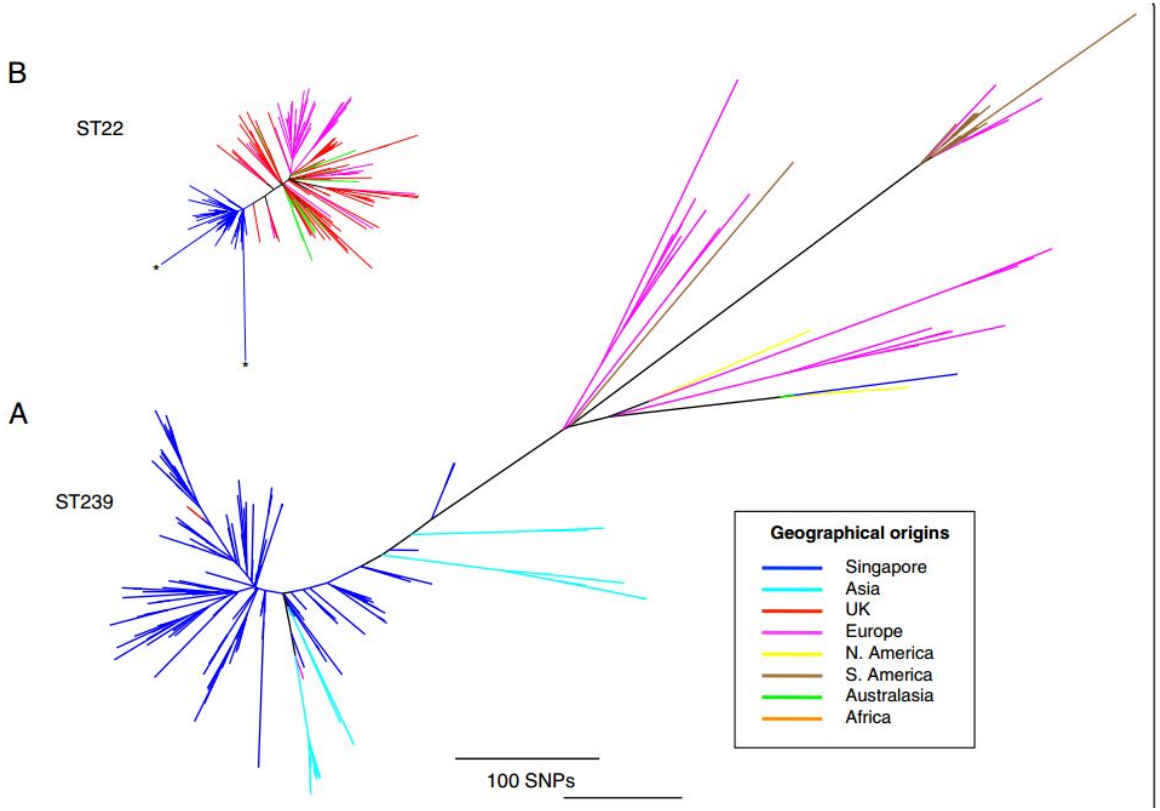
- 进化树分析

三种方式：

基于单基因或多基因建树

基于全基因组中单拷贝基因建树

基于全基因组SNP建树



参考文献：Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system

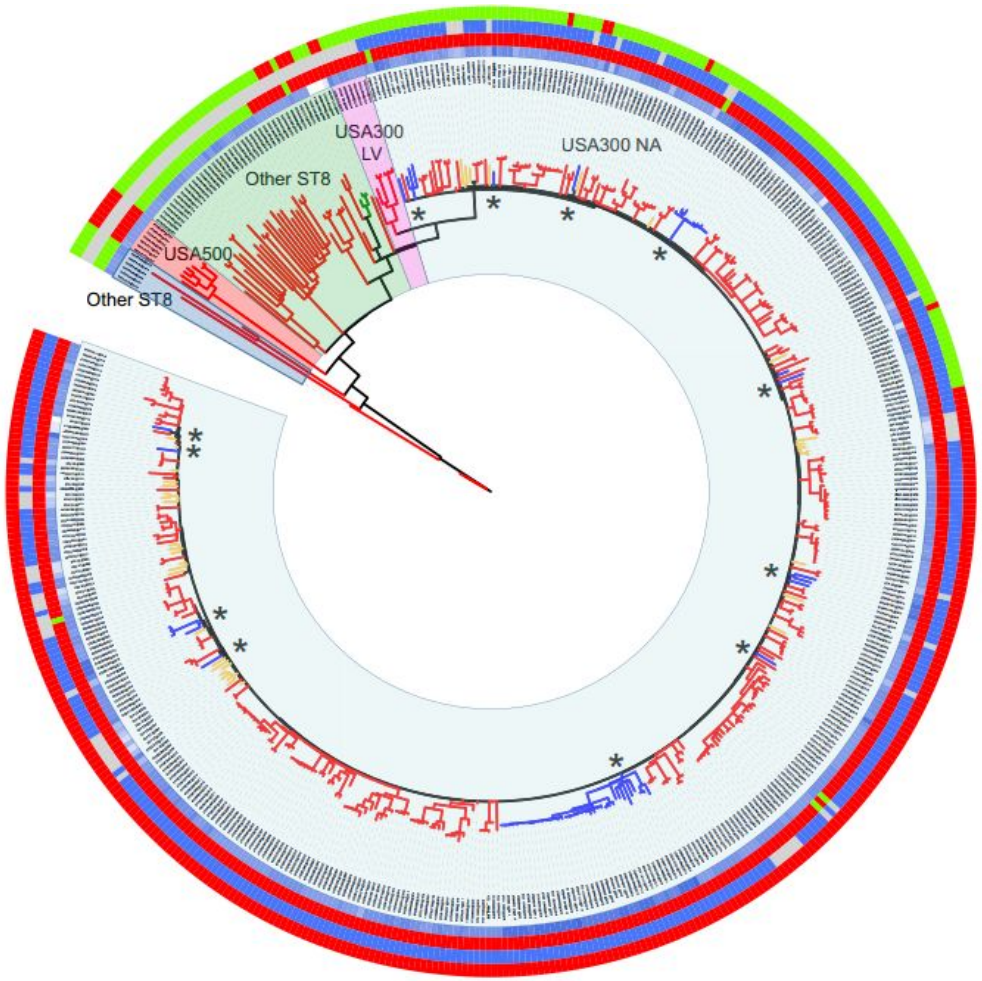
- 进化树分析

三种方式：

基于单基因或多基因建树

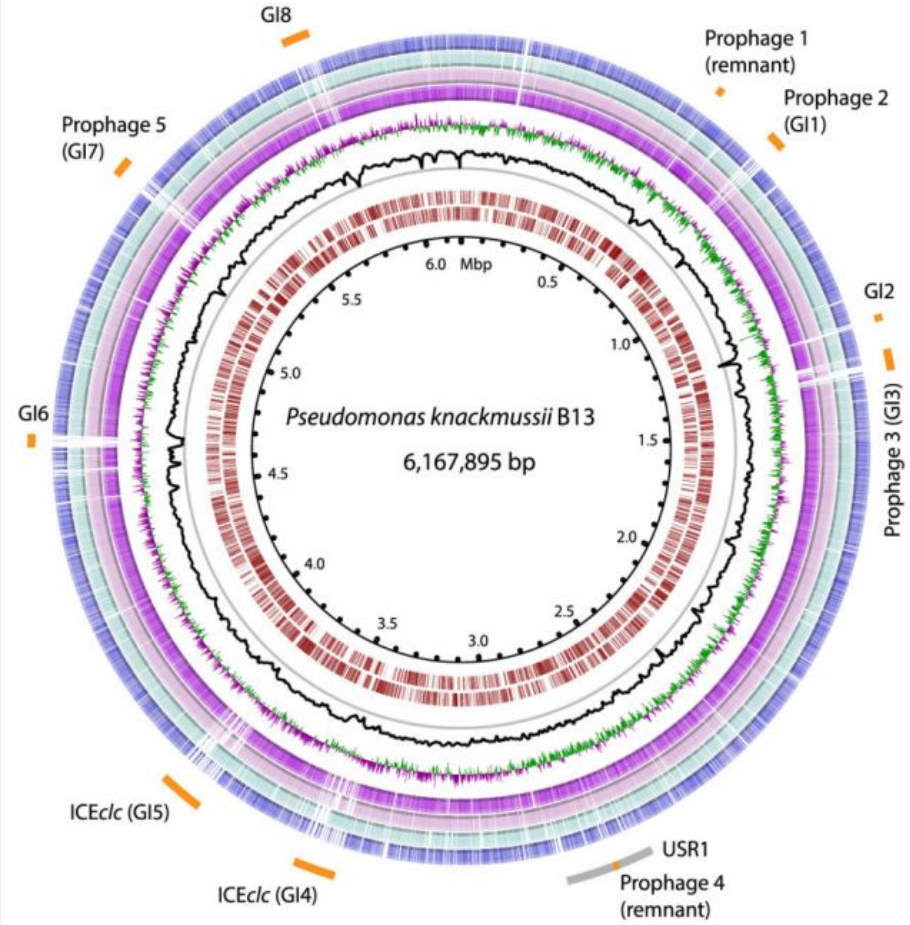
基于全基因组中单拷贝基因建树

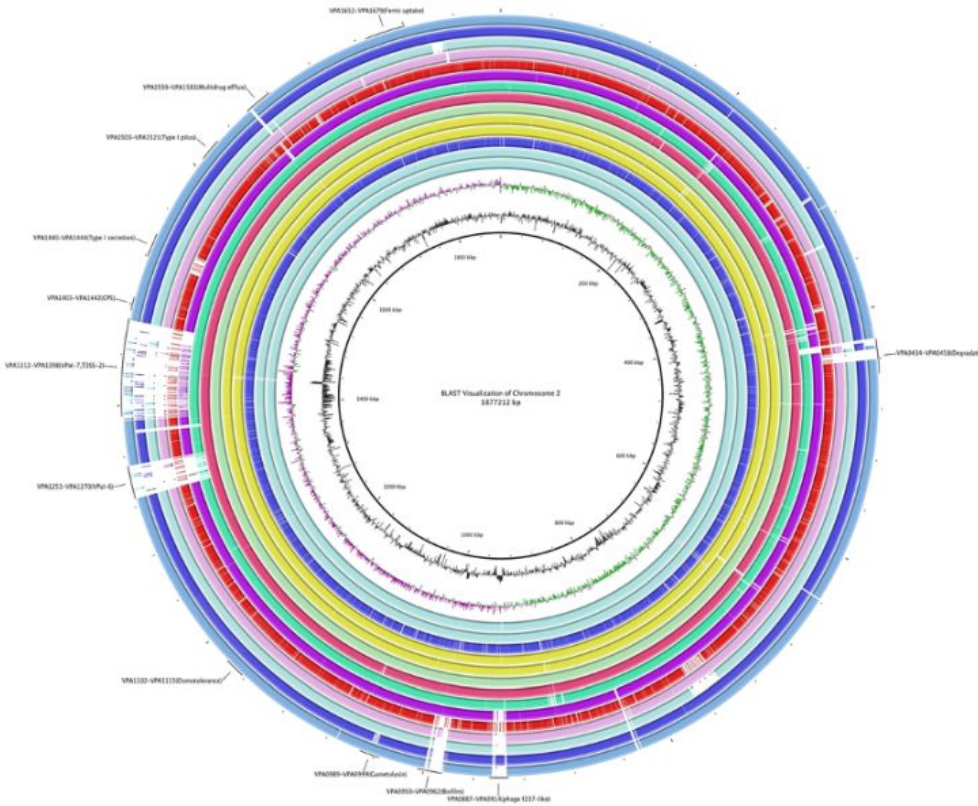
基于全基因组SNP建树



- 多基因组比较分析

基于基因组或基因簇在氨基酸水平上、核苷酸水平上的整体比较分析





参考文献：Comparative genomic analysis of clinical and environmental strains provides insight into the pathogenicity and evolution of Vibrio parahaemolyticus

3 微生物基因组研究方案设计

病原微生物组学分析

基因岛分析

效应物分析

病原微生物基因组分析

人源、动物、植物病原菌

耐药基因分析

毒力基因分析

致病系统
分析

对病原微生物进行分类鉴定、致病机理解析、调控系统分析等

非病原微生物组学分析

CRISPR分析

抗逆系统分析

环境微生物基因组分析

工业菌、古生菌、藻类等

转运系统分析

代谢系统分析

产量、环境适应性

解析微生物的环境适应性和重要代谢系统的作用机制

比较基因组分析研究思路

① 比较基因组之----功能研究

a. 表型异同的菌株之间比较分析，挖掘表型异同的基因组根源；

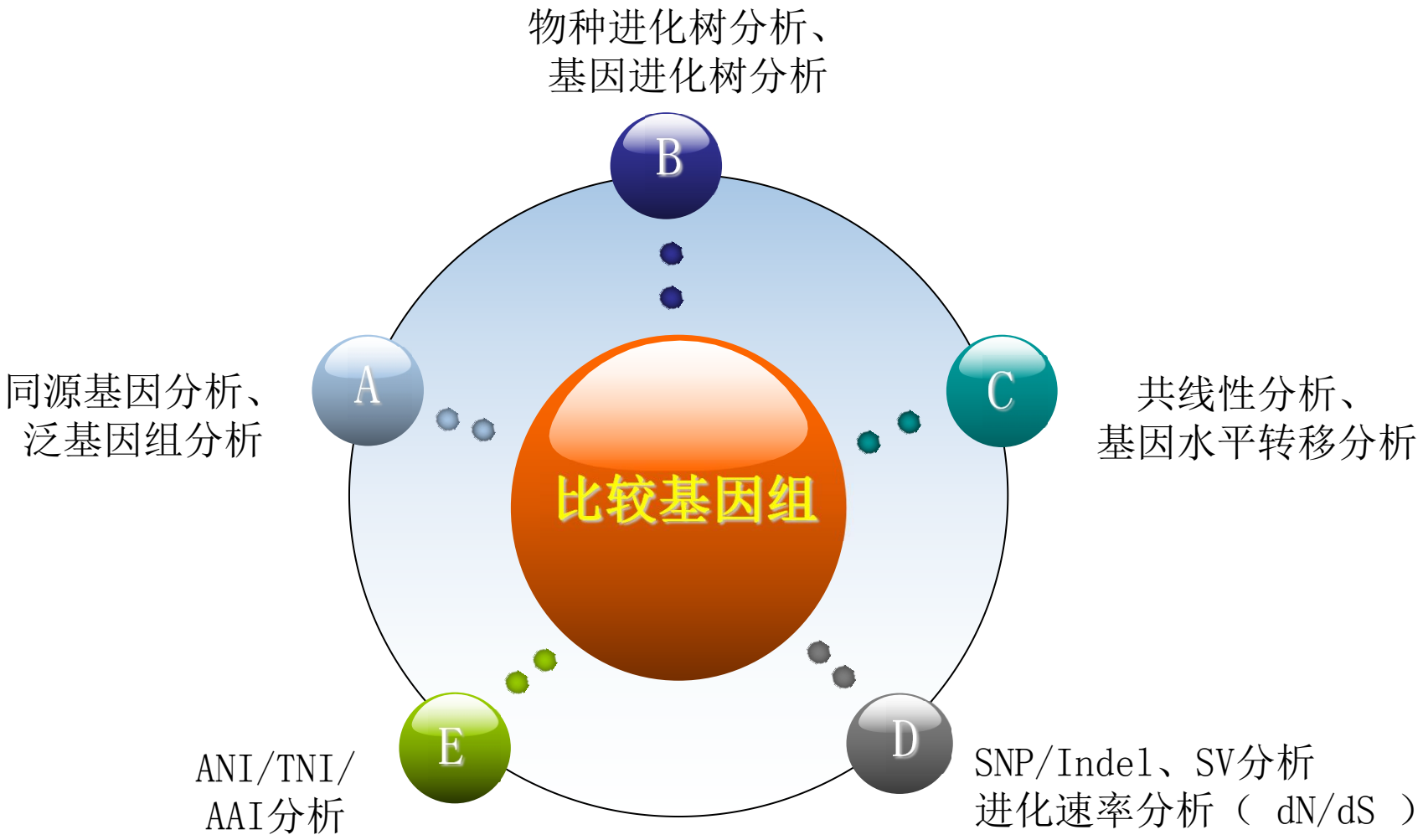
b. 强环境耐受株/环境变量敏感株、高产株/低产株……

② 比较基因组之----进化分类、功能研究

a. 一株/几株菌测序，与已报道的同种/属的比较分析，同源基因、共线性、进化树、泛基因组、基因簇比较揭示进化分类、功能机制；

b. 基因组测序文章研究中最常见、数量最多的类型；

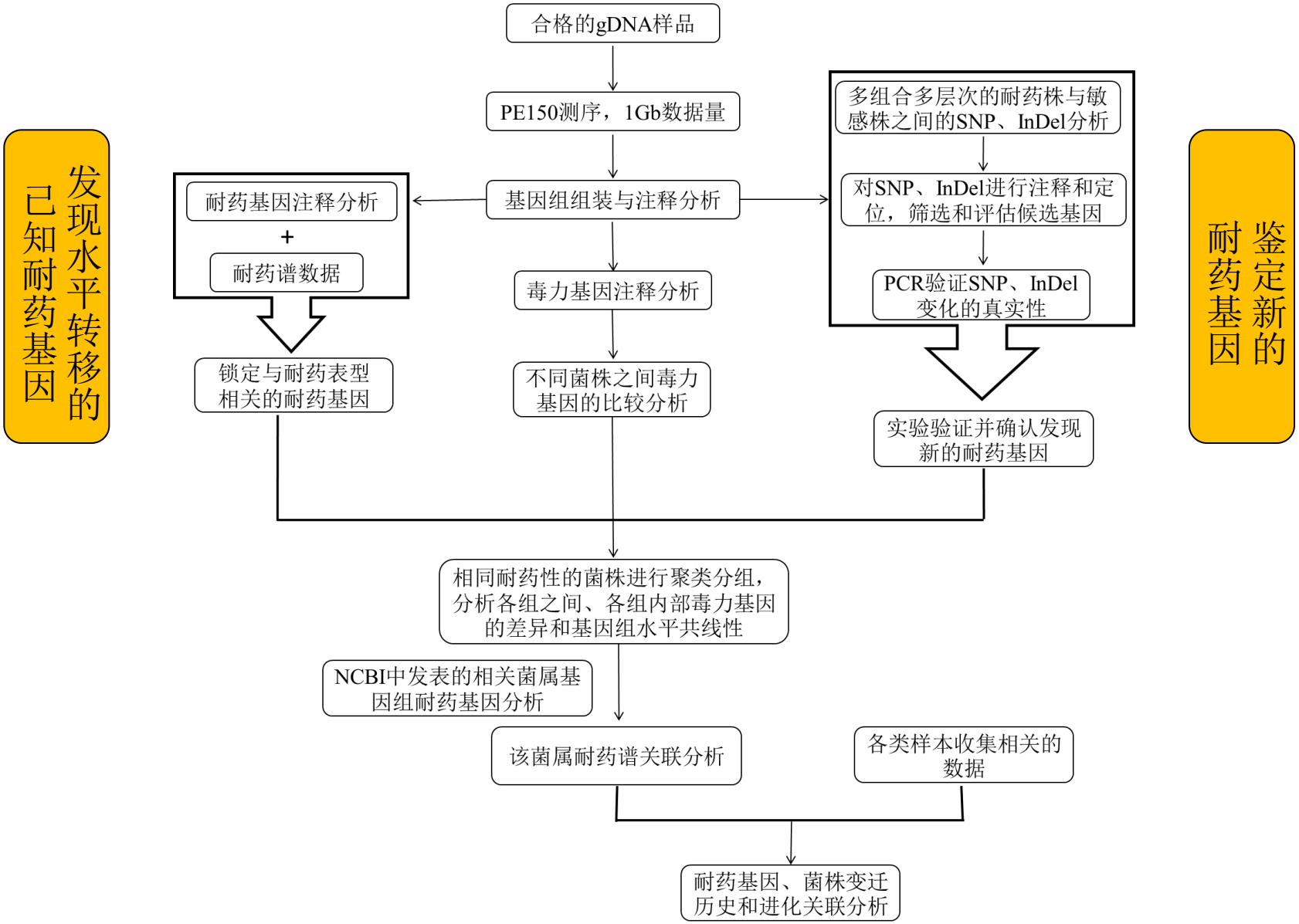
比较基因组分析研究思路



病原类微生物基因组分析

- 耐药基因挖掘
- 致病相关基因挖掘：胞外多糖、胞外蛋白酶、脂多糖、效应物（毒力基因+无毒基因）、分泌系统、调控系统（群体感应、双组分调控系统、环二鸟苷酸(c-di-GMP)分析）
- 基因岛分析
- 关键系统、基因簇的比较分析、进化分析
- 与宿主互作的相关基因研究（转录组为主）

对耐药基因的研究思路


合格的gDNA样品
PE150测序，1Gb数据量
基因组组装与注释分析
耐药基因注释分析
+
耐药谱数据
锁定与耐药表型相关的耐药基因
毒力基因注释分析
不同菌株之间毒力基因的比较分析
多组合多层次的耐药株与敏感株之间的SNP、InDel分析
对SNP、InDel进行注释和定位，筛选和评估候选基因
PCR验证SNP、InDel变化的真实性
实验验证并确认发现新的耐药基因
发现水平转移的已知耐药基因
鉴定新的耐药基因
相同耐药性的菌株进行聚类分组，分析各组之间、各组内部毒力基因的差异和基因组水平共线性
NCBI中发表的相关菌属基因组耐药基因分析
该菌属耐药谱关联分析
各类样本收集相关的数据
耐药基因、菌株变迁历史和进化关联分析

非病原微生物组学分析

CRISPR分析

抗逆系统分析

环境微生物基因组分析

工业菌、古生菌、藻类等

转运系统分析

代谢系统分析

产量、环境适应性

解析微生物的环境适应性和重要代谢系统的作用机制

基于测序的环境微生物研究方向

- 重金属富集与降解相关基因分析
- 某些次级代谢产物合成相关基因分析
- 耐药相关基因分析
- 毒素合成相关基因分析
- 降解某种化合物的相关基因分析
- 环境适用性相关基因分析
- 高底物耐受性、高产相关基因

合成生物学的基础

基于测序技术的工业微生物研究思路

基于测序技术的工业微生物研究思路

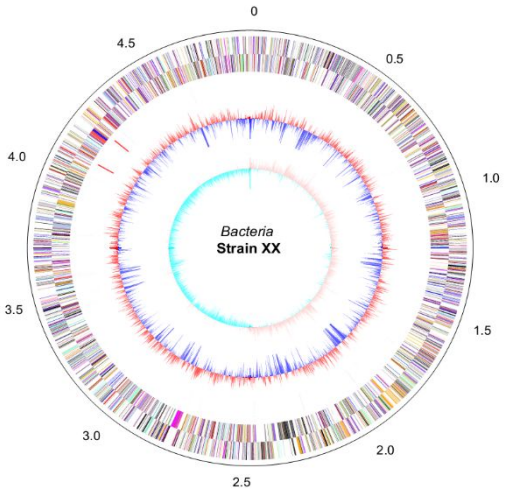
01 高产能菌株的筛选

自然筛选所得菌株

1、完成图测序获取全部基因信息和基因组结构信息、代谢通路信息；

2、根据表型反推代谢通路，锁定候选基因；

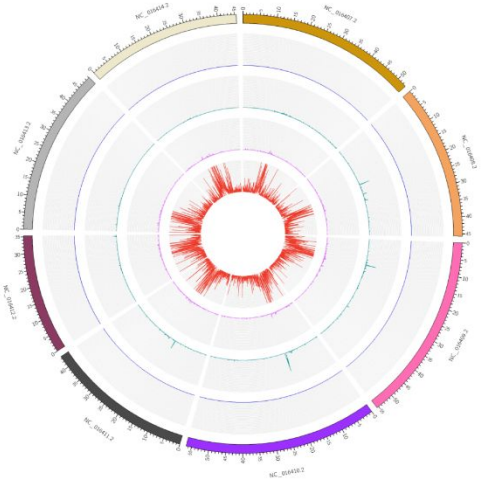
3、与其他已测菌株做差异基因比较分析，锁定候选基因。



诱变筛选所得菌株

1、扫描图测序或者重测序；

2、研究微进化，寻找结构变异（SNP、InDel、SV等）及其分布，与差异基因关联分析锁定候选基因和关键调控位点；



基于测序技术的工业微生物研究思路

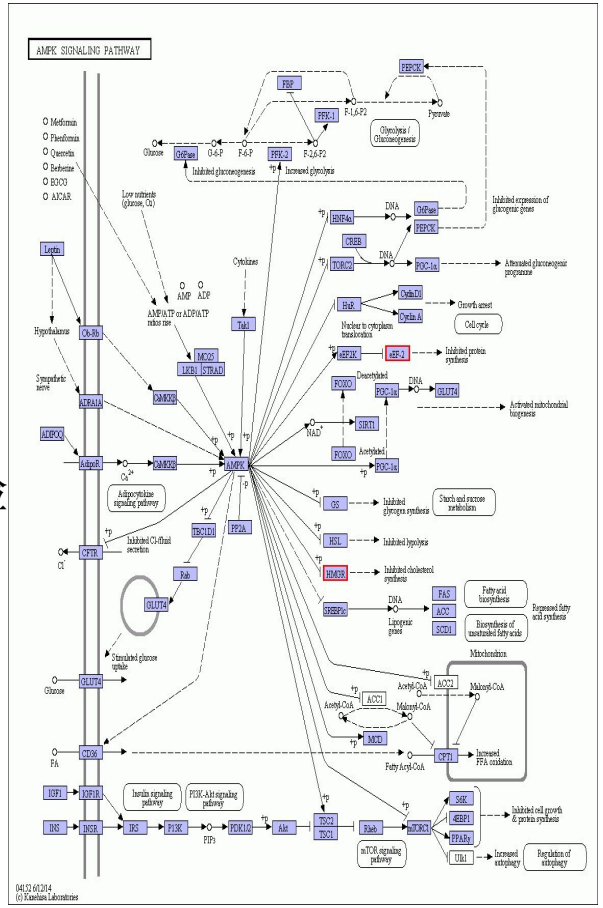
02高产菌株的调控研究

代谢通路分析

1、锁定与表型相关的多条代谢通路；

2、寻找各代谢基因、调控基因在代谢通路中的地位，推测可能采用的通路；

3、通过基因敲除、q-PCR、转录组等验证研究。

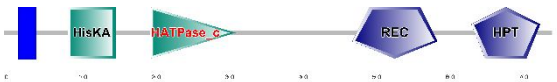


调控基因分析

1、通过文献报道、通路搜索获得调控相关候选基因集；

2、分析调控基因的结构域、启动子和调控方式；

3、如有必要对调控基因进行突变或高表达。



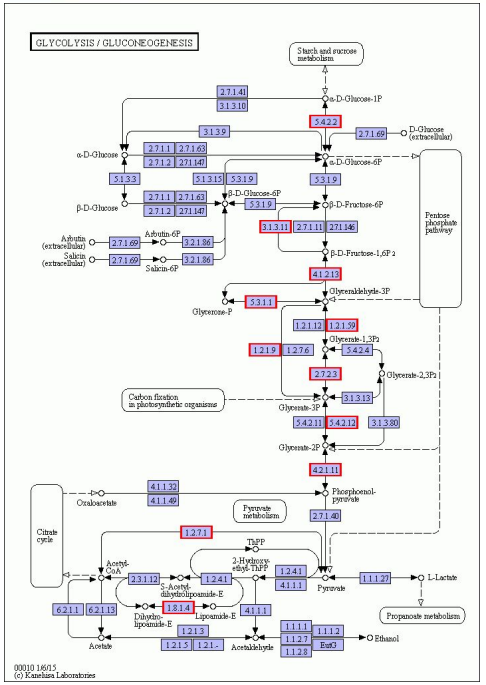
基于测序技术的工业微生物研究思路

03新物质合成基因筛选

前体合成相关基因

1、通过分析代谢通路锁定前体合成候选基因；

2、通过基因敲除等功能实验切断其中某些途径以提高产物合成。

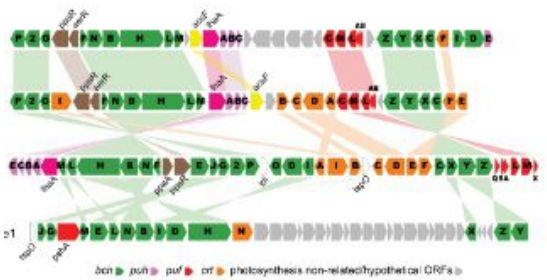


基因簇比较分析

1、与近缘物种进行基因簇比较分析获得彼此之间的差异；

2、分析基因簇两侧的正向和反向重复序列；

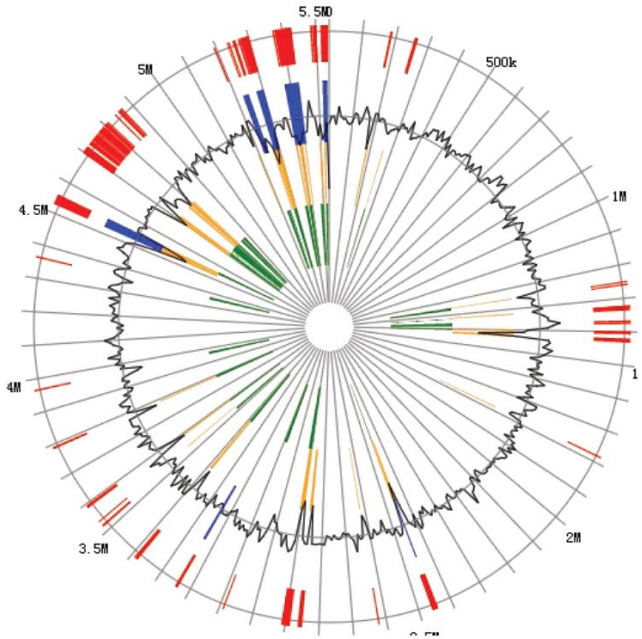
3、有助于解析基因簇的进化。

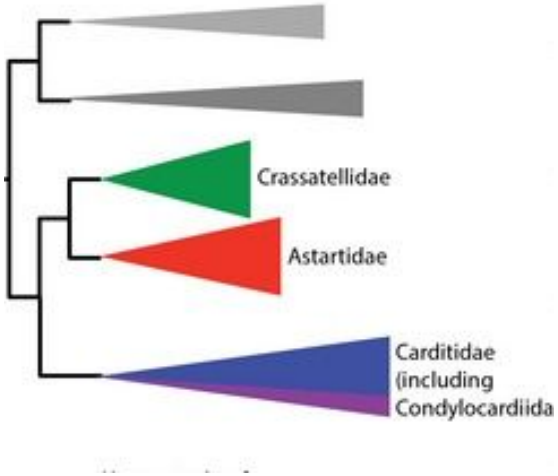


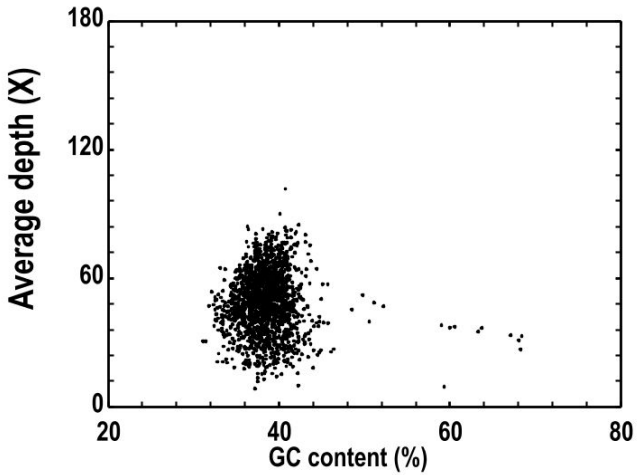
基于测序技术的工业微生物研究思路

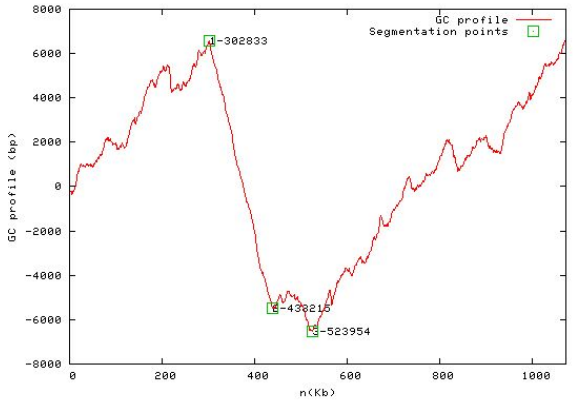
04其他分析

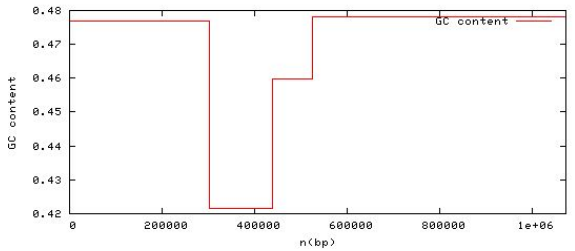
- 基因岛分析
- 前噬菌体预测
- CRISPR预测
- 高浓度底物耐受性相关基因分析
- 基因或基因簇水平转移分析分析











案例

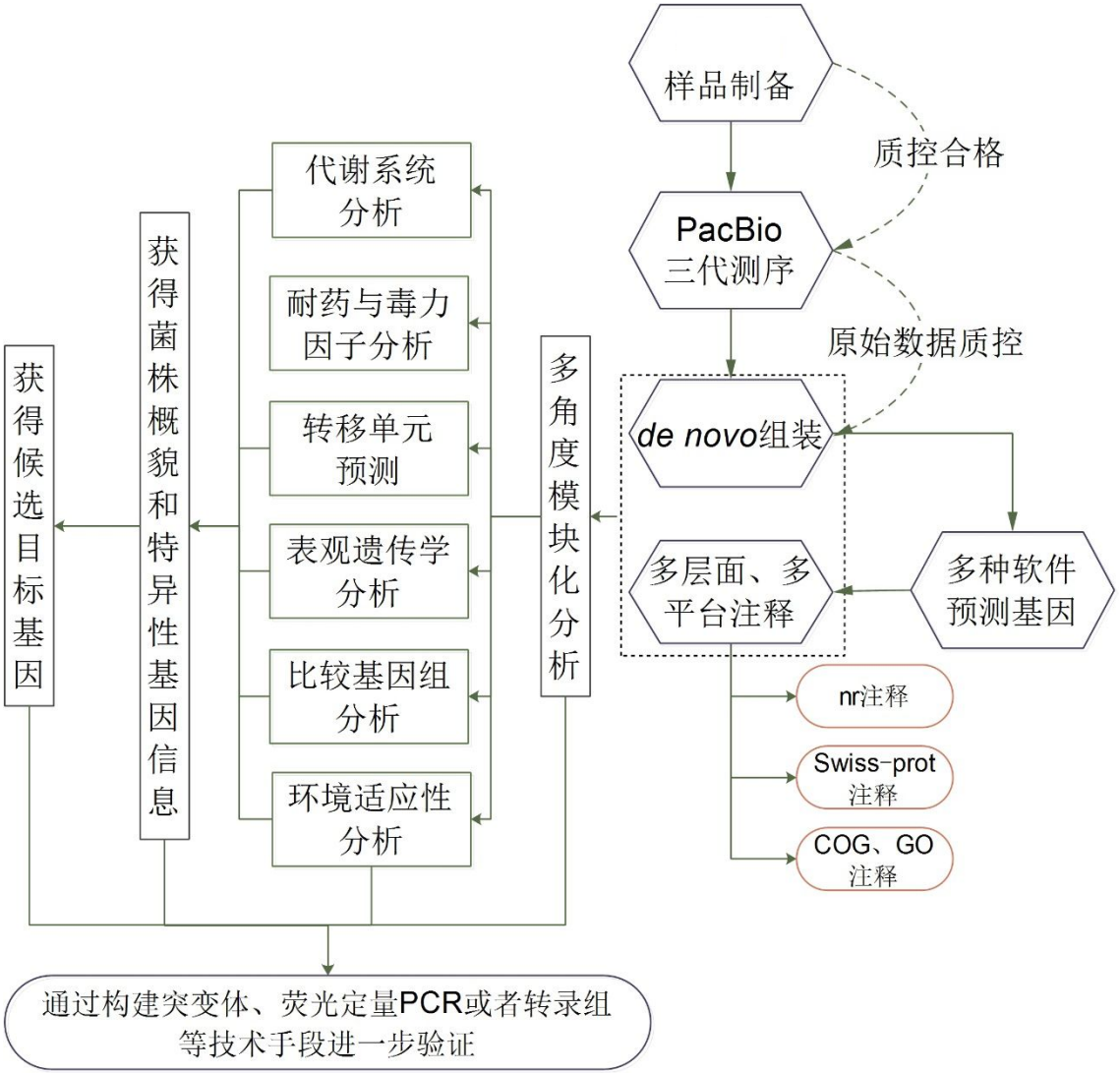
研究内容：

借助高通量测序技术和生物信息分析手段在组学层面对分离获得的一株菌进行深入的数据挖掘，寻找与温度、特定化合物代谢相关的基因或基因簇，同时在组学层面上了解该菌株与其他已测序菌株之间的异同点。

研究方案：

本研究可以在基因组特征序段层面、代谢系统层面、表观遗传学层面、比较基因组学层面等进行深入的数据挖掘和分析。

方案设计


样品制备
质控合格
PacBio
三代测序
原始数据质控
de novo组装
多层面、多
平台注释
多种软件
预测基因
nr注释
Swiss-prot
注释
COG、GO
注释
多
角
度
模
块
化
分
析
代谢系统
分析
耐药与毒力
因子分析
转移单元
预测
表观遗传学
分析
比较基因组
分析
环境适应性
分析
获
得
菌
株
概
貌
和
特
异
性
基
因
信
息
获
得
候
选
目
标
基
因
通过构建突变体、荧光定量PCR或者转录组
等技术手段进一步验证

美吉生物
Majorbio

021-31050576

mdna@majorbio.com

2406642459

Majorbio

美吉生物官方微信平台

美吉生物 Majorbio

讲座资料获取方式：

1.发送 **您的姓名+单位名称+联系电话** ☞mdna@majorbio.com即可获取讲座PPT

2.免费注册云平台账号（www.i-sanger.com），点击首页美吉大讲堂，免费观看讲座视频，http://www.i-sanger.com/video/index.html

美吉生物 Majorbio

生信培训班开课啦！

微生物基因组测序与分析数据挖掘精品培训班（偏产品班）　11月21日-24日

日程		主题
Day 1	上午	微生物基因组基础研究和微生物产业
	下午	微生物基因组测序的各类分析数据高级解读
		高通量测序技术原理、发展历程和在微生物研究中的应用
Day 2	上午	环境类、工业生产微生物在组学层面的研究方法和文章设计思路
		Linux 操作基础入门与实操练习
	下午	微生物组学层面差异位点分析（理论+操作）
		代谢相关基因的挖掘（上机操作）
Day 3	上午	病原细菌在组学层面的研究方法和文章设计思路
		质粒的特征和研究方法（一）
	下午	质粒的特征和研究方法（二）
		病原微生物的数据挖掘（上机操作）
Day 4	上午	微生物基因组高级作图分析（上机操作）
	下午	课程答疑
		发结业证书、参观美吉生物总部实验室

美吉生物 Majorbio

生信培训班开课啦！

微生物基因组测序与分析生信培训班（偏生信班）　　11月27日-12月1日

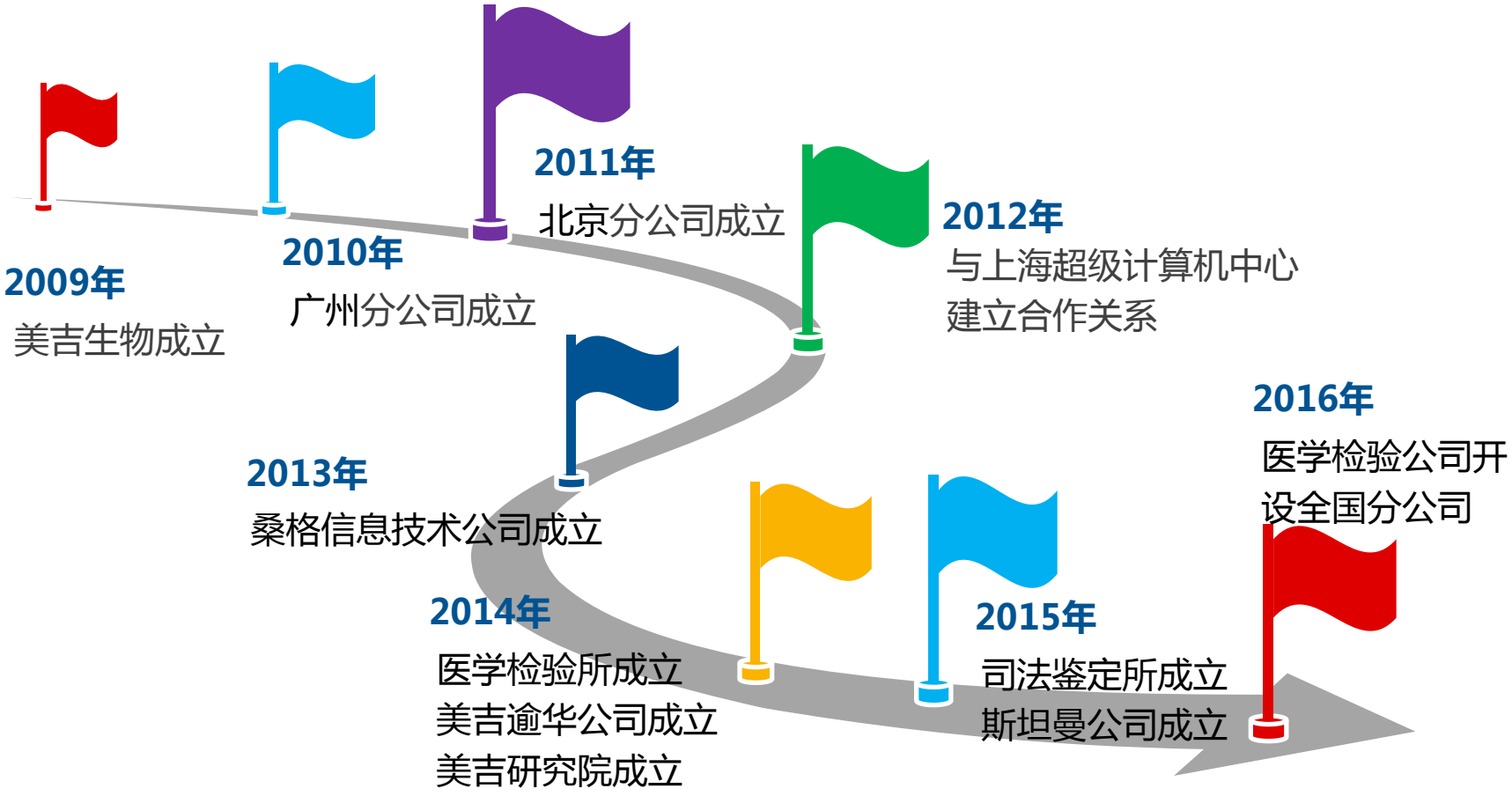
日程		主题
Day 1	上午	高通量测序原理和在微生物研究领域的应用
		基于高通量测序技术的微生物基因组研究流程
	下午	Linux 操作基础入门与实操练习（实战）
		测序数据的质控操作
Day 2	上午	二代测序的数据组装和基因预测
	下午	主流数据库的使用详解
		基因组注释
Day 3	上午	基因组重测序分析流程（理论+实战）
	下午	数据上传 NCBI 操作
		各类实用脚本的操作
Day 4	上午	基因组学水平上的数据高级挖掘 （一）
	下午	基因组学水平上的数据高级挖掘 （二）
		微生物基因组结题报告高级解读
Day 5	上午	微生物基因组研究热点和文章设计思路
	下午	课程答疑
		颁发结业证书、参观美吉生物总部实验室

美吉生物
Majorbio

4 美吉与高通量测序

关于美吉

发展历程


2009年
美吉生物成立
2010年
广州分公司成立
2011年
北京分公司成立
2012年
与上海超级计算机中心
建立合作关系
2013年
桑格信息技术公司成立
2014年
医学检验所成立
美吉逾华公司成立
美吉研究院成立
2015年
司法鉴定所成立
斯坦曼公司成立
2016年
医学检验公司开
设全国分公司

美吉生物
Majorbio

-Sanger

美吉医学检验
MAJORDX

上海美吉医学检验有限公司
法医物证司法鉴定所

美吉谕华
YH·Health

斯坦曼生物医药
Steinman-Biomed

美吉生物信息研究院
Bio-Informatics Institute

美吉生物 Majorbio

关于美吉

高通量业务 —— 一站式服务

迄今为止，我们的合作单位已达1900余家，合作的课题组已超过6800个……

我们已助力各位客户发表了600+篇SCI论文，总影响因子超过1700……

"M" (majorbio)

"W" (we)

美吉生物® Majorbio 美吉

生命 科技

- 上半部分是M(majorbio)象征美吉，下半部分是W（we）象征我们，寓意我们众志成城、齐心协力撑起一个伟大的美吉。
- 绿色代表生命，蓝色代表科技，寓意我们从事的是基因科技与人类健康事业。

谢谢您的欣赏

美吉生物 Majorbio

地址/Addr:上海市浦东新区国际医学园区康新公路3399号3号楼
电话/Tel：021-51875086　　服务热线：400 660 1216
网址/Web：www.majorbio.com　　传真/Fax：021-51875086-8002