基于二代测序技术的宏基因组学分析

G05成员：王磊 刘唐 李春梅 王晨曲

汇 报 人：王晨曲

Our Team

G05-A 王磊
环境科学与工程学院

群落结构分析、ORF预测

G05-B 刘唐
环境科学与工程学院

群落结构分析，代谢途径分析

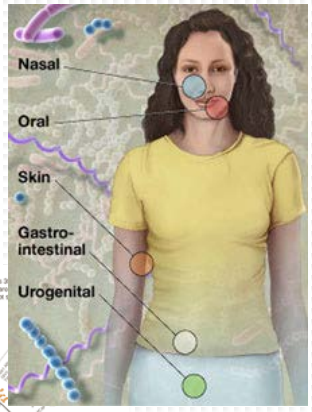
G05-C 李春梅
生命科学学院-BIOPIC

宏基因组测序文库构建、测序数据拼接

G05-D 王晨曲
生命科学学院-分子医学研究所所

测序数据拼接、COG聚类

人类基因组计划
Human Genome Project



人类微生物宏基因组计划
Human Microbiome Project

肠道宏基因组计划
Metagenomics of the Human Intestinal Tract

地球微生物宏基因组计划
Earth Microbiome Project

宏基因组学（Metagenomics）又叫微生物环境基因组学，宏基因组学通过直接从环境样品中提取全部微生物的DNA,构建宏基因组文库,利用基因组学的研究策略研究环境样品所包含的全部微生物的遗传组成及其群落功能.

宏基因组学的研究对象是特定环境中的总DNA，不是某特定的微生物或其细胞中的总DNA，不需要对微生物进行分离培养和纯化，这对我们认识和利用95%以上的未培养微生物提供了一条新的途径。

THE METAGENOMICS PROCESS

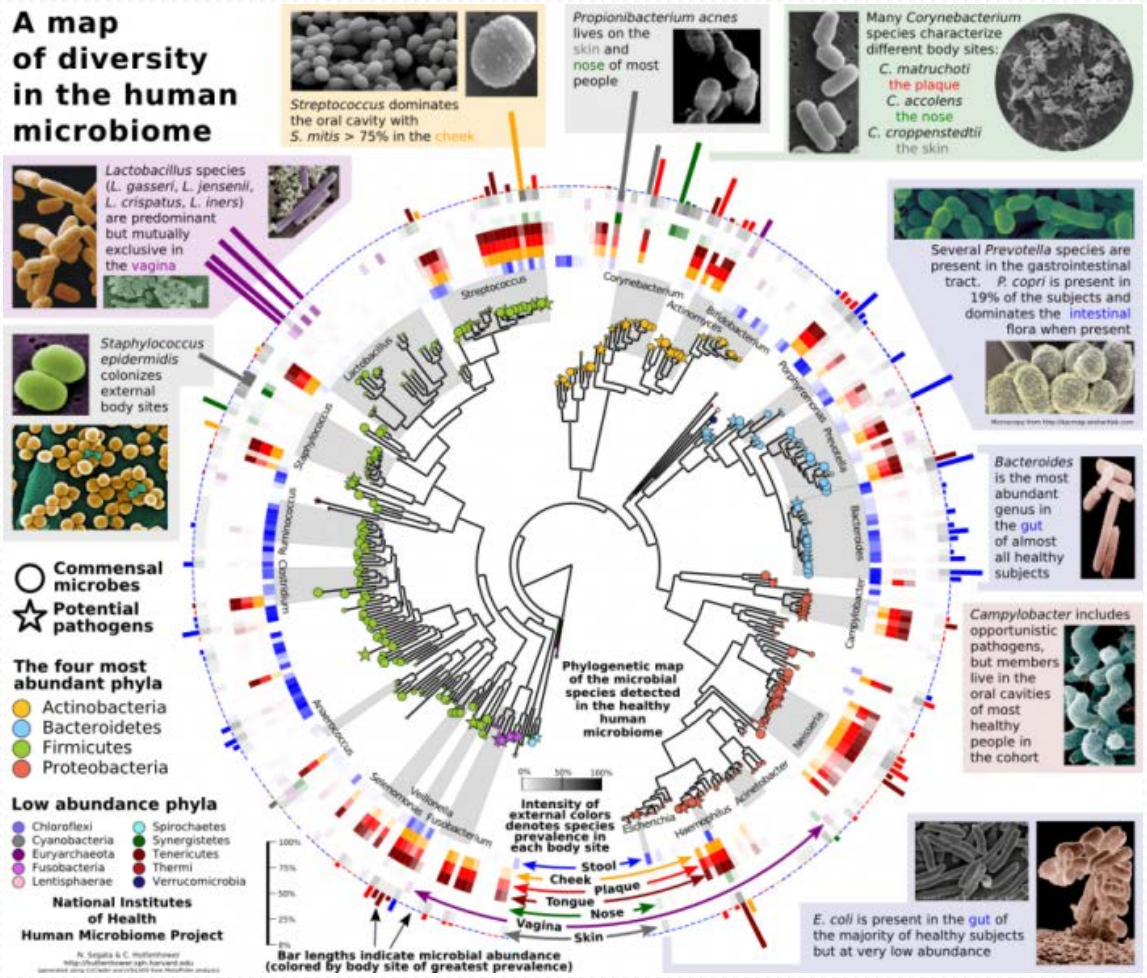
Extract all DNA from microbial community in sampled environment

DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

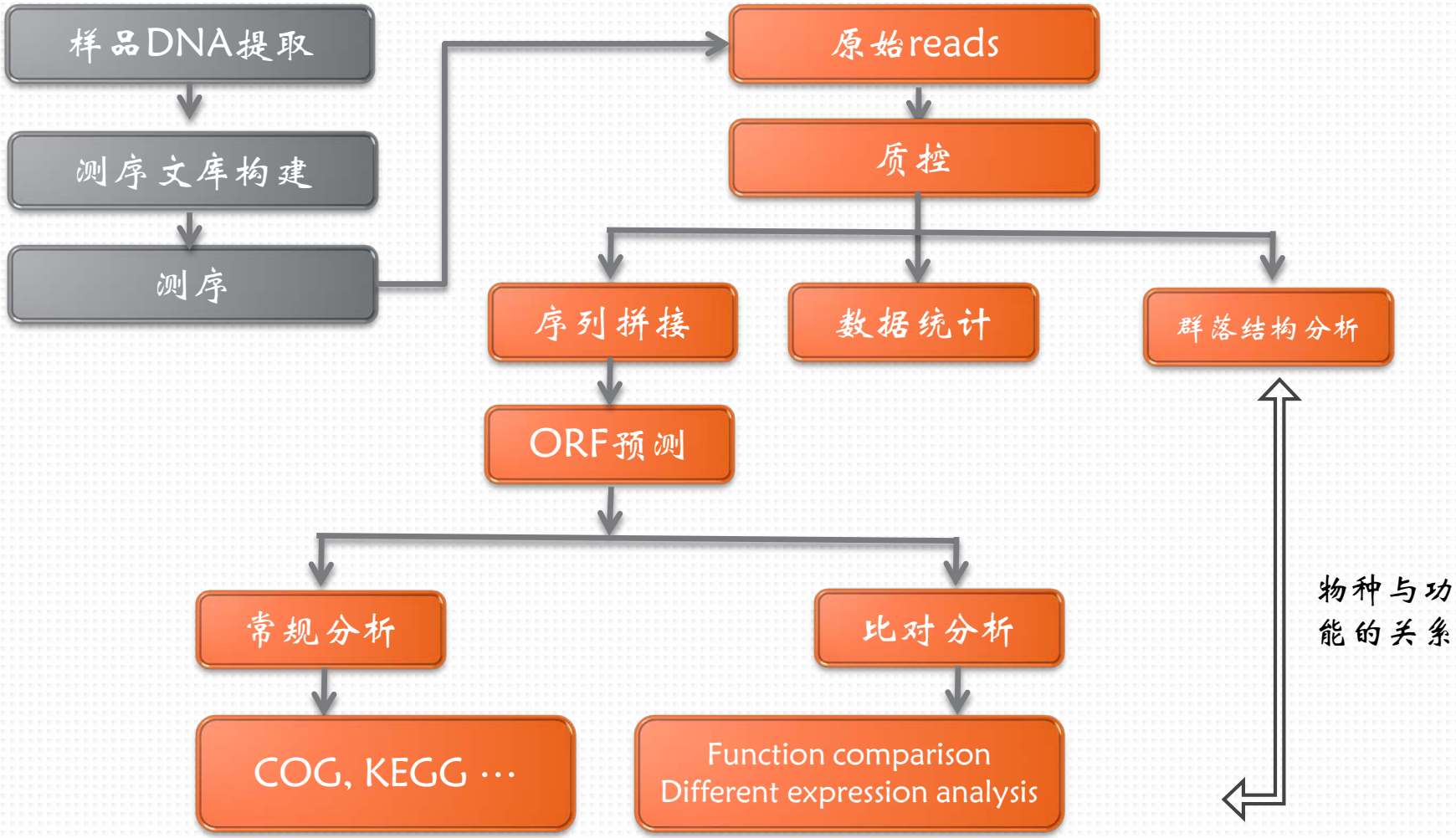


群落结构复杂
基因组信息不完整…

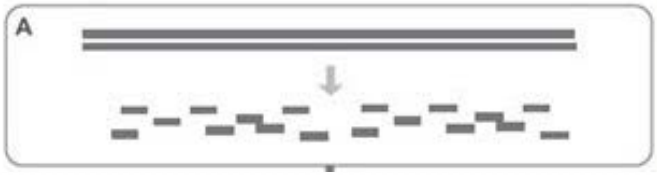
如何高效分析宏基因组测序产生的海量数据？

微生物多样性
群落结构
进化关系
功能活性
相互协作关系
与环境之间的关系

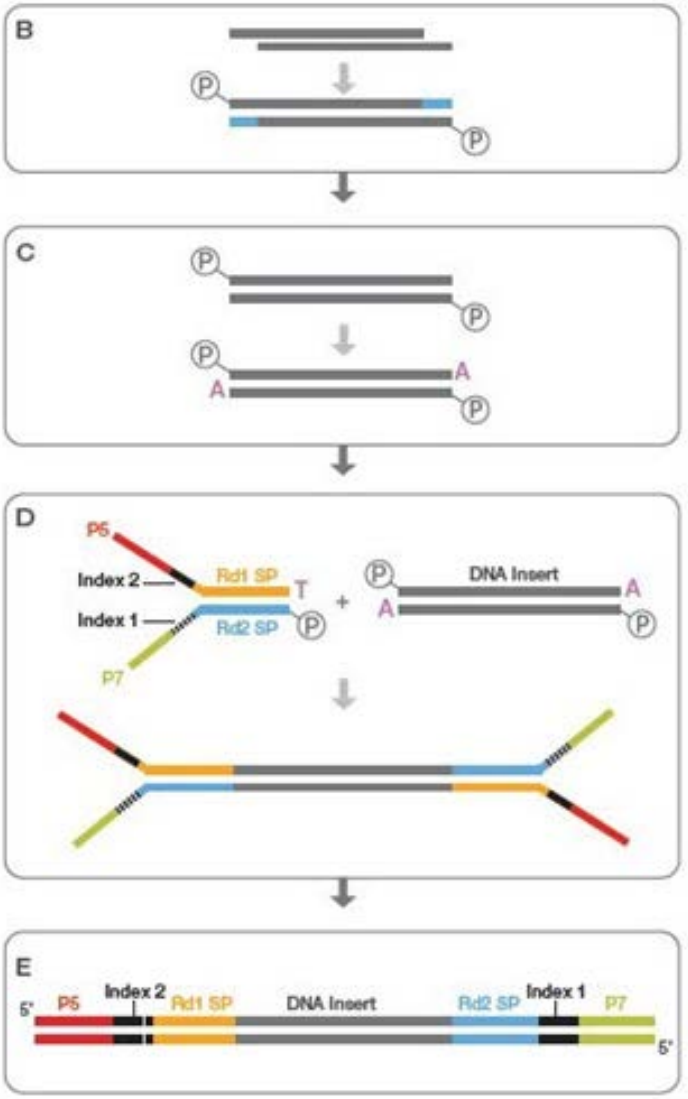
基于二代测序技术的宏基因组学分析流程



Illumina 测序文库构建流程



Illumina HiSeq 2000 Platform

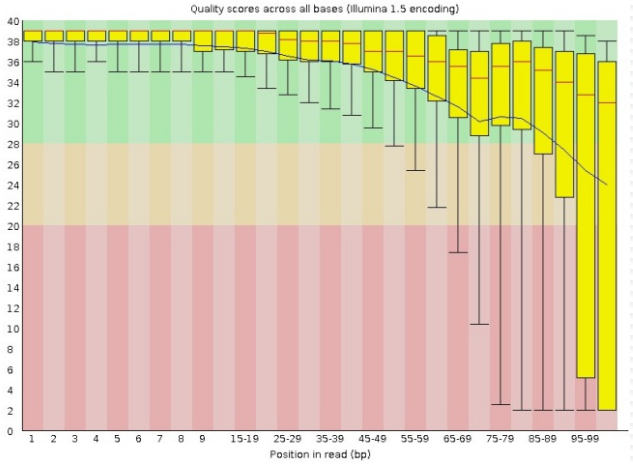


数据统计与预处理

- 数据统计 （FastQC）

```
/rd1/user/wangcq/project/metagenome-ABC/rawdata $ fastqc -f fastq -o ./FastqQC SRR609198_1.fastq SRR609198_2.fastq
```

Measure	Value
Filename	SRR609198_1.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	16350599
Filtered Sequences	0
Sequence length	100
%GC	51

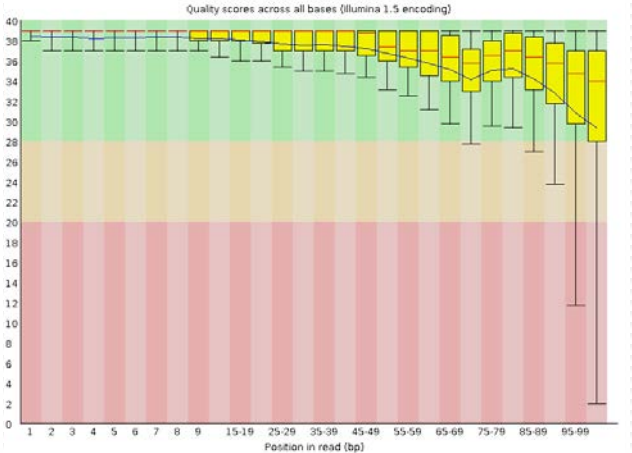


- 数据质控 （DynamicTrim.pl , LengthSort.pl)

```
/rd1/user/wangcq/project/metagenome-ABC/rawdata $ perl DynamicTrim.pl -o ./fqfilter SRR609198_1.fastq SRR609198_2.fastq
```

```
/rd1/user/wangcq/project/metagenome-ABC/rawdata $ perl LengthSort.pl -o ./fqfilter SRR609198_1.fastq SRR609198_2.fastq
```

Measure	Value
Filename	read_1.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	11628541
Filtered Sequences	0
Sequence length	100
%GC	49



序列De novo assembly (SOAPdenovo2)

- 建立Contig 文件

```
#maximal read length
max_rd_len=100
[LIB]
#average insert size
avg_ins=200
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#use only first 100 bps of each read
rd_len_cutoff=100
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
#a pair of fastq file, read 1 file should always be followed by read 2 file
q1=/path/**LIBNAMEA**/fastq1_read_1.fq
q2=/path/**LIBNAMEA**/fastq1_read_2.fq
```

```
/rd1/user/wangcq/project/metagenome-ABC/result $ less ref.config
# maximal read length
max_rd_len=100
[LIB]
# average insert size
avg_ins=200
# if sequence needs to be reversed
reverse_seq=0
# in which parts the reads are used 1=contig; 2=scaffold;3=both
asm_flags=3
rd_len_cutoff=100
rank=1
pair_num_cutoff=3
map_len=32
q1=/rd1/user/wangcq/project/metagenome-ABC/rawdata/read_1.fastq
q2=/rd1/user/wangcq/project/metagenome-ABC/rawdata/read_2.fastq
```

序列De novo assembly (SOAPdenovo2)

- Assembly

```
${bin} all -s config_file -K 63 -R -o graph_prefix 1>ass.log 2>ass.err
```

User can also choose to run the assembly process step by step as:
step1:

```
${bin} pregraph -s config_file -K 63 -R -o graph_prefix 1>pregraph.log 2>pregraph.err
OR
${bin} sparse_pregraph -s config_file -K 63 -z 5000000000 -R -o graph_prefix 1>pregraph.log 2>pregraph.err
```

step2:

```
${bin} contig -g graph_prefix -R 1>contig.log 2>contig.err
```

step3:

```
${bin} map -s config_file -g graph_prefix 1>map.log 2>map.err
```

step4:

```
${bin} scaff -g graph_prefix -F 1>scaff.log 2>scaff.err
```

```
/rd1/user/wangcq/project/metagenome-ABC/result $ SOAPdenovo all -s ref.config -K 35 -O 16S 1>ass.log 2>ass.err
```

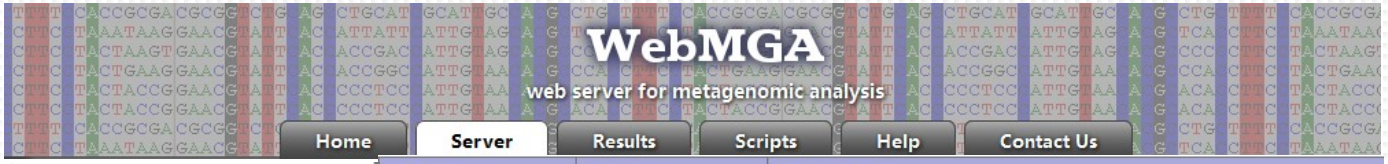
	Contig	Scaffold
Size_includeN	269472205	270556143
Size_withoutN	269472205	268838613
Number	1753401	1712316
Mean_Size	153	158
Median_Size	100	100
Longest_Seq	98903	98903
Shortest_Seq	100	100

Length	Contig	Scaffolds
>100	480523	444655
>500	45519	50481
>1K	14122	15908
>10K	317	361
>100K	0	0
>1M	0	0

- 筛选长度大于500bp的contig进行下列分析

功能分析

聚类分析
rRNA预测
tRNA预测
ORF预测
功能注释
通路分析
序列信息
质量控制
序列过滤
种群分类
OUT识别
格式转换



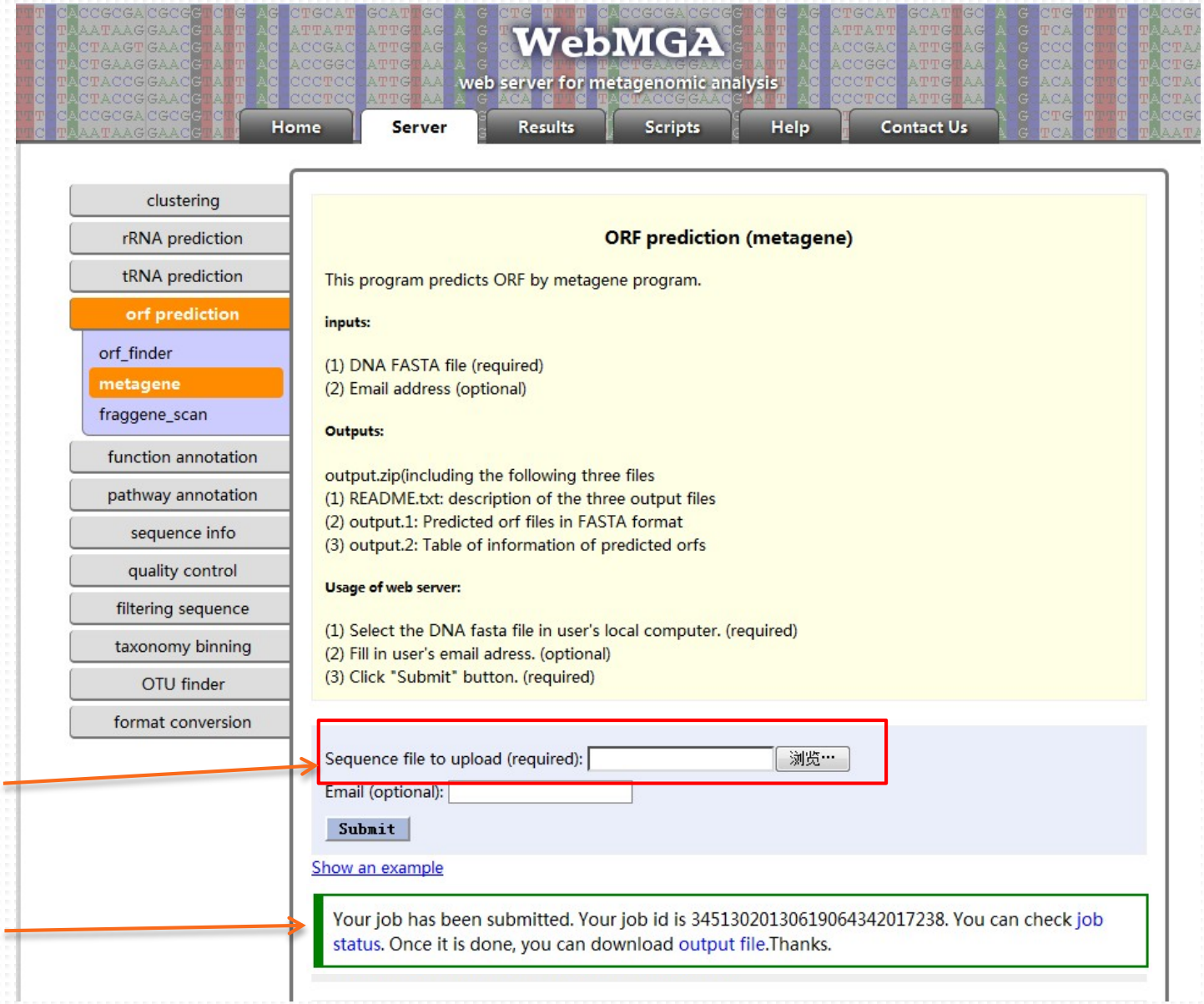
clustering
rRNA prediction
tRNA prediction
orf prediction
function annotation
pathway annotation
sequence info
quality control
filtering sequence
taxonomy binning
OTU finder
format conversion

Category	Name	Description
clustering	cd-hit-est	fast DNA clustering
	cd-hit	protein clustering
	h-cd-hit	hierarchical protein clustering
	cd-hit-454	filtering 454 duplicate reads
rRNA prediction	blastn_rRNA	rRNA prediction by blastn program
	hmm_rRNA	rRNA prediction by hmmer 3.0 program
tRNA prediction	tRNA	tRNA prediction by tRNAscan-SE program
orf prediction	orf_finder	orf prediction by six-reading-frame technique
	metagene	orf prediction by metagene program
	fraggene_scan	orf prediction by fraggene_scan program
function annotation	cog	protein function annotation by COG database
	kog	protein function annotation by KOG database
	prk	protein function annotation by NCBI PRK database
	pfam	protein function annotation by pfam database
	tigrfam	protein function annotation by tigrfam database
pathway annotation	kegg	pathway annoation by KEGG database
sequence info	fna_stat	statistics of DNA sequences including length and GC content
	faa_stat	statistics of protein sequences including length
quality control	qc_filter_fastq	removing reads with low average quality for a read with fastq format
	qc_filter_fasta_qual	removing reads with low average quality for a read with fasta+quality format
	trimm	trimming the low-quality tail of illumina reads
filtering sequence	filter_human	filtering human sequences from reads
taxonomy binning	rdp_binning	taxonomic binning by rdp classifier program
	frhit_binning	taxonomic binning by frhit program
OTU finder	cd-hit-otu	OTU finder by cd-hit-otu program
format conversion	fastq2fasta	converting fastq file to fasta file

输入文件：Fasta格式

输出文件：gz格式

功能分析—ORF预测



输入contig >500bp 的Fasta文件

上传文件后，点击job status，下载数据。

功能分析—ORF预测

下载数据

Output 1

```
>4675112.1 /source=4675112 /start=1 /end=501 /frame=1 /length=167
KNRIHEFLNILWDGGAIYTTGRQGPNISKGLLIKENVATGKRPSGGGNTFYTDGGSRNIRLEKNVSLNNP
IGVTDYGPPSPVGDPLPYPPYSIANDVPYGSDMGGCITYGDLEYRENYWLEGSLPNTIVFNNFLINSLVG
FPPYSYQGFFDICPAVIDGVSYPKNLS
>4675114.1 /source=4675114 /start=49 /end=501 /frame=-1 /length=150
SSPKYAYNFDHYNYQSGKKAKVVTAEHTSLPATQPELLVASNDDSFVPVAEPTKEELETAVAEVRKTYLQ
MDKAERKEFRRDVKNFIKEKKESIKAVKKTNAMDNDLKLAAIFGAVGIVALIISGDVFYIIGGIALIIGV
VFFVKWLVRQ*
>4675116.1 /source=4675116 /start=3 /end=98 /frame=3 /length=31
KLIVLSHDGYLFLVDVFIRRLSVKNVITNFK*
>4675116.2 /source=4675116 /start=133 /end=501 /frame=-1 /length=122
AFTTNNLILHNIEELEPLKNLLMLYNSKLNILHFDDIEGVGDNKKKNKAFLKTHFSDITYENIDSYSKDL
YKTVAGYIHANDIKLIAMMRKKHSFLERLFNRHPEETLAFNIDVPFLVMENS*
>4675118.1 /source=4675118 /start=3 /end=500 /frame=3 /length=166
PIKNRVLIDINELSKETAGMLAKFYTSNKNFIINKAYDPSFYSKPIEIAEDARLNEENTYWDTLRHEPLS
ETEKSVYRMIDTLRNIPVVKTYTDIIKVLIDGYHSVGKVELGPYLGLVAWNTMEGVRVQGGLKTNYKFSK
NWIYTGIGAYGFDDERFKYFLSAQRI
```

Output 2

#ORF	source_DNA	start	end	strand	length
4675112.1	4675112	1	501	+	167
4675114.1	4675114	49	501	-	150
4675116.1	4675116	3	98	+	31
4675116.2	4675116	133	501	-	122
4675118.1	4675118	3	500	+	166
4675120.1	4675120	3	467	+	154
4675122.1	4675122	3	500	+	166
4675124.1	4675124	1	501	-	167
4675126.1	4675126	1	501	-	167
4675128.1	4675128	2	499	+	166
4675130.1	4675130	1	501	+	167
4675132.1	4675132	1	501	+	167
4675134.1	4675134	3	89	+	28
4675134.2	4675134	76	501	+	142
4675136.1	4675136	2	295	+	97
4675136.2	4675136	292	501	+	70

57,312 ORF (>100nt)

功能分析—功能注释

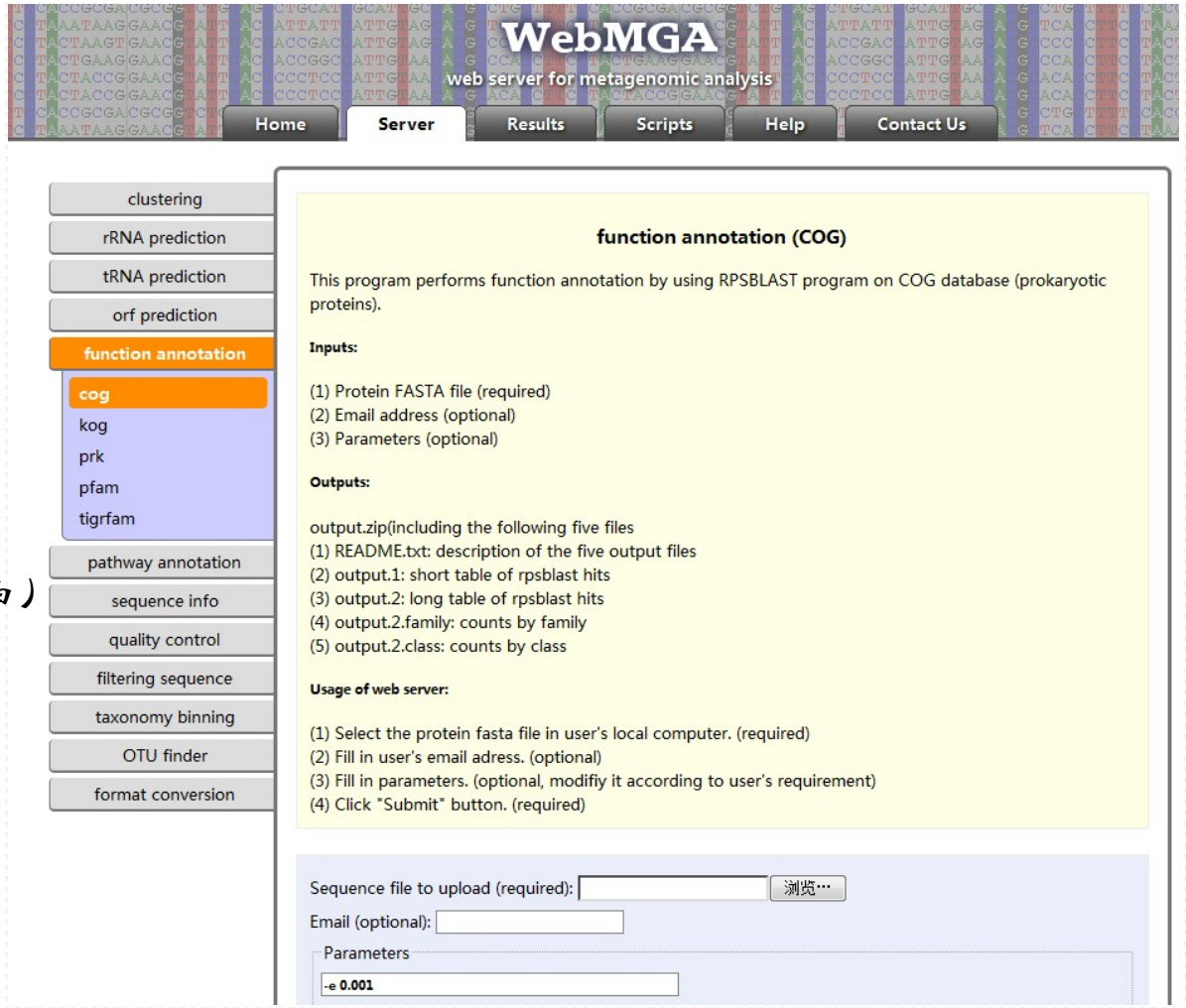
COG (Clusters of orthologous Groups)

KOG (EuKaryotic Orthologous Groups)

PRK (Protein K(c)lusters)

Pfam (Protein family)

TIGRfam (主要对象是细菌和古细菌蛋白)



功能分析—功能注释

COG (Clusters of orthologous Groups)

D:\wangchenqu\slides\ABC\big homework\cog\output.1

```
#Query  Hit     E-value Identity        Score   Query-start     Query-end       Hit-start       Hit-end Hit-length
4684316 COG3889 1e-05   26      45.0    18      140     758     864     872
4685340 COG3889 4e-06   25      47.0    305     411     761     855     872
```

Output1: COG hits

D:\wangchenqu\slides\ABC\big homework\cog\output.2

```
#Query  Hit     E-value Identity        Score   Query-start     Query-end       Hit-start       Hit-end Hit-length      description     class   class description
4684316 COG3889 1e-05   26      45.0    18      140     758     864     872     Predicted solute binding protein        R       General function prediction only
4685340 COG3889 4e-06   25      47.0    305     411     761     855     872     Predicted solute binding protein        R       General function prediction only
4698652 COG3846 3e-04   28      40.7    247     372     282     413     452     Type IV secretory pathway, TrbL components      U       Intracellular trafficking, secret
```

Output2: COG hits and classify

D:\wangchenqu\slides\ABC\big homework\cog\output.2.class

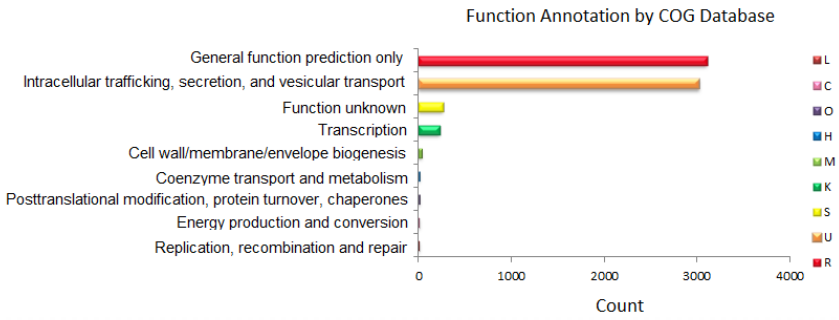
```
#COG class      count   description
C       8       Energy production and conversion
H       14      Coenzyme transport and metabolism
K       227     Transcription
L
M
O
R
S
U
```

Output3: COG class

D:\wangchenqu\slides\ABC\big homework\cog\output.2.family

```
#COG    count   description     class   class description
COG0243 1       Anaerobic dehydrogenases, typically selenocysteine-containing   C       Energy production and conversion
COG0484 13      DnaJ-class molecular chaperone with C-terminal Zn finger domain O       Posttranslational modification, p
COG1107 2       Archaea-specific RecJ-like exonuclease, contains DnaJ-type Zn finger domain     L       Replication, recor
COG1251 6       NAD(P)H-nitrite reductase       C       Energy production and conversion
COG1429 14      Cobalamin biosynthesis protein CobN and related Mg-chelatases   H       Coenzyme transport and metabolism
COG1512 1       Beta-propeller domains of methanol dehydrogenase type   R       General function prediction only
COG3064 37      Membrane protein involved in colicin uptake     M       Cell wall/membrane/envelope biogenesis
COG3846 3017    Type IV secretory pathway, TrbL components      U       Intracellular trafficking, secretion, and vesicula
COG3883 1       Uncharacterized protein conserved in bacteria   S       Function unknown
```

Output4: COG family



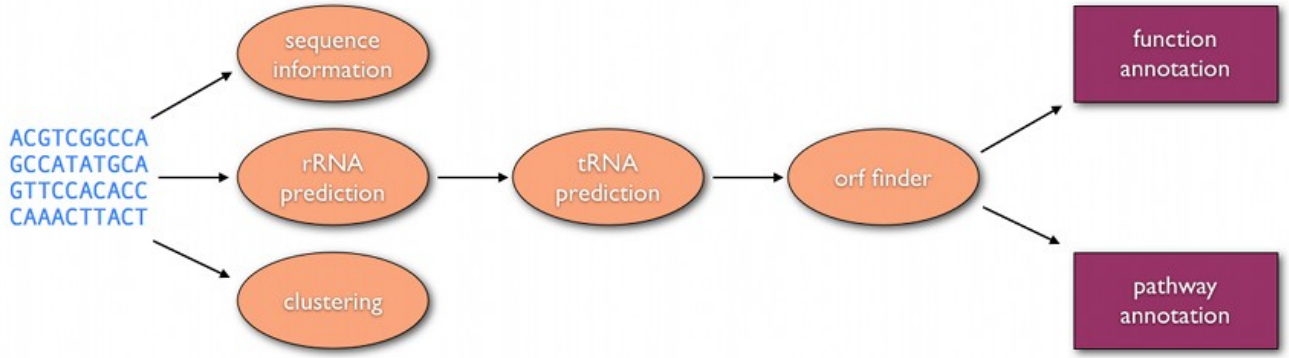
#COG	count	description	class	class description
COG3889	3107	Predicted solute binding protein	R	General function prediction only
COG3846	3017	Type IV secretory pathway, TrbL components	U	Intracellular trafficking, secretion, and vesicular transport
COG5164	227	Transcription elongation factor	K	Transcription
COG4278	200	Uncharacterized conserved protein	S	Function unknown
COG4260	53	Putative virion core protein (lumpy skin disease virus)	S	Function unknown
COG3064	37	Membrane protein involved in colicin uptake	M	Cell wall/membrane/envelope biogenesis
COG1429	14	Cobalamin biosynthesis protein CobN and related Mg-chelatases	H	Coenzyme transport and metabolism
COG0484	13	DnaJ-class molecular chaperone with C-terminal Zn finger domain	O	Posttranslational modification, protein turnover, chaperones
COG5283	8	Phage-related tail protein	S	Function unknown
COG1251	6	NAD(P)H-nitrite reductase	C	Energy production and conversion
COG1107	2	Archaea-specific RecJ-like exonuclease, contains DnaJ-type Zn finger domain	L	Replication, recombination and repair
COG0243	1	Anaerobic dehydrogenases, typically selenocysteine-containing	C	Energy production and conversion
COG1512	1	Beta-propeller domains of methanol dehydrogenase type	R	General function prediction only
COG3883	1	Uncharacterized protein conserved in bacteria	S	Function unknown
COG4263	1	Nitrous oxide reductase	C	Energy production and conversion
COG4386	1	Mu-like prophage tail sheath protein gpL	R	General function prediction only
COG4907	1	Predicted membrane protein	S	Function unknown
COG5281	1	Phage-related minor tail protein	S	Function unknown

功能分析

WebMGA

web server for metagenomic analysis

Multiple metagenomics analyses



Download scripts and examples

1. Scripts: (a) client_submit_job.pl for single metagenomic analysis; (b) Rammcap_submit_job.pl for multiple metagenomic analyses.

2. Examples for single metagenomic analysis: for each analysis, users can download examples from our web site and modify it to be used for their own analysis.

3. Example for multiple metagenomic analyses: For complicated metagenomics analyses including several above-mentioned single analyses, we provide an example and users can modify it to be used for their own analysis.

```
./client_submit_job.pl input_fasta_file_name program_name output_name "[email]" "[parameter_set_1]" "[parameter_set_2]" [input_fasta_file_name2] [job_id]
```

```
./Rammcap_submit_job.pl input_fasta_file_name program_path
```

群落结构分析

- 识别16S rRNA (blastN，数据库:Ribosomal Database Project)

```
mol@mol-virtual-machine:~/abc$ ./blast/bin/blastn -query as_q.f
asta -db release10_32_unaligned -out as_q_16s -evalue -0.0001 -
task blastn -outfmt 0
```

注：MEGAN软件只能识别blast+中输出格式设为0、5、6的三中文件格式

D:\wangchenqu\slides\ABC\big homework\contig

```
BLASTN 2.2.28+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: /home/mol/abc/blast/16s/release10_32_unaligned.fa
           2,765,278 sequences; 2,818,065,686 total letters



Query= 4311582 length 250 cvg_1.0_tip_1

Length=250


***** No hits found *****
```

- 过滤No-hits 条目
- 导入MEGAN4（已自动导入了NCBI-taxonomy）

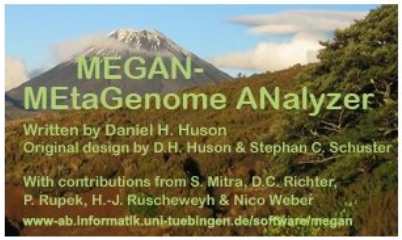
群落结构分析

MEGAN 4 - MEtaGenome ANalyzer

Software for analyzing metagenomes.

(Download here.)

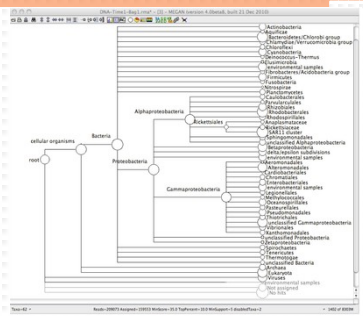
Over 7000 registered users.



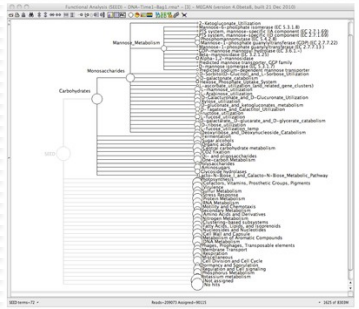
Program installers:

MEGAN_macos_4_70_4.dmg (64-bit, MacOS X)
MEGAN_windows_4_70_4.exe (32-bit, Windows-XP)
MEGAN_windows-x64_4_70_4.exe (64-bit, Windows 7)
MEGAN_unix_4_70_4.sh (64-bit, Linux, Unix)

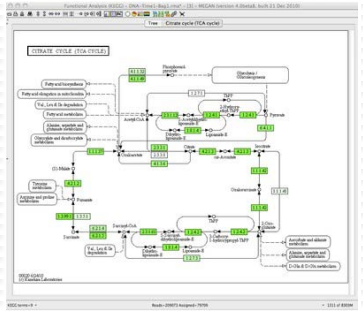
Taxonomic analysis



Functional analysis using the SEED classification

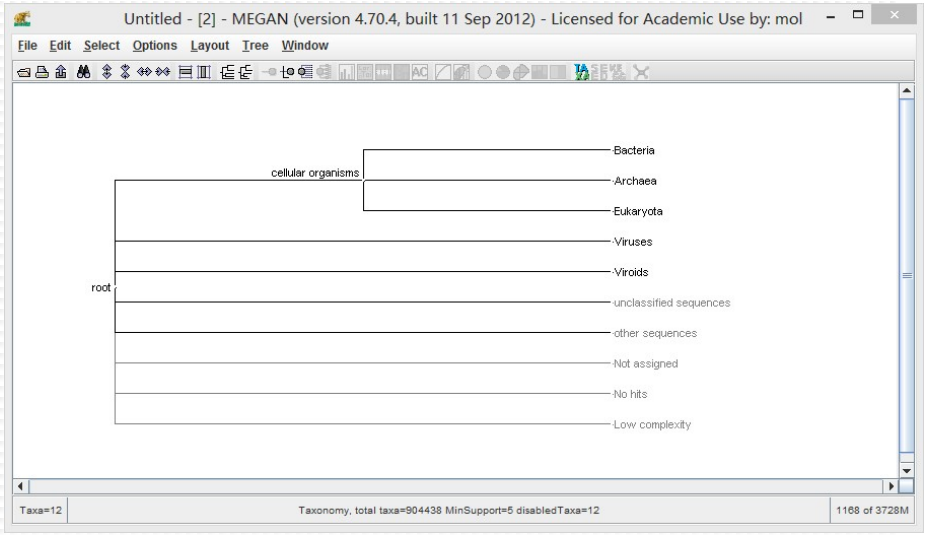


Functional analysis using the KEGG classification



Comparative visualization

群落结构分析


Untitled - [2] - MEGAN (version 4.70.4, built 11 Sep 2012) - Licensed for Academic Use by: mol
File Edit Select Options Layout Tree Window
Bacteria
cellular organisms
Archaea
Eukaryota
Viruses
Viroids
root
unclassified sequences
other sequences
Not assigned
No hits
Low complexity
Taxa=12
Taxonomy, total taxa=904438 MinSupport=5 disabledTaxa=12
1168 of 3728M

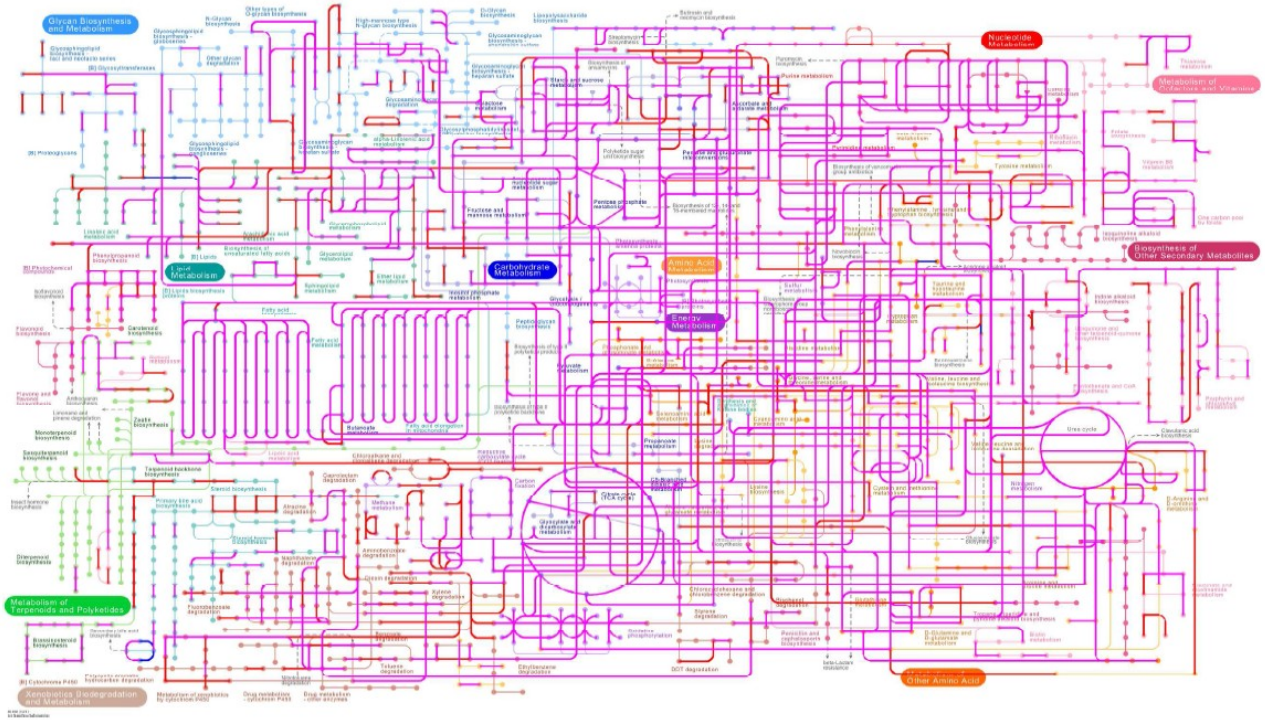

root
cellular organisms
Bacteria
Actinobacteria
Actinomycetales
Corynebacterineae
Gordonia
Mycobacterium
Tsukamurella
Frankia
Micrococcineae
Tetrasphaera
Clavibacter
Propionibacterineae
Streptomyces
environmental samples <Actinobacteria>
Bacteroidetes
Bacteroidales
Bacteroides
Parabacteroides
Cytophagaceae
environmental samples <Bacteroidetes>
Flavobacterium
Sphingobacteriales
Chitinophagaceae
environmental samples <Cytophagales>
Sphingobacterium
Verrucomicrobia
environmental samples <Green non-sulfur bacteria>
Cyanobacteria
Chroococcales
environmental samples <Cyanobacteria>
Oscillatoriales
environmental samples <Bacteria>
environmental samples <Acidobacteria>
Firmicutes
Bacilli
Bacillus
Lactobacillales
Lactobacillus
Streptococcus
Clostridiales
environmental samples <clostridial firmicutes>
Planctomycetales
Proteobacteria
Alphaproteobacteria
environmental samples <alpha subdivision>
Rhizobiales
environmental samples <Rhizobiaceae group>
Methylocystaceae
Rhodobacteraceae
Acetobacter
Sphingomonadaceae
Betaproteobacteria
Burkholderiales
Comamonadaceae
Aquabacterium
environmental samples <beta subdivision>
Dechloromonas
Kinetoplastibacterium
environmental samples <delta subdivision>
Gammaproteobacteria
Pectobacterium
environmental samples <gamma proteobacteria>
Pseudomonas
Francisella
Aliivibrio
Xanthomonadaceae
environmental samples <Xanthomonas group>
Xanthomonas
Spirochaetia
Candidatus Phytoplasma
unclassified Bacteria
candidate division JS1
candidate division OD1
Archaea
Crenarchaeota
environmental samples <Archaea>
Euryarchaeota
environmental samples <Euryarchaeota>
Methanomicrobia
environmental samples <Methanomicrobia>
Methanosarcina
unclassified Thermococcales
Thermoplasmata
environmental samples <Thermoplasmata>
environmental samples <Thermoplasmales>
Eukaryota
environmental samples <Eukaryotae>
Malawimonas
stramenopiles
unclassified sequences

代谢途径分析

- 对Contigs文件进行本地blastX，数据库为ncbi-nr

```
mol@mol-virtual-machine:~/abc$ ./blast/bin/blastx -query contig
_500 -db nr -out contig_500_bx -evalue -0.0001 -outfmt 0 -num_a
lignments 5 -num_descriptions 5
```

- blastX输出文件导入MEGAN，进行KEGG通路富集分析



常用软件汇总

	工具平台	数据库
测序	Illumina 454 pyrosequencing SOLID	
质控	fastx_toolkit Denoiser FastQC Perl ,Python	
群落结构分析	blast+ MEGAN4 Qiime ARB	RDP Silva grenngenes
序列拼接	SOAPdenovo Velvet Abyss	
ORF预测	WebMGA MetaGene Orphelia SoftBerry	
COG	MG-RAST CAMERA	SEED COG database
代谢途径分析	MEGAN4 WebMGA	KEGG

小结

1. 二代测序技术是分析微生物群落结构分析及功能研究的强有力工具

2. 掌握多项生物信息学分析软件可大大提高研究效率

3. 多学科交叉可为研究提供更多思路

致谢

感谢罗老师本学期的悉心教导，通过学习，我们掌握了生物信息学常用工具，对生物信息学应用有了更全面的认识。

感谢助教及班级同学在课程学习中给予的帮助。


THANK
YOU