松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

denovo 基因组

项目报告

合同编号	
客户单位	
报告时间	

适用范围

本项目分析报告适用于 denovo 基因组项目，不同样本数分析内容会略有差别。

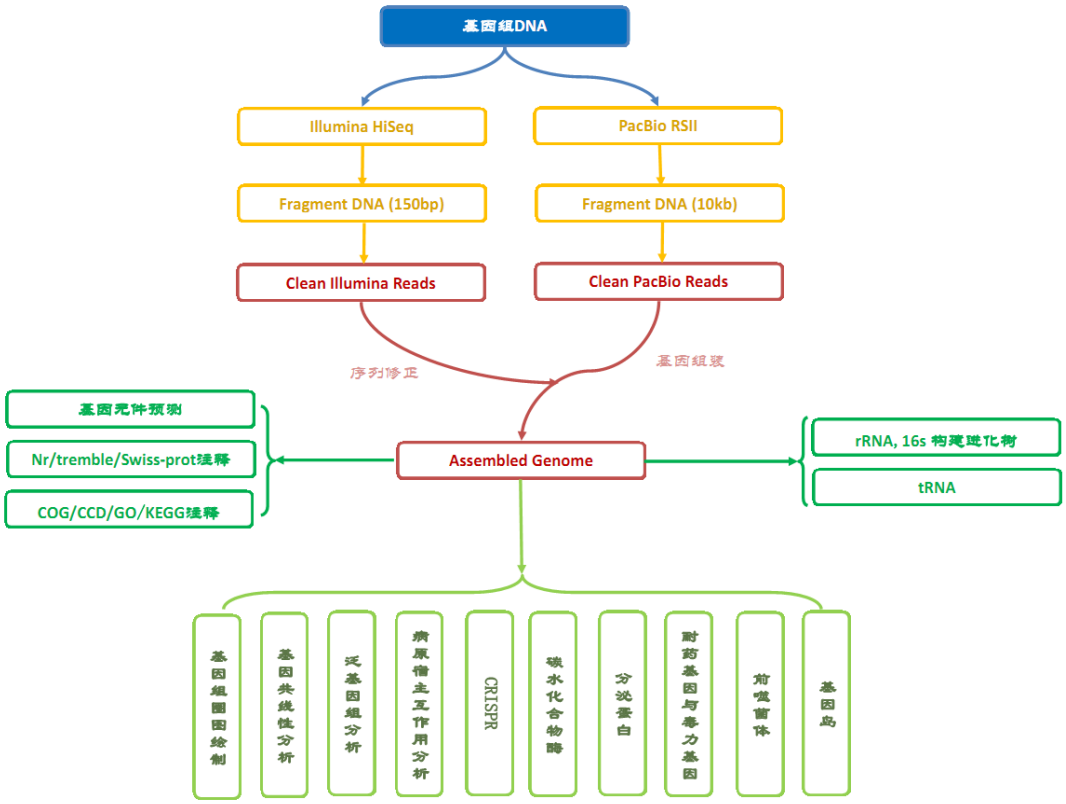
松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

目录

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

本项目利用 PacBio 公司的第三代单分子测序技术得到基因组序列，同时针对第三代测序技术错误率较高的缺点，使用 Illumina 公司的第二代数据提高序列质量，最终得到高质量的全基因组序列。



生物生物微生物基因组 Denovo 分析流程框架

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

1. 名词解释

Bp:base-pair，碱基对，读长的单位，每一个 bp 指一对互补的碱基。

Read：读长，测序数据中每一条序列就是一个 read。

Raw_reads:原始数据

Clean_reads：QC 之后的数据

Fastq: 序列数据存储的标准格式之一，每 4 行为一条 read 的信息。包含测序 read 名，序列，正反链标示，序列质量值

Pair-end 测序：双端测序，两端均测序，随后合并成一条 read。

Single-end 测序：单端测序，只测一端，即为一条 read。

质量评分：指的是一个碱基的错误概率的对数值，即质量评分越高，错误概率越小。

QC： Quality control，即质量控制。

滑窗法：检测一个窗口内的碱基质量值，如果满足条件则向前移动一个单位继续检测，如果不满足条件即做删除处理，随后继续移动到下一个单位进行检测，直到检测完所有的数据。

Denovo Assembly：基因组重头组装

基因组修正：指对初步组装成功的基因组进行的修正和优化。包括碱基修正，Gap 修补等。

Contig：拼接软件基于 reads 之间的 overlap 区，拼接获得的序列称为 Contig（重叠群）。三代数据由于长读长的优势，可以方便地得到较长的 Contig。原核生物基因组组装时，三代测序往往可直接得到染色体水平的 Contig，而二代数据会得到多条 Contig，每条 Contig 只是基因组的一部分，它们需要通过其他手段进一步组装才能拼成完整的基因组。

GC 含量：基因组中 G 碱基与 C 碱基的含量。不同物种基因组中 GC 含量可能差别较大。

ORF：Open Reading Frame 开放阅读框。

CDS：Sequence coding for aminoacids in protein 蛋白质编码区，是编码一段蛋白产物的序列。CDS 必定是一个 ORF，但也可能包括很多 ORF。反之，每个 ORF 不一定都是 CDS。原核基因组中没有真核基因中的外显子内含子之分，原核基因组 CDS 区段往往占整个基因组比例的 80%以上，多数情况下，原核生物中一个基因就会对应一个 CDS。

基因组注释：Genome annotation 是利用生物信息学方法和工具，对基因组所有基因的生物学功能进行高通量注释。基因组注释各个条目的解析详见对应章节，这里不再重复展开说明了。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2. 分析结果展示

2.1 测序质量评估及质控

2.1.1 三代 PacBio 单分子测序质量评估与控制

该项目利用高质量的 DNA，构建 2 个插入片段为 10kb 的 PacBio RS II 文库，利用第三代测序仪 Pacbio RS II 对 DNA 进行非扩增长片段测序。

测序数据的产生是经过了 DNA 提取、建库、测序多个步骤的，这些步骤中产生的无效数据会对后续信息分析带来严重干扰，如测序接头序列，建库长度的偏差，以及测序错误、低质量碱基、未测出的碱基（以 N 表示）等情况，须通过一些手段将上述无效数据过滤掉，以保证分析的正常进行。

过滤标准主要有：

1) 过滤掉含有接头序列的 reads；

2) 截掉 reads 部分区域的低质量碱基；

3）去除 adapter 序列

4) 根据预估基因组大小（3M），从最长的 Reads 开始依次将较短的 Reads 纳入待分析数据池，直到数据量达到预估基因组大小的 40x。

最终，用于分析的三代数据基本信息如下：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

Reads_Count:	13973
Base_Count:	116961580
Max_len:	29106
Average_len:	8370.54
N50:	9442
Min_len	1001

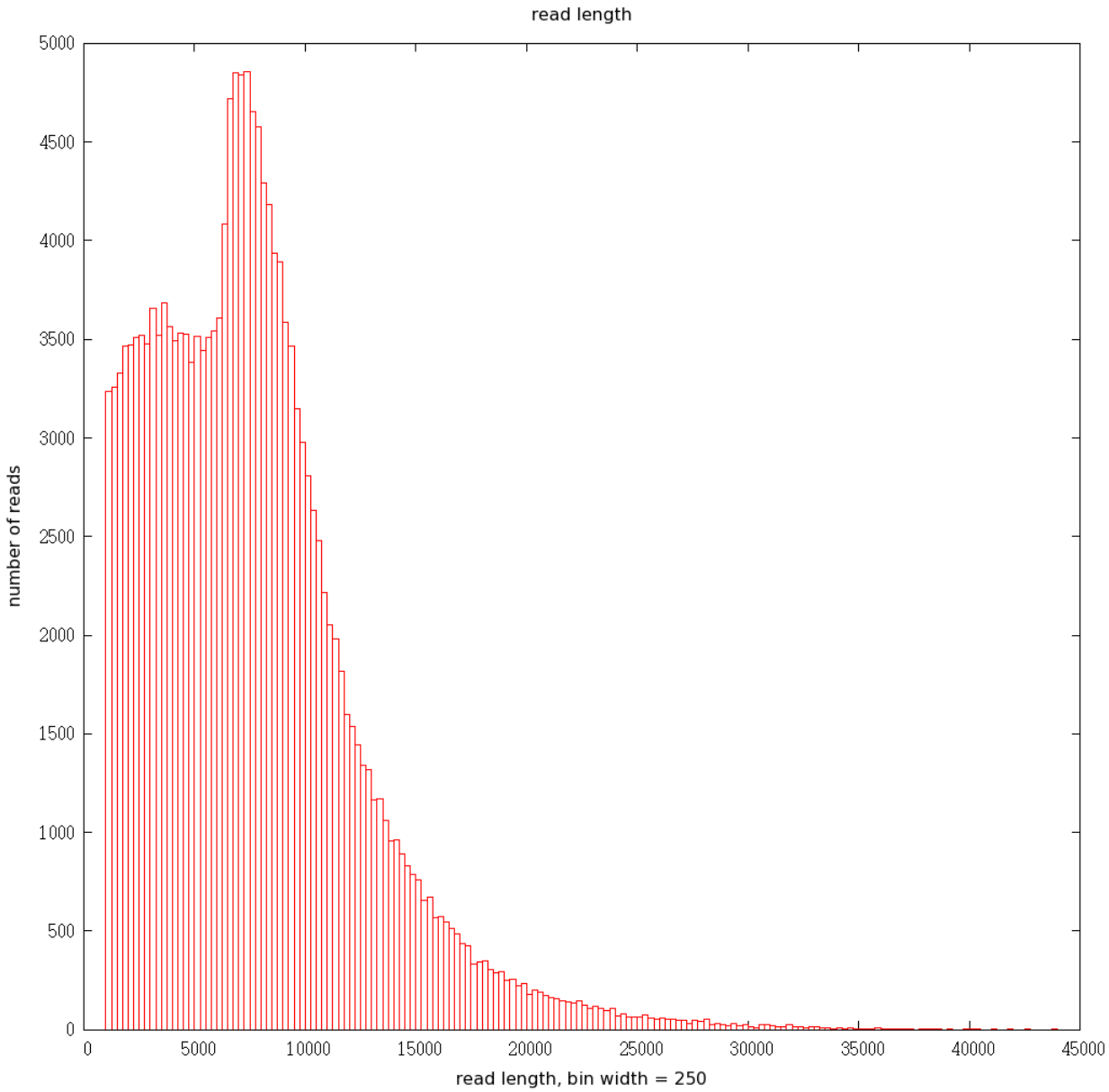


图 2.0 过滤后 PacBio 测序 Reads 长度分布统计

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.1.2 二代 Illumina Miseq 测序质量评估与控制

本次测序采用 Miseq PE300 模式（双端测序 PE：paired-end），每一个样本分别有 R1.fastq 和 R2.fastq 两个文件，分别代表 5' -> 3'和 3'->5'的测序结果。R1.fastq 与 R2.fastq 中的文件行数是一致的，且根据 reads name 一一对应。

FASTQ: Fastq 是 Solexa 测序技术中一种反映测序序列的碱基质量的文件格式。每条 read 包含 4 行信息。第一行以"@"开头，随后是序列标示和相关的描述信息，第三行以"+" 开头，随后是序列描述信息或者什么都不加；），第二行为碱基序列，第四行是质量信息，与第二行中的碱基序列一一对应，根据评分体系不同每个字符的含义所表示的数字有所差别。例如：

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

质量评分：质量评分指的是一个碱基的错误概率的对数值。其最初在 Phred 拼接软件中定义与使用，其后在许多软件中得到使用。其质量得分与错误概率的对应关系见下表：

Phred quality scores are logarithmically linked to error probabilities		
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %

对于每个碱基的质量编码标示，不同的软件采用不同的方案，本项目中使用的方案是，Phred quality score，值的范围从 0 到 62 对应的 ASCII 码从 64 到 126，得分在 0 到 40 之间；

软件：FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/（用于统计序列原始信息及绘图），

结果目录：1_QC/

All_sample_QC_infor.xls：所有样本原始数据统计，结果如下：

Table 2.1 原始数据统计

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

	R1	R2	ALL
Total Reads Count(#):	3290272	3290272	6580544
Total Bases Count(bp):	990371872	990371872	1980743744
Average Read Length(bp):	301	301	301
Q30 Bases Count(bp):	866910043	697318224	1564228267
Q30 Bases Ratio(%):	87.53	70.41	78.9717636
Q20 Bases Count(bp):	937298275	823482062	1760780337
Q20 Bases Ratio(%):	94.64	83.15	88.89490841
Q10 Bases Count(bp):	973261180	924124640	1897385820
Q10 Bases Ratio(%):	98.27	93.31	95.79158464
N Bases Count(bp):	6	14	20
N Bases Ratio(%):	0	0	0
GC Bases Count(bp):	438654767	440953892	879608659
GC Bases Ratio(%):	44.29	44.52	44.40799885

注：若样本数目较多，此处只会截取部分样本数据，完整数据请见结果文件夹中的对应文件。

Total Reads Count：样本所有 reads 数目，为 reads1 与 reads2 数目之和

Total Base Count：所有碱基数目，即数据量

Average Read Length：平均序列长度

Q30 Base Count：碱基质量在 30 以上的数目

Q30 Base Ratio：Q30 碱基比例

N Base Count：N 碱基的数目

N Base Ratio：N 碱基比例

GC Base Count：GC 碱基数目

GC Base Ratio：GC 含量

各样本碱基质量图如下：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

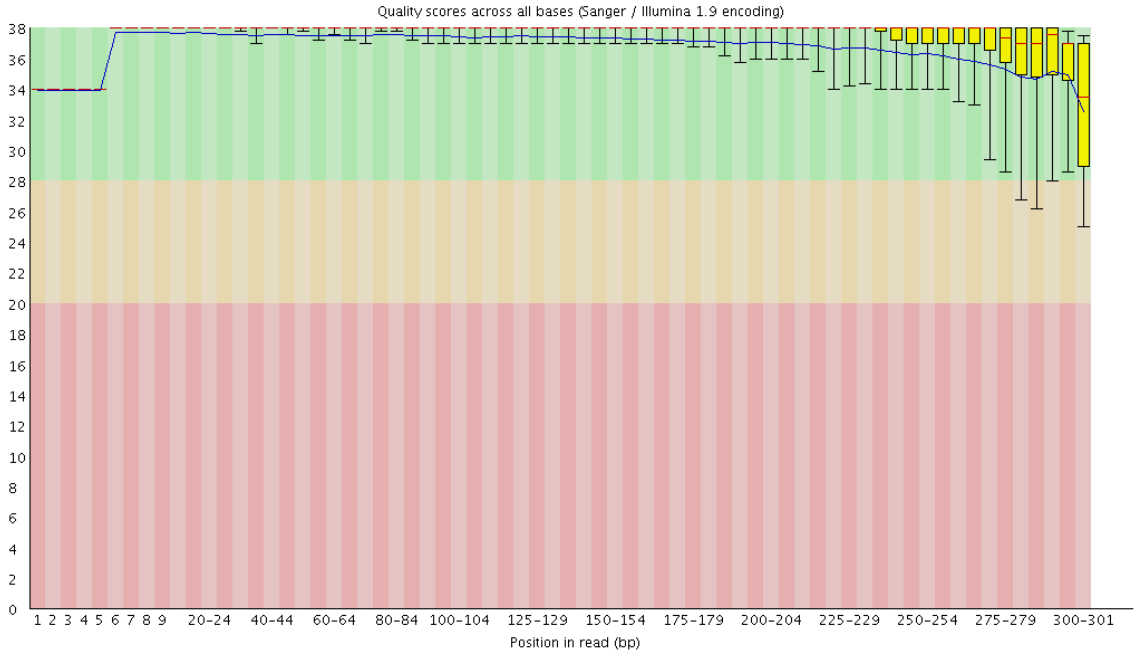


图 2.1 各位置碱基质量分布图

注：若样本数据较多，此处只展示某个样本的 Read1 质量分布，其它样本数据见 1_QC/Sample/*fastqc.zip 文件。

说明：横坐标表示测序位置，纵坐标为测序质量值图中，横轴代表位置，纵轴 quality。红色表示中位数，黄色是 25%-75%区间，触须是 10%-90%区间，蓝线是平均数。Hiseq 测序是双端测序，每条 read 长度 125bp。随着测序的进行，酶的活性会逐步下降，因此到达一定测序长度后，碱基质量值也会随之下降。从图 6.1 可知，中位值均在 Q20 以上，因此该文库碱基质量良好，可用于后续分析。本分析会对所有数据进行质控，后续只取 Q20 以上的数据进行分析。

各样本碱基 GC 含量分布图如下：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

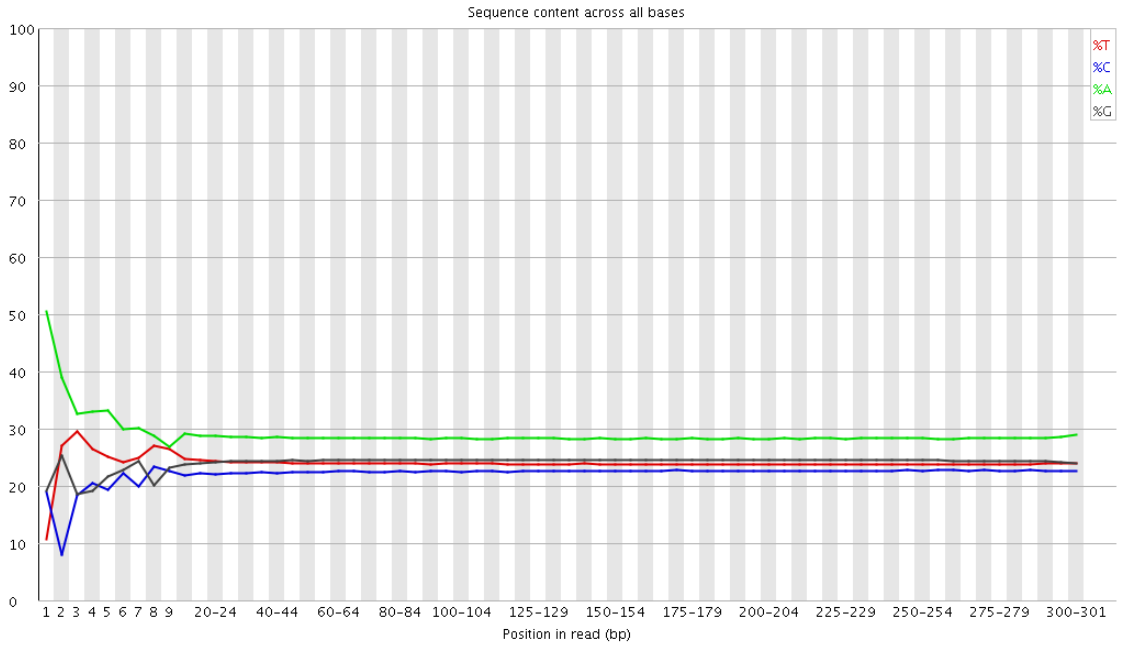


图 2.2 各位置碱基 GC 含量分布图

注：若样本数据较多，此处只展示某个样本的 Read1 质量分布，其它样本数据见 1_QC/Sample/*fastqc.zip 文件。

说明：横坐标是 reads 碱基坐标，纵坐标是所有 reads 的 A、C、G、T 碱基分别占的百分比。在文库较均匀随机的情况下，四种颜色的分界线应该波动极小，呈一条直线，但一般测序前几个碱基由于测序尚不大稳定，前几个碱基 ACGC 含量会有波动。

对于 Hiseq 双端测序原始序列 3’端可能带有 adaptor 接头序列，以及一些少量低质量序列和杂质序列，为了提高后续分析质量和可靠性，对原始序列进行去接头、质量剪切、污染评估等处理。

数据质控步骤：

1）去除 3’端测序接头，采用的软件为 cutadapt，Read1 3’端测序接头为 AGATCGGAAGAGCACACGTCTGAAC，Read2 3’端测序街头为 AGATCGGAAGAGCGTCGTGTAGGGA。

2）去除融合后的 reads 尾部质量值在 20 以下的碱基。设置 10bp 的端口，如果窗口内的平均质量值低于 20，从窗口开始去除后端的碱基

3）切除 reads 中含 N 部分序列：长度阈值 35bp

4）对序列进行污染评估，看其是否有污染，方法为：随机从 QC 之后序列中抽取 10000 条序列进行 blast 比对，比对数据库为 NCBI NT 数据库，取 evalue <=1e-10 并且相似度>90%,coverage >80%的比对结果，计算其物种分布。

去除测序接头软件：cutadapt（https://pypi.python.org/pypi/cutadapt/1.2.1）

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

主要参数设置：-O 10 -min_len 35 -a AGATCGGAAGAGCACACGTCTGAAC

质量控制使用软件：Prinseq（http://prinseq.sourceforge.net/）

主要参数设置：-trim_qual_left 20 -trim_qual_right 20 -trim_qual_window 10 -trim_qual_step 1 -min_len 35

污染评估软件：blast+

（http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download）

主要参数设置：-evalue 1e-10 -num_threads 40

结果目录：1_data_for_analysis/

All_sample_QC_infor.xls：所有样本 QC 之后结果统计，详细结果如下：

表 2.2 QC 之后结果统计

Raw_sequences	6580544
Raw_bases	1980743700
Raw_mean_length	301
Good_sequences	6518526
Good_ratio	99.06
Good_bases	1757120344
Good_mean_length	269.56

注：若样本数目较多，此处只会截取部分样本数据，完整数据请见结果文件夹中的对应文件。

Raw_sequences：原始序列数目，为 Read1 与 Read2 数目之和

Raw_bases：原始序列碱基数目

Raw_mean_length：原始数据序列平均长度

Good_sequences：QC 之后剩余的序列数目

Good_ratio：QC 之后剩余序列数目比例

Good_bases：剩余序列总碱基数目

Good_mean_length：QC 之后序列平均长度

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.2 基因组拼接

2.2.1 方法说明

三代测序拥有非常长的读长，可以方便地完成基因组拼接。但是，三代数据本身的错误率较高，于是当三代数据拼装成基因组框架后，引入二代测序数据对基因组进行修正，从而得到高质量的全基因组序列。

序列拼接采用的工具为 PacBio 公司用于三代数据拼装的专用流程 canu，之后引入二代测序数据，用 Gapcloser 及 GapFiller 对 scaffold 补 Gap，最后采用 PrInSeS-G 进行序列校正。

拼接软件：canu。该工具为 PacBio 公司专门为三代单分子测序基因组装开发的专用工具，为经典拼装工具 Celera Assembler 的升级版，预估基因组大小为 6M，组装参数为工具默认参数。

错误校正软件：PrInSeS-G，该软件用于测序错误的校正，可以很好的修正拼接过程中的碱基错误及小片段的插入缺失。

2.2.2 结果展示

结果目录：2_assembly/

Assembled_Geonme.fa：拼接后基因组序列文件

Assembled_Genome_Info.xls：拼接结果统计。

拼接最终得到 5 条 contig，由长到短：

表 2.3 拼接得到的 Contig 结果

ID	Base_number	GC_ratio
tig00000000	3185053	0.45
tig00000001	39587	0.42
tig00000002	14560	0.38

拼接结果得到了一条大 Contig 和两条小 Contigs。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.3 基因预测

2.3.1 方法说明

基因预测一般指预测 DNA 序列中编码蛋白质的部分。其方法主要有两大类：一类是基于相似性的预测方法，即利用已知的 mRNA 或蛋白质序列为线索在 DNA 序列中搜寻所对应的片段，达到基因预测的目的；另一类是基于统计学模型的从头预测方法，这种方法可不依赖已知的 DNA 序列进行，即利用统计学模型训练出相应参数，再对基因进行预测。

我们采用 Prokka 软件来对各样本的组装结果进行基因预测。Prokka 是一系列基因元件预测工具的集合，它调用 Prodigal 预测编码基因，Aragorn 预测 tRNA，barrnap 预测 rRNA，Infernal 预测 miscRNA，预测出的各类基因元件最终会加以汇总并完成初步注释。预测结果存放在目录 3_gene_prediction 内。

2.3.2 结果展示

结果目录：3_gene_predict/

gene_result.xls：所有样本基因预测结果统计，结果见下表：

表 2.5 基因预测结果统计

All_num	>=500bp	>=1000bp	N50	Max_len	Min_len	All_len	Mean_len
2830	2130	1072	1203	9291	74	2705722	956.09

Sample：样本名

All_num：预测到的基因数目

>=500bp：长度大于 500bp 的基因数目

>=1000bp：长度大于 1000bp 的基因数目

N50：N50 长度

Max_len：长度最长的基因长度

Min_len：最小的基因长度

All_len：所有基因总长度

Mean_len：基因平均长度

gene_prediction_stast.xls：基因预测统计表，展示如下：

表 2.6 基因预测总体统计

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

Class	Number
Size(base)	3239219
G+C content (%)	44
Protein-coding genes	2830
Min length (base)	74
Max length (base)	9291
Average length (base)	956.09
Total coding gene (base)	2705722
Coding ratio(%)	83.53007
tRNA	48
rRNA	12

***/Gene_Len_Dis.pdf**：单样本基因长度分布图，结果如下图：

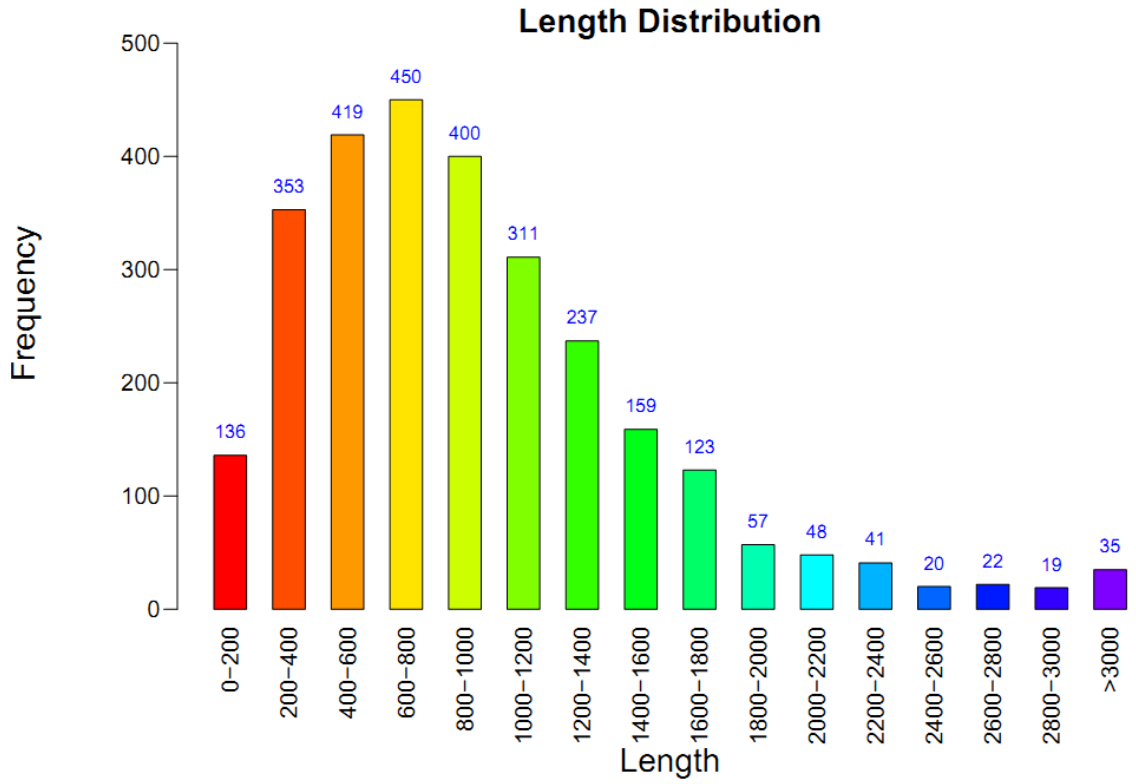


图 2.4 单样本基因长度分布图

***/Gene_GC_content.pdf**：基因 GC 含量分布图，结果展示如下：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

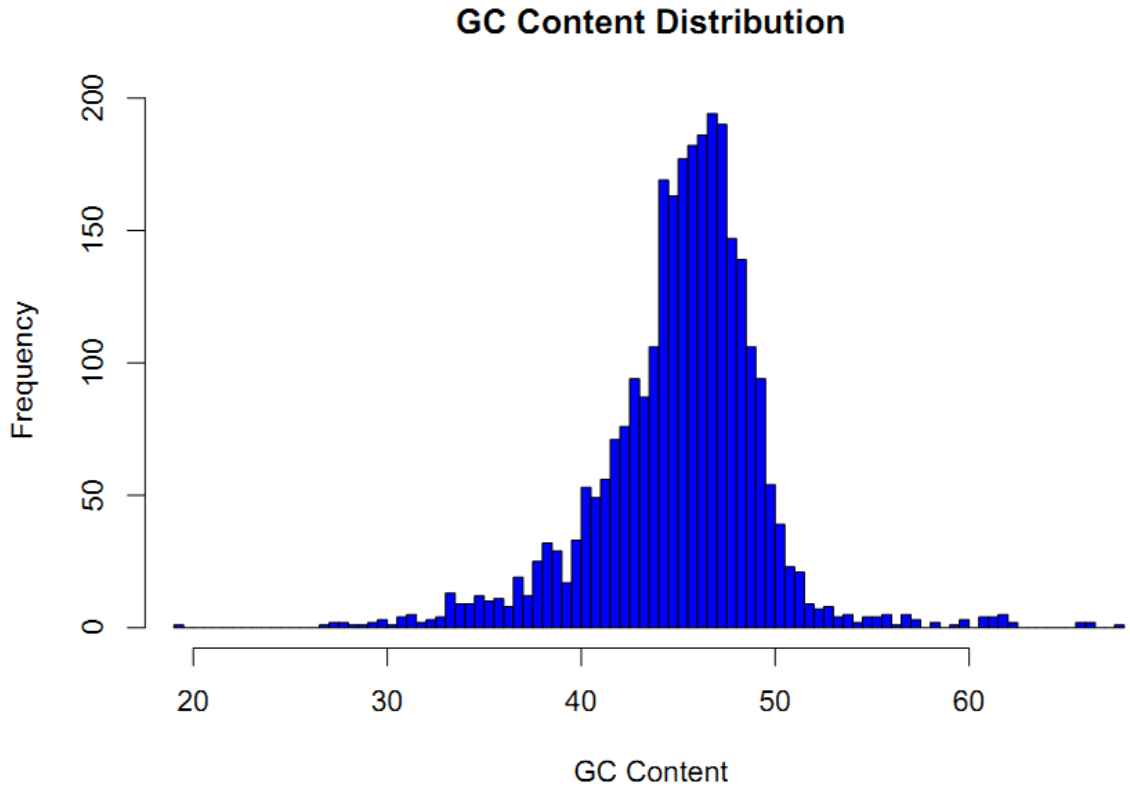


图 2.5 单样本基因 GC 含量分布图

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.4 比较基因组

2.4.1 根据 16s 序列进行菌种鉴定和系统发育树构建

16S rRNA 基因是细菌上编码 rRNA 相对应的 DNA 序列，存在于所有细菌的基因组中。16S rRNA 具有高度的保守性和特异性以及该基因序列足够长（包含约 50 个功能域）。随着测序技术的出现及核酸研究技术的不断完善，16S rRNA 基因检测技术已成为病原菌检测和鉴定的一种强有力工具。

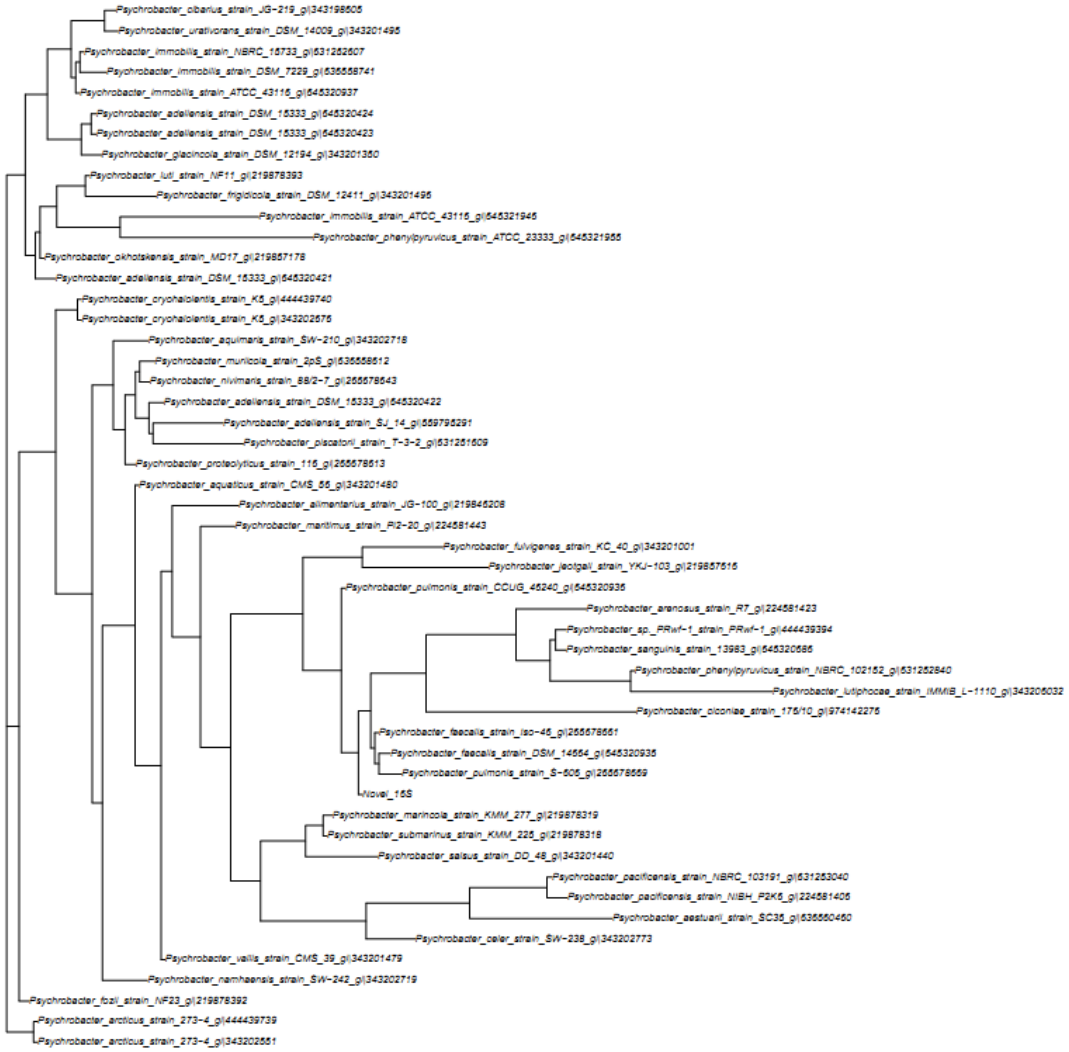
根据基因预测得到的 16sRNA 序列，将它与 NCBI 的 16s 数据库进行 blastn 比对，取比对的 identify 阈值为 95，其余参数默认，得到了 78 条与之高度相近的 rRNA。之后，利用 muscle 软件进行序列多重比对，用 FastTree 软件计算预测得到的 16s 与数据库 78 条序列的两两间遗传距离，并构建系统发育树：

表 2.7 新预测得到的 16s 与相近 16s 序列的遗传距离（前 10）

Genus	Species	Strain	GI	Distance
Psychrobacter	faecalis	Iso-46	265678661	0.002009
Psychrobacter	pulmonis	CCUG 46240	645320936	0.002045
Psychrobacter	faecalis	DSM 14664	645320935	0.003418
Psychrobacter	pulmonis	S-606	265678669	0.004422
Psychrobacter	maritimus	Pi2-20	224581443	0.019227
Psychrobacter	urativorans	DSM 14009	343201495	0.019236
Psychrobacter	cibarius	JG-219	343198605	0.01968
Psychrobacter	alimentarius	JG-100	219846208	0.01968
Psychrobacter	luti	NF11	219878393	0.021099
Psychrobacter	frigidicola	DSM 12411	343201496	0.021269

具体的近缘菌株列表，详见/report/3_gene_predict/16s/CloseStrains.xls

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司



松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

图 2.7 根据 16s 序列构建的系统发育树（novel 代表本项目组装出的基因组）

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.5 蛋白基本注释

2.5.1 各数据库比对

将样本物种的蛋白序列与公共数据 gene 进行比较，通过 gene 的相似性进行功能注释。基因相似性比对主要基于 BLAST 算法。BLAST，全称 Basic Local Alignment Search Tool，即"基于局部比对算法的搜索工具"，由 Altschul 等人于 1990 年发布。Blast 能够实现比较两段核酸或者蛋白序列之间的相似性的功能，它能够快速的找到两段序列之间的相似序列并对比对区域进行打分以确定相似性的高低。将蛋白序列分别与 KOG、Swissprot、TrEMBL、GO、KEGG 库进行比对，取相似度>30%，且 e<1e-5 的注释，合并基因得到的所有注释详细信息。

各数据库说明如下：

KOG/COG: COG 是 Clusters of Orthologous Groups of proteins 的简称，KOG 为 euKaryotic Ortholog Groups。这两个注释系统都是 NCBI 的基于基因直系同源关系，其中 COG 针对原核生物，KOG 针对真核生物。COG/KOG 结合进化关系将来自不同物种的同源基因分为不同的 Ortholog 簇，目前 COG 有 4873 个分类，KOG 有 4852 个分类。来自同一 ortholog 的基因具有相同的功能，这样就可以将功能注释直接继承给同一 COG/KOG 簇的其他成员。详见 http://www.ncbi.nlm.nih.gov/COG/。

Swiss-Prot（A manually annotated and reviewed protein sequence database）搜集了经过有经验的生物学家整理及研究的蛋白序列。详见 http://www.ebi.ac.uk/uniprot/。

KEGG 是 Kyoto Encyclopedia of Genes and Genomes 的简称，是系统分析基因产物和化合物在细胞中的代谢途径以及这些基因产物的功能的数据库。它整合了基因组、化学分子和生化系统等方面的数据,包括代谢通路（KEGG PATHWAY）、药物（KEGG DRUG）、疾病（KEGG DISEASE）、功能模型（KEGG MODULE）、基因序列（KEGG GENES）及基因组（KEGG GENOME）等等。KO（KEGG ORTHOLOG）系统将各个 KEGG 注释系统联系在一起，KEGG 已建立了一套完整 KO 注释的系统，可完成新测序物种的基因组或转录组的功能注释。详见 http://www.genome.jp/kegg/。

GO(Gene Ontology)是一套国际标准化的基因功能描述的分类系统。GO 分为三大类 ontology：生物过程（Biological Process）、分子功能（Molecular Function）和细胞组分(Cellular Component)，分别用来描述基因编码的产物所参与的生物过程、所具有的分子功能及所处的细胞环境。GO 的基本单元是 term，每个 term 有一个唯一的标示符（由“GO:”加上 7 个数字组成，例如 GO:0072669）；每类 ontology 的 term 通过它们之间的联系（is_a, part_of, regulate）构成一个有向无环的拓扑结构。详见 http://www.geneontology.org/。

各数据库及功能注释所用到的软件及方法：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

SwissProt、TrEMBLe 序列数据库的比对：NCBI blast 2.2.28+，blastx;

COG/KOG：NCBI blast2.2.28+，rpsblast;

GO 功能注释：基于 Swissprot 和 TrEMBL 两部分的蛋白注释结果及 GO 数据库通过自写脚本获取 GO 注释信息;

KEGG 相关注释：KAAS，KEGG Automatic Annotation Server。

结果目录：4_Annotation/

Annotation_statistics.xls：各数据库注释比例统计，结果如下：

表 2.8 各数据注释比例统计

Database	Number of Unigenes	Percentage(%)
Annotated in CDD	2389	84.42
Annotated in COG	2202	77.81
Annotated in NR	2733	96.57
Annotated in NT	1779	62.86
Annotated in PFAM	2323	82.08
Annotated in Swissprot	1971	69.65
Annotated in TrEMBL	2709	95.72
Annotated in GO	2063	72.9
Annotated in KEGG	1511	53.39
Annotated in at least one database	2782	98.3
Annotated in all database	1074	37.95
Total Unigenes	2830	100

Annotated in COG: COG 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in Swissprot: Swissprot 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in TrEMBLe: TrEMBLe 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in GO: GO 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in KEGG: KO 注释成功的蛋白数目及其占总蛋白数的比例

Annotated in at least one Database: 在以上 8 个数据库中至少 1 个数据库注释成功的蛋白数目及其

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

占总蛋白数的比例

Annotated in all Databases: 在以上 7 个数据库中都注释成功的蛋白数目及其占总蛋白数的比例

Total Unigenes: 总的蛋白条数,占总蛋白比例为 100%

Annotation_ratio.pdf：各数据库注释上的基因比例折线图，展示如下图：

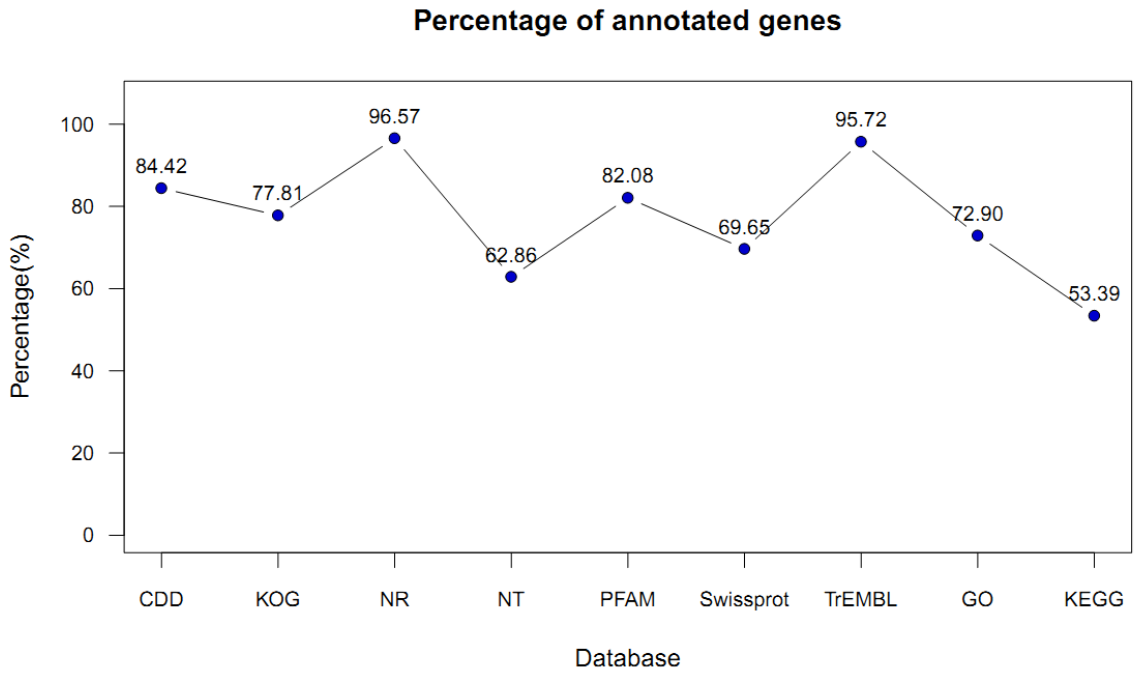


图 2.8 各数据库比例注释折线图

Venn_diagram_for_annotation.pdf：各数据库注释上的基因韦恩图，默认为绘制 NR、KEGG、Swissprot、KOG/COG 之间的韦恩图，展示如下图：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

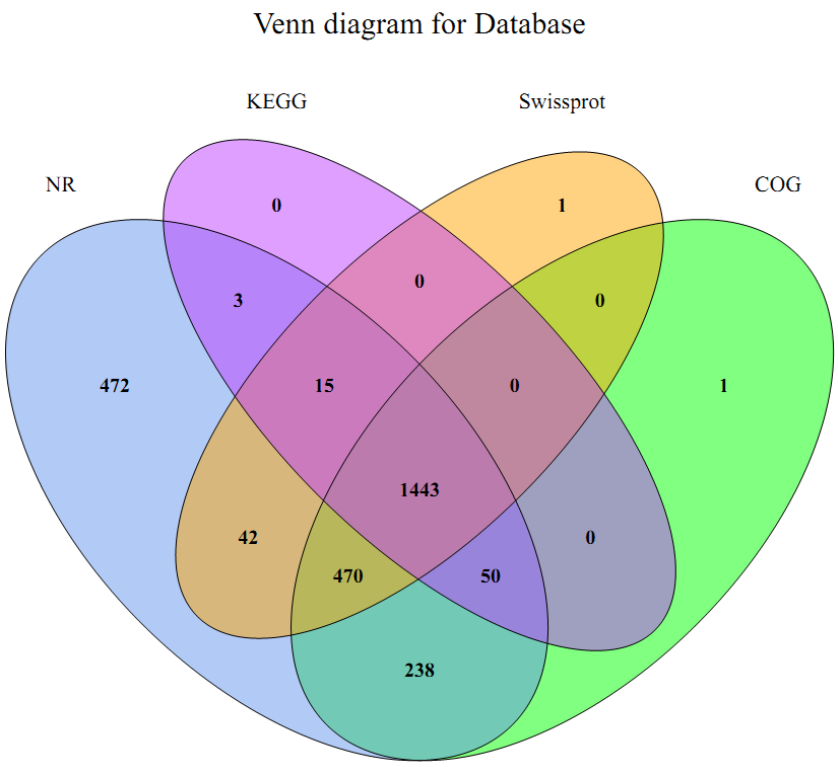


图 2.9 各数据库注释韦恩图

Unigene_annotation.xls：所有蛋白注释结果汇总表，表头注释如下：

Gene_id：基因 id

Length：基因长度

Sequence：基因序列

CDS_seq：编码区序列

Start：基因在基因组上开始的位置

End：基因在基因组上的终止位置

Strand：基因所在的正负链

Protein_sequence：蛋白序列

Protein_id：对应的蛋白序列 ID

Gene_name：基因名字

Gene_description：基因描述

Product：对应的蛋白产物

Protein_seq：蛋白序列

CDD：NCBI 保守序列库注释结果

COG：对应的 COG 注释结果

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

NR：NR 库注释结果

NT：NT 库注释结果

PFAM：Pfam 注释结果

Swissprot：Swissprot 注释结果

TrEMBL：TrEMBLe 注释结果

GO：GO 注释结果

KEGG：KEGG 注释结果

2.5.2 NR 库注释

将 Unigene 比对到 NCBI 的蛋白非冗余（non-redundant，NR）数据库中，得到的注释结果可给出基因基本功能，以及近缘物种等信息。

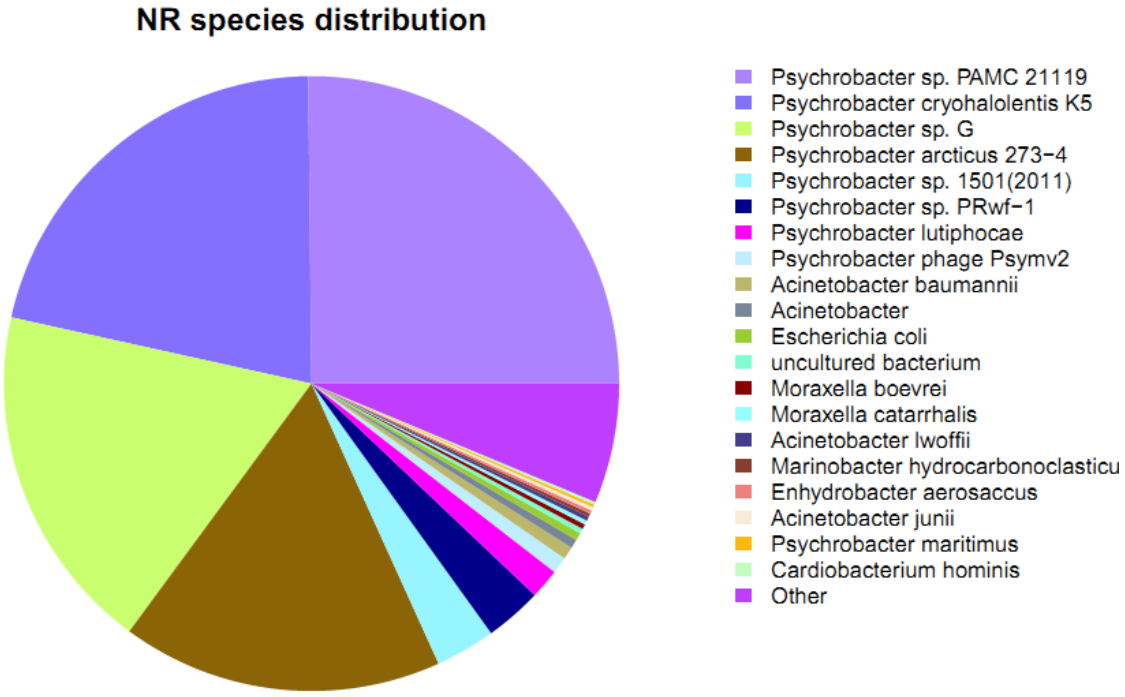


图 2.10 NR 库比对得到的序列物种信息饼图

2.5.3COG、KOG 注释

将 Unigene 序列比对到 COG/KOG 数据库中，原核物种比对 COG 数据库，真核物种比对到 KOG 数据库，基于比对结果计算各功能类下面注释上的基因数目。

COG/KOG 注释结果，为许多更深入注释分析的基础。

结果目录：4_Annotation/COG/

COG_code_count.xls/COG_code_count.xls：各功能类基因数目统计列表，结果如下表：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

表 **2.9**COG/KOG 各功能类基因数统计表

Code	Name	Gene_num	Gene_ratio
A	RNA processing and modification	1	0.05
C	Energy production and conversion	165	7.49
D	Cell cycle control, cell division, chromosome partitioning	27	1.23
E	Amino acid transport and metabolism	159	7.22
F	Nucleotide transport and metabolism	53	2.41
G	Carbohydrate transport and metabolism	48	2.18
H	Coenzyme transport and metabolism	111	5.04
I	Lipid transport and metabolism	75	3.41
J	Translation, ribosomal structure and biogenesis	165	7.49
K	Transcription	109	4.95
L	Replication, recombination and repair	235	10.67
M	Cell wall/membrane/envelope biogenesis	119	5.4
O	Posttranslational modification, protein turnover, chaperones	96	4.36
P	Inorganic ion transport and metabolism	130	5.9
Q	Secondary metabolites biosynthesis, transport and catabolism	29	1.32
R	General function prediction only	296	13.44
S	Function unknown	224	10.17
T	Signal transduction mechanisms	66	3
U	Intracellular trafficking, secretion, and vesicular transport	63	2.86
V	Defense mechanisms	31	1.41

KOG_Categories.pdf/COG_Categories.pdf：COG /KOG 注释功能类条形图，绘图源文件为 KOG_code_count.xls/COG_code_count.xls，结果展示如下：

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

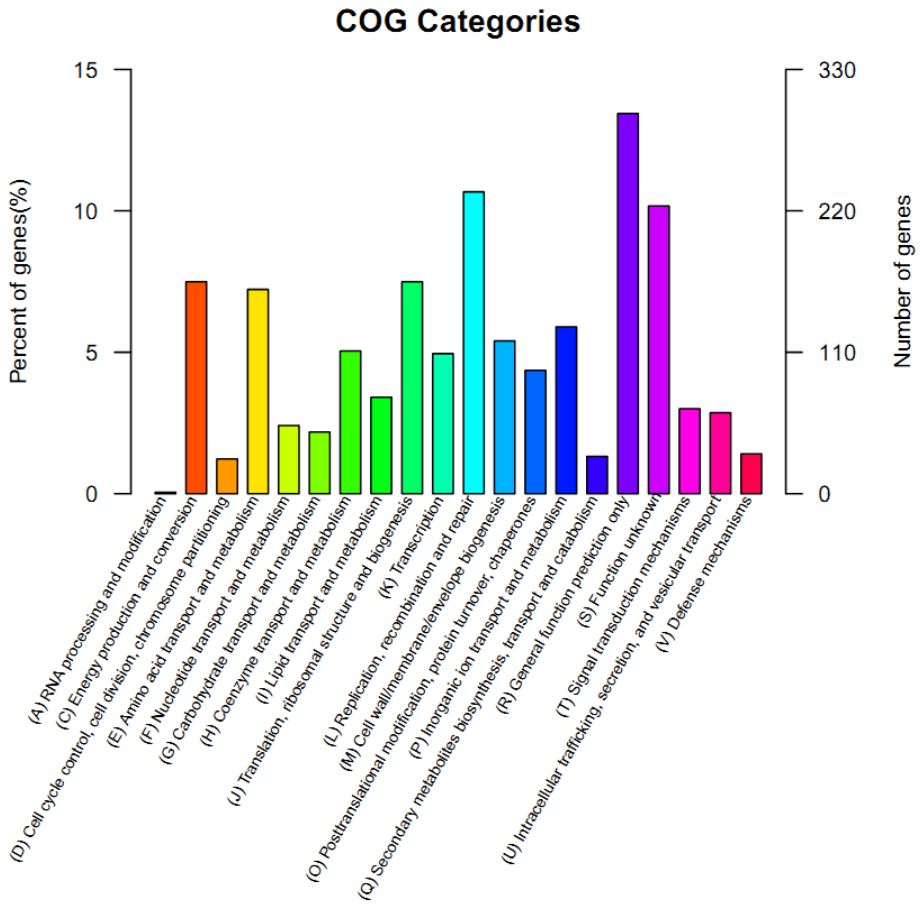


图 2.11 COG 注释条形图

注：各字母意义：

[S] Function unknown

[Z] Cytoskeleton

[Y] Nuclear structure

[W] Extracellular structures

[V] Defense mechanisms

[U] Intracellular trafficking, secretion, and vesicular transport

[T] Signal transduction mechanisms

[R] General function prediction only

[Q] Secondary metabolites biosynthesis, transport and catabolism

[P] Inorganic ion transport and metabolism

[O] Posttranslational modification, protein turnover, chaperones

[N] Cell motility

[M] Cell wall/membrane/envelope biogenesis

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

[L] Replication, recombination and repair

[K] Transcription

[J] Translation, ribosomal structure and biogenesis

[I] Lipid transport and metabolism

[H] Coenzyme transport and metabolism

[G] Carbohydrate transport and metabolism

[F] Nucleotide transport and metabolism

[E] Amino acid transport and metabolism

[D] Cell cycle control, cell division, chromosome partitioning

[C] Energy production and conversion

[B] Chromatin structure and dynamics

[A] RNA processing and modification

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.5.4 GO 注释

对得到的基因进行 GO 分类，统计基因在 Biological Process, Cellular Component, Molecular Function 三个类别的各 GO term。此分析是基于 blastuniprot 的结果(即合并与 swissprot 和 trembl 的结果)，利用得到的 uniprot 号比对 GO term。

所用软件：自写程序

结果目录：4_Annotation/GO/

GO_classification_level2.pdf：GO 分类在 level2 水平上基因分别条形图，结果如下图：

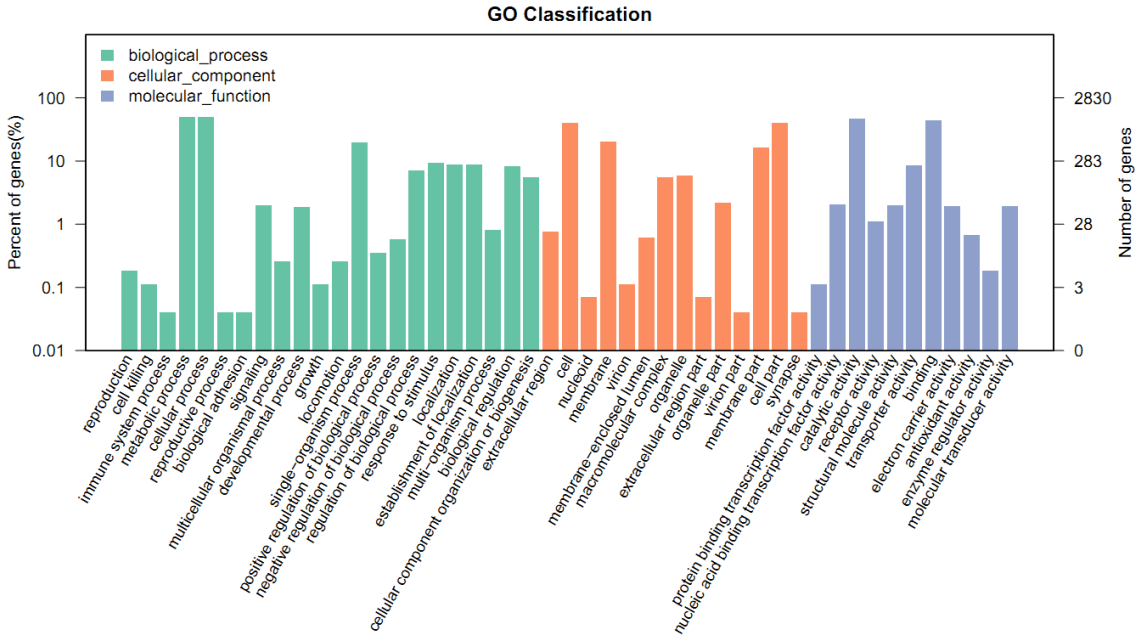


图 2.12 level2 水平 GO 注释上的基因分布

说明：横坐标表示 Level2 水平上 GO term，不同颜色代表不同 GO 类，分为三大类，纵坐标表示注释到该 term 上面的 Unigene 数目及比率，纵坐标采用对数坐标。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.5.5 KEGG 注释

对得到的基因进行 KEGG Pathway 分析，利用 KAAS 预测得到对应的 KO 号，然后利用 KO 号对应到 KEGG pathway 上，分析基因与 KEGG 中酶注释的关系文件以及映射到 pathway 的信息。

结果目录：4_Annotation/KEGG/

kegg_annot2.xls：各 pathway 注释上的 Unigene 数目，结果如下：

表 2.10 各 pathway 注释上的 Unigene 数目

Pathway_ID	Pathway_name	Group	Gene_num
ko00471	D-Glutamine and D-glutamate metabolism	Metabolism	3
ko00564	Glycerophospholipid metabolism	Metabolism	17
ko00030	Pentose phosphate pathway	Metabolism	10
ko00040	Pentose and glucuronate interconversions	Metabolism	6
ko00680	Methane metabolism	Metabolism	22
ko04724	Glutamatergic synapse	Organismal Systems	2
ko00270	Cysteine and methionine metabolism	Metabolism	17
ko00565	Ether lipid metabolism	Metabolism	2
ko00360	Phenylalanine metabolism	Metabolism	10
ko00910	Nitrogen metabolism	Metabolism	16

注：上表展示的仅为前 10 的 pathway，完整列表请看相关文件

KEGG_Categories.pdf：pathway 注释分类结果，绘图源文件为 kegg_annot_catar.xls，展示如下图：

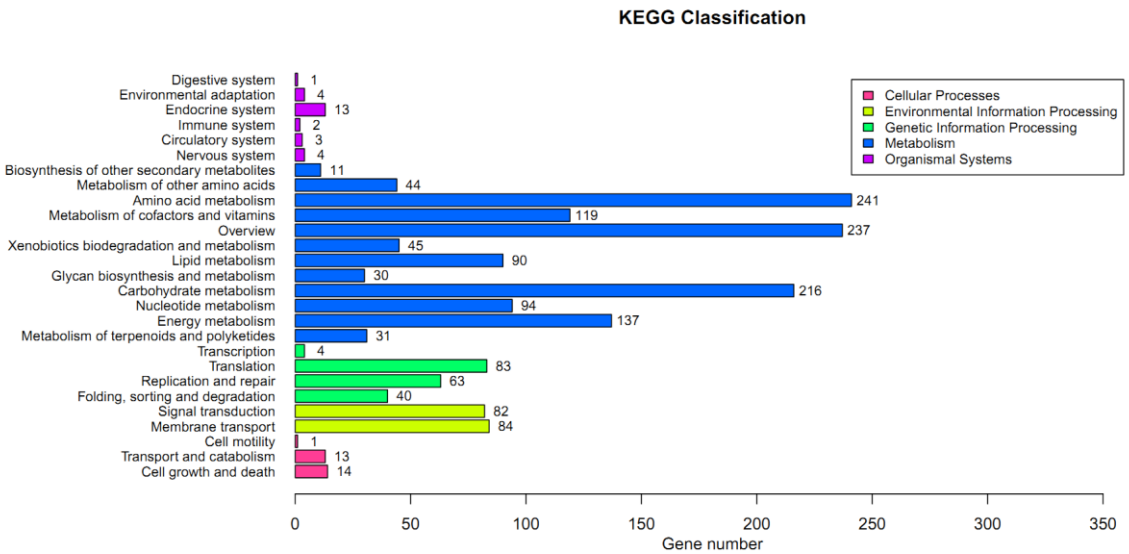


图 2.13pathway 注释分类条形图

说明：纵坐标为 KEGG 代谢通路的名称，横坐标为注释到该通路下的基因个数，将基因根据参与的 KEGG 代谢通路分为 5 个分支：细胞过程（A，Cellular Processes），环境信息处理（B，Environmental Information Processing），遗传信息处理（C，GeneticInformation Processing），代谢（D，Metabolism），有机系统（E，Organismal Systems）。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.6 基因组及基因高级注释

该部分分析，试图通过基因与基因组注释结果，来深入探究细菌多方面的特征，诸如毒力、耐药性、适应环境机理、乃至演进过程。客户根据自己研究的需要，重点关注其中的部分结果即可。

2.6.1 基因岛

基因岛（Genomics Islands，GIs）是一些细菌、噬菌体或质粒中有横向起源迹象的一部分基因组。目前在多种微生物基因组中都发现了多个基因岛，例如沙门氏菌、铜绿假单胞菌、鲍曼不动杆菌、金黄色葡萄球菌、肺炎杆菌等。基因岛大小通常从 10 到 500 Kbp，其 GC 含量和密码子选择与基因组的其余部分不同。一个基因岛可以与多种生物功能相关，能与共生或病原机理相关，与生物体的适应性相关等，因此一直以来都是研究的热点。

基因岛的许多亚分类基于它们所表现出来的功能，如病原性基因岛（Pathogenicity islands，PAIs）其上包含有毒力基因，与致病机理相关，抗生素抗性基因岛（Resistance islands，REIs）包含有许多抗生素抗性基因。

采用的工具：

软件 IslandPath-DIOMB 是基于序列组成的预测方法预测基因岛的，可识别病原岛，以及潜在的水平基因转移。主要是基于序列中含有二核苷酸偏向性（phylogentically bias）和移动性基因（mobility genes，如转座酶或整合酶）来判定基因组岛。

表 2.11 基因组岛预测结果

GeneIsland	length	CG_Ratio(%)	Gene_Count
tig00000000:2887878-2907131	19253	40.28463097	22
tig00000000:968153-982023	13870	39.84138428	12
tig00000000:2930125-2936176	6051	42.5384234	9
tig00000000:2850572-2865304	14732	43.23241922	11
tig00000000:1408712-1420397	11685	42.10526316	23
tig00000000:1340926-1350935	10009	42.16205415	7
tig00000000:585837-600020	14183	41.00683917	17
tig00000000:882508-893305	10797	42.502547	10
tig00000000:897942-915166	17224	40.85578263	13
tig00000000:2287497-2303933	16436	37.9532733	16
tig00000000:1572344-1579911	7567	39.98942778	9

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

tig00000000:217219-249656	32437	38.17862318	25
tig00000000:743254-747764	4510	39.57871397	7
tig00000000:1684786-1693973	9187	41.90704256	10
tig00000000:515672-559552	43880	40.8887876	37

结果详见/Report/5_AdvancedAnno/GenomeIslands，内包含 4 个文件：

GeneIsland.bed 基因岛位置的 bed 文件

GeneIsland.fa 基因岛的 fa 文件（核酸序列）

GeneIsland_gene.txt 基因岛上的基因（三列，第一列标记基因岛的方式为染色体:开始-结束第二列为预测得到的基因 ID，第三列为基因的产物，即简要注释信息）

GeneIsland_Summary.txt 基因岛信息汇总（四列，第一列为基因岛名称，第二列为长度，第三列为 GC 含量，方便与全基因组的 GC 含量对比，第四列为基因岛上基因的数目）

2.6.2 前噬菌体

整合在宿主基因组上的温和噬菌体的核酸称之为前噬菌体。噬菌体基因组与宿主菌染色体整合（或以质粒形式储存在细胞内）后，能随宿主细菌 DNA 复制而同步复制，并随细菌的分裂而传代，宿主细胞可正常繁殖，处于“溶原周期”。但在一定条件下，如紫外线、X 线、致癌剂、突变剂等作用下，噬菌体基因组可进行复制，产生并释放子代噬菌体，进入到“裂解周期”，此后噬菌体基因组即变为可增殖型而进行自主增殖，并使细胞裂解。

带有前噬菌体基因组的细菌称为溶原性细菌（lysogenic bacterium）。溶原性细菌具有抵抗同种或有亲缘关系噬菌体重复感染的能力，即使得宿主菌处在一种噬菌体免疫状态。

采用的工具

PhiSpy 考虑了前噬菌体上的 7 种特征：蛋白长度，转录区所在的正负链，AT-skew，GC-skew， unique phage words 的多寡，噬菌体插入位点以及与噬菌体蛋白的相似程度。该工具通过对基因组上不同区域噬菌体 trait 富集度的排序来确定可能的噬菌体位置，假阴性概率 6% ，假阳性概率 0.66%（PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies）。

上面的分析，得到了前噬菌体区域，并且可以划定出其中的基因。而为了对预测结果进行验证，我们采用 PHAST 数据库的已知噬菌体与前噬菌体序列进行 blast 比对，看看预测出的前噬菌体上，是否存在已

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

知的噬菌体相关的基因。

表 2.12 前噬菌体预测结果

proPhage	length	CG_Ratio(%)	Gene_Count	PHAST_Genes
tig00000000:2887878-2935669	47791	42.45569249	47	4
tig00000000:1401535-1419965	18430	43.7710255	30	20
tig00000000:146-17575	17429	42.01617993	17	2
tig00000001:11154-38105	26951	49.47868354	27	2

Blast.details.txt：blast 结果，仅仅针对落到 phiSpy 预测得到的前噬菌体区间内的基因

PHAST.blast.anno：以 blast 结果为基础，添加了蛋白条目的注释信息。（由于预测得到的前噬菌体上的基因为成功比对到 PHAST 的数据库，故本报告中，该文件未生成）

Prophage_protein_Anno.txt：根据上述 blast 结果对蛋白条目进行的注释。表头见上述表，同样仅针对落到 phiSpy 预测得到的前噬菌体区间内的基因

Prophage_summary.txt：前噬菌体汇总。额外多出一列，为每个前噬菌体区间中，多少蛋白被 PHAST 库注释到了。

Prophage.fa：前噬菌体 fa 序列

2.6.3、毒力因子

VFDB 数据库全称为 Virulence Factors of Pathogenic Bacteria，用于专门研究致病细菌、衣原体和支原体致病因子的数据库。其包含 75 个属，共 1800 个致病因子，30053 个与毒力因子相关的基因。VFDB 分为 SetA 和 SetB 两大部分，SetB（full dataset）中的 30053 基因包含有毒力因子相关基因，也有预测得到的可能的毒力因子基因，落入 2124 个毒力因子条目中。而 SetA 是 VFDB 的核心基因（core dataset），一共 2585 个，落入 487 个毒力因子条目中，这些基因的毒力均有实验证据。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毒力因子功能注释信息结合起来，得到注释结果。注释结果分为 SetA 和 SetB 两组。

毒力因子包括细菌毒素,调节细菌附着的细胞表面蛋白,保护细菌细胞表面的碳水化合物和蛋白质和水解酶等。在一些物种中，不同的独立因子组合能够引起不同的疾病。进一步的比较基因组分析能够发现更多关于生物和病原体进化的信息。

表 2.13 VFDB 注释毒力因子结果概况（分为 SetA 和 SetB 两组）

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

	SetA			SetB		
Total_Proteins	Predicted_VF_Proteins	Ratio(%)	VF_Terms	Predicted_VF_Proteins	Ratio(%)	VF_Terms
2753	101	3.6687	59	121	4.3952	83

结果详见/Report/5_AdvancedAnno/VFDB。由于分别针对 SetA 和 SetB 进行了比对分析，所以结果有相同的两套。以 SetA 为例：

A.blast.anno：对 blast 结果进行的注释，表头如下：

GeneName	基因名称（基因组）
GeneInfo	基因信息（基因组）
VFG_Name	毒力因子基因（VFG）名称
blast_identity	Blast Identity（均大于 20）
blast_evalue	Blast eValue （均小于 1e-5）
VFG_GI	毒力因子基因的 gi 号
VFG_Symbol	毒力因子基因的 Symbol
VF_Info	毒力因子信息
VF_Name	毒力因子名称
VF_term	毒力因子 ID
VF_Taxonomy	毒力因子基因所在的菌株

A.list：VFDB 注释汇总表

A.protein.fa：将注释到 VFDB 的蛋白单独提取出来

A.protein.summary：以蛋白为核心，统计每个蛋白有多少个 VF 与之对应。

A.VF.summary：统计每个 VF 中，落入了多少蛋白。将该文件拖拽到 excel 中，点击最末列的连接，可直接访问对应条目在 VFDB 官网上的详细信息。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.6.4 耐药基因

我们注释耐药基因选用 CARD 数据库。CARD 数据库全称 The Comprehensive Antibiotic Resistance Database，是经典的耐药基因数据库 ARDB（Antibiotic Resistance Genes Database）的升级版（ARDB 自 2009 年后就没有升级更新过）。数据库中，有 2359 条序列，3567 条 Antibiotic Resitance Ontology Term。

使用 BLAST 软件，把目标物种的氨基酸序列，与 CARD 数据库进行比对，把目标物种的基因和其相对应的耐药功能注释信息结合起来，得到注释结果。

表 2.14 CARD 注释耐药基因的结果

Total_Proteins	Predicted_AR_Proteins	Ratio(%)
5556	46	0.82793

结果详见/Report/5_AdvancedAnno/CARD。

CARD.blast.anno：以 blast 结果为基础，对相关的蛋白进行的详细注释，表头如下：

GeneName	基因名称（基因组）
GeneInfo	基因信息（基因组）
ARG_Protein	耐药基因（ARG）名称
blast_identity	Blast Identity（均大于 40）
blast_evalue	Blast eValue （均小于 1e-5）
ARG_DNA	ARG 的 GeneBank ID
ARG_Taxonomy	ARG 所在的物种
ARO_term	ARO 编号
ARO_Name	ARO 名称（部分 ARO 名称甚至等同于 GeneSymbol）
ARO_Description	ARO 详细信息

注：ARO 为 Antibiotic Resistance Ontology 的缩写，这个概念，是 CARD 汇总了耐药基因的生理功能后概括出的，但是，该概念并没有如 GO 那样被广泛的采用。所以，ARO 的 ID 仅仅在这张表中出现，而 ARO 的表述信息，则可视为一般的基因功能注释信息。

CARD.list：CARD 注释信息汇总

Protein.fa:注释到 CARD 的蛋白汇总 fa 文件

Protein.summary：蛋白为核心，统计每个蛋白有多少个 ARO 与之对应。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.6.5 病原宿主互作用

PHIbase 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。该数据库对寻找药物干预的靶基因研究有重要作用，同时该数据库还包括抗真菌化合物和相应的靶基因。

数据库中的每个基因都包含核酸和氨基酸序列，以及感染宿主过程中预测的蛋白功能的详细描述。

PHI-base 中的基因涉及的表型（phenotype）被分成 9 类，具体分类信息详见 PHI-base 的文献 The Pathogen-Host Interactions database (PHI-base):additions and future developments：

Table 3. Definitions for the nine high-level phenotype outcomes used in PHI-base

High-level phenotype outcome[a]	Definition
Loss of pathogenicity	The transgenic strain fails to cause disease that is observed in the wild type (i.e. qualitative effect).
Reduced virulence	The transgenic strain still causes some disease formation but fewer symptoms than the wild-type strain (i.e. a quantitative effect). Synonymous with the term reduced aggressiveness.
Unaffected pathogenicity	The transgenic strain which expresses altered levels of a specific gene product(s) causes the same level of disease compared to the wild-type reference strain.
Increased virulence (Hypervirulence)	The transgenic strain causes greater incidence or severity of disease than the wild-type strain.
Effector (plant avirulence determinant)	Some effector genes are required to cause disease on susceptible hosts but most are not. A plant pathogen-specific term which was previously referred to as a corresponding avirulence *(Avr)* gene. An effector gene is formally identified because its presence leads to the direct or indirect recognition of a pathogen in resistant host genotypes which possess the corresponding disease resistance *(R)* gene. Positive recognition leads to activation of plant defense and the pathogen either fails to cause disease or causes less disease. In the absence of the pathogen, effector delivery into a healthy plant possessing the corresponding *R* gene activates plant defense responses.
Lethal	The transgenic strain is not viable. The gene product is essential for life of the organism.
Enhanced antagonism	The transgenic strain shows greater endophytic biomass in the host and/or the formation of visible disease symptoms.
Resistant to chemical	The transgenic strain[b] grows and/or develops normally when exposed to chemistry concentrations that are detrimental to the wild-type strain.
Sensitive to chemical	The transgenic strain which expresses either no or reduced levels of a specific gene product(s) or possesses a specific gene mutation(s), has the same ability[c] as the wild-type strain to grow and develop when exposed to detrimental chemistry concentrations.

[a]Compared to wild-type reference strain (i.e. a direct isogenic strain comparison).
[b]Molecular studies on natural field isolate population are also considered, once the natural target site has been identified.
[c]On rare occasions increased sensitivity to chemistry has been observed.

表 2.16PHI 注释病原宿主互作用相关基因的结果

Total_Proteins	Predicted_PHI_Proteins	Ratio(%)
2753	63	2.288413

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

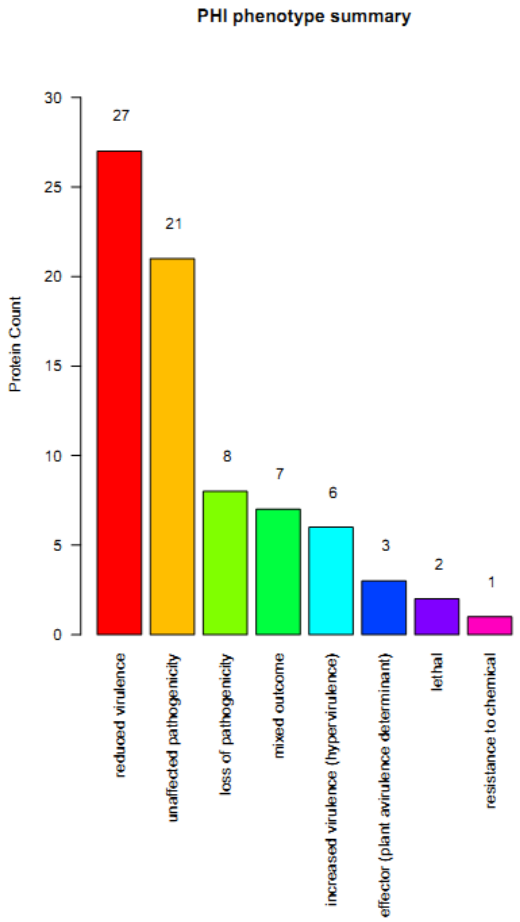


图 2.14 病原宿主互作用注释得到的表型汇总柱状图

结果详见/Report/6_AdvancedAnno/PHI，**采用该部分结果时，务必根据宿主物种信息对结果进行筛选**！

PHI.list：PHI 注释的概况

Prokaryote_PHI.blast.anno：blast 结果注释，表头如下

GeneName	基因名称（基因组）
GeneInfo	基因信息（基因组）
PHI_Protein(Identifier)	PHI 蛋白名称（SwissProt）
blast_identity	Blast Identity（均大于 40）
blast_evalue	Blast eValue （均小于 1e-5）
PHI_ID	PHI ID 号
PHI(GeneSymbol)	PHI 的 GeneSymbol
Pathogen_Organism	蛋白所在的病原菌
Host_Organism	病原菌所寄生的宿主
Host_Kingdom	宿主所在的界（**有动植物之分，老师根据这个信息，注意筛选 PHI 的结果**）
phenotype	表型信息
PHI_Link	PHI 注释信息的链接（未针对 PHI 专门做汇总，就把链接加到这里了）

Protein.summary：蛋白为核心，统计每个蛋白有多少个 PHI 条目与之对应。

pheno.summary.txt：注释上的 PHI 条目对应的表型，为 PHI.pdf 绘图的原始数据

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.6.6、分泌蛋白及信号肽

分泌蛋白是指在细胞内合成后，在信号肽的引导下穿过细胞膜分泌到细胞外起作用的蛋白质。分泌蛋白中有许多是生命活动所需的重要酶类。分泌蛋白的 N 端是由 15～30 个氨基酸组成的信号肽，对分泌蛋白的分泌起主导作用。使用信号肽预测工具 SignalP，采用神经网络和隐马氏模型的方法预测蛋白序列是否是分泌蛋白。

SignalP 根据不同物种类型，分别建立了不同的预测模型，以针对真核生物，革兰阳性菌和革兰阴性菌得到对应的预测结果。仅仅考虑前 70 个氨基酸，因为，信号肽仅位于蛋白链的开头，且长度极少长于 45 个氨基酸。

表 2.17 信号肽与分泌蛋白预测结果汇总

Organism Type	gram-
Total Proteins Number	2753
Signal Proteins from SignalP-TM	5
Signal Proteins from SignalP-noTM	184
Total Sigan Proteins	189
Signal Protein Ratio(%)	6.865237922

结果详见/Report/5_AdvancedAnno/signalP。

SignalP.txt 预测得到的分泌蛋白的汇总表，表头信息如下：

Protein	蛋白名称
Protein_Info	蛋白信息
Model	预测模型（两种，SignalP-TM 为跨膜蛋白模型预测成立，SignalP-noTM 为非跨膜蛋白模型预测成立。对革兰阳性菌，没有 SignalP-noTM，只有 SignalP-TM）
Signal_Peptide	信号肽序列
Cleavage_site	剪切位置，也可理解为信号肽长度。因为分泌蛋白在后期加工的时候，要把信号肽的部分给剪掉。

Predict_Info.txt： 预测结果详细信息，仅保留了预测得到的可能的信号肽的信息。

name	蛋白名称
Cmax	最大 C 值。C 值为剪切位点值。每个氨基酸会有一个 C 值，在剪切位点处 C 值是最高的。
pos	最大 C 值出现的位置
Ymax	最大 Y 值。Y 值是综合考虑 S 值和 C 值的一个参数，其比单独考虑 C 值要更精确。因为在一条系列中 C 值可能有不止一个较高的位点，但是剪切位点只有一个；此时的剪切位点就由 Y-max 值来推测的，为 S 值是陡峭的位置和具有高 C 值的位点。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

pos	最大 Y 值出现的位置
Smax	最大 S 值。每个氨基酸对应 1 个 S 值，在结果显示的图表中有一个曲线显示 S 值的变化趋势，（在 full 模式中可以看见具体数值），信号肽区域的 S 值较高。
pos	最大 S 值出现的位置
Smean	S-mean 是从 N 端氨基酸开始到剪切位点处各氨基酸的平均 S 值。
D	D 值是 S-mean 和 Y-max 的平均值，对区分是否为分泌蛋白具有重要作用
?	是否分泌蛋白，全部为是（Y）
Dmaxcut	D 值的阈值（不同模型设定不同）
Networks-used	采用的神经网络模型（两种，SignalP-TM 为跨膜蛋白模型预测成立，SignalP-noTM 为非跨膜蛋白模型预测成立。对革兰阳性菌，没有 SignalP-noTM，只有 SignalP-TM）

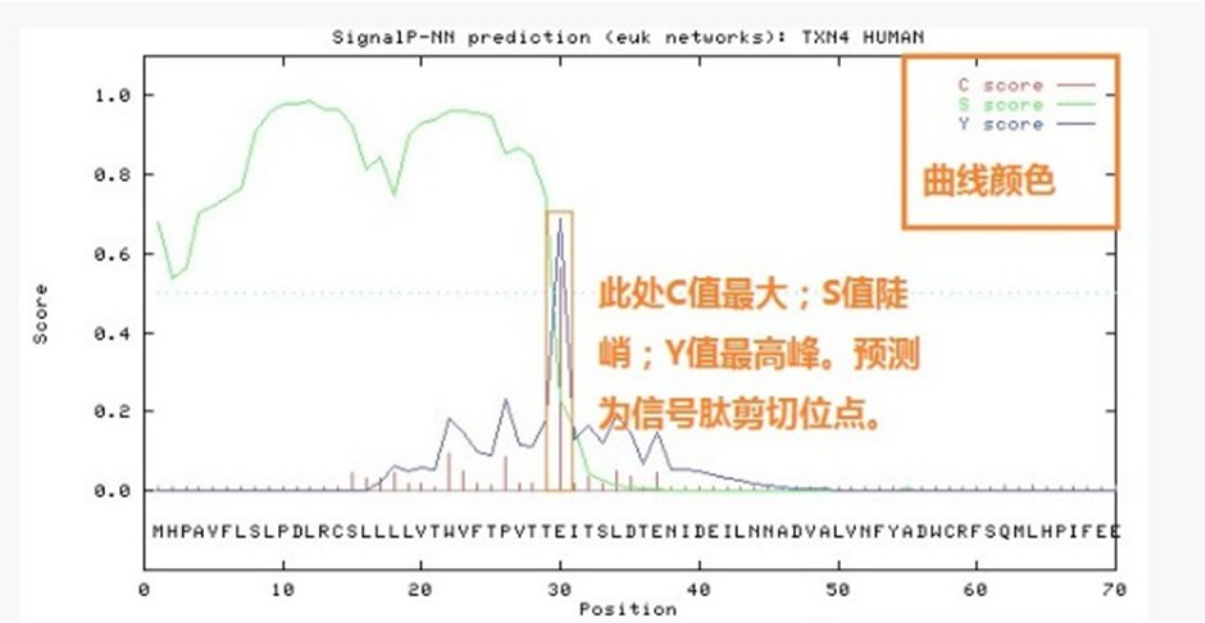
SignalP.summary：预测结果总结，包括 6 行内容

Organism Type	革兰阳性，格兰阴性还是真核生物
Total Proteins Number	总蛋白种类数
Signal Proteins from SignalP-TM	跨膜蛋白模型得到的分泌蛋白
Signal Proteins from SignalP-noTM	非跨膜蛋白模型得到的分泌蛋白（革兰阳性没有这个模型）
Total Sigan Proteins	总分泌蛋白数目
Signal Protein Ratio(%)	分泌蛋白占总蛋白数目的百分比

MatureProtein.fa 成熟分泌蛋白序列（即剪掉信号肽的蛋白序列）

SignalPeptide.fa 信号肽序列

Images 神经网络模型三个值（C，S，Y）在前 70 个氨基酸中变化趋势的情况。同样的图分 jpg 和 eps（矢量图格式，可用 ie 浏览器等打开，也可用 Adobe Illustrator 编辑，转格式）



松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

2.6.7、碳水化合物酶

CAZy 全称为 Carbohydrate-Active enZYmes Database，碳水化合物酶相关的专业数据库，内容包括能催化碳水化合物降解、修饰、以及生物合成的相关酶系家族。其包含五个主要分类：糖苷水解酶（Glycoside Hydrolases, GHs）、糖基转移酶（GlycosylTransferases, GTs）、多糖裂解酶（Polysaccharide Lyases, PLs）和糖类酯解酶（Carbohydrate Esterases, CEs）、氧化还原酶（Auxiliary Activities, AAs）。此外，还包含与碳水化合物结合结构域（Carbohydrate-Binding Modules，CBMs）。

碳水化合物活性酶具有降解、修饰及生成糖苷键的功能，常常具有多结构域的特点，除了其催化作用的催化结构域外，还包含有功能各异的其它结构域，碳水化合物结合结构域（Carbohydrate-Binding Module, CBM）就属于这些附属结构域的一种。碳水化合物结合结构域是一种非催化结构域，能折叠成特定的三维空间结构，具有结合碳水化合物的功能。近年来研究表明：碳水化合物结合结构域能通过结合碳水化合物活性酶的底物，高碳水化合物活性酶的催化结构域作用于底物的活性。

预测的数据库，来自 dbCAN，该数据库给出了各种碳水化合物酶结构域构建的隐马尔科夫模型索引文件，于是采用 hmmscan 对细菌中所有蛋白进行预测，预测的 eValue 阈值为 1e-5

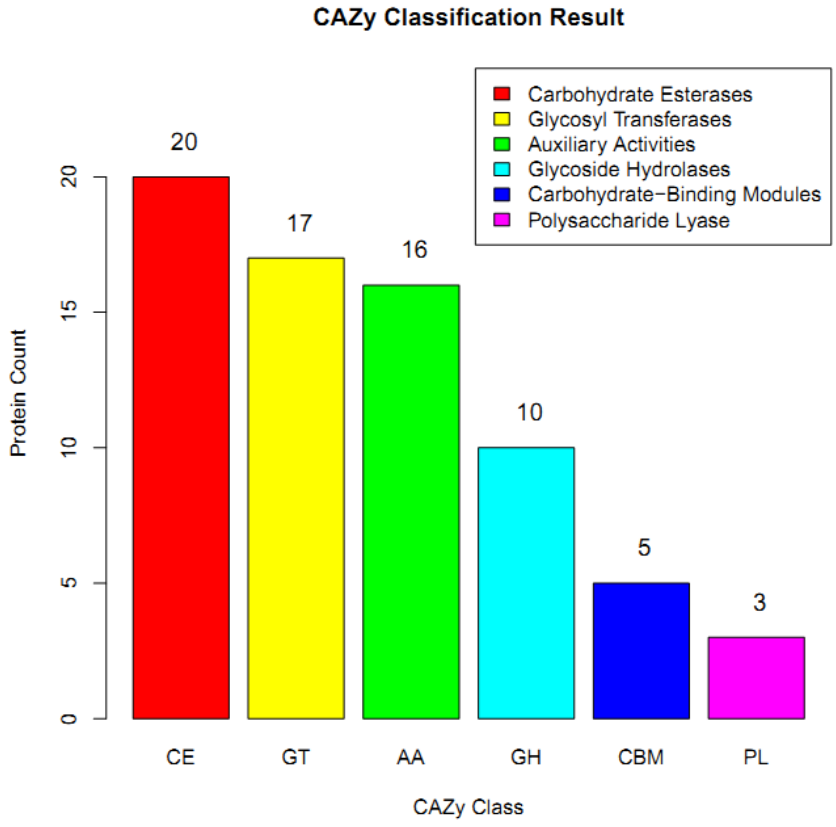


图 2.15 预测得到的碳水化合物酶所在大类数目柱状图

结果详见/Report/5_AdvancedAnno/signalP。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

hmmtable.txt hmmscan 主要运行结果，为成功预测到碳水化合物酶 domain 的蛋白及其对应 domain 的配对信息。

CAZy_All.xls 蛋白与碳水化合物酶 domain 的对应关系，5 列：

Protein	蛋白名称
CAZy_Family	碳水化合物酶家族
eValue	Hmm 搜索的 eValue
Gene_Info	基因信息
cazy-activities	碳水化合物酶活性的详细描述

Protein.xls 所有预测到的碳水化合物酶的蛋白的信息汇总，四列，蛋白名、蛋白信息、碳水化合物酶家族、以及家族数量

Predicted_CAZy.fa 预测到碳水化合物酶的蛋白详细序列

CAZy_family.xls 碳水化合物酶家族的详细信息。最后两列分别为来做 dbCAN 和 CAZy 官网上面碳水化合物酶家族的信息。

CAZy_Class.xls 碳水化合物酶大类统计信息，一共为 6 类。

CAZy.pdf 根据碳水化合物酶大类统计，绘制的柱状图：

2.6.8 短回文重复序列（CRISPR）搜索

CRISPR（clustered regularly interspaced short palindromic repeat sequences，成簇的、规律间隔的短回文重复序列）是在 40%已测序的细菌和 70%已测序的古细菌中都有报道过的、含有一个与噬菌体和质粒同源的短的重复序列，通过对外来同源的 DNA 作用对噬菌体有抗性，影响质粒的连接，是原核生物的免疫系统一部分。

CRISPR 相关基因（CRISPR-associated genes，Cas gene）位于 CRISPR 前端，与 CRISPR 结合行使免疫作用。在细菌和古细菌长期演化过程中形成的 CRISPR/Cas9 适应性免疫防御系统，可用来对抗入侵的病毒及外源 DNA，通过将入侵噬菌体和质粒 DNA 的片段整合到 CRISPR 中，并利用相应的 CRISPR RNAs（crRNAs）来指导同源序列的降解，实现了免疫的功能。

CRISPR Recongnision Tool 是一款经典的 CRISPR 搜索软件，输入全基因组的 DNA 序列来从中寻找 CRISPR。

结果详见：/Report/5_AdvancedAnno/CRISPR

CRISPR.All：CRT 直接输出的结果，内有 CRISPR 的详细信息，其中，重复出现的序列成为 Repeat，重复序列之间的部分成为 SPACER

CRISPR.summary：搜索结果概况

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

CRISRP.fa：所有 CRISPR 区域具体的序列

CRISPR.bed：标记 CRISPR 位置的 bed 文件

Pdf 文件（可能多个），为每个 CRISPR 的 SeqLogo 图

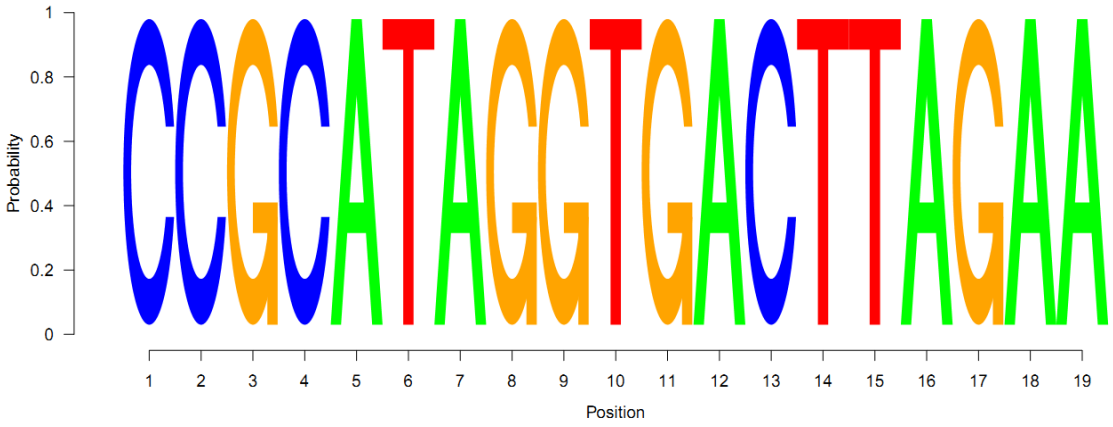


图 2.16 CRISPR 的 SeqLogo 图展示

2.6.9 近缘菌株泛基因组分析

前面的分析，根据 16s 序列得到了与测序菌株相近的一批菌株。从 NCBI 的细菌参考基因组数据库（ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria）中获取这些近缘的基因组，与测序得到的基因组中的编码基因进行比较，利用 P-GAP 工具（https://sourceforge.net/projects/pgap/）得到所有样品中的共有基因与特有基因，并进行深入分析。

所有样本中均存在的同源基因作为共有基因（Core gene），去掉共有基因后，得到非共有基因（Dispensable gene），特有基因（Specific gene）为只有该样品特异拥有的基因。所有非共有基因与共有基因合并作为泛基因集（Pan gene）。其中共有基因（Core gene）和特有基因（Specific gene）很可能与样品的共性和特性相对应，可以作为样本间功能差异的研究依据。

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

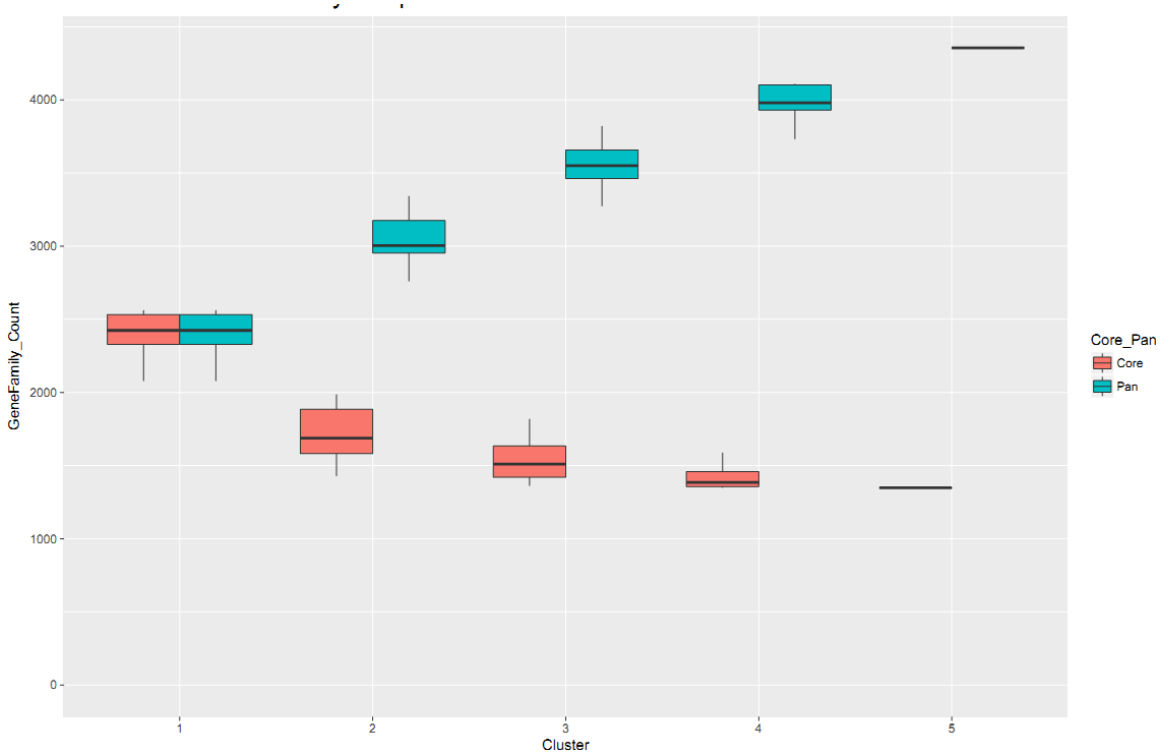


图 2-17 Core-Pan 基因稀释箱线图

绿色为 Pan 基因稀释曲线，红色为 Core 基因稀释曲线。横坐标表示每次统计选取的样品个数，纵坐标表示样品个数分布。

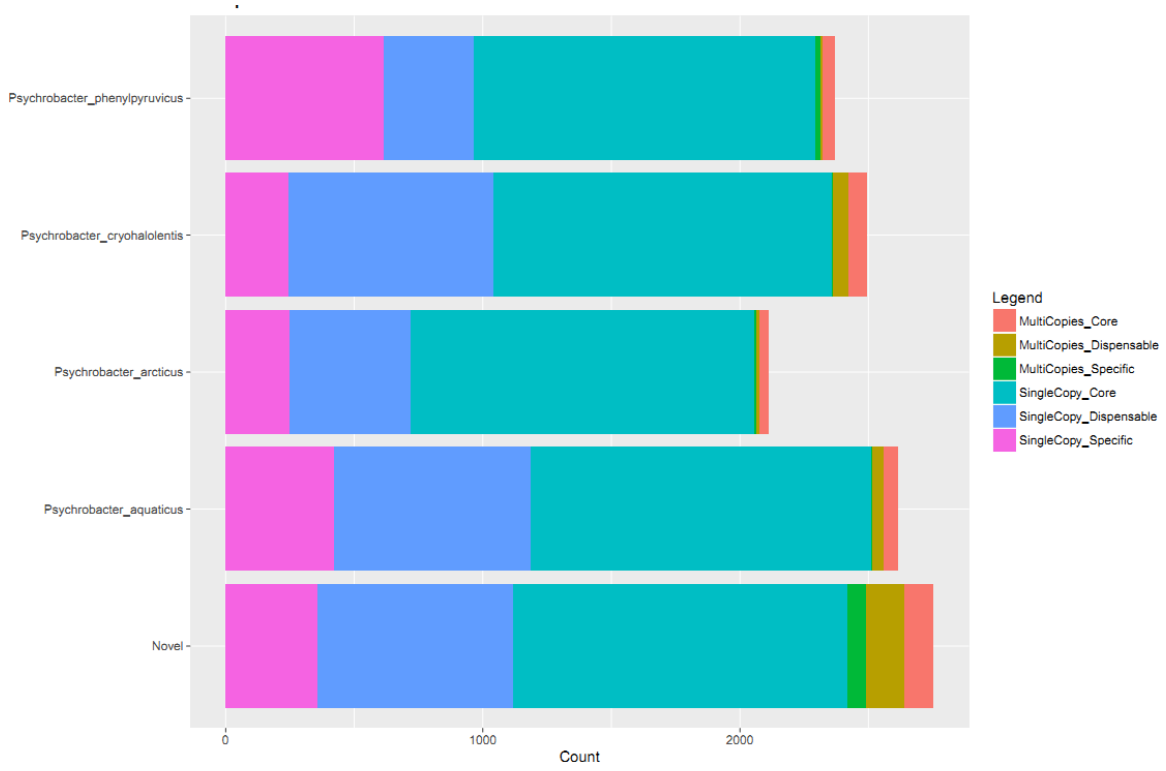


图 2-18 同源基因数目统计条形图

Single-copy：单拷贝同源基因，Multiple-copy：多拷贝同源基因数目，

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

Core：所有物种中，都出现的基因；Dispensable：部分物种（不止一种）中出现的基因；Specific：仅仅一个物种中出现的基因。

单拷贝基因：指在生物的一个染色体组中只有一份拷贝的基因，细胞中的大多数基因都是单拷贝的。多拷贝基因：进化过程中，微生物的基因组 DNA 序列会发生重复，这些重复有的继续发生进化歧异，成为与原来序列不同的新基因；而有的以结构和功能仍基本相同的形式保留下来成为多拷贝基因。

结果：

GenomeList.txt：参与分析的菌株序列列表，四列：属、种、株以及 RefSeq 的 ID

Gene_Distribution_By_Conservation.txt 每一个菌种根据基因保守性划分的统计结果。其中，每一列代表每个基因组中出现在不同数目菌株中的基因的数目（由于计数时同一菌株内多重拷贝的基因未去冗余，所以不同的菌株中 Core 基因的数量不全相同）。

PanGenome.Data.txt 不同数量的菌株，所有的组合方式得到的 Core 和 Pan 基因组的计数。若参与分析的菌株数量多，则不同基因的排列组合方式也会较多，这时这张表格的行数会越多。列数一共四列，从左到右依次是：菌株组合的数目，菌株总基因数（未去冗余），PanGenome（相似基因去除冗余），CoreGenome（相似基因去冗余）

Orthologs_Cluster.txt：泛基因组分析基因聚类与去冗余的详细结果。所有基因组中的全部基因根据序列相似性聚类成若干 Cluster，每一行代表一个 Cluster 中的所有基因。若一个菌株基因组中多条基因归到同一 Cluster 中，则该基因组中该 Cluster 的基因存在多重拷贝，这些基因 ID 以逗号间隔。若某基因组中无基因落入某 Cluster，则以-表示。

CorePanGeneStat.txt 与 panGene.pdf：统计每个菌株中划归到不同组别中基因的数目。Pdf 文件为表格的堆叠条形图。

CorePanGenome.txt 与 panGenome.pdf：与上一组不同的是，该统计针对 GeneCluster，合并了冗余的基因。

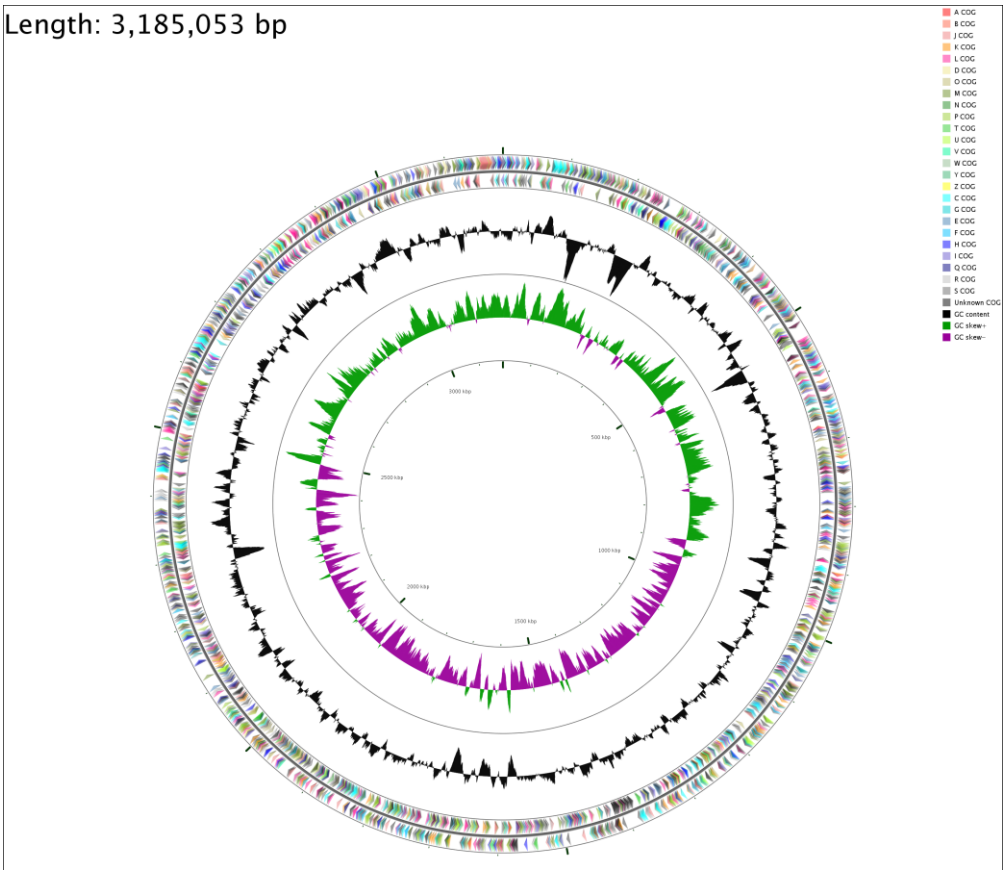
Core-Pan.pdf：泛基因组的稀释箱线图，该图的原始数据为 **Gene_Distribution_By_Conservation.txt**

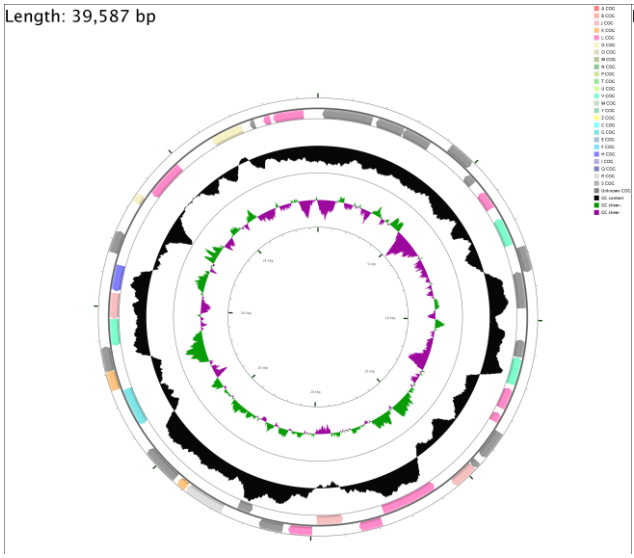
松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

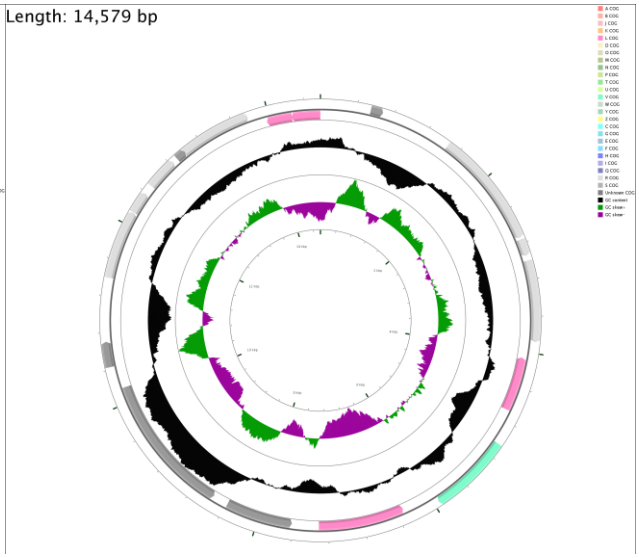
2.7 基因组圈图总结

每个 Contig 均绘制一幅圈图，作为分析的总结

圈图绘制在线工具 CGView Server：http://stothard.afns.ualberta.ca/cgview_server/index.html







上面为两条染色体下面为三条质粒

从外到内：圈 1 表示 cog 功能，一种颜色表示一种 cog 的功能

圈 2 表示 GC 含量；圈 3 表示 gcskew((G-C)/(G+C)

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

3. 结果说明

1_QC/原始数据及 QC 结果，此文件夹中会有对应样本的子文件夹，里面存放了各样本原始数据统计及 QC 结果。

All_sample_raw_data_infor.xls：所有样本原始数据统计结果，格式说明见上报告 6.1

All_sample_QC_infor.xls：所有样本 QC 之后结果统计，格式说明见上报告 6.1

*/PE_trimed_infor.xls：样本对应的 QC 之后结果统计，格式同上

*/Raw_data_infor.xls：样本对应原始数据统计，格式同 All_sample_raw_data_infor.xls

*/*_rand_100000.fa_blast_out.best_species_count.xls：样本污染比对结果，为每个样本随机挑选 100000 条序列 NT 数据库对应的物种统计结果

*/*_R1_fastqc.zip：样本 Read1 序列对应的 FASTQC 分析结果

*/*_R2_fastqc.zip：样本 Read2 序列对应的 FASTQC 分析结果，详细 FASTQC 结果解释见 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

2_assembly/ 目录下是基因组拼接及比较基因组分析结果

contigs.prinses.fa：原始拼接 scaffold 结果

HF563562.fasta：参考基因组

ref_query.pdf：与参考序列比对 nucmer 图

ref_query.coords：与参考序列比对信息文件，各列信息如下

S1、E1：参考基因组中比对起始位置及终止位置

S2、E2：scaffold 中起始及终止位置，当为反向比对时 S2>E1

LEN1/LEN2：参考基因组及 scaffold 比对上的长度

%IDY：相似度

LENR、LENQ：参考序列长度及 scaffold 长度

COVR、COVQ：比对长度占参考序列及 scaffold 长度的比例

TAGS：参考基因组及 scaffold ID

Assembly_result.xlsx：拼接结果统计

3_gene_prediction/ 基因预测结果

*.gbk：预测结果 gbk 格式文件

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

*.xls：预测结果信息文件

*.fna：orf 核酸序列

*.faa：orf 蛋白序列

Gene_GC_content.xls：orf GC 含量统计

Gene_len_distribution.xls：orf 长度分布统计

Gene_Len_Dis.pdf：orf 长度分布图

Gene_GC_content.pdf：orf GC 含量分布图

gene_prediction_stast.xlsx：预测结果统计文件

gene_result：基因统计结果

Predicted_rRNA.fa：预测的 rRNA

Predicted_tRNA.fa：预测的 tRNA

4_Annotation/ 蛋白注释结果

Annotation_ratio.pdf：各数据库注释比例图

Annotation_statistics.xls：各数据库注释统计表

nr_species_count.pdf：NR 数据库注释物种分布图

Unigene_annotation.xls：所有注释结果整合文件

nr_species_count.xls：NR 注释物种统计表

Venn_diagram_for_annotation.pdf：注释上的基因在各数据库韦恩图

blast_best_hit/ 各数据库最佳比对结果，各列说明如下：

Query_ID	比对的蛋白 ID
Query_len	Query 序列长度
Sbjct	比对上的数据中序列 ID
Sbjct_len	Sbjct 序列长度
Bitscore	分数，越高比对结果越好
Evalue	E 值，越小比对结果越好
Identity	相似度
Align_len	比对上的序列长度
Query_ratio	比对上的长度占 Query 序列总长度比例
Sbjct_ratio	比对上的长度占 Sbjct 序列总长度比例

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

COG/ 目录下是注释 COG 分类结果文件

gene_to_COG.xls：各基因注释上的 COG 详细信息

COG_Categories.pdf：各 COG code 注释上的基因数条形图

COG_code_count.xls：各 COG code 注释上的基因数统计表

COG_blast.out.infor.best.xls：COG 注释最佳比对结果

KEGG/ 目录下是注释 KEGG 结果文件：包括基因注释信息、ko 注释基因列表、ko 图

kegg_annot1.xls：基因 pathway 注释结果

kegg_annot2.xls：注释上的 pathway 统计结果

kegg_categories.xlsx：pathway 注释分类结果

KEGG_Categories.pdf：pathway 注释分类条形图

GO/ 目录下是注释 GO 功能分类的结果文件

full_GO_annot.xls：基因注释上 GO 详细信息

part_GO_annot.xls：GO 注释信息，每一行代表一条基因注释上的所有 GO 信息

full_GO_annot _all_level.txt：所有 GO 树上被注释到的 GO 统计结果

full_GO_annot _level2.xlsx：level2 水平上 GO 统计结果

GO_classification_level2.pdf：level2 水平上 GO 注释结果分布图

WEGO_annot.xls：WEGO 格式 GO 注释结果

5_AdvancedAnno/ 高级注释结果

这部分的文件内容，详见报告第五章每一小节下面的文件结果介绍

6_Circos/

每一条 Contig 的圈图展示

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

4. 参考文献：

1. Chin CS, Alexander DH, Marks P, et al. **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** Nat Methods. 2013 Jun;10(6):563-9. [PubMed]
2. Martin.**Cutadapt removes adapter sequences from high-throughput sequencing reads**. EMBnet.(2011). [EMBnet]
3. Schmieder R, Edwards R.**Quality control and preprocessing of metagenomic datasets.** *Bioinformatics. 2011 Mar 15;27(6):863-4.* [PubMed]
4. Berlin K, Koren S, Chin CS, et al. **Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.** *Nat Biotechnol. 2015 Jun;33(6):623-30.* [PubMed]
5. Koren S, Walenz BP, Berlin K, et al. **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. bioRxiv. (2016). [bioRxiv]
6. Massouras A, Hens K, Gubelmann C,et al. **Primer-initiated sequence synthesis to detect and assemble structural variants.** *Nat Methods. 2010 Jul;7(7):485-6.* [PubMed]
7. Edgar RC. **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research.* 2004;32(5):1792-1797. [PubMed]
8. Price MN, Dehal PS, Arkin AP. **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Mol Biol Evol. 2009 Jul;26(7):1641-50.* [PubMed]
9. Tatusov RL, Fedorova ND, Jackson JD, et al. **The COG database: an updated version includes eukaryotes.** *BMC Bioin-formatics. 2003 Sep 11;4:41. Epub 2003 Sep 11.* [PubMed]
10. Gene Ontology Consortium. **Gene Ontology annotations and resources.** *Nucleic Acids Res. 2013 Jan;41(Database issue):D530-5.* [PubMed]
11. Ogata H1, Goto S, Sato K, Fujibuchi W, et al. **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res. 1999 Jan 1;27(1):29-34.* [PubMed]
12. Hsiao W, Wan I, Jones SJ, et al. **IslandPath: aiding detection of genomic islands in prokaryotes.** *Bioinformatics. 2003 Feb 12;19(3):418-20.* [PubMed]
13. Akhter S, Aziz RK, Edwards RA. **PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies.***Nucleic Acids Res. 2012 Sep;40(16):e126.* [PubMed]
14. Arndt D, Grant JR, Marcu A, et al. **PHASTER: a better, faster version of the PHAST phage search tool.***Nucleic Acids Res. 2016 Jul 8;44(W1):W16-21.* [PubMed]
15. Chen L, Xiong Z, Sun L,et al. **VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors.** *Nucleic Acids Res. 2012 Jan;40(Database issue):D641-5.* [PubMed]
16. Jia B, Raphenya AR, Alcock B, et al. **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database.** *Nucleic Acids Res. 2016 Oct 26. pii: gkw1004.* [PubMed]

松江区香闵路698号
中国 上海
T: 400-821-0268; 800-820-1016
F: +86 021 37772170
生工生物工程（上海）股份有限公司

17. Urban M, Pant R, Raghunath A, et al. **The Pathogen-Host Interactions database (PHI-base): additions and future developments.** Nucleic Acids Res. 2015 Jan;43(Database issue):D645-55. [PubMed]
18. Petersen TN, Brunak S, von Heijne G, et al. **SignalP 4.0: discriminating signal peptides from transmembrane regions.** *Nat Methods. 2011 Sep 29;8(10):785-6.* [PubMed]
19. Lombard V, Golaconda Ramulu H, Drula E, et al. **The carbohydrate-active enzymes database (CAZy) in 2013.** *Nucleic Acids Res. 2014 Jan;42(Database issue):D490-5.* [PubMed]
20. Bland C, Ramsey TL, Sabree F, et al. **CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.***BMC Bioinformatics. 2007 Jun 18;8:209.* [PubMed]
21. Zhao Y, Wu J, Yang J, et al. **PGAP: pan-genomes analysis pipeline.** *Bioinformatics. 2012 Feb 1;28(3):416-8.* [PubMed]
22. Grant JR, Stothard P., et al. **The CGView Server: a comparative genomics tool for circular genomes.** *Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W181-4.* [PubMed]