# Solexa 基因组测序技术在食源性蜡样芽胞杆菌基因组研究中的应用

李琳[1]，姬莉莉[1]，段静静[1]，金河坡[1]，李冰[1]，徐振波[1,2]

（1.华南理工大学轻工与食品学院，广东广州 510640）（美国马里兰大学微生物病理系，巴尔的摩 MD21201）

**摘要：**本文拟通过对 2 株从酱油渣中分离获得的蜡样芽孢杆菌用 DNA 纯化试剂盒提取全基因组 DNA，并用胶回收试剂盒纯化后，构建质量合格的测序文库并进行 Solexa 基因组重测序，获得基因组序列后对其测序质量进行评估，并对基因组信息进行分析，其中包括利用 mapping 分析研究 2 株菌相对于参考基因组的各自的变异基因，并找到其共有变异基因，并对共有变异基因进行 COG 注释、GO 功能注释及 KEGG 生物学通路分析，通过对变异基因的这些分析，结合文献中其他对蜡样芽孢杆菌耐盐的研究，在基因功能分析的基础上找出决定其耐盐能力的关键基因，本研究共找到 49 个耐盐相关基因，这些基因很可能在菌株耐盐中起着关键作用。本研究探索其耐盐机制，为酱油渣的综合利用创造条件，同时进一步为微生物在食品发酵过程中的各种耐盐机制研究提供重要依据。

**关键词：**基因组重测序；蜡样芽孢杆菌；测序质量评价；基因组测序结果分析

**文章篇号：**1673-9078(2015)8-143-152 DOI: 10.13982/j.mfst.1673-9078.2015.8.024

# Genomic Study of Foodborne *Bacillus cereus* using Solexa Genome Sequencing Technology

**LI Lin[1], JI Li-li[1], DUAN Jing-jing[1], JIN He-po[1], LI Bing[1], XU Zhen-bo[1,2]**

(1.College of Light Industry and Food Sciences, South China University of Technology, Guangzhou 510640, China)

(2.Department of biomedical science of university of Maryland, Baltimore 21201, United States)

**Abstract:** The DNA of two *Bacillus cereus* strains, isolated from soy sauce residue, was extracted using a DNA extraction kit and purified using a gel recovery kit. A high-quality sequencing library was constructed, followed by whole-genome resequencing using Solexa sequencing technology. The quality of genome sequencing was assessed and the genomic information was analyzed. Mapping analysis was used to compare the mutant genes of two strains using the reference genome, the common mutant genes were singled out, and Clusters of Orthologous Groups (COG) annotation, Gene Ontology (GO) functional annotation, and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathway analysis were conducted on the common mutant genes. Using the results of these analyses of common mutant genes combined with previous literature related to salt tolerance of *Bacillus cereus*, the key genes that determine salt tolerance were singled out based on the analysis of gene function. In this study, 49 genes were found to be associated with salt tolerance that may play an important part in the process. The mechanism of salt tolerance was explored in this study, creating conditions for comprehensive utilization of soy sauce residue and providing an important basis for further research on the mechanism of salt tolerance in microorganisms during the fermentation of food.

**Key words:** genome resequencing; *Bacillus cereus*; evaluation of sequencing quality; analysis of genome sequencing results

在生物界，DNA 转录成 RNA，再翻译成氨基酸及形成蛋白质，是一切生命活动的体现。生物体的一切遗传信息，均保留在物种的基因组中；因此，对基因组的了解与探索，是生物学研究的基础。基因组测序是研究微生物的常用手段，其中全基因组重测序通过对已知基因组序列的物种进行不同个体的基因组测序，进一步对测序结果进行 mapping 分析，找出可靠的变异位点，在此基础上对个体或群体进行差异性分析。近年来，基因组测序技术在食源性致病微生物的基因组研究中得到广泛应用。

蜡样芽孢杆菌是一种革兰氏阳性菌，在不适宜的生长条件下可以形成芽孢，在环境中分布很广泛，易在食品中污染并产生毒素，是一种易引起食物中毒的病原菌。在酱油加工过程中，酱油原料经蒸煮后每克

收稿日期：2014-09-12

基金项目：国家 973 计划项目(2012CB720800)；国家自然科学基金青年基金项目(31201362)；中央高校基本科研业务费面上项目(2012ZM0060)

作者简介：李琳，教授，研究方向：糖类物质及其药物制备与生物利用

通讯作者：徐振波（1982-），男，博士，讲师，研究方向：食源性致病微生物安全研究

原料中的细菌数骤减至 $10^2$~$10^3$ 个，存活菌群大多为芽孢杆菌，为酱油中最原始的菌种。芽孢在适当的生长条件下很快开始萌发，在适宜温度下可以几何级数急剧生长繁殖，造成成曲细菌数剧增。在发酵过程中加大发酵基质盐度，不利于芽孢杆菌生长[1]。发酵后得到的酱油渣食盐含量约 10%，有的高达 25%，也经常会污染蜡样芽孢杆菌，直接将酱油渣饲养家畜，会导致家畜中毒等[1]。酱油渣盐度较高，芽孢杆菌仍能正常生长存活，然而其耐盐机制尚未明确。

本文以酱油渣中分离获得的 2 株蜡样芽孢杆菌为研究对象，通过对其 Solexa 基因组测序，分析序列信息，获得与耐盐相关的功能基因和关键调控因子，研究芽孢杆菌的耐盐机制，为酱油渣综合利用的安全性提供参考，同时进一步对微生物在食品发酵过程中的耐盐机制研究提供研究依据。

## 1 材料与方法

### 1.1 实验材料

本研究中的 2 株蜡样芽孢杆菌（B25 与 B26）均是从酱油渣中分离得到，用营养琼脂固体培养基 37 ℃培养或者营养琼脂液体培养基 37 ℃ 150 r/min 震荡培养，菌株已经使用常规生化鉴定试剂盒鉴定（MID 芽孢杆菌生化鉴定条，广东环凯微生物科技有限公司）；细菌基因组快速提取试剂盒和凝胶回收试剂盒，均购自广州东盛生物科技有限公司。

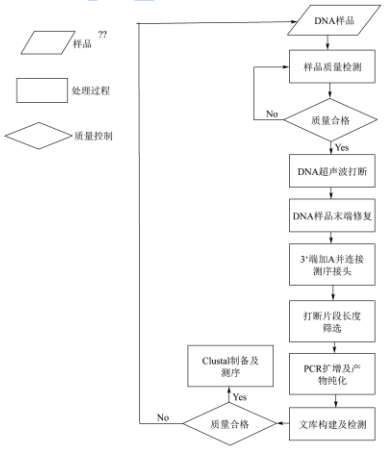
### 1.2 基因组 DNA 的提取与纯化

对 2 株芽孢杆菌进行基因组 DNA 提取与纯化。首先，利用细胞裂解液裂解细胞释放基因组 DNA，吸附柱硅基质材料在高盐、低 pH 情况下吸附 DNA，低盐、高 pH 情况下释放 DNA 这一特性，先将 DNA 吸附到柱子上，经蛋白酶消化、漂洗液清洗除去蛋白质、脂质等杂质后，用纯化液洗脱获得基因组 DNA。对于芽孢杆菌要利用溶菌酶裂解细胞释放基因组 DNA 后，用可逆吸附柱吸附 DNA，蛋白酶 K 消化蛋白质，乙醇沉淀 DNA，用纯化液洗脱获得基因组 DNA。获得基因组 DNA 后，进行核酸电泳验证，并进行胶回收纯化 DNA。胶回收需要在 360 nm 长波紫外线光盒照射下快速切胶，溶解胶，利用硅胶可逆吸附柱实现 DNA 的纯化。

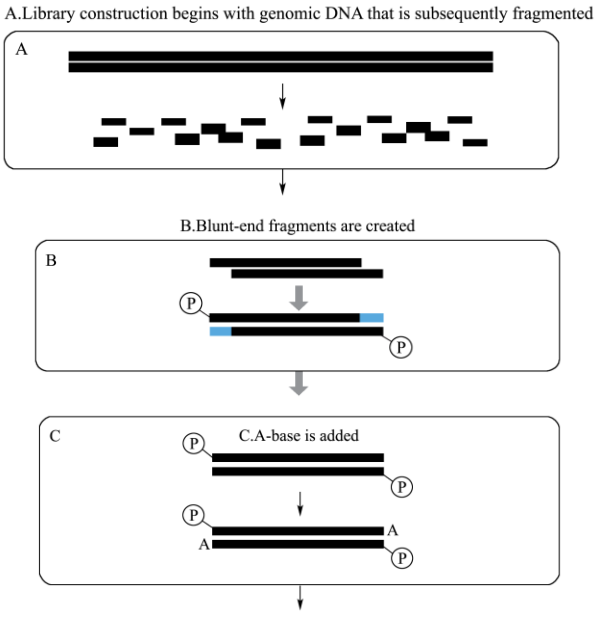
### 1.3 测序文库的建立

获得基因组 DNA 后，需要对 2 株蜡样芽胞杆菌（25 与 26）进行测序文库的构建。该过程可分为 DNA 样品的打断、修复平末端、3'端加‘A’以及 PCR 扩增等步骤（图 1A）。具体流程包括：首先，对所提取的 DNA 样品进行质量检测，质量合格的样品用于构建测序文库；采用超声法 Covaris 或者 Bioruptor 将大片段 DNA 样品（如基因组 DNA 或长片段 PCR 产物）随机打断，产生条带小于或等于 800 bp 的一系列不同长度的 DNA 片段（图 1B-A）；其次，用 T4 DNA 聚合酶、Klenow DNA 聚合酶和 T4 PNK 将打断后形成的粘性末端全部修复成平末端（图 1B-B）；随后将 DNA 片段 3'端加上碱基‘A’，使得 DNA 片段能与 3'端带有‘T’碱基的特殊接头连接（图 1B-C），用电泳法选择并筛选需回收的目的片段连接产物（图 1B-D），再使用 PCR 技术扩增两端带有接头的 DNA 片段（图 1B-E）；最后，对构建文库的质量进行检测，用合格的文库进行 cluster 制备和测序[2]。在文库构建过程中有两次的质量控制，以获得高质量的测序文库及后续更准确可靠的测序结果。

A




B



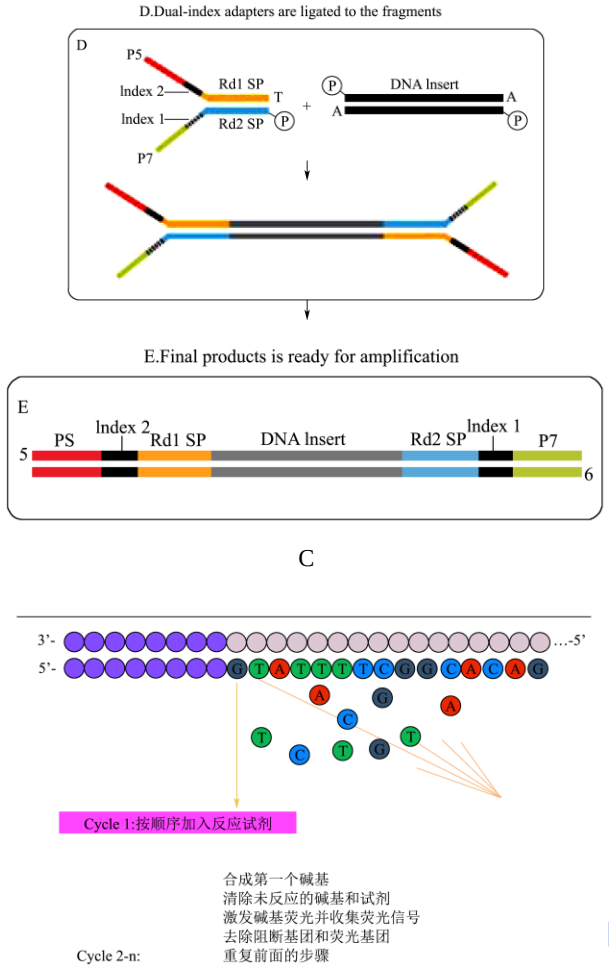



图 1 (A-C)测序文库构建流程与测序原理图

**Fig.1 (A-C) Schematic illustration of sequencing library construction and sequencing principle**

## 1.4 Solexa 测序

在构建蜡样芽胞杆菌的测序文库后，进行 Solexa 测序。Solexa 测序技术最核心的是 DNA 簇（DNA Cluster）和可逆性末端终结（Reversible terminator）。测序的基本原理是边合成边测序（SBS, Sequencing by Synthesis），其过程如下（图 1C）：首先将大片段 DNA 随机打断成小的片段，加上‘A’，加上测序接头，并经过电泳筛选所需要的片段，构建质量合格的测序文库；然后将单链 DNA 片段两端固定在芯片（flow cell）上，形成桥式结构，进行桥式 PCR 扩增。扩增后形成数百万条待测序的片段，每条模板扩增出一个分子簇，主要是为了测序时放大信号，这个过程被称为簇形成过程；将荧光标记的 dNTP、聚合酶和引物加入到测序通道启动测序。DNA 合成时，伴随着碱基的加入会有焦磷酸被释放，从而发出荧光。不同碱基用不同荧光标记，读取到核苷酸发出的荧光后，将 3’羟基末端切割，随后加入第 2 个核苷酸，重复第一个核苷酸的步骤，直到模板序列全部被合成双链 DNA[3~4]。

## 1.5 测序数据的质量控制流程

获得下机数据之后，需要对原始数据进行质量控制(quality control)即 QC 分析。原始的下机数据（*.bcl file）为荧光图像信号根据不同碱基标记的荧光信息，可得到基因组的碱基序列，这个过程为碱基识别，得到 Raw data（*fastq），去除之前测序加上的接头信息后，进行质量控制，如果质量合格，就可以与参考基因组比对，比对结果中去除一些冗余序列，质量合格，就得到了 BAM 文件，得到的去低质量数据后得到 clean data (*filter.fastq)（图 2A）。

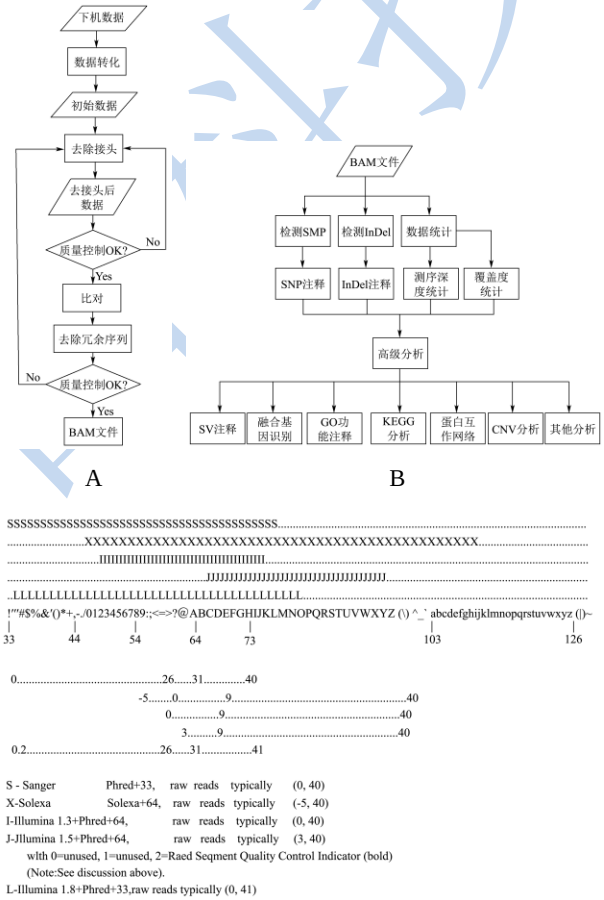



图 2 (A-C)测序数据分析流程及质量控制编码

**Fig.2 (A-C) Schematic illustration of sequencing data analysis and quality control coding processes**

## 1.6 测序质量编码

测序质量是评估测序碱基准确性很重要的参数，而根据不同的软件，测序质量编码采用不同的方案，一共有五种方案（图 2C），其中，如 Phred+33 编码方式，则表示采用 33 至 126 对应的 ASCII 字符对应于 0 至 93 之间的 Phred 质量得分，定义是$Q=-10\lg p$，其中

p 是碱基测序错误的概率，44 对应 ASCII 字符“，”，Phred 质量得分为 44-33-1=10，$10=-10\lg p$，$p=10^{-1}=0.1$；那么 54 对应的 ASCII 码“6”的测序序列错误概率为 0.01。与前边类似，测序质量值（$Q$）与测序错误率（E）相对应，其相应关系为：

$$Q=-10*\log_{10}E,$$

即如果错误率为 1%，带入上式可得

$$Q=-10*\log_{10}(0.01)=20,$$

也就是说如果 90%的测序信息是准确的，测序数据质量得分为$Q20$。因此，当测序错误率分别为 10%、1%、0.1%和 0.01%时，其碱基质量分别对应为 Q10、Q20、Q30 和 Q40。而在实际应用中一般选择 Q20 及 Q30 来评价测序质量[5]。

## 1.7 高通量测序数据 mapping 分析及分析软件

在获得高通量测序数据并对其测序质量评价后，要想了解基因组数据与参考基因组的变异位点及与功能的关系，需对基因组测序序列进行 mapping 分析，其原理与 NCBI 中 BLAST 程序类似，由于重测序基因组数据巨大，其比对速度更快，耗时较短，其主要把测序得到的序列与参考基因组进行比对，通过将获得的每一个序列高速和参考基因组序列比对，得到 read 在参考基因组上的匹配位置及匹配质量等相关信息[5]。

根据参考基因组在 NCBI 里的功能注释等已知信息，可预测序列的相关功能等信息，这对于下游功能分析非常重要，序列的 mapping 分析是测序数据分析的关键和必备步骤。mapping 分析代表软件有：SSAHA（英国剑桥大学韦尔科姆基金会桑格学院），BWA（英国剑桥大学韦尔科姆基金会桑格学院），Bowtie（美国马里兰大学）等。在千人基因组计划项目实施过程中，研究小组发现不同 mapping 软件的输出格式的不同，对后续分析造成了困难，研究小组开发了目前标准比对数据格式 SAM 及其二进制格式 BAM（Binary Alignment/Map），并且提供了下游分析工具 SAMtools，目前绝大多数 mapping 软件都支持比对数据格式输出 SAM 格式[5]。

## 1.8 全基因组测序数据 mapping 分析

在获得蜡样芽胞杆菌的高通量测序数据之后，通过对蜡样芽胞杆菌的测序序列进行 mapping 分析，应用 BWA 软件，以 *B. cereus* ATCC14579 为参考基因组进行比对，获得 SAM（Sequence Alignment/Map）输出文件。

## 1.9 SAM 注释

在蜡样芽胞杆菌的 mapping 分析后，根据 SAM 文件对序列比对结果的注释，可获得测序序列来自于参考序列的位置（position），根据参考基因组的功能注释等已知信息，可预测序列的相关功能等信息。理解和使用 SAM 格式的软件及 SAMtools 中的工具，要理解 SAM 格式每列所代表的意义。

比对数据输出的 SAM 格式文件总共分为两部分，第一部分为注释信息，头片段（header section），可有可无，以符号“@”开始，用不同的 tag 表示不同的信息，比如，@SQ，参考序列说明。@CO，任意的说明信息。第二部分为比对结果部分，其中每一行均表示一个片段的比对信息，不同的行用“\t”分隔，每行一共有 11 个必须字段及多个可选字段组成。可选字段则纪录了比对软件或测序仪器的一些附属信息，由“标签：类型：值”（“TAG:TYPE:VALUE”）构成。11 个必须字段分别为：比对片段的编号（QNAME）；位标识（FLAG），每一个数字代表一种比对情况，得到的值是所有情况的数字相加总和；参考序列的编号（RNAME），没有比对上的序列就没有参考序列，这里用‘*’表示；比对上的位置（POS），从 1 开始计数，若没有比对上，记为‘0’；mapping 的质量（MAPQ）；比对信息简要表达式（CIGAR），使用数字加字母表示比对结果，其中常用的字母有 M-Match/Mismatch，I-Insertion，D-deletion，P-padding，H-hard clipping，N-skipped bases，S-soft clipping，所以 2S5M1P1I4M 就表示前 2 个碱基被剪切去除了，然后 5 个比对上了，然后打开了一个缺口，有一个碱基插入，最后是 4 个比对上了，顺序对应与比对片段；下一个片段比对上的参考序列的编号（RNEXT），如果没有另外的片段，这里是‘*’，同一个片段，用‘=’；下一个片段比对上的位置(PNEXT)，如果没有比对上，仍记为‘0’；Template 的长度(TLEN)；序列片段的序列信息(SEQ)，如果没有信息，此处为‘*’。序列的质量信息（QUAL），格式同 FASTQ 一样[5]。CIGAR 中 M/I/S/=/X 等对应数字的和要等于序列长度。值得注意的是，这 11 个字段的顺序是固定的。

## 1.10 基因组数据高级分析

通过对测序结果进行 mapping 分析，找出可靠的变异位点，并在此基础上对个体或群体进行差异性分析。通过这种方法，可以准确的寻找出大量的可靠的单核苷酸多态性位点（SNP, Single Nucleotide

Polymorphism），插入缺失位点（InDel，Insertion Deletion），结构变异位点（SV，Structure Variation），拷贝数变异（CNV，Copy Number Variation）等变异信息，从而获得生物群体的遗传特征，利用这些变异我们可以寻找与其功能相关的信息。我们从酱油渣中分离到两株菌，经生化试验得知其为芽孢杆菌，为了探究在高盐环境中生存中耐盐机制，选择对这两株菌进行基因组重测序，对其测序结果进行分析得到其SNP、InDel位点，并对其共有变异位点进行统计，对其功能进行注释。

## 2 结果与讨论

### 2.1 测序技术的选择

对食源性微生物进行基因组分析，首先必须进行全基因组测序，而测序技术的选择是关键。基因组测序从1970年开始经历了三代的发展，第一、二代测序技术已经商业化，并在基因组重测序，RNA-Seq，ChIP-Seq等分子生物学领域广泛应用，以测序为关键词发表的文章数量也在飞速增加[5]。从以Sanger双脱氧核苷酸末端终止法为代表的第一代测序技术，到现在研究者测序时经常会使用到的第二代测序技术（以Illumina公司的Solexa技术、ABI公司的SOLiD技术及Roche公司的454技术为代表）以及正在不断发展完善的第三代测序技术（以Helicos公司的单分子测序技术，Pacific Biosciences公司的单分子实时检测技术和Oxford Nanopore Technologies公司的纳米孔测序技术为代表），高通量测序技术伴随人类基因组计划的完成得到了飞速发展。而各个测序技术都有其不同的特点也都有其关键技术，主要从以下两方面讨论：从测序建库方法来说[5]，第二代高通量基因组测序技术需要将基因组DNA随机打断成片段，经电泳筛选、PCR扩增进行建库，不同的公司扩增方法不同，Solexa技术采用桥式PCR，而454技术和SOLiD技术均采用微乳液PCR。从测序方法来说，第二代测序技术需要在flow cell上边合成边测序而第一代Sanger测序技术是通过电泳的方法实现的。而第三代测序方法不需要PCR扩增，减少了非理想PCR扩增导致的测序质量下降[3~5]。

在选择使用测序技术时，需根据不同的研究内容和研究目的，选择经济又相对准确的测序技术。传统的第一代基于Sanger的测序方法准确性非常高，但是因为其操作复杂、价格高、测序速度慢等缺点，无法满足高通量全基因组重测序的需要，而第三代测序技术正处于发展和完善中。相比较这三代测序技术，第二代高通量测序技术不仅能达到较高准确度，而且测序成本也大大降低、测序速度也很高，而Solexa测序技术，因为其测序片段长度提高以及测序准确性的提升，成为了研究者经常会选择的测序方法[5]。而Sanger的测序方法，因其高的准确性，现在常用于基因变异的验证，这样避免了可能因为测序准确性问题引起的对变异基因的错误估计。综上所述，本研究选择Illumina公司的Solexa测序技术进行测序。

将测序数据与参考基因组进行比对分析，对测序数据覆盖区域、覆盖深度等做出综合评价。本研究以芽孢杆菌 ATCC 14579的基因组序列（NC_004722.1 GI: 30018278，全长为5,411,809 bp）作为参考基因组[23]，选用BWA作为比对软件。BWA是目前普遍使用的比对软件，具有速度快，结果准确，错误率低等优点。比对后获得SAM格式的结果文件。对SAM文件进行处理，移除潜在的PCR重复片段，得到BAM格式文件，BAM文件为SAM文件的二进制格式，BAM文件可用SAMTOOLS或IGV软件（Integrative Genomics Viewer）查看[6]。

*B. cereus* 25、26全长均为5.41 Mb，覆盖率分别为89.7%，89.74%；平均测序深度为177.84、98.85。其中测序深度（Sequencing Depth）是指测序得到的碱基总量（bp）与基因组大小（Genome）的比值，是评价测序量的指标之一。覆盖度是指测序获得的序列占整个待测区域的比例。如果测序的覆盖度是98%，则表示仍有2%的序列区域是通过测序没有获得的。测序深度与基因组覆盖度之间是一个正相关的关系，测序带来的错误率或假阳性结果会随着测序深度的提升而下降。重测序的个体，如果采用的是双末端或Mate-Pair方案，当测序深度在10~15X以上时，基因组覆盖度和测序错误率控制均得以保证。比对结果统计显示*B. cereus* 25、*B. cereus* 26两样品覆盖率均在89%左右，而覆盖深度均在90 X以上。

### 2.2 测序质量的评价

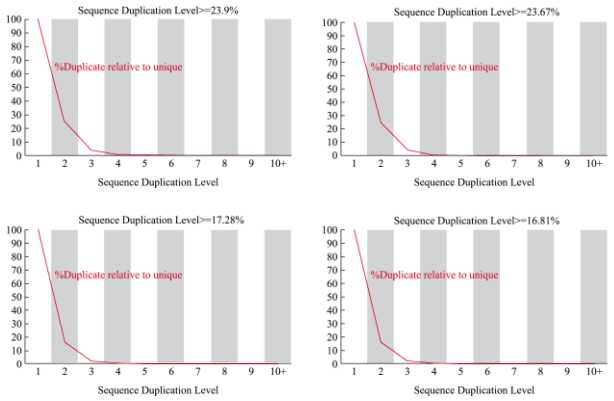
在基因组重测序后，需对原始测序数据进行质量分析以评估测序数据是否适合进行生物信息学分析，质量分析主要包括测序质量及碱基组成分析。原始数据质量评估包括每个测序循环的碱基组成、每个测序循环的质量分布和每个测序循环的碱基质量波动图。

#### 2.2.1 重复序列水平（Duplicate Sequences）

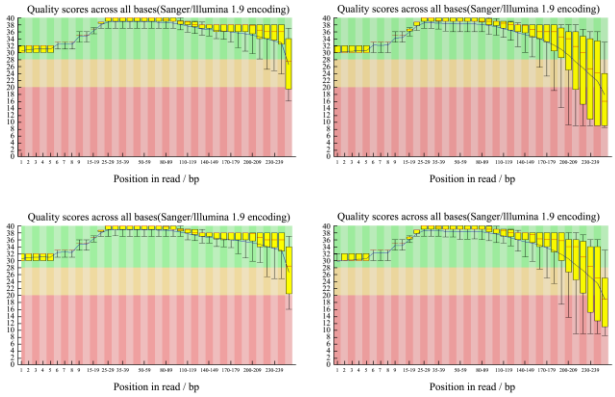
对测序质量进行评价，首先通过统计序列完全一样的片段的频率，评价重复序列的情况（图3A）。在基因组测序过程中，随着测序深度增加，可能会产生重复，若重复程度较高则说明测序可能出现了偏差，

如测序建库过程中 PCR 扩增问题。

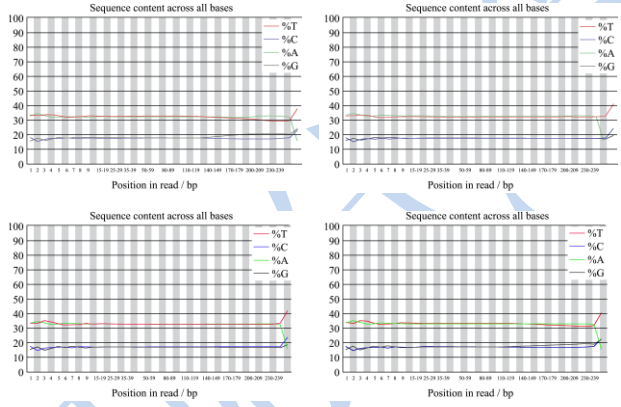
A：重复序列水平（Sequence duplication level）



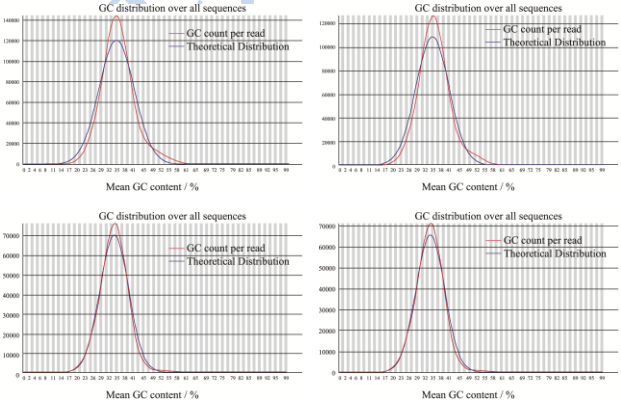

B：碱基测序质量（Per Base Sequence Quality）



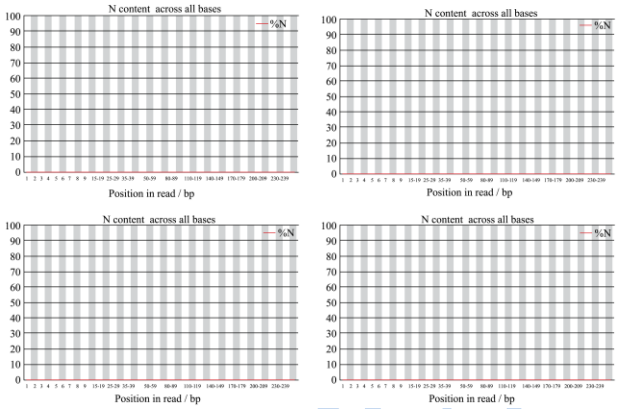

C：序列各碱基的含量（Per Base Sequence Content）



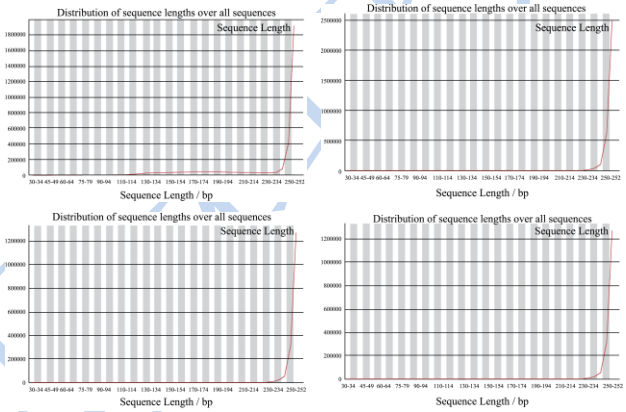

D：序列的 GC 含量（Per Sequence GC Content）




E：未识别碱基 N 含量（Per Base N Content）




F：序列长度分布（Sequence Length Distribution）




G：测序序列质量得分（Per Sequence Quality Scores）

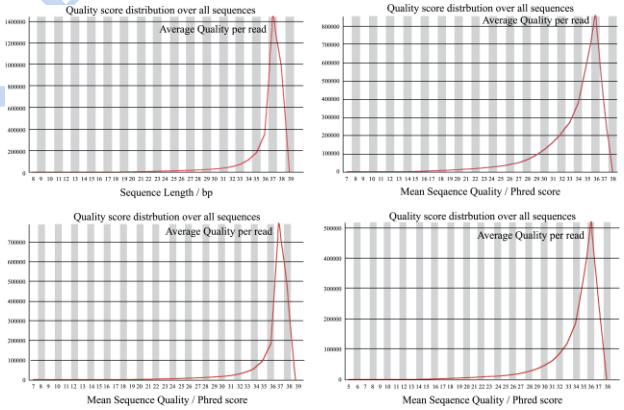



图 3 （A-G）测序质量各指标的汇总图

**Fig.3 Summary graph of each index of sequencing quality**

如图 3A 所示，通过对重复的次数（横坐标）和重复读长的数目（纵坐标），以特异性序列的总数为 100%进行统计。由于重测序原始数据量大，若对全部序列进行统计则工作量较大，因此通过统计测序数据的前 200000 条片段代表全基因组数据的重复情况。

在片段长度方面，一般只选取一定长度，对于超过设定长度的，则只统计其中一部分碱基；同时，重复数目大于等于 10 的 reads 被合并统计。由于随着片段长度增加测序质量下降，较长的片段可信度较低，造成只统计一定长度的片段可能低估其重复程度。当

非特异性（unique）的读长(reads)占总数的比例大于20%时，报"WARN"；当非 unique 的 reads 占总数的比例大于 50% 时，报"FAIL“。在本研究中，重测序原始数据的 reads 值>20%，报"WARN"，这可能是测序深度的增加造成的。从后边的测序质量评价得知，本次测序的测序质量很高达到 Q30，测序结果可信。所以这个指标只能作为一个参考的指标，不能仅根据这个指标来判断测序质量的好坏。

### 2.2.2 碱基测序质量（Per Base Sequence Quality）

测序碱基质量是评价测序质量的一个重要的指标，碱基质量$Q=-10*log_{10}E$，E 为测错的概率。

如果一条 reads 某位置出错概率为 0.01 时，其 quality 就是 20。

如图 3B，横轴代表碱基位置，纵轴代表质量。其中红色表示中位数，黄色是质量分数四分位 25%-75%区间，蓝线是平均数。在设置中，任一位置的下四分位数低于 10 或中位数低于 25，报"WARN"；任一位置的下四分位数低于 5 或中位数低于 20，报"FAIL"。从这个图中我们可以看到这次测序的质量几乎均在 Q30 范围内，测序质量较高，可信度比较高。

### 2.2.3 序列各碱基含量（Per Base Sequence Content）

对所有 reads 的每一个位置，统计 ATCG 四种碱基的分布。如图 3C，横轴代表位置，纵轴为百分比。正常情况下，无论在什么位置四种碱基的出现频率应该是接近，四条线应该平行且很接近。而实际测序时可能会有以下偏差情况：四条线在某些位置纷乱交织，即部分位置碱基出现偏差（bias），这代表可能有 over-represented sequence 的污染；四条线平行但是分开，这说明所有位置的碱基比例一致的表现出 bias，往往代表文库有 bias 或者测序中的系统误差。根据设定，当任一位置的 A/T 比例与 G/C 比例相差超过 10%，报"WARN"，A/T 比例与 G/C 比例相差超过 20%，报"FAIL"。在本次实验中该比例相差超过 20%，报"FAIL"，但这个指标不能代表本次测序整体的趋势，这主要是因为随着测序读长的增加，测序质量逐渐下降，在测序读长的最后出现了偏差，这次测序基本上处于比较稳定的情况，这次测序整体上还是比较可信，测序随机性很好。

### 2.2.4 序列 GC 含量（Per Sequence GC Content）

如图 3D，红线是实际情况，蓝线是理论分布（符合正态分布，其均值由平均 GC 含量推断的，本图中大概 35%，不一定是 50%）。当形状接近正态但偏离理论分布提示可能有系统偏差。偏离理论分布的 reads 超过 15%时，报"WARN"；偏离理论分布的 reads 超过 30%时，报"FAIL"。本图中虽然也符合正态分布，但偏离理论分布 15%，报"WARN"，可能测序中存在一些系统偏差。

### 2.2.5 未识别碱基 N 含量(Per Base N Content)

当测序仪器不能辨别某条 reads 的某个位置到底是什么碱基时，就会产生“N”。对所有 reads 的每个位置，统计 N 的比率。如图 3E，正常情况下 N 的比例是很小的，所以图上常常看到一条直线，但放大 Y 轴之后会发现还是有 N 的存在，这是可以接受的。而当 Y 轴在 0%~100%的范围内也能看到“鼓包”时，说明测序系统出了问题。当任意位置的 N 的比例超过 5%，报"WARN"；当任意位置的 N 的比例超过 20%，报"FAIL"。本图为一条直线，N 的比例是很小的。

### 2.2.6 测序长度分布（Sequence Length Distribution）

如图 3F，当 reads 长度不一致时报"WARN"；当有长度为 0 的 read 时报"FAIL"。该结果显示长度不一致，但总体还是集中在 250 bp。

### 2.2.7 测序质量得分（Per Sequence Quality Scores）

如图 3G，横轴为 quality，纵轴是 reads 数目。当峰值小于 27（错误率 0.2%）时报"WARN"，当峰值小于 20（错误率 1%）时报"FAIL"。在本次实验中，峰值出现在 37，测序质量较高。

对于其他碱基质量评估如 Overrepresented sequences&Kmer Content 都是 pass 状态。

## 2.3 基因组测序数据的分析

当前高通量基因组测序数据分析主要，包括匹配（mapping），拼接（assembly），测序序列定量，富集（peak）分析以及下游功能分析等。其中下游功能分析包括基因功能注释，Gene Ontology 分析，pathway 分析等。根据这些数据，更大程度上挖掘并解释 DNA，RNA 及表观三个水平上对应的高通量基因组测序数据，更加全面系统的研究生物学问题成为生物信息学，计算生物学，统计学及计算机科学等领域科研人员的研究方向，逐渐更进一步提高高通量基因组测序数据的利用率。

### 2.3.1 变异基因检测及注释

#### 2.3.1.1 . SNP 位点检测及注释

首先对测序结果与参考基因组比对得到变异位点即 SNP（single nucleotide polymorphism）及 InDel

（Insertion/Deletion）位点。单核苷酸多态性指个体间基因组 DNA 序列同一位置单个核苷酸变异所引起的多态性，是指不同物种个体基因组 DNA 序列同一位置上的单个核苷酸存在差别的现象。

通过检测样品中每个位点上所具有的最高概率的基因型，获得样品与参考基因组之间的一致性文件（CNS 文件)，并对一致性文件进行筛选和过滤，获得高可信度的多态性位点，最后对其进行分类和及注释。使用 SAMTOOLS 软件检测 SNP 位点，并使用 ANNOVAR 对 SNP 位点进行注释[7]。SNP 位点统计主要统计以下四项内容：（1）SNP 位点为纯合/杂合位点；（2）SNP 分布区域（编码区、基因上游、基因下游、基因间隔区）；（3）是否改变基因功能：同义突变、错义突变、无义突变等。对于 *B. cereus* 25 发现有 16,207 个纯合位点，325 个杂合位点，*B. cereus* 26 有 29,567 个纯合位点，776 个杂合位点；其中

*B. cereus* 25 有 13042 个位点位于基因编码区，剩余大部分是同时位于基因上游及下游区，

*B. cereus* 26 有 24621 个位于编码区。与 *B. cereus* 25 相同，大部分同时位于基因上游及下游；*B. cereus* 25 有 6,076 个错义突变位点，*B. cereus* 26 有 11,580 错义突变位点，对于两株菌剩余大部分为同义突变，还有少部分的终止子获得和缺失。

2.3.1.2 InDel 位点检测及注释

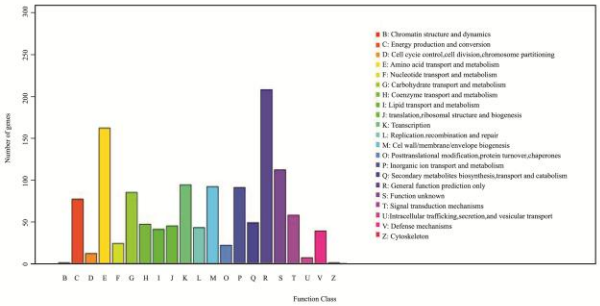
InDel 即插入/缺失，是指两种亲本在全基因组中的差异，相对另一个亲本而言，其中一个亲本的基因组中有一定数量的核苷酸插入或缺失。对比后获得的一致性文件（CNS 文件）进行筛选和过滤，获得高可信度的 Indel 位点，并对其进行分类和注释。使用 SAMTOOLS 软件检测 InDel 位点，并使用 ANNOVAR 对 InDel 位点进行注释[6~7]。InDel 位点相对与 SNP 位点少了很多，与 SNP 类似，两株菌大部分的 InDel 位点位于基因编码区及同时位于上游及下游区。对于 *B. cereus* 25 分别为编码区为 142 个，基因上游、下游区为 288 个，而 *B. cereus* 26 分别为 219、439 个。对于这些插入或缺失位点，大部分为移码突变，*B. cereus* 25 为 87 个移码突变位点，*B. cereus* 26 为 137 个移码突变位点。

2.3.1.3 共有变异位点汇总

这两株菌均为耐盐菌株，所以对其共有的 SNP 和 InDel 位点及其相应基因进行统计。由于分析的物种并非常见模式物种，因此基因注释要基于基因序列与已知数据库的比对。将 SNP 位点及 InDel 位点分析的样品间共有变异基因结果进行合并，将 *B. cereus* 25、*B. cereus* 26 两样品共有基因序列与已知数据库进行比对，并依据比对信息对分析的基因进行功能注释（COG、GO 及 KEGG）。共有变异基因汇总得到共有 434 个共有 InDel 位点，其中大部分均位于上游区和基因编码区且绝大部分都属于纯合变异位点，得到 13646 个共有 SNP 位点，大部分处于基因编码区，少部分处于基因上游区，绝大部分也都属于纯合变异位点[8]。

### 2.3.2 样品间共有变异基因 COG 注释

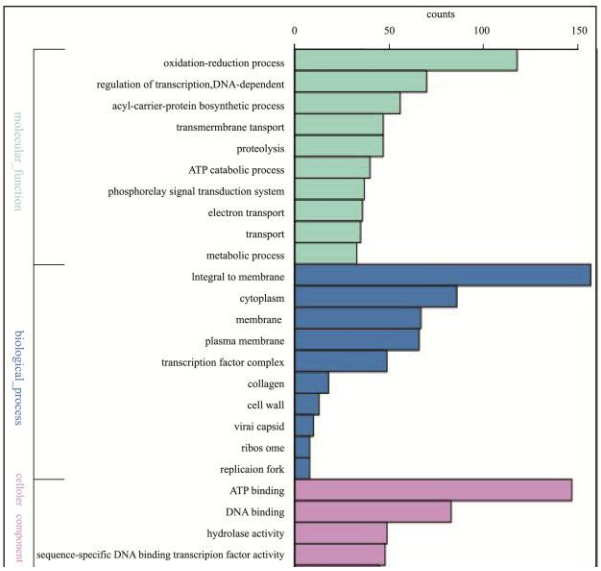
蛋白质直系同源簇（COGs）数据库是对细菌、藻类和真核生物的 21 个完整基因组的编码蛋白，根据系统进化关系分类构建而成。COG 库对于预测基因的功能和整个新基因组中蛋白质的功能都很有用。在分析中我们使用 blastall 软件将两样品共有的基因与 COG 数据库进行比对，并对比对结果进行注释。统计了每个功能类别富集的基因数目，每个基因的功能注释所属功能类别[9]。比对结果注释结果为，general function only 超过 200 个基因，Amino transport and metabolism 有超过 150 多个基因，Energy production and conversion, Carbohydrate transport and metabolism, Transcription, Cell membrane/walls/envelope biogenesis, Inorganic ion transport and metabolism, Function unknown 的数目都是在 50~100 个，Chromatin structure and dynamics，Cytoskeleton 仅有个别的基因。这说明耐盐性的发挥可能与一些一般功能或者氨基转移无机离子转运及能量的产生等相关的基因的表达有关。




### 2.3.3 共有变异基因 GO 功能注释

GO（Gene Ontology）功能注释基于基因与 NR 蛋白库的比对，NR 数据库为所有非冗余的的 GenBank CDS 区的翻译序列+参考序列的蛋白+PDB 数据库+SwissProt 蛋白数据库+PRF 蛋白数据库。GO 功能注释可分为分子功能（Molecular Function），生物过程（Biological Process）和细胞组成（Cellular Component）三个部分，主要提供基因功能分类标签和基因功能研究的背景知识。通过使用 Blast2go 软件将样品共有基因与 NR 蛋白库进行比对，获得基因的 GO ID，并进一步对基因功能进行注释[10]。GO 功能里的每一部分都分为 10 个项目，在分子功能里氧化还

原过程占最大的比例，大概 120 个，其他有酰基载体蛋白生物合成、跨膜转运、蛋白水解、ATP 分解过程、磷酸转移信号转导、电子转移等；在生物过程中，integral to membrane 占最大的比例，超过 150 个，其他为细胞膜、质粒等；在细胞组成中，ATP binding 占最多，接近 150 个，其他的为 DNA binding hydrolase activity 等。




### 2.3.4 共有变异基因 KEGG 注释

生物通路（Biological Pathway）分析基于 Kyoto encyclopedia of genes and genomes（KEGG）生物学通路数据库(http://www.genome.jp/)，其中，基因组信息存储在 GENES 数据库里，包括完整和部分测序的基因组序列；更高级的功能信息存储在 PATHWAY 数据库里，包括图解的细胞生化过程如代谢、膜转运、信号传递、细胞周期，还包括同系保守的子通路等信息；KEGG 的另一个数据库是 LIGAND，包含关于化学物质、酶分子、酶反应等信息。本研究从复杂调控网络的角度出发，对基因集合进行生物通路富集分析，将候选的突变/变异基因放到生物通路中进行综合分析，分析功能性变异对生物通路的影响程度及规律[15]。对每条生物通路显著性进行计算，通过 Fisher Exact Test 计算 p 值，并用 FDR 方法对显著性进行校正，得到校正后的 Corrected P-Value（Q-value）。以 0.05 为显著性阈值得到基因集合相对于背景具有统计意义的 KEGG 通路[11]。

### 2.3.5 耐盐基因通路及功能

根据文献报道，*B. cereus* ATCC 细菌中参与 $Na^+/H^+$、$K^+$运输、二肽或三肽转运、参与应激反应（如氧化性应激反应）等过程的基因变异均可能与其耐盐性相关，通过综合基因的 GO、KEGG 及 COG 注释结果，对于参与上述三类过程的基因进行筛选，初步获得与细菌耐盐性相关的基因[12~14]。耐盐结果分析得到 49 个与耐盐相关的基因，这些基因在 COG 注释里主要有与 $Na^+/H^+$、$K^+$ 转运及过氧化氢激酶丙酮酸激酶等基因，GO 功能注释主要与 ATP binding、跨膜转运等基因相关[14]，而基于 COG/GO 注释的这 49 个基因中在 KEGG 通路中大部分是 null 的状态，很多在 KEGG 数据库中找不到，而其他的三个通路在 KEGG 通路中有两相系统、RNA 降解和 ABC transporter。

## 3 结论

本研究通过对 2 株分离于酱油渣的蜡样芽孢杆菌进行全基因组 Solexa 测序技术，获得其全基因组信息；通过以基因组 *B. cereus* ATCC 14579 作为参考基因组进行比对，得到样本的 SNP 及 InDel 位点；同时，基于菌株的耐盐特性，对其共有变异进行统计并对其功能进行注释，并对其共有变异进行 COG 注释、GO 功能注释及 KEGG 生物通路富集统计。通过综合以上分析结果对三类过程的基因进行筛选，初步获得与细菌耐盐性相关的基因。下一步将通过对筛选的耐盐基因进行 Sanger 测序验证以及转录组学研究，以进一步在转录组或蛋白组等水平上深入阐述其耐盐机制，为食源性微生物的耐盐特性研究提供科学依据。

## 参考文献

[1] 郑海燕.酱油生产中的芽孢杆菌及其防治[J].中国调味品,1989,12:1-7
ZHENG Hai-yan. *Bacillus cereus* and its prevention during the procee of say souce production [J]. China Condiment, 1989, 12: 1-7

[2] Li Heng, Durbin Richard. Fast and accurate short read alignment with Burrows-Wheeler transform [J]. Bioinformatics, 2009, 25(14): 1754-1760

[3] 于聘飞,王英,葛芹玉.高通量 DNA 测序技术及其应用进展[J].南京晓庄学院学报,2010,3:1-5
YU Pin-fei, WANG Ying, GE Qin-yu. High-fluxed DNA sequencing technology and its application development [J]. Nanjing Xiao Zhuang University, 2010, 3: 1-5

[4] 王兴春,杨致荣,王敏,等.高通量测序技术及其应用[J].中国生物工程杂志,2012,32(1):109-114
WANG Xing-chun, YANG Zhi-rong, WANG Min, et al. High-throughput sequencing technology and its application [J]. China Biotechnology, 2012, 32(1): 109-114

[5] 王从茂.高通量基因组数据的处理、分析与建模[D].上海:上海交通大学生命科学技术学院,2012
WANG Cong-mao. Processing, analysis and modeling on

high-throughput genomic data [D]. Shanghai: School of Life Science and Biotechnology Shanghai Jiao Tong University, 2012

[6] Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools [J]. Bioinformatics, 2009, 25: 2078-2079

[7] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data [J]. Nucleic Acids Research, 2010, 38: 164-164

[8] Krzywinski M, Schein J, Birol İ, et al. Circos: an information aesthetic for comparative genomics [J]. Genome Research, 2009, 19(9): 1639-1645

[9] Roman L Tatusov, Natalie D Fedorova, John D Jackson, et al. The COG database: an updated version includes eukaryotes [J]. BMC Bioinformatics, 2003, 4(41): 1-14

[10] Stefan Götz, Juan Miguel García-Gómez, Javier Terol, et al. High-throughput functional annotation and data mining with the Blast2GO suite [J]. Nucleic Acids Research, 2008, 36(10): 3420-3435

[11] Xie Chen, Mao Xi-zeng, Huang Jia-ju, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases [J]. Nucleic Acids Research, 2011, 39(2): 316-322

[12] den Besten, Heidy M. W, van der Mark, Eric-Jan, Hensen, Lonneke, et al. Quantification of the effect of culturing temperature on salt-induced heat resistance of Bacillus species [J]. Applied and Environmental Microbiology, 2010, 76(13): 4286-4292

[13] den Besten, Heidy M W, Mols, Maarten, Moezelaar, Roy, et al. Phenotypic and transcriptomic analyses of mildly and severely salt-stressed *Bacillus cereus* ATCC 14579 cells [J]. Applied and Environmental Microbiology, 2009, 75(12): 4111-4119

[14] Sarah Hahnkea, Daniel Wibbergb, Geizecler Tomazetto, et al. Whole genome sequence of clostridium bomimense strain M2/40 isolated from a lab-scale mesophilic two-phase biogas reactor digesting maize silage and wheat straw [J]. Biotechnology, 2014, 184: 199-200