# De novo 测序技术在啤酒易感乳杆菌全基因组研究中的应用

刘君彦[1]，李琳[1]，李冰[1]，邓阳[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）（2.美国马里兰大学微生物病理系，巴尔的摩 MD21201）

**摘要：**本文应用 De novo 测序技术对一种未经测序的啤酒易感乳杆菌进行全基因组研究，包括全基因组遗传信息的首次获取以及基因功能的生物信息学分析与注释。通过对 1 株从啤酒中分离获得的乳杆菌基因组 DNA 进行提取与纯化，构建了符合质量要求的测序文库，并进行了 De novo 测序；通过对 De novo 测序数据进行筛选与质量评价，进一步对筛选后的数据进行了 De novo 组装与结果评价，最后获得乳杆菌全基因组序列；在获得全基因组序列的基础上，对全基因组序列进行基因预测，并对预测基因进行 GO 基因功能注释、COG 基因功能注释与 KEGG 生物通路注释，获取了该乳杆菌的基因序列与基因功能信息。该乳杆菌全基因组测序数据与组装结果良好，获得的序列中共有 1,824 个预测基因，其中分别有 545、1,303、120 条预测基因有相应的 GO、COG、KEGG 注释。本研究为该啤酒易感乳杆菌功能基因的研究与生物信息学分析提供了基础数据和技术参考。

**关键词：**啤酒易感乳杆菌；De novo 测序；全基因组信息；基因功能注释
**文章篇号：**1673-9078(2015)11-155-162 DOI: 10.13982/j.mfst.1673-9078.2015.11.025

# Application of De novo Sequencing in the Whole Genomic Study of Beer-spoilage *Lactobacilli*

**LIU Jun-yan[1], LI Lin[1], LI Bing[1], DENG Yang[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, China)

**Abstract:** The whole genome of a typical hard-to-culture beer-spoilage *Lactobacilli* strain which had never been sequenced was analysized. The *Lactobacilli* strain was isolated from beer and its genomic DNA was extracted and purified. A high-quality sequencing library was constructed and the whole genome sequencing was then performed using De novo system. The sequences were further analyzed and the predicted genes were obtained, including GO gene function, COG gene function and KEGG pathway. In the whole genome sequences of this *Lactobacilli* strain, 1,824 genes were predicted and 545, 1,303, and 120 of them had GO, COG, and KEGG annotation, respectively. This study may provide more references for the study of functional gene and bioinformatics analysis of the beer-spoilage lactobacilli strains .

**Key words:** beer-spoilage *Lactobacilli*; De novo sequencing; whole genome information; gene function annotation

基因是编码蛋白质或 RNA 分子遗传信息的基本遗传单位，一个生物体的全部遗传信息都表现在基因组中。要想详细了解 DNA 编码蛋白质的情况，以及 DNA 与基因的关系等等，就必须首先弄清楚 DNA 核酸序列的整体结构。基因组 DNA 测序是对一个生物体基因组认识的第一步。全基因组测序是对未知基因组序列的物种进行个体的基因组测序。De novo 测序也叫从头测序，不需要任何基因序列信息即可对某个物种进行测序，是指对基因组序列未知或没有近源物种基因组信息的某个物种，对其不同长度基因组 DNA 片段及其文库进行序列测定，然后用生物信息学方法进行拼接、组装和注释，从而获得该物种完整的基因组序列图谱。Illumina 测序平台是目前应用最为广泛的测序平台，通量大且可控制，数据精确，操作简单、自动化，所需样品少，能够在极短时间内获得数十亿高精确度的碱基序列信息，是目前具有较强实用性的一种测序技术。近年来，De novo 测序技术广泛应用

收稿日期：2015-01-27
基金项目："十二五"国家科技支撑项目（2012BAD37B01）；广东省科技计划项目（E8140475）；国家 973 计划项目（2012CB720800）；国家自然科学基金青年基金项目（31201362）；广东省优秀博士学位论文作者资助项目（K3140030）；中央高校基本科研业务费面上项目（2012ZM0060）
作者简介：刘君彦（1992-），女，硕士研究生，研究方向：食品微生物安全
通讯作者：徐振波（1982-），男，博士，讲师，研究方向：食源性微生物安全控制与致毒机理研究；邓阳（1986-），男，博士，助理研究员，研究方向：食品微生物及生物技术

于微生物的基因组研究中。

耐酸乳杆菌是一种典型的难培养啤酒易感微生物，其代谢产物能够损害啤酒的风味和口感，并能通过形成混浊、沉淀等使啤酒外观发生变化。作为啤酒易感微生物，该菌最大的特点是首次检测时间长，在常规乳酸菌检测培养基 MRS（de Man Rogosa Sharpe）上生长非常缓慢，需 14 d 以上才肉眼可见。耐酸乳杆菌属乳杆菌属，其细胞呈杆状，菌落呈不规则圆形、粗糙微凸起、不透明，可承受 4~5%的乙酸浓度，能在 pH 3.3~6.6 之间于 23~40 ℃范围内生长，低于 15 ℃不生长，属于兼性厌氧菌，革兰氏阳性菌，同型发酵型，过氧化氢酶阴性、氧化酶阴性。[1~3]迄今为止，关于耐酸乳杆菌的相关研究报道仍不够深入，仍停留在分离鉴定的程度，其全基因组遗传信息仍然未知状态。

本文以啤酒中分离获得的一株乳杆菌为研究对象，通过 De novo 测序技术对其进行全基因组测序，分析序列信息，获得其全基因组遗传信息，并对所获得的基因信息进行功能分析，确定各基因的功能。通过基因组学建立乳杆菌基因信息库，为该啤酒易感乳杆菌功能基因的研究与生物信息学分析提供基础数据和技术参考。

## 1 材料与方法

### 1.1 实验材料

乳杆菌（BM-LA14527），分离自啤酒；M.R.S 培养基，OXOID LTD.，BASINGSTOKE，HAMPSHIRE，ENGLAND；溶菌酶、RNA 酶、细菌基因组 DNA 快速提取试剂盒、凝胶回收试剂盒，广州东盛生物科技有限公司；Gel Extraction Kit 试剂盒，TaKaRa 公司。

### 1.2 基因组 DNA 的提取

分别取 2 mL 处于对数生长期乳杆菌的新鲜培养物，12,000 r/min 离心 1 min，弃上清，向菌体沉淀中加入 20 μL 溶菌酶溶液（20 mg/mL）悬浮细胞，于 37 ℃处理 30 min 以上，使细胞裂解释放基因组 DNA，然后通过加入 RNA 酶去除 RNA，再经蛋白酶、去蛋白液和漂洗液的作用，将蛋白、脂质等杂质洗脱，以获得高质量基因组 DNA。

### 1.3 菌株种属鉴定

采用 16S 测序法对实验菌株进行鉴定。首先根据乳杆菌的 *16Sr DNA* 基因序列设计上游引物 16S1（5’-AGAGTTTGATCCTGGCTCAG-3’）和下游引物 16S2（5’-CTACGGCTACCTTGTTACGA-3’），进行 PCR 扩增反应。PCR 反应条件为 94 ℃预变性 3 min，然后 94 ℃ 30 s、50 ℃ 30 s、72 ℃ 40 s 进行 30 个循环，最后 72 ℃延伸 10 min。PCR 产物经 1%琼脂糖凝胶电泳，于凝胶成像系统下观察并拍照记录结果。PCR 产物通过电泳，利用 Gel Extraction Kit 试剂盒割胶回收纯化，将回收的片段进行 16S 测序，测序引物为 16S PCR 扩增引物。得到 *16Sr DNA* 基因片段序列后，进行 blast 比对。

### 1.4 基因组 DNA 的纯化

将获得基因组 DNA 进行核酸电泳验证，并进行胶回收纯化 DNA。在短波 360 nm 光盒照射下，快速切取含有目的 DNA 片段的琼脂糖凝胶条带，55 ℃~65 ℃水浴至胶全部融化，降至室温后将溶液加入 DNA 纯化柱中，将蛋白、盐等杂质洗脱后，获得高质量的 DNA 片段，实现基因组 DNA 的纯化。

### 1.5 测序文库的构建

获得纯化的高质量乳杆菌基因组 DNA 后，需要对其进行测序文库的构建。首先对样品质量进行检测，主要采用 1%琼脂糖凝胶电泳、紫外分光光度计检测样品的浓度与纯度并扩增 16S 全长序列进行验证。然后用检测合格的样品构建文库：首先采用超声法 Covaris 或者 Bioruptor 将大片段 DNA（如基因组 DNA、BAC 或长片段 PCR 产物）随机打断并产生主带小于等于 800 bp 的一系列 DNA 片段，然后用 T4 DNA Polymerase、Klenow DNA Polymerase 和 T4 PNK 将打断形成的粘性末端修复成平末端，再通过 3' 端加碱基“A”，使得 DNA 片段能与 3'端带有“T”碱基的特殊接头连接，用电泳法选择需回收的目的片段连接产物，再使用 PCR 技术扩增两端带有接头的 DNA 片段并进行 PCR 产物纯化，即获得构建好的文库；对文库进行质量检测，检测合格后即可用合格的文库进行 cluster 制备和上机测序。

### 1.6 Illumina 测序与 De novo 组装

测序文库构建完成后，进行上机测序。本研究所采用的是 Illumina 测序技术，其采用的 HiSeq 2000 测序系统，其测序读长达到 100 个碱基，单次运行可产生 600 GB 的数据，是目前通量较高的测序系统之一。Illumina 测序通量大且可控制，数据精确，操作简单、自动化，所需样品少，能够在极短时间内获得数十亿高精确度的碱基序列信息，最大的缺点是其读长短，不适用于大基因组的 de novo 测序，但是正适用细菌

这种较小基因组的测序。Illumina 测序的原理是“边合成边测序”，其过程为：首先将基因组 DNA 打碎成约 100~200 bp 的小片段，在片段的两个末端连接上特定的 DNA 接头（adapter）；将 DNA 片段变成单链后通过与芯片表面的引物碱基互补的一端被固定在芯片上，另外一端随机和附近的另外一个引物互补，也被固定住，从而形成桥状结构；进行桥型扩增，所有单链桥型待测片断被扩增成为双链桥片断，通过扩增反应，将会获得待测的上百万条 DNA 簇；加入 DNA 聚合酶和被荧光标记的 dNTP 和接头引物进行扩增，在 DNA 合成过程中，每一个核苷酸加到引物末端时都会释放出焦磷酸盐，激发生物发光蛋白发出荧光，测序仪通过捕获、采集、统计荧光信号，就可以得知每个模板 DNA 片段的序列[4,5]。

获得测序数据，首先要对原始的测序数据进行一系列预处理。对测序得到的原始数据（reads）进行质量评价、过滤及统计，去除接头序列、多 N 序列及质量值过低的序列，得到过滤后的数据（clean reads），并将过滤后的数据进行质量评价，本次质量评价所用的工具为常用的 FastQC 软件。

序列组装是高通量测序数据处理中的一个非常重要的环节。前期的测序环节，采用的是将全基因组 DNA 打断成小片段，再对小片段分别进行测序，所以所得的测序结果是许多小片段的基因序列，需要对过滤后的数据经过一系列手段进行 de novo 组装，得到 contig 及 scaffold 序列的 fasta 文件，最终得到全基因组序列。一般需要根据不同测序平台选择最适合软件进行初步 de novo 组装（454 或 Ion Torrent 平台采用 OLC 算法编写的软件，Illumina、Solid、Sanger 等采用 DBG 算法的软件），本研究采用的 velvet 短序列组装软件（http://www.ebi.ac.uk/~zerbino/velvet/；版本：1.2.08）。Velvet 是由欧洲生物信息中心(EMBL-EBI)开发的一个软件，主要用于拼接测序长度短的序列，例如 Illumina、Solid 测得的序列。Velvet 是目前广泛使用的拼接短 reads 的首选拼接工具，非常适合于拼接细菌、病毒等基因组。数据组装过程一般主要分为以下几个步骤：首先利用所有 reads 的 K-mer 序列构建 De Brujin 图；其次利用处理序列的前后关系，将所有 overlap 的序列组装成 contig 序列集，并调整主要参数，得到最优的组装结果；最后是构建 scaffold 序列，即把 reads 比对到 contig 上，根据 reads 的 paired-end 和 overlap 关系，统计覆盖到不同 contig 序列上的成对的 reads 信息，构建 scaffold 序列。在 scaffold 序列构建过程中，主要是对其中的一些片段进行 blast 比对，寻找是否存在同源性高的参考序列。如果有同源性很高的参考基因组序列，那么可以利用参考基因组序列做参考，组装获得一系列片段，并和 de novo 组装得到的片段相互填补 gap。如果没有同源性高的参考序列，可以使用 Mauve 软件（http://gel.ahabs.wisc.edu/mauve/）对 de novo 组装得到的 contig 序列进行定位，然后使用末端延伸测序进行填补 gap。当然在对同一个全基因组序列组装的过程中，也有部分序列有同源性高的参考序列，而部分序列没有同源性高的参考序列的情况，这种情况下则可以综合使用这两种方法。若用现有的方法数据无法再进行填补 gap，则对可以通过设计引物测序、采用不同的高通量测序平台再次测序等方法完成。[6,7]

整个过程中非常重要的组装环节完成后将进行对组装结果的评价，影响组装最大的数据指标是 reads 的长度和数据量。当测序的 reads 越长且数量越多时，reads 间的重叠部分就越多，在不考虑测序质量的情况下得到的 contigs 就越长。目前所有第二代测序平台中，单次运行产生数据量最大的是本研究所用的 Illumina 测序平台。

首先统计 scaffold 序列及 contig 序列的数量及 N50、N90 来评估组装结果的质量。其中 N50、N90 是用于评价组装结果好坏的一个标准。Read 组装后会获得一系列长短不同的片段，将组装后的片段按照从长到短排序，并按照这个顺序将序列长度依次相加，当相加长度达到总长度的一半时，最后一个加上序列的长度即为 N50，相加长度为总长度的 90%时，最后一个加上序列的长度即为 N90。为对组装结果进行进一步评估，利用测序 reads 与组装结果比对，计算组装结果 Depth 值和 GC 含量，进而反应 GC 含量与测序深度的分布，对组装结果进行质控。以 500bp 为窗口无重复的计算组装基因组的 GC 含量和平均深度，可以根据 GC-depth 图分析测序数据是否存在 GC 偏向性。最后进行组装覆盖度统计，将 read 与组装的 scaffold 序列进行 BWA 比对分析，得到比对上的 reads 数目与比对上的 scaffold 序列的长度，进而估计整个基因组的覆盖度，覆盖度=比对上的长度之和/组装结果长度。

### 1.7 基因预测

目前基因预测的方法主要有 3 种，第一种是直接分析法，即分析 mRNA 和 EST 数据直接得到结果；第二种是相似性比对法，即通过相似性比对从已知基因和蛋白质序列得到间接证据；第三种是统计模型法，即基于各种统计模型和算法从头预测，比如隐马可夫（Hidden Markov Model，HMM）模型。其中比较常

用的是相似性比对法。现在常用的做法是先通过 SNAP、Glimmer、Genscan 等软件预测出基因组的开放阅读框，然后将预测出的开放阅读框与其他近缘物种的基因组进行 blast 比对，比对出有同源基因的开放阅读框被注释为同样的功能，没有同源基因的开放阅读框则被舍去或注释为假说蛋白（hypothetical protein）[8]。

本研究采用 SNAP 软件（2013-11-29 版本），以其近缘物种 coprinopsis cinerea 基因组作为训练集，采用 HMM 模型从组装结果中预测基因序列，用于后续的基因功能注释等分析。

### 1.8 基因功能分析

基因功能分析是利用生物信息学方法，对基因组所有基因的生物学功能进行功能注释。主要的基因功能分析方法有基因本体论（Gene Ontology，GO）注释、直系同源序列聚类分析（Cluster of Orthologous Group，COG）、京都基因和基因组百科全书（Kyoto Encyclopedia of Genes and Genomes）生物学通路分析等。

#### 1.8.1 GO 功能分析

GO 基因注释主要基于蛋白序列比对，GO 注释的数据库为 NR 数据库，分析所用软件为 Blast2GO。NR 数据库即所有非冗余的 GenBank CDS 区的翻译序列、参考序列的蛋白、PDB 数据库、SwissProt 蛋白数据库与 PRF 蛋白数据库的整合体。GO 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，它常用于提供基因功能分类标签和基因功能研究的背景知识，通过物种和基因信息，用 GO 数据库进行查找，从而得到基因的 GO 注释信息，即基因的一部分功能信息。[9]根据 GO 数据库的注释信息，对预测基因进行蛋白序列比对，从分子功能、生物过程与细胞组成三个方面获得基因的功能信息。

#### 1.8.2 COG 功能分析

蛋白质直系同源簇（COGs）数据库是对细菌、藻类和真核生物的 15 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 库是一个预测单个蛋白质的功能和整个新基因组中蛋白质的功能的有价值的平台。[10]对本研究根据 COG 数据库中的蛋白信息，对预测基因进行 COG 功能分析，注释相关的基因编码蛋白信息。

#### 1.8.3 KEGG 生物通路分析

KEGG 是系统分析基因功能、联系基因组信息和功能信息的知识库，包括一些不同功能的数据库。其中基本基因组信息存储在 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、细胞周期等，还包括同系保守的子通路等信息；LIGAND 数据库则包含关于化学物质、酶分子、酶反应等信息。KEGG 提供了 Java 的图形工具来访问基因组图谱，比较基因组图谱和操作表达图谱，以及其他序列比较、图形比较和通路计算的工具。[11]本研究将预测的基因进行基于 KEGG 数据库(http://www.genome.jp/)的生物通路富集分析，提取出相关的生物通路上的基因。

## 2 结果与分析

### 2.1 菌株鉴定结果

PCR 扩增产物经电泳检测后条带如图 1 所示，比对结果显示，测序结果与耐酸乳杆菌的 *16Sr DNA* 基因序列之间存在 100%同源性。以上结果说明，实验菌株为耐酸乳杆菌。

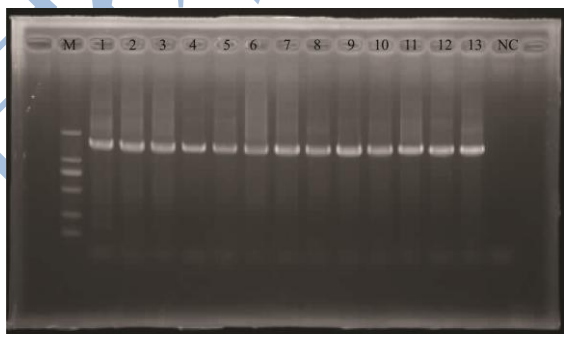



图 1 试验菌株 16Sr DNA 基因扩增图谱

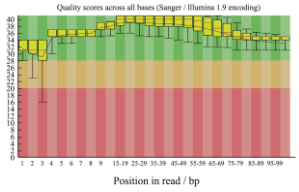
**Fig.1 Determination for presence of *16Sr DNA* gene**

注：M：DNA Marker(DS$^{TM}$2000)；1-13：实验菌株；NC：空白对照

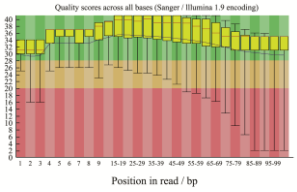
### 2.2 测序质量评价

测序质量评价决定了测序结果的可信度，其指标有碱基质量（Per base sequence quality）、序列质量（Per sequence quality scores）、碱基含量（Per base sequence content）、碱基 GC 含量（Per base GC content）、序列 GC 含量（Per sequence GC content）、未知碱基含量（Per base N content）、序列长度分布（Sequence Length Distribution）、序列重复水平（Sequence Duplication Levels）等。一、测序碱基质量（Per base sequence quality），是测序质量评价的一个重要指标，碱基质量的计算公式为：$Q=-10\times\log_{10}(p)$，p 为测错的概率。即当一条 reads 某位置出错概率为 0.01 时，其碱基质量就是 20。如图 2a，横轴代表 reads 位置，纵轴代表碱

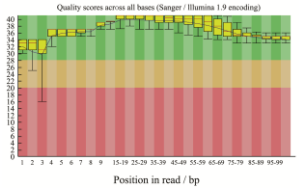
基质量。红线表示中位数，黄色区域是 25%~75%区间，触须是 10%~90%区间，蓝线是平均数。在测序质量检测过程中，若任一位置的下四分位数（位于 25%）低于 10 或中位数低于 25，报"WARN"；若任一位置的下四分位数低于 5 或中位数低于 20，报"FAIL"。本次测序的下四分位数与中位数均合格，且均高出标准值较高水平，可知重测序碱基质量较高，测序结果可信度较高。二、测序序列质量（Per sequence quality scores），如图 2b，横轴代表序列质量，纵轴代表 reads 数目。当峰值小于 27（错误率为 0.2%）时报"WARN"，当峰值小于 20（错误率为 1%）时报"FAIL"。本次测序的峰值均出现在 38，大于 27，序列质量较高，测序结果可信度较高。三、碱基含量（Per base sequence content），对所有 reads 的每个位置，统计 ATCG 四种碱基的分布。如图 2c，横轴代表 reads 位置，纵轴代表百分比。正常情况下四种碱基的出现频率应该是接近的，且没有位置差异，因此质量高的样本中四条线应该平行且接近。当部分位置碱基的比例出 bia 时，图中代表四种碱基的四条线就会在某些位置交错，提示我们有 overrepresented sequence 的污染；当所有位置的碱基比例一致地表现出 bias 时，即图中四条线平行但分开，往往代表文库本身或在建库过程中有 bias，或者是测序过程中的系统误差。当任一位置的 A/T 比例与 G/C 比例相差超过 10%，报"WARN"；当任一位置的 A/T 比例与 G/C 比例相差超过 20%，报"FAIL"。本次测序原始数据 read2 中该比例相差超过 20%，报"FAIL"，原始数据 read1 与过滤后数据的该比例超过 10%，报"WARN"。但这个指标不能代表本次测序整体的趋势，碱基含量分布只是在测序初期，在测序读长的最开始时出现了偏差，后期出现频率比较接近且较稳定，所以本次测序的碱基含量分布整体上还是比较均匀，测序结果比较可信。四、碱基 GC 含量（Per base GC content），对所有 reads 的每个位置，统计 GC 含量。如图 2d，横轴代表 reads 位置，纵轴代表百分比，红线代表碱基 GC 含量。如果建库足够均匀，reads 的每个位置是没有差异的，此时红线应该平行于横轴；当部分位置碱基 GC 含量出现 bias 时，提示我们有 overrepresented sequence 的污染；当所有位置的碱基 GC 含量一致的表现出 bias 时，则代表文库本身或在建库过程中有 bias，或者是测序中的系统误差。当任一位置的碱基 GC 含量偏离均值的 5%时，报"WARN"；当任一位置的碱基 GC 含量偏离均值的 10%时，报"FAIL"。本次测序中碱基 GC 含量在 40%左右，在测序初期有些许摆动，但并未偏离均值的 5%，所以本次测序的碱基 GC 含量质量较高，测序结果的可信度较高。五、序列 GC 含量（Per sequence GC content），统计 reads 的平均 GC 含量的分布。如图 2e，横轴代表平均 GC 含量，纵轴是数量，红线是实际分布，蓝线是理论分布（呈正态分布，但均值不一定在 50%，而是由平均 GC 含量推断的，本图中均值为 40%左右）。若曲线有形状上的偏差，是由于文库的污染或是部 reads 构成的子集有偏差（overrepresented reads）；若形状接近正态分布但偏离理论分布的情况，提示我们可能有系统偏差。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"。本次测序的实际分布与理论分布十分接近，说明本次测序的可信度较高。六、未知碱基含量（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”，即未知碱基。对所有 reads 的每个位置，统计未知碱基含量。如图 2f，横轴代表 reads 的位置，纵轴代表百分比。正常情况下未知碱基含量是很小的，所以图上常常看到一条接近横轴的直线，但放大纵轴后会发现还是有未知碱基的存在，这是正常的。当纵轴在 0~100%的范围内也能看到凸起时，说明测序系统出了问题。当任意位置未知碱基的含量超过 5%时，报"WARN"；当任意位置的未知碱基的含量超过 20%时，报"FAIL"。本图中纵轴在 0~100%范围内并未看到凸起，且未知碱基的含量未超过 5%，说明未知碱基含量正常，测序结果可信度较高。七、序列长度分布（Sequence Length Distribution），统计序列长度的分布。如图 2g，横轴代表序列长度，纵轴代表数量。当序列长度不一致时报"WARN"；当有长度为 0 的 read 时报“FAIL”。本次结果显示序列长度分布较一致，集中在 100 bp 左右，表明测序结果可信度较高。八、序列重复水平（Sequence Duplication Levels），统计序列完全一样的 reads 的频率。测序深度越高，越容易产生一定程度的重复，这是正常的现象，但如果重复的程度很高，就提示我们可能有 bias 的存在。如图 2 h，横坐标是重复的次数，纵坐标是重复序列的数目百分比，以 unique reads 的总数作为 100%。当非 unique 的 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"。本图中显示，非 unique 的 reads 占总数的 80%以上，报"FAIL"。本次实验中由于测序深度的增加，使重复序列的数量增多。序列重复水平只是众多测序质量评价的指标之一，只能作为参考，并不能因为重复序列较多而否定整个测序结果，由于前面各指标都表明本次测序结果可信度较高，所以依然可以认为本次测序结果的可信度较高。综上质量评价可知，本次测序结果可信度较高。



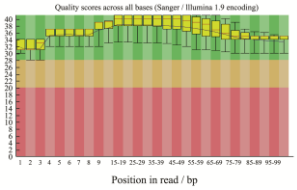

A




B




C

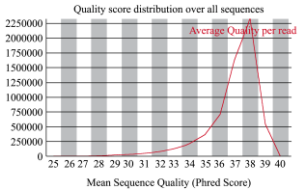



D

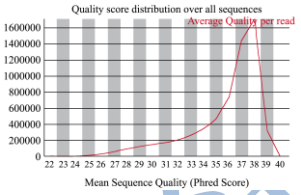
**a. Per base sequence quality**


A


B

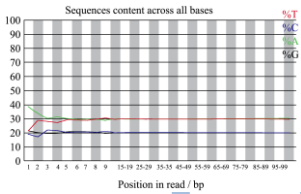



C

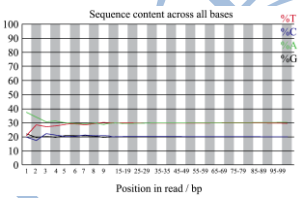



D

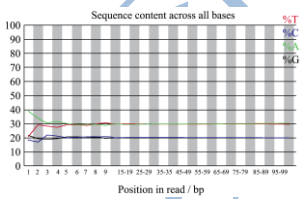
**b. Per sequence quality scores**



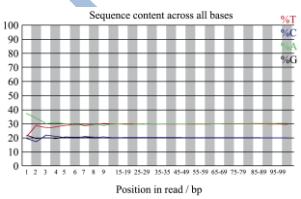

A




B

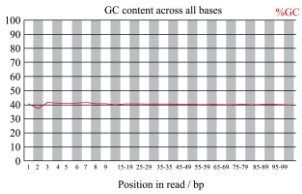



C

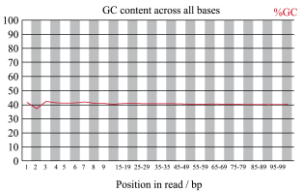



D

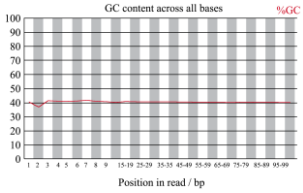
**c. Per base sequence content**



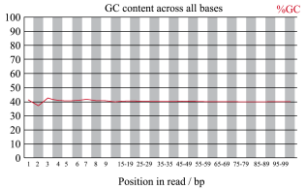

A




B

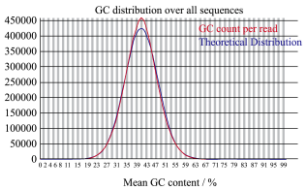



C

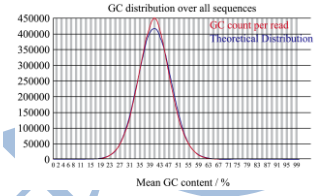



D

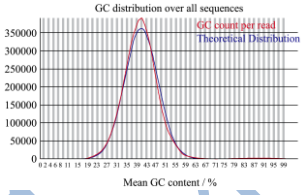
**d. Per base GC content**



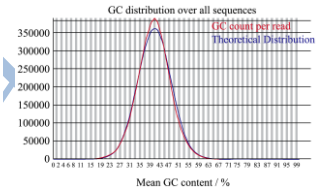

A




B

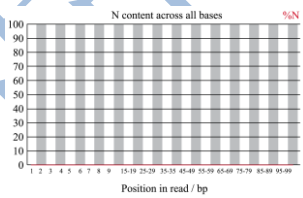



C

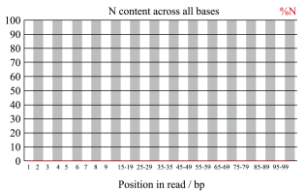



D

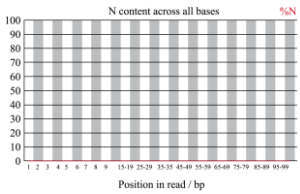
**e. Per sequence GC content**



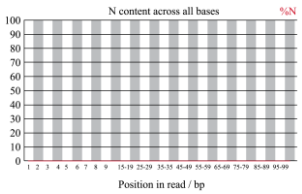

A




B

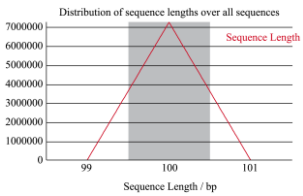



C

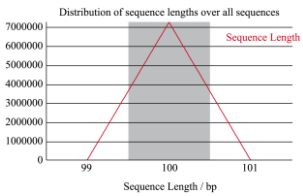



D

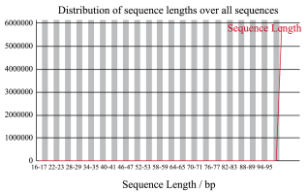
**f. Per base N content**



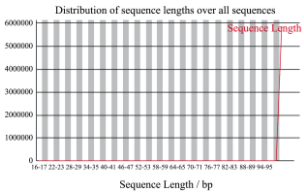

A




B




C




D

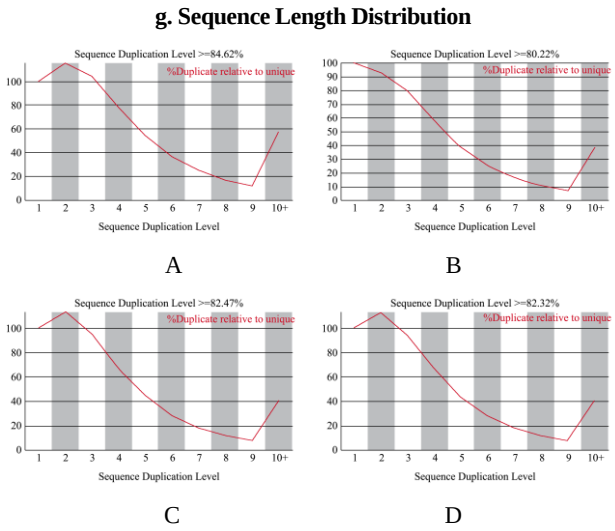



图 2 测序质量评价各指标汇总图

**Fig.2 Sequence quality evaluate index summarize**

注：图中A、B为原始数据，C、D为过滤后数据。

## 2.3 组装结果

### 2.3.1 数据组装分析

本次实验统计了长度大于 500 bp 的 scaffold 序列及 contig 序列，结果见表 1。从表中可以看出，scaffold 序列的数目为 100 条，N50 长度约 32 kb，N90 长度约 9 kb，说明组装结果较好。

表 1 组装结果统计表

| 项目 | Scaffold | Contig |
|---|---|---|
| 总数目 | 100 | 182 |
| 总长度 | 1,542,492 | 1,537,201 |
| N50 | 32,098 | 17,914 |
| N90 | 9,105 | 4,401 |
| 最大长度 | 114,884 | 59,278 |
| 最小长度 | 503 | 503 |
| 序列 GC 含量/% | 36.35 | 36.47 |

### 2.3.2 质控统计及组装覆盖度统计

本次结果显示，经过组装分析之后，序列已经去除污染。组装覆盖度统计如表 2 所示，组装序列总长度为 1,542,493 bp，覆盖长度高达 1,541,178 bp，覆盖率为 99.91%，说明组装结果较好。

表 2 组装覆盖度统计

**Table 2 Assembly coverage**

| 组装结果 | 覆盖长度 | 覆盖度/% | 平均覆盖深度 |
|---|---|---|---|
| 1,542,493 | 1,541,178 | 99.91 | 743.37 |

## 2.4 基因预测

本研究中预测基因的长度从 9 bp~49,997 bp，基因预测的结果统计见表 3，由表中可以看出，预测基因总数为 1,824 个，基因总长度为 1,592,677 bp。以 1,000 bp 为一个长度单位，将基因长度分为 11 个区间，统计预测基因的长度分布及比例，如图 4 所示，横轴代表基因长度，纵轴代表数量。由图 3 可以看出，长度为 100~200 bp 的基因最多，且随后的变化规律大概为基因数量随着基因长度的增加而减少。

表 3 预测基因统计表

**Table 3 Gene prediction**

| 项目 | 数据 |
|---|---|
| 基因数目 | 1,824 |
| 基因长度 | 1,592,677 |
| 基因 GC 含量 | 37.22 |
| 基因平均长度 | 873 |

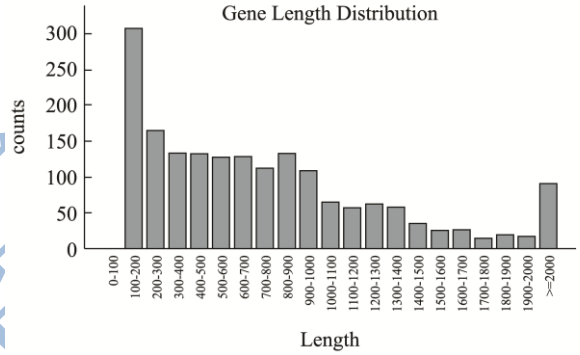



图 3 预测基因长度分布

**Fig.3 The length distribution of predicted gene**

## 2.5 基因功能注释

### 2.5.1 GO 功能注释

本次分析中共有 545 条预测基因有相应的 GO 基因功能注释。对基因数目大于 20 的 GO 类别进行汇总，结果显示分子功能注释中，参与结合 ATP 的基因占的比例较高，数目约为 260 个，其余类别按比例从高到低有结合 DNA、结合金属离子、核糖体的结构组分、水解酶活性、结合 RNA、ATP 酶活性等；生物过程注释中，参与氧化还原过程的基因占得比例最高，数目约为 90 个，其他为新陈代谢过程、翻译过程、磷酸化作用、蛋白质水解作用、ATP 分解代谢过程、转运过程、细胞分裂过程等；细胞组成注释中，参与组成细胞质的基因比例最高，数目约为 200 个，其余比例较高的细胞组成有质膜、核糖体、细胞膜等。本研究获得了耐酸乳杆菌一系列的 GO 基因功能注释，从分子功能、生物过程和细胞组成三个方面明确了乳杆菌基因发挥的各种功能作用，为对乳杆菌的深入研究提供了基础数据与理论依据。

### 2.5.2 COG 功能注释

本次分析中共有 1,303 条预测基因有相应的 COG 功能注释。对各功能类别的基因数目进行统计，其中基因数目较多的类别主要有通用功能预测、翻译、各种结构的生物转化、复制、重组和修复、翻译、转录、各种物质的转运与代谢、各种机制的发生等。本研究获得了乳杆菌一系列的 COG 基因功能注释，为乳杆菌功能基因的深入研究提供了基础数据与理论依据。

### 2.5.3 KEGG 生物通路注释

本次分析中共有 120 条预测基因有相应的 KEGG 生物通路注释，并统计了富集基因最多的 20 类 KEGG 生物通路，其中有约 50 条基因与核糖体代谢有关，其余富集基因较多的 KEGG 生物通路如下：嘌呤代谢、嘧啶代谢、氨基糖与核苷酸糖代谢、肽聚糖生物合成、DNA 复制、磷酸戊糖途径、果糖和甘露糖代谢、脂肪酸生物合成、细胞周期、糖酵解、丙酮酸代谢、丙氨酸、天冬氨酸、谷氨酸代谢、碱基切除修复等。本研究获得了乳杆菌一系列的 KEGG 生物通路注释，明确了乳杆菌基因参与的各类代谢情况，为对乳杆菌全基因组信息的深入研究提供了基础数据与理论依据。

### 2.5.4 全基因组功能分析

通过对该乳杆菌进行全基因组研究，结合预测基因的 GO 功能注释、COG 功能注释及 KEGG 生物通路注释，并进行基因对比，发现该乳杆菌存在能够编码依赖于 ATP 的多药转运蛋白的基因，该转运蛋白能将苦味酸排除细胞，使细菌能够在啤酒中的无氧或微氧环境下生长[12]。此基因的存在能够引起啤酒中双乙酰和有机酸含量显著升高和导致啤酒浑浊。在啤酒行业中，消费者评价啤酒质量的标准是口味新鲜、酒体澄清、泡沫丰富以及色度适中。由于存在编码排出苦味酸蛋白的基因，该乳杆菌能使啤酒酸度增加并导致啤酒浑浊，从而导致啤酒风味改变、酒体浑浊，严重影响啤酒产品质量，最终会造成重大的经济损失。该啤酒易感乳杆菌的全基因组信息为啤酒中微生物安全控制提供了理论基础，可根据该基因设计相应的解决方案，控制啤酒的微生物安全。

## 3 结论

本研究通过对 1 株分离于啤酒的乳杆菌进行 De novo 测序，获取其全基因组序列，进行基因预测后将预测基因进行功能注释，获得一系列基因功能信息，为乳杆菌功能基因的研究与生物信息学分析提供基础数据和技术参考。同时根据基因功能信息，筛选出造成啤酒腐败变质的相关基因，下一步将选用不同状态的乳杆菌进行转录组学研究与蛋白组学研究，从而在转录水平与蛋白水平上阐述乳杆菌导致啤酒腐败变质的分子机制，为啤酒微生物安全控制提供理论基础和科学依据。

## 参考文献


[1] Entani E, Masai H, Suzuki K-I. *Lactobacillus acetotolerans*, a new species from fermented vinegar broth [J]. Int. J. Syst. Bacteriol., 1986, 36: 544-549

[2] Deng Y, Liu J, Li H, et al. An improved plate culture procedure for the rapid detection of beer-spoilage lactic acid bacteria [J]. Journal of the Institute of Brewing, 2014, 120(2): 127-132

[3] 邓阳,刘君彦,房慧婧,等.VBNC 状态啤酒易感乳杆菌的诱导及复苏[J].现代食品科技,2014,30(4):154-159
DENG Yang, LIU Jun-yan, FANG Hui-jing, et al. Induction and resuscitation of vbnc state beer-spoilage *Lactobacilli* [J]. Modern Food Science and Technology, 2014, 30(4): 154-159

[4] Mardis E R. The impact of next-generation sequencing technology on genetics [J]. Trends in Genetics, 2008, 24: 133-141

[5] Shendure J, Ji H. Next-generation DNA sequencing [J]. Nature Biotechnology, 2008, 26: 1135-1145

[6] Zerbino D R, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs [J]. Genome Research, 2014, 18: 821-829

[7] Guistini D S, Liao N Y, et al. Platt D. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data [J]. Genome Biology, 2009, 10(9): R94

[8] 禹胄,李涛,蔡涛,等.微生物基因组注释系统 MGAP[J].微生物学报,2003.43(6):805-808
YU Zhou, LI Tao, CAI Tao, et al. The annotation system of microorganism genome, MGAP [J]. Acta Microbiologica Sinica, 2003. 43(6): 805-808

[9] Ashburner M, Ball C, Blake J, *et al*. Gene ontology: tool for the unification of biology [J]. Nature Genetics, 2000, 25(1): 25-29

[10] Tatusov B, Fedorova N, Jackson J, et al. The COG database: an updated version includes eukaryotes [J]. BMC Bioinformatics, 2003, 4:41

[11] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes [J]. Nucleic Acids Research, 2000, 28(1): 27-30

[12] Juvonen R, Satokari R. Detection of spoilage bacteria in beer by polymerase chain reaction [J]. Journal of the American Society of Brewing Chemists, 1999, 57: 99-103