

# 醋酸菌中 CRISPR 位点的比较基因组学与进化分析

夏凯, 梁新乐, 李余动

浙江工商大学食品与生物工程学院, 杭州 310018

**摘要:** CRISPR (Clustered regularly interspaced short palindromic repeats)是近几年发现的一种广泛存在于细菌和古菌中, 能够应对外源 DNA 干扰(噬菌体、病毒、质粒等), 并提供免疫机制的重复序列结构。CRISPR 系统通常由同向重复序列、前导序列、间隔序列和 CRISPR 相关蛋白组成。本研究以醋酸发酵中常见 3 个属醋杆菌属 (*Acetobacter*)、葡糖醋杆菌属(*Gluconacetobacter*)和葡糖杆菌属(*Gluconobacter*)的 48 个菌株为研究对象, 通过其基因组上 CRISPR 相关基因序列的生物信息学分析, 探索 CRISPR 位点在醋酸菌中的多态性及其进化模式。结果表明 48 株醋酸菌中有 32 株存在 CRISPR 结构, 大部分 CRISPR-Cas 结构属于 type I-E 和 type I-C 类型。除了葡糖杆菌属外, 葡糖醋杆菌属和醋杆菌属中的部分菌株含有 II 类的 CRISPR-Cas 系统结构(CRISPR-Cas9)。来自不同属菌株的 CRISPR 结构中重复序列具有较强的保守性, 而且部分菌株 CRISPR 结构中的前导序列具有保守的 motif (与基因的转录调控有关)及启动子序列。进化树分析表明 *casI* 适合用于醋酸菌菌株的分类, 而不同菌株间 *casI* 基因的进化与重复序列的保守性相关, 预示它们可能受相似的功能选择压力。此外, 间隔序列的数量与噬菌体数量及插入序列(Insertion sequence, IS)数量有正相关的趋势, 说明醋酸菌在进化过程中可能正不断受新的外源 DNA 入侵。醋酸菌中 CRISPR 结构位点的分析, 为进一步研究不同醋酸菌株对醋酸胁迫耐受性差异及其基因组稳定性的分子机制奠定了基础。

**关键词:** 醋酸菌; CRISPR; 重复序列; 遗传进化

## Comparative genomics and evolutionary analysis of CRISPR loci in acetic acid bacteria

Kai Xia, Xinle Liang, Yudong Li

College of Food and Biotechnology, Zhejiang Gongshang University, Hangzhou 310018, China

**Abstract:** The clustered regularly interspaced short palindromic repeat (CRISPR) is a widespread adaptive immunity system that exists in most archaea and many bacteria against foreign DNA, such as phages, viruses and plasmids. In general, CRISPR system consists of direct repeat, leader, spacer and CRISPR-associated sequences. Acetic acid bacteria (AAB) play an important role in industrial fermentation of vinegar and bioelectrochemistry. To investigate the polymorphism and evolution pattern of CRISPR loci in acetic acid bacteria, bioinformatic analyses were performed on 48 species from three main genera (*Acetobacter*, *Gluconacetobacter* and *Gluconobacter*) with whole ge-

收稿日期: 2015-06-01; 修回日期: 2015-09-07

基金项目: 国家自然科学基金(编号: 3117175)和浙江省自然科学基金(编号: LY14C060001)资助

作者简介: 夏凯, 硕士研究生, 专业方向: 醋酸菌分子生物学。E-mail: xiakai333@126.com

通讯作者: 李余动, 博士, 副教授, 研究方向: 微生物基因组学。E-mail: lyd@zjsu.edu.cn

DOI: 10.16288/j.ycz.15-244

网络出版时间: 2015-10-8 15:57:12

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20151008.1557.002.html>

nome sequences available from the NCBI database. The results showed that the CRISPR system existed in 32 species of the 48 strains studied. Most of the CRISPR-Cas system in AAB belonged to type I CRISPR-Cas system (subtype E and C), but type II CRISPR-Cas system which contain *cas9* gene was only found in the genus *Acetobacter* and *Gluconacetobacter*. The repeat sequences of some CRISPR were highly conserved among species from different genera, and the leader sequences of some CRISPR possessed conservative motif, which was associated with regulated promoters. Moreover, phylogenetic analysis of *cas1* demonstrated that they were suitable for classification of species. The conservation of *cas1* genes was associated with that of repeat sequences among different strains, suggesting they were subjected to similar functional constraints. Moreover, the number of spacer was positively correlated with the number of prophages and insertion sequences, indicating the acetic acid bacteria were continually invaded by new foreign DNA. The comparative analysis of CRISPR loci in acetic acid bacteria provided the basis for investigating the molecular mechanism of different acetic acid tolerance and genome stability in acetic acid bacteria.

**Keywords:** acetic acid bacteria; CRISPR; repeat sequence; genetic evolution

醋酸菌隶属于  $\alpha$  变形杆菌纲, 红螺菌目, 醋酸杆菌科<sup>[1]</sup>。醋酸菌一般可以从水果、花卉以及发酵食品中分离获得, 由于其能在乙醇脱氢酶和乙醛脱氢酶的作用下将乙醇氧化为乙酸, 因而在醋的工业化生产过程中扮演着非常重要的角色<sup>[2]</sup>。截止到目前, 醋酸菌主要包括 12 个属, 其中研究最多的是醋酸杆菌(*Acetobacter*)、葡糖醋杆菌(*Gluconacetobacter*)及葡糖杆菌(*Gluconobacter*)。醋酸杆菌和葡糖醋杆菌拥有较高的耐酸性因而被广泛地用于醋的发酵<sup>[3]</sup>。醋酸菌容易受到外源 DNA 的干扰, 特别是噬菌体的侵染。虽然噬菌体在基因水平转移和遗传多样性方面发挥着重要作用, 但在实际的发酵生产过程中, 噬菌体的侵染将会造成巨大的经济损失<sup>[4]</sup>。另外, 噬菌体的侵染也造成了醋酸菌基因组的不稳定, 进而影响到菌种的性能<sup>[5]</sup>。为了抵御来自外源 DNA 的干扰, 特别是噬菌体、病毒以及质粒, 自然界的细菌在进化过程中形成了一系列的防御机制, 包括流产性感染、表面排斥以及限制性调节系统等<sup>[6]</sup>。CRISPR (Clustered regularly interspaced short palindromic repeats)是近几年发现的一种广泛存在于细菌和古菌中, 能够对外源 DNA 干扰并提供免疫机制的重复序列结构<sup>[7]</sup>。因此, 醋酸菌中 CRISPR 结构位点的分析对于研究醋酸菌耐酸能力差异的遗传机制具有重要的意义<sup>[8]</sup>。

CRISPR 系统发挥作用主要分为 3 步: 第一步是适应阶段, 通过将外源核酸片段整合到 CRISPR 位点形成新的间隔序列; 第二步是合成 CRISPR RNA (crRNA); 最后一步是干扰, 在 crRNA 的指引下 Cas 核酸酶可以将特定新侵入的外源核酸同源片段序列

切割<sup>[9]</sup>。根据 CRISPRdb<sup>[10]</sup>数据库的统计信息, 目前已报道的细菌和古菌的基因组序列中约 45% 的细菌基因组和 83% 的古菌基因组中都含有 CRISPR 结构。一般微生物中 CRISPR 结构的数量有 1~2 个, 而多可达 10 或 30 个<sup>[9]</sup>。CRISPR 系统主要由同向重复序列 (Direct repeat, DR)、前导序列 (Leader sequence)、间隔序列 (Spacer sequence) 和 CRISPR 相关 (CRISPR-associated, Cas) 蛋白组成<sup>[11]</sup>。一般来说, CRISPR-Cas 系统及其各个组分在不同微生物基因组中的数量、基因序列、出现的频率以及大小等方面具有较高的变异<sup>[12]</sup>。Makarova 等<sup>[13]</sup>根据 *cas* 基因的结构排列顺序, 将 CRISPR-Cas 系统主要分为 3 个大类型 type I (含有 *cas3* 和 *cas7*), type II (含有 *cas9*) 和 type III (含有 *cas10*) 以及 10 个亚类型 (I-A、I-B、I-C、I-D、I-E、I-F、II-A、II-B、III-A 和 III-B)<sup>[13]</sup>。

重复序列的长度在不同物种中差别较大, 而在同一物种中保守性较高<sup>[14]</sup>。一般重复序列的长度为 23~55 bp, 重复序列中存在着短的回文序列, 可以形成发夹结构。间隔序列长度一般为 21~72 bp, 间隔序列来自外源核酸 DNA, 具有很高的多态性, 有时在同一种的不同菌株之间也会出现差异<sup>[9]</sup>。间隔序列的重复数在不同的菌种中差异明显, 据统计最高的重复数可以达到 588 (*Haliangium ochraceum*), 但一般都在 50 以下, 较低的只有 2。前导序列在 CRISPR 结构中发挥着重要的作用, 其通常位于 CRISPR 结构第一个重复序列的上游一段 AT 富含区域, 该段区域在不同物种间的保守性不高<sup>[15]</sup>。研究发现, 新插入的间隔序列总是会加在前导序列和第一个重复序列之间<sup>[16]</sup>。同时由于前导序列的特殊结构,

前导序列通常含有下游重复序列的转录启动子<sup>[17]</sup>。*cas* 基因位于 CRISPR 结构中重复序列的上游或下游区域,据 Daniel 等的研究,目前共有 45 个 *cas* 基因家族,并大致可分为 8 个亚型,每个亚型均由该亚型特定的基因组成<sup>[18]</sup>。另外,有 6 个基因(*cas1-6*)与大多数不同亚型的基因相关联,被认为是 *cas* 核心基因,尤以 *cas1* 为 CRISPR 结构的标志性基因<sup>[18]</sup>。研究表明,CRISPR 系统中 *cas* 基因和重复序列相互关联起作用,存在协同进化现象<sup>[19]</sup>。

目前有关细菌 CRISPR 结构的研究主要集中在大肠杆菌(*Escherichia coli*)、乳酸菌(*Lactic acid bacteria*)、嗜热链球菌(*Streptococcus thermophilus*)等<sup>[20]</sup>,但有关醋酸菌基因组中 CRISPR 结构的分析未见有相关报导。现有研究表明<sup>[1]</sup>,醋杆菌基因组序列中存在大量的前噬菌体 DNA 和转座子,导致醋杆菌基因组的不稳定性。与醋杆菌相比,葡糖杆菌和葡糖醋杆菌的基因组序列中含有较少的前噬菌体 DNA 序列,推测可能与其基因组中的 CRISPR 结构及数量有关。本文对醋酸菌中的 CRISPR 结构进行了比较分析,探索 CRISPR 位点在醋酸菌不同属中的变异规律,并对醋酸菌基因组中的前噬菌体序列及插入序列(Insertion sequence, IS)进行了预测,揭示噬菌体数量与 CRISPR 位点之间的关系<sup>[8]</sup>。

## 1 材料和方法

### 1.1 材料

*A. pasteurianus* Ab3 基因组序列的测序与分析是由浙江工商大学梁新乐实验室完成(GenBank 编号:CP012111),其余 27 株醋杆菌,9 株葡糖醋杆菌以及 11 株葡糖杆菌的基因组序列 GenBank 编号见附表 1。上述基因组序列均可从 NCBI 网站下载(<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>)。

### 1.2 方法

#### 1.2.1 醋酸菌中 CRISPR 位点查找

完整基因组中 CRISPR 结构的查找主要根据 CRISPRdb 数据库<sup>[10]</sup>。Draft 基因组序列中 CRISPR 的查找主要通过 CRISPRfinder(<http://crispr.u-psud.fr/>)以及 CRT 软件完成<sup>[21]</sup>。

#### 1.2.2 CRISPR 位点的生物信息学分析

同向重复序列的比对通过 ClustalW 完成, Draft

基因组中的 *cas* 基因的查找通过本地 BLAST 完成。前导序列的多序列比对分析同过 ClustalW 以及 BLAST 和 WEBLOGO<sup>[22]</sup>完成。对获得的 CRISPR 位点中的同向重复序列进行归类分析,共分为 8 个(CRISPR 1~8)。通过 MEGA 6.0 对不同种类的同向重复序列进行遗传进化聚类分析<sup>[23]</sup>。选取 CRISPR 结构位点中第一个重复序列上游 1000 bp 序列作为前导序列查找的对象,通过 ClustalW 反复比对分析,找出保守的 motif 区。*cas1* 基因序列的聚类分析方法和同向重复序列相同。

#### 1.2.3 基因组中噬菌体区域及 IS 序列预测

醋酸菌基因组中噬菌体区域的预测通过 PHAST<sup>[24]</sup>以及 Prophinder<sup>[25]</sup>共同完成。插入序列由 ISfinder(<https://www-is.biotoul.fr/>)<sup>[26]</sup>完成( $E=1e^{-10}$ )。原核生物启动子的预测由在线软件([http://www.fruitfly.org/seq\\_tools/promoter.html](http://www.fruitfly.org/seq_tools/promoter.html))进行(采用默认参数)。

#### 1.2.4 CRISPR 与噬菌体及 IS 序列的统计分析

通过非参数检验 Wilcoxon rank sum test 比较存在 CRISPR 与不存在 CRISPR 位点的菌株之间的噬菌体及 IS 数量的差异。由 Spearman 相关性检验进行 CRISPR 的间隔序列或重复序列与噬菌体及 IS 序列数量的相关性分析。所有统计分析由软件 Origin 8.5 完成。

## 2 结果与分析

### 2.1 醋酸菌 CRISPR 结构位点的基因组分布概况

48 株醋酸菌 CRISPR 位点分析结果表明,32 株醋酸菌含有至少一个 CRISPR 结构(附表 2),其中醋杆菌属 19 株,葡糖醋杆菌属 6 株,葡糖杆菌属 7 株。根据 CRISPRdb 数据库信息,4 株具有完整基因组序列菌株的 9 个 CRISPR 结构位点如图 1 所示。醋杆菌属 CRISPR 的平均个数是 1.8 个(*A. pasteurianus* Ab3 基因组序列中不含有 CRISPR 结构),低于葡糖醋杆菌属的 2.7 个以及葡糖杆菌属的 3.3 个。醋杆菌属中,CRISPR 数量最少的是 *A. nitrogenifigens* DSM 23921 和 *A. pomorum* DMCS\_004 以及巴氏醋杆菌(*A. pasteurianus*),都只有 1 个 CRISPR,特别是巴氏醋杆菌的 CRISPR 结构均位于质粒上,在其染色体上未发现 CRISPR 结构的存在;CRISPR 结构数最多的

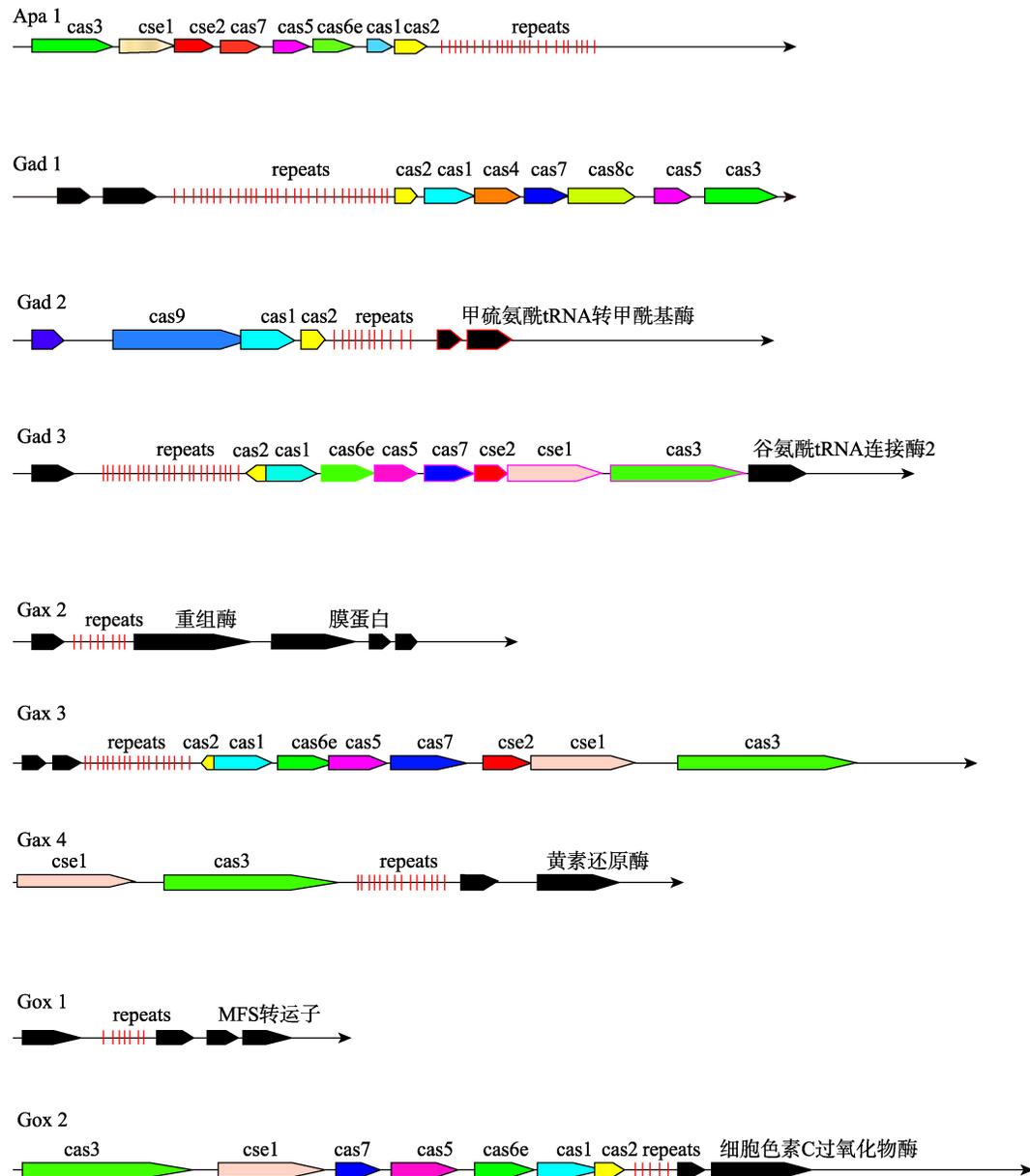


图 1 4 株醋酸菌中 9 个 CRISPR 位点结构示意图

Apa : 巴氏醋杆菌 *A. pasteurianus* IFO 3283-01 ; Gad : *Ga. diazotrophicus* PAI 5 ; Gax : *Ga. xylinus* NBRC 3288 ; Gox : *G. oxydans* H24。不同颜色的箭头代表不同的基因 ; 红色线条代表重复序列 (Repeats)。

是 *A. cibinongensis* 4H-1 和 *A. tropicalis* NBRC 16470 (各 4 个)。葡糖醋杆菌属中, CRISPR 数最少的是 *Ga. xylinus* E25, 只有 1 个, 其他葡糖醋杆菌则含有 2~4 个 CRISPR 结构。葡糖杆菌属中, CRISPR 结构数量相对最多, 几乎都含有 3~4 个。

对 32 株醋酸菌的 CRISPR-Cas 系统进行分类, 发现醋酸杆菌属中的 CRISPR-Cas 系统主要属于 I-E、I-F、I-C 等类型(包含 *cas3* 和 *cas7* 基因), 其中 *A. aceti*

ATCC 23746 的基因组序列中存在着 II 类 Cas 系统(附表 2, 附图 1)。葡糖醋杆菌属中的 CRISPR-Cas 系统主要属于 I-E 和 I-D, 其中在 *Ga. diazotrophicus* PAI 5 的基因组序列中存在着 3 种 CRISPR-Cas 系统: I-E 类, I-D 类以及 II 类。而在葡糖杆菌属的 7 个基因组序列中, CRISPR-Cas 系统均属于 I-E 类。3 个属所有菌株的基因组序列中皆不含 III 类 CRISPR-Cas 系统, 不同菌株基因组中的 CRISPR-Cas 结构情况

如附图 1 所示。不同醋酸菌属的菌种基因组序列中具有相同类型的 CRISPR-Cas 系统结构。

本研究所用的 48 个菌株共含有 73 个 CRISPR, 其中确定位点有 51 个, 可疑的有 22 个。为了便于分析, 根据重复序列的相似情况将 CRISPR 分为 9 组, 分别为 8 个类别(CRISPR 1~8)及其他(表 1)。其中在 CRISPR 1~8 中重复序列的结构具有较强的规律性, 而且在各个类型中重复序列具有较好的保守性。在其他类型中的重复序列多是可疑位点, 保守性较差, 没有一定的规律性。CRISPR 1~7 在醋杆菌属、葡糖醋杆菌属和葡糖杆菌属中都存在, 而 CRISPR 8 只存在于 *G. thailandicus* NBRC 3255 和 *G. thailandicus* NBRC 3257 中。另外, 具有相同类型 CRISPR-Cas 系统的菌株中, CRISPR 的数量和结构

却存在着较大的差异(附表 2)。

## 2.2 重复序列与 Cas 蛋白的进化分析

根据 48 株醋酸菌重复序列的比对分析发现, 相同的重复序列在不同属的菌株中广泛分布且具有一定的保守性(附表 2)。而在同种的不同菌株中, 重复序列也存在着差异性, 这种差异性体现在部分碱基突变以及 CRISPR 结构的不同(表 1)。以重复序列构建的系统进化树表明醋杆菌属、葡糖醋杆菌属和葡糖杆菌属的菌种互相混杂, 不能彼此分开, 因此不能利用该序列对不同属的菌种进行分类(图 2A)。每一个 CRISPR 类型中(1~8)都包含了来自不同属的菌种, 这说明菌种在遗传进化过程中可能形成了相同的防御体系。聚类分析同时也显示, 同一种内的

表 1 CRISPR 重复序列的分类分析

CRISPR	类型	重复序列(5'→3')	数量	间隔序列平均长度(bp)
CRISPR 1	重复序列变异型	TGAGAATATCCCCGCGCATGCGGGGAACAT	1	31
		CGCTTTATCCCCGCGTGTGCGGGGAACAC	1	32
		CGCTTTATCCCCGCGTGTGCGGGGAACAC	1	32
		GAGAGTACCCCCGCTCACGCGAGGAACAC	1	32
CRISPR 2	典型重复序列	GTGTTCCCCGCACACGCGGGGATGAACCG	18	32
	重复序列变异型	GTGTTCCCCGCACACGCGGGGATGAACCC	2	32
		GTGTTCCCCGCACACGCGGGGAT - - - - -	1	38
		GTGTTCCCCGCACACGCGGGGATGAACCGC	1	31
CRISPR 3	典型重复序列	AGCCTACCACTGGCAAATCGGTAGGGAAACCACGGC	2	30
CRISPR 4	典型重复序列	TTTCTGAGCTGCCTGTGCGGCAGTGAAC	2	32
	重复序列变异型	CATATAAACTGCCTGTGCGGCAGTGAAA	1	32
CRISPR 5	典型重复序列	GTTTCTATCCACGCTCCCGCACGGGGAGCGAC	7	34
	重复序列变异型	GTTTTATCCACGCTCCCGCACGGGGAGCGAC	1	34
		GTTTCTATCCACGCCCCGCACGGGAGGCGACC	1	34
		ATTCAATCCACGCTCCCGCACGGGAGGCGACC	1	32
CRISPR 6	典型重复序列	GTCGCTCCCCGTGCGGGAGCGTGGATAGAAAC	2	34
CRISPR 7	典型重复序列	CGGCTCATCCCCGCACGTGCGGGGAACAC	5	32
	重复序列变异型	CGGCTCATCCCCGCACGTGCGGGGAACACC	3	31
		GCCGACATCCCCGCGTGTGCGGGGAACAC	1	32
		AGTTTCACCCCCGCGTGTGCGGGGAACAC	1	32
		AGATTACCTCCGCGTGTGCGAGGAACAC	1	32
		CGGTTTCATCCCCGCGTGTGCGGGGAACAC	1	32
CGGTTTCATCCCCGCACGTGCGGGGAACACG	1	31		
CRISPR 8	典型重复序列	GGTGGATCACCTCGCGGGGTGAAGTG	2	33
其他			16	38

注: 下划线代表重复序列变异型中突变的碱基(和典型重复序列相比)。

CRISPR 具有完全不同的重复序列, 如 *G. thailandicus* NBRC 3255 中的 Gth 5 和 Gth 6(附表 2, 图 2A)。

重复序列的重复数在不同的物种间的差异明显, 即使在不同的种或者同一种内不同分离株的 CRISPR 结构中, 重复数的差异依然存在。醋杆菌属中 34 个 CRISPR 位点共有 443 个重复序列, 长度在 23 bp 到 38 bp 不等(包括确定的和可疑的, 确定的 CRISPR 中重复序列长度在 30 bp 左右)。在这 34 个 CRISPR 中重复数最高可达 38 个(*A. aceti* ATCC 23746), 最低只有 2 个。葡糖醋杆菌属中的 16 个 CRISPR 位点中共有 131 个重复序列, 长度在 23 bp 到 36 bp 不等(同样, 确定的 CRISPR 中重复序列长度在 30 bp 左右)。在这些 CRISPR 中重复数最高可达 30 个(*Ga. diazotrophicus* PAI 5), 最低也只有 2 个。葡糖杆菌属中共有 23 个 CRISPR 结构, 共有 136 个重复序列, 长度在 23 bp 到 30 bp 不等, 大多数在 30 bp 左右, 而 CRISPR 中重复数最高只有 10 个, 远低于醋杆菌属以及葡糖醋杆菌属中的菌株, 最低的重复数只有 2 个(附表 2)。

Cas 蛋白的比较基因组分析表明, 在绝大多数的 CRISPR 结构重复序列上游或者下游位置存在着和 Cas 蛋白编码相关的基因序列。根据 CRISPRdb

数据库中的数据分析, 巴氏醋杆菌(*A. pasteurianus*) CRISPR 结构 Apa 1 中重复序列的上游存在着 *cas3*、*cas7*、*cas5*、*cas6e*、*cas1* 以及 *cas2* 基因(图 1)。而葡糖醋杆菌中的 Gax 2 和葡糖杆菌中的 Gox 1 (图 1)不存在 *cas* 基因。同时, 在 Gad 3 和 Gax 3 中 *cas* 基因的排列顺序高度一致, 依次是 *cas2-cas1-cas6e-cas5-cas7-cse2-cse1-cas3*。*cas1* 基因的进化聚类分析发现 *cas1* 基因的保守性和重复序列相关, 重复序列在不同物种间的保守性也使得 *cas1* 基因保守(图 2B), 与前人的研究结果相吻合<sup>[19]</sup>。

### 2.3 间隔序列与前导序列的比较分析

在 48 株醋酸菌的 73 个 CRISPR 位点中共含有 637 个间隔序列, 长度从 18~60 bp 不等。虽然在同一物种中不同 CRISPR 中间隔序列长度不一, 但分析结果表明在 CRISPR1~8 分类中, CRISPR 1、CRISPR 2、CRISPR 3、CRISPR 4、CRISPR 6 以及 CRISPR 7 的间隔序列长度几乎都在 32 bp, CRISPR 5 的间隔序列长度在 34 bp, CRISPR 8 中的间隔序列长度在 33 bp(表 1)。

对 48 株醋酸菌的 51 个确定的 CRISPR 结构中的前导序列进行比对分析, 比对时选取各个位点第

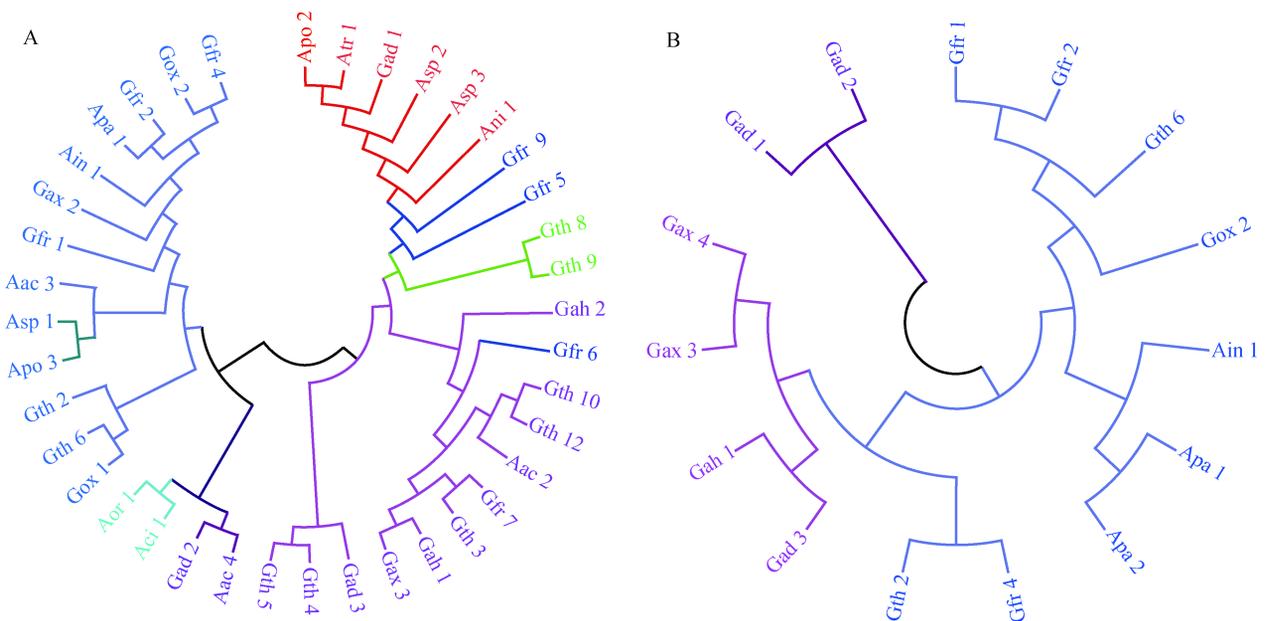


图 2 利用醋杆菌属(A)、葡糖醋杆菌属(Ga)和葡糖杆菌属(G)中的重复序列以及 *cas1* 基因序列构建的系统发生树拓扑结构图

A: 利用重复序列构建的系统发生树; B: 利用 *cas1* 基因序列构建的系统发生树。不同颜色的分支代表来源于不同种属的菌株。

一个重复上游 1000 bp 富含 AT 区域作为研究对象。经过 ClustalW 以及本地 BLAST 的反复比对分析,在 7 个菌种的 13 个 CRISPR 位点(Aac 3、Gox 1、Gox 2、Gfr 1、Gfr 2、Gfr 3、Gfr 4、Gth 2、Gth 4、Gth 5、Gth 6、Gth 7、Gth 12)的上游序列中发现一段较为保守的短 motif 序列(图 3A)。该序列靠近第一个重复序列,以“AAATCGG”碱基序列开始,以“GAAGAG”结束。保守结构中存在着众多的连续的“TTTT”结构,并有较明显的碱基突变,说明这些富含 AT 的区域与基因启动子的转录调控有关。进一步启动子预测分析,表明大部分前导序列中含有原核生物的启动子序列及其典型的转录因子结合位点(图 3B)。

### 2.4 间隔序列数量与前噬菌体数量、IS 序列数量的相关性分析

对 32 株醋酸菌基因组中的前噬菌体序列预测发现,在醋酸菌的基因组序列中皆存在着前噬菌体

(至少 1 个,最多 6 个)。醋杆菌属 15 个基因组序列中的前噬菌体平均数量为 2.49 个,低于葡糖醋杆菌属(7 个基因组序列)中的 3.86 个,高于葡糖杆菌属(10 个基因组序列)中的 1.8 个(附表 3)。在研究前噬菌体数量与 CRISPR 数量之间的关系发现,在不存在 CRISPR 结构的基因组序列中的前噬菌体数量并未多于其他含有多个 CRISPR 结构的基因组序列中的前噬菌体个数,32 株醋酸菌基因组序列中的前噬菌体个数与 CRISPR 未有显著的相关关系( $P=0.58$ )。而醋酸菌基因组序列中的前噬菌体个数与间隔序列个数在整体上存在着正相关趋势(菌株 *Ga. oboediens* 174Bp2、*Ga. europaeus* LMG 18494 和 *Ga. europaeus* LMG 18890 不能在其基因组序列中找到 CRISPR-Cas 系统的相关编码基因(附表 2),为了减少统计结果产生的误差,因此将这 3 株菌排除在外)(图 4A)(Spearman 相关性检验:  $\rho=0.40$ ,  $P=0.04$ )。对醋酸菌基因组中 IS 序列的预测发现,葡糖杆菌属基因

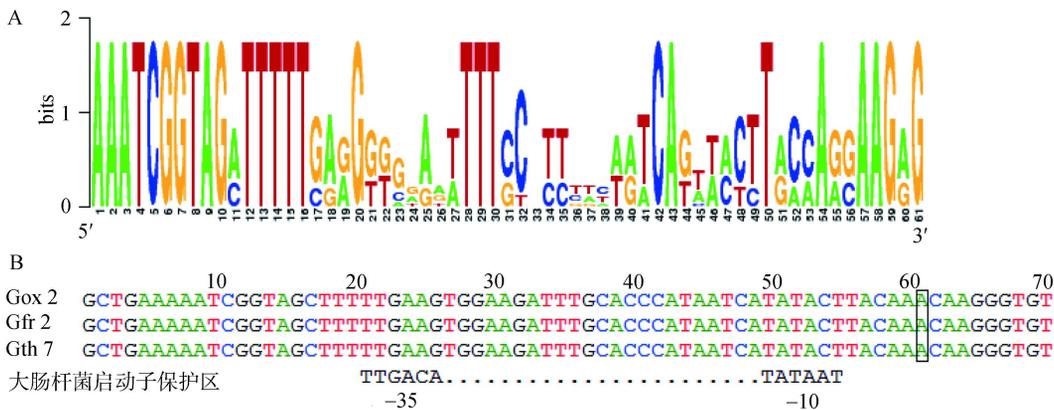


图 3 CRISPR 前导序列的保守 motif 区域与启动子位点比较分析

A: 13 个 CRISPR 前导序列具有保守 motif; B: 3 个菌株预测启动子的序列比对,黑框代表可能的转录位点。

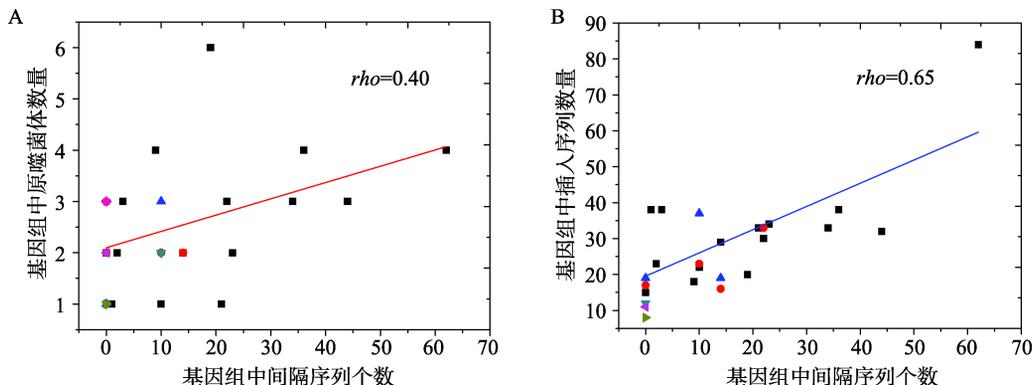


图 4 噬菌体数量及 IS 序列与 CRISPR 数量关系图

A: 噬菌体数量与间隔序列数量的关系; B: IS 序列数量与间隔序列数量的关系。

组序列中的 IS 序列个数是最多的(平均 53.71), 远高于醋杆菌属的 22.33 个以及葡糖醋杆菌属的 24.2 个, 并与前噬菌体数量呈正相关关系。基因组序列中 IS 序列个数在 *Ga. diazotrophicus* PAI 5 中最多(84 个), 在 *A. orientalis* 21F-2 中最少(4 个)。同时分析 IS 序列与间隔序列个数的关系发现, 插入序列和间隔序列间存在着显著的正相关趋势(图 4B)(Spearman 相关性检验:  $\rho = 0.65$ ,  $P = 4.32E-4$ ), 与前噬菌体数量和间隔序列个数之间的关系时相一致。另外, 葡糖醋杆菌属中的部分菌株拥有较多的插入序列可能与其基因组整体大小较大相关。

### 3 讨论

CRISPR 结构作为一种新发现的抵御噬菌体、病毒等外源核酸入侵的细胞防御机制, 越来越受到人们的关注。微生物在发酵过程中容易受到噬菌体的污染, CRISPR 可以抵抗噬菌体的侵染, 这给微生物发酵工业解决噬菌体污染带来全新的解决办法。随着人们对 CRISPR 功能的深入理解, CRISPR-Cas9 基因编辑系统正被广泛应用于微生物的基因组改造<sup>[17]</sup>。而对醋酸菌 CRISPR 的研究, 将有助于醋酸发酵菌种的进一步改造与控制。传统醋酸发酵过程中产生很高的酸度环境, 用于醋酸发酵生产的醋酸菌, 特别是醋杆菌属和葡糖醋杆菌属, 通常具有较高耐酸性能, 而其基因组如何适应高酸度发酵环境的进化机制还有待研究。比较基因组研究发现醋杆菌属和葡糖醋杆菌属的基因组序列中存在大量的噬菌体核酸序列<sup>[1]</sup>, 这与醋酸菌种性能的稳定性及醋酸发酵质量都有重要的影响。

本文对醋酸菌 3 个主要属中的 48 株菌进行了 CRISPR 位点分析, 发现 48 株菌中有 32 株含有 CRISPR 结构, 而 CRISPR-Cas 系统的分类表明大部分 CRISPR-Cas 系统属于 I 类中的 I-E 和 I-F 亚类型, 少数菌株的基因组序列中存在着 II 类的 Cas 系统, 而所有菌株中皆未发现 III 类 Cas 系统。从 CRISPR 的分布情况来看, 葡糖杆菌属和葡糖醋杆菌属的 CRISPR 数量多于醋杆菌属, 这与醋杆菌基因组序列中含有较多的噬菌体序列和转座子的现象相对应。相同的 CRISPR 在不同属的各菌株中均有分布, 具有一定的保守性, 推测这 3 个菌属的遗传性质比较接近。而重复序列进化分析表明重复序列不适合用

于不同属的菌株之间的分类, 说明 CRISPR 主要是通过水平转移机制在菌株间传递并独立进化。不同 CRISPR 类别(CRISPR 1~8)中的间隔序列长度几乎相同, 预示醋酸菌在发酵过程中受到相同或相似的噬菌体入侵。另外, CRISPR 结构在同属而不同种中的差异性可能揭示了不同菌种对外源 DNA 干扰的抵御能力的差异性, 而 CRISPR 结构在同种不同菌株中的差异性也提示外源 DNA 入侵的多样性。

通过 *cas* 基因的研究发现, *casI* 基因在不同属中存在着保守性, 而这一现象刚好和重复序列的保守性相吻合。此结果与文献报道的 *cas* 基因的作用和重复序列相关相一致<sup>[19]</sup>, 说明它们可能受相似的功能选择压力。前导序列在不同物种间并不保守<sup>[27]</sup>, 对醋酸菌 51 个已确定的 CRISPR 位点的前导序列进行分析后发现, 在 7 个种的 13 个 CRISPR 位点序列上游有一段短相对保守的 motif。该段序列在不同的物种中碱基突变率较大, 没有明显的回文特征, 推测可能是由于生长环境不同, 导致醋酸菌基因组部分序列在遗传进化过程中发生了较大的变化, 进而形成了菌株之间的差异, 这也与 Yoshinao 等<sup>[2]</sup>全基因组比较分析结果相似。

Takashi 等<sup>[28]</sup>研究发现在 *Streptococcus pyogenes* 中, 前噬菌体的数量与 CRISPR 的数量存在着负相关关系, CRISPR 限制了噬菌体的侵入。然而, 对醋酸菌基因组序列中前噬菌体区域数量与 CRISPR 以及间隔序列数量关系的研究发现, 醋酸菌基因组中前噬菌体区域数量与间隔序列数量呈现正相关趋势, 即前噬菌体的数量并不会因为间隔序列数量增多而减少。出现这个现象的原因可能是醋酸菌在进化过程中, 由于生存环境的变化, 噬菌体的侵染概率以及侵染噬菌体种类都会有差异, 因此当外界噬菌体以及病毒的侵染数量和种类增加时, 醋酸菌基因组中的 CRISPR 位点会相应的增加。另外, 这种现象可能也与醋酸菌基因组有较强的不稳定性相关, 容易受外源 DNA 的侵入。因为在 IS 序列分析时发现, 醋酸菌基因组中 IS 序列与间隔序列数量也同样呈现正相关趋势。

总之, 本文首次对醋酸菌 3 个主要属中的 CRISPR 结构进行了生物信息学分析, 其 CRISPR 结构的阐明为今后抗噬菌体工程醋酸菌的构建以及醋酸菌 CRISPR-Cas9 介导的基因敲除研究提供了思路。

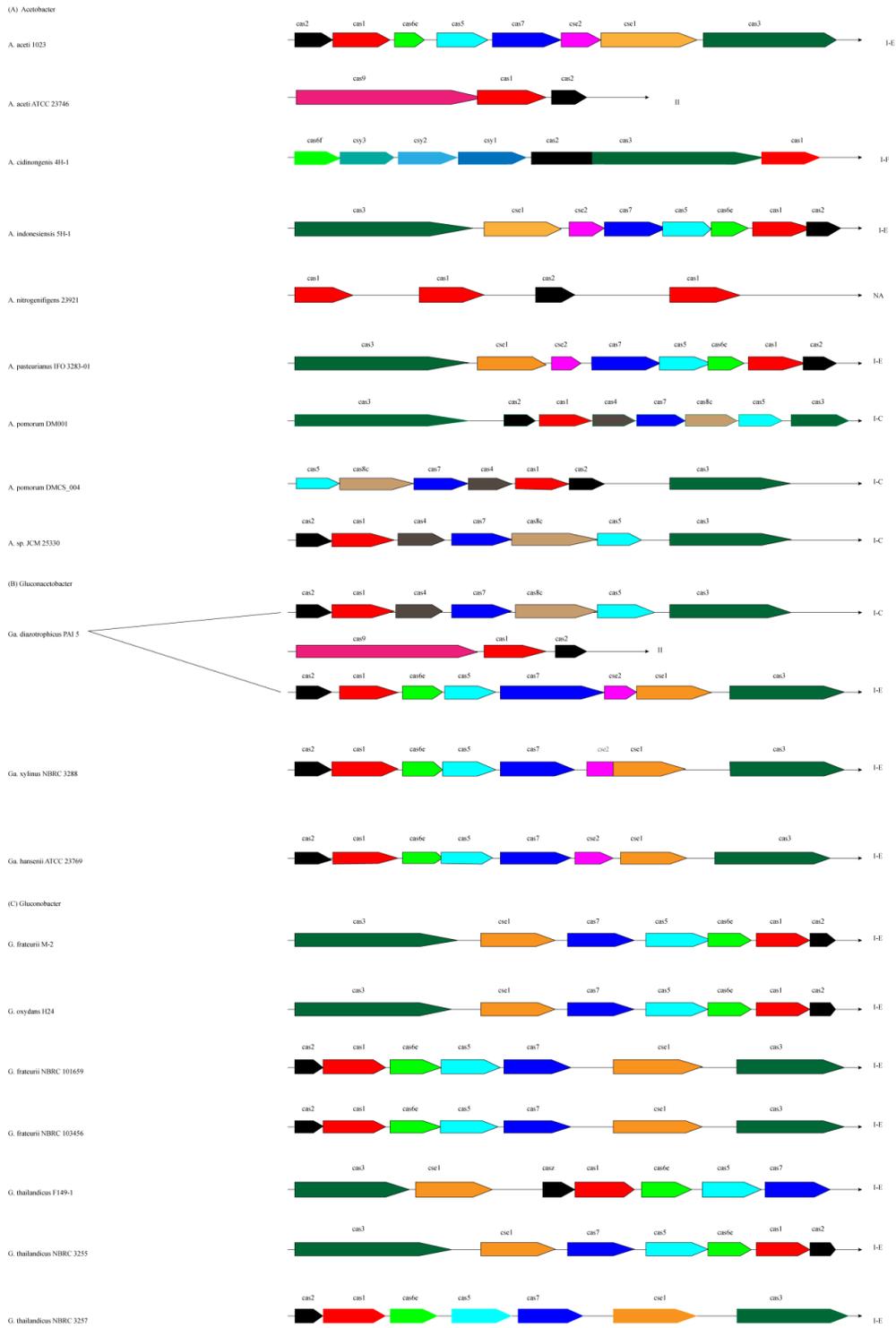
同时 CRISPR 位点的研究有助于揭示醋酸菌基因组的不稳定性及其对高酸环境适应性进化的分子遗传机制。

附录: 附图和附表见电子版 [www.chinagene.cn](http://www.chinagene.cn)。

## 参考文献

- [1] Illegheems K, Vuyst LD, Weckx S. Complete genome sequence and comparative analysis of *Acetobacter pasteurianus* 386B, a strain well-adapted to the cocoa bean fermentation ecosystem. *BMC Genomics*, 2013, 14: 526. [DOI]
- [2] Azuma Y, Hosoyama A, Matsutani M, Furuya N, Horikawa H, Harada T, Hirakawa H, Kuhara S, Matsushita K, Fujita N, Shirai M. Whole-genome analyses reveal genetic instability of *Acetobacter pasteurianus*. *Nucl Acids Res*, 2009, 37(17): 5768–5783. [DOI]
- [3] Matsutani M, Hirakawa H, Saichana N, Soemphol W, Yakushi T, Matsushita K. Genome-wide phylogenetic analysis of differences in thermotolerance among closely related *Acetobacter pasteurianus* strains. *Microbiology*, 2012, 158(Pt 1): 229–239. [DOI]
- [4] Hambly E, Suttle CA. The virosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol*, 2005, 8(4): 444–450. [DOI]
- [5] Matsutani M, Hirakawa H, Yakushi T, Matsushita K. Genome-wide phylogenetic analysis of *Gluconobacter, Acetobacter*, and *Gluconacetobacter*. *FEMS Microbiol Lett*, 2011, 315(2): 122–128. [DOI]
- [6] Heler R, Marraffini LA, Bikard D. Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Mol Microbiol*, 2014, 93(1): 1–9. [DOI]
- [7] 李君, 张毅, 陈坤玲, 单奇伟, 王延鹏, 梁振, 高彩霞. CRISPR/Cas 系统: RNA 靶向的基因组定向编辑新技术. *遗传*, 2013, 35(11): 1265–1273. [DOI]
- [8] 李铁民, 杜波. CRISPR-Cas 系统与细菌和噬菌体的共进化. *遗传*, 2011, 33(3): 213–218. [DOI]
- [9] Barrangou R, Marraffini LA. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol Cell*, 2014, 54(2): 234–244. [DOI]
- [10] Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, 2007, 8: 172. [DOI]
- [11] Millen AM, Horvath P, Boyaval P, Romero DA. Mobile CRISPR/Cas-mediated bacteriophage resistance in *Lactococcus lactis*. *PLoS One*, 2012, 7(12): e51663. [DOI]
- [12] Terns RM, Terns MP. CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends Genet*, 2014, 30(3): 111–118. [DOI]
- [13] Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol*, 2011, 9(6): 467–477. [DOI]
- [14] Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, 2005, 151(Pt 3): 653–663. [DOI]
- [15] Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol*, 2008, 190(4): 1401–1412. [DOI]
- [16] Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 2007, 315(5819): 1709–1712. [DOI]
- [17] Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 2008, 321(5891): 960–964. [DOI]
- [18] Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*, 2005, 1(6): e60. [DOI]
- [19] Horvath P, Coûté-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, Barrangou R. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol*, 2009, 131(1): 62–70. [DOI]
- [20] Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJ. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology*, 2010, 156(Pt 5): 1351–1361. [DOI]
- [21] Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, 2007, 8: 209. [DOI]
- [22] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*, 2004, 14(6): 1188–1190. [DOI]
- [23] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*, 2013, 30(12): 2725–2729. [DOI]
- [24] Zhou Y, Liang YJ, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. *Nucleic Acids Res*, 2011, 39(Web Server issue): W347–W352. [DOI]
- [25] Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, 2008, 24(6): 863–865. [DOI]
- [26] Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucl Acids Res*, 2006, 34(Database issue): D32–D36. [DOI]
- [27] Jiang FG, Doudna JA. The structural biology of CRISPR-Cas systems. *Curr Opin Struct Biol*, 2015, 30: 100–111. [DOI]
- [28] Nozawa T, Furukawa N, Aikawa C, Watanabe T, Haobam B, Kurokawa K, Maruyama F, Nakagawa I. CRISPR inhibition of prophage acquisition in *Streptococcus pyogenes*. *PLoS One*, 2011, 6(5): e19543. [DOI]

(责任编辑: 谢建平)



附图 1 19 株醋酸菌中 21 个 CRISPR-Cas 系统的分类和分布情况

NA 代表 not accessible, 相同的颜色代表同一基因或者同源基因, 颜色箭头之间的距离代表的是基因间的相对位置, 箭头的长短代表的是基因的相对大小。

附表 1 醋酸菌基因组序列信息汇总

菌株	基因组编号	菌株	基因组编号
<i>A. aceti</i> 1023	GCA_000691125	<i>A. aceti</i> ATCC 23746	GCA_000379545
<i>A. aceti</i> NBRC 14818	GCA_000193495	<i>A. cibinongensis</i> 4H-1	GCA_000963925
<i>A. indonesiensis</i> 5H-1	GCA_000963945	<i>A. malorum</i> DmCS_005	GCA_000743885
<i>A. nitrogenifigens</i> DSM 23921	GCA_000429165	<i>A. okinawensis</i> JCM 25146	GCA_000613865
<i>A. orientalis</i> 21F-2	GCA_000963965	<i>A. orleanensis</i> JCM 7639	GCA_000964205
<i>A. pasteurianus</i> 386B	GCA_000723785	<i>A. pasteurianus</i> IFO 3283-01	GCA_00010825
<i>A. pasteurianus</i> IFO 3283- 01-42C	GCA_00010945	<i>A. pasteurianus</i> IFO 3283-03	GCA_00010845
<i>A. pasteurianus</i> IFO 3283-07	GCA_00010865	<i>A. pasteurianus</i> IFO 3283-12	GCA_00010965
<i>A. pasteurianus</i> IFO 3283-22	GCA_00010885	<i>A. pasteurianus</i> IFO 3283-32	GCA_00010925
<i>A. pasteurianus</i> IFO 3283-26	GCA_00010905	<i>A. papaya</i> JCM 25143	GCA_000613285
<i>A. pomorum</i> DM001	GCA_000193245	<i>A. pomorum</i> DMCS_004	GCA_000755675
<i>A. sp.</i> JCM 25330	GCA_000613905	<i>A. syzygii</i> 9H-2	GCA_000964225
<i>A. tropicalis</i> NBRC 101654	GCA_000225485	<i>A. tropicalis</i> NBRC 16470	GCA_000755665
<i>Ga. diazotrophicus</i> PAI 5	GCA_000067045	<i>Ga. europaeus</i> LMG 18494	GCA_000227545
<i>Ga. europaeus</i> LMG 18890	GCA_000285295	<i>Ga. europaeus</i> 5P3	GCA_000285335
<i>Ga. hansenii</i> ATCC 23769	GCA_000164395	<i>Ga. oboediens</i> 174Bp2	GCA_000227565
<i>Ga. sp.</i> SXCC-1	GCA_000208635	<i>Ga. xylinus</i> E25	GCA_000550765
<i>Ga. xylinus</i> NBRC 3288	GCA_000182745	<i>G. oxydans</i> H24	GCA_000311765
<i>G. frateurii</i> M-2	GCA_000964445	<i>G. frateurii</i> NBRC 101659	GCA_000284875
<i>G. frateurii</i> NBRC 103456	GCA_000509445	<i>G. morbifer</i> G707	GCA_000234355
<i>G. oxydans</i> 621H	GCA_000011685	<i>G. oxydans</i> DSM 2003	GCA_000507285
<i>G. oxydans</i> DSM 3504	GCA_000583855	<i>G. thailandicus</i> F149-1	GCA_000964465
<i>G. thailandicus</i> NBRC 3255	GCA_000344115	<i>G. thailandicus</i> NBRC 3257	GCA_000576285

附表 2 32 株醋酸菌基因组中 CRISPR 结构概况

菌株	CRISPR 数量	CRISPR 名称	位置	重复序列数量	重复序列长度 (平均)	间隔序列数量	间隔序列长度 (平均)	代表性的重复序列 (5'-3')	CRISPR-Cas 系统类型
<i>A. aceti</i> 1023	2	Aac 1 <sup>a</sup>	chr	2	28	1	33	AAGAGCAGCGCGCAGGAAGA GGCAAC	I-E 类
		Aac 2 <sup>b</sup>	chr	34	29	33	32	CGGTTTCATCCCCGCGTGTGCGGG GAACAC	
<i>A. aceti</i> ATCC 23746	2	Aac 3 <sup>b</sup>	chr	8	29	7	32	GTGTTCCCCGCGTAGGCGGGGA TGAACCC	II 类
		Aac 4 <sup>b</sup>	chr	38	36	37	30	AGCCTACCCTGGCAAATCGGT AGGGAAACCACGGC	
<i>A. cibinongensis</i> 4H-1	4	Aci 1 <sup>b</sup>	chr	4	28	3	32	TTTCTGAGCTGCCTGTGCGGCAG TGAAC	I-F 类
		Aci 2 <sup>b</sup>	chr	4	28	3	32	TTATGAGTCTTCTGTGCGGCAG TGAAC	
		Aci 3 <sup>a</sup>	chr	2	25	1	35	ACTGCCGCACAGGCAGCTCAGA AAA	
		Aci 4 <sup>a</sup>	chr	3	23	2	27	TATCGTCATATTAGATTCAGTCT	
<i>A. indonesiensis</i> 5H-1	3	Ain 1 <sup>a</sup>	chr	4	29	3	32	GTGTTCCCCGCACACGCGGGGA TGAACCG	I-E 类
		Ain 2 <sup>a</sup>	chr	7	29	6	32	GTGTTCCCCGCATACGCGGGGA TGAACCG	
		Ain 3 <sup>a</sup>	chr	5	29	4	32	GTGTTCCCCGCATACGCGGGGA TGAACCG	
<i>A. nitrogenifigens</i> DSM 23921	1	Ani 1 <sup>b</sup>	chr	4	33	3	32	ATTTCAATCCACGCCTCCGCACG GGAGGCGACC	NA
<i>A. orientalis</i> 21F-2	2	Aor 1 <sup>b</sup>	chr	9	28	8	32	CATATAAACTGCCTGTGCGGCA GTGAAA	NA
		Aor 2 <sup>a</sup>	chr	3	28	2	33	CTTCACTGCCGCACAGGCAGCTC AGAAA	

续表 2

菌株	CRISPR 数量	CRISPR 名称	位置	重复序列数量	重复序列长度 (平均)	间隔序列数量	间隔序列长度 (平均)	代表性的重复序列 (5'-3')	CRISPR-Cas 系统类型
<i>A. orleanensis</i> JCM 7639	2	Aor 3 <sup>a</sup>	chr	2	33	1	54	TGTCCTTGGCGTTCTGCTCAGCCTTG GTCAGCT	NA
		Aor 4 <sup>a</sup>	chr	2	39	1	32	TTTTCAGGGGTAAATCCTTGGTCCAA GGAACCCCTATA	
<i>A. pasteuriana</i> IFO 3283-01	1	Apa 1 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-01-42C	1	Apa 2 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-03	1	Apa 3 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-07	1	Apa 4 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-12	1	Apa 5 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-22	1	Apa 6 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-26	1	Apa 7 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pasteuriana</i> IFO 3283-32	1	Apa 8 <sup>b</sup>	plasmid	24	29	23	32	GTGTTCCCGCACACGCGGGGATGA ACCG	I-E 类
<i>A. pomorum</i> DM001	2	Apo 1 <sup>a</sup>	chr	2	38	1	49	TTGTTACACGATCGCGCTGAGCCTG TAGCTTCTCAG	I-C 类
		Apo 2 <sup>b</sup>	chr	14	32	13	34	GTTTCTATCCACGCTCCCGCACGGGG AGCGAC	
<i>A. pomorum</i> DMCS_004	1	Apo 3 <sup>b</sup>	chr	15	32	14	34	GTCGCTCCCGTGC GG GAGCGTGGA TAGAAAC	I-C 类
<i>A. sp.</i> JCM 25330	3	Asp 1 <sup>b</sup>	chr	6	32	5	34	GTCGCTCCCGTGC GG GGGCGTGGA TAGAAAC	I-C 类
		Asp 2 <sup>b</sup>	chr	12	32	11	34	GTTTTATCCACGCTCCCGCACGGGG AGCGAC	
		Asp 3 <sup>b</sup>	chr	4	33	3	34	GTTTCTATCCACGCCCCGCACGGGA GGCGACC	
<i>A. tropicalis</i> NBRC 16470	4	Atr 1 <sup>b</sup>	chr	19	32	18	34	GTTTCTATCCACGCTCCCGCACGGGG AGCGAC	NA
		Atr 2 <sup>b</sup>	chr	16	32	15	34	GTTTCTATCCACGCTCCCGCACGGGG AGCGAC	
		Atr 3 <sup>a</sup>	chr	3	23	2	33	GTTTCTATCCACGCTCCCGCACGGGG AGCGAC	
		Atr 4 <sup>b</sup>	chr	29	32	28	34	GTTTCTATCCACGCTCCCGCACGGGG AGCGAC	
<i>Ga. diazotrophicus</i> PAI 5	4	Gad 1 <sup>b</sup>	chr	30	32	29	34	GTTTCAATCCACGCTCCCGCACAGGG AGCGAC	I-D 类, II 类, I-E 类
		Gad 2 <sup>b</sup>	chr	11	36	10	30	AGCCTACCATCGCAAATCGGTAGG GAAACCACGGC	
		Gad 3 <sup>b</sup>	chr	23	29	22	31	CGGTTTATCCCGCACGTGCGGGAA CACG	
		Gad 4 <sup>a</sup>	chr	2	32	1	42	GCAGCAACGCGCCGATCCAGGCC TCGAGCGAGACCACGCGTCGCCGT AGACCTGCG	
<i>Ga. europaeus</i> LMG 18494	2	Gae 1 <sup>a</sup>	chr	3	34	2	20	CGCCAAAGCCGCTTCGCGGCCCTC CCGCTGTGGCTGAAACGGTTGC	NA
		Gae 2 <sup>a</sup>	chr	3	25	2	38	CGCCAAAGCCGCTTCGCGGCCCTC CCGCTGTGGCTGAAACGGTTGC	
<i>Ga. europaeus</i> LMG 18890	2	Gae 4 <sup>a</sup>	chr	2	28	1	18	TGTGCGGATGTTCCGATAAGGCGT GCG	NA

续表 2

菌株	CRISPR 数量	CRISPR 名称	位置	重复序列数量	重复序列长度 (平均)	间隔序列数量	间隔序列长度 (平均)	代表性的重复序列 (5'-3')	CRISPR-Cas 系统类型
<i>Ga. hansenii</i> ATCC 23769	3	Gah 1 <sup>b</sup>	chr	7	29	6	32	CGGTTTCATCCCCGCGTATGCGGGGA ACAC	I-E 类
		Gah 2 <sup>b</sup>	chr	4	29	3	32	GCCGACATCCCCGCGTGTGCGGGGA ACAC	
		Gah 3 <sup>a</sup>	chr	2	24	1	39	CCGGGGCCTCCGCCACCGGGGCGG	
<i>Ga. xylinus</i> E25	1	Gax 1 <sup>a</sup>	chr	2	24	1	36	TGCCCCCGCAACCAATGAATCCTG	NA
<i>Ga. xylinus</i> NBRC 3288	4	Gax 2 <sup>b</sup>	chr	8	29	7	32	GTGTTCCCGCACGCGCGGGGATGA ACCG	I-E 类
		Gax 3 <sup>b</sup>	chr	16	29	15	32	CGGTTCCAGCCTCGCGTATGCGGGGA ACAC	
		Gax 4 <sup>b</sup>	chr	14	29	13	32	CCGTTTAGCCTGCGGGTGTGCGGGGA ACAC	
		Gax 5 <sup>a</sup>	chr	2	25	1	60	CTGCCCGACATGCCCGACTGCCCGA	
<i>G. oxydans</i> H24	2	Gox 1 <sup>b</sup>	chr	7	29	6	32	GTGTTCCCGCATGTGCGGGGATGA ACCG	I-E 类
		Gox 2 <sup>b</sup>	chr	5	30	4	31	GTGTTCCCGCACACGCGGGGATGA ACCGC	
<i>G. frateurii</i> M-2	3	Gfr 1 <sup>b</sup>	chr	7	29	6	32	GTGTTCCCGCACGTGCGGGGATGA ACCG	I-E 类
		Gfr 2 <sup>b</sup>	chr	4	23	3	38	GTGTTCCCGCACACGCGGGGAT	
		Gfr 3 <sup>b</sup>	chr	6	29	5	32	GTGTTCCCGCATGTGCGGGGATGA ACCG	
<i>G. frateurii</i> NBRC 101659	3	Gfr 4 <sup>b</sup>	chr	7	29	6	32	GTGTTCCCGCACACGCGGGGATGA ACCC	I-E 类
		Gfr 5 <sup>b</sup>	chr	5	30	4	31	TGAGAATATCCCCGCGCATGCGGGG AACAT	
		Gfr 6 <sup>b</sup>	chr	5	29	4	32	CGCTTTATCCCCGCGTGTGCGGGGAA CAC	
<i>G. frateurii</i> NBRC 103456	3	Gfr 7 <sup>b</sup>	chr	4	29	3	32	AGTTTACCCCCGCGTGTGCGGGGA ACAC	I-E 类
		Gfr 8 <sup>b</sup>	chr	5	29	4	32	CGCTTTATCCCCGCGTGTGCGGGGAA CAC	
		Gfr 9 <sup>b</sup>	chr	4	29	3	32	GAGAGTACCCCCGCTCACGCGAGGA ACAC	
<i>G. thailandicus</i> F149-1	4	Gth 1 <sup>a</sup>	chr	2	28	1	33	GGTGGATCACCTCGCGGGGTGAA GTG	I-E 类
		Gth 2 <sup>b</sup>	chr	8	29	7	32	GTGTTCCCGCATGTGCGGGGATGA ACCG	
		Gth 3 <sup>b</sup>	chr	10	29	9	32	AGATTTACCTCCGCGTGTGCGAGGA ACAC	
		Gth 4 <sup>b</sup>	chr	6	30	5	31	CGGCTCATCCCCGCACGTGCGGGGA ACACC	
<i>G. thailandicus</i> NBRC 3255	4	Gth 5 <sup>b</sup>	chr	6	30	5	31	CGGCTCATCCCCGCACGTGCGGGGA ACACC	I-E 类
		Gth 6 <sup>b</sup>	chr	8	29	7	32	GTGTTCCCGCATGTGCGGGGATGA ACCG	
		Gth 7 <sup>b</sup>	chr	10	29	9	32	GTGTTCCCGCACACGCGGGGATGA ACCG	
		Gth 8 <sup>a</sup>	chr	2	28	1	33	GGTGGATCACCTCGCGGGGTGAA GTG	
<i>G. thailandicus</i> NBRC 3257	4	Gth 9 <sup>a</sup>	chr	2	28	1	33	GGTGGATCACCTCGCGGGGTGAA GTG	I-E 类
		Gth 10 <sup>b</sup>	chr	10	29	9	32	CGGTTTCATCCCCGCGTGTGCGGGGA ACAC	
		Gth 11 <sup>b</sup>	chr	7	29	6	32	CGGTTTCATCCCCGCACATGCGGGGA ACAC	
		Gth 12 <sup>b</sup>	chr	6	30	5	31	CGGTTTCATCCCCGCACGTGCGGGGA ACACC	

附表 3 不同醋酸菌基因组序列中噬菌体、插入序列概况

菌株	基因组大小(Mb)	CRISPR 数量	间隔序列数量	原噬菌体数量	插入序列数量
<i>A. aceti</i> 1023	2.91	2	34	3	33
<i>A. aceti</i> ATCC 23746	3.57	2	44	3	32
<i>A. aceti</i> NBRC 14818	3.46	0	0	2	28
<i>A. cibinongensis</i> 4H-1	3.00	4	9	4	18
<i>A. nitrogenifigens</i> DSM 23921	4.13	1	3	3	38
<i>A. orleanensis</i> JCM 7639	2.92	2	2	2	23
<i>A. pasteurianus</i> 386B	2.73	0	0	2	15
<i>A. pomorum</i> DM001	2.79	2	14	2	19
<i>A. sp.</i> JCM 25330	7.10	3	19	6	20
<i>A. papaya</i> JCM 25143	2.94	0	0	1	8
<i>A. tropicalis</i> NBRC 10654	3.59	0	0	2	17
<i>Ga. diazotrophicus</i> PAI 5	3.82	4	62	4	84
<i>Ga. europaeus</i> LMG 18494	3.86	2	4	5	60
<i>Ga. europaeus</i> LMG 18890	4.09	2	2	5	67
<i>Ga. xylinus</i> E25	3.34	1	1	1	38
<i>Ga. xylinus</i> NBRC 3288	3.04	4	36	4	38
<i>Ga. oboediens</i> 174Bp2	4.04	0	0	6	66
<i>G. frateurii</i> M-2	3.14	3	14	2	29
<i>G. thailandicus</i> NBRC 3255	3.20	4	22	3	33
<i>G. oxydans</i> 621H	2.61	0	0	1	12
<i>G. oxydans</i> DSM 3504	2.79	0	0	1	11
<i>A. orientalis</i> 21F-2	2.82	2	10	1	4
<i>A. pasteurianus</i> IFO 3283-01	2.85	1	23	2	34
<i>A. syzygii</i> 9H-2	2.60	0	0	3	15
<i>A. malorum</i> DCMS_005	3.40	0	0	3	31
<i>Ga. hansenii</i> ATCC 23769	3.52	3	10	2	23
<i>G. oxydans</i> H24	3.49	2	10	2	37
<i>G. frateurii</i> NBRC 101659	3.05	3	14	1	16
<i>G. frateurii</i> NBRC 103456	3.28	3	10	3	22
<i>G. thailandicus</i> F149-1	3.33	4	22	1	30
<i>G. thailandicus</i> NBRC 3257	3.33	4	21	1	33
<i>G. morbifer</i> G707	2.79	0	0	3	19