学术论文

# 痢疾杆菌全基因组序列及基因组岛的分析

刘 红[1]，杨 帆[1]，张笑冰[1]，张继瑜[1]，杨国威[1]，董 杰[1]，薛 颖[1]，侯云德[1]，袁正宏[2]，闻玉梅[2]，徐建国[3]，陈洪松[4]，马大龙[4]，王 宇[4]，杨 剑[5]，沈 岩[5]，强伯勤[5]，吴洪涛[6]，贺秉坤[6]，吕渭川[6]，金 奇[1,7]

（1. 病毒基因工程国家重点实验室，北京 100052；2. 复旦大学分子病毒学实验室，上海 200032；3. 中国预防医学科学院流行病学和微生物学研究所，北京 102206；4. 北京大学医学部肝病研究所，北京 100083；5. 人类基因组北方研究中心，北京 100176；6. 华北制药集团，石家庄 050000；7. 卫生部微生物基因组研究中心，北京 100052）

[摘要] 福氏 2a 志贺菌（*Shigella flexneri serotype* 2a）是引起人类细菌性痢疾的主要病原体。本文在国际上首次完成了福氏 2a 志贺菌 301 株（Sf301）（我国细菌性痢疾的优势流行株）的全基因组核苷酸序列测定和初步分析。该基因组包括一条由 4 607 203 个碱基对（bp）组成的环状染色体和一个含 221 618 bp 的侵袭性大质粒 pCP301 以及另外两个小质粒。通过将 Sf301 的染色体序列与其亲源关系相近的非致病性大肠杆菌 K－12 菌株 MG1655 进行比较基因组学研究，发现 Sf301 的染色体上有 572 Kb 特异性序列，并形成了 320 个长度大于 50 bp 的“痢疾岛”（*Shigella*-island，SIs），其中大于 1 Kb 的共计 131 个。这些岛共包含 519 个开放读码框架（Open Reading Frames，ORFs），多数 SIs 的一侧或两侧均伴有插入序列元件、转座子或者 tRNAs。G＋C 含量及密码子使用频率等分析显示出部分 SIs 的外源性。通过结构及 ORF 编码产物功能的分析，鉴别出 9 个可能与痢疾杆菌致病性有关的“毒力岛”，其中 7 个可能的毒力岛为首次发现。

[关键词] 福氏 2a 志贺菌 301 株；全基因组序列测定；痢疾岛；毒力岛

[中图分类号] Q93-331；R733；R378.2$^{+}$4 [文献标识码] A [文章编号] 1009－1742（2002）10－0040－08

## 前 言

微生物基因组计划（Microbial Genome Program，MGP）及后基因组研究是继人类基因组计划以外生命科学领域启动的又一巨大工程。通过对微生物基因组及功能基因的研究，不仅能够使人们更深入地了解病原微生物的繁殖代谢、致病和耐药等机理，寻找更灵敏及特异的用于诊断和分型的分子标记以及有效的药物作用靶标，而且可为临床筛选有效的药物及发展疫苗提供基础。

志贺菌属（*shigella*）细菌通称痢疾杆菌，是一类具有高度传染性和危害严重的革兰氏阴性肠道致病菌，临床感染可以导致痢疾（shigellosis），其症状以发热、脱水和便血为特征。痢疾是世界上，尤其是发展中国家重要的传染病之一。全球每年的病例超过 1.6 亿，并导致 110 万患者死亡，其中绝大多数为 5 岁以下的儿童，因此痢疾是世界上造成婴幼儿死亡的主要原因之一[1]。福氏志贺菌是发展中国家也是我国流行的优势株，每年因痢疾造成的死亡人数中有 50％～70％由福氏志贺菌引起[2]。在中国，福氏 2a 志贺菌曾多次引起痢疾的大流行，并呈经常性散在爆发，使得痢疾成为我国发病率居第一位的传染病，每年约有上千万人次患病，对公共卫生与健康造成了巨大威胁；此外，目前临床上 95％以上的分离株对多种抗生素不敏感，已研制的疫苗不理想，目前尚没有有效的防治手段。据此，

[收稿日期] 2002－07－06

[基金项目] 国家重大基础研究规划（“973”计划 G1999054105）；高新技术研究发展计划（“863”计划 Z19-02-05-01）；北京市科委支持项目（955020700）

[作者简介] 刘 红（1973－），女，山东鱼台县人，病毒基因工程国家重点实验室博士研究生

开展痢疾杆菌基因组研究，阐明其结构与功能的关系具有十分重要的意义。

福氏 2a 志贺菌 301 株（*Shigella flexneri* 2a strain 301，Sf301）全基因组序列测定是我国第一个在国际上率先发布并完成的微生物基因组计划，所得到的大量重要数据和信息将为相关研究提供重要线索。本文重点介绍痢疾杆菌基因组中的特异性序列以及 SIs 的结构特征及功能预测，以期有助于更好地理解痢疾杆菌的进化和致病机理，为进一步研究痢疾的预防和治疗措施奠定基础。

毒力岛（Pathogenicity Island，PAIs）是医学细菌学领域的新名词，是指一类编码成簇毒力相关基因的、相对分子量较大、其 DNA 片段的 G + C 含量及密码子使用情况与宿主细菌染色体明显不同，且遗传相对不稳定的染色体 DNA 片段[3]；而基因组岛（Genomic islands，GIs）是指，当两个或多个亲缘关系较近的物种进行全基因组序列比较时所产生的、各基因组特异的 DNA 片段，是发现和鉴别新的毒力岛、代谢岛、共生岛等的基础和源泉[4]。本文所讨论的 SIs 即属于痢疾杆菌的基因组岛。

## 1　材料与方法

### 1.1　菌株和生长条件

福氏 2a 志贺菌 301 株：于 1984 年从北京市昌平县痢疾患者的粪便标本中分离。该菌株一直被作为中国福氏志贺菌的参考株，由中国预防医学科学院流行病学研究所提供。

菌株在常规刚果红培养基中 37℃ 培养过夜，挑取红色菌落在不含抗生素的 LB 培养基中于 37℃ 振荡培养过夜，用于分离大质粒和染色体 DNA。

### 1.2　鸟枪法序列测定与拼接

分别用改良 CTAB 法[5]及 QIAGEN 的 MIDI 试剂盒（QIAGEN 公司，德国）提取痢疾杆菌的染色体和大质粒 DNA；以 pBluescript KS（－）（Strategene 公司，美国）为载体，分别构建染色体和大质粒 Shotgun 随机文库，转化大肠杆菌 DH5α，用 Millipore（Millipore 公司，美国）试剂盒在 96 孔板中大规模制备测序模板。序列测定及数据收集采用 BigDye 试剂盒在 ABI3700 及 ABI377（Perkin Elemer，美国）自动测序仪上进行。共对 50 200 个克隆进行了双向测定，所测定序列碱基数约覆盖 10 倍的基因组总长度。

序列拼接应用 Phred/Phrap[6]软件，选择优化的参数和分值（≥20）在 SGI 工作站上运行。当染色体和质粒分别拼接生成 318 和 50 个 Contigs 时，利用 Consed 程序进行序列补平[7]。Gaps 的补平采用通过 Consed 对 Contigs 末端进行编辑，通过对 perl/Tk 程序鉴别的跨 Gap 质粒进行引物延伸以及对 PCR 扩增产物的直接测序来完成。

### 1.3　基因组注释

采用 Glimmer2.0 软件[8]预测基因组的 ORFs，选择长度大于 30 个氨基酸的 ORF，并对重叠及聚簇的 ORFs 进行手工检查，去除一些不合理的；利用 BLASTP 在非冗余蛋白库（The Non－Redundant protein database，NR）和直系同源群集合蛋白库（Clusters of Orthologous Groups of proteins，COGs）[9]中进行同源性搜索，依照其结果进行功能注释。ORF 同源的标准为：参与比对的提问序列和目标序列均大于全长的 60% 且一致性高于 30%。

tRNA 基因的识别采用 tRNAscan-SE 程序[10]，其他小 RNA 的鉴别是通过 BLAST 程序将已知小 RNA 序列在 Sf301 基因组序列进行比对搜索；重复序列是指在用 BLASTN 进行基因组自身以及与已知插入序列（Insertion Sequences，IS）元件数据库的两两比对中，显著性达到 $e^{-10}$ 且长于 200 碱基对的区域。

### 1.4　比较基因组分析

与大肠杆菌 K-12MG1655（GenBank 登记号为 U00096）的全基因组[11]比较采用杨剑等编写的程序 GenomeComp 。所有移位和倒置的接合部位以及所有假基因的序列，均经过 PCR 扩增及其产物的再次测序而检查核实，并对以上部位及痢疾岛与保守序列结合部的 ORFs 进行了手工检查与鉴别。

### 1.5　核酸序列登记号

Sf301 的染色体及大质粒 pCP301 的核苷酸序列已提交 GenBank，登记号分别为 AE005674 和 AF386526。

## 2　结果

### 2.1　基因组基本特征

Sf301 株的全基因组主要包括染色体和侵袭性大质粒（简写为 pCP301）两部分，其基本特征列于表 1。环型染色体全长 4 607 203 bp，G + C 含量为 0.5089 mol，包含 7 套 16S－23S－5S rRNA 操

纵子，1个 tmRNA 和包括 RNase P，6S RNA 及4.5S RNA 在内至少9个未分类 RNA；染色体全长的80.4%为蛋白编码区，0.8% 编码稳定的 RNA，而314个完整或截短拷贝的 IS 元件占到全长的3.6%。预测的开放读码框总数为4 434个，除去由于碱基插入、缺失、读框内终止以及移码突变而导致的254个假基因外，ORFs 的平均读长为891 bp。在所有的4 180个完整 ORFs 中，有2 895个(65%）可以分类到 COGs 库的不同生物功能种类，268个（6.4%）只有简单的功能预测，而另外1 195个（28.6%）则只是推断的功能未知的编码序列，其中有179个 ORF 与目前已知的任何蛋白质都没有显著的同源性。编码蛋白在 COGs 库中的功能分类情况列于表2。

**表1 Sf301 基因组基本特征**

Table 1 General features of Sf301 genome

| | 染色体 | 大质粒 |
|---|---|---|
| 总长（碱基对） | 4 607 203 | 221 618 |
| 开放读框总数 | 4 434 | 267 |
| 开放读框平均长度（碱基对） | 891 | 658 |
| 序列编码率/% | 80.4 | 76.24 |
| G+C 含量/% | | |
| 全长 | 50.89 | 45.77 |
| 蛋白编码区 | 51.96 | 46.13 |
| RNA 基因 | 54.79 | – |
| 基因间区域 | 46.07 | 44.59 |
| 插入序列元件 | 314 | 88 |
| 其中不完整拷贝数 | 67 | 62 |
| 核糖体 RNA | | |
| 16S | 7 | – |
| 23S | 7 | – |
| 5S | 8 | – |
| 转运 RNA 数目 | 97 | – |
| tmRNA | 1 | – |
| 未分类 RNA 数目 | 9 | – |

大质粒 pCP301 全长221 618 bp，共编码267个 ORFs，除去6个假基因，其平均读长为658 bp，序列编码率为76.24%，编码区的 G+C 含量为0.4613 mol。pCP301 中含有68 kb 的 IS 元件，占到总长度的30%，包括18个 IS 元件种类共计111个片段。

### 2.2 复制起始与终止区

沿用大肠杆菌 K-12 MG1655 确定的染色体的第一个碱基，痢疾杆菌福氏2a 301株的第一个碱基选定在无明显特征的 lasT 和 thrL 两基因之间。通过 GC 偏斜（GC skew）[（G-C）/（G+C)] 分析以及与大肠杆菌 K-12 MG1655 的相似性，确定了 Sf301 染色体复制原点 *oriC*。我们选择 *gidA* 和 *mioC* 基因间 Bgl Ⅱ内切酶位点的鸟嘌呤残基作为 Sf301 染色体复制原点的第一个核苷酸。GC 偏斜分析显示，Sf301 染色体的复制终止点（*terC*）位于染色体的1.60 Mb 附近。

**表2 Sf301 染色体蛋白在 COGs 系统中的分类**

Table 2 Distribution of Sf301 chromosomal proteins among COGs functional groups

| Functional class (COGs) | Number | percent of total/% |
|---|---|---|
| 翻译、核糖体结构及生物发生 | 159 | 3.80 |
| 转录 | 185 | 4.43 |
| DNA 复制、重组与修复 | 615 | 14.71 |
| 细胞分裂与染色体分离 | 28 | 0.67 |
| 翻译后修饰、蛋白周转、伴侣 | 108 | 2.58 |
| 细胞外套的生物合成、外膜 | 168 | 4.02 |
| 细胞动力和分泌 | 106 | 2.54 |
| 无机离子的转运和代谢 | 153 | 3.66 |
| 信号转导系统 | 89 | 2.13 |
| 能量产生与转化 | 217 | 5.19 |
| 碳水化合物的转运与代谢 | 264 | 6.32 |
| 氨基酸的转运与代谢 | 294 | 7.03 |
| 核苷的转运与代谢 | 74 | 1.77 |
| 辅酶的代谢 | 117 | 2.80 |
| 类脂化合物的代谢 | 68 | 1.63 |
| 次级代谢产物的生物合成，、转运及分解代谢 | 72 | 1.72 |
| 预测的一般功能 | 268 | 6.41 |
| 功能未知和未分类的 | 1 195 | 28.59 |
| 总数 | 3 180 | 100 |

### 2.3 痢疾杆菌特异性的 DNA 片段——SIs 的分析

Sf301 的染色体与大肠杆菌 K-12 MG1655 株染色体长度非常接近（4 639 221 bp）[11]。基因组比较结果显示，两者拥有4.03 Mb 保守的“共有序列”，但这种“共线性”却由于 Sf301 存在的 DNA 片段的移位或倒置而多次被中断。基因组的重排大多都与 IS 元件相关且总伴随有基因组序列的缺失或获得，从而形成本文所讨论的 K-岛（K-12-islands，KIs）和 SIs。痢疾杆菌特异性572

Kb 的序列共形成了 320 个大于 50 bp 的 SIs，大于 1Kb 的为 131 个（其中 10 Kb 以上的 8 个），其长度占特异性序列总长的 87.8%，该序列编码了痢疾杆菌独特的 519 个 ORF 中的 468 个。

SIs 的平均 G+C 含量为 0.4825 mol，显著低于保守序列的 0.5124 mol。如图 1 所示，有 41 个岛的 G+C 含量明显高于（>0.55 mol）或低于（<0.45 mol）基因组 0.5089 mol 的平均含量，其中 13 个是独立的 IS 元件岛；有 54 个岛的 G+C 含量则在 0.48 mol 到 0.53 mol 之间。SIs 和基因组在某些密码子的使用频率上也存在明显的差异（图 2），进一步分析发现这一差异与该岛的 G+C 含量有关，即 G+C 含量与基因组明显不同的岛，其密码子的使用频率也差异较大。

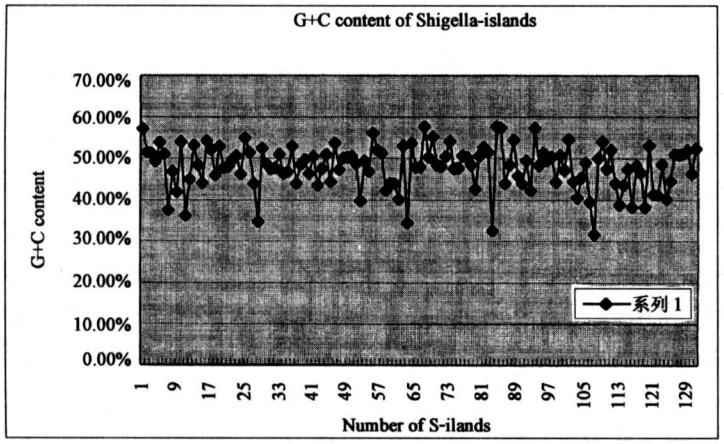


**图 1　痢疾杆菌岛的 G+C 含量**

Fig.1　G+C content of 131 *Shigella*-islands

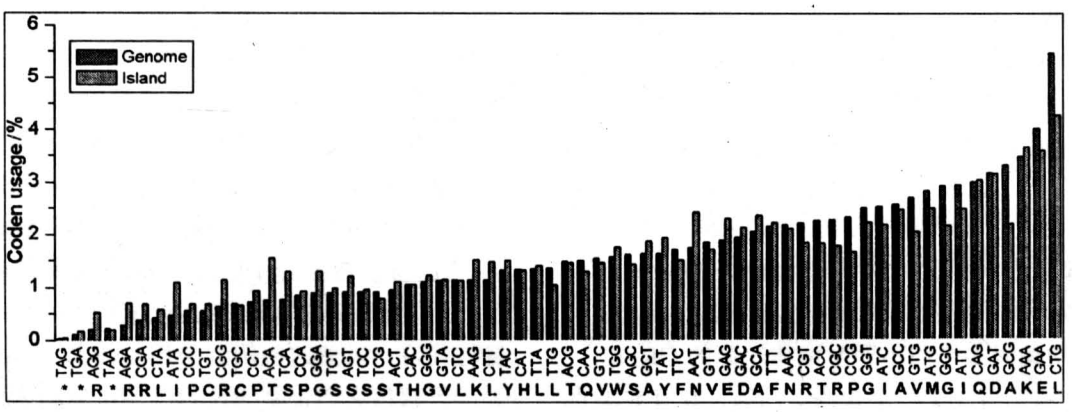


**图 2　Sf301 保守序列与岛上基因的密码子使用频率的分析**

（密码子按照在保守序列上的使用频率排列）

Fig.2　Codon usage analysis of Sf301 chromosomal genes which on "backbone sequences" (genome) and on SIs (island), respectively. Codons are put in the order of frequency in the genes on the conserved backbone.

2.3.1　已经报道的 PAIs　按照我们的计算方法，此前在痢疾杆菌中已经鉴定的两个 PAIs——SHI-2 岛[12]和 she 岛[13]均由多个连续的特异性片段组成，但本文仍将其各自作为一个完整的 SI。Sf301 中，she 岛插入在一个 Phe-tRNA 基因位点，全长 49 163 bp，G+C 含量为 0.4891 mol，共编码 41 个 ORF，而 SHI-2 岛毗邻 selC-tRNA 基因位点，全长 23 937 bp，G+C 含量为 0.482 mol，共编码 23 个 ORF。这两个片段与已经报道的 PAIs 的结构完全一致，其差异主要在于各自 IS 元件的种类、数量及位置不尽相同

2.3.2　各 SIs 的结构分析　按照其 ORF 的组成，

SIs可以分为三组，第一组有44个，它们完全或几乎完全由IS元件组成，称为“IS岛”。该组岛全长81 488 bp，占SIs的14%；第二组称作“噬菌体相关岛（phage-related islands）”，共31个，全长近160 Kb，它们或者完全是噬菌体的残余序列，或者部分携带有噬菌体相关的ORFs；其余56个岛组成了第三组，它们均编码其它已知或未知功能的ORFs。

除了有24个SIs的侧翼序列既不是某种氨基酸的tRNA也不是重复序列外，其余所有107个SIs均与IS元件有关：它们的一侧或者双侧连接有或相同或不同类型的IS元件。所以按照结构特征，SIs可以分为两大类：与IS元件相关的岛和与IS元件无关的岛 。

2.3.3 几个可能的PAIs的预测 侵袭质粒抗原ipa（invasion plasmid）是痢疾杆菌主要的表面抗原。染色体上有7个ipaH基因的同源序列，其中2个由于读框内IS元件的插入成为假基因，另外5个位于我们已经鉴别的岛上。按照在基因组上的排列位置，我们分别命名为ipaH－岛1－5。

ipaH－岛1由SI 21－26组成，全长约38 Kb，插入在Gly－tRNA基因位点；ipaH－岛2由SI 48－51组成，全长约25 Kb，两侧均为IS1插入元件，该岛上有4个连续的ORF与沙门氏菌铁转运系统的sitABCD操纵子相同[14]；ipaH－岛3由SI 67－69组成，全长约22 Kb，3’末端紧邻一个Gly－tRNA位点，其序列与大肠杆菌O157：H7中噬菌体的部分序列高度同源[15]；ipaH－岛4和5分别由SI 75－80和SI 94－96组成，分别长约23和11 Kb，均编码多个与不同噬菌体同源的ORFs。

SI 1和2组成了一个长约23 Kb，具有典型毒力岛结构的区域——插入在asp－tRNA位点，另一侧为IS629插入元件。主要携带有沙门氏菌 *sci* 操纵子及一些功能未知的噬菌体ORF的同源序列。

SI 4－10：全长约21 Kb，虽然序列上有很大的变异，但仍然保留了痢疾杆菌溶原性噬菌体SfⅡ的主要基因——编码细菌萜醇葡糖基转移酶的bgt和编码葡糖基转移酶Ⅱ的 *gtr* Ⅱ，它们是Ⅱ型抗原表达所必须的[16]。

SI 83：全长8 178 bp，两侧均没有特征性结构，其G+C含量仅为32.49%。该岛主要携带有与福氏痢疾杆菌型特异性O－抗原合成相关的 *rfb* 基因簇。

SI 116：全长8 115 bp，其3’末端紧邻一个IS 1插入元件，G+C含量为38.25%，主要编码与福氏痢疾杆菌菌体抗原核心多糖生物合成相关的 *rfa*（*waa*）基因簇。

## 3 讨论

通过GC偏斜分析以及与大肠杆菌K－12 MG1655的相似性，确定了Sf301染色体复制原点oriC位点，它毗邻基因 *gidB*，*gidA*，*mioC* 和 *asnC*，该区域内包含一段细菌复制特征性的AT－丰富序列。同样的分析显示，Sf301染色体的复制终止点（*terC*）位于染色体的1.60 Mb附近，在环型染色体上，该区域位于复制原点的对面，从而将其分成两个大致相等的复制子，表明了染色体的对称结构。虽然在大肠埃希氏属中可以允许有低水平的染色体非对称性[17]，但作为细菌长期进化和自然选择的结果，染色体的对称性有利于维持其结构的平行和稳定，这在肠道细菌的进化中起着重要作用[18]。全序列测定与分析发现，痢疾杆菌染色体上含有长达165 Kb的总计247个完整拷贝的各类IS元件，使得痢疾杆菌成为目前已知IS元件最多的基因组，无论是其种类还是数量都远远高于大肠杆菌K－12MG1655。Sf301的基因组中富含多种可移动的遗传成分，提示其遗传物质可能是活跃而不稳定的，这样不利于染色体结构的稳定性。研究发现，核糖体RNA操纵子（rrn operon）与肠道菌的基因组重排从而与染色体结构的对称性密切相关[19]。通过与已公布的大肠杆菌基因组的比较显示，Sf301的染色体上有多处大的DNA片段的倒置与移位，虽然这些倒置与移位均与rrn操纵子无关，但相信它们也是通过相同的机制将水平转移获得的外源片段在基因组内部进行重排，以保证染色体结构的平行和稳定。

自1950以来，志贺菌一直被作为一个属来研究，并被分成了四种：痢疾志贺菌（*S. dysenteriae*）、福氏志贺菌（*S. flexneri*）、宋内氏志贺菌（*S. sonnei*）及鲍氏志贺菌（*S. boydii*）[20]。但最近的遗传学分析提出，志贺菌是3.5万到27万年前从大肠杆菌起源，并经过了7～8次独立的进化而形成的，似乎还不能形成一个独立的属[21]。全基因组比较结果证实，Sf301似乎比O157：H7更为接近大肠杆菌K－12。痢疾杆菌与大肠杆菌从共同的祖先分化以后，各自在进化的过程中都可能

会丢掉或获得一些 DNA 片段，从而更利于各自的生存和繁殖，其结果导致 K－12 以正常菌群的方式共生于人类的肠道中，而痢疾杆菌则成为致病菌。比较分析发现，部分 SIs 与保守序列之间在 G＋C 含量与密码子使用频率上存在明显的差异，提示这些序列在进化上的外源性；结合 SIs 中丰富的噬菌体残余序列及基因组 IS 元件的丰富性，推测这些 SIs 可能是经由噬菌体或 IS 元件从其他物种水平转移而来。还有的 SIs 在 G＋C 含量与密码子使用频率上均与保守序列无明显差异，考虑到二者的进化关系，推测它们的产生可能是由于 K－12 基因组在相应位点的置换（replacement）或缺失造成的；与此相关的对于 K－岛（KIs）分析将为此提供进一步的佐证。

痢疾杆菌的致病过程主要包括细菌到达结肠粘膜，侵入粘膜上皮细胞并在细胞内繁殖，同时扩散到相邻细胞，引起程序性细胞死亡，最终造成肠粘膜水肿破坏并脱落。侵袭性大质粒是痢疾杆菌最主要的毒力因子，与其致病性密切相关[22]。在大质粒 DNA 中，编码 Ipa，VirG 和 Mxi－Spa 类蛋白的约 30 多个基因紧密连接在一起，形成一个长 31 kb 的区域，被称为“侵入区”，负责痢疾杆菌对于粘膜上皮细胞的侵袭，与志贺菌的致病作用直接相关[23]。IpaH 作为一类侵袭质粒抗原基因，同时存在于肠侵袭性大肠杆菌和志贺菌中[24]。IpaH 多基因家族具有共同的结构特点：5’端为 600～700 bp 的可变区，3’端为 839 bp 的恒定区。IpaH 基因的恒定区和可变区 GC 含量明显不同，说明它们具有不同的来源，恒定区可能是通过某种机制，由 ipaH 基因的早期基因衍生而来。IpaH 蛋白的来源和功能仍不清楚，但由于目前发现它只存在于痢疾杆菌和肠侵袭性大肠杆菌中，一般认为可能是 III 型分泌系统所分泌的，与细菌的侵入有关。我们发现在 Sf301 中，ipaH 基因共同存在于染色体和大质粒中，其中 pCP301 中有 5 个拷贝，而染色体上有 7 个 ipaH 的同源序列，其中 5 个位于我们鉴别并命名的 ipaH－岛上。值得注意的是 ipaH－岛 1 中除了同大肠杆菌 O157 中的噬菌体同源的 ORF 外，该岛还编码一个螺旋酶、一个推断的复制蛋白以及一个抗终止因子基因，提示该岛可能通过噬菌体转移获得，并且可能与痢疾杆菌在特定环境中的 DNA 复制及细胞分裂有关。所有的 ipaH－岛均携带有完整或截短的可移动遗传元件以及数目不等的未知功能的 ORFs。对于它们编码基因的体外表达及产物功能的研究将可能揭示 ipaH 蛋白在细菌侵袭中的作用，也将有助于阐明痢疾杆菌染色体与侵袭性大质粒之间的进化及调控关系。

脂多糖（lipopolysaccharide，LPS）是革兰氏阴性菌的主要表面成分，由类脂 A、核心多糖和 O－特异性多糖链即 O－抗原三部分连接而成，其中 O－抗原的特异性是革兰氏阴性菌血清型分型的基础。已知染色体上与 *his* 操纵子连锁的 *rfb* 基因簇与福氏痢疾杆菌型特异性 O－抗原合成相关[25]；而与 *mtl* 连锁的 *rfa* 基因簇则负责菌体抗原核心多糖的生物合成[26]，这两个区域均是福氏痢疾杆菌表达完整毒力所必须的决定因子[27,28]。我们所鉴别的 SI 83 和 SI 116，具有显著的外源性特征，即其 G＋C 含量及密码子使用均与保守的染色体序列明显不同；除了已知与 LPS 特异性成分的生物合成有关的 ORFs 外，它们还编码有其他未知功能的 ORFs，推测这些 ORFs 的功能也与完整 LPS 的生物合成密切相关。对于这两个 SIs 功能及来源的进一步研究将不仅有助于阐明痢疾杆菌与其他微生物的进化关系，并且可以作为筛选新的免疫组分的源泉。由于痢疾杆菌的 LPS 既是重要的毒力决定簇，也是重要的保护决定簇，阐明其生物合成及完整表达的遗传机制将为研制高效痢疾疫苗提供理论基础和实验依据。

痢疾杆菌主要的致病特点是侵袭结肠粘膜上皮细胞，虽然侵袭相关基因都定位于毒力大质粒上，但其毒力基因的完全表达却受染色体上多个基因的调控[20]。值得注意的是，虽然其突变后可以利用不同的机制降低或减轻痢疾杆菌的毒力，以前鉴定的所有染色体毒力位点都并非编码真正的毒力因子，因为 BLAST 搜索显示所有的这些位点在 K－12 基因组中都存在同源的甚至一致的序列。所以通过对所有 SIs 结构的分析，及其已经展开的对于其编码基因功能的研究，将可能会鉴别出真正的毒力及耐药相关基因，为开发预防和治疗痢疾的新策略奠定坚实的基础。

## 参考文献

[1] Sansonetti P J. Microbes and microbial toxins: paradigms for microbial-mucosal interactions Ⅲ. Shigellosis: from symptoms to molecular pathogenesis [J]. Am J Physiol Gastrointest Liv Physiol, 2001, 280, G319～G323

[2] Mei Y, Liu H, Xu J. Cloning and Application of genus specific DNA probes for shigella [J]. Chinese J Epidemiol, 1989, 10: 167～170

[3] Lee C A. Pathogenicity islands and the evolution of bacterial pathogens[J]. Infect Agents Dis. 1996 Jan, 5 (1):1～7

[4] Karlin S. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes[J]. Trends Microbiol 2001 Jul,9(7):335～343

[5] Ausubel F M, et al. 颜子颖 等译. 精编分子生物学实验指南[M]. 北京:科学出版社,1998,39

[6] Ewing B, Hillier L, Wendl M C, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment[J]. Genome Res, 1998, (8):175～185

[7] Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing[J]. Genome Res, 1998, (8):195～202

[8] Salzberg S L, Delcher A L, Kasif S, et al. Microbial gene identification using interpolated Markov models [J]. Nucleic Acids Res, 1998, 26, 544～548

[9] Tatusov R L, Galperin M Y, Natale D A, et al. The COG database: a tool for genome-scale analysis of protein functions and evolution[J]. Nucleic Acids Res, 2000,28,33～36

[10] Lowe T M, Eddy S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence[J]. Nucleic Acids Res, 1997, 25, 955～964

[11] Plattner F R., Plunkett G III, Bloch C A, et al. The complete genome sequence of Escherichia coli K－12 [J]. Science, 1997, 277, 1453～1474

[12] Moss J E, Cardozo T J, Zychlinsky A, et al. The selC-associated SHI-2 pathogenicity island of *shigella flexneri*[J]. Mol Microbiol, 1999, 33, 74～83

[13] Rajakumar K, Sasakawa C, Adler B. Use of a novel approach, termed island probing, identifies the Shigella flexneri she pathogenicity island which encodes a homolog of the immunoglobulin A protease-like family of proteins[J]. Infect Immun, 1997, 65, 4606～14

[14] Janakiraman A, Slauch J M. The putative iron transport system SitABCD encoded on SPI1 is required for full virulence of salmonella typhimurium[J]. Mol Microbiol 2000 Mar;35(5):1146～55

[15] Perna E S, Plunkett G Ⅲ, Burland V, et al. Genomic sequence of enterohaemorrhagic Escherichia coli 0157: H7[J]. Nature, 2001,409, 529～533

[16] Mavris M, Manning P A, Morona R. Mechanism of bacteriophage SfII-mediated serotype conversion in *shigella flexneri*[J]. Mol Microbiol, 1997,26, 939～950

[17] Tetsuya H, Kozo M, Makoto O, et al. Complete genome sequence of enterohemorrhagic eschelichia coli O157: H7 and genomic comparison with a laboratory strain K-12[J]. DNA Research 2001, 8, 11～22

[18] Bergthorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of Escherichia coli[J]. Mol Biol Evol 1998 Jan,15(1):6～16

[19] Liu G R., Rahn A, Liu W Q, et al. The evolving genome of Salmonella enterica serovar Pullorum[J]. J Bacteriol 2002 May,184(10):2626～33

[20] Hale T L. Genetic basis of virulence in shigella species [J]. Microbiol Rev, 1991, 55, 206～224

[21] Pupo G M, Lan R, Reeves P R. Multiple independent of origins of shigella clones of escherichia coli and convergent evolution of many of their characteristics[J]. Proc Natl Acad Sci USA, 2000,97, 10567～10572

[22] Sansonetti P J, Kopecko D J, Formal S B, et al. Involvement of a plasmid in the invasive ability of *shigella flexneri*[J]. Infect Immun 1982 Mar, 35(3): 852～60

[23] Sasakawa C, Kamata K, Sakai T, et al. Virulence-associated genetic regions comprising 31 kilobases of the 230-kilobase plasmid in Shigella flexneri 2a [J]. J Bacteriol, 1988, 170: 2480～2484

[24] Buysse J M, Hartman A B, Strockbine N, et al Genetic polymorphism of the ipaH multicopy antigen gene in shigella spps. and enteroinvasive escherichia coli [J]. Microb Pathog, 1995,19:335～349

[25] Rajaumar K, Jost BH, Sasakawa C, et al. Nucleotide sequence of the rhamnose biosynthetic operon of Shigella flexneri 2a and role of lipopolysaccharide in virulence [J]. J Bacteriol 1994 Apr, 176(8):2362～2373

[26] Okada N, Sasakawa C, Tobe T, et al. Virulence-associated chromosomal loci of *shigella flexneri* identified by random Tn5 insertion mutagenesis[J]. Mol Microbiol 1991 Jan,5(1):187～95

[27] Formal S B, Gemski P, Jr, et al. Genetic transfer of shigella flexneri antigens to escherichia coli K-12[J]. Infect. Immun, 1970, 1, 279～287

[28] Sansonetti P J, Hale T L, Dammin G J, et al. Alterations in the pathogenicity of escherichia coli K-12 after transfer of plasmid and chromosomal genes from shigella flexneri[J]. Infect Immun 1983 Mar, 39(3): 1392～402

# Complete Genome Sequence of *Shigella flexneri* 2a 301 Strain and Analysis of "*Shigella*-islands"

Liu Hong[1], Yang Fan[1], Zhang Xiaobing[1], Zhang Jiyu[1], Yang Guowei[1], Dong Jie[1], Xue Ying[1], Hou Yunde[1], Yuan Zhenghong[2], Wen Yumei[2], Xu Jianguo[3], Cheng Hongsong[4], Ma Dalong[4], Wang Yu[4], Yang Jian[5], Shen Yan[5], Qiang Boqin[5], Wu Hongtao[6], Lü Weichuan[6], Jin Qi[1,7]

(1. *State Key Laboratory for Molecular Virology and Genetic Engineering, Beijing* 100052, *China*; 2. *Laboratory of Molecular Virology, Fudan University, Shanghai* 200032, *China*; 3. *Institute of Epidemiology and Microbiology, Chinese Academy of Preventive Medicine, Beijing* 102206, *China*; 4. *Peking University Health Science Center, Beijing* 100083, *China*; 5. *National Center of human Genome research, Beijing* 100176, *China*; 6. *Huabei Pharmaceutical Co., Ltd. Shijiazhuang* 050000, *China*; 7. *Microbial Genome Center, Ministry of Public Health, Beijing* 100052, *China*)

[**Abstract**] *Shigella flexneri* serotype 2a are the most prevalent species and serotype that cause bacillary dysentery or shigellosis in man. This paper presents the complete genome sequence of a *Shigella flexneri* 2a strain which isolated from the Beijing outbreak, and the primary analysis of "*Shigella*-genomic islands (SIs)" that means *Shigella flexneri* 2a 301 strain-specific genome fragments. The whole genome is composed of a 4,607,203 bp chromosome and a 221,618 bp virulence plasmid, designated pCP301. The chromosome shares a conserved 'backbone' sequence about 4.03 Mb with those of a benign laboratory strain *E. coli* K12 (MG1655) which is essentially collinear. Sf301 has 572 Kb specific-sequence which form into 320 SIs with sizes greater than 50 bp and encoding in total 519 *Shigella*-specific Open Reading Frames (ORFs). Among these SIs, there are 131 islands with sizes greater than 1 Kb with repeated sequences of transposable elements, transposons or tRNAs flanking on one or both sides. The average G + C content of the SIs is 48.25%, significantly lower than that of the conserved backbone. Frequency of codons such as ACA, AAT, GCG, CTG, etc., on SIs are quite distinct from that on backbone sequences. All above observations together suggest that many of the SIs are foreign origin. Among them, the authors identified 7 putative SIs with typical structure of pathogenicity islands (PAI) and 2 SIs harbor some ORFs related to biosynthesis of lipopolysaccharide (LPS) have implications in virulence, in addition to the previously identified PAIs, SHE and SHI-2. The other SIs are mostly a mosaic of genes of known function and ORFs encoding polypeptides sharing none or low homology with known proteins from one or more bacterial species. All of these could be subjected to investigations towards novel preventive and treatment strategies against shigellosis.

[**Key words**] *Shigella flexneri* 2a 301 strain; genome sequence; genomic island; pathogenicity island