DOI:10.12113/202304017

Y1 转座酶关联转座子在大肠杆菌和沙门氏菌基因组中的分布和结构特征

王冰清,关中夏,王亚丽,石莎莎,郭梦可,宋成义,高　波*
(扬州大学 动物科学与技术学院,江苏 扬州 225002)

摘　要:Y1 转座酶关联转座子(Y1ATs)的活性催化位点为一个酪氨酸,能够切割和连接单链 DNA,在原核生物分布广泛。为探究 Y1 转座酶关联转座子在大肠杆菌(*Escherichia coli*, *E. coli*)与沙门氏菌(*Salmonella enterica*, *S. ente*)基因组中系统进化特性,通过 Hmmsearch 程序对 Y1 转座酶关联转座子进行了挖掘分析。结果表明,Y1 转座酶关联转座子广泛分布于96.84%大肠杆菌基因组和 80.4%沙门氏菌基因组。根据序列比对和蛋白结构域预测将 Y1 转座酶关联转座子分为 10 类,均隶属于 IS200/IS605 超家族,其中 11 645 个属于 IS200 家族,4 811 个属于 IS605 家族。IS200 家族广泛分布于 *S. ente* 基因组中(72.24%),而 IS605 家族广泛分布于 *E. coli* 基因组中(89.38%)。IS200 拷贝数以及完整拷贝数显著高于 IS605。IS200 家族仅含有一个 Y1 转座酶编码区,而 IS605 家族含两个开放阅读框,分别编码 Y1 转座酶和 TnpB 蛋白。IS200 家族的 Y1 氨基酸序列高度保守(95.3%),而 IS605 家族的 Y1 和 TnpB 具有较高遗传多样性,为研究转座子在原核生物的遗传进化模式提供重要参考。IS200 家族具有高度保守的 Y1 转座酶,且完整拷贝数比例较高,提示该类转座子可能具有转座活性,对其活性的挖掘有利于研制转座子介导的新型高效基因编辑工具。

关键词:转座子;Y1;TnpB;IS200;IS605;大肠杆菌(*E. coli*);沙门氏菌(*S. ente*)
中图分类号:Q311　　**文献标志码**:A　　**文章编号**:1672-5565(2024)03-217-08

Distribution and structural characteristics of Y1 transposase associated transposons in *Escherichia coli* and *Salmonella enterica* genomes

WANG Bingqing, GUAN Zhongxia, WANG Yali, SHI Shasha, GUO Mengke, SONG Chengyi, GAO Bo*
(*College of Animal Science and Technology*, *Yangzhou University*, *Yangzhou* 225002, *Jiangsu*, *China*)

Abstract: Y1 is a transposase that contains one tyrosine in its catalytic center, and it is capable of cleaving and ligating single-stranded DNA. Y1-associated transposons (Y1ATs) are widely distributed among prokaryotes. To explore the systematic evolutionary characteristics of Y1ATs in the genomes of two bacterial strains, *Escherichia coli* (*E. coli*) and *Salmonella enterica* (*S. ente*), which is done by mining and analyzing Y1ATs using the Hmmsearch program. The results indicate that Y1ATs are widely distributed in 96.84% of the *E. coli* genome and 80.4% of the *S. ente* genome. Based on sequence alignment and protein structure domain prediction, Y1ATs are classified into 10 classes, all belonging to the IS200/IS605 superfamily, with 11 645 belonging to the IS200 family and 4,811 belonging to the IS605 family. The IS200 family is widely distributed in the *S. ente* genome (72.24%), while the IS605 family is highly distributed in the *E. coli* genome (89.38%). The number of IS200 copies and intact copies are significantly higher than that of IS605. The IS200 family has only one Y1 transposase coding region, while the IS605 family has two open reading frames, encoding Y1 transposase and TnpB protein, respectively. The amino acid sequence of Y1 in the IS200 family is highly conserved (95.3%). In contrast, Y1 and TnpB in the IS605 family exhibit genetic diversity, which can provide significant insights into the genetic evolutionary pattern of transposons

收稿日期:2023-04-22;**修回日期**:2023-08-12;**网络首发日期**:2023-09-12.
网络首发地址:https://link.cnki.net/urlid/23.1513.Q.20230911.1032.004
基金项目:国家自然科学基金项目(No.32271508).
* **通信作者**:高波,女,教授,博导,研究方向:转座子挖掘及应用.E-mail:bgao@yzu.edu.cn.
引用格式:王冰清,关中夏,王亚丽,等.Y1 转座酶关联转座子在大肠杆菌和沙门氏菌基因组中的分布和结构特征[J].生物信息学,2024,22(3):217-224.
WANG Bingqing, GUAN Zhongxia, WANG Yali, et al.Distribution and structural characteristics of Y1 transposase associated transposons in *Escherichia coli* and *Salmonella enterica* genomes[J].Chinese Journal of Bioinformatics,2024,(3):217-224.

among prokaryotes. The IS200 family containing a highly conserves Y1 transposase with a high proportion of intact copies suggests that this type of transposon may have transpositional activity. Exploring its activity could be beneficial in developing novel and efficient transposon-mediated gene editing tools.

Keywords: Tansposon; Y1; TnpB; IS200; IS605; *Escherichia coli* (*E. coli*); *Salmonella enterica* (*S. ente*)

转座子为基因组中可以改变自身位置的独特 DNA 片段。研究表明,转座子几乎存在于所有生物的基因组中,可以发生转座并不断扩张,是基因组扩张的决定性因素,同时也对生物基因组结构和进化有着重要的影响[1]。原核转座子主要分为插入序列(Insertion sequence, IS),复合转座子及 TnA 家族。IS 成员众多、结构简单,仅携带与转座和调节有关的基因[2]。其中 IS200/IS605 家族广泛分布于细菌和古菌,迄今已鉴定出 153 余种成员[3]。IS200 作为家族创建者,最早发现于鼠伤寒沙门氏菌(*Salmonella typhimurium*)[4]。研究表明,IS200 有着稳定的分布和高拷贝数量[2]。IS605 最早发现于幽门螺旋杆菌(*Helicobacter pylori*),其编码的基因与 IS200 转座酶同源,因此统称为 IS200/IS605 家族[5]。该家族分子结构包括转座子左侧末端(Left end, LE)、右侧末端(Right end, RE),转座酶 TnpA 以及 TnpB 蛋白。TnpA 不具有经典的 IS 转座酶催化结构域 DDE 特征,而是 HuH 核酸内切酶家族的一员,包含一个保守的氨基酸三联体,由组氨酸(H)-巨型疏水残基(u)-组氨酸(H)构成。TnpA 转座酶能够切割和连接单链 DNA,其催化中心为单个酪氨酸(Y),因此又称为 Y1 转座酶[3]。与 CRISPR-Cas 系统中的 Cas9 和 Cas12 不同,Y1 转座酶不需要 RNA 引导转座过程。相反,Y1 通过自身的结构域识别并选择其靶标 DNA,并通过嵌入 DNA 和剪接目标 DNA 来实现转座。LE 和 RE 含有回文序列,可形成亚末端发卡结构[6]。Y1 特异地识别并结合这些短的末端二级结构,在特异位点剪切,形成环形单链 DNA 中间体,然后将其 3' 端插入目标单链 DNA 上富含 AT 的四核苷酸或五核苷酸位点,而 5' 端插入位点无特异性,靶位点不发生复制[3]。

IS200/IS605 家族编码多种 RNA 导向的核酸酶,目前已发现的包括 IscB 家族、IsrB 家族和 TnpB 家族[7]。研究表明,CRISPR-Cas9 起源于 IscB[8];TnpB 与 IscB 进化关系较远,被认为是 Cas12 的祖先[9-10]。另外,TnpB 可能也是一种较大蛋白质 Fanzors 的祖先,这种蛋白质被发现于多种真核转座子[11]。目前这三种核酸酶相关转座子系统的生物学功能仍未知,但推测这些核酸酶有利于 Y1 等转座酶催化反应、RNA 导向的转座,或者和转座子一起发挥抗毒素等作用,从而确保 IS200/IS605 插入基因组[7]。此外,也有研究发现,在耐辐射球菌以及大肠杆菌中,TnpB 对 IS*Dra*2 的切除具有抑制作用[12]。

目前研究表明,IS200/IS605 超家族分布广泛,但其在大肠杆菌(*Escherichia coli*)和沙门氏菌(*Salmonella enterica*)的分布种类、数量、结构和进化特性仍不明确,尤其是 Y1 关联转座子的挖掘,对于开发遗传编辑和流行病学研究工具潜力巨大。本研究分析了 Y1 关联转座子系统在 *E. coli* 和 *S. ente* 的遗传结构进化,可为挖掘 RNA 导向的转座子提供重要参考。

1 材料与方法

1.1 Y1 关联转座子挖掘

在美国国家生物技术信息中心(NCBI)网站(https://www.ncbi.nlm.nih.gov/)所提供的 whole-genome shotgun conting 数据库(WGS)中下载细菌蛋白 NR(非冗余)序列,收集 ISfinder(https://www-is.biotoul.fr/blast.php)中 TnpA 转座酶序列作为参考序列,使用 Hmmsearch 软件(v3.3.2)搜寻所有符合条件的 Y1 转座酶蛋白序列,设置 E 期望值为 1×10^{-4}。对于 Y1 转座酶含量最高的 *E. coli* 和 *S. ente*,使用 Usearch 程序对所获序列进行聚类,要求相似性 > 80%,并获得代表序列的 CDS 序列。在 Hmmsearch 软件中根据代表 CDS 序列收集获得 *E. coli* 及 *S. ente* 的所有基因组拷贝(E 期望值为 1×10^{-4}),前后延伸 1.8 kb 侧翼序列以保证序列的完整性。

1.2 插入序列识别

使用 MAFFT 程序进行多重序列比对[13],将 Y1 相关转座子进行分类。在 ISfinder 网站提供的数据库中进行序列 Blast 比对,通过与注释序列的比对获得转座子信息。使用 BioEdit 软件(v7.2.0)确定转座子的边界并进行序列截取[14],前后保留 50 bp 以研究其插入位点及侧翼序列特征。对于部分相似性高却未能在 ISfinder 上找到注释序列的转座子,使用 Hmmscan 网站(https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan)对其蛋白结构域进行分析,以确定其分类。若序列开放阅读框(orf)无 Y1 同源性,则舍弃该序列。将具有完整 LE、RE 且能够编码

100 aa 以上转座酶的序列视为完整转座子(IS605 中 *TnpA*>100 aa,*TnpB*>300 aa)。

1.3　转座子结构预测与进化分析

使用 Bioedit 软件提取多拷贝转座子的 LE,RE 以及 CDS。其中 LE 与 RE 使用 EMBOSS explorer 网站(https://www.bioinformatics.nl/emboss-explorer)构建一致序列,并在 Oligo Analyzer 网站(https://sg.idtdna.com/pages/tools/oligoanalyzer)中预测转座子 DNA 序列末端二级结构。通过 BioEdit 软件将 CDS 翻译为 Y1 转座酶和 TnpB 蛋白的序列,然后使用 EMBOSS explorer 网站构建它们的一致序列。同时根据得到的所有 Y1 转座酶蛋白序列,在 Weblogo(https://weblogo.threeplusone.com/create.cgi)网站中绘制序列 Logo 图,并对其进行结构与变异位点分析。使用 IBS 软件(v1.0.3)绘制转座子结构[15]。通过 Alpha-Fold 网站(https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb? authuser=0#scrollTo=kOblAo-xetgx)预测 Y1 转座酶结构,使用 PyMol 软件(v2.5.5)对其关键氨基酸残基及结构进行标注。

2　结果与分析

2.1　Y1ATs 广泛分布于 *E. coli* 和 *S. ente*

采用上述 *Hmmsearch* 序列收集方法,在 2 467 个 *E. coli* 基因组中挖掘到 8 645 条 Y1 关联序列(Y1ATs),分布于 2 389 个基因组(占比 96.84%),其中序列完整的 Y1ATs 占 55.62%;在 1 495 个 *S. ente*基因组中收集到 8 316 条 Y1ATs 序列,分布于 1 202 个基因组(占比 80.4%),完整序列占比高达93.13%(表 1)。Y1ATs 在 *E. coli* 中的含量略高于*S. ente*,但在 *S. ente* 中的完整序列比例远高于 *E. coli*。此外,相对于 *E. coli*,Y1ATs 在 *S. ente* 基因组中有着更高的拷贝数(表 1)。

2.2　Y1ATs 在 *E. coli* 和 *S. ente* 中的分类及完整拷贝的分布

通过 ISfinder 已注释序列比对,以及 Hmmscan 网站蛋白结构域预测,剔除非 Y1 相关序列后,在 *E. coli*和 *S. ente* 共鉴定到了 10 类 Y1ATs,均属于 IS200/IS605 超家族。2 类为 IS200(IS200C 和 IS200F),6 类为 IS605(IS609, IS*Ec*46, IS*Ec*41, IS*Sen*6,IS*Ec*44 和 IS*Kpn*69)。其中 7 种含有完整拷贝(*E. coli* 中的 IS200C, IS609, IS*Ec*46, IS*Ec*44, IS*Ec*41 以及*S. ente*中的 IS200F,IS*Sen*6),1 种仅有残缺拷贝(IS*Kpn*69)。还有两类序列在 ISFinder 中尚未注释,但在基因组中有着较高的序列一致性,长度在 2 500 bp 左右。虽然在 ISfinder 网站的 Blast 比对中难以找到其同源序列,但通过 Hmmscan 网站的蛋白分析,发现这两个类群的序列都包含 Y1 转座酶,IS*Ec*94 包含一个 165 aa 的 Y1 蛋白和一个 458 aa 的 FIhA 家族蛋白,而 IS*Ec*95 包含一个 200 aa 的 Y1, Y2 融合蛋白和一个 343 aa 的 dipZ 家族蛋白。因此将这两个类群的序列判定为新的 Y1 关联插入序列,根据 ISfinder 的命名规则将其命名为 IS*Ec*94 和 IS*Ec*95,拷贝数分别为 466 和 21,由于两者左右末端序列尚不清楚,难以界定是否有完整拷贝。

表 1　*E. coli* 和 *S. ente* 中 Y1ATs 的分布概况

Table 1　Distribution overview of Y1ATs in *E. coli* and *S. ente*

菌属	Y1ATs 个数	完整 Y1ATs 个数 /%	基因组数	含有 Y1ATs 的基因组数/%	单个基因组 Y1ATs 拷贝数	Y1ATs 基因组平均拷贝数
E. coli	8 645	4 808/55.62	2 467	2 389/96.84	1-31	3.62
S. ente	8 316	7 745/93.13	1 495	1 202/80.40	1-43	6.92

分析汇总含 Y1 完整拷贝的 IS200 和 IS605 转座子,发现在 *E. coli* 的 2 467 个基因组中 89.38%的基因组含有 IS605,33.04%的基因组含有 IS200,642 个基因组共同存在着 IS200 和 IS605 转座子(图 1(a))。*E. coli* 基因组中存在 4 种 IS605 转座子的重叠插入(图 1(b))。在 *S. ente* 的1 495个基因组中,含 IS200 的基因组占比达到 72.24%,含 IS605 的基因组占 20.27%,195 个基因组同时存在 IS200 和 IS605 转座子(图 1(c))。

统计分析 IS200 和 IS605 在 *E. coli* 和 *S. ente* 基因组中的平均拷贝数,结果如图 2 所示,IS200 的基因组拷贝显著高于 IS605。IS200 在 *S. ente* 和*E. coli* 基因组平均拷贝分别为 7.2 和 4.75。IS605 的平均拷贝均在 1 左右。根据拷贝数分布图可知,除 IS200 在少数基因组存在较高拷贝(约 35),Y1ATs 在大多数基因组中的拷贝数均较低(图 3)。

2.3　Y1ATs 的组成结构

IS200C 和 IS200F 全长大多在 710 bp 左右,中间为 Y1 转座酶编码序列(152 aa)(图 4(a))。个别序列会因为插入外源序列从而导致长度达到 2 287 bp,编码的蛋白也增加到 378 aa(表 2)。两端结构分为左末端(The left IS end, LE)和右末端(The right IS

end, RE),均含有回文序列,形成发卡和茎环结构,而不是经典 IS 的末端反向重复序列(TIR)(图 4(b),图 5)。IS200C 与 IS200F 左侧切割位点分别为 TTGT 和 TTTT,未包含在转座子中,位于 LE 的左侧;右侧切割位点均为 TTAT,包含在转座子中,位于 RE 末端。根据序列比对结果,推断 IS200 3'端偏好插入 T 富集区(图 6)。

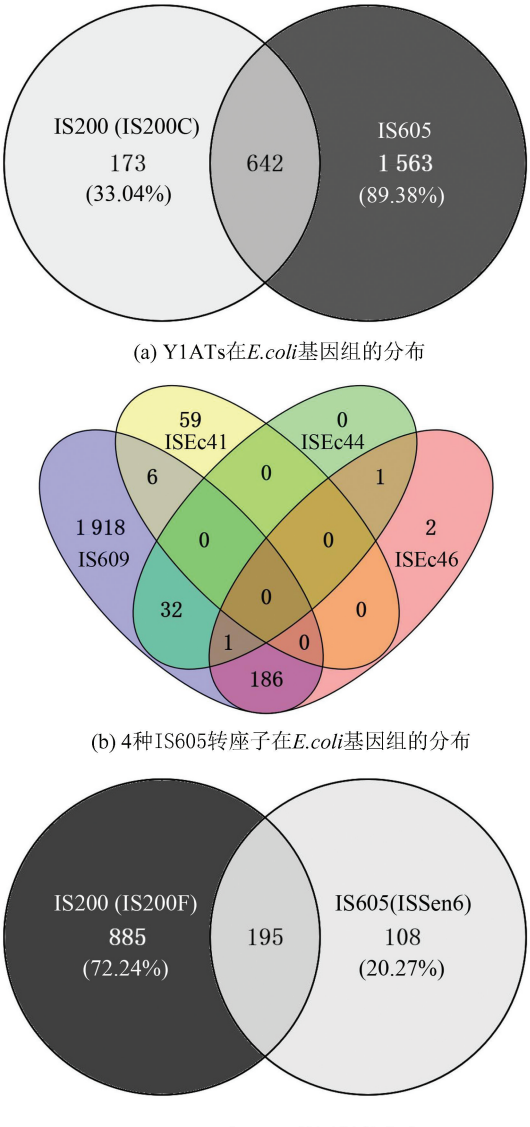


图 1 Y1ATs 完整拷贝在 *E. coli* 和 *S. ente* 基因组的分布概况

Fig.1 The distribution overview of complete Y1ATs copies in the genomes of *E. coli* and *S. ente*

注:重叠部分表示含有两种或多种 *Y1* 变体的基因组.

IS609,IS*Ec*41,IS*Ec*44,IS*Ec*46 和 IS*Sen*6 的完整拷贝总长为1 748-1 879 bp,中间为 Y1 和 TnpB 两个基因的编码序列,两者方向反向且无重叠序列(图 4(a))。两端分别为 LE 和 RE。Y1 编码约 143 aa,TnpB 编码约 400 个 aa。由于插入突变与重组的存在,个别转座子长达 3 220 bp,Y1 编码蛋白可增加到 332 个 aa。IS*Sen*6、IS*Ec*44 与 IS*Ec*41 三者左、右侧切割位点相同,分别为 CCAT 和 TCAA。IS609 左、右侧切割位点分别为 TTAT,TCAA;IS*Ec*46 左右切割位点分别为 TTAG,TCAC(表 2)。

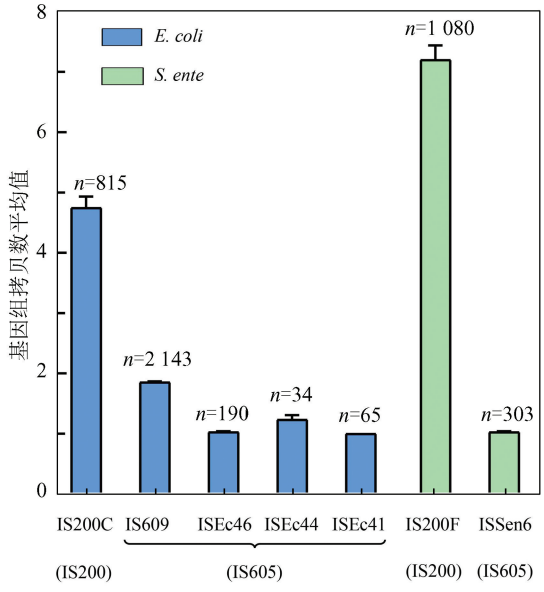


图 2 IS200 和 IS605 转座子在两个菌属的平均拷贝数比较

Fig. 2 Comparison of the average copy number of IS200 and IS605 transposons in two bacterial genera

注:X 轴代表 Y1ATs,Y 轴代表 Y1ATs 在基因组中拷贝数的平均值;n 表示基因组数量.

根据已报道的 Y1 转座酶晶体结构,们通过 Alpha-Fold网站以及 PyMol 软件对拷贝数及序列完整性较高的 IS200C 和 IS200F 进行了 Y1 转座酶蛋白结构图绘制和关键氨基酸残基的标注(图 4(c)、4(d))。Y1 转座酶的 HuH 基序由 H61 和 H63 组成,位于链 β5 上(图 4(c))。螺旋 α4 上的 Y125 是 Y1 转座酶中唯一严格保守的酪氨酸残基。HuH 基序和 Y125 是 TnpA 与 TIR 结合的核心结构域,它们形成催化位点,对于催化 DNA 单链中磷酸二酯键的断裂至关重要[16]。E56、H63 和四个水分子与 Mn2+离子配位,形成催化结构中必须的金属配位球[17]。H17,D60 和 H61 在维持金属结合位点方面也发挥着重要作用。K82 和 G83 对于茎环 DNA 的结合至关重要。此外,保守的 R25 可能在 Y1 转座酶发生构象变化时与 DNA 相互作用,因此可能是必不可少的[16]。在转座过程中,Y1 转座酶通过 β 折叠的合并形成二聚体(图 4(d))。

2.4 Y1 和 TnpB 蛋白的系统进化分析

使用 EMBOSS explorer 网站构建 *E. coli* 和 *S. ente* 中 IS200/IS605 超家族的 Y1 转座酶及 TnpB 蛋白一致序列(所用序列均为完整拷贝),并通过 Bioedit 软件计算各一致序列间相似性,结果如图 7 所示。IS200 家族

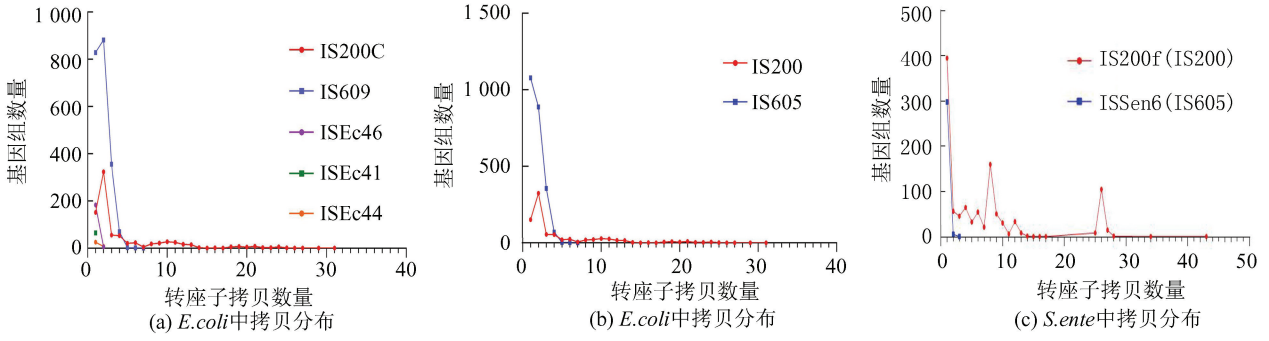
Y1 序列相似性高达 95.3%,说明其在进化过程中转座酶区域高度保守,可能来自同一个古老祖先。IS605 家族 Y1 平均相似性为47.92%±19.86%(26.8%~90.2%),TnpB 蛋白序列平均相似性为 51.42%±21.6%(25.9%~88.7%),提示两者遗传多样性均较高。此外,由表 2 可以看出,IS200 在基因组中的衍生以完整拷贝为主,而 IS605 在基因组中的衍生以残缺拷贝居多。使用 WebLogo 网站绘制完整拷贝 IS200 的 Y1 转座酶一致序列变异 Logo 图,其中 IS200C 约 3 000 条,遗传组成相对较为多样;IS200F 约 7 000 条,序列一致性较高。序列间变异情况见图 8。

表 2　*E. coli* 和 *S. ente* 中 Y1ATs 的分类和结构组成

Table 2　Classification and structural composition of Y1ATs in *E. coli* and *S. ente*

菌属	转座子家族	转座子名称	Y1ATs 数量			含有 Y1ATs 的基因组数/%	序列长度			LE 切割位点	RE 切割位点	序列长度变化		
			完整序列	残缺序列	总数		IS (bp)	Y1 (aa)	TnpB (aa)			IS (bp)	Y1 (aa)	TnpB (aa)
E. coli	IS200	IS200C	3 163	706	3 869	815/33.04	709	152	—	TTGT	TTAT	704-2 287	100-378	—
E. coli	IS605	IS609	1 424	2 551	3 975	2 143/86.87	1 748	143	382	TTAT	TCAA	1 743-3 220	100-332	305-499
E. coli	IS605	IS*Ec*46	186	10	196	190/7.70	1 763	141	389	TTAG	TCAC	1 755-1 771	102-137	335-389
E. coli	IS605	IS*Ec*41	35	30	65	65/2.63	1 841	150	401	CCAT	TCAA	1 828-1 842	117-150	316-401
E. coli	IS605	IS*Ec*44	40	2	42	34/1.38	1 879	134	401	CCAT	TCAA	1 740-1 975	134-138	401-410
E. coli	—	IS*Ec*94	—	—	466	463/18.77	2 606	165	—	—	—	1 810-2 609	—	—
E. coli	—	IS*Ec*95	—	—	21	8/0.32	2 462	—	—	—	—	1 652-2 462	—	—
E. coli	IS605	IS*Kpn*69	0	4	4	4/0.16	—	—	—	—	—	—	—	—
E. coli	—	other Y1ATs	0	7	7	7/0.28	—	—	—	—	—	—	—	—
S. ente	IS200	IS200F	7 602	174	7 776	1 080/72.24	710	152	—	TTTT	TTAT	703-2 073	103-293	—
S. ente	IS605	IS*Sen*6	143	170	313	303/20.27	1 818	146	402	CCAT	TCAA	1 814-1 857	104-143	315-402
S. ente	IS605	IS*Ec*46	0	207	207	204/13.65	—	—	—	—	—	—	—	—
S. ente	IS605	IS609	0	9	9	6/0.40	—	—	—	—	—	—	—	—
S. ente	—	other Y1ATs	0	11	11	11/0.74	—	—	—	—	—	—	—	—

注:-表示未发现或条件尚不足以定义.

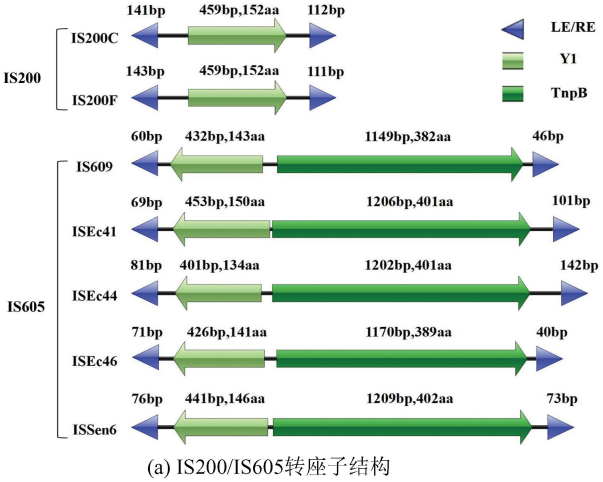


(a) *E.coli*中拷贝分布　(b) *E.coli*中拷贝分布　(c) *S.ente*中拷贝分布

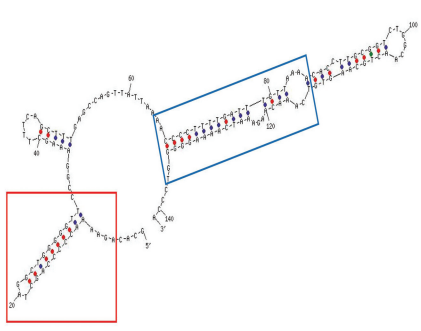
图 3　*E. coli* 和 *S. ente* 中 Y1ATs 各家族拷贝分布

Fig.3　Copy distribution of Y1ATs families in *E. coli* and *S. ente*

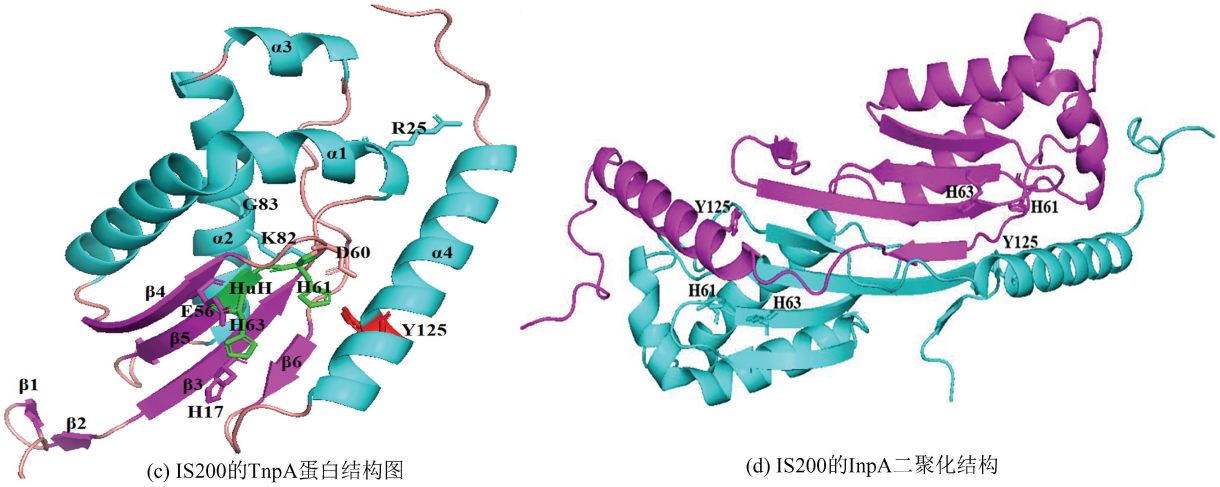
注:图(*a*)中 *IS200C* 属于 *IS200* 家族,*IS609*、*ISEc46*、*ISEc41*、*ISEc44* 属于 *IS605* 家族.



(a) IS200/IS605转座子结构



(b) IS200C的LE亚末端发夹与茎环结构,红框表示为5'端发卡结构,蓝框表示3'端茎环结构



(c) IS200的TnpA蛋白结构图　　(d) IS200的InpA二聚化结构

图 4　IS200 和 IS605 家族转座子结构示意图

Fig. 4　Schematic diagram illustrating the structure of the IS200 and IS605 transposon families

注:图(c)中 α 螺旋和 β 折叠分别用蓝色、粉色标出.绿色代表 HuH 基序,红色代表保守的 Y125 酪氨酸残基.

>IS200C LE

GCACAGAAAACCCCCCAGCTAGGCTGGGGGTTCCGGAAAGCTTTCAGCTTTGAGCCAGTTATTAAAACCCCTTTTGA

TTTGTTAAAACACCTTGCGGTCTGGCAACTGCAAGTGTCAAACAAGAAATCAAAAGGGGGTCCCA

>IS200F LE

GTCTATGGAAAACCCCCCAGCTAGGCTGGGGGTTCCGGAAAGCTTTCAGCTTTAAGCCAGTTATTAAAACCCCTTTTG

ATTTGTTAAAACATCTTGCGGTCTGGCAACTGCAAAAGTTCAACAAGAAATCAAAAGGGGGTCCCA

图 5　IS200C、IS200F 的 LE 亚末端发夹与茎环结构 DNA 序列图

Fig. 5　DNA sequence diagram of LE subterminal hairpin and stem-loop structure of IS200C and IS200F

注:蓝色部分配对形成发卡结构,绿色部分配对形成茎环结构.

图 6　IS200 转座子左侧末端 LE 的偏好插入位点

Fig. 6　The preferred insertion site of LE at the left end of IS200 transposon

注:黑色区域为 IS200 转座子序列起始位置,ttgt 为左侧切割位点,LE 偏好插入 T 富集区.

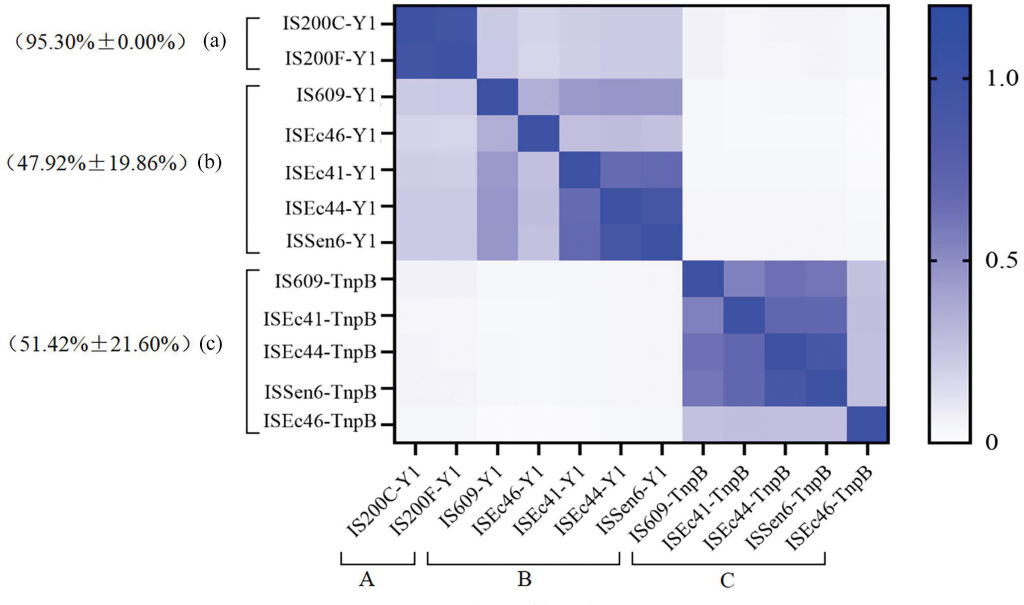


图 7　*E. coli* 和 *S. ente* 中的 Y1 转座酶和 TnpB 蛋白进化特性

Fig. 7　Evolutionary characteristics of Y1 transposases and TnpB proteins in *E. coli* and *S. ente*

注:(*a*)代表 *IS*200 的 *Y1* 转座酶相似性;(*b*)代表 *IS*605 的 *Y1* 转座酶相似性;(*c*)代表 *TnpB* 蛋白相似性.蓝色从深到浅表示序列相似性从高到低,序列平均相似性在左侧标出.

图 8　IS200 的 Y1 转座酶变异 Logo 图

Fig.8　Logo diagram of IS200 Y1 transposases mutation

注:字母大小表示其在序列中的占比.

3　讨　论

本研究结果表明,Y1ATs 广泛分布于 *E. coli* 和 *S. ente* 两个菌属(占比>80%),提示 Y1ATs 存在大规模传播,并可能存在水平传播现象。IS200 和 IS605 的拷贝数在两个菌属存在着显著差异,IS200 在 *S. ente* 基因组中传播更广泛,而 IS605 在 *E. coli* 基因组中传播更为广泛,这说明在 *E. coli* 和 *S. ente* 中存在着 Y1ATs 进化动力学的差异[18]。IS 通过基因组重排和有益基因的转移增加了遗传多样性和基因组可塑性[19]。IS200 在基因组上的拷贝数显著高于 IS605,其完整拷贝数也显著高于 IS605。但研究表明 IS200 对其原始宿主沙门氏菌造成的自发变异很小,转座较为罕见[20]。原因之一可能是由于 IS200LE 端发夹结构与启动子重叠,导致了启动子的低表达[21]。IS200 最初被视为惰性转座子的典例,但它在细菌属中的分布表明其有很长的进化历史[22],推测是 IS200 低频次的转座给自身提供了一定的进化优势。IS200 的稳定分布与高拷贝与前人的研究结果一致,其高丰度提示可能存在功能性转座元件,为进一步挖掘活性 Y1ATs 奠定基础。

转座子的拷贝数与转座子的进化模式及插入年龄有重要关系[23]。在物种进化过程中,伴随着时间的推移,转座子累积了突变、重排等,从而导致转座酶的逐渐失活,转座子的拷贝数也随之降低[24]。IS200C 与 IS200F 较高的序列一致性(95.3%)表明其很有可能由同一个祖先进化而来(图 7)。相较 IS200C,IS200F 在转座酶区域更为保守(图 8),这一定程度上解释了 IS200F 的基因组拷贝数量高于 IS200C 的现象。IS200 转座子在转座酶区域较为保守,这一点可以作为 IS200 是活性转座子的依据之一。相比而言,IS605 累积了较多的突变、缺失和重组。IS609 和 IS*Sen*6 分别是 *E. coli* 和 *S. ente* 两个物种中含量最高的 IS605 转座子,但它们的残缺拷贝数量甚至都超过了完整拷贝(表 2),推测是 IS605 转座子在进化过程中的不断突变与非法重组导致的。

进一步研究单编码框 IS200 家族和双编码框 IS605 家族,有利于了解细菌转座元件进化过程、转座及调控机制。本研究将 *E. coli* 和 S. *ente* 中 Y1ATs 共分为 10 类,IS200 仅编码 Y1 转座酶,是已知最小的自主 IS。IS605 则编码 Y1 和 TnpB。研究表明,Y1 催化 IS 的裂解、连接等转座活动,而 TnpB 对转座活动的作用机制尚不清楚,甚至可能起抑制作用[12],这可能与 IS605 低拷贝特性相关。研究表明 TnpB 具有 RNA 导向的核酸酶活性,本研究对其序列结构的遗传进化进行了初步分析,为挖掘靶向整合转座子提供参考。

在序列的 Mafft 多重比对中,发现 *E. coli* 中的 IS*Ec*94,IS*Ec*95 高度相似,但这些序列在 ISfinder 的 Blast 比对中难以找到相应的注释序列。将这部分序列翻译为蛋白质,在 Hmmscan 网站上进行转座酶蛋白序列的比对,比对结果显示这些序列编码>100 aa 的 Y1 蛋白,属于 IS200 家族成员。由此可见,一些 Y1 关联转座子难以被插入序列注释软件精准识别,有着未经注释的独特序列片段,此现象一定程度上表明了序列间可能正在发生着重组[25]。

4　结　论

本研究在 *E. coli* 和 *S. ente* 两个菌属中挖掘到大量 Y1 关联转座子,其在进化关系上隶属于 IS200/IS605 家族。IS200 和 IS605 存在不同的进化模式,IS200 存在较多拷贝及完整拷贝,且各亚家族间 Y1 转座酶序列高度保守;IS605 完整拷贝数较低,各亚家族间 Y1 转座酶和 TnpB 变异较大。本研究为揭示原核生物转座子进化机制和进一步挖掘活性 Y1 关联转座子提供重要参考。

参考文献(References)

[1]沈丹, 陈才, 王赛赛, 等. Tc1/Mariner 转座子超家族的研究进展[J]. 遗传, 2017, 39(1): 13. DOI: 10.16288/j.yczz.16-160.
SHEN Dan, CHEN Cai, WANG Saisai, et al. Research progress on the Tc1/Mariner transposon superfamily[J]. Genetics, 2017, 39(1): 13. DOI: 10.16288/j.yczz.16-160.

[2]BEUZÓN C R, CHESSA D, CASADESÚS J. IS200: An old and still bacterial transposon [J]. International Microbiology, 2004, 7(1): 3-12. DOI: 10.2436/im.v7i1.9438.

[3] HE S, CORNELOUP A, GUYNET C, et al. The IS200/IS605 family and "peel and paste" single-strand transposition mechanism[J]. Microbiology Spectrum, 2015, 3(4): 609-630. DOI:10.1128/microbiolspec.MDNA3-0039-2014.

[4] LAM S, ROTH J R. IS200: A salmonella-specific insertion sequence[J]. Cell, 1983, 34(3): 951-960. DOI: 10.1016/0092-8674(83)90552-4.

[5] KERSULYTE D, AKOPYANTS N S, CLIFTON S W, et al. Novel sequence organization and insertion specificity of IS605 and IS606: Chimaeric transposable elements of Helicobacter pylori[J]. Gene, 1998, 223(1-2): 175-186. DOI: 10.1016/s0378-1119(98)00164-4.

[6] BEUZÓN C R, CASADESÚS J. Cloning with Mud-P22 hybrid prophages: Mapping of IS200 elements on the chromosome of Salmonella typhimurium LT2[J]. Molecular and General Genetics, 1997, 256(5): 586-588. DOI: 10.1007/s004380050605.

[7] ALTAETRAN H, KANNAN S, DEMIRCIOGLU F E, et al. The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonu-cleases[J]. Science, 2021, 374(6563): 57-65. DOI: 10.1126/science.abj6856.

[8] KAPITONOV V V, MAKAROVA K S, KOONIN E V. ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs[J]. Journal of Bacteriology, 2016, 198(5): 797-807. DOI: 10.1128/JB.00783-15.

[9] GUERILLOT R, SIGUIER P, GOURBEYRE E, et al. The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries[J]. Genome Biology and Evolution, 2014, 6(2): 260-272. DOI: 10.1093/gbe/evu010.

[10] SHMAKOV S, SMARGON A, SCOTT D, et al. Diversity and evolution of class 2 CRISPR-Cas systems[J]. Nature Reviews Microbiology, 2017, 15(3): 169-182. DOI: 10.1038/nrmicro.2016.184.

[11] BAO W, JURKA J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements[J]. Mobile DNA, 2013, 4(1): 12. DOI: 10.1186/1759-8753-4-12.

[12] PASTERNAK C, DULERMO R, TON-HOANG B, et al. ISDra2 transposition in Deinococcus radiodurans is down-regulated by TnpB[J]. Molecular Microbiology, 2013, 88(2): 443-455. DOI: 10.1111/mmi.12194.

[13] TORRE E, THRELFALL E J, HAMPTON M D, et al. Characterization of Salmonella virchow phage types by plasmid profile and IS200 distribution[J]. Journal of Applied Bacteriology, 1993, 75(5): 435-440. DOI: 10.1111/j.1365-2672.1993.tb02799.x.

[14] YANG Peng, CRAIG P A, GOODSELL D, et al. BioEditor-simplifying macro-molecular structure annotation[J]. Bioinformatics, 2003, 19(7): 897-898. DOI: 10.1093/bioinformatics/btg103.

[15] LIU Wenzhong, XIE Yubin, MA Jiyong, et al. IBS: an illustrator for the presentation and visualization of biological sequences[J]. Bioinformatics, 2015, 31(20): 3359-3361. DOI: 10.1093/bioinformatics/btv362.

[16] RONNING D R, GUYNET C, TON-HOANG B, et al. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases[J]. Molecular Cell, 2005, 20(1): 143-154. DOI: 10.1016/j.molcel.2005.07.026.

[17] LEE H H, YOON J Y, KIM H S, et al. Crystal structure of a metal ion-bound IS200 transposase[J]. The Journal of Biological Chemistry, 2006, 281(7): 4261-4266. DOI: 10.1074/jbc.M511567200.

[18] LIU Yibing, ZONG Wencheng, DIABY M, et al. Diversity and evolution of pogo and Tc1/mariner transposons in the apoidea genome[J]. Biology, 2021, 10(9): 940. DOI: 10.3390/BIOLOGY10090940.

[19] CERVEAU N, LECLERCQ S, BOUCHON D, et al. Evolutionary dynamics and genomic impact of prokaryote transposable elements[M]. Berlin: Springer-Verlag Berlin Heidelberg, 2011, 291-312. DOI: 10.1007/978-3-642-20763-1_17.

[20] LAM S, ROTH J R. Structural and functional studies of insertion element IS-200[J]. Journal of Molecular Biology, 1986, 187(2): 157-167. DOI: 10.1016/0022-2836(86)90225-1.

[21] CALVA E, ORDOÑEZ L G, FERNANDEZ-MORA M, et al. Distinctive IS200 insertion between gyrA and rcsC genes in Salmonella typhi[J]. Journal of Clinical Microbiology, 1997, 35(12): 3048-3053. DOI: 10.1128/jcm.35.12.3048-3053.1997.

[22] VAN-VALEN L.Evolutionary Genetics[J]. Science,1962,138(3538):424. DOI:10.1126/science.138.3538.424.

[23] 沈丹. Tc1/mariner转座子挖掘,高活性成员鉴定及其在增强子捕获中的应用[D]. 扬州: 扬州大学, 2021. DOI: 10.27441/d.cnki.gyzdu.2020.000081.
SHEN Dan. Tc1/Mariner transposon mining, high-activity member identification and its application in enhancer capture[D]. Yangzhou: Yangzhou University, 2021. DOI: 10.27441/d.cnki.gyzdu.2020.000081.

[24] HE Susu, GUYNET C, SIGUIER P, et al. IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer[J]. Nucleic Acids Research, 2013, 41(5): 3302-3313. DOI: 10.1093/nar/gkt014.

[25] SADLER M, MORMILE M R, FRANK R L. Characterization of the IS200/IS605 insertion sequence family in halanaerobium hydrogeniformans[J]. Genes, 2020, 11(5): 484. DOI: 10.3390/genes11050484.