岳　鹏,高海琴,李　哲,等. 2 种常见 CRISPR – Cas 系统在苹果中的适用性特征[J]. 江苏农业科学,2022,50(7):43 – 51.
doi:10.15889/j.issn.1002 – 1302.2022.07.006

2 种常见 CRISPR – Cas 系统在苹果中的适用性特征

岳　鹏[1], 高海琴[1], 李　哲[2], 米志波[3], 郑　博[1], 赵彦敏[1]
(1. 张家口开放大学,河北张家口 075000; 2. 张家口市第六中学,河北张家口 075000;
3. 河北省张家口市涉密载体管理中心,河北张家口 075000)

摘要:运用生物信息学方法初步分析 CRISPR – Cas9 和 Cpf1 在苹果中的整体适用性特征,以期为苹果基因组编辑和 CRISPR – Cas 在苹果研究中的推广使用提供一定的参考和便利。结果表明,苹果染色体中有数量可观的 PAM,平均间隔 7 bp 碱基有 1 个 5′ – NGG、间隔 3 bp 有 1 个 5′ – TTN;也就是说 5′ – TTN 比 5′ – NGG 的出现频率高。SpCas9、FnCpf1 分别有 29.0% 、26.9% 的作用位点几乎覆盖了所有染色体基因,个别不能被 SpCas9 识别的基因能被 FnCpf1 识别,反之亦然。苹果的 CRISPR 靶序列有大量重复,单一靶序列被视为能被 Cas 蛋白特异识别并有效编辑。在靶序列长度为 20 nt 时,99.5% 的染色体基因可至少被其中 1 种 Cas 蛋白编辑,分别具有不同的可编辑度搭配;其中的 237 个基因只能被 1 种 Cas 蛋白编辑,填补了另一种 Cas 蛋白留下的编辑空白,另有 220 个染色体基因(0.5%)不能被任一种 Cas 蛋白编辑,即 2 种 Cas 蛋白同时留下编辑空白,没有互补。

关键词:苹果;CRISPR – Cas;适用性特征;PAM 出现频率;基因编辑
中图分类号:S661.101　　**文献标志码**:A　　**文章编号**:1002 – 1302(2022)07 – 0043 – 08

苹果(*Malus domestica*)是世界上主要果树作物之一,其产量和质量易受生物和非生物胁迫的影响[1]。因此,了解抗逆性相关基因的功能及调控规律,对培育抗逆性强的品种至关重要[2]。与其他作物相比,果树具有高度杂合的多倍体基因组且繁育周期长,导致传统的育种研究进展缓慢[3];而随着基因组测序工作的完成,对其基因结构、基因通路和基因功能的认识为基因编辑奠定了基础[4]。高效、易用、省时、低成本的基因编辑技术是赋予果树重要经济性状的捷径。常间四文重复序列丛集关联蛋白[CRISPR(clustered regularly interspaced short palindromic repeat) – Cas (CRISPR – associated proteins)]系统具备以上优势,在研究生物多种类的精确分子机制方面具有极高的应用价值[5]。

当前,基因组编辑技术主要集中在第 2 类 CRISPR – Cas 系统[6];它可以使用单个效应蛋白剪切 DNA,包括Ⅱ型、Ⅴ型、Ⅵ型[7]。其中,Ⅱ型的 Cas9 及其失去剪切活性的衍生品 dCas9 被广泛应用于多种生物体的基因组操作,包括靶向基因干扰、转录激活抑制、表观遗传修饰及目标碱基对转换[8-9]。经典的 Cas9 蛋白须识别衔接在靶序列(+)3′端形如 5′ – NGG 的前间隔序列邻近基序(protospacer – adjacent motif, PAM, +),并在 PAM 上游第 3 个碱基处切割双链,形成平头末端;在人类基因组中约平均间隔 8 bp 就会有 1 个 5′ – NGG[10]。不同于此,Ⅴ型的 Cpf1(Cas12a)要识别毗邻在靶序列(+)5′端形似 5′ – TTN 的 PAM(+),并在 PAM 下游的同向链(+)第 18、互补链(–)第 23 碱基处交错切割,形成有 5 个突出碱基的黏性末端[11]。作为新型且更小的 CRISPR 效应蛋白,Cpf1 的开发利用有利于突破和克服 Cas9 在使用中的一些限制,尤其是拓宽了 CRISPR – Cas 系统的识别范围,使之能更有效地编辑富含 AT 碱基的基因组[12]。

CRISPR – Cas9 已被大量使用在植物基因功能分析及育种研究中[13-15],其中包括多种果树[14]。有关苹果的几项研究,依次是首次利用 CRISPR – Cas9 敲除内源 *PDS*(phytoene desaturase)基因[16],高效传送 CRISPR – Cas9 核糖核蛋白到苹果原生质体操纵 *DIPM1*(*DspA/E* – interacting proteins of *M.* ×

收稿日期:2021 – 06 – 04
基金项目:中国成人教育协会“十四五”成人继续教育科研规划重点课题(编号:2021 – 323ZB);河北省张家口市科技项目(编号:18110300043);河北省张家口市社会科学立项研究课题(编号:2021121)。
作者简介:岳　鹏(1984—),男,河北张家口人,硕士,讲师,从事农林生命科学类和开放教育研究。E – mail:yppolymerase@foxmail.com。
通信作者:赵彦敏,博士,副教授,从事农林生命科学类和开放教育研究。E – mail:zym319@163.com。

domestica)、*DIPM2* 和 *DIPM4* 基因[17],优化 CRISPR – Cas9 的应用条件并敲除 *PDS* 和 *TFL1*(terminal flower)基因[3],运用 CRISPR – Cas9 敲除 *DIPM4* 基因的同时又减少外源 DNA 的残留[18]。相对地,CRISPR – Cpf1 的技术特点更加鲜明,但在 2016 年首次应用于水稻和烟草后[19],在植物基因组项目中使用的报道较少,目前还未见于果树研究[2]。

已有的研究结果表明,CRISPR – Cas 系统可实际应用于苹果基因编辑,但只关注了少数几个基因,且仅限于使用 CRISPR – Cas9。本研究对苹果全基因组序列进行初步分析,尝试探索 2 种流行的 CRISPR – Cas 系统,即 CRISPR – Cas9 和 CRISPR – Cpf1,在苹果基因组编辑中的整体适用性,首次从 PAM 的数量和频率、靶序列的重复和分布及 2 种 Cas 蛋白的互补等几个方面开展讨论,并形成 PAM 位点和靶序列信息库,以期为苹果基因组编辑和 CRISPR – Cas 系统在苹果研究中的推广使用提供一定的参考和便利。

1　材料与方法

1.1　数据获取

苹果全基因组代表性数据 ASM211411v1 下载自 NCBI 网站[20],17 条染色体(chromosome,chr)和 1 条线粒体(mitochondrion,mt)的 DNA 序列以 FASTA 格式分别存储在文本文件中;所有下载的 DNA 都只有单链(+)。相应的基因文件也从 NCBI 网站下载,其中以列表形式记录基因在 DNA 序列上的起止位点等信息。最新发布在 NCBI 网站上的苹果基因组数据 ASM411538v1 质量更高[21],但相应的基因信息不够完善,只在本研究中作对比和补充。在诸多 CRISPR – Cas 系统中,效应蛋白 SpCas9[化脓性链球菌(*Streptococcus pyogenes*,Sp)]和 FnCpf1[弗郎西斯菌(*Francisella novicida*,Fn)]识别的 PAM 较有代表性,分别为 5′ – NGG 和 5′ – TTN[10-11];本研究围绕这 2 种典型的 PAM,尝试使用个人计算机分析苹果全基因组。

1.2　PAM 计数和间距计算

对苹果各 DNA 序列出现的 5′ – NGG 或 5′ – TTN 计数,并记录中间碱基在 DNA 序列上的位点作为 PAM 位点;还需要同时累计 5′ – CCN 或 5′ – NAA,以实现对互补链的搜索。计数期间,单独计算 N 所代表的各种碱基占比,并用 PAM 位点数量与 DNA 长度的比值表示序列的 PAM 密度。除线粒体外,合计染色体 DNA 的各项数据以考量全基因组。

同样地,分别计算各 DNA 序列和全基因组的 PAM 出现频率。将每 2 个相邻 PAM 位点之差作为间距(用字母 d 表示),其意义为间隔 d 个碱基对存在 1 个 PAM,代表 PAM 的出现频率;并累计每个 d 的出现次数(用字母 n 表示),再把全部 d 升序排列(以 t 为排列序号),记录不同 d 的数量(用字母 m 表示)。PAM 出现频率的均值(d_{mean})和中值(d_{median})计算如下:

$$d_{\mathrm{mean}}=\sum_{i=1}^{m}d_in_i\Big/\sum_{i=1}^{m}n_i;\text{若}\sum_{i=1}^{t}n_i\geqslant\sum_{i=1}^{m}n_i\Big/2>\sum_{i=1}^{t-1}n_i,d_{\mathrm{median}}=d_t。$$

1.3　计算剪切位点

在 DNA 序列上从 5′ – NGG 前溯 3 个碱基或从 5′ – CCN 后推 3 个碱基,获得 Cas9 剪切位点。通过判断剪切位点是否处于基因的起止位点之间,将对应的 PAM 位点划入相应的基因范畴;不属于任何基因范畴的 PAM 位点不做标记。

依据 Cpf1 的剪切特征,在 DNA 序列上从 5′ – TTN 后推 18 个碱基或从 5′ – NAA 前溯 23 个碱基,获得同向剪切位点,用来判断相应的 PAM 位点是否属于同向链基因;从 5′ – TTN 后推 23 个碱基或从 5′ – NAA 前溯 18 个碱基,获得逆向剪切位点,用来判断相应的 PAM 位点是否属于互补链基因。

1.4　靶序列的截取

根据 2 种 PAM 的特点,在 DNA 序列上分别截取长度为 20 nt(不含 PAM)的靶序列(target),用 PAM 位点命名,以 FASTA 格式储存;PAM 在互补链上的,还需要按照碱基配对原则转换碱基并逆序。对具有同种 PAM 的靶序列进行重复性搜索,找到单一序列(singleton)和相同序列簇(cluster);簇中的 PAM 位点有属于染色体基因的计 1 分,有属于染色体基因间隔的计 2 分,有属于线粒体基因的计 4 分,有属于线粒体基因间隔的计 8 分,4 种分值任意组合可将所有簇归入 15 个重复类(用 repeat N 命名,N 为不大于 15 的正整数)。重点关注靶序列单一且属于基因范畴的 PAM 位点,其数量与基因长度的比值表示基因的可编辑度;按照“基因 ID – PAM 位点 – 靶序列”的模式建立简易信息库,找到可编辑度最高和最低的基因。

1.5　程序实现

以上操作已被整合到几个 Perl 脚本中,并尽量优化算法降低时间复杂度和空间复杂度[22]。其中,为避免耗费大量时间,在判断 PAM 位点是否影响基

因时,用数组模拟 DNA 序列,数组索引表示位点,基因区间内的数组元素由基因 ID 填充,其他为未定义(undef)值;当剪切位点对应的数组元素为基因 ID 时,将 PAM 位点划入此基因范畴。为避免占用大量内存空间导致程序卡顿,在搜索重复靶序列时,把总文件分割成大小合适的几个子文件,使用散列快速剔除子文件内的相同序列;然后利用每个子文件中序列唯一的特点,使用散列剔除子文件间的相同序列,最终合并成没有重复序列的总文件。

2　结果与分析

2.1　PAM 含量分析

苹果基因组(ASM211411v1)中 5′ – NGG 总量为 48 368 223,chr15 的数量最多(4 179 083),chr01 的最少(2 291 346),各染色体间差异较大;再结合序列长度估算序列的 PAM 密度,基因组是 0.074,chr10 最高(0.078),chr06 最低(0.069),相差不大(表 1)。在基因组中,CGG、GGG、AGG 和 TGG 在 NGG 总量中的比例分别为 11.9%、21.4%、30.4%、36.3%(图 1 – A);在各染色体中的比例与此一致,略有微小波动(表 1)。线粒体中的 5′ – NGG 含量为 46 468,但密度达到 0.117,高于基因组最高值;CGG、GGG、AGG 和 TGG 的比例分别为 19.2%、25.9%、30.2%、24.7%,与基因组中的同项数据也有明显差异(图 1 – B)。

基因组中 5′ – TTN 的总数为 127 635 154,同样是 chr15 最多(11 092 974),chr01 最少(5 902 134);估算基因组的密度是 0.194,依然是 chr10 最高(0.206),chr06 最低(0.180),两者比较接近(表 2)。在基因组中,TTC、TTA、TTG 和 TTT 在 TTN 总量中的比例分别为 19.6%、21.2%、22.2%、37.0%(图 1 – C);同样,在各染色体中的比例与此一致(表 2)。线粒体中 5′ – TTN 含量为 69 296,密度只有 0.174 6,低于基因组最低值;TTC、TTA、TTG 和 TTT 的比例分别为 28.4%、18.7%、20.7%、32.2%,与基因组中的同项数据也差异明显(图 1 – D)。

作为对比,另一基因组数据(ASM411538v1)的 5′ – NGG 密度为 0.084,CGG、GGG、AGG 和 TGG 的比例分别为 11.7%、21.5%、30.6%、36.3%;5′ – TTN 的密度为 0.219,TTC、TTA、TTG 和 TTT 的比例分别是 19.6%、21.2%、22.4%、36.8%。使用 2 个不同版本的基因组数据计算出的结果很接近,特别是 NGG 和 TTN 的构成比例几乎完全一致,可以用同一组饼状图表示(图 1 – A、图 1 – C)。

2.2　PAM 出现频率

在苹果基因组(ASM211411v1)中,相邻5′ –

表 1　苹果基因组中的 5′ – NGG

染色体	L_{seq} (bp)	N_{NGG}	N_{NGG}/L_{seq}	P_{CGG} (%)	P_{GGG} (%)	P_{AGG} (%)	P_{TGG} (%)	d_{mean} (bp)	d_{median} (bp)	d_{max} (bp)
chr01	32 709 648	2 291 346	0.070	12.0	21.4	30.4	36.2	11.838	7	760
chr02	37 631 755	2 846 361	0.076	12.0	21.4	30.3	36.4	12.065	7	1 069
chr03	37 690 471	2 757 325	0.073	11.8	21.5	30.3	36.4	12.092	7	999
chr04	32 357 154	2 344 270	0.072	11.9	21.4	30.4	36.2	11.979	7	629
chr05	48 068 851	3 462 956	0.072	11.9	21.4	30.4	36.4	11.975	7	1 045
chr06	37 231 166	2 557 225	0.069	11.8	21.4	30.5	36.3	11.948	7	641
chr07	36 738 692	2 805 236	0.076	12.0	21.5	30.2	36.3	12.006	7	1 347
chr08	31 666 303	2 426 929	0.077	12.1	21.4	30.4	36.1	11.980	7	482
chr09	37 676 754	2 744 725	0.073	11.9	21.5	30.4	36.2	11.964	7	987
chr10	41 841 605	3 254 974	0.078	11.9	21.5	30.3	36.3	12.038	7	952
chr11	42 925 075	3 050 822	0.071	11.7	21.4	30.4	36.5	12.105	7	842
chr12	33 134 071	2 471 075	0.075	11.9	21.4	30.4	36.3	12.078	7	1 105
chr13	44 437 459	3 232 321	0.073	11.6	21.4	30.7	36.3	11.822	7	670
chr14	32 560 231	2 332 518	0.072	11.8	21.4	30.5	36.3	12.099	7	1 012
chr15	55 080 361	4 179 083	0.076	11.9	21.4	30.2	36.4	12.035	7	1 002
chr16	41 441 581	3 044 581	0.073	11.6	21.5	30.6	36.3	11.847	7	627
chr17	34 817 048	2 566 476	0.074	11.8	21.5	30.4	36.3	12.100	7	855

注:L_{seq} 为序列长度;N_{NGG} 为 5′ – NGG 的数量;N_{NGG}/L_{seq} 表示 PAM 密度;P_{NGG} 表示各 NGG 在总数中的占比;d_{mean} 为间距平均值;d_{median} 为间距中值;d_{max} 为间距最大值。

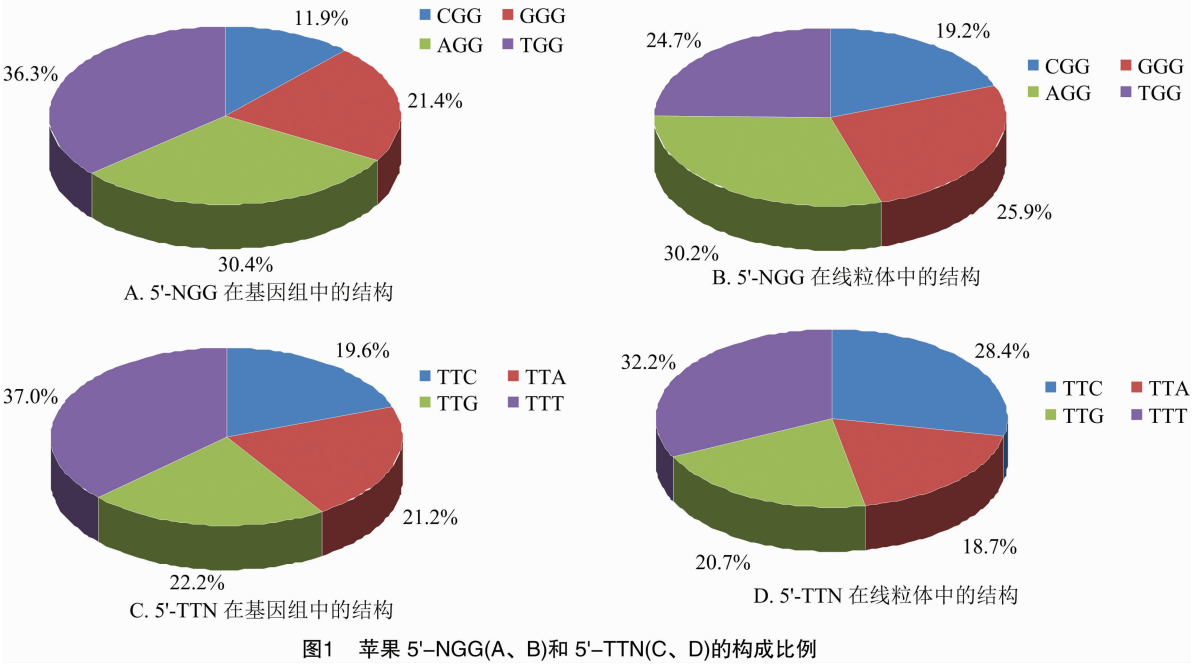


图1　苹果 5'–NGG(A、B)和 5'–TTN(C、D)的构成比例

表 2　苹果基因组中的 5′ – TTN

chromosome	L_{seq} (bp)	N_{TTN}	N_{TTN}/L_{seq}	P_{TTC} (%)	P_{TTA} (%)	P_{TTG} (%)	P_{TTT} (%)	d_{mean} (bp)	d_{median} (bp)	d_{max} (bp)
chr01	32 709 648	5 902 134	0.180	19.8	21.2	22.4	36.6	4.596	3	730
chr02	37 631 755	7 569 396	0.201	19.5	21.3	22.1	37.1	4.537	3	526
chr03	37 690 471	7 380 535	0.196	19.4	21.3	22.1	37.2	4.518	3	576
chr04	32 357 154	6 172 848	0.191	19.6	21.2	22.2	37.0	4.550	3	865
chr05	48 068 851	9 096 002	0.189	19.7	21.1	22.4	36.9	4.560	3	603
chr06	37 231 166	6 694 220	0.180	19.6	21.3	22.3	36.7	4.565	3	569
chr07	36 738 692	7 428 228	0.202	19.5	21.2	22.1	37.1	4.534	3	1 022
chr08	31 666 303	6 394 957	0.202	19.6	21.2	22.2	37.1	4.547	3	738
chr09	37 676 754	7 207 518	0.191	19.7	21.3	22.2	36.9	4.556	3	608
chr10	41 841 605	8 626 455	0.206	19.5	21.3	22.1	37.1	4.542	3	907
chr11	42 925 075	8 185 633	0.191	19.4	21.3	22.1	37.2	4.512	3	387
chr12	33 134 071	6 590 948	0.199	19.5	21.3	22.1	37.1	4.529	3	584
chr13	44 437 459	8 305 900	0.187	19.8	21.2	22.6	36.4	4.601	3	955
chr14	32 560 231	6 246 774	0.192	19.5	21.3	22.0	37.2	4.518	3	749
chr15	55 080 361	11 092 974	0.201	19.4	21.4	22.1	37.1	4.534	3	577
chr16	41 441 581	7 869 911	0.190	19.7	21.1	22.5	36.6	4.584	3	512
chr17	34 817 048	6 870 721	0.197	19.5	21.3	22.0	37.2	4.520	3	338

注：L_{seq}为序列长度；N_{TTN}为 5′ – TTN 的数量；P_{TTN}表示各 TTN 在总数中的占比；d_{mean}为间距平均值；d_{median}为间距中值；d_{max}为间距最大值。

NGG 的间距最小为 1 bp，出现次数最多(图 2 – A)；最大为 1 347 bp，出现在 chr07 的 19 065 740 ~ 19 067 087位点间(表 1)。间距中值是 7 bp，表示平均约间隔 7 bp 碱基就有 1 个 5′ – NGG，也说明半数以上的间距不大于 7 bp；间距均值为 12.0 bp，是受到最大值的影响而产生了偏移(图 2 – a)。各染色体情况与此高度一致(表 1)。而在线粒体中，间距中值和均值分别为 5、8.5 bp，与基因组完全不同(图 2 – b)。

基因组中，相邻 5′ – TTN 的间距最小为 1 bp，出现次数最多(图 2 – c)；最大为 1 022 bp，出现在 chr07 的 26 831 153 ~ 26 832 175 位点间(表 2)。间距中值是 3 bp，表示平均约间隔 3 bp 碱基就有 1 个 5′ – TTN，也说明半数以上的间距不大于 3 bp；受最大值影响，间距均值是 4.5 bp(图 2 – c)。同样，各

染色体情况与此一致(表 2)。线粒体的间距中值和均值分别为 4、5.7 bp,也与基因组完全不同(图 2 - d)。使用另一版本的基因组数据(ASM411538v1)计算,可得出类似的结果:5′ - NGG 的间距中值和均值分别为 7、11.9 bp,5′ - TTN 的间距中值和均值分别为 3、4.5 bp;所展示出的变化趋势也可以绘成同样的图像(图 2 - a、图 2 - c)。

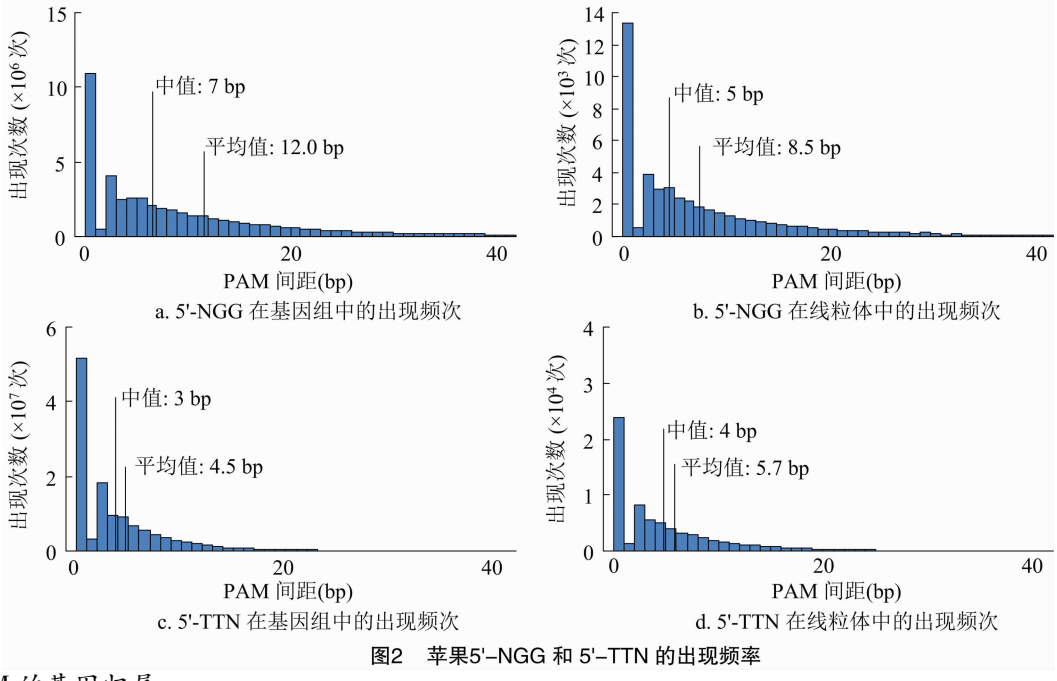


图2　苹果5'-NGG 和 5'-TTN 的出现频率

2.3　PAM 的基因归属

在苹果基因组中,29.0% 的 5′ - NGG 和 26.9% 的 5′ - TTN 能影响到基因,基本覆盖了全部 43 464 个基因(表 3)。其中,chr06 上 ID 为 114825448 的基因含 5′ - NGG 最多,有 14 620 个;chr16 的 114822391(表示基因 ID,下同)和 chr09 的 114827208 基因不含 5′ - NGG。chr06 的 114825448 基因含 5′ - TTN 最多,有 36 069 个;chr16 的 108169786、chr12 的 108174957、chr11 的 108174696、chr02 的 114823832、chr03 的 114824289 和 108171505 基因都不含 5′ - TTN。

在线粒体中,15.6% 的 5′ - NGG 能作用于全部 70 个基因,13.5% 的 5′ - TTN 覆盖了 98.6% 的基因(表 3)。其中,ID 为 13630194 的基因含 5′ - NGG 最多,有 1 186 个;13630239(121 892 ~ 121 913 位点)和 13630229 基因含 5′ - NGG 最少,只有 5 个。13630194 基因含 5′ - TTN 最多,有 1 337 个;13630239 基因(121 892 ~ 121 913 位点)不含 5′ - TTN。与各染色体相比,线粒体中属于基因间隔的 PAM 占比明显更大(表 3)。

表 3　带有 5′ - NGG 或 5′ - TTN 的基因在苹果中的分布

序列名称	基因数量(个)	带有 5′ - NGG 的基因数量(个)	带有 5′ - TTN 的基因数量(个)	属于基因范畴的 5′ - NGG 在总量中的占比(%)	属于基因范畴的 5′ - TTN 在总量中的占比(%)
chr01	2 048	2 048	2 048	27.4	25.4
chr02	2 723	2 723	2 722	32.6	30.3
chr03	2 527	2 527	2 525	29.9	28.2
chr04	2 158	2 158	2 158	29.4	27.2
chr05	3 097	3 097	3 097	29.2	27.5
chr06	2 074	2 074	2 074	26.0	24.2
chr07	2 648	2 648	2 648	31.7	29.4
chr08	2 113	2 113	2 113	29.7	27.6
chr09	2 601	2 600	2 601	29.5	27.9
chr10	2 998	2 998	2 998	28.9	27.0
chr11	2 717	2 717	2 716	30.9	28.7
chr12	2 273	2 273	2 272	29.0	26.8
chr13	2 451	2 451	2 451	23.9	22.2
chr14	2 159	2 159	2 159	28.4	25.7
chr15	3 894	3 894	3 894	29.4	27.1
chr16	2 467	2 466	2 466	25.8	23.9
chr17	2 516	2 516	2 516	30.7	28.2
基因组	43 464	43 462	43 458	29.0	26.9
线粒体	70	70	69	15.6	13.5

2.4　靶序列的重复簇

苹果的 CRISPR 靶序列有大量重复,共计 15 种重复簇(表4)。其中,重复序列只属于染色体基因的簇划入 repeat1,只属于染色体基因间隔的为 repeat2,只属于线粒体基因的为 repeat4,只属于线

粒体基因间隔的为 repeat8，repeat15 是重复的靶序列同时出现在上述 4 个区域中。可以发现，带有 5′－NGG 的靶序列，染色体基因的 648 条与线粒体基因中的 620 条重复；带有 5′－TTN 的靶序列，染色体基因中的 1 038 条与线粒体基因中的 966 条重复（表 4，repeat5、7、13、15）。只属于染色体基因的，带有 5′－NGG 的重复靶序列共有 3 883 583 条，带有 5′－TTN 的重复靶序列共有 84 71 496 条（表 4，repeat1、3、5、9、7、11、13、15 括号中的数字）；而在线粒体基因中，这 2 个数值分别为 980、1 087（表 4，repeat4、5、6、12、7、13、14、15 括号中的数字）。

表 4　苹果靶序列的数量

重复簇类别	带有 5′－NGG 的靶序列数量				带有 5′－TTN 的靶序列数量			
	chrI	chrO	mtI	mtO	chrI	chrO	mtI	mtO
repeat1	1 875 683（1 875 683）	0	0	0	4 023 260（4 023 260）	0	0	0
repeat2	0	6 344 821（6 344 821）	0	0	0	17 116 392（17 116 392）	0	0
repeat4	0	0	670（670）	0	0	0	751（751）	0
repeat8	0	0	0	737（737）	0	0	0	1 054（1 054）
repeat3	2 687 858（2 007 446）	10 667 688（10 227 423）	0	0	6 250 491（4 447 352）	26 875 192（25 724 035）	0	0
repeat5	434（63）	0	429（56）	0	699（146）	0	685（122）	0
repeat6	0	1 616（424）	1 494（208）	0	0	2 569（847）	2 183（174）	0
repeat9	1 371（183）	0	0	1 299（50）	2 178（261）	0	0	2 108（122）
repeat10	0	7 199（1 966）	0	6 338（388）	0	13 678（4 357）	0	11 583（693）
repeat12	0	0	110（12）	104	0	0	191（30）	177（2）
repeat7	192（58）	226（105）	166（22）	0	290（113）	315（134）	234（10）	0
repeat11	552（146）	653（288）	0	472（27）	1 119（360）	1 361（713）	0	895（43）
repeat13	11（2）	0	12（4）	10	33（4）	0	31	31
repeat14	0	42（23）	30（2）	33（6）	0	91（53）	60	79（30）
repeat15	11（2）	16（9）	13（6）	14（5）	16	28（22）	16	21（7）
单一序列	9 436 659	17 341 912	4 344	30 191	24 083 878	49 254 078	5 220	43 971

注：chrI 表示染色体基因区域；chrO 表示染色体基因间隔区域；mtI 表示线粒体基因区域；mtO 表示线粒体基因间隔区域；括号中的数字表示同一区域内重复靶序列的数量。

2.5　基因可编辑度

在苹果各染色体中，带有 5′－NGG 的单一靶序列数量为 26 778 571，其中属于基因的有 9 436 659，其余都处于基因间隔中；带有 5′－TTN 的单一靶序列数量为 73 337 956，属于基因的有 240 838 78（表 4）。SpCas9 对 chr04 上的 103432805 基因有最高的可编辑度，为 0.226，含有 PAM 最多的 114825448 基因的可编辑度仅为 0.020（图 3）；另有 372 个基因的可编辑度为 0。FnCpf1 对 chr08 上的 114826681 基因有最高的可编辑度，为 0.344，含有 PAM 最多的 114825448 基因的可编辑度仅为 0.049（图 3）；另有 305 个基因的可编辑度为 0。

在线粒体中，带有 5′－NGG 的单一靶序列数量为 34 535，属于基因的有 4 344；带有 5′－TTN 的单一靶序列数量为 49 191，属于基因的有 5 220（表 4）。SpCas9 对 13630239 基因（121892～121913 位点）可能有最高的可编辑度，为 0.227，含有 PAM 最多的 13630194 基因的可编辑度为 0.092（图 3）；有 4 个基因的可编辑度为 0。FnCpf1 对 13630216 基因可能有最高的可编辑度，为 0.224，含有 PAM 最多的 13630194 基因的可编辑度为 0.095（图 3）；有 6 个基因的可编辑度为 0。

2.6　Cas 蛋白的编辑互补

苹果中的大多数基因可同时被 2 种 Cas 蛋白编辑，分别具有不同的可编辑度搭配（图 3）。可编辑度为 0 的基因在全部基因中的占比较小，在各 DNA 序列上均有分布；其中，chr04 上的数量最多，有 66 个不能被 SpCas9 编辑、有 62 个基因不能被 FnCpf1 编辑（图 4）。经过筛选，共有 237 个（0.5%）染色体基因、2 个（2.9%）线粒体基因能被 1 种 Cas 蛋白编

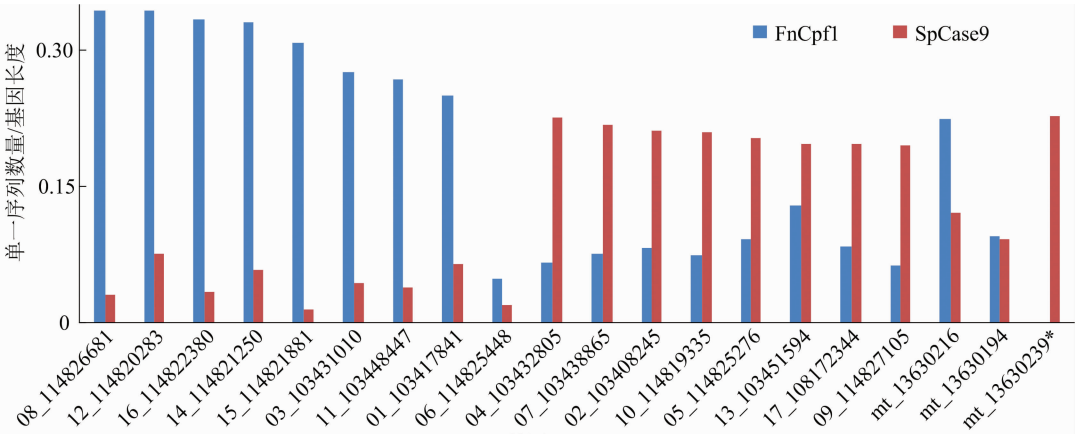
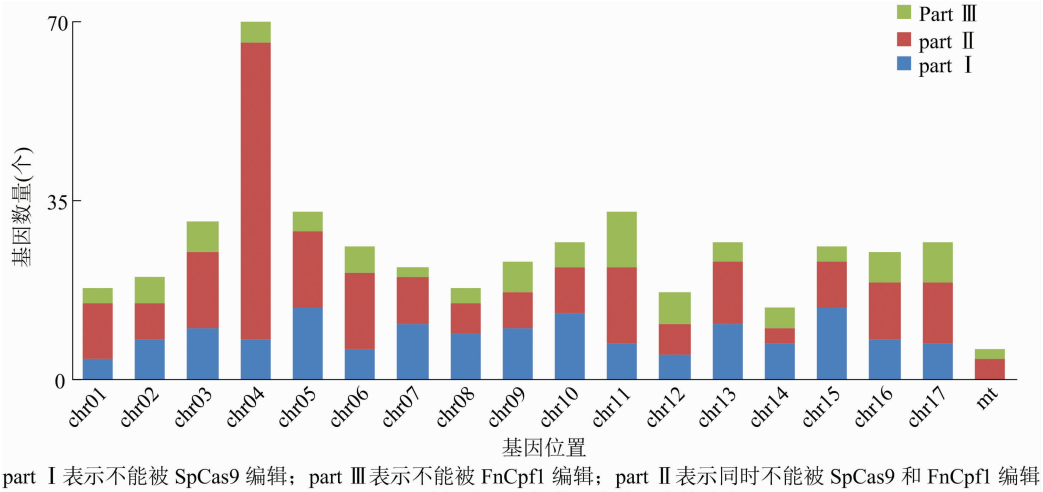


图3　2 种 Cas 蛋白在部分苹果基因中的可编辑度

辑，填补了另一种 Cas 蛋白留下的编辑空白（图 4，part Ⅰ、part Ⅲ）；共有 220 个染色体基因（0.5%）、4 个（5.7%）线粒体基因不能被任一种 Cas 蛋白编辑，即 2 种 Cas 蛋白同时留下编辑空白，没有互补（图 4，part Ⅱ）。



part Ⅰ表示不能被 SpCas9 编辑；part Ⅲ表示不能被 FnCpf1 编辑；part Ⅱ表示同时不能被 SpCas9 和 FnCpf1 编辑

图4　可编辑度为 0 的基因在苹果中的分布

3　讨论

作为重要的基因编辑工具[23]，CRISPR－Cas 系统在苹果基因组中有较好的整体适用性，主要表现在 3 个方面。一是有数量可观的 PAM 分散在苹果 DNA 序列的各个角落，出现频率很高，平均间隔很短。二是 Cas 蛋白的作用位点几乎覆盖了所有基因，个别不能被 SpCas9 识别的基因却含有 FnCpf1 的识别位点，反之亦然。三是拥有单一靶序列的基因占大多数，99.5% 的染色体基因和 94.3% 的线粒体基因都能至少被其中 1 种 Cas 蛋白编辑。苹果 DNA 序列的测序结果表明，AT 碱基的含量高于 CG 碱基[20－21]，导致 5′－TTN 的数量远超 5′－NGG、5′－TTN 的出现频率更高、带有 5′－TTN 的单一靶序列数量更多，也就是说 Cpf1 在苹果基因编辑中有更大的可挖掘潜力。

各染色体的 PAM 密度、组成和出现频率几乎一致，可视作苹果基因组的整体特征之一。虽然也存在于苹果细胞中，但线粒体通常被认为是有益的共生生物[24]，其 DNA 不被计入基因组；这一点在本研究中也有突出体现，即与染色体的同项数据相比，线粒体都有明显差异。目前，对苹果线粒体基因的编辑还未见报道，其操作过程是否与基因组相同还需要更深入的研究验证，在本研究中仅是预测性的初步探讨；且线粒体 DNA 体量小、基因少[20]，对基因组编辑的影响不大，在特定环境中可不做考虑。叶绿体也有类似情况[25]，可在条件成熟时进一步研究讨论。

已有的测序结果含有未知碱基,在数亿碱基的苹果 DNA 序列中比例微小,对多项计算结果的影响可忽略不计[20-21]。但如果 2 个 PAM 之间存在未知碱基且结合上下游无法判断是否存在另一个 PAM,就在确定 PAM 频率时摒弃这 2 个 PAM 的间距,避免出现超长间距的同时,也保证了是在计算相邻 2 个 PAM 的间距。对比人类基因组取间距中值作为 PAM 的出现频率,本研究也采用了同样的取值方法;苹果基因组平均间隔 7 bp 碱基就有 1 个 5′-NGG,频率高于人类基因组的 8 bp[10]。

一般,判断 PAM 是否属于基因依据的是其位点是否在基因起止位点间,临近基因边缘的 PAM 就有可能实际作用到了间隔区。不同于此,本研究将判断依据改进为剪切位点是否在基因起止位点间,既避免了上述问题,也充分挖掘了隐藏的基因 PAM。在此基础上截取的靶序列都具有明确的基因归属。靶序列的长度按常规被设定为 20 nt,初步分析了因序列重复导致的脱靶情况;根据序列越短重复率越高的共识,可适当增加靶序列的长度提高基因的可编辑度。此外,脱靶的原因还包括相似匹配和种子序列的长度[11],可在未来的研究中做更深入的分析。

基因可编辑度与单一靶序列的数量成正比,与基因自身的长度成反比,表示的是单位长度内含有的备选靶序列密度。可编辑度为 0 的基因,小部分是因为不含有 PAM,大多数是在屏蔽了重复靶序列后,备选数量为 0。本研究采用了较严格的屏蔽标准,凡是在苹果 DNA 序列中出现的重复靶序列均计入重复簇;重复簇的类别划分较细,15 个类别涵盖了靶序列所在 4 个区域的所有搭配,方便在试验设计时有侧重地取舍。在苹果全部基因中,2 种 Cas 蛋白都有占比很小的编辑盲区,FnCpf1 要好于 SpCas9。盲区重叠的部分所含的 224 个基因不适宜使用这 2 种 Cas 蛋白编辑,可考虑换用识别不同 PAM 的其他 Cas 蛋白;其中超过半数的基因(139 个)编码多种 RNA,通常在实际研究中较少涉及到。

在 Perl 脚本的帮助下,各步骤的运算结果都在文本文件中详细列表构成了信息库,可直接打开查询感兴趣的信息;也可导入数据库加以专业化的管理和维护,成为网络服务平台的构建基础,这是开展下一步研究的一个重要方向。

参考文献:

[1] Arzani A, Ashraf M. Smart engineering of genetic resources for enhanced salinity tolerance in crop plants[J]. Critical Reviews in Plant Sciences,2016,35(3):146-189.

[2] Wang X H,Tu M X,Li Z,et al. Current progress and future prospects for the clustered regularly interspaced short palindromic repeats (CRISPR) genome editing technology in fruit tree breeding[J]. Critical Reviews in Plant Sciences,2018,37(4):233-258.

[3] Charrier A,Vergne E,Dousset N,et al. Efficient targeted mutagenesis in apple and first time edition of pear using the CRISPR-Cas9 system[J]. Frontiers in Plant Science,2019,10:40.

[4] Zhou J H,Li D D,Wang G M,et al. Application and future perspective of CRISPR/Cas9 genome editing in fruit crops[J]. Journal of Integrative Plant Biology,2020,62(3):269-286.

[5] Yan F C,Wang W,Zhang J Q. CRISPR-Cas12 and Cas13:the lesser known siblings of CRISPR-Cas9[J]. Cell Biology and Toxicology,2019,35(6):489-492.

[6] Strecker J,Jones S,Koopal B,et al. Engineering of CRISPR-Cas12b for human genome editing[J]. Nature Communications,2019,10:212.

[7] Makarova K S,Zhang F,Koonin E V. Snapshot:class 2 CRISPR-Cas systems[J]. Cell,2017,168(1/2):328-328e1.

[8] Hu J H,Miller S M,Geurts M H,et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity[J]. Nature,2018,556(7699):57-63.

[9] Moradpour M,Abdulah S N A. CRISPR/dCas9 platforms in plants:strategies and applications beyond genome editing[J]. Plant Biotechnology Journal,2020,18(1):32-44.

[10] Hsu P D,Lander E S,Zhang F. Development and applications of CRISPR-Cas9 for genome engineering[J]. Cell,2014,157(6):1262-1278.

[11] Zetsche B,Gootenberg J S,Abudayyeh O O,et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system[J]. Cell,2015,163(3):759-771.

[12] 杨　帆,李　寅. 新一代基因组编辑系统 CRISPR/Cpf1[J]. 生物工程学报,2017,33(3):361-371.

[13] Ma X L,Zhu Q L,Chen Y L,et al. CRISPR/Cas9 platforms for genome editing in plants:developments and applications[J]. Molecular Plant,2016,9(7):961-974.

[14] Wang X H,Tu M X,Wang D J,et al. CRISPR/Cas9-mediated efficient targeted mutagenesis in grape in the first generation[J]. Plant Biotechnology Journal,2018,16(4):844-855.

[15] Rodríguez-Leal D,Lemmon Z H,Man J,et al. Engineering quantitative trait variation for crop improvement by genome editing[J]. Cell,2017,171(2):470-480,e8.

[16] Nishitani C,Hirai N,Komori S,et al. Efficient genome editing in apple using a CRISPR/Cas9 system[J]. Scientific Reports,2016,6:31481.

[17] Malnoy M,Viola R,Jung M H,et al. DNA-free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins[J]. Frontiers in Plant Science,2016,7:1904.

[18] Pompili V,Costa L D,Piazza S,et al. Reduced fire blight susceptibility in apple cultivars using a high-efficiency CRISPR/

卢姊豪,贡　嘎,常　攀,等. 西藏牦牛肠出血性大肠杆菌 HPI 基因调查及耐药性分析[J]. 江苏农业科学,2022,50(7):51 – 58.
doi:10.15889/j.issn.1002 – 1302.2022.07.007

西藏牦牛肠出血性大肠杆菌 HPI 基因调查及耐药性分析

卢姊豪, 贡　嘎, 常　攀, 武　琦, 李天骄, 吴　丹, 罗润波, 索朗斯珠
(西藏农牧学院动物科学学院,西藏林芝 860000)

摘要:为明确西藏牦牛源肠出血大肠杆菌强毒力岛(HPI)基因以及 ESBLs 耐药基因携带情况,复壮笔者所在实验室保存的 14 株牦牛源肠出血性大肠杆菌,采用麦康凯培养基和伊红美蓝培养基进行筛选,镜检确定复壮成功后,应用 PCR 方法检测 HPI 携带情况,并对 HPI 相关毒力基因进行克隆与序列分析,测序结果同时采用 K – B 纸片法对被检菌株进行药敏试验和 ESBLs 表型检测,最后应用 PCR 方法进行 ESBLs 耐药基因检测。结果显示,14 株被检菌株中 HPI 标志性基因 *fyuA* 的携带率最高,达 42.86%(6/14),其次分别为 *ybtA*、*irp2* 和 *irp8* 基因,携带率均为 21.43%(3/14),*irp3* 携带率为 14.2%(2/14),而未检测到 *irp5* 阳性基因。分离菌株对青霉素、苯唑西林、氨苄西林、头孢氨苄、头孢唑林、头孢拉定、阿米卡星的耐药率均≥57.14%,对派拉西林、羧苄西林、头孢他啶、环丙沙星耐药率均 ≥21.43%,并存在多种耐药现象,耐药谱集中在 3 ~ 10 耐,其中 6 耐菌株最多,有 5 株,其次为 8 耐 3 株、7 耐 2 株,10 耐、9 耐、5 耐和 3 耐均 1 株。对被检菌株进行 14 种 ESBLs 类基因检测,结果发现被检菌株只携带 1 种 ESBLs 类基因,为 bla_{TEM} 基因,携带率 21.43%(3/14)。综上,西藏牦牛源肠出血性大肠杆菌中存在 HPI 基因。对常用抗菌药产生较高的耐药性并存在多种耐药现象,但对 ESBLs 类药物具有较高的敏感性。

关键词:肠出血性大肠杆菌;强毒力岛;耐药性;西藏牦牛
中图分类号: S855.1⁺2　　**文献标志码:** A　　**文章编号:**1002 – 1302(2022)07 – 0051 – 08

目前,出血性大肠杆菌(enterohaemorrhagic *Escherichia coli*,EHEC)是 6 种大肠杆菌中重要的食源性病原菌,给食品健康带来前所未有的挑战。牛源 EHEC 是引起人类大肠杆菌病的主要来源之一。据报道,国外健康牛以及牛肉中分离出 EHEC(O46 血清型)并给人类健康带来一定的危害[1-2]。赵燕娟等 2019 年从西藏牦牛体内分离鉴定出牦牛源肠出血性大肠杆菌并确定主要毒力基因为 *ehxA*[3],提示西藏牦牛群中确实存在肠出血性大肠杆菌。

强毒力岛(high pathogeniticity island,HPI)也称为耶尔森菌强毒力岛。它的核心基因主要有 6 种,分别为 *fyuA*、*irp2*、*irp8*、*irp5*、*ybtA* 和 *irp3* 组成,其中 *irp2* 和 *fyuA* 是 HPI 基因簇的核心基因[4]。耶尔森菌强毒力岛在大肠杆菌和耶尔森菌之间进行水平传播的相关报道很多,并与致病性大肠杆菌的毒力

收稿日期:2021 – 07 – 04
基金项目:国家肉牛/牦牛产业技术体系建设专项(编号:CARS – 37);西藏自治区科技厅 2019 年度重点项目(编号:201901)。
作者简介:卢姊豪(1998—),女,北京人,硕士研究生,主要从事高原动物传染病研究,E – mail:670630168@qq.com;共同第一作者:贡嘎(1983—),男,西藏拉萨人,博士研究生,主要从事高原动物传染病研究,E – mail:xzlzgg@163.com。
通信作者:索朗斯珠,博士,教授,主要从事高原动物传染病研究。E – mail:xzslsz@163.com。

Cas9 – FLP/FRT – based gene editing system[J]. Plant Biotechnology Journal,2020,18(3):845 – 858.
[19] Endo A,Masafumi M,Kaya H,et al. Efficient targeted mutagenesis of rice and tobacco genomes using Cpf1 from *Francisella novicida*[J]. Scientific Reports,2016,6:38169.
[20] Daccord N,Celton J M,Linsmith G,et al. High – quality de novo assembly of the apple genome and methylome dynamics of early fruit development[J]. Nature Genetics,2017,49(7):1099 – 1106.
[21] Zhang L Y,Hu J A,Han X L,et al. A high – quality apple genome assembly reveals the association of a retrotransposon and red fruit colour[J]. Nature Communications,2019,10:1494.
[22] Alsuwaiyel M H. Algorithms: design techniques and analysis (revised edition)[M]. Singapore: World Scientific Publishing,2016:20 – 34.
[23] Koonin E V,Makarova K S,Zhang F. Diversity,classification and evolution of CRISPR – Cas systems[J]. Current Opinion in Microbiology,2017,37:67 – 78.
[24] Nykky J,Vuento M,Gilbert L. Role of mitochondria in parvovirus pathology[J]. PLoS One,2014,9(1):e86124.
[25] Waters M T,Langdale J A. The making of a chloroplast[J]. The EMBO Journal,2009,28(19):2861 – 2873.