

# 水稻基因组中 MITEs 的偏爱位置

张显亮 洪国藩\*

(中国科学院国家基因研究中心, 上海 200233)

**摘要** 已知微型倒转重复转座因子 (MITEs) 广泛分布于真核生物基因附近的非编码区。为探讨 MITEs 在基因组一级结构中的分布规律, 通过对 82 kb 的水稻 BAC 克隆 H1332 全序列分析, 精确鉴定了 10 个蛋白质编码基因的位置和转录方向以及 7 个 MITEs 的位置, 毫无例外的观察到, MITEs 只要存在, 总是位于最近的基因的上游, 不管该基因与之相距多远、转录方向如何。另外还发现, 1 个 MITE 被 2 个基因共享。由于 MITEs 易于鉴定, 因而这种位置模式可以用作探测其下游基因存在的工具。

**关键词** 水稻基因组; 位置模式; 微型倒转重复转座因子; 基因 5' 侧区

微型倒转重复转座因子 (miniature inverted-repeat transposable elements, MITEs) 是最近被鉴定出来的一大类真核转座因子, 对基因的顺式调控因子有所贡献<sup>[1-4]</sup>。它们似乎在从植物到动物的基因组中呈弥散型广泛分布在基因附近的非编码区<sup>[2]</sup>, 以其长度小 (100 ~ 500 bp)、一般富含 T/A、含末端倒转重复序列 (terminal inverted repeats, TIRs)、有形成二级结构潜力、无编码能力和插入时引起靶序列倍增, 又称直向重复 (direct repeats, DRs) 为特征。现已鉴定的不同家族的 MITEs 都拥有上述结构上的特征, 但只有同一家族内部的成员之间才拥有序列上的保守性。而在同一家族的不同成员之间, 序列最保守的部分往往是 TIRs, 所以, 保守的 TIRs 成为界定 MITEs 不同家族的主要依据。因为在不同物种的基因组之间可以发现具有相似 TIR 序列的 MITEs<sup>[2,4]</sup>, 所以 MITEs 的类型并不以物种为界来作划分。由于 MITEs 是最近才发现的一类转座因子, 目前关于这种转座子的认识仍限于表层, 不外乎不同物种中新 MITEs 的发现、分类及其可转移性 (mobility) 证据的发掘<sup>[1-4]</sup> 等; 对于 MITEs 的作用及其转座机制, 有的只是推测。大家都讲 MITEs 与基因密切相关, 这是沿用大量单个基因序列片段上发现 MITEs 时的最初说法; 至于 MITEs 如何与基因相关, 就作者所知, 一直没有确切报道。

本实验室较早地进行了大规模水稻基因组测序, 并对所获得的其中一段 DNA 长程连续序列进行了精细分析。很幸运, 这段 DNA 上诸基因转录方向各异, 因而得以观察到 MITEs 是在基因上游而不是下游。所以在所研究的这段水稻基因组序列上第一次看到了一种 MITEs 和基因之间在 DNA 一维序列上的相对位置模式。关于这种位置模式的生物学意义及其应用, 本文做了进一步讨论。

## 1 材料和方法 (Materials and Methods)

### 1.1 材料及仪器

水稻第 4 号染色体 BAC 克隆 H1332, 经鸟枪法全序列测定并组装 (测序仪为 PE 公司所产的 ABI377; 测序载体采用 pBluescript KS; 随机序列组装使用 Stadern Package), 得全长序列 82 218 bp (GenBank 登录号: AJ243961), 用于全序列分析。相关 cDNA 克隆由日本 Sasaki 博士提供。

### 1.2 方法

1.2.1 蛋白质编码基因的鉴定 搜索水稻 EST<sup>[5]</sup> 数据库中与 BAC H1332 序列完全匹配的条目, 确定相应的 cDNA 序列, 由此界定的区域被认为是蛋白质编码基因; 搜索拟南芥蛋白数据库 (<http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-blast2atdb>), 高度同源区被认为是蛋白质编码区。综合上述两种来源的资料, 以基因预测软件 GenScan<sup>[6]</sup> 为辅助, 在绝大多数情况下可以获得蛋白质编码区的准确位置信息, 包括终止密码子, 内含子/外显子边界和起始密码子等。GenBank 搜索

收稿日期: 1999-11-26 接受日期: 1999-12-24

\* 联系人: Tel, 021-64516371; Fax, 021-64825775; e-mail, gfhong@newnetra.ncgr.ac.cn

用美国国家生物技术信息中心 (National Center for Biotechnology Information, Bethesda, Md) 的 BLAST 程序进行。

1.2.2 MITEs 的鉴定 用 BLAST 程序进行序列同源性搜索, 可以检测到基因组序列中的已知 MITEs, 新的类 MITE 序列也可以根据它们极富特征的微型末端倒转重复结构被检测到<sup>[2]</sup>。

## 2 结果 (Results)

### 2.1 BAC I1332 上编码的 10 个蛋白质基因

总长为 82 kb 的 BAC I1332 上可以准确鉴定 10 个蛋白质编码基因 (表 1, 图 1)。其中 7 个 (I1332.1, I1332.2, I1332.3a, I1332.3, I1332.4, I1332.6, I1332.8) 基因的编码区由水稻 cDNA 序列确证; 另外 3 个基因 (I1332.5, I1332.7, I1332.9) 虽无 cDNA 序列确证, 但所翻译的氨基酸序列都能找到与之高度同源的拟南芥蛋白, BLAST 结果,  $P$  值分别为  $3.8 \times 10^{-142}$ 、

Table 1 Genes encoded by BAC I1332

Number <sup>b</sup>	Gene	BLASTP match	Organism	GenBank accession no.	$P$ value	EST match	Tissue expressing
1	I1332.1	AtMRP2	<i>A. rabidopsis thaliana</i>	AC003096 (gi 3132479)	0	C72419	panicle
2	I1332.2	unknown protein	<i>A. rabidopsis thaliana</i>	AF049236 (gi 3068705)	$7 \times 10^{-34}$	AU058040	callus
3	I1332.3a					C74465	panicle
3	I1332.3					AU029363 C74272 D48443	panicle panicle green shoot
4	I1332.4	contains 6 PF 0040 WD40 G-beta repeat domain	<i>A. rabidopsis thaliana</i>	AC007519 (gi 5668798)	$6.9 \times 10^{-9}$	D43026 C26424 C98576	callus callus panicle
5	I1332.5	contains PF 00069 eukaryotic protein kinase domain	<i>A. rabidopsis thaliana</i>	AC007060 (gi 4587513)	$3.8 \times 10^{-142}$		
6	I1332.6					AU030332 AU030331 C91820 D41592	immature leaf immature leaf panicle green shoot
7	I1332.7	Ser/ Thr kinase protein	<i>A. rabidopsis thaliana</i>	AC006569 (gi 4512705)	$6 \times 10^{-55}$		
8	I1332.8	hypothetical protein	<i>A. rabidopsis thaliana</i>	AC004138 (gi 3461814)	$2 \times 10^{-101}$	AU029996 C97109 C27719 C71934	immature leaf callus callus panicle
9	I1332.9	NAC domain protein NAM	<i>A. rabidopsis thaliana</i>	AF123311 (gi 4325286)	$2.2 \times 10^{-42}$		

<sup>b</sup> Numbers in this column correspond to those indicated in Fig. 1.

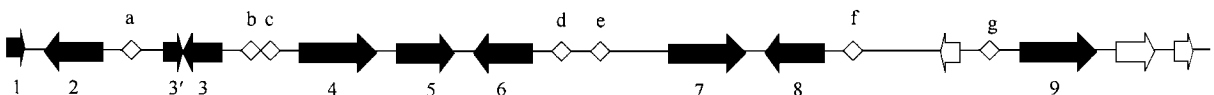


Fig. 1 Schematic representation of the positional pattern of miniature inverted transposable elements (MITEs) relative to genes in the rice genomic sequence of BAC I1332 (under GenBank/ DDBJ/ EMBL accession no. AJ243961)

The closed arrows with numbers beneath denote genes and their directions of transcription precisely determined with similarity match in public database available, and open arrows indicate genes mainly predicted by program GenScan<sup>[6]</sup>. The open diamonds with letters above indicate MITEs, all detected by similarity search. The numbers 1—9 denote gene I1332.1, I1332.2, I1332.3a, I1332.3, I1332.4, I1332.5, I1332.6, I1332.7, I1332.8, I1332.9, respectively, and the letters a, b, c, d, e, f, g, denote MITE *Tnr1-Os1*, *Suqir-Os1*, *Tnr1-Os2*, *Tourist-Os1*, *Zhangyi-Os1*, *Tourist-Os2*, *Zhangyi-Os2*, respectively. MITEs identified were all located upstream of the genes regardless of their directions of transcription.

$6 \times 10^{-55}$ 、 $2.2 \times 10^{-42}$ 。

### 2.2 BAC I1332 上的 7 个 MITEs

在 BAC I1332 序列上所检测到的 7 个 MITE 样序列因子的位置和特征总结于图 1 和表 2。其中 4 个属于已经鉴定的 MITE 家族,包括 2 个 *Tourist* 因子<sup>[7,8]</sup>和 2 个 *Tnr1* 因子<sup>[9]</sup>。2 个 *Tourist*, *Tourist-Os1* 和 *Tourist-Os2*, 分别位于 I1332.6 和 I1332.8 的 5 侧区,拥有相同碱基达 82%。虽然由于退化导致的 TIRs 和 DRs 不可鉴别,但多重序列比较依然可以显示其中所含的 *Tourist* 因子共同核心基序(图 2)。2 个 *Tnr1* 因子, *Tnr1-Os1* 和 *Tnr1-Os2*, 分别位于 I1332.2 和 I1332.4 的 5 侧区(图 1,表 2),二者拥有 84% 的相同碱基,与已知的 *Tnr1*<sup>[10]</sup> 因子进行多重序列比较,表明它们属于 *Tnr1* 因子(图 3)。

除已知的 MITEs 外,我们还在 BAC I1332 上

发现两族新型类 MITE 序列, *Suqin* 和 *Zhangyi*, 它们与已知的 MITEs 家族没有任何序列相似性。*Suqin* 族发现一个,命名 *Suqin-Os1*, 位于 I1332.3 的 5 侧区,与 *Tnr1-Os1* 靠得很近(图 1,表 2)。*Suqin-Os1* 具有近乎完美的倒转重复结构特征和 9

Table 2 MITEs and MITElike elements in BAC I1332

Name	Class	Characteristics	Position(nt) <sup>i</sup>
<i>Tnr1-Os1</i>	MITE	TIRs, 2 bp DRs	a(15920—16154)
<i>Tnr1-Os2</i>	MITE	TIRs, 2 bp DRs	c(25239—25466)
<i>Tourist-Os1</i>	MITE	<i>Tourist</i> core sequence	d(40888—41038)
<i>Tourist-Os2</i>	MITE	<i>Tourist</i> core sequence	f(58738—58887)
<i>Suqin-Os1</i>	novel	TIRs, 9 bp DRs	b(24496—24898)
<i>Zhangyi-Os1</i>	novel	TIRs	e(42389—42588)
<i>Zhangyi-Os2</i>	novel	TIRs	g(69773—69980)

<sup>i</sup> The letters a—g in this column correspond to the MITEs, as indicated in Fig. 1.



Fig. 2 Multiple alignment of *Tourist-Os1*, *Tourist-Os2* and *Tourist-OsSH2* [ GenBank accession no. U70541 (21644—21977) ]

Dashed lines represent gaps introduced to optimize alignment. Nucleotides that are identical to those of *Tourist-OsSH2* are indicated in bold.



Fig. 3 Multiple alignment of *Tnr1-Os1*, *Tnr1-Os2* and *Tnr1D-Og*<sup>[9]</sup>

Dashed lines represent gaps introduced to optimize alignment. Nucleotides that are identical to those of *Tnr1D-Og* are indicated in bold. The 2 bp DRs "TA" are underlined.

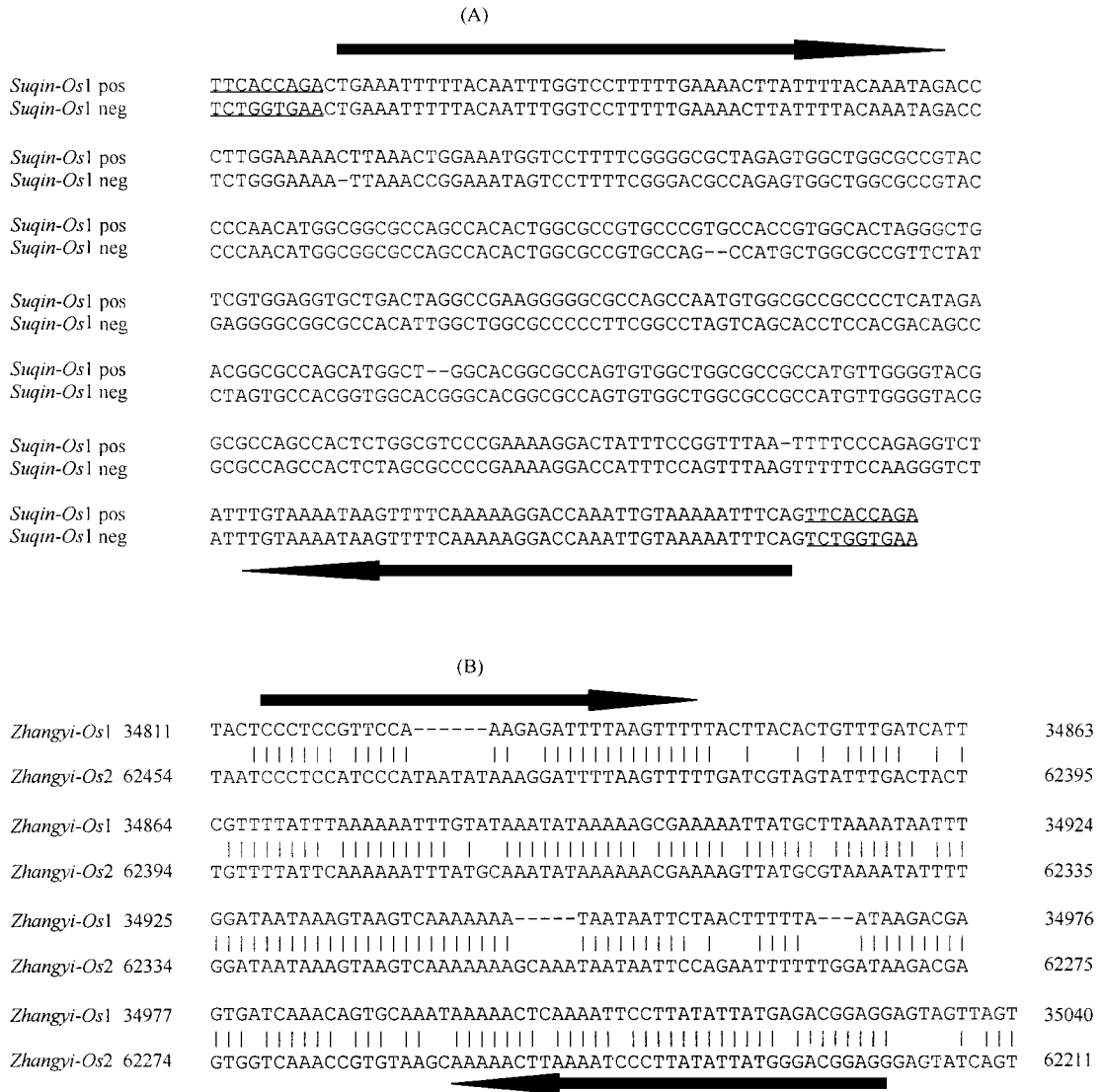


Fig. 4 The new family of MITEs identified in BAC 11332

(A) Opposite strands (*Suqin-Os1* pos and *Suqin-Os1* neg) alignment of the putative novel MITE-like element *Suqin-Os1*; closed arrows indicate the position of TIRs, and underlined regions indicate the putative DRs. (B) Alignment between the two members (*Zhangyi-Os1* and *Zhangyi-Os2*) of MITE-like family *Zhangyi*; closed arrows indicate the position of TIRs. Dashed lines represent gaps introduced to optimize alignment. The numbers on left and right indicate the position of the MITEs in the sequence of BAC 11332.

bp 的 DRs [图 4 (A)]。 *Zhangyi* 族的 2 个成员, *Zhangyi-Os1* 和 *Zhangyi-Os2*, 分别位于 11332.6 和 11332.9 的 5 侧区 (图 1, 表 2), 各自全长约 240 bp, 二者拥有 81% 的相同碱基, 具有很明显的末端倒转重复序列, 但 DRs 由于退化而不可鉴别 [图 4 (B)]。

### 2.3 BAC 11332 上 MITEs 及类 MITE 序列因子相对于基因的位置模式

将上述基因在 BAC 11332 上的位置及转录方向, 以及上述 MITEs 的位置直观地总结在图 1 中, 不难发现, 如果基因附近存在 MITEs, 后者总是位

于基因编码区的上游区, 不管基因的转录方向如何。其中 *Tnr1-Os1* 位于 11332.2 和 11332.3a 两个位置相邻但转录方向相反基因的公共 5 侧区。

### 3 讨论 (Discussion)

MITEs 与植物基因密切相关, 但迄今为止没有关于它们如何相关的确切报道。在本研究中, 对 82 kb 的水稻 BAC 11332 全序列测定和分析, 使得这段序列成为观察 MITEs 集中分布于基因上游的第一个例证。这个结果的可靠性依赖于对蛋白质编码基因和 MITEs 鉴定的准确性。由于 MITE 类型

的转座子不同家族之间具有保守的末端倒转重复结构特征和比较保守的 DRs, 长度一般在几百碱基对之内, 以及同一家族不同成员之间具有显著的序列同源性, 所以容易保证 MITE 鉴定的准确性。在本研究中, 有一些潜在的类 MITE 序列因子, 可能由于序列相似性搜索程序 BLAST 本身的敏感度不足而未能检测到, 但经证实, 它们都同处于基因的上游区(未发表数据)。

基因预测软件所给出结果的不可靠是众所周知的<sup>[11]</sup>, 所以在本研究中确定的 10 个蛋白质编码基因都依赖与 cDNA 序列和(或)蛋白质氨基酸序列的同源匹配(表 1)。有 cDNA 序列匹配的基因的位置和方向自然可以确定; 只有 3 个基因无 cDNA 序列匹配, 但它们的概念翻译与已知蛋白质氨基酸序列匹配  $P$  值最大只有  $2.2 \times 10^{-42}$ , 远小于随机概率。这说明, 我们在 BAC 11332 中鉴定的 10 个蛋白质编码基因的位置和转录方向是正确的。另外 3 个预测基因(在图 1 中由空心箭头表示)由 GenScan 预测, 虽无同源信息确证, 但它们的位置和转录方向与前述 MITEs 的位置模式毫不矛盾。

随着各种真核基因组分析的迅速进展, 不久我们将可以验证本文所看到的局部模式是否具有普遍性。拟南芥菜(*A. thaliana*)是已知高等植物中基因组最小的一个, 已有的全序列分析结果也最丰富, 但在该基因组中, 散在重复 DNA 含量很低, 大约只占总序列的 4%。在这 4% 的序列中, 除去广为熟知的逆转座子和 DNA 介导的转座子类型, 本来已经所剩无几。所以直到 1998 年才在拟南芥菜基因组中发现第一个(也是目前所知的唯一的一个)MITEs 家族; 而且它在该基因组中拷贝数很低, 在定位上与基因也没有明显的近程关联<sup>[4]</sup>。这些事实暗示, 也许拟南芥基因密度较高, 因而与其他植物基因组, 如玉米和水稻, 有所不同。另外, 我们已经了解, 有少数 MITEs 位于基因的内含子<sup>[3, 10, 12, 13]</sup>和 3' UTR 中<sup>[14]</sup>。但这些情况只限于少数倒转重复特征已经明显退化的族类<sup>[2]</sup>, 并不影响 MITEs 偏爱位于基因编码区上游的统计模式。

最近有人提出, MITEs 有可能充当基质附着区(MARs)的作用<sup>[15]</sup>, 这与本文所报道的 MITEs 位置模式是一致的, 因为功能型 MARs 多位于基因的 5' 上游<sup>[16~27]</sup>, 不但为染色体结构单元(loop)提供结合于核骨架的锚点<sup>[28]</sup>, 而且还是复制原点和转录增强子的聚藪<sup>[29]</sup>。另外, MITEs 序列本身的倒转重复结构及富含 T/A 的特征也与它充当功能

型 MARs、基因上游调控元件的角色相适应。

不管怎样, 在水稻基因组中, MITEs 偏爱位于基因上游的事实, 意味着其下游不远处至少存在一个定向转录的基因的可能性将高于随机取样的区域, 这将有利于基因的分选。

## References

- 1 Wessler S R, Bureau T E, White S E. LTR-retrotransposons and MITEs—important players in the evolution of plant genomes. *Curr Opin Genet Dev*, 1995, 5: 814—821
- 2 Bureau T E, Ronald P C, Wessler S R. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc Natl Acad Sci USA*, 1996, 93: 8524—8529
- 3 Tu Z. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. *Proc Natl Acad Sci USA*, 1997, 94: 7475—7480
- 4 Casacuberta E, Casacuberta J M, Puigdomenech P, Monfort A. Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: Characterization of the *Emigrant* family of elements. *Plant J*, 1998, 16: 79—85
- 5 Sasaki T *et al.* Toward cataloguing all rice genes: Large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J*, 1994, 6: 615—624
- 6 Burge C B, Karlin S. Finding the gene in genomic DNA. *Curr Opin Struct Biol*, 1998, 8: 346—354
- 7 Bureau T E, Wessler S R. *Tourist*: A large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, 1992, 4: 1283—1294
- 8 Bureau T E, Wessler S R. Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. *Proc Natl Acad Sci USA*, 1994, 91: 1411—1415
- 9 Tenzen T, Matsuda Y, Ohtsubo H, Ohtsubo E. Transposition of *Tnrl* in rice genomes to 5'-PuTAPy-3' sites, duplicating the TA sequence. *Mol Gen Genet*, 1994, 245: 441—448
- 10 Nelson A J, Doerner P W, Zhu Q, Lamb C J. Isolation of a monocot 3-hydroxy-3-methylglutaryl coenzyme A reductase gene that is elicitor-inducible. *Plant Mol Biol*, 1994, 25: 401—412
- 11 Fickett J W. Finding genes by computer: The state of the art. *Trends in Genet*, 1996, 12: 316—320
- 12 Zhao Y, Kung S D, Dube S K. Nucleotide sequence of rice 4-coumarate: CoA ligase gene, 4-CL. 1. *Nucleic Acids Res*, 1990, 18: 6144
- 13 Shah D M, Hightower R C, Meagher R B. Genes encoding actin in higher plants: Intron positions are highly conserved but the coding sequences are not. *J Mol Appl Genet*, 1983, 2: 111—126
- 14 Bugos R C, Thom M. Glucose transporter cDNAs from sugarcane. *Plant Physiol*, 1993, 103: 1469—1470
- 15 Avramova Z, Tikhonov A, Chen M, Bennetzen J L. Matrix attachment regions and structure colinearity in the genomes of two

- grass species. *Nucleic Acids Res*, 1998, **26**: 761—767
- 16 Fukuda Y. Characterization of matrix attachment sites in the upstream region of a tobacco chitinase gene. *Plant Mol Biol*, 1999, **39**: 1051—1062
- 17 Chattopadhyay S, Whitehurst C E, Chen J. A nuclear matrix attachment region upstream of the T cell receptor beta gene enhancer binds Cux/ CDP and SATB1 and modulates enhancer-dependent reporter gene expression but not endogenous gene expression. *J Biol Chem*, 1998, **273**: 29838—29846
- 18 van der Geest A H, Hall T C. The beta-phaseolin 5' matrix attachment region acts as an enhancer facilitator. *Plant Mol Biol*, 1997, **33**: 553—557
- 19 Phi-Van L, Stratling W H. Dissection of the ability of the chicken lysozyme gene 5' matrix attachment region to stimulate transgene expression and to dampen position effects. *Biochemistry*, 1996, **35**: 10735—10742
- 20 Yu J, Bock J H, Slightom J L, Villeponteau B. A 5' beta-globin matrix-attachment region and the polyoma enhancer together confer position-independent transcription. *Gene*, 1994, **139**: 139—145
- 21 Das A T, Luderus M E, Lamers W H. Identification and analysis of a matrix-attachment region 5' of the rat glutamate-dehydrogenase-encoding gene. *Eur J Biochem*, 1993, **215**: 777—785
- 22 de Moura Gallo C V, Vassetzky Y S, Huesca M, Scherrer K. A transcription-dependent DNase I-hypersensitive site in a far upstream segment of the chicken alpha-globin gene domain coincides with a matrix attachment region. *Biochem Biophys Res Comm*, 1992, **184**: 1226—1234
- 23 Phi-Van L, von Kries J P, Ostertag W, Stratling W H. The chicken lysozyme 5' matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes. *Mol Cell Biol*, 1990, **10**: 2302—2307
- 24 Porter S D, Hu J, Gilks C B. Distal upstream tyrosinase 5' MAR-containing sequence has regulatory properties specific to subsets of melanocytes. *Dev Genet*, 1999, **25**: 40—48
- 25 Yan Z J, Qian R L. The 5'-flanking cis-acting elements of the human epsilon-globin gene associates with the nuclear matrix and binds to the nuclear matrix proteins. *Cell Res*, 1998, **8**: 209—218
- 26 Scholz H, Bossone S A, Cohen H T, Akella U, Strauss W M, Sukhatme V P. A far upstream cis-element is required for Wilms' tumor-1 (WT1) gene expression in renal cell culture. *J Biol Chem*, 1997, **272**: 32836—32846
- 27 Razin S V, de Moura Gallo C V, Scherrer K. Characterization of the chromatin structure in the upstream region of the chicken alpha-globin gene domain. *Mol Gen Genet*, 1994, **242**: 649—652
- 28 Bonifer C, Hecht A, Saueressig H, Winter D M, Sippel A E. Dynamic chromatin: The regulatory domain organization of eukaryotic gene loci. *J Cell Biochem*, 1991, **47**: 99—108
- 29 Boulikas T. Nature of DNA sequence at the attachment regions of genes to the nuclear matrix. *J Cell Biochem*, 1993, **52**: 14—22

## Preferential Location of MITEs in Rice Genome

ZHANG Xian-Liang<sup>1</sup>, HONG Guo-Fan<sup>2\*</sup>

(<sup>1</sup> National Center for Gene Research, the Chinese Academy of Sciences, Shanghai 200233, China)

**Abstract** To observe the association between MITEs and genes in details, 82 kb of the rice genomic DNA were sequenced, and 10 genes were identified by similarity search. It was found that miniature inverted-repeat transposable elements (MITEs) were located upstream of the genes identified without exception, regardless of their directions of transcription. One MITE was found to be located between two 5' regions of the genes with opposite orientation of transcription, indicating that it was shared by them. The positional patterns of MITEs may therefore serve as an aid for preliminary screening of genes.

**Key words** rice genome; positional pattern; MITEs; 5'-flanking region

Received: November 26, 1999 Accepted: December 24, 1999

\* Corresponding author: Tel, 86-21-64516371; Fax, 86-21-64825775; e-mail, gfhong@newnetra.ncgr.ac.cn