专论与综述

可转座基因与植物基因组多样性

刘玉军*,** 李一勤* 刘 强*

(*清华大学生物科学与技术系 北京 100084 **北京林业大学生物学院 北京 100083)

高等植物基因组中存在许多各式各样的重复序列。按单倍体中的出现频率可将其分为高频度(10^5 以上)、中频度(10^2—10^5)和低频度(10^2 以下)重复序列。如按重复样式分类,可分为像卫星 DNA(satellite DNA)那样存在于基因组上的串联重复序列(tandem repeat sequence)和在基因组上分散存在的散布重复序列[1]。大多数散布重复序列为中频度并具有移动能力的可转座基因(transposable gene),它们在进化过程中发生转座(transposition)的结果,引起基因组中一些基因拷贝数的增加,导致植物基因组多样性(genome diversity)[2,3]。

植物基因组中的可转座基因包括两大类:一类是在自身编码的转座酶(transposase)的作用下,从基因组的某个位置直接转座到其他部位的由"DNA 到 DNA"的 DNA 型可转座基因,习惯上将这类可转座基因称为转座子(transposon);另一类是像反病毒(retrovirus)的原病毒(provirus)那样,一经转录后即成为 RNA,并进一步反转录成 cDNA,然后在以自身编码的整合酶(integrase)的作用下,插入到基因组中去。这是一类经 RNA 介导的由"DNA 到 RNA 再到 DNA"的返还型(retro-type)可转座基因[4]。

返还型可转座基因包括两种,一种是反转座子(retrotransposon),它的两端含有长度大且方向相同的长末端重复序列(long terminal repeat,LTR),一种是返座子(retroposon),像长散布核元件(long interspersed nuclear element,LINE)和短散布核元件(short interspersed nuclear element,SINE)等就属返座子。现阶段植物中含有 SINE 和 LINE 这类返座子的例子还不多见[5],但无论是被子植物还是裸子植物中,反转座子都广泛存在[6,7]。水稻和玉米中就存在许多反转座子,而且这些反转座子的长度和碱基序列都表现出不同程度的多样性[8-10]。

一、W_x 基因中的可转座基因

亚洲稻(*Oryza sativa*)是非洲稻(*Oryza glaberrima*)W_x 基因的第 10 个内含子的长度不同。序列分析比较表明,亚洲稻 W_x 基因的碱基序列中有一包括 125 个碱基对的插入序列,而且这一序列具有 SINE 的特征。利用它们的这一差异,仅水稻 W_x 基因中就发现有 2 个拷贝的 p-SINE1(r1 和 r2),而且还发现第 13 个内含子中存在一个类似转座子的基因 *Tnr*1[11]。此外,在 3′端下游还发现了以 Stowaway 命名的基因(图 1A)[6]。图 1B 所示为玉米的 W_x 基因[12],其内含子中多处插有反转座子。

1. p-SINE1:植物中首次发现的返座子

从水稻 W_x 基因的第 1 和第 10 内含子中找到的 125 bp 的插入序列,是在植物中首次发现的 SINE,定名为 p-SINE1[11,13]。人类 SINE 的 Alu序列最初可能起源于7S的RNA,然后

本文由国家重点基础研究专项经费资助(批准号:G1999011700)。

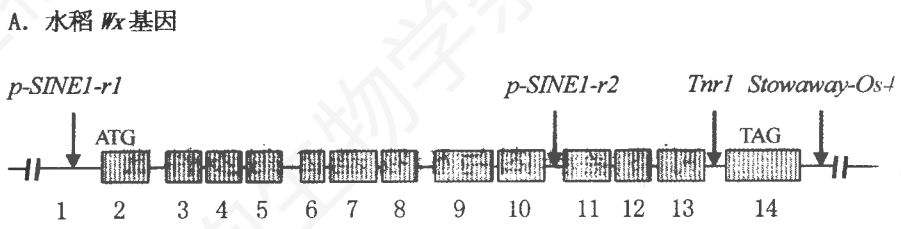

图 1 插入水稻和玉米 W_x 基因内含子中的多种可转座基因
线框表示外显子。

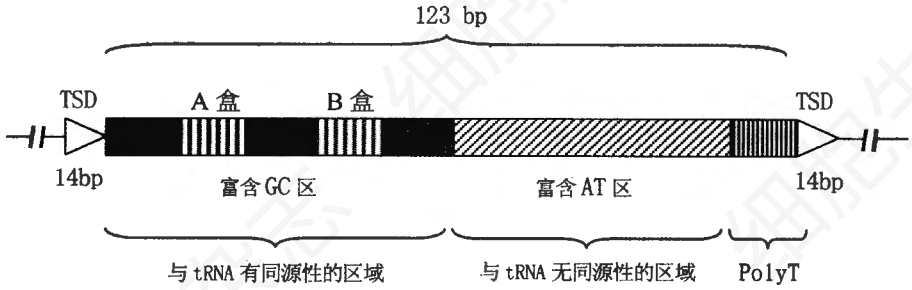

图 2 p-SINE1-r1 的结构
TSD 表示插入的两端靶位点序列，A 盒和 B 盒表示 RNA 聚合酶的启动子区域。

不断进化而来。而 p-SINE1 的 5′末端拥有与 tRNA 同源性的区域(图 2)[14]，推测它可能起源于 tRNA。p-SINE1 在亚洲稻的 W_x 基因中仅有 2 个拷贝，在水稻基因组上竟存在高达 4 500个以上的拷贝，而经多克隆(multiclone)后更可见到各种各样的部位都有它的出现。Motohashi 等采用 IPCR(inverse PCR)法和散弹枪(shot-gun)法从质粒基因库中分离出了近60个存在于不同基因位点的 p-SINE1[13,15]。

如前所述，p-SINE1-r1 和 r2 插入亚洲稻 W_x 基因的第 1 和第 10 内含子中，而非洲稻 W_x 基因的第 10 个内含子中却没有 p-SINE1-r2。这表明亚洲稻和非洲稻的基因型不相同。同样的情况在 p-SINE1-r6 上也出现。r6 在水稻的 RFLP 图谱上应该是位于第 11 染色体上的 p-SINE1。采用 PCR 法对 p-SINE1 存在与否的分析结果，在第 11 染色体上的对应位置上，亚洲稻存在 SINE，而一部分非洲稻上却找不到这样的 SINE。至于基因组上由 RNA 反转录来的返还型可转座基因，由于它一旦插入就无法再将其切除，因此不仅 SINE，就连 LINE 和反转座子也可以成为系统分化的良好指标。像 r32 和 r102 那样，由于在 SINE 的插入位点附近可以找到$(CAT)_n$重复和$(AG)_n$重

复,它们在系统分类上也将非常有用。例如,联结在 p-SINE1-r12 上的$(CAT)_n$重复的数目,非洲稻有 5 个、印度稻有 11 个、日本稻有 13 个,可见$(CAT)_n$重复的数目在品种之间存在明显差异。

2. *Tnr*1:一种新发现的散布性转座子

如前面图 1 所示,水稻 W_x 基因的内含子上存在着另一个类似转座子的 *Tnr*1 基因。它与 p-SINE1 一样,同属散布重复序列,在基因组上也以多拷贝的方式分散存在,每个单倍体上的数目高达 3,500 多个[16]。对克隆出的多个 *Tnr*1 的靶序列进行比较时发现,6 个靶序列中,5 个具有 TA 重复。由此可以推断,*Tnr*1 极有可能是一个靶序列特异性非常高的基因。与此相反的是,末端同样具有较长反向重复序列的 *Tnr*2 和 *Tnr*5,靶序列长度仅分别为 9 bp 和 4 bp(表 1),碱基序列中没有发现任何特异性。

表 1 水稻染色体上的一些转座子及类似于转座子的基因

基 因	长 度 (bp)	末端重复序列(bp)	靶序列 (bp)	家族名称	插入位点
*Tnr*1	235	75	2(TA)	*Tnr*1-*Stowaway*	W_x 基因的第 13 内显子
*Tnr*2	157	56	9	*Mu*?	*p*-SINE1-r4
*Tnr*3	1536	13	3	*En*/*Spm*	*p*-SINE1-r38
*Tnr*4	1768	53	9	(New)	*Tnr*1*G*
*Tnr*5	208	35	4	(New?)	*Tnr*1*H*

*Tnr*1 毕竟只是一个非常短小的、全长仅为 235bp 的可转座基因,以 *Tnr*1 的长度,不可能像 *Ac*、*Spm* 或 *Mu* 等大型转座子那样自己编码转座所必须的酶(转座酶)。*Tnr*1 的转座机制到底属经由 RNA 介导的返还型,还是直接由"DNA 到 DNA"的 DNA 型实际上仍未清楚。该基因不存在类似 SINE 或反转座子那样对反转录来说必不可少的区域,加之末端上具有反向重复序列(IR)等,似乎具有 DNA 型转座子的特征,但目前还没有发现其自主性因子(autonomous element)的存在。考虑到大量 *Tnr*1 插在基因组上这一事实,我们认为它在基因组上可能拥有隶属于自己的自主性因子。

另一方面,Bureau 等最近发现插入玉米 W_x5 基因位点的一个类似于转座子的 128 bp 的序列,并将其命名为 *Tourist*[17]。利用这一序列在基因数据库中进行检索时发现,在玉米、高粱、水稻、大麦等物种中都发现大量 *Tourist* 的同源性序列。这些序列与前面提到的 *Tnr*1 一样,为两端拥有长 IR 的短基因群(113－334 bp),其靶序列重复为 3 bp(TAA)。已经鉴定出的 36 个 *Tourist* 当中,绝大多数插入附近的 5′端上游、3′端下游或内含子中。所占比例分别为 1/2、1/4 和约 1/3[17,18]。此外,在高粱的 *Tourist*-sb5 上插有一个新基因 *Stowaway*[6]。该基因是在插入时 2 bp 的靶序列 TA 出现重复的结果,与 Tenzen 等发现的 Tnr1[19]属于同一家族。

二、活性反转座子

与酵母和果蝇的反转座子不同,已知植物反转座子的调控主要发生于转录水平。正常生长条件下,植物反转座子处于非活化状态,只有在环境胁迫条件下,如组织培养和被细菌、真菌或病毒感染等,才变为活化状态。

反转座子是水稻的一类主要可转座基因,水稻基因组中反转座子的全部拷贝估计约为 1000 个,其中有 32 个反转座子家族得到了分离鉴定。虽然这些反转座子在正常生长条件下似乎并无活性,但 32 个家族当中的 5 个家族至少在组织培养条件下处于活性状态,其中 *Tos*17 是最具活性的元件,对它的研究也最为

详细。Hirochika 的研究表明，*Tos*17 的活性调节主要发生在转录水平[20]。转座靶位点分析表明，*Tos*17 活性的发挥是水稻中组织培养诱发突变的重要原因。组织培养条件下 *Tos*17 的活化已经被用于开发位点选择诱变系统(site-selected mutagenesis system)，它的这一性质还可用于基因组分析，比如作为 RFLP 标记物和用作基因标记及功能分析的工具等。

烟草反转座子 *Tto*1 是植物中少有的几个活性反转座子之一，其完全核苷酸序列已经确定。序列分析表明，*Tto*1 具有经反转录完成自主性转座所要求的全部功能。对 *Tto*1 基因的构成及转录产物的性质进行分析则会发现，Tto1 起用的基因表达调控机制非常特殊。为研究 Tto1 在异源寄主中的自主性转座机理，将其引入水稻，并在水稻细胞中观察到了 *Tto*1 的转录和转座。为通过反转录方式探测到自主性转座的发生，对引入水稻细胞中的 *Tto*1 反转座子进行修饰，将其中一个反转录酶基因的一部分替换成含有内显子的抗潮霉素(hygromycin)基因。结果发现，内显子丢失只有在完整的 *Tto*1 发生共转染(cotransfection)时才能观察到[21]。上述结果表明，*Tto*1 可借助于反转录而进行自主性转座。根据这样的结果我们可以认为，转座所必需的寄主因子无论是在单子叶植物还是在双子叶植物中都得到了很好的保存，尽管两者远在大约 200 万年之前就已经分化开来。*Tto*1 的 LTR 在再生水稻植株叶片中具有转录活性，而在烟草叶片中则完全没有活性，表明反转座子在新寄主细胞中的行为发生了某种变化。对这些变化的进一步研究无疑将有利于加深对反转座子、寄主、及进化和基因组多样性之间可能存在的内在联系的理解。

Vicient 等对大麦属(*Hordeum*)植物 *BARE*-1 的整合酶、反转录酶和 LTR 结构域拷贝数量进行了比较，以评估具有转录活性的 *BARE*-1 反转座子家族在基因组进化中的作用。结果发现，基因组中平均拥有 13.7×10^3 个 *BARE*-1 的全长拷贝(full-length copy)，长度占整个基因组的 2.9%，而且拷贝数量还随着基因组的增大而增加[22]。每个完整反转座子的内部结构域(internal domain)通常与两个 LTR 相连，而 *BARE*-1 反转座子家族的内部结构域显然不遵守这一规则，LTR 的数量远远超出，并随着基因组大小的增大和 *BARE*-1 在基因组中所占比例的减小而增加。可以认为，不同 LTR 之间在同一染色体上发生的同源性重组可能是导致其数量超出的原因，因为这样的重组剔除了 *BARE*-1 元件而将 solo LTR 留下，从而减小了基因组中功能性反转座子的补充，并至少部分提供了从基因组肥胖症(genomic obesity)中解脱出来的可能途径。

三、采取重叠基因构造的反转座子

水稻的反转座子中，众所周知的是在分离原生质体或组织培养过程中被活化了的一个由 Tos 家族成员构成的群体[23]，上面提到的 *Tos*17 是这家族的典型代表。Noma 等最近在一种大洋洲野生稻(*Oryza australiensis*)的基因组发现一个拷贝数高达近 1 万的高频高反转座子 *RIRE*1，并确定了其全部构造。如图 3 所示，*RIRE*1 的两端具有较长的 LTR，内部则拥有典型的重复型反转座子构造。可能是由于以高达一万个拷贝存在于基因组上的缘故，*RIRE*1 反转座子在大洋洲野生稻的基因组上，在自己的序列中发生再度插入，也就是说它以一种重叠基因(nested gene)的构造形式出现的频度很高。这种大洋洲野生稻与栽培稻相比，基因组的大小增加了两倍[24]。基因组大小的变化很大程度上归因于 *RIRE*1 的增加[25]。用 PCR 法对 *RIRE*1 的 LTR 进行分析，发现图 3 *RIRE*1 上游(左侧)的 LTR 中有许多长短不一的 *RIRE*1 片段插入。可以认为，正是因为这种对部位或区域不加任何选择的插入变异的出现，可能使基因组上存在着多重拷贝的 *RIRE*1 中的大部分失去了活性。

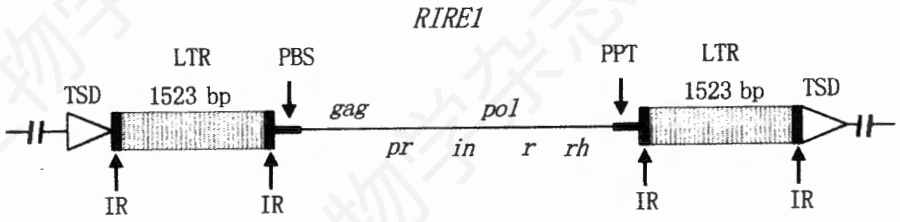

图 3 水稻反转座子 *RIRE*1 的构造

TSD 表示 *RIRE*1 转座子插入时产生的两端靶序列；
LTR 表示两侧的长末端重复序列；IR 表示两侧的反向重复序列。

上述重叠基因构造在其他反转座子或转座子中也经常出现。如水稻 p-SINE1 返座子中插有转座子 *Tnr*2 和 *Tnr*3[14,16]、*Tnr*1 转座子中插有转座子 *Tnr*4 和 *Tnr*5[26] 以及 *Tnr*2 转座子中插有反转座子 *RIRE*4[26] 等。可见在可转座基因中再插入可转座基因的例子很多。

四、可转座基因上的变异累积

将像 p-SINE1 和 *Tnr*1 这类散置型重复序列经多数分离以后比较其碱基序列，发现它们之间有 10% - 20% 的差异性[11,14,19]。不难看出，其中有一部分短小的插入、缺失或串联重复序列，但大部分是因碱基置换而产生的变异，而且所产生的这些变异的绝大部分源于由 C 开始到 T、或由 G 开始到 A 结尾的转换(transition)。这种转换在诸如 p-SINE1 或 *Tnr*1 这些以高频度散置于基因组上的重复序列中可以见到。例如，对日本稻的 W_x 基因和非洲稻的 W_x 基因的序列进行比较时，可以看出在 W_x 基因中的 p-SINE1 或 *Tnr*1 上存在比较集中的碱基置换。就其出现频度来说，散布重复序列同相邻的序列相比可高出 4 - 7 倍。从水稻基因组中单独克隆出的 11 个 *Tnr*1 基因，就其碱基置换变异所占的比例而言，有约 70% 出现的是由 C/G 开始到 T/A 结束的转换方式。即便是这样，只要简单计算一下便可发现，发生于 *Tnr*1 上的置换变异的频度比发生于相邻区域的置换变异频度要高出 3.4 倍。

由 C/G 开始到 T/A 结尾的这一转换，通常认为是由其中的 C 经脱氨基化而变换为 T 引起的，但正如利用脉孢菌(*Neurospora*)之 RIP(repeat-induced point mutation)体系的研究报道[27]的那样，因发生碱基置换而导致重复序列失去活性这样一种机制有可能起一定作用。此外，像在 p-SINE1-*r*29 或 *r*103 中见到的那类在 SINE 内部发生的串联重复，或者像发现转座子 *Tnr*2 或 *Tnr*3 插入 p-SINE1 中这样的例子，还有像反转座子 *RIRE*1 中再度以高频度插入 *RIRE*1 这样的现象，同样也可以成为重复序列上累积变异的例证。

五、可转座基因与异染色质

根据最近的报道，源于串联重复序列的异染色质(heterochromatin)区域中也累积着可转座基因。其中一些可转座基因是 Pimpinelli 等在果蝇(*Drosophila*)的唾液腺染色体上发现的。他们采用 FISH(fluorescence *in situ* hybridization)法成功地鉴定了插入异染色质区域上的、果蝇的主要可转座基因 *copia*、*gypsy*、*P*、*hobo* 和 *Bari-l* 等[28]。由于采用 FISH 这样的方法都能将这些可转座基因鉴定出来，可以断定它们在其所处区域内拥有的拷贝数相当多。Pelissier 等也曾报道在拟南芥菜的异染色质上发现插有反转座子[29]。以水稻为材料的研究也有一定进展，采用 PCR 法鉴定出染色体 sub-端粒区域的串联重复序列 TrsA 中插有反转座子 *RIRE*2 或 *RIRE*4 和转座子 *Tnr*2，相信这一研究今后会取得更大的进展。

以上分析了内含子、5′端上游和 3′端下游的非转录区域中可转座基因的累积，以及可转

的非转录区域中可转座基因的累积,以及可转座基因在其自身中出现的累积。所涉及的可转座基因主要包括反转座子、SINE 以及一些新发现的转座子等,它们均是在基因组上以高频度出现的散布重复序列。这些可转座基因中,累积了各种各样的因碱基置换和缺失等导致的变异,因此碱基序列出现多样化也就是情理之中的事情。就连原有的转移(metastasis)活性,恐怕大部分也已经丢失殆尽。Riggs 等曾将人类相似于 Tc1 的 *Tigger* 基因称之为"化石"(fossile)[30],现在看来对 *Tigger* 基因的这一称谓可以说既形象又恰如其分。虽然目前仍有许多奥秘尚未揭晓,但随着对这些"fossile"存在方式的不断探索和解明,或许能够搞清楚可转座基因与寄主基因组之间共生方式的来龙去脉。曾有推测认为,一个物种的转座子,伴随着启动子步移(promotor scrambling)的发生,一旦基因表达数量发生改变,反转座子的 LTR 有可能出来充当或扮演某个基因的启动子的角色。由此可见,生物在进化过程中似乎在巧妙地利用可转座基因可以移动的特点来应付必须面对的外界环境的变化。

摘 要

高等植物基因组含有大量各式各样的串联重复序列和出现频率很高的散布重复序列,如转座子、反转座子、短散布核元件和一些新发现的小型转座子等,它们当中的大多数是具有移动能力的可转座基因。这些可转座基因在漫长的进化过程中对基因和基因组多样性的形成所起的作用,成为近年来分子生物学领域中的重要研究内容。

参 考 文 献

[1] Nakajima, R. et al., 1996, *Genes Genet. Syst.*, **71**: 373 – 382.
[2] Bennetzen, J. L., 1996, *Trends Microbiol.*, **4**: 347 – 353.
[3] Wright, D. A., et al., 1996, *Genetics*, **142**: 569 – 578.
[4] Wright, D. A. and Voytas, D. F., 1998, *Genetics*, **149**: 703 – 715.
[5] Noma, K., et al., 1999, *Mol. Gen. Genet.*, **261**: 71 – 79.
[6] Voytas, D. F., et al., 1992, *Proc. Natl. Acad. Sci. USA*, **89**: 7124 – 7128.
[7] Hirochika, H. and Hirochika, R., 1993, *Jpn. J. Genet.*, **68**: 35 – 46.
[8] White, S. E., et al., 1994, *Proc. Natl. Acad. Sci. USA*, **91**: 11792 – 11796.
[9] Kumekawa, N., et al., 1998, *Mol. Gen. Genet.*, **260**: 593 – 602.
[10] Suoniemi, A., et al., 1998, *Plant J.*, **13**: 699 – 705.
[11] Umeda, M., et al., 1992, *Jpn. J. Genet.*, **66**: 569 – 586.
[12] Varagona, M. J., et al., 1992, *Plant Cell*, **4**: 811 – 820.
[13] Motohashi, R., et al., 1997, *Theor. Appl. Genet.*, **95**: 359 – 368.
[14] Mochizuki, K., et al., 1992, *Jpn. J. Genet.*, **67**: 155 – 166.
[15] Motohashi, R., et al., 1996, *Mol. Gen. Genet.*, **250**: 148 – 152.
[16] Tenzen, T. and Ohtsubo, E., 1991, *J. Bacteriol.*, **173**: 6207 – 6212.
[17] Bureau, T. E. and Wessler, S. R., 1992, *Plant Cell*, **4**: 1283 – 1294.
[18] Bureau, T. E. and Wessler, S. R., 1994, *Proc. Natl. Acad. Sci. USA*, **91**: 1411 – 1415.
[19] Tenzen, T., et al., 1994, *Mol. Gen. Genet.*, **245**: 449 – 455.
[20] Hirochika, H., 1997, *Plant Mol. Biol.*, **35**: 231 – 240.
[21] Hirochika, H., et al., 1996, *Plant Cell*, **8**: 725 – 734.
[22] Vicient, C. M., et al., 1999, *Plant Cell*, **11**: 1769 – 1784.
[23] Hirochika, H., et al., 1996, *Proc. Natl. Acad. Sci. USA*, **93**: 7783 – 7788.
[24] Martinez, C. P., et al., 1994, *Jpn. J. Genet.*, **69**: 513 – 523.
[25] Noma, K., et al., 1997, *Genes Genet. Syst.*, **72**: 131 – 140.
[26] Kachroo, P., et al., 1994, *Mol. Gen. Genet.*, **245**: 339 – 348.
[27] Selker, E. U., 1990, *Annu. Rev. Genet.*, **24**: 579 – 613.
[28] Pimpinelli, S., et al., 1995, *Proc. Natl. Acad. Sci. USA*, **92**: 3804 – 3808.
[29] Pelissier, T., et al., 1995, *Plant Mol. Biol.*, **29**: 441 – 452.
[30] Smit, A. F. A. and Riggs, A. D., 1996, *Proc. Natl. Acad. Sci. USA*, **93**: 1443 – 1448.