

DOI:10.11913/PSJ.2095-0837.2021.50543

郑永强, 邹洁, 温兴鹏, 吴志国. 植物转座元件注释及生物学功能研究进展[J]. 植物科学学报, 2021, 39(5): 543-551

Zheng YQ, Zou J, Wen XP, Wu ZG. Research progress on annotation and biological function of plant transposable elements[J]. *Plant Science Journal*, 2021, 39(5): 543-551

# 植物转座元件注释及生物学功能研究进展

郑永强<sup>1</sup>, 邹洁<sup>1</sup>, 温兴鹏<sup>2</sup>, 吴志国<sup>1\*</sup>

(1. 武汉大学生命科学学院, 武汉 430070; 2. 武汉大学高等研究院, 武汉 430070)

**摘要:** 转座元件是指在基因组中能够移动、复制并重新整合到基因组新位点的 DNA 片段。在植物中, 多种类型的转座元件, 特别是占比较高的 LTR 类逆转录转座元件, 可以通过产生新基因和转录本、提供调节元件、改变基因结构等多种途径广泛调控基因表达, 最终多维度有效推动基因组进化。同时, 基因组测序组装技术的快速发展也为转座元件的检测、注释提供了良好契机。本文从结构分类、全基因组检测、功能研究、基因组进化 4 个方面对当前植物转座元件的研究进展进行综述, 同时对今后的研究方向进行了展望。

**关键词:** 植物转座元件; 基因组进化; 检测注释; 功能研究

中图分类号: Q943.2

文献标识码: A

文章编号: 2095-0837(2021)05-0543-09

## Research progress on annotation and biological function of plant transposable elements

Zheng Yong-Qiang<sup>1</sup>, Zou Jie<sup>1</sup>, Wen Xing-Peng<sup>2</sup>, Wu Zhi-Guo<sup>1\*</sup>

(1. School of Life Science, Wuhan University, Wuhan 430070, China;

2. Institute for Advanced Studies, Wuhan University, Wuhan 430070, China)

**Abstract:** Transposable elements (TEs) are fragments of DNA that can move, copy, and re-integrate into a new site in the genome. In plants, many types of TEs, especially abundant long terminal repeat (LTR) retrotransposons, can regulate gene expression in multiple ways, such as producing new genes and transcripts, providing regulatory elements, and changing gene structure. Ultimately, these changes promote plant genome evolution. At the same time, the rapid development of genome sequencing and assembly technology has provided a good opportunity for TE detection and annotation. In this review, we summarize current research progress on plant TEs from the perspectives of structural classification, whole-genome detection, functional research, and genome evolution. We also discuss future research directions and provide references for further plant TE study.

**Key words:** Plant transposable elements; Genome evolution; Detection and annotation; Functional research

转座元件 (Transposable elements, TEs) 作为动植物基因组中神奇的“暗物质”, 曾一度被认为是垃圾 DNA (Junk DNA)。随着 70 年前 Barbara McClintock 在玉米 (*Zea mays* L.) 中发现转座元

件的 Ac-Ds 系统基础性作用以来, 这种“跳跃”的基因不断被科学家们所熟知, 大肠杆菌的“乳糖操纵子模型”也从侧面证明了转座元件的功能。由此, 这种“暗物质”的面纱逐渐被科学家们揭开。

收稿日期: 2021-02-23, 修回日期: 2021-04-12。

基金项目: 国家自然科学基金项目 (21602162)。

This work was supported by a grant from the National Natural Science Foundation of China (21602162).

作者简介: 郑永强 (1997-), 男, 硕士研究生, 研究方向为植物基因组学 (E-mail: zhengyq@whu.edu.cn)。

\* 通讯作者 (Author for correspondence. E-mail: wu.zhiguo@whu.edu.cn)。

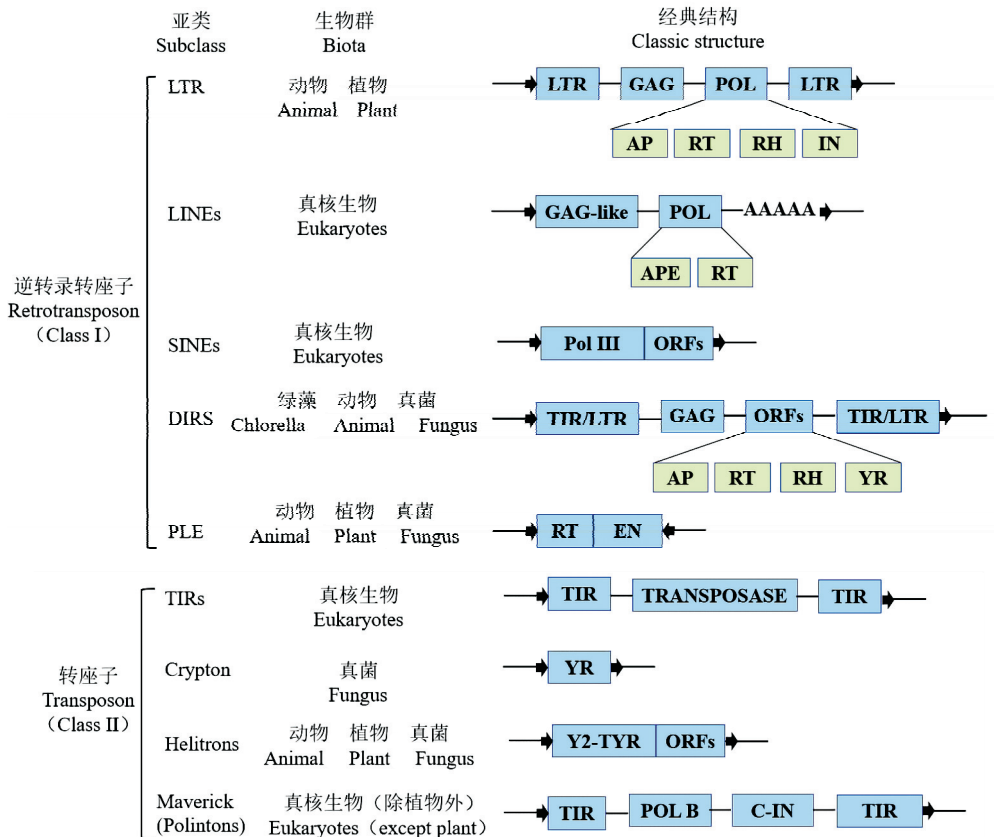
在作物中研究发现，转座元件可以通过顺式或反式作用方式调控部分基因表达，同时对基因组整体进化进程产生影响，有效提高了植物对环境的适应性及遗传多样性。此外，重复序列的发现和识别技术也成为基因组学的研究热点。本文针对高等植物转座元件的结构分类、注释策略、构建基因调控网络、推动基因组进化 4 个方面的研究进行总结，同时提出今后转座元件领域研究方向，以期为今后开展转座元件生物学意义发掘，种质资源鉴定和应用，遗传育种等方面的研究提供借鉴。

### 1 转座元件分类、结构特征

真核生物基因组中存在大量的重复序列，根据结构特征，研究人员将其分为 3 大类型<sup>[1]</sup>。第 1

类是拷贝数较少的片段重复 (Segmental duplications)。第 2 类是串联重复序列 (Tandem repeats)，包括微卫星 DNA、简单序列重复、小卫星 DNA、卫星 DNA 4 种。第 3 类是含量最多的散在重复 (Interspersed repeats)，这类型重复序列大多数起源于转座元件。

转座元件是指在基因组中能够移动、复制并重新整合到基因组新位点的 DNA 片段，它是基因组的重要组成部分，有学者认为可能起源于病毒或逆转录病毒退化<sup>[2]</sup>。按照经典的分类学方法，可将转座元件首先分为两大类<sup>[3]</sup>：RNA 类型的逆转录转座子 (Retrotransposon) 和 DNA 类型的转座子 (Transposon) (图 1)。其中逆转录转座子根据不同的转座机制又可以分为不同的亚类<sup>[4]</sup>，包括长末



LTR: 长末端重复序列; LINEs: 长散在核元件; SINEs: 短散在核元件; DIRS: 双歧杆菌中间重复序列; PLE: Penelope 类元件; GAG: 衣壳蛋白; AP: 天冬氨酸蛋白酶; RT: 逆转录酶; RH: RNA 酶 H; IN: 整合酶; APE: 无嘌呤核苷酸内切酶; ORFs: 开放阅读框; TIR: 末端反向重复序列; YR: 酪氨酸重组酶; EN: 核酸内切酶; TRANSPOSASE: 转座酶; Y2-YR: Y2 类型的酪氨酸重组酶; POL B: DNA 聚合酶 B; C-IN: c-int 类型整合酶; Pol III: RNA 聚合酶 III; POL: 编码多功能蛋白基因。

LTR: Long terminal repeats; LINEs: Long interspersed nuclear elements; SINEs: Short interspersed nuclear elements; DIRS: Dictyostelium intermediate repeat sequence; PLE: Penelope-like elements; GAG: Capsid protein; AP: Aspartic proteinase; RT: Reverse transcriptase; RH: RNase H; IN: Integrase; APE: Apurinic endonuclease; TIR: Terminal inverted repeats; YR: Tyrosine recombinase; EN: Endonuclease; Y2-YR: Y2 type tyrosine recombinase; POL B: DNA polymerase B; C-IN: c-int type integrase; Pol III: RNA polymerase III; POL: Genes encoding multifunctional proteins.

图 1 转座元件分类及结构特征

Fig. 1 Classification and structure of transposable elements

端重复序列 (Long terminal repeat, LTR)、长散在核元件 (Long interspersed nuclear elements, LINEs)、短散在核元件 (Short interspersed nuclear elements, SINEs)、双歧杆菌中间重复序列 (Dictyostelium intermediate repeat sequence, DIRS)、Penelope 类元件 (Penelope-like elements, PLE)。其中 LINEs 相关转录序列可能参与了 SINEs 的转座过程<sup>[5]</sup>，但是 LINEs 和 SINEs 的起源进化关系目前还不清楚<sup>[6]</sup>。DNA 类型的转座子根据不同的结构特征又进一步分为两大类，第 1 类包括末端反向重复转座子 (Terminal inverted repeats, TIRs) 和 Crypton 转座子，第 2 类包括 Helitrons 转座子和 Maverick 转座子 (也称 Polintons 转座子)。

不同类型转座元件转座机制各不相同。RNA 类型逆转录转座子的运动过程有点类似于文本编辑中的“复制+粘贴”。在转座时会先以 DNA 为模板，在 RNA 聚合酶 II 和逆转录酶的作用下，把 DNA 转录为 RNA 后再反转录为 cDNA，最后通过整合酶整合到基因组中完成转座。DNA 类型的转座子的转座机制被称为“剪切+粘贴”，此类转座元件会直接在转座酶 (Transposase) 的作用下，通过交错双链裂解方式从染色体上切除下来，然后重新插入到基因组的其他位置<sup>[4]</sup>。逆转录转座子中的 LTR 类转座元件在植物中占比较高，转座过程也较为复杂。这类转座子会首先经过转录形成成熟的 mRNA，其中大多数 mRNA 编码转座活动所必须的蛋白，少部分 mRNA 则包裹在 gag 编码的衣壳蛋白里共同构成病毒样颗粒 (Virus like particles, VLPs)。负责编码蛋白的 mRNA 会结合逆转录酶产生 cDNA 并最终形成双链，由 pol 所编码的核酸内切酶切割后再被 IN 结构域的整合酶整合进入基因组完成转座<sup>[7]</sup>。此外，Helitrons 转座元件是近年来发现的一种新型 DNA 转座元件，和大多数转座元件不同，Helitrons 通过滚环的方式进行转座<sup>[8]</sup>。

## 2 全基因组注释转座元件策略

目前，基因组测序技术日新月异，测序成本不断降低，单次读取长度从数十个碱基增加到数万个碱基。1977 年 Sanger 和 Coulson 发明一代测序技术，得到噬菌体 X174 的完整基因组序列，标志

着物种基因测序时代的到来<sup>[9]</sup>。以荧光信号获得测序信息的二代测序技术<sup>[10, 11]</sup>，特别是 Illumina 公司的 Solexa 平台<sup>[12, 13]</sup>的广泛使用，更是推动了测序技术的普及化发展。而第三代测序，即单分子测序技术 (Single molecule sequencing) 读长已经达到 10 kb，不需要 PCR 扩增的情况下就可以对高 GC 区域进行测序<sup>[14, 15]</sup>。此外，2009 年由 Job Dekker 和 Eric Lander 开发出的 Hi-C 技术<sup>[16]</sup>，能捕获全基因组范围的相互作用，分辨率在 1 Mb 左右，为三维组装基因组提供了可能。测序技术的不断革新推动了高质量基因组的组装，同时为转座元件的检测、注释提供了良好机遇。目前，多数研究者认为注释转座元件主要有以下 3 种策略<sup>[17]</sup> (表 1)。

第 1 种是根据转座元件数据库中的共识序列和基序进行比对查询。例如当前广泛使用的主流软件 RepeatMasker<sup>[18]</sup> 和 Greedier<sup>[19]</sup> 便是利用 RepBase、Dfam 数据库对真核生物基因组进行序列比对，将转座元件进行同源预测注释。CLARI-TE<sup>[20]</sup> 可以用 ClariTERep 数据库对植物基因组，特别是小麦 (*Triticum aestivum* L.) 基因组中多种重叠重复序列进行鉴定；TESeeker<sup>[21]</sup> 可以通过反复调用 ClustalW2、BLAST、CAP3 等多序列比对工具，得到所要搜索基因组中的转座元件。MGEScan-LTR<sup>[22]</sup> 可以利用数据库中 LTR 类转座元件序列构建 HMM 特征谱搜索，特别是在 LTR 逆转座子家族与查询序列同源性较低时，该软件相对于传统搜索引擎具有显著优势。EDTA<sup>[23]</sup> 是一个综合性的流程工具，它可以整合当前多个主流软件的预测结果，生成过滤后的非冗余转座元件数据库，同时还可以识别高度重复的基因组区域中常见的转座元件嵌套。

第 2 种方法是根据已组装基因组的从头 (de novo) 预测。这种预测方法的核心在于利用比对软件将基因组与基因组进行比对，根据基因组序列比对结果转换成多序列比对，最后聚类获得转座元件家族<sup>[24]</sup>。例如 Repeat Pattern Toolkit<sup>[25]</sup>、RECON<sup>[26]</sup>、RepeatScout<sup>[27]</sup> 等可以通过寡核苷酸或种子序列聚类，随后通过调用 RepeatMasker 对转座元件进行分类。RepeatModeler2<sup>[28]</sup> 是 2020 年最新开发的软件，可以调用 RECON 和 RepeatScout 组成管道流程对转座元件进行注释，相对于初始版本 RepeatModeler 而言，在 LTR 类转座元

表 1 转座元件注释资源特点  
Table 1 Characteristics of transposable element annotation resources

注释策略 Annotation strategy	软件名称 Software	预测转座元件类别 Predict category	适用物种 Species	开发时间 Start year	参考文献 Reference
数据库 比对	RepeatMasker	全部类别	真核生物	2016	[ 18 ]
	Greedier	全部类别	真核生物	2008	[ 19 ]
	CLARI- TE	LTR 类	植物	2014	[ 20 ]
	TESeeker	全部类别	真核生物	2011	[ 21 ]
	MGEscan-LTR	LTR 类	植物	2007	[ 22 ]
	EDTA	全部类别	动物、植物	2019	[ 23 ]
已组装基 因组的从 头预测	RECON	全部类别	真核生物	2002	[ 26 ]
	RepeatScout	全部类别	真核生物	2005	[ 27 ]
	RepeatModeler2	全部类别	真核生物	2020	[ 28 ]
	phRAIDER	全部类别	真核生物	2016	[ 29 ]
	TIR-Learner	TIR 类	植物	2019	[ 30 ]
	PILER	全部类别	动物、植物	2005	[ 31 ]
	LtrDetector	LTR 类	植物	2019	[ 32 ]
	LTRdigest	LTR 类	植物	2008	[ 33 ]
	LTRharvest	LTR 类	植物	2009	[ 34 ]
	LTR_FINDER	LTR 类	植物	2007	[ 35 ]
	测序 直接组装	REPuter	全部类别	真核生物	2001
RepeatExplorer2		全部类别	真核生物	2020	[ 37 ]
dnaPipeTE		全部类别	真核生物	2015	[ 38 ]
Tedna		全部类别	真核生物	2018	[ 39 ]
RepLong		长重复序列类	真核生物	2018	[ 40 ]

件注释敏感性方面有了长足的进步。phRAIDER<sup>[29]</sup>则是用 PatternHunter 间隔种子序列训练方法对高度重复的寡核苷酸和种子序列进行聚类。TIR-Learner<sup>[30]</sup>可以将同源性检测方法与基于结构的方法相结合,同时采用机器学习(ML)算法,对植物的 TIR 类转座子进行从头识别。PILER<sup>[31]</sup>可以用 PILER-DF、PILER-TA、PILER-TR 等不同比对模式对基因组中散在重复、串联重复、末端重复等不同类型进行鉴定。LtrDetector、LTRdigest、LTRharvest、LTR\_FINDER<sup>[32-35]</sup>等主要依据 LTR 序列的大小范围、每个末端出现的靶位点重复 TSD 等 LTR 结构特征对转座元件进行鉴定。

第 3 种是直接根据原始测序读长(reads)组装转座元件。这种方法的思路是根据 reads 检索重复出现的定长 Kmer 种子短序列,然后再将它扩展为更长的序列。例如 REPuter、RepeatExplorer2、dnaPipeTE<sup>[36-38]</sup>等软件可以直接利用测序 reads 对转座元件进行定量注释,特别是可以计算出转座元件插入时间(年龄)。与此同时,各个软件又各有特色,REPuter 可以将用户输入的 fasta 序列转换成 ASCII 码进行存储计算,其最大的优势在于计

算的快速高效和计算资源的节约利用;RepeatExplorer2 可以基于图聚类算法对真核生物重复序列 DNA 进行鉴定,同时提供转座元件的自动注释和分析结果的可视化等多项功能;dnaPipeTE 可以利用二代测序 reads,采用 reads 和 contig 比较方式对转座元件进行检索。REPdenovo、RepARK、Tedna<sup>[39]</sup>可以用原始 reads 对所有类型转座元件进行序列组装。特别是 RepLong<sup>[40]</sup>软件,作为最新开发的专门处理三代测序 Pacbio 长 reads 的注释工具,具有特异性高、灵敏性好等多种优势,可以通过重叠鉴定、网络构建、聚类检测、重复序列提取、比对 RepARK 数据库等 5 个关键步骤对植物转座元件进行全面注释。

### 3 转座元件影响基因表达

转座元件可以通过基因组杂交、水平转移等方式进入宿主。但是插入后的转座元件能否调控基因表达与多种因素相关。大多数转座元件在进入宿主基因组后会被 DNA 甲基化、组蛋白修饰等表观遗传手段沉默掉<sup>[41]</sup>,本能地抑制或降低转座元件对自身基因组的影响。此外,不同类型转座元件插入

位置的差异也会对邻近基因表达产生不同影响。植物基因组内 LTR 亚家族的优先插入位置也有很大不同, Ty1-Copia 类逆转录转座子优先分布在转录活性强的常染色质区域, 而 Ty3-Gypsy 类逆转录转座子则优先分布在距离着丝粒较近的异染色质区域<sup>[42]</sup>, 不同的分布会直接导致这两种转座元件调控基因表达能力的差异。再如转座元件在玉米基因间隔区的分布明显靠近基因区域, 该特征可能与基因转录调控有关, 而在内含子区域, 分布偏向于内含子两端, 可能与内含子剪切有关<sup>[43]</sup>。

不同类型转座元件调控基因表达的方式不尽相同, 转座元件和转座元件衍生物可以通过多种途径影响宿主基因表达。

### 3.1 产生新基因和转录本

光、温、水、肥等胁迫性环境刺激可以激活转座元件, 随后转座元件会开始新的“跳跃”。如果插入到基因外显子内部就可能形成新的基因, 许多转座元件衍生序列会被转录并影响附近基因的表达<sup>[44]</sup>。转录后的转座元件产生的小 RNA 可以通过表观遗传调控的方式调节基因的表达, 例如可以参与 small RNA (包括 siRNA、miRNA、piRNA) 和 lncRNA 等非编码 RNA 的生成, 从而进一步参与基因表达、个体发育、表观遗传等多种生物学过程<sup>[45]</sup>。拟南芥 (*Arabidopsis thaliana* (L.) Heynh.) 中 LTR 反转座元件被激活时, 其转录本会被加工产生大量 siRNA854, 随后参与基因调控网络构建<sup>[46, 47]</sup>。转座元件的复制转位也会促成了许多 lncRNA 基因结构的变异及新生。例如在水稻 (*Oryza sativa* spp. *japonica* cv. Nipponbare) 中, MITEs 可以产生 siRNA441、siRNA446 这两种小 RNA, 它们除了参与自身 MITEs 转座活性的调控外, 还正向调节了水稻对脱落酸信号 (ABA) 的响应<sup>[48]</sup>。

### 3.2 提供调节元件, 调节基因表达

转座元件整合到不同调控元件的作用不同, 例如插入到基因增强子 (Enhancer) 中相当于抑制基因表达, 而插入抑制子 (Inhibitor) 则产生促进基因表达的效果。当转座元件本身含有增强子或抑制子等调控元件时, 插入基因组后可以直接调控基因表达<sup>[45]</sup>。转座元件插入顺式调控元件的序列之中会破坏转录因子的结合位点来阻止它结合到对应 DNA 序列上, 造成基因的异常表达, 或者转座元

件上存在一个新的顺式调控元件结合位点, 其它的转录因子结合在该位点上亦能促进基因表达。研究发现, 在紫色葡萄 (*Vitis vinifera* L.) 中一个名为 *Gret1* 的 Ty3-Gypsy 逆转录转座子整合到了 *VvmybA1* 基因上游的调控序列中, 直接造成了这个基因无法表达, 而这个基因与花青素合成密切相关, 由此产生了白色葡萄这个新品种<sup>[49]</sup>。再如水稻中, 一个名为 *Renovator* 反转座元件的 LTR 区含有一个启动子, 直接提高了稻瘟病抗性基因 *Pit* 的转录水平, 使水稻重新表现出相应的抗病活性, 极大提高了水稻的环境适应性<sup>[50]</sup>。

### 3.3 改变基因结构及转录本变异

转座元件可以在植物生殖细胞或者体细胞基因转录过程中, 产生新可变剪切来嵌入基因组中影响基因表达。这种调控方式同样已在多个物种中得到证实。例如橙子 (*Citrus sinensis*) 中有一种 MYB 转录激活花青素生产的基因被命名为 *Ruby*, 它是组织特异性和冷应答性表达的, 在其启动子区域插入一个类似于 Ty1-Copia 的逆转录转座子, 西西里岛血橙 (*Citrus sinensis* “blood orange”) 便产生了<sup>[51]</sup>。因此, 橙子和血橙的品种变异便是转座元件导致基因组重编程的结果。再如玉米中 *b1* 基因可以编码转录因子调控花青素合成途径, 在它的第一个外显子插入了高拷贝的逆转录转座子后, 它的两个等位基因 *B-Peru* 和 *B-Bolivia* 的部分表观遗传出现差异, 最终改变玉米籽粒颜色<sup>[52]</sup>。当豌豆 (*Pisum sativum* L.) 的 *SBE1* 基因外显子中插入一个 0.8 kb 的 Ac/Ds 家族类转座元件时, 会造成该基因转录水平大幅降低、淀粉合成受阻、渗透压改变, 于是便产生了孟德尔遗传实验中的经典材料皱粒豌豆<sup>[53]</sup>。

此外, 宿主基因组对转座元件的沉默也会对本身基因的表达产生影响。例如宿主在 DNA 甲基化沉默外来转座元件时, 可能会引发异染色质形成的扩散, 进而导致附近基因的沉默<sup>[45]</sup>。总之, 转座元件在新基因的形成、转录过程调控、转录后修饰和可变剪切等多方面均可影响基因的表达, 这为遗传材料的变异和多样性提供了丰富资源。

## 4 转座元件推动基因组进化

基因组的扩张和删减丢失这两种方向的作用力, 共同推动了物种基因组规模的进化, 而转座元

件积累和植物基因组多倍化则是导致基因组增大的主要动力<sup>[54-57]</sup>。转座元件可以通过频繁“跳跃”来扩增基因组规模，而周期性短时间内转座元件大规模拷贝数的扩增称为转座元件的“爆发”。此类爆发事件具有扩增基因组规模、增加物种内和物种间多样性、增强环境适应性潜力等多种生物学作用。

对于植物来讲，逆转录转座子由于其“复制+粘贴”的转座原理，使得该类转座元件对于基因组规模的影响尤为明显，特别是位于基因间隔区内的转座元件短时间内就可以增加大量拷贝，扩大基因组规模。此外，植物基因组中 LTR 转座元件占比颇高也是该类转座元件易导致基因组规模扩大的原因之一。例如玉米超过 75% 基因组由 LTR 类逆转录转座子构成<sup>[58]</sup>，小麦中该比例达到 56.9%，陆地棉 (*Gossypium hirsutum* L.) 和亚洲棉 (*G. arboreum* L.) 中分别有 60.8%、72.3% LTR 类转座元件存在于基因组中<sup>[59]</sup>。

转座元件含量与基因组规模的相关性在多个物种中均得到了证实。在距今约 5.7 ~ 0.61 Ma，棉花中发生了几次 LTR 型转座元件爆发事件，这对棉花 A 基因组 (草棉 *Gossypium herbaceum* L.、亚洲棉、At 亚基因组) 的扩张和物种多倍化形成做出了重要贡献<sup>[60]</sup>；在距今 0.5 ~ 0.3 Ma 范围内，Gypsy 类逆转录转座子 Daniela、Sumaya 和 Sumana 等家族的爆发，直接导致了黑麦 (*Secale cereale* L.) 基因组的扩张<sup>[61]</sup>；在近 3 Ma 内，水稻 (亚洲栽培稻 *O. sativa* 的野生近缘野生稻) 中拷贝数高达 90 000 的 3 个 LTR 家族成员发生了大量扩增，直接导致水稻基因组规模的倍增<sup>[62]</sup>。此外，研究人员通过分析小麦染色质空间结构，发现亚基因组内部染色体间高级结构互作与转座元件偏好密切相关，表明特异性转座元件参与维持了多倍体亚基因组结构的稳定<sup>[63]</sup>。

转座元件对基因组的影响同样为许多假设提供了解释。基因组大小可以用 C 值来衡量，即指物种单倍体或配子所含的 DNA 量。研究发现，基因组大小与生物体复杂性之间没有相关性，即所谓“C 值悖论”<sup>[64]</sup>。相反，转座元件含量与基因组大小有明显关联，这为“C 值悖论”的解释提供了依据。研究表明，多倍化后亲本亚基因组组分偏差 (subgenome fractionation bias)<sup>[65]</sup>，即亚基因组优势的成因是 24 nt 长小分子 RNA 抑制了转座元

件的活性，造成邻近基因表达受限，进而全基因组基因表达形成了差异<sup>[66, 67]</sup>。这些实例证明了转座元件在植物基因组进化系统中的重要作用。

## 5 展望

植物转座元件不论从微基因水平还是宏基因组上都对物种形成和进化起到了关键作用。转座元件的“跳跃”在单个基因和基因调控网络的形成中都扮演了重要角色，而转座元件的积累也同时扩张了基因组规模，特别是扩张了常染色质规模，有效推动了植物基因组的进化。不断迭代更新的长 reads 测序技术必将为转座元件的注释和功能揭示提供更好的平台，当前转座元件的研究已经逐渐受到了学者的广泛重视，不论是端粒酶活性研究还是转座元件参与的表观遗传调控都成为了研究热点。

但是，目前对于转座元件的研究尚存在不足。从转座元件家族分类来说，现有分类基本上涵盖了绝大多数的转座元件。但是随着基因组的不断进化，由于环境适应性等诸多因素导致的遗传积累仍然会产生新的转座元件家族，这就需要研究人员不断开发新的转座元件检测工具。同时，对于非模式物种而言，现有的重复序列数据库无法识别那些物种中未被认识的新的重复类型，需要利用脚本对重复序列进行训练，构建本物种的数据库。因此，提高转座元件数据库的精确度和覆盖度也同样迫在眉睫。此外，除了依赖于转座元件数据库的 DNA 水平注释方法，我们是否可以从注释原理方面另辟蹊径？例如借鉴转录组注释基因的方法，当前研究人员已经开始尝试在拟南芥 (*Arabidopsis thaliana* (L.) Heynh.)、水稻、大豆 (*Glycine max* (L.) merr.) 等植物中利用 24 nt-siRNAs 等小 RNA 对转座元件进行从头组装并取得了一定进展<sup>[68]</sup>。转座元件的另一个重要研究方向是功能发掘。转座元件爆发大多只涉及到少数几个类型，对于不同物种而言，究竟是哪类转座元件导致了祖先物种的分化，这类转座元件是否与当时极端特殊气候下环境适应性有关，是否适时调节了相关基因的表达？这些科学问题都是未来植物学转座元件研究领域的重要课题。

随着研究的不断深入，70 年前 Barbara McClintock 提出的转座元件具有调控功能的假设正在不断被证实。面向未来我们需要做的是让基因组中

转座元件注释更加精准全面,调控功能网络更加清晰,以此充分发掘植物遗传资源的多样性,增加植物环境适应的可塑性,逐步使其在面向国家战略需求的同时,更加广泛地应用在植物品种鉴定、作物分子育种等领域。

### 参考文献:

- [ 1 ] Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes [ J ]. *Nature*, 1994, 371(6494): 215–220.
- [ 2 ] 李宏. 转座子的起源及其和物种进化的关系 [ J ]. 渝州大学学报, 1993, 25(1): 47–56.  
Li H. The origin of transposons and their relation to species evolutions [ J ]. *Journal of Yuzhou University*, 1993, 25(1): 47–56.
- [ 3 ] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al.* A unified classification system for eukaryotic transposable elements [ J ]. *Nat Rev Genet*, 2007, 8(12): 973–982.
- [ 4 ] Jönsson ME, Garza R, Johansson PA, Jakobsson J. transposable elements: a common feature of neuro developmental and neuro degenerative disorders [ J ]. *Trends Genet*, 2020, 36(8): 610–623.
- [ 5 ] Ohshima K, Hamada M, Terai Y, Okada N. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements [ J ]. *Mol Cell Biol*, 1996, 16(7): 3756–3764.
- [ 6 ] 程旭东, 凌宏清. 植物基因组中的非 LTR 反转录转座子 SINEs 和 LINEs [ J ]. 遗传, 2006, 28(6): 731–736.  
Cheng XD, Ling HQ. Non-LTR Retrotransposons: LINEs and SINEs in plant genome [ J ]. *Hereditas*, 2006, 28(6): 731–736.
- [ 7 ] Roller M, Stamper E, Villar D, Izuogu O, Martin F, *et al.* LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions [ J ]. *Genome Biol*, 2021, 22(1): 62.
- [ 8 ] Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts [ J ]. *Nat Rev Genet*, 2011, 12(9): 615–627.
- [ 9 ] Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, *et al.* Characterization of the yeast transcriptome [ J ]. *Cell*, 1997, 88(2): 243–251.
- [ 10 ] Schuster SC. Next-generation sequencing transforms today's biology [ J ]. *Nat Methods*, 2008, 5(1): 16–18.
- [ 11 ] Shendure J, Ji H. Next-generation DNA sequencing [ J ]. *Nat Biotechnol*, 2008, 26(10): 1135–1145.
- [ 12 ] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry [ J ]. *Nature*, 2008, 456(7218): 53–59.
- [ 13 ] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning [ J ]. *Genome Res*, 2008, 18(7): 1051–1063.
- [ 14 ] Ardui S, Ameer A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics [ J ]. *Nucleic Acids Res*, 2018, 46(5): 2159–2168.
- [ 15 ] Rhoads A, Au KF. PacBio sequencing and its applications [ J ]. *Genom Proteom Bioinf*, 2015, 13(5): 278–289.
- [ 16 ] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome [ J ]. *Science*, 2009, 326(5950): 289–293.
- [ 17 ] Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements [ J ]. *Nat Rev Genet*, 2018, 19(11): 688–704.
- [ 18 ] Tempel S. Using and understanding RepeatMasker [ J ]. *Methods Mol Biol*, 2012, 859: 29–51.
- [ 19 ] Li X, Kahveci T, Settles AM. A novel genome-scale repeat finder geared towards transposons [ J ]. *Bioinformatics*, 2008, 24(4): 468–476.
- [ 20 ] Daron J, Glover N, Pingault L, Theil S, Jamilloux V, *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B [ J ]. *Genome Biol*, 2014, 15(12): 546.
- [ 21 ] Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR. An automated homology-based approach for identifying transposable elements [ J ]. *BMC Bioinformatics*, 2011, 12(1): 130.
- [ 22 ] Rho M, Choi JH, Kim S, Lynch M, Tang H. De novo identification of LTR retrotransposons in eukaryotic genomes [ J ]. *BMC Genomics*, 2007, 8: 90.
- [ 23 ] Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline [ J ]. *Genome Biol*, 2019, 20(1): 275.
- [ 24 ] Ewing AD, Kazazian HH Jr. High-throughput sequencing reveals extensive variation in humanspecific L1 content in individual human genomes [ J ]. *Genome Res*, 2010, 20(9): 1262–1270.
- [ 25 ] Agarwal P, States DJ. The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome [ J ]. *Proc Int Conf Intell Syst Mol Biol*, 1994, 2: 1–9.
- [ 26 ] Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes [ J ]. *Genome Res*, 2002, 12(8): 1269–1276.
- [ 27 ] Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes [ J ]. *Bioinformatics*, 2008, 24(10): 1717–1724.

- 2005, S1: i351–i358.
- [28] Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, *et al.* RepeatModeler2 for automated genomic discovery of transposable element families [J]. *Proc Natl Acad Sci USA*, 2020, 117(17): 9451–9457.
- [29] Schaeffer CE, Figueroa ND, Liu X, Karro JE. phRAIDER: Pattern-Hunter based rapid ab initio detection of elementary repeats[J]. *Bioinformatics*, 2016, 32(12): i209–i215.
- [30] Su W, Gu X, Peterson T. TIR-Learner, a New ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome[J]. *Mol Plant*, 2019, 12(3): 447–460.
- [31] Edgar RC, Myers EW. PILER: identification and classification of genomic repeats [J]. *Bioinformatics*, 2005, 21(S1): i152–i158.
- [32] Valencia JD, Girgis HZ. LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo[J]. *BMC Genomics*, 2019, 20(1): 450.
- [33] Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons[J]. *BMC Bioinformatics*, 2008, 9(1): 18.
- [34] Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Finegrained annotation and classification of de novo predicted LTR retrotransposons[J]. *Nucleic Acids Res*, 2009, 37(21): 7002–7013.
- [35] Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons [J]. *Nucleic Acids Res*, 2007, 35(S2): W265–W268.
- [36] Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale[J]. *Nucleic Acids Res*, 2001, 29(22): 4633–4642.
- [37] Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using Repeat Explorer2[J]. *Nat Protoc*, 2020, 15(11): 3745–3776.
- [38] Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*) [J]. *Genome Biol Evol*, 2015, 7(4): 1192–1205.
- [39] Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements[J]. *Nat Rev Genet*, 2018, 19(11): 688–704.
- [40] Guo R, Li YR, He S, Ou-Yang L, Sun Y, Zhu Z. Replong: de novo repeat identification using long read sequencing data[J]. *Bioinformatics*, 2018, 34(7): 1099–1107.
- [41] Kelleher ES, Barbash DA, Blumenstiel JP. Taming the turmoil within: new insights on the containment of transposable elements[J]. *Trends Genet*, 2020, 36(7): 474–489.
- [42] Caspi A, Pachter L. Identification of transposable elements using multiple alignments of related genomes[J]. *Genome Res*, 2006, 16(2): 260–270.
- [43] Rho M, Choi JH, Kim S, Lynch M, Tang H. De novo identification of LTR retrotransposons in eukaryotic genomes [J]. *BMC Genomics*, 2007, 8(1): 90.
- [44] Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits [J]. *Nat Rev Genet*, 2017, 18(2): 71–86.
- [45] 崔懿奎, 曹晓风. 高等植物转座元件功能研究进展[J]. *生物化学与生物物理进展*, 2015, 42(11): 1033–1046.
- Cui XK, Cao XF. Overview of the function of transposable elements in higher plants [J]. *Progress in Biochemistry and Biophysics*, 2015, 42(11): 1033–1046.
- [46] McCue AD, Nuthikattu S, Reeder SH, Slotkin RK. Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA [J]. *PLoS Genet*, 2012, 8(2): e1002474.
- [47] Mccue AD, Slotkin RK. Transposable element small RNAs as regulators of gene expression [J]. *Trends Genet*, 2012, 28: 616–623.
- [48] Yan Y, Zhang Y, Yang K, Sun Z, Fu Y, *et al.* Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice [J]. *Plant J*, 2011, 65(5): 820–828.
- [49] Kobayashi S, Gotoyamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color [J]. *Science*, 2004, 304(5673): 982.
- [50] Shure M, Wessler S, Fedoroff N. Molecular identification and isolation of the Waxy locus in maize [J]. *Cell*, 1983, 35(1): 225–233.
- [51] Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, *et al.* Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges [J]. *Plant Cell*, 2012, 24(3): 1242–1255.
- [52] Selinger DA, Chandler VL. *B-Bolivia*, an allele of the maize *b1* gene with variable expression, contains a high copy retrotransposon-related sequence immediately upstream [J]. *Plant Physiol*, 2001, 125(3): 1363–1379.
- [53] Bhattacharyya MK, Smith AM, Ellis TH, Hedley C, Martin C. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme [J]. *Cell*, 1990, 60(1): 115–122.
- [54] Wendel JF, Cronn RC, Johnston JS, Price HJ. Feast and famine in plant genomes [J]. *Genetica*, 2002, 115(1): 37–47.
- [55] Petrov DA. Mutational equilibrium model of genome size



- evolution[J]. *Theor Popul Biol*, 2002, 61(4): 531–544.
- [56] Gregory TR. Insertion-deletion biases and the evolution of genome size[J]. *Gene*, 2004, 324: 15–34.
- [57] Hawkins JS, Grover CE, Wendel JF. Repeated big bangs and the expanding universe: directionality in plant genome size evolution[J]. *Plant Sci*, 2008, 174(6): 557–562.
- [58] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, *et al*. The B73 maize genome: complexity, diversity, and dynamics[J]. *Science*, 2009, 326(5956): 1112–1115.
- [59] Lin J, Cai Y, Huang G, Yang Y, Li Y, *et al*. Analysis of the chromatin binding affinity of retrotransposases reveals novel roles in diploid and tetraploid cotton [J]. *J Integr Plant Biol*, 2019, 61(1): 32–44.
- [60] Huang G, Wu Z, Percy RG, Bai M, Li Y, *et al*. Genome sequence of *Gossypium herbaceum* and genome updates of *Gossypium arboreum* and *Gossypium hirsutum* provide insights into cotton A-genome evolution [J]. *Nat Genet*, 2020, 52(5): 516–524.
- [61] Li G, Wang L, Yang J, He H, Jin H, *et al*. A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes[J]. *Nat Genet*, 2021, 53(4): 574–584.
- [62] Plegu B, Guyot R, Picault N, Roulin A, Sanyal A, *et al*. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice [J]. *Genome Res*, 2006, 16(10): 1262–1269.
- [63] Jia J, Xie Y, Cheng J, Kong C, Wang M, *et al*. Homology-mediated inter-chromosomal interactions in hexaploid wheat lead to specific subgenome territories following polyploidization and introgression [J]. *Genome Biol*, 2021, 22(1): 26.
- [64] Thomas CA. The genetic organization of chromosomes [J]. *Annu Rev Genet*, 1971, 5: 237–256.
- [65] Freeling M, Thomas BC. Gene-balanced duplications, like Tetraploidy, provide predictable drive to increase morphological complexity[J]. *Genome research*, 2006, 16(7): 805–814.
- [66] Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids[J]. *Proc Natl Acad Sci USA*, 2014, 111(14): 5283–5288.
- [67] Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression [J]. *Genome Res*, 2009, 19(8): 1419–1428.
- [68] El Baidouri M, Kim KD, Abernathy B, Arikat S, Maumus F, *et al*. A new approach for annotation of transposable elements using small RNA mapping [J]. *Nucleic Acids Res*, 2015, 43(13): e84.

(责任编辑:周媛)