文章编号:1000－2367(2017)06－0072－07 DOI:10.16366/j.cnki.1000－2367.2017.06.012

陆地棉 LTR 反转录转座子的数量分布与功能分析

刘震[1],卢全伟[1],张国强[1],彭仁海[1,2]

(1.安阳工学院 生物与食品工程学院,河南 安阳 455000;2.中国农业科学院 棉花研究所,河南 安阳 455000)

摘 要:LTR(Long terminal repeat)反转录转座子是真核生物基因组中普遍存在的一类遗传因子,它们以 RNA 为媒介在基因组中不断自我复制.在高等植物中,LTR 反转录转座子是基因组的重要成分之一.本研究通过多种方法挖掘并注释了陆地棉基因组中的 LTR 反转录转座子,结果表明陆地棉基因组 LTR 反转录转座子的 Gypsy 超家族与基因的分布呈近似的反比关系,而 Copia 超家族在各染色体的起始端有较多的分布.通过皮尔森相关系数发现陆地棉 LTR 反转录转座子的拷贝数与染色体大小之间有强相关性.在 LTR 反转录转座子上游和下游分布的基因具有类似的富集特征,其分子功能主要集中在结合和催化活性等方面.本研究结果加深了对陆地棉 LTR 反转录转座子的认识,为深入研究棉花基因组提供了重要数据支撑.

关键词:陆地棉;LTR 反转录转座子;分布;功能

中图分类号:Q341 **文献标志码**:A

转座子是基因组中一类可以移动的 DNA 片段,转座子首先在玉米中发现[1].通过分析越来越多物种的基因组序列数据,人们逐渐发现转座子在基因组中广泛分布,而且随着研究的深入,发现转座子对其宿主有着极为重要的作用,具体表现在基因组扩增、新基因形成、基因破坏以及基因表达活性等多个方面[2-5].转座子的发现打破了遗传物质在染色体上呈线性固定排列的传统观念,对遗传学和分子生物学的发展具有深远意义.

转座子依据结构和转座特征分为两大类:Ⅰ型转座子,又称反转录转座子,以 RNA 为中间媒介进行转座,属于“复制－粘贴”型,每转座一次增加一个拷贝;Ⅱ型转座子,也称 DNA 转座子,以 DNA 为中间体进行转座,属“剪切－粘贴”型.反转录转座子又可以进一步分为五大类,其中长末端重复(Long terminal repeat,LTR)类型在植物基因组中的数量最多,是植物基因组中最主要的转座子[6].

植物基因组中的大部分反转录转座子不具有转座功能,但也有小部分具有转座活性,它们可能插入到功能基因内或周边,从而影响这些基因的表达[7].研究表明,红皮葡萄的色泽是由果皮中花青苷的积累形成的,花青苷的合成受到 *Vvmyb* A1 基因的调控,一个反转录转座子插入到 *Vvmyb* A1 基因的启动子区域,导致花青苷不能合成,从而形成白皮葡萄[8].拟南芥控制成花的 *LEAFY* 基因中插入一个反转录转座子使该基因不能正常表达,可导致个体无雄性花器官[9].DNA 甲基化酶可以抑制反转录转座子的迁移和插入.另外,在逆境、激素处理、机械伤等情况下,反转录转座子的转座功能可能被激活[10-11].

棉花是世界上最重要的经济作物之一,提供了大量的天然纺织纤维.据细胞学研究,二倍体棉种分为 8 个基因组(即 A～G 和 K),四倍体棉种为异源四倍体,由 A 和 D 两个二倍体组成[12].二倍体亚洲棉(*Gossypium arboretum*,A 组)[13]和雷蒙德氏棉(*Gossypium raimondii*,D 组)[14],四倍体陆地棉(*Gossypium hirsutum*,AD 组)[15]和海岛棉(*Gossypium barabdense*,AD 组)[16]的基因组已经公布,为棉属转座子的鉴定和分析提供了绝佳的机会.陆地棉由于产量高和纤维品质较为优良,成为目前世界范围内种植面积最大的棉

收稿日期:2017-07-10;**修回日期**:2017-10-10.

基金项目:国家自然科学基金(31471548);棉花生物学国家重点实验室开放课题(CB2015A21);河南省高等学校重点科研项目(18A180010);安阳工学院校科研基金(YJJ2017007).

作者简介:刘震(1982－),男,河南安阳人,安阳工学院讲师,博士,研究方向:生物信息学.

通信作者:彭仁海,教授,博士,E-mail:aydxprh@163.com.

花栽培种.研究表明，陆地棉基因组的重复序列约占 67.2%，而且 D 组来源的转座子要比 A 组来源的转座子更活跃[15].本研究基于生物信息学方法从全基因组层面对陆地棉的 LTR 反转录转座子进行挖掘和分类，并分析它们的数量、分布特征、周边基因的功能富集及其与染色体大小的关系.

1　材料与方法

1.1　LTR 反转录转座子的挖掘与分类

陆地棉的基因组序列、基因注释和 GO 注释文件从 COTTONGEN(https://www.cottongen.org/)下载.首先使用不同算法的软件挖掘陆地棉基因组中的 LTR 反转录转座子，这些软件包括：LTR_STRUC[17]，LTRharvest[18]，PILER[19] 和 RepeatModeler[20-22]；然后使用 REPCLASS[23] 软件将上述结果序列归类到相应的超家族，将各软件的结果序列按超家族的不同分别合并，利用 CD-HIT[24] 软件去除每个超家族中的冗余序列，从而得到陆地棉特异的 LTR 反转录转座子序列库；再将该库与公共重复序列数据库 Repbase[25] 进行比较，删除该库与 Repbase 库中任一序列相似性大于 80%的序列，之后，将该库与 Repbase 库合并成一个库，命名为 Gh_LTR.依据 Gh_LTR 库，通过 RepeatMasker[20-22] 注释陆地棉基因组中的 LTR 反转录转座子.

1.2　LTR 反转录转座子的数量与分布

通过 Perl 脚本从 RepeatMasker 结果文件中收集 LTR 反转录转座子在陆地棉基因组的数量和位置，并利用 gff 注释文件的数据收集基因的数量和位置.统计分析染色体中每 10 kb 范围内的 LTR 反转录转座子与基因的数量，并通过 Circos 绘制分布图.LTR 反转录转座子拷贝数与染色体大小之间的皮尔森相关系数由 R 语言计算获得.

1.3　LTR 反转录转座子周边基因的 GO 富集分析

查找 LTR 反转录转座子内部及其上、下游 10 kb 范围内的基因，利用基因组 GO 注释文件确定这些基因的 GO 注释条目，并使用 WEGO[26](http://wego.genomics.org.cn/)进行富集分析.

2　结　果

2.1　陆地棉 LTR 反转录转座子的数量与分布特征

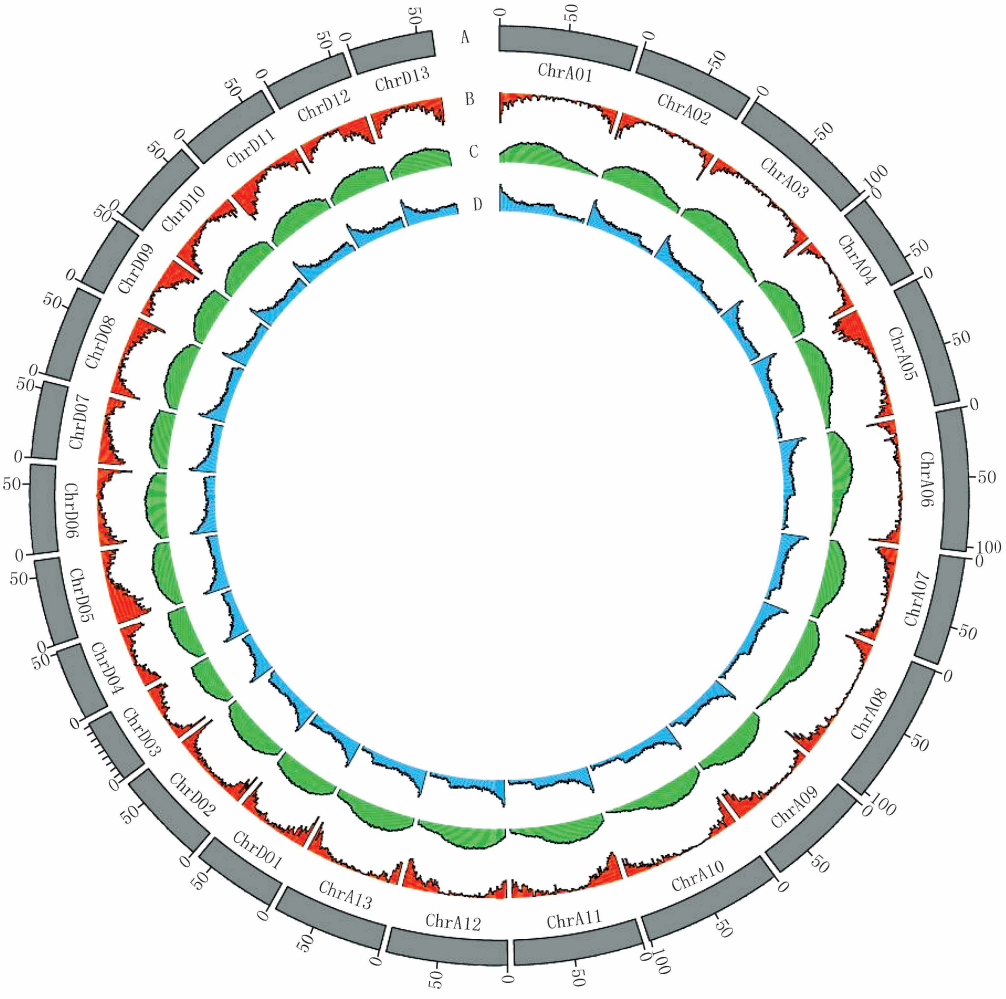
陆地棉基因组的 LTR 反转录转座子超家族主要分为 Copia 和 Gypsy 两类，其中，Gypsy 超家族共有 485 174 个拷贝，Copia 超家族共有 144 888 个拷贝，其它家族共 405 个拷贝，未能成功分类的共 102 299 个拷贝.

Gypsy 超家族和 Copia 超家族在陆地棉染色体上的分布明显不同，而同一超家族在不同染色体上具有类似的分布特征(图 1).Gypsy 超家族在染色体中部附近有较多的分布，其中，在 A 组染色体的分布波峰略偏向于染色体的起始端.整体来说，Gypsy 超家族与基因的分布成反比关系.

Copia 超家族在陆地棉 A 组和 D 组染色体的共同分布特征是在染色体起始端有一个明显的波峰，Copia 超家族与基因的分布没有明显的关系.然而，Copia 超家族在陆地棉 A 组染色体的中后部有一个明显的下降点(4 号染色体除外)，D 组染色体则没有该特征.

2.2　染色体大小与 LTR 反转录转座子的关系

已有研究表明，重复序列与物种基因组的大小有直接关系[27-29].本研究中，我们进一步调查了陆地棉不同染色体的大小与 LTR 反转录转座子的关系.陆地棉基因组包含 26 条染色体，平均每 Mb 染色体含有 381.330 3个转座子(表 1).从图 2 中可以看出，陆地棉染色体大小与 LTR 反转录转座子的拷贝数有明显的线性关系.经计算，陆地棉染色体大小与 LTR 反转录转座子拷贝数的皮尔森相关系数为 0.982 347 5，与 LTR 反转录转座子长度皮尔森相关系数为 0.978 702，均为极强相关.这也就表明，LTR 反转录转座子不仅对基因组大小有重要影响，在同一物种内，与不同染色体的大小也有明显的关联.

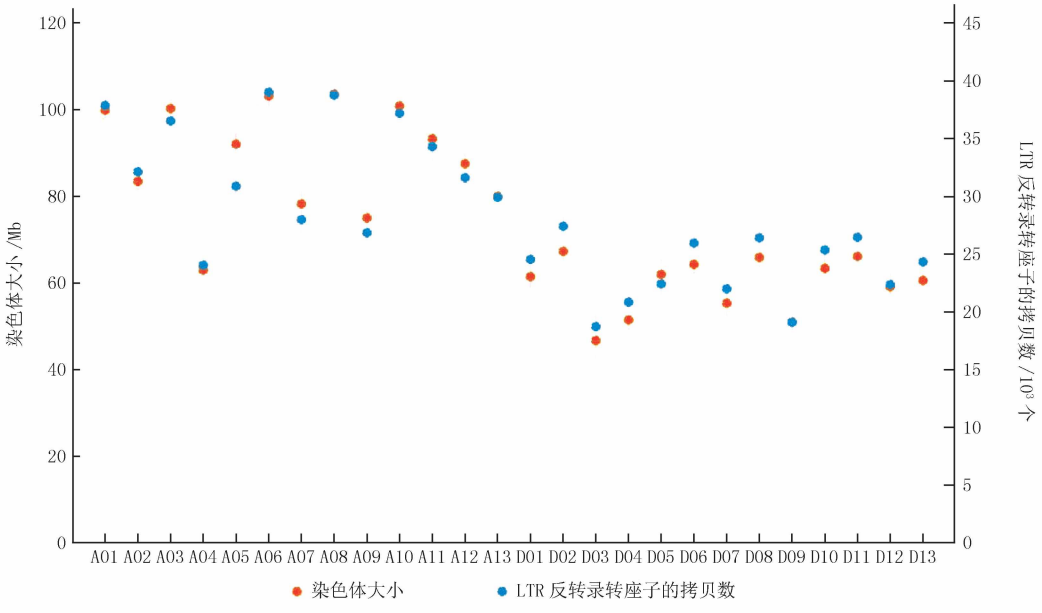


A: 染色体 ;B: 基因 ;C: Gypsy 超家族 ;D: Copia 超家族

图 1 LTR 反转录转座子在陆地棉染色体上的分布

表 1 染色体大小与 LTR 反转录转座子的关系(LTR 反转录转座子与染色体的长度单位是 Mb)

	A 组				D 组			
染色体	拷贝数/个	LTR 转座子长度	染色体长度	拷贝数/Mb 染色体	拷贝数/个	LTR 转座子长度	染色体长度	拷贝数/Mb 染色体
1	37 856	42.69	99.88	379.00	24 527	22.95	61.46	399.10
2	32 107	36.68	83.45	384.76	27 409	25.69	67.28	407.36
3	36 523	41.01	100.26	364.27	18 705	17.18	46.69	400.62
4	24 032	26.38	62.91	381.98	20 832	19.47	51.45	404.87
5	30 868	32.43	92.05	335.35	22 409	19.39	61.93	361.83
6	39 009	44.24	103.17	378.10	25 948	24.53	64.29	403.58
7	27 974	30.27	78.25	357.49	21 985	19.00	55.31	397.47
8	38 747	43.56	103.63	373.91	26 407	24.30	65.89	400.75
9	26 844	28.55	75.00	357.92	19 080	16.99	51.00	374.15
10	37 193	40.93	100.87	368.73	25 352	22.71	63.37	400.03
11	34 309	35.79	93.32	367.66	26 456	23.09	66.09	400.32
12	31 612	35.15	87.48	361.34	22 360	20.10	59.11	378.28
13	29 899	32.52	79.96	373.92	24 323	22.39	60.53	401.81



橘黄色为染色体大小，对应左侧坐标轴；蓝色为相应染色体上 LTR 反转录转座子的拷贝数，对应右侧坐标轴 .

图 2　染色体大小与 LTR 反转录转座子拷贝数的关系

2.3　LTR 反转录转座子周边基因的功能富集分析

Gypsy 超家族的上游、下游和内部分别找到 48 123、47 930 和 3 078 个基因，Copia 超家族的上游、下游和内部分别找到 38 774、38 264 和 2 762 个基因.可以看出，在陆地棉 LTR 反转录转座子上、下游分布的基因数量相当.利用这些基因的 GO 注释分别从细胞组成、分子功能和生物学过程 3 个方面分析 LTR 反转录转座子周边基因的功能富集情况.结果表明，Gypsy 超家族与 Copia 超家族周边基因具有类似的富集特征(图 3).细胞组成中注释较多的是细胞(cell)、细胞组分(cell part)、细胞器(organelle)和大分子复合物(macromolecular complex)；分子功能的注释主要集中在结合(binding)和催化活性(catalytic)；生物学过程的注释则主要集中在细胞过程(cellular process)、代谢过程(metabolic process)、生物调节(biological regulation)、建立定位(establishment of localization)、定位(localization)和色素(pigmentation).

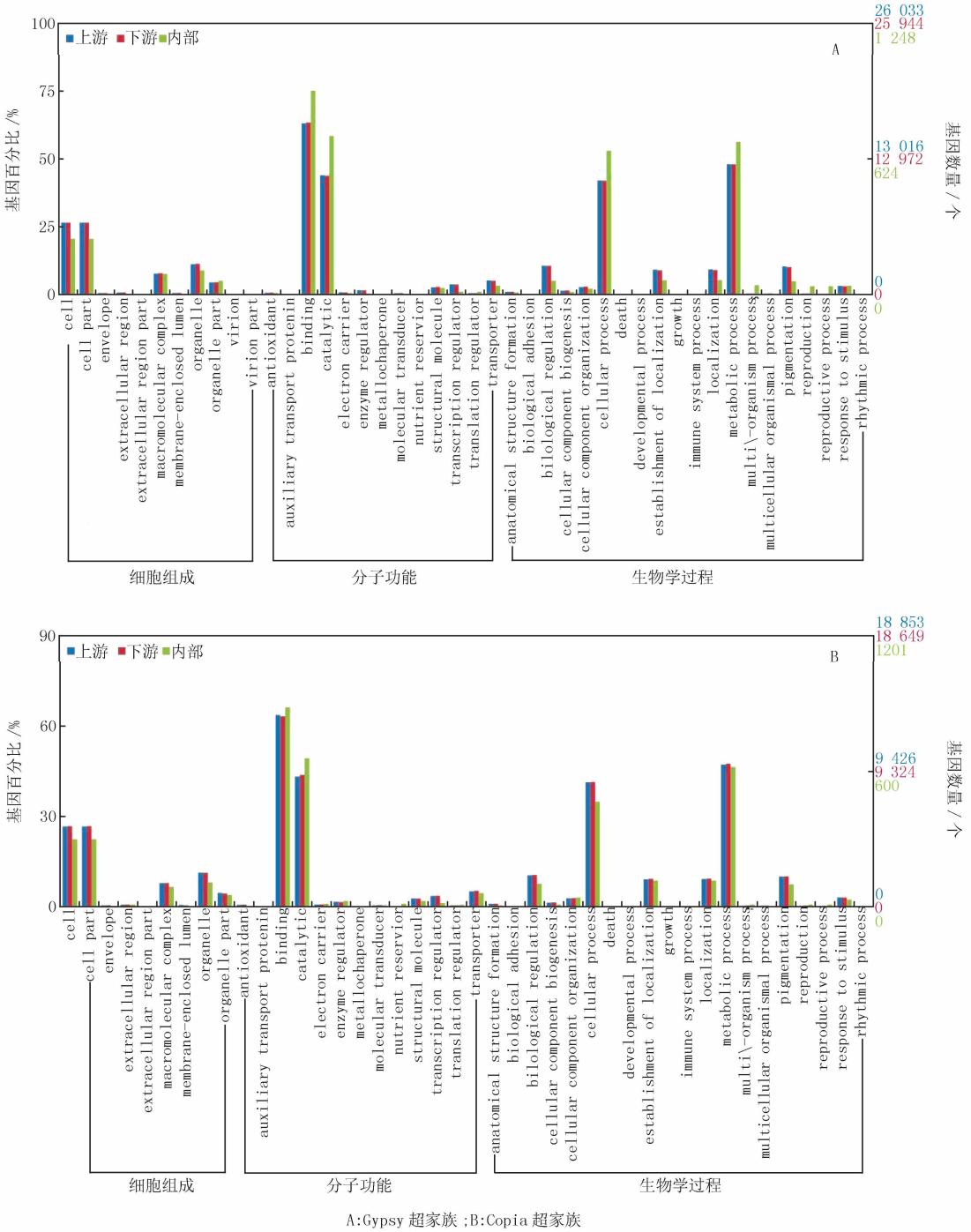
在两类超家族周边基因功能富集的结合和催化活性方面，LTR 反转录转座子序列内部基因的百分比大于上、下游基因.在细胞组成部分富集的细胞、细胞组分和细胞器方面，LTR 反转录转座子序列内部基因的百分比则小于上、下游基因.在生物学过程部分富集的细胞过程方面，Gypsy 超家族序列内部基因的百分比高于上、下游基因，而 Copia 超家族序列内部基因的百分比则低于上、下游基因(图 3).

3　讨　论

反转录转座子是高等植物基因组的重要组分之一.在不同物种的基因组中，转座子与基因的分布关系有很大差别.在玉米和高粱中，转座子偏向于分布在基因附近，与基因的分布成正相关[26]；而在拟南芥中，转座子与基因的分布则为负相关[30].陆地棉基因组测序的完成为从全基因组层面对其转座子的研究提供了可能.本研究结果表明，在陆地棉基因组中，Gypsy 超家族与基因的分布为负相关，而 Copia 超家族与基因的分布没有明显的关系.此外，研究结果也表明 Gypsy 超家族与 Copia 超家族在陆地棉不同染色体上具有稳定的分布特征，因此，我们推测 LTR 反转录转座子在陆地棉基因组上的分布不是随机的，而这种非随机性也从侧面体现了 LTR 反转录转座子的重要性.

已有研究表明，转座子积累是导致物种基因组大小差异的重要原因[27－29]，本研究进一步分析了陆地棉不同染色体的大小与 LTR 反转录转座子的关系，LTR 反转录转座子拷贝数与染色体大小的皮尔森相关系数比 LTR 反转录转座子总长度与染色体大小的皮尔森相关系数略大，这可能与 LTR 反转录转座子在进化过程中的缺失方式有关.LTR 反转录转座子在基因组中以“复制－粘贴”的形式进行扩增，这种扩增会受到

甲基化等方式的制约,形成扩增与缺失同时存在的动态特征.转座子的缺失,并非缺失完整的序列,而是缺失部分结构[31],这种方式就可能造成拷贝数与染色体的大小更相关.



A:Gypsy 超家族 ;B:Copia 超家族

图 3 陆地棉 LTR 反转录转座子周边基因的功能富集分析

在陆地棉 Gypsy 超家族与 Copia 超家族上游、下游和内部分布的基因的 GO 富集情况非常类似,这就表明特定基因周边的 LTR 反转录转座子没有家族特异性,特定基因的周边只是倾向于分布 LTR 反转录转座子.此外,GO 富集分析结果也表明,在两类超家族上游和下游分布的基因数量相当,而在其内部分布的基因约比上下游小一个数量级.

众多植物全基因组测序的完成,将有利于获得更多种类的转座子家族以及全长序列,从而有助于进一步了解它们在植物基因组结构和进化中的作用,拓展人们对转座子的认识.

参 考 文 献

[1] Shepherd N S, Schwarz-Sommer Z, Blumberg V S J, et al. Similarity of the Cin1 repetitive family of Zea mays to eukaryotic transposable elements [J]. Nature, 1984, 307(5947): 185-187.

[2] Negi P, Rai A N, Suprasanna P. Moving through the stressed genome: Emerging regulatory roles for transposons in plant stress response [J]. Front Plant Sci, 2016, 7(6): 1448.

[3] Chuong E B, Elde N C, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits [J]. Nat Rev Genet, 2017, 18(2): 71-86.

[4] Bennetzen J L. Transposable element contributions to plant gene and genome evolution [J]. Plant Mol Biol, 2000, 42(1): 251-269.

[5] Bennetzen J L. Transposable elements, gene creation and genome rearrangement in flowering plants [J]. Curr Opin Genet Dev, 2005, 15(6): 621-627.

[6] Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements [J]. Nat Rev Genet, 2007, 8(12): 973-982.

[7] 蒋爽，滕元文，宗宇，等. 植物 LTR 反转录转座子的研究进展 [J]. 西北植物学报，2013(11)：2354-2360.

[8] Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color [J]. Science, 2004, 304(5673): 982.

[9] Mirouze M, Reinders J, Bucher E, et al. Selective epigenetic control of retrotransposition in Arabidopsis [J]. Nature, 2009, 461(7262): 427-430.

[10] Tapia G, Verdugo I, Yanez M, et al. Involvement of ethylene in stress-induced expression of the TLC1.1 retrotransposon from Lycopersicon chilense Dun [J]. Plant Physiol, 2005, 138(4): 2075-2086.

[11] De Felice B, Wilson R R, Argenziano C, et al. A transcriptionally active copia-like retroelement in Citrus limon [J]. Cell Mol Biol Lett, 2009, 14(2): 289-304.

[12] 王坤波，刘旭. 棉属多倍化研究进展 [J]. 中国农业科技导报，2013(02)：20-27.

[13] Li F, Fan G, Wang K, et al. Genome sequence of the cultivated cotton Gossypium arboreum [J]. Nat Genet, 2014, 46(6): 567-572.

[14] Paterson A H, Wendel J F, Gundlach H, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres [J]. Nature, 2012, 492(7429): 423-427.

[15] Zhang T, Hu Y, Jiang W, et al. Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement [J]. Nat Biotechnol, 2015, 33(5): 531-537.

[16] Liu X, Zhao B, Zheng H J, et al. Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites [J]. Sci Rep, 2015, 5: 14139.

[17] Mccarthy E M, Mcdonald J F. LTR_STRUC: a novel search and identification program for LTR retrotransposons [J]. Bioinformatics, 2003, 19(3): 362-367.

[18] 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.

[19] Edgar R C, Myers E W. PILER: identification and classification of genomic repeats [J]. Bioinformatics, 2005, 21 Suppl 1: i152-i158.

[20] Huda A, Jordan I K. Analysis of transposable element sequences using CENSOR and RepeatMasker [J]. Methods Mol Biol, 2009, 537: 323-336.

[21] Tempel S. Using and understanding RepeatMasker [J]. Methods Mol Biol, 2012, 859: 29-51.

[22] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences [J]. Curr Protoc Bioinformatics, 2009, Chapter 4: 4-10.

[23] Feschotte C, Keswani U, Ranganathan N, et al. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes [J]. Genome Biol Evol, 2009, 1: 205-220.

[24] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data [J]. Bioinformatics, 2012, 28(23): 3150-3152.

[25] Bao W, Kojima K K, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes [J]. Mob DNA, 2015, 6: 11.

[26] Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations [J]. Nucleic Acids Res, 2006, 34(Web Server issue): W293-W297.

[27] Tenaillon M I, Hufford M B, Gaut B S, et al. Genome size and transposable element content as determined by high-throughput sequencing in maize and Zea luxurians [J]. Genome Biol Evol, 2011, 3: 219-229.

[28] 陈建军，王瑛. 植物基因组大小进化的研究进展 [J]. 遗传，2009, 31(5)：464-470.

[29] Tenaillon M I, Hollister J D, Gaut B S. A triptych of the evolution of plant transposable elements [J]. Trends Plant Sci, 2010, 15(8): 471-478.

[30] Wright S I, Agrawal N, Bureau T E. Effects of recombination rate and gene density on transposable element distributions in Arabidopsis

thaliana [J].Genome Res,2003,13(8):1897-1903.

[31] Pereira V.Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome [J].Genome Biol,2004,5(10):R79.

Quantitative Distribution and Functional Analysis of LTR Retrotransposons in *Gossypium hirsutum*

Liu Zhen[1],Lu Quanwei[1],Zhang Guoqiang[1],Peng Renhai[1,2]

(1.School of Biotechnology and Food Science,Anyang Institute of Technology,Anyang 455000,China;
2. State Key Laboratory of Cotton Biology,Institute of Cotton Research of Chinese Academy of Agricultural Science, Anyang 455000,China)

Abstract:LTR (Long terminal repeat) retrotransposons in eukaryotes are mobile genetic elements with ubiquitous distribution, these elements can amplify themselves via RNA intermediates and increase their copy numbers in genome. In high plants, LTR retrotransposon is an important component of genome. In this study, LTR retrotransposons of *Gossypium hirsutum* were excavated and annotated by a variety of methods, the results showed that the Gypsy superfamily of LTR retrotransposons in *Gossypium hirsutum* was approximately inversely related to the distribution of genes, while Copia superfamily distributes more at the starting terminal of each chromosome. Pearson correlation coefficients showed that there was a strong correlation between the LTR retrotransposons copy number and chromosome size in *Gossypium hirsutum*. Furthermore, the genes enrichment in either upstream or downstream of the LTR retrotransposons exhibit similar characteristics, and the molecular functions mainly focus on binding and catalytic activity etc. Our research will lead to a better understanding of LTR retrotransposons in *Gossypium hirsutum* genome and provide important data support for further study of cotton genome.

Keywords:*Gossypium hirsutum*; LTR Retrotransposons; distribution; function

[责任编辑 王凤产]