遗传 HEREDITAS (Beijing) 2013 年 10 月, 35(10): 1226－1236
ISSN 0253-9772 www.chinagene.cn

研究报告

DOI: 10.3724/SP.J.1005.2013.01226

雷蒙德氏棉和拟南芥基因启动子中顺式作用元件的分布

孙高飞[1,2], 何守朴[1], 杜雄明[1]

1. 中国农业科学院棉花研究所, 棉花生物学国家重点实验室, 安阳 455000;
2. 安阳工学院计算机科学与信息工程学院, 安阳 455000

摘要: 随着雷蒙德氏棉(*Gossypium raimondii*)基因组草图的完成, 相关的基因组学研究已经全面展开。文章利用已公布的雷蒙德氏棉和拟南芥基因组序列, 结合顺式作用元件(*cis*-regulatory element, CRE)数据库 PLACE 中的 CRE 序列信息, 对两个物种中带有 5′UTR 注释的基因启动子上游 1 000 bp 序列进行 CRE 扫描和统计。结果表明, 雷蒙德氏棉和拟南芥基因组中分别有 44(12.3%)和 57(15.5%)个 CRE 在启动子的特定位置呈峰状分布, 其中在两个基因组均呈峰状分布的有 34 个, 这些 CRE 又可以根据核心序列分为 4 大类。TATABOX 类 CRE 顶峰在启动子中出现的位置和其真实位置(~ –30 bp)具有一致性, 预示 CRE 真实位置在不同基因启动子中相对保守, 从而推测本研究中呈峰状分布 CRE 的顶峰位置可能就是转录因子和该 CRE 结合的真实位置。而同一 CRE 在两个基因组中存在的位置差异则主要源于雷蒙德氏棉基因的 5′UTR 长度变异大于拟南芥。另外, 文章还发现绝大多数峰状分布的 CRE 的位置都集中在–110 bp~0 bp 之间, 这种集中的分布可能更有利于转录因子之间相互作用, 从而调控下游基因的表达。

关键词: 雷蒙德氏棉; 全基因组; 顺式作用元件

Analysis of *cis*-regulatory element distribution in gene promoters of *Gossypium raimondii* and *Arabidopsis thaliana*

SUN Gao-Fei[1,2], HE Shou-Pu[1], DU Xiong-Ming[1]

1. *State Key Laboratory of Cotton Biology*, *Institute of Cotton Research*, *Chinese Academy of Agricultural Sciences*, *Anyang* 455000, *China*;
2. *Department of Computer Science and Information Engineering*, *Anyang Institute of Technology*, *Anyang* 455000, *China*

Abstract: Cotton genomic studies have boomed since the release of *Gossypium raimondii* draft genome. In this study, *cis*-regulatory element (CRE) in 1 kb length sequence upstream 5′ UTR of annotated genes were selected and scanned in the *Arabidopsis thaliana* (*At*) and *Gossypium raimondii* (*Gr*) genomes, based on the database of PLACE (Plant *cis*-acting Regulatory DNA Elements). According to the definition of this study, 44 (12.3%) and 57 (15.5%) CREs presented

收稿日期: 2013–04–07; 修回日期: 2013–07–22
基金项目: 国家科技支撑计划项目(编号：2013BAD01B03)资助
作者简介: 孙高飞, 硕士, 副教授, 研究方向：棉花生物信息学。E-mail: sungaofei@sina.com
何守朴, 硕士, 助理研究员, 研究方向：棉花种质资源学。E-mail: zephyr0911@126.com
孙高飞和何守朴同为第一作者。
通讯作者: 杜雄明, 博士, 研究员, 研究方向：棉花种质资源学。E-mail: dujeffrey8848@hotmail.com
网络出版时间: 2013-8-6 18:48:36
URL: http://www.cnki.net/kcms/detail/11.1913.R.20130806.1848.002.html

"peak-like" distribution in the 1 kb selected sequences of both genomes, respectively. Thirty-four of them were peak-like distributed in both genomes, which could be further categorized into 4 types based on their core sequences. The coincidence of TATABOX peak position and their actual position (~ –30 bp) indicated that the position of a common CRE was conservative in different genes, which suggested that the peak position of these CREs was their possible actual position of transcription factors. The position of a common CRE was also different between the two genomes due to stronger length variation of 5′ UTR in *Gr* than *At*. Furthermore, most of the peak-like CREs were located in the region of –110 bp~0 bp, which suggested that concentrated distribution might be conductive to the interaction of transcription factors , and then regulate the gene expression in downstream.

Keywords: *Gossypium raimondii*; genome-wide; *cis*-regulatory element (CRE)

棉花是我国最重要的经济作物之一, 在国民经济中占有重要地位。我国主要的栽培棉种是四倍体的陆地棉(*G. hirsutum*, AD_1)和海岛棉(*G. barbadense*, AD_2), 南方少数民族地区有极零星的二倍体亚洲棉(*G. arboretum*, A_2)种植。和其他棉种相比, 陆地棉具有显著的产量优势, 然而纤维品质和抗逆性却存在较大缺陷。目前, 传统的育种理论和方法对陆地棉的产量、纤维品质和抗逆性改良收效甚微, 因此迫切需要通过分子生物学手段突破育种瓶颈。随着测序技术的高速发展, 越来越多作物的全基因组序列被相继破解, 为作物的分子改良和分子育种提供了绝佳的条件和平台。由于四倍体陆地棉基因组较大, 结构复杂, 遗传图谱质量不佳。2012 年棉属中基因组较小的二倍体雷蒙德氏棉(*G.raimondii*, D_5)全基因组物理草图首先绘制完成, 共有 40 000 多个基因获得注释, 其中超过 90%获得了转录验证[1], 这标志着对棉花的分子改良真正开始迈入功能基因组学时代, 从全基因组水平上开展棉花基因表达调控机理研究成为未来棉花基因组学研究的重要研究方向。

转录因子(Transcription factor, TF)作为调控基因表达的关键因子, 通常和基因启动子内的顺式作用元件(*cis*-regulatory element, CRE)结合, 实现对下游基因的转录调控。这些转录因子结合位点(Transcription factor binding site, TFBS)的长度一般在 5~20 bp 左右, 和转录因子的结合在不同基因中也具有相对的保守性。顺式作用元件按照功能可以分为启动子元件、增强子元件及沉默子元件。启动子元件和 RNA 聚合酶结合, 精确控制基因的转录和转录效率; 增强子元件和转录因子结合则能够增强启动子的转录活性; 沉默子元件则和增强子元件相反, 和转录因子结合后阻遏基因的转录。因此, 可以通过分析基因启动子中 CRE 特征序列, 预测可能调控相关基因的转录因子, 特别对基因组水平上理解基因的转录调控机制, 建立基因调控网络具有重要意义。

目前已经有多种方法对已测序基因组进行全基因组CRE的扫描分析[2,3], 同时收集建立了各类在线数据库[4], 如PLACE、TRANSFAC和PlantCARE等[5], 结合这些方法和数据, 已经有大量的围绕CRE和基因调控关系的研究。Molina等[6]应用Gibbs抽样法对拟南芥(*Arabidopsis thaliana*)基因组进行全基因组CRE扫描分析, 发现包含TATA元件的启动子所占比例远少于预料的比例。Ding等[7]对拟南芥和白杨(*Populus trichocarpa*)的基因组进行了CRE扫描和对比, 解析出 796 在两个物种中共有的CRE组合。Civán等[8]利用软件MotifScanner对水稻基因组中的TATABOX和Y Patch两类CRE进行扫描, 分析了这两类CRE的分布规律。对拟南芥中胁迫响应的顺式调控元件进行扫描和分析, 发现生物胁迫和非生物胁迫主要通过两种特异的pCRE(putative CRE)家族来调控[9]。对水稻生殖细胞中特异或高表达的基因上游启动子序列进行CRE扫描和分析, 发现了一些新的基序, 可能是转录因子调控生殖细胞基因表达的结合位点[10]。张梅等[11]综述了DREBs 类转录因子能够通过与含有DRE/CRT顺式作用元件的抗逆相关基因启动子区相互作用, 进而调控一系列抗逆基因的表达。侯琳等[12]重点评述了描述TFBS的模型以及预测TFBS的多种软件以及TFBS生物信息学研究的发展。

本研究通过获取雷蒙德氏棉和拟南芥基因组全部有 5′UTR 注释的基因上游 1 000 bp 序列，利用 PLACE 数据库的 CRE 信息，对两个物种基因组的这些序列进行了 CRE 扫描，通过比较两个基因组中 CRE 的类型和特征，对部分分布特征明显的 CRE 进行分析。本研究证明了通过计算机预测基因上游序列中 CRE 位置的可行性，为进一步实验验证提供证据，同时对深入揭示基因调控机理具有参考意义。

1 数据来源与研究方法

1.1 数据来源

本研究中使用的拟南芥基因组序列和注释信息来自 http://www.arabidopsis.org(TAIR10)，棉花基因组序列和基因注释信息来自 ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/plants/Gossypium_raimondii/[13]，CRE 序列和相关注释来自 PLACE (http://www.dna.affrc.go.jp/PLACE/index.html)[14]。PLACE 数据库是一个植物顺式调控元件基序数据库，目前包含 469 个顺式调控元件基序。所有的序列信息来自于已经发表的研究论文。该数据库提供了扫描页面，能够为提供的序列进行顺式调控元件扫描。

1.2 方法

1.2.1 基因组基因信息分析和整合

将拟南芥和雷蒙德氏棉基因组序列和注释信息进行整理，从 gff 注释文件中提取 5′UTR 和 CDS 的起止位置信息，储存于 Microsoft SQL Server 2005 数据库。根据 gff 注释文件对于基因的位置注释和基因不同部分的位置注释可以看出，对于有 5′UTR 的基因，其转录起始位点(TSS)和 5′UTR 的起点相同，对于没有 5′UTR 的基因，其转录起始位点和第一个 CDS 的起始位点相同。

由于 5′UTR 在基因中的起始位置，因此对已有 5′UTR 注释的基因，5′UTR 的 5′端第一个碱基为 0 位置，取其上游 1 000 bp 序列作为启动子序列(图 1)。

5′UTR 的注释信息和转录因子起始位点是紧密相关的，而且直接影响到上游启动子序列的确定，继而影响 CRE 相对于起始位点的位置。对于没有 5′UTR 注释的基因，我们不能确定该基因有无 5′UTR。因此，本文只对有 5′UTR 注释的基因 CRE 分布进行统计和分析。

图 1 本研究所选取的启动子上游 1 kb 序列位置示意图

1.2.2 提取启动子序列

利用 PERL 语言编写脚本，根据基因组序列和数据库中整理的基因在染色体上的位置信息，以基因的起始位点为基点，提取拟南芥和雷蒙德氏棉基因上游 1 000 bp 序列供扫描使用。

1.2.3 CRE 扫描

根据 PLACE 数据库中提供的 CRE 的序列，使用 PERL 语言编写脚本，通过正则表达式对拟南芥和雷蒙德氏棉全部注释基因上游 1 000 bp 序列进行扫描，转录起始位点上游的第一个碱基的位置记为 −1，将 CRE 序列第一个碱基在上游序列中的位置记为该 CRE 的位置，扫描结果导入数据库。将上游序列从−1 000 到−1 每 10 bp 划分为一个区间(Section)，如果一个 CRE 的位置落入某个区间(即序列的第一个碱基位于该区间)，则认为该区间包含该 CRE。通过 SQL 语句将 CRE 在不同区间的数据进行统计，获得每个 CRE 在不同区间的分布数量。比较相同 CRE 在拟南芥和雷蒙德氏棉基因启动子中的数量分布情况。

1.2.4 CRE 峰定义

通过对 CRE 扫描数据的分析，我们发现某些 CRE 会在特定的启动子区间内聚集，形成一个峰状的分布。为了统一标准，本研究这样来定义峰状分布：当 CRE 在连续的 5 个 section 中的均值超过剩余所有 section 的均值，达到一定的比例时，我们称该 CRE 呈峰状分布。具体定义如下：

定义 1：CI 代表 CRE 在区间 I 中出现的次数。

定义 2：$\mathrm{R5I}=\sum_{i=J}^{J+4} CI$；其中 $J=-100\sim-5$; R5I 代表连续的 5 个区间的 CI 的和。

定义 3：$\mathrm{R95I}=\sum_{i=-100}^{J-1} CI+\sum_{i=J+1}^{-1} CI$；其中 $J=-100\sim-5$; R95I 代表去除以上 5 个连续区间后剩余其他区间内的 CRE 出现的次数。

定义 4：NZ95 代表其他 95 个区间中 CRE 出现次数不为零的区间的个数。

定义 5：$\mathrm{M5I}=\max_{I}\left[\left(\frac{\mathrm{R5I}}{5}\Big/\left(\frac{\mathrm{R95I}}{\mathrm{NZ95}}\right)\right)\right]$。(I= −100~−5)M5I 代表连续的 5 个区间的 CRE 数值在启动子的 100 个 section 中形成最高点。

对于同样比例的 CRE 分布，总量越大，其峰值分布的所反映的趋势就越强，因此，需要将总的 CRE 数量引入公式，从而使 CRE 数量较多的峰值分布具有更高的分值。

定义 6：$P=\mathrm{M5I}\times\log_{10}\left(\sum_{-100}^{-1}\mathrm{CI}\right)$。$P$ 是评价 CRE 峰状分布强度的评价指数，P 值越大，表示其峰的突起越明显。

2　结果与分析

2.1　基因数量的对比

首先对雷蒙德氏棉和拟南芥基因组中有 5′UTR 和无 5′UTR 的基因数量进行对比(表 1)。结果表明雷蒙德氏棉中有无 5′UTR 的基因数量比例和拟南芥基本一致，基因组拼接注释完整度较高。

2.2　启动子中不同位点碱基含量

提取雷蒙德氏棉和拟南芥基因组中有 5′UTR 的基因上游启动子 1 000 bp 长度序列，对 A/C/T/G 4 种碱基的分布进行统计比较(图 2)。

从图 2 可以看出，所有基因启动子中 A/T 含量要明显高于 C/G 含量，这个规律与整个基因组中碱基分布规律基本一致。另外，由于基因总数更多，雷蒙德氏棉基因启动子中 A/T 出现的频率要高于拟南芥，但 C/G 含量基本一致，说明雷蒙德氏棉基因启动子中 A/T 含量更高。

进一步比较两个图发现，拟南芥和雷蒙德氏棉基因启动子在大约–350 bp 之前，分布规律稳定一致，–350 bp 以后，T 出现频率开始下降，C 开始上升。在–30~–25 bp 左右位置，拟南芥中 T/A 出现频率突然出现一个尖锐的高峰(图 2A)，在雷蒙德氏棉中就没有这么明显(图 2B)，而是在更上游一点出现一个较缓和的小峰。

表 1　两个基因组有无 **5′UTR** 基因数量统计

物种	无 5′UTR	有 5′UTR	合计
雷蒙德氏棉 (*G. raimondii*)	9387 (25.0%)	28118 (75.0%)	37505 (100.0%)
拟南芥 (*A. thaliana*)	9948 (29.6%)	23654 (70.4%)	33602 (100.0%)

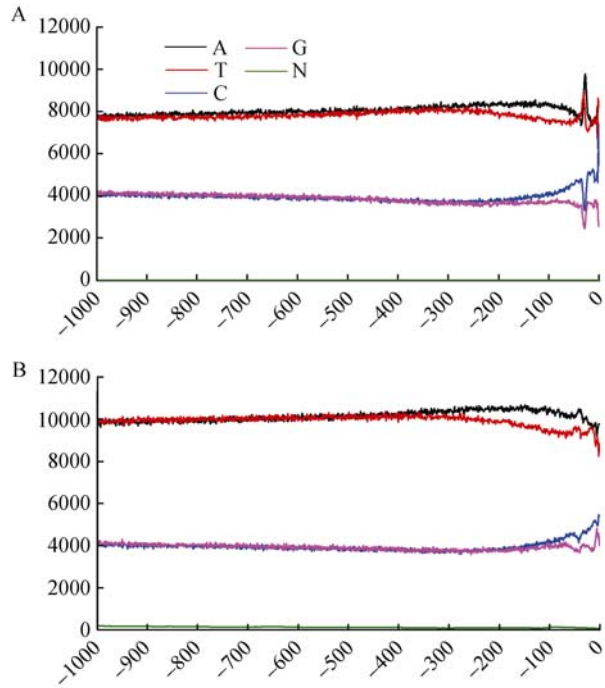

图 2　不同基因组中全部注释基因的启动子碱基分布
A：拟南芥基因组中有 5′UTR 基因启动子序列中的碱基分布; B：雷蒙德氏棉基因组中有 5′UTR 基因启动子序列中的碱基分布。X 轴均代表碱基的位置, Y 轴均代表所有启动子中某个碱基在这个位置上出现的频率。

2.3　CRE 在染色体上的分布

根据全基因组扫描定位，统计拟南芥和雷蒙德氏棉所有染色体上 CRE 数量和基因数量，结果表明拟南芥中 CRE 和基因在各条染色体上的分布比例差异在 0.16%以内，而雷蒙德氏棉在 0.05%以内，两个基因组各条染色体上的 CRE 和基因所占的比例非常接近(图 3)，说明 CRE 在不同染色体的启动子中数量整体分布均匀。

2.4　CRE 在启动子中的分布

利用 PLACE 数据库，对所有 469 个 CRE 在两个基因组的注释基因上游 1 000 bp 的启动子序列中分别进行扫描，在雷蒙德氏棉基因组启动子中能扫描到的 CRE 有 357 个(76.1%)，在拟南芥中有 368 个(78.4%)。两个物种共有的 CRE 有 350 个(74.6%)。

根据 1.2.4 定义，对 CRE 分布的峰值进行统计，当 P 值>6 时，其峰状分布特征比较明显。因此在本

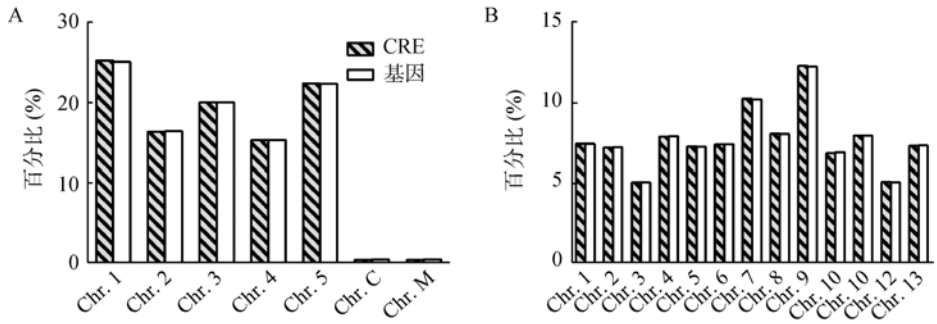

图 3　两个基因组上 CRE 和基因数量所占的比例

A：拟南芥基因组；B：雷蒙德氏棉基因组。X 轴表示染色体的编号(雷蒙德氏棉基因组的染色体编号为该基因组测序的编号)，Y 轴表示每条染色体上 CRE 或基因数量所占总数的比例。

研究中，满足 *P* 值>6，且 CRE 总数>100 的 CRE 定义为具有峰状分布的 CRE(当 CRE 总数<100 时，在整个启动子上分布比较分散，即使 *P* 值>6 也不能呈现出明显的峰)。按照这一标准，在雷蒙德氏棉上游启动子中呈峰状分布的 CRE 有 44 个(44/357=12.3%)，拟南芥中有 57 个(57/368=15.5%)。

2.4.1　两个基因组有 5′UTR 注释基因共有的呈峰状分布的 CRE

拟南芥和雷蒙德氏棉在注释基因上游启动子中共有的呈峰状分布的 CRE 有 34 个，根据核心序列可分为以下 4 类。

(1) TATABOX 类

TATABOX 类 CRE 是生物启动子中最重要的 CRE 之一，是 RNA 聚合酶的识别位点。共有 4 个 TATABOX 类 CRE 在两个物种启动子中有明显的峰状分布(表 2)。其中 TATABOX2(TATAAAT)和 TATABOX4 (TATATAA)在拟南芥分布特征最典型，形成一个突出高点(图 3A)。从峰的位置来看，拟南芥在−40 bp~−30 bp 之间，而雷蒙德氏棉在−50 bp~−40 bp 之间，说明雷蒙德氏棉的 TATABOX 类 CRE 在位置分布上比拟南芥要略为分散；从 CRE 总数来看，TATABOX2 和 TATABOX4 要明显高于另外两个 CRE，而 TATABOX1 的 *P* 值要大于其他 3 个 CRE，说明该 CRE 的峰状分布强度最高，而且这 4 个 CRE 的峰状分布强度都是拟南芥高于雷蒙德氏棉。TATABOX 在两个物种中的最高点位置和图 2 中 A/T 出现峰值的位置相符，因此 TATABOX 的富集应该是该位置 A/T 碱基出现峰值的原因。

(2) CT 富集类

CT 富集的 CRE 共有 3 个(表 3)，这 3 个 CRE 均从−60 bp 位置开始形成最高峰，并且高点位置在

表 2　TATABOX 类 CRE

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P* 值
TATABOX4	TATATAA	−7/−7	857/414	−4/−5	12183/19998	12.25/6.21
TATABOX2	TATAAAT	−7/−6	849/464	−4/−5	10753/21349	13.38/6.4
SORLREP3AT	TGTATATAT	−8/−6	99/48	−4/−4	1402/1330	8.58/7.72
TATABOX1	CTATAAATAC	−7/−6	97/25	−4/−4	223/181	63.14/27.48

表 3　CT 富集类 CRE

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P* 值
NODCON2GM	CTCTT	−6/−6	925/459	−2/−2	39447/30095	8.86/6.27
PYRIMIDINEBOXOSRAMY1A	CCTTTT	−6/−6	214/291	−4/−2	12426/16497	6.29/7.02
CTRMCAMV35S	TCTCTCTCT	−6/−6	318/88	−2/−3	3888/1566	27.25/19.72

–20 bp和–40 bp, 比TATABOX类更靠近TSS(表 3)。其中NODCON2GM的总数最大, 而CTRMCAMV35S的*P*值最大, 分布强度高。该类CRE在启动子中分布也非常广泛, 同时位置也比较保守。根据已有的研究, CT富集CRE可能是除TATABOX之外的另外一个重要的典型CRE[8,20]。

(3) ACGTG 类

ACGTG类是指包含核心序列ACGTG的CRE。ACGTG是G-box(CACGTG)的核心序列, 而G-box则是BHLH转录因子的一个主要的绑定序列[15]。一部分bZIP转录因子在拟南芥中也绑定G-box[16]。该类总共包含 18 种CRE(表 4), 不同类型CRE在启动子中的数量相差很大, 在拟南芥相差可达 280 多倍, 在雷蒙德氏棉相差可达 174 倍, 两物种排名前 5 位的CRE是 :ACGTATERD1, ABRELATERD1, ABRERATCAL, CACGTGMOTIF和ACGTABREMOTIFA2OSEM。拟南芥中AUXRETGA2GMGH3 的*P*值最大, 雷蒙德氏棉中EMBP1TAEM的*P*值最大。此类CRE在两个物种中峰的起始位置均在–110~–90 bp之间, 分布形态也非常相似, 说明这些CRE在两个物种中的分布具有很强的保守性, 此类CRE所结合的转录因子大部分与脱落酸、植物激素等生长发育调控有关。

(4) 其他峰状分布 CRE

除了上述 3 种能够明显划分为一类的 CRE, 还有其他一些共有的 CRE, 这些 CRE 序列间没有明显相似, 因此没有再进行系统归类(表 5)。其中MYBCOREATCYCB1 的总数最大, 而拟南芥中UP1ATMSD 的 *P* 值最大, 雷蒙德氏棉中 UPRMOTIFIIAT 的 *P* 值最大。

将表 2~表 5 中 CRE 随机挑选两个进行作图, 可以更加直观地反映出峰的分布和两个物种之间的区别(图 4)。

2.4.2　两个基因组特异的峰状分布 CRE

根据本文的定义, 在拟南芥中呈峰状分布而在雷蒙德氏棉中未呈峰状分布的 CRE 有 22 个(表 6), 其中 POLASIG1 的总数最多, 而拟南芥中 TELOBOXATEEF1AA1 的 *P* 值最大。以 ACGTCBOX 和 TGACGTVMAMY 为例作图, 其在拟南芥中分布峰状分布非常明显, 而在雷蒙德氏棉中在同样位置的峰则要平缓的多(图 5)。

表 4　ACGTG 类 CRE

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P*
ACGTATERD1	ACGT	–10/–11	893/642	–7/–9	54106/43205	7.76/6.65
ABRELATERD1	ACGTG	–10/–11	316/248	–9/–7	13729/12902	9.74/7.6
ABRERATCAL	MACGYGB	–10/–10	232/175	–7/–7	8208/7227	11.2/9.02
CACGTGMOTIF	CACGTG	–10/–11	179/123	–7/–9	4983/4452	12.88/10.24
ACGTABREMOTIFA2OSEM	ACGTGKC	–10/–10	114/72	–8/–9	2654/2406	15.51/9.47
GADOWNAT	ACGTGTC	–11/–10	63/37	–9/–7	1451/1170	13.48/9.17
IRO2OS	CACGTGG	–10/–10	47/41	–6/–9	1287/1165	11.65/8.74
BOXIIPCCHS	ACGTGGC	–10/–10	57/40	–6/–9	1203/1236	15.2/8.11
ABREATCONSENSUS	YACGTGGC	–10/–10	33/28	–10/–9	713/680	14.32/8.74
LRENPCABE	ACGTGGCA	–10/–10	32/23	–6/–9	585/750	16.54/6.43
ABREOSRAB21	ACGTSSSC	–13/–10	17/18	–9/–6	536/536	6.41/7.03
EMBP1TAEM	CACGTGGC	–11/–10	24/21	–8/–9	520/410	13.07/10.82
ABREZMRAB28	CCACGTGG	–10/–10	20/15	–7/–10	444/415	12.61/8.41
HEXAT	TGACGTGG	–10/–22	20/21	–6/–22	413/824	13.46/6.29
ABREATRD22	RYACGTGGYR	–12/–11	14/15	–10/–9	325/313	9.94/7.35
ABREMOTIFAOSOSEM	TACGTGTC	–11/–12	11/10	–11/–10	222/229	10.6/6.69
ACGTABREMOTIFAOSOSEM	TACGTGTC	–11/–12	11/10	–11/–10	222/229	10.6/6.69

AUXRETGA2GMGH3	TGACGTGGC	–10/–26	13/10	–7/–24	192/248	17.68/6.19

表 5 其他峰状分布的 CRE

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P*
MYBCOREATCYCB1	AACGG	–10/–8	180/148	–10/–8	11335/9040	6.38/6.01
SORLIP2AT	GGGCC	–11/–11	290/168	–8/–7	6744/9955	16.91/6.64
CGCGBOXAT	VCGCGB	–9/–6	142/107	–9/–6	6353/4545	8.12/7.52
MYBPLANT	MACCWAMC	–9/–10	95/83	–2/–8	4923/4438	6.09/6.25
UP1ATMSD	GGCCCAWWW	–9/–8	190/46	–7/–8	2584/1458	28.86/9.17
GAGA8HVBKN3	GAGAGAGAGAGAGAGA	–44/–17	20/14	–41/–92	653/341	6.42/6.02
UPRMOTIFIIAT	CCNNNNNNNNNNNNCCACG	–10/–11	27/22	–9/–9	481/483	13.74/10.93
GGTCCCATGMSAUR	GGTCCCAT	–11/–10	5/6	–36/–7	155/149	6.45/7.84

在雷蒙德氏棉中呈峰状分布而在拟南芥中未呈峰状分布的 CRE 有 12 个(表 7), 其中总数最多的 CRE 是 MARTBOX, 而雷蒙德氏棉中 *P* 值最大的 CRE 是 E2FANTRNR。以 E2FCONSENSUS 和 MYBPZM 为例, 其在雷蒙德氏棉中峰状分布非常明显, 而在拟南芥中同样位置的峰则相对平缓(图 6)。

从雷蒙德氏棉中未呈峰状分布的 CRE(表 7)以及在拟南芥中未呈峰状分布的 CRE(表 7)的 *P* 值可以进一步看出, 当 *P* 值>6 时, CRE 峰状分布特征比较明显, 当 *P* 值<6 时, CRE 未呈峰状分布。

3 讨 论

3.1 CRE 在雷蒙德氏棉和拟南芥基因启动子中的保守性

CRE序列普遍较短, 直接从启动子上进行扫描假阳性很高, 考虑到基因组序列的保守性, 如果CRE在所有已注释基因启动子中某个位置出现显著富集, 则说明这个位置极有可能是CRE与转录因子结合的位置。根据本研究定义, 在所有调查的CRE中, 有部分CRE在启动子中特定位置呈现明显的富集, 形成峰状分布, 说明有一定数量的基因启动子在此位置均含有该CRE。对比较典型的TATABOX类CRE进行分析发现, 拟南芥中顶峰的位置出现在–40 bp~–30 bp之间, 在雷蒙德氏棉中则出现在–50 bp~–30 bp之间, 这个位置正是TATABOX结合蛋白与TATABOX类CRE的结合位点[6]。因此, 我们推测CRE在启动子中富集的位置, 可能就是相关转录因子的实际结合位点, 这种明显的峰状分布在水稻启动子中也得到了验证[8]。

从 TATABOX 类 CRE 的情况, 我们推测其他具有类似峰状分布规律的 CRE 如 CG 富集类和 ACGTG 类 CRE 的顶峰所在位置也是转录因子的实际结合位点(图 3), 这些在 CRE 所结合的转录因子是植物正常生长发育和器官组织功能正常发挥所必须的, 所以这类具有重要生物学意义的 CRE 在不同基因启动子中的作用和位置都相对保守的。

此外, 本研究还发现在拟南芥和雷蒙德氏棉中有部分 CRE 的峰状分布呈物种特异性, 在拟南芥中呈峰状分布, 但雷蒙德氏棉中未呈峰状分布的 CRE 有 22 个(表 6), 在雷蒙德氏棉中呈峰状分布而在拟南芥中未呈峰状分布的 CRE 有 12 个(表 7), 这些 CRE 可以作为研究拟南芥和雷蒙德氏棉物种差异的参考。

3.2 雷蒙德氏棉和拟南芥顺式作用元件在启动子中的位置差异

同样以TATABOX2 和TATABOX4 为例(图 3A), 其在拟南芥中的峰值明显高于在雷蒙德氏棉的峰值, 形成这种明显对比的原因是这两个CRE在雷蒙德氏棉启动子的–50~–40 bp和–40~–30 bp两个连续的区间都富集, 而在拟南芥中则只是集中在–40~–30 bp区间。雷蒙德氏棉最高位点的数量明显小于拟南芥最高位点的数量, 但是雷蒙德氏棉最高位和相邻的次高点数量之和与拟南芥最高点的数量相近。TATABOX2/TATABOX4 在拟南芥中大量富集在–40~–30 bp区间, 而在雷蒙德氏棉中则分布在–50~–30 bp区间, 分布范围扩大了整整一倍, 从而

导致雷蒙德氏棉中的峰显得相对平缓，这种情况同样在水稻中也有发现[8]。

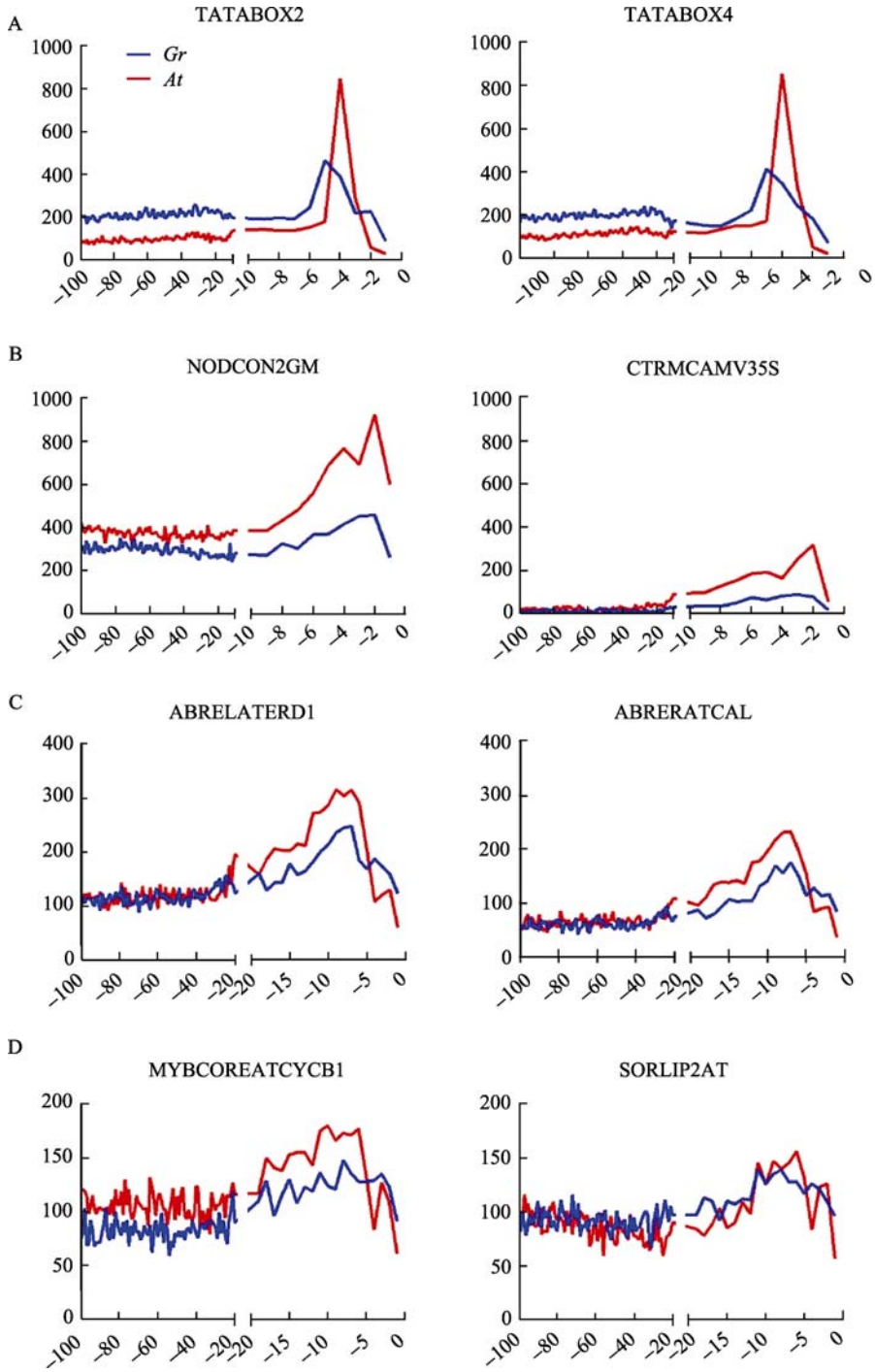

图 4　几个不同类型 **CRE** 在启动子区的峰状分布

A：2 个 TATA box 类 CRE; B：CT 富集类 CRE; C：ACGT 类 CRE; D：其他类 CRE。X 轴均代表启动子的位置，其中-1=10 bp, Y 轴均代表在当前位置具有该 CRE 出现的频率(图 5、图 6 的结构与此相同)。

为了进一步分析产生这种差异原因，本研究对两个基因组中所有 5′UTR 长度进行了简单的统计(图 7)，结果表明雷蒙德氏棉的 5′UTR 的平均长度和分布范围均大于拟南芥，因为本文所分析的序列均为 5′UTR 上游 1 000 bp 序列，以起始密码子为准，将所有的基因对齐，那么影响同一个 CRE 位置的可

能就是 5′UTR 的长度。因此，我们推测 5′UTR 长度在雷蒙德氏棉基因组中的更广泛变异可能是造成其

表 6 拟南芥中呈峰状分布的 CRE

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P*
POLASIG1	AATAAA	–10/–16	748/1150	–4/–5	44002/89994	6.81/5.88
INRNTPSADB	YTCANTYY	–6/–46	549/445	–2/–2	25521/33777	6.68/5.04
CCAATBOX1	CCAAT	–9/–6	425/469	–6/–3	29147/34239	6.34/5.84
DPBFCOREDCDC3	ACACNNG	–10/–12	225/155	–8/–12	14311/11327	6.19/4.83
POLASIG2	AATTAAA	–15/–16	228/515	–14/–16	13797/39504	6.23/5.91
SORLIP1AT	GCCAC	–9/–11	156/139	–6/–11	9045/9673	6.5/5.61
LTRECOREATCOR15	CCGAC	–9/–10	120/100	–8/–7	6494/6088	6.81/5.23
TATAPVTRNALEU	TTTATATA	–8/–38	194/146	–4/–5	4437/9578	6.6/4.8
UP2ATMSD	AAACCCTA	–6/–6	114/32	–3/–3	2030/1819	14.47/5.32
TGACGTVMAMY	TGACGT	–10/–25	77/50	–8/–8	3302/3072	7.93/5.01
HEXMOTIFTAH3H4	ACGTCA	–10/–10	94/60	–6/–7	3544/3214	8.44/5.55
WUSATAg	TTAATGG	–11/–20	63/61	–9/–19	2574/3883	6.72/5.22
ACGTCBOX	GACGTC	–9/–11	64/24	–6/–8	1927/978	8.54/5.83
TELOBOXATEEF1AA1	AAACCCTAA	–6/–57	70/22	–3/–60	1184/1224	13.18/4.21
DRE2COREZMRAB17	ACCGAC	–9/–12	49/31	–8/–10	2232/1925	6.33/4.41
MARABOX1	AATAAAYAAA	–11/–6	59/76	–4/–5	2045/4312	7.12/5.89
BOXCPSAS1	CTCCCAC	–6/–10	31/15	–2/–7	697/732	7.3/5.48
TCA1MOTIF	TCATCTTCTT	–6/–9	22/5	–4/–49	406/144	10.57/5.07
CDA1ATCAB2	CAAAACGC	–6/–8	18/8	–2/–40	435/296	6.96/4.71
ANAERO4CONSENSUS	GTTTHGCAA	–99/–94	14/7	–99/–91	359/208	6.96/3.42
UPRMOTIFIAT	CCACGTCA	–9/–10	22/24	–6/–7	433/755	12.24/5.76
PALINDROMICCBOXGM	TGACGTCA	–10/–11	18/6	–6/–43	341/161	8.57/5.88
ZDNAFORMINGATCAB1	ATACGTGT	–11/–18	10/7	–11/–14	329/280	6.62/4.36

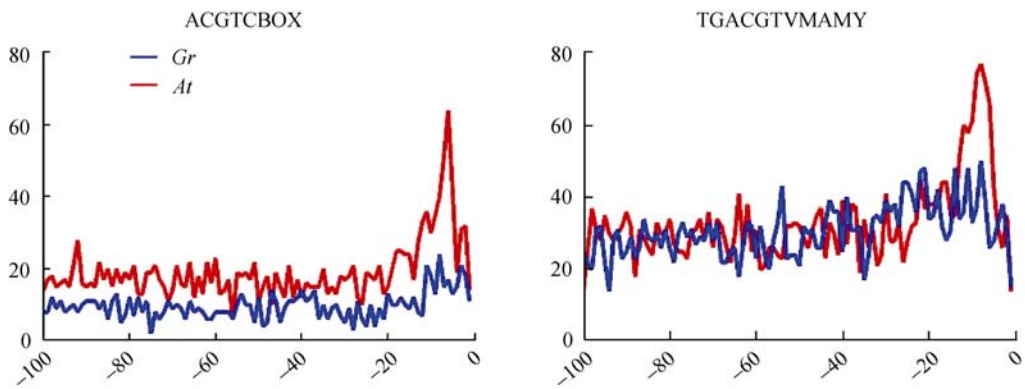

图 5 ACGTCBOX 和 TGACGTVMAMY 在拟南芥和雷蒙德氏棉启动子中的分布

CRE相对分散的重要原因，在对酵母菌启动子的研究中发现，5′UTR的长度变异与转录因子结合位点的分布模式具有相关性，同时可能影响基因的转录和可塑性[17]。5′UTR的长度在不同基因组中本身存在着较大变异，这种差异可能是由于物种基因组本身结构的进化所造成，对物种基因表达调控和表型变异具有深远的影响[18,19]。5′UTR长度的差异可能是物种之间的差异，也有可能是因为雷蒙德氏棉基因

组的注释不够精确造成的，但无论哪种原因，根据本研究的结果可以推测，5′UTR长度的变异程度

表 7　雷蒙德氏棉中呈峰状分布的 **CRE**

CRE 名称	CRE 序列	拟南芥/雷蒙德氏棉				
		最高峰起始位置	最高点数值	最高点位置	CRE 总数	*P*
MARTBOX	TTWTWTTWTT	–25/–6	616/991	–2/–2	25399/35048	5.47/7.02
SEF3MOTIFGM	AACCCA	–6/–6	134/184	–3/–4	7266/9806	5.85/6.88
E2FCONSENSUS	WTTSSCSS	–10/–6	80/122	–6/–2	4157/4533	5.77/7.08
MYBPZM	CCWACC	–9/–8	83/147	–47/–7	5409/7333	5.45/6.65
BOXLCOREDCPAL	ACCWWCC	–6/–10	75/76	–2/–8	3847/3998	5.7/6.32
PALBOXAPC	CCGTCC	–16/–6	28/27	–12/–2	1157/1062	5.69/6.51
ACGTOSGLUB1	GTACGTG	–52/–13	12/11	–6/–13	536/372	3.62/6.22
E2FANTRNR	TTTCCCGC	–10/–6	9/9	–6/–3	347/145	4.9/10.27
E2F1OSPCNA	GCGGGAAA	–83/–9	7/5	–89/–9	281/117	4.15/7.62
CACGCAATGMGH3	CACGCAAT	–19/–10	5/7	–99/–10	128/114	4.98/7.32

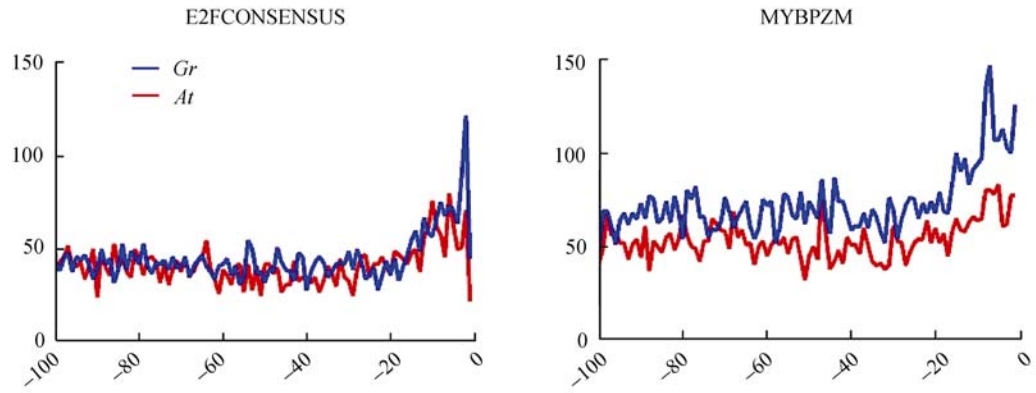

图 6　**E2FCONSENSUS** 和 **MYBPZM** 在拟南芥和雷蒙德氏棉启动子中的分布

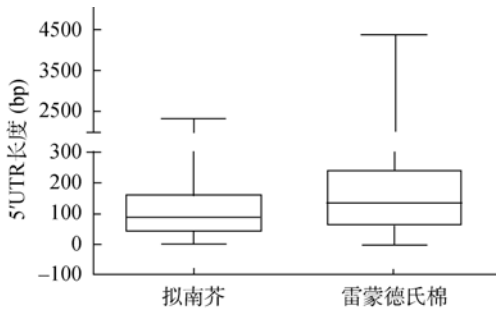

图 7　两个基因组中 **5′UTR** 长度比较

和 CRE 分布具有相关性。

3.3　重要 CRE 在启动子中的相对位置关系

在拟南芥启动子所有具有峰状分布规律的CRE中，约 93%(53/57)的峰位于–110 bp~–1 bp之间，非常靠近TSS，而在启动子的上游没有发现明显的峰。在雷蒙德氏棉中对应的比例约为 87% (39/44)，这可能意味着实际的转录因子结合位点更加倾向于靠近转录起始位点[18]。同时可以看到，ACGTG类、TATABOX类、CT富集类CRE在启动子中峰的位置分别为–110 bp~ –90 bp、–50 bp~ –30 bp、–30 bp~–10 bp，暗示了这几类重要的CRE之间位置相对保守，这种位置特征可能意味着不同的转录因子可以在紧靠TSS聚集形成多聚体，从而发挥相应功能。

参考文献(References):

[1] Wang KB, Wang ZW, Li FG, Ye WW, Wang JY, Song GL, Yue Z, Cong L, Shang HH, Zhu SL, Zou CS, Li Q, Yuan YL, Lu CR, Wei HL, Gou CY, Zheng ZQ, Yin Y, Zhang XY, Liu K, Wang B, Song C, Shi N, Kohel RJ, Percy RG, Yu JZ, Zhu YX, Wang J, Yu SX. The draft genome of a diploid cotton *Gossypium raimondii*. *Nat Genet*, 2012, 44(10): 1098–1103. DOI

[2] Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, van de Peer Y. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol*, 2003, 132(3): 1162–1176. DOI

[3] Su J, Teichmann SA, Down TA. Assessing computational methods of cis-regulatory module prediction. *PLoS Comput Biol*, 2010, 6(12): e1001020. DOI

[4] 陈鸿飞, 王进科. 转录因子相关数据库. 遗传, 2010, 32(10): 1009–1017. DOI

[5] Priest HD, Filichkin SA, Mockler TC. *Cis*-regulatory elements in plant cell signaling. *Curr Opin Plant Biol*, 2009, 12(5): 643–649. DOI

[6] Molina C, Grotewold E. Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics*, 2005, 6(1): 25. DOI

[7] Ding J, Hu HY, Li XM. Thousands of cis-regulatory sequence combinations are shared by *Arabidopsis* and poplar. *Plant Physiol*, 2012, 158(1): 145–155. DOI

[8] Civán P, Svec M. Genome-wide analysis of rice (*Oryza sativa* L. subsp. *japonica*) TATA box and Y Patch promoter elements. *Genome*, 2009, 52(3): 294–297. DOI

[9] Zou C, Sun KL, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu SH. *Cis*-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA*, 2010, 108(36): 14992–14997. DOI

[10] Sharma N, Russell SD, Bhalla PL, Singh MB. Putative cis-regulatory elements in genes highly expressed in *rice* sperm cells. *BMC Res Notes*, 2011, 4(1): 319. DOI

[11] 张梅, 刘炜, 毕玉平. 植物中DREBs类转录因子及其在非生物胁迫中的作用. 遗传, 2009, 31(3): 236–244. DOI

[12] 侯琳, 钱敏平, 朱云平, 邓明华. 转录因子结合位点生物信息学研究进展. 遗传, 2009, 31(4): 365–373. DOI

[13] Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu SQ, Udall J, Yoo MJ, Byers R, Chen W, Doron-Faigenboim A, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu GJ, Lee TH, Li JP, Lin LF, Liu T, Marler BS, Page JT, Roberts AW, Romanel E, Sanders WS, Szadkowski E, Tan X, Tang HB, Xu CM, Wang JP, Wang ZN, Zhang D, Zhang L, Ashrafi H, Bedon F, Bowers JE, Brubaker CL, Chee PW, Das S, Gingle AR, Haigler CH, Harker D, Hoffmann LV, Hovav R, Jones DC, Lemke C, Mansoor S, ur Rahman M, Rainville LN, Rambani A, Reddy UK, Rong JK, Saranga Y, Scheffler BE, Scheffler JA, Stelly DM, Triplett BA, Van Deynze A, Vaslin MF, Waghmare VN, Walford SA, Wright RJ, Zaki EA, Zhang T, Dennis ES, Mayer KF, Peterson DG, Rokhsar DS, Wang X, Schmutz J. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*, 2012, 492(7429): 423–427. DOI

[14] Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*, 1999, 27(1): 297–300. DOI

[15] Toledo-Ortiz G, Huq E, Quail PH. The *Arabidopsis* basic/helix-loop-helix transcription factor family. *Plant Cell*, 2003, 15(8): 1749–1770. DOI

[16] Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F. bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci*, 2002, 7(3): 106–111. DOI

[17] Lin ZG, Wu WS, Liang H, Woo Y, Li WH. The spatial distribution of *cis* regulatory elements in yeast promoters and its implications for transcriptional regulation. *BMC Genomics*, 2010, 11(1): 581. DOI

[18] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 2003, 20 (9): 1377–1491. DOI

[19] Lynch M, Douglas Scofield DG, Hong X. The evolution of transcription-initiation sites. *Mol Biol Evol*, 2005, 22 (4): 1137–1146. DOI

[20] Bernard V, Brunaud V, Lecharny A. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics*, 2010, 11(1): 166. DOI