雷蒙德氏棉 EST-SSRs 分布特征及开发与利用

王长彪 郭旺珍* 蔡彩平 张天真

(南京农业大学作物遗传与种质创新国家重点实验室, 南京 210095. * 联系人, E-mail: moelab@njau.edu.cn)

摘要 微卫星或简单重复序列(simple sequence repeats, SSR)存在于表达序列标签(expressed sequence tags, ESTs)中. 为了在棉花中开发 EST-SSR 功能性标记, 利用生物信息学方法对 NCBI 网上公开的 63485 条雷蒙德氏棉(*Gossypium raimondii* Ulbrich) ESTs 序列进行 EST-SSRs 特征分析. 剔除冗余序列, 得到非冗余序列 58906 条. 在非冗余序列中发现含不同重复基元 SSRs 的 EST 序列有 2620 条, 共 2818 个 EST-SSRs, EST-SSRs 序列的频率是 4.45%, 平均相隔 14.8 kb 出现一个 SSR. 在 1~6 bp 的重复基元中, 三核苷酸重复基元的 SSRs 出现频率最高(38.31%), 其次是二核苷酸(24.09%)、单核苷酸(23.35%). 对所有的重复基元类型进行统计分析发现, 所占比例最大的是 A/T(18.67%), 其次是 AT/TA(14.83%). 在复合型(compound)中发现三核苷酸串联三核苷酸的重复基元出现频率最高, 为 48.65%. 利用 Prime 3 软件, 设计了 1554 对 EST-SSRs 引物, 随机选用 300 对对本室四倍体作图亲本陆地棉 TM-1 和海岛棉海 7124 进行多态性检测, 其中 129 对有多态性, 多态性频率为 43%. 这些 EST-SSRs 将有效用于不同棉种间的分布特征比较及染色体定位等方面研究.

关键词 雷蒙德氏棉 **EST-SSRs** 分布 分子标记 多态性

棉花是世界性重要的经济作物. 目前, 世界上已有多个实验室利用分子标记技术进行了棉花遗传图谱构建[1~4]、品种纯度检测[5]、遗传多态性分析[6]、分子标记辅助选择育种[7,8]和基因定位[9]等方面研究. 然而, 相对于较大的棉花基因组, 可有效利用的标记数量非常有限, 从深度和广度进一步开展棉花基因组研究, 需要开发新的分子标记. 众多分子标记中, SSR(simple sequence repeat)分子标记从实用性和有效性方面均优越于其他标记类型. 近年来, 许多作物大规模的cDNA单边测序并在网上公开释放, 大大增加了基于ESTs (expressed sequence tags)的SSR标记开发能力. 自 2000 年起, 已相继在葡萄(*V. vinifera*)[10]、甘蔗(*Saccharum* spp)[11]、硬粒小麦(*Triticum durum*)[12]、黑麦(*Secale cereale* L)[13]、大麦(*Hordeum uhulgare* L)[14]、亚洲棉(*Gossypium arboreum* L)[4]、小麦(*Triticum* L)[15]和马铃薯(*Solanum tuberosum* L)[16]等作物开展了EST-SSRs标记的开发并广泛用于基因组研究和分子育种.

截止 2005 年 8 月 26 日, 在GenBank (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)中释放有 5 个棉种的棉花EST共 141267 条. 其中最多的 3 个棉种是雷蒙德氏棉(*G. raimondii* Ulbrich) 63577 条, 亚洲棉(*G. arboreum* L) 39216 条和陆地棉(*G. hirsutum* L)38093 条. 基于上述EST序列信息, 国内外棉花遗传育种研究人员已开展了初步研究. Chee等人[3]利用部分亚洲棉的ESTs序列开发出STS标记并定位了 10 个功能基因. Han等人[4]从亚洲棉的ESTs序列中开发了 554 对EST-SSRs引物, 其中 99 对定位到四倍体栽培棉种的遗传图谱上. 雷蒙德氏棉属于D_5基因组的二倍体野生棉种, 与四倍体棉种的供体亲本之一亲缘关系最近. 本研究从GenBank公布的雷蒙德氏棉ESTs中发掘出2620条含SSR的EST序列, 并研究了这些EST-SSRs在棉花转录组中的分布特征, 进一步设计EST-SSRs引物 1554 对, 进行海陆栽培棉种间的多态性分析. 为构建饱和遗传图谱、发掘基因和分子标记辅助选择育种奠定了坚实基础.

1 材料与方法

(ⅰ) EST序列来源. 2005 年 2 月 18 日从 dbEST/GenBank (http://www.ncbi.nlm.nih/entrez)中以 FASTA 格式下载了 63485 条 ESTs 序列, 它们来自雷蒙德氏棉开花前 3 d 到开花后 3 d 和第一片真叶的 ESTs.

(ⅱ) EST-SSRs 的开发. 采用 Clastal X 1.81 (http://www.digitalgene.net/Soft/Sequences/lignment/200409/0.html)、Treeview (ver 1.61) (http://www.taxonomy.zool-ogy.gla.ac.nkrod/od.html 和 Genedoc (ver 2.6.02) (http://www.psc.edu/biomed/genedoc/gddl.htm)软件对 63485 条ESTs序列进行冗余性查找, 然后利用在线软件 SSRIT (http://arsgenome.cornell.edu/cgi-

bin/rice/ssrtool.pl)在非冗余序列中查找SSRs. SSRIT是美国USDA-ARS生物信息技术中心和Cornell大学比较基因组学研究中心开发的SSRs查找工具, 以Pel脚本编写. 利用该软件查找了二、三、四、五、六核苷酸 5 种类型的SSRs, 利用TRF (tandem repeats finder, http://tandem.bu.edu/trf/trf.html)查找单核苷酸类型的SSRs. SSRs的查找标准为: 单核苷酸重复≥25, 二核苷酸重复≥14 bp, 三核苷酸重复≥15 bp, 四核苷酸重复≥16 bp, 五核苷酸重复≥20 bp, 六核苷酸重复≥24 bp. 其中二、三、五核苷酸重复基元的查找标准参照Cardle等人[17]的报道.

(iii) EST-SSRs的引物开发. 利用Primer 3.0 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi)对包含有SSRs的2620条EST进行引物设计, 得到 1554 对EST-SSRs引物. 引物设计的主要参数是: 引物长 18~20 bp, 最适为 20 bp; PCR产物长 100~ 250 bp; 最适T_m值为 57℃; GC含量为 35%~65%, 最适 50%. 引物合成由上海英俊生物技术有限公司完成.

(iv) DNA提取、SSR扩增和电泳. 海岛棉(*G. barbadense* L)品种海 7124 和陆地棉遗传标准系TM-1是本室用于四倍体栽培棉种遗传图谱构建的作图亲本. 其DNA提取方法参照Paterson等人[18]的报道. SSR-PCR扩增在美国MJ Research公司的PTC-225 上完成. 扩增产物的电泳检测参照Zhang等人[19]的报道.

2 结果和讨论

2.1 源于雷蒙德氏棉 ESTs 的 SSRs 发掘

用 Clastal X 1.81, Treeview 和 Genedoc 软件对63485 条 ESTs 序列比较分析, 去除冗余的序列, 得到58906 条非冗余 ESTs 序列, SSRIT 查找获得 2818 个EST-SSRs, 分布于 2620 条 EST 序列中. 包含 SSR 的ESTs 频率是 4.45%. 在 58906 条雷蒙德氏棉非冗余ESTs 序列中, 拼接总长度为 41612.63 kb, 平均相隔14.8 kb 就出现一个 SSR.

用Weber[20]的分类标准可将 2818 个EST-SSRs分成精密型(perfect)、非精密型(imperfect)和复合型(compound) 3 种类型, 其中精密型SSRs有 2632 个(93.40%), 非精密型SSRs有 75 个(2.66%), 复合型SSRs有 111 个(3.94%)(表 1). 在精密型SSRs中, 三核苷酸重复基元类型的SSRs出现频率最多, 而在非精密型 SSRs 中, 则二核苷酸重复基元类型最多. 复合型 SSRs 又分成精密复合型(perfect compound)和非精密复合型(imperfect compound)两类.

表 1 EST-SSRs 类型及分布频率

	重复类型	数量	百分率(%)
精密型	单核苷酸	632	22.43
	二核苷酸	583	20.69
	三核苷酸	1031	36.59
	四核苷酸	188	6.67
	五核苷酸	82	2.91
	六核苷酸	116	4.12
	合计	2632	93.40
非精密型	二核苷酸	69	2.45
	三核苷酸	6	0.21
	合计	75	2.66
复合型	精密型	72	2.56
	非精密型	39	1.38
	合计	111	3.94

基于不同的EST-SSRs搜寻标准, 不同研究者已经对许多作物中的EST-SSRs分布特征进行研究. 按照Cardle等人[17]的统计标准, 在拟南芥的ESTs序列中, 平均每 13.8 kb出现一个SSR, 水稻中出现频率为3.4 kb, 玉米为 8.1 kb, 大豆为 7.4 kb, 西红柿为 11.1 kb, 棉花为 20.0 kb, 杨树为 14.0 kb. Morgante等人[21]将EST-SSR的统计标准定为 1~5 碱基重复至少 3 次, 总长度为 12 个碱基, 结果显示, 在拟南芥中每 2.1 kb出现一个SSR. Gao等人[22]将不同重复基元的总长度定为 18 bp以上, 得出水稻中每 11.81 kb出现一个SSR的结果. 不同研究结果的差异主要归因于发掘EST-SSRs时所采用的标准不一样, 同时也和不同作物EST序列来源的组织、器官不同有关. 在本研究中, 与Cardle等人[17]的分类标准相比, 搜索单、四核苷酸重复基元重复次数的标准提高, 增加了对六核苷酸重复基元的筛选, 结果显示在雷蒙德氏棉中每 14.8 kb的EST出现一个SSR.

2.2 雷蒙德氏棉 EST-SSRs 的分布特征

2818 个SSR中共有 170 种重复基元(motif). 在精密型和非精密型SSRs中, 单、二、三、四、五和六碱基重复基元中出现频率最多的重复基元分别是$(A/T)_n$, $(AT/TA)_n$, $(AAG/TTC)_n$, $(ATAC/TATG)_n$, $(GAAAA/CTTTT)_n$和$(AAAAAT/TTTTTA)_n$(表 2). 它们在各自重复基元类型中的比例分别是 83.23%, 64.11%, 26.13%, 21.81%, 20.37%和 4.31%. 在所有类

型的重复基元中，三核苷酸重复基元出现的频率最高为 38.31%，其次分别为二、单、四、六和五核苷酸重复基元(表 2).

在检测到的 170 种 SSR 重复基元中，所占比例最高的是 A/T(18.67%)，其次分别是 AT/TA(14.83%)等. 不同类型重复基元的 EST-SSRs 分布见图 1.

在复合型 SSRs 中，至少每一串联重复基元的长度大于 10 bp. 根据串联重复基元的不同长度把 SSRs 分成 11 种类型，即 m1:2-2(两核苷酸和两核苷酸重复基元串联，依此类推)，m2:2-3，m3:3-3，m4:4-2，m5:5-2，m6:3-4，m7:6-3，m8:4-4，m9:1-2，m10:6-6 和 m11:5-4. 这 11 种类型中出现频率最高的是 m3，即在复合型 SSRs 中出现频率最高的是三核苷酸串联三核苷酸的 SSRs，其频率是 48.65%，其次是 m1 为 34.23%.

虽然在不同作物的 EST-SSRs 序列分布特征研究中，不同学者搜索SSRs采用的标准不同. 但对大部分植物的EST-SSRs序列综合调查，均表明重复基元为三核苷酸的EST-SSRs出现频率最高，在雷蒙德氏棉中也是如此. 主要原因是EST是cDNA单边测序的一段序列，除 3′或 5′的非翻译区外，其余属于编码区. 三核苷酸重复出现频率高是与其编码区三联体密码的编译相对应的. 在蔷薇(Rosa)中，二核苷酸重复基元频率高可能与所调查的ESTs序列主要来自cDNA的 3′转录非翻译区(UTR)有关[23]. 进一步的分析表明，不同植物中不同大小的重复基元出现丰度有一定程度相同.

在小麦、水稻、玉米、大豆中二核苷酸重复基元(motif)出现频率最多的都是$(AG/TC)_n$[20,24]，三核苷酸重复基元出现频率最多的分别是$(AAC/TTG)_n$，$(AGG/TCC)_n$，$(CCG/GGC)_n$和$(AAG/TTC)_n$[20]；在大麦和高粱中都是$(CCG/GGC)_n$[14,24]. 在本研究中，雷蒙

表 2 精密型和非精密型 SSRs 中不同重复基元(motif)出现的频率

重复基元类型	数量	频率(%)	最多的重复基元及其数量和百分率
单核苷酸	632	23.35	A/T (526, 83.23%)
二核苷酸	652	24.09	AT/TA(418, 64.11%)
三核苷酸	1037	38.31	AAG/TTC (271, 26.13%)
四核苷酸	188	6.94	ATAC/TATG (41, 21.81%)
五核苷酸	82	3.03	GAAAA/CTTTT(17, 20.37%)
六核苷酸	116	4.29	AAAAAT/TTTTTA(5, 4.31%)

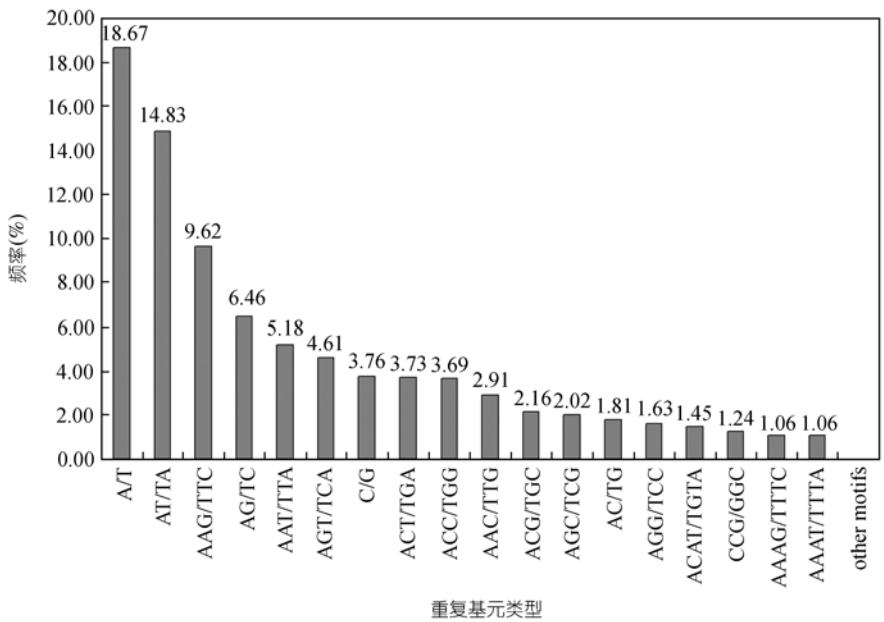

图 1 基于重复基元(motif)类型的 EST-SSRs 分布

other motifs 表示频率小于 1.00%的重复基元类型

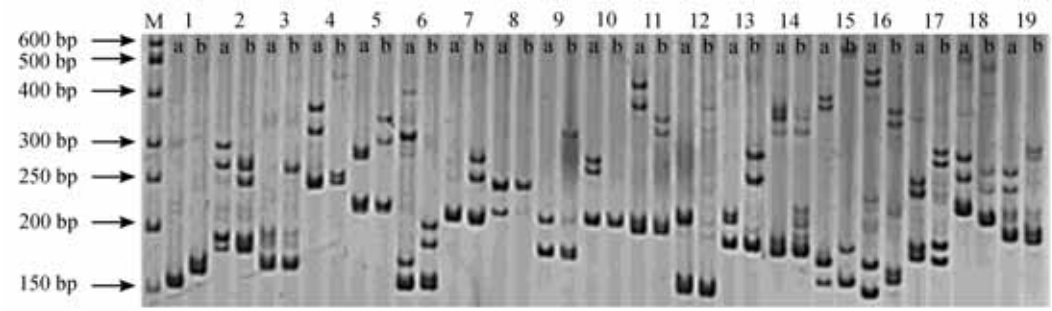

图 2 19 对引物在海陆棉种间的多态性

1~19 为 NAU2556, NAU2565, NAU2626, NAU2639, NAU2680, NAU2692, NAU2697, NAU2701, NAU2706, NAU2717, NAU2723, NAU2730, NAU2733, NAU2649, NAU2650, NAU2679, NAU2715, NAU2801, NAU2803; a, 海 7124; b, TM-1

德氏棉中二核苷酸重复基元出现频率最多的是 $(AT/TA)_n$, 三核苷酸重复基元中出现频率最多的是 $(AAG/TTC)_n$.

关于 EST-SSRs 分布特征, Clemson 大学基因组研究所棉花中心(Clemson University Genomics Institute, CUGI)对不同来源 ESTs 数据进行分析, 发现 34819 条亚洲棉 ESTs 序列中包含 SSRs 的 EST 频率是 2.73%, 1716 条陆地棉 ESTs 序列中其分布频率是 2.16%, 而对开花后 7~10 d, 开花前 3 d 到开花后 15 d 以及开花后 6 d 的未成熟纤维 3 种不同来源的 EST 序列综合分析, 发现 EST-SSRs 的分布频率为 2.92% (http://www.genome.clemson.edu/projects/cotton/ssr). 本研究中, 58906 条雷蒙德氏棉 EST 序列中, 包含 SSR 的 EST 分布频率是 4.45%.

2.3 雷蒙德氏棉 EST-SSRs 标记开发及其在海、陆四倍体栽培棉种间的多态性

利用Primer 3.0软件, 对2818个EST-SSRs序列进行EST-SSRs的引物设计, 共开发了 1554 对EST-SSRs引物. 随机选取 300 对引物对本室用于构建异源四倍体栽培棉种遗传图谱的 2 个亲本海岛棉海 7124 和陆地棉标准系TM-1 进行了遗传多态性分析, 发现有 129 对引物在TM-1 和海 7124 亲本间检测到多态性, 多态性率为 43%. 检测到的多态性差异明显, 一些引物还检测到 2 个或 2 个以上的位点差异(图 2), 易于作为多态性标记用于棉花基因组研究和分子标记辅助育种. 一般认为, EST-SSRs标记揭示的种间多态性低于源于基因组的SSR标记. Han等人[4]在 554 对源于亚洲棉的EST-SSRs引物中, 有 99 对引物在TM-1 和海 7124 间检测到多态性, 其多态性率是 17.9%. Chee等人[3]利用亚洲棉的ESTs也开发了 89 对EST-SSRs引物, 但仅 13 对在陆地棉和海岛棉中产生多态性, 其多态性率是 14.6%. 相比较而言, 本实验中利用雷蒙德氏棉ESTs开发的EST-SSRs标记在陆地棉和海岛棉间的多态性大大高于前人的研究. 因此, 源于雷蒙德氏棉转录组中的这些EST-SSRs标记可更有效用于棉花基因组研究. 本文中开发的其他 1254 对雷蒙德氏棉EST-SSRs引物也正在进行海岛棉和陆地棉DNA多态性分析及四倍体栽培棉种遗传图谱构建.

EST本身是功能基因的一部分序列, 从EST中开发出的EST-SSRs标记理论上可为功能基因提供“绝对”的标记. 雷蒙德氏棉种子虽然不长有价值的纤维但种皮密布绒毛. 作为四倍体棉种的供体种之一, 许多研究已经证明, 在D-亚基因组上有与纤维发育相关的QTLs[9,25,26]. 本研究利用的EST序列大部分来源于雷蒙德氏棉开花前 3 d到开花后 3 d胚珠cDNA序列, 因此, 利用本研究开发的雷蒙德氏棉EST-SSRs标记可有效用于棉纤维发育的基因组和分子育种研究.

致谢 本工作受国家自然科学基金(批准号: 30471104, 30270806)、教育部长江学者和创新团队项目、教育部新世纪优秀人才项目(批准号: NCET-04-0500)、江苏省人才基金(批准号: BK2003414)和江苏省高技术项目(批准号: BG2004305)资助.

参 考 文 献

1 Reinisch A, Dong J M, Brubaker C L, et al. A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*, chromosome organization and evolution in a disomic polyploid genome, Genetics, 1994, 138: 829~847

2 Zhang J, Guo W, Zhang T. Molecular linkage map of allotetraploid cotton (*Gossypium hirsutum* L × *Gossypium barbadense* L) with a haploid population. Theor Appl Genet, 2002, 105: 1166~1174[DOI]

3 Chee P, Rong J K, Williams-Coplin D, et al. EST derived PCR-based markers homologues in cotton. Genome, 2004, 47: 449~462[DOI]

4 Han Z G, Guo W Z, Song X L, et al. Genetic mapping of EST-derived

microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol Gen Genomics, 2004, 272: 308~327[DOI]

5 Yi C X, Zhang T Z. Preliminary studies on molecular marker used in purity test of hybrid seeds in upland cotton. Acta Gossypii Sinica, 1999, 11(6): 318~320

6 Liu S, Cantrell R G, Macarty J C, et al. Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Sci, 2000, 4: 1459~1469

7 Zhang T Z, Yuan Y L, Yu J, et al. Molecular tagging of a major QTL for fiber strength in upland cotton and its marker-assisted selection. Theor Appl Genet, 2003, 106: 262~268

8 Guo W Z, Zhang T Z, Zhu X F, et al. Modified backcross pyramiding breeding with molecular marker-assisted selection and its applications in cotton. Acta Agron Sinica, 2005, 31(8): 963~970

9 Shen X L, Guo W Z, Zhu X F, et al. Molecular mapping of QTLs for fiber qualities in three diverse lines in upland cotton using SSR markers. Mol Breed, 2005, 15: 169~181[DOI]

10 Scott K D, Eggler P, Seaton G, et al. Analysis of SSRs derived from grape ESTs. Theor Appl Genet, 2000, 100: 723~726[DOI]

11 Cordeiro G M, Casu R, McIntyre C L, et al. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci, 2001, 160: 1115~1123[DOI]

12 Eujayl I, Sorrells M E, Baum M, et al. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet, 2002, 104: 399~407[DOI]

13 Hackauf B, Wehling P. Identification of microsatellite polymorphisms in an expressed portion of the rye genome. Plant Breed, 2002, 121: 17~25[DOI]

14 Thiel T, Michalek W, Varshney R K, et al. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L). Theor Appl Genet, 2003, 106: 411~422

15 Peng J H, Nore L, Lapitan V. Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct Integr Genomics, 2005, 5: 80~96[DOI]

16 Feingold S, Lloyd J, Norero N, et al. Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L). Theor Appl Genet, 2005, 111: 456~466[DOI]

17 Cardle L, Ratnsay L, Milbourne D, et al. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics, 2000, 156: 847~854

18 Paterson A H, Brubaker C, Wendel J F. A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol Biol Rep, 1999, 11: 122~127

19 Zhang J, Wu Y T, Guo W Z, et al. Fast screening of microsatellite markers in cotton with PAGE/silver staining. Cotton Sci Sin, 2000, 12: 267~269

20 Weber J L. Informativeness of human (dC-dA)n·(dG-dT)n polymorphisms. Genomics, 1990, 7: 524~530[DOI]

21 Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet, 2002, 30: 194~200[DOI]

22 Gao L F, Tang J F, Li H W, et al. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed, 2003, 12: 245~261[DOI]

23 Sook J, Albert A, Christopher J, et al. Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics, 2005, 5: 136~143[DOI]

24 Kantety R V, La R M, Matthews D E, et al. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol, 2002, 48: 501~510[DOI]

25 Jiang C X, Wright R J, El-Zik K M, et al. Polyploid formation created unique convenues for response to selection in *Gossypium* (cotton). Proc Natl Acad Sci USA, 1998, 95: 4419~4424[DOI]

26 Kohel R J, Yu J, Park Y H, et al. Molecular mapping and characterization of traits controlling fiber quality in cotton. Euphytica, 2001, 121: 163~172[DOI]

(2005-09-30 收稿, 2005-12-12 接受)