甜瓜EST序列中微卫星的分布特征

胡建斌, 刘颖, 王兰菊, 李建吾*
河南农业大学园艺学院, 郑州 450002

提要: GenBank中35 547条甜瓜EST经去冗余处理后, 得到总长度为250.3 Mb的无冗余EST 34 438条。这些序列中有2 813个微卫星简单重复序列(Simple sequence repeat, SSR), 分布于2 107条EST中, 出现频率为8.16%, 平均分布距离为8.90 kb。三核苷酸重复是主导重复类型, 占SSR总数的47.14%; 其次是二核苷酸和单核苷酸重复, 分别占SSR总数的20.72%和16.99%。AAG/TTC是优势重复基元, 占微卫星总数的29.26%, AG/CT和A/T分别占14.61%和16.25%。在所有的SSR中, 重复次数为4~10次的占70.32%, 长度为12~20 bp的占51.12%。并对这些SSR的多态性潜能进行了评价。
关键词: 甜瓜; EST序列; 微卫星; 特征

Characteristics of Microsatellites in Melon (*Cucumis melo* L.) EST Sequences

HU Jian-Bin, LIU Ying, WANG Lan-Ju, LI Jian-Wu*
College of Horticulture, Henan Agricultural University, Zhengzhou 450002, China

Abstract: 34 438 melon expressed sequence tags (ESTs) were obtained after the removal of redundant sequences from 35 547 melon ESTs deposited in GenBank. In these ESTs, 2 813 microsatellites (simple sequence repeat, SSR) were tested and distributed in 2 107 ESTs. The occurrence frequency of SSR was 8.16% and the average distribution distance was 8.90 kb. Trinucleotide repeats were dominant repeat types and accounted for 47.14%. Dinucleotide and mononucleotide repeats accounted for 20.72% and 16.99%, respectively. AAG/TTC was most frequent repeat motif and accounted for 29.26% in all SSRs. AG/CT and A/T accounted for 14.61% and 16.25%, respectively. Among all SSRs, SSRs which repeated 4 to 10 times accounted for 70.32%, and SSRs which was 12 to 20 bp long accounted for 51.12%. The potential polymorphism of these SSRs was also analyzed.
Key words: melon (*Cucumis melo* L.); EST sequence; microsatellite; characteristic

微卫星或简单重复序列(simple sequence repeat, SSR)广泛存在真核生物的基因组中, 因其数量丰富、多态性高、多等位性、共显性等特点(Powell等1996; Varshney等2005a), SSR标记被公认为目前遗传学研究中最令人信赖的分子标记之一。然而, SSR引物的开发通常需构建文库、筛选、测序等工作, 费时费力。公共数据库中数量不断增加的表达序列标签(expressed sequence tag, EST)极大地增强了对基于EST的SSR标记开发能力。迄今, 在多种农作物中, 大量的EST-SSR标记已经得到开发, 并广泛用于遗传图谱的构建、基因发掘、遗传进化及比较基因组的研究(李永强等2004; 李小白等2006)。

甜瓜(*Cucumis melo* L.)为葫芦科(Cucurbitaceae)黄瓜属植物, 是一种世界性的园艺作物, 栽培历史悠久, 种质资源丰富, 世界各国都非常重视其遗传学研究。甜瓜基因组学的研究进展较快, 现已构建了12张遗传图谱, 一些控制重要经济性状和抗病性的基因或QTL已定位在图谱中(苏芳等2007), 为甜瓜基因结构和功能的研究奠定了基础。然而, 现已发表的甜瓜遗传图谱主要由同工酶、RAPD、RFLP、AFLP等标记构成, 与小麦、水稻和棉花等农作物的图谱相比, 甜瓜遗传图谱所包含的SSR标记较少。另一方面, RAPD、AFLP、基因组SSR等分子标记或是扩增非编码区域, 或是随机在基因组中扩增, 得到的位点一般与目标性状基因的距离较远, 以致分子标记在应用上与其目标有一定的偏差。EST-SSR来源于编码区DNA, 代表基因表达信息, 这使得直接鉴定决定重要表型性状的等位基因的工作成为可能(Kumpatla和Mukhopadhyay 2005)。到现在为止, 甜瓜基因组测序已得到3万多条EST, 为EST-SSR标记的开发提供了丰富的资

收稿 2008-11-27 **修定** 2009-01-08
资助 国家"863"计划(2007AA10Z100)和河南农业大学博士启动基金(30400247)。
* 通讯作者(E-mail: lijw555@sohu.com; Tel: 0371-63554959)。

源。本文从GenBank公布的EST中查找甜瓜SSR, 分析这些SSR在甜瓜转录组中的特点和分布规律, 从而为EST-SSR标记的开发建立基础。

材料与方法

以“melon mRNA”为关键词采用FASTA格式从GenBank/dbEST (http://www.ncbi.nlm.nih.gov/entrez)下载35 547条甜瓜(*Cucumis melo* L.) EST序列(以2008年5月20日NCBI公布的数据为准), 它们主要来自甜瓜的果实、叶片、根等组织。

采用EST-trimmer (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl)、cross-match (www.phrap.org)等网络软件除去EST中重复、ployA/T“尾巴”、载体等赝象序列。用MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html)在无冗余EST中搜索SSR, 并结合手工查寻。搜索条件为: 含有单、二和三核苷酸基元的最小重复数分别为20、8和5, 四核苷酸或四核苷酸以上的SSR最小重复数均为4。分别统计甜瓜不同组织中SSR的数目和出现频率。为了便于统计和分析SSR序列, 将重复基元所有循环序列和互补序列视为一类。如ACT、CTA、TAC、TGA、GAT和ATG均归为一类。

实验结果

1 甜瓜EST-SSR的频率和分布密度

从NCBI网站下载的35 547条甜瓜EST序列, 主要来源于甜瓜的根(13 853条)、果实(10 585条)、子叶(5 664条)、叶片(3 212条)和韧皮部(1 800条), 另外还有少量来自悬浮细胞和愈伤组织。经净化处理(除去重复、ployA/T、载体等)后共获得34 438条无冗余的EST, 序列总长度为250.3 Mb。按照查找标准, 共发现2 107条至少含有1个SSR的EST序列, 占无冗余EST总数的6.12%, 表明甜瓜EST中微卫星含量较为丰富。在2 107条EST中, 含有单个SSR的EST为1 458条, 含有2个或2个以上SSR的EST为649条, 其中还有69条序列出现两个SSR串联。共检出2 813个精确重复的SSR, 占无冗余EST的8.16%, 即甜瓜基因组中EST-SSR的出现频率。

甜瓜EST-SSR的优势重复基元为单、二和三核苷酸, 三者共占EST-SSR总数的84.85%, 其中又以三核苷酸重复所占的比例最大(47.14%), 二核苷酸重复次之(20.72%), 单核苷酸重复最少(16.99%)。基元长度大于等于4的重复序列所占的比例较小, 共计15.15%。从SSR的分布密度来看, 甜瓜EST中平均每8.90 kb就出现1个SSR, 但不同重复单元出现的平均距离各不一致, EST-SSR出现的频率越高, 其平均距离越小(表1)。甜瓜不同组织中SSR的分布密度不尽相同, 韧皮部中SSR密度最大(1/6.14 kb), 子叶次之(1/8.44 kb), 果实中最少(1/10.53 kb)。

表1 甜瓜中EST-SSR的数量、比例和平均距离

Table 1 Number, percentage and mean distance of EST-SSRs in melon

重复类型	数目	所占比例 /%	出现频率 /% 1)	平均距离 /kb 2)
单核苷酸	478	16.99	1.39	52.37
二核苷酸	583	20.72	1.69	42.94
三核苷酸	1326	47.14	3.85	18.88
四核苷酸	115	4.09	0.33	217.67
五核苷酸	180	6.40	0.52	139.07
六核苷酸	131	4.66	0.38	191.08
总计	2813	100.00	8.16	8.90

1) 出现频率 = 检出的SSR数目 / 无冗余EST总数;
2) 平均距离 = 无冗余EST总长度 /SSR总数。

2 甜瓜EST-SSR中基元类型及比例

在搜索到的2 813个甜瓜EST-SSR中共观察到69种重复基元, 一至六核苷酸重复分别有2、3、10、14、22和18种。不同类型的基元的出现频率不一致, 存在明显的偏倚性。单核苷酸重复基元中A/T占绝对优势, 所占比例为16.25%, C/G极少(18个, 占0.64%); 二核苷酸重复基元中AG/CT比例最高, 达14.61%, AT次之(3.63%), AC/GT较少(2.49%), 而CG则没有出现; 三核苷酸重复基元中AAG/TTC比例高达29.26%, 除CCG/CGG外(15个, 占0.53%), 其他各类三核苷酸重复基元的比例均在1%~5%之间; 四核苷酸重复基元中比例最高的是AAAG/CTTT (1.78%); 五核苷酸重复基元AAAAG/CTTTT比例最高(2.13%); 六核苷酸中AAAAAG/CTTTTT和ACCACG/CTGGTG的比例分别为0.64%和0.60% (表2)。

表2 甜瓜EST中主要重复基元

Table 2 The major repeat motifs in melon ESTs

重复类型	重复基元	数量	所占比例/%	出现频率/%
单核苷酸	A/T	457	16.25	1.33
	C/G	18	0.64	0.05
二核苷酸	AG/CT	411	14.61	1.19
	AT/AT	102	3.63	0.30
	AC/GT	70	2.49	0.20
三核苷酸	AAG/CTT	823	29.26	2.39
	AGG/CCT	116	4.12	0.34
	AAC/GTT	97	3.45	0.28
	AGC/CGT	67	2.38	0.19
	AGT/ATC	54	1.92	0.16
	AAT/ATT	39	1.39	0.11
	ACC/GGT	40	1.42	0.12
	ACT/ATG	38	1.35	0.11
	ACG/CTG	37	1.32	0.11
	CCG/CGG	15	0.53	0.04
四核苷酸	AAAG/CTTT	50	1.78	0.15
	AAAT/ATTT	30	1.07	0.09
	AAGG/CCTT	10	0.36	0.03
	AAAC/GTTT	6	0.21	0.02
	AATG/ACTT	5	0.18	0.01
五核苷酸	AAAAG/CTTTT	60	2.13	0.17
	AAAAT/ATTTT	17	0.60	0.05
	AATAC/ATGTT	16	0.67	0.05
	AAGAG/CTCTT	15	0.53	0.04
	AAAAC/GTTTT	12	0.43	0.03
	AAAGG/CCTTT	9	0.32	0.03
	AACCT/ATTGG	9	0.32	0.03
	AATCCAGGTT	8	0.28	0.02
	AAACC/GGTTT	5	0.18	0.01
	AATTC/AAGTT	5	0.18	0.01
六核苷酸	AAAAAG/CTTTTT	18	0.64	0.05
	ACCACG/CTGGTG	17	0.60	0.05
	AAGAGG/CCTTCT	10	0.36	0.03
	AAAAAT/ATTTTT	7	0.25	0.02
	AGGTAT/ATATCC	6	0.21	0.02
	AACACG/CTTGTG	6	0.21	0.02
	AAGATG/ACTTCT	5	0.18	0.01
	AACCTC/AGTTGG	5	0.18	0.01
	AAGACG/CTTCTG	5	0.18	0.01

出现频率小于0.01%的重复基元未列出。

3 甜瓜EST-SSR重复次数和长度

SSR重复次数的变异引起位点长度的变化是产生SSR多态性的主要原因。对2813个甜瓜EST-SSR以及主要重复类型(二核苷酸和三核苷酸重复)进行分类统计, 结果表明, 随着重复次数的增加, SSR数量迅速减少(图1)。甜瓜EST-SSR按重复次数可分为3个区间, 即4次至10次重复为第一个区间, 11次至20次重复为第二区间, 20次重复以上为第三区间。统计结果发现, 甜瓜EST-SSR主要分布在第一区间, 这一区间共有1 978个微卫星, 约占全部微卫星的70.32%, 一至六核苷酸重复基元均有分布, 其中三核苷酸重复比例最大; 第二区间有437个微卫星, 约占总数的15.54%, 主要为二核苷酸和三核苷酸重复; 第三区间有398个微卫星, 主要是

单核苷酸和部分二核苷酸重复, 占14.15%。

重复基元长度变化是EST-SSR位点多态性的主要表现形式。甜瓜EST-SSR长度分布情况见图2。由于搜索标准的严格度, 部分SSR被过滤掉(特别是部分长度小于20 bp的单核苷酸重复), 因此EST-SSR长度分布是不连续的, 如长度为13 bp、17 bp和19 bp的SSR没有出现。总的来说, 大部分甜瓜EST-SSR长度集中在12~20 bp范围内(1 466个SSR, 占52.12%), 几乎全为二核苷酸和三核苷酸重复; 其次是21~30 bp (920个SSR, 占32.71%), 这一区间三核苷酸所占比例最大; 30 bp以上的SSR数量相对较少(427个SSR, 占15.18%), 主要由单

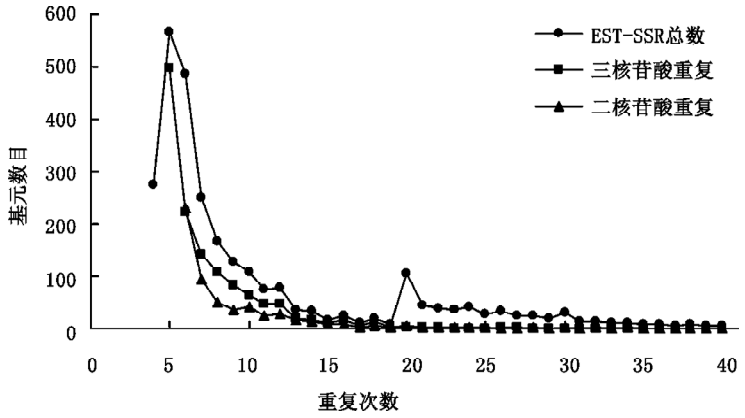

图1 甜瓜EST-SSR重复次数分布

Fig.1 Distribution of EST-SSR repeat numbers in melon

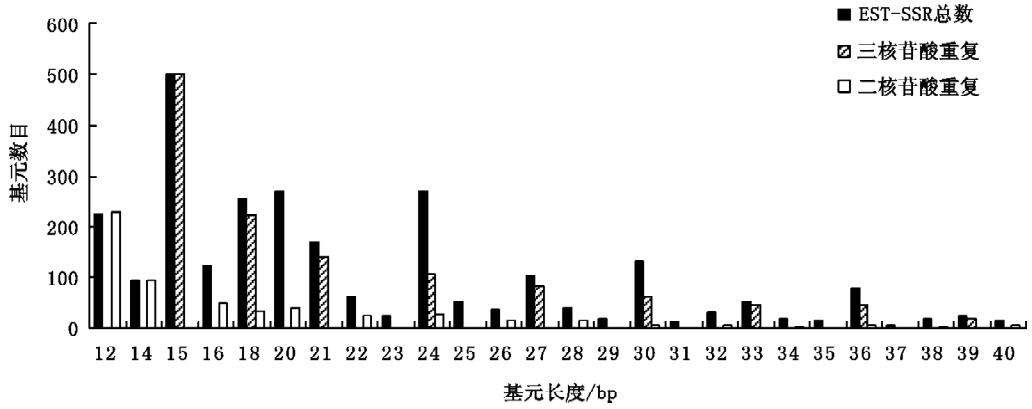

图2 甜瓜EST-SSR长度分布

Fig.2 Distribution of EST-SSR length in melon

核苷酸和三核苷酸重复组成。

讨 论

公共数据库中数量不断增加的EST信息极大地增强了SSR标记的开发能力。国际葫芦科基因组计划已全面启动, 甜瓜作为葫芦科研究作物之一, 已经产生3万多条EST序列, 这为EST-SSR标记的开发提供了宝贵的资源。但现已开发的甜瓜EST-SSR标记引物却不足30对(Danin-Poleg等2001; Kong等2007), 这与EST数量极不相配, 因此甜瓜EST-SSR标记还有进一步开发的潜力。明确甜瓜基因组中EST-SSR的分布特征是开发新的SSR标记的前提。本文发现, 甜瓜EST-SSR基元类型丰富、出现频率高、密度大, 平均距离仅为8.17 kb, 高于拟南芥(14.9 kb)、小麦(15.6 kb)、棉花(20.0 kb)和番茄(11.1 kb)等作物(李永强等2004), 这说明大规模开发EST-SSR标记的可能性较大。例如, 雷蒙德氏棉EST中SSR的平均距离是14.8

kb, 目前已从58 906条非冗余雷蒙德氏棉EST中成功开发1 554对EST-SSR引物(王长彪等2006)。Kong等(2007)从5747条甜瓜EST (2006年7月31日前NCBI公布的数据)中发现了383个SSR位点, 主要是二核苷酸重复(49.9%)和三核苷酸重复(43.6%), EST-SSR出现频率是1/4.7 kb, 高于本研究结果(1/14.8 kb), 这主要是由于搜索标准不同(最小重复数为5)所致。本文中甜瓜EST-SSR中主要重复类型为单、二和三核苷酸(约占总数的85%), 这与大多数植物中的报道相似(李永强等2004; Kumpatla和Mukhopadhyay 2005)。甜瓜EST-SSR中出现频率最高的基元是AAG/CTT (29.26%), 这与柑橘(14.2%) (Jiang等2006)、棉花(26.13%) (王长彪等2006)和拟南芥(29.00%) (Cardle等2000)中的报道相似, 进一步验证了Gao等(2003)认为AAG/TTC是双子叶植物中优势重复基元的观点。甜瓜EST中AAG/TTC的高频率出现可能与其作为三联体密码编码相应蛋白质时的高频率使用有关。

SSR位点多态性主要是因重复基元数量和基元碱基数不同所产生的简单序列长度多态性和随机扩增微卫星多态性。一般认为, SSR位点的变异频率与基元重复数存在一定的正相关, 即重复次数越多SSR产生变异的可能性越大(Schlötterer 2000)。而本文结果表明, 甜瓜70.32%的EST-SSR是低重复数基元(4至10次重复), 只有29.68%的EST-SSR重复数在10以上。从这个角度讲, 本研究所发掘的EST-SSR中, 仅有少部分具有多态性潜能。而Xu等(2000)认为微卫星变异依赖于等位基因的长度, 等位基因的重复序列扩张和收缩的频率在总体上是相等的, 即重复序列长度存在一个阈值, 长度在阈值以下的SSR倾向扩张, 而长度在阈值以上的SSR倾向收缩, 大多数真核生物阈值约为20 bp。甜瓜EST-SSR中, 长度在20 bp以下的占42.48%, 长于20 bp的占47.88%, 这些位点因未达到或超过阈值长度而倾向于扩张或收缩。因此, 按照Xu等(2000)观点, 我们所发掘的甜瓜EST-SSR位点大部分具有多态性潜能。在Kong等(2007)所开发的22对甜瓜EST-SSR引物中, 约1/4 SSR位点长度小于或等于20 bp, 但在品种间均能表现出多态性, 充分说明SSR多态性潜能与其长度(或重复次数)没有直接关系。

由于EST来源于编码区DNA, EST-SSR代表了基因表达的信息, 能为功能基因提供“绝对”标记。此外, EST-SSR标记具有物种间通用性, EST-SSR遗传作图将使物种之间连锁信息的转换更快, 实现多个图谱整合, 从而更有利于比较基因组学的研究(Lan等2000; Varshney等2005b)。现在, 依据本文结果所进行的甜瓜EST-SSR引物大规模开发, 已在本实验室全面展开。

参考文献

李小白, 崔海瑞, 张明龙(2006). EST分子标记开发及在比较基因组学中的应用. 生物多样性, 14 (6): 541~547

李永强, 李宏伟, 高丽锋, 何蓓如(2004). 基于表达序列标签的微卫星标记(EST-SSRs)研究进展. 植物遗传资源学报, 5 (1): 91~95

苏芳, 郭绍贵, 宫国义, 张海英, 许勇(2007). 甜瓜基因组学研究进展. 分子植物育种, 5 (4): 540~547

王长彪, 郭旺珍, 蔡彩平, 张天真(2006). 雷蒙德氏棉EST-SSRs分布特征及开发与利用. 科学通报, 51 (3): 316~320

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics, 156: 847~854

Danin-Poleg Y, Reis N, Tzuri G, Katzir N (2001). Development and characterization of microsatellite markers in *Cucumis*. Theor Appl Genet, 102: 61~72

Gao LF, Tang JF, Li HW, Jia JZ (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol Breed, 12: 245~261

Jiang D, Zhong GY, Hong QB (2006). Analysis of microsatellites in citrus unigenes. Acta Genet Sin, 33 (4): 345~353

Kong Q, Xiang C, Yu Z, Zhang C, Liu F, Peng C, Peng X (2007). Mining and charactering microsatellites in *Cucumis melo* expressed sequence tags from sequence database. Mol Ecol Notes, 7: 281~283

Kumpatla SP, Mukhopadhyay S (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome, 48: 985~998

Lan TH, DelMonte TA, Reischmann KP, Hyman J, Kowalski SP, McFerson J, Kresovich S, Paterson AH (2000). An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res, 10: 776~788

Powell W, Machray GC, Provan J (1996). Polymorphism revealed by simple sequence repeats. Trends Plant Sci, 1: 215~222

Schlötterer C (2000). Evolutionary dynamics of microsatellite DNA. Chromosoma, 109: 365~371

Varshney RK, Graner A, Sorrells ME (2005a). Genic microsatellite markers in plants: features and applications. Trends Biotechnol, 23: 48~55

Varshney RK, Sigmund R, Börner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A (2005b). Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Sci, 168: 195~202

Xu X, Peng M, Fang Z, Xu X (2000). The direction of microsatellite mutations is dependent upon allele length. Nat Genet, 24: 396~399