

花生 EST—SSR 分子标记的开发

王金彦, 潘丽娟, 杨庆利, 禹山林

(山东省花生研究所, 山东 青岛 266100)

摘要: 利用 NCBI 的 Genbank 数据库中公布的花生 53 177 条 EST 序列以及本实验室创造的花生栽培品种 E12 (*Arachis hypogaea* L.) 所构建的 Unigene 文库中的 4 074 条 EST 序列, 对这些序列进行前期处理(去除冗余序列, 对含有重叠区域的 EST 序列进行拼接), 总共获得非冗余且拼接较长的序列 11 260 条。通过软件分析发现两个 EST 库中共包含有 1 323 个 SSR 位点, 主要是 2 个和 3 个核苷酸重复, 除此之外也有少量的 4 核苷酸重复以及复合重复。这些 EST—SSR 平均长度为 18.88 bp, 平均每 8.5 条 EST 序列就包含有一个 SSR 位点。其中 AG/TC、CTT/GAA 重复出现的频率最大, 分别占到 2 个核苷酸重复和 3 个核苷酸重复的 39.3% 和 22.6%。

关键词: 花生; EST; SSR; 频率; 特点

中图分类号: S565.03 文献标识码: A 文章编号: 1000—7091(2009)增刊—0042—04

Development of EST—SSR Markers in Peanut (*Arachis hypogaea* L.)

WANG Jin—yan, PAN Li—juan, YANG Qing—li, YU Shan—lin

(Peanut Research Institute, Shandong Academy of Agriculture Sciences, Qingdao 266100, China)

Abstract: 53 177 ESTs from Genbank database in NCBI and 4 074 ESTs from Unigene library of E12 (*Arachis hypogaea* L.) in our laboratory are analysed. Totally 11 260 non—redundant ESTs contigs are detected. 1 323 SSR loci are searched by software analysis in NCBI database and Unigene library, mainly di— and tri—nucleotide repeat motif except of few tetra—nucleotide repeat motif and compound repeat. In these EST—SSR loci, average length is 18.88 bp and average 8.5 ESTs contain one SSR loci. The AG/TC and CTT/GAA motif are most frequent types, accounting for 39.3% and 22.6% in dinucleotide and trinucleotide repeats, respectively.

Key words: Peanut; EST; SSR; Frequency; Characteristics

花生是我国重要的油料作物和经济作物。近年来, 国内外在花生 (*Arachis hypogaea* L., $2n = 40$, AABB) 分子领域的研究已经悄然兴起。目前, 在花生中可以利用的标记有 RFLP、SSR、SCAR 等。其中以 SSR 标记应用最为广泛。SSR (Simple sequence repeat) 分子标记又称为简单序列重复或者微卫星 DNA (Microsatellite DNA), 是指以少数几个核苷酸 (2~6 个) 为单位串联重复的 DNA 序列, 通常为 2~3 个碱基。该标记是基于 PCR 技术产生的, 具有多态性高、共显性、重复性好等特点^[1]。SSR 标记在遗传图谱构建、分子标记辅助选择、遗传多样性分析、种质资源的鉴定等方面都有着广泛的应用^[2-4]。

EST (Expressed sequence tags), 又称为表达序列

标签, 是通过 cDNA 克隆的 5 端和 3 端测序所获得的序列信息, 这些长约 150~500 bp 的序列来自于基因的表达式产物^[5]。目前, 花生 EST 计划虽然起步较晚, 但随着 EST 数量的迅速发展, 截止到 2008 年 11 月 7 日, NCBI 的 Genbank 数据库中的花生 EST 数目已经达到了 53 177 条, 这些 EST 资源丰富了花生的基因组序列信息。EST—SSR 是通过 EST 序列进行分析, 找到含有重复单元的位置, 在两侧设计引物而开发得到的。因此, EST—SSR 反映基因的编码部分, 可直接获得基因的表达信息, 并可对一些重要性状进行直接鉴定等。本研究利用 NCBI 中公布的花生 EST 序列信息以及本实验室建立的花生 Unigene 文库中的 EST 序列信息, 对所包含的 SSR 位点进行

收稿日期: 2009—05—03

基金项目: 国家重点基础研究发展计划(2007CB116200); 国家高技术研究发展计划(2006AA10A114; 2007AA10Z189); 农业公益性行业科研专项(NYHYZX07—14)

作者简介: 王金彦(1982—), 男, 山东青岛人, 硕士, 主要从事花生分子生物学方面研究。

通讯作者: 禹山林(1956—), 男, 山东莱州人, 研究员, 主要从事花生遗传育种方面科研工作。

了分析,该研究为丰富花生 SSR 分子标记以及为育种的应用奠定基础。

1 材料和方法

1.1 花生 EST 序列来源

来自于 NCBI 的 Genbank 数据库中的 53 177 条序列 (<http://www.ncbi.nlm.nih.gov/>)以及本实验室利用花生栽培品种 E12 所构建的 Unigene 文库中的 4 074 条 EST 序列,共计 57 253 条。

1.2 EST 序列的前期处理

采用 EST—trimmer 软件 (<http://pgrc.jp/gatersleben.de/misa/download/est-trimmer.pl>)去除 3' 端的 PolyA 结构以及载体序列,去除小于 100 bp 的低质量 EST 序列,并对大于 700 bp 的 EST 序列的 3' 端进行截短处理以防止存在质量问题。

1.3 去除冗余、拼接 EST 序列

利用 StackPACK2.0 系统^[9] (<http://www.eugenetic.com>)对序列进行同源聚类以降低冗余度^[7]。对 EST 序列进行网上 BLAST 检索,获得其同源序列,再利用 CAP3 (<http://genome.cs.mtu.edu/sas.html>)软件将同源序列进行重叠群拼接,以延伸 EST 序列或获得较长 cDNA。

1.4 SSR 位点的筛选

利用 MISA 软件 (<http://pgrc.ipk-gatersleben.de/misa/>)对拼接后的 EST 序列进行 SSR 搜索,查找以 2, 3, 4, 5, 6 个核苷酸为重复单元的 SSR 基序。SSR 查找标准为不同重复基序的 SSR,其重复序列总长度不低于 18 个核苷酸。同时也筛选一些被若干碱基序列打断,但同时包含两种核苷酸重复的复

合 SSR。

1.5 SSR 引物的设计

利用 Primer3 软件 (<http://frodo.wi.mit.edu/cgi-bin/primer3/primer3-www.cgi>)对所获得的包含有 SSR 位点的 EST 序列进行引物设计,设计的条件为碱基数目在 18~25 个核苷酸,GC 含量在 40%~60%,退火温度在 55~65℃。

2 结果和分析

2.1 EST 序列的处理

对 NCBI 的 Genbank 数据库以及 Unigene 文库中总共 57 253 条序列进行前期处理,发现 NCBI 数据库中的 53 177 条 EST 序列经过去除冗余并拼接后,得到 7 270 条 EST 序列,序列重复性为 86.3%。在本实验室构建的 Unigene 文库中,经过前处理得到了 3 990 条无冗余的 EST 序列,重复性仅为 2%。这表明在 NCBI 中花生 EST 序列存在着大量的重复,而本实验室所构建的 EST 数据库重复性较低。

2.2 SSR 位点在 EST 中出现的频率

对两个库中总共 11 260 条 EST 序列进行 SSR 位点的搜索,在 NCBI 数据库中,发现有 437 个 SSR 位点,占总序列的 6.0%;在 Unigene 文库中,共搜索到 641 个 SSR 位点,占总序列的 16.1%。这两个库中共发现 1 323 个 SSR 位点,占总拼接后序列的 11.7%。其中在 NCBI 数据库中,除了 24 条 EST 存在两个 SSR 位点,3 条 EST 存在 3 个 SSR 位点之外,其余 EST 均只包含一个 SSR 位点。在 Unigene 库中,全部的 641 个 SSR 位点各自来自于一条序列。从长度来看,两个库的 SSR 平均为 18.30 bp 和 19.47 bp (表 1)。

表 1 SSR 位点的分布

Tab. 1 Distribution of SSR loci

	NCBI 数据库 NCBI database	Unigene 文库 Unigene library
SSR 数目 Number of SSR loci	437	641
含有单个 SSR 位点的序列数 Number of sequences including one SSR loci	410	641
含有两个 SSR 位点的序列数 Number of sequences including two SSR loci	24	0
含有三个 SSR 位点的序列数 Number of sequences including three SSR loci	3	0
平均长度/bp Average length	18.30	19.47

2.3 花生 EST—SSR 的特点

在 NCBI 数据库中,以 3 个核苷酸重复出现的次数最多,占到了总数的 46.3%,其次是 2 个核苷酸和复合核苷酸重复,分别有 30.7%和 16.2%。在 Unigene 基因文库中,3 个核苷酸出现的最多,占到总数的 56.8%,其次 2 个核苷酸重复,占到总数的 27.1%,复合重复类型也有 53 个,占到总数的 8.3% (表 2)。

在这些 EST—SSR 中,除去复合重复之外,总共有 122 种不同类型的重复单元,其中 2 个核苷酸重复有 10 种,3 个核苷酸重复有 51 种,4 个核苷酸重复的有 13 种,5 个核苷酸重复有 23 种,6 个核苷酸重复的有 25 种。从出现的频率来看,两个库中都是 AG/TC 出现的频率最大,分别占到总共 SSR 位点的 11.2%和 13.2%。在 NCBI 数据库中,CT/GA 的数量仅次于 AG/TC,为 10.07%,其次为 AT/TC 和

CTT/GAA。在 Unigene 文库中, CTT/GAA 重复仅比 AG/TC 重复少 2 次, 占总 SSR 位点的 11.9%(图 1)。在 4 个核苷酸重复中, 两个库中 AAAG/TTTC 占的比例最大, 而在 5 个核苷酸中则是 AAAAG/TTTTC 重复的数量最多。

表 2 EST—SSR 出现的频率

Tab. 2 Occurrence of EST—SSR

重复基序类型 Type of repeat motif	NCBI 数据库 NCBI database		Unigene 文库 Unigene library	
	数目 Number	频率/% Frequency	数目 Number	频率/% Frequency
2个核苷酸 Dinucleotide	134	30.7	174	27.1
3个核苷酸 Trinucleotide	207	47.3	364	56.8
4个核苷酸 Tetranucleotide	7	1.6	10	1.6
5个核苷酸 Pentanucleotide	17	3.9	16	2.5
6个核苷酸 Hexanucleotide	1	0.2	24	3.7
复合 Compound	71	16.2	53	8.3
总计 Total	437	100	641	100

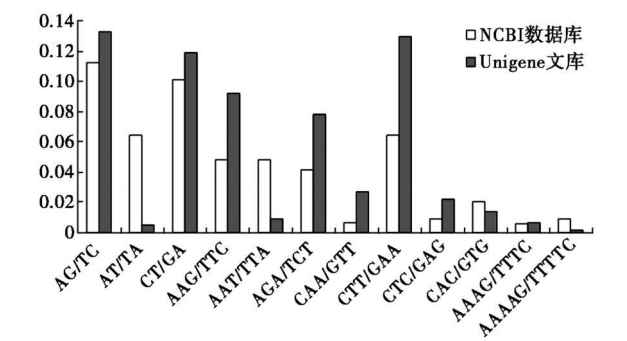


图 1 EST—SSR 在两个库中主要重复基序的分布频率

Fig. 1 Frequency distribution of EST—SSR based on main repeat motif in two libraries

3 讨论

近年来, 花生 SSR 分子标记的研究虽然已经取得了一定的进展, 但目前所利用的 SSR 标记还远远满足不了花生分子生物学研究的需要。SSR 分子标记可以分为基因组 SSR 标记和 EST—SSR 分子标记。基因组 SSR 分子标记的开发是建立在基因组文库开发的基础上, 这种方法耗时、费力、花费较大。EST—SSR 分子标记则是利用目前日益增多的花生 EST 序列资源开发得到的, 具有省时、简便、成本较低的特点, 而且 EST 本身是基因的一部分, 其变异可能直接与基因的功能改变相关。

在 NCBI 数据库中, 共获得了 437 个 SSR, 占非冗余序列的 6.0%, 在 Unigene 基因文库中, 共获得

了 641 个 SSR, 占非冗余序列的 16.1%。尽管我们从 NCBI 中获得的无冗余序列要比 Unigene 文库中要多, 但所获得的含有 SSR 位点的序列却不及 Unigene 文库。这可能是由于在 NCBI 数据库中, 大多数序列都是由较短的并且重复性较大的序列组成, 而在 Unigene 文库中, 由于重复性较少, 而且序列也较长, 因此含有的 SSR 位点的数目也较多。这个比率不同于小麦 (2.1%)^[8]、棉花 (7.15%)^[9]、油菜 (13.58%)^[10]、芝麻 (8.68%)^[11] 等物种中筛选出来的 EST—SSR。这些差异可能是由于物种之间序列的差别所造成的。Luo 等^[12] 开发的花生 EST—SSR 标记中所用的标准为 2 个核苷酸重复的次数不少于 7 次, 3 个核苷酸重复的次数不少于 5 次, 4 个核苷酸重复的次数不少于 4 次, 通过筛选, 总共在 1, 345 条 EST 中开发了 44 个 EST—SSR 引物, 仅占到全部序列的 3.27%。

EST 序列来自于基因的表达产物, 在品种间基因的序列相对保守, 而基因组 SSR 则来自于整个基因组, 包含着非编码区域, 因此 EST—SSR 标记的多态率要低于基因组 SSR 标记, 这在一定程度上造成了 EST—SSR 在品种间多态率较低。He 等^[13] 通过技术优化开发得到了 56 对基因组 SSR 标记, 其中有 19 对在不同的花生基因型中表现多态, 多态率为 33.9%。Ferguson 等^[14] 在建立的两个栽培花生基因组文库中, 总共获得了 348 个 SSR 标记, 其中有 110 个标记在 24 个花生地方品种间出现多态, 多态率为 31.6%。Moretzsohn 等^[1] 通过构建栽培品种 UF91108 基因组文库开发了 159 对 SSR 标记, 其中有 66 对在 2 个野生品种间表现多态, 57 对在 6 个栽培品种间表现多态。

由于 EST—SSR 多态性较低, 因此可以采用其他方法来寻找多态性的标记。根据 EST 5 端或 3 端的非编码区开发的标记多态性要稍好些^[15], 因此, 在设计引物时, 应尽量使引物靠近 3 或 5 端非编码区。为了提高 EST 标记的多态性, 还可以对无多态扩增产物进行酶切, 以揭示酶切扩增多态性 (Cleavage amplified polymorphism, CAPs); 或改进对 PCR 产物的分析手段, 如采用分辨率较高的变性梯度胶分离 PCR 产物, 提高多态检测率^[16]。如在洋葱^[17]、白云杉等^[18] 都建立 EST—SSCP (EST 单链构象多态性) 检测分离技术。

在 NCBI 数据库和 Unigene 基因文库中, 都是 2 个和 3 个核苷酸重复的 EST 数目最多, 这也表明这两种重复的出现的频率在两个库中较大。从重复类型来看, AG/TC、CT/GA 和 CTT/GAA 在两个库中出

现的频率较大。目前,花生基因组 SSR 分子标记较多,但 SSR 重复单元在不同的研究也各不相同。在 Ferguson 等^[14]在所获得的 348 个花生基因组 SSR 标记中,ATT 和 GA 重复单元的最多,分别占到了 29% 和 28%。Moretzsohn 等^[1]在用于分析栽培花生品种遗传多样性的基因组 SSR 中,大多数都是 AAC 和 GA 基序的重复。Cuc 等^[4]在栽培花生品种 TMV2 构建的基因组分库中开发了 170 对 SSR 标记,有 104 对标记能扩增出清晰的条带,其中 GT/CA 的重复单元出现的次数最多,占到了 37.6%,其次是 GA/CT 重复,占总数的 25.9%。Luo 等^[12]在获得的花生 EST-SSR 中发现 AT 和 AAT 重复出现的次数最多。本研究所获得的 1 323 条 EST-SSR 中,以 3 核苷酸重复出现的次数最多,而 AG/TC、CT/GA、CTT/GAA 重复则是出现较多的类型。由于基因组 SSR 的筛选有赖于基因组文库的构建以及探针的筛选,因此不同研究者探针的选取也造成了基序重复单元的差异。

花生 EST 序列不仅可以为花生功能基因组的研究创造可以利用的资源,而且为分子标记的开发提供了丰富的序列信息。本研究利用 NCBI 数据库以及本实验室创造的 Unigene 文库中的 EST 序列,寻找到大量的含有 SSR 位点的 EST 序列,将这些序列开发成 EST-SSR 标记,不仅可以丰富花生分子标记类型,而且在花生遗传图谱构建、遗传多样性分析、分子标记辅助选择等方面都有极其重要的意义。

参考文献:

- [1] Moretzsohn M C, Hopkins M S, Mitchell S E, *et al.* Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome[J]. BMC Plant Biol, 2004, 4: 11—20.
- [2] Moretzsohn M C, Leoi L, Proite K, *et al.* A microsatellite-based gene-rich linkage map for the AA genome of *Arachis* (*Fabaceae*)[J]. Theor Appl Genet, 2005, 111: 1060—1071.
- [3] 唐荣华, 贺梁琼, 高国庆, 等. 多粒型花生的 SSR 分子标记[J]. 花生学报, 2004, 33(2): 11—16.
- [4] Cuc L M, Mace E S, Crouch J H, *et al.* Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (*Arachis hypogaea* L.)[J]. BMC Plant Biol, 2008, 8: 55—65.
- [5] Adams M D, Kelly J M, Gocayne J D, *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project[J]. Science, 1991, 252: 1651—1656.
- [6] Miller R T, Christoffels A G, Gopalakrishnan C, *et al.* A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base[J]. Genome Res, 1999, 9: 1143—1155.
- [7] Kantety R V, La Rota M, Matthews D E, *et al.* Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat[J]. Plant Mol Biol, 2002, 48: 501—510.
- [8] Yu Ju-Kyung, Dake T M, Singh S, *et al.* Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat[J]. Genome 47: 805—818.
- [9] Han Z, Wang C, Song X, *et al.* Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton[J]. Theor Appl Genet, 2006, 112(3): 430—439.
- [10] 李小白, 张明龙, 崔海瑞. 油菜 EST 资源的 SSR 信息分析[J]. 中国油料作物学报, 2007, 29(1): 20—35.
- [11] 魏利斌, 张海洋, 郑永战, 等. 芝麻 EST-SSR 标记的开发和初步研究[J]. 作物学报, 2008, 34(12): 2077—2084.
- [12] Luo M, Dang P, Guo B Z, *et al.* Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut[J]. Crop science, 2005, 45: 346—353.
- [13] He G H, Meng R H, Newman M, *et al.* Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.) [J]. BMC Plant Biol, 2003, 3(3): 1—6.
- [14] Ferguson M E, Burow M D, Schulze S R, *et al.* Microsatellite identification and characterization in peanut (*Arachis hypogaea* L.) [J]. Theor Appl Genet, 2004, 108: 1064—1070.
- [15] Thiel T, Michalek W, Varshney R K, *et al.* Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (*Hordeum vulgare* L.) [J]. Theor Appl Genet, 2003, 106: 411—422.
- [16] Temesgen B, Brown G R, Harry D E. Genetic mapping of expressed sequence tag polymorphism (ESTP) marker in loblolly (*Pinus taeda* L.) [J]. Theor Appl Genet, 2001, 102: 664—675.
- [17] McCallum J, Leite D, Pither-Joyce M, *et al.* Expressed sequence markers for genetic analysis of bulb onion (*Allium cepa* L.) [J]. Theor App Genet, 2001, 103: 979—991.
- [18] Gosselin I, Zhou Y, Bousquet J, *et al.* Megagametophyte-derived linkage maps of white spruce (*Picea glauca*) based on RAPD, SCAR and ESTP markers[J]. Theor App Genet, 2002, 104: 987—997.