植物遗传资源学报 2010, 11(6): 772-776
Journal of Plant Genetic Resources

芸薹属作物 EST-SNP的发掘与分析

李雪姣[1],张 耿[2],顾爱侠[1],轩淑欣[1],王彦华[1],赵建军[1]
([1]河北农业大学园艺学院,保定 071000; [2]浙江大学农业与生物技术学院作物所,杭州 310029)

摘要:芸薹属作物是十字花科中重要的蔬菜和油用作物,经济价值和食用价值比较高。本研究通过生物信息学手段,从 3 个芸薹属物种 EST序列中获得了大量的 SNP和 Indel资源,对 SNP与 EST的数量及拼接结果之间的关系、SNP碱基置换的偏好性、Indel碱基数量与发生频率之间的关系进行了分析。分析结果对深入研究芸薹属基因组特点具有一定的参考价值,为进一步通过试验验证预测的 SNP位点、开发该作物高通量 SNP分子标记奠定了基础。

关键词:芸薹属;序列表达标签(EST);单核苷酸多态性(SNP);分子标记

Detection and Analysis EST-SNP in *Brassica*

LI Xue-jiao[1], ZHANG Geng[2], GU Ai-xia[1], XUAN Shu-xin[1], WANG Yan-hua[1], ZHAO Jian-jun[1]
([1] *College of Horticulture, Agricultural University of Hebei, Baoding 071000;* [2] *College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029*)

Abstract: As one of vegetable and oil crops in Cruciferae, genus *Brassica* has highly economic and edible values. In this research a number of SNPs and indels were obtained from EST sequences of three *Brassica* species by bioinformatic tools. The relationship between spelling results and the number of SNP and EST, preference of nucleotide replacement, and the relationship between amount and frequency of nucleotides were further analyzed. The results will be useful to gain insight into the characteristics of *Brassica* genome. The information obtained in the present study is of great importance to test SNP polymorphism experimentally and to develop high throughout SNP molecular markers in *Brassica*.

Key words: *Brassica*; Expressed sequence tags; Single nucleotide polymorphism; Molecular marker

单核苷酸多态性(Single nucleotide polymorphisms, SNP)是目前基因组中数量最丰富的 DNA多态性类型[1]。SNP提供了一种重要的分子标记资源,可以用于遗传作图、图位克隆、QTL定位以及确定个体间遗传距离等,并且很快成为农业研究中所选择的标记,尤其是用于高通量的分子标记辅助育种[2]。此外,SNP低突变率的特性,使其可以作为研究复杂遗传性状以及基因组进化的工具[3]。但是,作为主要的分子标记类型之一,前期的测序成本是 SNP相关标记开发的主要限制因素。利用已有的序列数据,结合生物信息学方法,进行 SNP位点的查找,再利用试验加以验证,可以有效降低 SNP开发的成本[4]。

序列表达标签(Expressed sequence tags, EST)是识别转录区域多态性的重要资源。公共数据库平台中的 DNA序列信息,EST数据增长速度最快,EST数据内在的冗余性使其成为潜在的 SNP位点探测资源[5]。而且,EST序列中的 SNP资源极其丰富。例如:在人类基因组中,预测每 1.3kb就有一个多态性位点[6];在栽培番茄中,每 7kb就有一个 SNP位点[7]。

因此,在未获得全基因组序列的动植物物种中,发掘 EST-SNP位点就成为一项很有意义的工作[8]。因为 EST-SNP的优势在于可以评估等位基因频率,分析目标性状和标记的关联性,与串联重复序列标记相比更加稳定。近年来,发掘高通量 EST-SNP位

收稿日期: 2010-01-26 **修回日期**: 2010-05-28
基金项目:国家自然科学基金(30871713);教育部重点项目(2009);国家科技支撑子课题(2009BADB8B03-1)
作者简介:李雪姣,硕士研究生,主要从事蔬菜生物技术与遗传育种研究。E-mail: lilinghan2002she@163.com
通讯作者:赵建军,研究员。E-mail: jjz1971@yahoo.com.cn

点的技术发展迅速[9-11]。目前已有多种方法和软件有效分析 EST序列的聚类、拼接、SNP的探测以及数据的储存、分析和可视化等步骤,并已应用到多个物种中[12-18]。

芸薹属(*Brassica*)是十字花科中经济价值最大的一个属,包含了油料作物、蔬菜、饲料以及调味品作物等。芸薹属作物包括 3个基本的二倍体种:白菜型油菜 *B. rapa* (AA, n = 10)、甘蓝 *B. oleracea* (CC, n = 9)、黑芥 *B. nigra* (BB, n = 8),以及 3个复合种:甘蓝型油菜 *B. napus* (AACC, n = 19)、芥菜 *B. juncea* (AABB, n = 18)、埃塞俄比亚芥 *B. carinata* (BBCC, n = 17)。甘蓝型油菜、白菜型油菜和甘蓝在我国栽培面积大,经济价值和食用价值高[19]。白菜型油菜和甘蓝还包含了多种蔬菜,且这些蔬菜富含纤维、维他命 C等有益健康的物质[20]。此外,芸薹属作物与模式植物拟南芥同属于十字花科,是比较基因组学研究中很有价值的材料体系。因此,针对芸薹属作物进行 EST-SNP的挖掘与分析,无论是在分子育种还是基因组进化研究方面,均有较高的参考价值。

1　材料与方法

1.1　EST序列的获得

芸薹属的 EST序列从美国国立生物技术信息中心(NCBI, The National Center for Biotechnology Information)的 GenBank数据库[21]的 FTP站点(ftp://ftp.ncbi.nih.gov/genbank/)获得,版本号为 169.0。从中获得了甘蓝型油菜、白菜型油菜和甘蓝相关的 EST序列。

1.2　EST序列的聚类和拼接

参考 PlantGDB(http://www.plantgdb.org/)拼接 PUT(PlantGDB-assembled Unique Transcripts)序列[22]的方法,对 EST序列进行聚类和拼接处理。为了保证 SNP位点预测的准确性,必须对 EST序列进行处理,使用 Vmatch(http://www.vmatch.de/)软件,将 EST序列与克隆载体、细菌以及重复元件数据库(表 1)进行比较,剔除被污染的序列以及冗余的重复元件。使用 PaCE[23]对过滤后的序列进行初步的聚类,再使用 CAP3[24]完成序列的聚类和拼接工作,要求序列相似度大于 95%,长度超过 100bp。

表 1　用于 EST序列分析的数据库资源

Table 1　The database resource for analysis of EST sequences

类型 Database type	名称 Name	下载地址 URL
克隆载体序列	UniVec	ftp://ftp.ncbi.nih.gov/pub/UniVec/
大肠杆菌基因组序列	*E. coli* genome sequence	ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/
植物重复元件数据库	Plant Repeat Databases	ftp://ftp.plantbiology.msu.edu/pub/data/TIGR_Plant_Repeats/

1.3　SNP和 Indel资源的发掘

利用 QualitySNP软件包[25]进行芸薹属 EST-SNP的发掘,选择 QualitySNP软件包的原因是由于该软件无需序列的质量信息,并且对多倍体有很好的支持。从 CAP3拼接结果中,选择包含 4条及 4条以上 EST序列的 contig用于 SNP和 Indel的发掘,核心命令为"QualitySNP 2 30 0.75 0.8 0.2 0.5 2",结合软件包其他程序,将结果导入到 MySQL(http://www.mysql.com/)数据库中,对软件包内提供的 PHP(http://www.php.net/)代码进行修改,建立 SNP检索系统供他人使用。并使用自行编写的 PERL程序脚本,对结果进行整理归纳和分析。

2　结果与分析

2.1　EST拼接结果

NCBI的 EST数据库中,甘蓝型油菜 EST序列数量最多,为 567177条,而甘蓝 EST序列数量很少,只有 26692条(表 2)。

聚类和拼接的分析结果表明,甘蓝型油菜、白菜型油菜和甘蓝中用于拼接的 EST数目,分别为 364964(64.35%)、107615(58.84%)和 19882(74.48%),其中分别有 30.54%、29.59%和 58.09%的 EST序列所组成 contig规模少于 4条。可以看出,对于 EST-SNP位点的检测,需要大量冗余的 EST序列作为基础,如果 EST数量较少,得到的结果则不甚理想。

表 2 芸薹属 EST-SNP和 EST-InDel发掘概况

Table 2 The survey of EST-SNP and EST-InDel in Brassica

	甘蓝型油菜 *B. napus*	白菜型油菜 *B. rapa*	甘蓝 *B. olereacea*
EST总量 Total EST	567177	182890	26692
用于拼接的 EST数量 Assembled EST	364964	107615	19882
Contig数量 Contig number	44381	16726	2956
包含 4条或以上序列的 Contig size≥4	18836	6619	667
包含潜在 SNP位点的 Candidate SNP contig	3261	2276	100
SNP总数 Total SNP	20830	10111	847
Indel总数 Total indel	1337	847	70

2.2 EST-SNP位点分析

EST-SNP位点的分析结果表明 (表 2),甘蓝型油菜 SNP总数最多 (20830个),是白菜型油菜的 2.06倍,是甘蓝的 24.59倍。

对 SNP发掘效率与 contig规模 (contig包含 EST数量)之间的关系进行了分析,在包含 4条及 4条以上 EST序列的 contig中,包含潜在 SNP位点的 contig比例为:甘蓝型油菜 17.31%,白菜型油菜 34.39%,甘蓝 14.99%。包含潜在 SNP的 contig中,平均每个 contig拥有的 SNP位点数目为:甘蓝型油菜 6.39个,白菜型油菜 4.44个,甘蓝 8.47个。

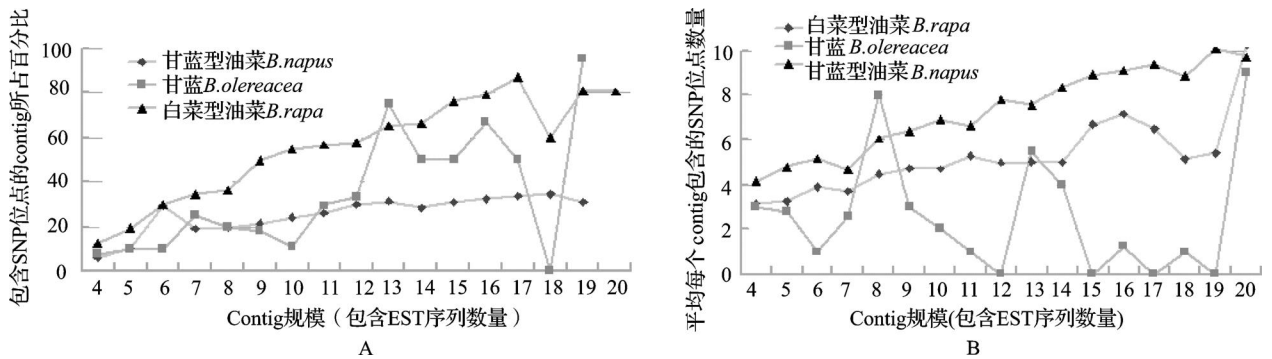

图 1 SNP位点与 contig规模之间的关系

Fig. 1 The relationship between SNP productivity and contig size

甘蓝的 EST序列很少,导致随机误差增大。从甘蓝型油菜和白菜型油菜的统计数据可清楚看出,随着 contig规模的增大 (所包含 EST数目的增加),包含潜在 SNP的 contig所占比例显著上升 (图 1A),contig包含的 SNP数量也大幅增加 (图 1B)。但是,在超过 20条以上 EST序列的 contig中,所包含的 SNP数量与 contig规模并不相称。推测原因是由于 EST为单向测序,包含有测序产生的错误,对序列联配和 SNP位点的查找有影响作用。当 contig包含 EST序列数量过多时,这种误差的积累极大地干扰了 SNP位点的查找,导致 SNP数量降低。

2.3 碱基置换类型分析

针对 SNP的碱基置换类型进行了统计分析 (表 3)。可以看出,芸薹属 3个物种的碱基置换存在明显且一致的偏好性,C/T和 A/G置换是最主要的两种类型,占到总量的一半左右。3种芸薹属作物的 G/C置换类型最少,小于总量的 10%。

2.4 插入 缺失分析

除了分析 SNP位点,还对插入 缺失 (Indel: insertion/deletion)进行了统计 (表 4)。可以看出,Indel的数量与 Indel碱基数成反比。但是,甘蓝型油菜的 6、9碱基 Indel以及白菜型油菜的 3、6碱基 Indel比例与整体趋势相反,可能是由于 3/6/9是密码子的整数倍,该类型的 Indel对基因编码蛋白影响相对较小,更容易在进化选择的过程中被保留下来。另外,芸薹属 EST序列的单碱基 Indel也存在一致的偏好性。甘蓝型油菜单碱基 Indel中,有 306个 (38.01%)是 A,280个 (34.78%)是 T;白菜型油菜中分别有 190个 (34.48%)A和 171个 (31.03%)T;甘蓝的单碱基 Indel没有明显的规律。

表 3　SNP碱基置换类型统计

Table 3　Nucleotide substitution frequencies of SNP

SNP类型 SNP type	甘蓝型油菜 *B. napus*		甘蓝 *B. olereacea*		白菜型油菜 *B. rapa*	
	数目 Amount	比例(%) Proportion	数目 Amount	比例(%) Proportion	数目 Amount	比例(%) Proportion
C/T	6702	32.17	59	28.50	3714	36.73
A/G	5246	25.18	38	18.36	2297	22.72
A/C	2248	10.79	30	14.49	1039	10.28
G/T	2260	10.85	32	15.46	1020	10.09
A/T	2329	11.18	26	12.56	1111	10.99
G/C	2045	9.82	22	10.63	930	9.20
总数 Total	20830		207		10111	

表 4　Indel类型统计

Table 4　Survey of indel in Brassica EST data

Indel碱基数 Indel size	甘蓝型油菜 *B. napus*		白菜型油菜 *B. rapa*		甘蓝 *B. olereacea*	
	数目 Amount	比例(%) Proportion	数目 Amount	比例(%) Proportion	数目 Amount	比例(%) Proportion
1	805	60.21	551	65.05	55	78.57
2	202	15.11	104	12.28	8	11.43
3	172	12.86	119	14.05	5	7.14
4	47	3.52	22	2.60	2	2.86
5	35	2.62	7	0.83		
6	37	2.77	24	2.83		
7	9	0.67	6	0.71		
8	5	0.37	5	0.59		
9	13	1.61	4	0.47		
10	1	0.07				
11	4	0.30	3	0.35		
12	4	0.30	2	0.24		
16	2	0.15				
18	1	0.07				
总数 Total	1337		847		70	

3　讨论

SNP是分子遗传分析所选择的高通量分子标记类型之一,SNP标记的开发需要高额的投入以及大量的时间。但是如果以大量的冗余 EST序列为基础,结合生物信息学手段,EST-SNP就成为一种非常廉价和高效的方法[26-28]。当然,EST-SNP的发掘也受到很多因素的限制,比如 EST序列单向测序导致的低质量,进而导致 SNP位点预测查找错误;EST序列的来源对 SNP位点的查找也有很大的影响;水平同源基因也对 SNP的预测有一定影响。但是,可以通过方法的改进,进行更准确和高效的 EST-SNP挖掘。

通过 EST总量可以看出,人们对芸薹属几个物种的关注度存在巨大差异,作为世界主要油料作物的甘蓝型油菜,受到的关注度最高、相关投入最大、获得的 EST序列数量也最多。而甘蓝有关基因和基因组研究的投入则相对较少、获得的 EST序列数量很少,但从另一个角度来说,甘蓝基因组相关研究的发展空间更大。而且,每个物种的 EST大部分集中在少数的栽培类型中,如白菜型油菜主要是大白菜(*B. rapa* subsp. *pekinensis*)的 EST数据,针对每个物种的不同栽培型开发 EST序列更有必要。

为了保证 SNP位点的准确性,满足 SNP位点查找的 contig必须是包含 4条以上(含 4条)的 EST序列,这势必会有大量的 EST序列被舍弃。但是,当

EST序列相对较少的时候,会有大量的 EST序列无法被利用,例如甘蓝利用 19882条序列拼接了 2956个 contig,但是其中 77.44%的 contig和 58.09%的 EST序列无法用于 SNP位点的查找。当 EST数目积累到一定程度后,利用率大大增加。

SNP位点发掘的同时,还进行了 EST-Indel资源的挖掘。由于 Indel突变对基因编码蛋白的影响大于 SNP突变,所以相对 SNP位点数量而言, Indel数目较少,但是仍获得了可观的数目,并且这些 Indel具有转化为分子标记的潜力,对于遗传变异、基因功能等方面的研究,也有较高的参考价值。此外,碱基数为 3/6/9的 Indel所占比例偏高的特性,与 Batley等[4]以及 Bhattramakki等[29]的发现相似,他们还发现了 12和 15碱基 Indel所占比例高的现象。出现该现象的原因可能是由于这些类型的 Indel对基因编码蛋白影响相对较小,更容易在进化选择的过程中被保留下来。

对芸薹属的几个有代表的物种甘蓝型油菜、白菜型油菜以及甘蓝进行了 EST-SNP位点以及 Indel的查找,发掘出了大量的 SNP位点和 Indel,并对其变异规律进行了初步的分析。这些资源对更加深入地研究芸薹属植物的基因组特点具有一定的参考价值。

参考文献

[1] Cho R J, Mindrinos M, Richards D R, et al. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana* [J]. Nat Genet, 1999, 23: 203-207

[2] Rafalski A. Applications of single nucleotide polymorphisms in crop genetics [J]. Curr Opin Plant Biol, 2002, 5: 94-100

[3] Syvanen A C. Accessing genetic variation Genotyping single nucleotide olymorphisms [J]. Nat Rev Genet, 2001, 2: 930-942

[4] Batley J, Barker G, O'Sulliuan H, et al. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data [J]. Plant Physio, 2003, 132: 84-91

[5] Lopez C, Piegu B, Cooke R, et al. Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz) [J]. Theor Appl Genet, 2005, 110: 425-431

[6] Sachidanandam R, Marth G, Mullikin J C, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms[J]. Nature, 2001, 409: 928-933

[7] Nesbitt T C, Tanksley S D. Comparative sequencing in the genus *Lycopersicon*. Implications for the evolution of fruit size in the domestication of cultivated tomatoes[J]. Genetics, 2002, 162: 365-379

[8] Lakshmi K M, John J G, David L H, et al. SNP-PHAGE-High throughput SNP discovery pipeline[J]. BMC Bioinformatics, 2006, 7: 468

[9] Ching A, Rafalski A. Rapid genetic mapping of ESTs using SNP pyrosequencing and indel analysis [J]. Cell Mol Biol Lett, 2002, 7: 803-810

[10] Gotoh K, Oishi M. Screening of gene-associated polymorphisms by use of ingel competitive reassociation and EST (cDNA) array hybridization [J]. Genome Res, 2003, 13: 492-495

[11] Pacey M T, Henry R. Single-nucleotide polymorphism detection in plants using a single-stranded pyrosequencing protocol with a universal biotinylated primer [J]. Anal Biochem, 2003, 317: 166-170

[12] Dantec L L, Chagne D, Pot D, et al. Automated SNP detection in expressed sequence tags: Statistical considerations and application to maritime pine sequences[J]. Plant Mol Biol, 2004, 54: 461-470

[13] Stephens M, Sloan J S, Robertson P D, et al. Auto-mating sequence-based detection and genotyping of SNPs from diploid samples [J]. Nat Genet, 2006, 38: 375-381

[14] Weckx S, Del F J, Rademakers R, et al. NovoSNP, a novel computational tool for sequence variation discovery[J]. Genome Res, 2005, 15: 436-442

[15] Zhang J, Wheeler D A, Yakub I, et al. SNP detector: A oftware ool for ensitive and accurate SNP detection [J]. PLoS Comput Biol, 2005, 53: 395-404

[16] Diego L, José A C, Ana I, et al. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a resequencing approach and SNP lex technology [J]. BMC Genomics, 2007, 8: 424-435

[17] Chen K, Michael D M, Ding L, et al. PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data[J]. Genome Res, 2007, 17: 659-666

[18] Nathalie P, Lee S P, Charles P, et al. Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs [J]. BMC Genomics, 2006, 7: 174

[19] Dorette T H, Verhoeven R, Goldbohm A, et al. Epidemiological studies on *Brassica* vegetables and cancer Risk [J]. cancer Epidemiology, Biomarkers and Prevention, 1996, 5: 733-748

[20] Fahey J W, Talalay P. The role of Crucifers in cancer chemoprotection in Phytochemicals and Health, D L Gustine and H E Flores[J]. American Society of Plant Physiologists, Rockville, MD, 1995: 87-93

[21] Dennis A B, Ilene K M, David J L, et al. GenBank[J]. Nucleic Acids Res, 2008, 36: 25-30

[22] Dong Q F, Carolyn J L, Shannon D S, et al. Comparative plant genomics resources at plant GDB [J]. Plant Physiol, 2005, 139: 610-618

[23] Anantharaman K, Srinivas A, Suresh K V. Efficient clustering of large EST data sets on parallel computers [J]. Nucleic Acids Res, 2003, 31: 2963-2974

[24] Huang X Q, Madan A. CAP3: A DNA sequence assembly program [J]. Genome Res, 1999, 9: 868-877

[25] Tang J F, Ben V, Roeland E V, et al. Quality SNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species [J]. BMC Bioinformatics, 2006, 7: 438

[26] Gu Z, Hillier L, Kwok P Y. Single nucleotide polymorphism hunting in cyberspace [J]. Hum Mutat, 1998, 12: 221-225

[27] Buetow K H, Edmonson M N, Cassidy A B. Reliable identification of large numbers of candidate SNPs from public EST data [J]. Nat Genet, 1999, 21: 323-325

[28] Picoult N L, Ideker T E, Pohl M G, et al. Mining SNPs from EST databases [J]. Genome Res, 1999, 9: 167-174

[29] Bhattramakki D, Dolan M, Hanafey M, et al. Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers [J]. Plant Mol Biol, 2002, 48: 539-547