赵　涛,王静毅,刘菊华,等. 香蕉 EST－SNP 标记的开发[J]. 江苏农业科学,2019,47(21):107－110.
doi:10.15889/j.issn.1002－1302.2019.21.024

香蕉 EST－SNP 标记的开发

赵　涛[1,2], 王静毅[1], 刘菊华[1], 徐碧玉[1], 金志强[1,3]
(1. 中国热带农业科学院热带作物生物技术研究所/农业部热带作物生物学与遗传资源利用重点实验室,海南海口 571101;
2. 南京农业大学园艺学院,江苏南京 210095; 3. 中国热带农业科学院海口实验站/海南省香蕉遗传改良重点实验室,海南海口 570102)

摘要:为发掘出一批香蕉的 SNP 位点、进一步研究香蕉的遗传关系、相关性状的定位等打下基础,从美国国立生物技术信息中心(National Center for Biotechnology Information,NCBI)的 dbEST 数据库下载 46 665 条香蕉 EST 序列,经生物信息学方法分析发掘 EST－SNP 位点,并对其所在核酸序列进行功能注释分析。通过对 46 665 条 EST 进行拼接,共得到 3 490 条重叠群(contigs),在含有 4 条以上重叠群中发现有 39 条重叠群中含有 SNP 位点,从中筛选出 127 个候选 SNP 位点,其碱基突变类型中转换、颠换分别占 SNP 位点总数的 63.78%、36.22%。通过序列比对分析发现了 34 个与香蕉相关基因,证明 NCBI 中的香蕉 EST 数据库数据量大,能够发掘出 SNP 标记对香蕉进行品种鉴定、分类和遗传多样性分析。

关键词:香蕉;EST 序列;SNP 位点;重叠群;转换;颠换;序列比对分析;遗传多样性
中图分类号: S668.101　　**文献标志码**: A　　**文章编号**:1002－1302(2019)21－0107－04

单核苷酸多态性(single nucleotide polymorphisms,SNPs)是指在基因组水平上,由单个核苷酸的变异导致等位基因的多态性,不同的等位基因在特定位置上含有不同的碱基对,等位基因频率一般要大于 1%。SNP 变异类型有转换(transition)、颠换(transversion)、插入(insert)和缺失(deletion)4 种,通常只分析颠换和转换。如果一个 SNPs 的次等位基因频率大于 0.1,便可用于关联或者连锁研究。单核苷酸多态性不仅分布在非编码区,在编码区也有分布,存在于编码区的 SNP 称为 cSNP,这为研究者提供了丰富的生物信息。同时,SNP 相比 SSR 具有更高的遗传稳定性。因此,现在人们广泛的将其称为第 3 代分子标记,同时被认为是应用前景最好的遗传标记[1－3]。

表达序列标签(expressed sequence tags,EST)是来源于功能基因表达的 cDNA 片段,是转录区域多态性识别的重要资源。随着公共数据库中 EST 序列的暴发式增长,以 EST 序列为基础开发分子标记变得越来越方便;同时,EST 标记还具有通用性好、信息量大、开发方法简单快捷以及成本低等优点。利用 EST 开发分子标记可直接用于动植物分子育种等相关领域的研究[4]。

香蕉(*Musa acuminata*)属于芭蕉科芭蕉属,单子叶草本植物。目前,香蕉已经成为我国热带地区主要农业支柱产业,同时也是世界 6 亿人口的主食作物[5],更是世界四大水果之一。然而,近年来环境气候的变化导致我国香蕉主产区经常遭受冷、干旱等逆境胁迫,同时香蕉枯萎病使得香蕉产业正遭受着毁灭性威胁[6]。目前,香蕉主栽品种大多是三倍体,基因组高度复杂,通常状况下都是高度不育的,难以通过传统的杂交育种得到优良品种。现在香蕉育种中如何进行品种鉴定是难点之一。近年来,SNP 已广泛应用于品种鉴定和重要性状的基因定位、遗传连锁图谱构建、遗传多样性分析等相关研究领域[7－13]。同时,国内外在香蕉方面进行开发 SNP 的文章鲜有报道。本研究利用 NCBI 中的 dbEST 数据库,通过生物信息学分析开发 SNP,以期获得合适的分子标记,为香蕉育种株系鉴定提供技术支持。

1　材料与方法

1.1　香蕉 EST 序列的获取

从 NCBI 网站(http://www.ncbi.nlm.nih.gov/genbank/)通过关键词“MUSA”搜索下载,共得到 46 665 条香蕉 EST,所有 EST 序列均以 FASTA 格式保存。

1.2　香蕉 SNP 的挖掘

利用 SeqClean(http://compbio.dfci.harvard.edu/tgi/software)去除载体序列及冗余序列,之后使用 CD－HIT(http://www.bioinformatics.org/cd－hit)和 CAP3(http://seq.cs.iastate.edu/cap3.html)进行序列的聚类与拼接。利用 QualitySNP(http://www.bioinformatics.nUtools/snpweb/)寻找 SNP 位点。

1.3　筛选原则

香蕉 SNP 位点筛选原则:(1)规定候选 SNP 位点两侧至少有 5 bp 碱基要完全保守;(2)候选 SNP 位点中的次要等位基因频率至少为 30%[14];(3)碱基判读质量与其所在的位置相关,测序所得的序列前区段质量普遍偏低,应选择序列

收稿日期:2018－08－03
基金项目:海南省重点研发计划(编号:ZDYF2018097);国家自然科学基金(编号:31501043);国家现代农业产业技术体系建设专项(编号:CARS－31);中央级公益性科研院所基本科研业务费项目(编号:1630052017018)。
作者简介:赵　涛(1990—),男,江苏徐州人,硕士研究生,研究方向为园艺学。Tel:(0898)66890772;E－mail:2532450562@qq.com。
通信作者:金志强,博士,研究员,博士生导师,研究方向为热带果树分子遗传学,E－mail:zhiqiangjin2001@yahoo.com.cn;徐碧玉,博士,研究员,研究方向为热带园艺植物基因工程,E－mail:biyuxu@126.com。

100 bp 以后的候选 SNP 位点。

1.4　BLAST 比对

提取含有 SNP 位点的重叠群(contigs)在 NCBI 的 BLASTn 数据库中进行序列比对,提取与序列相似性最高的序列注释信息,对 SNP 靶向基因产物和物种来源进行分析。

2　结果与分析

2.1　EST 文库来源

由表 1 可知,香蕉 EST 文库数量多,但其序列主要来源于 14 个 EST 文库,其数量为 44 829 条,占总 EST 的 96.06%。香蕉 EST 文库主要来源于香蕉 A 基因组,在所有的 EST 文库中,来源于香蕉叶片组织的高达 49.48%,来源于菜花样芽分生组织的占 23.72%,来源于香蕉根系的占 11.09%,来源于香蕉果实的仅占 5.41%。在香蕉 EST 文库中源于 Cachaco 品种的最多,高达 23.72%,其次为 Calcutta 4 – AA,占比为 20.00%,Grande Naine 品种占 14.05%,Pisang Awak(ABB) Sukari Ndizi(AB) Mpologoma(AAA) 占 11.77%,Pisang Klutug Wulung(PKW) – BB 仅占 11.33%,其品种和主要组织来源见表 1。

表 1　NCBI dbEST 数据库中香蕉主要的 EST 文库

文库名	物种	品种	组织	数量(条)
LIBEST_026748	*M. acuminata* AAA Group	Manoranjitham(AAA)	叶	596
LIBEST_025117	*M. acuminata* AAA GrouP	Manoranjitham	叶	241
LIBEST_017061	*M. acuminata* subsp. *burmannicoides*	Variety Calcutta 4(AA)	叶	1 143
LIBEST_017062	*M. acuminata* subsp. *burmannicoides*	Variety Calcutta 4(AA)	叶	1 143
LIBEST_027517	*M. acuminata*	Calcutta 4 – AA	叶	9 333
LIBEST_027525	*M. acuminata* AAA Group	Cavendish Grande Naine – AAA	叶	3 962
LIBEST_022976	*M. acuminata*	Pisang Awak (ABB), Sukari Ndizi (AB), Mpologoma(AAA)	根	2 535
LIBEST_022975	*M. acuminata*	Pisang Awak (ABB), Sukari Ndizi (AB), Mpologoma(AAA)	叶和芽分生组织	2 959
LIBEST_021167	*M. acuminata* AAA Group	Grande Naine	叶	4 030
LIBEST_021166	*M. acuminata* AAA Group	Grande Naine	果	2 528
LIBEST_023613	*M. balbisiana*	Pisang Klutug Wulung(PKW) – BB	根	2 644
LIBEST_023614	*M. balbisiana*	Pisang Klutug Wulung(PKW) – BB	叶	2 645
LIBEST_023617	*Musa* ABB Group	Cachaco	菜花样芽分生组织	2 502
LIBEST_023616	*Musa* ABB Group	Cachaco	菜花样芽分生组织	8 568

2.2　香蕉 EST 序列 SNP 频率分析

如表 2 所示,在 GenBank 数据库中下载到 46 665 条香蕉 EST 序列,通过 SeqClean 去除序列冗余,得到有效的 EST 序列 46 056 条。使用 CD – HIT 和 CAP3 进行序列的聚类与拼接,获得 3 490 条重叠群,为了提高 SNP 位点的可靠性,本研究所用的重叠群 EST 条数均大于 4,经过 QualitySNP 软件发掘 SNP 位点,在 456 条重叠群中发现 39 条中含有 SNP 位点,总计 127 个 SNP 位点。39 条重叠群的碱基总数为 35 743 bp,SNP 出现的频率为 0.35%,即平均每 281 bp 含有 1 个 SNP 位点。39 条重叠群中平均 1 条重叠群中含有 3.2 个 SNP 位点,含有 SNP 位点数最多的重叠群有 14 个位点,具体见表 3。

表 2　香蕉序列拼接结果

指标	数量(条)
EST 序列总数	46 665
参与拼接的 EST 序列数	3 490
重叠群总数	3 490
至少含有 4 条 EST 序列的重叠群数	456
包含 SNP 候选位点的重叠群数	39
可靠 SNP 位点数	127

表 3　重叠群中含有 SNP 位点

重叠群含 SNP 位点数(个)	重叠群(条)	占总重叠群的比(%)
1	17	43.58
2	4	10.26
3	4	10.26
4	3	7.70
5	2	5.13
6	5	12.82
7	2	5.13
10	1	2.56
14	1	2.56

如表 4 所示,本研究使用的 EST 序列包含 SNP 位点碱基转换占比 63.78%,颠换占比 36.23%,碱基的插入、缺失不统计。在不同重叠群中不同突变类型 SNP 位点的数量差异较大,其分布密度变化也很大。

2.3　SNP 位点所在核苷酸序列同源性比对结果分析

提取 39 个含有 SNP 位点的重叠群在 NCBI 的 BLASTn 数据库中进行比对。本研究发现 3 个未知蛋白,可能是香蕉特有或尚未被发现的基因(表 5),但须进一步验证。其他基因包括1个与抗逆有关的类热休克蛋白,3个与蛋白质降解、

表 4　不同碱基突变类型 SNP 的数量及其所占比例

SNP 变异类型	碱基	数量（个）	占比（%）
碱基转换	A/G	42	33.07
	C/T	39	30.71
碱基颠换	A/C	4	3.15
	A/T	17	13.39
	G/C	17	13.39
	G/T	8	6.30

DNA 损伤修复有关的泛素蛋白，1 个 CBS（cystathionine – beta – synthase）编码胱硫醚 –β– 合成酶基因，4 个与蛋白质合成相关的核糖体蛋白，1 个与信号传导相关的钙调蛋白，1 个参与真核翻译起始进程的真核翻译起始因子，1 个含 LIM 结构域的 LIM 蛋白，1 个与 DNA 结合的组蛋白，1 个参与细胞内物质运输和信号转导的 ADP – 核糖基化因子，1 个运输蛋白，1 个过氧化物酶基因，1 个韧皮部蛋白以及 1 个磷脂酰肌醇转移蛋白质家族成员等，其具体的 SNP 位点的比对结果见表 5。

表 5　香蕉 SNP 所在核苷酸序列同源性分析结果

重叠群	基因	相似度（%）	*E* 值	登录号
97	类热休克蛋白	99	0	XM_009411485.2
143	聚泛素	97	0	XM_009414391.2
149	CBS 含域蛋白 CBSX3，线粒体样蛋白	98	0	XM_009394619.2
151	丙二酸 – 羧酸酯 – 羧化酶	98	0	XM_009384199.2
155	未知蛋白	99	0	XM_009417049.2
171	麦角菌素、类根 R – B2	96	0	XM_009421347.2
172	核糖核酸酶 3	93	0	XM_009397530.2
196	60S 核糖体蛋白 L38 – like	96	0	XM_009384655.2
217	钙调蛋白 – 3 – like	99	0	XM_009412168.2
221	真核翻译起始因子 5a – 2	98	0	XM_009392566.2
231	伸长因子 1 – delta 1 – like	98	0	XM_009391379.2
236	26S 蛋白酶体非 ATP 酶调节亚基 8 同源亚群 A	99	0	XR_670687.2
244	未知蛋白	98	0	XM_009401309.2
251	PGRP – D mRNA 用于肽聚糖识别蛋白 D	97	3.00×10^{-59}	AB291943.1
264	类泛素蛋白 5	99	0	XM_009399402.2
278	LIM 含域蛋白 wlim2b – like	99	0	XM_009406291.2
287	二甲基丙基化因子 1 型	99	0	XM_009395629.2
299	组蛋白 H3.3	98	0	XM_009393049.2
300	蛋白 lurp – 1 相关	97	0	XM_009386660.2
308	磷脂酰肌醇转移蛋白质 – like	99	0	XM_009403089.2
316	初生的多肽相关复合物亚基类	98	0	XM_009406436.2
318	PHLOEM 蛋白 2 – LIKE A4 – like	99	0	XM_009385904.2
327	类 60S 核糖体蛋白 l28 – 2	98	0	XM_009386569.2
334	ATP – 核糖基化因子 1 – like	99	0	XM_009422684.2
347	类 60S 核糖体蛋白 L32 – 1	99	0	XM_009405954.2
351	蛋白质 GOS9 – like	99	0	XM_018818163.1
373	40S 核糖体蛋白 S15a	95	0	XM_009420996.2
399	异黄酮还原样蛋白	99	0	XM_009410410.2
402	ADP – 核糖基化因子 1 – like	93	0	XM_018827673.1
403	肽前体顺反式异构酶 FKBP12	98	0	XM_009401416.2
427	易位子相关蛋白亚单位 alpha – like	99	0	XM_009393701.2
434	蛋白运输蛋白 SEC13 同源 b – like	99	0	XM_018829157.1
442	细胞色素 b – c1 复合亚单元 6 – like	99	0	XM_009400936.2
444	预测 E3 泛素蛋白连接酶 RHC1A	99	0	XM_009413443.2
449	类 GTPase 活化蛋白 1	99	0	XM_009394592.2
450	未知蛋白	99	0	XM_009385717.2
456	过氧化物酶 – 2B	99	0	XM_009382289.2

3　讨论与结论

目前，开发 EST – SNP 的软件众多，软件的选取以及如何设置参数都是影响试验结果的关键因素。如 PolyPhred 只能预测某一核苷酸位点上单个碱基的替换，SNPdetector 假阳性率和假阴性率均低，novoSNP 的假阳性率明显偏高；在具有可靠的参考序列时，SOAPsnp 正确率较高；AutoSNP 正确率低；QualitySNP 预测位点少但正确率高于 AutoSNP，且 QualitySNP

运行速度更快[15];因此,本研究应选取 QualitySNP 开发 SNP。

在 EST 序列中进行 SNP 位点开发时,研究者应当注意影响 SNP 开发质量的各种筛选参数。其中最主要的因素为重叠群的规格(重叠群所包含 EST 序列的数量)和次要等位基因(等位基因中出现次数较少的碱基)的出现次数。李猛利用 QualitySNP 软件对葡萄 EST 序列进行候选 SNP 位点分析时发现,为了得到高质量的候选 SNP 位点,重叠群规格应选择拼接 EST 数量≥4 条以上,同时次要等位基因至少出现 2 次[16]。因为错配仅出现 1 次的话很可能是由序列差错引起的,而同一碱基位置上发生 2 次序列差错的概率则很小。因此在规格为 4 条,主次等位基因出现次数比为 1∶1,即次要等位基因出现 2 次的重叠群中开发的候选 SNP 其可靠度较高。在规格大于 4 条的重叠群中,也应当尽量保证主次等位基因出现次数比近似为 1∶1,即在规格为 5 ~6 条的重叠群中,次要等位基因应至少出现 2 次。一般在聚类时为得到高的比对分值,通常须要在 1 条序列中加入空格,但这样会被误判为插入或缺失,为避免出现这种情况,在处理结果时可以不考虑插入或缺失,而只分析替换类型。

本研究从 NCBI 中 dbEST 公共数据库下载 46 665 条 EST 序列,共有 46 056 条 EST 序列参与拼接,总计拼接成 3 490 条重叠群,所含 EST 序列≥4 条的重叠群共 456 条,在 39 个重叠群中发现 SNP 位点。同时大于 4 条以上的重叠群主要由 4 ~7 条 EST 序列拼接而成,最多的 1 条重叠群也只有 13 条 EST,8 条以上 EST 拼接的重叠群比较少。同时,本研究中重叠群主要长度在 800 ~1 500 bp,长度在 1 500 bp 以上的较少。一般为了提高 SNP 的可靠性,用于 SNP 分析的重叠群至少包含 4 条以上。

在 39 条重叠群中筛选出 127 个候选 SNP 位点,SNP 频率为 0.35%,较甘蔗[14]、茶树[17]等其他物种的 SNP 频率低,可能是由于香蕉是三倍体植物自交高度不育,在生产上主要依靠吸芽和组培苗进行繁殖生产,香蕉无法通过基因交流产生新的基因变化,所以自身遗传差异变化小,SNP 位点相比其他植物少。

一般情况下碱基转换的 C/T 比 A/G 更常发生。CpG 二核苷酸的胞嘧啶(C)在基因组中最易发生突变,其中大多数是甲基化的,可自发地脱去氨基而形成胸腺嘧啶(T),因此转换型变异的 SNP 约占 2/3[17]。在本研究中,香蕉 SNP 位点碱基变异类型以 G/A 为主,占 33.07%,C/T 占 30.70%,与甘蔗[14]、栉孔扇贝[18]碱基变异类型相同,与小麦[19]、大麦[20]、辣椒[21]等物种的 SNP 碱基变异类型不符。转换类型和颠换类型的数量分别占候选 SNP 位点总数的 63.78% 和 36.22%,转换与颠换比为 1.76∶1.00,即转换类型的数量明显高于颠换,与檀小辉等的研究结果[14]存在差异。

本研究中,含有 SNP 位点最多的重叠群 Contigs402 和 Contigs373 分别有 14、11 个 SNP 位点,其 EST 构成分别为 5、4 条,长度分别为 852、863 bp。而只含有 1 个位点的 Contigs97、Contigs287 的 EST 组成分别为 6、6 条,长度分别为 766、901 bp。由此看出,香蕉重叠群中 EST 序列数量与包含的 SNP 位点数量并无明显规律,这可能与不同物种间 SNP 位点的分布差异有关。

参考文献:

[1]Collins F S, Guyer M S, Charkravarti A. Variations on a theme: cataloging human DNA sequence variations[J]. Science, 1997, 278(5343): 1580 - 1581.

[2]Harding R M, Fullerton S M, Griffiths R C, et al. Archaic African and Asian lineages in the genetic ancestry of modern humans[J]. American Journal of Human Genetics, 1997, 60(4): 772 - 789.

[3]Nickerson D A, Taylor S L, Weiss K M, et al. DNA sequence diversity in a 9.7kb region of the human lipoprotein lipase gene[J]. Nature Genetics, 1998, 19(3): 233 - 240.

[4]梁 芳,张 继,吕 平,等. 基于 EST 序列的玫瑰 EST - SNP 位点发掘与分析[J]. 南方农业学报,2016,47(3):325 - 331.

[5]张 静,孙秀秀,徐碧玉,等. 香蕉分子育种研究进展[J]. 分子植物育种,2018,16(3):914 - 923.

[6]窦同心. 香蕉抗寒、抗病相关基因的遗传转化验证[D]. 广州:华南农业大学,2016:1 - 2.

[7]孟 霞,曾兴权,其美旺姆,等. 西藏冬青稞种质资源 SNP 标记的遗传多样性分析[J]. 现代农业科技,2018(1):40 - 41,43.

[8]姚丹青,楼坚锋,朱文莹,等. 基于 SNP 标记的黄瓜遗传多样性分析[J]. 上海农业学报,2017,33(1):21 - 30.

[9]刘 凯,邓志英,李青芳,等. 利用高密度 SNP 遗传图谱定位小麦穗部性状基因[J]. 作物学报,2016,42(6):820 - 831.

[10]杨润婷,吴 波,李 翀,等. 两种 SNP 分型方法的比较及其在柚品种鉴定中的应用[J]. 园艺学报,2013,40(6):1061 - 1070.

[11]毛建军. 杂交水稻品种鉴定的 SNP 研究及东乡野生稻两个 NBS 序列的分析[D]. 长沙:湖南农业大学,2005:44 - 45.

[12]李胜杰,白俊杰,赵 荦,等. 大口黑鲈 EST - SNP 标记开发及其与生长性状的相关性分析[J]. 海洋渔业,2018,40(1):38 - 46.

[13]阴长发. 甘蓝型油菜 EST - SNP 开发及花色性状的 QTL 定位[D]. 长沙:湖南农业大学,2013:38 - 40.

[14]檀小辉,张 继,梁 芳,等. 基于 EST 序列的甘蔗 SNP 发掘及分析[J]. 江苏农业科学,2016,44(7):64 - 66,67.

[15]李 猛,郭大龙,刘崇怀,等. EST - SNP 开发软件特性分析及比较[J]. 生命的化学,2011,31(6):906 - 911.

[16]李 猛. 葡萄 EST - SNP 标记的开发及应用[D]. 洛阳:河南科技大学,2012:24 - 25.

[17]王丽鸳,张成才,成 浩,等. 茶树 EST - SNP 分布特征及标记开发[J]. 茶叶科学,2012,32(4):369 - 376.

[18]李纪勤,包振民,李 玲,等. 栉孔扇贝 EST - SNP 标记开发及多态性分析[J]. 中国海洋大学学报(自然科学版),2013,43(1):56 - 63.

[19]Chao S, Zhang W J, Akhunov E, et al. arker polymorphism in US wheat (*Triticum aestivum* L.) cultivars[J]. Molecular Breeding, 2009, 23(1): 23 - 33.

[20]Sato K, Close T J, Bhat P, et al. Single nucleotide polymorphism mapping and alignment of recombinant chromosome substitution lines in barley[J]. Plant & Cell Physiology, 2011, 52(5): 728 - 737.

[21]刘 峰,谢玲玲,弭宝彬,等. 辣椒转录组 SNP 挖掘及多态性分析[J]. 园艺学报,2014,41(2):343 - 348.