

香菇基因组中 EST-SSR 的构成和分布*

林范学^{1,2} 程水明³ 李安政¹ 林芳灿^{1**}

(华中农业大学应用真菌研究所 武汉 430070) (济宁师范专科学校生物系 济宁 272125) (黄冈师范学院生物系 黄冈 438000)

摘要 :从真菌基因组计划网站(FGP)和NCBI网站数据库下载了符合条件的总长度为 8.1×10^6 bp的11150条香菇的EST(包括10条cDNA)序列,通过SSRhunter 1.3软件结合手工查找,从中发现2.83%即316条EST含有一共469个SSR,平均每17.3kb出现一个EST-SSR。在所有EST-SSR中,三碱基和六碱基SSR出现最多,分别占EST-SSR总数的38.00%和20.00%。出现较多的基序为 $(A)_n$ 、 $(T)_n$ 、 $(GA)_n$ 、 $(AG)_n$ 、 $(TGA)_n$ 、 $(GAT)_n$ 和 $(TCTTT)_n$,占所有EST-SSR的35.39%。

关键词 :EST-SSR,基序,构成,分布,香菇

中图分类号 :Q939.5 S646.12 文献标识码 :A 文章编号 :0253-2654(2007)03-0438-05

Compose and Distribution of EST-SSR in Genomes of *Lentinula edodes**

LIN Fan-Xue^{1,2} CHENG Shui-Ming² LI An-Zheng¹ LIN Fang-Can^{1**}

(The Institute of Applied Mycology, Huazhong Agricultural University, Wuhan 430070)

(Department of Biology, Jining Teachers school, Jining 272125)

(Department of Biology, Huanggang Normal University, Huanggang 438000)

Abstract :In this study, 11,150 qualified ESTs(including 10 cDNAs), about 8.1×10^6 bp in length, were downloaded from ESTs databases of *Lentinula edodes* in Fungal Genomics Project(FGP) and NCBI. By searching with SSRhunter 1.3 and handwork, 469 SSRs were identified in 316 ESTs, approximately 2.83% of total ESTs. The average distance between SSRs was about 17.3kb. Trinucleotide and hexanucleotide ESR-SSRs were found to be the two most abundant repeats, accounting for 38.00% and 20.00% of the total ESR-SSRs respectively. The most familiar motifs are $(A)_n$, $(T)_n$, $(GA)_n$, $(AG)_n$, $(TGA)_n$, $(GAT)_n$ and $(TCTTT)_n$, which comprised about 35.94% of all EST-SSRs.

Key words :ESR-SSR, Motif, Compose, Distribution, *Lentinula edodes*

简单重复序列(simple sequence repeats, SSR)又称微卫星DNA、短串联重复序列,一般为以1~6个碱基为核心序列,首尾相连组成的串联重复序列^[1]。SSR广泛分布于真核生物基因组中,对SSR的构成、长度和分布等各种性质的研究可以了解染色体片段之间的特殊关系,有助于分析种或属之间的亲缘关系。研究表明,在系统分类中,物种基因组中SSR的数量、长度、构成、突变率和在染色体上的分布等特性有很大的不同^[2]。在功能基因组学研究中,许多证据表明,SSRs具有调节基因表达的作用^[3,4]。SSR两翼具有相对保守的单拷贝序列,根据该序列设计引物,对模板DNA进行PCR扩增,

所检测到的扩增片段的长度多态性可作为分子标记^[5]。与其它分子标记相比,SSR标记具有数量丰富、多态性高、共显性遗传、DNA用量少、操作简便、重复性好等优点,在发展高密度遗传图、植物品种鉴定、基因定位、群体遗传、物种进化等研究中广为应用^[6~10]。开发SSR标记的方法主要有构建与筛选基因组文库法、微卫星富集法、省略筛库法和数据库搜索法等四种^[11],其中数据库搜索法,尤其是从表达序列标签EST(expressed sequence tags)数据库中搜索,由于具有省时、省力、低成本等优点,已经成为一种广为采用的方法^[12~14]。

作为第二代分子标记^[15],SSR标记在病原菌中

* 高校博士点专项基金资助项目(No. 20020504021)

** 通讯作者 Tel:027-87287348 E-mail: Linfangcan@mail.hzau.edu.cn

收稿日期:2006-07-07,修回日期:2006-08-25

的应用较多^[16-18],对 SSR 在基因组或开放阅读框中的数量、构成和分布等特点也有研究^[19-20],而在世界第二大栽培蕈菌香菇(*Lentinula edodes*)等食用蕈菌中的应用鲜有报道。Barroso 等 2000 年首次报道了双孢蘑菇(*Agaricus bisporus*)及侧耳(*Pleurotus* spp.)中存在(TATG)_n 基序^[21],秦莲花等也发现香菇基因组中存在(TATG)_n 基序^[22];Rosa 等利用(GT)_n和(GA)_n 做探针,分别在*P. eryngii*和*P. ferulae*中鉴定了GT和GGAA 2个微卫星标记^[23]。这在食用蕈菌遗传育种研究中发展SSR标记做了有益的探索。本研究从真菌基因组计划网站FGP(<https://fungalgenomics.concordia.ca>)和NCBI(<http://www.ncbi.nlm.nih.gov>)网站下载了截止到2005年12月25日所有符合条件的香菇EST序列,利用软件查找了所有香菇的EST-SSR,分析了香菇EST-SSR的数量、构成、分布等特性,为在香菇遗传育种中进一步开发利用SSR标记做了初步的研究。

1 材料与方 法

本研究所用EST序列是从FGP网站和NCBI网站数据库下载的长度大于100bp、5'和3'末端50bp碱基内不包含(T)₅或(A)₅的香菇EST序列。EST数据截止到2005年12月25日。按照单至六碱基重复次数至少为15、8、5、4、3、3次的标准,利用SSRHunter1.3软件^[24],并结合人工进行SSR查找并统计。

2 结果与分析

2.1 香菇 EST-SSR 的数量和分布

从FGP网站和NCBI网站的EST数据库分别下载了符合条件的9767条和1383条香菇EST(包括10条cDNA),共计11150条,序列总长度为 8.1×10^6 bp。按照查找标准,从上述两个数据库中共查找出316条含SSR的EST,占EST总数的2.83%,其中从FGP中查找出293条,占该数据库EST的3.00%;从NCBI中查找出23条,占该数据库EST的1.66%。从上述两个数据库中共查找出469个SSR(FGP中421个,NCBI中48个),占EST总数的4.21%,平均每17.3kb出现一个SSR。所有SSR碱基数共12215 bp,占EST碱基数总长度的0.15%。无论EST总数、含SSR的EST(SSR-EST)的总数还是SSR总数,

都以FGP数据库为多。

表1 香菇基因组中EST-SSR的数量和分布

重复次数	单碱基	二碱基	三碱基	四碱基	五碱基	六碱基
3					44	84
4				5	1	8
5			120			
6			30		1	
7			20			1
8		22	7			
9		6				
10			1			
11		3				
12		2				
13		1				
14		1				
15	8	1				
16	6	1				
17	12					
18	14					
19	3					
21~30	12	5				1
31~40	13	3				
41~50	5	1				
>50	5	22				
Total	78	68	178	5	46	94
Proportion (%)	16.63	14.50	38.00	1.07	9.81	20.00
Total length (bp)	2097	4462	2958	80	710	1908
Average length (bp)	26.88	65.62	16.62	16.00	15.43	20.30

香菇EST-SSR的构成和分布列于表1。从SSR数量上看,在所有469个SSR中,出现最多的是三碱基重复,有178个,占总数的37.95%;其次为六碱基重复,有94个,占20.00%;再后依次为单碱基、二碱基、五碱基重复,分别有78、68、46个;四碱基重复出现最少,仅有5个。数量最多的前3位SSR与粗糙脉孢菌(*Neurospora crassa*)中的结果一致^[25],灰盖鬼伞(*Coprinus cinereus*)中也是以三碱基和六碱基重复居前两位^[19]。在SSR总长度上,二碱基、三碱基和单碱基SSR最长,分别有4462bp、2958bp和2097bp。在SSR平均长度上,前三位为二碱基、单碱基和六碱基重复,分别为65.62bp、26.88bp和20.30bp。

在重复次数上,单碱基SSR有30种重复,主要集中在15~40次重复之间,共有68条,50次以上重复的只有10条。出现SSR数量最多的为17和18次重复,分别有12和14条。二碱基SSR有32种重复,其中37条出现在8~16次重复,8次重复的数量最多,有22条;21~50次重复有9条,50次以上重复的有22条。三碱基SSR有5种重复次数,都集

中在5~10次重复,其中5次重复的最多,有120条。四碱基SSR只有1种即4次重复的5条。五碱基SSR有3种重复,以3次重复为最多,有44条;4次和6次重复的各有1条。六碱基SSR有4种重复,数量最多的也是3次重复,有84条,另外6次重复的有8条,7次和27次重复的各有1条。

从分析结果中可知,单碱基和二碱基SSR在重复次数的种类和跨度上均比四到六碱基SSR大许多。这似乎说明,单碱基和二碱基SSR易于变异,而四、五、六碱基SSR相对保守。但并不绝对,如六碱基SSR中就有27次重复。

2.2 香菇 EST-SSR 中基序的特点和性质

香菇基因组中各种EST-SSR基序的数目、种类见表2。单碱基SSR中主要基序为A和T,分别有34条和29条,二者占到该类SSR的80.77%,而G

和C基序只有13条和2条。这一特点与其它真菌相一致,灰盖鬼伞中多聚A和多聚T二者共有208个,占单碱基重复的73.5%;构巢曲霉(*Aspergillus nidulans*)中两者数量合计达429个,占单碱基重复的92.5%^[26]。该特点与其它物种也一致^[27,28]。香菇中单碱基SSR重复次数最多的(T)₁₀₀,出现次数较多的有(A)₈、(C)₇、(A)₇和(C)₅,分别出现9、6、5和5次。

二碱基SSR中以GA、AG基序最多,共有40条,占该类总数的58.82%。这一现象与在部分真菌中的分析结果一致^[19,25,29],而与构巢曲霉和蜜蜂(*Apis* spp.)有所差异,这二者中以AT基序最多,AG基序次之^[26,27]。香菇中该类SSR重复次数最多的是(GA)₉,出现次数较多的有(TA)₈和(GT)₈,均有7条。

表2 香菇基因组中EST-SSR的主要基序

基序	数量	基序	数量	基序	数量	基序	数量	基序	数量	基序	数量
A	34	GA	24	TGA	23	CCAA	2	TCTTT	19	GGCAAA	7
T	29	AG	16	GAT	21	AAAG	1	CAAAG	6	GGTGA	6
C	13	GT	8	CGA	13	CTCA	1	CAACG	4	CTCAAG	5
G	2	TA	8	ACG	12	TCTA	1	TTCCC	3	AGGACA	4
		TC	6	GCT	11			AGGAT	2	CCAAGG	4
		TG	3	GGA	11			CTTCT	2	GGAAGA	4
		AC	2	CIT	10			TATCC	2	GGTCGA	4
		CT	1					TTTTC	2	AGTTGG	3
										ATACCA	3
										CGAGGA	3
										GAGGTG	3

注:表中所列三碱基SSR中的基序为10次以上、占该类全部SSR的56.74%,五碱基SSR在2次以上、占该类全部SSR的82.61%,六碱基SSR在3次以上、占该类全部SSR的48.94%

三碱基SSR有31种基序,其中TGA、GAT、CGA、ACG、GCT、GGA、CIT等7种基序较多,计占该类SSR总数的56.74%。该结果与其它真菌不同,后者中AGG、AGC、ACC和AAG等基序较丰富^[19,20,25]。在香菇中该类SSR以(TGA)₃、(GAT)₃等基序次数较多,分别为16和14次,重复最多的为(CAC)₁₀。

四碱基SSR是香菇所有SSR中最少的一类,均为4次重复,共5条。这与其它真菌不同,如稻瘟病菌、粗糙脉孢菌和构巢曲霉中二碱基SSR数量最少^[25,26,29],灰盖鬼伞中单碱基SSR数量最少^[19]。蜜蜂中五碱基SSR的数量最少^[27]。但在小麦中,四碱基SSR数量也最少^[14];在茶树中,四碱基SSR数量也较少^[30]。香菇中该类SSR基序有CCAA 2条,

AAAG、CTCA、TCTA各1条,出现了其它物种中数量较多的AAAG基序。

香菇中五碱基SSR有46条,基序数目有15种,出现2次以上的占到该类基序的82.61%。数量最多的为(TCTTT)₃,出现了19次,其次为(CAAAG)₃和(CAACG)₃,分别出现6次和4次。研究过的其它真菌中AAAAG基序的数量最多。由于相关研究者将所有可循环的及互补的序列归为一类,所以AAAAG基序中包含TCTTT基序,可见香菇中五碱基SSR主要基序与其它真菌相一致。

香菇六碱基SSR是所有基序中最为丰富的一类,共有44种。较多的基序为(GGCAAA)₃、(CTCAAG)₃和(GGTGGA)₃,分别有7、5和5次。重

复次数最多的是(AACCCT)₂₇。这一点与其它真菌不同,后者六碱基 SSR 多以 AAAAAG、AAAAAT 为主^[26,29];蜜蜂中六碱基 SSR 也是以 AAAAAG、AAAAAT 为主^[27]。

总体上看,香菇中 EST-SSR 以(A)_n、(T)_n、(GA)_n、(AG)_n、(TGA)_n、(GAT)_n 和(TCTTT)_n 出现较多,分别有 34、29、24、16、23、21 和 19 次,合计占有 EST-SSR 的 35.39%。在所有 EST-SSR 中,六碱基 SSR 的基序种类最多,有 44 种,其次为三碱基 SSR,有 31 种基序,五碱基 SSR 有 15 种基序。

综上,在香菇基因组中,单碱基、二碱基和五碱基 SSR 中的主要基序和数量与其它真菌基本一致,三碱基、四碱基和六碱基 SSR 与其它真菌不同。由此可见,在 SSR 的构成和分布上,香菇与其他真菌既有共性,也有区别;在某些方面,如四碱基 SSR 的数量,类似小麦和茶树(*Camellia sinensis*)等高等植物^[14,30]。

3 讨论

根据序列结构,SSR 可分为完全、不完全和复合型等 3 种重复类型。搜索 SSR 的软件有多种,本研究使用的 SSRHunter 可精确查出所有序列中的完全重复 SSR,且给出 SSR 上下游各 150bp 的序列^[24],以利于设计引物。对于 SSR 长度,有些学者认为碱基数等于或大于 12bp 较为合适^[31,32];Cardle 等(2000)限定单至五碱基重复 SSR 碱基数要分别大于 15bp、14bp、15bp、16bp 和 20bp^[33]。本研究采取 Cardle 的标准并稍作修改,将单至六碱基重复 SSR 最低碱基限定为 15bp、16bp、15bp、16bp、15bp 和 18bp,从而使查找结果较为严谨。通过软件查找,本研究从 11150 条香菇 EST 序列中发现 316 条 EST 含有 SSR,占 EST 总数的 2.83%。该比例与从其它物种中 EST-SSR 数出现比例大体相符,如苜蓿(*Medicago* spp.)、面包小麦(*Triticum aestivum* L.)和棉花分别有 2.98%、6.00% 和 1.1% ~ 4.8% 的 EST 含有 SSR^[34~36]。

灰盖鬼伞和粗糙脉孢菌中三碱基 SSR 数目最多,六碱基 SSR 次之^[19,25]。本研究结果与此相符。说明近缘物种的 SSR 相关信息比较接近。香菇中每 17.3 kb 出现一个 EST-SSR,频率较灰盖鬼伞(1 SSR/4.61 kb)和粗糙脉孢菌(1 SSR/2.57 kb)均很低。

经过对香菇 EST-SSR 的分析,发现香菇中的 SSR 以(A)_n、(T)_n、(GA)_n、(AG)_n、(TGA)_n、(GAT)_n 和(TCTTT)_n 出现较多,与对其它真菌的研究结果基本一致,且有植物中常见的(AG)_n。对人基因组中 SSR 的研究表明,SSRs 具有调节基因表达的作用^[3];对稻瘟病菌的研究也发现,含有 SSR 的基因多为转录蛋白、信号传导的调节基因,所以 SSR 在物种形成中具有重要作用^[20]。由于有关 SSR 在香菇各种已知功能基因中的分布、数量等均未知,对 EST-SSR 这种直接来源于 cDNA 的功能标记尚需要进一步研究。

参考文献

- [1] Beckmann J S, Soller M. *Bio Technology*, 1990, **8**: 930 ~ 932.
- [2] Kashi Y, King D, Soller M. *Trends Genet*, 1997, **13**: 74 ~ 78.
- [3] Kunzler P, Matsuo K, Schaffner W. *Biol Chem Hoppe Seyler*, 1995, **4**: 201 ~ 211.
- [4] Moxon E R, Wills C. *Sci Am*, 1999, **280**: 94 ~ 99.
- [5] 张增翠, 侯喜林. *遗传*, 2004, **26**(5): 763 ~ 768.
- [6] Weissenbach J. *Clin Chem Lab Med*, 1998, **36**: 511 ~ 514.
- [7] 苏顺宗, 黄玉碧, 杨俊品, 等. *种子*, 2003, **123**(1): 23 ~ 25.
- [8] 任丽娟, 沈晓蓉, 周森平, 等. *中国农业科学*, 2003, **36**(10): 1150 ~ 1155.
- [9] Franck P, Gamery L, Celebrano G, et al. *Molecular Ecology*, 2003, **9**(2): 907 ~ 921.
- [10] Sunmucks P, England P R, Taylor A C, et al. *Genetics*, 1996, **144**(2): 747 ~ 756.
- [11] 李明芳, 郑学勤. *遗传*, 2004, **26**(5): 769 ~ 776.
- [12] Provan J, Powell W, Waugh R. *Theor Appl Genet*, 1996, **92**(8): 1078 ~ 1084.
- [13] Huang S W, Zhang B X, Millbourne D. *Euphytica*, 2000, **117**: 163 ~ 167.
- [14] 陈军方, 任正隆, 高丽锋, 等. *作物学报*, 2005, **31**(2): 154 ~ 58.
- [15] Weissenbach J, Gyapay G, Dib C, et al. *Nature*, 1992, **359**: 794 ~ 801.
- [16] Kaye C, Milazzo J, Rozenfeld S, et al. *Fung Genet & Biol*, 2003, **40**: 207 ~ 214.
- [17] Kaszubiak A, Klein S, de Hoog G S, et al. *Genetics and Evolution*, 2004, **4**: 179 ~ 186.
- [18] Steimel J, Chen W, Harrington T C. *Molecular Ecology Notes*, 2005, **5**: 484 ~ 486.
- [19] 刘林, 李成云, 杨静, 等. *西南农业学报*, 2006, **19**(1): 131 ~ 135.
- [20] 李成云, 李进斌, 周晓晔, 等. *中国水稻科学*, 2005, **19**(2): 167 ~ 173.
- [21] Barroso G, Sonnenberg A S M, van Griensven L J L D, et al. *Fung Genet & Biol*, 2000, **31**: 115 ~ 123.

478.

[23] Rosa V D , Reverberi M , Fabbri A A , *et al.* Isolation and characterization of microsatellite loci in the species *Pleurotus eryngii* (DC :Fr) Quel and *P. ferulae* (DC :Fr) Quel using a biotin/ streptavidin enrichment technique. In :6th European Conference on Fungal Genetics abstract book ,Pisa ,Italia 2002 ,P :IVp8.

[24]李 强 ,万建民.遗传 2005 **27**(5) 808 ~ 810.

[25]李成云 ,李进斌 ,周晓罡 ,等.中国农业科学 2004 **37**(6) :851 ~ 858.

[26]李成云 ,李进斌 ,周晓罡 ,等.遗传学报 2005 **32**(5) 538 ~ 554.

[27]李 斌 ,夏庆友 ,鲁 成 ,等.遗传学报 2004 **31**(10) :1089 ~ 1094.

[28] Katti M V ,Ranjekar P K ,Gupta V S. Mol Biol Evol 2001 **18**(7) :1161 ~ 1167.

[29]李成云 ,李进斌 ,周晓罡 ,等.中国水稻科学 2004 **18**(3) :269 ~ 273.

[30]金基强 ,崔海瑞 ,陈文岳 ,等.茶叶科学 2006 **26**(1) :17 ~ 23.

[31] Toth G ,Gaspari Z ,Jurka J. Genome Res 2000 **10**(7) 967 ~ 981.

[32] Morgante M ,Hanafey M ,Powell W. Nat Genet 2002 **30**(2) :194 ~ 200.

[33] Candle L ,Ramsay L ,Milbourne D ,*et al.* Genetics 2000 **156** :847 ~ 854.

[34] Eujayl I ,Sledge M K ,Wang L ,*et al.* Theor appl genet 2004 **108**(3) : 414 ~ 422.

[35] Gupta P K ,Rustgi S ,Sharma S ,*et al.* Mol gen genomics 2003 **270** 315 ~ 323.

[36] Saha S ,Karaca M ,Jenkins J N ,*et al.* Euphytica 2003 **130**(3) :355 ~ 364.