

## SSR 在黑木耳和毛木耳转录组中的分布和序列特征

周雁 范秀芝 陈连福 边银丙\*

华中农业大学应用真菌研究所 湖北 武汉 430070

**摘要:** 以黑木耳和毛木耳转录组测序获得的转录本序列为基础, 采用软件 SSR Locator 对两个物种中 SSR 的数量、分布和结构特征等方面进行分析, 发现毛木耳中 SSR 的丰度和密度较黑木耳小。在两者的 SSR 中, 三碱基和六碱基重复的出现最多。对含量最多的三碱基重复序列的基序进行分析表明富含 GC 的 SSR 在这两个种的转录组中占优势地位。

**关键词:** 黑木耳, 毛木耳, 转录组, 简单重复序列, 基序

## Distribution and sequence characteristics of SSR in the transcriptomes of *Auricularia auricula-judae* and *Auricularia polytricha*

ZHOU Yan FAN Xiu-Zhi CHEN Lian-Fu BIAN Yin-Bing\*

Institute of Applied Mycology of Huazhong Agricultural University, Wuhan, Hubei 430070, China

**Abstract:** Illumina Solexa sequencing technology was used to generate transcript sequences of *Auricularia auricula-judae* and *A. polytricha*. The number, distribution and motif characteristic of SSR in this two species were analyzed using the software SSR Locator. The SSR abundance and density in *A. polytricha* were less than that in *A. auricula-judae*. However, tri- and hexa-nucleotide SSR were two kinds of significantly abundant repeats in both transcriptomes. The motif analysis of tri-nucleotide SSR which was the most popular repeats indicated the GC-rich repeats were the dominant SSR in both transcriptomes.

**Key words:** *Auricularia auricula-judae*, *A. polytricha*, transcriptome, SSR, motif

简单重复序列 (simple sequence repeats, SSR) 广泛分布于真核生物基因组中, SSR 在基因组进化过程中扮演着十分重要的角色, 故对 SSR 的数量、构成和分布等特性的分析, 有助

基金项目: 中国博士后科学基金 (No. 20110491164)

\*Corresponding author. E-mail: bianyb.123@163.com

收稿日期: 2013-12-26, 接受日期: 2014-01-20

于了解物种之间的亲缘关系 (Li *et al.* 2009)。在许多动物、植物和少数真菌中已开发出大量 SSR 分子标记 (Dutech *et al.* 2007), 为构建分子标记遗传连锁图、种质资源多样性分析、比较基因组学和分子标记辅助选择育种等领域的发展提供了大量的分子标记资源。在植物中, 由于 SSR 具有多态性高、共显性遗传、易于用 PCR 检测和在基因组上分布均匀等特点, 以 EST 为基础开发的分子标记中应用最多的是 SSR, 而且由于 EST-SSR 来源于转录区, 当它们用于种质资源评价时表现的是转录区的差异, 因而更能反映出真实的遗传多样性; 应用于分子标记辅助选择时, 可以直接进行等位基因选择; 此外, 由于 EST-SSR 的高转移性, 使之非常适于比较作图研究和合并不同遗传连锁图 (Powell *et al.* 1996)。

随着第二代测序技术的飞速发展, 基因组和转录组测序被广泛应用于各种真核和原核生物的分子生物学研究, 为各种生物的基础研究奠定了技术基础。随着裂褶菌 *Schizophyllum commune* Fr.、灰盖拟鬼伞 *Coprinopsis cinerea* (Schaeff.) Redhead *et al.*、灵芝 *Ganoderma lucidum* (Curtis) P. Karst.、双色蜡蘑 *Laccaria bicolor* (Maire) P.D. Orton 和皱木耳 *Auricularia delicata* (Mont.) Henn. 等蕈菌基因组测序的完成, 这使得 SSR 的查找和分子标记的开发更加高效便利 (Floudas *et al.* 2012; Labbé *et al.* 2011; Qian *et al.* 2013)。对于非模式生物而言, 无参转录组测序在基因克隆、基因表达谱分析和分子标记开发等方面的应用已非常广泛 (Zalapa *et al.* 2012)。

黑木耳 *Auricularia auricula-judae* (Bull.) Quél. (戴玉成和李玉 2011) 和毛木耳 *A. polytricha* (Mont.) Sacc. 是木耳属中栽培最广泛

的两个种, 是重要的食药两用真菌 (戴玉成和杨祝良 2008; 戴玉成等 2010)。本研究采用 Illumina 的 Solexa 测序技术分别对这两个物种的菌丝体和子实体进行转录组测序, 并在无参转录组拼接产生的 EST 基础上, 利用软件 SSR Locator 进行 SSR 的查找, 分析了黑木耳和毛木耳的 EST-SSR 的数量、构成、分布等特性, 为这两个物种的遗传连锁图谱的构建和分子标记辅助选择中进一步开发利用 SSR 标记提供了序列基础。

## 1 材料与方法

### 1.1 供试菌株

用于黑木耳转录组测序的供试菌株为 Au916, 是保藏于华中农业大学应用真菌研究所的双核体栽培菌株。

毛木耳野生菌株 APTJ6101 引自四川省农业科学院土壤肥料研究所, 通过原生质体分离获得原生质体单核体 App7。毛木耳栽培菌株 MHJY002 来源于福建省福州市闽侯县嘉永食用菌有限公司, M2S16 是通过担孢子单孢分离获得的 MHJY002 的孢子单核体后代之一。App7 和 M2S16 经杂交配对产生的杂交子 APM2-16 是毛木耳转录组测序的供试菌株, 现保藏于华中农业大学应用真菌研究所。

### 1.2 RNA 样本的准备

分别将黑木耳菌株 Au916 和毛木耳菌株 APM2-16 接种于液体完全培养基 (liquid complete yeast medium, LCYM) 中 (Horgen *et al.* 1989), 25℃ 培养 7d 后收集菌丝体并保存于 -80℃ 超低温冰箱中保存备用。黑木耳 Au916 的子实体样本为人工接种段木上生长的成熟耳片; 而毛木耳 APM2-16 的子实体样本是通过代料栽培获得的成熟耳片, 栽培料中含有 80% 木

屑、19%麸皮和 1%石膏,含水量 55%–60%。采集到的新鲜子实体经无菌水洗净表面杂质后立即保存于–80℃超低温冰箱中保存备用。

### 1.3 RNA 的分离提取和转录组测序

联合使用 RNAiso-mate for Plant Tissue、RNAiso Plus 和 High-Salt Solution for Precipitation (TaKaRa) 进行黑木耳菌丝体和子实体 RNA 的分离提取。而毛木耳菌丝体和子实体的 RNA 则使用 TRIzol 试剂盒(Invitrogen Life Technologies) 和 High-Salt Solution for Precipitation (TaKaRa) 进行分离提取。获得的总 RNA 用 Agilent 2100 生物分析仪(Agilent Technologies) 进行 RNA 完整性检测和浓度测定。

符合转录组测序建库要求的总 RNA 在北京六合华大基因科技股份有限公司深圳分公司的 Illumina HiSeq™ 2000 测序平台上完成转录组测序,所用原始数据提交到 NCBI 的 SRA(Sequence Read Archive) 数据库中,黑木耳菌丝体测序数据的登录号为 No. SRX318366,黑木耳子实体的登录号为 No. SRX318367,而毛木耳菌丝体和子实体转录组测序数据的登录号分别是 No. SRX319468 和 No. SRX319472。

每个样品经双末端测序产生的 reads 在去除含接头的序列后用组装软件 SOAPdenovo 进行转录组的从头组装(Li *et al.* 2010)。首先将存在部分重叠的 reads 组装在一起获得重叠群(contig);然后将 reads 比对回重叠群,通过双末端测序的结果确定来自同一转录本的不同重叠群和它们之间的距离,获得含未知序列的 scaffold;再进一步用双末端的信息对 scaffold 作补洞处理,最后产生两端不能再延长的 EST 序列。菌丝体和子实体的 reads 分别完成组装后,再用 TGICL 和 CAP3 对两个样品的 EST 序列

混合后作进一步序列拼接和去冗余处理,最后得到非冗余的 EST 序列(Perteau *et al.* 2003; Huang & Madden 1999)。

### 1.4 SSR 的查找和分析

用软件 SSR Locator 分别对拼接完成的黑木耳和毛木耳非冗余 EST 序列进行单至六碱基重复 SSR 的查找和统计分析(da Maia *et al.* 2008),其中单碱基重复次数至少为 20 次,二碱基重复至少 8 次,三碱基重复至少 5 次,四碱基重复至少 4 次,而五碱基和六碱基的重复次数在 3 次以上。同一条序列中,100bp 范围内包含用非重复碱基连接的多个 SSR 作为复合型 SSR 进行统计。

## 2 结果与分析

### 2.1 转录组测序和拼接

分别提取黑木耳、毛木耳的菌丝体和成熟子实体的总 RNA,经 Agilent 2100 生物分析仪(Agilent Technologies) 检测,当 RIN(RNA Integrity Number) 值≥6.5,浓度达到 100ng/μL 以上时构建代表不同物种不同生育期的 4 个测序文库,在 Illumina HiSeq™ 2000 测序平台上完成转录组测序。每个样本都获得了 1G 以上的原始 reads,通过去除含接头的 reads,进行转录组的从头组装。黑木耳的菌丝体和子实体转录组经组装分别获得 31 233 个和 33 371 个 EST,用 TGICL 和 CAP3 软件对这两个阶段的 EST 作进一步组装最终产生了 32 753 个非冗余 EST,序列总长度约为 19.6Mb, N50 值为 808bp(表 1)。毛木耳的菌丝体和子实体转录组则分别获得 43 941 个和 50 658 个 EST,用同样的方法进一步拼接共获得 48 963 个非冗余 EST,序列总长度约为 19.8Mb,与黑木耳转录组结果相似, N50 值仅有 457bp,远低于黑木耳转录组的 N50 值(表 2)。

表 1 黑木耳 Au916 转录组的基本信息  
Table 1 General features of the *Auricularia auricula-judae* Au916 transcriptome

	菌丝体 Mycelium	子实体 Fruiting body	综合 Total
总原始 reads 数 Number of raw reads	13 333 334	13 333 334	26 666 668
总碱基数 Total nucleotides (bp)	1 200 000 060	1 200 000 060	2 400 000 120
Q20 的比例 <sup>a</sup> Q20 percentage <sup>a</sup>	88.88%	88.00%	-
N 的比例 <sup>b</sup> N percentage <sup>b</sup>	0.00%	0.00%	-
平均 reads 长度 Average reads length (bp)	90	90	90
重叠群数 Number of contigs	138 842	159 208	-
重叠群长度的 N50 值 N50 length of contigs (bp)	192	202	-
Scaffolds 数 Number of scaffolds	44 488	49 896	-
Scaffolds 长度的 N50 值 N50 length of scaffolds (bp)	470	548	-
ESTs 数 Number of ESTs	31 233	33 371	32 753
ESTs 的总长度 Length of ESTs (bp)	13 679 486	16 212 364	19 644 658
ESTs 长度的 N50 值 N50 length of ESTs (bp)	540	645	808
GC 含量 GC content	56.50%	57.67%	-

注：<sup>a</sup>，Q20 的比例表示测序错误率低于 1% 的序列百分比；<sup>b</sup>，N 的比例是测序失败的碱基占所有序列的百分比。  
Note: <sup>a</sup>，Q20 percentage indicates the percentage of sequences at a sequencing error rate lower than 1%; <sup>b</sup>，N percentage indicates the percentage of the nucleotides which could not be sequenced.

表 2 毛木耳 APM2-16 转录组的基本信息

Table 2 General features of the *Auricularia polytricha* APM2-16 transcriptome

	菌丝体 Mycelium	子实体 Fruiting body	综合 Total
总原始 reads 数	12 627 212	13 266 670	25 893 882
Number of raw reads			
总碱基数	1 136 449 080	1 194 000 300	2 330 449 380
Total nucleotides (bp)			
Q20 的比例 <sup>a</sup>	90.88%	90.37%	-
Q20 percentage <sup>a</sup>			
N 的比例 <sup>b</sup>	0.04%	0.01%	-
N percentage <sup>b</sup>			
平均 reads 长度	90	90	90
Average reads length (bp)			
重叠群数	244 024	307 887	-
Number of contigs			
重叠群长度的 N50 值	112	104	-
N50 length of contigs (bp)			
Scaffolds 数	75 041	83 591	-
Number of scaffolds			
Scaffolds 长度的 N50 值	270	284	-
N50 length of scaffolds (bp)			
ESTs 数	43 941	50 658	48 963
Number of ESTs			
ESTs 的总长度	13 864 206	16 448 330	19 803 591
Length of ESTs (bp)			
ESTs 长度的 N50 值	332	346	457
N50 length of ESTs (bp)			
GC 含量	61.27%	61.02%	-
GC content			

注: <sup>a</sup>, Q20 的比例表示测序错误率低于 1% 的序列百分比; <sup>b</sup>, N 的比例是测序失败的碱基占所有序列的百分比。

Note: <sup>a</sup>, Q20 percentage indicates the percentage of sequences at a sequencing error rate lower than 1%; <sup>b</sup>, N percentage indicates the percentage of the nucleotides which could not be sequenced.

## 2.2 SSR 的分布和数量

用软件 SSR Locator 分别对拼接完成的 32 753 个黑木耳非冗余 EST 序列和 48 963 个毛

木耳非冗余 EST 序列进行 SSR 的查找。按查找标准, 从黑木耳转录组中共查找到 1 767 个 SSR 存在于 1 558 条 EST 中, 占整个转录组 EST 总数

的 4.76%，其中 161 条 EST 序列中存在多个 SSR，这些 SSR 中有 83 组 SSR 是复合型 SSR，平均 1Mb 中存在大约 90 个 SSR，所有 SSR 碱基数共 30 755bp，占转录组序列总长度的 0.16%(表 3)。用同样的方法在毛木耳转录组序列中查找到 1 550 个 SSR 存在于 1 445 条 EST，占整个转录组 EST 序列总数的 2.95%，其中 98 条 EST 序列中存在多个 SSR，这些 SSR 中有 56 组 SSR 是复合型 SSR，平均 1Mb 中存在大约 78 个 SSR，所有 SSR 碱基数共 26 956bp，占转录组序列总长度的 0.14%（表 3）。总体而言，黑木耳转录组中 SSR 的数量要略多于毛木耳转录组。

表 3 黑木耳和毛木耳转录组中 SSR 的分布  
Table 3 SSR distribution in the transcriptome of *Auricularia auricula-judae* and *A. polytricha*

分布特征 Characteristic of distribution	黑木耳 <i>A. auricula</i>	毛木耳 <i>A. polytricha</i>
总碱基数 Total nucleotides (bp)	19 644 658	19 803 591
SSR 总量 Number of identified SSR	1 767	1 550
含 SSR 的序列数 Number of SSR containing sequences	1 558	1 445
含多个 SSR 的序列数 Number of sequences containing more than one SSR	161	98
复合 SSR 数量 Number of SSRs present in compound formation	83	56
SSR 总长度 Total SSR length (bp)	30 755	26 956
SSR 密度 SSR density (per Mbp)	90	78

黑木耳和毛木耳转录组中 SSR 的构成和分布是基本相似的（表 4）。在所有的 SSR 中，出现最多的三碱基重复，在黑木耳转录组中有 825 个，占总数的 46.69%，毛木耳转录组中有 741 个，占总数的 47.81%；其次为六碱基重复，在黑木耳和毛木耳中分别含 596 个（33.73%）和 487 个（31.42%）；再次为五碱基重复和四碱基重复；单碱基重复和二碱基重复在黑木耳和毛木耳的转录组中存在差异，黑木耳中以二碱基重复最少，只有 19 个，而毛木耳中则以单碱基重复最少，为 28 个。另一方面，除了黑木耳的三碱基重复中存在 102bp 的 SSR 外，一到六碱基重复的 SSR 在两个物种中普遍较短，平均都为 17.4bp（表 4）。

2.3 SSR 基序的特点和性质

在黑木耳和毛木耳转录组中，SSR 基序的数目和种类也基本相同。在单碱基重复的基序数量上，两个不同种的转录组中是存在差异的，其中黑木耳转录组中以 A/T 重复较多，而 C/G 重复略少一些，而毛木耳中则相反，C/G 重复的 SSR 占单碱基重复的绝大多数（图 1）。在二碱基重复的 SSR 中，TA/TA 基序在两个物种的转录组中均不存在，CG/CG 基序在毛木耳中是特异存在的，而 AT/AT 和 GC/GC 两种基序在毛木耳转录组中是不存在的（图 1）。从数量上而言，在二碱基重复的 SSR 中，黑木耳转录组中不同类型基序的 SSR 均在 5 条左右，AT/AT 和 GC/GC 两种基序只有 1 条；而毛木耳转录组中的变化幅度相对较大，以 GA/TC 基序的 SSR 最多，有 16 条，占有二碱基重复 SSR 的 45.71%，CG/CG 基序则只有 1 条。三碱基重复 SSR 是转录组中含量最丰富的，黑木耳和毛木耳中分别存在 29 和 26 种不同的基序，其中 ACG/CGT、AGC/GCT、CAG/CTG、CCG/CGG、CGA/TCG、CGC/GCG、GAC/GTC、GCA/TGC、GCC/GGC 和 GGA/TCC 等



表 4 黑木耳和毛木耳转录组中 SSR 的数量和长度

Table 4 Abundance and length of SSR in the transcriptome of *Auricularia auricula-judae* and *A. polytricha*

重复单元 Repeat unit	黑木耳 <i>A. auricula-judae</i>			毛木耳 <i>A. polytricha</i>		
	SSR 数量 SSR count	平均长度 Average length (bp)	最大长度 Maximum length (bp)	SSR 数量 SSR count	平均长度 Average length (bp)	最大长度 Maximum length (bp)
单碱基 Mono-	31	24.6	32	28	23.1	40
二碱基 Di-	19	17.2	22	35	19.2	19
三碱基 Tri-	825	16.6	102	741	16.6	39
四碱基 Tetra-	93	17.0	24	75	17.4	32
五碱基 Penta-	203	16.0	30	184	15.5	25
六碱基 Hexa-	596	18.7	30	487	18.9	42
总数 Total	1 767	17.4	102	1 550	17.4	42

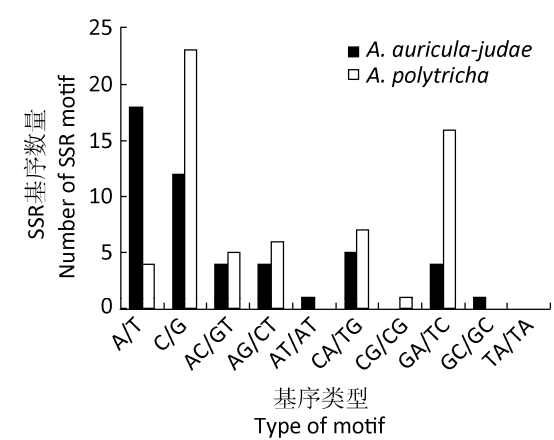


图 1 黑木耳和毛木耳转录组中单碱基重复 SSR 和二碱基重复 SSR 不同基序的频率分布

Fig. 1 Frequencies of different mono- and di-nucleotide repeats SSR motif in the transcriptomes of *Auricularia auricula-judae* and *A. polytricha*.

10 种基序都较多，分别在黑木耳和毛木耳中占这类 SSR 的 84.73%和 86.23%。在黑木耳转录组中以 CAG/CTG 基序的 SSR 数量最多，有 124 个；而毛木耳中则以 CGA/TCG 基序的 SSR 数量最多，有 98 个（图 2）。

黑木耳转录组中，基序为 TGCA/TGTA 和 ATCC/GGAT 的四碱基重复 SSR 在该类 SSR 中最丰富，分别有 5 个；而毛木耳中则以 CCTC/GAGG 和 CGAG/CTCG 最丰富，分别有 6 个。基序为 CGCGC/GCGCG 的五碱基重复 SSR 在黑木耳转录组中最多，共 6 条；而毛木耳中则以 GCCGC/GCGGC 最多，共 7 条。六碱基重复的基序在转录组中是最丰富的，在黑木耳和毛木耳中分别有 362 种和 314 种不同类型。在黑木耳转录组中，基序为 CTTCTC/GAGAAG 的六碱基重





- Dai YC, Yang ZL, 2008. A revised checklist of medicinal fungi in China. *Mycosystema*, 27: 801-824 (in Chinese)
- Dai YC, Zhou LW, Yang ZL, Wen HA, Bau T, Li TH, 2010. A revised checklist of edible fungi in China. *Mycosystema*, 29: 1-21 (in Chinese)
- Dutech C, Enjalbert J, Fournier E, Delmotte F, Barrès B, Carlier J, Tharreau D, Giraud T, 2007. Challenges of microsatellite isolation in fungi. *Fungal Genetics and Biology*, 44(10): 933-949
- Floudas D, Binder M, Riley R, Barry K, Blanchette RA, Henrissat B, Martínez AT, Otillar R, Spatafora JW, Yadav JS, Aerts A, Benoit I, Boyd A, Carlson A, Copeland A, Coutinho PM, de Vries RP, Ferreira P, Findley K, Foster B, Gaskell J, Glotzer D, Górecki P, Heitman J, Hesse C, Hori C, Igarashi K, Jurgens JA, Kallen N, Kersten P, Kohler A, Kües U, Kumar TK, Kuo A, LaButti K, Larrondo LF, Lindquist E, Ling A, Lombard V, Lucas S, Lundell T, Martin R, McLaughlin DJ, Morgenstern I, Morin E, Murat C, Nagy LG, Nolan M, Ohm RA, Patyshakuliyeva A, Rokas A, Ruiz-Dueñas FJ, Sabat G, Salamov A, Samejima M, Schmutz J, Slot JC, St John F, Stenlid J, Sun H, Sun S, Syed K, Tsang A, Wiebenga A, Young D, Pisabarro A, Eastwood DC, Martin F, Cullen D, Grigoriev IV, Hibbett DS, 2012. The paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science*, 336(6089): 1715-1719
- Horgen PA, Kokurewicz KF, Anderson JB, 1989. The germination of basidiospores from commercial and wild collected isolates of *Agaricus bisporus* (= *A. brunnescens*). *Canadian Journal of Microbiology*, 35(4): 492-498
- Huang X, Madan A, 1999. CAP3: a DNA sequence assembly program. *Genome Research*, 9(9): 868-877
- Labbé J, Murat C, Morin E, Le Tacon F, Martin F, 2011. Survey and analysis of simple sequence repeats in the *Laccaria bicolor* genome, with development of microsatellite markers. *Current Genetics*, 57(2): 75-88
- Li CY, Liu L, Yang J, Li JB, Su Y, Zhang Y, Wang YY, Zhu YY, 2009. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. *Interdisciplinary Sciences: Computational Life Sciences*, 1(2): 141-150
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J, 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20(2): 265-272
- Metzgar D, Bytof J, Wills C, 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research*, 10(1): 72-80
- Perlea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J, 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5): 651-652
- Powell W, Machray GC, Provan J, 1996. Polymorphism revealed by simple sequence repeats. *Trends in Plant Science*, 1(7): 215-222
- Qian J, Xu H, Song J, Xu J, Zhu Y, Chen S, 2013. Genome-wide analysis of simple sequence repeats in the model medicinal mushroom *Ganoderma lucidum*. *Gene*, 512(2): 331-336
- Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P, 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany*, 99(2): 193-208
- [附中文参考文献]
- 戴玉成, 李玉, 2011. 中国六种重要药用真菌名称的说明. 菌物学报, 30: 516-518
- 戴玉成, 杨祝良, 2008. 中国药用真菌名录及部分名称的修订. 菌物学报, 27: 801-824
- 戴玉成, 周丽伟, 杨祝良, 文华安, 图力古尔, 泰辉, 2010. 中国食用菌名录. 菌物学报, 29: 1-21