绒毛状烟草和林烟草全长 cDNA 文库构建及 EST 序列分析

孙 榕[1,2]，陈明丽[1]，解敏敏[1]，孙玉合[1]，龚达平[1*]

（1.中国农业科学院烟草研究所，青岛 266101；2.中国农业科学院研究生院，北京 100081）

摘 要：表达序列标签（EST）广泛应用于基因功能研究和分子标记开发。以普通烟草两个二倍体祖先种绒毛状烟草（*Nicotiana tomentosiformis*, TT）和林烟草（*Nicotiana sylvestris*, SS）多个组织为试验材料，使用 CloneMiner cDNA 文库构建方法构建了均一化全长 cDNA 文库，测序并进行序列拼接、功能注释、进化分析和标记开发。绒毛状烟草和林烟草均一化全长 cDNA 文库容量分别为 0.72×10^6 和 1.12×10^6 pfu/mL，重组率分别约为 94%和 93%，插入片段平均长度为 1.4 kb。测序获得 20 953 条 EST 序列，拼接为 10 504 个 unigenes。与普通烟草 EST 序列混合拼接，产生 34 450 条 contigs，123 511 条 singletons，烟草异源四倍体中 T 和 S 基因与绒毛状烟草、林烟草之间的相似性远高于两个二倍体祖先种之间。预测获得 104 915 个编码序列，其中 73 670 个序列包含功能结构域，81% unigenes 在番茄中具有同源基因。鉴定了 11 869 个微卫星位点（SSR）和 25 209 个单核苷酸多态性位点（SNP）。这些数据信息对于烟草基因功能研究和分子育种具有重要价值。

关键词：烟草；cDNA 文库；EST；SNP；SSR

中图分类号：S572.03 **文章编号：**1007-5119（2018）05-0009-08 DOI：10.13496/j.issn.1007-5119.2018.05.002

Full-Length cDNA Library Construction of *Nicotiana tomentosiformis* and *Nicotiana sylvestris* and ESTs Analysis of Tobacco

SUN Rong[1,2], CHEN Mingli[1], XIE Minmin[1], SUN Yuhe[1], GONG Daping[1*]

(1. Tobacco Research Institute of CAAS, Qingdao 266101, China; 2. Graduate School of Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: Expressed sequence tags (EST) are widely used in gene function research and molecular marker development. In order to obtain a large number of EST sequences of tobacco, a variety of tissues and organs from *Nicotiana tomentosiformis* and *Nicotiana sylvestris* were taken as plant materials, and two full-length enriched cDNA libraries were constructed using the CloneMiner cDNA method. The EST sequences were used for sequence assembly, functional annotation, phylogenetic analysis and molecular marker development. The normalized full-length cDNA libraries were constructed successfully from *Nicotiana tomentosiformis* and *Nicotiana sylvestris*. The titer were 0.72×10^6 and 1.12×10^6 pfu/mL, respectively, and the recombination rates were approximately 94% and 93%, respectively. The average length of inserted cDNA fragments was 1.4 kb. 20 953 ESTs were generated from the full-length enriched cDNA libraries, and assembled into 10 504 unigenes. All of the ESTs from allopolyploid tobacco (*Nicotiana tabacum*) and its two diploid progenitors, *Nicotiana tomentosiformis* and *Nicotiana sylvestri*s were assembled, resulting in 34 450 contigs and 123 511 singletons. The global assembly showed that the transcripts from the resident T- and S-genomes in the allotetraploid nucleus were more closely related to their diploid homologs than they were to each other. In total, 104 915 coding sequences were identified, of which 73 670 sequences contained functional domains. Approximately 81% of the unigenes had homologs in tomato. Furthermore, we found 11 869 putative simple sequence repeats (SSR) and 55 369 single nucleotide polymorphisms (SNPs). The EST resources have important implications for gene function research and molecular breeding.

Keywords: tobacco; full-length cDNA; EST; SNP; SSR

烟草属于茄科植物，不仅是重要的经济作物，也是研究异源多倍体的重要模式物种[1-3]。普通烟草为异源四倍体（2n=48），可能是由两个二倍体野生烟草绒毛状烟草（*N. tomentosiformis*，TT, 2n=24）

基金项目：中国农业科学院烟草研究所青年基金项目“烟草腺毛 NtTe 基因的精细定位”（2015B01）；
中国烟草总公司烟草基因组计划重大专项项目“K326 和红花大金元蚜虫抗性的定向改良”（110201801024JY-01）

作者简介：孙 榕（1994-），女，硕士研究生，主要从事烟草遗传育种研究。E-mail：934928377@qq.com。*通信作者，E-mail：gongdaping@caas.cn

收稿日期：2018-05-05 **修回日期：**2018-08-14

和林烟草（*N. sylvestris*, SS, 2n=24）种间杂交后通过染色体加倍形成[4-5]。普通烟草基因组大小约 4.5 G[6]，重复序列比例高[7-8]，具有高度的复杂性。基于 TGI 的基因标签数据，BINDLER 等[9]开发 SSR 标记构建了一张高密度的普通烟草遗传图谱。WU 等[10]利用单拷贝的直系同源基因（single-copy conserved ortholog set，COSII）和简单重复序列（simple sequence repeat，SSR）标记，比较了两个二倍体祖先烟草和普通烟草的遗传图谱，发现普通烟草相比二倍体经历了更多的染色体重排。最近，采用二代测序技术相继完成了林烟草、绒毛状烟草和普通烟草的基因组测序[11-14]。

全长 cDNA 序列分析是获得基因全长信息的最直接有效的一种方法。全长 cDNA 克隆具有完整的转录序列, 包括编码区（CDSs）和非翻译区（UTRs），这有助于基因组测序完成后基因的准确预测。烟草 EST 序列信息是开展基因功能研究[15-16]、分子标记开发[17]、基因芯片开发[18]等的重要基础。美国烟草基因组计划、欧洲烟草测序项目组和日本烟草公司等构建了普通烟草品种不同发育时期的多个 cDNA 文库, 分别获得 80 783 条、58 969 条和 65 102 条 EST 序列。二倍体祖先烟草的基因组草图已经测序完成, 但是关于全长转录本的信息还很少，因此开展烟草全长 cDNA 研究具有重要价值。本研究构建了两个二倍体祖先种绒毛状烟草和林烟草的全长 cDNA 文库，高效测序并获得了大量 EST，利用生物信息学工具结合普通烟草 EST 进行了拼接, 并对序列分化和基因功能进行了分析, 开发了大量的 SNP 和 SSR 标记，为烟草的基因功能研究奠定了基础。

1 材料与方法

1.1 cDNA 文库构建

绒毛状烟草和林烟草种植在中国农业科学院烟草研究所温室。收集幼苗、根、茎、叶和花等组织储存在–80 ℃备用。使用 Trizol 提取(Invitrogen)各组织总 RNA，全长 cDNA 采用 GenRacerTM 试剂盒富集(Invitrogen)。cDNA 文库使用 CloneMiner II 试剂盒 (Invitrogen) 构建并插入到载体 pDONR222，再通过基因组饱和杂交，构建均一化文库。为了评估文库的质量，随机挑取 60 个克隆通过载体特异性引物 PCR 扩增检测, 通过琼脂糖凝胶电泳估计插入片段大小。

1.2 EST 测序、组装和功能注释

随机挑取 cDNA 克隆在 ABI3730DNA 分析仪（Applied Biosystems 公司）上从 5′进行单向测序。利用 PHRED[19]将峰值信息转换为序列文件。利用 LUCY 和 Seqclean 程序过滤低质量区域、载体序列和 poly(A)序列，得到高质量序列[20]。使用 CAP3 程序在 95%以上相似性和 50 bp 以上重叠的参数下进行聚类拼接[21]。将起始密码子上游的序列定义为 5′-UTRs。使用 BLASTN 软件将 unigenes 与 Sol 数据库中绒毛状烟草和林烟草的 CDS 序列进行比对，分析 unigenes 的 5′-UTR 长度。所有的 unigenes 使用 ESTScan 程序搜索开放阅读框（open reading frames, ORF）[22-23]。采用 interproscan 程序进行 GO 注释[24]。拟南芥基因的 GO 注释信息来自 TAIR 数据库[25]。

1.3 SSR 标记和 SNP 标记挖掘

使用 MISA 程序分析微卫星序列（microsatellite, SSR）。重复单元为 2~6 个核苷酸碱基，最小重复次数分别是：2 碱基重复单元为 6 次，3 碱基重复单元为 5 次，4 碱基重复单元为 4 次，5 碱基重复单元为 3 次, 6 碱基重复单元为 3 次。两个相邻的 SSR 重复单元之间间隔最大长度设置为 100 bp。利用 Perl 脚本 AutoSNP 对 CAP3 拼接产生的比对序列进行 SNP 鉴定[26]。

2 结 果

2.1 cDNA 文库构建和 EST 测序组装

利用不同组织的混合 RNA, 构建了二倍体野生烟草绒毛状烟草和林烟草的全长均一化文库，文库容量分别为 0.72×10^6 和 1.12×10^6 pfu/mL，重组率分别约为 94%和 93%，文库插入片段平均大小约为 1.4 kb。随机挑选 22 525 个克隆进行 5′末端测序，共产生 20 953 条 EST 序列。EST 序列的长度在 100～

908 bp 之间，平均长度为 671 bp。其中 10 450 条 EST 序列来源于绒毛状烟草（GenBank 序列号：JZ959365-JZ969801），10 503 条 EST 序列来源于林烟草（GenBank 序列号：JZ948884-JZ959364）。通过聚类拼接，绒毛状烟草获得 4805 个 unigenes，包括 1202 个 contigs 和 3603 个 singletons ；林烟草获得 5699 个 unigenes，包括 1410 个 contigs 和 4289 个 singletons。其中，88%的绒毛状烟草 unigenes 和 84%林烟草 unigenes 序列中 EST 序列数量都少于 6 条, 说明均一化文库的冗余度相对较低。将 unigenes 与基因组中预测的 CDS 序列比较来鉴定 UTR 区域。绒毛状烟草和林烟草中，2917 个基因和 3838 个基因具有完整的 ATG 起始序列，UTR 长度分布如图 1，大多数基因的 UTR 长度小于 200。

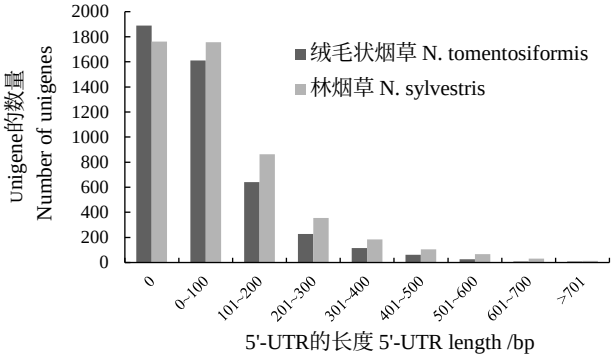

图 1　Unigenes 5′**非翻译区长度分布**

Fig. 1　Distributions of 5′-UTR lengths of unigenes

2.2　EST **组装**

从 Genbank 数据库中, 搜集获得了普通烟草 40 多个文库共计 334 382 条 EST 序列，多于 1000 条 EST 序列的文库信息如表 1[15-16,18,27-30]。这些文库来源包括烤烟品种 K326、Hicks Broadleaf、白肋烟 Burley 21、TN86、香料烟 Samsun 和雪茄烟 Petit Havana 等在不同条件下的根、花、叶、腺毛、花等组织和细胞。

将普通烟草与林烟草和绒毛状烟草共计 347 022 条 EST 序列进行组装，获得了 34 450 条 contigs 和 123 511 条 singletons（共 157 961 个 unigenes）。contigs 的长度从 107 到 3502 bp, 平均长度为 879.6 bp。每个 contig 包含的 EST 数量平均为 14.8 条（最多的为 1158 条），平均每个碱基的覆盖度为 4.4 倍。

3 个烟草属物种混合组装构成的 contig 数量分布如图 2。普通烟草、绒毛状烟草和林烟草共有 EST 成员的 contig 是 376 条；绒毛状烟草与普通烟草共同组成的 contig 是 3103 条；林烟草与普通烟草共同组成的 contig 是 2761 条 ;普通烟草特异的 contigs 有 28 210 条。二倍体烟草没有特异的 contig，说明二倍体烟草没有与普通烟草分化程度很大的特异基因序列。这些结果表明，普通烟草中 T 基因组和 S 基因组与二倍体基因组的相似程度远比它们之间的相似程度更高。但是，在绒毛状烟草和林烟草中仍然分别有 2904 和 2225 条 EST 序列属于单一序列, 没有与普通烟草的 EST 构成 contig。可能与文库质量、表达差异和测序错误等原因造成的文库偏好有关。

2.3　**基因预测、功能注释及与其他植物基因的比较分析**

通过 ESTScan 程序在 104 915 个 unigenes（66.4%）中发现了开放阅读框（ORF），平均长度为 524 bp(最小片段为 150 bp, 最长片段为 3486 bp)。这些预测的烟草 ORF 序列长度相比基因组预测的 CDS 长度要短, 可能是没有拼接出基因全长造成的。小部分 unigene 可以与 NR 数据库有较好的比对结果, 说明 ESTScan 并没有鉴定准确, 可能与 ESTScan 程序使用拟南芥的参数有关。

在 104 915 个 ORF 序列中，73 670 个蛋白产物至少包含一个注释的 Pfam 结构域。其中，出现次数最多的结构域是 Pkinase (Pfam: PF00069)，这是一个包含蛋白激酶催化功能的结构保守的蛋白结构域。其他的结构域包括 RNA 识别基序（RNA recognition motif）、亮氨酸重复序列（leucine rich repeat）、WD40 重复序列（WD40 repeat）、锌指结构域（Zinc finger domains）[32]和细胞色素 P_{450}。大量研究表明，这些结构域在其他植物中也广泛存在。GO 注释将基因功能分为细胞组分、分子功能和生

表 1 烟草 EST 文库信息

Table 1 Summary of tobacco EST libraries

文库编号 Lib. ID	品种信息 Tobacco accession	EST 数量 Numbers of ESTs	类型 Nature	文库组织描述 Library description
Lib.14090	Bright Yellow 2	13 700		延滞期，对数期，稳定期细胞
Lib.16492	Bright Yellow 2	5512		激素处理的细胞
Lib.20805	Petit Havana	2668		继代后第 3 天的细胞
Lib.16263		5927	均一化	根，茎，叶
Lib.18153	K326	4361		萌发的种子
Lib.18173	K326	4692		萌发后 2 周的幼苗
Lib.18782	K326	4540		打顶前的叶片
Lib.18783	K326	5288		打顶后的根
Lib.19510	Burley 21	1938		打顶后的叶片
Lib.19511	K326	4422		打顶前的花
Lib.19512	K326	3755		打顶后叶片
Lib.19513	K326	5150		打顶前的叶脉
Lib.19514	K326	4351		打顶前的根
Lib.19515	K326	4844		打顶前的茎
Lib.19516	TN86	3205		打顶后叶片
Lib.23331	K326, Samsun, Burley 21	2145		叶片
Lib.23332	Samsun	3183		腺毛
Lib.23333	Burley 21	1950		腺毛
Lib.23334	K326	1163		腺毛
Lib.23335	K326	1690		成熟早期叶片
Lib.23336	K326	2308		成熟晚期叶片
Lib.21783	Samsun NN	28 477	均一化	萌发后 5 d 的幼苗
Lib.21784	Samsun NN	22 418	均一化	5 周后的整株
Lib.21785	Samsun NN	18 052	均一化	5 周后冷胁迫下的整株
Lib.23036	Hicks Broadleaf	8264		叶
Lib.23037	Hicks Broadleaf	29 526	均一化全长	根
Lib.23038	Hicks Broadleaf	7845		根
Lib.23040	Hicks Broadleaf	24 044	均一化全长	根、花、叶
Lib.20776	Hicks Broadleaf	8048		成熟叶片
Lib.20778	Hicks Broadleaf	1500		冷胁迫
Lib.20779	Hicks Broadleaf	1085		花
Lib.23330	Petit Havana	11 216		柱头
Lib.24954		65 102	均一化全长	移栽 11、24、48 d 后的叶、顶芽、根
Lib.26713	SR1	2356		合子
Lib.26715	SR1	2149		二胞原胚
Lib.26714	SR1	1964		卵细胞
Lib.26545	SR1	1864		精细胞

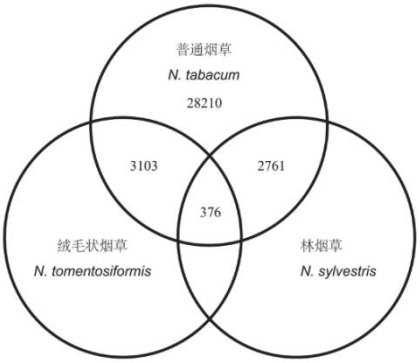

图 2 普通烟草和绒毛状烟草、林烟草重叠群数据维恩图

Fig. 2 Venn diagram of the contig contents of the three *Nicotiana* species

物学过程 3 个部分和各个小类。将烟草和拟南芥的 GO 注释结果比较发现，整体上两个物种的基因功能分类是相似的，但在辅助转运蛋白(auxiliary transport protein)、蛋白标签(protein tag)、金属伴侣(metallochaperone)等功能类别中具有显著差异（图 3）。将预测的 104 915 个烟草基因的氨基酸序列与拟南芥、水稻、杨树、大豆、葡萄和番茄的蛋白质数据库进行比对分析（E 值>1E-10）。结果表明（图 4），预测的烟草基因氨基酸序列与茄科植物，特别

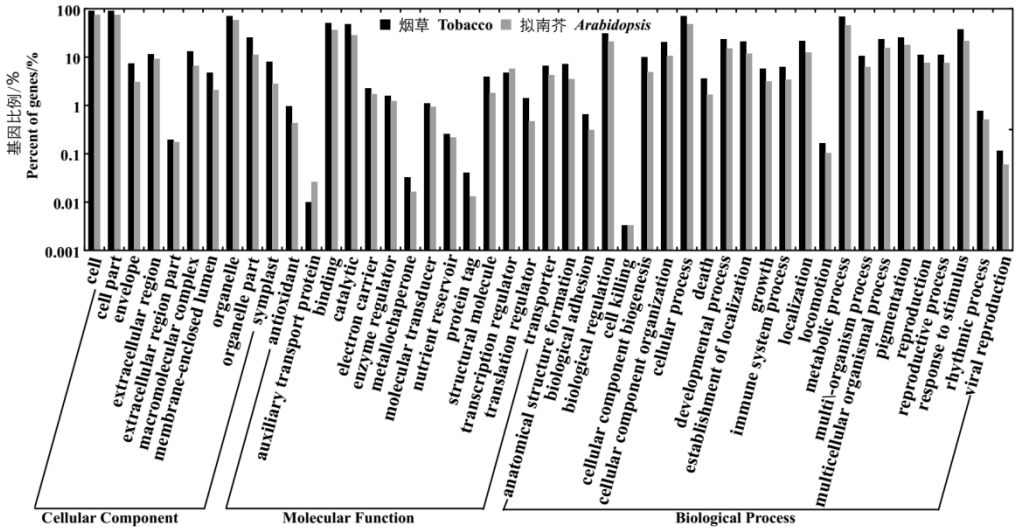

图 3　烟草和拟南芥基因 GO 注释比较

Fig. 3　Profile of GO annotations for tobacco and Arabidopsis

是番茄基因具有高度的相似性，仅有 19%的烟草蛋白序列在番茄中没有相似序列；而在水稻中烟草有 33%的蛋白没有序列相似性。图 4 的曲线也反应出了烟草与其他物种的分化程度，与同属于茄科的番茄关系最近。

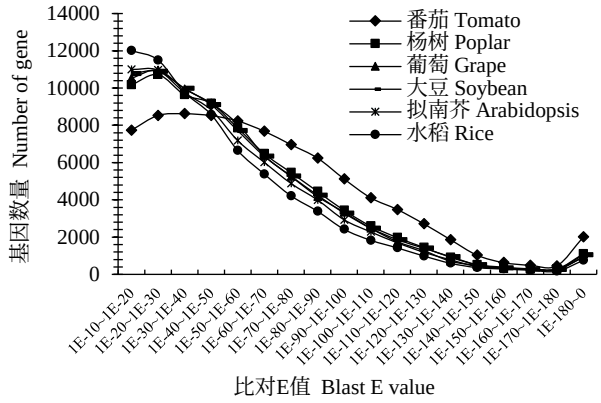

图 4　烟草蛋白序列与其他植物同源性比较

Fig 4　Comparison of deduced tobacco peptide sequences with genes of other plants

2.4　SSR 标记的鉴定

从 152 891 个 unigenes 中共鉴定了 10 424 个 SSR 位点。绝大多数序列仅含有 1 个 SSR 位点，1209 个基因具有多个 SSR 位点。共计开发了 11 869 个烟草 EST-SSR 标记，包括二核苷酸、三核苷酸、四核苷酸、五核苷酸和六核苷酸的 SSR 类型。其中，945 个 SSR 位点是复合形式。二核苷酸和三核苷酸的 SSR 类型为主要类型，共占 72.51%。二核苷酸重复序列有 4434 条、三核苷酸重复序列有 4054 条、四核苷酸重复序列有 516 条、五核苷酸重复序列有 1273 条，六核苷酸重复序列有 1592 条。烟草 EST-SSR 重复单元类型统计发现，在二核苷酸 SSRs 中 AG/CT 类型的重复序列最多，占二核苷酸重复序列的 56.2%。在三核苷酸 SSRs 中 AAG/CTT 类型的重复序列是最丰富的，占总数的 38.9%（表 2）。

重复单元重复次数最多的是二核苷酸 SSRs，AG/CT 类型的重复单元最大重复数量为 53 次(表 3)。

表 2　SSR 重复单元的频率

Table 2　SSR motifs and their frequency

SSR 类型	SSR 基元	SSR 数量
SSR	SSR Motif	SSR Number
Di	AC/GT	1173
	AG/CT	2494
	AT/AT	758
	CG/CG	9
	Subtotal	4434
Tri	AAC/GTT	648
	AAG/CTT	1576
	AAT/ATT	301
	ACC/GGT	247
	ACG/CGT	50
	ACT/AGT	167
	AGC/CTG	320
	AGG/CCT	281
	ATC/ATG	293
	CCG/CGG	171
	Subtotal	4054
Tetra		516
Penta		1273
Hexa		1592

表 3 烟草 SSR 不同重复单元和重复长度的频率分布

Table 3 Frequency distribution of SSRs in tobacco by motif and repeat length

类型	重复次数 Numbers of repeats																										
Type	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
AG	—	846	505	322	201	161	81	83	36	50	30	27	16	35	15	12	6	5	6	6	5	5	1	10	3	4	1	2
AC	—	443	245	203	105	51	46	11	15	24	6	7	3	1	2	1	0	0	1	0	2	0	1	0	2	1	2	1
AT	—	249	105	87	56	38	31	18	16	10	10	11	16	6	6	4	7	6	7	7	8	6	2	9	5	5	7	0
CG	—	6	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAC	408	142	51	28	11	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AAG	856	382	114	59	44	30	12	12	7	10	7	6	5	9	1	4	1	0	2	1	1	2	1	0	0	5	0	1
AAT	142	66	23	16	16	6	1	2	8	6	4	1	3	1	0	0	0	3	0	3	0	0	0	0	0	0	0	0
ACC	155	59	22	9	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ACG	36	7	5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ACT	85	58	18	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGC	185	85	33	9	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGG	173	77	16	6	5	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATC	187	63	16	10	6	7	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCG	75	64	22	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

本研究获得的烟草 EST 序列共计 94.75 Mb，SSR 的密度为 10 kb 分布 12.5 个 SSRs, 频率约为 6.8%。这些新的标记可用于遗传连锁图谱构建和比较基因组研究。

2.5 SNP 标记的鉴定

从 33 734 个 contigs 中共鉴定 103 027 个 SNP 和 24 760 个插入/删除（insertions/deletions，indels）位点。测序错误导致基于 EST 序列鉴定的 SNP 可信度比较低，而具有较高冗余度和共分离值的 SNP 最有可能是真实的遗传变异[31-32]。使用两个或两个以上的冗余评分进行过滤后，获得 42 355 个 SNP 标记和 11 014 个 InDel 标记, 其中 SNP 标记类型包括 26 294 个转换和 16 061 个颠换。大多数过滤后的 SNPs 是从包含 4 个以上序列的 contigs 中鉴定的。由于普通烟草是异源四倍体，基因组包含许多高度同源的基因。通过删除 SNP 频率大于 20 SNP/kb 的位点，共获得 25 209 个 SNP。

3 讨 论

由于烟草是异源四倍体，基因组重复序列比例高, 2013 年烟草科研工作者采用新一代测序技术结合 BAC-to-BAC 策略完成了普通烟草及其二倍体祖先绒毛状烟草和林烟草的基因组测序。但对烟草基因组中基因的预测和功能的注释远落后于基因组测序。烟草 EST 数据是烟草基因组进行基因预测和功能注释的重要信息，目前还没有开展绒毛状烟草和林烟草全长 cDNA 文库构建及 EST 测序的研究报道。本研究采用 CloneMiner cDNA 文库构建方法构建了两个二倍体祖先绒毛状烟草和林烟草的均一化全长 cDNA 文库。相比转录组测序，更有利于发现基因的全长信息，为研究烟草基因功能提供了一条重要途径。此外，全长基因是评估烟草基因组测序组装质量的重要指标之一。本研究获得的绒毛状烟草和林烟草 EST 数据有效的评估了中国烟草基因组计划组装拼接的二倍体烟草基因组，基本覆盖了所有基因[12-14]。

通过 EST 测序和拼接, 绒毛状烟草和林烟草分别获得 4805 个和 5699 个 unigenes。与烟草基因组预测的 7 万个基因相比[12-14]，本研究获得的基因数量还相对有限，两个全长 cDNA 文库还需要挑选更多的克隆进行 EST 测序。Genbank 数据库中，报道了大量的普通烟草 EST 数据, 为烟草基因组提供了大量的基因信息。通过与普通烟草 EST 序列混合拼接，获得了 157 961 个 unigenes，预测获得 104 915 个编码序列。这些数据对于烟草基因组的基因预测和功能研究具有非常重要的价值。

分子标记在烟草基因定位和品种改良上发挥着重要作用[33-36]。本研究鉴定了 11 869 个微卫星位点（SSR）和 25 209 个单核苷酸多态性位点（SNP）。相比传统的分子标记, SSR 和 SNP 标记具有数量多、分布广泛的优点。目前，最高密度的两个烟草遗传图谱分别是用 SSR 和 SNP 标记构建的[17,37]。本研

究基于 EST 数据鉴定的大量 SSR 和 SNP 标记，为开展遗传图谱构建、基因定位克隆和分子标记辅助育种打下了良好基础。

4 结　论

本研究构建了烟草二倍体祖先种绒毛状烟草和林烟草的均一化全长 cDNA 文库，获得了超过 2 万条EST序列信息和10 504个unigenes, 并且64.3%的 unigenes 具有 5'-UTR。将这些序列与以前报道的普通烟草 EST 序列进行组装和基因注释，解析了异源四倍体同源基因间的分化程度，鉴定了 10 424 个 SSR 标记和 25 209 个 SNP/indel 标记。这些数据为烟草基因组注释、基因功能研究和分子育种工作提供了非常有价值的信息。

参考文献

[1] LAYTEN DAVIS D, NIELSEN M T: Tobacco: production, chemistry and technology[M]. Malden: Blackwell Science Ltd, 1999.

[2] MUELLER L A, SOLOW T H, TAYLOR N, et al. The SOL genomics network: a comparative resource for solanaceae biology and beyond[J]. Plant Physiology, 2005, 138(3): 1310-1317.

[3] TREMBLAY R, WANG D, JEVNIKAR A M, et al. Tobacco, a highly efficient green bioreactor for production of therapeutic proteins[J]. Biotechnology Advances, 2010, 28(2): 214-221.

[4] LIM K Y, MATYASEK R, KOVARIK A, et al. Genome evolution in allotetraploid Nicotiana[J]. Biological Journal of the Linnean Society, 2004, 82(4): 599-606.

[5] CLARKSON J J, LIM K Y, KOVARIK A, et al. Long-term genome diploidization in allopolyploid Nicotiana section Repandae (Solanaceae) [J]. The New Phytologist, 2005, 168(1): 241-252.

[6] ARUMUGANATHAN K, EARLE E D. Nuclear DNA content of some important plant species[J]. Plant Molecular Biology Reporter, 1991, 9(3): 208-218.

[7] ZIMMERMAN J L, GOLDBERG R B. DNA sequence organization in the genome of *Nicotiana tabacum*[J]. Chromosoma, 1977, 59(3): 227-252.

[8] KENTON A, PAROKONNY A S, GLEBA Y Y, et al. Characterization of the *Nicotiana tabacum L.* genome by molecular cytogenetics[J]. Molecular and General Genetics MGG, 1993, 240(2): 159-169.

[9] BINDLER G, PLIESKE J, BAKAHER N, et al. A high density genetic map of tobacco (*Nicotiana tabacum L.*) obtained from large scale microsatellite marker development[J]. Theoretical and Applied Genetics, 2011, 123(2): 219-230.

[10] WU F, EANNETTA N T, XU Y, et al. COSII genetic maps of two diploid Nicotiana species provide a detailed picture of synteny with tomato and insights into chromosome evolution in tetraploid *N. tabacum*[J]. Theoretical and Applied Genetics, 2010, 120(4): 809-827.

[11] RENNY-BYFIELD S, CHESTER M, KOVARIK A, et al. Next generation sequencing reveals genome downsizing in allotetraploid Nicotiana tabacum, predominantly through the elimination of paternally derived repetitive DNAs[J]. Molecular Biology and Evolution, 2011, 28(10): 2843-2854.

[12] SIERRO N, BATTEY J N, OUADI S, et al. The tobacco genome sequence and its comparison with those of tomato and potato[J]. Nature Communications, 2014, 5: 3833.

[13] SIERRO N, BATTEY J N, OUADI S, et al. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*[J]. Genome Biology, 2013, 14(6): R60.

[14] 中国烟草基因组重大专项首席科学家团队. 战略与机遇：迈进烟草基因组时代[J]. 中国烟草学报，2017，23（3）：8-13.

CHIEF SCIENTIST TEAM OF CHINA TOBACCO GENOME PROJECT. Strategy and opportunity: striding into tobacco genome era[J]. Acta Tabacaria Sinica, 2017, 23(3): 8-13.

[15] QUIAPIM A C, BRITO M S, BERNARDES L A S, et al. Analysis of the *Nicotiana tabacum* stigma/style transcriptome reveals gene expression differences between wet and dry stigma species[J]. Plant Physiology, 2009, 149(3): 1211-1230.

[16] ZHAO J, XIN H, QU L, et al. Dynamic changes of transcript profiles after fertilization are associated with de novo transcription and maternal elimination in tobacco zygote, and mark the onset of the maternal-to-zygotic transition[J]. The Plant Journal, 2011, 65(1): 131-145.

[17] BINDLER G, HOEVEN R V D, GUNDUZ I, et al. A microsatellite marker based linkage map of tobacco[J]. Theoretical and Applied Genetics, 2007, 114(2): 341-349.

[18] EDWARDS K D, BOMBARELY A, STORY G W, et al. TobEA: an atlas of tobacco gene expression from seed to senescence[J]. BMC Genomics, 2010, 11(1): 142.

[19] EWING B, HILLIER L, WENDL MC, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment[J]. Genome Research, 1998, 8(3): 175-185.

[20] LI S, CHOU H H. LUCY2: an interactive DNA sequence quality trimming and vector removal tool[J]. Bioinformatics, 2004, 20(16): 2865-2866.

[21] HUANG X, MADAN A. CAP3: A DNA sequence assembly program[J]. Genome Research, 1999, 9(9): 868-877.

[22] ISELI C, JONGENEEL C V, BUCHER P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences[C]. Heidelberg: the Seventh International Conference on Intelligent Systems For Molecular Biology, 1999, 99: 138-

148.

[23] LOTTAZ C, ISELI C, JONGENEEL C V, et al. Modelingsequencing errors by combining Hidden Markov models[J]. Bioinformatics, 2003, 19(2): 103-112.

[24] QUEVILLON E, SILVENTOINEN V, PILLAI S, et al. InterProScan: protein domains identifier[J]. Nucleic Acids Research, 2005, 33: W116-120.

[25] CAMON E, BARRELL D, LEE V, et al. The Gene Ontology Annotation (GOA) Database-an integrated resource of GO annotations to the UniProt Knowledgebase[J]. Silico Biology, 2004, 4(1): 5-6.

[26] BARKER G, BATLEY J, O'SULLIVAN H, et al. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP[J]. Bioinformatics, 2003, 19(3): 421-422.

[27] MATSUOKA K, DEMURA T, GALIS I, et al. A comprehensive gene expression analysis toward the understanding of growth and differentiation of tobacco BY-2 cells[J]. Plant & Cell Physiology, 2004, 45(9): 1280-1289.

[28] 李文正，宋利民，李永平，等. 烟草 EST 数据库构建及 cDNA 阵列技术建立[J]. 农业生物技术学报, 2008, 16（4）：662-669..

LI W Z, SONG L M, LI Y P, et al. Construction of EST database and establishment of cDNA array[J]. Journal of Agricultural Biotechnology, 2008, 16(4): 662-669.

[29] LEIN W, USADEL B, STITT M, et al. Large-scale phenotyping of transgenic tobacco plants (*Nicotiana tabacum*) to identify essential leaf functions[J]. Plant Biotechnology Journal, 2008, 6(3): 246-263.

[30] GADANI F, HAYES A, OPPERMAN CH, et al. Large scale genome sequencing and analysis of *Nicotiana tabacum*: the tobacco genome initiative[C]. Bergerac: 5th Bergerac Tobacco Scientific Meeting, 2003: 117-130.

[31] BATLEY J, BARKER G, O'SULLIVAN H, et al. Mining for single nucleotide polymorphisms and insertions/ deletions in maize expressed sequence tag data[J]. Plant Physiology, 2003, 132(1): 84-91.

[32] WANG S, SHA Z, SONSTEGARD T S, et al. Quality assessment parameters for EST-derived SNPs from catfish[J]. BMC Genomics, 2008, 9(1): 450.

[33] 朱承广，任民，蒋彩虹，等. 以关联分析发掘烟草抗赤星病基因分子标记[J]. 中国烟草科学, 2017, 38（1）：68-72.

ZHU C G, REN M, JIANG C H, et al. Identification of molecular markers for tobacco brown spot resistant genes through association analysis [J]. Chinese Tobacco Science, 2017, 38(1): 68-72.

[34] 肖炳光. 烟草基因组计划进展篇：3.烟草分子标记遗传连锁图构建和重要抗病基因定位[J]. 中国烟草科学, 2013, 34（3）：118-119.

XIAO B G. Construction of molecular marker genetic linkage map and location of important disease resistance gene in tobacco[J]. Chinese Tobacco Science, 2013, 34(3): 118-119.

[35] 文轲，张志明，任民，等. 烤烟 CMV 抗性基因 QTL 定位[J]. 中国烟草科学，2013，34（3）：55-59.

WEN K, ZHANG ZH M, REN M, et al. QTL analysis of the resistance gene to CMV in flue-cured tobacco[J]. Chinese Tobacco Science, 2013, 34(3): 55-59.

[36] 龚达平, 王鲁, 李凤霞, 等. 基于 EST 序列的烟草 cSNP 发掘[J]. 中国烟草科学，2012，33（6）： 61-65.

GONG D P, WANG L, LI F X, et al. Tobacco cSNP mining based on expressed sequence tag[J]. Chinese Tobacco Science, 2012, 33(6): 61-65.

[37] GONG D, HUANG L, XU X, et al. Construction of a high-density SNP genetic map in flue-cured tobacco based on SLAF-seq[J]. Molecular Breeding, 2016, 36(7): 1-12.