大刍草苗期转录组 RNA-Seq 数据的 *de novo* 拼接

肖之夏　郑用琏

华中农业大学作物遗传改良国家重点实验室,武汉 430070

摘要　采用 solexa 测序技术,对大刍草苗期全株各组织转录组进行了 RNA-Seq 测序、*de novo* 拼接和信息比对研究。结果表明:转录组测序共得到了 46.4 GB 的原始数据,归并整理后获得长 76 bp 的序列有 175 101 250条,经质量控制和 *de novo* 拼接后,共获得了 58 147 条大刍草转录本,其平均长度为 1 335 bp。比对分析发现其中 94.3%的转录本和玉米 B73 自交系的 cDNA 序列有较好的匹配,与水稻匹配的有 84.1%,高粱 84.6%,短柄草 83.9%,共 56 036 条转录本。

关键词　大刍草;转录组;RNA-Seq;*de novo* 拼接

中图分类号　S 519.24　　**文献标识码**　A　　**文章编号**　1000-2421(2014)02-0015-07

玉米属于禾本科(Poaceae)玉蜀黍属(*Zea*),玉蜀黍属又被划分成 5 个种,分别是 *Z. diploperennis*、*Z. perennis*、*Z. luxurians*、*Z. nicaraguensis* 和 *Z. mays*,其中 *Z. mays* 有 4 个亚种,分别是 *Z. mays* ssp. *mexicana*、*Z. mays* ssp. *parviglumis*、*Z. mays* ssp. *huehuetenangensis* 和 *Z. mays* ssp. *mays*,前 3 个亚种被统称为大刍草,第 4 个是经人类长期栽培驯化的玉米。果穗化石证据的缺失以及玉米和大刍草果穗上的巨大差异,一度让科学家无法确定玉米的直系祖先。Longley[1] 通过对 *Z. mays* ssp. *mexicana* 染色体臂的长度、中心粒的位置、结节大小和位置的研究指出,*Z. mays* ssp. *mexicana* 可能就是玉米的祖先,Kato 等[2-3] 通过对染色体形态的研究则将 *Z. mays* ssp. *parviglumis*、*Z. mays* ssp. *mexicana* 和 *Z. mays* ssp. *mays* 聚在一类,后续对叶绿体、核糖体的研究也得到了相似结果[4-5],而 Doebley 等[6] 基于同工酶以及 Matsuoka 等[7] 基于 SSR 标记的遗传多态性研究,则认为 *Zea mays* ssp. *parviglumis* 是玉米的祖先。

转录组是一个细胞中的所有转录本信息,包括转录本的数量、特定发育阶段的表达动态、转录后的修饰以及 non-coding RNA 的调控表达情况等。转录组研究的目的是通过收集所有 RNA 的信息(广义上包括了 mRNA、non-coding RNA 等),进而推断完整的基因结构,确定选择性剪切事件,研究在不同组织、不同发育阶段、不同实验处理中的相关基因在转录水平上的表达水平。近年来,随着转录组学研究的迅速展开,在研究方法上也有较多的发展与创新,主要包括基于杂交和基于测序的方法。“杂交法”主要是利用高密度的商业化芯片进行研究,但限制这一方法广泛应用的瓶颈是必须预先得到试材对象的全基因组序列信息,检测杂交信号误差也较大[8];“测序法”包括基于标签序列代表基因的 SAGE、CAGE、MPSS 以及基于高通量测序的 RNA-Seq 技术[9]。杂交法和标签序列法都只对转录组的部分序列进行分析,无法得出完整的基因结构信息,也无法分辨出选择性剪切产生的不同转录本;而基于高通量测序的 RNA-Seq 技术则可以较为全面地、对几乎全部的 RNA 转录本进行分析。

根据所研究的物种是否有参考基因组信息,RNA-Seq 序列拼接的策略也分为基于参考基因组的序列拼接和 *de novo* 序列拼接,或当基因组信息不完整时,将二者结合起来进行序列拼接。基于参考基因组的拼接方法是将测序得到的序列匹配到基因组上去,根据 reads 的重叠情况和是否跨越剪切位点等信息推断出每个可能的剪切本,该方法使用

收稿日期:2013-06-13
基金项目:国家“863”计划项目(2012AA10307)
肖之夏,硕士研究生. 研究方向:玉米分子生物学. E-mail:zhixia_xiao@webmail.hzau.edu.cn
通信作者:郑用琏,博士,教授. 研究方向:玉米分子生物学. E-mail:zhyl@mail.hzau.edu.cn

的软件包括了 Cufflinks、Scripture 等。*de novo* 方法则不依赖于参考基因组信息，仅使用测序的 reads 单独进行拼接，需要消耗的计算资源更大，该技术的局限在于：对不同品系之间 SNP 的检测、对 mRNA 选择性剪切的检测以及对基因结构变异的检测都很难达到理想的效果。*de novo* 策略使用的软件有 Trinity、Oases 等，其中 Trinity 可以发现更多的全长转录本，其灵敏度甚至能够近乎于基于基因组信息的拼接。目前已经发表的利用 *de novo* 拼接方法进行的非模式生物转录组研究涵盖了昆虫、植物等许多物种[10-12]。

Emrich 等[13]最早开展了高通量测序技术在玉米转录组研究中的应用。Thomas 实验室开发了 Illumina 平台进行玉米叶片转录组研究，推断与叶夹角和光合合成相关的基因[14]。目前已经有 4 个独立的玉米自交系和 1 个地方品种的 RNA-Seq 数据[15-16]公布，对于大刍草功能基因组的研究却很少。相较于玉米地方品种和自交系，野生近缘种大刍草具有更为复杂的遗传多样性。本研究利用 Illumina 平台对大刍草的转录组进行序列分析，并对 RNA-Seq 数据进行拼接和注释，旨在为后续开展玉米进化研究提供参考，为发掘玉米重要驯化基因提供新的序列资源。

1 材料与方法

1.1 大刍草材料

选取 6 个大刍草品系为试材，在温室采用沙土作为基质育苗，待其长至 3 叶期时全株取样，每个品系取 1 株提取 RNA 用于 cDNA 文库构建。

1.2 DNA/RNA 提取

DNA 抽提采用传统的 CTAB 法[17]。总 RNA 的抽提是采用 Invitrogen 公司的 Trizol 试剂提取，经过分光光度计检测纯度后，取 $D_{260}/D_{280}=1.9\sim2.1$、$D_{260}/D_{230}=2.0\sim2.5$、RIN（RNA integrity number）大于 8 的样本用于建立文库。

1.3 cDNA 文库的构建和 Illumina 测序

样品总 RNA 经过纯化后，利用连接了 poly-T 的磁珠富集 RNA 样本中的 mRNA。将富集得到的 mRNA 利用超声波打成小段，电泳后选择 200～500 bp 的片段回收，用随机引物和反转录酶进行 cDNA 第 1 链的合成，然后补齐成双链，加上接头后再进行 PCR 扩增，以确保低丰度转录本模板的含量。

Illumina 测序交由 LC Science 公司（LLC，Texas，USA）完成。采用 Genome Analyzer II 测序仪，两端 76 bp 测序，每个样品单独使用 1 个 lane。

1.4 质量控制和序列的拼接

使用 fastx-toolkit 软件对原始数据的质量（$Q_{phred}=-10\log10e$，其中 e 为测序错误率）进行控制。从 reads 的 3′端开始，对低质量的碱基进行剪切，切到第 1 个质量大于 15 的碱基为止；然后将切除后 reads 长度小于 35 bp 以及质量大于 15 的碱基比例小于 95%的 reads 剔除；最后将剩余 reads 中质量小于 15 的碱基用 N 替代，以减少其对拼接过程的影响。同时，对质量控制前后整体序列的质量进行统计，质量控制前后 reads 的数量和长度通过 perl 脚本进行统计。

序列的 *de novo* 拼接采用基于构建 De Bruijn 图方法的 Trintiy[18]和基于重叠拼接的 TGICL。首先将 6 个样本库的序列单独进行 Trinity 拼接，然后使用 TGICL 将得到的 contigs 进行聚类整合和延长。根据大刍草转录组的情况对 Trinity 法进行优化；TGICL 的聚类过程使用 B73 的 cDNA 作为种子序列，然后根据序列的同源性进行聚类。拼接完毕后，对所有拼接使用的 reads 针对拼接结果进行 mapping，并将没有 reads 匹配上的转录本剔除。

1.5 大刍草转录本的比对

首先使用 Blast 将拼接得到的转录本和 B73 的 cDNA 以及高粱、短柄草、水稻的蛋白序列进行比对，BlastN 的 cut-off 值取 e-value$<10^{-10}$，BlastX 的 cut-off 值取 e-value$<10^{-5}$。同时，对匹配片段（HSP）的长度进行限定，BlastN 中必须大于 200 bp，BlastX 中必须大于 60 个氨基酸。

1.6 拼接结果的验证

本研究共选取了 37 条大刍草转录本进行验证，其中 15 条的序列与 B73 cDNA 完全匹配，12 条是完全没有匹配的序列中表达量最高的转录本，另 10 条则是已经被剔除掉的转录本（最后没有 reads 匹配的序列）。利用反转录 PCR 的方法扩增转录本片段，回收后进行 TA 克隆，再用 Sanger 测序验证。引物序列是用 primer 3 设计的，见表 1。

2 结果与分析

2.1 测序及质量控制

Solexa 测序共得到 46.4 GB 的数据，共 175 101 250条 76 bp 的双端 reads。由于需用于 *de*

novo 拼接，故原始的数据量应该保证至少覆盖转录组 10 倍，同时因没有大刍草的基因组和转录组的参考序列，故只能使用玉米的转录组大小作出估计。为了确保能得到完整的转录组，我们在用 Trinity 拼接后将所有样本得到的 contigs 进行了整合和延长拼接。

保证输入序列的质量对于拼接来说是至关重要的。本研究中采用的质量控制方法，对序列的长度和质量都进行了限定，在经过质量控制后仍保留了原始数据量的 77%，约 35.7 GB 的数据。虽然设定

表 1　反转录 PCR 验证所用的引物序列和扩增的长度

Table 1　Primer information of RT-PCR validation

ID 号 No. ID	转录本名称 Transcript name	左引物 Primer-F	右引物 Primer-R	扩增的位置 Position	扩增长度/bp Length expected
A1	MblContig1214	CTTGCGTTACTACCACAGAG	AACTTATCTCCTTCACTTTCCC	1 143～1 901	758
A2	AorContig21228	CACACTTCCTCATGCTATTCTC	ACTTCATGTGGACAATGCCC	780～1 393	613
A3	MblContig1959	CCACCTTCAGCGAGTACAC	AAATCTAGTTCTCCATGCGG	549～1 253	704
A4	AorContig2265	AGCCCATCTTCCACTTCGTC	ATGAACTTCTCCAGCTCCAG	619～1 252	633
A5	MblContig12215	GTGCCCTCCAGTACATCAG	CCTTCTCCAAGTTCTCCAC	670～1 431	761
A6	MblContig7972	GATGATGTAGTCCAGCGAG	CCTACACCTACACCCTCAG	269～945	676
A7	AorContig10402	TGAGTTTGCTGAGAAACCGA	GTGTGAATTCTTCCCAATACC	698～1 307	609
A8	AorContig11509	CCTCATCTCCAAGAACATCGG	TATACCCAATACCTCACCACC	1 415～2 086	671
A9	MblContig915	TGACCATGAGGATAAGTTTGAC	CCCAGAAGAGAAGAGCATACAC	1 047～1 882	835
A10	MblContig1067	GTGGACAACTTCTTCAACGA	TTATTTCTCTTGTTCCTGGCGA	1 036～1 651	615
A11	MblContig15160	GAAAGTGCAGATCCCTCGGT	CCACCCAACATCCCTTTCCT	188～599	411
A12	MblContig21994	GAACGACAGGCTACCTAACCT	TATCATGCGAACTGGTCTTGG	186～714	528
A13	MblContig11219	ATCCATCATCACAGAACCCA	GAGTTATCAGCAAACAAGCAG	165～689	524
A14	AorContig18961	AGGATGCGACAAATGTTGG	GTGCCCTGAACTTTGCTCC	892～1 593	701
A15	AorContig12474	TTAACCCTCGCTCTTCCTTTCC	CATACCAACATCGCTTCTCAC	118～701	583
B1	MblContig2	ATCGAAAGTTGATAGGGCAG	GGAGGGATGCTTTGGATGG	5 617～6 415	798
B2	MblContig1352	CGGAGGAAGGAGAGGATGAG	ACAACGACGCAATTATCAGG	1 234～1 957	723
B3	MblContig4719	GATTCTAACCTTGTGTCAGACCC	CCGCTACCTTATCCTATTTCCA	410～1 117	707
B4	MblContig10530	GGCTGATCGAGGTGAAGGT	GCTAGGCAAGCAGCATACTG	164～605	441
B5	MblContig13317	ATGGGAACGAATGAACAGG	GGCATTGAGAAGGAAGGAC	135～596	461
B6	MblContig10605	GGCTGATCGAGGTGAAGGT	GCATTCATATTCGTTACAGGCAGG	176～639	463
B7	MblContig10529	GGCTGATCGAGGTGAAGGT	CACTGCTGATGCATGGGAG	161～663	502
B8	MblContig9636	GGCTGATCGAGGTGAAGGT	CACTGCTGATGCATGGGAG	153～772	569
B9	MblContig11216	TCTACTACACGACGACAAACATGG	AAGGTGCAGACGACACTGG	51～720	669
B10	MblContig26252	CATGAGGAAGCAATACTCCC	AAGTCGGTTCGATCTTTCGT	53～514	461
B11	MblContig34542	AGTTTGTTGAGAAGGAAGAGGT	CCAGCCAAGTATTACAGTATCAC	117～888	771
B12	MblContig14311	CACACTGGGACTGAGACAC	ACAGACCAAGGGCGAACAC	0～506	506
D1	AorContig11381	GGTTCATTCCCGACCACGA	GAGTACTGCCATGATTGTCTCC	2～307	305
D2	AorContig12070	TTTAGCCTCCTTACATCCGA	TAGGCCAATCTCCAGATCC	11～316	305
D3	AorContig12523	TGAGTTTGCCATCACCTCC	GGCATAGAGAATTTCACAAGG	5～541	536
D4	AorContig13663	AAAGTATTAGCACTGGTAGACTGG	GAGATCGCAATGCGTTACC	10～335	325
D5	AorContig14479	AGTTGAGACCAGGATTGTTCCA	TTTACCCAACGACGAACGA	0～574	574
D6	AorContig14670	CTGTCTCAAATCCCAGCAC	GAGTCTTGAGAGTAGAAAGCAG	12～515	503
D7	AorContig14761	TTTCTCGGTGTCAAACGAC	TCGGTTTAGGGTTTAGCTATCAG	90～558	468
D8	AorContig14808	CAGATCAGCATCCAACAAATCC	GACTCTACATGAGCCCTACAG	14～355	341
D9	AorContig14853	AGCGTAAACCTAGTGGAATGAG	CAGACGATGCACGGATTGG	0～548	548
D10	AorContig15035	TGAAATTGAACAGGCAGTTGAC	CGACGACGTAGGTGATATAAACAG	44～207	163

最小长度的过滤条件为35 bp,但经过质量控制后,仍有60.4%的reads长度为75 bp。利用fastx-toolkit工具对质量控制前后序列的质量进行统计后,发现质量控制前,许多在50 bp后的reads其碱基的质量就不能满足拼接的需要了,而质量控制后,即使是在75 bp,也能保证75%的reads在其位置的质量高于20,这表明质量控制是非常有效的。

经过Trinity拼接后,共得到了505 092条contig,平均长度为734 bp,N50值为560,最长的contig为8 151 bp,最短的也有305 bp。考虑到单个样本的数据量可能不足以覆盖整个转录组,又利用TGICL将这些contigs进行了聚类和延长,共得到58 147条大刍草转录本。聚类时采用了B73 cDNA作为种子序列,其中21 466条转录本是能与B73转录本聚在一起的contigs整合的结果,而另36 681条转录本则是剩下的contigs相互比较后聚类拼接的结果(表2)。TGICL得到的序列平均长度为1 335 bp,N50为1 122 bp。B73 cDNA共有63 540条(v5b.60),其N50值为1 419,在数量和长度上与本研究得到的大刍草转录本没有很大差别,其整体的长度比大刍草的更长,二者转录本长度的比较见表3。

表2 拼接结果的整体数据统计

Table 2 Statistic of assembly result

项目 Items	统计 Statistic
原始reads总数量 Total number of raw reads	175 101 250
原始reads总长度/bp Total base-pairs	13 307 695 000
质量控制后reads的数量 Total number of reads after QC	134 827 962
质量控制后reads的总长度/bp Total base-pairs after QC	10 246 925 150
reads的平均长度/bp Average read length	76
contig的数量 Total number of contigs	505 092
contig的平均长度/bp Average length of contigs	734
单个contig的最大长度/bp Largest length of contigs	8 151
contig的N50值 N50 value of contigs	560
转录本的总数量 Total number of assembled transcripts	58 147
转录本的平均长度/bp Average length of transcripts	1 335
单个转录本的最大长度/bp Largest length of transcript	14 668
转录本的N50值 N50 value of transcripts	1 122
转录组总体大小/bp Total size of transcriptome	77 621 529

表3 大刍草和玉米B73转录本长度分布比较

Table 3 Comparison of the transcripts length of tesotinte and B73 bp

转录本长度 Transcripts length	<500	500~999	1 000~1 499	1 500~1 999	2 000~2 999	≥3 000
大刍草 Teosinte	7 616	18 212	12 562	9 172	8 006	2 785
玉米B73 Maize B73	4 020	14 789	15 651	14 113	11 389	3 578

2.2 转录本注释

通过与玉米、高粱、水稻、短柄草进行序列比对完成了对转录本的功能注释。利用Blast将大刍草转录本和B73 cDNA序列比对后,发现有54 878条都有很好的匹配结果,占所有序列的94.3%。这些序列共匹配上26 377条B73 cDNA,约涵盖40%的B73转录组。而与高粱、水稻、短柄草等3种单子叶植物的蛋白质序列的比对结果分别为:与水稻匹配的有84.1%,高粱84.6%,短柄草83.9%,共56 036条转录本,说明在进化过程中,大刍草中也保留了许多的保守序列。

2.3 拼接结果的验证

利用反转录PCR对27个转录本的片段扩增并纯化后,重新测序。结果有4条转录本扩增片段的测序结果与拼接结果在长度上有较大差别,占14.8%;有4条没有扩增成功,占14.8%;其余的都有十分完好的匹配,占70.4%。4条长度不一致的片段与原序列的相似度接近0.99,其长度上的差异可能是由于插入或缺失等遗传事件造成的;另外,由于拼接过程将所有样本序列进行了整合,所以某些转录本在不同样本之间的差异可能会对拼接的结果有所影响,导致转录本长度的差异。而4条扩增失败的片段都是没有reads匹配上的序列,这些序列可能是拼接的错误,也可能是因为原转录本表达量过低造成的。从扩增验证的结果(表4)看,本研究的拼接结果的准确性高于85%。

表 4　RT-PCR 验证的结果

Table 4　Result of RT-PCR validation

ID 号 No. ID	转录本名称 Transcript name	预计扩增长度/bp Length expected	实际扩增长/bp Length obtained	Blast 相似度 Blast identity
A1	MblContig1214	758	759	0.99
A2	AorContig21228	613	618	0.99
A3	MblContig1959	704	707	0.99
A4	AorContig2265	633	633	0.99
A5	MblContig12215	761	762	0.98
A6	MblContig7972	676	677	1.00
A7	AorContig10402	609	616	0.99
A8	AorContig11509	671	675	0.95
A9	MblContig915	835	847	0.99
A10	MblContig1067	615	616	0.94
A11	MblContig15160	411	413	0.99
A12	MblContig21994	528	528	0.98
A13	MblContig11219	524	539	0.96
A14	AorContig18961	701	702	0.99
A15	AorContig12474	583	584	0.97
B1	MblContig2	798	400	0.99
B2	MblContig1352	723	725	0.99
B3	MblContig4719	707	708	1.00
B4	MblContig10530	441	443	0.96
B5	MblContig13317	461	462	1.00
B6	MblContig10605	463	466	0.95
B7	MblContig10529	502	506	0.95
B8	MblContig9636	569	477	0.97
B9	MblContig11216	669	677	0.98
B10	MblContig26252	461	505	0.89
B11	MblContig34542	771	792	0.90
B12	MblContig14311	506	507	0.99
D1	AorContig11381	305	Amplifying failed	
D2	AorContig12070	305	308	0.99
D3	AorContig12523	536	Match failed	
D4	AorContig13663	325	326	0.99
D5	AorContig14479	574	Match failed	
D6	AorContig14670	503	138	1.00
D7	AorContig14761	468	484	0.91
D8	AorContig14808	341	342	0.99
D9	AorContig14853	548	550	0.96
D10	AorContig15035	163	Amplifying failed	

3　讨　论

模式植物拟南芥、水稻等的转录组信息较为丰富，而非模式植物的序列信息却十分匮乏。高通量测序技术为非模式植物全基因组序列分析、高通量转录组的分析提供了可能，RNA-Seq 技术已逐步成为一种常规的研究方法。虽然高通量测序技术能够让研究者以较低的成本获得大量的序列数据，但是数据分析和挖掘仍然是一个重要的制约因素，特别是对于基因组和转录组的拼接，尤其是 *de novo* 拼接，对于研究者来说是一个新的挑战。前人的研究均表明，对于没有基因组序列信息的植物进行转录

组的 *de novo* 拼接非常困难，但随着更优化的拼接方法和软件相继被公布，加之能够获得较长测序片段的第三代测序技术日趋成熟，这可能又会是基因组学和转录组学研究的一次革命。

在我们的拼接过程中，首先利用了 B73 的 cDNA作为种子序列对 Trinity 的结果进行了整合。由于玉米和大刍草的近缘关系，我们认为应该有部分大刍草序列和 B73 的 cDNA 比较相似，这样依据种子序列聚类后再进行拼接能够极大地提高准确性。而另一部分无法和这些种子序列聚在一起的 contigs 只能在进行相互比对后进行拼接。这样的拼接策略保证了序列的准确性，使得验证结果较好。设计的 37 对引物中，正式作为大刍草转录本的 27 条序列的比对结果除了 2 条为 90%左右外，其他的都超过了 95%，这对于 *de novo* 拼接来说准确度是较高的。

在统计表达量时，研究者普遍使用 RPKM 值。但在计算转录本的表达量时，许多影响因素会导致统计误差的产生。首先是 cDNA 文库是否经过均一化处理，即用 PCR 等方法对其中的转录本片段进行扩增。均一化处理可以使表达量的转录本更容易被检测到，但由于扩增的过程使低表达和高表达的转录本增加的倍数不尽相同，高表达转录本很可能会达到扩增的饱和态，所以也很难保证检测的每个转录本表达量的真实性。目前，研究者们已经开发出了许多不同的均一化方法[19]，并且证明了均一化后的文库更有利于后续拼接得到的转录组的完整性[20]。本研究的目的是希望能够拼接出较为完整的大刍草转录组，所以我们对 cDNA 文库进行了均一化。另外一个原因就是 RPKM 值的表示方法。RPKM 值是根据匹配到一个转录本上的 reads 数量来计算的，对于能够完整匹配到不同序列上的 reads 只能剔除，但这些 reads 则可能是同时匹配到同一个转录本的不同剪切本上或者是 2 个基因同源性很高的转录本上，剔除不能匹配的 reads 后，转录本的表达量就会有偏差，若 reads 长度较短，则会出现大量这样的偏差，那么检测的转录本的表达量就会整体偏低。故在本研究中我们没有分析转录本的表达量。

de novo 拼接的转录组对后续的结果分析的另一限制是：由于没有基因结构信息，对 mRNA 选择性剪接的分析难以实施；另外，对于 2 个同源性很高的转录本也很难判断出它们是来源于同一基因还是 2 个同源基因。最终也不能很准确地检测出在进化中发生的基因结构变异事件。本研究中我们选用已于 2008 年公布且每年不断完善的 B73 基因组序列为参考序列，提出了一个较为完整清晰的大刍草转录组，经与 B73 cDNA 序列比对后，检测到 54 878 条具有很好匹配效果的序列，占所有序列的94.3%，共匹配上 26 377 条 B73 cDNA，约涵盖 40%的 B73 转录组。而且与高粱、水稻、短柄草等 3 种单子叶植物的蛋白质序列也有约 84%的匹配。

本研究采用基于 RNA-Seq 数据的 *de novo* 分析与拼接策略，通过对 6 个大刍草品系苗期的 RNA 分析获得了较好质量的大刍草转录组信息，可为后续对大刍草功能基因的发掘和与玉米的比较基因组研究提供参考，同时也可为非模式作物的 RNA-Seq 研究和 *de novo* 拼接提供新思路。

参 考 文 献

[1] LONGLEY A E. Chromosome morphology in maize and its relatives[J]. Bot Rev, 1941, 7: 263-289.

[2] KATO T A. Cytological studies of maize (*Zea mays* L.) and teosinte (*Zea mexicana* Schrader Kuntze) in relation to their origin and evolution[J]. Mass Agric Exp Stn Bull, 1976, 635: 1-186.

[3] KATO T A, LOPEZ R A. Chromosome knobs of the perennial teosintes[J]. Maydica, 1990, 35: 125-141.

[4] DOEBLEY J F, RENFROE W, BLANTON A. Restriction site variation in the *Zea* chloroplast genome[J]. Genetics, 1987, 117: 139-147.

[5] BUCKLER E S, HOLTSFORD T P. *Zea* systematics: ribosomal its evidence[J]. Mol Bio Evolution, 1996, 13: 612-622.

[6] DOEBLEY J F, GOODMAN M M, STUBER C W. Isoenzymatic variation in *Zea* (Gramineae)[J]. Systematic Botany, 1984, 9: 203-218.

[7] MATSUOKA Y Y, VIGOUROUX M M, GOODMAN J, et al. A single domestication for maize shown by multilocus microsatellite genotyping[J]. Proc Natl Acad Sci USA, 2002, 99: 6080-6084.

[8] OKONIEWSKI M J, MILLER C J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations[J]. BMC Bioinformatics, 2006, 7: 276.

[9] WANG Z, GERSTEIN M, SNYDER M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nature Reviews, 2009, 10: 57-63.

[10] 陈小平，洪彦彬，张二华，等. 高通量花生转录组分析系统的构建与应用[J]. 中国油料作物学报，2011，33(3)：235-241.

[11] 何涛，王端青，胡亚欧，等. 基于 RNA-Seq 数据识别果蝇剪接位点和可变剪接事件[J]. 中国科学：生命科学，2011(10)：1016-

1023.

[12] 王志伟，应正宙. 利用 RNA-Seq 法对人真菌病原体白色念珠菌转录组进行广泛注释[J]. 农业生物技术学报，2010(5)：937.

[13] EMRICH S J，BARBAZUK W B，LI L，et al. Gene discovery and annotation using LCM-454 transcriptome sequencing[J]. Genome Res，2007，17：69-73.

[14] LI P，PONNALA L，GANDOTRA N，et al. The development dynamics of the maize leaf transcriptome[J]. Nat Genetics，2010，42：1060-1067.

[15] WANG X F，ELLING A A，LI X Y，et al. Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize[J]. Plant Cell，2009，21：1053-1069.

[16] JIA Y，LISCH D R，OHTSU K，et al. Loss of RNA-dependent RNA polymerase2（RDR2）function causes wide-spread and unexpected changes in the expression of transposons，genes，and 24-nt small RNAs[J]. PLoS Genet，2009，5：e1000737.

[17] 王芳，王汉宁，张金文，等. 玉米基因组 DNA 的快速高效提取[J]. 草业科学，2006，23(12)：65-67.

[18] GRABHERR M G，HAAS B J，YASSOUR M，et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome[J]. Nat Biotechnol，2011，29：644-652.

[19] ZHULIDOV P A，BOGDANOVA E A，SHCHEGLOV A S，et al. Simple cDNA normalization using Kamchatka crab duplex-specific nuclease[J]. Nucleic Acids Res，2004，32：e37.

[20] OGURA A，LIN M，SHIGENOBU Y，et al. Effective gene collection from metatranscriptome of marine microorganisms[J]. BMC Genomics，2011，12：S15.

De novo assembly of teosinte seedling transcriptome defined by RNA-Seq

XIAO Zhi-xia　ZHENG Yong-lian

National Key Laboratory of Crop Genetic Improvement，Huazhong Agricultural University，Wuhan 430070，*China*

Abstract　Teosinte is the ancestor of maize，and plays an important role in maize domestication process and gene cloning. Solexa RNA-Seq was used to *de novo* assembly and analyze the transcriptome of teosintes. 40. 6 GB raw data were produced，including 175 101 250 reads of 76 bp length. After quality control and *de novo* assembly，58 147 teosinte transcripts with an average length of 1 335 bp were obtained. After bioinformatically comparing，it was found that 94. 3% of teosinte transcripts had good matching with B73 cDNAs，and that 84. 1% of the transcript had good matching with rice，84. 6% with sorghum and 83. 9% with brachypodium at protein level. This research will provide a reference for subsequent studies on maize evolution and gene discovery.

Key words　teosinte；transcriptome；RNA-Seq；*de novo* assembly

（责任编辑：张志钰）