研究报告
Research Report

五种豆科植物物种基于 De Novo 序列装配和注释的比较研究

Sagar S. Patel[1], Dipti B. Shah[1], Hetalkumar J. Panchal[2]

1. G. H. Patel Post Graduate Department of Computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat-388120, India
2. Gujarat Agricultural Biotechnology Institute, Navsari Agricultural University, Surat, Gujarat- 395007, India
通讯作者: sgr308@gmail.com; 作者
计算分子生物学, 2014 年, 第 3 卷, 第 12 篇 doi: 10.5376/cmb.cn.2014.03.0012

建议最佳引用格式：

Patel et al., 2014, Comparative study of five Legume species based on De Novo Sequence Assembly and Annotation, Computational Molecular Biology, Vol.4, No.9, 1-6 (doi: 10.5376/cmb.2014.04.0009)

摘 要 豆科植物是世界热带和亚热带地区的一种重要的油料作物。最近，名为 RNA-seq 的新一代测序技术为转录组分析提供了强有力的方法。这项研究是集中在 RNA 序列对五种豆科植物，分别是来自 NCBI 数据库的花生 *SRR1212866*、鹰嘴豆 *SRR527764*、菜豆 *SRR1283084*、葫芦巴 *SRR066197* 和豌豆 *SRR403901*，比较研究测重于各种重要特征如：用 N50，序列组装重叠群产生 reads，用已知的蛋白质和基因进一步搜索；其中，许多基因是根据 GO 功能分类注释并通过搜索 KEGG 数据库将序列定位到途径中。这些数据将用于基因发现和功能研究，并且在当前研究中报道的大量转录物将作为这五种豆科物种的有价值的遗传资源。

关键词 De Novo 组装; 生物信息学; 豆科植物; 序列组合和注释

Comparative study of five Legume species based on De Novo Sequence Assembly and Annotation

Sagar S. Patel[1], Dipti B. Shah[1], Hetalkumar J. Panchal[2]

1. G. H. Patel Post Graduate Department of Computer Science and Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat-388120, India
2. Gujarat Agricultural Biotechnology Institute, Navsari Agricultural University, Surat, Gujarat- 395007, India
Corresponding author, sgr308@gmail.com; Authors

Abstract Legume species are an important oilseed crop in tropical and subtropical regions of the world. Recently, next-generation sequencing technology, termed RNA-seq, has provided a powerful approach for analysing the Transcriptome. This study is focus on RNA-seq of five legume species which are *Arachis hypogaea* L. (The peanut) of *SRR1212866*, *Cicer arietinum* L. of *SRR527764*, *Phaseolus vulgaris* L. of *SRR1283084*, *Trigonella foenum-graecum* L. of *SRR066197* and *Vicia sativa* L. of *SRR403901* from NCBI database. Comparative study focuses on various important features like; reads were generated with N50, sequence assembly contigs which is further searched with known proteins and genes; among these, how many genes were annotated with gene ontology (GO) functional categories and sequences mapped to pathways by searching against the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG). These data will be useful for gene discovery and functional studies and the large number of transcripts reported in the current study will serve as a valuable genetic resource of these five legume species.

Keywords De Novo assembly; Bioinformatics; Legume species; Sequence assembly and annotation

介绍

新一代测序方法——高通量RNA测序(转录组)正越来越多地应用在植物的检测和定量已知或新型转录物的选择技术。这种转录组分析方法是快速和简单的，因为它不需要cDNA的克隆。这些cDNA的直接测序可以深度产生短reads。测序后，得到的reads可以组装成基因组规模的转录配置文件。它是一种更全面和有效的方法来测量转录组组成，获得RNA表达模式，并发现新的外显子和基因(Mortazavi et al., 2008; Wang et al.,2009);

收稿日期：2014 年 12 月 24 日
接受日期：2014 年 12 月 24 日
发表日期：2014 年 12 月 25 日

使用各种装配工具，基因的功能注释和用各种生物信息学工具携带的途径分析来组装转录组的测序数据。本研究报告的大量转录本可以作为描述五种豆类物种的宝贵的遗传资源。

高通量短read测序是基因组学界公开的最新测序技术之一。例如，在Illumina基因组分析仪上的平均单次运行可以得到超过3000至4000万个单端序列(~35 nt)。然而，输出结果可以轻松超过为传统Sanger测序的长度设计的基因组分析系统，甚至获得较小体积数据的454(Roche)测序技术。通常，短read测序初期使用是局限于与参考基因组几乎相同的基因组数据的匹配。全基因表达水平的转录组分析全是短read测序的理想应用。传统上，这种分析包括互补DNA(cDNA)文库构建，EST的Sanger测序和微阵列分析。与传统的Sanger方法相比，新一代测序已经成为增加测序深度和覆盖范围，同时减少时间和成本的可行方法(L J Collins et al.)。

1 方法

1.1 序列检索

本研究的重点是五种豆科植物的de novo测序和序列注释，分别是来自NCBI数据库的花生*SRR1212866*、鹰嘴豆*SRR627764*、菜豆*SRR1283084*、葫芦巴*SRR066197*和豌豆*SRR403901*，用于de novo转录组分析。从Illumina HiSeq 2000平台和LS454-454GS FLX平台来源的NCBI SRA中下载的原始数据(http://trace.ncbi.nlm.nih.gov/Traces/sra/)。使用NCBI的SRA TOOL KIT将原始序列转化为fastq文件格式用于进一步注释(http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software)。

1.2 NGS QC 工具包

NGS QC工具包，它是用于高质量数据的质量检查和过滤的应用程序。此工具包是独立的应用程序，可从http://www.nipgr.res.in/ngsqctoolkit.html免费获得。该工具包包括用Roche 454和Illumina平台生成测序数据的用户容易掌握的QC工具以及用于辅助QC(序列格式转换器和修整工具)和分析(统计工具)的附加工具.提供了各种选项便于用户定义的QC参数.该工具包预计对利于较好下游分析的NGS数据的QC非常有用(Patel RK, et al)。

1.3 通过 CLC GENOMICS WORKBENCH 7 进行 De novo 测序

用于分析，比较和下一代测序数据可视化的全面和用户容易掌握的分析包。这个软件包用de novo测序工具的默认参数对序列进行de novo测序(http://www.clcbio.com/products/clc-genomics-workbench/)。

1.4 BLASTX

对测序文件进一步注释，第一步是从序列中鉴定出翻译的蛋白质序列。改变几个参数后在NCBI上进行BLASTX比对，如选择非冗余蛋白质数据库(nr)作为数据库；双子叶植物在生物选择和算法参数的最大目标序列设置为10，期望阈值设置为6。

1.5 Blast2GO

Blast2GO(http://www.blast2go.com/b2ghome)是一个用于(新)序列的功能注释和注释数据分析的ALL in ONE工具。基于蛋白质数据库注释的结果，Blast2GO被用于获得基于GO项的unigenes的功能分类。转录序列根据三个GO term分类，如分子功能，细胞过程和生物过程(Ness et al., 2011; Shi et al., 2011; Wang et al., 2010)。用WEGO(http://www.wego.genomics.org.cn)工具对所有unigenes进行GO功能分类，并在宏观层面了解该物种的基因功能的分布。用KEGG数据库(http://www.genome.jp/kegg/pathway.html)注释这些unigenes的途径。

1.6 SSR 挖掘

我们使用MIcroSAtellite(MISA)(http://pgrc.ipk-gatersleben.de/misa/)进行微卫星挖掘，统计输出其产生有用转录组的信息。

1.7 植物转录因子

PlantTFcat：在线植物转录因子和转录调节因子分类和分析工具，用于鉴定序列中的植物转录因子(http://plantgrn.noble.org/PlantTFcat/)。

2 结果与讨论

2.1 序列比较

(表1)。

2.2 NGS QC 工具包

通过去除衔接子和其他污染的材料用该工具过滤序列，然后用该工具检查序列的质量，最终用de novo序列组装的高质量过滤序列文件(表2)。

2.3 De novo 序列组装

CLC GENOMICS WORKBENCH 7 考虑用于De novo序列组装，使用默认参数比如Mismatch Cost = 2, Insertion Cost = 3, Deletion Cost = 3, Length Fraction = 0.5, Similarity Fraction = 0.8, Word size = 21，由本软件产生的序列平均值和其他细节列于表3。

表 1 物种序列比较

Table 1 Species comparison based on sequence

Species	SRR Number	Reads	%GC Content	Platform
Arachis hypogaea L.	SRR1212866	7.3 M spots	48.5	Illumina HiSeq 2000
Cicer arietinum L.	SRR627764	36 M spots	41.8	Illumina
Phaseolus vulgaris L.	SRR1283084	20.4 M spots	46.4	Illumina HiSeq 2000
Trigonella foenum-graecum L.	SRR066197	627,117 spots	45.2	454 GS FLX
Vicia sativa L.	SRR403901	12.4 M spots	42.4	Illumina HiSeq 2000

表 2 NGS QC 工具包结果

Table 2 NGS QC Toolkit Result

Species	Total number of reads (Original File)	Total number of reads (High Quality (HQ) Filter file)	Total number of bases (Original File)	Total number of bases (High Quality (HQ) Filter file)	Percentage of HQ reads
Arachis hypogaea L.	7300624	7216150	365031200	360807500	98.84%
Cicer arietinum L.	19422297463	1942030133	1904983	1904959	99.99%
Phaseolus vulgaris L.	20444892	13418027	1042689492	684319377	65.63%
Trigonella foenum-graecum L.	627117	609237	146335656	141577237	97.15%
Vicia sativa L.	12427455	12131939	608945295	594465011	97.62%

表 3 序列测量长度

Table 3 Contig measurement in Length

Species	N50	Minimum	Maximum	Average	Count (Contigs)
Arachis hypogaea L.	448	199	6635	425	10824
Cicer arietinum L.	1239	179	8439	805	34678
Phaseolus vulgaris L.	293	187	5386	302	6999
Trigonella foenum-graecum L.	470	86	3231	445	7256
Vicia sativa L.	588	197	6080	503	22748

2.4 BLASTX 和 blast2GO 的功能注释

2.4.1 BLASTX

使用10-6的E值阈值进行BLASTX以将序列与非冗余序列数据库比对。BLAST结果的各种统计信息列于表4。

2.4.2酶代码(EC)分类

表5是酶序列分类，进一步分为六类，氧化还原酶，转移酶，水解酶，裂解酶，异构酶和连接酶。

2.4.3基因本体(GO)分类

为了将各种豆科植物的转录序列进行功能分

类， 分配GO terms组装转录序列。转录序列按照GO功能类别进行分组(http://www.geneontology.org)，其分布在分子功能，生物过程和细胞组分的三个主要类别(表6)。

表 4 BLAST 结果的比较

Table 4 Blast Result comparison

Species	Without Results	BlastWithout Hits	BlastWith Results	BlastWith Results	MappingAnnotated Sequences	Total Sequences
Arachis hypogaea L.	60	688	4789	568	4719	10824
Cicer arietinum L.	3492	3996	25459	786	945	34678
Phaseolus vulgaris L.	102	2601	1988	629	1679	6999
Trigonella foenum-graecum L.	167	2656	1983	192	2258	7256
Vicia sativa L.	0	1114	13482	500	7652	22748

表5 EC分类

Table 5 Enzyme Code (EC) Classification

Species	Oxidoreducta--ses	Transferases	Hydrola-ses	Lyases	Isomera--ses	Ligases	Total
Arachis hypogaea L.	301	614	431	78	51	71	1546
Cicer arietinum L.	51	92	76	20	5	4	248
Phaseolus vulgaris L.	110	232	147	23	20	31	563
Trigonella foenum-graecum L.	148	149	179	34	38	27	575
Vicia sativa L.	429	927	718	80	82	100	2336

表6 GO 分类

Table 6 Gene Ontology (GO) Classification

Species	Molecular Function	Biological Process	Cellular Components	Total
Arachis hypogaea L.	4512 (43%)	3352 (32%)	2607 (25%)	10471
Cicer arietinum L.	916 (40%)	734 (32%)	654 (28%)	2304
Phaseolus vulgaris L.	1727 (47%)	1168 (31%)	829 (22%)	3724
Trigonella foenum-graecum L.	2792 (28%)	4407 (43%)	2980 (29%)	10179
Vicia sativa L.	7026 (37%)	5815 (31%)	5920 (32%)	18761

图1是WEGO工具的输出结果：它表明，在分子功能类别中，编码与催化活性相关的结合蛋白和蛋白质的基因是最富集的。与代谢过程和细胞过程相关的蛋白质富集在生物过程类别中。关于细胞成分类别，细胞和细胞部分是最高度表示的类别。我们发现在所有其他豆类种类相同，所以我们只用此图说明WEGO工具。

许多基因在KEGG数据库中用不同的途径注释(http://www.genome.jp/kegg/pathway.html)。进一步的比较结果显示在表7中。许多转录物包括各种途径，如代谢途径，植物-病原体相互作用途径，脂肪酸代谢途径和脂肪酸生物合成。

表7 KEGG结果

Table 7 KEGG Result

Species	Genes	KEGG Pathway
Arachis hypogaea L.	568	109
Cicer arietinum L.	786	78
Phaseolus vulgaris L.	629	89
Trigonella foenum-graecum L.	192	87
Vicia sativa L.	500	122

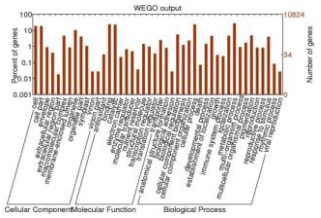

图1 花生WEGO工具结果

Figure 1 WEGO Tool Result of *Arachis hypogaea* L.

2.5 SSR 挖掘

微卫星标记(SSR标记)是构建花生遗传图谱和多样性分析中最成功的分子标记(Zhang et al)。为了鉴定SSR，用per脚本MISA搜索所有转录物。表8描述了SSR搜索结果，其显示了每个物种SSR结果的详细信息。最大部分的SSR是单核苷酸SSR，随后是三核苷酸SSR和二核苷酸SSR。虽然在转录物中只鉴定了一小部分四核苷酸SSR、五核苷酸SSR和六核苷酸SSR，但在大多数物种中该数目是相当显著的。

表 8 转录组中鉴定出的 SSR 统计

Table 8 Statistics of SSRs identified in transcripts

SSR Mining	Species				
	Arachis hypogaea L.	Cicer arietinum L.	Phaseolus vulgaris L.	Trigonella foenum-graecum L.	Vicia sativa L.
Total number of sequences examined:	10824	34678	6999	7256	22748
Total size of examined sequences (bp):	4605095	27932177	2110290	3226271	11444673
Total number of identified SSRs:	742	5228	1405	3107	1150
Number of SSR containing sequences:	649	4391	1304	2191	1055
Number of sequences containing more than one SSR:	74	681	86	747	92
Number of SSRs present in compound formation:	48	337	64	747	48
Distribution to different repeat type classes:					
Mono-nucleotide	265	2019	1218	2589	362
Di-nucleotide	164	1271	87	235	243
Tri-nucleotide	299	1818	90	243	529
Tetra-nucleotide	10	78	7	28	10
Penta-nucleotide	2	17	2	10	3
Hexa-nucleotide	2	25	1	2	3

2.6植物转录因子

此外，通过与已知的转录因子基因家族的序列比较来鉴定转录因子编码转录物。表9中的结果显示，确定了转录因子基因分布在与家族中，并且表9和图2中描述的是Trigonella foenum-graecum L.的植物转录因子结果。转录因子编码转录物在各种已知蛋白质家族中的整体分布与早期预测的其他豆类非常相似(Libault et al., 2009)。

表9 植物转录因子结果

Table 9 Plant Transcription Factor Result

Species	At least different families
Arachis hypogaea L.	70
Cicer arietinum L.	97
Phaseolus vulgaris L.	43
Trigonella foenum-graecum L.	45
Vicia sativa L.	82

3 结论

本研究侧重于NCBI数据库中五种不同豆科植物的de novo测序和分析，通过用新一代Illumina和454测序进行RNA-seq分析。转录组测序使得能够对生物体进行各种功能性基因组学研究。虽然已经开发了用于快速测序和表观转录组的几种高通量技术，但是表达的序列数据仍然不能用于许多生物体，包括许多作物植物。在这项研究中，我们对五种不同豆科植物进行了de novo功能注释，不考虑任何参考物种具有显着的非冗余集的成绩单没有考虑任何具有显着的非冗余集转录本的参考物种。根据五种植物数据的详细分析得到了几个重要的特征如GC含量，豆科植物和其他植物物种的保守基因，通过GO term区分功能类别和通过MISA工具鉴定SSR。值得注意的是，对花生、鹰嘴豆、菜豆、葫芦巴和豌豆这五种不同豆科植物的比较研究将有利于进一步的功能基因组研究，因为它包括每个物种完整注释的有用信息。

致谢

感谢 Prof. (Dr.) P.V. Virparia, Director, GDCST, Sardar Patel University, Vallabh Vidyanagar 为研究工作提供设备。

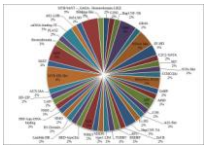

图 2 葫芦巴转录因子结果

Figure 2 Plant Transcription Factor Result of *Trigonella foenum-graecum L.*

参考文献

Collins J. L., Biggs J. P., Voelckel C. and Joly S., 2008, An approach to transcriptome analysis of non-model organisms using short-read sequences, Genome Informatics 21:3-14
http://dx.doi.org/10.1142/9781848163324_0001

Jianan Zhang, Shan Liang, Jialei Duan, Jin Wang, Silong Chen, Zongshu Cheng, Qiang Zhang, Xuanqiang Liang and Yurong Li, 2012, De novo assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (Arachis hypogaea L.), BMC Genomics 2012 13:90
http://dx.doi.org/10.1186/1471-2164-13-90

Libault, M., Joshi, T., Benedito, V.A., Xu, D., Udvardi, M.K., and Stacey, G., 2009, Legume Transcription Factor Genes: What makes legumes so special?. Plant Physiology 151: 991-1001
http://dx.doi.org/10.1104/pp.109.144105

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5(7): 621-8
http://dx.doi.org/10.1038/nmeth.1226

Ness, R.W., Siol, M., and Barrett S.C.H., 2011, De novo sequence assembly and characterization of the floral transcriptome in cross and self-fertilizing plants, BMC Genomics 12: 298
http://dx.doi.org/10.1186/1471-2164-12-298

Patel RK, Jain M, 2012, NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data, PLoS ONE 7(2): e30619. doi:10.1371/journal.pone.0030619
http://dx.doi.org/10.1371/journal.pone.0030619

Rohini Garg, Ravi K. Patel, Akhilesh K. Tyagi, and Mukesh Jain., 2011, De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification, DNA RESEARCH 18, 53-63; doi:10.1093/dnares/dsq028
http://dx.doi.org/10.1093/dnares/dsq028

Shi, C.Y., Yang, H., and Wei, C.L., 2011, Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds, BMC Genomics 12: 131
http://dx.doi.org/10.1186/1471-2164-12-131

Vaidya K., Ghosh A., Kumar V, Chaudhary S, Srivastava N, Katudia K, Tiwari T and Chikara K., 2012, De novo transcriptome sequencing in Trigonella foenum-graecum to identify genes involved in the biosynthesis of diosgenin. The Plant Genome
http://dx.doi.org/10.3835/plantgenome2012.08.0021

Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., and Liu, S.S., 2010, De novo characterization of a whitefly transcriptome and analysis of its gene expression during development, BMC Genomics 11: 400
http://dx.doi.org/10.1186/1471-2164-11-400

Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet. 10(1): 57-63
http://dx.doi.org/10.1038/nrg2484