基于RNA-Seq技术苹果基因结构优化与新转录本预测

倪伟, 高付凤, 杨恒峰, 张佳腾, 徐金, 毛志泉, 陈学森, 沈向*
山东农业大学园艺科学与工程学院, 作物生物学国家重点实验室, 山东泰安271018

摘要: 苹果(*Malus domestica*)已经完成全基因组测序并公布全基因组图谱, 但是基因注释信息很不完善, 因此本试验利用RNA-Seq技术对苹果已注释基因结构进行优化和新转录本预测。以‘富士’花后15、30、45、60 d的果肉为材料提取总RNA, 构建文库并使用Illumina双末端测序HiSeq 2500平台进行测序, 获得高质量的序列(clean reads)。将得到的结果与苹果‘金冠’参考基因组进行比对, 共优化了13 272个已注释基因的结构, 优化基因5′端和3′端数目分别为8 843个和8 955个, 另外, 共鉴定得到1 604个新的转录本, 部分基因参与了苹果的转录表达调控、生长调节、氨基酸和糖类等的代谢过程。

关键词: 苹果; RNA-Seq; 基因结构优化; 新转录本预测

苹果(*Malus domestica*)是蔷薇科(Rosaceae)仁果亚科(Pomideae)苹果属(*Malus* Mill.), 苹果共有17对染色体, 而与苹果同属蔷薇科的草莓、桃等果树的染色体数在7~9之间。绝大多数苹果品种为二倍体(2*n*=34), 少数和极少数品种为三倍体(2*n*=51)和四倍体(2*n*=64)。2010年以苹果‘金冠’为材料进行的全基因组测序, 为苹果功能基因组学的研究奠定了重要的基础。测序发现苹果基因组大小约为742.3 Mb, 总基因数为95 216/57 386个(包含/不包含转座子、等位基因等), 基因组上42.4%为转座因子, 共发现4 021个转录因子、992个抗性因子, 苹果染色体组上的基因数超过82%以上, 90%左右的基因定位在原来确定的位置, 对其基因组进行分析发现500.7 Mb的序列为重复序列, 占整个苹果基因组的67% (Velasco等2010)。基因信息公开后, 注释文件版本为1.0, 经过不断地补充修正, 当前版本为3.0.a1, 基因组增大为1 874.77 Mb, 共注释58 380个基因位点和60 549个蛋白编码转录本(https://www.rosaceae.org/species/malus/malus_x_domestica/genome_v3.0.a1)。但是, 苹果基因仍存在注释信息不完善, 测序覆盖度不全面和基因功能不清等问题, 比如对苹果基因克隆时缺少非转录区域(untranslated region, UTR)或非转录区域基因功能不清等, 这些问题都使苹果基因组功能的研究受到限制。

转录组是指特定组织或细胞在某一发育阶段或功能状态下转录出来的所有编码RNA (mRNA)和非编码RNA (ncRNA)的总和(Costa等2010; Wang等2009)。由于近年来测序技术的不断优化, 测序成本的不断下降, 数据量更加全面, 基因组学分析软件的开发应用, 高通量测序及生物信息学分析技术成为生命科学研究的重要工具和手段(Pareek等2011)。苹果因为品种繁多、来源复杂、生物学性状特异、生物信息学研究不够系统和完善, 基因组学研究仍面临较大挑战。而转录组测序(RNA sequencing)是利用第二代高通量测序技术进行cDNA测序, 能够全面快速地获取研究材料在某一处理条件下的全部转录本信息, 可获得大量的转录本信息, 从中挖掘重要功能基因, 是揭示植物优良特性的重要研究手段(Dassanayake等2010; Li等2013)。本试验利用Illumina HiSeq 2500测序平台对苹果进行了高通量转录组测序, 预测得到1 604个新转录本并对13 272个基因结构进行了优化, 补充并完善了苹果基因组信息, 为之后的苹果组学功能研究奠定基础。

材料与方法

1 试验材料

对苹果(*Malus domestica* Mill.)进行转录组测序, 品种选用‘富士’ (‘Fuji’)作为材料, 该品种果实大、肉脆汁多、风味浓郁芳香、树势中庸丰产、是生产中的主栽品种。试验于2016年3~5月在山东省泰安市岱岳区滩清湾村20年生苹果园中进行, 对‘富士’统一用自选育海棠高效授粉树优系1379花粉进行人工授粉, 花粉采自于山东农业大学观赏果树实验站。每株随机选取150朵铃铛花进行

收稿 2016-12-26 修定 2017-07-04
资助 山东省现代农业产业技术体系创新团队(SDAIT-06-07)、现代农业产业技术体系建设专项经费(CARS-28)、公益性行业(农业)科研专项经费(201303093)和支撑计划专项经费(2014BAD6B102)。
* 通讯作者(E-mail: guanshangguoshu@163.com)。

去雄授粉, 每个花序保留1~2朵花, 随后套以无纺布, 防止外源花粉影响, 3 d后除去无纺布, 保证其正常的生长发育。

对授粉后15 d的果实进行采样, 在液氮中保存并带回, 保管放置在–80°C条件下, 每15 d采样1次, 共采样4次, 并设置3个重复。果肉组织RNA提取参照改良CTAB法, 并严格按照说明进行操作, 利用Nanodrop检测提取的RNA纯度($OD_{260/280}$比值)、Qubit对RNA浓度进行精确定量、Agilent 2100精确检测RNA的完整性, 样品检测合格后进行文库构建。

2 文库构建和测序

样品检测合格后, 用带有Oligo(dT)的磁珠富集mRNA, 随后加入fragmentation buffer将mRNA打断成短片段, 以mRNA为模板合成一链cDNA, 然后加入缓冲液、dNTPs和DNA polymerase I合成二链cDNA, 随后利用AMPure XP beads纯化双链cDNA。纯化的双链cDNA再进行末端修复、加A尾并连接测序接头, 最后进行PCR富集得到最终的cDNA文库。文库构建完成后, 先使用Qubit 2.0进行初步定量, 稀释文库至1 ng·μL^{-1}, 随后使用Agilent 2100对文库的插入片段长度(insert size)进行检测, 使用Q-PCR方法对文库的有效浓度进行准确定量(文库有效浓度>2 nmol·L^{-1}), 质检合格后, 使用Illumina双末端测序HiSeq 2500平台进行测序。

3 新转录本预测和基因结构优化

对测序得到的原始序列数据(raw reads)去除带接头(adapter)的reads、去除N (N表示无法确定碱基信息)的比例大于10%的reads、去除低质量reads得到clean reads, 因为后续分析都基于clean reads。将苹果‘金冠’基因组序列作为参考基因组进行, 用Cufflinks (2.1.1版)进行拼装, 然后用Cuffcompare和已知的基因模型进行比较, 可以发现新基因(相对于原有基因注释文件)以及已知基因新的外显子区域; 并对已知基因的起始和终止位置进行优化。

4 功能注释

采用序列对比的方法对基因数据库进行序列相似性分析, 使用BLAST程序将拼接得到基因数据库并与NCBI数据库进行比对, 选取最佳的功能注释。根据NCBI数据库的功能注释信息, 通过hmmscan软件得到新基因的GO (Gene Ontology)注释文件, 在分析中GO富集分析采用的软件为GOseq (Release 2.12版) (Young等2010; Langmead等2009)对所有的基因数据库进行GO功能注释并分类统计。KEGG (Kyoto Encyclopedia of Genes and Genomes)是系统分析基因功能、基因组信息的数据库、是进行生物体内代谢分析、代谢网络研究的强有力工具(Kanehisa等2008; Apweiler等2004), 根据NCBI数据库的功能注释信息使用KOBAS (V2.0版)软件进行Pathway富集分析。

实验结果

1 RNA-Seq测序数据质量与分析

对花后15、30、45和60 d (PT1、PT2、PT3和PT4)的苹果果实提取其总RNA, 利用 Nanodrop对RNA的浓度和纯度进行检测、并用Agilent 2100检测RNA的完整性, 检测得到的RNA浓度均在200~700 ng·μL^{-1}、总量≥9.5 μg、$OD_{260/280}$值为1.9~2.1、25S与18S rRNA的比值范围在1.5~3.0之间、RNA完整性计数(RNA integrity number, RIN)为8.0~10.0。以上结果表明样品质量满足建库测序要求, 可以进行RNA-Seq测序。

测序数据质量情况如表1所示, 4个样品(PT1、PT2、PT3和PT4)经过RNA-Seq共产生约21 800万的原始序列(raw reads)。测序得到的原始序列, 里面含有带接头的、低质量的reads, 为了保证信息分析质量, 必须对原始序列进行过滤, 得到过滤后

表1 RNA-Seq 测序质量评估

Table 1 RNA-Seq sequence quality assessment

样品名称	原始序列数据	过滤后数据	过滤后数据量	Q20/%	Q30/%	GC含量/%
PT1	60 877 858	59 047 506	8.86 G	98.05	95.26	47.34
PT2	56 200 559	54 660 780	8.20 G	98.11	95.40	47.56
PT3	52 364 578	50 859 698	7.63 G	98.09	95.35	47.69
PT4	49 274 590	47 739 653	7.16 G	98.08	95.34	48.07

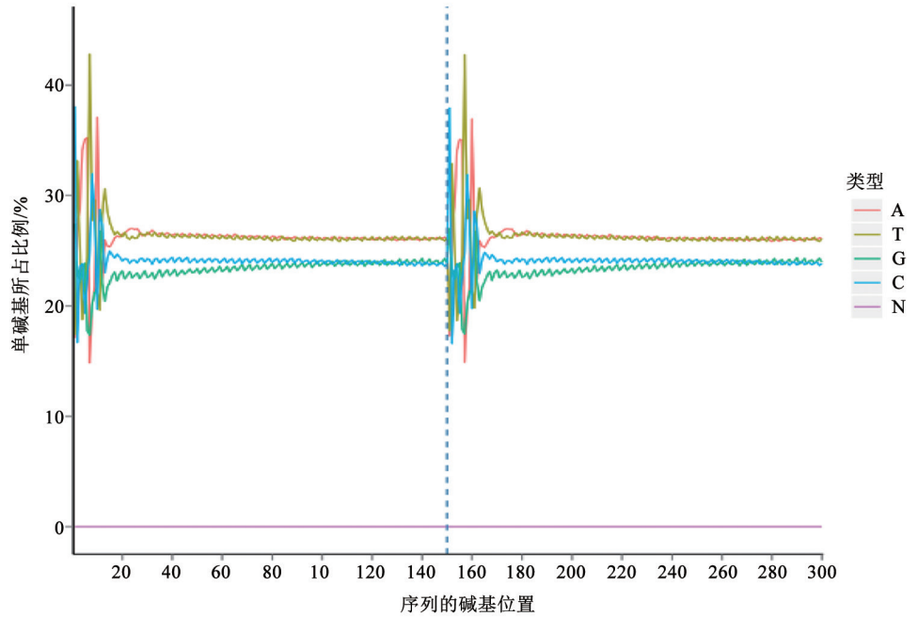

图1 原始序列碱基含量

Fig.1 Content of the original base

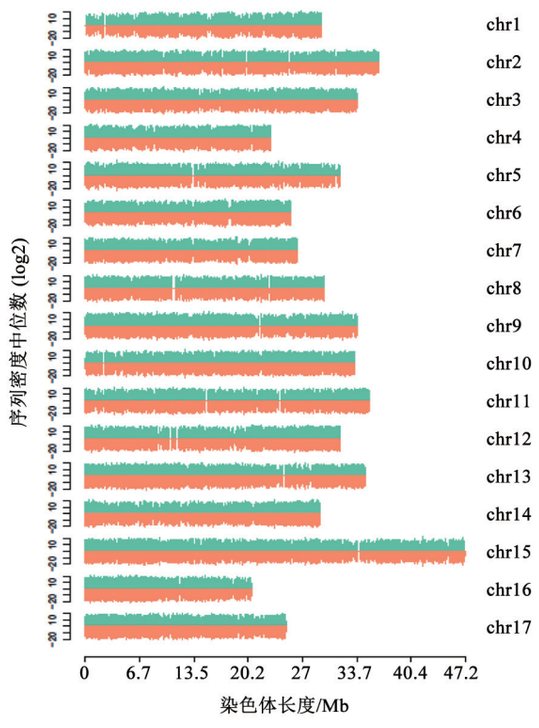

图2 样品在染色体上的密度分布图

Fig.2 Statistics of sample density distribution in chromosomes

横坐标为染色体的长度信息(以百万碱基为单位), 纵坐标为$\log_2$(reads的密度的中位数), 绿色为正链, 红色为负链。

数据(clean reads), 后续分析都基于clean reads。RNA-Seq测序的碱基质量值是碱基识别出错概率的整数映射, 使用Phred碱基质量值公式计算(Ewing等1998)。碱基质量值越高表明碱基识别准确度越高, 例如碱基质量值10 (Q10)、20 (Q20)、30 (Q30)和40 (Q40)分别表示碱基识别出错的概率为10%、1%、0.1%和0.01%。

由图1可知碱基分布类型情况, 这表明各个碱基占的比例约为25%, 含量相差不大, 由于5′端6 bp的碱基具有随机性, 测序所得每个reads前6~7个碱基出现较大波动属正常情况, G和C碱基及A和T碱基含量每个测序循环上分别相等, 且整个测序过程稳定不变, 呈水平线, 不存在碱基分离现象。

将过滤后的干净序列在基因组上的位置信息, 选取TopHat2软件(v2.0.12版), 设置参数为mismatch=2, 将过滤后的测序序列进行基因组定位分析。序列与参考基因组比对情况如表2所示, 4个样品整体数据都约有71%左右匹配到苹果参考基因组上, 其中在参考序列上有单一比对位置的测序序列(uniquely ummmapped)约为61%, 而在参考序列上有多个比对位置的测序序列(multiple mapped)约占10%, 这可能是由于这些序列对应的基因为多拷贝基因。将比对到基因组上的序列分布情况进行统

表2 样品和参考基因组比对情况

Table 2 Statistics of sample mapped to genome

样品名称	经过过滤的序列	匹配序列	匹配序列比例/%	多位置匹配序列比例/%	单一位置匹配序列比例/%	外显子比例/%	内含子比例/%	基因间区比例/%
PT1	59 047 506	42 274 443	71.59	10.51	61.12	93.43	5.40	1.17
PT2	54 660 780	39 237 274	71.78	10.78	61.11	95.03	4.20	0.73
PT3	50 859 698	36 287 392	71.35	10.29	61.18	94.60	4.67	0.73
PT4	47 739 653	34 248 247	71.74	9.94	61.84	95.33	4.03	0.60

计, 定位区域分别为外显子(exon)、内含子(intron)和基因间区(intergenic)。

将序列比对到基因组上各个染色体(分正负链)的密度情况进行统计, 如图2所示, 从定位到染色体上的序列数与染色体长度的关系中, 可以更加直观看出染色体长度序列总数的关系。在正常状态下, 整个染色体长度越长, 该染色体内部定位的序列总数会越多(Marquez等2012)。

2 新转录本预测与功能注释

利用转录组测序技术不但可以优化已知基因的结构, 而且能预测新的基因。将测序得到的序列进行拼接, 获得非重复序列基因(unigenes), 并将其与原有基因组注释信息进行比较, 找到未注释功能的片段, 将其与GO、KEGG数据库进行序列对比, 鉴定的新的转录区进行翻译, 过滤掉少于50个氨基酸残基的过短序列和只包含单个外显子的序列, 共获得1 604个新的转录本。由表3可知, 新转录本的长度大都在500 bp以上, 表示这些基因序列主要为编码蛋白的基因, 而这些基因的注释信息表明, 大部分基因参与编码未知功能的蛋白, 剩余基因主要参与苹果的转录表达调控、生长调节、抗病、氨基酸糖类等的代谢过程。

3 已知基因结构优化

由于不同测序分析软件的手段的局限性, 测序结果往往不够全面精确, 所以我们利用转录组测序结果对已知基因的结构进一步的优化和精确。在已注释基因边界之外的区域有连续的匹配读段支持, 则将基因的非翻译区域向上下游延伸, 修正基因的边界(穆彩琴等2016)。如表4和5所示, 本试验中共对13 272个转录本进行了结构优化, 其中5′端为8 843个, 而3′端为8 955个。

讨 论

由于苹果基因组测序的初步完成, 其作为果树的模式植物受到重视, 并且苹果功能基因组学的研究已经成为研究的热点。公开的苹果数据库GDR (Genome Database for Rosaceae)、IAMSA (FEMIASMA Computational Biology Web Resources)

表3 利用RNA-Seq技术鉴定的苹果新转录本

Table 3 Novel transcripts in apple identified by RNA-Seq technology

基因编号(ID)	染色体编号	基因位置/bp	+/−链	长度/bp	功能预测
Apple_new00081	4	527 575~528 438	−	571	类抗病蛋白At4g27220
Apple_new00093	12	705 587~706 860	−	792	类生长调节因子8
Apple_new00128	17	135 907~138 910	−	2 435	类反转录酶
Apple_new00166	16	592 842~594 353	−	948	类三酰基甘油脂酶1
Apple_new00167	16	1 530 149~1 531 182	−	851	类β-淀粉酶2
Apple_new00246	9	800 403~901 543	+	1 055	类F-box/花萼重复蛋白At3g06240
Apple_new00329	2	355 831~357 241	−	1 225	类真核翻译起始因子5B
Apple_new00387	9	4 513 057~4 513 820	−	764	转录因子bHLH78
Apple_new00442	8	15 944~16 755	−	467	LRR受体丝/苏氨酸蛋白激酶At1g56140
Apple_new00635	9	1 342 367~1 345 444	−	1 131	翻译起始因子IF-3
Apple_new00708	11	2 386 992~2 387 780	+	565	DNA聚合酶
Apple_new00875	10	10 818 578~10 819 833	−	1 256	生长素诱导蛋白15A

部分数据未列出。表4和5同此。

表4 部分基因3′和5′端延伸情况

Table 4 Genes with 3′ and 5′ end extended

染色体编号	5′端/个	3′端/个	转录本/个	+链/个	−链/个
1	317	330	486	242	244
2	405	421	614	302	312
3	359	370	563	283	280
4	298	276	436	222	214
5	425	390	597	303	294
6	281	304	420	195	225
7	328	346	488	234	254
8	324	299	454	214	240
9	403	415	585	273	312
10	416	421	602	283	319
11	381	396	575	277	298
12	338	361	546	258	288
13	418	382	616	319	297
14	263	274	423	217	206
15	522	522	813	409	404
16	247	286	404	194	210
17	290	301	456	233	223
其他	2 828	2 861	4 194	2 058	2 136
共计	8 843	8 955	13 272	6 516	6 756

表5 3′或5′端的延伸基因

Table 5 3′ or 5′ ends extension of selected genes

基因编号(ID)	染色体编号	+/−链	3′/5′ UTR	原位置/bp	注释后位置/bp
103400004	15	+	5′	8 327 325~8 328 937	8 327 292~8 328 937
103400123	2	-	5′	11 389 248~11 394 169	11 388 783~11 394 169
103400221	1	-	3′	8 644 185~8 648 127	8 644 185~8 648 133
103403983	16	+	5′	16 748 206~16 755 087	16 748 164~16 755 087
103405884	17	-	3′	23 512 417~23 515 744	23 512 417~23 515 750
103432349	3	+	5′	29 849 733~29 854 827	29 849 616~29 854 827
103433106	4	-	3′	6 434 985~6 437 245	6 434 985~6 437 259
103439340	7	-	5′	16 936 387~16 939 355	16 934 493~16 939 355
103440773	8	+	3′	8 192 139~8 195 524	8 192 139~8 195 546
103451146	12	+	3′	31 289 069~31 293 340	31 289 069~31 293 407

和Plaza (Jung等2008; Proost等2009)等, 提供了苹果基因组核酸或蛋白质序列、表达序列标签(EST)、Blast搜索和基因组浏览等信息, 但是缺少全面的基因组信息和基因家族分类, 以及苹果表达图谱及microRNA信息等。本试验利用RNA-Seq技术对苹果花后4个时期的果肉进行测序, 鉴定并预测新的转录本和对已知基因结构的优化, 全面补充基因组的信息, 促进基因结构的进一步优化和基因功能注释更加准确。

转录组学是反映生物个体或特定器官、组织或结构在某一特定发育、生理过程中所有基因表达水平的研究(Trapnell等2010), RNA-Seq有高通量、重复性强、检测范围广、定量精确等优点, 不仅对于已知基因组序列的物种适用, 对于未知基因组序列的物种同样适用。有学者对野生草莓25个不同器官的组织进行了转录组测序, 利用得到的数据对草莓已知基因重新进行注释, 提高了草莓基因注释的准确性和精确性, 并发现了部分新的转录本, 增长并扩充了基因的外显子(Darwish等2015)。有研究利用Solexa测序技术对盐胁迫下

的陆地棉种子进行了转录组分析, 发现了12个可能与盐胁迫有关的基因, 验证了激素信号系统与盐胁迫的相关性(Wang等2011)。有学者将数字基因表达谱应用于玉米雌穗发育的研究, 并发现了一批调控花序的新基因(Eveland等2015)。本试验中应用RNA-Seq技术鉴定了苹果1 604个新的基因, 注释信息对应苹果发育前期转录调控、氨基酸糖类代谢、生长调节等过程, 对于研究调控苹果果实发育的分子机制有重要意义。RNA-Seq不仅能鉴定并发现新的转录本, 并在进一步完善基因结构信息等方面也有重要的作用。有研究利用RNA-Seq技术对谷子已注释的基因结构进行进一步确认, 共对7 175个基因的结构进行修正, 其中3′端为4 330个、5′端为5 362个, 为更全面地理解谷子抗旱、耐瘠薄和高光效的分子机制提供了基础信息(穆彩琴等2016)。有学者将RNA-Seq测序结果与羊布鲁氏菌16 M现有基因组上基因进行比对后, 发现共有773个基因的5′或3′端在原有基础上发生了延伸, 这对后续的研究进行了进一步的验证(郭英飞等2015)。

综上所述, 我们将RNA-Seq技术成功地用于苹果的研究中, 优化了苹果已注释基因的结构, 并发掘了新的转录本。这些结果更加全面地揭示了苹果全基因组, 也为更好地了解苹果果实发育的规律奠定了基础。

参考文献

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Res, 32 (S1): D115–D119

Costa V, Angelini C, De Feis I, Ciccodicola A (2010). Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol, 2010 (1): 853916

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2010). Shedding light on an extremophile life style through transcriptomics. New Phytol, 183: 764–775

Darwish O, Shahan R, Liu ZC, Slovin JP, Alkharouf NW (2015). Re-annotation of the woodland strawberry (*Fragaria vesca*) genome. BMC Genomics, 16 (1): 1–9

Ewing B, Hillier L, Wendl MC, Green P (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res, 8 (3): 175–185

Eveland AL, SatohNagasawa N, Goldshmidt A (2010). Digital gene expression signatures for maize development. Plant Physiol, 154 (3): 1024–1039

Guo YF, Wang YF, Gong CL, Yang MJ, Yuan JY, Zhuang YB, Ke YH, Du XY, Wang ZJ, Chen ZL (2015). Identification of novel transcripts and sRNA of *Brucella melitensis* by RNA-Seq. Chin J Zoonoses, 31 (3): 216–221 (in Chinese with English abstract) [郭英飞, 王玉飞, 龚春丽, 杨明娟, 袁久云, 庄妤冰, 柯跃华, 杜昕颖, 汪舟佳, 陈泽良(2015). 基于RNA-Seq的羊种布鲁氏菌新转录本与非编码RNA鉴定. 中国人兽共患病学报, 31 (3): 216–221]

Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D (2008). GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res, 36: 1034–1040

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36: 480–484

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 10 (3): 1–10

Li CQ, Wang Y, Huang XM, Li J, Wang HC, Li JG (2013). De novo assembly and characterization of fruit transcriptome in *Litchi chinensis* Sonn and analysis of differentially regulated genes in fruit in response to shading. BMC Genomics, 14 (1): 21–43

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res, 22 (6): 1184–1195

Mu CQ, Zhang RJ, Qu CL, Han YH, Wang XC, Yang ZR (2016). Identification of novel genes and optimization of annotated genes in foxtail millet by RNA-Seq technology. Plant Physiol J, 52 (7): 1066–1072 (in Chinese with English abstract) [穆彩琴, 张瑞娟, 屈聪玲, 韩渊怀, 王兴春, 杨致荣(2016). 基于RNA-Seq技术的谷子新基因发掘及基因结构优化. 植物生理学报, 52 (7): 1066–1072]

Pareek CS, Smoczynski R, Tretyn A (2011). Sequencing technologies and genome sequencing. J Appl Genet, 52 (4): 413–435

Proost S, Van BM, Sterck L, Billiau K, Van PT, Vande PY, Vandepoele K (2009). PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell, 21 (12): 3718–3731

Trapnell C, Williams B A, Pertea G, Mortazavi A, Kwan G, van Baren M J, Salzberg S L, Wold B J, Pachter L (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Bio Technol, 28 (5): 511–515

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, et al (2010). The genome of the domesticated apple (*Malus Domestica* Borkh). Nat Genet, 42 (10): 833–839

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10 (1): 57–63

Wang C, Zhang LJ, Huang RD (2011). Cytoskeleton and plant salt stress tolerance. Plant Signal Behav, 6 (1): 29–31

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol, 11 (2): R14

Identification of novel transcripts and optimization of annotated genes in apple (*Malus domestica*) by RNA-Seq

NI Wei, GAO Fu-Feng, YANG Heng-Feng, ZHANG Jia-Teng, XU Jin, MAO Zhi-Quan, CHEN Xue-Sen, SHEN Xiang*

State Key Laboratory of Crop Biology, College of Horticultural Science and Engineering, Shandong Agricultural University, Taian, Shandong 271018, China

Abstract: Although the apple (*Malus domestica*) genome sequencing has been successfully accomplished and previous promulgated, it is incomplete in terms of the genic annotation information. In this study, we used the RNA-Seq technology to optimize the annotated gene organization and predicate the novel transcript. We use fruit after blooming 15, 30, 45, 60 days as material to isolated total RNA by 'Fuji'. The cDNA library was constructed for the RNA samples and sequenced on the Illumina HiSeq 2500 platform. The obtained clean reads of high quality were assembled compared with the reference genome sequence, we had amended the 5′ ends or 3′ ends of the 13 727 genes in the original genome level, and optimized 8 843 5′ ends of genes, and 8 955 3′ ends of genes, respectively. Furthermore, 1 604 novel transcripts were identified, and the partial genic annotation information participated the apple metabolic process of transcription regulation, growth regulation, disease resistance, amino acids, and glycometabolism.

Key words: apple (*Malus domestica*); RNA-Seq; gene structural optimization; novel gene prediction

Received 2016-12-26 Accepted 2017-07-04

This work was supported by the Innovative Team of Modern Agricultural Industry Technology Systerm of Shandong (Grant No. SDAIT-06-07), the Construction of Modern Agricultural Industry Technology System (Grant No. CARS-28), the Public Welfare Industry (Agriculture) (Grant No. 201303093), and the Support Plan (Grant No. 2014BAD6B102).

*Corresponding author (E-mail: guanshangguoshu@163.com).