科学通报
Chinese Science Bulletin

基于PacBio平台的全长转录组测序

任毅鹏, 张佳庆, 孙瑜, 吴振峰, 阮吉寿, 贺秉军, 刘国卿, 高山 and 卜文俊

Citation: 科学通报 **61**, 1250 (2016); doi: 10.1360/N972015-01384

View online: http://engine.scichina.com/doi/10.1360/N972015-01384

View Table of Contents:http://engine.scichina.com/publisher/scp/journal/CSB/61/11

Published by the 《中国科学》杂志社

科学通报 2016 年 第 61 卷 第 11 期：1250 ~ 1254
论 文
《中国科学》杂志社
SCIENCE CHINA PRESS

基于 PacBio 平台的全长转录组测序

任毅鹏[①], 张佳庆[①], 孙瑜[①], 吴振峰[②], 阮吉寿[②], 贺秉军[①], 刘国卿[①], 高山[①*], 卜文俊[①*]

① 南开大学生命科学学院, 天津 300071;
② 南开大学数学科学学院, 天津 300071
* 联系人, E-mail: gao_shan@mail.nankai.edu.cn; wenjunbu@nankai.edu.cn

2015-12-14 收稿, 2016-01-12 修回, 2016-01-13 接受, 2016-03-03 网络版发表
南开大学 2015 年研究生科研创新计划和国家自然科学基金(31371974, 31201738)资助

摘要 当前, 绝大多数的转录组数据都是基于以Illumina平台为代表的第二代高通量测序技术获得的, 但是第二代测序技术无法提供大量的长转录本并且丢失可变剪接等重要信息, 因而大大制约了转录组数据的深度利用. 通过以PacBio为代表的第三代测序技术, 可以获得更长乃至全长转录组, 但由于PacBio转录组测序近几年才刚兴起, 只有少量的物种基于PacBio平台获得了转录组. PacBio全长转录组测序, 在国际上才刚开展但发展很快, 其实验与数据分析标准和质量控制方面的研究对于未来的大规模应用至关重要. 本研究在国际上首次尝试依据PacBio平台最新试剂(P6/C4)优化实验参数, 设计质量控制指标并使全长转录组测序标准化. 本文基于一组昆虫(麻皮椿)全长转录组数据, 对已取得的部分结果进行报告.

关键词 全长转录组, 单分子测序, PacBio, 质量控制, 标准流程

基因组和转录组测序是生命科学领域的基础性工作. 由于绝大部分非模式生物缺乏基因组数据, 全长转录组测序就变得尤为重要, 全长转录本可以大大促进这些物种的基因功能、基因表达调控和进化关系等多方面的基础与应用研究. 但是, 由于RNA易降解和不同的转录本表达量差异巨大等多种原因, 从总RNA中获取尽量多的全长转录本难度巨大. 当前, 绝大多数的转录组数据都是基于第二代高通量测序技术获得的, 第二代测序技术测序序列短, 短序列拼接无法提供大量的长转录本并且丢失可变剪接等重要信息, 因而转录组的从头测序开始采用PacBio第三代测序技术. PacBio RS系列测序仪, 是基于单分子实时(single molecule real time, SMRT)测序技术的单分子测序仪, 由美国Pacific Biosciences (PacBio)公司设计制造[1]. PacBio RS Ⅱ型测序仪结合最新的P6/C4试剂(2014年8月15日推出), 可以获得高达12000 bp平均长度的测序读段, 大大促进了获得完整基因组和全长转录组的能力. 根据公开发表的文献, 只有少量物种基于PacBio平台获得了转录组: 这其中包括第二、三代测序混合拼接或第二代矫正第三代技术获得的人的类淋巴母细胞[2]和丹参(*Salvia miltiorrhiza*)[3]转录组数据; 完全基于PacBio平台获得的转录组绝大部分来自人类[4,5], 另外也有HIV病毒[6]、牛(*Bovine*)[7]、小鼠(*Mus musculus*)[8]和克氏冕狐猴(*Propithecus coquereli*)[9]等; 然而, 基于PacBio平台的全长转录组测序方面的研究, 国际上才刚开展, 直到2015年才有真菌(*Fungi*)[10]、四倍体棉花(*Gossipium hirsutum*)[11]和欧洲乌贼(*Sepia officinalis*)[12]等物种测序. 本研究在大量实验与数据分析的基础上, 在国际上首次尝试依据PacBio平台最新试剂(P6/C4)优化实验参数, 设计质量控制指标并使全长转录组测序标准化. 本文报道的部分研究结果基于

引用格式: 任毅鹏, 张佳庆, 孙瑜, 等. 基于 PacBio 平台的全长转录组测序. 科学通报, 2016, 61: 1250–1254
Ren Y P, Zhang J Q, Sun Y, et al. Full-length transcriptome sequencing on PacBio platform (in Chinese). Chin Sci Bull, 2016, 61: 1250–1254, doi: 10.1360/N972015-01384

一组昆虫(麻皮椿(*Erthesina fullo* Thunberg))全长转录组数据但不限于昆虫, 具有更为普遍的意义.

1 材料与方法

本研究使用的麻皮蝽于2015年5月购自江西神农生物技术公司, 饲养到11月20日选取雄性和雌性麻皮蝽各1只提取总RNA. 首先, 在无菌环境下去除麻皮蝽整个腹部; 而后将腹部以外的全部组织(约200 mg)混合, 用UNIQ-10 Trizol Total RNA Extraction Kit (Biotech, 上海)试剂盒提取总RNA; 最后, 加灭菌的焦碳酸二乙酯(diethy pyrocarbonate, DEPC)水溶解得到20 μL总RNA溶液. 取1 μL总RNA溶液用Qubit 2.0(Life Tech, 美国)测定RNA浓度为1.72 μg/μL, 另取1 μL总RNA溶液根据琼脂糖凝胶电泳结果确定RNA无降解. 剩余18 μL(30.96 μg)总RNA溶液, 参照SMARTer® PCR cDNA Synthesis Kit(Clontech, 美国)试剂盒说明合成cDNA, 根据优化后的参数对cDNA进行PCR扩增和纯化(图1). 上述产物根据PacBio标准流程建库测序.

2 结果

全长转录组测序的基本原则是尽量最大限度获取含有5′端帽子结构的完整RNA, 两步关键因素: 第1步就是cDNA合成中的全长反转录; 第2步就是cDNA的PCR扩增. 为了保证第2步扩增的目标尽量覆盖各种长度的转录本(即不漏), 需要在多个环节进行质量控制, 这里汇报3个环节的工作. 第1个环节是cDNA扩增循环次数的确定: 首先从总cDNA中取少量样品配制50 μL反应液, 在98℃温度下进行2 min的初始变性, 而后进行10个PCR循环(98℃ 20 s; 65℃ 30 s; 72℃ 3 min 30 s), 最后在72℃温度下进行5 min的延伸; 从50 μL反应液中取出5 μL产物, 剩下的45 μL继续进行2个PCR循环; 以此类推, 最后分别得到10, 12, 14, 16, 18和20个循环的产物, 进行琼脂糖凝胶电泳供人工判读(图2). 人工判读的标准是尽量选择高产量、低循环数的产物, 要求主带清晰, 避免产生多余的条带或者分布向小片段偏移, 麻皮蝽全长转录组测序最终确定为12个循环进行后续实验. 第2个环节是对总cDNA扩增产物取样进行琼脂糖凝胶电泳(图3(a)), 查看条带分布, 标准是亮度中心区与已知的mRNA长度分布的峰值接近, 亮度向上下两个方向逐渐递减且不出现大面积的缺失. 第3个环节就是通过片段筛选进行分级扩增, 由于全长转录本的跨度非常大(可以从200 bp到10 Kbp以上), 根据某

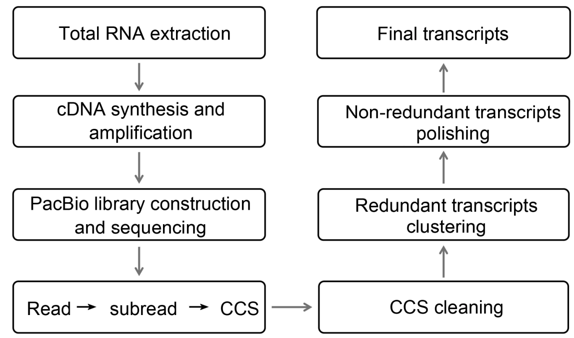

图 1 全长转录组测序的标准化流程. Read: 测序读段; Subread: 测序子读段; CCS: 环形一致读段; Redundant transcripts: 冗余转录本; Non-redundant transcripts: 非冗余转录本; Final transcripts: 最终转录本

Figure 1 Standardized protocol of full-length transcriptomic sequencing

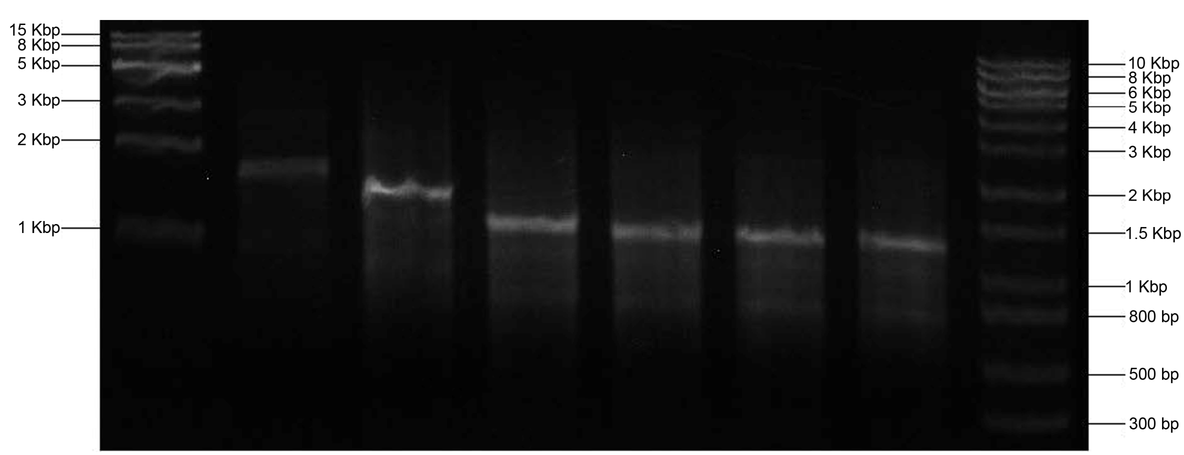

图 2 麻皮蝽cDNA扩增前循环数的确定

Figure 2 Optimation of PCR parameters for *E. fullo* cDNA amplification

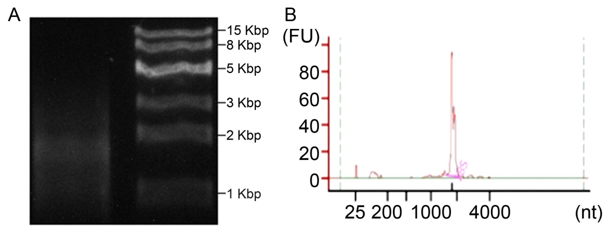

图 3 麻皮蝽mRNA与cDNA扩增产物长度分布

Figure 3 Size distribution of *E. fullo* mRNA and amplificated cDNA product

一个条件对cDNA进行PCR扩增, 会导致较短的或高丰度转录本抢占大部分扩增机会, 从而导致更长的转录本没有得到扩增, 进而无法通过测序获得其序列. 片段筛选分为不筛选、琼脂糖凝胶电泳筛选(manual agarose-gel size selection)以及通过BluePippin系统筛选(BluePippin size-selection system). 以上3个环节互相影响, 其参数的最终确定严重依赖样品RNA总量及RNA分布等多个因素. 这部分的工作还需要更大量数据积累经验, 将实验参数的确定进一步公式化, 以避免过多人工经验判定. 根据此前麻皮蝽及大量昆虫mRNA分布研究, 麻皮蝽mRNA长度集中以1.6 Kbp为中心的一个狭小区域, 因此全长转录组测序不采用分级扩增策略, 总cDNA扩增产物显示的结果(图3(a))与Agilent 2100 Bioanalyzer (Agilent, 美国)测定的麻皮蝽mRNA长度分布一致(图3(b)).

本次实验共使用7个芯片(SMRT cell)进行测序, 由于PacBio RS系列测序仪不集中输出质量控制报告, 本研究利用测序输出文件获得一些与测序数据质量相关的信息, 并输出报告(表1). 表1中, 第2列表示此次测序每个芯片共有150292个零模波导孔(zero-mode waveguides, ZMW)可以产生测序读段; 第3~5列分别表示产出0, 1和2条以上测序读段的ZMW的比例, ZMW(1)越高有效数据就越多, ZMW(2)越低越好, ZMW(2)过高可能由于样品过载, 产生的数据就不可靠; 第6和7列, 是测序得到的读段的平均和总长度, 前者越长测序效果越好, 总长度越大, 数据量越大; 第8和9列, 是测序得到的子读段的平均和总长度, 子读段反映了实际RNA的长度, 其长度越长表示cDNA文库扩增越好, 总长度越大, 有效数据量越大. 更多的有关质量控制的指标, 参见高山等编著的《PacBio单分子测序指南》. 质量控制信息提取与分析的脚本已整合进公开发表的测序质控软件Fastq_clean[13].

3 讨论和结论

PacBio平台发展迅猛, 其测序长度和通量的增加潜力依然巨大, 有可能替代第二代测序技术, 成为基因组和转录组从头测序的首选. 未来的转录组测序优选方案是基于PacBio平台获得样品的全长转录组, 而后以全长转录组为参考应用Illumina双端测序数据定量. PacBio全长转录组测序在建立标准流程, 优化实验参数和质量控制方面的研究将是未来基因组学领域的一个重要研究方向.

表 1 麻皮蝽全长转录组测序部分质量控制信息[a)]

Table 1 Data quality control information of *E. fullo* full-length transcriptome[a)]

Cell	ZMW (Total)	ZMW (0)	ZMW (1)	ZMW (2)	ReadLength Mean (bp)	ReadLength Sum (bp)	SubreadLen Mean (bp)	SubreadLen Sum (bp)
1	150292	2.49%	63.40%	34.11%	18197.92	1733989042	1343.04	1677077878
2	150292	55.93%	33.80%	10.28%	16244.45	825153167	1262.81	798051033
3	150292	51.59%	39.16%	9.25%	15689.09	923271331	1209.24	891651900
4	150292	69.72%	24.07%	6.21%	14383.92	520352809	1336.60	504494097
5	150292	66.05%	29.85%	4.10%	15815.45	709576198	1171.01	684698375
6	150292	73.37%	22.99%	3.65%	16809.27	580709969	1187.09	560512107
7	150292	53.95%	40.51%	5.54%	16690.74	1016065750	1155.56	980046359

a) ZMW: 零模波导孔; ReadLength Mean: 测序读段的平均长度; ReadLength Sum: 测序读段总长度; SubreadLen Mean: 测序子读段的平均长度; SubreadLen Sum: 测序子读段总长度

致谢 感谢天津大学药学院张耀洲教授和浙江理工大学生命科学学院陈健教授在全长转录组建库方面的帮助, 同时感谢国家人口与健康科学数据共享平台肿瘤转化医学专题服务项目提供计算资源.

参考文献

1 Gao S, Ou J H, Xiao K. Using R and Bioconductor in Bioinformatics (in Chinese). Tianjin: Tianjin Science and Technology Translation Publishing Co, 2014. 33–34 [高山, 欧剑虹, 肖凯. R 语言与 Bionconductor 生物信息学应用. 天津: 天津科技翻译出版有限公司, 2014. 33–34]

2 Hagen T, Fabian G, Donald S, et al. Defining a personal, allelespecific, and singlemolecule longread transcriptome. Proc Natl Acad Sci USA, 2014, 111: 9869–9874

3 Xu Z, Peters R J, Weirather J, et al. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of salvia miltiorrhiza and tanshinone biosynthesis. Plant J, 2015, 82: 951–961

4 Kin F A, Vittorio S, Pegah T A, et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proc Natl Acad Sci USA, 2013, 110: E4821–E4830

5 Sharon D, Tilgner H, Grubert F, et al. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol, 2013, 31: 1009–1014

6 Ocwieja K E, Sherrill-Mix S, Mukherjee R, et al. Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. Nucleic Acids Res, 2012, 40: 10345–10355

7 Larsen P A, Smith T P. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. BMC Immunol, 2012, 13: 52

8 Barbara T, Ozgun G, Stephen R Q, et al. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proc Natl Acad Sci USA, 2014, 111: E1291–E1299

9 Larsen P A, Campbell C R, Yoder A D. Next-generation approaches to advancing eco-immunogenomic research in critically endangered primates. Mol Ecol Resour, 2014, 14: 1198–1209

10 Sean P G, Elizabeth T, Asaf S, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One, 2015, 10: e0132628

11 van Eijk M. Genome assembly and Iso-Seq transcriptome sequencing of tetraploid cotton. Plant and Animal Genome XXIII Conference. Plant and Animal Genome, San Diego, 2015

12 Worley K C. European cuttlefish whole transcriptome sequencing: A single-molecule full length transcript survey with Iso-Seq method. Plant and Animal Genome XXIII Conference. Plant and Animal Genome, San Diego, 2015

13 Zhang M, Sun H, Fei Z J, et al. Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control. Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on. IEEE, Belfast, 2014. 44–48

Full-length transcriptome sequencing on PacBio platform

REN YiPeng[1], ZHANG JiaQing[1], SUN Yu[1], WU ZhenFeng[2], RUAN JiShou[2], HE BingJun[1], LIU GuoQing[1], GAO Shan[1] & BU WenJun[1]

[1] *College of Life Sciences, Nankai University, Tianjin 300071, China;*
[2] *College of Mathematics, Nankai University, Tianjin 300071, China*

The Next Generation Sequencing (NGS) technology, particularly the Illumina platform now has produced most of the animal and plant transcriptomes, but the short reads from NGS sequencers result in incompletely assembled transcripts which are lack of some important information (e.g. alternative splicing). This limits better understanding of transcriptome data. Based on the single-molecule real-time (SMRT) sequencing technology, the PacBio platform can provide longer and even full-length transcripts that originate from observations of single molecules without assembly. The full-length transcripts can be used to investigate alternative splicing, alternative polyadenylation, novel genes, non-coding RNAs and fusion transcripts, *et al*. Until the end of 2015, transcriptomes of a few species have been sequenced using the PacBio platform. They are classfied into three groups. The first group includes human lymphoblastoid and *Salvia miltiorrhiza* using a combination of NGS short reads and SMRT technology. The second group includes HIV-1, bovine immunoglobulin G, human embryonic stem cells, mouse neurexins and *Propithecus coquereli* using SMRT. The third group includes european cuttlefish, tetraploid cotton and fungi using SMRT with the latest PacBio full-length transcriptome data analysis pipeline IsoSeq.

The use of SMARTer PCR cDNA Synthesis Kit and the IsoSeq data analysis pipeline was recommended to facilitate full-length transcriptome sequencing. However, the transcriptome data quality could be affected by ribosomal RNA contamination, cross-contamination on agarose gel, the effect of size selection using gel or BluePippin, prevalence of PCR chimera products and the wrong removal of SMRT bell adapters. Although IsoSeq can remove artificial concatemers that are produced due to insufficient SMRT bell amount during the sequencing library preparation step, some problems still exists. For example, IsoSeq can not distinguish PCR chimeras from true fusion genes. Another critical problem is the misidentification of 5′ and 3′ primers due to sequencing errors or partial trimming of them as the SMRT bell adapters. This could provide the wrong strand information of transcripts for further analysis. In addtion, transcripts of the same gene are difficult to be clustered without the genome guide. Therefore, it is necessary to standardize the experiment and data analysis protocols and design quality control measures of the full-length transcriptome sequencing technology for its application in a large scale.

In this study, we sequenced the first full-length insect transcriptome using the *Erthesina fullo* Thunberg as material. Seven SMRT cells on PacBio RS II sequencer were used to produce 381,394 reads with 16,262 bp average size. Totally 6 Gbp effective data was used for further analysis on the optimization of experimental parameters, design of quality control measures and standardization of protocols using the new PacBio reagents (P6/C4). Some of results in this study were reported to provide useful information to help better understanding the full-length transcriptome sequencing technology and designing experiments.

full-length transcriptome, single molecule sequencing, PacBio, quality control, standard protocol

doi: 10.1360/N972015-01384