

DOI: 10.3724/SP.J.1005.2013.00983

dRNA-seq 原理及其在原核生物转录组研究中的应用

侯志伟^{1,2}, 王赞³, 高宏¹, 侯圣伟¹

1. 中国科学院水生生物研究所, 淡水生态与生物技术国家重点实验室, 武汉 430072;
2. 中国科学院大学, 北京 100049;
3. 河南城建学院生物工程系, 平顶山 467036

摘要: 第二代测序技术不仅推进了基因组学的研究, 而且也广泛应用到转录组学的研究中。原核生物转录组学的研究方法主要有 Tiling 芯片和 RNA-seq, 后者因其较高的覆盖率和分辨率以及越来越低廉的成本而逐渐被更多的研究单位采用。根据对 mRNA 富集方法或 rRNA 去除方法的不同, 目前 RNA-seq 方法主要有 6 种, 其中 dRNA-seq 由于采用了 5' 单磷酸依赖的外切酶, 可以特异性地降解加工过的转录本, 从而使初始转录本得到富集, 已被广泛应用到原核生物转录起始位点、小的调控 RNA、启动子及操纵子等研究中。文章对 dRNA-seq 的原理、技术流程及其在原核生物转录组研究中的应用进行了综述, 并对这一研究方法在现阶段应用的优势和局限性进行了讨论, 也进一步对该方法在未来的发展和应用进行了展望, 期望为国内相关领域研究人员提供参考。

关键词: dRNA-seq; 转录组测序; 小 RNA; 转录起始位点; 非编码 RNA

The principle of dRNA-seq and its applications in prokaryotic transcriptome analyses

HOU Zhi-Wei^{1,2}, WANG Yun³, GAO Hong¹, HOU Sheng-Wei¹

1. The State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Department of Bioengineering, Henan University of Urban Construction, Pingdingshan 467036, China

Abstract: The genomic and transcriptomic studies have been greatly accelerated by the next-generation sequencing technology. Currently, there are two main methods in transcriptomic studies, namely, Tiling microarray and RNA-seq. Comparatively, because of its overwhelming superiority in transcript coverage and resolution, as well as decrease of costs, RNA-seq has been employed in transcriptome analyses by an increasing number of research institutes. According to the mRNA enrichment or rRNA depletion methods, RNA-seq can be performed in six ways. Among them, dRNA-seq can discriminate the original transcripts from the processed RNAs based on a differential exonuclease treatment. This kind of method has been broadly applied in studies of prokaryotic transcriptional start sites, small regulatory RNAs, promoters and

收稿日期: 2013-03-20; 修回日期: 2013-05-16

基金项目: 国家自然科学基金项目(编号: 31270132)资助

作者简介: 侯志伟, 硕士研究生, 专业方向: 蓝藻分子遗传学。Tel: 027-68780171; E-mail: zuazhiwei@126.com

通讯作者: 高宏, 副研究员, 研究方向: 蓝藻分子遗传学。E-mail: whihb2006@gmail.com

侯圣伟, 硕士研究生, 专业方向: 蓝藻分子遗传学。E-mail: housw2010@gmail.com

网络出版时间: 2013-6-19 16:29:02

URL: <http://www.cnki.net/kcms/detail/11.1913.R.20130619.1629.002.html>

operons, etc. Here, we reviewed the detailed principle and technical process of dRNA-seq and its applications in prokaryotic transcriptome analyses. Besides, we also summarized the advantages and limitations of dRNA-seq at present stage, and then gave our perspective of its future development and potential applications, expecting to present useful references for civil researchers in related fields.

Keywords: dRNA-seq; transcriptome sequencing; sRNA; TSS; ncRNA

基于 Solexa、454 和 SOLiD 平台第二代测序技术的发明和广泛应用极大地促进了基因组结构和功能元件的研究,在基因组测序、基因组比较、转录组分析、基因表达谱分析和小分子 RNA 测定等方面表现出强大的技术优势^[1,2]。截至 2013 年 5 月 15 日,已有 4 152 种细菌和 254 种古生菌的基因组测序完成(数据来源于 <http://www.ncbi.nlm.nih.gov/genomes>)。然而,这些基于 DNA 序列的基因组数据并不能反映基因组上编码区和非编码区的转录水平和转录调控机制。因此,从 RNA 水平研究原核生物在特定生长时期和培养条件下全基因组转录情况的转录组学为这一问题提供了解决方案。它通过研究特定环境条件下基因编码 RNA(mRNA)和非编码 RNA(Non-coding RNA, ncRNA)的表达水平及其调控规律,进一步推动了基因功能和 ncRNA 调控机制的研究。

目前,广泛用于原核生物转录组分析的方法主要有 2 种,即基于杂交技术的 Tiling 芯片法(Tiling microarray)和基于第二代高通量测序平台的转录组测序方法(RNA-seq)。Tiling 芯片是由许多相互重叠的短的寡核苷酸探针组成,这些寡核苷酸片段拼接起来能够覆盖部分或整个基因组序列。将标记的转录本与 Tiling 芯片上的这些寡核苷酸杂交,根据杂交释放出的荧光信号来推断转录本的序列和表达水平。这种方法是基于成熟的基因芯片技术发展起来的,有完善的实验操作和数据分析方法,常被用来大规模分析基因表达水平和寻找具有调控作用的非编码小 RNA(Small regulatory RNA, sRNA)^[3,4]。由于寡核苷酸探针的设计必须有相应物种的基因组序列作参照,因此该方法只能用于基因组序列已知的物种的转录组分析。另外, Tiling 芯片的分辨率与寡核苷酸探针的密度相关,想要精确鉴定转录本的序列就必须

提高探针的数量和重叠度,这使得 Tiling 芯片的成本大大增加。此外,由于 Tiling 芯片是通过判断杂交信号的强弱来间接反映转录本表达水平,因此容易受到背景信号和交错杂交的干扰,也不能用于低丰度转录本的检测^[5]。第二代测序技术的出现使得测序成本大大降低,促进了其在转录组学研究领域的应用。RNA-seq 就是在这种背景下应运而生,它既不需要预先设计探针,也不需要参考基因组序列,同时也实现了以更高的分辨率和覆盖度对细胞内所有转录本进行直接检测。

RNA-seq, 即 RNA 测序又称转录组测序,是最近发展起来的利用深度测序方法进行转录组分析的技术^[6]。RNA-seq 方法有很多种,本文对一种能够区分初始转录本和加工过的转录本的新的 RNA-seq 技术—differential RNA sequencing(dRNA-seq)进行综述,着重介绍 dRNA-seq 的原理及其在原核生物转录组学研究中的应用。

1 dRNA-seq 的原理

1.1 RNA-seq 概述

RNA-seq 主要分为 4 步^[7]: (1)从待测组织或细胞样品中提取高质量的总 RNA, 去除 rRNA; (2)根据测序平台读长的要求将 RNA 片段化, 反转录为 cDNA; (3)将 cDNA 文库进一步扩增至测序需求(单分子测序不需要扩增 cDNA 文库); (4)高通量测序和转录组数据分析。由于原核生物 mRNA 和具有调控功能的 sRNA 只占细胞内总 RNA 的 2%~5%, 其余 95%为 rRNA 和 tRNA^[8], 而且这些 mRNA 的 3'末端没有 poly(A)尾巴, 因此对 mRNA 的富集方法或对 rRNA 和 tRNA 去除方法的研究成为 RNA-seq 探讨的热点。

常见的原核生物 rRNA 和 tRNA 去除方法主要

有 6 种: (1) rRNA 消减杂交法(rRNA subtractive hybridization)^[9-14], 通过使用与 rRNA 互补的寡核苷酸探针与总 RNA 样品杂交去除 16S 和 23S rRNA。目前, 已有多家公司开发了相应的试剂盒: MICROBExpress bacterial mRNA enrichment kit(Ambion), RiboMinus bacteria transcriptome isolation kit(Invitrogen) 和 Ribo-Zero rRNA removal kit(Epicentre)。(2) 5'单核苷酸依赖的外切酶处理法(Terminator 5'-phosphate-dependent exonuclease treatment, TEX treatment)主要使用 Epicentre 开发的 mRNA-ONLY prokaryotic mRNA isolation kit。这两种方法相对比较成熟, 使用比较普遍。(3) 选择性引物扩增法(Not-so-random primers amplification, NSR amplification)^[15], 通过预先设计一系列的具有非 rRNA 偏好性的 NSR 引物, 然后使用这些 NSR 引物选择性扩增 rRNA 以外的转录本。相应的试剂盒有 Ovation prokaryotic RNA-seq system(NuGEN)。(4) 依赖于双链特异核酸酶的 cDNA 均一化法(Duplex-specific nuclease normalization, DSN normalization)^[16, 17], 这种方法是从真核生物借鉴而来, 是在 cDNA 水平上降低 rDNA 和 tDNA 所占的比例, 在原核生物中应用的还不多。相应的试剂盒有 Evrogen 公司的 trimmer-direct cDNA normalization kit。(5) 大肠杆菌 poly(A)聚合酶加尾法^[18], 通过使用可以选择性地对 mRNA 加尾的细菌 poly(A)聚合酶来相对偏好性地获取 mRNA, 但与前几种方法相比, 这种方法去除 rRNA 的效率较低, 常结合其他方法一起使用^[19]。相应的试剂盒有 Ambion 公司的 MessageAmp II-bacteria kit。(6) 与 RNA 结合蛋白 Hfq 等免疫共沉淀法^[20]。由于 Hfq 能够高效地结合 sRNA, 并能辅助它们与靶标 mRNA 结合, 因此常用于 sRNA 及其靶标 mRNA 的研究^[20, 21]。每种方法各有优劣, 对这些方法的比较可以参考相关文献^[8, 14, 19, 22], 本文只介绍 dRNA-seq 中用到的 TEX 处理法的原理和 cDNA 文库制备流程。因为相对于其他方法, 这种方法能够区分初始转录本和加工过的转录本, 可以用来在全基因组水平上鉴定转录起始位点(Transcriptional start site, TSS)、转录本加工位点和 ncRNA, 对分析启动子区域、操纵子结构和 ncRNA 调控机制非常有帮助, 而这些正是当前基因组注释、功能基因组学研究及 RNA 调控机制研究急需的信息。

1.2 dRNA-seq 的原理和技术流程

在原核细胞中, 根据 5'端带有的化学基团的不同, 可以将 RNA 分为 3 类: 即带有 5'三磷酸的 RNA 分子(5'-PPP-RNA), 带有 5'单磷酸的 RNA 分子(5'-P-RNA)和带有 5'羟基的 RNA 分子(5'-OH-RNA)。通常情况下, 初始转录本都是 5'-PPP-RNA, 包括初始的 mRNA 和 ncRNA。而经过加工后成熟的 5S rRNA、16S rRNA、23S rRNA 和 tRNA 都是 5'-P-RNA。另外, 5'-OH-RNA 通常是由 5'-PPP-RNA 和 5'-P-RNA 部分降解形成的。dRNA-seq 中用到的 cDNA 文库通常由两部分构成, 其中一部分使用 5'单磷酸依赖的外切酶(Terminator 5'-phosphate-dependent exonuclease, TEX)处理, 选择性的降解加工过的转录本, 只保留初始转录本的信息(TEX+); 另一部分不经 TEX 处理, 保留初始转录本和加工过的转录本

信息(TEX-)^[23, 24]。通过对两个文库中转录本的比较就能够在全基因组水平上鉴定转录起始位点和加工位点。典型的 dRNA-seq 技术流程如图 1 所示^[25]。

1.2.1 高质量总 RNA 样品的制备

首先, 收集特定培养条件或生长时期的原核细胞, 提取高质量总 RNA。为了尽量降低 rRNA 降解片段的干扰, 整个操作过程中一定要避免污染外源 RNase 和剧烈震荡, 以确保提取的总 RNA 的完整性。然后使用 DNase I 去除残留的基因组 DNA, PCR 检测是否消化干净, 再进一步检测 RNA 的完整性。然后将高质量的总 RNA 分为两份, 分别用于构建 TEX-和 TEX+ cDNA 文库。此时这两个文库主要由 5'-PPP-RNA、5'-P-RNA 和 5'-OH-RNA 构成。

1.2.2 5'单磷酸外切酶处理

将制备的 RNA 样品置于 90 °C 变性 2 min, 冰上冷却 5 min, 充分破坏 RNA 的二级结构。然后向 TEX+文库中加入适量 5'单磷酸依赖的外切酶降解 5'-P-RNA, 同时向 TEX-文库中加入等量的 RNase-free 水做对照。TEX+文库中 5'-P-RNA 在这一步会被降解, 5'-PPP-RNA 和 5'-OH-RNA 则被保留下来。

1.2.3 烟草酸焦磷酸酶处理

有机溶剂抽提去除 TEX 后, 使用烟草酸焦磷酸酶(Tobacco acid pyrophosphatase, TAP)处理 TEX+和

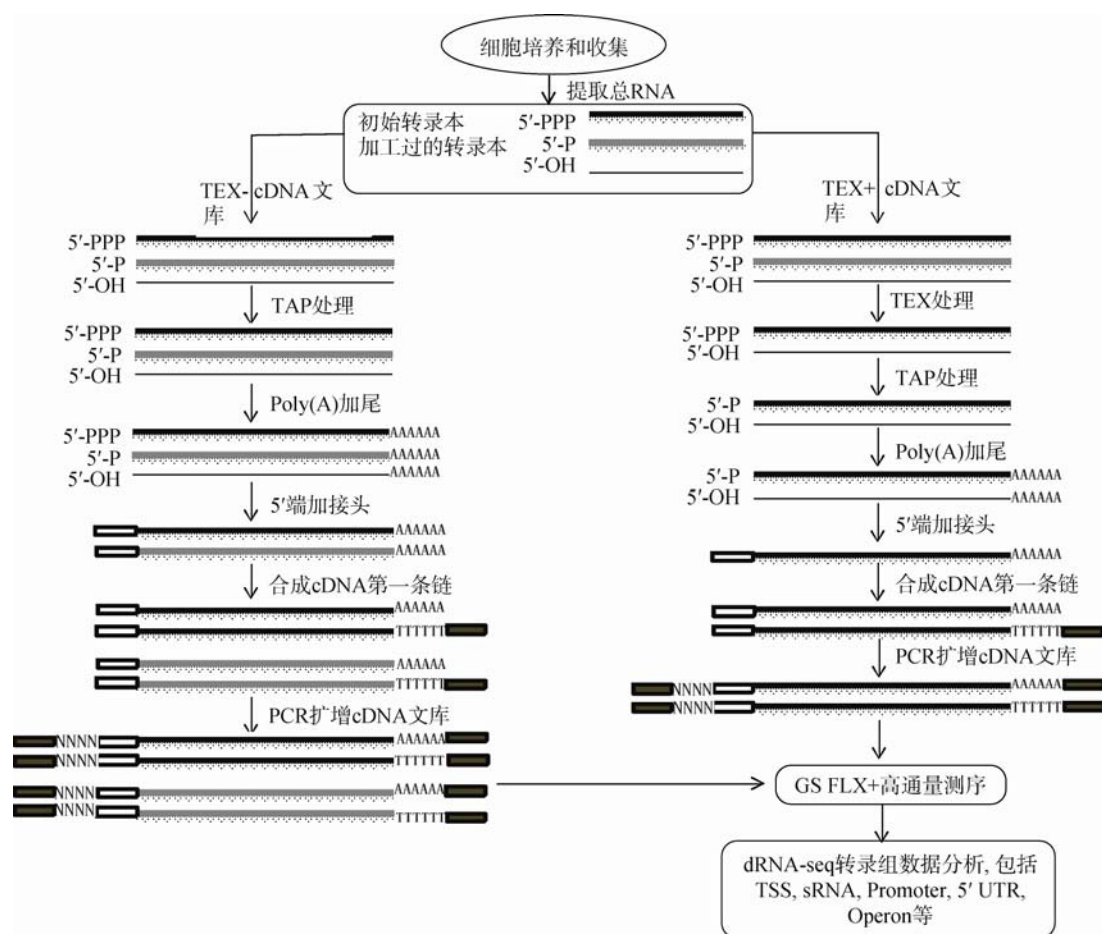


图 1 dRNA-seq 技术流程

TEX-文库。TAP 能够将初始转录本上的 5'-PPP 基团转化为 5'-P 基团, 为后续的 5'端加 RNA 接头做准备。再次有机抽提去除 TAP。

1.2.4 3'末端加 poly(A)尾, 5'末端加 RNA 接头处理

分别使用细菌 poly(A)聚合酶和 RNA 连接酶对 TEX+和 TEX-文库中的 RNA 样品进行 poly(A)加尾和 5'端加 RNA 接头处理, 为后续 cDNA 文库的构建做准备。由于 5'-OH-RNA 不能与 RNA 接头相连, 因此不会出现在最终的 cDNA 文库中。

1.2.5 合成 cDNA 第一条链

根据 Roche/454 Amplicon 测序平台设计合成带有 oligo(dT)的反向引物, 以 TEX+/TEX-文库中的 RNA 样品为模板, 使用 M-MLV 反转录酶(RNaseH)合成 cDNA 的第一条链。

1.2.6 PCR 扩增 TEX+和 TEX- cDNA 文库

使用基于 Roche/454 Amplicon 测序平台的含有

5'端接头序列的正向引物 PCR 扩增 TEX+和 TEX-cDNA 文库。为了区分不同培养条件或生长时期下的转录本, 扩增 cDNA 文库时在引物的 5'端分别加上不同的 4 个碱基的条码序列(Barcode)。

1.2.7 高通量测序和数据分析

由于 dRNA-seq 是为绘制全基因组水平的 TSS 和 sRNA 而设计的, 因此没有 RNA 片段化这一步, 在高通量测序平台的选择上更倾向于读长较长的平台, 如 454 的 GS FLX+(最大读长约 1 000 bp)。与其他 RNA-seq 相比, dRNA-seq 在转录组数据分析过程中也更倾向于在全基因组水平上研究 TSS、sRNA、启动子(Promoter)、5'非转录区域(5' untranslated region, 5' UTR)及操纵子(Operon)等。

2 dRNA-seq 在原核生物转录组学研究中的应用

dRNA-seq 目前主要用于在全基因组水平上绘

制 TSS、鉴定 sRNA 和反义转录本。对细胞内所有转录本转录起始位点和表达水平的研究不仅能增进人们对功能已知的转录本表达调控机制的理解,也可用于分析在某些环境条件下特异表达的转录本及其表达情况,从而促进对这些转录本功能的研究。sRNA 在细菌的转录调控中扮演了重要角色,然而 sRNA 通常转录自非编码区且与 mRNA 相比长度较短,因此很难被鉴定,在原核生物基因组上已经被注释的 sRNA 也非常有限。dRNA-seq 由于能够在全基因组水平鉴定 TSS,因此,极大便利了 sRNA 的研究。由于 dRNA-seq 基于反转录 cDNA 文库测序,相比基因芯片技术,需要更少的细胞作为起始材料,因此可以应用于一些难以进行遗传操作的物种转录组的研究。本文从比较转录组学研究、启动子和 5' UTR 分析、sRNA 的研究以及难以进行遗传操作物种的研究 4 个方面来列举 dRNA-seq 的应用。

2.1 dRNA-seq 在比较转录组学研究中的应用

通过对不同生长环境下或者不同生长状态下的原核生物进行全基因组水平的 TSS 分析,能够快速的全基因组水平鉴定差异表达转录本,使人们对原核生物在特定环境条件下的应激调控有更加深刻的认识。此外,通过对比某些基因突变株和野生型转录组水平的表达差异,也有助于全面了解该基因的功能及所在的调控网络,寻找受其调控的下游基因。

2009 年, Jager 等^[26]第一次使用 dRNA-seq 对在不同含氮生长条件下的甲烷八叠球菌 GO1 株 (*Methanosarcina mazei* sp. Stain GO1) 的转录组进行测序,在全基因组水平上绘制了 876 个 TSS,其中 586 个起始 mRNA 的转录,208 个 TSS 起始 ncRNA 的转录。在所有的初始转录本中,有 287 个转录本只在有铵盐的环境条件下表达,有 253 个转录本只在固氮的条件下表达。并且确认了许多个在氮胁迫环境下特异表达的 sRNA,在它们的启动子区域还鉴定了一个响应氮胁迫的保守序列,可能是某种转录因子的结合位点。

蓝藻又称蓝细菌(Cyanobacteria),是一类能够进行产氧光合作用的原核生物。蓝藻除了通过光合作用参与地球上碳源的固定外,还为地球上氮源的固定做出了重要贡献^[27,28]。鱼腥藻 PCC7120(*Anabaena* sp. PCC 7120)是研究原核生物细胞分化和固氮的模

式藻。在缺氮诱导条件下沿着丝体每间隔约 10 个营养细胞形成一个高度特化的细胞,称为异形胞。*hetR* 和 *ntcA* 是异形胞发育和图式形成的主调控基因。Mitschke 等^[29]对鱼腥藻 PCC7120 的转录组进行了测序,通过检测加氮和缺氮培养条件下野生型和 *hetR* 突变株的转录情况,在全基因组水平上绘制了 10 000 多个 TSS,并鉴定了 900 多个与固氮相关的 TSS,其中 158 个可能受 NtcA 直接调控,209 个与 HetR 表达水平相关。这一研究成果为鱼腥藻 PCC7120 异形胞发育相关基因调控关系的研究提供了极大便利,将进一步推动原核细胞分化和固氮调控机制的研究。

2.2 dRNA-seq 在启动子和 5' UTR 分析中的应用

dRNA-seq 可以以单核苷酸分辨率在全基因组水平上鉴定 TSS,能够在全基因组水平绘制启动子图谱,从而对基因的调控结构等有更深刻的认识。鼠伤寒沙门氏菌(*Salmonella Typhimurium*)是一种能够引起人和其它哺乳动物肠胃炎的革兰氏阴性菌,其基因组上有两个编码毒力基因的毒力岛,即沙门氏菌毒力岛 SPI1(*Salmonella* pathogenicity island 1, SPI1)和沙门氏菌毒力岛 SPI2,分别负责沙门氏菌在宿主细胞外和细胞内毒力基因的表达。这两个毒力岛上几乎所有环境依赖型的毒力基因的表达是由细菌信号素鸟苷四磷酸(ppGpp)介导,鸟苷四磷酸由 RelA 和 SpoT 两个酶合成,因此 *relA* 和 *spoT* 基因突变株不能产生 ppGpp。2012 年, Ramachandran 等^[30]通过 dRNA-seq 对鼠伤寒沙门氏菌的野生型 SL1344 和 RelA/SpoT 缺失突变株 SPI1 诱导阶段的转录组进行测序,鉴定了 3 306 个 TSS。根据这些 TSS,进行了全基因组水平的启动子分析,在 1 932 个启动子上鉴定了 σ^{70} 的保守结合位点(-10 区),发现其中 92%的启动子负责启动维生素和辅酶因子编码基因的转录,42%的启动子负责启动与运动性和趋药性相关的基因。他们还在 365 个启动子中鉴定了保守的-35 区域发现其中 41%的基因与细胞分化相关。此外,他们还预测了 169 个依赖于其他 σ 因子(σ^{24} 、 σ^{28} 、 σ^{32} 、 σ^{38} 和 σ^{54})的启动子。对操纵子的分析鉴定了 1 538 个单顺反子转录本和 625 个多顺反子转录结构。

幽门螺旋杆菌(*Helicobacter pylori*)是一种能生活在人胃粘膜上的革兰氏阴性致病菌,基因组只有 1.67 M,包含 1 576 个开放阅读框(Open reading

frame, ORF)。2010 年 Sharma 等^[23]使用 dRNA-seq 对 5 种环境条件或生长时期下的幽门螺旋杆菌 26 695 株的转录组进行测序,在全基因组水平上鉴定了 1 907 个 TSS(在这之前只有 55 个 TSS 被鉴定)。通过对这些 TSS 位置的分析并与传统 RNA-seq 相结合,将 87.5%的幽门螺旋杆菌的基因归到了 337 个操纵子里,第一次在全基因组水平上绘制了幽门螺旋杆菌的操纵子图谱。通过对 5' UTR 的分析,发现约 50%的 5' UTR 长度分布在 20~40 个碱基,并且发现 5' UTR 的长度与这些基因的功能相关,例如与核苷酸合成相关基因几乎都有约 30 nt 长的 5' UTR,而与细胞分化、致病性相关基因具有明显更长的 5' UTR。

2.3 dRNA-seq 在全基因组水平鉴定 sRNA 的应用

sRNA 是一类在进化上高度保守的非编码小分子单链 RNA,在基因表达调控中发挥至关重要的作用^[31]。由于没有典型的开放阅读框,sRNA 在基因组上少有注释,研究难度较大。dRNA-seq 能快速在全基因组水平鉴定 sRNA,预测可能参与调控的反义 RNA(Anti-sense RNA, asRNA)。枯草芽孢杆菌(*Bacillus subtilis*)是厚壁菌门的模式生物,属于革兰氏阳性菌。对该物种转录组学的探索将对革兰氏阳性菌转录后调控的研究具有重要意义。2010 年, Irnov 等^[32]使用 dRNA-seq 对生长在稳定期的枯草芽孢杆菌 168 株进行了转录组测序,在全基因组水平上鉴定了大约 600 个 TSS。他们重点分析了这些 TSS 中可能具有顺式(Trans-acting)和反式(Cis-acting)调控作用的 RNA,鉴定了 40 个带有长的 5' UTR 的 TSS(潜在的顺式调控 RNA),验证了部分通过基因芯片或其他方法发现的 sRNA,同时鉴定了 50 个新的编码 sRNA 的 TSS(潜在的反式调控 RNA)。进一步研究发现,16 个 sRNA 位于基因组上的原噬菌体区域,其中有 3 对 sRNA 可以两两反义相互作用,可能是类型 毒素-抗毒素系统(Toxin-antitoxin system)的新成员。发现了 29 个可能与靶 mRNA 结合的 asRNA,这些靶 mRNA 广泛参与了枯草芽孢杆菌孢子的形成、毒素的合成等生理活动中。部分 asRNA 能与靶 mRNA 完全配对,其他则通过头对头或者尾对尾的方式与靶 mRNA 部分配对。asRNA 与 mRNA 配对后,形成的双链 RNA 可能被双链 RNA 依赖的 RNase 降解,或者通过影响核糖体的结合影响靶

mRNA 的翻译,从而调控目标基因的表达。

Kroeger 等^[33]通过 dRNA-seq 对鼠伤寒沙门氏菌 SL1344 株和其亲本菌株 ST4/74 的转录组进行了测序,在全基因组水平上鉴定了 1 873 个 TSS,其中 13%的 TSS 起始反义转录。他们还进一步鉴定了在鼠伤寒沙门氏菌刚进入生长稳定期表达的 140 个 sRNA,其中 60 个为新鉴定的 sRNA。还发现了 9 个受环境调节的 sRNA,这些 sRNA 可以在细胞生长的不同状态特异性的表达。此外,通过鼠伤寒沙门氏菌与其他肠杆菌 sRNA 的同源性分析,发现几乎一半的 sRNA 特异的存在于鼠伤寒沙门氏菌中,仅有不到 20% 的 sRNA 是所有肠杆菌共有的。还有一些 sRNA 是沙门氏菌属特有的,很可能与沙门氏菌感染有关。

2011 年, Mitschke 等^[34]通过 dRNA-seq 对单细胞模式藻-集胞藻 PCC6803(*Synechocystis* sp. PCC 6803)的转录组进行了测序,在全基因组水平上绘制了 3 527 个 TSS,其中约 1/3 的 TSS(1 165 个)位于编码基因的上游,约 1/3 的 TSS(1 112 个)位于 866 个基因的反义链上。另有 821 个 TSS 位于基因编码框的内部,429 个 TSS 位于基因间隔区。通过对位于基因间隔区的 TSS 分析,预测了 314 个 ncRNA,并用 Tiling 芯片对部分 asRNA 和 ncRNA 的转录水平进行了验证。此外,他们还对全基因组水平的 5' UTR 进行了注释,并重新注释了 58 个蛋白编码基因。

2.4 dRNA-seq 在难以进行遗传操作的物种转录组研究中的应用

有些致病菌虽然基因组测序已经完成,但由于不能被离体培养,很难进行大规模基因功能研究。dRNA-seq 仅需要少量的起始样本,能够用于这些物种的转录组分析,丰富人们对这些物种遗传调控的认识。肺炎衣原体(*Chlamydia pneumoniae*)是人类呼吸道疾病的重要病原体,是专性细胞内寄生的革兰氏阴性菌,具有两相发育周期,即存在细胞外的具有感染性的原体(Elementary body, EB)和在宿主细胞内繁殖的无感染性的网状体(Reticulate body, RB, 亦称始体 initial body)。由于缺乏体外培养方法和遗传操作工具,肺炎衣原体的生物学研究主要通过基因组序列分析。2011 年, Albrecht 等^[35]通过 dRNA-seq 对肺炎衣原体 CWL-029 株的原体和网状体进行了

转录组测序,首次在全基因组水平上绘制了肺炎衣原体的 565 个 TSS。预测了 75 个 ncRNA 并对其中的 54 个进行了实验验证。通过对 799 个相邻基因共转录情况的分析,将其中 752 个基因归到了 246 个多顺反子操纵子中。结合半定量分析发现,共有 288 个基因在原体和网状体中差异表达,其中 83 个基因(56 个基因功能未知)和 8 个 sRNA 在原体中高表达,192 个基因(46 个基因功能未知)和 5 个 sRNA 在网状体中高表达。此外,他们还在 18 个蛋白编码基因的 5'端发现了串联启动子区域。对这些串联启动子分析发现,部分串联启动子在 EB 和 RB 中分别起始转录不同的 TSS,产生不同长度的 5' UTR。不同长度的 5' UTR 可能会形成不同的二级结构,进一步影响 mRNA 的稳定性和翻译活动,从而使肺炎衣原体在不同的发育阶段具有不同的基因调控模式。

3 分析与展望

相对于基于 rDNA 探针杂交去除 rRNA 的 RNA-seq, dRNA-seq 使用 5'单磷酸依赖的外切酶降解加工过的转录本去除 rRNA,这种方法不受物种亲缘关系远近的限制,具有更广的适用范围。另外, dRNA-seq 还实现了在全基因组水平上鉴定 TSS 和加工位点,不再需要使用传统的 RACE 技术对转录本逐个鉴定,这对于鉴定启动子区域和研究基因和 ncRNA 的表达调控具有重要意义。但也有很多问题需要特别注意,如为了保留初始转录信息,整个 dRNA-seq 文库构建过程中没有将 RNA 或 cDNA 片段化,且在 TEX-文库中并没有去除 rRNA,因此在提取样品总 RNA 和构建 cDNA 文库的过程中要尽量避免污染 RNA 酶和剧烈震荡,以免造成初始转录本的断裂,导致低丰度的 mRNA 和 ncRNA 难以检测,并产生大量的 rRNA 小片段增加测序难度。此外,5S rRNA 是 5'-P-RNA,但在 TEX+文库中通常并没有被降解掉,可能是其稳定的二级结构把 5'端的磷酸分子隐藏起来导致的^[25]。在 dRNA-seq 中由于 5'-OH-RNA 不能连 5' RNA 接头,这部分转录本并没有出现在 cDNA 文库中,如果要得到这部分数据,可以先将其磷酸化,再加 5' RNA 接头和 poly(A)尾,然后再进行 dRNA-seq。

在 dRNA-seq 中,通过在转录本 3'末端加 poly(A)尾巴和在 5'末端加 RNA 接头,保留了转录本的方向

信息,对鉴定 ncRNA 和反义转录本非常重要。sRNA 是在各种生物中广泛存在的调控分子,它们可能通过影响 mRNA 的转录、稳定性、翻译起始和终止及蛋白产物的活性在转录后水平上调控基因的功能^[35]。由于传统的基于生物信息学预测的方法偏好于发现一些在相近种属里高度保守的 sRNA,而基于 TSS 的 sRNA 的研究没有这种偏好性,可以发现大量不是很保守的、新进化出的及高度退化的 sRNA,这将极大地促进对 sRNA 介导的基因表达调控的研究。因此,寻找新的 sRNA 及其靶标 mRNA,以及鉴定介导 sRNA 与 mRNA 相互作用的蛋白等逐渐成为转录组研究的热点。

dRNA-seq 已应用于古细菌、真细菌、蓝细菌、衣原体等的转录组学研究中,纵观这些研究可以发现,asRNA 和 ncRNA 在原核生物转录组中占了很大的比例,可能在基因的转录后调控中发挥重要作用。原核生物的转录组织结构和调控机制远比以前认为的复杂的多,对 RNA 介导的调控作用的研究将进一步推进原核生物基因表达调控网络的研究。另外, dRNA-seq 在全基因组水平上对原核生物 TSS、启动子区域、5' UTR 和操纵子的注释极大地丰富了现有基因组的信息,将为功能基因组学的研究提供重要参考依据。

此外,测序技术的不断发展,也将促进 dRNA-seq 技术的完善和发展。例如近年出现的直接 RNA 测序技术以及单细胞测序技术都可能应用到 dRNA-seq 中^[36-40]。与其他 RNA-seq 一样, dRNA-seq 在现阶段并没有实现对 RNA 的直接测序,而是将 RNA 分子反转成 cDNA 后进行高通量测序,在反转过程中可能会出现错误的转录本。另外,有些转录本的复杂二级结构阻止了这些 RNA 反转录成 cDNA,因此基于 cDNA 的高通量测序不能检测到这部分转录信息。直接 RNA 测序技术将可能为这些问题提供解决方法,使得测序数据更为准确。自然界大约有 5×10^{30} 的微生物,而寄生在人体的微生物数是人体自身细胞总数的十倍以上(构成“人体微生物组”),与人类健康息息相关。但绝大多数微生物目前并不能在实验室培养生长,很难对它们进行遗传学分析^[41]。单细胞测序技术是一种以单细胞或者微量细胞为材料的测序技术。随着这种技术的日益成熟, dRNA-seq 可能应用于上述物种的转录组学研究,因此单细胞

测序技术被认为将是不可被培养物种转录组学研究的一个里程碑。此外, 由于现阶段高通量测序读长的限制, 导致 dRNA-seq 带有明显的 5'末端偏好性, 测序结果不能均匀地覆盖整个转录本, 随着测序技术的不断改善, 测序长度的不断增长, 未来测序技术将有可能覆盖全长的转录本, dRNA-seq 即可在全基因组水平上精确鉴定 3' UTR 和分析操纵子结构。

参考文献(References):

- [1] 李湘龙, 柏斌, 吴俊, 邓启云, 周波. 第二代测序技术用于水稻和稻瘟菌互作早期转录组的分析. *遗传*, 2012, 34(1): 102–112. [DOI](#)
- [2] 梁烨, 陈双燕, 刘公社. 新一代测序技术在植物转录组研究中的应用. *遗传*, 2011, 33(12): 1317–1326. [DOI](#)
- [3] Gregory BD, Yazaki J, Ecker JR. Utilizing tiling microarrays for whole-genome analysis in plants. *Plant J*, 2008, 53(4): 636–644. [DOI](#)
- [4] Yazaki J, Gregory BD, Ecker JR. Mapping the genome landscape using tiling array technology. *Curr Opin Plant Biol*, 2007, 10(5): 534–542. [DOI](#)
- [5] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 2008, 18(9): 1509–1517. [DOI](#)
- [6] 祁云霞, 刘永斌, 荣威恒. 转录组研究新技术: RNA-Seq 及其应用. *遗传*, 2011, 33(11): 1191–1202. [DOI](#)
- [7] Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*, 2012, 10(9): 618–630. [DOI](#)
- [8] Sorek R, Cossart P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, 2009, 11(1): 9–16. [DOI](#)
- [9] Yoder-Himes DR, Chain PSG, Zhu Y, Wurtzel O, Rubin E, Tiedje JM, Sorek R. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc Natl Acad Sci USA*, 2009, 106(10): 3976–3981. [DOI](#)
- [10] Stewart FJ, Ottesen EA, DeLong EF. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J*, 2010, 4(7): 896–907. [DOI](#)
- [11] Ottesen EA, Young CR, Eppley JM, Ryan JP, Chavez FP, Scholin CA, DeLong EF. Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Natl Acad Sci USA*, 2013, 110(6): E488–E497. [DOI](#)
- [12] Dodd D, Moon YH, Swaminathan K, Mackie RI, Cann IK. Transcriptomic analyses of xylan degradation by *Prevotella bryantii* and insights into energy acquisition by xylanolytic *Bacteroidetes*. *J Biol Chem*, 2010, 285(39): 30261–30273. [DOI](#)
- [13] Chen ZT, Duan XP. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol Biol*, 2011, 733: 93–103. [DOI](#)
- [14] Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, Livny J, Earl AM, Gevers D, Ward DV, Ward D V, Nusbaum C, Birren B W, Gnirke A. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol*, 2012, 13(3): r23. [DOI](#)
- [15] Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods*, 2009, 6(9): 647–649. [DOI](#)
- [16] Bogdanova EA, Shagin DA, Lukyanov SA. Normalization of full-length enriched cDNA. *Mol Biosyst*, 2008, 4(3): 205–212. [DOI](#)
- [17] Yi H, Cho YJ, Won S, Lee JE, Yu HJ, Kim S, Schroth GP, Luo S, Chun J. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res*, 2011, 39(20): e140–e140. [DOI](#)
- [18] Frias-Lopez J, Shi YM, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA*, 2008, 105(10): 3805–3810. [DOI](#)
- [19] Warnecke F, Hess M. A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J Biotechnol*, 2009, 142(1): 91–95. [DOI](#)
- [20] Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet*, 2008, 4(8): e1000163. [DOI](#)
- [21] Zhang AX, Wassarman KM, Rosenow C, Tjaden BC, Storz G, Gottesman S. Global analysis of small RNA and mRNA targets of Hfq. *Mol Microbiol*, 2003, 50(4): 1111–1124. [DOI](#)
- [22] He SM, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods*, 2010, 7(10): 807–812. [DOI](#)
- [23] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Reinhardt R, Stadler P F, Vogel J. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 2010, 464(7286): 250–255. [DOI](#)
- [24] Peano C, Pietrelli A, Consolandi C, Rossi E, Petiti L, Tagliabue L, De Bellis G, Landini P. An efficient rRNA

- removal method for RNA sequencing in GC-rich bacteria. *Microb Inform Exp*, 2013, 3(1): 1–11. [DOI](#)
- [25] Harbers M, Kahl G. Tag-based next generation sequencing. In: Borries A, Vogel J, Sharma CM, eds. *Differential RNA Sequencing (dRNA-Seq): Deep-Sequencing-Based Analysis of Primary Transcriptomes*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA., 2012: 109–121. [DOI](#)
- [26] Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. Deep sequencing analysis of the *Methanosarcina mazei* Gol transcriptome in response to nitrogen availability. *Proc Natl Acad Sci USA*, 2009, 106(51): 21878–21882. [DOI](#)
- [27] Pisciotta JM, Zou YJ, Baskakov IV. Light-dependent electrogenic activity of cyanobacteria. *PLoS ONE*, 2010, 5(5): e10821. [DOI](#)
- [28] Tamagnini P, Leitão E, Oliveira P, Ferreira D, Pinto F, Harris DJ, Heidorn T, Lindblad P. Cyanobacterial hydrogenases: diversity, regulation and applications. *FEMS Microbiol Rev*, 2007, 31(6): 692–720. [DOI](#)
- [29] Mitschke J, Vioque A, Haas F, Hess WR, Muro-Pastor AM. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. *Proc Natl Acad Sci USA*, 2011, 108(50): 20130–20135. [DOI](#)
- [30] Ramachandran VK, Shearer N, Jacob JJ, Sharma CM, Thompson A. The architecture and ppGpp-dependent expression of the primary transcriptome of *Salmonella* Typhimurium during invasion gene expression. *BMC Genomics*, 2012, 13: 25. [DOI](#)
- [31] 汤海明, 陈红, 张静, 任景怡, 许宁. 新一代测序技术应用于 microRNA 检测. *遗传*, 2012, 34(6): 784–792. [DOI](#)
- [32] Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res*, 2010, 38(19): 6637–6651. [DOI](#)
- [33] Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, Chao YJ, Sittka A, Hebrard M, Haendler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JCD. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci USA*, 2012, 109(20): E1277–E1286. [DOI](#)
- [34] Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J, Hess WR. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci USA*, 2011, 108(5): 2124–2129. [DOI](#)
- [35] Albrecht M, Sharma CM, Dittrich MT, Müller T, Reinhardt R, Vogel J, Rudel T. The transcriptional landscape of *Chlamydia pneumoniae*. *Genome Biol*, 2011, 12: R98. [DOI](#)
- [36] Oszolák F. Third-generation sequencing techniques and applications to drug discovery. *Expert Opin Drug Discov*, 2012, 7(3): 231–243. [DOI](#)
- [37] Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. *Genome Biol*, 2011, 12(2): 217. [DOI](#)
- [38] Oszolák F, Milos PM. Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev: RNA*, 2011, 2(4): 565–570. [DOI](#)
- [39] Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet*, 2010, 19(R2): R227–R240. [DOI](#)
- [40] Oszolák F, Platt AR, Jones DR, Reifemberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. *Nature*, 2009, 461(7265): 814–818. [DOI](#)
- [41] 潘星华, 朱海英, Marjani SL. 单细胞基因组学分析的技术前沿. *遗传*, 2011, 33(1): 17–24. [DOI](#)