DOI: 10.13376/j.cbls/2014040
文章编号：1004-0374(2014)03-0270-06

刘晓实验室从功能基因组学和系统生物学角度，主要以线虫神经系统为模型，研究细胞命运决定和进化的分子机制。目前重点在线虫建立单细胞精度的基因表达谱，构建以转录因子和 lncRNA 为核心的基因调控网络，并揭示它们在神经细胞命运决定中的作用，认识神经系统多样性的分子发育基础。为实现这一目标，我们开发和利用多种高通量实验方法，包括组织特异转录组技术和单细胞精度线虫影像分析系统。同时，我们开发基因组学技术，高效组装非模式动物的基因组。通过开发非模式动物的基因组和遗传资源，开展进化发育生物学研究。

刘晓 教授

高通量组织特异表达谱分析技术

刘　晓
(清华大学生命科学学院，北京 100084)

摘　要：测量基因表达谱是研究动物如何发育和应对刺激的重要途径。基因表达谱实验检测的 RNA 通常来自多种细胞的混合体。这样的数据具有较低的分辨率，不能区分动物体内不同细胞类型的转录组，而且会偏袒 RNA 含量占优势的组织以及上调的基因。这个问题可以通过获得特异组织的 RNA 进行表达谱分析来实现。这十几年来已有多种方法被开发出来获取特异的细胞类型或它们的 RNA，并与高通量 RNA 分析技术结合起来生成各种特异组织或细胞的转录组。本综述将介绍这些方法的基本原理和在大尺度表达谱分析中的应用，并讨论它们各自的优点和局限性。

关键词：组织特异；基因表达谱；功能基因组技术；细胞标记；RNA 标记
中图分类号：Q786　　**文献标志码：**A

Technologies of high-throughput tissue-specific gene expression profiling

LIU Xiao
(School of Life Science, Tsing-Hua University, Beijing 100084, China)

Abstract: Gene expression profiling is an important strategy to study animal development and response to stimuli. RNAs measured in gene expression profiling experiments are frequently purified from mixture of multiple cell types. The resultant data have low resolution, incapable of distinguishing transcriptome of different cell types and likely biased to up-regulated genes in dominant tissues. These problems can be solved by obtaining tissue-specific gene expression profile. For a dozen years, there have been several strategies developed to isolate specific tissues or purify RNAs from tissue of interest, and combined with high-throughput RNA assays to generate transcriptome of various specific tissues or cell types. This review will introduce basic principles of these methods and their application in large-scale transcriptome analysis, and discuss on their advantages and limitation.

Key words: tissue-specific; gene expression profile; functional genomic technique; cell labeling; RNA labeling

收稿日期：2013-10-08； 修回日期：2014-03-05
基金项目：国家重点基础研究发展计划(“973”项目)(2013CB945600)
*通信作者：E-mail: xiaoliu@tsinghua.edu.cn

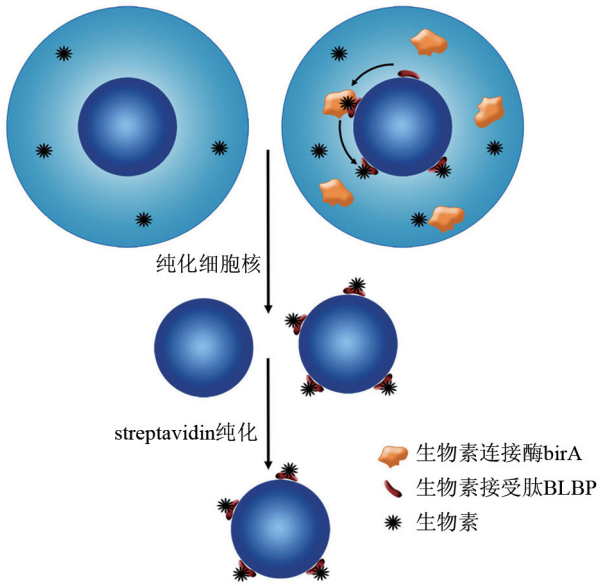
动物是由多种组织和细胞类型构成的有机体。这些细胞的分裂分化、相互作用和对环境的反应需要通过改变它们的基因表达调控程序改变转录组。因此获得各组织的转录组是认识动物如何发育和产生疾病所必需的。虽然原位杂交、免疫组织荧光和报告基因能够高分辨率地鉴定基因表达的组织特异性，这些实验花费高、工作量大，尤其是影像分析尚不能高度自动化，很难高通量获得准确量化的基因表达数据[1-3]。随着基因组学和功能基因组学的兴起，各种大尺度表达谱分析方法被开发出来，包括基因芯片 (microarray)、RNA 深度测序 (RNA-seq) 和染色质免疫共沉淀并深度测序 (ChIP-seq)。这些方法使得研究人员可以在全基因水平分析基因表达谱和基因表达的转录启始调控机制[4-6]。虽然这些方法可以鉴定出成百上千差异表达或可能被特定转录因子调控的基因，这些候选基因的生物学意义和调控机制经常很模糊。这其中的一个重要原因是这些实验的样品通常是动物的某一部分，甚至是整个个体，换句话说是多种类型细胞的混合体。因此，研究人员无法轻易区分这些基因表达变化发生在哪些细胞。如果某类细胞在样品中占的比例少，它们的基因表达变化则不易被检测到，尤其是那些表达量下调的基因。

这些年功能基因组学领域开发出一些技术能够有效地获得组织特异的转录组，包括细胞分选、以及共价和非共价 RNA 标记。单细胞测序在本专刊由汤富酬撰写，本文不详细讨论。除了介绍这些实验技术的基本原理和应用状况。本文还会讨论这些组织特异转录组技术各自的优点和局限性，以及它们的适用范围。

1　细胞分选

获取动物组织特异转录组最常用的方法是把样品组织块解离成单细胞，再利用组织特异表达的荧光蛋白转基因或者是细胞表面特异抗原的抗体通过荧光激活细胞分类术 (fluorescence-activated cell sorting, FACS) 收集目的细胞，最后对纯化的细胞进行表达谱分析。利用细胞分选策略的方法还包括人工分离[7]、激光显微切割[8]和单细胞深度 RNA 测序[9]。细胞分选法不但在体外培养的细胞系和哺乳动物样品中广泛使用，对小型模式动物也可以成为一种有效的手段。例如为了鉴定负责秀丽杆线虫机械感受神经元分化的基因，Martin Chalfie 实验室尝试对比野生型线虫和机械感受神经元缺陷型线虫的转录组[10]。然而一条成年秀丽杆线虫雌雄同体有 959 个体细胞和数千生殖细胞，而其机械感受神经元只有六对。所以分析整个虫体的 RNA 无法鉴定出在野生型和突变型间表达有显著差异的基因。为了特异地获得机械感受神经元的表达谱，Martin Chalfie 实验室生成了特异在机械感受神经元表达绿色荧光蛋白 (GFP) 的转基因线虫，分别把携带这个报告基因的野生型线虫品系和机械感受神经元缺陷型线虫品系的胚胎解离成单细胞。通过体外培养过夜，部分被培养的胚胎细胞分化成类似体内机械感受神经元的表达 GFP 的细胞。用 FACS 分选出表达 GFP 的细胞，对其 RNA 进行生物芯片分析，鉴定出 19 个非常显著地在突变型细胞里下调的基因，包括已知的 11 个机械感受神经元表达基因中的 7 个。对 12 个未知基因的进一步研究鉴定出负责维持机械感受神经元基因 mec-17[10]。

虽然细胞分选法非常有效，但其需要把样本组织块分解为大量单个细胞。而某些样品不能满足这一点，例如成体秀丽杆线虫。为此，Steven Henikoff 开发了特异细胞类型标记细胞核分离法 (isolation of nuclei tagged in specific cell types, INTACT)(图 1) 直接从动物体内通过亲和纯化收集特异组织的细胞核以分析转录组和染色质结构[11]。INTACT 利用了大肠杆菌的生物素连接酶 (BirA) 能将生物素连接到生



生物素连接酶birA和与生物素接受肽BLBP融合的细胞核表面蛋白表达在特异细胞里，使得目的细胞核表面特异地共价连接生物素。利用生物素亲和层析收集的细胞核可以进行高通量RNA和DNA分析。

图1　特异细胞类型标记细胞核分离法示意图

物素连接酶识别肽链 (Biotin ligase recognition peptide, BLRP) 这一生化特性，将 BLRP 融合在表达在核包被的蛋白上，如暴露在细胞质的核膜孔复合物蛋白。这个融合蛋白就是细胞核标记融合蛋白 (nuclear tagging fusion, NTF)。通过转基因，利用组织特异启动子把细胞核标记融合蛋白 NTF 和生物素连接酶 BirA 共表达在目的组织中，使其细胞核外表面表达 NTF 并且连接上生物素。提取这些转基因动物的完整细胞核，利用亲和层析把携带生物素标记 NTF 的细胞核纯化出来。这种组织特异的细胞核不但可以通过分析其 RNA 研究其转录组，还可以对其基因组进行表观遗传学分析 [11]。组织特异转录组和表观遗传组的结合为研究基因表达调控机制提供了独特的资源。虽然实验过程比较复杂，INTACT 已成功地在多种模式生物中使用，包括线虫、果蝇和拟南芥 [11-13]。

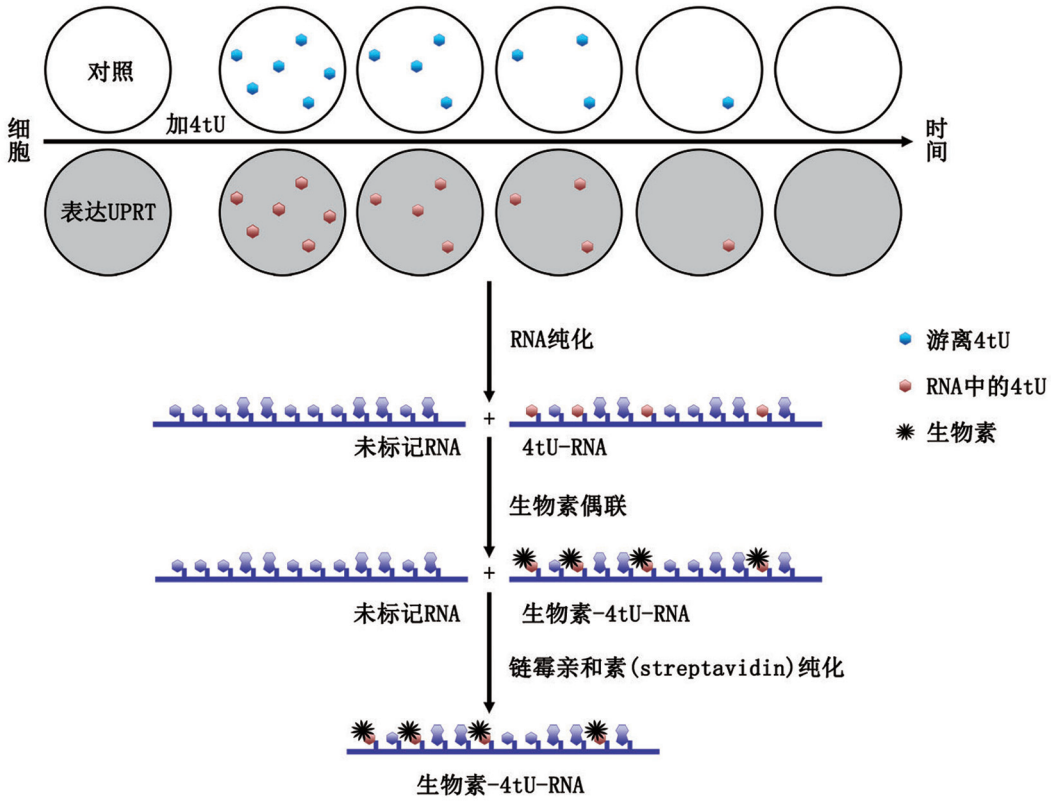
2 RNA免疫共沉淀(RNA immunoprecipitation, RIP)

RIP 是一种研究 RNA 蛋白质在体内相互作用的方法。通过交联剂处理把细胞内空间距离很近的大分子交联起来，再利用抗体把特异的蛋白质和与其交联的 RNA 通过亲和层析分离出来，从而鉴定出与特异蛋白结合的 RNA[14]。多聚 A 结合蛋白 (Poly A-binding protein, PAB) 可以结合所有 mRNA，把 PAB 表达在特异组织并作 RIP 分析就可鉴定目的细胞的转录组。这种 PAB-RIP 法是第一个通过标记 RNA 研究组织特异转录组的方法，最先用来研究秀丽杆线虫肌肉特异转录组 [15]。PAB-RIP 法非常灵敏，甚至可以研究单细胞精度的基因表达谱。秀丽杆线虫有两个神经元，ASEL 和 ASER 负责味觉感受。虽然这对细胞在解剖位置上左右对称，但是它们表达不同的化学受体，而且感受不同的小分子。ASEL 感受钠离子而 ASER 感受氯离子。为了研究 ASE 神经元左右不对称的分子机制，Lino 实验室构建了两个转基因线虫品系，分别在 ASEL 和 ASER 神经元表达有 FLAG 肽链标记的 PAB。利用 FLAG 抗原决定簇亲和层析，与 FLAG-PAB 交联的 mRNA 被纯化。生物芯片分析鉴定出 188 个左右不对称表达的基因，包括 13 个已知在 ASEL 和 ASER 差异表达的基因中的 8 个。进一步实验核实了 9 个新的左亚型和右亚型特异基因，包括倾向于在 ASER 表达的神经肽类基因 *nlp-5* 和 *nlp-7*。分别检测 *nlp-5* 和 *nlp-7* 单突变体没有发现显著表型，然而 *nlp-5* 和 *nlp-7* 双突变体线虫表现出一定的盐趋化性缺陷，提示这两个基因功能冗余。另外，对倾向于在 ASEL 表达基因的启动子序列分析，发现它们的表达都受细胞特异转录因子 CHE-1 调控 [16]。

PAB-RIP 法简单灵敏，在秀丽杆线虫功能基因组研究中广泛使用。比如模式生物 DNA 因子大百科全书计划 (model organism ENCyclopedia Of DNA Element, modENCODE) 用 PAB-RIP 测量了 15 个在不同发育时期的组织转录组 [17]。对海量的组织特异转录组的计算分析揭示了基因调控网络新性质。例如，75% 的基因被发现具有不同程度的时空表达特异性，包括大量持家基因。以这套数据为基础还鉴定出超过 200 个长链非编码 RNA (long non-coding RNA)，而且绝大部分都显示组织表达特异性 [18]。对 modENCODE 的组织特异转录组数据的自组织图谱分析 (self-organizing map) 鉴定出大量共表达基因。对它们启动子和 mRNA 的 3' 非翻译区的序列分析准确预测了 DNA 调控序列和 miroRNA 结合位点 [17]。总之，通过 PAB-RIP 产生的大尺度组织特异转录组数据结合计算分析构建了高精密度基因表达图谱，为构建基因的调控网络和鉴定其生物功能奠定了基础。

3 共价RNA标记

原生动物存在一些多细胞动物没有的核苷酸合成代谢酶，可以把核苷类似物加工成 RNA 可以识别的核苷酸类似物，从而整合到 RNA 当中。通过转基因把这些核苷酸合成代谢酶表达在真核生物细胞内并施加核苷类似物，这些核苷类似物就可以通过参与真核细胞的 RNA 合成代谢从而掺入其 RNA。这样，这些核苷类似物可以用作 RNA 标记以区别于不表达这些核苷酸合成代谢酶的细胞合成的 RNA。目前使用最广泛的是 *Toxoplasma gondii* 的尿嘧啶磷酸核苷转移酶 (uracil phosphoribosyltransferase, UPRT)。UPRT 能够把尿嘧啶类似物 4 硫尿嘧啶 (4-thiouracil, 4tU) 转化为 4tUMP，从而最终掺入到 RNA 中 [19]。哺乳动物和昆虫没有 UPRT 活性，因此 4tU 可以在这些动物中用作标记 (图 2)。因为不需要分离单个细胞，4tU 标记法首先使用在不易解离的果蝇组织上。比如果蝇的神经胶质细胞在脑中分散，而且具有复杂的形态，很难通过组织解离收获大量完整的细胞体。为了鉴定神经胶质细胞特异的基因，Chris Doe 实验室生成由神经胶质细胞特异启动子驱动的 UPRT 转基因果蝇 [20]。实验



在特异细胞表达UPRT，并且在特定时间加入4tU使得只有目的细胞在限定时间能将4tU掺入RNA。对纯化的RNA进行化学处理使4tU与生物素耦联，再利用生物素亲和层析收集掺入了4tU的RNA。

图2　4tU RNA脉冲标记法获得时空特异转录组

表明，在果蝇细胞里仅表达 UPRT 就足以改变其核苷酸合成代谢途径，使转基因细胞可以利用 4tU 作为合成 RNA 的原料。而且 UPRT 转基因的表达和 4tU 的掺入没有显著影响果蝇的生长发育和行为。给转基因幼虫喂食含有 4tU 的食物融合，然后纯化整个果蝇的 RNA，通过硫酯键使生物素与 4tU 共价结合使掺有 4tU 的 RNA 标记上生物素，最后通过对生物素亲和层析纯化的 biotin-4tU-RNA 进行生物芯片分析。由于 4tU 只掺入到 RNA 序列中 U 的位置，富含 U 的 RNA 序列就倾向于被富集。计算分析时通过对 U 的数目进行线性回归过滤掉了这种倾向性。最后，这个实验鉴定出大量幼虫神经胶质细胞富集的基因，其中已知的 4 个神经胶质细胞特异基因属于前 3.2% 富集的基因 [20]。

4tU 标记不但需要组织特异的 UPRT 表达，还依赖外源的 4tU 给药。这使研究者可以结合 UPRT 表达的组织特异性和 4tU 给药的时间特异性研究特异组织转录组的动态变化 (图 2)。即使对组织解离相对容易的哺乳动物，细胞分离法也很难获得大量真实反映体内应对某种刺激的特异细胞类型。比如，对注射了 LPS 的小鼠脾脏进行转录组分析可以鉴定出大量 LPS 反应基因，但不知道哪些基因代表哪些细胞的反应。为了分析小鼠脾内皮细胞对 LPS 的反应，Chris Doe 实验室构建了特异在脾内皮细胞表达 UPRT 的转基因小鼠 [21]。给小鼠施加 4tU 后 1 h 注射 LPS，3 h 后收集脾 RNA，生物素纯化 4tU 标记的 RNA 后进行 RNA-seq 分析，鉴定出 97 个脾内皮细胞的 LPS- 诱导基因，包涵 24 个以前全脾转录组分析鉴定出的被 LPS 诱导的基因。对显著被 LPS 诱导的脾内皮细胞进行 GO 分析，发现这些基因高度富集免疫反应相关基因，包括“防卫反应”、“先天免疫反应”及“细胞因子 (cytokine) 激活反应”[21]。

4　讨论

动物体是由多种细胞构成的混合体。因此对动物整体或大组织块样品的 RNA 分析会失去组织特异的信息。为此，科研人员开发了各种检测组织特异转录组的功能基因组学方法。对于可以在体外培养的细胞和易解离的哺乳动物组织样品，通过

FACS 分选特异细胞类型是最广泛使用的手段之一。单细胞 RNA 深度测序目前在哺乳动物细胞的研究发展很快。但由于线虫和果蝇细胞的 RNA 丰度低，目前还不能实现有效的单细胞测序。而显微切割和特异细胞类型标记细胞核分离法可以用于不易解离成单细胞的组织。但这种细胞分离策略都需要长时间处理组织，改变其微环境，很可能改变了细胞的基因表达程序。还有，细胞分离法容易损伤细胞的特化结构，如神经的轴突和树突，从而丢失位于这些结构的 RNA。而共价或非共价标记纯化特异组织的 RNA 则可以很好地解决这些问题。RNA 标记更大的优势在于其不需要解离组织块，所以能在线虫果蝇这些小型模式动物中广泛使用，与它们高度发达的遗传学研究结合起来构建基因调控网络，探索其生物功能。由于其操作简单、检测灵敏，PAB-RIP 是线虫功能基因组研究的主要手段之一。但是 RIP 实验需要交联这一步骤，其数据不可避免地具有较高的噪音。而且 PAB-RIP 只能研究具有 PolyA 的 RNA 分子，不能检测大部分非编码 RNA。以 4tU 为代表的代谢物共价标记法则没有这些限制。而且由于代谢物共价标记需要施药这一步骤，使研究者能够检测特异组织应对特异刺激的转录组反应以及 RNA 的生成和降解等动态过程。不过，4tU 共价标记法需要研究对象能够摄入核苷类似物，所以不能应用到所有模式动物上。比如目前秀丽杆线虫没有成功的组织特异 4tU 共价标记的报道。最近，我们实验室开发了一种利用反式拼接标记 mRNA 获得线虫组织特异转录组的方法。线虫 70% 编码基因的 pre-mRNA 会与拼接导链 (splicing leader, SL) RNA 发生反式拼接，使 mRNA 的 5’ 端获得 SL-RNA 的 22 个碱基。我们在 SL-RNA 基因的 5’ 端加一个标记序列，用组织特异启动子驱动被标记的 SL-RNA 基因的表达，使得目标组织的 mRNA 拼接上标记序列。对携带标记序列的 mRNA 高通量测序，就获得了目标组织的转录组 (未发表)。总之，这些功能基因组方法各有千秋，需针对研究对象的特点和生物学问题选取合适的组织特异转录组检测方法。

[参 考 文 献]

[1] Sunkin SM, Ng L, Lau C, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res, 2013, 41(Database issue): D996-D1008

[2] Murray JI, Boyle TJ, Preston E, et al. Multidimensional regulation of gene expression in the *C. elegans* embryo. Genome Res, 2012, 22(7): 1282-94

[3] Liu X, Long F, Peng H, et al. Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. Cell, 2009, 139(3): 623-33

[4] mod EC, Roy S, Ernst J, et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science, 2010, 330(6012): 1787-97

[5] Gerstein MB, Lu ZJ, Van Nostrand EL, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science, 2010, 330(6012): 1775-87

[6] Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) project. Science, 2004, 306(5696): 636-40

[7] Factor DC, Najm FJ, Tesar PJ. Generation and characterization of epiblast stem cells from blastocyst-stage mouse embryos. Methods Mol Biol, 2013, 1074: 1-13

[8] Sanna PP, Repunte-Canonigo V, Guidotti A. Gene profiling of laser-microdissected brain regions and individual cells in drug abuse and schizophrenia research. Methods Mol Biol, 2012, 829: 541-50

[9] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet, 2013, 14(9): 618-30

[10] Zhang Y, Ma C, Delohery T, et al. Identification of genes expressed in *C. elegans* touch receptor neurons. Nature, 2002, 418(6895): 331-5

[11] Steiner FA, Talbert PB, Kasinathan S, et al. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. Genome Res, 2012, 22(4): 766-77

[12] Deal RB, Henikoff S. The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. Nat Protoc, 2011, 6(1): 56-68

[13] Deal RB, Henikoff S. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev Cell, 2010, 18(6): 1030-40

[14] Gilbert C, Svejstrup JQ. RNA immunoprecipitation for determining RNA-protein associations *in vivo*. Curr Protoc Mol Biol, 2006, Chapter 27: Unit 27 4

[15] Roy PJ, Stuart JM, Lund J, et al. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. Nature, 2002, 418(6901): 975-9

[16] Takayama J, Faumont S, Kunitomo H, et al. Single-cell transcriptional analysis of taste sensory neuron pair in *Caenorhabditis elegans*. Nucleic Acids Res, 2010, 38(1): 131-42

[17] Spencer WC, Zeller G, Watson JD, et al. A spatial and temporal map of *C. elegans* gene expression. Genome Res, 2011, 21(2): 325-41

[18] Nam JW, Bartel DP. Long noncoding RNAs in *C. elegans*. Genome Res, 2012, 22(12): 2529-40

[19] Cleary MD, Meiering CD, Jan E, et al. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. Nat Biotechnol, 2005, 23(2): 232-7

[20] Miller MR, Robinson KJ, Cleary MD, et al. TU-tagging:

cell type-specific RNA isolation from intact complex tissues. Nat Methods, 2009, 6(6): 439-41
[21] Gay L, Miller MR, Ventura PB, et al. Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. Genes Dev, 2013, 27(1): 98-115

刘晓报告讨论

付向东 (加州大学圣地亚哥分校)

Q：建立秀丽隐杆线虫的组织特异性 mRNA 标记时，需要将启动子扩增多少碱基的长度才能保证组织特异性？

A：2~2.5 kb。因为线虫的基因组有 1 亿个碱基，2 万个基因，那么平均到每个基因就是 5 kb，除去内含子和其他非编码序列，那么大概是 2~2.5 kb 的大小。而从我们实验的结果来看，对于那些表达比较复杂的基因，如转录因子，一般需要 5~7 kb 才能比较完整的鉴定出来。

Q：那么对于不同的基因，你们就需要去尝试扩增不同长度的启动子，以保证能检测出组织特异性？

A：我们扩增的时候有两种解决的方案，一般我们都尽量扩增出长些的片段，当然受限于 PCR 扩增的效率，我们能保证的是在 5 kb 内都有良好的扩增效果，对于超过 5 kb 长度的序列，扩增效果就不是很好。另一个就是根据已有的测序资料，找到那些比较保守的区域。

Q：那么当你扩增的片段过长的时候也有一个问题，因为启动子自身的分布比较分散，就可能两个启动子相距比较近的时候，你都把这个两个基因扩增进去，那么你测序的结果就不能反映这两个基因的组织特异性？

A：我的这个系统中，在报告基因的上游只有一个启动子，那么我们就能保证绝大多数的基因都是组织特异性的表达，当然有可能存在就是由于扩增太长而添入其他启动子，但这毕竟是极少数的，我们只要能保证我们获得的绝大多数结果都是组织特异性的就可以了。

沈晓骅 (清华大学生命科学学院)

Q：其实现在已经有了单细胞的测序方法。而你使用的 SL-RNA 方法现需要去鉴定启动子，然后又将其导入线虫的不同阶段的 1 000 多个细胞，再去鉴定它的基因表达，操作就很复杂。那单细胞的测序方法不是更简单方便？

A：对于单细胞测序，我也想过这个问题，但是线虫有个特点，细胞只有哺乳细胞的 1/9，那么同样的细胞就需要有 9 个，那么就需要去从 9 个线虫中，取 9 个相同的细胞，这本身就是一件很有难度的事，最重要的是还需要保证这 9 个细胞的同步性，这样 9 个细胞内的转录组才会一致，否则就是将处在不同发育状态的细胞合并测序，获得的结果也就不能很好地反映组织特异性。