论文

利用转录组数据鉴定和分析小鼠非编码 RNA

赵宇慧^{①②}, 刘万飞^{①③}, 曾潞瑶^{①②}, 刘守成^{①②}, 谈馨煜^①, HASAN AwadAljohi^③,
胡松年^{①*}

① 中国科学院北京基因组研究所, 中国科学院基因组科学与信息重点实验室, 北京 100101;

② 中国科学院大学, 北京 100049;

③ Joint Center for Genomics Research (JCGR), King Abdulaziz City for Science and Technology and Chinese Academy of Sciences, Riyadh 11442, Saudi Arabia

* 联系人, E-mail: husn@big.ac.cn

收稿日期: 2015-06-23; 接受日期: 2015-07-22

国家自然科学基金(批准号: 31271385)和中国科学院知识创新工程(批准号: KSCX2-EW-R-01-04)资助项目

摘要 基因转录表达是一个复杂、精确并具有时空特异性的过程。目前对转录组的研究主要集中在蛋白编码基因上。近几年, 一个新的转录组研究工具——大规模并行 cDNA 测序技术(RNA-seq)为更深入地研究转录组带来了希望。利用 RNA-seq 数据, 鉴定出大量的非编码 RNA, 特别是 lincRNA, 并且发现这些非编码 RNA 是多个生物学过程中重要的调控因子。利用深度测序获得的 15 个小鼠组织 RNA-seq 数据探索非编码 RNA 在小鼠不同组织中的多样性和动态变化。依据自定的标准, 在这 15 个组织中共鉴定出 16249 个非编码基因(对应 21569 个非编码 RNA)。研究这些非编码 RNA 的多种特征, 可以发现与蛋白编码基因相比, 非编码 RNA 通常比较短, 外显子个数少, 表达量低, 组织特异性强。而且, 这些非编码 RNA 有明显的转录起始和转录延伸信号(H3K4me3, H3K27me3, H3K36me3 修饰, RNAP II 结合位点以及 CAGE)的富集。基因集富集分析结果表明, lincRNA 与多个生物学过程相关, 如免疫反应、肌肉发育和有性生殖等。本研究提供了更加全面的对小鼠非编码 RNA 的注释信息, 为小鼠非编码 RNA 的功能和进化研究奠定了基础。

关键词
非编码 RNA
RNA-seq
转录组
lincRNA
小鼠

早期的研究已经广泛证明哺乳动物的基因组是普遍转录的^[1,2]。例如, 人的基因组中有超过 80% 的序列是转录的, 而在小鼠基因组中也有大约 60% 的序列是转录的^[3-6]。最近的一项研究通过分析小鼠 123 个不同细胞类型和组织的 1000 多个转录组数据集, 进一步证实了小鼠内含子和基因间区的转录活性^[7]。哺乳动物基因组不仅可以转录成 mRNA, 也可

以产生很多非编码 RNA^[6,8-10]。近年来, 数以千计的非编码 RNA 被鉴定出来, 而且这些非编码 RNA 在许多生物过程中起重要作用, 如基因印记、X 染色体失活、细胞周期和发育过程, 特别是多能性调控^[11-19]等。

基于第二代测序技术的大规模并行 RNA 测序技术(RNA-seq)的发展为人们提供了一个前所未有的方法来无偏差地鉴定非编码 RNA, 特别是在哺乳动物

引用格式: 赵宇慧, 刘万飞, 曾潞瑶, 等. 利用转录组数据鉴定和分析小鼠非编码 RNA. 中国科学: 生命科学, 2015, 45: 845-859

英文版见: Zhao Y H, Liu W F, Zeng J Y, et al. Identification and analysis of mouse non-coding RNA using transcriptome data. Sci China Life Sci, 2015, 58, in press, doi: 10.1007/s11427-015-4929-x

中组织特异性表达的低丰度非编码 RNA^[20~22]. 大幅度地提高了测序通量的 RNA-seq 技术促进了后续转录本丰度的测量. 与早期的基因表达系列分析(serial analysis of gene expression, SAGE)和表达序列标签(expressed sequence tag, EST)等测序方法相比, RNA-seq 具有捕获几乎全部转录本的能力. 另外, 与多聚腺苷酸 RNA-seq(mRNA-seq)方法相比, 去核糖体 RNA 的 RNA-seq(rmRNA-seq)被认为具有更高的准确性与全面性^[23].

随着 RNA-seq 数据的快速增长, 非编码 RNA 特别是 lincRNA 也被快速鉴定出来. 目前, 在小鼠(*Mus musculus*)^[9,24~27]、人(*Homo sapiens*)^[25,28,29]、黑猩猩(*Pan troglodytes*)^[30]、斑马鱼(*Danio rerio*)^[13,31]、爪蟾(*Xenopus laevis*)^[32]、线虫(*Caenorhabditis elegans*)^[33]和拟南芥(*Arabidopsis thaliana*)^[34]中已发现了几千个 lincRNA. 一些新发现的 lincRNA 的功能已经得到了很好的验证. Linc-HOXA1 通过招募富含嘌呤元件结合蛋白 B(purine-rich element binding protein B, PURB), 作为转录辅助因子来抑制同源基因 A1(homeobox A1, *Hoxa1*)的表达^[35]. 一类叫做 ncRNA-activating(ncRNA-a)的 lincRNA 有类似增强子的功能, 它可以促进邻近蛋白编码基因的表达^[36]. 同源基因转录本的反义 RNA(HOX transcript antisense RNA, HOTAIR)通过招募表观遗传复合体多梳抑制复合体 2(polycomb repressive complex 2, PRC2)使染色质变为抑制状态, 从而反式调控 40 kb 的同源基因群 D(*HOXD*)位点基因的表达, 使之沉默^[37,38]. 总之, lincRNA 通过多种作用机制来发挥各种功能, 包括表观修饰和基因表达调控, 以及作为蛋白信号复合物的“支架”^[39~42]等.

最近研究报道, 大多数剪接事件是高度组织特异的^[43], 并且 lincRNA 在特定组织有严格的表达模式^[44]. 为了阐释小鼠非编码 RNA 的多样性和动态变化, 本课题组开发了一套流程来鉴定小鼠非编码 RNA. 通过覆盖了主要组织类型的 15 个小鼠组织数据(14 个公共数据库中来自 ENCODE 的 mRNA-seq 数据和 1 个自产的 rmRNA-seq 数据), 鉴定出了 16249 个非编码基因. 与蛋白编码基因相比, 这些新鉴定的非编码 RNA 具有表达量低、组织特异性强和保守性低的特征. 本研究不但扩展了小鼠非编码 RNA 数据集, 而且为小鼠非编码 RNA 的功能研究提供了数据来源. 期望

通过此工作为非编码 RNA 研究提供一个新的视角.

1 材料方法

1.1 数据集

从美国国家生物技术信息中心(National Center of Biotechnology Information, NCBI)下载了 ENCODE 测序产生的 14 个小鼠组织链特异的 mRNA-seq 数据(GSM900188, GSM900198, GSM900199, GSM900194, GSM900189, GSM900195, GSM900196, GSM900184, GSM900183, GSM900186, GSM900197, GSM900185, GSM900193 和 GSM900192)^[45,46], 平均每个组织有 147 M 双端测序片段(表 1), 读段长度为 76 bp. 此外, 还下载了小鼠心脏的 RNA 聚合酶 II(RNAP II), H3K4me3, H3K27me3 和 H3K36me3 的 ChIP-seq 数据^[47,48]. 另外, 从 NCBI 下载了一个大脑的 mRNA-seq 数据(SRX191149)来验证分析得到的部分结果^[49]. 对这个 mRNA-seq 数据的处理方法与前面 15 个组织的数据处理方法相同.

1.2 RNA 测序

取 8 周大的成年雄性 C57BL/6J 小鼠的大脑组织, 用 Trizol 法(Invitrogen, 美国)提取总 RNA, 然后用 Ribo-minus Eukaryote kit(cat.10837-08, Invitrogen, 美国)试剂盒去除核糖体 RNA. 实验遵照实验动物管理条例(中国科技部, 2004 年)进行, 并且获得中国科学院北京基因组研究所伦理委员会许可. 利用全转录组提取试剂盒(RNA-Seq Library Preparation Kit-Illumina Compatible Kit, 吴江汇杰, 苏州)处理 500 ng 去除核糖体的 RNA 来构建转录组文库. cDNA 文库经过扩增, AMPure XP beads 清洗后, 在 Illumina HiSeq 2000(Illumina, 美国)的 GA-analyzer 上进行测序. 最终, 得到了约 211 M 长 101 bp 的链特异性双端测序片段(SRX806806).

1.3 序列比对与转录本构建

去掉比对到核糖体 RNA 上的片段. 这一步是通过把测序片段直接比对到小鼠核糖体 RNA 序列上来完成的. 随后, 用自己编写的 Perl 脚本去掉了低质量的片段. 整合 UCSC(2013.5.6), NCBI(Refseq v58)和 Ensembl(v72)数据库中存储的所有小鼠转录本数据, 获得一个由所有已知转录本构成的完整的注释基因

表 1 RNA-seq 数据概况^{a)}

组织	序列片段归档 (sequence read archive, SRA)号	原始数据	高质量数据	百分比	比对上数据	比对上百分比
肾上腺	SRX135155	148292125	139090674	93.80%	127574606	91.72%
结肠	SRX135165	131005753	127438807	97.28%	111576426	93.50%
心脏	SRX135166	155581190	149271569	95.94%	119156550	94.85%
肾脏	SRX135161	211079100	204777760	97.01%	190296905	92.93%
大肠	SRX135156	148616147	145407725	97.84%	134480910	92.49%
肝脏	SRX135162	162688171	158919893	97.68%	150081337	94.44%
肺	SRX135163	133066159	128994546	96.94%	118492917	91.86%
乳腺	SRX135151	147002618	140730296	95.73%	128854541	91.56%
卵巢	SRX135150	105677057	101955077	96.48%	98032852	96.16%
小肠	SRX135153	138056928	134988269	97.78%	123869804	91.76%
脾脏	SRX135164	152594296	148439097	97.28%	79641990	93.39%
胃	SRX135152	168550565	158880743	94.28%	140181599	88.23%
睾丸	SRX135160	150763638	147359308	97.74%	138020994	93.66%
胸腺	SRX135159	117387927	114297797	97.37%	103907569	90.91%
大脑*	SRX806806	211809646	107740712	50.87%	65484766	60.78%

a)*: 此数据为本实验室自己测序获得

集. 接着, 用 GSNAP (2013.7.16 版本)把 15 个小鼠组织(14 个公共数据库下载的组织数据和一个新测的组织数据)的 RNA-seq 数据比对到小鼠基因组(mm10)上, 所用参数是“-N 1, -force-xs-dir 和-s 小鼠所有转录本的剪接位点”(其他参数均为默认值). 最后, 用 Cufflinks v2.1.1 把每个组织比对上的片段拼接成转录本.

1.4 非编码 RNA 鉴定流程

为了准确完整地鉴定转录本, 用 Cuffmerge 把 15 个组织的所有转录本整合成一个非冗余的唯一的转录本集合. 然后, 把与 RefSeq, UCSC 和 Ensembl 完全重叠的位点过滤掉. 为了精确地区分编码和非编码位点, 用 CPC 和 CPAT-1.2.1 来预测每个位点的编码可能性. 只有当两个软件对某个位点的预测结果一致时, 该位点才会被归类为编码/非编码位点. 接着, 过滤掉每百万比对上的读段中比对到 1 kb 外显子上的读段数(reads per kilobase of exon model per million mapped reads, RPKM)<0.1 的位点. 最后, 为了鉴定出可信的单外显子非编码基因, 进行了如下分析: (i) 通过 BLASTN 把这些单外显子的非编码基因比对到小鼠, 人和其他物种的 EST 数据上; (ii) 用这些非编码基因扫描 Rfam 协方差模型. 最终只保留了至少有一种类型数据支持的基因. 要求单外显子的非编码基因至少在 2 个组织表达. 根据这些非编码基因

在基因组上的位置, 把它们分成了重叠(overlapping)、反义(antisense)、内含子区(intronic)和基因间区(intergenic)的非编码基因.

1.5 组蛋白甲基化和 RNAP II 结合位点分析

下载 H3K4me3, H3K27me3, H3K36me3 和 RNAP II 的 ChIP-seq 数据比对结果和峰值文件. 将启动子界定为转录起始位点(transcription start site, TSS)上下游 5000 bp 范围内. 根据 H3K4me3, H3K27me3 和 RNAP II 数据的峰是否与非编码基因的启动子区域重叠, 获得了非编码基因的染色质状态. 对于 H3K36me3 修饰来说, 取决于 H3K36me3 数据的峰是否与非编码基因的转录区域重叠.

1.6 TSS 和 CAGE 证据

Fantom5 项目公布的小鼠 129466 个 TSS 样 CAGE 峰值位点用来为鉴定的非编码基因提供 TSS 支持^[50](<http://fantom.gsc.riken.jp/5/tet/>). 用 UCSC 的 LiftOver 工具把非编码基因在 mm10 基因组上的位置转化成 mm9 基因组上的位置. 定义非编码基因的 TSS 区间是 TSS 上下游 500 bp. 当 TSS 样 CAGE 峰值位点与非编码基因的 TSS 区间有重叠时, 就认为该非编码基因有 CAGE 数据支持.

从 DDBJ 数据库下载了小鼠睾丸的 5'CAGE 标签比对结果 (ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/)

DRA000/DRA000991/DRZ001400/), 并计算了每个核苷酸位点的 CAGE 覆盖度. 然后, 计算所有非编码基因 TSS 上下游 1 kb 范围内每个位点的累积 CAGE 标签覆盖度.

1.7 保守性分析

为了评估非编码基因的保守性, 从 UCSC 下载了小鼠基因组(mm10) PhastConsElements60way 数据^[51]. PhastConsElements60way 数据是利用 phyloP (phylogenetic *P*-values)和 PhastCons(<http://compugen.bscb.cornell.edu/phast/>)算法对 59 个脊椎动物基因组与小鼠基因组进行多序列比对而获得的基因组保守性概况. 每个区间的保守性计算方式如下: 从区间的起点开始, 取 12 bp 的窗口, 计算窗口的保守性分数, 然后每次 1 bp 移动窗口, 依次计算各窗口的保守性分数, 选保守性分数最大的窗口对应的值作为该区间的保守性值. 保守性值越高, 区间保守性越强.

1.8 基因集富集分析

采用与以前研究^[9,13]相同的方法做基因集富集分析(gene set enrichment analysis, GSEA). 根据 lincRNA 和蛋白编码基因在 15 个组织的表达量, 计算每个 lincRNA 和所有蛋白编码基因的皮尔森相关系数. 然后, 按照相关性从大到小把蛋白编码基因排序, 每个 lincRNA 都有一个按相关性排序的蛋白编码基因列表, 该列表就是 GSEA^[52]的输入文件. 最后, GSEA 输出一个 lincRNA 和 GO 分类的相关性矩阵($P < 0.01$), 接下来用这个矩阵进行了聚类. 为了分析所有正相关的 GO 分类在每个集群(cluster)的富集情况, 用二项分布检验对所有正相关的 GO 分类进行了排序.

1.9 lincRNA 和它们邻近蛋白编码基因的共表达谱

根据 lincRNA 和它们邻近的蛋白编码基因的位置, 获得了距离小于 10 kb 的相邻 lincRNA-蛋白编码基因对. 用类似的方法, 也得到了相邻的蛋白编码基因-蛋白编码基因对. 通过随机选择 2 个基因获得了随机基因对. 然后, 用 R 来计算它们之间表达量的皮尔森相关系数并分析相关性值的分布. 另外, 用 GOstat(<http://gostat.wehi.edu.au/cgi-bin/goStat.pl>)分析了 lincRNA 邻近蛋白编码基因的功能富集情况.

1.10 RT-PCR 验证

根据厂商提供的技术流程, 用 Trizol 法(Invitrogen)提取了小鼠组织的 RNA, 然后用 Dnase I (1 U/ μ L; Invitrogen)除去可能的基因组 DNA 污染. 接着用 Superscript III(Invitrogen)把 DNA-free 的 RNA 逆转录成 cDNA, 然后 cDNA 被稀释 20 倍用来做 PCR. PCR 是用 10 μ L 体系在 Applied Biosystems 9700 PCR System(Life Technologies, 美国)完成的, 10 μ L 体系具体组成如下: 10 $\times$ PCR buffer 1 μ L, dNTPmix(2.5 mmol/L) 1.5 μ L, 引物(每个 10 μ mol/L) 0.5 μ L, Taq DNA polymerase(5 U/ μ L TaKaRa, 大连) 0.1 μ L, 20-fold diluting cDNA 1 μ L, nuclease-free water 5.9 μ L. PCR 条件是 94 $^{\circ}$ C 3 min; 接着 35 个循环, 每个循环 98 $^{\circ}$ C 10s, 60 $^{\circ}$ C 30s, 72 $^{\circ}$ C 30s; 然后 72 $^{\circ}$ C 5 min. 最后, 用 2.0%琼脂糖凝胶回收 PCR 产物. 扩增 lincRNA 的引物是用在线工具 GenScript (<https://www.genscript.com/ssl-bin/app/primer>)设计的, 并且用小鼠的 RefSeq mRNA 检查了引物的特异性.

2 结果

2.1 从 15 个组织中鉴定新的小鼠非编码 RNA

为了全面探索小鼠非编码 RNA, 本实验室开发了一套鉴定非编码 RNA 的流程(图 1). 用来鉴定的 15 个组织包括从 NCBI 下载的 14 个小鼠组织(包括肾上腺、结肠、心脏、肾脏、大肠、肝脏、肺、乳腺、卵巢、小肠、脾脏、胃、睾丸和胸腺)的 RNA-seq 数据, 以及自产的利用 Illumina HiSeq 2000 对小鼠大脑进行测序获得的 RNA-seq 数据. 两方面的数据合计共得到了 22.8 亿个链特异的双端测序片段(178.74 Gb, 65.58 倍基因组覆盖度). 在对原始数据进行预处理后, 用 GSNAP 将所有高质量的数据比对到小鼠基因组上(mm10)获得了 18.6 亿个比对上的片段(142.97 Gb, 52.46 倍基因组覆盖度)(表 1). 之后, 利用 Cufflinks 软件将每个组织比对上的片段拼接成转录本, 用 Cuffmerge 把 15 个组织的转录本进行整合, 最终得到 75749 个位点. 去掉与已知的 RefSeq, UCSC 和 Ensembl 完全重合的位点后, 共得到了 44420 个新位点, 这一步过滤掉的位点为所有已知基因的 79.7%. 随后, 根据 CPC 和 CPAT 软件的预测结果, 其中 32862 个位点被鉴定为非编码基因, 730 个位点是蛋

图 1 非编码 RNA 鉴定流程图

为了鉴定非编码基因, GSNAP 和 Cufflinks 分别用来做序列比对和转录组拼接; CPC 和 CPAT 用来区分非编码基因和蛋白编码基因. 过滤掉没有 Rfam 或 EST 支持的单外显子非编码基因后, 最终得到了 16249 个非编码基因

白编码基因. 此外, 将判断一个基因表达与否的标准定义为其 RPKM 值至少在一个组织中大于 0.1. 依据是 15 个组织基因表达的假阳性(false positive rate, FPR)和假阴性(false negative rate, FNR)^[53]的交点在 0.05~0.13(网络版附图 1). 基于这一标准, 31066 个非编码基因和 722 个蛋白编码基因通过了筛选. 为了得到可信的单外显子非编码基因, 过滤了没有 EST 数据或 Rfam 支持的单外显子非编码基因, 最终得到 16249 个表达的非编码基因(网络版数据集 1)和 722 个表达的蛋白编码基因. 根据非编码基因和已知基

因注释的位置关系, 把这些非编码基因分成了 4 类: 与注释重叠的非编码基因、基因间区的非编码基因、内含子区的非编码基因^[54]和反义基因, 各类基因在所有非编码基因中所占的比例分别为 23.06%, 50.50%, 12.66%和 13.78%(表 2). 根据分类, 为每类基因在基因号前加了不同前缀, 它们分别是 XLOC, INTE, INTR 和 ANTI, 代表的意义分别是与注释重叠的非编码基因、基因间区的非编码基因、内含子区的非编码基因和反义基因.

2.2 非编码基因的基因组特征

为了研究非编码基因的基因组特征, 将非编码基因和蛋白编码基因进行了比较. 非编码基因的外显子个数比蛋白编码基因少(每个非编码基因平均 2.4 个外显子, 而每个蛋白编码基因平均 9.7 个外显子). 非编码基因一般比蛋白编码基因短(非编码基因平均长度是 11.2 kb, 而蛋白编码基因是 42.3 kb). 这些结果与 Pauli 等人^[13]和 Cabili 等人^[28]的研究结果一致. 此外, 本实验室研究了非编码基因的外显子个数少是否导致它长度短的主要原因. 分析比较在外显子个数相同情况下, 非编码基因和蛋白编码基因的全基因、外显子和内含子的平均长度, 结果显示非编码基因的基因和内含子长度比蛋白编码基因长, 但外显子长度比蛋白编码基因短(图 2A~C). 这些结果说明与蛋白编码基因相比, 非编码基因的长度相对短是因为它们外显子个数少并且外显子短. 进一步分析显示, 非编码基因的内含子区包含的重复元件比蛋白编码基因多(64.7%的非编码基因内含子有重复元件而只有 53.8%的蛋白编码基因内含子有重复元件; 非编码基因每个内含子平均含 6.96 个重复元件, 而蛋白编码基因内含子平均含 3.93 个重复元件). 内含子区的重复元件大部分是短散在元件(short interspersed element, SINE)、简单重复序列(simple repeat)、长末端重复序列(long terminal repeat, LTR)和长散在元件(long interspersed nuclear elements,

表 2 非编码基因分类^{a)}

类型	重叠	反义	基因间区	内含子区	合计
非编码基因	3747(5249)	2239(3118)	8206(11046)	2057(2156)	16249(21569)
长非编码 RNA 基因	3521(5023)	2217(3096)	8182(11022)	2024(2123)	15944(21264)

a) 将非编码基因分成了 4 类, 分别是与注释重叠的非编码基因、基因间区的非编码基因、内含子区的非编码基因和反义基因; 第 1 个数字是鉴定出的基因个数, 第 2 个数字是这些基因对应的转录本个数

图2 非编码基因和蛋白编码基因的长度和重复序列比较

A-C: 有相同外显子个数的非编码基因和蛋白编码基因的基因长度、内含子长度和外显子长度比较; D: 8种重复元件的数量百分比, 非编码基因内含子区主要的重复元件类型和蛋白编码基因类似, 非编码基因内含子区包含的 LTR 和 LINE 比蛋白编码基因多, 蛋白编码基因内含子区包含的 SINE 比非编码基因多; E: 8种重复元件的长度百分比, 类似地, LTR 和 LINE 对非编码基因内含子区的重复元件长度贡献较大

LINE)(图 2D). 虽然重复元件类型相似, 但是非编码基因中存在的 LTR 和 LINE 的个数和长度的比例都比蛋白编码基因中的高(图 2D 和 E). 当查看重复元件的长度时, 发现 LTR 和 LINE 比其他的重复元件长(网络版附图 2). 为了进一步确认是否是 LTR 和 LINE 导致非编码基因内含子较大, 比较了包含 LTR 和 LINE 的内含子和不包含 LTR 和 LINE 的内含子的长度. 结果显示, 包含 LTR 和 LINE 的内含子比不包含 LTR 和 LINE 的内含子长(网络版附图 3). 这个分析结果意味着高比例的 LTR 和 LINE 使非编码基因的内含子比较大, 大内含子进一步导致在外显子个数相同的情况下, 非编码基因长度比蛋白编码基因长. 此外, 还从总体上比较了 4 类非编码基因的基因长度, 结果发现与注释重叠的非编码基因最长, 反义 RNA 次之. 最有趣的现象是基因间区的非编码基因的长度和内含子区的非编码基因的长度类似(网络版附图 4A). 为了进一步探索它们长度的差异, 在外显子个数相同的情况下比较了它们的长度, 发现单外显子时, 与注释重叠的非编码基因和内含子区非编码基因长度最短和次短, 而多外显子时, 基因间区非编码基因最短, 这可能解释了与注释重叠的非编码基因和内含子区非编码基因长度分布图中相对短的第一个峰和长的最后一个峰(网络版附图 4). 另外, 非编码基因的 GC 含量与蛋白编码基因类似.

2.3 非编码基因的转录活性和保守性分析

之前的研究表明, 蛋白编码基因和非编码基因都与多种染色质修饰相关^[3,55]. 众所周知, 启动子区 H3K4me3 和 H3K27me3 修饰分别是基因表达激活和抑制的标志^[56], 而 H3K36me3 修饰常位于转录延伸区域^[55]. 检查所有非编码 RNA 的这些相关转录元件的结果显示, TSS 周围有 H3K4me3 和 H3K27me3 的显著富集, 而基因区 H3K36me3 的富集程度明显比基因上下游高(图 3A~C).

此外, 研究 TSS 附近的 CAGE 标签和 RNAP II 结合位点的分布情况发现, CAGE 标签和 RNAP II 在 TSS 周围有明显的峰(图 3D 和 E). 这说明非编码基因启动子区这两种标记的显著富集与 CAGE 是启动子标志这一事实相符^[57]. 数据集 S2 列出了至少有组蛋白甲基化修饰、RNAP II 和 CAGE 这 3 种数据中一种支持的非编码基因.

为了评估非编码 RNA 的进化保守性, 计算了非

编码基因、蛋白编码基因和基因间区随机区域的外显子和启动子的保守性分数. 结果显示, 非编码 RNA 的外显子保守性比蛋白编码基因低. 随机区间是这 3 组中保守性分数最低的, 这可能反映了非编码 RNA 结构的约束比氨基酸密码子低, 此结果与以前研究一致(图 4)^[9]. 不像外显子的保守性, 非编码基因的启动子保守性与蛋白编码基因相当. 另外, 比较了 lncRNA 和小 RNA 的保守性, 结果发现 lncRNA 比小 RNA 更保守. 总之, 非编码基因明显比随机区间保守, 这些结果为在哺乳动物中进一步进行非编码 RNA 的功能研究奠定了基础.

这一系列的证据说明非编码基因有它们自己的转录标志, 能够独立转录, 而且它们在进化上是保守的, 以上这些结果都意味着这些非编码 RNA 是有功能的而不是前期研究报道说的转录噪声^[58].

2.4 非编码 RNA 的表达水平和组织特异性

为了研究非编码基因的表达动态, 计算了每个基因在各个组织的 RPKM 值. 结果发现非编码基因的表达量比蛋白编码基因低(表 3). 蛋白编码基因的平均 RPKM 值是非编码基因的 8 倍(蛋白编码基因平均 RPKM 是 17.18, 而非编码基因是 2.18). 此外, 蛋白编码基因的表达量跨越了 7 个数量级(10^{-1} ~ 10^5)而非编码基因只跨越了 6 个数量级(10^{-1} ~ 10^4). 此外, 还比较了 lncRNA 的 4 个子类的表达量(网络版附图 5A), 与注释重叠的 lncRNA(LOWA)是表达量最高的, lincRNA 和反义 RNA 在所有组织中表达量都较低, 内含子区 lncRNA 的表达量在所有 lncRNA 的表达范围内上下波动.

基于非编码基因和蛋白编码基因在各个组织的表达量, 发现它们的表达模式明显不同. 简言之, 非编码基因比蛋白编码基因的组织特异性更显著(表 4, 网络版附图 5B). 从量上来看, 5535(34.06%)个非编码基因只在一个组织中表达, 而在 15 个组织中都表达的非编码基因只有 448(2.8%)个. 前一部分中, 413(7.46%)个非编码基因是大脑特异的, 3812(68.87%)个非编码基因是睾丸特异的; 这些严格意义上特异表达的非编码基因可能在组织特异的过程中有它们特殊的功能.

同时, 通过分析不同组织中差异表达的非编码基因, 得到了一个基于任意 2 个组织相似性的层次聚类图, 结果发现睾丸和大脑与其他组织差异最大(图 5).

图3 非编码基因在 TSS 附近有 H3K4me3, H3K27me3, RNAP II 和 CAGE 富集, 在基因区有 H3K36me3 富集

A 和 B: 非编码基因的 TSS 区富集的 H3K4me3 和 H3K27me3 显著高于平均水平; C: 在基因上下游 5 kb 内研究 H3K36me3 序列的分布, 底部行说明基因区富集的 H3K36me3 序列明显高于上下游区域; D 和 E: RNAP II 和 CAGE 标签显著富集在非编码基因和蛋白编码基因的 TSS 周围

表3 表达量比较(RPKM)

组织	最大 RPKM 值		平均 RPKM 值		非编码基因/蛋白编码基因
	非编码基因	蛋白编码基因	非编码基因	蛋白编码基因	
大脑	27579.40	2904.23	9.51	6.56	1.5/1
卵巢	7815.00	7094.96	2.17	15.53	1/7
乳腺	11543.73	9819.95	1.67	15.64	1/9
胃	25140.23	143063.11	2.78	27.34	1/10
小肠	58749.37	20770.18	5.28	20.03	1/4
肾上腺	23602.23	26078.88	4.71	28.08	1/6
大肠	12512.71	17495.67	1.88	20.96	1/11
胸腺	28372.71	5872.64	4.22	13.42	1/3
睾丸	6580.29	9895.02	2.65	14.93	1/6
肾脏	14664.59	9877.03	2.12	16.66	1/8
肝脏	18602.28	21279.78	2.77	19.50	1/7
肺	16647.83	44579.06	2.67	16.28	1/6
脾脏	52702.23	134225.30	7.56	47.05	1/6
结肠	17686.47	10574.05	2.68	16.74	1/6
心脏	10907.67	26518.70	1.47	16.19	1/11

图 4 非编码基因和蛋白编码基因的外显子和启动子区保守性比较

A: 非编码基因、蛋白编码基因和随机区间的外显子的保守性分数累积分布图. 外显子保守性分数是根据 59 种脊椎动物多序列比对结果获得的保守性分数计算得到的; B: 相似地, 非编码基因、蛋白编码基因和随机区间的启动子区保守性分数累积分布图. P_i 值越大区间越保守

表 4 组织特异性表达非编码基因统计^{a)}

类型	基因数	比例	重叠	基因间区	内含子区	反义
组成型	448	2.76%*	210	51	145	42
组织特异性总和	5535	34.06%*	1548	3006	220	761
大脑特异表达	413	7.46%**	165	175	41	32
卵巢特异表达	115	2.08%**	38	42	18	17
乳腺特异表达	81	1.46%**	16	45	8	12
胃特异表达	40	0.72%**	9	14	12	5
小肠特异表达	23	0.42%**	2	8	11	19
肾上腺特异表达	97	1.75%**	33	33	20	11
大肠特异表达	53	0.96%**	13	31	3	6
胸腺特异表达	120	2.17%**	35	47	22	16
睾丸特异表达	3812	68.87%**	1054	2184	21	553
肾脏特异表达	177	3.20%**	29	115	6	27
肝脏特异表达	110	1.99%**	24	55	11	20
肺特异表达	118	2.13%**	24	66	16	12
脾脏特异表达	146	2.64%**	56	55	20	15
结肠特异表达	61	1.10%**	10	39	5	7
心脏特异表达	169	3.05%**	40	97	6	26

a) *: 这些数据是与所有 16249 个非编码基因比较计算的比例; **: 这些数据是与所有 5535 个组织特异性非编码基因比较计算的比例

综上所述, 非编码基因的表达动态突出了两个特点: 与蛋白编码基因相比, 非编码基因表达量低并且组织特异性强, 而且这两点之间不存在因果关系.

2.5 lincRNA 邻近蛋白编码基因的 GO 分析

除了传统的 tRNA, rRNA 和 snRNA 等非编码 RNA, 新鉴定的 lincRNA 被认为在哺乳动物中有多种生物功能. 例如, 一类称为 ncRNA-activating

(ncRNA-a)的 lincRNA 有类似增强子的功能, 可以正调控邻近蛋白编码基因的表达^[36]. 为了研究新鉴定的 lincRNA 是否有这种调控功能, 重点关注了 1612 个有邻近蛋白编码基因的 lincRNA 的表达模式, 结果发现, 74.44%的 lincRNA-蛋白编码基因对呈现正相关而 25.56%呈现负相关. 而且, 20.22% lincRNA-蛋白编码基因对是高度正相关, 皮尔森相关系数在 0.8~1, 这显著高于随机的基因对 ($P < 2.2 \times 10^{-16}$, t 检

图5 组织层次聚类

A: 用 JS(Jensen-Shannon)距离来衡量任意两个组织的相似性. JS 距离值越大, 组织相关性越差, 反之亦然. 大脑和睾丸都与其他组织相关性较远; B: 用层次树重新进行了层次聚类, Y 轴标出了不同的相关性值. 所有组织按照它们所属的生理系统(如消化系统和免疫系统)聚在了一起; C 和 D: 表达相关性例子. 大肠和小肠的表达相关性(左边)及大脑和睾丸的表达相关性(右边). 显而易见, 大肠和小肠的相关性比大脑和睾丸强, 大脑和睾丸的相关系数偏离了中轴线很多

验). 把 lincRNA-蛋白编码基因对与 3690 对蛋白编码基因-蛋白编码基因对进行了比较, 发现共表达的趋势在一定程度上相似, 蛋白编码基因-蛋白编码基因对也比随机基因对相关性强 ($P < 2.2 \times 10^{-16}$, t 检验)(网络版附图 6). 这些结果与以前的研究结果一致^[9,28]. 总体来说, 可以得出这样的结论, lincRNA 与它们邻近的蛋白编码基因存在一定的共表达关系, 这种形式或许对它们的调控功能很重要. 利用 Gostat 对这些 lincRNA 邻近的蛋白编码基因进行了功能聚类, 结果显示它们显著富集在转录调控、细胞内部分和代谢过程.

2.6 有义-反义基因对

基于每个反义基因对应的有义基因的个数, 新鉴定出的 2239 个有义-反义基因对可以分为一对一和一对多两种类型. 根据基因组上的位置, 共找到 2099 个一对一的有义-反义基因对, 其中 449 个是头对头(convergent), 370 个是尾对尾(divergent), 1280 个是外显子不重叠(non-overlap)^[59]. 评估基因对之间的表达相关性的结果显示, 1040(49.55%)对是正相关, 905(43.12%)对是负相关并且有义-反义基因对表达相关的程度显著高于随机基因对 ($P < 2.2 \times 10^{-16}$, t 检验)(网络版附图 7). 这与反义 RNA 的调控功能相符.

这些有义基因的功能主要富集在信号转导和神经系统发育相关的功能类别中(网络版附图 8). 为了更好地理解它们的调控机制研究人员还需要做许多获得性功能和缺失性功能实验.

2.7 lincRNA 和组织特异性非编码基因的功能富集

本研究的目标是推测 lincRNA 的生物功能, 为进一步实验验证提供理论基础. 为此, 根据 lincRNA 和蛋白编码基因在 15 个组织的表达谱, 计算了每个 lincRNA 和每个蛋白编码基因的爱尔森相关系数, 然后用 GSEA^[9,52]构建了一个 lincRNA 和蛋白编码基因的相关性矩阵(检错率<0.01). 随后, 通过双向聚类成 10 个类, 发现多个 lincRNA 子集与不同的功能类别相关, 包括免疫应答、DNA 复制与起始、肌肉发育和有性生殖等(网络版附图 9, 网络版附表 1).

根据 GSEA 的结果, 每个组织的 lincRNA 都与该组织的生理功能密切相关. 例如, 睾丸中的 lincRNA 主要在减数分裂、第一性征发育、有性生殖和配子形成等生殖发育中起重要作用, 而大脑的 lincRNA 主要参与大脑发育、突触发生、轴突形成和信号转导等.

2.8 lincRNA 的 RT-PCR 验证结果

为了验证分析流程鉴定出的 lincRNA 的真实性, 随机选择了 16 个组成型表达的 lincRNA(即在 15 个组织中都表达)并在 12 个组织进行了 RT-PCR 验证. 16 个 lincRNA 中有 7 个在 12 个组织中的表达量都是可以检测到的, 有 15 个符合至少在 10 个组织中的表达量是可以检测到的(图 6). 另外, 42 个组织特异性 lincRNA 中有 31 个可以用 RT-PCR 检测到(网络版附图 10). 大脑中 lincRNA 的验证率是最高的, 13 个大脑特异的 lincRNA 有 11 个被检测到. 这些验证结果表明分析流程是可信的, 可以产出高质量的 lincRNA 列表. 网络版数据集 3 中列出了用来验证 lincRNA 的 RT-PCR 引物序列等详细信息.

3 讨论

3.1 非编码 RNA 的鉴定与特征描述

本实验用 15 个深度测序的小鼠组织 RNA-seq 数据鉴定出大量非编码 RNA. 这些 RNA-seq 数据允许检测到那些低表达或者严格意义上的组织特异性表

图 6 组成型 lincRNA 的 RT-PCR 结果

M: marker; C: 对照; 1~16: lincRNA. 星号标记的 lincRNA 是在实验中失败的 lincRNA

达基因.

最终, 得到了 16249 个非编码基因, 其中 6315 个是单外显子非编码基因. 为了证明这些非编码基因是有功能的, 不仅研究了它们的基因组特征, 同时也研究了很多其他的辅助特征, 如组蛋白修饰 (H3K4me3, H3K27me3 和 H3K36me3)、RNAP II、CAGE 以及进化保守性. 与蛋白编码基因相比, 这些非编码基因外显子个数少, 基因和外显子短, 但是由于高比例的 LTR 和 LINE 的存在导致了非编码基因的内含子偏长. 这些鉴定出的非编码基因的 TSS 周围有组蛋白修饰、RNAP II 和 CAGE 等数据的支持.

3.2 非编码 RNA 的组织特异性

从 15 个组织的表达谱可以看出, 非编码基因表达量低, 组织特异性强. 有趣的是大脑中非编码基因的表达量整体比其他组织高(表 3). 进一步计算了 NCBI 数据库中大脑 mRNA 数据的非编码基因/编码基因平均表达量比值来揭示这个差异. 分析结果显示大脑中非编码基因/编码基因的比值是 1:1.4, 其比值依然比其他 14 个组织高. 由此可以推断非编码基因/编码基因的表达量比值高可能是大脑组织本身的特性而不是由于不同的建库方法导致的. 为了研究非编码基因的组织特异性是否是由于它们的表达量

低造成的, 比较了表达水平相似的非编码基因和蛋白编码基因的表达宽度. 结果发现, 组织特异性与表达量密切相关(网络版附图 11). 虽然随着表达量升高, 组织特异性变弱, 但是非编码基因的组织特异性仍然比蛋白编码基因明显. 非编码基因的高组织特异性表明它们在组织特异过程中发挥重要的生物功能, GSEA 功能富集分析的结果也证明了这点.

3.3 不同研究之间的非编码 RNA 比较分析

与之前基于小鼠细胞系数据注释的 lincRNA^[9,24]和基于多组织 EST 数据获得的 lincRNA^[60]相比, 本研究鉴定出的 lincRNA 与它们之间的重叠不到 10%. 考虑到本研究用的是组织样本数据, 在已发表的文章中找到一个利用组织样本数据鉴定的 lincRNA 数据集且这些组织包含于本研究的组织列表中^[26], 比较后发现这 6755 个 lincRNA(3965 新的 lincRNA 基因)有 31.16% 与本研究的 lincRNA 重叠. 不同研究鉴定出的 lincRNA 重叠较少可能是由于注释标准和阈值设定的不同导致的^[61]. 仔细检查这些流程, 发现本研究开发的分析流程与其他研究的分析流程之间有很多不同之处. 主要差异包括以下几方面: 不同研究的组织、细胞类型不同; 不同研究转录本测序方法不同, EST, ChIP-seq 或者 RNA-seq 这些都有; 测序数据特征不同, 如链特异性、数据深度和组织数量. 例如, 本研究的数据是链特异的, 链特异性让鉴定反义 RNA 成为可能, 而普通测序方法是没法区分读段来

自哪条链的; 本研究共得到 2282 M 片段, 而其他研究数据量跨度从 4853460 条 ESTs 到 1936 M 片段; 不同研究用了不同的软件来重构转录本和预测新转录本的编码可能性(表 5), 软件之间的特异性和敏感性差别很大. 例如, Luo 等人^[26]鉴定出的新 lincRNA 数据集中, 有 31% lincRNA 只能通过 Scripture 重构出来. 不同研究用了不同阈值来过滤 lincRNA.

3.4 非编码 RNA 直系同源序列

为了进一步证明本研究鉴定出的 lincRNA 是功能元件, 评估了这些 lincRNA 的进化起源. 其中 85.26%(11022 个中的 9397 个)的 lincRNA 在人的基因组上有直系同源区间(网络版数据集 4). 为了进一步估计这些 lincRNA 在其他物种中表达的直系同源转录本, 调查了由 TransMap 共线性比对到小鼠基因组上的哺乳动物和非哺乳脊椎动物的转录本^[28,62]. 该分析鉴定出 1415 个 lincRNA 有来自 TransMap 的共线性直系同源转录本(网络版数据集 4, 网络版附图 12), 占小鼠所有 lincRNA 的比例为 12.84%. 同源性比例低意味着 lincRNA 的保守性低于蛋白编码基因.

3.5 poly(A)-非编码 RNA

目前, 大多数转录组研究都主要关注 poly(A)+ 转录本, 而 poly(A)-转录本有待进一步探索. 在人中, 有约 20% 的转录本是 poly(A)-转录本或二态转录本,

表 5 lincRNA 研究总结

数据集	样品	数据类型 & 数据量	转录本重构软件	预测编码可能性软件	参考文献
~1600 多外显子 lincRNAs	4 种小鼠细胞类型 小鼠胚胎干细胞(mouse embryonic stem cells, ESC)、小鼠胚胎成纤维细胞(mouse embryonic fibroblasts, MEF)、小鼠肺成纤维细胞(mouse lung fibroblasts, MLF)和神经前体细胞(neural precursor cells, NPC)	ChIP-seq	作者自己研发的基于滑动窗口的程序	密码子替换频率方法(codon substitution frequency, CSF)	Guttman 等人 ^[9]
1140 多外显子 lincRNAs	3 种小鼠细胞类型(ESC, NPC 和 MLF)	RNA-Seq (493 M)	Scripture	CSF 和 ORF	Guttman 等人 ^[24]
9490 lincRNAs	多种组织细胞类型	所有公共数据库中可下载的小鼠 EST(4853460)	TGICL	Blastx 和 EMBOSS	Qu 和 Adelson ^[60]
6755 个新 lincRNAs	6 个组织	RNA-Seq (1936 M)	Cufflinks 和 Scripture	CPC 和 CNCI	Luo 等人 ^[26]
11022 lincRNAs	15 个组织	RNA-Seq (2282 M)	Cufflinks	CPC 和 CPAT	本研究

并且一些重要的 lincRNA 是 poly(A)-转录本. 例如, 功能研究相对透彻的 lincRNA *MALAT1* 和 *NEAT1*^[3,63]. 为了挖掘 poly(A)-转录本, 采用去核糖体 RNA 的方法在 HiSeq2000 上对小鼠大脑组织进行了测序, 该方法可以同时捕获 poly(A)+转录本和 poly(A)-转录本. 鉴定 poly(A)-lincRNA 具体步骤如下: 找到 255 个只在大脑组织去核糖体 RNA 的 RNA-seq 数据中表达的 lincRNA. 利用 NCBI 下载的大脑组织 mRNA-seq 数据(SRX191149)去掉可能的 poly(A)+转录本, 发现这 255 个 lincRNA 中有 8 个在其中表达. 这暗示剩下的 247 个 lincRNA(占大脑中所有表达 lincRNA 的 18.70%)可能是 poly(A)-转录本. 分析发现这 247 个大脑潜在的 poly(A)-转录本, 分别有 82(1.43%), 57(2.27%)和 92(4.20%)个 lincRNA 在本实验室前期利用 SOLiD 方法测定的睾丸^[27]、乳腺^[64]和卵巢^[65]去核糖体 RNA-seq 数据中是表达的. 上述

结果暗示大约 1.43%~18.70% lincRNA 可能是 poly(A)-lincRNA. 进一步的分析发现, 这些潜在的 poly(A)-lincRNA 没有 RNAP II 标签的显著富集, 而且 GC 含量也比 poly(A)+lincRNA 低.

4 小结

本研究提供了一个鉴定和分析非编码 RNA 特别是 lincRNA 的方法. 更重要的是, 本研究拓展了目前小鼠的 lincRNome, 为今后 lincRNA 的生理学功能研究奠定了基础. 本研究获得的这些非编码 RNA 的注释信息可以为功能实验提供数据源. 每个 lincRNA 特定的功能还需要进一步的实验验证, 就像验证生存和大脑发育所需的 lincRNA 那样^[66]. 希望将来对 lincRNA 的生理学和病理学研究可以在人类疾病治疗中得到应用.

参考文献

- Clark M B, Amaral P P, Schlesinger F J, et al. The reality of pervasive transcription. *PLoS Biol*, 2011, 9: e1000625
- Ewan Birney J A S, Anindya Dutta R G, Thomas R G, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 2007, 447: 799–816
- Djebali S, Davis C A, Merkel A, et al. Landscape of transcription in human cells. *Nature*, 2012, 489: 101–108
- Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 2002, 420: 563–573
- Katayama S, Tomaru Y, Kasukawa T, et al. Antisense transcription in the mammalian transcriptome. *J Biol Chem*, 2005, 280: 1564–1566
- Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. *Science*, 2005, 309: 1559
- Yue F, Cheng Y, Breschi A, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, 2014, 515: 355–364
- Zheng D, Frankish A, Baertsch R, et al. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res*, 2007, 17: 839–851
- Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 2009, 458: 223–227
- Sati S, Ghosh S, Jain V, et al. Genome-wide analysis reveals distinct patterns of epigenetic features in long non-coding RNA loci. *Nucleic Acids Res*, 2012, 40: 10018–10031
- Pauli A, Rinn J L, Schier A F. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet*, 2011, 12: 136–149
- Hu W, Alvarez-Dominguez J R, Lodish H F. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep*, 2012, 13: 971–983
- Pauli A, Valen E, Lin M F, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*, 2012, 22: 577–591
- Brown C J, Hendrich B D, Rupert J L, et al. The human *XIST* gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 1992, 71: 527–542
- Heard E, Distech C M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev*, 2006, 20: 1848–1867
- Yang P K, Kuroda M I. Noncoding RNAs and intranuclear positioning in monoallelic gene expression. *Cell*, 2007, 128: 777–786
- Dinger M E, Amaral P P, Mercer T R, et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res*, 2008, 18: 1433–1445

- 18 Hawkins P G, Morris K V. Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. *Transcription*, 2010, 1: 165–175
- 19 Guttman M, Donaghey J, Carey B W, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 2011, 477: 295–300
- 20 Haas B J, Zody M C. Advancing RNA-seq analysis. *Nat Biotechnol*, 2010, 28: 421
- 21 Cloonan N, Forrest A R, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 2008, 5: 613–619
- 22 Yassour M, Kaplan T, Fraser H B, et al. *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci USA*, 2009, 106: 3264–3269
- 23 Cui P, Lin Q, Ding F, et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 2010, 96: 259–265
- 24 Guttman M, Garber M, Levin J Z, et al. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*, 2010, 28: 503–510
- 25 Sigova A A, Mullen A C, Molinie B, et al. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci USA*, 2013, 110: 2876–2881
- 26 Luo H, Sun S, Li P, et al. Comprehensive characterization of 10,571 mouse large intergenic noncoding RNAs from whole transcriptome sequencing. *PLoS One*, 2013, 8: e70835
- 27 Liu W, Zhao Y, Cui P, et al. Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. *Front Genet*, 2011, 2: 93
- 28 Cabili M N, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 2011, 25: 1915–1927
- 29 Hangauer M J, Vaughn I W, Mcmanus M T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*, 2013, 9: e1003569
- 30 Wetterbom A, Ameur A, Feuk L, et al. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. *Genome Biol*, 2010, 11: R78
- 31 Ulitsky I, Shkumatava A, Jan C H, et al. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 2011, 147: 1537–1550
- 32 Tan M H, Au K F, Yablonovitch A L, et al. RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. *Genome Res*, 2013, 23: 201–216
- 33 Nam J W, Bartel D. Long non-coding RNAs in *C. elegans*. *Genome Res*, 2012, 22: 2529–2540
- 34 Liu J, Jung C, Xu J, et al. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. *Plant Cell*, 2012, 24: 4333–4345
- 35 Maamar H, Cabili M N, Rinn J, et al. *linc-HOX1* is a noncoding RNA that represses *Hox1* transcription in *cis*. *Genes Dev*, 2013, 27: 1260–1271
- 36 Ørom U A, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 2010, 143: 46–58
- 37 Rinn J L, Kertesz M, Wang J K, et al. Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell*, 2007, 129: 1311–1323
- 38 Gupta R A, Shah N, Wang K C, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 2010, 464: 1071–1076
- 39 Mattick J S. The genetic signatures of noncoding RNAs. *PLoS Genet*, 2009, 5: e1000459
- 40 Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: insights into functions. *Nat Rev Genet*, 2009, 10: 155–159
- 41 Wang K C, Chang H Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 2011, 43: 904–914
- 42 Sasaki Y T, Ideue T, Sano M, et al. MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc Natl Acad Sci USA*, 2009, 106: 2525–2530
- 43 Brown J B, Boley N, Eisman R, et al. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, 2014, 512: 393–399
- 44 Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 2012, 22: 1775–1789
- 45 Consortium E P. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489: 57–74
- 46 Stamatoyannopoulos J A, Snyder M, Hardison R, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol*, 2012,

13: 418

- 47 Pervouchine D D, Djebali S, Breschi A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nat Commun*, 2015, 6: 5903
- 48 Karolchik D, Barber G P, Casper J, et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*, 2014, 42: D764–D770
- 49 Barbosa-Morais N L, Irimia M, Pan Q, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 2012, 338: 1587–1593
- 50 FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest AR, Kawaji H, et al. A promoter-level mammalian expression atlas. *Nature*, 2014, 507: 462–470
- 51 Meyer L R, Zweig A S, Hinrichs A S, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*, 2013, 41: D64–D69
- 52 Subramanian A, Tamayo P, Mootha V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 2005, 102: 15545–15550
- 53 Ramskold D, Wang E T, Burge C B, et al. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 2009, 5: e1000598
- 54 Nakaya H I, Amaral P P, Louro R, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol*, 2007, 8: R43
- 55 Lv J, Liu H, Huang Z, et al. Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. *Nucleic Acids Res*, 2013, 41: 10044–10061
- 56 Li G, Ruan X, Auerbach R K, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 2012, 148: 84–98
- 57 Faulkner G J, Forrest A R, Chalk A M, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, 2008, 91: 281–288
- 58 Hüttenhofer A, Vogel J. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*, 2006, 34: 635–646
- 59 Werner A. Biological functions of natural antisense transcripts. *BMC Biol*, 2013, 11: 31
- 60 Qu Z, Adelson D L. Identification and comparative analysis of ncRNAs in human, mouse and zebrafish indicate a conserved role in regulation of genes expressed in brain. *PLoS One*, 2012, 7: e52275
- 61 Ulitsky I, Bartel D P. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 2013, 154: 26–46
- 62 Kuhn R M, Karolchik D, Zweig A S, et al. The UCSC genome browser database: update 2009. *Nucleic Acids Res*, 2009, 37: D755–D761
- 63 Yang L, Duff M O, Graveley B R, et al. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol*, 2011, 12: R16
- 64 周媛媛, 龚未, 肖景发, 等. 小鼠乳腺发育的转录组学研究——怀孕哺乳周期乳腺的关键调控基因. *中国科学: 生命科学*, 2014, 57: 340–355
- 65 Pan L, Gong W, Zhou Y, et al. A comprehensive transcriptomic analysis of infant and adult mouse ovary. *Genomics Proteomics Bioinformatics*, 2014, 12: 239–248
- 66 Sauvageau M, Goff L A, Lodato S, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *ELife*, 2013, 2: e01749