论文

哺乳动物长链非编码 RNA 的进化注释

卜德超^{①②†}, 罗海涛^{①②†}, 焦飞^③, 方双桑^{①②}, 谭程夫^{①②}, 刘志勇^①, 赵屹^{①*}

① 中国科学院计算技术研究所生物信息学研究组智能信息处理重点实验室, 北京 100190;

② 中国科学院大学研究生院, 北京 100190;

③ 滨州医学院生物化学与分子生物学系, 烟台 264003

† 同等贡献

* 联系人, E-mail: biozy@ict.ac.cn

收稿日期: 2014-12-06; 接受日期: 2015-02-12; 网络版发表日期: 2015-12-01

国家自然科学基金(批准号: 91229120)资助

摘要 哺乳动物的基因组中能转录出成千上万的长链非编码 RNA(lncRNA), 这些新发现的 RNA 分子已经被证实参与了各种各样复杂的生物过程. 目前大多数哺乳动物的 lncRNA 的参考注释并不完全, 这限制了对 lncRNA 的进化以及接下来的功能研究. 本研究组利用 9 个物种 6 种组织的转录组测序数据, 成功地构建出一个完备的哺乳动物 lncRNA 的参考注释集合(约 4142~42558 条 lncRNA). 基于这些 lncRNA, 做了一系列的进化研究, 发现 30%~99% 的 lncRNA 在基因组上是保守的, 这些保守的 lncRNA 中仅有 20%~27% 也能在转录层面保守, 这与编码基因的保守性形成鲜明对比: 保守的编码基因约有 48%~80% 能在转录层面保守. 随后, 基于 lncRNA 的表达量数据成功地构建出其在 9 个物种中的系统发生树, 通过对系统发生树的对比分析发现, lncRNA 的进化速率明显比编码基因快, 且在不同的组织间有很大差别. 将此项研究中得到的 lncRNA 的集合以及其保守和进化的数据收集到 PhyloNONCODE 数据库中(<http://www.bioinfo.org/phyloNoncode>), 这将成为研究非编码 RNA 进化及功能的非常有用的资源.

关键词 长链非编码 RNA, 保守性, 进化

长链非编码 RNA(lncRNA) 是指长度在 200~100000 nt、不编码蛋白质的 RNA. 哺乳动物的基因组上大部分的区域能转录出 lncRNA^[1]. 这一类的非编码 RNA 在很多生物过程中都有调节作用, 例如, 细胞的发育、分化、衰老^[2~5], 并且与许多疾病相关^[6~10]. lncRNA 的表达水平较编码基因更低^[1,11], 并且经常表现出组织特异性增强^[11]. 与编码基因及一

些小的非编码 RNA(如 miRNA 和 snoRNA)相比, 大部分 lncRNA 的序列保守性更低^[12]. 尽管整体上不保守, 许多 lncRNA 内部含有局部特别保守的单元, 这预示它们很可能有潜在的功能^[13~16]. 整体来说, lncRNA 的序列保守性要比随机的更高^[12,17]. 除此之外, lncRNA 常表现出高的物种表达特异性, 并且它们的突变频率更低, 表明它们受到的选择压力更小.

引用格式: 卜德超, 罗海涛, 焦飞, 等. 哺乳动物长链非编码 RNA 的进化注释. 中国科学: 生命科学, 2016, 46: 167~177, doi: 10.1360/05SCLS-2014-0214.R2
英文版见: Bu D C, Luo H T, Jiao F, et al. Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Sci China Life Sci, 2015, 58: 787~798, doi: 10.1007/s11427-015-4881-9

转录层面的快速进化影响其在谱系间的快速诞生和消失^[18], 低表达的特性与其快速进化速率具有相关性^[19]. 这些发现暗示 lncRNA 在不同物种间表达水平的变化或许能解释不同物种间的性状差异, 从某种程度上讲, 它在物种的形成及进化过程上, 可能起与编码基因同等重要的作用. 然而 lncRNA 整体的保守性, 包括其基因组层面保守性和转录组层面的表达进化模式, 这些都没有明确地被评估过. 之所以存在这些问题, 很大一部分原因是由于 lncRNA 的注释集合还没有构建完全.

本研究收集了 9 个物种(人(*Homo sapiens*)、黑猩猩(*Pan troglodytes*)、大猩猩(*Gorilla beringei*)、红毛猩猩(*Pongo pygmaeus*)、恒河猴(*Macaca mulatta*)、小鼠(*Mus musculus*)、负鼠(*Possum*)、鸭嘴兽(*Ornithorhynchus anatinus*)、鸡(*Gallus domestica*)), 每个物种取 6 种组织(大脑、小脑、心脏、肝脏、肾脏、睾丸)的转录组深度测序数据^[20], 这 9 个物种代表了哺乳动物中的主要模式生物, 包括胎盘类哺乳动物、有袋动物和单孔目动物, 鸡在此作为进化中的外群. 采用生物信息学分析方法, 从测序数据中首先重构成转录本集合, 并从中挑选不具备编码能力的 RNA, 经过一系列特征的筛选, 得到 lncRNA. 这批测序数据共鉴定出了 4141~13709 条 lncRNA. 通过与已知注释集合 Ensembl^[21]和其他著名的 lncRNA 数据库^[22-26], 获得了哺乳动物 lncRNA 的一个完整注释集合. 与之前的研究相符合, lncRNA 的表达量更低, 组织特异性更强^[11,22,27]. 为了评估 lncRNA 在整个哺乳动物中的保守性, 对基因组层面和转录组层面的保守性做了研究, 并与编码基因做了相应的对比分析. 结果显示从基因组层面上来看, 30%~99% 的人的 lncRNA 在其他物种中保守, 但在转录组层面, 这些保守的 lncRNA 仅有相当小部分发生转录. 之后用这里鉴定到的保守 lncRNA 的表达量, 构建了 lncRNA 在 9 个物种间的表达系统发生树, 大部分组织构建的树结构都与已知的系统发生树相一致. 通过对系统发生树的分析发现, lncRNA 表达谱的进化模式要与编码基因类似^[20], 表达进化速率在不同组织间有很大不同, 明显比编码基因快, 这说明 lncRNA 可能经历了更强的正向选择.

本研究中所公布的 lncRNA 参考注释集, 以及它的保守和进化信息, 是重要的科学研究数据, 把这些注释和分析结果全部整理到了 PhyloNONCODE 数据

库中, 这是一个专门用于存储哺乳动物 lncRNA 的基因组和转录组进化信息的数据库. PhyloNONCODE 允许用户方便地浏览和查询 lncRNA 多个物种基因组和转录组层面的保守信息, 并且配置了 UCSC 基因组浏览器来进行直观的可视化展示.

1 材料与方法

1.1 数据收集

为了得到哺乳动物的比较转录组, 本研究收集了 GEO(Gene Expression Omnibus)中多个物种的转录组测序数据^[20](accession 号: GSE30352)并做了深度的解析. 这套数据共含有 29 亿 Illumina Genome Analyser IIx 的 76 读长的 reads, 涵盖了 8 个哺乳动物的 6 种组织(大脑、小脑、心脏、肾脏、肝脏和睾丸). 8 个哺乳动物包括: 胎盘哺乳动物(人类、黑猩猩、大猩猩、红毛猩猩、恒河猴和小鼠), 有袋动物(负鼠)和单孔目动物(鸭嘴兽), 鸟类(鸡)作为进化外群. 为了进一步验证分析流程的准确性, 另外收集了大鼠(*Rattus norvegicus*) 5 种组织(大脑、心脏、肾、肝、睾丸)的转录组测序数据用作验证数据集, 采用相同分析流程进行分析. 大鼠的转录组数据下载来自 GEO 数据库(accession 号: GSE41637)^[28].

从 UCSC 基因组浏览器下载参考基因组如下: 人类 *hg19* 基因组、黑猩猩 *panTro3* 基因组、大猩猩 *gorGor3* 基因组、红毛猩猩 *ponAbe2* 基因组、恒河猴 *rheMac3* 基因组、小鼠 *mm9* 基因组、大鼠 *rn4* 基因组、负鼠 *monDom5* 基因组、鸭嘴兽 *ornAna1* 基因组和鸡 *galGal3* 基因组^[29]. 相应的注释基因也从 UCSC 基因组浏览器下载.

1.2 lncRNA 鉴定

基于转录组测序数据, 将短 reads 片段利用比对工具 Tophat(version V1.4.1)^[30]回贴到基因组, Tophat 使用以下参数: min-anchor=5, min-isoform-fraction=0, 其余设置为默认值. 为了方便进行统一量, 同一组织的生物学重复被合并成一个 BAM 文件, 然后利用从头组装工具 Cufflinks(V1.3.0)^[31]用缺省参数为每个组织进行转录本的重构.

为了获得每个物种全面的 lncRNA 注释集合, 本研究开发了一种生物信息学的计算流程包括以下 5 个步骤:

(i) 对每个物种, 利用 cuffcompare 将每个物种的所有重构转录本合并成为一个初始转录本集合. (ii) 合并后的转录本集合与 Ensembl 基因进行比较, 去除已知的蛋白编码基因、部分假基因、microRNA, tRNAs, snoRNAs, rRNAs, snRNAs, scoRNAs; 这里要去除的假基因是所谓的“多态性假基因”, 这些假基因在 Ensembl 数据库都有一个蛋白号注释. 去除这类假基因是为了避免它们可能在 ENCODE 细胞系或组织中有编码的可能性. (iii) 每个转录本的编码能力使用 CNCI(coding noncoding index)软件来进行评估^[32]. CNCI 是一个强大的工具, 可以不依赖于已知的注释有效区分蛋白编码序列和非编码序列. CNCI 软件使用网址 <http://www.bioinfo.org/software/cnci>. (iv) Brawand^[20]的数据并非链特异性. 对于链非特异性数据, Cufflinks 重构时不能决定单外显子转录本的方向. 所以在此方法中, 只保留了多外显子且长度大于 200 nt 的转录本. (v) 蛋白编码基因的基因间、内含子和反义转录区域的转录本被保留, 这些转录本是借助 cuffcompare 与已知编码基因进行比较, 分类代码分别为“i”, “u”或“x”. 保留的转录本最终与已知的 Ensembl 注释进行合并(GENCODE^[23], NONCODE^[24-26]和人类健康组织测序(Human Body Map) lncRNAs^[11]中的已知 lncRNA 被加入这里的人的注释中, NONCODE 中的 lncRNA 被加入小鼠的注释). 9 个物种通过这 5 个步骤共产生了 4141~42558 条 lncRNAs.

1.3 表达和组织特异性分析

利用 Cufflinks^[31]计算 lncRNA 和蛋白编码基因的表达值 FPKM(fragments per kilobase of exons per million fragments mapped). 为了研究 lncRNA 表达的组织特异性, 每个基因计算基于 Jensen-Shannon(JS)散度^[11]的组织特异性度量得分. 特异性度量(0~1)量化组织转录表达模式. 在极端情况下若基因只在一个组织中表达, 那么 JS 值将为 1.

1.4 成对保守性分析

为了评估基因的保守性, 利用 UCSC 的 LiftOver 工具^[29,33], 对一个物种的所有基因, 寻找其在其他物种的同源基因. 即 LiftOver 利用 Gapped BLAST 算法的独立实现 BLASTZ^[34], 专门用于 2 个长的基因组序列之前的比对. 借助 BLASTZ 为核心算法可以寻找 2

个基因组中的同源区域. 它的算法思路是: 把原始基因组打碎为基因小片段, 然后借助线性的 BLASTZ 算法将其比对到目标基因组. 考虑到 LiftOver 算法的优势在于大片段区域之间的同源性的查找, 这里只关注 lncRNA 基因的区域保守性而非内部的转录结构的保守性. 这里所定义的基因组上的同源, 是指某个基因在其目标基因组上能够寻找到. 如果某个基因在某一特定谱系的所有物种均有同源基因, 这个基因被视为某一谱系的保守基因.

本研究定义 CPCC 的数值来衡量 lncRNAs 的保守性程度. 具体定义如下: 假设 X 物种的 A 基因在 Y 物种有保守基因 B, B 在 X 物种中的保守基因为 C. 如果 A 与 C 没有重叠, CPCC 数值为 0, A 称为单向保守 lncRNAs; 如果 A 与 C 部分重叠, CPCC 数值为 A, C 重合部分与 A, C 之和的比值, A 称为双向部分保守 lncRNAs; 如果 A 与 C 完全重叠, CPCC 数值为 1, A 称为双向完全保守 lncRNAs.

不是所有基因的保守基因都能在目标物种中保守转录. 在查看保守基因是否在目标物种表达时, 采取以下准则: 至少一个转录本完全或部分(>80%)落入此同源基因区域时, 此基因才被定义为表达. 保守基因在目标物种中的表达量定量值为落入此基因区域的所有转录本的表达量的总和, 这种定量策略类似于 Cufflinks 计算某一基因的表达量^[35]: 将某一基因区域的所有转录本的表达量加和得到基因的表达量.

1.5 哺乳动物表达系统发生树的构建

lncRNA 和编码基因的表达系统发生树都通过计算两两物种间的表达量间的距离, 采用 neighbor joining 的方法重构出树状结构^[36,37]. 样本之间的距离来自于计算基因表达谱的 $1-\rho$, ρ 是 Spearman 相关系数. Spearman 相关系数是 2 个变量之间统计相关的非参数度量, 它评估 2 个变量之间的关联程度. 采用这个方法是因为它能尽可能地消除异常值和潜在数据归一不准确造成的误差. 低表达量的基因在第一步中会被先去掉. 系统发育树的重构使用 R 的“ape”包^[38]完成. 经过上千次随机抽样来评估分支模式的可靠性(Bootstrap 值).

2 结果

2.1 9 个物种 lncRNA 的参考注释的构建

为了构建哺乳动物全面的 lncRNA 的参考注释,

本研究组做了如下工作:

(1) 对 8 种哺乳动物细胞加 1 种鸟类的共 9 个物种 6 种组织的转录组测序数据, 利用生物信息学方法重构出 RNA, 并按照一系列的筛选条件进一步挑选出 lncRNA(详见材料和方法); 每个物种约有 14294~17202 蛋白质编码基因和 4141~13709 条 lncRNAs 被鉴定出(图 1A, 网络版附表 1)。这些 lncRNA 又按照其与编码基因的位置关系被分为 3 类, 即基因间区域 lncRNA(lincRNA), 内含子区域 lncRNA(intronic lncRNA)和反义 lncRNA (antisense lncRNA)。这些分类信息可以很容易从 Phylo-NONCODE 数据库获取。通过与 ENSEMBL 数据库

中已知的蛋白编码基因和 lncRNA^[21]进行比较, 本研究分别鉴定出人类和小鼠的大约 76%~86%的蛋白质编码基因和 40%~50%的 lncRNA。此外, 通过与 GENCODE V12^[23]进行比较, 共鉴定了其 49% (6426/13249)的编码基因(6426/13249); 与人类健康组织测序^[11]的进行比较, 鉴定了约 37%(3023/8195)的 lncRNA。

(2) 通与整合各种 lncRNA 的数据库, 包括 Ensembl 数据库^[21]中的相应物种的 lncRNA 注释, 其中由于人和小鼠有更丰富的已知注释集, GENCODE^[23], NONCODE^[24~26]和人类健康组织测序^[11]中的人类 lncRNAs 被进一步加入人的参考注释集中,

图 1 lncRNA 的注释集合

A: 8 个哺乳动物和 1 个鸟类中的 lncRNA 注释集合的统计。Y 轴代表 lncRNA 的数目。鉴定到的已知的 lncRNA 用红色柱子表示, 新鉴定出的 lncRNA 用蓝色柱子表示; B: 小鼠的 lncRNA 和编码基因的外显子和转录本长度分布。X 轴是长度的自然对数值, Y 轴是概率密度; C: 小鼠的 lncRNA 和编码基因的表达谱。X 轴是 cufflinks 计算出的 FPKM 的自然对数值; D: 小鼠的 lncRNA(红色)和编码基因(黑色)最大特异性分值(JS score)分布

NONCODE^[24-26]的小鼠 lncRNAs 被加入小鼠的参考注释集中. 最终得到的参考注释集中包括人类、黑猩猩、红毛猩猩、恒河猴、小鼠、负鼠、鸭嘴兽、鸡的 lncRNAs 分别为 42558, 9347, 6517, 4141, 8094, 25464, 5964, 7872 和 6804 个(图 1B, 网络版附图 1).

(3) 对每一个物种 lncRNAs 进行定量分析, 计算相应的 FPKM 值^[35]. 与以往的研究相符^[11,18], 这里得到的哺乳动物 lncRNAs 的平均表达水平比蛋白质编码基因低(图 1C, 网络版附图 2). 为了进一步评估 lncRNA 表达的组织特异性, 对每一个转录本计算其在某一个组织特异性得分(JS score)^[11]. 与以往的报道相吻合^[11,27], 相比蛋白编码基因 10%~17%的比例, 组织特异性表达的 lncRNA 所占的比例明显升高(21%~50%, JS score>0.9)(图 1D, 网络版附图 2 和 3). 因此, 在本研究的所有物种中, lncRNAs 明显比蛋白编码基因有更高的组织特异性, 反映出其在复杂的生物系统中的独特功能.

2.2 哺乳动物 lncRNA 的保守性分析

为了系统地评估哺乳动物 lncRNAs 的保守性, 分析它们从低等到高等不同进化阶段的保守性水平. 针对每一个哺乳动物物种, 对 lncRNAs 和蛋白质编码基因的基因组层面和转录层面的保守性进行了对比. 以人为例, 本研究检测到平均 94.4%(40185/42558)的人的 lncRNA 在胎盘哺乳动物中是保守的, 但在有袋动物和单孔类, 这个比例分别下降到 70.1% (29849/42558)和 54.8%(23303/42558)(图 2A, 网络版附表 4). 3 个谱系的蛋白质编码基因可观察到相似的保守模式(网络版附图 4). 根据 lncRNA 基因组层面的保守性, 定义 CPCC(coverage pattern of conserved counterparts)分值(见材料和方法), 并将 lncRNA 的保守性分为 3 个水平: (i) 单向保守 lncRNAs(CPCC score=0)代表最低保守水平; (ii) 双向和部分保守 lncRNAs (0<CPCC score<1)所代表中间保守水平; (iii) 双向和完全保守 lncRNAs(CPCC score=1)代表最高保守水平.

根据这种打分的度量, 观察到人的 lncRNA 在灵长类双向和完全保守的数目明显比在其他谱系物种中的更高(图 2A). 用 circos 展示的成对保守谱图也表现出相同的模式(图 2B, 网络版附图 5). 与之前的研究相符合^[40,41], 按照 CPCC 打分, *hotair* 基因在整个哺乳动物中广泛存在(网络版附图 11), *xist* 基因在人到

负鼠的基因中存在, 在更低等的鸭嘴兽和鸡物种中不存在(网络版附图 11). 62%和 59%的人的 lncRNA 分别在小鼠和大鼠中有 1:1 的同源基因(CPCC 的打分在 0~1 之间), 这与之之前 6 个哺乳动物的进化研究相符: 它们的研究中这个数字分别是 58%和 54%^[27]. 作为进一步对 CPCC 的准确性验证, 对之前报道的 18 个与眼相关的保守的 lncRNA 基因^[42]进行查看, 发现按照 CPCC 计算的分值, 这 18 个中大部分在哺乳动物中是进化上保守的(网络版附表 7).

lncRNA 在基因组上的保守性仅是第一个层次的保守性, 转录组层面上的保守性更可能与其生物学功能有重要的关联, lncRNA 位点转录层面的保留或缺失对研究哺乳动物的进化及功能有非常重要的意义. 为了查看转录保守的比例, 基于转录组测序数据进行了物种间成对的和谱系内保守基因的转录状态的评估(详见材料与方法). 结果显示 lncRNA 和编码基因具有相当大差异的转录保守模式. 借助成对分析的结果, 尽管有比较大比例的 lncRNA 表现出很强的基因组保守性, 但只有 19.3%~26.3%的保守基因能保守地转录(图 2A, 网络版附表 4). 相反, 70.1%~79.6%的编码基因能保守地转录(网络版附图 4). 谱系内的保守性分析结果中, 这个差别更加明显: 比例分别为 45.9%(18434/40160), 51.0%(18428/36131)和 68.3% (13804/20202)的灵长类、胎盘哺乳类和哺乳类 lncRNAs 在相应的谱系中保守表达(图 2C). 而相应 86.8%(17050/19635), 88.6%(17156/19371)和 92.8% (15235/16417)的编码基因都能保守地表达(图 2C). 这些结果都显示了 lncRNA 的转录模式是更加特异的, 既包含物种特异, 也包含谱系特异, 这比它们的基因组层面的保守性更能反应进化上的特点.

2.3 lncRNA 表达值的进化趋势

越来越多的证据支持这一概念: 即基因调控模式的改变可能导致主要表型差异, 构成自适应的进化. 研究结果显示, lncRNA 的表达变化对物种和谱系特异性有重要贡献. 为了跟踪 lncRNA 表达的进化模式, 本研究为每个组织计算表达距离矩阵来, 并构建相应的基因表达系统发生树. 为了深入了解蛋白质编码基因和 lncRNA 表达式之间的进化模式的差异, 为 9 个哺乳物种保守的 10061 个蛋白质编码基因和 10770 个 lncRNA, 分别构建了不同的进化树, 并做了

图 2 lncRNA 和编码基因的保守性分析

A: 成对保守性搜索结果中人的 lncRNA 在其他物种中保守的数目。柱子的高度代表在目标物种中的保守的数目。折线代表的是这些保守的 lncRNA 中表达的数目。已知的系统发生树在底部标出, CPCC 的定义详见材料和方法; B: 人的 1 号染色体上的 lncRNAs 保守性图谱。图中的圆圈部分展示的是保守性结果的 circos 图^[39]。外圈到内圈量度分别展示了人的 lncRNA 到鸡中保守的 lncRNA。颜色的深浅代表保守性的强弱。下面展示了 3 个例子, 分别对应了灵长类保守, 胎盘哺乳类保守和脊椎动物保守。lncRNA 的结构(黑色)和在 46 个脊椎动物中多序列比对的结果(绿色)在下方展示; C: lncRNA(红色)和编码基因(蓝色)在不同谱系中保守的数目。红线和蓝线分别代表了 lncRNA 和编码基因中至少在所在谱系的某一个物种中表达的数目; D: phastcons 值的累积分布, 红色代表所有编码基因, 绿色代表哺乳动物保守的编码基因, 黑色代表所有的 lncRNA, 蓝色代表哺乳动物保守的 lncRNA

相互间的对比。与之前对编码基因的研究相一致^[20], lncRNA 表达系统发生树与哺乳动物的已知的进化树高度的一致(图 3A 和 B, 网络版附图 6 和 7)。lncRNA 表达系统发生树成功地分开了哺乳动物的几个谱系(胎盘哺乳类, 有袋目和单孔目), 灵长类跟啮齿类也被分开, 人和几个猩猩在系统发生树上被聚在一起, 而与之最近的是恒河猴。这项观测暗示 lncRNA 的表达变化伴随着时间的不断积累, 这样距离越近的物种间的表达差异会越小。为了测试 lncRNA 分析流程的鲁棒性, 本研究另外收集了一套独立的数据集, 取自大鼠的 5 种组织^[28]。这套数据集中共有 18330 个 lncRNAs 被鉴定出, 其中 16163 个是第一次被发现。这 18330 个大鼠 lncRNA 基因采用了相同的进化分析方法。结果显示, 当把大鼠的表达信息加入后, 重构出的 10 个物种的系统发生树, 所有的物种都有其对应的正确的进化位置(网络版附图 8)。这也说明了上

面采取的基于转录组测序数据的一整套分子进化分析方案具有很广的适用性, 能被用于不同的数据源的转录组测序数据。

2.4 不同组织的 lncRNA 进化速率

随后研究 lncRNAs 在不同组织间的表达进化速率。通过对系统发生树的分析发现, 构建的 lncRNA 系统发生树枝长度总和比编码基因更快, 暗示 lncRNA 的表达进化速率更快(图 3C), 这也与之前的研究吻合^[17]。对编码基因和 lncRNA 来说, 树的枝长总和在不同的组织间表现出相似的趋势: 在不同的组织间差异变化很大(图 3C); 与之前编码基因的研究一致^[20], lncRNA 的表达系统发生树在 2 个神经组织中进化得更慢, 这预示着在哺乳动物的进化中神经系统中的 lncRNA 有很强的选择压力。相似的, lncRNA 在睾丸组织中进化速率更快, 说明睾丸组织

图3 哺乳动物 lncRNA 的表达系统发生树

A 和 B: 小脑和睾丸组织的系统发生树. Bootstrap 值是 10770 条 lncRNA 随机取样 1000 次计算得出, 在图中用圆圈表示, 白色 ≥ 0.9 ; 黄色 < 0.9 ; C: 6 种组织的表达系统发生树的分支长度和的比较(br: 大脑; cb: 小脑; ht: 心脏; kd: 肾脏; lv: 肝脏; ts: 睾丸), 箱形图中编码基因用实线表示, lncRNAs 用虚线表示. 错误率: 基于 bootstrap 分值计算的 95% 置信区间(1000 次重复取样); D: 人和其他物种表达量的 Spearman 相关系数. C: 编码基因; N: lncRNA; br: 大脑; cb: 小脑; ht: 心脏; kd: 肾脏; lv: 肝脏; ts: 睾丸. 如 C-br 代表编码基因在大脑的表达量的 Spearman 相关系数

中的 lncRNA 经历了更弱的选择压力(图 3C). 此外, 无论是成对的分析还是表达系统发生树的分析都显示 lncRNA 的表达变化比编码基因的表达变化更大(图 3D).

2.5 Phylo NONCODE 是研究 lncRNA 的保守和进化的综合性数据源

基于 lncRNA 的保守和进化数据, 本研究开发了 PhyloNONCODE 数据库. 前期的工作中, 已经搭建了非编码知识平台包括了一些非编码著名的数据, 如 NONCODE^[24-26], NPInter^[43], anti-ODE^[44]和一些在线的工具和分析服务, 如 ncRNAs^[45]等. Phylo-

NONCODE 是对非编码知识平台的一项重要补充. 作为此知识平台中的重要的一分子, Phylo NONCODE 是第一个提供了大规模的 lncRNA 的鉴定和进化注释信息的资源平台, 可以方便地一站式地进行 lncRNA 的进化分析. 目前来说, Phylo NONCODE 涵盖了 10 个物种(包含 9 个哺乳动物、1 个鸟类)的 lncRNA 的集合. 用户可以方便地对不同谱系的保守或特异的基因进行查询. Phylo NONCODE 提供了浏览功能来显示基因在其他物种的保守状态. 在浏览页或快速检索结果页面包括到每个 lncRNA 的详细信息页的链接, 以使用户很容易获取 lncRNA 的基本信息, 并进一步进行可视化分析(图 4). 借助 UCSC 基因组浏

图4 Phylo NONCODE 中 lncRNA 的浏览和可视化

在浏览页, 每个 lncRNA 在其他 9 个物种中的保守性打分用 CPCC 表示. 浏览页中有转录本/基因明细页的链接. 用户可以进一步查看相关的保守性信息并通过系统的可视化工具做更深入的分析

览器作为可视化工具, 已知的或更新的数据都可以从此获取. 除此之外, 用户还可以上传自己的数据来查看他们感兴趣的 lncRNA 基因上下文关系. 对独立来源的大鼠数据的分析结果显示此流程可以很方便地扩展到各种数据来源的转录组测序数据上. 这也使 Phylo NONCODE 平台具有很强的扩展性.

3 讨论

lncRNA 的已知注释情况在不同的物种间差别很大, 人和小鼠具有最多的 lncRNA 的注释. 但在其他物种中, 即使是在人们熟悉的哺乳动物中, lncRNA 的参考注释仍没有被系统地建立. 本研究提供了一

个哺乳动物完备的 lncRNA 的参考注释集. 不同于之前的研究, 借助 CSF(codon substitution frequency)打分或是与已知蛋白序列的相似性来区分编码和非编码基因^[11,17], 本研究新开发的 CNCI^[32]是一个新的有效的分类工具, 它借助相邻密码子频率来区分编码序列和非编码序列. CNCI 能够有效地对非全长转录本或是反义链的转录本进行编码能力的评估, 且通过训练得到的模型能很好地应用到其他物种, 具有扩物种预测的优势. 之前的研究结果显示 CNCI 非常适用于从转录组测序数据重构出来的转录本, 较其他工具对于非模式生物有非常明显的优势.

对人和小鼠的数据集来说, 比较有趣的一点是: 大部分已知的编码基因都能被重构出来, 但仅有一小部分的 lncRNA 能被重构出来. 这也反应了 2 者之间不同的表达模式: 与编码基因相比, lncRNA 更加组织特异性地表达^[11]. 很显然所鉴定的人的已知编码基因的比例要与之前的利用 24 个正常组织和细胞系的鉴定率是一致的^[11]. 与 Necseulea 等人^[17]的研究进行比较, 发现 2 者之间的一致性较小. 拿人为例, 2 者得到的 lncRNA 的注释集合的重合度约为 50%. 考虑到这里所用的 lncRNA 主要来自 NONCODE 和通过测序数据预测得到, 说明不同的数据源和不同的预测工具对结果的影响很大. 目前很多 lncRNA 的可信度还未知, 本研究尝试提供一个 lncRNA 的子集, 作为一个黄金数据集, 这将在下一版的 Phylo NONCODE 中提供.

基因组上的弱保守性使得对 lncRNA 的进化及功能的研究要困难不少, 但长 lncRNA 的独特的组织特

异性的表达模式显示它们很可能具有重要的进化上的意义. 二代测序技术的出现可以从转录层面来研究 lncRNA 的特性. 本研究整合了 lncRNA 的基因组和转录组的保守性, 形成了哺乳动物进化的一个宏观的展示. 之前的结果已经显示编码基因能广泛地表达并且非常保守, 而 lncRNA 具有很强的时空表达特异性并且快速地进化^[17,18,22,27]. 猜测这是由于 2 者的不同角色造成的: 编码基因作为表型的直接调控者, 而 lncRNA 作为中间调控者, 更倾向于通过调节编码基因来控制表型^[46-49].

值得注意的是, 与编码基因有重叠的 lncRNAs 的基因组的保守性并不能像 lincRNAs 那样定义那么准确^[14]. 然而为了提供一个全面的 lncRNA 的注释及保守信息, 所有类别的 lncRNA 都被收录在了 Phylo NONCODE. 事实上, 当对 lincRNA 做同样的分析时, 仅有极小的差别被观测到(鸭嘴兽和负鼠在系统发生树上的位置发生了互换)(网络版附图 9 和 10).

两种层面的保守性: 基因组保守性和转录组保守性被第一次收录在一个数据库 Phylo NONCODE 中. CPCC 打分值代表基因组上的保守性, FPKM 代表转录组的信息. Phylo NONCODE 提供给了生物学家一个探索 lncRNA 特性的很重要的资源, 来进一步帮助对 lncRNA 重要功能及进化意义的探究. 越来越低的价格也使越来越多的 lncRNA 以一个前所未有的速度被鉴定出来. 这里所提供的 lncRNA 的分析流程将能适用于各种类型的数据, Phylo NONCODE 数据库将成为研究 lncRNA 功能的重要数据源.

参考文献

- 1 Mattick J S, Makunin I V. Non-coding RNA. *Hum Mol Genet*, 2006, 15: R17-R29
- 2 Ponting C P, Oliver P L, Reik W. Evolution and functions of long noncoding RNAs. *Cell*, 2009, 136: 629-641
- 3 Wilusz J E, Sunwoo H, Spector D L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev*, 2009, 23: 1494-1504
- 4 Moran V A, Perera R J, Khalil A M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res*, 2012, 40: 6391-6400
- 5 Puvvula P K, Desetty R D, Pineau P, et al. Long noncoding RNA PANDA and scaffold-attachment-factor SAFA control senescence entry and exit. *Nat Commun*, 2014, 5: 5323
- 6 Chen G, Wang Z, Wang D, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*, 2013, 41: D983-D986
- 7 Kung J T, Colognori D, Lee J T. Long Noncoding RNAs: past, present, and future. *Genetics*, 2013, 193: 651-669
- 8 Liu Q, Huang J, Zhou N, et al. LncRNA loc285194 is a p53-regulated tumor suppressor. *Nucleic Acids Res*, 2013, 41: 4976-4987
- 9 Xue Y, Ma G, Gu D, et al. Genome-wide analysis of long noncoding RNA signature in human colorectal cancer. *Gene*, 2015, 556: 227-234

- 10 Chakravarty D, Sboner A, Nair S S, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat Commun*, 2014, 5: 5383
- 11 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
- 12 Marques A C, Ponting C P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol*, 2009, 10: R124
- 13 Ponjavic J, Ponting C P, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, 2007, 17: 556–565
- 14 Mitchell Guttman I A, Garber M, French C, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 2009, 458: 223–227
- 15 Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*, 2010, 20: 1313–1326
- 16 Liz J, Portela A, Soler M, et al. Regulation of pri-miRNA processing by a long noncoding RNA transcribed from an ultraconserved region. *Mol Cell*, 2014, 55: 138–147
- 17 Necsulea A, Soumillon M, Warnefors M, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 2014, 505: 635–640
- 18 Kutter C, Watt S, Stefflova K, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet*, 2012, 8: e1002841
- 19 Managadze D, Rogozin I B, Chernikova D, et al. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol*, 2011, 3: 1390–1404
- 20 Brawand D, Soumillon M, Necsulea A, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 2011, 478: 343–348
- 21 Flicek P, Amode M R, Barrell D, et al. Ensembl 2011. *Nucleic Acids Res*, 2011, 39: D800–D806
- 22 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
- 23 Harrow J, Frankish A, Gonzalez J M, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 2012, 22: 1760–1774
- 24 Liu C, Bai B, Skogerbo G, et al. NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res*, 2005, 33: D112–D115
- 25 He S, Liu C, Skogerbo G, et al. NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res*, 2008, 36: D170–D172
- 26 Bu D, Yu K, Sun S, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res*, 2012, 40: D210–D215
- 27 Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res*, 2014, 24: 616–628
- 28 Merkin J, Russell C, Chen P, et al. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 2012, 338: 1593–1599
- 29 Fujita P A, Rhead B, Zweig A S, et al. The UCSC genome browser database: update 2011. *Nucleic Acids Res*, 2011, 39: D876–D882
- 30 Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 2009, 25: 1105–1111
- 31 Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 2012, 7: 562–578
- 32 Sun L, Luo H, Bu D, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*, 2013, 41: e166
- 33 Kent W J, Baertsch R, Hinrichs A, et al. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA*, 2003, 100: 11484–11489
- 34 Schwartz S, Kent W J, Smit A, et al. Human–mouse alignments with BLASTZ. *Genome Res*, 2003, 13: 103–107
- 35 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 2010, 28: 511–515
- 36 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 1987, 4: 406–425
- 37 Studier J A, Keppler K J. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*, 1988, 5: 729–731
- 38 Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 2004, 20: 289–290
- 39 Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*, 2009, 19: 1639–1645
- 40 Rens W, O'Brien P C, Grutzner F, et al. The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol*, 2007, 8: R243
- 41 He S, Liu S, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol*, 2011, 11: 102

- 42 Mustafi D, Kevany B M, Bai X, et al. Evolutionarily conserved long intergenic non-coding RNAs in the eye. *Hum Mol Genet*, 2013, 22: 2992–3002
- 43 Wu T, Wang J, Liu C, et al. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res*, 2006, 34: D150–D152
- 44 Yin Y, Zhao Y, Wang J, et al. antiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics*, 2007, 8: 319
- 45 Liao Q, Xiao H, Bu D, et al. ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res*, 2011, 39: W118–W124
- 46 Roberts T C, Morris K V, Weinberg M S. Perspectives on the mechanism of transcriptional regulation by long non-coding RNAs. *Epigenetics*, 2014, 9: 13–20
- 47 Pennisi E. Lengthy RNAs earn respect as cellular players. *Science*, 2014, 344: 1072–1072
- 48 Lau E. Non-coding RNA: zooming in on lncRNA functions. *Nat Rev Genet*, 2014, 15: 574–575
- 49 Alvarez-Dominguez J R, Hu W, Yuan B, et al. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood*, 2014, 123: 570–581