**人类信使RNA和长链非编码RNA基本属性比较**

**Comparison of human mRNA and lncRNA basic properties**

马旭营 伊现富

**摘 要:**

长链非编码RNA(long non-coding RNA，lncRNA)是指一类长度大于200nt，几乎不具有蛋白编码功能的RNA。近年来大量的研究证明lncRNA在生物体中广泛地参与各种生理及病理过程，具有促进细胞凋亡、病毒入侵、多能干细胞去分化等多种功能，与生物体的生活息息相关。本文主要比较了lncRNA与mRNA的一些基本属性，重点比较了两者的编码潜能，探究lncRNA的功能，为进一步深入研究lncRNA提供参考。

**Abstract:**

Long non-coding RNA (lncRNA) refers to a class of RNA with a length greater than 200 nt and almost no protein coding function. In recent years, a large number of studies have shown that lncRNA is widely involved in various physiological and pathological processes in the organism. It has many functions, such as promoting apoptosis, virus invasion and pluripotent stem cell dedifferentiation, and is closely related to the life of the organism. In this paper, we compared some basic properties of lncRNA and mRNA, focusing on the coding potential of both, exploring the function of lncRNA, and providing a reference for further study of lncRNA.

关键词：

长链非编码RNA；mRNA；编码潜能

1. 前言

1.1 lncRNA的概念

长链非编码RNA(long non－coding RNA 或long noncoding RNA，lncRNA) 亦称长非编码RNA，是一类不有相同的转录机制，即需要RNA聚合酶Ⅱ的参与以及与转录起始和延伸相关的组蛋白的修饰。这些lncRNA具有５’端的甲基鸟苷帽子，相当一部分lncRNA的３’端具有多聚腺苷尾。

最初研究认为，这些由不编码蛋白质的DNA转录出的产物没有生物学作用，仅仅是“转录噪声”。近期研究发现，lncRNA虽然不能被翻译成蛋白质，但它参与许多生命活动在基因转录调控、翻译后修饰和表观遗传学调控等方面有十分重要的作用，并且和疾病的发生、发展、诊断及治疗有密切的关系[1]。

1.2 lncRNA的分类

对lncRNA进行科学的分类是有效地研究lncRNA的前提，目前比较公认的是根据其来源于基因组的不同位置分为五类：①正义的(sense)，蛋白编码基因转录产物经过剪接产生的无编码功能的转录本；②反义的(antisense，AS)，由与蛋白编码基因互补的DNA链转录产生，且和该基因有一个或数个外显子的重叠；③双向的(bidirectional)，基因组邻近区域互补的两条链上的基因同时表达的转录本；④内含子的(intronic)，即完全来源于基因的内 含子；⑤基因间的(intergenic)，位于两个基因间的区域[2]。

1.3 lncRNA的分子结构

lncRNA的一级结构即为lncRNA的核苷酸排列顺序。lncRNA调节基因功能的途径多种多样，其中最为重要的一种方式便是通过碱基互补配对方式与靶基因结合来直接调节靶基因的转录翻译或间接调节靶基因上游或下游基因的转录翻译，碱基配对的基础便是其一级结构。研究报道lncRNA Gas5可直接与糖皮质激素上的DNA结合域(DNAbinding domain)结合，进而与含有糖皮质激素反应元件(glucocorticoid response elements)目的基因竞争并调节其表达，而与靶标碱基互补配对的基础便是lncRNA的一级结构。

lncRNA二级结构及三级结构(空间结构) 是lncRNA发挥其功能的中枢。2012年，Novikova等报道了人类 SRAlncRNA(steroid receptor RNA activator lncRNA )二级结构信息。SRA能够激活数种性激素受体，并与乳腺癌的发病密切相关。

目前，还没有更多关于lncRNA空间结构(三级结构及四级结构)的研究报道，现有的关于lncRNA高级结构的认识仅来源于NEAT1。NEAT1的两个亚基拥有相同的启动子及相似表达量(人NEAT1亚基为: NEAT\_V1∶3.7kB，NEAT\_V2∶22.7 kB )，二者均参与特异性核腔隙(specific nuclear compartments) -paraspeckles的形成[3]。

1.4 lncRNA的作用机制

研究表明,lncRNA发挥生物学功能的主要机制有基因印记(Genetic imprinting)、染色质重塑、细胞周期调控、剪接调控、mRNA降解和翻译调控等(表1)。

表1 部分 lncRNA 的生物学功能

|  |  |
| --- | --- |
| lncRNA名称 | 生物学功能 |
| H19 | 基因印记 |
| Xist | X染色体失活 |
| Tsix | 阻断Xist积累,维持X染色体的活性 |
| HOTAIR | 组蛋白修饰复合体的骨架分子 |
| ANRIL | 抑制转录 |
| Gas5 | 糖皮质激素受体的诱饵 |
| lincRNA-p21 | 通过结合转录因子抑制基因表达 |
| PANDA | 转录因子的诱饵 |
| hsrω-n | 调控mRNA前体的剪接 |
| sat III | 调控mRNA前体的剪接 |
| MALAT1 | 调控丝氨酸/精氨酸剪接因子磷酸化 |
| 1/2-sbsRNA | 介导mRNA降解 |
| BACE1-AS | 增加mRNA的稳定性 |

1.5 lncRNA研究中存在的问题

当前, lncRNA研究正处于起步阶段, 因此面临着许多亟待解决的问题：(1)lncRNA的定义仍存在争议。一般认为, lncRNA是长度大于200个核苷酸的非编码RNA。但是, Spizzo等认为, 以200个核苷酸作为界定lncRNA过于武断, 因为小于200个核苷酸的非编码RNA中还存在很多非编码RNA, 它们既不属于小RNA(Small RNA)也不属于结构RNA (Structural RNA)。(2)lncRNA生物学功能的阐明并非易事。一方面, 如何区分功能性和非功能性非编码转录物依然存在困难, 另一方面, 由于lncRNA种类和功能的多样性, 致使不同的lncRNA研究结果之间的借鉴意义并不高。(3)尚无统一的命名原则。目前lncRNA还没有一个规范的命名方法, 只是研究者根据其功能、结构特点、作用方式等进行命名, 有时很难从名称中了解其真正含义和功能。 (4)lncRNA数据库不够全。相对于其他非编码RNA数据库, lncRNA相关数据库的内容还不够全, 对lncRNA的注释远远不够丰富。(5)lncRNA功能预测的工具不多。针对lncRNA的生物信息学工具仍然极少, 例如, 目前难以对lncRNA二级结构和靶标等进行有效地预测。(6)研究领域有待拓展。目前, 有关lncRNA的研究主要集中于肿瘤、神经、发育、植物等领域, 在其他领域和对其他疾病的研究依然欠缺。(7)用于lncRNA研究的新技术并不多。因此, 需要建立更多、更有效的研究方法用于系统地研究lncRNA的结构和功能[4]。

本次研究希望通过对mRNA与lncRNA的基本属性进行对比，从而加深对lncRNA的了解，为以后对lncRNA的深入探索打下基础。

2. 原理和方法

本次实验使用的数据是人类的部分mRNA和全部lncRNA，基因组版本为hg19(GRCh37)。

2.1人类mRNA与lncRNA编码潜能的比较

首先从NONCODE(http://www.noncode.org/)中下载人类全部lncRNA的“bed”格式文件，然后打开CPAT(http://lilab.research.bcm.edu/cpat/)上传数据。CPAT是一款预测RNA编码潜能的在线工具，与其他的预测编码潜能工具不同，CPAT不是基于比对来搜索蛋白质证据或多重比对来计算系统发育保守性评分，而是使用了一种以4种序列特征：开放阅读框大小，开放阅读框覆盖，Fickett TESTCODE统计和六联体使用偏倚构建的逻辑回归模型来预测编码潜能。这种逻辑回归模型使得CPAT具有很高的准确性，同时也使得CPAT比Coding-Potential Calculator和Phylo Codon Substitution Frequencies快约4个数量级，使得用户能够在数秒钟内处理数以千计的转录本。CPAT允许用户上传“bed”格式或“fasta”格式的数据，也可以直接粘贴数据或数据所在网址，支持的人类基因组版本为hg19(GRCh37)。但CPAT也有一个小缺陷，那就是能够承载的数据量较小，仅能接受10M以内的数据。由于lncRNA的数据文件大于10M，所以不能直接上传，分三次将其手动粘贴至接受数据区域，选择物种基因组版本为Human (hg19, GRCh37)，然后提交即可。将跳转页面的预测结果全部粘贴至Excel，使用函数“COUNTIF”统计其中可以编码的lncRNA数目，计算其所占比例，再将原始数据上传至Galaxy(https://usegalaxy.org/)，使用Text Manipulation工具集中的“Text reformatting with awk ”工具将每条染色体上的lncRNA分离出来再进行每条染色体上lncRNA的编码潜能预测。由于人类全部基因组的mRNA数据量实在太大，本次实验在每条染色体上随机选取了2000条mRNA进行预测，然后在Excel中进行统计，再将两者进行比较。

2.2人类mRNA与lncRNA的SNP比较

首先将从NONCODE中下载的lncRNA数据上传至Galaxy，使用“Text reformatting with awk ”提取出每条染色体上的lncRNA。然后从Galaxy的数据库中搜索人类基因组中每条染色体的SNP，基因组版本选择hg19(GRCh37)。使用Operate on Genomic Intervals工具集中的“Join”工具，以lncRNA数据为第一套数据集，以SNP数据为第二套数据集，其他选项默认，通过坐标比较提取含有SNP的lncRNA。再使用Join, Subtract and Group中的“Group”工具，以提取过SNP的lncRNA为数据集，对lncRNA的ID进行计数，“Group by column”中选择“column4”，点击“Insert Operation”，“Type”中选择“Count”，“On column”选择“1”，其余参数默认，得到的结果中第一列为lncRNA的ID，第二列为每条lncRNA上的SNP数目。然后使用Filter and Sort工具集中的“Sort”工具，对第二列SNP数目采用降序排列，所有参数默认即可。将所得出的结果文件下载复制到Excel中，统计每条lncRNA上SNP的数目，再计算平均数目。

对mRNA的操作与上述步骤相同，在此不再赘述。最后将两者进行比较。由于mRNA的数据量过大，Galaxy运行时间过长，而时间又有些紧迫，所以本实验只选取了Y染色体作为例子，以后如有机会，在进行其他染色体的操作。

3. 结果

3.1人类的mRNA与lncRNA编码潜能比较

根据CPAT工具的预测以及Excel分析的结果，mRNA的平均编码潜能约为3.48%，也就是说，在本次实验所选择进行统计的48000条mRNA中，约有1670条mRNA可以用于编码蛋白质，具有编码潜能；而lncRNA的平均编码潜能为2.85%，换句话说，在所有的227332条lncRNA中，约有6479条lncRNA可以用于编码蛋白质，具有编码潜能(图1)。由此可以看出mRNA的编码潜能略高于lncRNA，但是差异并不显著，仅为0.63%。

图1 人类mRNA与lncRNA平均编码潜能比较

关于每条染色体上mRNA与lncRNA编码潜能的差异，在1-6、10-15、17、20、Y染色体上较为显著，即mRNA的编码潜能高于lncRNA，但是差异并不显著，分别为1.97%、 1.76%、 0.24%、0.86%、5.45%、3.10%、1.91%、1.19%、0.14%、1.81%、2.35%、0.19%、1.05%、0.15%、0.77%；在7-9、16、18、19、21、22、X染色体上，lncRNA的编码潜能与mRNA相比较高一些，然而差异也不显著，分别为2.07%、0.37%、0.74%、1.10%、3.15%、0.20%、0.33%、0.05%、0.02%(表2，图2)。

表2 每条染色体上人类mRNA和lncRNA编码潜能比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 染色体号 | 基因总数目 | 可编码数目 | 基因总数目 | 可编码数目 | 编码潜能 | |
|  | mRNA | | lncRNA | | mRNA | lncRNA |
| 1 | 2000 | 88 | 19407 | 472 | 4.40% | 2.43% |
| 2 | 2000 | 73 | 19662 | 372 | 3.65% | 1.89% |
| 3 | 2000 | 42 | 13209 | 246 | 2.10% | 1.86% |
| 4 | 2000 | 58 | 11340 | 231 | 2.90% | 2.04% |
| 5 | 2000 | 157 | 13241 | 318 | 7.85% | 2.40% |
| 6 | 2000 | 106 | 14176 | 312 | 5.30% | 2.20% |
| 7 | 2000 | 22 | 11405 | 362 | 1.10% | 3.17% |
| 8 | 2000 | 41 | 10852 | 263 | 2.05% | 2.42% |
| 9 | 2000 | 42 | 9642 | 274 | 2.10% | 2.84% |
| 10 | 2000 | 79 | 10527 | 215 | 3.95% | 2.04% |
| 11 | 2000 | 93 | 9939 | 344 | 4.65% | 3.46% |
| 12 | 2000 | 51 | 10921 | 263 | 2.55% | 2.41% |
| 13 | 2000 | 60 | 7584 | 90 | 3.00% | 1.19% |
| 14 | 2000 | 90 | 7503 | 161 | 4.50% | 2.15% |
| 15 | 2000 | 46 | 7150 | 151 | 2.30% | 2.11% |
| 16 | 2000 | 71 | 7713 | 359 | 3.55% | 4.65% |
| 17 | 2000 | 98 | 8952 | 345 | 4.90% | 3.85% |
| 18 | 2000 | 33 | 5643 | 271 | 1.65% | 4.80% |
| 19 | 2000 | 118 | 6781 | 414 | 5.90% | 6.10% |
| 20 | 2000 | 63 | 6196 | 186 | 3.15% | 3.00% |
| 21 | 2000 | 45 | 3720 | 96 | 2.25% | 2.58% |
| 22 | 2000 | 98 | 4383 | 218 | 4.90% | 4.95% |
| X | 2000 | 56 | 6232 | 176 | 2.80% | 2.82% |
| Y | 2000 | 38 | 1154 | 130 | 1.90% | 1.13% |
| 总计 | 48000 | 1668 | 227332 | 6479 | 3.48% | 2.85% |

图2 每条染色体上人类mRNA和lncRNA编码潜能比较

3.2 人类mRNA和lncRNA上SNP数目比较

根据Galaxy运行出的结果以及Excel分析出的结果，从人类Y染色体上lncRNA中提取出的SNP数目为12060个，Y染色体上的lncRNA条数为1157条，平均每条lncRNA上的SNP数目为10.45个；从人类Y染色体上mRNA中提取出的SNP数目为174429个，Y染色体上的mRNA条数为60583条，平均每条mRNA上的SNP数目为2.88个(表3，图3)。由此可以看出在人类Y染色体上lncRNA的平均SNP数目要显著高于mRNA的平均SNP数目，且差距并不算小，为7.57，接近mRNA平均SNP数目的三倍。

表3 人类Y染色体上mRNA与lncRNA的平均SNP数目比较

|  |  |  |  |
| --- | --- | --- | --- |
| RNA种类 | RNA数目 | SNP数目 | 平均SNP数目 |
| mRNA | 60583 | 174429 | 2.88 |
| lncRNA | 1157 | 12060 | 10.45 |

图3 人类Y染色体上mRNA与lncRNA的平均SNP数目比较

4. 结论

根据上述结果进行分析，大致可以得到如下结论：

(1)人类的mRNA编码潜能要略大于lncRNA，因为mRNA要编码蛋白质，而lncRNA并不编码蛋白质，故而lncRNA的编码潜能要比mRNA的编码潜能小一些，这一点与已知的研究进展是一致的。

(2)人类的Y染色体上mRNA的SNP数目要小于lncRNA的SNP数目，这说明mRNA的保守性要大于lncRNA，因为mRNA要编码蛋白质，而lncRNA一般不编码蛋白质，所以lncRNA的变异位点多一些对生物体的表型影响并不算大，这一点也与研究进展及实验之前的推论是相同的。

5. 讨论

由于一些客观上的因素以及主观上的阻力，本次实验尚存在以下问题：

(1)由于人类的全部基因组mRNA数据量过于庞大，CPAT这个工具无法承载如此巨大的数据量，故而在每条染色体上挑选了2000条mRNA，由于样本数目较小，导致mRNA编码潜能与lncRNA编码潜能的差距并不显著，如果样本数据量大一些，或使用整个基因组mRNA，差异可能会更大一些。

(2)由于实验者的能力有限，未能利用现有知识随机挑选样本，所以挑选mRNA的过程有些随意，不够专业。

(3)由于Galaxy的运行时间过长，而实验时间有限，所以SNP的比较只选取了人类Y染色体作为例子，以后要如有时间，应使用所有的染色体进行分析。

6. 参考文献

[1] 井深，张惠荣.长链非编码RNA 研究现状与趋势的文献计量分析.中华医学图书情报杂志,2012,21(9)：54-56

[2] 王国强，卫宁，王禹等.长链非编码RNA的生物学功能研究进展.家畜生态学报,2014,35(3)：1-5

[3] 王婷梅，曲丽娜，李莹辉.lncRNA的结构、功能及其与疾病的关系.中国生物化学与分子生物学报,2015,31(7):659-666

[4] 夏天,肖丙秀,郭俊明.长链非编码RNA的作用机制及其研究方法.HEREDITAS (Beijing),2013,35(3):269―280