PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2014, 41(2): 153~162
www.pibb.ac.cn

一个肝癌相关长链非编码 RNA 的克隆及序列分析 *

唐 珂[1, 2)**] 魏 芳[2)**] 李 昊[2)] 黄宏斌[2, 4)] 张文玲[2)] 龚朝建[2)]
李夏雨[3)] 宋亚莉[2)] 廖前进[1)] 彭淑平[2)] 向娟娟[2)] 周 鸣[2)] 马 健[2)]
李小玲[2)] 熊 炜[1, 2, 3)] 李 勇[5)] 曾朝阳[1, 2, 3)***] 李桂源[1, 2, 3)***]

([1)] 中南大学湘雅医学院附属肿瘤医院，湖南省肿瘤医院，长沙 410013；
[2)] 中南大学肿瘤研究所，卫生部癌变原理重点实验室及教育部癌变与侵袭原理重点实验室，长沙 410078；
[3)] 中南大学湘雅三医院疾病基因组研究中心，湖南省非可控性炎症与肿瘤重点实验室，长沙 410013；
[4)] 国防科技大学信息系统工程重点实验室，长沙 410073；
[5)] *Department of Biochemistry and Molecular Biology, Center for Genetics and Molecular Medicine, School of Medicine, University of Louisville, Louisville, KY* 40202, USA)

摘要 最近我们利用新一代测序(next generation sequencing，NGS)技术对肝细胞癌(hepatocellular carcinoma，HCC)患者活检标本及正常对照肝组织样品进行高通量 RNA 测序(RNA-sequencing，RNA-Seq)，在肝癌样品中染色体 11q13.1 区域检测到几个相邻的 RNA-Seq 信号峰，而在正常对照组织中没有检测到，且该染色体区域目前尚无已知基因登录，提示这几个 RNA-Seq 峰可能代表一个或多个未知的新基因．以此为线索，证实这几个 RNA-Seq 峰来自同一个新基因，并克隆了该基因全长序列，在克隆该基因全长序列时，发现该基因编码的 RNA 存在多种剪接形式，最长的转录本为 3 562 bp．将该基因编码的 12 条代表性 RNA 转录本序列递交到美国国立生物技术信息中心(National Center for Biotechnology Information，NCBI)的 GenBank 数据库中，GenBank ID 号分别为 KC136297～KC136308．该基因编码的 RNA 没有发现明显的开放阅读框(open reading fragment，ORF)，提示该基因可能编码长链非编码 RNA(long non-coding RNA，lncRNA)．为了探讨该 lncRNA 基因可能的转录调控机制，我们用生物信息学方法预测了该 lncRNA 基因潜在启动子区域，发现在其转录起始位点上游–719～–469 bp 处有一个潜在的启动子，其中包含 7 个 Sp1、1 个 STAT5 和 1 个 EGR1 转录因子结合位点．该 lncRNA 在肝细胞癌发生发展过程中的作用机制值得进一步深入研究．

关键词 长链非编码 RNA，肝细胞癌，染色体 11q13.1，基因克隆，生物信息学
学科分类号 Q61 **DOI**: 10.3724/SP.J.1206.2012.00613

原发性肝癌是亚洲与部分非洲地区的高发肿瘤，其病死率居我国恶性肿瘤的第二位，五年生存率只有 5%左右，肝癌的防治及其发病机制研究一直是我国医学科学研究中的重要课题．肝细胞癌(hepatocellular carcinoma，HCC)是原发性肝癌的一种主要类型，尽管目前在肝细胞癌的诊断和治疗方面有了新的进展，但是对于其发生发展及转移的确切机制仍然存在大量未知的问题．

新一代测序(next generation sequencing，NGS)技术的迅猛发展将基因组水平的研究带入了一个新的阶段，许多基于全基因组的研究成为可能，RNA 测序(RNA-sequencing，RNA-Seq)[1]就是利用新一代测序技术对生物样品转录的全部 RNA 进行高通量测序，以获得被测样品转录组(transcriptome)数据的一种重要新技术．

* 国家自然科学基金(81372907, 81172189, 81171930, 81272298, 81272255 和 91229122), 湖南省自然科学基金(14JJ1010), 霍英东高校青年教师基金(121036), 中央高校基本科研业务费专项基金(2011JQ020)及中南大学博士后科学基金资助项目．
** 共同第一作者．
*** 通讯联系人．Tel: 0731-84805383
曾朝阳．E-mail: zengzhaoyang@xysm.net
李桂源．E-mail: ligy@xysm.net
收稿日期：2013-03-20，接受日期：2013-07-11

最近我们利用新一代测序技术对肝细胞癌患者活检标本及正常对照肝组织样品进行 RNA-Seq，在肝癌样品中检测到染色体 11q13.1 区域存在相邻的几个明显的 RNA-Seq 信号峰，而在正常对照组织中却没有，且该染色体区域目前尚无已知基因登录，提示可能存在一个或多个未知的新基因．我们以此为线索，证实这几个 RNA-Seq 峰来自同一个新基因，并克隆了该基因全长序列，同时对新克隆的基因序列进行了生物信息学分析．

1 材料与方法

1.1 新一代测序平台

RNA-Seq 所用新一代测序平台为 illumina solexa，按照该平台说明书要求进行文库制备及测序．

1.2 临床样品

用于 RNA-Seq 及新基因克隆的肝癌活检组织及正常对照肝组织样本来自中南大学湘雅三医院．活检标本的采集经中南大学伦理委员会批准，并获得病人知情同意．RNA-Seq 检测了 1 例肝癌组织和 1 例正常对照样品，后续验证使用了 10 对肝癌和癌旁组织样品．

1.3 试剂及试剂盒

琼脂糖(agrose)等常用生化试剂、胶回收试剂盒购自上海华舜生物工程有限公司；1 kb ladder DNA 分子质量 Marker 购自 Invitrogen 公司；利用 RNeasy® Mini Kit (Qiagen)抽提试剂盒抽提高质量的 RNA，SuperScript™Ⅲ(Invitrogen)试剂盒将 RNA 逆转录成 cDNA．

为了验证在 11 号染色体所检测到的各个 RNA-Seq 峰是否来自同一个 RNA 序列，我们针对各个 RNA-Seq 峰分别设计了正向(Forward，F)引物和反向(Reverse，R)引物，利用相邻峰对应引物组合进行 PCR 扩增．各引物所在的大致位置和方向见图 1c，对应的序列如下：1F，5′ CCCTTGTTTAGCCTCTGCTG 3′，2F，5′ CCTCTGTCACTGCCACAAAA 3′，2R，5′ CATGTGGTGAGTGGCTGTG 3′，3F，5′ TCACTTACCCCTGCGACTCT 3′，3R，5′ CATCACAGGGTGTGTTTTGC 3′，4F，5′ GCCCTGCTCTATCAGCAAAC 3′，4R，5′ GCCTGCTGAGTTTGCTGATA 3′，5R，5′ CCCAAACCAAGCTGACAAAT 3′．

利用 GeneRacer™ 试剂盒(Invitrogen)扩增未知 RNA 的 5′端及 3′端全长序列．针对本项目 RNA 序列用于 GeneRacer™ 反应的基因特异性引物(gene specific primer，GSP)大致位置和方向见图 1c，序列如下：Racer R1，5′ GCAGAGGCTAAACAAGGGGGTCCTG 3′，Racer F5，5′ CACTGAATGCCCCACGTCAAAGAAA 3′．

RT-PCR 或者 GeneRacer™ PCR 的产物均用 TA clone 试剂盒(Invitrogen)进行连接、转化，挑选阳性克隆用 Sanger 法测序，以获得插入片段的序列．

1.4 序列分析及生物信息学预测所用软件

新一代测序技术进行 RNA-Seq 发现的新基因，我们用传统的 Sanger 法测序进行了验证，并用 Sanger 法测序克隆了该基因的全长序列．Sanger 法测序所获得的序列用 DNAStar 软件(DNASTAR Inc.)进行组装和校对；新克隆基因潜在开放阅读框(open reading frame，ORF)采用美国国立生物信息研究所(National Center for Biotechnology Information, NCBI)开发的在线分析软件 ORF Finder (www.ncbi.nlm.nih.gov/gorf/gorf.html)进行预测；潜在开放阅读框编码的多肽序列输入 Pfam (http://pfam.sanger.ac.uk/)数据库与已知蛋白结构域是否有同源性；目前国际 DNA 元件百科全书 (encyclopedia of DNA elements, ENCODE)计划的 GENCODE 项目旨在利用计算分析、人工标注和实验验证来鉴定出人基因组中所有的基因特征．2013 年 1 月 22 日公布的最新第 15 版 GENCODE 包含 20 447 个蛋白编码基因、13 249 个 lncRNA 基因、9 173 个小的非编码基因 (small non-coding RNA genes)及 13 447 个假基因 (pseudogenes) 总共 195 433 种转录本．我们利用加州大学圣克鲁斯分校 (University of California Santa Cruz, UCSC) 的基因组在线分析工具 (genome.ucsc.edu) 将我们克隆的 lncRNA 序列与 ENCODE 计划 GENCODE 项目已注册的 lncRNA 序列进行了比较，以确证我们克隆了一个新的 lncRNA 基因．lncRNA 进化上的保守性通过 NCBI 的 Blast (blast.ncbi.nlm.nih.gov/Blast.cgi) 在线程序进行分析；lncRNA 的二级结构预测使用在线软件 FoldRNA (linux1.softberry.com/cgi-bin/programs/rnastruct/foldrna.pl)；新克隆的 lncRNA 第一外显子及其上游 2 000 bp 序列使用 Promoter Scan (www-bimas.cit.nih.gov/molbio/proscan)在线分析软件预测潜在的启动子区域；新克隆 lncRNA 潜在启动子区域存在的

转录因子结合位点使用 Genomatix 公司(www.genomatix.de)开发的在线分析软件 MatInspector 进行分析.

2　结　　果

2.1　RNA-Seq 在染色体 11q13.1 区域发现多个相邻的 RNA 信号峰

通过新一代测序技术对肝细胞癌活检组织及正常对照样品中的 RNA 进行高通量测序(RNA-Seq)，并以人类基因组参考序列(Build 37.3，Hg19)为模版对 RNA-Seq 序列进行组装，我们发现，肝癌组织中在染色体 11q13.1 区域(图 1a)距 11 号染色体短臂末端物理距离 65.6Mb 附近(图 1b)存在 6 个相邻的较明显的 RNA 信号峰(图 1c)，除第 5 个峰只有 50 个左右测序读长(reads)支持外，其余每个峰都有超过 100 个 reads 支持，最高达到近 400 个 reads 支持(图 1c 下半部分的信号峰)，而同一染色体区域正常对照样品中则没有或仅有痕量的 reads (图 1c 上半部分)，表明在所检测的肝癌组织中这一染色体区段转录出了一条或多条 RNA 序列，且转录水平较高.

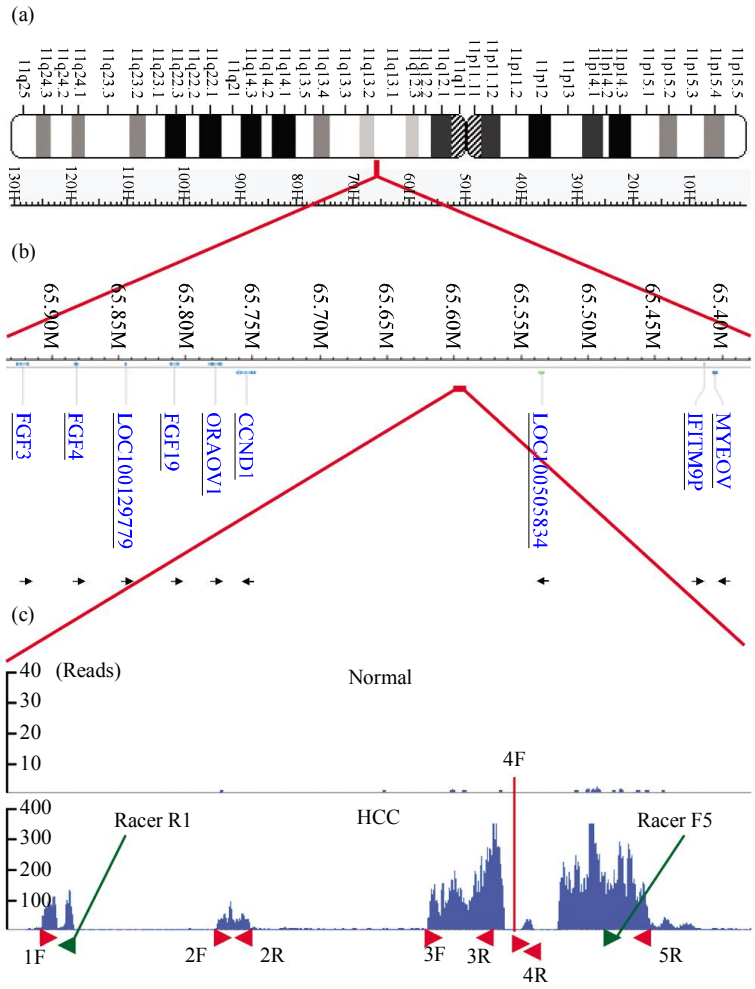

Fig. 1　RNA-Seq peaks on chromosome 11q13.1 and primers used in this study

(a) Several RNA-Seq peaks found in HCC samples located on chromosome 11q13.1. (b) RNA-Seq peaks and their adjacent genes. (c) RNA-Seq results of normal liver tissue and HCC sample, primers used in this study were also marked.

这些 RNA 信号峰所在的染色体部位没有已知基因登录(图 1b)，在它们下游(11 号染色体长臂末端方向)150 kb 左右为重要的瘤基因 *CCND1*(编码 cyclin D1 蛋白)，下游 150～350 kb 范围内还分布

着 3 个成纤维细胞生长因子(fibroblast growth factor，FGF)基因 *FGF19*、*FGF4* 和 *FGF3*，口腔肿瘤过表达基因 1 (oral cancer overexpressed 1，*ORAOV1*)以及一个假基因[DnaJ (Hsp40) homolog, subfamily B, member 7 pseudogene] *LOC100129779*；在上游约 50 kb 有一功能未知(uncharacterized)的 miscRNA 基因 *LOC100505834*，上游 190 kb 附近有一个干扰素诱导的跨膜蛋白 9 假基因(interferon induced transmembrane protein 9 pseudogene，*IFITM9P*) 及骨髓瘤过表达基因 (myeloma overexpressed，*MYEOV*). 为了验证这几个 RNA-Seq 峰是来自几个独立的 RNA 序列还是来自同一条 RNA 序列，我们针对各个 RNA-Seq 峰分别设计了正向和反向引物(图 1c，第 1、2 个峰较窄，故只分别设计了一条正向或一条反向引物)，然后以肝癌组织 cDNA 为模板，用相邻峰对应的引物配对进行 PCR 扩增.

2.2　RT-PCR 证实 11q13.1 区域多个 RNA-Seq 峰转录自同一个基因

以肝癌组织 cDNA 为模版，用引物 1F 和 2R 进行 PCR 扩增，PCR 产物经琼脂糖凝胶电泳发现扩增出多条目的带，引物 1F 和 3R 也同样得到了多条带(图 2a). 将目的条带分别割胶纯化、TA 克隆，挑选阳性克隆进行测序，发现引物 1F 和 2R 获得的 PCR 产物覆盖了第 1、2 和 3 个 RNA-Seq 峰，且存在至少 2 种转录后剪接形式，同样的，引物 1F 和 3R 扩增的 PCR 产物覆盖了第 1~4 个 RNA-Seq 峰，存在至少 3 种转录后剪接形式(图 2b).

用引物 3F 和 4R、4F 和 5R 等组合 PCR 扩增也得到了类似的结果(图 2c 和图 2d)，证实染色体 11q13.1 区段测得的这几个相邻的 RNA-Seq峰实际上转录自同一个基因，它们对应的 DNA 片段为同一个基因的不同外显子，且由 PCR 和测序结果可知该基因转录的 RNA 存在多种剪接方式. 利用引物 1F 和 5R 我们扩增了该基因更多种类型的转录本.

由于 PCR 扩增得到的是 DNA 双链，仅从测序结果我们还无法判断该 RNA 的实际方向，也还没得到该 RNA 5′和 3′端全长序列，因此我们进一步利用 GeneRacer™ 试剂盒扩增该 RNA 的 5′和 3′端全长.

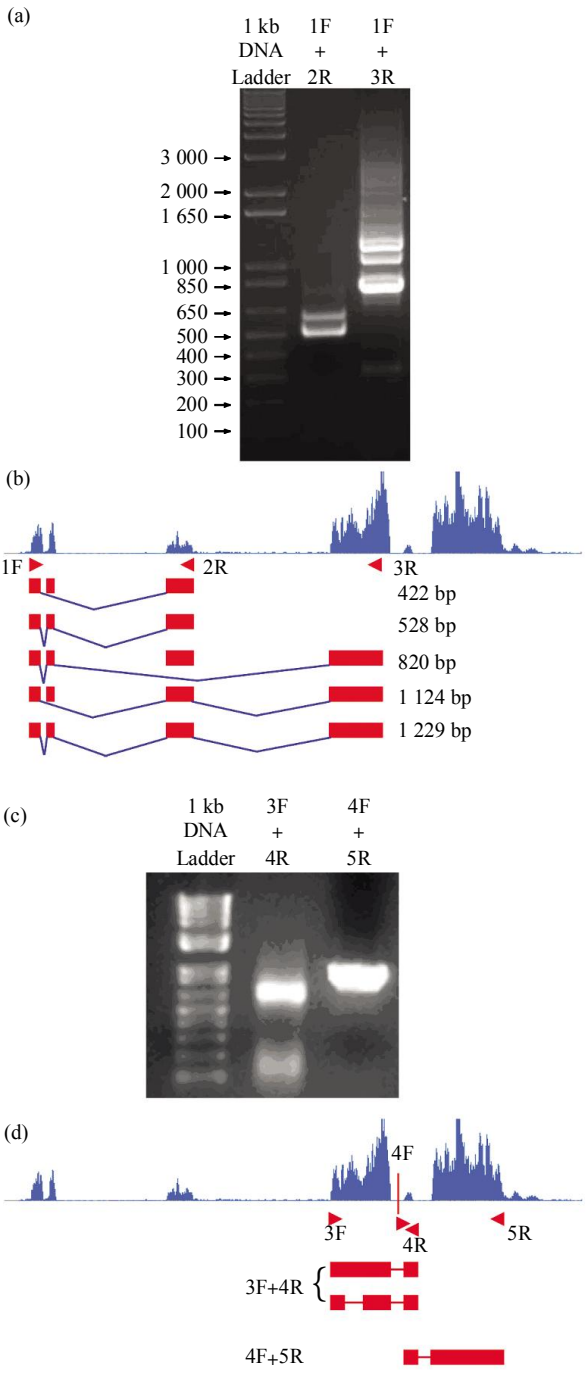

Fig. 2　RT-PCR confirmed that RNA-Seq peaks represent a novel gene

(a) Using primer pairs 1F & 2R or 1F & 3R, HCC cDNA as template, RT-PCR amplified multiple bands. (b) RT-PCR products of Figure 2a were ligated into TA-clone and sequenced, five transcript isoforms were found. (c) RT-PCR amplification results of primer pairs 3F & 4R or 4F & 5R. (d) Three transcript isoforms were found from RT-PCR products of Figure 2c.

2.3　GeneRacer™ 克隆 RNA 全长序列

GeneRacer 所用引物见图 3a，首先用烟草酸性焦磷酸酶(tobacco acid pyrophosphatase，TAP)去除 RNA 5′端帽子(cap)结构，然后在去除帽子结构的 RNA 5′端用 RNA 连接酶(RNA ligase)连接一段特异性 RNA 接头序列(图 3a 中左端黑色粗线条部分)，接下来用 5′端增加了另一段特异性接头的 oligo-dT 为引物(图 3a 中右端以 TTTTT-NNNNN 表示)，对 RNA 进行逆转录，获得 cDNA 的第一条链，此时 cDNA 就在两头都分别加上了两段特异性的接头序列．针对这两端的接头序列，试剂盒分别提供了 5′和 3′端的 GeneRacer 公共引物，再在我们感兴趣的目的基因序列中设计基因特异性引物(gene specific primer，GSP)与公共的 5′或 3′GeneRacer 引物组合进行 PCR 扩增得到感兴趣的基因 5′和 3′端全长序列．将肝癌组织 RNA 用 GeneRacer 试剂盒逆转录的 cDNA(两头已加接头)为模板，利用图 1c 中设计的 2 条基因特异性 GeneRacer 引物 Racer R1 和 Racer F5 与试剂盒中的 GeneRacer 3′公共引物进行 PCR 扩增(图 3b)，结果 Racer R1 与 GeneRacer 3′公共引物未能扩增出目的条带，而 Racer F5 与 GeneRacer 3′引物则扩增出了非常清楚的目的条带，表明 Racer F5 靠近目的基因的 3′端，其对应的图 1c 中第 6 个 RNA-Seq 峰就是该新基因的第 6 号外显子．利用 GeneRacer 试剂盒中提供的 HeLa 细胞 cDNA 为模板，Racer F5 与 GeneRacer 3′引物也未扩增出特异性条带，表明该基因在 HeLa 细胞中不表达或者表达很弱．图 3b 中 Racer F5 与 GeneRacer 3′公共引物扩增产物电泳结果除了有一条很强的目的条带外，在该条带上方还有两条较弱的条带(图中箭头所示)，表明该基因 3′端可能同样存在可变剪接．将 PCR 产物 TA 克隆，Sanger 法测序，得到了该新基因 3′端 3 种不同转录本的序列(图 3c 右边部分)．接下来用 Racer R1 与 GeneRacer 5′公共引物进行扩增，获得了该新基因 5′端 6 种不同转录本的序列(图 3c 左边部分)．

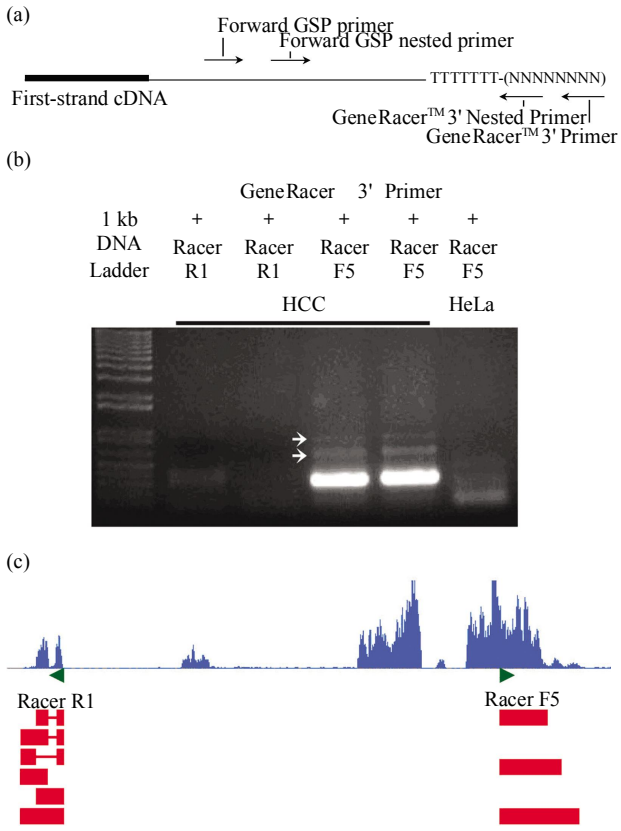

Fig. 3　Full length of novel gene′s 5′ & 3′ sequence were cloned using GeneRacer kit

(a) Sketch of GeneRacer experiment designment. (b) 3' GeneRacer PCR result. (c) Sequencing of 5' & 3' GeneRacer PCR products got variable transcripts of the novel gene.

2.4　新克隆的基因有多个转录本且没有明显的开放阅读框

在其他肝癌组织样本中进一步大量 PCR、TA 克隆和测序分析，得到了该基因更多的转录本序列，图 4a 是在肝癌组织中表达频率比较高的 12 种转录本剪接模式，我们已将这 12 条代表性的转录本序列递交到美国国立生物技术信息中心(NCBI)的 GenBank 数据库中，GenBank ID 号分别为 KC136297～KC136308．将该基因最长的转录本(图 4a 中的 isoform-1，3 526 bp)序列输入 NCBI 的 ORF finder 软件进行开放阅读框预测，发现最长的开放阅读框仅有 429 bp，预测的蛋白仅 142 个氨基酸残基(图 4b)，其他 11 条代表性转录本预测出的开放阅读框最长的也仅 570 bp，预测的蛋白仅 189 个氨基酸残基(isoform-5，结果未显示)．将所有潜在 ORF 编码的多肽序列与 Pfam 数据库中已知的蛋白质或者结构域进行比对，未发现有同源性．提示该基因可能编码一个长链非编码 RNA (long non-coding RNA，lncRNA).

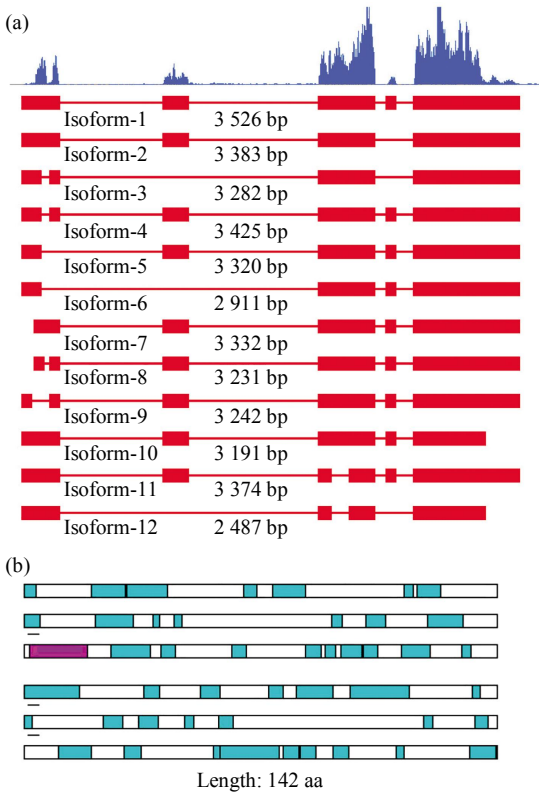

Fig. 4　Representative transcripts and predicted open reading fragments of the novel gene

(a) 12 representative transcripts of the novel HCC associated gene were cloned and deposited into the GenBank database, their GenBank IDs were from KC136297 to KC136308. (b) The open reading fragment (ORF) of the longest transcript, Isoform-1 in (a), was predicted by ORF Finder, the longest ORF was only 142 amino acids.

2.5　新克隆的 lncRNA 在 ENCODE 数据库中未发现同源基因

通过 UCSC 的基因组在线分析工具 (genome.ucsc.edu) 将我们新克隆的 lncRNA 序列与 ENCODE 数据库中的 GENCODE (Version 15.0)进行了比对，在该染色体区域无论是蛋白质编码基因还是非编码基因都没有已知基因登录 (图 5)，进一步证实了我们克隆的是一个新基因.

2.6　新克隆的 lncRNA 进化保守性分析

对新克隆的 lncRNA 在不同物种中进行了同源性和进化上的保守性分析. 发现该 RNA 序列仅在黑猩猩(chimpanzee，学名 *Pan troglodytes*，同源性高达 98%)、苏门答腊猩猩(Sumatran orangutan, *Pongo abelii*，同源性 95%)、猕猴(rhesus monkey，*Macaca mulatta*，同源性 91%)、北方白颊长臂猿(northern white-cheeked gibbon, *Nomascus leucogenys*，994～1 389 bp 未匹配到同源序列，其他部分同源性 95%)、白耳狨猴 (white-tufted-ear marmoset，*Callithrix jacchus*，仅 1～374，1 806～2 083 和 2 833～3 187 bp 三段序列同源性约 86%)等灵长类动物中发现了同源序列(图 6)，而在猪(*Sus scrofa*)、狗 (*Canis lupus familiaris*)、牛 (*Bos taurus*)、绵羊 (*Ovis aries*)、马(*Equus caballus*)、家猫(*Felis catus*)、兔(*Oryctolagus cuniculus*)、大鼠(*Rattus norvegicus*)、小鼠(*Mus musculus*)、中国仓鼠(*Cricetulus griseus*)、

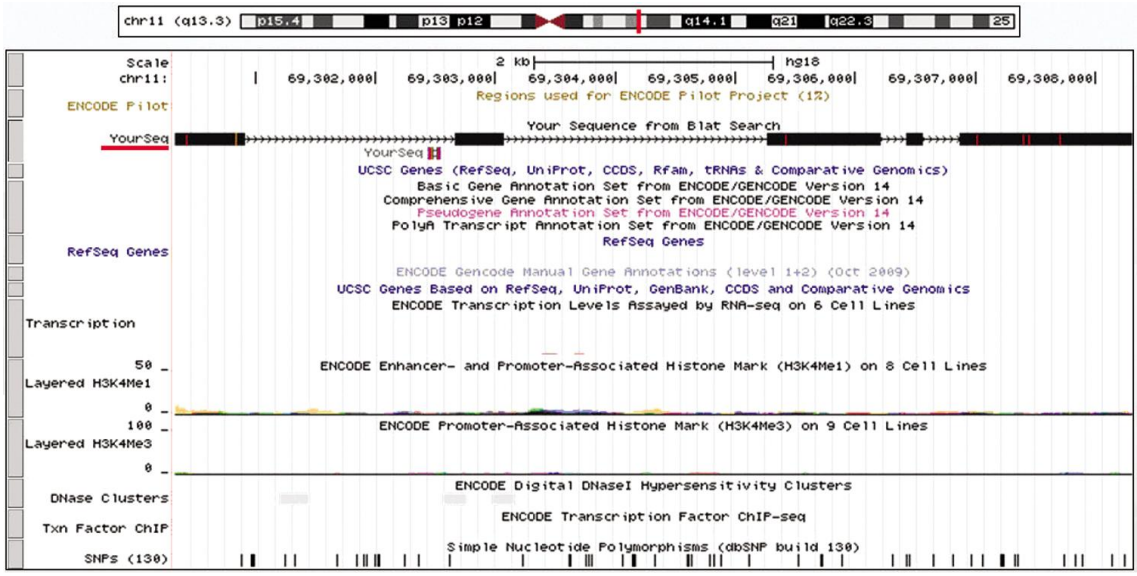

Fig. 5　Mapping the novel lncRNA gene onto the ENCODE database

Using USCS genomic online analysis tools, the longest isoform of the novel lncRNA, KC136297, was mapped onto the ENCODE database, "Your seq" indicated 5 exons of the novel lncRNA gene, there was none known gene at this chromosome region registered in the ENCODE database.

灰短尾负鼠(*Monodelphis domestica*)、大熊猫(*Ailuropoda melanoleuca*)及鸭嘴兽(*Ornithorhynchus anatinus*)等哺乳动物中均未发现有同源序列，表明该基因在进化上不保守.

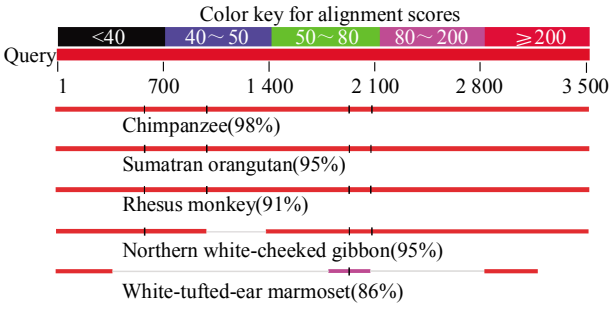

Fig. 6　The homology genes of the novel lncRNA were only found in some primates

2.7　新克隆的 lncRNA 二级结构预测

长链非编码 RNA 可折叠成许多有功能的二级结构而发挥功能. 因此我们用 FoldRNA 在线分析软件预测了图 4a 中新克隆的 lncRNA 全部 12 种代表性转录本的二级结构，发现它们均有着相似的二级结构(图 S1，见网络版附件)，即折叠成 3 个主要的分支，呈三叶草样，相对于最长的转录本(Isoform-1，图 S1a)，其他不同剪接型式缺失的 RNA 片段对该 lncRNA　二级结构的维持没有显著影响(代表性预测结果见图 S1b 和 c).

在对新克隆的 lncRNA 进行了同源性和进化上的保守性分析时，我们发现，该 lncRNA 第 1～374，1 806～2 083 和 2 833～3 187 bp 三段序列相对保守 性较高，我们进一步预测了这三段序列的缺失对 该 lncRNA 二级结构的影响，发现单独缺失 2 833～3 187 bp 序列，lncRNA 二级结构没有明显变化(结果未显示)，而单独缺失 1～374(图 S1d)和 1 806～2 083 bp(图 S1e)则可明显改变该 lncRNA 的二级结构，特别是同时缺失 1～374，1 806～2 083 和 2 833～3 187 bp 序列时该 lncRNA 二级结构改变更为明显(图 S1f).

2.8　新克隆的 lncRNA 启动子预测和转录结合位点分析

为了探讨该 lncRNA 在肝细胞癌中表达调控的可能机制，我们进一步预测了该 lncRNA 基因上游启动子序列. 将该 lncRNA 第一号外显子及其上游 2 000 bp 序列输入 Promoter Scan 在线分析软件，发现其转录起始位点上游–719～–469 bp 处有一个潜在的启动子(图 7)中绿色标记，预测分值为 91.46，该区域存在 7 个 Sp1 转录因子结合位点(图 7 中用橙色标记)，此外启动子区域还存在 1 个 STAT5 (图 7 中用红色标记位点)和 1 个 EGR1(图 7 中用紫色标记)结合位点.

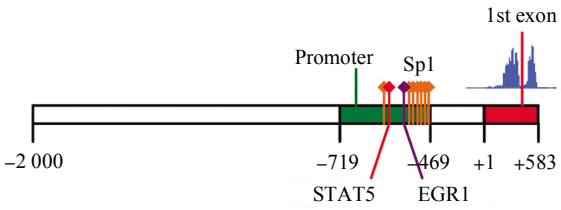

Fig. 7　Some important transcription factor binding sites on the novel lncRNA promoter

A 2000 bp sequence up-stream from the transcript initiation site of the novel lncRNA gene was used for predicting potential promoter and transcription factor binding sites, there is a potential promoter from –719 to –469 bp, and there are 1 STAT5, 1 EGR1, 7 Sp1 binding sites in the potential promoter region.

3　讨　　论

LncRNAs 是一类长度超过 200nt，缺少特异性的完整开放阅读框，没有或很少有蛋白质编码功能的非编码 RNAs 分子[2-5]，在真核生物转录组中占有非常大的比例. 最初 lncRNAs 被认为是 RNA 聚合酶Ⅱ转录的副产物，是基因组转录的“噪音”，不具有生物学功能，但近年的研究结果表明，lncRNAs 与 miRNAs[6-7]一样广泛参与了细胞内的基因表达调控，具有重要的生物学功能. 大多数目前已经发现的 lncRNAs 由 RNA 聚合酶Ⅱ转录、经可变剪接形成，并通常被多聚腺苷酸(poly A)化. 本文中我们所克隆的 lncRNA 可被 GeneRacer 试剂盒获得 5′和 3′端全长序列，表明它和 mRNA 一样，具有 5′帽子(cap)和 3′端 poly A 尾结构.

LncRNAs 在表观遗传、转录及转录后水平等多个层面调节基因的表达：有的 lncRNAs 通过介导染色质重塑和组蛋白修饰干扰转录；有的 lncRNA 通过调节选择性剪接模式，生成小 RNAs，调节蛋白质活性，改变蛋白质定位等方式发挥功能；还有的 lncRNAs 通过与 miRNAs 相互作用，竞争性抑制 miRNAs 与靶 mRNAs 结合的能力，调节基因表达[8-11]. lncRNAs 通过上述多种途径参与

机体生长、发育、衰老及死亡等重要生命活动过程的调控，因而越来越受到关注[12-16]. LncRNAs比mRNAs和小分子RNAs的序列保守性较差，承受的进化压力较小，但有的lncRNAs某些区域保守性相对较高，这些保守元件的特定核苷酸序列对维持其二级结构的稳定或对lncRNAs功能的发挥可能具有重要作用. 已有资料表明，lncRNAs可通过折叠成具有功能的二级结构，与细胞内相关蛋白质和核酸结合从而发挥调控功能，与蛋白质相比，lncRNAs空间结构稳定性较差，可快速产生和分解，为有机体提供更加敏感的调节[17].

我国学者在肝癌相关lncRNA方面已经获得了一些重要的科学发现，发表了一系列高水平研究论文[18-22]，本课题组通过RNA-Seq技术在染色体11q13.1区域发现并随后克隆了一个新的lncRNA，该lncRNA在正常对照肝组织中没有表达，而在肝癌组织中高表达，表明该lncRNA在肝细胞癌发病过程中有可能发挥类似癌基因(oncogene)的作用. 该lncRNA没有明显的开放阅读框，存在多个转录本，且不同转录本均预测出了相似的二级结构. 另外，该lncRNA且在进化上保守性差，但其第1～374、1 806～2 083和2 833～3 187 bp三段序列的保守性相对较高，生物信息学预测其中第1～374及1 806～2 083 bp，特别是1 806～2 083 bp序列，对该lncRNAs二级结构的维持具有较重要的作用. 该lncRNA有可能通过形成特定的二级结构，与细胞内其他蛋白质结合而发挥生物学功能. 本文克隆的lncRNA在肝细胞癌中如何参与细胞内的基因表达调控及如何影响其他蛋白质的功能，该lncRNA发挥生物学功能是否依赖于它折叠产生的构象，1 806～2 083 bp序列对该lncRNA生物学功能的影响等，均值得进一步深入研究.

染色体11q13.1是在多种肿瘤中常见的染色体扩增区域，也是一个瘤基因(oncogene)富集区[23]，在我们克隆的lncRNA基因附近就分布着几个重要的瘤基因如*CCND1*[24-25]、*ORAOV1*[26]、*MYEOV*[27]、*FGF19*[28]、*FGF4*和*FGF3*[29]等，最近的研究表明lncRNA有可能影响其临近基因的表达调控，我们新克隆的lncRNA对其相邻基因的表达调控是否有影响也值得深入探讨. 另外，染色体11q13.1在肝细胞癌中存在高频扩增，可能会影响到该染色体区域中包括我们新克隆的lncRNA在内相关基因的表达上调，但新克隆的lncRNA在肝细胞癌中表达上调是否还存在其他途径？我们通过启动子分析和预测，发现该lncRNA上游–719～–469 bp有一个潜在的启动子，启动子区域存在Sp1[30]、STAT5[31-32]及EGR1[33]等转录因子结合位点，这些转录因子及其相关的信号通路是否参与了该lncRNA的表达调控也是一个新的科学问题. 总之，深入探讨该lncRNA在肝细胞癌发生发展过程中的作用及其机制，将为进一步了解肝细胞癌的发病机制，为肝癌的临床诊断、治疗和预后判断等提供新的线索.

附件 图S1，见论文网络版 http:// www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120613&flag=1

参 考 文 献

[1] Qing T, Yu Y, Du T T, *et al*. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci, 2013, **56**(2): 134-142

[2] Gong Z, Zhang S, Zhang W, *et al*. Long non-coding RNAs in cancer. Sci China Life Sci, 2012, **55**(12): 1120-1124

[3] Ponting C P, Oliver P L, Reik W. Evolution and functions of long noncoding RNAs. Cell, 2009, **136**(4): 629-641

[4] Zhang W, Huang C, Gong Z, *et al*. Expression of LINC00312, a long intergenic non-coding RNA, is negatively correlated with tumor size but positively correlated with lymph node metastasis in nasopharyngeal carcinoma. J Mol Histol, 2013, **44**(5): 545-554 (DOI: 10.1007/s11427-013-4577y)

[5] He S, Liu C, Skogerbo G, *et al*. NONCODE v2.0: decoding the non-coding. Nucleic Acids Res, 2008, **36**(Database issue): D170-172

[6] 龚朝建, 黄宏斌, 徐 柯, 等. microRNAs与TP53基因调控网络研究进展. 生物化学与生物物理进展, 2012, **39**(12): 1133-1144
Gong Z J, Huang H B, Xu K, *et al*. Prog Biochem Biophys, 2012, **39**(12): 1133-1144

[7] Zeng Z Y, Huang H B, Huang L L, *et al*. Expression profiles and regulation network of Epstein-Barr virus-encoded microRNAs and their potential target host genes in nasopharyngeal carcinoma. Sci China Life Sci, 2014, **57** (DOI:10.1007/s11427-013-4577y)

[8] Wilusz J E, Sunwoo H, Spector D L. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev, 2009, **23**(13): 1494-1504

[9] Tsai M C, Manor O, Wan Y, *et al*. Long noncoding RNA as modular scaffold of histone modification complexes. Science, 2010, **329**(5992): 689-693

[10] 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**(7291): 1071-1076

[11] He S, Su H, Liu C, *et al*. MicroRNA-encoding long non-coding RNAs. BMC Genomics, 2008, **9**: 236

[12] Li L, Feng T, Lian Y, *et al*. Role of human noncoding RNAs in the control of tumorigenesis. Proc Natl Acad Sci USA, 2009, **106**(31):

12956−12961

[13] Gutschner T, Diederichs S. The Hallmarks of Cancer: a long non-coding RNA point of view. RNA Biol, 2012, **9**(6): 1076−1087

[14] 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**(Web Server issue): W118−124

[15] Guo X, Gao L, Liao Q, *et al*. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res, 2013, **42**(2): e35

[16] Li A, Wei G, Wang Y, *et al*. Identification of intermediate-size non-coding RNAs involved in the UV-induced DNA damage response in *C. elegans*. PLoS One, 2012, **7**(11): e48066

[17] Pang K C, Frith M C, Mattick J S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet, 2006, **22**(1): 1−5

[18] Yang Z, Zhou L, Wu L M, *et al*. Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. Ann Surg Oncol, 2011, **18**(5): 1243−1250

[19] Yuan S X, Yang F, Yang Y, *et al*. Long noncoding RNA associated with microvascular invasion in hepatocellular carcinoma promotes angiogenesis and serves as a predictor for hepatocellular carcinoma patients' poor recurrence-free survival after hepatectomy. Hepatology, 2012, **56**(6): 2231−2241

[20] Zhu Z, Gao X, He Y, *et al*. An insertion/deletion polymorphism within RERT-lncRNA modulates hepatocellular carcinoma risk. Cancer Res, 2012, **72**(23): 6163−6172

[21] Yang F, Zhang L, Huo X S, *et al*. Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. Hepatology, 2011, **54**(5): 1679−1689

[22] Huang J F, Guo Y J, Zhao C X, *et al*. HBx-related lncRNA Dreh inhibits hepatocellular carcinoma metastasis by targeting the intermediate filament protein vimentin. Hepatology, 2013, **57**(5): 1882−1892

[23] Ying J, Shan L, Li J, *et al*. Genome-wide screening for genetic alterations in esophageal cancer by aCGH identifies 11q13 amplification oncogenes associated with nodal metastasis. PLoS One, 2012, **7**(6): e39797

[24] Che Y, Ye F, Xu R, *et al*. Co-expression of XIAP and cyclin D1 complex correlates with a poor prognosis in patients with hepatocellular carcinoma. Am J Pathol, 2012, **180**(5): 1798−1807

[25] Yang Y, Liao Q, Wei F, *et al*. LPLUNC1 inhibits nasopharyngeal carcinoma cell growth *via* down-regulation of the MAP kinase and cyclin D1/E2F pathways. PLoS One, 2013, **8**(5): e62869

[26] Xavier F C, Rodini C O, Paiva K B, *et al*. ORAOV1 is amplified in oral squamous cell carcinoma. J Oral Pathol Med, 2011, **41**(1): 54−60

[27] Takita J, Chen Y, Okubo J, *et al*. Aberrations of NEGR1 on 1p31 and MYEOV on 11q13 in neuroblastoma. Cancer Sci, 2011, **102**(9): 1645−1650

[28] Sawey E T, Chanrion M, Cai C, *et al*. Identification of a therapeutic strategy targeting amplified FGF19 in liver cancer by Oncogenomic screening. Cancer Cell, 2011, **19**(3): 347−358

[29] Arao T, Ueshima K, Matsumoto K, *et al*. FGF3/FGF4 amplification and multiple lung metastases in responders to sorafenib in hepatocellular carcinoma. Hepatology, 2013, **57**(4): 1407−1415

[30] Yin P, Zhao C, Li Z, *et al*. Sp1 is involved in regulation of cystathionine gamma-lyase gene expression and biological function by PI3K/Akt pathway in human hepatocellular carcinoma cell lines. Cell Signal, 2012, **24**(6): 1229−1240

[31] Yu J H, Zhu B M, Riedlinger G, *et al*. The liver-specific tumor suppressor STAT5 controls expression of the reactive oxygen species-generating enzyme NOX4 and the proapoptotic proteins PUMA and BIM in mice. Hepatology, 2012, **56**(6): 2375−2386

[32] Liao Q, Zeng Z, Guo X, *et al*. LPLUNC1 suppresses IL-6-induced nasopharyngeal carcinoma cell proliferation via inhibiting the Stat3 activation. Oncogene, 2013(DOI: 10.1038/onc.2013.161)

[33] Zwang Y, Oren M, Yarden Y. Consistency test of the cell cycle: roles for p53 and EGR1. Cancer Res, 2012, **72**(5): 1051−1054

Cloning and Functional Characterization of a Novel Long Non-coding RNA Gene Associated With Hepatocellular Carcinoma*

TANG Ke[1,2)**], WEI Fang[2)**], BO Hao[2)], HUANG Hong-Bin[2,4)], ZHANG Wen-Ling[2)], GONG Zhao-Jian[2)], LI Xia-Yu[3)], SONG Ya-Li[2)], LIAO Qian-Jin[1)], PENG Shu-Ping[2)], XIANG Juan-Juan[2)], ZHOU Ming[2)], MA Jian[2)], LI Xiao-Ling[2)], XIONG Wei[1,2,3)], LI Yong[5)], ZENG Zhao-Yang[1,2,3)***], LI Gui-Yuan[1,2,3)***]

([1)] *Hunan Provincial Tumor Hospital and the Affiliated Tumor Hospital of Xiangya School of Medicine, Central South University, Changsha* 410011, China;

[2)] *Key Laboratory of Carcinogenesis of Ministry of Health and Key Laboratory of Carcinogenesis and Cancer Invasion of Ministry of Education, Cancer Research Institute, Central South University, Changsha* 410078, China;

[3)] *Hunan Key Laboratory of Nonresolving Inflammation and Cancer, Disease Genome Research Center, The Third Xiangya Hospital, Central South University, Changsha* 410013, China;

[4)] *Key Laboratory of Information System Engineering, National University of Defense Technology, Changsha* 410073, China;

[5)] *Department of Biochemistry and Molecular Biology, Center for Genetics and Molecular Medicine, School of Medicine, University of Louisville, Louisville, KY* 40202, USA)

Abstract Recently, we sequenced the transcriptomes of a hepatocellular carcinoma biopsy and a normal liver tissue using the RNA-Sequencing (RNA-Seq) strategy based on the Next Generation Sequencing (NGS) technique, and identified several adjacent high RNA-Seq signal peaks on chromosome 11q13.1 in the hepatocellular carcinoma biopsy, while not in the normal control tissue. In this chromosome region, there is no characterized genes have been identified, implying that these RNA-Seq peaks may represent one or more novel genes. Further study was confirmed that these RNA-Seq peaks were transcribed by one novel gene. Through cloning the full length of this novel gene, we found that this novel gene transcribed many splicing isoforms, and the longest isoform is 3 562 bp. Then we deposited twelve representative RNA isoforms into the GenBank database of the National Center for Biotechnology Information (NCBI), and created the GenBank IDs from KC136297 to KC136308 for these isoforms. None significant open reading fragment (ORF) was found in any transcripts of this novel gene, implying that this gene may encodes long non-coding RNAs (lncRNAs). To further elucidate the potential transcriptional regulation mechanism of this lncRNA gene, we predicted the promoter from the upstream sequence of the lncRNA gene using bioinformatic tools, and found that there is one potential promoter in –719 to –469 bp from the transcript start site of the lncRNA gene, and there are seven Sp1, one STAT5 and one EGR1 transcription factor binding sites in the promoter region. The molecular mechanisms of the lncRNA gene in carcinogenesis and progression of hepatocellular carcinoma are worthful for further investigation.

Key words long non-coding RNA, hepatocellular carcinoma, chromosome 11q13.1, gene cloning, bioinformatics
DOI: 10.3724/SP.J.1206.2012.00613

*This work was supported by grants from The National Natural Science Foundation of China (81372907, 81172189, 81171930, 81272298, 81272255, 91229122), The Hunan Province Natural Sciences Foundation of China (14JJ1010), The Fok Ying Tong Education Foundation (121036), the Fundamental Research Funds for the Central Universities (2011JQ020) and the Postdoctoral Science Foundation of Central South University.

**These authors contributed equally to this work.
***Corresponding author. Tel: 86-731-84805383
ZENG Zhao-Yang. E-mail: zengzhaoyang@xysm.net
LI Gui-Yuan. E-mail: ligy@xysm.net

Received: March 20, 2013 Accepted: July 11, 2013

附　　录

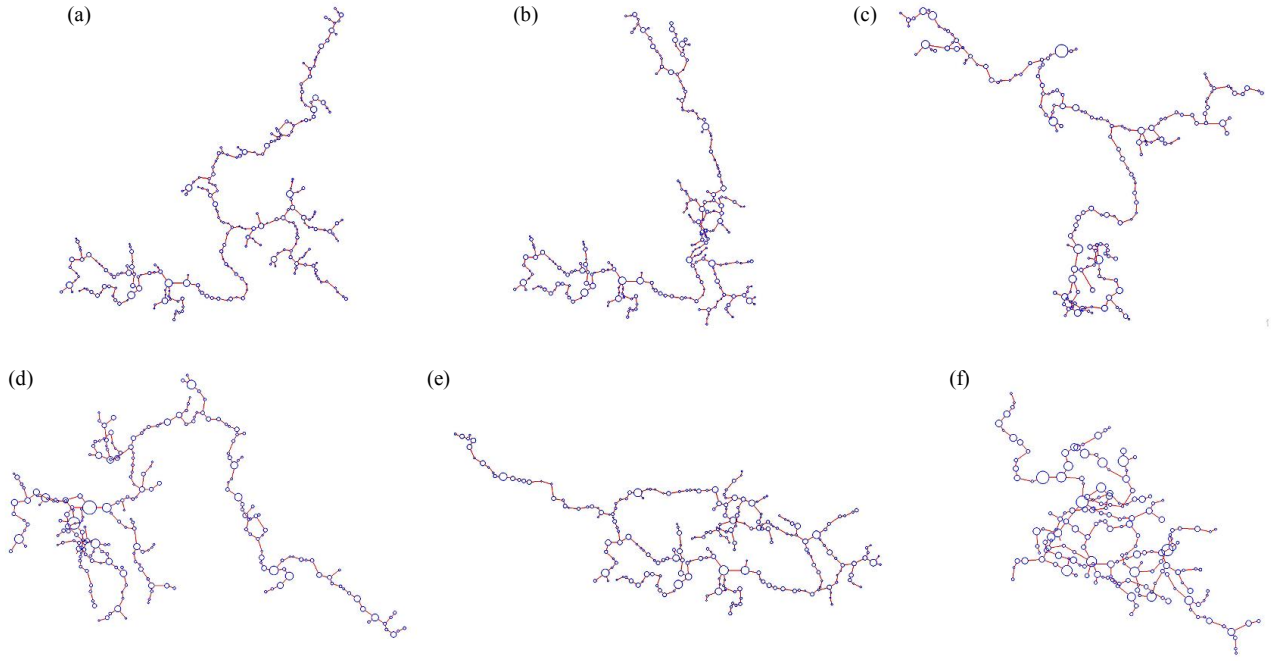

Fig. S1　The predicted secondary structure of the novel lncRNA

(a) Isoform-1. (b) Isoform-2. (c) Isoform-12. (d) Isoform-1, deleted 1～374 bp. (e) Isoform-1, deleted 1 806～2 083 bp. (f) Isoform-1, deleted 1～374, 1 806～2 083 and 2 833～3 187 bp.