Silicon Cloning for the Unknown Functional Genes Related to Esophageal Carcinoma and ncRNA Finding

WU Bing-li[1,3], XU Li-yan[2], YING Xiao-min[3], NIU Yong-dong[1], LI Wu-ju[3,*], LI En-min[1,*]
(*1. Biochemistry and Molecular Biology Department, Medical College of Shantou University, Shantou 515041; 2. Department of Tumor Pathology, Medical College of Shantou University, Shantou 515041; 3. Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850*, China)

食管癌相关功能未知基因的电子克隆延伸与ncRNA的发现

吴炳礼[1,3]/许丽艳[2]/应晓敏[3]/牛永东[1]/李伍举[3,*]/李恩民[1,*]
(1. 汕头大学医学院生化与分子生物学教研室，汕头 515041；2. 汕头大学医学院肿瘤病理研究室，汕头 515041；3. 军事医学科学院基础医学研究所，北京 100850)

【摘要】背景与目的：运用电子克隆等生物信息学方法研究筛查出的48个与食管癌相关功能未知的DNA序列片段，为食管癌相关研究提供指导。**材料与方法：**以48个DNA序列片段为核心，运用BioEdit建立本地数据库；通过电子克隆的方法对48个DNA序列中功能未知的基因片段进行序列延伸；通过BLAST同源分析搜索48个基因的内含子以及上下游基因间隔区中存在的非编码RNA(noncoding RNA, ncRNA)。**结果：**48个DNA序列中功能未知的基因片段通过电子克隆的方法平均能够延伸190 bp以上；在48个基因的内含子以及上下游基因间隔区存在着与已知ncRNA相似性很高的片段。**结论：**运用电子克隆的方法可以使某些食管癌相关功能未知基因的序列得以明显延伸；一些食管癌相关基因所在的染色体区段存在着某些与ncRNA高度相似的片段，这提示我们，ncRNA可能参与食管癌的发生过程，其具体功能有待深入研究。

【关键词】生物信息学；序列分析；非编码RNA；电子克隆
中图分类号：R818.03 **文献标识码：**A **文章编号：**1004-616X(2008)02-0085-04

[ABSTRACT] **BACKGROUND AND AIM:** In order to provide further support for the experimental verification and functional exploration of 48 DNA sequences related to esophageal carcinoma found in our lab, these sequences were investigated systematically using bioinformatics methods such as silicon cloning and BLAST search tool. **MATERIALS AND METHODS:** A local database including the 48 sequences was constructed by the BioEdit software; the 48 DNA sequences coding unknown genes were extended by the method of silicon cloning; the potential ncRNAs, located in the introns, the upstream and downstream of intergenic regions of the above 48 DNA sequences, were analyzed by BLAST tool. **RESULTS:** The sequences coding unknown genes were extended to more than 190 bp on average using silicon cloning. Some sequence segments with high similarity to the known noncoding RNA(ncRNAs) were found in the intronic or intergenic regions. **CONCLUSION:** Some esophageal carcinoma related genes can be extended obviously by the method of silicon cloning, some sequences with high similarity to known ncRNAs were found in the chromosome region where the esophageal carcinoma related genes are located. The information gives us potential clues that these ncRNAs maybe participate in the development process of esophageal carcinoma, and the functions of these ncRNAs needs further study.

[KEY WORDS] *bioinformatics; sequence analysis; noncoding RNA; silicon cloning*

在以往的两次食管癌相关基因的筛查中，本研究小组曾获得48个DNA序列片段，其中17个来自于食管癌细胞差异表达基因cDNA寡核苷酸芯片分析[1]，另31个来自于食管癌相关基因转录激活元件结合蛋白酵母单杂交筛选实验[2-3]。因此，我们通过生物信息学的方法研究上述48个食管癌相关基因DNA序列片段中所

收稿日期：2007-10-25 **修回日期：**2007-11-10
基金项目：国家自然科学基金项目(30370641, 30570829, 30672376)，广东省自然科学基金重点项目(05104541, 7118419)，教育部高等学校博士点重点学科专项基金项目(20050560002, 20050560003)
作者简介：吴炳礼(1979-)，男，硕士，研究方向：生物信息学。
* Correspondence to: LI En-min, Tel: 0754-8900847, E-mail: nmli@stu.edu.cn; LI Wu-ju, E-mail: wujuli@yahoo.com

蕴藏的结构信息,以期对它们的生物医学功能的深入认识,特别是与食管癌的关系提供指导。

1 材料与方法

1.1 序列片段 来自于汕头大学医学院生物化学与分子生物学教研室以往的两次食管癌相关基因筛查实验,共48个DNA序列片段。

1.2 网络数据库的使用 在NCBI网站 http://www.ncbi.nlm.nih.gov/BLAST/ 使用BLAST程序进行序列同源性分析,采用默认参数[4-5]。用UCSC数据库 http://genome.ucsc.edu/ 的Blat[6]程序分析、界定并获得每个已知基因的内含子以及与上下游基因间区的序列及其在人类基因组中的具体位置。

1.3 BioEdit软件的使用 从 http://www.mbio.ncsu.edu/BioEdit/bioedit.html 下载并安装BioEdit;再从 http://www.noncode.org/搜索中科院计算所生物信息学实验室建立的非编码RNA数据库NONCODE并下载该数据库中所搜集的所有ncRNA序列,再用BioEdit软件将其格式化,建立本地数据库;然后使用BioEdit软件的BLAST程序对已知基因的内含子及其上下游基因间区与ncRNA进行同源分析,采用默认参数。

1.4 电子克隆 电子克隆(silicon cloning)也称电子延伸,是以目的表达序列标签(expression sequence tag, EST)作为种子序列,在EST数据库进行同源性搜索,选择高度同源的EST进行序列拼装,构建序列重叠群(contig),这相当于实验中的cDNA扩增,再以此重叠群为种子序列重复进行同源性搜索、拼接直至序列不能再被延长,以获得部分乃至全长的cDNA序列。电子克隆原理示意图如图1所示。在本文中电子克隆操作如上述进行,所用网页服务器为 http:// pbil.univ-lyon1.fr/cap3.php [7]。延伸后的序列仍用Blat程序进行染色体定位,观察延伸的长度。

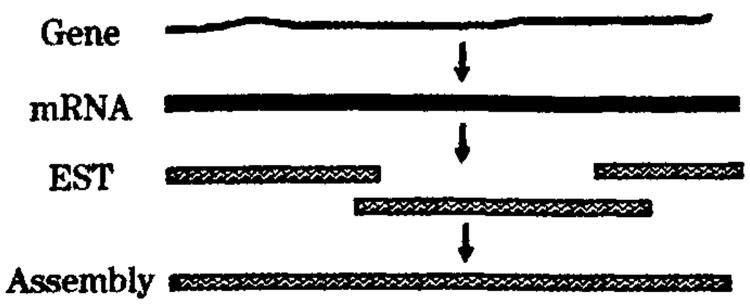

图1 序列拼接的原理示意图
Figure 1 The schematic representation of sequence assembly

2 结 果

2.1 网络数据库分析结果 经过NCBI和UCSC两个数据库相关程序的分析,48个目标序列编码基因中大部分为蛋白编码基因,这些编码基因及其相应的上下游基因在染色体上的具体物理位置见表1,为2006年5月发布的人类基因组NCBI Build 36.1版本。其中:①21、31和37为功能未知基因的序列;②3、15和47为杂合序列,由定位在不同染色体上的两条序列构成,反映出在起源细胞中可能存在着相应染色体易位现象;③19、20、28和34同为 *PPEF2* 基因的序列;④26和35同为 *PSMB7* 基因的序列;⑤25、33、44和48同为18s rRNA基因的序列;⑥18、24、27、36、38、39、40和42同为28s rRNA基因的序列。

表1 序列的网络数据库分析结果
Table 1 Sequence analysis results from database in the internet

Sequences	Chromosomes	Upstream genes(positions)	The sequence coding genes(positions)	Downstream genes(positions)
1	9	PTGES2(129,922,793 - 129,930,295)	NGAL(129,951,553 - 129,955,555)	C9orf16(129,962,360 - 129,966,028)
2	6	ITPR3(33,697,322 - 33,772,317)	C6orf125(33,773,324 - 33,787,482)	IHPK3(33,797,421 - 33,822,660)
3	-	-	chimeric	-
4	6	CMAH(25,213,948 - 25,274,742)	LRRC16(25,387,756 - 25,728,737)	SCGN(25,760,408 - 25,809,987)
5	12	LYZ(68,028,431 - 68,034,280)	YEATS4(68,039,799 - 68,070,843)	FRS2(68,150,396 - 68,259,829)
6	4	USP53(120,353,230 - 120,436,121)	LOC401152(120,438,237 - 120,441,336)	FABP2(120,457,854 - 120,462,766)
7	21	GABPA(26,028,752 - 26,066,212)	APP(26,174,732 - 26,465,003)	CYYR1(26,760,399 - 26,867,452)
8	2	CYR23B1(72,209,875 - 72,228,471)	SEC15B(72,542,113 - 72,906,662)	SPR(72,968,056 - 72,972,794)
9	9	KLF9(72,189,335 - 72,219,393)	TRPM3(72,339,786 - 72,926,334)	TMEM2(73,488,103 - 73,573,228)
10	5	IL17B(148,734,025 - 148,739,031)	CSNK1A(148,855,038 - 148,911,200)	FLJ41603(148,941,328 - 148,994,720)
11	8	VCIP135(67,705,042 - 67,742,006)	FLJ11267(67,742,406 - 67,755,789)	PTTG3(67,842,186 - 67,842,794)
12	2	AK128219(18,998,650 - 19,001,974)	DB018395(19,031,210 - 19.409,991)	BC025712(19,414,728 - 19,421,853)
13	1	FCER1G(157,998,160 - 158,002,111)	APOA2(158,005,156 - 158,006,491)	FLJ12770(158,008,926 - 158,026,160)
14	2	A4GALT(41,412,626 - 41,441,374)	ARFGAP3(41,517,031 - 41,577,798)	PACSIN2(41,590,275 - 41,735,649)
15	-	-	chimeric	-
16	5	COL4A3BP(74,710,821 - 74,843,196)	POLK(74,843,337 - 74,930,889)	FLJ35779(74,006,012 - 75,044,006)
17	16	LOC55565(70,451,090 - 70,474,913)	KIAA01741(70,486,947 - 70,520,407)	PKD1L3(70,520,942 - 70,591,378)
18	2	-	28s rRNA	-
19	4	VDP(76,897,674 - 76,954,390)	PPEF2(77,000,050 - 77,042,705)	ASAHL(77,050,832 - 77,081,190)
20	4	VDP(76,897,674 - 76,954,390)	PPEF2(77,000,050 - 77,042,705)	ASAHL(77,050,832 - 77,081,190)
21	M	-	unknown sequence862 - 1,178	-

续表 1

Sequences	Chromosomes	Upstream genes(positions)	The sequence coding genes(positions)	Downstream genes(positions)
22	M	DQ582201(236 – 368)	HN1(1,756 – 4,264)	BC018860(5,810 – 7,515)
23	20	C20orf198(36,508,742 – 36,512,978)	KIAA1219(36,534,900 – 36,640,916)	SMAF1(36,643,252 – 36,650,518)
24	2	–	28s rRNA	–
25	16	–	18s rRNA	–
26	9	NEK6(124,099,083 – 124,194,271)	PSMB7(124,195,299 – 124,257,275)	GPR144(124,292,977 – 124,318,933)
27	19	–	28s rRNA	–
28	4	VDP(76,897,674 – 76,954,390)	PPEF2(77,000,050 – 77,042,705)	ASAHL(77,050,832 – 77,081,190)
29	4	SLC25A31(128,871,024 – 128,914,896)	HSPA4L(128,922,903 – 128,973,976)	PLK4(129,021,495 – 129,039,802)
30	16	SPG7(88,102,306 – 88,151,674)	RPL13(88,154,591 – 88,157,349)	CPNE7(88,169,677 – 88,191,154)
31	1	–	unknown sequence(26,214,075 – 26,214,316)	–
32	M	BC018860(5,810 – 7,515)	OK/SW – cl.16(7,644 – 9,869)	STRF6(10,642 – 12,138)
33	16	–	18s rRNA	–
34	4	VDP(76,897,674 – 76,954,390)	PPEF2(77,000,050 – 77,042,705)	ASAHL(77,050,832 – 77,081,190)
35	9	NEK6(126,060,070 – 126,154,538)	PSMB7(126,155,565 – 126,217,542)	GPR144(126,254,610 – 126,256,669)
36	19	–	28s rRNA	–
37	1	–	unknown sequence(26,214,074 – 26,214,316)	–
38	2	–	28s rRNA	–
39	2	–	28s rRNA	–
40	19	–	28s rRNA	–
41	11	DKFZp686B0790(65,022,410 – 65,030,999)	SCYL1(65,049,124 – 65,062,758)	LTBP3(65,062,852 – 65,082,006)
42	1		28s rRNA	–
43	7	FASTK(150,211,356 – 150,215,599)	C7orf21(150,215,821 – 150,217,736)	CENTG3(150,221,474 – 150,279,717)
44	16	–	18s rRNA	–
45	7	HEATR2(732,864 – 792,642)	UNC84A(822,778 – 881,083)	C7orf20(882,717 – 902,597)
46	X	ATG4A(107,221,590 – 107,284,551)	COL4A6(107,285,502 – 107,568,316)	COL4A5(107,569,810 – 107,827,431)
47	–	–	chimeric	–
48	16	–	18s rRNA	–

Note: M for human mitochrodrial genome

2.2 ncRNA搜索结果 本地BLAST序列同源性 为功能未知基因的序列(21、22、31、32和37)以及功能已知基因内含子的序列(9、12、19和29)。这些序列可能编码新的转录本,即代表着新基因。它们的电子克隆延伸结果见表3。从表3可见,其中5个序列均被成功延伸,从几个 bp 至600多bp不等,平均为190 bp。

2.3 电子克隆延伸结果 进行电子克隆的序列搜索结果显示,在目标基因的某些内含子以及上下游基因的间隔区内存在与已知ncRNA序列相似性很高的片段,而且部分与ncRNA高度相似的片段重复出现,如与ncRNA KLHL1 antisense RNA高度相似的片段,结果见表2。另外,在第10序列编码基因 *CSNK1A1*(casein kinase 1, apha 1)的上游,发现了两个微小RNA(microRNA),分别为miR-143和miR-145。

表 2 在已知基因的内含子和上下游基因间区中寻找ncRNA的结果

Table 2 ncRNA finding results from introns and interval regions of known genes

Sequences	Positions	Homologous ncRNA	EV
1	Upstream intergenic region	KLHL1 antisense RNA	3e – 57
	Downstream intergenic region	CMPD associated ncRNA	8e – 76
2	The first intron	KLHL1 antisense RNA	1e – 61
5	The first intron	His-1	2e – 57
	The secondary intron	DISC2	2e – 75
	Upstream intergenic region	His-1	1e – 56
	Downstream intergenic region	KLHL1 antisense RNA	6e – 68
8	The first intron	KLHL1 antisense RNA	7e – 67
	The secondary intron	His-1	4e – 66
	The sixth intron	KLHL1 antisense RNA	e – 120
	The twelfth intron	KLHL1 antisense RNA	e – 127
9	The sixth intron	KLHL1 antisense RNA	e – 147
	Upstream intergenic region	KLHL1 antisense RNA	e – 163
	Downstream intergenic region	KLHL1 antisense RNA	8e – 73

3 讨 论

利用互联网上的生物信息学资源,通过电子克隆手段可以使目标基因的片段得以有效延伸,借此可以促进人类基因组上未知功能基因的发现[8],与此同时还可以进一步通过验证实验纠正以往人类基因组测序中的错误[9]。通过电子克隆使目标基因片段延伸的长度取决于相关EST数量的多少以及它们覆盖目标基因全长的程度。本文采用电子克隆方法使5个功能未知基因的序列片段平均延伸190 bp以上,这可以为进一步设计PCR引物进行基因扩增,或设计基因探针进行杂交验证实验提供更大的可操作空间。然而,特别值得注意的是,NCBI数据库中的EST可能仅仅是一次测序得到的序列结果,其精确度往往较差,这提示电子克隆延伸的序列可能与实际情况存在着较大的差异,而且还要将所延伸的序列再次进行染色体定位,避免拼接来自其它染色体上的具有部分相同序列所导致的错误延伸。然而,与传统的实验室基因

表 3 电子克隆延伸长度

Table 3 The length of silicon cloning results

Sequences	Chromosomes	Self locations	The precise locations of the extended sequences	Extension length (bp)
9	9	72,381,245 - 72,383,709	72,381,245 - 72,383,709	0
12	2	19,119,549 - 19,120,340	19,119,545 - 19,120,340	4
19	4	77,026,215 - 77,026,578	77,026,211 - 77,026,684	113
21	M	862 - 1178	603 - 1517	600
22	M	2642 - 3185	2070 - 3235	624
29	4	128,940,508 - 128,940,542	128,940,508 - 128,940,542	0
31	1	26,214,075 - 26,214,316	26,214,075 - 26,214,316	0
32	M	9225 - 9812	9123 - 9990	182

Note: M for human mitochrodrial genome

克隆方法，如 cDNA 文库筛查或 cDNA 末端快速扩增(RACE)等相比，电子克隆方法依靠电脑和网络资源，速度快、成本低是其最大的优点，因此依然有继续发挥作用的价值。

ncRNA 除了以往所知的 rRNA 和 tRNA 外，近几年还发现了 microRNA 和 siRNA 等，它们在基因表达调控、RNA 的加工与修饰、蛋白质的运输与稳定性调节等方面发挥着非常重要的作用[10-12]。本文在 48 个食管癌相关的 DNA 序列片段中搜索已知 ncRNA。经本地 BLAST 同源性分析，结果在基因的内含子以及上下游基因间隔区中发现了与已知 ncRNA 相似性很高的片段。值得注意的是，本文所发现的 ncRNA 尽管长度较短，仅 100~200 bp，但重复出现。我们推测这些 ncRNA 片段可能是人类基因组中某些 ncRNA 的早期重组进化的产物。然而，为什么这些 ncRNA 会在人类基因组的多个位点出现，它们的生物学意义如何，特别是与食管癌等肿瘤究竟有怎样的功能联系等问题尚有待进一步研究。再者，需要特别指出的是，本文在 *CSNK1A1* 基因的上游发现存在 microRNA，miR-143 和 miR-145。我们推测这可能与 *CSNK1A1* 基因的表达调控有关。

参考文献：

[1] 许丽艳，李恩民，熊华淇，等．NGAL 基因在永生化食管上皮细胞恶性转化中过表达的研究 [J]．生物化学与生物物理进展，2001,28(6):839 - 843.

[2] 许丽艳，李恩民，牛永东，等．食管癌细胞 NGAL 基因 -152~-60 区段存在 TPA 反应元件 [J]．生物化学与生物物理进展，2006,33(2):140 - 148.

[3] 牛永东，许丽艳，韩 溟，等．酵母单杂交筛选人食管癌细胞系 SHEEC 中 NF-kappa B 元件结合蛋白 [J]．癌变·畸变·突变，2006,18(4):310 - 313.

[4] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool[J]. J Mol Biol, 1990, 215(3):403 - 410.

[5] Zhang JH, Madden TL. PowerBLAST: A New Network BLAST Application for Interactive or Automated Sequence Analysis and Annotation[J]. Genome Res, 1997, 7(6): 649 - 656.

[6] Kent WJ. BLAT-the BLAST like alignment tool[J]. Genome Res, 2002, 12(4):656 - 664.

[7] Huang, X, Madan A. CAP3: A DNA sequence assembly program[J]. Genome Res, 1999, 9(9), 868 - 877.

[8] 张德礼．电子克隆新基因 [J]．中国高校科技与产业化，2002,9(1):40 - 42.

[9] 张德礼，李衍达，季梁．用电子克隆新基因 C17orf32 和 ZNF362 对 NCBI 人类基因数据库模式参考序列 5 种错误类型的分析与纠正[J]. 遗传学报, 2004, 31(4):325 - 334.

[10] 肖章奎，薛良义．ncRNA 研究技术进展 [J]．生命科学，2007, 19(2):122 - 126.

[11] Storz G. An expanding universe of noncoding RNAs[J]. Science, 2002, 296(5571):1260 - 1263.

[12] Storz G, Opdyke JA, Zhang A. Controlling mRNA stability and translation with small, noncoding RNAs[J]. Curr Opin Microbiol, 2004, 7(2):140 - 144.

中国环境诱变剂学会风险评价第九届学术交流会征稿通知(第二轮)

定于 2008 年 5 月 17~21 日在湖北省十堰市召开《中国环境诱变剂学会风险评价全国第九届学术交流会暨第四届换届大会》。会议征文内容如下：①致癌物、生殖毒物和致突变物的新进展、新方法、新规范；②食品、化妆品、药品、农药和各类产品的安全性评价；③健康风险、暴露风险、技术壁垒和其他因素风险的原理与评估方法；④相关的临床和基础研究的试验方法、检测报告、试验研究和综述性论文。论文内容应为尚未发表过的研究成果。论文限 5 000 字或摘要限 1 000 字。

会议设优秀论文奖。评选条件为：我学会会员，尚未公开发表过的论文全文，写作格式要求同本学会主办的《癌变·畸变·突变》杂志的要求一致，同时邮寄纸质全文用于初评(首页右上角注明会员证号)。

请于 4 月 1 日前将论文及参会回执 Email 发至：linfeibj@yahoo.com.cn ；也可寄至：100050 北京市天坛西门 中国药品生物制品检定所 林飞收 (参会回执请在本学会的网站上下载，网址为：www.egh.net.cn) 联系电话：010 - 67095576。

中国环境诱变剂学会风险评价委员会

2008 年 2 月 28 日