人类新基因 C17orf32 的电子克隆和编码区序列 RT-PCR 验证*

张德礼[1)]** 丁培国[1)] 凌伦奖[2)] 陈润生[2)] 马大龙[1)]

（[1)]北京大学人类疾病基因研究中心，国家人类基因组北方研究中心，北京 100083；
[2)]中国科学院生物物理研究所，北京 100101）

摘要 利用生物信息学与实验验证的技术路线，成功地克隆了人类新基因 C17orf32 的 cDNA（GenBank 登记号：AY074907 和 TPA：BK000260），发现 C17orf32 的完整开放阅读框架（ORF，31～657 bp）cDNA（627 bp）与人类假定基因 LOC124919 ORF（25～807 bp）的 25～651 位只有一个碱基不同. 经 RT-PCR 验证并 cDNA 测序、人类表达序列标签（EST）数据库的 BLAST 检索和基因组成规律分析三方面的结果，均支持 C17orf32 的序列，而不支持 LOC124919 的编码序列. C17orf32 基因组序列全长 4.610 kb，含有 6 个外显子和 5 个内含子，cDNA 序列全长1 679 bp，ORF 横跨全部 6 个外显子. 该基因 ORF 翻译起始处符合 Kozak 规则，ORF 起始码上游同一相位有终止码，ORF 后有 2 个加尾信号和 PolyA 尾. C17orf32 基因的成功克隆表明，NCBI GENOME Annotation Project 在 2001 年 12 月预测的人类假定蛋白 XP_058865 编码基因 LOC124919 的模式参考序列 XM_058865 中存在偏差，即在 C17orf32 基因 cDNA 的 406 与 407 位碱基之间错误插入一个碱基 G，从而导致在插入位点后，ORF 编码 125 位氨基酸以后蛋白质序列的改变，出现 260 个氨基酸的多肽. 因此，应慎重看待计算机注释的人类基因组编码序列. 建立的技术路线有助于发现更多新的人类功能基因.

关键词 C17orf32，LOC124919，XM_058865，XP_058865，生物信息学，电子克隆，RT-PCR，人类基因组注释

学科分类号 Q811.4

随着基因定位（连锁图谱、物理图谱、转录图谱）和人类基因组测序及生物信息技术的迅猛发展，特别是人类基因组工作框架图和精细图谱及其初步分析结果的先后公布[1,2]，表达序列标签（EST）已成为人类寻找未知功能的新基因，以及克隆不同时空差异表达基因和疾病相关基因的重要标志物[3～6]. 基于 EST 的电子克隆（in silico cloning）和定位候选克隆策略已成为克隆新基因的主要方法，并在虚拟的网上空间将模式生物基因组的研究成果成功地应用于人类基因组的研究[7～10]. C17orf32 是单纯采用生物信息学方法，电子克隆成功并初步实验验证的人类新基因之一.

1 材料和方法

1.1 计算机克隆的策略和技术路线

本研究利用人类 EST 公共数据库，采用解放军参谋部三部的神威Ⅳ型超级计算机运行 SiClone 软件（中国科学院生物物理研究所凌伦奖与陈润生教授设计）进行 EST Contig 序列的全自动化拼接和校对，以 BLAST 比对为基本方法进行人类新基因大规模搜寻工作，力求至少搜寻到 100～300 个新的人类 cDNA（400～3 000 bp，包含编码 100～300 个氨基酸残基的完整开放读码框架），采用 cDNA 序列基因预测软件、包含校对过的 cDNA 的复原基因组序列预测基因软件、蛋白质功能结构预测软件，筛选出 30～50 个具有典型基因特点和一定功能结构域的人类新基因候选对象，再通过逆转录聚合酶链式反应（RT-PCR）克隆 cDNA 序列，得到重组质粒后，通过基因表达谱分析等实验技术验证其客观存在性. 技术路线同文献［11］：a. 下载人类 EST 公共数据库的心脏表达 cDNA 序列. b. 相同 Unigene 或 Cluster 的 EST 片段归类在一起，并初步拼接和校对（SiClone 软件完成）. c. 基于 GenBank 完整 EST 数据库的 EST 拼接、校对并尽可能延长获得大片段或全长 cDNA（SiClone 软件完成），获得 Contig. d. Contig 输入 nr 数据库通过 BLAST 查新选留新颖 Contig，通过人类基因组图谱验证 Contig. e. 通过 GenBank 完整 EST 数据库和人类基因组图谱双重比对按权重校对确定最终 cDNA Contig；本实验 Contig 015574 源1 119 bp新 cDNA 片段经延长和校对，最终成为 1 679 bp cDNA Contig，并受人类基因组图谱的完全支持. f. 分析包含候选开放读码框架（ORF）的 Contig

* 中国博士后科学基金资助项目（2920011217608062000）.
** 通讯联系人. Tel：010-62092541
E-mail：delizhang@bjmu.edu.cn；delizhang2000@sohu.com
收稿日期：2002-03-08，接受日期：2002-03-28

cDNA 典型基因特征，如 ORF 前有无终止码和启动子，Kozak 规则，加尾信号，polyA 等；如有基因组序列存在，用 Contig cDNA 替换基因组序列的相应部分，获得以最终 cDNA 校对的人类基因组序列. g. 启用蛋白质功能结构和/或三维结构预测软件，分析基于序列或三维结构的 ORF 相应氨基酸序列的独特功能结构域. h. 新基因实验克隆、确认并为功能研究提供线索. 所用生物信息学软件包括 Pcgene，DNAsis 和 ORF finder 等.

1.2 实验克隆

1.2.1 质粒和菌株：质粒 pGEM-T-easy 载体购自 Promega 公司；*E. coli* XL1-blue 由王琰教授馈赠.

1.2.2 工具酶及试剂：限制性内切酶，DNA 修饰酶，T4 DNA 连接酶，Taq DNA 聚合酶购于 Boehringer Mannheim，GIBCO-BRL，Promega，TaKaRa 等公司. DNA 分子质量标准 DL2000 是 TaKaRa 公司产品. 其他常规试剂均按《分子克隆手册》配制.

1.2.3 引物的合成：根据计算机克隆所得到的完整 cDNA ORF（编码 208 个氨基酸残基）设计 5′和 3′扩增引物. 5′引物：5′ATG GCG TCC TCT TTG CTT G 3′；3′引物：5′CGG GAT CCA GTT CTC CCA GCT CGC C 3′.

采用 Net Primer 引物分析软件评价设计引物. 引物由北京赛百盛公司合成. 该 RT-PCR 引物覆盖 C17orf32 基因完整 ORF，仅不含终止码 TAA，即从 ORF 终止码 TAA 开始序列改为 GATCCCG，以便融合表达进行细胞定位.

1.2.4 PCR 反应：采用 RT-PCR 方法，以 MGC803 人胃腺癌细胞系[12] cDNA 文库为模板，在2400 PCR 仪（RPE 公司）上进行扩增30个循环[11]. 反应完毕后取10 μl于1%琼脂糖凝胶进行电泳分析，确定后，再取 10 μl PCR 产物用 Silica 回收亚克隆化.

1.2.5 重组测序质粒的构建：载体 pGEM-T-easy 和 PCR 产物（摩尔比为 1∶3）用 DNA 连接试剂盒溶液水浴 14～16 h.

1.2.6 测序策略及方法：由本室用 ABI PRISM 3100 Genetic Analyzer 进行测序. 用 T7 和 Sp6 通用引物进行正反向测序.

2 结果与讨论

2.1 RT-PCR 结果

2.1.1 PCR 的扩增：PCR 结果经 1%琼脂糖凝胶电泳分析出现单一条带，约 631 bp，与预计大小相符（图 1）.

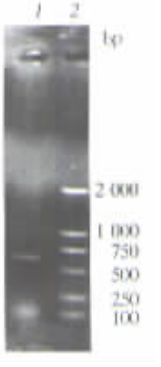

Fig.1 Identification of the PCR product of 627 bp ORF of human C17orf32 cDNA by DNA agarose gel electrophoresis

1：PCR product；*2*：DNA molecular mass marker.

2.1.2 重组质粒的构建及酶切鉴定：将末端带 A 的 PCR 产物，插入 pGEM-T-easy 载体的多克隆位点的 LacZ 基因中，获得的重组质粒命名为 pGEM-Teasy-RC208. 该重组质粒经 *Eco*R Ⅰ双酶切后，电泳显示的片段与预计的相符.

2.2 测序结果及分析

将重组的 pGEM-Teasy-RC208 重组质粒，用全自动测序仪进行序列分析（图 2 和图 3）. 测序

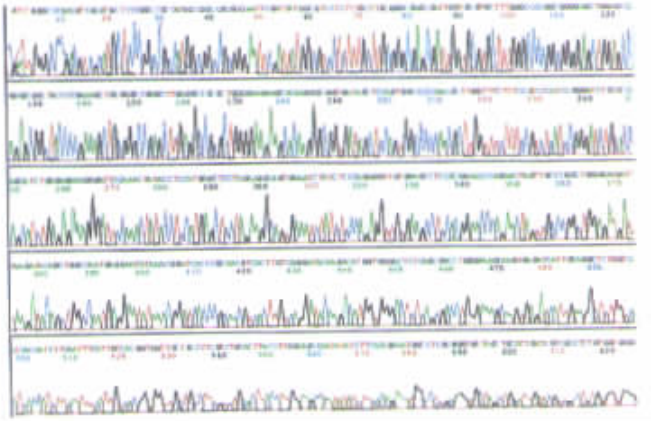

Fig.2 Result of 621 bp ORF sequencing of C17orf32 cDNA

GenBank accession number：AY074907 and TPA：BK000260.

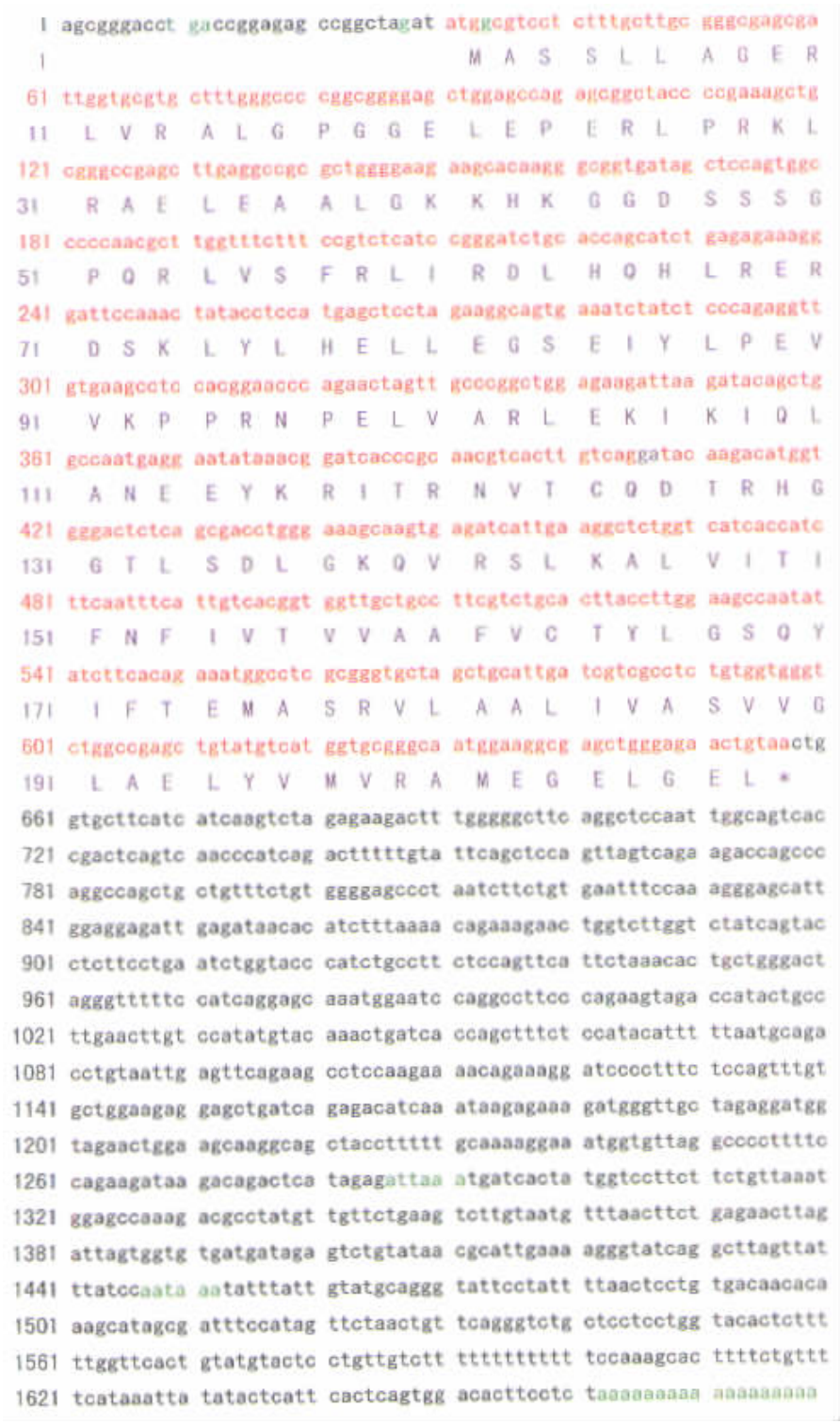

Fig.3 The full-length 1 679 bp cDNA sequence of C17orf32 and its coding protein of 208 amino acids

GenBank accession number：AY074907 and TPA：BK000260.

结果与计算机克隆 RC208 cDNA 全长序列1 679 bp 的全程比对结果表明，用 MGC803 人胃癌细胞系扩出的 cDNA 片段（包含编码 208 个氨基酸残基的最大 ORF）与电子克隆结果一致. 人类 C17orf32 cDNA 的完整序列（图 3）及其编码蛋白的氨基酸序列的 GenBank 登记号：AY074907 和 TPA：BK000260.

2.3 EST 数据库的检索结果和意义

经 EST 数据库（http://www.ncbi.nlm.nih.gov/BLAST/）比对，1 679 bp C17orf32 cDNA 全长受人类 EST 的完全支持（图 4）.

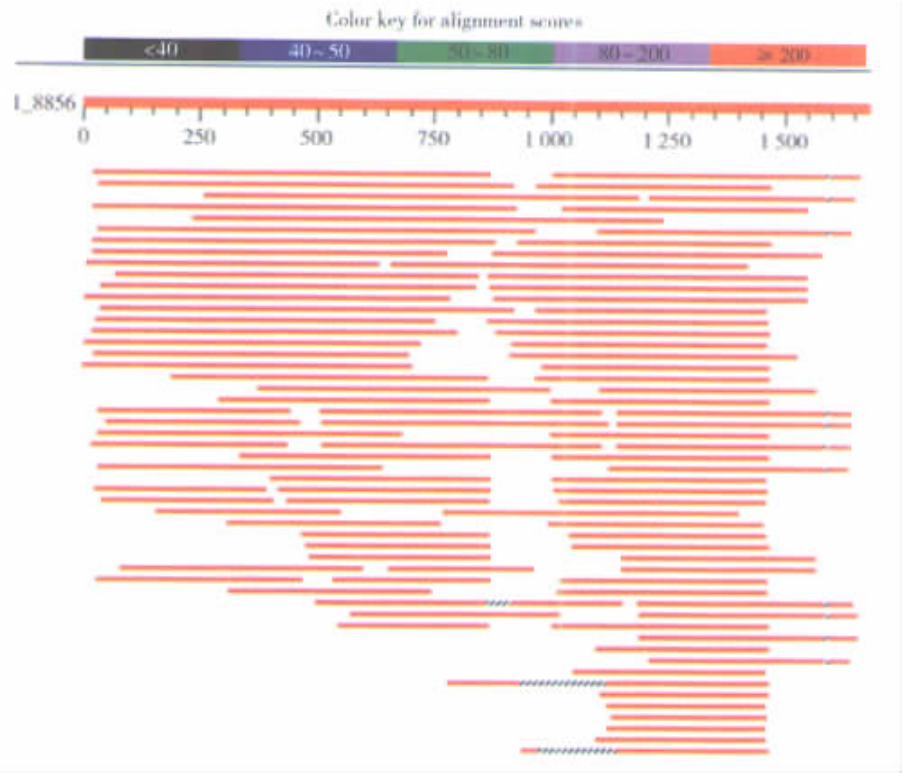

Fig.4 Alignment of C17orf32 cDNA of 1 679 bp in human EST database
GenBank accession number：AY074907 and TPA：BK000260.

2.4 nr 数据库的检索结果和意义

该基因 1 679 bp cDNA，只与 889 bp 的人类假定基因 LOC124919（supported by XM_058865）mRNA 同源（99%）. 该基因所编码 208 个氨基酸的蛋白质与全长 260 个氨基酸的人类假定蛋白 XP_058865（supported by XM_058865）共有 57% 匹配，同源性亦为 57%，其中从头开始的 125 个上游氨基酸残基完全相同（图 5）.

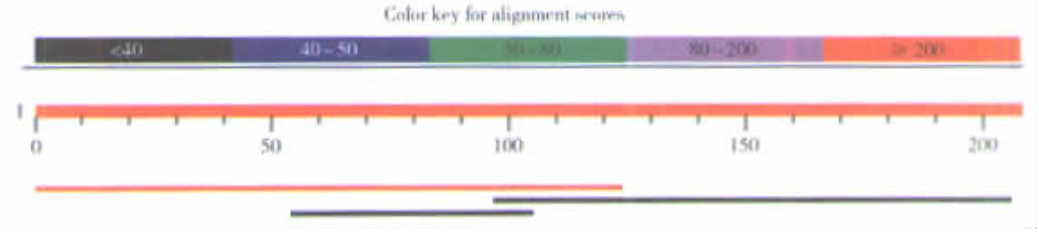

Fig.5 Alignment of 208 amino acids encoded by human C17orf32 of 1 679 bp cDNA in nr database

2.5 基因组数据库的检索结果和意义

该1 679 bp cDNA 与人类 17 号染色体NT 010808.7 完全匹配（100%）（>ref|NT_010808.7|Hs17_10965），与人类其他染色体无同源性.

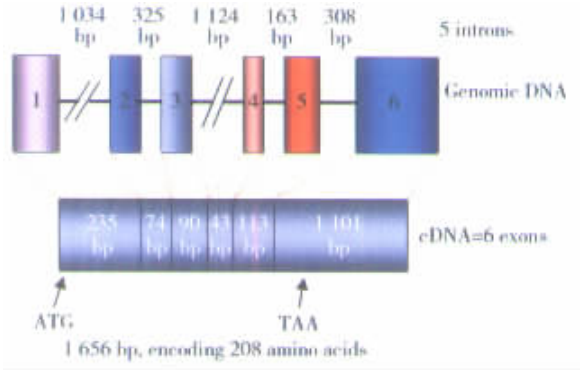

Fig.6 Chromosomal mapping and genomic organization of C17orf32, a novel human gene encoding a protein of 208 amino acid residues

GenBank accession number：AY074907 and TPA：BK000260.

该基因含有 6 个外显子，其 ORF 627 bp（31～657 bp）横跨全部 6 个外显子. ORF 下游 622～631 bp为5′～3′（+）方向的 SAGE（serial analysis of gene expression）标签 cDNA 序列 GTGCGGGCAA. 该基因的染色体结构见图 6.

2.6 启动子的检索结果和意义

将 ORF 前 2 000 bp 基因组 DNA 序列输入 Berkeley Drosophila Genome Project（BDGP）网址（http://www. fruitfly. org/seq _ tools/promoter. html）检索，结果存在 16 个启动子，可能性高达 58%～97%的有 9 个，可能性为 51%～57%的有 7 个，并且存在 2 个 TATA（图 7）.

Start	End	Score	Promoter sequence
674	724	0.53	TTAAATACTGTAAATTCCCTGGGAGCAATGCCTGAGTCTGATTAGCTGGT
736	786	0.67	GAGCCTAGTGTACAAGAGGTGCTCAAAAAATATTTGTCCTGTATTATAAT
949	999	0.57	GTGTCTAATATCAGCCACTAAAATGTCAGTCACCACTCTGAACAGGGCTG
1307	1357	0.97	TAGGGCTGAATTAAAAAGCTTCCCTCTGACATCGGTGCTGGGGACCCCAG
1403	1453	0.90	GGGCCAGCCCTGTACAACCAGGGCTCTGGGCCGCCTGGCTTCCCTGCTC
1436	1486	0.56	CGCTGGCTTCCCTGCTCCTGGCACATGTCCCAGACAGCGCAGGTCGGCTC
1665	1715	0.63	GAAGGCTCCAGCTCCGGCCGATGGCACAGAGATGGCTTTGCCACCGGGCTT
1679	1729	0.65	CGGCCGCCGCACAGAGATGGCTTTGCCATCGGGCCCCAGTCAGCGACCGGG
1722	1772	0.59	GACCGGGACCACCAGCGGCTGCCCACGGCCAGGGGCTGCCGGCTGTACA
1757	1807	0.70	CTGCCGGGCTGTACAGGCTGCCATGTCCCGACCGAGCTGCCGCCCGGCTG
1788	1838	0.58	CCGAGCGCCCGCCCGGCTGCTGACACAGAGCCCGACCCGCGCCGGGGGG
1796	1846	0.55	CCGCCCGGCTGCTGACACAGAGCCCGACCCGCGCCGGGGCGGTTCCGCCC
1807	1857	0.53	CTGACACAGAGCCCGACCCGCGGCCGGCGGTTCCGCCCCAGTCCCACAC
1831	1881	0.58	CGGGCGGTTCCGCCCCAGTCCCACACTGCCCTGCGCGGGCTCCCGGCCCT
1841	1891	0.52	CGCCCCAGTCCCACACTGCCCCGCGCGGCGTCCCGGCCCTCTTGGCCCGC
1863	1913	0.51	GCGCGGCGTCCCGGCCCTCTTGGCCCGCCTGGTCCCACTCGGAGCTCAG

Fig.7 Predicted potential promoter sequences of C17orf32 gene in 2 000 bp genomic DNA before the start codon of the ORF of C17orf32

2.7 蛋白质特点、功能区的分析

人类 C17orf32 编码蛋白由 208 个氨基酸组成，等电点 p*I* = 9.35，相对分子质量为 23.129. 输入清华大学生物信息研究所网站（http:// www. bioinfo. tsinghua. edu. cn/SubLoc/）预测，结果表明可能为胞浆蛋白（cytoplasmic reliability index：*RI* =1；expected accurcy=56%）. 输入 Pcgene 软件和 DDBJ（http://www. isrec. isb-sib. ch/software/PFSCAN _ form. html）进行 Prosite 分析，结果表明 16～169 位为降血钙素（calcitonin/CGRP/IAPP family signature）家族标签，43～117 位为二清乳清酸酶 1 标签（dihydroorotase signature 1），68～75 位为酪氨酸激酶磷酸化位点，28～45 位为双向核定位信号，116～119 位为 cAMP 和 cGMP 依赖性蛋白激酶磷酸化位点，56～144 位为 2 个蛋白激酶 C 磷酸化位点，123～135 位为 2 个酪蛋白激酶Ⅱ磷酸化位点，121～124 位为 N 端糖基化位点，44～135 位为 2 个 N 端 myristoylation 位点，38～41 位为 *Amidation* 位点.

2.8 C17orf32 与 LOC124919 序列的比较

2001 年 12 月 5 日 NCBI GENOME Annotation Project 提交了一个人类假定基因 LOC124919 的模式参考序列 XM _ 058865. LOC124919 的 ORF 为 25～807 bp，实际上是在我们克隆的 C17orf32 基因 cDNA 的 406 与 407 位碱基之间（测序峰图的 432 与 433 位碱基之间）插入了一个碱基 G，因而二者

在78%范围内有99%的同源性，从而导致在插入位点后ORF编码蛋白质的氨基酸序列改变. 基因组组成分析表明，C17orf32包括6个外显子和5个内含子，内含子与外显子交界的头尾处都符合GT/AG规律，LOC124919只是在C17orf32的第三外显子和第三内含子交接处AGGT中间多出一个碱基G，仅此一点导致不符合GT/AG这一真核生物遗传学的基本规律. 在人类EST数据库的BLAST结果表明，C17orf32基因cDNA的406位侧翼序列受EST的完全支持，而LOC124919在插入碱基G处完全不受EST的支持，因为有26条EST反对插入G，尽管华盛顿大学医学院Wilson提交的3条EST（R56811，R13211，R47783）的尾部不精确部位支持插入G，1条EST（R13877）的尾部不精确部位支持插入A，而由同一作者提交的这4条EST尾部不精确部位支持插入碱基G或A是无效的. 这些结果预示二者本身就应该是同一个基因，C17orf32基因的成功克隆证实NCBI GENOME Annotation Project预测的人类假定基因LOC124919存在序列偏差.

2.9 种属间蛋白质比较

人类C17orf32的全长序列与其他种属已知蛋白质无论在核酸水平还是在氨基酸水平几乎均无整体的同源性. 与线虫假定蛋白F09E5.11.p[*Caenorhabditis elegans*，NP_495004]和念珠藻属[13]多磷酸盐激酶（polyphosphate kinase，NP_487633）的相似性（图4）分别为31%（匹配范围29%）和18%（匹配范围18%），与线虫假定蛋白F09E5.11.p的基因F09E5.11（NM_062603）完整开放读码框架的核酸相似性为53%（52%），这预示该基因是存在进化痕迹的. 此外，它不与其他蛋白质有任何同源性. 经本实验室申请，国际人类基因命名委员会已将其命名为chromosome 17 open reading frame 32，缩写为C17orf32.

参 考 文 献

1 Lander E S，Linton L M，Birren B，*et al*. Initial sequencing and analysis of the human genome. Nature，2001，**409**（6822）：860～921

2 Venter J C，Adams M D，Myers E W，*et al*. The sequence of the human genome. Science，2001，**291**（5507）：1304～1351

3 Wiemann S，Weil B，Wellenreuther R，*et al*. Toward a catalog of human genes and proteins：sequencing and analysis of 500 novel complete protein coding human cDNAs. Genome Research，2001，**11**（3）：422～435

4 Blumberg H，Conklin D，Xu W F，*et al*. Interleukin 20：discovery，receptor identification，and role in epidermal function. Cell，2001，**104**（1）：9～19

5 Huminiecki L，Bicknell R. In silico cloning of novel endothelial-specific genes. Genome Research，2000，**10**（11）：1796～1806

6 Schultz J，Doerk T，Ponting C P，*et al*. More than 1 000 putative new human signalling proteins revealed by EST data mining. Nat Genet，2000，**25**（2）：201～204

7 Capone M C，Gorman D M，Ching E P，*et al*. Identification through bioinformatics of cDNAs encoding human thymic shared Ag-1/stem cell Ag-2. A new member of the human Ly-6 family. J Immunol，1996，**157**（3）：969～973

8 Camargo A A，Samaia H P B，Dias-Neto E，*et al*. The contribution of 700 000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci USA，2001，**28**（21）：12103～12108

9 Hogenesch J B，Ching K A，Batalov S，*et al*. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell，2001，**106**（4）：413～415

10 Gopal S，Schroeder M，Pieper U，*et al*. Homology-based annotation yields 1 042 new candidate genes in the *Drosophila melanogaster* genome. Nature Genetics，2001，**27**（3）：337～340

11 张德礼，孙晓静，凌伦奖，等. 人类SR蛋白超家族新成员——SFRS12（SRrp508）的基因克隆和特性分析. 遗传学报，2002，**29**（5）：377～383
Zhang D L，Sun X J，Ling L J，*et al*. Acta Hereditas Sinica，2002，**29**（5）：377～383

12 Tian X，Song S，Wu J，*et al*. Vascular endothelial growth factor：acting as an autocrine growth factor for human gastric adenocarcinoma cell MGC803. Biochem Biophys Res Commun，2001，**286**（3）：505～512

13 Kaneko T，Nakamura Y，Wolk C P，*et al*. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res，2001，**8**（5）：205～213

In Silico Cloning of C17orf32，a Novel Human Gene and Verification of Its Coding Region by RT-PCR*

ZHANG De-Li[1)]**，DING Pei-Guo[1)]，LING Lun-Jiang[2)]，CHEN Run-Sheng[2)]，MA Da-Long[1)]

([1)] *Peking University Center for Human Disease Genomics，China National Center for Human Genome Research，Beijing* 100083，China；
[2)] Institute of Biophysics，The Chinese Academy of Sciences，Beijing 100101，China)

Abstract A novel human gene encoding a protein of 208 amino acids is identified and characterized，which has been offered by HGNC with symbol of C17orf32 and name of chromosome 17 open reading frame 32. The full-length cDNA of 1 679 bp for C17orf32 was cloned through a blast search of public databases following the

identification of 1 119 bp cDNA obtained by EST assembly with full robotization of SiClone software (created by Chen RS and Ling LJ, and will be released on their website) in ShenWei Ⅳ-type supercomputer. Structurally, C17orf32 has one calcitonin／CGRP／IAPP family signature from amino acid 16 to 169, one dihydroorotase signature from amino acid 43 to 117, one tyrosine kinase phosphorylation site from amino acid 68 to 75, and one bipartite nuclear localization signal from amino acid 28 to 45. These motifs imply the potential biological importance of this gene. Genomic organization analyses show that C17orf32 gene is comprised of six exons, in the size ranging from 43 to 1 101 bp, and five introns, in the size ranging from 163 to 1 124 bp, and spanning 4.61 kb. All of the exon／intron boundaries are consistent with the GT／AG rule, and consensuses surrounding the splice boundaries are found as well. The C17orf32 gene is located on accession NT_010808.7 in the human chromosome 17, and is only linked with LOC124919, a hypothetical human gene of 889 bp mRNA encoding hypothetical protein XP_058865 of 260 amino acids supported by XM_058865. The sequence of LOC124919 has not been verified experimentally. Furthermore, the full-length ORF of 627 bp cDNA from 31 to 654 bp by RT-PCR from the single-stranded human gastric adenocarcinoma MGC803 cell line are cloned and sequenced, which is fully identical with that of the in silico cloning determined by the nucleotide sequencing. Thus, in silico cloning of C17orf31 gene with GenBank accession number of AY074907 and TPA：BK000260 is identified solely by bioinformatics analyses. The full-length cDNA sequence of 1 679 bp exhibits very good overall homology to that of LOC123722 of 899 bp mRNA, with matching percentage of 99% in 78% of total window and 57% in 57% of total window over the full-length nucleotide and protein, respectively. However, the base G in the No. 401 position of LOC123722 cDNA is a redundant insert, which causes a reading frame shift in the translation of an alternative protein. The insert G of LOC123722 is not supported by the experimental clone, and is fully rejected by human EST alignment, and is shown as a redundance by genomic GT／AG organization analysis. C17orf32 gene has 9 putative promoters with possibility of 58% ～97%, two TATAs, a stop codon in the upstream of ORF, two PolyA signals and a PolyA tail in the downstream of ORF, and accords with Kozak rule around the translation start of the ORF. Based on the above results, it can be concluded that a complete novel human gene is obtained. The full-length gene sequence exhibits little overall homology to any known protein at either the nucleotide or the amino acid level. The two related proteins, with 31% (in 29% of total window) and 18% (in 18% of total window) identity over the full-length protein, respectively, are hypothetical caenorhabditis elegans protein F09E5.11.p of 221 amino acids and polyphosphate kinase [the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120] of 736 amino acids. Taken together, by combining bioinformatics analyses with experimental verification, a novel human gene C17orf32 is successfully cloned, verified by a series of theoretical and experimental evidence. The strategy will be helpful in discovering more novel human genes, even in correcting errors appeared in NCBI GENOME ANNOTATION PROJECT REFSEQs, such as LOC124919, a model reference sequence predicted from NCBI contig NT_010808 by automated computational analysis using gene prediction method. Therefore, human genome coding region annotated by computer should be used with caution.

Key words C17orf32, LOC124919, XM_058865, XP_058865, bioinformatics, in silico cloning, RT-PCR, human genome annotation

* This work was supported by a grant from The China Postdoctoral Science Foundation (2920011217608062000).

** Corresponding author. Tel：86-10-62092541, E-mail：delizhang@bjmu.edu.cn；delizhang2000@sohu.com

Received：March 8, 2002 Accepted：March 28, 2002