表达序列标签数据库搜索鉴定小鼠 UBAP1 基因及其数字化表达分析*

钱 骏 董 利 张必成 王洁如 周 鸣 李忠花 李伟芳 李小玲 李桂源**
(中南大学湘雅医学院肿瘤研究所, 教育部癌变与侵袭原理重点实验室, 长沙 410078)

摘要 UBAP1 (ubiquitin associated protein 1) 基因是最近克隆的一个定位于人类染色体 9p21-22 鼻咽癌杂合性丢失高频区的泛肽相关蛋白家族新成员. 为了深入研究 UBAP1 基因的功能, 利用计算机对表达序列标签 (expressed sequence tag, EST) 、UniGene 等数据库进行综合搜索分析, 结合 cDNA 克隆测序的方法, 成功地获得了 UBAP1 基因在小鼠中的同源基因. 小鼠 UBAP1 基因 cDNA 全长为 2 676 bp, 编码一个由 441 个氨基酸组成的蛋白质, 在其蛋白质 C 端只有一个泛肽相关功能域 (UBA domain). 与人 UBAP1 基因相比, 两者编码的氨基酸序列有 89% 相同. 基于 EST 的数字化表达分析显示 UBAP1 基因在小鼠正常组织中广泛高表达.

关键词 UBAP1 基因, 基因克隆, 表达序列标签, 组织表达
学科分类号 Q78, R739

泛肽相关蛋白是真核生物中一个重要基因家族. 在目前发现的一百多个家族成员中, 大多数成员主要参与泛肽介导的蛋白质水解途径, 如泛肽羧端水解酶类 (ubiquitin carboxy-terminal hydrolases), 泛肽结合酶 (Ub-conjugating enzyme) E2 和泛肽连接酶 (Ub-ligating enzyme) E3 等, 在转录水平的调节、蛋白质降解、细胞凋亡、细胞周期调控等过程中具有重要作用[1,2]. 该家族成员的共同特征是具有一个或多个在进化中高度保守的 UBA 功能域 (ubiquitin associated domain). UBA 功能域是新近发现的能与泛肽非共价结合的结构域, 靶蛋白通过 UBA 功能域与泛肽结合, 发生泛肽化 (ubiquitination), 是进入泛肽途径的起始信号[2]. 近年来关于泛肽系统在细胞恶性转化和癌变发生发展中的作用日益受到重视. 泛肽系统不仅参与了细胞周期和多种癌/抑癌基因的控制, 一些泛肽相关基因本身就是潜在的癌/抑癌基因[3]. 最近, 在人类染色体 9p21-22 鼻咽癌杂合性丢失 (loss of heterozygosity, LOH) 高频区, 应用表达序列标签 (expressed sequence tags, EST) 介导的定位-候选克隆策略, 我们获得了一个泛肽相关蛋白家族的新成员 UBAP1 (ubiquitin associated protein 1) 基因. UBAP1 基因在人正常组织中广泛表达, 初步研究显示该基因在鼻咽癌中存在显著的表达下调和/或缺失, 是鼻咽癌抑瘤基因的有力候选者[4,5]. 利用已有的生物信息学工具, 对 EST 和 UniGene 等数据库进行同源搜索分析, 我们进一步克隆了人 UBAP1 基因在小鼠中的同源基因, 为今后以小鼠动物模型深入研究 UBAP1 基因的功能奠定了基础. 基于 EST 的数字化表达分析, 显示 UBAP1 基因在小鼠组织中广泛高表达.

1 材料与方法

1.1 软件与数据库

核酸数据库为 GenBank 118.0 版, EMBL 61.0 版, 蛋白质数据库为 Swiss-Prot 39.0 和 TrEMBL 14.0 版, 通过 INTERNET 查询美国国家生物信息中心 (NCBI) 的 EST 数据库和 UniGene 数据库 (www.ncbi.nlm.nih.gov). 对 cDNA 序列以及推导蛋白的氨基酸序列采用 Blast 程序与 GenBank 和 Swiss-Prot/TrEMBL 数据库进行同源性搜索. 蛋白质基序、结构域的查询和家族分析使用 InterPro 数据库(http://www.ebi.ac.uk/interpro/scan.html).

1.2 小鼠 EST 数据库搜索与鉴定

以人 UBAP1 基因的全长 cDNA 序列为查询序列对鼠 EST 数据库进行 Blastn 分析, 对匹配结果的 E 值小于 10^{-10}的小鼠 EST 进行 EST 重叠群构建. 各重叠群的同一序列用 FASTA 和 TFASTA 程序与人 UBAP1 基因比较, 根据相似区段的位置绘

* 国家高技术“863”计划资助项目 (102-10-01-05, Z19-01-01-03)、国家重点基础研究发展规划项目 (973) (G1998051008) 及美国中华医学基金资助 (96655).
** 通讯联系人.
Tel: 0731-4805383, E-mail: ligy@cs.hn.cn
收稿日期: 2001-06-27, 接受日期: 2001-07-30

制同源重叠群图. 同时对所获得的重叠群与 NCBI 的 UniGene 数据库进行比较, 采用位于意大利遗传及医学研究所 (TIGEM) 的 EST 拼接机器 (www. tigem. it/ESTmachine. html) 及英国人类基因组作图项目资源中心 (HGMP-RC) 的 ESTblast 软件 (http://www. hgmp. mrc. ac. uk/Registered/Webapp/estblast/) 进行全长 cDNA 的拼接和组装.

1.3 电子杂交和数字化组织表达分析

以小鼠 UBAP1 基因的全长 cDNA 序列为“探针”, 采用 Blastn 程序对鼠 EST 数据库进行匹配分析, 所获得的一致性大于 95%, E 值小于 10^{-10} 的 EST 被认为是代表小鼠 UBAP1 基因的 EST. 将 Blast 匹配得到的 EST 与 UniGene 数据库里所收录的代表小鼠 UBAP1 基因的 EST 重叠群里所有的 EST 逐一进行比较, 对每个 EST 来源的 cDNA 文库进行分析, 判别 EST 的组织来源. 由于每个 cDNA 文库都有详细的构建方法、组织来源和所测 EST 数量的介绍, 对于某一 UniGene 的 EST 重叠群来说, 在理论上均可用于虚拟的数字化组织表达分析, 也即俗称的电子杂交 (electronic or virtual Northern). 通过计算各类组织的 cDNA 文库中代表小鼠 UBAP1 基因的 EST 数量, 即可反映该基因在不同组织中的表达丰度[6,7].

2 结 果

2.1 小鼠 UBAP1 的全长 cDNA 克隆及序列分析

采用 EST 策略进行基因的全长克隆, 对 Blast 同源性检索要求比较严格, 通常 E 值要求小于 10^{-4}[8]. 以人 UBAP1 基因的全长 cDNA 序列为查询序列, 对 GenBank 的鼠源 EST 数据库进行 Blast 检索, 发现共有 76 条 EST 序列与人 UBAP1 基因的同源性满足搜索参数. 通过 TIGEM 的 EST 拼接机器及 HGMP-RC 的 ESTblast 服务器对小鼠 UniGene 数据库进行 Blast 检索, 检出与人 UBAP1 基因同源的小鼠 EST 族 Mm29877. 小鼠 EST 族 Mm.29877 由 136 条 EST 序列构成, 此基因族可被组装为三个长度分别为 2 672 bp、384 bp、425 bp的连续 EST 重叠群. 将这三个重叠群序列分别与人 UBAP1 基因作 Blast2 比较分析, 发现二个小片段重叠群与人 UBAP1 基因无任何同源性, 而长度为 2 672 bp 的重叠群序列共有四个部分序列分别与人 UBAP1 基因的相应片段高度同源, 尤其是与人 UBAP1 基因编码区片段同一性达 89%, 同时该重叠群序列基本覆盖了人 UBAP1 的全长 cDNA 序列 (图 1).

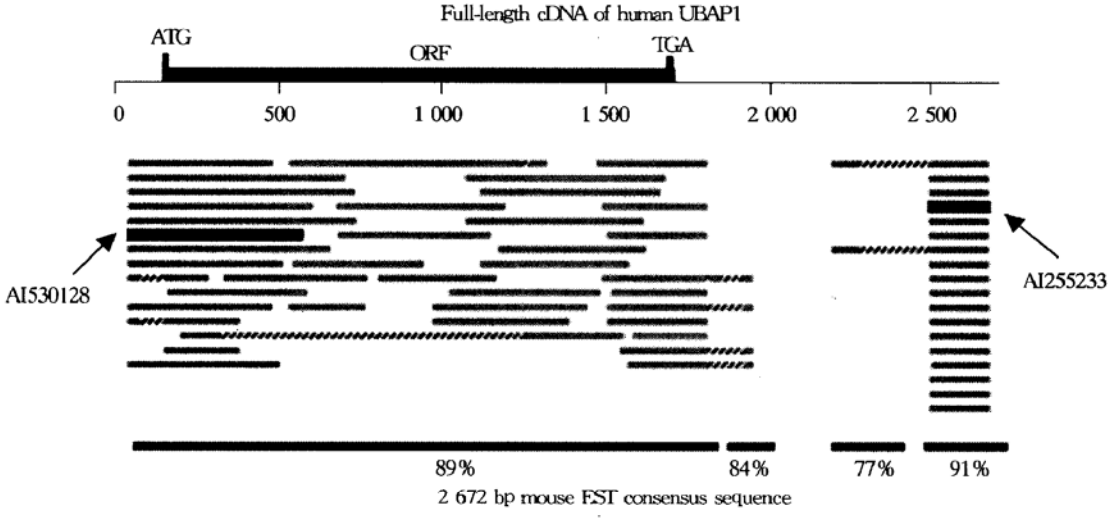

Fig. 1 Homologous comparison of human UBAP1 cDNA against 2 672 bp mouse EST contig

Percent showed identities between 4 homologous fragments of mouse consensus cDNA and human UBAP1 cDNA. The ORF region and start/stop codon of human UBAP1 were in bold line. The mouse EST AI530128 and EST AI255233 overlapping the 5′ and 3′ end of hUBAP1 were pointed out by the arrows.

为了获得小鼠 UBAP1 基因的全长 cDNA 序列, 将 2672 重叠群序列与人 UBAP1 基因作比较分析, 发现该重叠群中的 EST AI530128 和 AI255233 分别与人 UBAP1 基因的全长 cDNA 序列的 5′端和 3′端高度同源 (图 1), 同时这两个 EST 都来源于同一个 cDNA 克隆 IMAGE1889594 的 5′

和3′端，显示该克隆的插入片段可能包含小鼠ubap1基因的cDNA全长序列. 从IMAGE国际合作组购买得到cDNA克隆IMAGE1889594，对该克隆进行直接测序得到一长为2 676 bp的cDNA插入序列，含有一完整的长为1 326 bp的开放阅读框. 起始密码子ATG位于203碱基，在1 528碱基处有终止密码子和位于2 636碱基的AATAAA Poly (A) 加尾信号. 与人UBAP1基因相比较，两者编码的氨基酸序列有89%相同，且两者的UBA结构域氨基酸序列100%相同（图2）. 与人UBAP1基因不同的是，小鼠UBAP1基因编码的蛋白质只含一个UBA功能域. 小鼠UBAP1基因的GenBank编号为AF 275549.

```
human_UBAP1   MASKKLGADFHGTFSYLDDVPFKTGDKFKTPAKVGLPIGFSLPDCLQVVREVQYDFSLEK 60
mouse_UBAP1   MASKKLGTDVHGTFSYLDDVPFKIGDKFKTPAKVGLPIGFSLPDCLQVVREMQYDFSLEK 60
              *******:*.************ ****************************:******** 

human_UBAP1   KTIEWAEEIKKIEEAERKAECKIAEAEAKVNSKSGPEGDSKMSFSKTHSTATMPPPINPI 120
mouse_UBAP1   KTIEWAKDIKLIQEAQREAERKAEEAEAKVNSKSGPEGDSKVSFPKTHNTATMPPPINPI 120
              *******:** *:**:**** *  *****************:**.***.********** 

human_UBAP1   LASLQHNSILTPTRVSSSATKQKVLSPPHIKADFNLADFECEEDPFDNLELKTIDEKEEL 180
mouse_UBAP1   LASLQHNNILTPTRVSSSATKQKVLSPPHTKADFNPADFECEEDPFDNLELKTIDEKEEL 180
              *******.******************** ***** ************************ 

human_UBAP1   RNILVGTTGPIMAQLLDNNLPRGGSGSVLQDEEVLASLERATLDFKPLHKPNGFITLPQL 240
mouse_UBAP1   RNILVGTTGPIMAQLLDSNTARGSSGAVLQDEEVLASLEQATLDFKPLHKPNGFITLPQL 240
              ****************.* .**.**:***********:******************** 

human_UBAP1   GNCEKMSLSSKVSLPPIPAVSNIKSLSFPKLDSDDSNQKTAKLASTFHSTSCLRNGTFQN 300
mouse_UBAP1   GNCEKMSLSSKVSLPPIPTVSNIKSLSFPKLDSDDSNQKTVKLASTFHSTSCLRSGASRS 300
              ******************:**********************.*********.*: :.  

human_UBAP1   SLKPSTQSSASELNGHHTLGLSALNLDSGTEMPALTSSQMPSLSVLSVCTEESSPPNTGP 360
mouse_UBAP1   LLKPSTQSSASELNGDHTLGLSALNLNSGTEVPTLTSSQMPSLSSVSVCTEESSPPDPCP 360
               ************.**********:****:*:********** :**********:. * 

human_UBAP1   TVTPPNFSVSQVPNMPSCPQAYSELQMLSPSERQCVETVVNMGYSYECVLRAMKKKGENI 420
mouse_UBAP1   TVTPLNFSVSQVPNMPSCPQAYLELQALSPSERQCVETVVNMGYSYDCVLRAMRKKGENI 420
              **** ***************** *** ********************:******:****** 

human_UBAP1   EQILDYLFAHGQLCEKGFDPLLVEEALEMHQCSEEKMMEFLQLMSKFKEMGFELKDIKEV 480
mouse_UBAP1   EQILDYLFAHGQLCEKGFDLF---------------------------------------- 441
              ******************* :

human_UBAP1   LLLHNNDQDNALEDLMARAGAS 502
mouse_UBAP1   ---------------------
```

Fig. 2 Homologous comparison of predicted amino acid sequence between human UBAP1 and mouse UBAP1 gene

Conserved residues are indicated by "* ": UBA domain is underlined and shaded.

2.2 小鼠UBAP1基因的数字化组织表达分析

以小鼠UBAP1基因的cDNA全长为查询序列，对鼠源EST数据库进行Blastn搜索，共获得136个高度同源的EST. 剔除来自相同克隆的冗余EST，共有121条EST分别来自于64个不同组织的cDNA文库，其中87条EST来自于22种小鼠正常组织. 电子RNA印迹显示，UBAP1基因在小鼠这22种正常组织中广泛表达，其中除了肾、口腔、胰腺中表达较弱外，在其他组织中均有较强的表达（图3）.

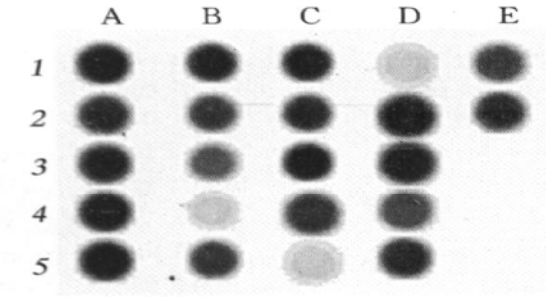

Fig. 3 Tissue expression analysis of mouse UBAP1 gene by virtual Northern

*1*A: bone, *1*B: cecum, *1*C: lung, *1*D: pancreas, *1*E: thymus, *2*A: brain, *2*B: head_ neck, *2*C: lymph node, *2*D: pituitary gland, *2*E: vascular, *3*A: colon, *3*B: heart, *3*C: mammary gland, *3*D: skin, *4*A: ear, *4*B: kidney, *4*C: nerves, *4*D: stomach, *5*A: eye, *5*B: liver, *5*C: ovary, *5*D: testis.

3 讨 论

人类基因组草图的完成意味着对人类3万多基因的功能研究将成为后基因组时代的核心任务. 利用小鼠动物模型是从整体上研究这些基因功能及相关疾病发生发展分子机制的理想方法. 而建立转基因或基因剔除小鼠模型的前提是需要获得人类基因在小鼠中的同源基因. 截至2001年6月15日，在公共EST数据库中，已录入的来自不同物种不同组织的EST有800多万条. 其中，大多数来自于人和小鼠（69%），小鼠已成为数据库中EST增长最快的模式生物(http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). 因此，利用EST

数据库进行同源搜索分析是鉴定人类基因在小鼠中同源基因最快捷的方法. 目前国际上相当数量的基因克隆研究都是从同源 EST 分析开始的, 而利用 EST 数据库里丰富的组织来源和疾病状态等信息, 进行基因的大规模表达分析, 则是近年兴起的基因表达研究强有力的工具和方法, 已成为与基因表达系列分析 (serial analysis of gene expression, SAGE)、基因芯片等并架齐驱的高通量表达分析方法之一[7,9,10].

EST 是来自随机选取的 cDNA 克隆的末端序列. 简单地说, 一个 EST 就是对应于某一种组织 mRNA 的一个 cDNA 克隆的一段序列. 一般长度大于 150bp 的 EST 在同源查找和基因作图中的作用较大. 但也不尽然, 事实上, 在我们获得人 UBAP1 基因的全长 cDNA 过程中, 包含了人 UBAP1 全长 cDNA 的克隆 2428252 的 3′EST 只有 88 bp, 虽然两者序列的同一性为 100%, 但如果未加仔细分析, 盲目地按照所谓的筛选原则和默认 BLAST 搜索参数就很容易忽略此类过短的 EST[4]. 对绝大部分克隆来说, EST 取自其 cDNA 两端, 通过 EST 分析还可以较容易地钓取该 EST 相应克隆的另一端 EST 序列. 本研究中, 我们正是通过分析覆盖人 UBAP1 基因全长的小鼠 EST 重叠群的分布, 很幸运地找到了与人 UBAP1 全长 cDNA 两端高度同源的来自同一克隆的小鼠 EST. 这样的 cDNA 克隆往往意味着包含全长 cDNA 插入片段[8]. 对该克隆的直接测序结果也证实了我们的分析. 随着 EST 数据库的剧增和 IMAGE 协定的扩展, 越来越多新基因的 cDNA 全长可以通过索取包含该基因相应 EST 的 IMAGE 克隆而得到, 避免或减轻了采用 RACE、文库筛选等实验室方法筛选全长基因的麻烦[8,11].

对于从不同组织、不同病理状态下所构建的 cDNA 文库来说, 代表某一基因的 EST 被检出的数量不仅可以反映该基因在不同组织间的表达分布, 还可以反映该基因在特定组织的表达丰度. NCBI 于 1996 年提出的癌症基因组解剖计划 (cancer genome anatomy project, CGAP), 就是针对大量 cDNA 文库和 cDNA 序列的快速发展所提出的利用低消耗、高通量的 EST 技术绘制不同基因在癌变发生发展过程中的差异表达谱[6]. 其中 CGAP 的分支之一, 小鼠肿瘤基因索引 (Mouse Tumor Gene Index, Mtgi) 就是与人类肿瘤基因索引 (Human Tumor Gene Index, Htgi) 相辅相成, 致力于建立和利用小鼠模型, 促进肿瘤相关基因的鉴定, 发现癌变过程中的新基因和新通路, 同时积累经验, 加快肿瘤基因功能的研究. 为此, CGAP 计划发展了一系列新兴的生物信息学工具, 其中最重要的就是包括 virtual Northern 在内的利用 EST 数据库进行大规模表达研究的数字化差异表达系统 (digital differential display, DDD)[6,12,13]. 利用这些工具, 我们很方便地获得了小鼠 UBAP1 基因在正常组织中的广泛表达图谱. 值得一提的是, UBAP1 基因除了在鼻咽癌中表达下调外[4,5], 数字化表达研究显示 UBAP1 基因在人类其他肿瘤如结肠癌、直肠癌、脑瘤及肝癌中也存在着可能的表达下调, 提示 UBAP1 基因很可能与人类多种肿瘤的发生发展密切相关. 有关 UBAP1 基因在上述肿瘤中的表达分析实验正在进行中. 同时我们先前采用差异 RT-PCR 的方法检测了小鼠的 9 个组织中 UBAP1 基因的表达, 显示小鼠 UBAP1 基因在这 9 种组织中都有广泛表达, 而且在肾中的表达的确比其他组织要弱, 与本研究中的电子杂交结果一致[4]. 有趣的是, 随着 EST 数据库的剧增, 在最新 GneBank 123.0 版的小鼠 EST 库中, 已经可以搜索到 226 个与小鼠 UBAP1 基因同源的 EST, UBAP1 基因在小鼠中的表达也扩大到了 38 个组织. 目前的研究充分证明了 UBAP1 基因在人和鼠中高度保守并广泛表达. 一般认为进化保守的基因可能参与重要生理活动调控, 人和鼠 UBAP1 蛋白的高度同源及它们在组织的泛表达说明这一基因功能的重要性. 小鼠 UBAP1 基因的克隆为深入研究 UBAP1 基因的功能提供了坚实的实验材料.

参 考 文 献

1 Varshavsky A. The ubiquitin system. Trends Biochem Sci, 1997, **22** (10): 383~387

2 Hofmann K. The uba domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. Trends Biochem Sci, 1996, **21** (5): 172~173

3 Spataro V, Norbury C, Harris A L. The ubiquitin-proteasome pathway in cancer. Br J Cancer, 1998, **77** (3): 448~455

4 Qian J, Yang J B, Zhang X H, *et al*. Isolation and characterization of a novel cDNA, UBAP1, derived from the tumor suppressor locus in human chromosome 9p21-22. J Cancer Res Clin Oncol, 2001, **127** (10): 613~618

5 钱 骏, 王洁如, 向 秋, 等. 一个定位于染色体 9p21-22 的新基因 UBAP1 的克隆及在鼻咽癌中的表达分析. 中国生物化学与分子生物学报, 2001, **17** (3): 299~305
Qian J, Wang J R, Xiang Q, *et al*. Chin J Biochem Mol Biol, 2001, **17** (3): 299~305

6 Krizman D B, Wagner L, Lash A, *et al*. The cancer genome anatomy project: EST sequencing and the genetics of cancer

progression. Neoplasia, 1999, **1** (2): 101~ 106
7 Colantuoni C, Purcell A E, Bouton C M, *et al*. High throughput analysis of gene expression in the human brain. J Neurosci Res, 2000, **59** (1): 1~ 10
8 Gill R W, Sanseau P. Rapid in silico cloning of genes using expressed sequence tags (ESTs). Biotechnol Annu Rev, 2000, **5**: 25~ 44
9 Pandey A, Lewitter F. Nucleotide sequence databases: a gold mine for biologists. Trends Biochem Sci, 1999, **24** (7): 276~ 280
10 Carulli J P, Artinger M, Swain P M, *et al*. High throughput analysis of differential gene expression. J Cell Biochem Suppl, 1998, **30-31**: 286~ 296
11 Lennon G, Auffray C, Polymeropoulos M, *et al*. The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. Genomics, 1996, **33** (1): 151~ 152
12 Lal A, Lash A E, Altschul S F. A public database for gene expression in human cancers. Cancer Research, 1999, **59** (21): 5403~ 5407
13 Riggins G J, Strausberg R L. Genome and genetic resources from the Cancer Genome Anatomy Project. Hum Mol Genet, 2001, **10** (7): 663~ 677

Identification and Digitalized Expression Analysis of Murine UBAP1 Gene by Means of EST Database Searching*

QIAN Jun, DONG Li, ZHANG Bi-Cheng, WANG Jie-Ru, ZHOU Ming, LI Zhong-Hua, LI Wei-Fang, LI Xiao-Ling, LI Gui-Yuan**

(Cancer Research Institute, Xiang-Ya School of Medicine, Central South University, Key Laboratory of Carcinogenesis, Chinese Ministry of Education, Changsha 410078, China)

Abstarct Ubiquitin associated protein 1 (UBAP1) is a novel member of UBA domain protein family, located at human chromosome 9p21-22 where loss of heterozygosity frequently occurs in nasopharyngeal carcinoma. Through integrated analysis of public database such as EST, Unigene in silico and direct sequencing of cDNA clone, full-length cDNA sequence of murine homologoue of human UBAP1 gene was identified, which is 2 676 bp long and encodes a putative protein of 441 amino acids with a UBA domain. There is an average of 89% identical residues between mouse and human UBAP1 protein. Digitalized expression analysis based on EST data showed that mouse UBAP1 gene expressed ubiquitiously and strongly in most of mouse normal tissues.

Key words UBAP1, gene cloning, expressed sequence tag (EST), tissue expression

* This work was supported by grants from state 863 High Technolog R & D Project of China (102-10-01-05, Z19-01-01-03), the National Basic Research Programs of China (G1998051008) and Chinese Medicine Board (96655).

** Corresponding author. Tel: 86-731-4805383, E-mail: ligy@cs.hn.cn

Received: June 27, 2001 Accepted: July 30, 2001