技术与方法

DOI:10.3969/j.issn.0253-3626.2011.10.023

舌鳞癌干细胞相关差异表达序列标签的生物信息学分析

邹　波，季　平，孙守娟，漆小娟
（重庆医科大学附属口腔医院颌面外科，重庆　400015）

【摘　要】目的：对利用抑制消减杂交技术获得的舌鳞癌干细胞相关表达序列标签（Expressed squence tag，EST）进行电子克隆及序列分析。方法：利用互联网上的公共数据库及生物信息学分析软件对获得的舌鳞癌干细胞相关差异表达序列标签进行核酸序列和蛋白质序列分析，探索克隆差异表达基因的研究方法和策略。结果：通过电子延伸得到 1 个 1 705 bp 的 cDNA 序列；核酸序列分析显示，该 cDNA 定位在 1 号染色体，含有 1 个编码 146 个氨基酸的完整开放阅读框；对氨基酸序列进行结构域预测显示其存在 1 个与 Translin 超家族同源的区域；BLAST 比对显示该氨基酸序列与 Translin 相关蛋白 X（Translin associated protein X，TRAX）具有较高的同源性。结论：利用生物信息学技术获得了舌鳞癌相关表达序列标签的基本信息，为深入研究该表达序列标签的功能奠定基础。

【关键词】肿瘤干细胞；舌鳞癌；表达序列标签；生物信息学分析
【中国图书分类法分类号】R739.8　　【文献标志码】A　　【收稿日期】2011-07-01

Bioinformatics analysis of expressed sequence tag associated with stem cells in tongue squamous carcinoma

ZOU Bo，JI Ping，SUN Shou-juan，QI Xiao-juan
（*Department of Oral and Maxillofacial Surgery，the Affiliated Stomatology Hospital，Chongqing Medical University*）

【Abstract】*Objective*：To analyze the expressed squence tag（EST）associated with stem cells in tongue squamous carcinoma obtained by suppression subtractive hybrization（SSH）. *Methods*：EST was acquired from stem cells in tongue squamous carcinoma by SSH. Bioinformatics analysis of stem cell related genes was done by the bioinformatic database and biosoftwares such as BLAST and Ex-Pasy analyse soft box. *Results*：A 1 705 bp complete cDNA sequence was obtained by electronic clone. This cDNA sequence was sited in homo sapiens 1q42，and included a complete open reading frame（ORF）coding 146 amino acids. The analysis of its protein se quence indicated that this protein contained a structural domain similar to Translin superfamily. BLAST analysis revealed that this protein was homologous with Translin associated protein X（TRAX）. *Conclusion*：Through bioinformatics analysis，we obtain the basic information about stem cell related EST in tongue squamous carcinoma，and this may provide potentially important insights into the function analysis of these ESTs.

【Key words】tumor stem cell；tongue squamous carcinoma；expressed sequence tag；bioinformatics analysis

肿瘤干细胞（Cancer stem cell，CSC）是肿瘤组织内的一小群特殊的细胞，具有正常组织干细胞的生物学特性，如自我更新、多向分化能力等[1]。越来越多的证据证实了肿瘤干细胞的存在，并且与肿瘤的发生、复发和转移相关[2-4]。目前对于肿瘤干细胞的研究主要集中于细胞分选和鉴定，这也是肿瘤干细胞研究的难点之一，由于缺乏对特异性标志物的认识，现阶段并不能精确的识别肿瘤干细胞，更多的是富集具有干细胞性质的肿瘤细胞。在口腔鳞癌中，利用不同的方法也都成功富集了肿瘤干细胞样细胞[5-7]。为了深入研究舌鳞癌干细胞生物学特性的相关机制，进一步探寻干细胞特异标志物，本实验室利用醛脱氢酶（Aldehyde dehyclrogenase，ALDH）为标志物从舌鳞癌细胞株Tca8113 中富集肿瘤干细胞样细胞，并采用抑制消减杂交技术（Suppression subtractive hybridization，SSH）筛选干细胞样细胞与普通肿瘤细胞的差异表基因，构建消减文库。经初步分析消减文库中的 cDNA 片段，我们获得多条已知基因以及一些与已知基因同源性较低的但与表达序列标签（Expressed sequence tag，EST）有较高相似性的 cDNA 片段，其中一条与透明细胞癌来源的 EST 序列高度相

作者介绍：邹　波（1986-），男，硕士，
研究方向：口腔颌面部肿瘤。
通信作者：季　平，男，教授，Email：jiping_62@hotmail.com。
基金项目：重庆市卫生局科研基金资助项目（编号：2008-2-232）。

似，本实验对此 cDNA 片段实施电子克隆并进行相关的生物信息学分析。

1 材料与方法

1.1 GenBank 数据库检索和 EST 拼接

以测序所得 EST 序列作为种子序列，采用 GenBank 的 dbEST 数据库进行 BLAST 检索，检索出与种子序列有片段重叠、同源性高的 EST 序列，以 FASTA 格式下载到本地，利用 CAP3 软件将匹配序列拼接为连接体序列（Contig），再以此为种子序列进行 BLAST 检索，重复多次直至连接体序列无法再延伸，最后生成的 cDNA 序列即为此 EST 代表的未知基因序列。

1.2 开放阅读框分析和染色体定位

登陆 NCBI 开放阅读框（Open reading frame，ORF）分析系统（http://www.ncbi.nlm.nih.gov/gorf/gorf.html），采用标准遗传密码，对最终所得的 cDNA 序列进行开放阅读框分析；将此 cDNA 序列与基因组数据库比对，运用 Map Viewer 软件进行染色体定位。

1.3 基因编码蛋白的理化性质与分析

利用 ExPaSy 软件包中的 ProtParam 软件分析该 cDNA 编码蛋白的氨基酸组成、分子量和等电点等理化性质；利用 ProtScale 软件进行蛋白疏水性分析。

1.4 基因编码蛋白二级结构和功能区分析

登陆蛋白二级机构在线预测系统（http://npsa-pbil.ibcp.fr/cgi-bin/secpred_gor4.pl），对编码蛋白的二级结构进行预测。利用 SignalP-HMM 在线软件进行蛋白结构信号肽分析。利用 NCBI 中 CDD 数据库（Conserved domain database）预测基因编码蛋白的结构功能域。

1.5 同源性检索

将拼接后的 cDNA 序列与非冗余数据库进行 BLAST 比对，检索同源序列。

2 结 果

2.1 干细胞相关差异表达序列标签的电子克隆

经 BLAST 比对 ESTs 数据库得到差异表达序列标签的同源 EST，CAP3 软件拼接后得到 1 个长 1 705 bp 的连接 cDNA 序列；将此序列 BLAST 检索基因组数据库，此 cDNA 在染色体上跨度 100 kbp（231,649 K-231,749 kbp），基因组序列（NT_167186.1）涵盖此 cDNA 序列。利用 Map Viewer 软件将此 cDNA 序列在染色体上定位于 1q42（图 1）。

2.2 连接体序列的开放阅读框

ORF FINDER 在线软件分析结果显示，最大开放阅读框长度为 441 bp，编码 146 个氨基酸，其起始密码子 ATG 位于第 40 位核苷酸，下游第 480 位为终止密码子 TAG，此开放阅读框起始密码子 ATG 的+4 位碱基为鸟嘌呤 G，符合 Kozak 规律（图 2、3）。

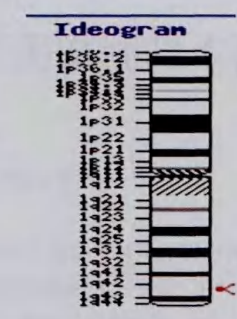


图 1 cDNA 染色体定位

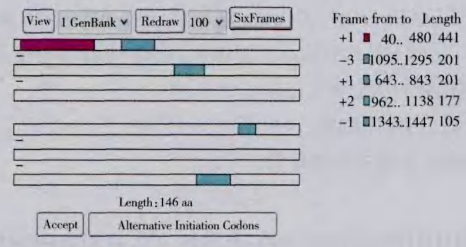


图 2 cDNA 开放阅读框预测

```
40  atggatgaaattaataaacaattgatatttacgactgaagacaat
    M  D  E  I  N  K  Q  L  I  F  T  T  E  D  N
85  gggaaagaaaataaaactccctcctctgatgcacaggataagcag
    G  K  E  N  K  T  P  S  S  D  A  Q  D  K  Q
130 tttggtacttggagactgagagtcacacctgtcgattaccttctg
    F  G  T  W  R  L  R  V  T  P  V  D  Y  L  L
175 ggagtggctgacttaactggagaattgatgcggatgtgtattaac
    G  V  A  D  L  T  G  E  L  M  R  M  C  I  N
220 agtgtgggggaatgggggacattgataccccctttgaagtgagccag
    S  V  G  N  G  D  I  D  T  P  F  E  V  S  Q
265 tttttacgtcaggtttatgatgggttttcattcattggcaacact
    F  L  R  Q  V  Y  D  G  F  S  F  I  G  N  T
310 ggaccttacgaggtttctaagaagctgtataccttgaaacaaagt
    G  P  Y  E  V  S  K  K  L  Y  T  L  K  Q  S
355 ttggccaaagtggagaatgcttgttatgccttgaaagtcagaggg
    L  A  K  V  E  N  A  C  Y  A  L  K  V  R  G
400 tcagaaattccaaaacatatgttggcagatgtgtttttcagttaaa
    S  E  I  P  K  H  M  L  A  D  V  F  S  V  K
445 acagaatgatagatcaagaagagggcatttcttag 480
    T  E  M  I  D  Q  E  E  G  I  S  -
```

图 3 最大阅读框的碱基序列

2.3 连接体 cDNA 编码蛋白的分子量和等电点

利用 ExPaSy 软件包中的 ProtParam 软件对连接体 cDNA编码蛋白的氨基酸组成、分子量和等电点等进行分析，结果显示，该蛋白分子质量为 16 397.5，等电点为 4.80，是一酸性蛋白；氨基酸组成以亮氨酸（Leu）居多，占 8.2%，其次天冬氨酸（Asp）、谷氨酸（Glu）、甘氨酸（Gly）分布较平均，各占 7.5%。

2.4 连接体 cDNA 编码蛋白的亲/疏水性分析

利用 ProtScale 软件的 Kyte and Doolittle 算法对编码蛋白进行亲/疏水性分析。正值越大表示越疏水，负值越大表示越亲水，介于+ 0.5~-0.5 之间的主要为两性氨基酸，结果提示大约在 42~51 区域的疏水性最强，其次为 130~134 区域；12~33 区域的亲水性最强，其次 92~97 区域具有较强的亲水性（见图 4）。

2.5 连接体 cDNA 编码蛋白的二级结构

利用 NPS@ 服务器上的 GOR4 程序对连接体 cDNA 编码蛋白的氨基酸序列进行二级结构预测（如图 5）发现：该蛋

白主要存在 3 种二级结构，即 43.84%的无规则卷曲(Random coil)，38.36%的 α-螺旋(Alpha helix)以及 17%的延伸链(Extended strand)。

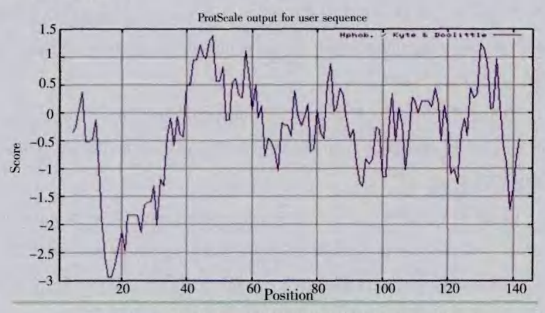


图 4　cDNA 编码蛋白的亲/疏水性分析图

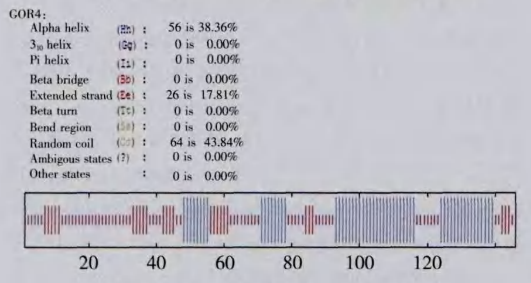


图 5　cDNA 编码蛋白的二级结构预测

2.6　编码蛋白的结构域与信号肽预测

CDD 数据库预测编码蛋白的结构域显示 38~121 位氨基酸存在与 Translin 超家族同源的区域(图 6)；信号肽分析显示(图 7)编码蛋白在信号肽、锚定蛋白及最大分割位点的概率都为 0.000，推测该蛋白不存在信号肽。

2.7　编码蛋白的同源性

经 NCBI 网站上 BLASTp 程序以及 SMART 数据可进行同源性蛋白质分析，结果显示，此编码蛋白与 Translin 相关蛋白 X(Translin associated protein X，TRAX)具有较高的同源性(图 8)。

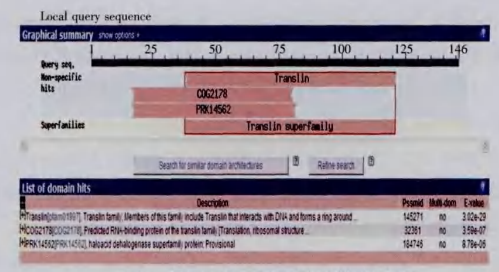


图 6　cDNA 编码蛋白的结构域预测

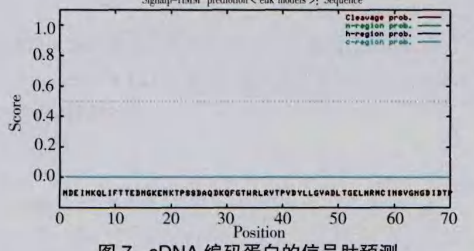


图 7　cDNA 编码蛋白的信号肽预测

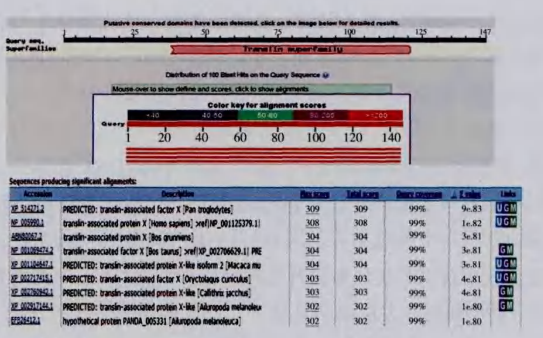


图 8　cDNA 编码蛋白的同源比对

3　讨　论

舌鳞状细胞癌是头颈部鳞癌的常见类型之一，近几十年来，尽管手术、放、化疗等治疗方法都取得了长足的发展，但常规治疗后仍然极易复发，对患者的生活质量造成严重影响，而且长期生存率也一直没有明显提高[8]。为了改善舌鳞癌的治疗效果，需要对其发生、发展的内在机制进行深入的了解，“肿瘤干细胞”理论的提出为理解肿瘤的发生提供了新的理念，该理论认为肿瘤可以看做一个异常发育的器官，其内部存在严格的等级体系，位于这个体系最顶端的即肿瘤干细胞，它们是肿瘤发生、复发及转移的罪魁祸首，开发针对肿瘤干细胞的新型治疗方法或许可以从根本上改变肿瘤治疗的现状。

现有的证据虽已证实多数的肿瘤中都有肿瘤干细胞的存在，但是由于缺乏特异的分子标志而不能精确地定位，目前对于肿瘤干细胞的研究更多的是从生物功能方面入手，如耐药、抗凋亡、sp 表型等。在口腔鳞状细胞癌中，利用不同的实验方法也已都富集出了具有干细胞性质的肿瘤细胞，最近的研究发现肿瘤组织中高表达 ALDH 的细胞都具有干细胞的性质，醛脱氢酶是一个解毒酶家族，可以将细胞内多种醛类氧化成相应的弱羧酸，从而减弱醛类对细胞的毒性，也是肿瘤细胞耐药的机制之一。本实验室利用 ALDH 作为标志物筛选舌鳞状细胞癌细胞株 Tca8113 中高表达醛脱氢酶的亚群细胞并证实了其具有干细胞的性质，为了进一步研究这群细胞表现特殊生物学特性的机制以及寻找肿瘤干细胞的特异标志物，我们利用抑制消减杂交技术构建了舌鳞癌干细胞样细胞与普通肿瘤细胞的差异表达基因文库，经过初步分析获得了 62 个已知基因以及一些与已知基因同源性较低的 EST 片段。本次实验我们对其中一个 EST 片段进行电子克隆，该片段与来源于透明细胞癌组织的一个 EST 片段具有较高的相似性。电子克隆又称“虚拟克隆”，是指利用计算机技术和网络资源(如核苷酸数据库、基因组数据库、ESTs 数据库、蛋白质数据库等)，采用生物信息学方法(如序列拼接、聚类分析、同源检索等)延伸 ESTs 序列，以期获得部分乃至全长 cDNA 序列的方法。与传统的原位杂交筛选 cDNA 文库或利用特异引物进行 cDNA 末端快速扩增相比，通过电子克隆

获得整个编码区、新基因或可变间接体的主要优势在于节省时间和节约经费，能够起到事半功倍的效果，可以加快对基因结构及功能的研究进程，推动比较基因组学的发展。虽然我们未检索到评估电子克隆与实际结果之间差异的相关文献，但其他实验室利用电子克隆技术获得的 cDNA 都得到了相关生物学实验的验证。如 Stolle 等[9]利用电子克隆的方法成功获得一个肿瘤相关基因 RASL11B，Yousef 等[10]利用此方法也成功获得了前列腺癌及乳腺癌相关基因 CAG。国内的管冬元等[11]利用此方法成功克隆了大鼠肝癌组织中高表达的 4 个新基因，作者还通过 RT-PCR 等方法对电子拼接的结果进行了验证，结果均支持电子拼接的结果。这些表明，EST 的电子克隆是克隆全长 cDNA 序列的有效途径[12,13]。本实验采用的 EST 经连续拼接后得到一个 1 705 bp 的 cDNA 序列，具有一个完整的开放阅读框，该开放阅读框编码 146 个氨基酸，功能域预测显示有一个与 Translin 超家族高度同源的区域，提示此蛋白也许是该家族成员或与之有密切关联，经 BLAST 同源比对，发现该蛋白与 TRAX 具有较高的同源性，这与功能域的预测结果一致。TRAX 是 Translin 蛋白的结合配偶体，二者是一对同源配对保守蛋白，有紧密的功能联系，TRAX 的稳定性依赖于 Translin 的存在，它们共同参与多种生理活动的调节，包括细胞生长、mRNA 加工、精子发生、神经发育及功能、基因组稳定调节和致癌等[14]。有研究发现 Translin 与寡聚核苷酸 d(GT)n 及 d(TTAGGG)n 有高度的亲和力，提示其参与端粒或微随体的新陈代谢[15,16]。还有研究报道Translin 和 TRAX 在 DNA 损伤修复中起作用，Translin 缺失的小鼠在接受非致死剂量辐射后，脾脏的造血集落形成延迟，而且Translin 和TRAX与DNA损伤应答相关蛋白结合[17]。因此我们推测该 EST 片段所代表的相关基因或许与干细胞端粒的新陈代谢、DNA 损伤修复等维持其自我更新能力的生理过程相关。因为 EST 电子克隆是以 EST 数据库为基础的，而数据库中 EST 数据的精确度为 97%[18]，意味着通过电子克隆获得的序列与实际序列可能存在一定的差异，所以该 EST 所代表的全长 cDNA 序列及其相关生物学功能还需要在后续的实验中进一步验证。

总之，我们对舌鳞癌干细胞相关差异表达基因文库中的一条 EST 序列进行了一系列的生物信息学分析，并获得了相应核酸及蛋白的基本信息，为进一步研究该 EST 在舌鳞癌中的表达及其生物功能提供了帮助。

参 考 文 献

[1] Tannishtha R, Lrving L.Weissman.Stem cell, cancer and cancer stem cell[J].Nature, 2001, 414(6859): 105-111.

[2] Al-Hajj M, Wicha M S, Benito-Hernandez A, et al.Prospective identification of tumorigenic breast cancer cells[J].Proc Natl Acad Sci U S A, 2003, 100(7): 3983-3988.

[3] Singh S K, Hawkins C, Clarke I D, et al.Identification of human brain tumour initiating cells[J].Nature, 2004, 432(7015): 396-401.

[4] O'Brien C A, Pollett A, Gallinger S, et al.Man colon cancer cell capable of initiating tumour growth in immunodeficient mice[J].Nature, 2007, 445(7123): 106-110.

[5] Zhang P, Zhang Y, Mao L, et al.Side population in oral squamous cell carcinoma possesses tumor stem cell phenotypes[J].Cancer Letter, 2009, 277(2): 227-234.

[6] Chen Y C, Chen Y W.Aldehyde dehydrogenase 1 is a putative marker for cancer stem cells in head and neck squamous cancer[J].Biochemical and Biophysical Research Communications, 2009, 385(3): 307-313.

[7] Chiou S H, Yu C C, Huang C Y, et al.Positive Correlations of Oct-4 and nanog in oral cancer stem-like cells and high-grade oral squamous cell carcinoma[J].Clin Cancer Res, 2008, 14(13): 4085-4095.

[8] Haddad R I, Shin D M.Resent advances in head and neck cancer [J].The New England Journal of Mdeicine, 2008, 359: 1143-1154.

[9] 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[J].Genome Research, 2001, 11(3): 422.

[10] Claverie J M.Computational methods for the identification of differential and coordinated gene expression[J].Hum Mol Genet, 1999, 8(10): 1821-1832.

[11] 管冬元，方肇勤，朱 希，等，大鼠肝癌高表达基因 cDNA 序列的克隆[J].实用肿瘤杂志，2008，23(5)：409-414.

[12] 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[J].Cell, 2001, 106(4): 413-415.

[13] Dworkin J, Losick R.Differential gene expression governed by chromosomal spatial asymmetry[J].Cell, 2001, 107(3): 339-346.

[14] Jaendling A, McFarlane R J.Biological roles of translin and translin-associated factor-X: RNA metabolism comes to the fore[J].Biochem.J, 2010, 429(2): 225-234.

[15] Jacob E, Pucshansky L, Zeruya E, et al.The human protein translin specifically binds single-stranded microsatellite repeats, d(GTn), and G-strand telomeric repeats, d (TTAGGG)n: a study of binding parameters[J].J.Mol.Biol, 2004, 344(4): 939-950.

[16] Cohen S, Jacob E, Manor H.Effects of single-stranded DNA binding proteins on primer extension by telomerase[J].Biochim.Biophys. Acta, 2004, 1679(2): 129-140.

[17] Fukuda Y, Ishida R, Aoki K, et al.Contribution of translin to hematopoietic regeneration after sublethal ionizingirradiation[J].Biol. Pharm.Bull, 2008, 31(2): 207-211.

[18] 何志英，姚玉成，胡以平.EST 技术及其在基因全长 cDNA 克隆上的应用策略[J].国外医学遗传分册，2002，25(2)：67-69.

（责任编辑：冉明会）