

人端粒重复序列结合因子基因组和假基因研究

孙 洁¹,黄 河¹,宋怀东²,吴昕彦²

(1. 浙江大学医学院 附属第一医院,浙江 杭州 310003;2. 上海瑞金医院,上海 200025)

[摘 要] 目的:确定人端粒重复序列结合因子(TERF1)的基因组结构和假基因的定位与结构。方法:利用 GenBank 等生物信息学资源,NCBI 提供的 BLAST 及其他相关的生物信息学工具(Sequencher,DNA Strider 和 Autoassembler 等),确定 TERF1 基因组和假基因的定位与结构,并用 PCR 和测序证实。结果:定位在 8q13 上的 TERF1 基因组由 10 个外显子构成。它有 4 个假基因分别分布在第 13、18、21 和 X 染色体上(Ψ TERF1-13、 Ψ TERF1-18、 Ψ TERF1-21 和 Ψ TERF1-X),其结构均与 TERF1 mRNA 相似,但缺少部分外显子。 Ψ TERF1-13、 Ψ TERF1-18 和 Ψ TERF1-21 的周边序列之间存在长度至少 60 kb 的同源片段。结论:TERF1 基因组含 10 个外显子,它有 4 个加工过的假基因分布在第 13、18、21 和 X 染色体上。大片段同源序列属近期重复片段中的染色体间倍增。

[关键词] 端粒重复序列结合因子; 外显子; 假基因

[中图分类号] R 394.5 [文献标识码] A [文章编号] 1008-9292(2003)05-0407-05

Study on the genome structure of human telomeric repeat binding factor 1 and its pseudogenes
SUN Jie, HUANG He, SONG Huai-dong, et al (*The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou 310003, China*)

[Abstract] **Objective:** To determine the genome structure of human telomeric repeat binding factor 1 (TERF1) and its pseudogenes. **Methods:** Sequences were obtained from GenBank and analyzed using the BLAST program and other relevant biology program (Sequencher,DNA Strider and Autoassembler,etc) to determine the genome and pseudogenome structure of TERF1. PCR and sequencing were performed to verify the results. **Results:** TERF1 gene,which mapped to 8q13 was divided into 10 exons. It had four processed pseudogenes located on chromosome 13, 18, 21 and X respectively(Ψ TERF1-13, Ψ TERF1-18, Ψ TERF1-21 and Ψ TERF1-X). They were entire intronless TERF1 genes which lacked some exons. Three homologous fragments of at least 60 kb on the flanking region of Ψ TERF1-13, Ψ TERF1-18 and Ψ TERF1-21, respectively were noted. **Conclusion:** TERF1 gene has 10 exons. It has four processed pseudogenes which are located on chromosome 13, 18, 21, and X, respectively. Large homologous fragments that belong to the recently duplicated segments are transchromosomal duplications.

[Key words] Telomeric repeat binding factor; Exon; Pseudogene

[J Zhejiang Univ (Medical Sci), 2003,32(5):407-411.]

端粒是真核细胞染色体的自然末端,是一种特殊的核蛋白结构,它由染色体 3'末端几百到几千个“TTAGGG”重复序列和端粒结合蛋白构成,具有保护染色体末端和保证复制正常进行的作用,与细胞周期和衰老有密切的关系^[1]。端粒酶是一种核糖核蛋白,它可以其分子内部的一段 RNA 为模板合成端粒重复序列而延长端粒^[2]。大量研究表明,端粒酶活性与肿瘤

密切相关,85%的人类晚期恶性肿瘤和 90%的

收稿日期: 2003-03-27 修回日期: 2003-05-24
基金项目:国家自然科学基金(39870339);浙江省自然科学基金(399141)。
作者简介:孙 洁(1974—),女,博士生,从事血液病学临床及基础研究。
通讯作者:黄 河(1961—),男,博士,教授,博士生导师,从事血液病学临床及基础研究;E-mail:hehuang@hzcnc.com

肿瘤细胞株和永生细胞系都具有端粒酶活性^[3]。人端粒重复序列结合因子(telomeric repeat binding factor 1, TERF1)是最早发现的端粒结合蛋白^[4],是重要的端粒酶活性调控因子,它负反馈抑制端粒酶活性从而抑制端粒酶延长端粒。据研究表明,TERF1 在急性白血病和消化道肿瘤细胞中的表达明显减低^[5,6],并认为 TERF1 与细胞周期调控有关^[7],因此 TERF1 可能是参与衰老和肿瘤发生发展的重要因子,对它的研究有助于寻找诊断和治疗肿瘤的新途径^[8]。然而,明确 *TERF1* 基因组结构是寻找 TERF1 在白血病和肿瘤细胞中是否突变的前提。已有的研究显示 *TERF1* 基因组定位于 8q13^[9],可能有 3 个假基因分别定位于第 13、21^[10]和第 X^[11]号染色体上,但基因组结构和假基因的定位和结构均未确定。本研究利用生物信息学技术预测了 *TERF1* 的基因组结构并发现了 4 个假基因,基因组和假基因的结构均经 PCR 和测序获得了证实。其中定位于第 18 号染色体的假基因属首次报道。

1 材料和方法

1.1 *TERF1* 基因组和假基因结构的分析预测 用 *TERF1* 的全长 mRNA 序列(NM-017489),对 htgs 数据库作 BLASTN 分析(Request ID:995786588-3337-6002),将得到的与之显著相似且有明确定位的 contig 与该 mRNA 用 Sequencher 软件分析,根据 GT-AG 原则确定了 *TERF1* 的 10 个外显子,以及分布在第 13、18、21 和 X 染色体上假基因(分别命名为 Ψ *TERF1*-13、 Ψ *TERF1*-18、 Ψ *TERF1*-21 和 Ψ *TERF1*-X)的结构。用 DNA Strider 软件分析假基因的蛋白编码框架。用 BLAST TWO SEQUENCE 软件分析假基因周边序列的同源性。用 Autoassembler 软件分析测序结果。

1.2 *TERF1* 外显子和假基因的引物设计和 PCR 扩增 用传统的酚-氯仿-异戊醇方法抽取正常人外周血细胞 DNA 作为 PCR 模板。在 8 号染色体上跨越 *TERF1* 的每个外显子设计并合成特异性引物,在 13、18、21 和 X 染色体上设计并合成针对 *TERF1* 的 4 个假基因的特异性引物(表 1)。

表 1 *TERF1* 外显子和假基因的 PCR 引物
Table 1 Primers used for amplifying exons and pseudogenes of *TERF1*

Primer	Primer sequence(5'-3')	Length of amplicons (bp)
exon1	GGGATTCTGAACAAAGTTCAGC	689
	AGACTGGCGGCAAATTC AAC	
exon2	TATGCACCAGACACTGTAGGC	285
	TTACTCAACAAACCACACAGTGG	
exon3	CAAGAAGAGTACCTAGCACATAGTATACAC	376
	ACCAAAGGTCAAATTTCTGAAGAG	
exon4	CTCAACAGAGGATTTTCAGACTTACCTG	382
	AGGTGTAAGCTACTGCGTCCAGC	
exon5	TGAGGAGAGGCCAGAAGAAGTC	464
	CACACCTTTCTACAAACCACAATG	
exon6	AAAGACCACCTTCGTACGCTG	505
	TGCACAAGTCATGAAATAACTCTACAG	
exon7	TGCTTGACTACACCATTAATTATGC	430
	TTGGGCTAACTAGGACCCTGAG	
exon8	GGACACATCAATTCTCCCTTC	205
	TTTACCTGTTTTCCATTGCCTTAC	
exon9	CTTTAATGTCCATAAGCGTTATTTCG	318
	TGATCTTTGCACTATACCATACTGC	
exon10	ACCCTCATTAATTGATGGTCACAG	451
	GAAATGTTTAGAGAAGTCTTCCGTC	
Ψ <i>TERF1</i> -13	CCATGGGTCTCTCAGGTTAATAG	998
	GGGAAACTGGTCTAAAATACTACTGC	
Ψ <i>TERF1</i> -18	TCACCGTTTTAGCCGGGATG	1708
	AGCTGACAAACCTGCCCATG	
Ψ <i>TERF1</i> -21	GTGAAAGGAAGTGGAACATGTG	1635
	GTCGGGTTGTTGAATTTATAATTCCC	
Ψ <i>TERF1</i> -X	CTGTTCAGCTGCTGATAACATCTG	506
	ATTGAACCCCAAATCATCAGG	

PCR 反应总体积为 25 μ l,反应体系如下: 10 mmol/L Tris-HCl (pH 9.0)、10 mmol/L KCl、10 mmol/L (NH₄)₂SO₄、2 mmol/L MgCl₂、0.2 mmol/L dNTPs、0.1%的 Triton X-100、200 μ mol 引物、50 ng DNA 模板和 1.5 U 的高保真 Taq 酶(Taq plus polymerase,上海生工)。扩增及循环条件:解链 94℃ 30 s,退火温度在扩增 exon 2、exon 7、exon 9 和 exon 10 为 55℃,在 exon 1、exon 3、exon 6、exon 8、 Ψ *TERF1*-13、 Ψ *TERF1*-21 和 Ψ *TERF1*-X 为 56℃,在 exon5 为 58℃,exon 4 为 60℃, Ψ *TERF1*-18 为 65℃;退火时间均为 45 s,延长 72℃ 1 min,共 32~35 次循环。PCR 反应在 PE 9600 PCR 热循环仪上进行。

1.3 *TERF1* 外显子和假基因 PCR 产物的测序 外显子 PCR 反应产物切胶后用 QIAquick Gel Extraction Kit (QIAGEN) 纯化后直接测序。假基因 PCR 反应产物切胶纯化后克隆入 pGEMT-easy 测序载体 (Promega) 后测序。测序在 PE 公司的 ABI PRISM 377 DNA 测序仪 (Perkin Elmer) 上进行 (上海瑞金医院分子医学和人类基因中心提供)。

2 结 果

2.1 *TERF1* 的基因组结构 定位于 8q13 (AC022893) 上的 *TERF1* 基因组总长度 38.6 kb, 由 10 个外显子构成。外显子和内含子的接合边界已确定 (图 1)。PCR 反应进一步证实了 10 个外显子的存在 (图 2)。测序结果与 GenBank 预测的序列完全一致。

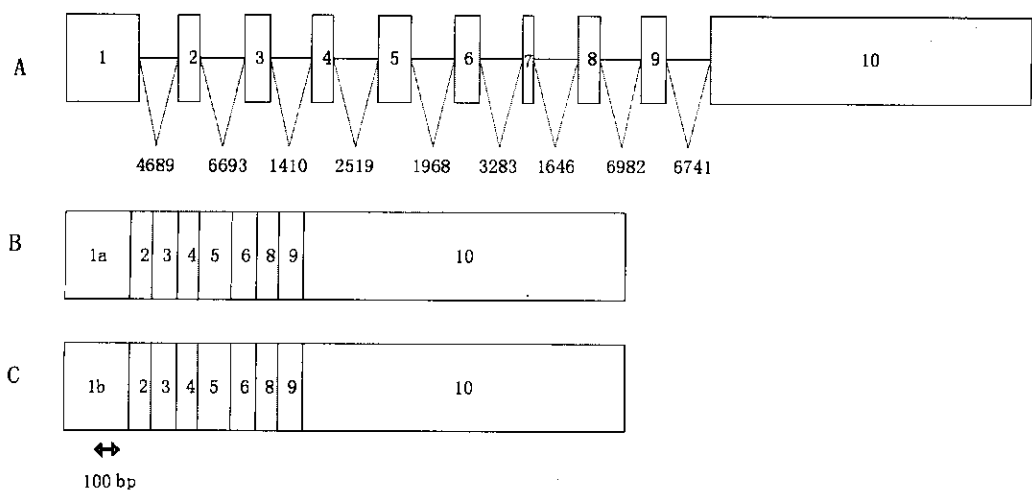


图 1 *TERF1* 基因组及其假基因的结构

Fig. 1 The structures of *TERF1* genome and its pseudogenes

A: The structure of *TERF1* genome. Exons are represented by numbered boxes, introns by solid lines, sizes of exon 1~10 is 332,96,122,87,150,113,60,92,104,1532 bp respectively; introns' sizes (bp) were shown below; B: The structure of Ψ *TERF1*-13, 18 and 21. 1a was long as about 271 bp; C: The structure of Ψ *TERF1*-X, 1b was long as about 296 bp

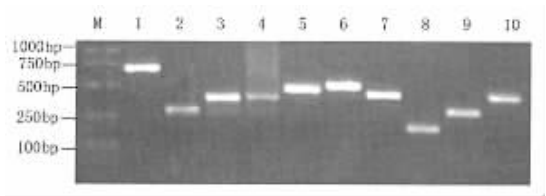


图 2 *TERF1* 的 10 个外显子的 PCR 扩增产物

Fig. 2 PCR amplicons from the 10 exons of *TERF1* gene
M: DNA marker DL 2 000 (TaKaRa); 1~10: Represent exon1~10, respectively

2.2 假基因定位与结构 *TERF1* 有 4 个假基因分别分布在第 13、18、21 和 X 染色体 (GenBank 号分别为: AL359737; AC053538; L050303 和 AL137013), 分别命名为 Ψ *TERF1*-13、 Ψ *TERF1*-18、 Ψ *TERF1*-21 和 Ψ *TERF1*-X。这 4 个染色体上的序列与

*TERF1*cDNA 相似, 均不含内含子, 同时缺乏部分外显子; 所有的同源序列都缺少第 7 号外显子, 长 60 bp, 是所有外显子中最短的; 同时还缺少 1 号外显子的部分序列, 而且同源序列与 1 号外显子的同源性较与其他外显子低; 此外, 这些同源序列还存在个别碱基的插入、缺失和替换 (图 1)。它们的蛋白质编码框架与 *TERF1* 不同, 没有一个序列能编码有功能的 *TERF1* 蛋白。此外, 本实验还发现在 X 染色体上的同源序列的 3' 端相邻区域有连续 25 个 A 的多聚腺嘌呤结构, 而其它序列的 3' 端相邻区域未发现有典型的多聚腺嘌呤结构, 但有富含 A 的序列。PCR 结果证实了假基因的存在 (图 3)。测序结果也与 GenBank 中的相应序列吻合。

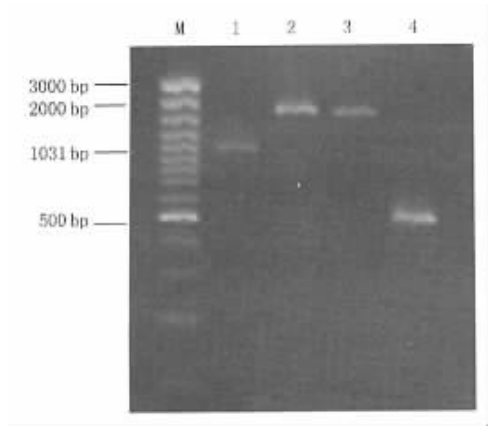


图 3 *TERF1* 的 4 个假基因的 PCR 结果

ig. 3 PCR results of the four pseudogenes of *TERF1* M; GeneRulerTM 100 bp DNA Ladder Plus (MBI Fermentas), 1 ~ 4; Represent PCR amplicons of Ψ *TERF1*-13, Ψ *TERF1*-18, Ψ *TERF1*-21 and Ψ *TERF1*-X, respectively

2.3 假基因周边的大片段同源序列 本实验用 BLAST TWO SEQUENCES 比较假基因的周边序列时,发现在 13 号染色体、18-contig 和 21-contig 上有大片段同源序列,其中包含 Ψ *TERF1*-13, Ψ *TERF1*-18 和 Ψ *TERF1*-21 (图 4)。其中在 13 与 18 号染色体之间及 21 与 18 号染色体之间有近 60 kb 左右的同源片段,13 和 21 号染色体之间有约 93 kb 的同源片段,同源性均在 96% 以上。

3 讨 论

TERF1 是重要的端粒结合蛋白,它具有负调控端粒酶活性和抑制端粒延长的作用,与肿瘤和衰老密切相关。但是 *TERF1* 基因组结构尚不清楚。GenBank 虽然通过计算机软件预测了 *TERF1* 的基因组结构,但是未获得证实,且不能排除现有的基因组数据库存在错误的可能性。本实验通过生物信息学软件预测发现 *TERF1* 的基因组由 10 个外显子和 9 个内含子组成,并通过 PCR 和测序证实和确定了外显子和内含子的接合边界。该结果与 GenBank 预测的结果一致。

在预测 *TERF1* 基因组结构的过程中本研究发现了分布在 13、18、21 和 X 染色体上的 *TERF1* 的 4 个假基因结构并通过测序证实了

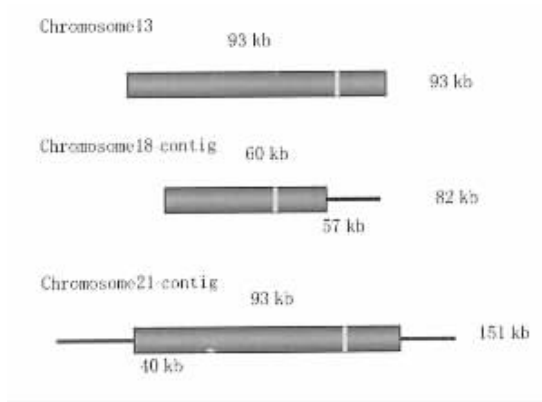


图 4 包含 *TERF1* 假基因的分布在第 13、18 和 21 染色体(或测序单位 contig)上的大片段同源序列

Fig. 4 Large homologous fragments on chromosome 13, 18-contig and 21-contig, which contain the pseudogenes of *TERF1*

Contigs of the chromosomes are shown as black lines; the length of the correspond contig is shown on the right. The homologous fragments are shown as black bars, their length is indicated by the numbers above, their location on the contigs is indicated approximately by the numbers at the initial or terminal sites. The white lines on the bar demonstrate the *TERF1* pseudogenes

假基因的存在。假基因是与功能性基因密切相关的 DNA 序列,但由于缺失、插入和无义突变失去阅读框架而不能编码蛋白质产物。假基因由于产生的机制不同分两型^[12]。I 型是基因倍增的产物,后因基因突变而失去功能;II 型是有功能基因的 mRNA 反转录成 cDNA 后插入基因组中形成的,后又逐渐被突变修饰,又称为被加工过的假基因(processed pseudogene),往往表现为不含内含子,3'末端有多聚腺嘌呤。*TERF1* 的 4 个假基因均不含内含子,缺乏部分外显子,此外还存在个别碱基的插入、缺失和替换,3'区有多聚腺嘌呤序列或富 A 序列。从它们的蛋白编码框架看所有的假基因均不能编码有功能的 *TERF1* 蛋白,属典型的被加工过的假基因。假基因由于不能翻译成有功能的蛋白质,一直被认为是无意义的序列,但最新的研究显示一些假基因具有保护功能基因转录的 RNA 免受破坏的作用^[13]。在本研究对所发现的假基因的功能还不明确,但了解假基因的结构有助于在对功能基因的研究工作中避免受假

基因的干扰。

本实验还发现在 $\Psi TERF1-13$, $\Psi TERF1-18$ 和 $\Psi TERF1-21$ 的周边序列上存在大片段的高同源序列。人类基因组计划的研究显示约有 5% 的人类基因组散布着起源在距今 35 百万年前的倍增序列^[14]。这种近期倍增序列分为 2 种:非同源染色体之间的片段倍增(染色体间的倍增序列)和仅限于染色体内的倍增(染色体特异性倍增)。这类倍增片段往往很大(1~100 kb),且往往具有高度的同源性(约 95%)。这类片段可能是因为不断发生的染色体重组而产生,是人类基因组进化的重要推动力。与染色体特异性倍增不同,多数染色体间的倍增片段并不编码有功能的蛋白,但是存在很多加工过的假基因,且多分布在染色体近中心或近端粒的区域。本实验发现的分布在第 13、18 和 21 号染色体上的同源序列长达 60 bp 或 93 kb,同源性为 96%。随着对 18 号染色体测序工作完成,在第 13 号染色体与 18-contig、21-contig 与 18-contig 之间的同源序列也可能达到 93 kb 长。所有的片段均分布在染色体的近中心区域。对片段做 BLASTX 后并未发现片段内存在功能基因,但是含有 *TERF1* 的加工过的假基因,认为分布在第 13、18 和 21 号染色体上的大片段同源序列属于染色体间的倍增序列。

References:

[1] BLACKBURN E H. Structure and function of telomeres [J]. *Nature*, 1991,350(6319):569—573.

[2] SHIPPEN-LENTZ D, BLACKBURN E H. Functional evidence for an RNA template in telomerase [J]. *Science*, 1990,247(4942):546—552.

[3] KIM N W, PIATYSZEK M A, PROWSE K R, et al. Specific association of human telomerase activity with immortal cells and cancer [J]. *Science*, 1994, 266 (5193):2011—2015.

[4] CHONG L, STEENSEL B V, BROCCOLI D, et al. A human telomeric protein [J]. *Science*, 1995, 270 (5242):1663—1667.

[5] YAMADA K, YAJIMA T, YAGIHASHI A, et al. Role of human telomerase reverse transcriptase and telomeric-repeat binding factor proteins 1 and 2 in human hematopoietic cells [J]. *Jpn J Cancer Res*, 2000, 91 (12):1278—1284.

[6] ARAGONA M, BUDA C A, PANETTA S, et al. Immunohistochemical telomeric-repeat binding factor-1 expression in gastrointestinal tumors [J]. *Oncol Rep*, 2000, 7(5):987—990.

[7] KISHI S, WULF G, NAKAMURA M, et al. Telomeric protein Pin2/TRF1 induces mitotic entry and apoptosis in cells with short telomeres and is down-regulated in human breast tumors [J]. *Oncogene*, 2001, 20 (12): 1497—1508.

[8] ARAGONA M, PONTORIERO A, PANETTA S, et al. The role of telomere-binding proteins in carcinogenesis [J]. *Minerva Med*, 2000, 91(11—12):299—304. [in Italian]

[9] BROCCOLI D, CHONG L, OELMANN S, et al. Comparison of the human and mouse genes encoding the telomeric protein, TRF1: chromosomal localization, expression and conserved protein domains [J]. *Hum Mol Genet*, 1997, 6(1):69—76.

[10] YOUNG A C, CHAVEZ M, GIAMBERNARDI T A, et al. Organization and expression of human telomere repeat binding factor genes [J]. *Somat Cell Mol Genet*, 1997, 23(4):275—286.

[11] WILSON S. Human DNA sequence from clone RP11-311P8 on chromosome Xq 13. 1—13. 3 [EB/OL]. 2000, Genbank AL137013. 4 [GI:8217489].

[12] VANIN E F. Processed pseudogenes: characteristics and evolution [J]. *Annu Rev Genet*, 1985, 19:253—272.

[13] HIROTSUNE S, YOSHIDA N, CHEN A, et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene [J]. *Nature*, 2003, 423(6935):91—96.

[14] EICHLER E E. Recent duplication, domain accretion and the dynamic mutation of the human genome [J]. *Trends in Genetics*, 2001, 17(11):661—669.

[责任编辑 黄晓花]