人类基因组中的反转录转座子*

刘新文 童坦君[1)] 张宗玉
(北京医科大学生物化学与分子生物学系，北京 100083)

摘要 人类基因组中有35%以上的序列为转座子序列. 反转录转座子是引起人类疾病的潜在病因. 人类基因组中的主导转座子——L1反转录转座子内部有二个开放读框，其编码蛋白具有RNA结合蛋白、反转录酶和内切酶活性. L1可能通过靶引物反转录机制整合到染色体中；Alu等非自主性反转录转座子可能利用L1反转录酶的反式互补作用进行转座.

关键词 反转录转座子，人类基因组，人类疾病，反转录转座
学科分类号 Q527

通过大规模的人类基因组序列测定及序列分析，已发现人类基因组中仅有3%的序列为外显子序列，其余均为“无用”DNA (junk DNA，包括内含子、简单重复序列和可转移成分等). 其中可转移成分占人类基因组总序列的35%以上[1~3].

1 人类基因组中的可转移成分

根据序列构成及转座方式不同，人类基因组中的可转移成分可划分为三类：DNA转座子、自主性反转录转座子和非自主性反转录转座子[1].

DNA转座子 (DNA-based transposable elements) 与细菌转座子结构类似，两端有反向重复序列 (inverted repeats, IR)，内部有转座酶编码序列，通过剪贴 (cut and paste) 或复制 (copy and paste) 方式进行转座. 在人类基因组中，DNA转座子约占1.6%[4]，其中以“水手”成分 (mariner element) 最为常见. 虽然人体内目前尚未发现活性的“水手”成分，Ivics等研究结果提示人体内可能存在活性DNA转座子. Ivics等[5]根据真骨类鱼不同种间转座子转座酶同源序列的系统比较，构建了一种新的人工“水手”转座子——“睡美人”(“sleeping beauty”). 他门利用“睡美人”的转座酶序列作为辅助载体 (helper) 的成分，与重组质粒pT/neo (含鲑鱼转座子序列IRs序列、启动子和报道基因neo) 共同转化HeLa细胞，在G418筛选出的转化细胞基因组中检测出neo基因序列，说明neo基因已“剪贴”到HeLa细胞染色体中.

自主性反转录转座子 (autonomous retrotransposons) 内部含逆转录酶编码序列，通过DNA-RNA-DNA方式进行转座，即转座子转录产生相应的RNA，再经逆转录生成新的转座子DNA并整合到基因组中. 根据结构中有无长末端重复序列 (long terminal repeats, LTRs)，自主性反转录转座子又可以分为二类：含LTRs的转座子和不含LTRs的转座子. 人体中含量最多的LTRs自主性反转录转座子为人类内源性逆转录病毒 (human endogenous retroviruses, HERVs)，约占1%~2%[6]. 虽然还未发现体内HERVs具有反转录转座能力，但已发现HERVs中有些元件活性表达，提示人体内可能存在活性HERVs. 不含LTRs的自主性反转录转座子主要为长散布元件 (long interspersed elements, LINE)，全长5~7 kb，称为L1. 人类基因组中有10万份以上的L1，约占总序列的15%左右[4]，主要分布于A-T富集区. L1是进化上保守的一个超基因家族，5′、3′端为非编码区 (untranslated region, UTR)，中间含二个开放读框 (open reading frame, ORF). 由于95%以上的L1 5′端存在缺失，其中约有10% L1还存在重排，因此具有序列全长及转座活性的L1非常少见[4]. 据估计人类基因组中约有3000~4000份L1具有全长序列，可能只有30~60份L1具有反转录转座能力. 目前已发现7份活性L1[7].

非自主性反转录转座子 (non-autonomous retrotransposons) 也通过DNA-RNA-DNA方式转座，但本身不能编码蛋白质，因而必须借助胞内酶才能实现转座. 这类转座子序列长100~300 bp，拷贝数达数十万份. 人类基因组中的非自主性反转录转座子主要为Alu家族 (70万份左右，占10%，主要分布于G-C富集区) 和加工后的假基因

* 国家自然科学基金资助重点项目 (39930170).
1) 通讯联系人.
Tel: (010) 62091454; E-mail: biochem@mail.bjmu.edu.cn
收稿日期：1999-01-04，修回日期：1999-05-15

(processed pseudogenes), 结构中均无 LTR, 无编码蛋白质能力, 但 3′端有长短不一的 poly A 尾序列[8,9]. 据推测, Alu 等非自主性反转录转座子转座所需的胞内逆转录酶很可能由活性 L1 编码[1,9].

鉴于人类基因组中含有丰富的反转录转座子, 1/4 以上的基因组序列来源于活性 L1 直接转座或由 L1 逆转录酶辅助 Alu 等非自主性反转录转座子转座[4], 目前认为 L1 是人类基因组中的主导转座子 (master human mobile element)[1].

转座子的移动与果蝇和酵母核型变化相关, 而且转座子通过特殊方式重新组织 DNA 分子, 提高生物的突变率, 因而有可能创建更好的遗传信息表达方式, 促进生物物种进化. 但转座子在进化上到底起多大作用仍然有待进一步研究[2,3].

2 反转录转座子与人类疾病

1988 年, Kazazian 等[10]首次证实反转录转座子是人类致病的潜在原因. 他们发现二段缺失突变的 L1 插入到凝血因子 Ⅷ基因中, 阻断该基因的表达, 引发了血友病 A. 随后, 相继在凝血因子 Ⅷ基因 (1 种)、Duchenne 型肌营养不良 (Duchenne muscular dystrophy, DMD) 基因 (3 种)、多发性结肠腺癌 (adenomatous polyposis coli, APC) 基因 (1 种) 和β珠蛋白基因 (1 种) 中发现了其他 6 种 L1 插入片段, 其中 5 种发生于生殖细胞或早期发育细胞, 1 种 (APC 中的 L1 片段) 发生于克隆的癌细胞, 提示 L1 反转录转座子有可能在体细胞中也起作用[1].

L1 插入片段有以下结构特点:

a. 各种插入片段的序列互不相同, 说明分别来源于不同的祖先;

b. 上述 8 种插入片段中, 7 种片段 5′端有 538 bp~ 3.8 kb 缺失, 1 种 (β珠蛋白基因中的 L1) 具有全长序列 (6.0 kb), 而所有 8 种 L1 片段的开放读框 (ORF) 均完好无损, 说明插入片段可能有反转录转座活性. 目前已经发现小鼠全长序列的 L1 插入片段能够进行转座;

c. 上述 8 种插入片段中, 7 种属于 L1 超家族中的 Ta 亚类, 表明活性 L1 可能主要来源于 Ta 亚类. Sassaman 等[7]从人类基因库中分离出 13 个全长序列的 Ta, 其中 3 种即具有转座活性.

除 L1 外, Alu 也是人类疾病的病因之一. 1994 年, Makalowski 等[11]从人类 DNA 序列数据库中寻找编码区中的 Alu 序列, 结果在 15 种不同外显子中找到了 17 种插入片段, 其中 3 种为已发现的致病 Alu. 到目前为止, 至少已发现 9 种全长 Alu 插入片段与人类疾病有关. Alu 插入片段发生于生殖细胞或早期发育细胞. Alu 家族中的 Alu-Sb1 和 Alu-Sb2 亚类是活性 Alu 的主要来源[1,11].

3 活性 L1 和 Alu 的结构

典型的活性 L1 结构如图 1a, 5′、3′端为非编码序列 (UTR), 内部有二个开放读框 ORF1 和 ORF2. 5′UTR 含有一个内部启动子, 3′UTR 含一个非典型的加尾信号和 polyA 尾序列. 二个开放读框之间为 66 bp 的间隔序列. ORF1 序列和反转录病毒 gag 基因 (group specific antigen gene) 相似, 编码一种 RNA 结合蛋白 P40, 特异地结合于 L1RNA ORF2 的 5′端附近. ORF2 上游序列与脱嘌呤嘧啶 (apurinic/apyrimidinic, AP) 内切核酸酶基因相似[12], 下游序列与反转录病毒及其他反转录转座子的逆转录酶序列同源. ORF2 蛋白是一种双功能蛋白质, 近 N 端有内切酶结构域 (EN), 近 C 端为逆转录酶结构域 (RT). 纯化的 EN 蛋白具内切酶活性, 产生 5′-PO_4、3′-OH 切刻. L1 内切酶具有靶位点优先性, 优先切割 $(Py)_n/(Pu)_n$ 位点 ("/" 为切点), 但该酶无 AP 位点优先且转换数很低. 有趣的是, L1 逆转录酶中的大量序列与端粒酶催化亚基序列相似, 但二者间关系如何有待研究. 此外, ORF2 蛋白 (EN/RT) 也可与 L1RNA 结合[1,7,9,12].

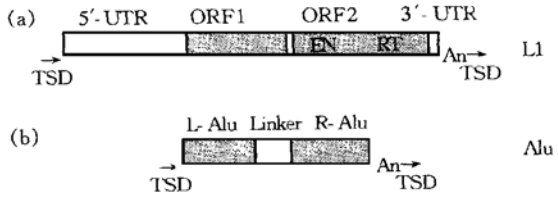

图 1 反转录转座子的结构

UTR: 非编码序列; TSD: 靶位点重复序列.

典型 Alu 长 282 bp, 由左右二部分同源序列组成 (图 1b), 其间为富含腺苷酸的连接区 (A-rich linker). Alu5′端含内部启动子, 3′端有 polyA 尾序列. 由于左右同源序列均来自内部缺失或点突变, 因而实质上 Alu 是 "获得 polyA 尾序列的 7sL RNA 假基因的二聚体"[9,11].

4 反转录转座机理

L1 是人类基因组中的主导转座子, 其转座机

理受到广泛重视. 目前认为, L1 通过 ORF 蛋白顺式结合 (*cis*-binding) 及靶引物反转录 (target-primed reverse transcription, TPRT) 机制转座, 其过程如下[1,9]:

a. 如图 2 所示, 活性 L1 由 5′端内部启动子起始转录, 生成相应的双顺反子转录本 (含 ORF1 和 ORF2 的对应序列). L1RNA 的长度及其 3′端的 poly A 尾结构表明 L1RNA 的生成可能由 RNA 聚合酶Ⅱ催化. 此外发现转录因子 YY1 结合于内部启动子, 其功能不明. 有文献报道, 分化细胞中 CpG 的甲基化可抑制转座子表达, 进而阻抑转座活性[2,3,13]. 因此目前认为 L1 转录仅限于低甲基化的未分化细胞, 如: 早期生殖细胞及未分化的肿瘤细胞.

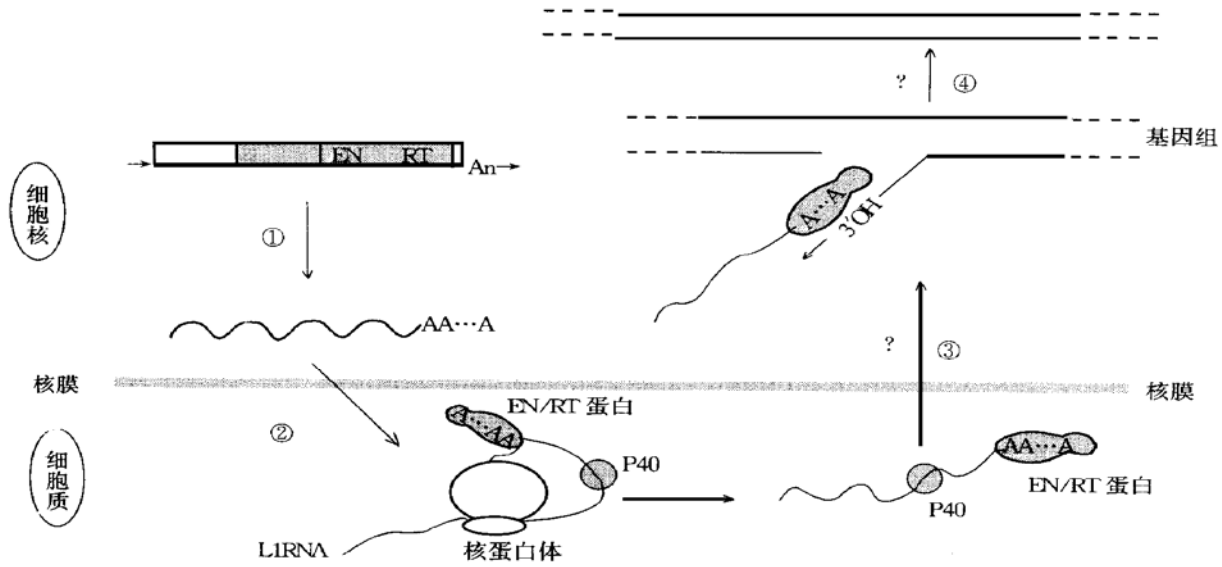

图 2 L1 的转座子机理

b. L1RNA 由核内进入胞质, ORF1、ORF2 转译分别生成 P40 和 EN/RT 蛋白. P40 与 L1RNA ORF2 相应序列的 5′端附近结合, EN/RT 蛋白可能与 L1RNA 3′端的 poly A 尾序列结合, 形成核蛋白颗粒 (ribonucleoprotein particles, RNPs). 利用多克隆抗体很容易从未分化培养细胞和肿瘤细胞中检测出 P40, 且已经从人和小鼠细胞中分离出 P40 和 L1RNA 结合形成的 RNP. 已证实在 HeLa 细胞 L1 转座中, P40、L1 内切酶活性及逆转录酶活性均起重要作用. Moran 等还发现 L1RNA 的 poly A 尾序列在 L1 转座中非常重要.

已有许多实验证明, L1 蛋白质优先与编码该蛋白质的同一 L1RNA 亲和结合, 称为顺式优先 (*cis* preference). 顺式优先允许 ORF 蛋白一边合成一边与 L1RNA 结合, 从而有利于 RNP 的形成, 提高了转座效率. 同时也发现少量蛋白质可与胞内的其他 RNA 结合. 在 HeLa 细胞中, 活性 L1.2 可以辅助胞内少量 mRNA 进行反转录.

c. RNP 由胞质到达靶 DNA 序列. 目前还不清楚在 RNP 转运过程中是否 P40 和 ORF2 蛋白同时起重要作用; 或是 L1RNA 和 ORF2 蛋白组成 RNP, 在细胞核分裂后被动到达染色质区. 由于 ORF2 蛋白难以检测, 目前还未证实核内 ORF2 蛋白的存在.

d. RNP 到达靶序列后, L1RNA 可能通过靶引物反转录生成 L1DNA 并整合到染色体中. 即, 首先由 L1 内切酶切割靶位点产生 3′-OH, 然后 L1 逆转录酶以 L1RNA poly A 为模板, 从 3′-OH 开始引发反转录 (靶缺刻引发, target nick-priming), 合成 L1DNA, 新合成的 DNA 链右端与染色体 DNA 相连. 已经有文献报道在体外实验中发现昆虫无 LTR 反转录转座子通过 TPRT 机制进行转座. 可能由于 L1 逆转录酶的持续合成能力差 (poor processivity), 新合成的转座子 5′端常出现缺失突变.

如上文所述, ORF 蛋白主要与 L1RNA 顺式结合, 但并不排除 ORF 蛋白与其他含 poly A 尾序列的 RNA 结合, 通过反式互补方式辅助其他 RNA 反转录形成假基因而转座[9].

已经发现, Alu 和 L1 在结构上有许多相似之处, 除 3′端具有 poly A 尾序列之外, 二者末端均具有相似的靶位点重复序列 (target-site

duplications, TSDs), 提示二者由相同的内切酶参与转座. 此外, Alu 侧翼序列含有保守的内切酶切点基序 (motif), 该基序与 L1 内切酶优先切割序列相似, 可能形成与 L1RNA 相似的二或三级结构, 从而与 L1 蛋白识别结合. 但是低效结合的反式互补模式难以解释 Alu 在人类基因组中的拷贝数达 50~ 100 万份, 比 L1 的拷贝数 (1~ 8 万份) 还多. 为此, Boeke 提出了另一种作用模式, 认为 Alu RNA 通过拓朴异构上的顺式作用与 L1 ORF2 蛋白结合而实现转座. 如前所述, Alu RNA 实质上为获得 poly A 结构的 7sL RNA 二聚体, 而 7sL RNA 是信号识别颗粒 (signal recognition particle, SRP) 的重要组分. 有文献报道, Alu RNA 与二种胞内 SRP 蛋白 SRP9 和 SRP14 高亲和结合 (SRP9/14 结合位点可能位于 Alu 的左侧同源区), 形成核蛋白复合物. Alu RNA 很可能借助 SRP9/14与核蛋白体的相互作用而结合于核蛋白体大亚基的特定位置, 使 Alu RNA 的 poly A 尾得以在核蛋白体出口附近摆动, 从而以近似顺式的方式与 L1 ORF2 蛋白高效结合 (图 3)[9]. Alu RNA 与 L1 ORF2 蛋白结合后如何转座到染色体中则可能有多种机制. Makalowski 等[11]从数据库中找出的 17 种插入片段中, 2 种 Alu 片段经反转录直接插入基因组, 而 12 种 (71%) 证明为剪接介导 (splice-mediated events) 的插入事件, 3 种机制未明.

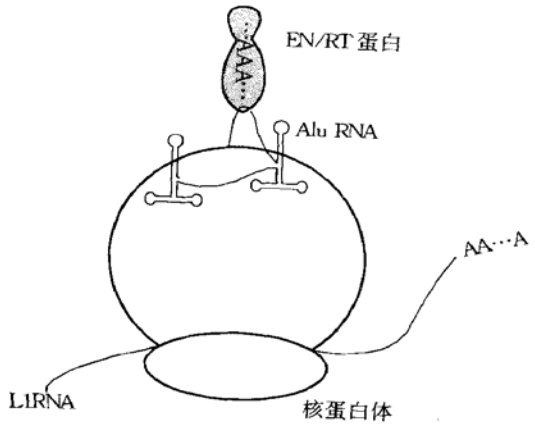

图 3 Alu RNA 与 L1 ORF2 蛋白结合模式

5 展 望

L1 与人类疾病的关系值得重视. 人类 L1 在小鼠 LTK 细胞中具有高转座活性, 因而有可能利用 L1 筛选出具有理论和实践意义的转基因小鼠. 此外, L1 作为载体在基因治疗与基因转移中的作用亟待开拓.

参 考 文 献

1 Kazazian H H, Moran J V. The impact of L1 retrotransposons on the human genome. Nature Genetics, 1998, **19** (1): 19~ 24
2 Kidwell M G, Lisch D R. Transposons unbound. Nature, 1998, **393** (6680): 22~ 23
3 刘新文, 童坦君. 真核生物中的基因流动现象. 生理科学进展 (Liu X W, Tong T J. Prog Physiol Sci), 1998, **29** (4): 324
4 Smit A F. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev, 1996, **6** (6): 743~ 748
5 Ivics Z, Hackett P B, Plasterk R H, *et al*. Molecular reconstruction of Sleeping Brauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell, 1997, **91** (4): 501~ 510
6 Lower R, Lower J, Kurth R. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Pro Natl Acad Sci USA, 1996, **93** (11): 5177~ 5184
7 Sassaman D M, Dombroki B A, Moran J V, *et al*. Many human L1 elements are capable of retrotransposition. Nature Genetics, 1997, **16** (1): 37~ 43
8 Britten R J. Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. Pro Natl Acad Sci USA, 1994, **91** (13): 6148~ 6150
9 Boeke J D. Lines and Alu——the poly A connection. Nature Genetics, 1997, **16** (1): 6~ 7
10 Kazazian H H, Wong C, Youssoufian H, *et al*. Heamophilia A resulting from de novo insertion of L1 sequences represents a novel mechamism for mutation in man. Nature, 1988, **332** (6160): 164~ 166
11 Makalowski W, Mitchell G A, Labuda D. Alu sequences in the coding regions of mRNA: a source of protein variablity. Trends in Genetics, 1994, **10** (6): 188~ 193
12 Feng Q, Moran J V, Kazazian H H, *et al*. Human L1 retrotransposon encodes a conserved endonuclease required to retrotransposition. Cell, 1996, **87** (5): 905~ 916
13 Yoder J A, Walsh C P, Bestor T H. Cytosine methylation and the ecology of intragenomic parasite. Trends in Genetics, 1997, **13** (8): 335~ 340

Human Retrotransposons. LIU Xin-Wen, TONG Tan-Jun, ZHANG Zong-Yu (*Department of Biochemistry and Molecular Biology, Beijing Medical University, Beijing* 100083, China).

Abstract At least 35% of the human genome is made up of transposon DNA. Retrotransposons are potential causal agents of human disease. The mast human mobile element, L1 retrotransposon, has 5′, 3′-UTR and two ORFs which encode a sequence-specific RNA-banding protein and a protein containing an endonuclease (EN) domain and a reverse transcriptase (RT) domain. It's likely that L1 undergoes target-primed reverse transcription in order to carry out retrotransposon. The mobilization of the non-autonomous retrotransposons, such as Alu and processed pseudogens, require a cellular source of reverse transcriptase, which is most likely encoded by L1.

Key words retrotransposon, human genome, human disease, retrotransposition