15. Spaargaren, M. & Bos, J. L. Rab5 induces Rac-independent lamellipodia formation and cell migration. *Mol. Biol. Cell* **10,** 3239–3250 (1999).
16. Benmerah, A., Bayrou, M., Cerf-Bensussan, N. & Dautry-Varsat, A. Inhibition of clathrin-coated pit assembly by an Eps15 mutant. *J. Cell Sci.* **112,** 1303–1311 (1999).
17. Joneson, T., White, M. A., Wigler, M. H. & Bar-Sagi, D. Stimulation of membrane ruffling and MAP kinase activation by distinct effectors of RAS. *Science* **271,** 810–812 (1996).
18. Wennstrom, S. *et al.* Activation of phosphoinositide 3-kinase is required for PDGF-stimulated membrane ruffling. *Curr. Biol.* **4,** 385–393 (1994).
19. Honda, K. *et al.* Actinin-4, a novel actin-bundling protein associated with cell motility and cancer invasion. *J. Cell Biol.* **140,** 1383–1393 (1998).
20. Djinovic-Carugo, K., Young, P., Gautel, M. & Saraste, M. Structure of the alpha-actinin rod: molecular basis for cross-linking of actin filaments. *Cell* **98,** 537–546 (1999).
21. Araki, N., Hatae, T., Yamada, T. & Hirohashi, S. Actinin-4 is preferentially involved in circular ruffling and macropinocytosis in mouse macrophages: analysis by fluorescence ratio imaging. *J. Cell Sci.* **113,** 3329–3340 (2000).
22. Tocque, B. *et al.* Ras-GTPase activating protein (GAP): a putative effector for Ras. *Cell. Signal.* **9,** 153–158 (1997).
23. Greenwood, J. A., Theibert, A. B., Prestwich, G. D. & Murphy-Ullrich, J. E. Restructuring of focal adhesion plaques by PI 3-kinase. Regulation by PtdIns (3,4,5)-p(3) binding to alpha-actinin. *J. Cell Biol.* **150,** 627–642 (2000).
24. Corgan, A. M., Singleton, C., Santoso, C. B. & Greenwood, J. A. Phosphoinositides differentially regulate alpha-actinin flexibility and function. *Biochem. J.* **378,** 1067–1072 (2004).
25. Fraley, T. S. *et al.* Phosphoinositide binding inhibits alpha-actinin bundling activity. *J. Biol. Chem.* **278,** 24039–24045 (2003).
26. Krueger, E. W., Orth, J. D., Cao, H. & McNiven, M. A. A dynamin-cortactin-Arp2/3 complex mediates actin reorganization in growth factor-stimulated cells. *Mol. Biol. Cell* **14,** 1085–1096 (2003).
27. Schafer, D. A. *et al.* Dynamin2 and cortactin regulate actin assembly and filament organization. *Curr. Biol.* **12,** 1852–1857 (2002).
28. Fazioli, F. *et al.* Eps8, a substrate for the epidermal growth factor receptor kinase, enhances EGF-dependent mitogenic signals. *EMBO J.* **12,** 3799–3808 (1993).
29. Seastone, D. J. *et al.* The WASp-like protein scar regulates macropinocytosis, phagocytosis and endosomal membrane flow in Dictyostelium. *J. Cell Sci.* **114,** 2673–2683 (2001).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements This work was supported by grants from the Associazione Italiana per la Ricerca sul Cancro and Human Science Frontier Program to P.P.D.F. and G.S., and by grants from the Association for International Cancer Research, European Community (VI Framework), Italian Ministry for University and Research and Consiglio Nazionale delle Ricerche to P.P.D.F. We thank E. Frittoli for technical assistance. L.L. was supported in part by a fellowship from the Associazione Sviluppo Piemonte.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.P.D.F. (difiore@ifom-firc.it).

A highly active synthetic mammalian retrotransposon

Jeffrey S. Han & Jef D. Boeke

Department of Molecular Biology and Genetics and High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

LINE-1 (L1) elements are retrotransposons that comprise large fractions of mammalian genomes[1]. Transcription through L1 open reading frames is inefficient owing to an elongation defect[2], inhibiting the robust expression of L1 RNA and proteins, the substrate and enzyme(s) for retrotransposition[3–5]. This elongation defect probably controls L1 transposition frequency in mammalian cells. Here we report bypassing this transcriptional defect by synthesizing the open reading frames of L1 from synthetic oligonucleotides, altering 24% of the nucleic acid sequence without changing the amino acid sequence. Such resynthesis led to greatly enhanced steady-state L1 RNA and protein levels. Remarkably, when the synthetic open reading frames were substituted for the wild-type open reading frames in an established retrotransposition assay[4], transposition levels increased more than 200-fold. This indicates that there are probably no large, rigidly conserved *cis*-acting nucleic acid sequences required for retrotransposition within L1 coding regions. These synthetic retrotransposons are also the most highly active L1 elements known so far and have potential as practical tools for manipulating mammalian genomes.

L1 retrotransposons account, directly or indirectly, for more than 30% of mammalian genomes by mass[1], by means of self-mobilization and trans-mobilization of *Alu* elements[6]. A full-length (about 6-kilobase) L1 (Fig. 1a) consists of two open reading frames, ORF1 and ORF2, that encode proteins required for retrotransposition[3,4]. Although ORF1 translation is assumed to occur by 5′ cap-binding and scanning, the mechanism for ORF2 translation initiation is unknown. ORF1 and ORF2 assemble into a ribonucleoprotein complex[7,8] that enters the nucleus and nicks the target site, priming the reverse transcription[5] of L1 RNA. L1 proteins show a strong preference for acting on the L1 RNA that encoded them, a phenomenon known as *cis*-preference[9,10]. Nicking and reverse transcription activities are provided by ORF2 protein[3,11], but the function of ORF1 is unknown.

A highly active L1 element would be potentially useful as a tool for mammalian genetics. However, ORF1 and ORF2 sequences are poorly transcribed, and consequently minuscule amounts of functional, full-length L1 RNA are produced in mammalian expression systems[2]. We reasoned that the retrotransposition frequency of L1 might be limited in part by this poor transcriptional elongation. To circumvent this problem, we designed a synthetic mouse ORF2 (smORF2) based on favoured codons in highly expressed mammalian genes. This altered 24% of the nucleic acid sequences but did

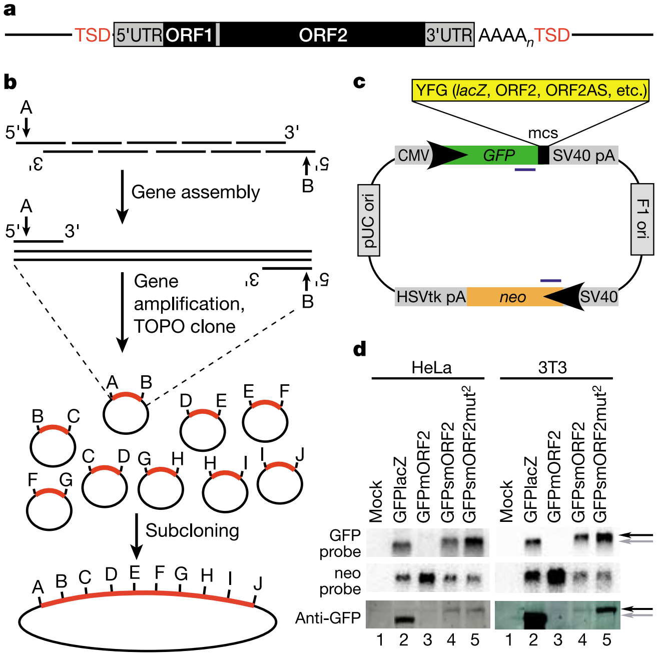


Figure 1 Synthesis and expression of synthetic mouse ORF2. **a**, L1 structure. TSD, target site duplication; UTR, untranslated region. **b**, Overview of gene synthesis. Oligonucleotides encoding each fragment were mixed in a PCR assembly reaction and subsequently used as template amplification. Amplification products were cloned and ligated together with unique restriction sites (labelled A to J). **c**, Plasmid structures. The test sequences (*lacZ*, mORF2 or smORF2) are fused, in frame, downstream of the GFP ORF. An independent *neo* transcript is used to monitor transfection efficiency and loading. Blue lines represent probes used in **d**. **d**, Analysis of smORF2 expression. Top, RNA expression of GFPlacZ, GFPmORF2 and GFPsmORF2. Middle, RNA expression of loading control. Bottom, protein expression of GFPlacZ, GFPmORF2 and GFPsmORF2. Grey and black arrows indicate the expected sizes of GFPlacZ and GFPmORF2/GFPsmORF2, respectively.

not change the protein (Supplementary Fig. S1). This was done in an attempt to destroy any *cis*-acting sequences responsible for poor transcription, including a previously described adenosine-rich bias that might be responsible for the transcription defect[2]. The adenosine content of the smORF2 sequence was decreased to 26% (in comparison with 40% for native mouse ORF2 (mORF2)). Using oligonucleotide-based gene synthesis[12] (Fig. 1b), the recoded gene was synthesized in nine roughly 500-base-pair (bp) fragments and was subsequently assembled by ligation into smORF2. The expression of smORF2 was tested in a fusion vector with green fluorescent protein (GFP) (Fig. 1c) as described previously[2]. In both human and mouse cells, transfection of GFPsmORF2 led to a massive increase in RNA compared with wild-type GFPmORF2 (Fig. 1d, top panel, lanes 3 and 4). The introduction of two mutations that abolish the endonuclease and reverse transcriptase activities of mORF2 provided a further slight increase in smORF2 RNA levels (Fig. 1d, top panel, lanes 5), which is consistent with the known toxicity of ORF2 overexpression in other organisms[5,11]. Probing for the vector-encoded *neo* transcript showed that these increases in RNA were not due to differences in transfection efficiency or loading (Fig. 1d, middle panel). Immunoblotting these samples with anti-GFP (Fig. 1d, bottom panel) showed that protein levels were correlated with RNA increase, marking the first instance of the reproducible expression of detectable amounts of recombinant full-length ORF2 protein in a mammalian system. However, GFPsmORF2 protein levels were still low relative to the control GFPlacZ protein, suggesting that the ORF2 sequence might also be poorly translated or unstable.

We next sought to determine whether the increased RNA levels led to altered retrotransposition efficiency. We used an established tissue culture assay (Fig. 2a) to measure relative retrotransposition frequencies in HeLa cells. mORF2 was replaced with smORF2 in a full-length mouse L1 to make a partly synthetic mouse L1 (psmL1). Because we were concerned that recoded mORF2 might lack potentially important *cis*-acting sequences required for retrotransposition (for example, an internal ribosomal entry site), we also constructed a partly synthetic version of ORF2 (psmL1-2) in which the first roughly 500 bp of mORF2 consisted of wild-type L1 sequence and the remainder was synthetic. In HeLa cells, both psmL1 and psmL1-2 were about 20–25-fold more active than wild-type mL1 (Fig. 2b). Synthesis and incorporation of a synthetic mORF1 (smORF1) and partly synthetic mORF1 variants led to further increases in retrotransposition, reaching a maximum of more than 200-fold increase over wild type (Fig. 2b) in the element with two fully synthetic ORFs.

To verify that these smL1 G418-resistant colonies resulted from authentic L1 retrotransposition, we characterized six smL1 insertions. The mutant loci were identified by inverse polymerase chain reaction (PCR), enabling the amplification of each complete insertion and flanking sequence. For each primer pair, parental HeLa cells produced only empty site products (Fig. 3a, odd-numbered lanes), whereas the respective G418-resistant clones produced both empty site and filled smL1 insertion products of predicted sizes (Fig. 3a, even-numbered lanes). Amplicons were cloned and sequenced to determine their general structures and genomic flanks, summarized in Fig. 3b. All showed a properly spliced *neo* gene, a poly(A) tail, and most (five of six) had target site duplications 5–108 bp long. Insertion no. 10 had a 10-bp target deletion and insertion no. 18 had a 5′ L1 inversion, features commonly found in L1 insertions[13–15]. In addition, various chromosomes served as targets, and the endonuclease cleavage sites inferred from target site duplications matched the previously reported degenerate consensus (5′-TTTT/AA-3′ on the bottom strand)[3,16] (Fig. 3c). Taken together, these results suggest that the synthetic L1 retrotransposes

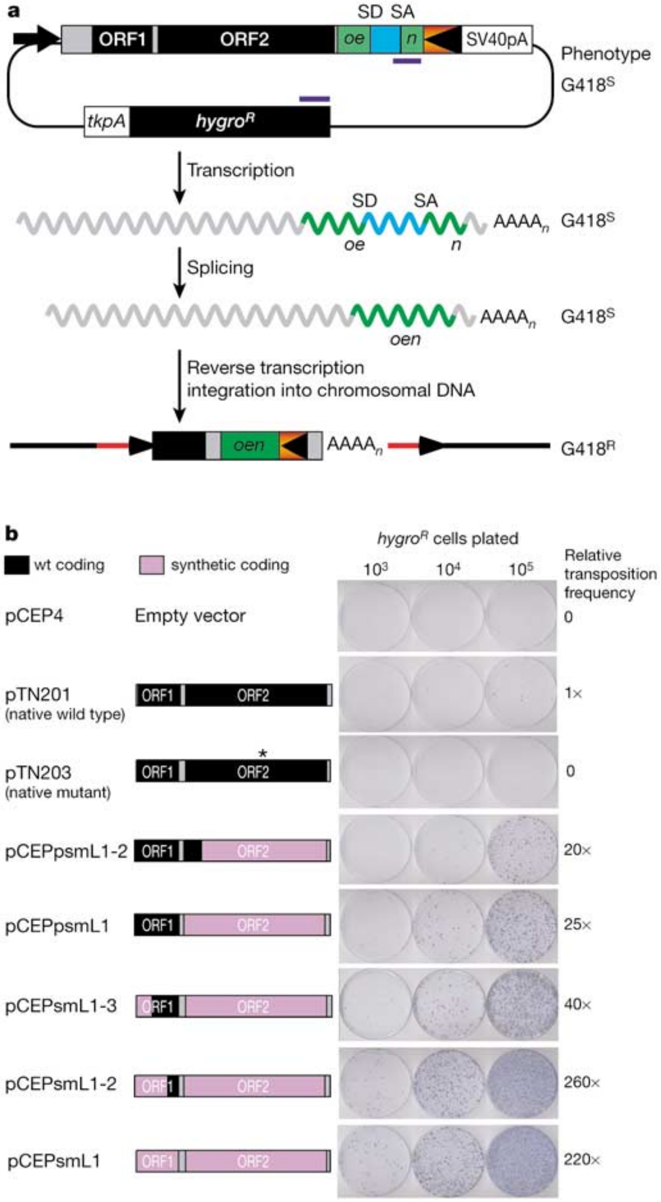
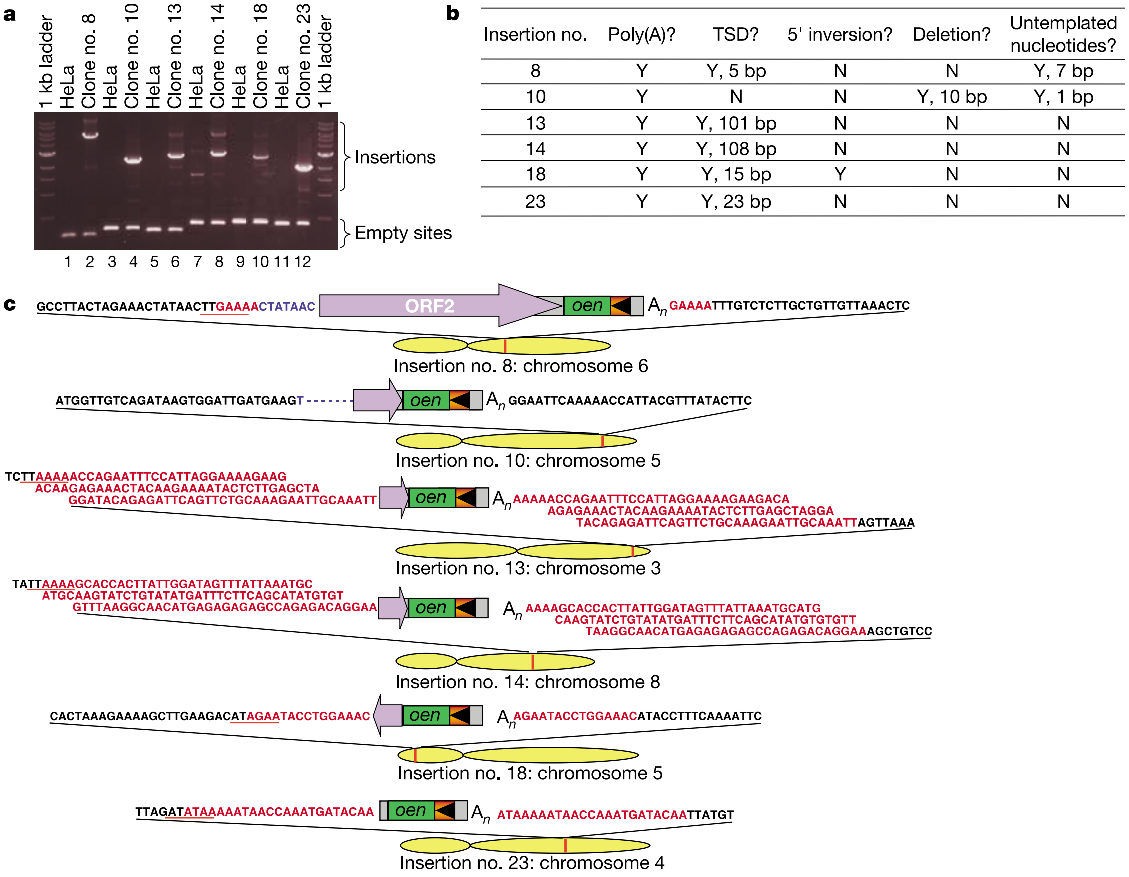


Figure 2 Retrotransposition of synthetic mL1. **a**, The retrotransposition assay. The L1 element contains an intron-interrupted *neo* reporter in the 3′ untranslated region with its own promoter and polyadenylation signal. Only when *neo* is transcribed from the L1 promoter, spliced, reverse transcribed and integrated into the genome does a cell become G418-resistant[4]. Blue lines represent probes for RNA analysis (Fig. 4). SD, splice donor; SA, splice acceptor. **b**, Retrotransposition was assayed in HeLa cells ($N = 3$). pTN201 contains only wild-type native mouse L1 sequence, and pTN203 contains wild-type native mouse L1 sequence with a D709Y reverse transcriptase point mutation[22]. The average absolute number of colonies for pTN201 was 440 events per 10^6 transfected cells.

Table 1 **High-frequency retrotransposition in mouse cells**

	Relative transposition frequency		
Plasmid	HeLa	3T3	L
pCEP4 (empty vector)	0	0	0
pTN201 (native mouse wild type)	<0.005	<0.002	<0.002
pTN203 (native mouse mutant)	0	0	0
pJM101L1 (native human wild type)	0.13	0.017	0.07
pCEPsmL1 (synthetic mouse wild type)	1	1	1
pCEPsmL1mut[2] (synthetic mouse mutant)	0	0	<0.002

With the use of the transient assay[17], synthetic mouse L1 (pCEPsmL1) retrotransposition frequency was compared with that of wild-type native human L1 and wild-type native mouse L1 ($N = 3$). The average absolute numbers of colonies of pJM101L1rp (colonies per 10^6 transfected cells) for HeLa, 3T3 and L cells were 2,904, 108 and 1,568, respectively.


a
1 kb ladder
HeLa
Clone no. 8
HeLa
Clone no. 10
HeLa
Clone no. 13
HeLa
Clone no. 14
HeLa
Clone no. 18
HeLa
Clone no. 23
1 kb ladder
Insertions
Empty sites
1 2 3 4 5 6 7 8 9 10 11 12
b
c
GCCTTACTAGAAACTATAACTTGAAAACTATAAC
ORF2
oen
An GAAAATTTGTCTCTTGCTGTTGTTAAACTC
Insertion no. 8: chromosome 6
ATGGTTGTCAGATAAGTGGATTGATGAAGT
oen
An GGAATTCAAAAACCATTACGTTTATACTTC
Insertion no. 10: chromosome 5
TCTTAAAAACCAGAATTTCCATTAGGAAAAGAAG
ACAAGAGAAACTACAAGAAAATACTCTTGAGCTA
GGATACAGAGATTCAGTTCTGCAAAGAATTGCAAATT
oen
An AAAAACCAGAATTTCCATTAGGAAAAGAAGACA
AGAGAAACTACAAGAAAATACTCTTGAGCTAGGA
TACAGAGATTCAGTTCTGCAAAGAATTGCAAATTAGTTAAA
Insertion no. 13: chromosome 3
TATTAAAAGCACCACTTATTGGATAGTTTATTAAATGC
ATGCAAGTATCTGTATATGATTTCTTCAGCATATGTGT
GTTTAAGGCAACATGAGAGAGAGCCAGAGACAGGAA
oen
An AAAAGCACCACTTATTGGATAGTTTATTAAATGCATG
CAAGTATCTGTATATGATTTCTTCAGCATATGTGTGTT
TAAGGCAACATGAGAGAGAGCCAGAGACAGGAAAGCTGTCC
Insertion no. 14: chromosome 8
CACTAAAGAAAAGCTTGAAGACATAGAATACCTGGAAAC
oen
An AGAATACCTGGAAACATACCTTTCAAAATTC
Insertion no. 18: chromosome 5
TTAGATATAAAAATAACCAAATGATACAA
oen
An ATAAAAATAACCAAATGATACAATTATGT
Insertion no. 23: chromosome 4

Insertion no.	Poly(A)?	TSD?	5' inversion?	Deletion?	Untemplated nucleotides?
8	Y	Y, 5 bp	N	N	Y, 7 bp
10	Y	N	N	Y, 10 bp	Y, 1 bp
13	Y	Y, 101 bp	N	N	N
14	Y	Y, 108 bp	N	N	N
18	Y	Y, 15 bp	Y	N	N
23	Y	Y, 23 bp	N	N	N

Figure 3 Synthetic mouse L1 uses the standard retrotransposition mechanism. **a**, Primers flanking each insertion were used for amplification from G418-resistant clones (see the text). **b**, Characteristics of cloned insertions. TSD, target site duplication. **c**, Structure and flanking sequence of cloned insertions are shown. Insertion no. 8 contained an additional 7 bp (highlighted in blue) not found in the human genome sequence (http://genome.ucsc.edu). Insertion no. 10 contained one untemplated base pair relative to the human genome sequence database followed by a 10-bp deletion (indicated in blue) immediately upstream of the L1 insertion. TSDs are highlighted in red, and presumptive endonuclease cleavage sites are underlined.

by target-primed reverse transcription and not some aberrant mechanism.

We also compared the activity of the synthetic mouse L1 retrotransposons with wild-type human and mouse L1 in mouse cells. Because episomal plasmids used to introduce marked retrotransposons do not replicate efficiently in mouse cells, we used a transient retrotransposition assay[17] in 3T3 and L cells. We also performed the transient assay in HeLa cells, verifying the relative retrotransposition frequencies obtained with the standard assay (compare pTN201 and pCEPsmL1 from Fig. 2b and Table 1).

The synthetic mouse L1 (pCEPsmL1) underwent retrotransposition at much higher frequencies (more than 200-fold) than its wild-type counterpart in mouse cells. In addition, we compared smL1 with a human L1 (pJM101L1rp) because L1rp has previously been used to generate transgenic mouse lines and thus serves as a benchmark for retrotransposition frequencies in mice. smL1 was significantly more active than L1rp in all cell types tested, making it the most active L1 element known so far. Introducing catalytic mutations into smL1 to produce smL1mut^2 essentially abolished retrotransposition, confirming that smL1 retrotransposition depends on ORF2 endonuclease and reverse transcriptase functions.

Northern blot analysis of wild-type full-length mL1 and its synthetic counterparts revealed that increasing lengths of synthetic L1 sequence led to increasing full-length L1 RNA levels (Fig. 4). pCEPsmL1mut^2 was used in place of pCEPsmL1 because pCEPsmL1 was difficult to maintain episomally, as determined by the *hygro* transfection/loading control (data not shown). This suggests, because the intact pCEPsmL1 plasmid is not maintained in transfected cells for long periods, that the reported increased retrotransposition frequencies of smL1 relative to native mouse or human L1 are underestimates. In addition, the increased RNA levels suggest that increased RNA expression is, at least in part, responsible for enhanced retrotransposition levels. However, because codon usage was optimized for mammalian cells, improved translational efficiency might also be significant. We also cannot rule out the possibility that recoding destroyed regulatory nucleic acid motifs that affect retrotransposition in multiple ways.

It is already known that both the 5′ and 3′ untranslated regions and the interORF region of L1 are not required for retrotransposition (refs 4, 18, and R. S. Alisch and J. V. Moran, personal communication). This indicates that if any essential nucleic acid

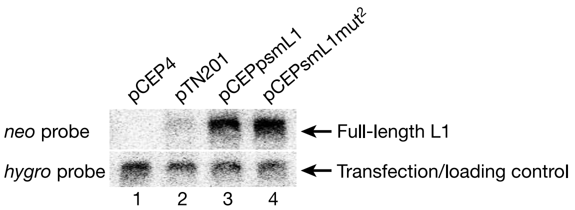


Figure 4 High-frequency retrotransposition in mouse cells: total RNA analysis of smL1 expression. Expression of native, partly synthetic, and completely synthetic mL1 was compared in HeLa cells.

motifs exist in L1, they are within the coding regions. Because our synthetic mouse L1 is extensively mutagenized throughout the coding regions but still retrotransposes with startlingly high efficiency, either L1 essential nucleic acid sequences are small and fortuitously preserved in smL1, or L1 essential nucleic acid sequences are highly tolerant to mutations. It is also plausible that, aside from encoding functional proteins, there are no special requirements for L1 nucleic acid sequence. Only further investigation will distinguish between these possibilities, but it is of interest to note that an absence of required nucleic acid motifs might necessitate unconventional explanations for mysterious aspects of the L1 life cycle such as the ORF2 translational initiation mechanism and *cis*-preference.

Finally, our synthetic L1 might represent a major step forward in efforts to design a useful random mutagenesis system in mice. Transposons are useful genetic tools because of their ability to produce mutations by adding new DNA sequence, 'tagging' the disrupted gene for easy cloning and identification[19]. However, in mammalian systems the low frequency of retrotransposition has precluded the practical large-scale use of retrotransposons for mutagenesis. For example, the theoretical potential of L1 as a mutagenic agent has been shown in mice, but so far the frequency of progeny carrying a new insertion has reached a maximum of less than 10%[20]. An ideal mutagenesis system would produce multiple new insertions per progeny animal such that each carries a new mutant gene. Thus, the retrotransposition rate of L1 in current mouse models is one to two orders of magnitude lower than desired. In mouse cells our synthetic L1 retrotransposes at a frequency ranging from 15-fold to 50-fold higher than L1rp. If this increase in L1 activity in mouse cells translates to animals, we envisage creating transgenic mouse lines that, once made, continuously generate easily clonable random mutations in each progeny animal, without embryonic stem cell manipulations. Such a line would be extremely useful for practical forward genetic screens. In addition, cataloguing and storing mutants (or mutant sperm) could allow the comprehensive knockout of mouse genes. □

Methods

Gene synthesis

smORF2 and ORF1 sequences were created by replacing each codon in the mouse L1 ORFs with the favoured codons in highly expressed human genes[21] (Supplementary Table S2). The sequence was further altered with silent mutations introducing unique cleavage sites and eliminating potential hairpins that might have inhibited gene assembly. 60-mer oligonucleotides collectively encoding both strands of smORF2 were ordered from Qiagen, and gene synthesis[12] was performed on each ~500-bp segment as shown in Fig. 1b. Assembly reactions contained each primer at 30 nM and 1 × ExTaq mix (Takara) in a total of 25 μl. Amplification reactions contained each outer primer at 0.5 μM, 2.5 μl assembly reaction, and 1 × ExTaq mix in a total volume of 25 μl. PCR conditions were 94 °C for 4 min, 25 cycles of 94 °C for 30 s, 65 °C for 30 s, and 72 °C for 30 s, followed by 72 °C for 7 min. PCR products were cloned into pCRII with the TOPO-TA cloning kit (Invitrogen). A total of 24–48 clones were sequenced for each fragment and mutations were removed by standard cloning techniques. Finally, synthesized fragments were ligated together in pBluescriptKS^{-}. Oligonucleotide sequences used are shown in Supplementary Table S1.

Plasmids

pGFPlacZ and pGFPmORF2 are described elsewhere[2]. pTN201 and pTN203 were gifts from H. Kazazian[22]. pJM101L1rp was provided by J. Moran[17]. Detailed descriptions of the construction of the plasmids are available from the authors on request.

Cell culture and transfection

HeLa cells, 3T3 cells and L cells were gifts from the laboratories of J. Moran, S. Desiderio and J. Nathans. Cells were grown in DMEM medium (Invitrogen) with 10% FBS (Invitrogen) and 1% penicillin/streptomycin (Invitrogen).

Transfections were performed with Fugene 6 (Roche) in six-well dishes. The transfection mix consisted of 100 μl Opti-MEM (Invitrogen), 3 μl Fugene and 2 μg DNA. For downstream northern or immunoblot analyses, cells were harvested 36–48 h after transfection.

Northern blot analysis

Total RNA was isolated with TRIzol reagent (Invitrogen) in accordance with the manufacturer's instructions. Total RNA (6 μg) from each sample was treated with 10 units of DNase I for 15 min at 37 °C, then run on a 0.8% agarose/formaldehyde gel, blotted overnight to a Genescreen plus nylon membrane (NEN) in 10 × SSC, and crosslinked by ultraviolet radiation. Prehybridizations and hybridizations were both performed in ULTRAhyb (Ambion) at 42 °C. The following $[\gamma\text{-}^{32}\mathrm{P}]$ATP end-labelled oligonucleotides were used as probes: GFP probe, JB4057; GFP plasmid *neo* probe, JB4059; transposition plasmid *neo* probe, JB4541; *hyg* probe, JB6341. Washes were performed in 2 × SSC, 0.1% SDS and in 0.2 × SSC, 0.1% SDS. Radioactive signal was detected with Fuji imaging plates and a Fuji scanner (BAS-1500). For subsequent reprobing, membranes were stripped with three 10-min washes in boiling 0.1 × SSC, 1% SDS.

Immunoblot analysis

Cells were harvested in 5% SDS/PBS; this was followed by sonication. Total lysates were subjected to 7.5% SDS–polyacrylamide-gel electrophoresis and transferred to poly(vinylidene difluoride) (Amersham). Antibody incubations were performed in PBS containing 0.05% Tween-20 and 5% milk. Washes were performed in PBS, 0.1% Tween-20. Anti-GFP(FL) antibody (Santa Cruz) was used at 1:250 dilution. Anti-rabbit IgG (Amersham) was used at 1:5,000 dilution. Blots were developed with ECL-plus (Amersham).

Retrotransposition assays

The standard retrotransposition assay in HeLa cells was performed essentially as described[4]. Transfected cells were selected with 200 μg ml^{-1} hygromycin for 10–12 days, then counted and seeded in 600 μg ml^{-1} G418 for 10 days. Colonies were stained with 0.4% Giemsa in PBS.

The transient retrotransposition assays in HeLa, 3T3 and L cells were performed essentially as described[17]. Each transposition construct was cotransfected with the GFP-expressing plasmid pTracerEF (Invitrogen) to normalize for transfection efficiency. At 24 h after transfection, cells were split 1:2, 1:20 and 1:200 into 100-mm dishes. At 36 h after transfection, the diluted cells were selected with G418 and the remaining cells were analysed for GFP expression by flow cytometry to normalize for transfection efficiency. 3T3 cells were selected in 1 mg ml^{-1} G418; L cells were selected in 400 μg ml^{-1} G418. Colonies were stained with 0.4% Giemsa or 0.5% Coomassie brilliant blue.

Cloning of retrotransposition events

Integration sites were determined by inverted PCR essentially as described[23]. Genomic DNA (5 μg) from each clone was digested with *Eco*RI, inactivated by heat, diluted to 1 ml and ligated overnight, precipitated with ethanol, resuspended in 30 μl water and subjected to two rounds of inverted PCR with oligos JB6466/JB6467 (round 1) and JB6468/JB6469 (round 2). Sequencing with JB3529, JB3530 and JB3531 identified the 3′ flanking sequences. Primers based on flanking sequence were used to amplify intact smL1 insertions, which were subsequently sequenced.

Received 22 January; accepted 30 March 2004; doi:10.1038/nature02535.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Han, J. S., Szak, S. T. & Boeke, J. D. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429,** 268–274 (2004).
3. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87,** 905–916 (1996).
4. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
5. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription *in vitro. EMBO J.* **21,** 5899–5910 (2002).
6. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35,** 41–48 (2003).
7. Martin, S. L. Characterization of a LINE-1 cDNA that originated from RNA present in the ribonucleoprotein particles: implications for the structure of an active mouse LINE-1. *Gene* **153,** 261–266 (1995).
8. Kolosha, V. O. & Martin, S. L. *In vitro* properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc. Natl Acad. Sci. USA* **94,** 10155–10160 (1997).
9. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
10. Wei, W. *et al.* Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell. Biol.* **21,** 1429–1439 (2001).
11. Mathias, S. L., Scott, A. F., Kazazian, H. H. Jr, Boeke, J. D. & Gabriel, A. Reverse transcriptase encoded by a human transposable element. *Science* **254,** 1808–1810 (1991).
12. Stemmer, W. P., Crameri, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164,** 49–53 (1995).
13. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17,** 915–928 (2000).
14. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110,** 315–325 (2002).
15. Symer, D. E. *et al.* Human L1 retrotransposition is associated with genetic instability *in vivo. Cell* **110,** 327–338 (2002).
16. Cost, G. J. & Boeke, J. D. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37,** 18081–18093 (1998).
17. Wei, W., Morrish, T. A., Alisch, R. S. & Moran, J. V. A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. *Anal. Biochem.* **284,** 435–438 (2000).
18. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr Exon shuffling by L1 retrotransposition. *Science* **283,** 1530–1534 (1999).
19. Hamer, L., DeZwaan, T. M., Montenegro-Chamorro, M. V., Frank, S. A. & Hamer, J. E. Recent advances in large-scale transposon mutagenesis. *Curr. Opin. Chem. Biol.* **5,** 67–72 (2001).
20. Ostertag, E. M. *et al.* A mouse model of human L1 retrotransposition. *Nature Genet.* **32,** 655–660 (2002).
21. Haas, J., Park, E.-C. & Seed, B. Codon usage limitation in the expression of HIV-1 envelope glycoprotein. *Curr. Biol.* **6,** 315–324 (1996).

22. Naas, T. P. *et al.* An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J.* **17,** 590–597 (1998).
23. Morrish, T. A. *et al.* DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nature Genet.* **31,** 159–165 (2002).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank Boeke laboratory members, especially Y. Aizawa, for helpful discussions and critical reading of the manuscript. We thank A. Yetil for assistance. This work was supported by the NIH (J.D.B.) and the Medical Scientist Training Program (J.S.H.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.D.B. (jboeke@jhmi.edu).

Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain

Jin-Biao Ma*, Keqiong Ye* & Dinshaw J. Patel

Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York 10021, USA

* These authors contributed equally to this work

Short RNAs mediate gene silencing, a process associated with virus resistance, developmental control and heterochromatin formation in eukaryotes[1–5]. RNA silencing is initiated through Dicer-mediated processing of double-stranded RNA into small interfering RNA (siRNA)[6,7]. The siRNA guide strand associates with the Argonaute protein in silencing effector complexes, recognizes complementary sequences and targets them for silencing[8–11]. The PAZ domain is an RNA-binding module found in Argonaute and some Dicer proteins and its structure has been determined in the free state[12–14]. Here, we report the 2.6 Å crystal structure of the PAZ domain from human Argonaute eIF2c1 bound to both ends of a 9-mer siRNA-like duplex. In a sequence-independent manner, PAZ anchors the 2-nucleotide 3′ overhang of the siRNA-like duplex within a highly conserved binding pocket, and secures the duplex by binding the 7-nucleotide phosphodiester backbone of the overhang-containing strand and capping the 5′-terminal residue of the complementary strand. On the basis of the structure and on binding assays, we propose that PAZ might serve as an siRNA-end-binding module for siRNA transfer in the RNA silencing pathway, and as an anchoring site for the 3′ end of guide RNA within silencing effector complexes.

siRNA is a 19–23-base-paired (bp) duplex with 2-nucleotide (nt) 3′ overhangs at both ends, containing 5′-phosphates and 3′-hydroxyls[7,15]. siRNA is not merely a consequence of Dicer processing, as both the length and ends are important for mediating RNA interference (RNAi)[7,10,15], whereby target messenger RNA is sequence-specifically degraded in an siRNA-programmed effector complex termed the RNA-induced silencing complex (RISC)[8,9,16]. This suggests that specific structural features of siRNA constitute recognition targets for the protein components within the RNAi machinery. The PAZ domain can bind to single-stranded (ss) RNAs and siRNA duplexes[12–14], and requires the 2-nt 3′ overhang for efficient complex formation[13,14]. We have determined a co-crystal structure of PAZ domain and a 9-mer RNA (5′-CGUGACUCU-3′) in order to illustrate the details of PAZ–RNA interaction and gain functional insights into the recognition process.

The structure of the PAZ–RNA complex reveals an unanticipated arrangement in which the 9-mer RNA, initially designed as a ssRNA ligand, forms a self-complementary siRNA-like A-form duplex, which is bound by the PAZ domain at each end (Fig. 1). The 3′-shifted pairing of each strand results in 2-nt, single-stranded 3′ overhangs at the duplex ends, which are characteristic of siRNA architecture. However, the short 7-bp duplex is unstable, due to the presence of three non-canonical pairs (Fig. 1a), raising the concern that the observed duplex formation could result from crystal-packing interactions. This is not the case, as the PAZ domain and the 9-mer RNA form a stable 2:2 complex in solution, as judged by gel filtration (Supplementary Fig. S1). Thus, PAZ binding apparently stabilizes siRNA-like duplex formation. Moreover, the absence of contacts between the two PAZ domains in the complex indicates independent binding of PAZ domains to the siRNA-like duplex, and suggests that such an arrangement should also hold for complex formation to the ends of a typical 21-nt siRNA.

The PAZ domain in the complex adopts a heart-shaped globular topology (Fig. 2a), with a twisted β-barrel consisting of six β-strands (β1–β3, β6–β8), capped by two amino-terminal α-helices (α1, α2) on one side and connected to an αβ module (β4–β5–α3) on the other side. The two strands of the siRNA-like duplex interact with a specific PAZ domain in a highly asymmetric manner. The strand bound with its 3′ end contacts the PAZ domain along its full 9-nt length, whereas the complementary strand makes contacts only with the 5′-terminal residue. The first 7 nucleotides constituting the RNA duplex structure, proceeding in the 5′ to 3′ direction, interact with the protein's carboxy-terminal tail and the positively charged surface formed by the strands β2, β3 and the β6–β7 loop (Fig. 2b). The bound RNA adopts a stacked 5′ to 3′ helical trajectory, except for a sharp turn (clockwise rotation of ~110° along the helical axis) in the phosphodiester backbone between the duplex and 2-nt 3′-overhang segments, thereby inserting the 2-nt ends into a central protein pocket formed between the barrel and αβ module. The same central RNA-binding pocket has been proposed previously from an evaluation of NMR chemical shift perturbations, mutagenesis and sequence conservation analysis[12–14]. The free[12–14] and RNA-bound

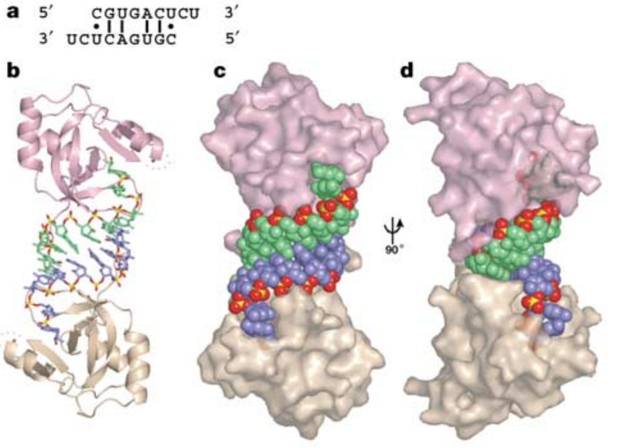


Figure 1 Overview of the PAZ–siRNA-like duplex structure **a**, The self-complementary siRNA-like duplex formed in the crystal. **b**, The entire complex of two PAZ domains bound to each end of an siRNA-like duplex. Protein and RNA are presented in ribbon and stick representations, respectively. **c**, The same view as **b**, but protein and RNA are presented in semi-transparent surface- and space-filling representations, respectively. **d**, 90° rotation of **c**. Note that PAZ domains predominantly contact the 3′-overhang-containing strands and dock with the 5′ ends of the complementary strands. The 2-nt 3′ overhangs are anchored within the pockets, with their base edges facing outwards towards solvent. The PAZ domains are coloured pink and beige and the RNA strands are blue and green, except for the phosphate groups of the RNA strands, in which phosphate is yellow and oxygen is red.