Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes

Jeffrey S. Han[1], Suzanne T. Szak[2] & Jef D. Boeke[1]

[1]*Department of Molecular Biology and Genetics and High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA*
[2]*Biogen, Inc., Cambridge, Massachusetts 02142, USA*

LINE-1 (L1) elements are the most abundant autonomous retrotransposons in the human genome, accounting for about 17% of human DNA. The L1 retrotransposon encodes two proteins, open reading frame (ORF)1 and the ORF2 endonuclease/reverse transcriptase. L1 RNA and ORF2 protein are difficult to detect in mammalian cells, even in the context of overexpression systems. Here we show that inserting L1 sequences on a transcript significantly decreases RNA expression and therefore protein expression. This decreased RNA concentration does not result from major effects on the transcription initiation rate or RNA stability. Rather, the poor L1 expression is primarily due to inadequate transcriptional elongation. Because L1 is an abundant and broadly distributed mobile element, the inhibition of transcriptional elongation by L1 might profoundly affect expression of endogenous human genes. We propose a model in which L1 affects gene expression genome-wide by acting as a 'molecular rheostat' of target genes. Bioinformatic data are consistent with the hypothesis that L1 can serve as an evolutionary fine-tuner of the human transcriptome.

L1 elements are the most abundant autonomous retrotransposons in the human genome, accounting for about 17% of human DNA[1]. A functional full-length L1 element, depicted in Fig. 1a, contains an internal promoter in the 5′ untranslated region (5′ UTR) that initiates transcription at base 1 (ref. 2). This is followed by two long open reading frames, ORF1 and ORF2, required for retrotransposition[3,4] *in cis*[5,6]. ORF1 encodes an RNA-binding protein that has nucleic acid chaperone activity *in vitro*[7–9], but no known specific role in the L1 replication mechanism. ORF2 encodes a protein with endonuclease and reverse transcriptase activities, both critical for retrotransposition[3,4]. A short 3′ UTR is followed immediately by a poly(A) tail, and the entire element is typically flanked by target site duplications. The L1 retrotransposition machinery is not only used for L1 mobilization itself but also assists in the retrotransposition of Alu retroelements[10].

L1 elements are mobilized through target-primed reverse transcription, in which ORF2 nicks target DNA, using the resultant 3′-OH to prime the reverse transcription of L1 RNA[11,12]. ORF2 is probably relatively non-processive and often fails to reach the RNA 5′ end during first-strand synthesis. This explains, at least in part, the distribution of predominantly non-functional, 5′ truncated L1s in the genome[13,14].

A puzzling feature of L1 is the difficulty in detecting L1 RNA and ORF2 protein in mammalian cells, even in the context of high-copy plasmids and overexpression. In contrast, recombinant ORF2 was expressed successfully in yeast and baculovirus systems[12,15,16]. This suggests a mammalian-specific mechanism for negatively regulating L1 expression, which, considering the mutagenic nature of transposition and dearth of L1 expression in somatic cells[17,18], is not unexpected. The difficulty in expressing L1 RNA suggests a transcriptional defect, but translation might also have a function in low ORF2 expression, because the ORF2 translation initiation mechanism is not understood. Recent data indicate that L1 transcription might be inhibited by cryptic premature polyadenylation signals[19], which would produce nonfunctional truncated L1 transcripts. Here we show that poor expression of ORF2 results from the inability of RNA polymerase to elongate efficiently through L1 coding sequences, with polyadenylation having a smaller role. We also show that L1 sequences in the antisense orientation inhibit transcription by producing premature polyadenylation; this is significant because most L1 elements within human genes are in the antisense orientation.

L1 forms a component of most mammalian transcription units, but the effects of these primarily intronic inserted sequences have not been studied carefully. About 79% of human genes are estimated to contain at least one segment of L1 in their transcription unit (see Methods), and L1 segments from pre-existing and newly derived insertions usually contain L1 ORF2 (refs 14, 20–22). As these sequences are mostly intronic, it has been assumed that the extra sequences are spliced out and do not affect target gene expression. Our findings indicate that, through a combination of transcriptional elongation inhibition and premature polyadenylation, L1 insertions in either orientation can affect the RNA production of endogenous genes, both qualitatively and quantitatively. We propose a human genome model in which L1 has led to numerous subtle but potentially significant transcriptome alterations.

ORF2 reduces RNA and protein amounts

To examine the effect of ORF2 sequences on expression, we fused either *lacZ* or L1 ORF2 coding regions downstream of the green fluorescent protein (GFP) ORF (Fig. 1b). Placing these test transcripts downstream of GFP permitted the examination of RNA and protein expression independently of translation initiation, because translation initiates in GFP. The presence of ORF2 sequence significantly lowered steady-state protein production and led to a 70-fold decrease in RNA relative to *lacZ* (Fig. 1c, d, lanes 2 and 3). Inserting ORF2 in the antisense orientation produced a similar, but less potent, decrease in full-length RNA. These phenomena also occur with the mouse ORF2 sequence (Fig. 1d, lanes 5 and 6) and are therefore not human L1-specific. The effect is limited neither to specific cell lines nor to the promoter used (Supplementary Fig. S2a, b).

When L1 ORF2 is inserted in the antisense orientation, most

of the decrease in full-length RNA is accounted for by lower-molecular-mass species. These are probably prematurely polyadenylated transcripts, because they exist in both total cytoplasmic and poly(A)-selected RNA and do not hybridize with a 3′ UTR probe (Supplementary Fig. S2c). Cloning and sequencing of these foreshortened RNAs identified the polyadenylation sites used (Fig. 1e). Thus, an ORF2 sequence placed in the antisense direction inhibits full-length transcript production primarily through premature polyadenylation. Some truncated, polyadenylated transcripts derived from GFPORF2 were also identified but these specific truncations have only a minor function in decreasing the amounts of full-length transcript. Polyadenylated transcripts are expected to be stable. If these shorter transcripts are primarily responsible for decreased full-length RNA amounts they should be detected at much higher concentrations, and they are in GFPORF2AS. After quantifying total hybridization signal and comparison with the control, only 15% of the 'missing' GFPORF2 RNA is accounted for by lower-molecular-mass species, whereas 87% is accounted for in GFPORF2AS (Supplementary Fig. S3). We also experimentally verified that inserting a poly(A) signal upstream of the L1 ORF2 sequence (immediately after GFP) leads to truncated species that can account for the lack of full-length transcript (Supplementary Fig. S2d). Thus, only some of the decrease in sense-strand GFPORF2 RNA is accounted for by premature polyadenylation, and most of the decrease results from another mechanism.

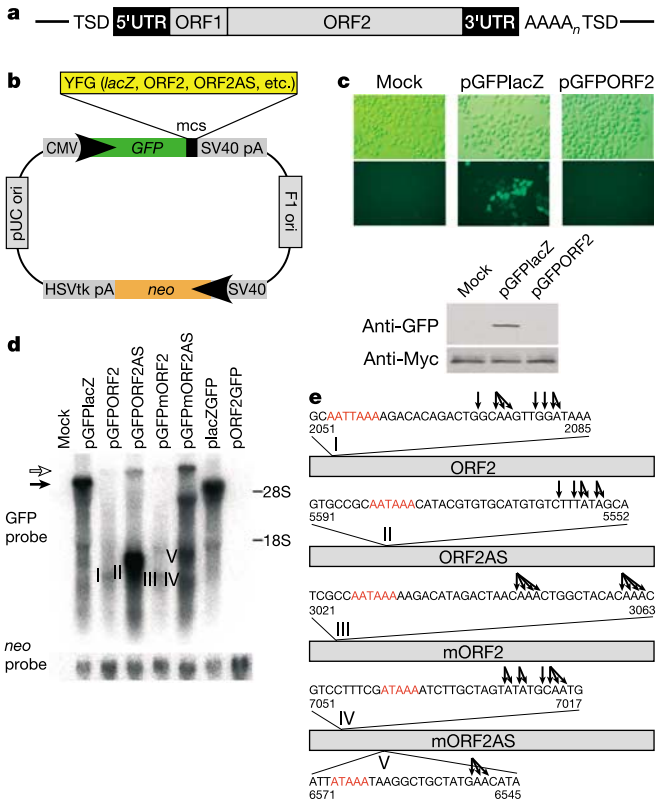

Figure 1 ORF2 sequence decreases expression. **a**, L1 structure (see the text). **b**, Plasmid structures. Fusions are in frame unless noted otherwise. *neo* is used for normalization of transfection. **c**, Immunoblotting of HeLa transfections. Anti-Myc shows equal loading. **d**, Total RNA analysis of HeLa transfections. mORF2, mouse ORF2; AS, antisense. Open and black arrows show the expected positions of GFPORF2 and GFPlacZ, respectively. **e**, Identification of lower-molecular-mass bands. Roman numerals refer to bands in Fig. 1d. The presumptive poly(A) signals are highlighted in red, and the polyadenylation sites are indicated by arrows. Numbers refer to L1.2 sequence (GenBank accession no. M80343) or L1$_{spa}$ sequence (GenBank accession no. AF016099).

Poor expression is not due to ORF2 protein feedback

Poor expression of ORF2 could result from some effect of ORF2 protein. pGFPORF2mut[2] contains two missense mutations (D205G and D702Y) that destroy the known catalytic activities of ORF2 (refs 3, 4) and alleviate its toxicity in yeast[23] and baculovirus[12]. pGFPORF2 was altered separately by adding a stop codon and four extra base pairs between GFP and ORF2; in this construct, GFPstop-ORF2, the GFPORF2 fusion protein is truncated and ORF2 is shifted out of frame with respect to GFP, and produces only GFP. Finally, we introduced two +1 frameshifts early in ORF2 by adding adenosines before nucleotide positions 2,134 and 2,629 (numbers relative to the L1.2 GenBank sequence, accession no. M80343) to eliminate the possibility of the production of ORF2 protein by the initiation of internal translation. Although these frameshifts are early in ORF2, they occur after a residue (N14) required for transposition[3], and thus must be translated in a full-length active L1 element, whether or not ORF2 is initiated internally. pGFPORF2, pGFPORF2mut[2], pGFPstopORF2 and pGFPORF2fs each show similar reductions in RNA (Fig. 2), confirming that the RNA deficit results from the ORF2 nucleotide sequence, not ORF2 protein.

Inhibitory effect does not map to a discrete sequence

We attempted to map a region of the ORF2 sequence responsible for this decrease in RNA. Starting with pGFPstopORF2, progressive amino-terminal and carboxy-terminal deletions, as well as several internal deletions, were tested (Fig. 3a). No single discrete nucleotide sequence block determined the RNA deficit. Rather, longer ORF2 sequences led to lower concentrations of RNA, which correlated with protein abundance (Supplementary Fig. S3). This is consistent with a repetitive sequence or sequences scattered throughout ORF2 that collectively inhibit ORF2 expression.

An interesting feature of L1 coding sequences is a strong adenosine-rich bias in the sense strand (Fig. 3b). This A-rich bias is absent from the 5′ UTR. We tested the hypothesis that ORF1 and ORF2 confer a similar length-dependent inhibition of expression. Because ORF1 is only 1,017 base pairs long, placing ORF1 downstream of

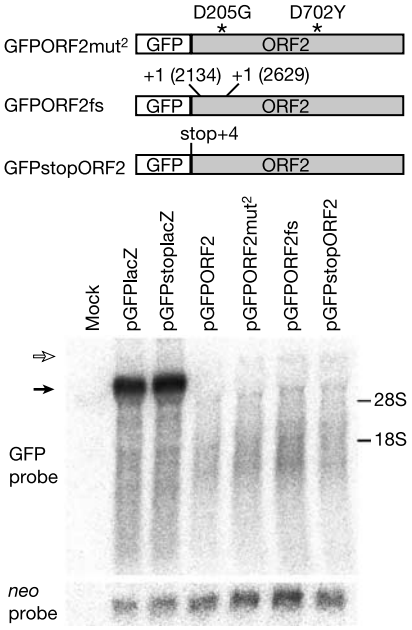

Figure 2 Decreased RNA amounts are not due to ORF2 protein. Top, structures of GFPORF2mut[2], GFPORF2fs and GFPstopORF2. Bottom, total RNA analysis of HeLa transfections. Open and black arrows show the expected positions of GFPORF2 and GFPlacZ, respectively.

GFP is not expected to interfere with its expression on the basis of the ORF2 deletion mapping experiments. However, if the A-rich bias governs poor expression, expanding the ORF1 sequence to ORF2 length (about 4 kilobases) should markedly decrease RNA amounts. A control construct (pGFPORF1) with a single copy of ORF1 is expressed efficiently, but a lengthened ORF1-containing construct (pGFPstop4ORF1) containing four tandem repeats of ORF1, like the ORF2 construct, led to low concentrations of GFP RNA. As with ORF2, the effect was sense-strand-specific (Fig. 3c, lanes 4–7). In contrast, tandemized L1 5′ UTR (5 × 800 base pairs long), lacking the A-rich bias, had no effect (Fig. 3c, lanes 8 and 9). The effect of unusual base composition on transcription might be analogous to a similar phenomenon in yeast, in which mutants in the THO complex decrease the transcription of long GC-rich DNA stretches[24]. Elsewhere we show that alteration of the L1 base content abolishes the transcription defect and results in high-frequency retrotransposition[25], further supporting the hypothesis that the unusual strand bias inhibits expression.

ORF2 transcripts are inefficiently elongated

Lower steady-state RNA concentrations could result from either increased RNA degradation or decreased RNA production. We measured the RNA half-life of the GFPlacZ and GFPORF2 transcripts after treatment with the transcriptional inhibitor actinomycin D. GFPlacZ and GFPORF2 transcript amounts decreased on a similar timescale relative to the *neo* transcript (Fig. 4a, left). The half-life of GFPORF2 relative to GFPlacZ is decreased at most twofold (Fig. 4a, right); quantification by real-time polymerase chain reaction with reverse transcription (RT–PCR) was consistent with this result (Supplementary Fig. S4). This minor change in RNA half-life cannot account for the 70-fold decrease in steady-state GFPORF2 RNA amounts (Fig. 1d, lanes 2 and 3), indicating that most of the GFPORF2 transcript decrease might result from some mechanism other than transcript instability.

A nuclear run-on assay was performed to evaluate RNA polymerase density along GFPlacZ and GFPmORF2 transcripts produced in human nuclei. Mouse ORF2 was used for this hybridization-based assay to avoid the high background expected from the 500,000 endogenous human L1 elements. These experiments show that ORF2 does not inhibit transcription initiation, because polymerase density in the early region of GFPmORF2 is similar to the same region in GFPlacZ (Fig. 4b, compare GFP1 probes). However, whereas RNA polymerase proceeds through the GFPlacZ transcript processively, functionally engaged RNA polymerase density decreases gradually as transcription is assayed along the ORF2 sequence (Fig. 4b). Thus, the ORF2 sequence seems to be a poor substrate for transcriptional elongation. This could be due to slow elongation rate, stalling of the RNA polymerase complex, or premature dissociation.

Evidence for L1 involvement in transcriptome evolution

Because L1 is a mobile element with relatively non-specific target site selection[3,12,14,21,22,26], the observed transcriptional properties of L1 sequences take on potentially great significance. When L1 elements insert into introns, the new L1 sequence inevitably becomes part of the target gene and its transcript. On the basis of our data, we predict that L1 insertions in either orientation could attenuate the expression of target genes by premature truncation of RNA (for antisense insertions) or a transcriptional elongation defect (for sense insertions), each of which decreases the production of full-length pre-messenger RNA. Because the L1 sequence in the sense orientation with respect to transcription of the target gene decreases the production of full-length products to a greater extent than the L1 antisense sequence (Fig. 1d), we expect that this effect would require somewhat more inserted antisense sequence within the target gene. However, decreased target gene expression is expected in both L1 orientations. To investigate these hypotheses, we used expression profiling data to select the 5% most highly and most poorly expressed genes in humans. We then examined the genomic loci of these two sets of genes and used RepeatMasker to search the predicted pre-mRNA transcripts for L1 sequence.

The average total amount of L1 sequence present per gene was markedly different for highly expressed and poorly expressed genes. Highly expressed genes had small amounts of L1 sequence (an average of 918 nucleotides of sense L1 per gene and 1,760 nucleotides of antisense L1 per gene), whereas poorly expressed genes had large amounts of L1 sequence (an average of 4,760 nucleotides of sense L1 per gene and 8,860 nucleotides of antisense L1 per gene; see Fig. 5a). In comparison with randomly selected populations of genes, these values represent the two possible extremes on the spectrum (Fig. 5b). When the total amount of L1 sequence present was normalized for total intron amount, L1 was still underrepresented in highly expressed genes (Fig. 5c), showing that this effect cannot be entirely explained by the greater total intron content of the poorly expressed genes. These data suggest that L1 sequence content is one of many contributing factors that determine the expression of a particular gene. Further analysis shows that this relationship holds even within specific isochores (L1-poor/highly expressed or L1-rich/poorly expressed) of the human genome (Fig. 5d).

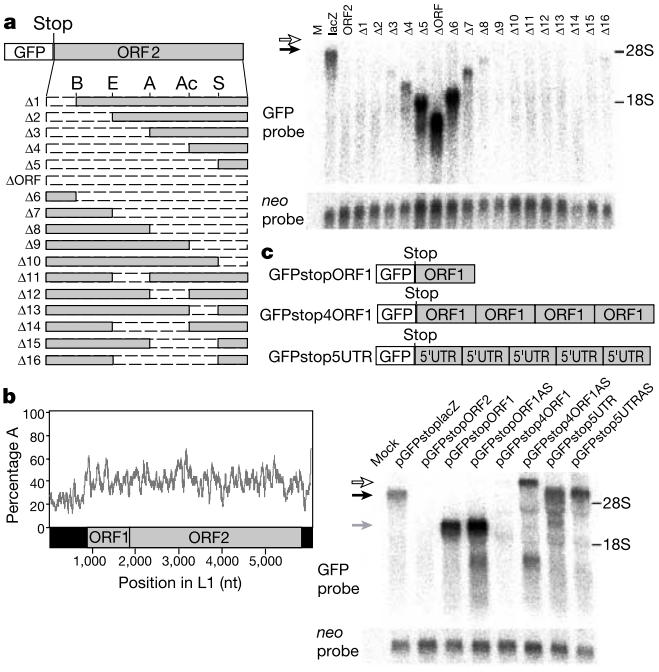

Figure 3 Decrease in L1 expression is dependent on length. **a**, The left panel depicts the structures of deletion constructs. Hollow regions represent deleted sequences. B, *Bbv*CI; E, *Eco*RI; A, *Afl*II; Ac, *Acl*I; S, *Spe*I. The right panel shows a total RNA analysis of HeLa transfections. Lanes: M, mock; lacZ, pGFPstoplacZ; ORF2, pGFPstopORF2. Open and black arrows show the expected positions of GFPstopORF2 and GFPstoplacZ, respectively. **b**, The adenosine base composition of the sense strand, in 50-nucleotide windows, was plotted for each position in L1.2 with MacVector 6.5.3 (Oxford Molecular). **c**, The top panel shows the structures of GFPstopORF1, GFPstop4ORF1 and GFPstop5UTR. The 4ORF1 repeat is about 4,500 nucleotides long and the 5′ UTR repeat is about 4,000 nucleotides long. The bottom panel shows a total RNA analysis of HeLa transfections. Open, black and grey arrows show the expected positions of GFPstop4ORF1, GFPstop5UTR and GFPstopORF1, respectively.

Discussion

L1 ORF sequences in the sense orientation serve as a poor substrate for transcription. The inhibition of transcription by the L1 ORFs could well serve as a negative regulatory mechanism evolved by the

host to prevent excessive retrotransposition. This is consistent with the ability to express L1 ORF2 in non-mammalian organisms that lack the L1 family of retrotransposons and therefore might not have evolved the required regulatory machinery. It is possible that in retrotransposition competent tissues (for example germ cells[17,18]) a more processive form of the RNA polymerase II complex is recruited to the L1 promoter and that this bypasses the elongation defect. This could be due either to germ-cell- or other cell-specific transcription factors that affect the type of RNA polymerase elongation complex formed at the L1 promoter[27–29], or germ-cell-specific elongation factors[30,31] that enhance active L1 transcription.

The changes in target gene expression and structure that could result from an L1 insertion are potentially of even greater significance. Because RNA polymerase gradually pauses and/or dissociates from the template as it encounters longer stretches of L1 sequence, we expect L1 insertions to attenuate expression of the target gene (Fig. 6a). Bioinformatic data presented here support the hypothesis that L1 insertions attenuate gene expression and have a profound impact on the human transcriptome. Our data are also consistent with examples of known *de novo* full-length L1 insertions into introns that lead to very significantly decreased RNA amounts of target genes in mammalian cells[32,33]. Further experiments will be needed to determine whether the inhibitory effect is cumulative over multiple segments of L1 ORF-derived sequence or whether all L1 sequence must be in a single contiguous block to inhibit elongation.

When molecular parasites invade a genome, they evolve to maximize their own survival, often to the detriment of their host[34,35]. Nevertheless, this unwelcome guest might become useful over evolutionary time, as has occurred with significant consequences in *Drosophila*, in which telomeres are produced by retrotransposition[36], and in the evolution of the vertebrate adaptive immune system[37]. In the present case, L1 elements might have had an integral role in the evolution of humans by repeatedly fine-tuning gene expression. Because L1s are found in high copy number in all mammals, and mouse ORF2 behaves similarly to human ORF2, this proposed model could apply throughout mammalian evolution. The fine-tuning could be a cumulative effect of small L1 insertions into introns that individually might have relatively minor effects on expression. When such expression adjustments are beneficial, the insertions will provide a selective advantage and become fixed in the host genome; when effects are deleterious, the insertion alleles will fail to become fixed in the gene pool. Because large insertions within genes are likely to have an immediate, marked impact on expression of the target, they are more likely to be selected against[13].

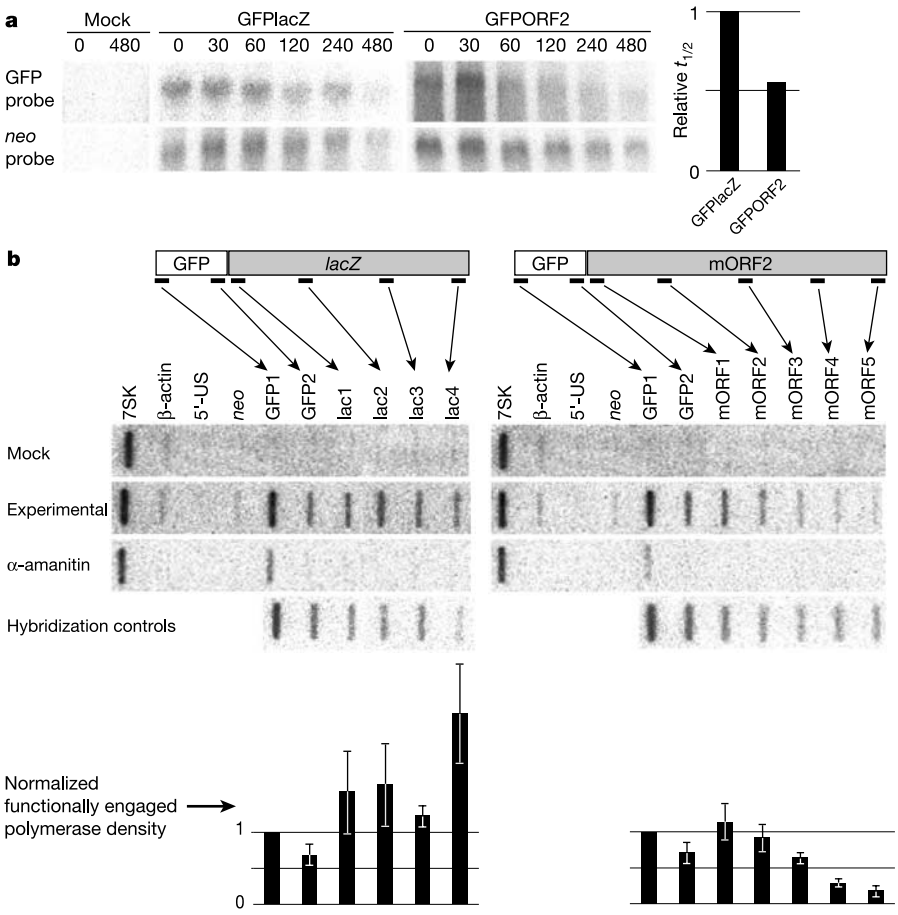

Figure 4 Analysis of ORF2 stability and transcription. **a**, Half-life measurements. In the left panel, transfected HeLa cells were treated with actinomycin D 48 h after transfection. Total RNA was collected 0, 30, 40, 120, 240 and 480 min after treatment and quantified by blotting. In the right panel, the half-lives ($t_{1/2}$) of GFPlacZ or GFPORF2 (relative to the $t_{1/2}$ of the *neo* control) were calculated and compared with $t_{1/2,GFPlacZ}/t_{1/2,neo}$ set to 1. **b**, Nuclear run-on analysis (NRO). Nuclei were isolated from HeLa cells 36 h after transfection and used for NRO. Bold lines under GFP, *lacZ* and mORF2 indicate probe positions. 7SK controls for RNA polymerase III transcription. β-actin is a control for RNA polymerase II transcription. 5′-US is a negative control that hybridizes to a region upstream of the cytomegalovirus (CMV) promoter. *neo* controls for transfection. Hybridization controls are described in Methods. Normalized functionally engaged polymerase density is the signal ($N = 3$) corrected for α-amanitin-resistant transcription and hybridization efficiency, with GFP1 set to 1. Error bars show the standard deviation.

Our findings also indicate that L1 insertions into endogenous genes could potentially generate novel mRNA, and consequently protein isoforms, by producing prematurely truncated, stable transcripts in addition to the original gene product (Fig. 6b). An example of a useful short mRNA/protein isoform generated from an L1 insertion has been shown for the human ATRN gene[38]. In this example, the L1 insertion is short (212 base pairs) and the polyadenylation signal used is the presumed native L1 polyadenylation signal. We predict that truncated mRNA isoforms can be produced from other polyadenylation signals in L1 coding sequences, particularly in the L1 antisense orientation. We have found examples of human transcripts that use precisely these cryptic poly(A) signals (for example, SPC25 gene transcripts, some of which end at poly(A) signal II from Fig. 1d) (S. Wheelan and J.D.B., manuscript in preparation). The cryptic poly(A) signals seem to occur at sites scattered throughout the element sequence. All of these contain the highly conserved AATAAA element or a subtle variant (AATTAAA or ATAAA) at the appropriate distance from the site of cleavage and polyadenylation (Fig. 1e). The downstream GU/U-rich regions of poly(A) sites are highly variable[39], and we suspect that the 40% U content of the ORF2 antisense strand might account for the stronger polyadenylation signals observed in the antisense case. The weakness of the sites in the sense strand presumably reflects ongoing selection on L1 to make full-length RNA copies for retrotransposition. Premature polyadenylation of L1 sense sequences has recently been noted by others[19] and is consistent with this aspect of the proposed model.

These models (Fig. 6a, b) predict that L1 insertions could have major effects on both the quality and quantity of genome-wide mRNAs. Longer L1 insertions should have more potent effects and are therefore more likely to be disastrous for target gene expression. The potential danger of long L1 insertions is consistent with the observed profile of L1 insertions in humans, in which short segments corresponding to 3′-terminal fragments of the L1 element dominate the length distribution[14,20–22]. This is probably due to a combination of non-processive reverse transcription and the sub-

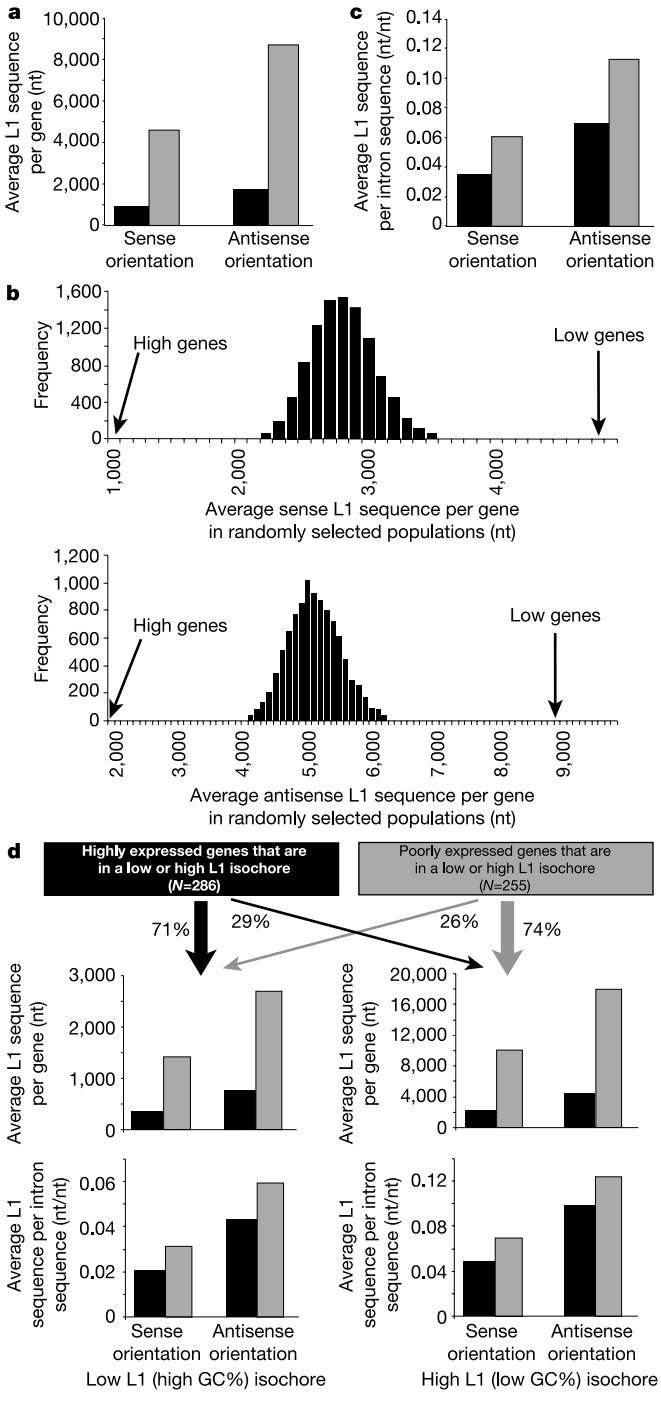

Figure 5 Bioinformatic analysis of L1 content in genes. **a**, Average L1 content of genomic loci of sets of highly (black bars) and poorly (grey bars) expressed genes (see Methods). **b**, Average L1 content in sets of randomly selected populations of genes (see Methods). Positions where the highly and poorly expressed genes would be (data superimposed from **a**) are indicated and are outside the random distribution ($P < 0.01$). **c**, Data from **a**, normalized to total intron content. **d**, Highly and poorly expressed genes were sorted into high GC, low L1 isochore or low GC, high L1 isochore[50] classes. The percentage of each expression class falling into each isochore is indicated. Subpopulations were analysed as described in **a** and **c**.

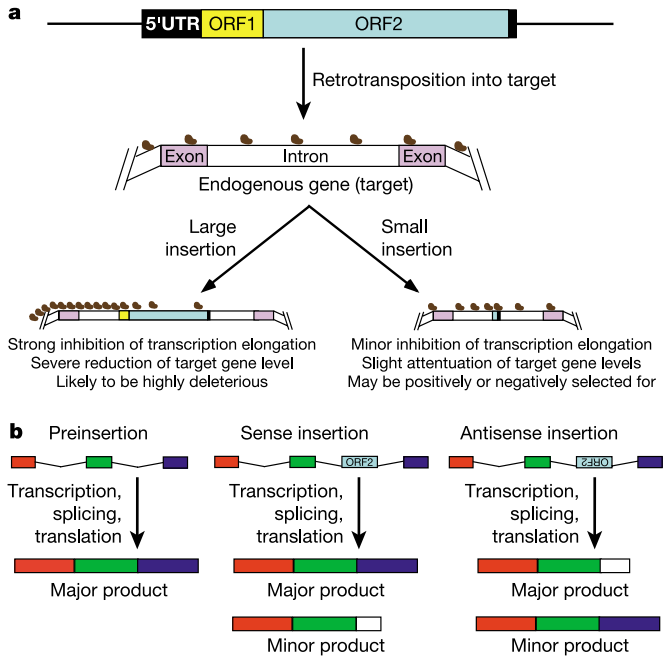

Figure 6 Models for L1-mediated modulation of gene expression/structure. **a**, Effects on transcription. Brown dots represent transcriptional complexes, which could be slowed, paused or dissociated from the templates on encountering significant amounts of L1 sequence. **b**, Effects on mRNA and protein structure. Left, hypothetical gene with three exons. Middle, intronic sense L1 insertions can produce a minor amount of prematurely polyadenylated mRNA, potentially giving rise to a truncated protein with additional, previously untranslated amino acids at the C terminus (white segment). Right, intronic antisense L1 insertions can produce a major amount of prematurely polyadenylated mRNA.

sequent 'purifying selection' against long insertions[13]. In addition, if L1 functions to fine-tune gene expression, retrotransposition-competent (and by definition full-length) L1 elements could provide a selective advantage to the host and might be present at a higher than expected frequency. This might help to explain a puzzling aspect of the L1 insertion profile—the overrepresentation of full-length L1 elements[13,14]. Although many such elements are currently non-functional because of mutation, most were probably previously active.

L1 elements and their kin in other eukaryotes have been shown to rearrange genes and chromosomes by a wide variety of mechanisms. In addition to gene inactivation through insertion into exons, insertions in tissue culture cells are often associated with genomic instability[21,22]. Homologous recombination between dispersed copies of retrotransposon sequences can also lead to chromosome rearrangements[40]. Finally, transcription past the L1 element polyadenylation site during retrotransposition can lead to the 'transduction' of 3′ sequences flanking the donor L1 element[41–43]. A symmetrical process involving 5′ sequences can occur when a cellular promoter upstream of an L1 element transcribes into an active L1 element[22]. Here we propose an addition to the retrotransposon repertoire of genome remodelling activities—the insertion of expression-modulating sequences into host gene introns to reprogram gene expression. As recent genome sequencing studies have shown, the gene complements of mammals can be remarkably similar[44]. Thus, a major component of mammalian speciation might result from subtle transcriptome reprogramming through alterations in exon usage patterns as well as amounts of gene expression. Even within a species, alleles that show *cis*-acting heritable variations in expression are relatively common in normal individuals[45,46] and might account for phenotypic differences. The proposed model suggests the possibility that L1 elements could have a major function in such variations, both within and between species. Rigorous experimental proof of this model requires knocking L1 sequences into the introns of specific mammalian genes, generating isogenic filled and empty alleles, and comparing the quantity and quality of the resultant gene products. □

Methods

Oligonucleotide sequences, plasmid construction

All oligonucleotide sequences described in this manuscript are shown in Supplementary Table S1. General structures of test expression reporters are shown (Supplementary Fig. S1); details of plasmid construction are available from the authors on request.

Cell culture and transfection

Cell culture and transfections were performed as described[25]. Proportions were scaled up linearly for 150-mm dishes. For downstream northern, immunoblot or nuclear run-on analysis, cells were harvested 36–48 h after transfection.

Northern blot analysis

Northern blots were performed essentially as described[25]. Prehybridizations and hybridizations were performed in 50% formamide, 5× SSC, 5× Denhardt's solution, 1% SDS, and 100 μg ml^{-1} boiled herring-sperm DNA at 42 °C. The following [γ-^{32}P]ATP end-labelled oligonucleotides were used as probes: GFP probe, JB4057; *neo* probe, JB4059; 3′ UTR probe, JB5574.

Image Gauge v. 3.0 (Fuji Photo Film Co.) was used to quantify the signal ($N = 3$). Each band was corrected by subtracting a lane-specific background. Corrected test transcripts were normalized to corrected *neo* transcripts.

For half-life measurements ($N = 2$–4), actinomycin D was used at 5 μg ml^{-1}. A radiolabelled 500-base-pair fragment of the enhanced GFP gene was used to detect GFP fusion transcripts. End-labelled JB4059 was used to detect *neo* transcripts. GFPORF2 samples were exposed 50-fold longer than GFPlacZ to make them visible. The quantified transcripts, normalized to *neo*, were plotted on a semi-logarithmic axis against time, where the slope from a best-fit line represents the decrease in relative transcript over time. The slope from these plots of GFPlacZ and GFPORF2 were compared to obtain the relative half-lives shown in Fig. 4a (right).

Western blot analysis

Immunoblots were performed as described elsewhere[25]. Anti-GFP(FL) antibody (Santa Cruz) was used at 1:500 dilution. Anti-rabbit IgG (Amersham) was used at 1:5,000 dilution. Anti-Myc antibody (Santa Cruz) was used at 1:1,000 dilution. Anti-mouse IgG (Amersham) was used at 1:5,000 dilution.

Fluorescence microscopy

Cells were visualized with a Nikon Eclipse TE300 microscope using a fluorescein isothiocyanate filter (Chroma B-2E/C) for GFP visualization. The same exposure time was used for all fluorescence pictures.

RT–PCR cloning

First-strand synthesis was performed with SUPERSCRIPT II RNaseH$^-$ Reverse Transcriptase (Invitrogen) in accordance with the manufacturer's instructions. For each first-strand synthesis reaction, 40 ng DNase I-treated total RNA was used. The primer 3RACERT was used for first-strand synthesis. A 1-μl portion of the first-strand synthesis reaction was used in a PCR reaction containing 100 mM dNTPs, each primer at 500 nM, and 5 units of Amplitaq (Perkin Elmer). Amplification primers were 3RACEAMP and JB4360. Amplification products were separated on a 1% agarose gel and dominant bands were excised and cloned using the TOPO TA cloning kit (Invitrogen). DNA sequencing was performed by Agencourt.

Nuclear run-on analysis

Isolation and transcription of nuclei. Nuclei were isolated and run-on reactions were performed essentially as described previously[47]. Reactions were stopped by the addition of TRIzol, and RNA was isolated in accordance with the manufacturer's instructions. RNA was hydrolysed and hybridized as described previously[47].

Filter preparation/hybridization. Each oligonucleotide (0.75 pmol) was diluted in 100 μl of 0.5 M NaOH and slot-blotted on a GeneScreen Plus nylon membrane. Filters were neutralized with 100 mM Tris-HCl, pH 8, crosslinked by ultraviolet irradiation, then prehybridized and hybridized in 50% formamide, 5× SSC, 5× Denhardt's solution, 1% SDS and 100 μg ml^{-1} boiled herring-sperm DNA.

Control hybridization transcripts. A T7 transcription reaction of each control transcript with a C-terminal ATP-binding aptamer[48] was performed *in vitro* with radiolabelled CTP. Full-length transcripts were purified on an ATP-agarose column as described previously[49]. These transcripts were hydrolysed and hybridized as described above.

Signal quantification. Each quantified slot-blot signal was corrected for membrane background by subtracting signal from an adjacent region of the membrane. Each slot-blot signal was also corrected for α-amanitin-resistant transcription. To measure elongation, the corrected experimental signal was divided by the hybridization correction factor (background-corrected control hybridization signal); results were then normalized to GFP1 by dividing all signals by the GFP1 signal. To measure initiation, the corrected GFP1 experimental signal was divided by the corresponding *neo* experimental signal.

Bioinformatic analysis of L1s in genes

The gene expression data of normal tissues profiled by Gene Logic Inc. were analysed to identify 983 genes with the highest expression and 866 genes with the poorest expression. The genomic coordinates (transcript start to transcript end) of 16,597 RefSeq genes and the genome RepeatMasker*.out files were downloaded from http://genome.ucsc.edu/ (June 2002 genome assembly, build 12). The RepeatMasker*.out files were parsed to identify the genomic coordinates of L1 elements. These L1 coordinates were used to calculate the amount of 'LINE1/L1' in each transcript.

The distributions of L1s in random genes were constructed by performing bootstrapping. In brief, a set of genes was randomly generated from the 16,597 RefSeq genes. The L1 content was recorded for each random gene set. This process was repeated for 10,000 iterations to generate a distribution.

Isochores[50] with low (40% or less) GC content and high (50% or more) GC content were obtained from http://bioinfo2.ugr.es/isochores/. The highly and poorly expressed genes belonging to these subsets were analysed for L1 content as described above. Genes spanning isochore boundaries were added to both isochores (this was rare). Some genes did not fall into either isochore subset and were therefore omitted from this analysis.

Received 22 January; accepted 30 March 2004; doi:10.1038/nature02536.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Swergold, G. D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol. Cell. Biol.* **10,** 6718–6729 (1990).
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. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
6. Wei, W. *et al.* Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell. Biol.* **21,** 1429–1439 (2001).
7. 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).
8. Martin, S. L. & Bushman, F. D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **21,** 467–475 (2001).
9. Kolosha, V. O. & Martin, S. L. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J. Biol. Chem.* **278,** 8112–8117 (2003).
10. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genet.* **35,** 41–48 (2003).
11. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72,** 595–605 (1993).

12. 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).
13. Boissinot, S., Entezam, A. & Furano, A. V. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* **18,** 926–935 (2001).
14. Szak, S. T. *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3,** research00521–researc005218 (2002).
15. 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).
16. Clements, A. P. & Singer, M. F. The human LINE-1 reverse transcriptase: effect of deletions outside the common reverse transcriptase domain. *Nucleic Acids Res.* **26,** 528–3535 (31998).
17. Branciforte, D. & Martin, S. L. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol. Cell. Biol.* **14,** 2584–2592 (1994).
18. Trelogan, S. A. & Martin, S. L. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc. Natl Acad. Sci. USA* **92,** 1520–1524 (1995).
19. Perepelitsa-Belancio, V. & Deininger, P. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genet.* **35,** 363–366 (2003).
20. 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).
21. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110,** 315–325 (2002).
22. Symer, D. E. *et al.* Human L1 retrotransposition is associated with genetic instability *in vivo*. *Cell* **110,** 327–338 (2002).
23. Feng, Q. *Mechanism of Human L1 Element Retrotransposition*. Thesis, Johns Hopkins Univ. (1996).
24. Chávez, S. *et al.* Hpr1 is preferentially required for transcription of either long or G + C-rich DNA sequences in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **21,** 7054–7064 (2001).
25. Han, J. S. & Boeke, J. D. A highly active synthetic mammalian retrotransposon. *Nature* **429,** 314–318 (2004).
26. 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).
27. Minakami, R. *et al.* Identification of a *cis*-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res.* **20,** 3139–3145 (1992).
28. Tchenio, T., Casella, J. F. & Heidmann, T. Members of the SRY family regulate the human LINE-1 retrotransposons. *Nucleic Acids Res.* **28,** 411–415 (2000).
29. Yang, N., Zhang, L., Zhang, Y. & Kazazian, H. H. Jr An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res.* **31,** 4929–4940 (2003).
30. Xu, Q., Nakanishi, T., Sekimizu, K. & Natori, S. Cloning and identification of testis-specific transcription elongation factor S-II. *J. Biol. Chem.* **269,** 3100–3103 (1994).
31. Miller, T., Williams, K., Johnstone, R. W. & Shilatifard, A. Identification, cloning, expression, and biochemical characterization of the testis-specific RNA polymerase II elongation factor ELL3. *J. Biol. Chem.* **275,** 32052–32056 (2000).
32. Schwahn, U. *et al.* Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nature Genet.* **19,** 327–332 (1998).
33. Yajima, I. *et al.* An L1 element intronic insertion in the black-eyed white (Mitf[mi-bw]) gene: the loss of a single Mitf isoform responsible for the pigmentary defect and inner ear deafness. *Hum. Mol. Genet.* **8,** 1431–1441 (1999).
34. Hickey, D. A. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101,** 519–531 (1982).
35. Bestor, T. H. Sex brings transposons and genomes into conflict. *Genetica* **107,** 289–295 (1999).
36. Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A. & Sheen, F. M. Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell* **75,** 1083–1093 (1993).
37. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394,** 744–751 (1998).
38. Tang, W. *et al.* Secreted and membrane attractin result from alternative splicing of the human ATRN gene. *Proc. Natl Acad. Sci. USA* **97,** 6025–6030 (2000).
39. Zarudnaya, M. I., Kolomiets, I. M., Potyahaylo, A. L. & Hovorun, D. M. Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res.* **31,** 1375–1386 (2003).
40. Burwinkel, B. & Kilimann, M. W. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J. Mol. Biol.* **277,** 513–517 (1998).
41. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr Exon shuffling by L1 retrotransposition. *Science* **283,** 1530–1534 (1999).
42. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9,** 653–657 (2000).
43. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10,** 411–415 (2000).
44. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420,** 563–573 (2002).
45. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297,** 1143 (2002).
46. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13,** 1855–1862 (2003).
47. Cuello, P., Boyd, D. C., Dye, M. J., Proudfoot, N. J. & Murphy, S. Transcription of the human U2 snRNA genes continues beyond the 3′ box *in vivo*. *EMBO J.* **18,** 2867–2877 (1999).
48. Dieckmann, T., Butcher, S. E., Sassanfar, M., Szostak, J. W. & Feigon, J. Mutant ATP-binding RNA aptamers reveal the structural basis for ligand binding. *J. Mol. Biol.* **273,** 467–478 (1997).
49. Sassanfar, M. & Szostak, J. W. An RNA motif that binds ATP. *Nature* **364,** 550–553 (1993).
50. Oliver, J. L. *et al.* Isochore chromosome maps of the human genome. *Gene* **300,** 117–127 (2002).

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

Acknowledgements We thank Y. Aizawa, J. Corden, J. Moran, J. Nathans, S.-L. Ooi and D. Valle for helpful discussions and critical reading of the manuscript, S. Wheelan for unpublished bioinformatics data, S. Murphy for providing a detailed nuclear run-on protocol, J. Moran for pY104, H. Kazazian for pTN201, and R. Bandaru for help with statistical analysis. 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).