30 min at 37°C, either with or without preheating at 90°C for 5 min, followed by electrophoresis in a 15% denaturing gel.

For detection of *rev*-EGFP mRNA, we used a 25-mer deoxyribonucleotide probe that was complementary to the EGFP mRNA of the *rev*-EGFP fusion protein. A 29-mer deoxyribooligonucleotide probe was used for detection of the GAPDH transcript.

HIV-1 antiviral assay. For determination of anti-HIV-1 activity of the siRNAs, transient assays were done by cotransfection of siDNAs and infectious HIV-1 proviral DNA, pNL4-3 into 293 cells as described[15]. Before transfection, the cells were grown for 24 h in six-well plates in 2 ml EMEM supplemented with 10% (vol/vol) FBS and 2 mM L-glutamine, and transfected using Lipofectamine Plus reagent (Life Technologies, GibcoBRL) as described by the manufacturer. The DNA mixtures consisting of 0.5 µg siDNAs or controls, and 0.5 µg pNL4-3 were formulated into cationic lipids and applied to the cells. After one, two, three, and four days, supernatants were collected and analyzed for HIV-1 p24 antigen (Beckman Coulter, Hialeah, FL). The p24 values were calculated with the aid of the Dynatech MR5000 ELISA plate reader (Dynatech Labs Inc., Chantilly, VA). Cell viability was also assessed using a Trypan Blue dye exclusion count at four days after transfection.

Acknowledgments
We thank Aaron Coleman for providing pIND and G. Pavlakis and A. Michienzi for the CMV-rev-EGFP vector. The authors also wish to thank Alessandra Poggi for help with some preliminary experiments. This research was supported by grants from the National Institutes of Health AI 29329, AI42552 and AI 46030.

Competing interests statement
The authors declare competing financial interests: see the Nature Biotechnology *website (http://biotech.nature.com) for details.*

Received 21 December 2001; accepted 15 March 2002

1. Hammond, S.M., Bernstein, E., Beach, D. & Hannon, G.J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
2. Fire, A. RNA-triggered gene silencing. *Trends Genet.* **15**, 358–363 (1999).
3. Svoboda, P., Stein, P., Hayashi, H. & Schultz, R.M. Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference. *Development* **127**, 4147–4156 (2000).
4. Sharp, P.A. RNA interference—2001. *Genes Dev.* **15**, 485–490 (2001).
5. Clemens, J.C. *et al.* Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl. Acad. Sci. USA* **97**, 6499–6503 (2000).
6. Elbashir, S.M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
7. Bernstein, E., Caudy, A.A., Hammond, S.M. & Hannon, G.J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
8. Dalmay, T., Horsefield, R., Braunstein, T.H. & Baulcombe, D.C. SDE3 encodes an RNA helicase required for post-transcriptional gene silencing in *Arabidopsis*. *EMBO J.* **20**, 2069–2078 (2001).
9. Nykanen, A., Haley, B. & Zamore, P.D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
10. Lipardi, C., Wei, Q. & Paterson, B.M. RNAi as random degradative PCR. siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
11. Smardon, A. *et al.* EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. *Curr. Biol.* **10**, 169–178 (2000).
12. Elbashir, S.M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
13. Caplen, N.J., Parrish, S., Imani, F., Fire, A. & Morgan, R.A. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl. Acad. Sci. USA* **98**, 9742–9747 (2001).
14. Bertrand, E. *et al.* The expression cassette determines the functional activity of ribozymes in mammalian cells by controlling their intracellular localization. *RNA* **3**, 75–88 (1997).
15. Good, P.D. *et al.* Expression of small, therapeutic RNAs in human cell nuclei. *Gene Ther.* **4**, 45–54 (1997).
16. Scherr, M. & Rossi, J.J. Rapid determination and quantitation of the accessibility to native RNAs by antisense oligodeoxynucleotides in murine cell extracts. *Nucleic Acids Res.* **26**, 5079–5085 (1998).
17. Scherr, M. *et al.* Detection of antisense and ribozyme accessible sites on native mRNAs: application to NCOA3 mRNA. *Mol. Ther.* **4**, 454–460 (2001).
18. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
19. Lee, N.S. *et al.* Functional colocalization of ribozymes and target mRNAs in *Drosophila* oocytes. *FASEB J.* **15**, 2390–2400 (2001).

Effective expression of small interfering RNA in human cells

Cynthia P. Paul[1], Paul D. Good[1], Ira Winer[2], and David R. Engelke[1,2]*

In many eukaryotes, expression of nuclear-encoded mRNA can be strongly inhibited by the presence of a double-stranded RNA (dsRNA) corresponding to exon sequences in the mRNA (refs 1,2). The use of this "RNA interference" (RNAi) in mammalian studies had lagged well behind its utility in lower animals because uninterrupted RNA duplexes longer than 30 base pairs trigger generalized cellular responses through activation of dsRNA-dependent protein kinases[3]. Recently it was demonstrated[4] that RNAi can be made to work in cultured human cells by introducing shorter, synthetic duplex RNAs (~20 base pairs) through liposome transfection. We have explored several strategies for expressing similar short interfering RNA (siRNA) duplexes within cells from recombinant DNA constructs, because this might allow long-term target-gene suppression in cells, and potentially in whole organisms. Effective suppression of target gene product levels is achieved by using a human U6 small nuclear RNA (snRNA) promoter to drive nuclear expression of a single RNA transcript. The siRNA-like parts of the transcript consists of a 19–base pair siRNA stem with the two strands joined by a tightly structured loop and a U_{1-4} 3′ overhang at the end of the antisense strand. The simplicity of the U6 expression cassette and its widespread transcription in human cell types suggest that this mode of siRNA delivery could be useful for suppressing expression of a wide range of genes.

The U6 snRNA promoter cassettes and si-like RNA inserts are shown in Figure 1. We previously showed that RNA expressed by RNA polymerase III from the U6+1 or U6+27 cassettes was expressed primarily as full-length transcripts and was located in the nucleus[5,6]. U6+27 transcripts, containing the first 27 nucleotides of human U6 RNA, were capped with γ-methyl phosphates and accumulated to higher levels than U6+1 transcripts. Cassettes are designed so that short RNA coding sequences are inserted between unique *Sal*I and *Xba*I sites. After the *Xba*I site, the cassette encodes a strong stem to protect the transcripts against 3′-5′ exonuclease attack, then a poly(U) transcription termination sequence. However, the insertion sequences discussed later also contain their own UUUU terminator at the 3′ end of the inserted sequences, terminating most transcription before the cassette-encoded stem/terminator region.

To test whether expressed si-like RNA is effective, we targeted a site in human lamin A/C mRNA that has been demonstrated to be vulnerable to synthetic siRNA[4]. The inserted sequences encoded several variants of siRNA duplexes and controls, shown in Figure 1B. Previous work on synthetic anti-lamin A/C siRNA used two independent strands with 3′ unpaired tails[4]. Although it would be theoretically possible to synthesize two strands independently *in vivo*, the need to anneal the two strands could make the production

[1]*Department of Biological Chemistry and* [2]*Medical Scientist Training Program, The University of Michigan, Ann Arbor, MI 48109-0606.*
**Corresponding author (engelke@umich.edu).*

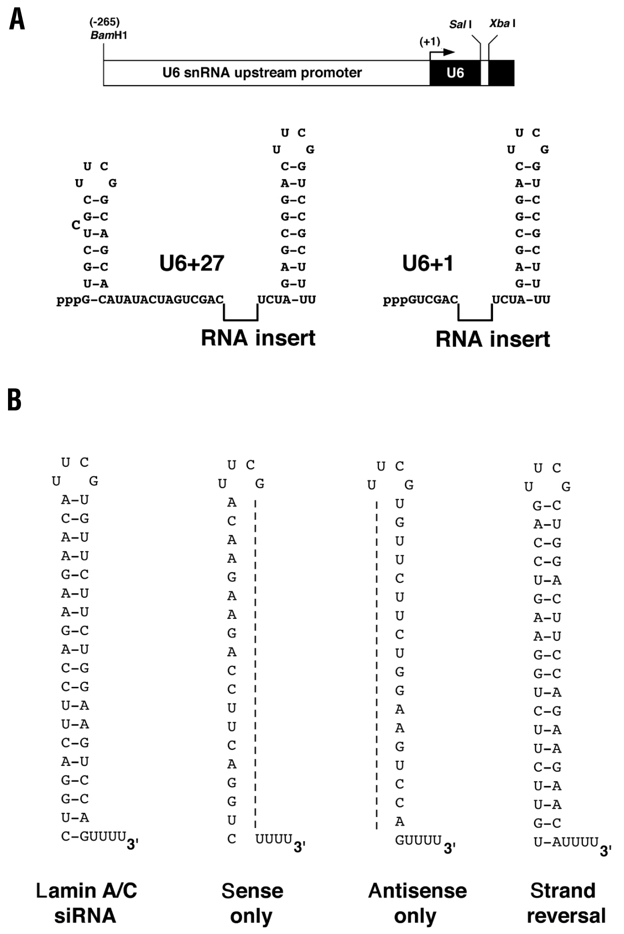

Figure 1. Expression cassettes and small RNA inserts. (A) The two U6 snRNA promoter[12–14] expression cassettes used to express siRNAs and controls are shown with the expected transcripts by RNA polymerase III, assuming no UUUU terminators in the RNA insert. Cassettes had either no remaining U6 snRNA sequences (U6+1) or the first 27 nucleotides of U6 snRNA (U6+27) to direct methylation of the 5′-γ-phosphate and stabilize the transcript[5]. With the inserts shown, most transcription terminates with the insert UUUU, but readthrough to the cassette stem terminator also occurs. (B) Four tested anti-lamin RNA inserts are shown. Each would begin immediately after the *Sal*I sequence from the cassette, and most termination occurs after the UUUU at the insert 3′ terminus (ref. 5 and data not shown).

of siRNA inefficient, and might complicate routine cloning and expression of the siRNA constructs. However, as seen elsewhere in this issue (Lee *et al.*, p. 500), the synthesis of the siRNA as independent strands from U6 promoter can also be effective.

To make the siRNA duplex as one short transcript, an RNA insert was used that contains the 19-nucleotide sense strand of the target, followed by a UUCG tetraloop sequence[7], the antisense strand, and a UUUU transcription terminator, in that order. This terminates a high percentage of the transcripts exactly at the end of the siRNA stem[5]. The 3′-UUUU overhang after the siRNA is attacked by 3′ exonucleases, leaving 1 to 4 U 3′-end overhangs[5]. Results with synthetic siRNA[4] suggest that such 3′ overhangs can increase efficacy. RNA blot analysis has shown that high levels (~10^4–10^5 RNA molecules/cell) of nearly full-length RNAs can be expressed from these cassettes[5]. The "hairpin siRNAs" give comparable expression levels, although there is a complex pattern of breakdown products as well as the full-length product (not shown).

Figure 2 shows elimination of lamin A/C protein when HeLa cells were transiently transfected with either synthetic siRNA or siRNA-expressing clones. Cells shown in Figure 2A underwent Oligofectamine-mediated transfection with either no RNA or a synthetic 19–base pair siRNA duplex with 3′-TT overhangs[4]. Nuclei in Figure 2A were visualized with 4,6-diamidino-2-phenylindole (DAPI) staining (blue). As expected from previous work, the lamin A/C signal (red) substantially disappears from most cells, presumably those that are transfected. For testing recombinant DNAs, cells were cotransfected with a plasmid (pCMVβ) expressing β-galactosidase (β-Gal) to mark transfected cells. Production of β-Gal also precludes the possibility that siRNA constructs nonspecifically obstruct protein synthesis. Figure 2B shows individual frames with staining for lamin A/C (red), β-Gal (green), or the overlay of the two signals. Without the siRNA inserts, cells transfected with any of the expression cassette plasmids do not have detectably reduced lamin A/C signal (shown only for U6+1 in Fig. 2B). When either U6+1 or U6+27 cassettes were used with anti-lamin hairpin siRNA inserts, dramatic reductions of lamin A/C signal were observed relative to the untransfected cells in the same fields. Transfected cells receiving the U6+27-siRNA expression cassettes gave the most consistent and greatest lamin A/C reductions (>90%, Table 1), similar to synthetic siRNA (~95%). This might reflect a threshold effect caused by lower levels of the U6+1-expressed siRNA[5].

Figure 2C shows lamin A/C–β-Gal overlay panels for control RNAs expressed from the U6+27 cassette. Expression of only the sense or only the antisense strands of the siRNA in U6+27 did not affect lamin levels, reinforcing the notion that the observed reduction in Figure 2B requires the duplex, a hallmark of siRNA action.

We next tested a U6+27 hairpin siRNA construct with the order of the strands reversed to determine the specific need for an accessible 3′ overhang on the antisense strand of the duplex. Some models for siRNA function predict that siRNA degradation of the target message is amplified by annealing of the antisense strand to the mRNA and extension to a longer duplex with an RNA-dependent RNA polymerase. This condition would indicate the need to have an accessible antisense 3′ terminus so that it can be extended. Surprisingly, there was a significant reduction of the lamin signal with the reversed-strand construct, although it was not as consistent or effective as the original orientation. It is not clear why the reversed-strand construct causes partial reduction of the lamin signal. It is possible that small amounts of breakdown products with 3′-UU overhang are created on the antisense strand of the reversed construct by 3′ exonuclease digestion or a discrete endonuclease cleavage between the strands. Alternatively, these hairpin siRNAs, when expressed within the cells, might not need to act exclusively through primer extension amplification. Although the active form of the nuclear-expressed RNAs will require long-term investigation, we recommend that siRNA transcripts have the sense strand first, followed by a tetraloop and antisense strand ending with a 3′ overhang created by the poly(U) terminator.

Previous studies of siRNA-mediated target cleavage by extracts *in vitro* suggested that the 5′ termini of one or both strands might need to be phosphorylated, and that this might be needed for efficient assembly into obligatory ribonucleoprotein complexes[8,9]. Results pre-

Table 1. Effect of siRNA and expression cassettes on the levels of the lamin A/C protein in transfected cell nuclei

Construct	Percentage lamin A/C in transfected vs. nontransfected cells
pAVU6+27 No insert	130 ± 5
Synthetic anti-lamin siRNA	5 ± 2
pAVU6+27 Anti-lamin siRNA hairpin	9 ± 5
pAVU6+27 Sense strand only	130 ± 40
pAVU6+27 Antisense strand only	130 ± 30
pAVU6+27 Reverse-strands hairpin	25 ± 14

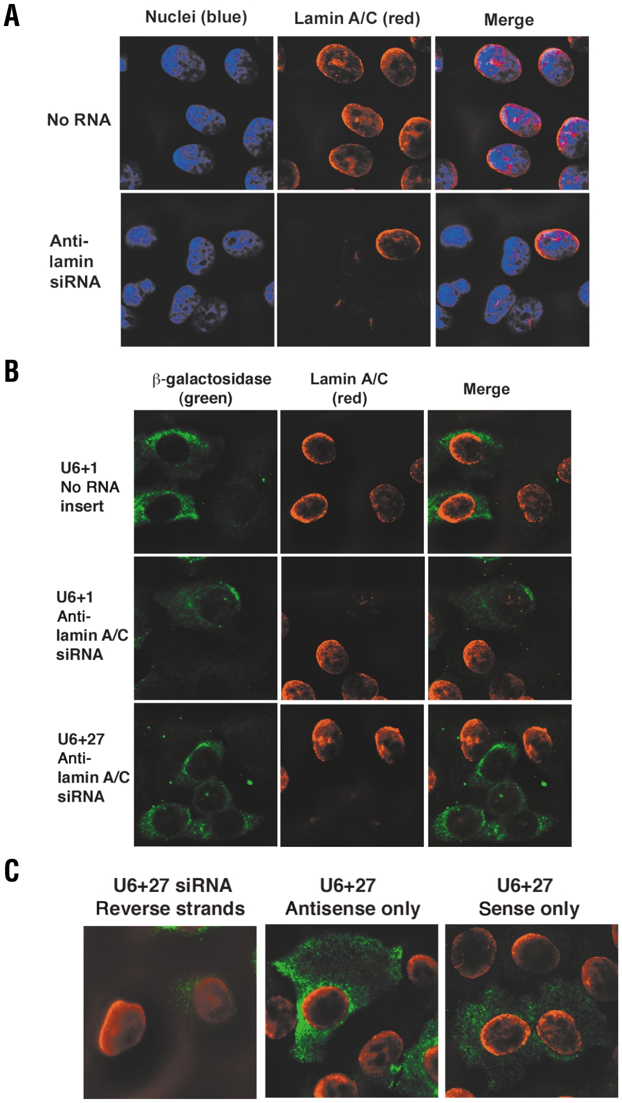

Figure 2. Effects of siRNA constructs on lamin A/C levels. HeLa cells were transfected with either synthetic siRNA or recombinant DNA cassettes expressing different small RNAs from different RNA polymerase III promoters. Cells were stained with DAPI (blue) or with antibodies to lamin A/C (red) or β-Gal (green). (A) Synthetic siRNA or no RNA transfections, showing that lamin A/C staining of the nuclear periphery is largely abolished in most cells, with only low levels of residual red staining in nuclear interiors. (B) Transfection with U6 promoter cassettes either without an siRNA insert (U6+1, no insert) or containing the anti-lamin siRNA shown in Figure 1B (U6+1 siRNA and U6+27 siRNA). Transfected cell cytoplasms are green, whereas nuclei from untransfected cells show no green cytoplasm. Empty expression cassettes have no apparent effect on lamin A/C levels (only empty U6+1 is shown), while transfected cells (green) using siRNA-expressing constructs have little remaining lamin A/C (red). (C) Overlay panels of β-Gal and lamin A/C signal after transfection with different control insertions shown in Figure 1B. Quantitative assessment of remaining lamin A/C signals in transfected cells compared to untransfected cells on the same slide is given in Table 1.

sented here suggest that the hairpin siRNAs might not need to have the 5′ end of either strand unblocked. These hairpin siRNA observations would be consistent with a mechanism in which the U6 transcript containing the hairpin duplex RNA is able to assemble into any necessary protein complexes. The moderate preference for the antisense strand at the end agrees with the prediction that the antisense strand is used as a primer for RNA-dependent RNA polymerase on the message target[10], but this appears not to be essential.

Another unexpected finding was that expression from the U6 snRNA promoter cassettes, which give primarily nucleoplasmic expression[5], would succeed in inhibiting target expression when a majority of the existing mRNA is cytoplasmic. To be certain that the long hairpin did not cause altered localization, we carried out *in situ* hybridization with fluorescent probes to the hairpin that showed nuclear localization very similar to that seen previously (Fig. 3). Admittedly, a small percentage of the U6-driven transcripts might exit to the cytoplasm and be active there. These results suggest that the U6-driven transcripts are suppressing pre-mRNAs before nuclear exit.

The U6 expression cassettes used in these studies are <400 base pairs long and should be relatively easy to incorporate into a variety of vectors. The siRNA inserts can be synthesized as complementary oligodeoxynucleotide pairs to rapidly create cassettes directed at multiple sites. It should even be possible to use several cassettes per vector, targeted at either multiple mRNAs or multiple sites on the same message. It is likely that the hairpin siRNA strategy will be applicable to many mRNA targets. Preliminary experiments targeting both an endogenous human splicing factor and HIV-1 reverse transcriptase coding region (A. Ehsani, S. Li, A. Kleihauer, and J.J. Rossi, personal communication) have shown the hairpin siRNA strategy to be effective. However, as with the synthetic siRNAs, it is sometimes necessary to test several target sites along an mRNA to find one that gives the strongest inhibition. While much remains to be learned about the mechanism by which these transcripts work, the results with the simple U6 cassettes suggest that they might be useful for diverse experimental applications.

Experimental protocol

Materials. Lipofectin, Plus reagent, and Oligofectamine were purchased from Invitrogen (Carlsbad, CA), as were synthetic DNA oligonucleotides for cloning and probes. Cy3-2′-*O*-methyl RNA oligonucleotide hybridization probes were from Trilink (San Diego, CA). Synthetic siRNA oligonucleotides were from Dharmacon (Lafayette, CO). Anti-lamin A/C monoclonal antibodies were purchased from Santa Cruz Biotechnology (Santa Cruz, CA) (sc-7292, used at 1 μg/ml); rabbit anti-β-Gal antibodies were from Molecular Probes (Eugene, OR) (A-11132, 1 μg/ml); Oregon green 488–labeled goat anti-rabbit secondary antibodies were from Molecular Probes (O-11038, 5 μg/ml); and cyanin-3 (Cy3)-labeled goat anti-mouse secondary antibodies were from Amersham-Pharmacia Biotech (Piscataway, NJ) (PA 43002,

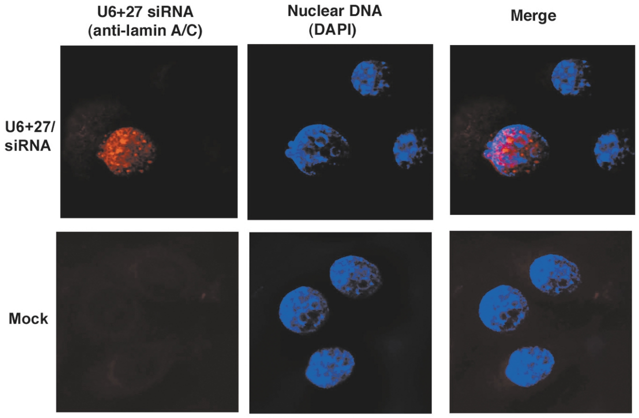

Figure 3. Localization of U6+27 siRNA transcripts. Two days after transfection with the U6+27 anti-lamin A/C cassette plasmid, cells were fixed and stained for nuclear DNA (DAPI, blue) and probed with a Cy3-labeled 2’-*O*-methyl oligoribonucleotide (red) complementary to the antisense strand of the siRNA. As expected from work with previous U6 expression constructs[5,6], the U6+27 siRNA pattern was primarily in a nuclear speckled pattern. Nuclear and cytoplasmic background staining by the Cy3-oligonucleotide in the absence of U6+27-siRNA (“Mock”) was minimal.

1 μg/ml). Cassettes[5] were cloned in pAV vectors, derived from pCWRSVN (ref. 11) by placing the promoter modules between *Bam*HI and *Hin*dIII sites, after modifying the vector. Modifications included destruction of the *Bam*HI site downstream of the Neo cassette, and removal of all sites between the original *Sal*I and *Xho*I sites, inclusive, by cleavage and religation. After inserting the cassettes, a new polylinker was created between the *Hin*dIII and *Sac*II sites. Sequences to be expressed were inserted as synthetic oligodeoxynucleotides precisely between the end of the unique *Sal*I site and the beginning of the unique *Xba*I site. Recombinant constructs were sequenced.

Transfections. Transient transfections were carried out on subconfluent HeLa cells. Synthetic RNA was transfected using Oligofectamine as described[4]. Recombinant DNA constructs were transfected using Lipofectin with Plus reagent according to the manufacturer's instructions. In transient transfections, cells were split after one day. Cells were fixed and examined for lamin protein after three days, and fixed and examined by *in situ* hybridization after two days.

Fluorescence microscopy. Transfected cells were fixed and subjected to previously described protocols for visualizing proteins[4] with antibodies (lamin A/C and β-Gal) or detecting small RNAs (http://singerlab.aecom.yu.edu/protocols) by hybridizing 5′-Cy3-labeled oligos (5′-Cy3-AAACUGGACUUCCAGAAGAACACGAA, 2′-*O*-methyl ribonucleotides) to the fixed preparations. Fluorescence was acquired with a Nikon Eclipse E800 (Tokyo, Japan) with a Hamamatsu Orca II camera (Hamamatsu-City, Japan). For each construct, hundreds of cells were examined to confirm that the selected images were representative. On multiple slides, lamin A/C fluorescence in transfected cells was deconvoluted and quantitated using Isee software (Inovision; Raleigh, NC) and is expressed in Table 1 as a percentage of lamin A/C signal from nontransfected cells on the same slides. Lamin signal was consistently higher in transfected cells than in untransfected cells on the same slide.

Acknowledgments

This work was supported by NIH grant AI40936 to D.R.E. and the Medical Scientist Training Program at the University of Michigan. We thank Gary R. Kunkel for the cloned, human U6 snRNA gene, and Michael Imperiale and Eric Fearon for other materials.

Competing interests statement

The authors declare competing financial interests: see the Nature Biotechnology *website (http://biotech.nature.com) for details.*

Received 28 December 2001; accepted 19 March 2002

1. Zamore, P.D. RNA interference: listening to the sound of silence. *Nat. Struct. Biol.* **8**, 746–750 (2001).
2. Bernstein, E., Denli, A.M. & Hannon, G.J. The rest is silence. *RNA* **7**, 1509–1521 (2001).
3. Gil, J. & Esteban, M. Induction of apoptosis by the dsRNA-dependent protein kinase (PKR): mechanism of action. *Apoptosis* **5**, 107–114 (2000).
4. Elbashir, S.M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
5. Good, P.D. *et al.* Expression of small, therapeutic RNAs in human cell nuclei. *Gene Ther.* **4**, 45–54 (1997).
6. Bertrand, E. *et al.* The expression cassette determines the functional activity of ribozymes in mammalian cells by controlling their intracellular localization. *RNA* **3**, 75–88 (1997).
7. Cheong, C., Varani, G. & Tinoco, I. Jr. Solution structure of an unusually stable RNA hairpin, 5′-GGAC(UUCG)GUCC. *Nature* **346**, 680–682 (1990).
8. Elbashir, S.M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**, 188–200 (2001).
9. Nykänen, A., Haley, B. & Zamore, P.D. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**, 309–321 (2001).
10. Lipardi, C., Wei, Q. & Paterson, B.M. RNAi as random degradative PCR: siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
11. Chatterjee, S., Johnson, P.R. & Wong, K.K. Jr. Dual-target inhibition of HIV-1 *in vitro* by means of an adeno-associated virus antisense vector. *Science* **258**, 1485–1488 (1992).
12. Kunkel, G.R., Maser, R.L., Calvet, J.P. & Pederson, T. U6 small nuclear RNA is transcribed by RNA polymerase III. *Proc. Natl. Acad. Sci.* USA **83**, 8575–8579 (1986).
13. Kunkel, G.R. & Pederson, T. Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used. *Genes Dev*. **2**, 196–204 (1988).
14. Danzeiser, D.A., Urso, O. & Kunkel, G.R. Functional characterization of elements in a human U6 small nuclear RNA gene distal control region. *Mol. Cell. Biol*. **13**, 4670–4678 (1993).

Using the transcriptome to annotate the genome

Saurabh Saha[1,2†], Andrew B. Sparks[1,3†], Carlo Rago[1],
Viatcheslav Akmaev[4], Clarence J. Wang[4],
Bert Vogelstein[1], Kenneth W. Kinzler[1*],
and Victor E. Velculescu[1*]

A remaining challenge for the human genome project involves the identification and annotation of expressed genes. The public and private sequencing efforts have identified ~15,000 sequences that meet stringent criteria for genes, such as correspondence with known genes from humans or other species, and have made another ~10,000–20,000 gene predictions of lower confidence, supported by various types of *in silico* evidence, including homology studies, domain searches, and *ab initio* gene predictions[1,2]. These computational methods have limitations, both because they are unable to identify a significant fraction of genes and exons and because they are unable to provide definitive evidence about whether a hypothetical gene is actually expressed[3,4]. As the *in silico* approaches identified a smaller number of genes than anticipated[5–9], we wondered whether high-throughput experimental analyses could be used to provide evidence for the expression of hypothetical genes and to reveal previously undiscovered genes. We describe here the development of such a method—called long serial analysis of gene expression (LongSAGE), an adaption of the original SAGE approach[10]—that can be used to rapidly identify novel genes and exons.

The LongSAGE method (Fig. 1) generates 21 bp tags derived from the 3′ ends of transcripts that can rapidly be analyzed and matched to genomic sequence data. The method is similar to the original SAGE approach[10], but uses a different type IIS restriction endonuclease (*Mme*I) and incorporates other modifications to produce longer transcript tags. The resulting 21 bp tag consists of a constant 4 bp sequence representing the restriction site at which the transcript was cleaved, followed by a unique 17 bp sequence derived from an adjacent sequence in each transcript. Theoretical calculations show that >99.8% of 21 bp tags are expected to occur only once in genomes the size of the human genome (Table 1A). Likewise, similar analyses based on actual sequence information from ~16,000 known genes suggest that >75% of 21 bp tags would be expected to occur only once in the human genome, with the remaining tags matching duplicated genes or repeated sequences (as discussed below). In contrast, conventional SAGE tags of 14 bp do not allow unique assignment of tags to genomic sequences, though they do allow such assignment to the much less complex compendium of expressed sequence tags (ESTs) and previously characterized mRNAs[10–12]. To optimize the quantification of transcripts, tags are ligated together to form "ditags," which are then concatenated and cloned. Sequencing tag concatemers in parallel allows the identification of up to ~30 tag sequences in each sequencing reaction. Matching tags to genome

[1]*Howard Hughes Medical Institute and the Sidney Kimmel Comprehensive Cancer Center, and* [2]*Program in Cellular and Molecular Medicine, Johns Hopkins Medical Institutions, Baltimore, MD 21231.* [3]*Current address: GMP Genetics, 200 Prospect Street, Waltham, MA 02451.* [4]*Genzyme Molecular Oncology, P.O. Box 9322, Framingham, MA 01701.* [†]*These authors contributed equally to this work.* **Corresponding authors (kinzlke@jhmi.edu and velculescu@jhmi.edu).*