Function-based gene identification using enzymatically generated normalized shRNA library and massive parallel sequencing

Michael Shtutman[a,1], Anil Maliyekkel[a,b,1], Yu Shao[c], C. Steven Carmack[c], Mirza Baig[a], Natalie Warholic[a], Kelly Cole[a], Eugenia V. Broude[a], Timothy T. Harkins[d], Ye Ding[c], and Igor B. Roninson[a,2]

[a]Cancer Center, Ordway Research Institute, Albany, NY 12208; [b]Department of Biochemistry and Molecular Genetics, University of Illinois at Chicago, Chicago, IL 60607; [c]Wadsworth Center, New York State Department of Health, Albany, NY 12208; and [d]454 Life Sciences, Branford, CT 06405

Communicated by Alexander Varshavsky, California Institute of Technology, Pasadena, CA, March 9, 2010 (received for review January 30, 2010)

As a general strategy for function-based gene identification, an shRNA library containing ≈150 shRNAs per gene was enzymatically generated from normalized (reduced-redundance) human cDNA. The library was constructed in an inducible lentiviral vector, enabling propagation of growth-inhibiting shRNAs and controlled activity measurements. RNAi activities were measured for 101 shRNA clones representing 100 human genes and for 201 shRNAs derived from a firefly luciferase gene. Structure-activity analysis of these two datasets yielded a set of structural criteria for shRNA efficacy, increasing the frequencies of active shRNAs up to 5-fold relative to random sampling. The same library was used to select shRNAs that inhibit breast carcinoma cell growth by targeting potential oncogenes. Genes targeted by the selected shRNAs were enriched for 10 pathways, 9 of which have been previously associated with various cancers, cell cycle progression, or apoptosis. One hundred nineteen genes, enriched through this selection and represented by two to six shRNAs each, were identified as potential cancer drug targets. Short interfering RNAs against 19 of 22 tested genes in this group inhibited cell growth, validating the efficiency of this strategy for high-throughput target gene identification.

cancer targets | functional genomics | RNA interference | siRNA design | lentiviral vectors

RNAi, a general tool for targeted gene knockdown in mammalian cells, is carried out using synthetic siRNA duplexes or vectors expressing shRNA. shRNA is processed by enzyme Dicer to yield siRNA, which is incorporated into the RNA-induced silencing complex (RISC). RISC containing the guide (antisense) strand of siRNA causes endonucleolytic cleavage of target mRNA (1). Synthetic siRNA duplexes and shRNA templates are usually developed through rational design based on sequence-based rules that undergo continuous modification and optimization (2). The efficacy of the published rules for siRNA design remains controversial, and their applicability to shRNA is still uncertain (3).

As a tool for identifying genes, inhibition of which confers a selectable phenotype, several groups have generated expression libraries comprising vectors that express synthetic shRNA sequences targeting thousands of genes, typically at three to six different shRNAs per gene (4–10). Such libraries have been used for several types of selection to identify genes, knockdown of which produces a selectable phenotype. Designed shRNA libraries, however, are very expensive and time-consuming to generate for any new organism, and are limited to known genes and splice variants. Additionally, such libraries are not necessarily able to inhibit every gene they target, because the current status of shRNA design provides no assurance that the small number of shRNAs per gene will include active inhibitors.

An alternative strategy for shRNA library construction is enzymatic conversion of randomly fragmented cDNA into templates for shRNA (11–17). A significant drawback of large-scale shRNA libraries derived from cellular cDNA is that the relative abundance of shRNA sequences is proportional to the starting mRNA they target. This variation among individual shRNAs makes it difficult to inhibit lower-abundance mRNAs. A second problem with enzymatically generated libraries is that their large size and undefined composition preclude the use of an efficient "molecular barcoding" approach (6–10) for rapid identification of enriched or depleted sequences.

In the present article, we describe a general strategy for function-based gene identification, applicable to essentially any organism and independent of the status of shRNA design rules. This strategy involves (*i*) enzymatic generation of transcriptome-scale shRNA libraries with relatively even representation of different genes, (*ii*) expression selection of functional shRNAs using a regulated lentiviral expression vector, (*iii*) identification of sequences enriched by selection through massive parallel sequencing, and (*iv*) validation of shRNA targets identified by selection using synthetic siRNAs. Structure-activity analysis of multiple clones from such a library enabled us to identify the significant parameters associated with determining shRNA activity. We also demonstrate the utility and effectiveness of this strategy for identifying genes required for breast carcinoma cell growth, for which this strategy yielded genes and pathways implicated in cell growth and cancer, potential targets for anticancer drugs.

Results and Discussion

shRNA Library Construction. Our strategy for enzymatic generation of shRNA templates is schematized in Fig. 1*A* and described in *SI Methods*. DNase I digestion of target DNA is used to generate random double-stranded fragments (step 1), followed by ligation of the ends of these fragments to a single-stranded adaptor that forms a hairpin (step 2). The hairpin adaptor (HA; Fig. S1*A*) contains the loop from mir-23 miRNA and a recognition site for restriction enzyme MmeI, which cuts within the cDNA sequence 18–20 nt away from its recognition site, producing a targeting sequence of a size suitable for shRNA (step 3). The MmeI-generated fragments with 3′ NN overhangs are then ligated to a second adaptor (the termination adaptor, TA; Fig. S1*A*) (step 4), which provides an internal primer for subsequent extension (step 5). Parts of TA sequences are then removed by restriction enzyme digestion (step 6) to generate shRNA templates containing an inverted repeat followed by Pol III termination signal; the templates are then ligated into an expression vector to produce a library (step 7).

Author contributions: M.S., A.M., T.T.H., and I.B.R. designed research; M.S., A.M., Y.S., C.S.C., M.B., N.W., K.C., E.V.B., and T.T.H. performed research; M.S., A.M., Y.S., C.S.C., E.V.B., Y.D., and I.B.R. analyzed data; and M.S., A.M., Y.D., and I.B.R. wrote the paper.

The authors declare no conflict of interest.

[1]M.S. and A.M. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: roninson@ordwayresearch.org.

This article contains supporting information online at www.pnas.org/cgi/content/full/1003055107/DCSupplemental.

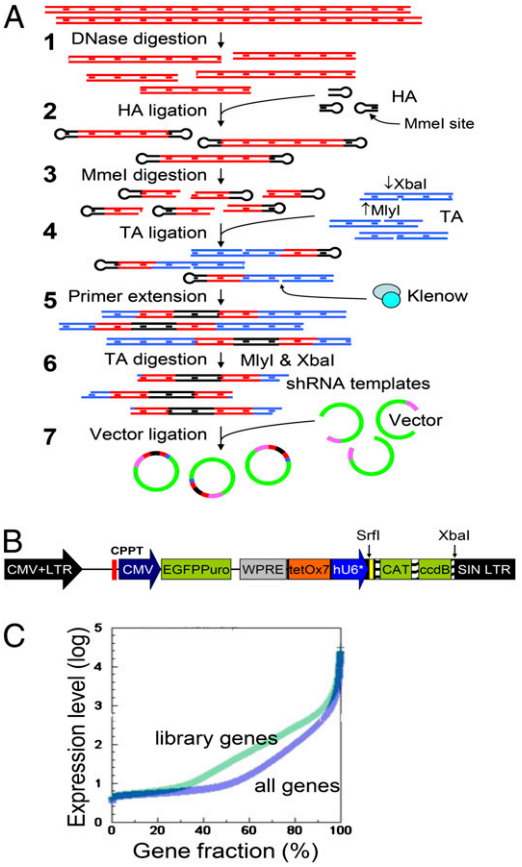


Fig. 1. shRNA library construction. (*A*) Scheme of shRNA library construction from randomly fragmented DNA. (*B*) Diagram of LLCEP TU6LX vector. CMV + LTR, LTR promoter with CMV enhancer; SIN LTR, self-inactivating LTR promoter; CPPT, central polypurine tract sequence; WPRE, woodchuck hepatitis posttranscriptional regulatory element; TetOx7, module of 7 *tet* operator repeats; CAT, chloramphenycol acetyltransferase; ccdB, cytotoxic protein (negative selection marker). (*C*) RNA expression levels in MCF7 cells, plotted in the order of increasing expression for all human UniGene entries and for genes identified in the shRNA library.

The principal differences between this strategy and earlier protocols (11, 12, 12–16) are as follows. (*i*) We use a much shorter HA, which is retained in the final shRNA library without truncation. The adaptor-derived stem sequence, joined to 19–21 bp of cDNA sequence, produces a total hairpin stem length of 27–29 bp, the size that produces more efficient knockdown (18). (*ii*) The design of the TA (Fig. S1) provides a Pol III termination signal and a 3′ (G/A)N overhang that places a purine at the +1 position relative to the promoter improving Pol III transcription (19) and includes a single-stranded nick that primes the extension with Klenow fragment (Fig. 1*A*, step 5), without the need for an external primer. (*iii*) The library is constructed in a tightly regulated lentiviral vector LLCEP TU6LX (Fig. 1*B*), an optimized version of our previously described vector LLCEP TU6X, which is regulated by tetracycline/doxycycline via tTR-KRAB repressor (20). The use of this vector for shRNA expression prevents the loss of growth-inhibitory sequences and allows for precisely controlled shRNA activity measurements.

The library construction strategy was first tested on the GL2 firefly luciferase gene, randomly fragmented with DNaseI. Approximately 90% of 700 clones sequenced from the luciferase-derived shRNA library contained proper shRNA structures comprising unique 19–21-bp sequences of the luciferase gene. To generate a human transcriptome library and to overcome the problem of uneven representation of different genes in conventional cDNA, we took advantage of the process of cDNA normalization that equalizes the abundance of different mRNA sequences (21). As the starting material for shRNA library construction, we used a library originally developed for the selection of genetic suppressor elements (GSE, short cDNA fragments encoding either antisense RNA or protein fragments acting as transdominant inhibitors). The GSE library was derived from randomly fragmented cDNA of MCF7 human breast carcinoma cells, normalized by $C_o t$ fractionation (21). The GSE library construction strategy (22) favored directional cloning of cDNA fragments flanked by different adaptors at the 5′ and 3′ ends relative to the original mRNA; sequence analysis of representative clones showed that approximately two thirds of the clones contained the adaptors in the preferred orientation. Partial directionality of the GSE library offered a possibility of creating an shRNA library where most of the shRNA sequences would have 5′-sense-loop-antisense-3′ (SA) strand orientation, which seemed desirable owing to the reports that shRNA sequences with 5′-antisense-loop-sense-3 (AS) orientation were less effective (11, 12). We replaced the 3′ adaptor of normalized cDNA fragments from the GSE library with the HA, with the subsequent steps of library construction the same as in Fig. 1*A*.

The shRNA library from normalized cDNA contained a total of 2.8×10^6 clones. Sequence analysis of 676 randomly picked clones showed that 632 of them (93.5%) contained proper stem-and-loop inserts. Of these, 500 clones (79.1%) matched with the UniGene transcript database and targeted 461 UniGene entries. Another 76 clones (12.0%) matched human genome sequences but not sequences within the UniGene database, and are likely to represent as-yet-unidentified transcripts. Of the sequenced clones, 68.2% had SA orientation and 31.8% had AS orientation, as in the starting GSE library. The length distribution of the targeted cDNA sequences was 52.4% 20 bp, 46.8% 21 bp, and 0.8% 19 bp, reflecting the heterogeneity of MmeI digestion.

To estimate the normalization and representativity of the shRNA library, we have analyzed the distribution of shRNA sequences identified by massive parallel sequencing of inserts recovered from genomic DNA of cells transduced with the library (see below). A total of 53,201 shRNAs corresponding to 14,699 UniGene entries were identified, with 72% of the entries differing in representation no more than 3-fold. Fig. 1*C* displays RNA expression [signal intensity in microarray hybridization (23)] in MCF7 cells for ≈24,000 probe sets representing all of the UniGene entries in the microarray (one probe set per entry), plotted in the order of increasing expression, as well as for probe sets representing UniGene entries identified in our library. Comparison of the two curves shows that genes expressed at the highest and the lowest levels in MCF7 cDNA have similar representation in the library, whereas the genes with intermediate levels of expression are moderately overrepresented (Fig. 1*C*). Hence, our library provides relatively uniform representation of at least ≈15,000 genes, at the average of >150 shRNAs per gene.

Activity Assays of Luciferase-Derived and cDNA-Derived shRNAs. The use of an inducible promoter for shRNA expression allows for precisely controlled measurement of RNAi activity, by comparing target expression levels in the presence and in the absence of the inducer. To generate data for shRNA structure-activity analysis, we carried out high-throughput assays of target knockdown by randomly picked clones from both luciferase- and cDNA-derived libraries, as schematized in Fig. S2*A*. The luciferase-derived clones were assayed for luciferase activity knockdown in HT1080 human fibrosarcoma cells expressing GL2 luciferase, as determined by comparing normalized luciferase activity in the presence and in the absence of doxycycline. Fig. 2*A* shows luciferase knockdown in 230 cell populations corresponding to 201 different shRNA sequences. Thirty-five percent and 11% of the clones produced >50%

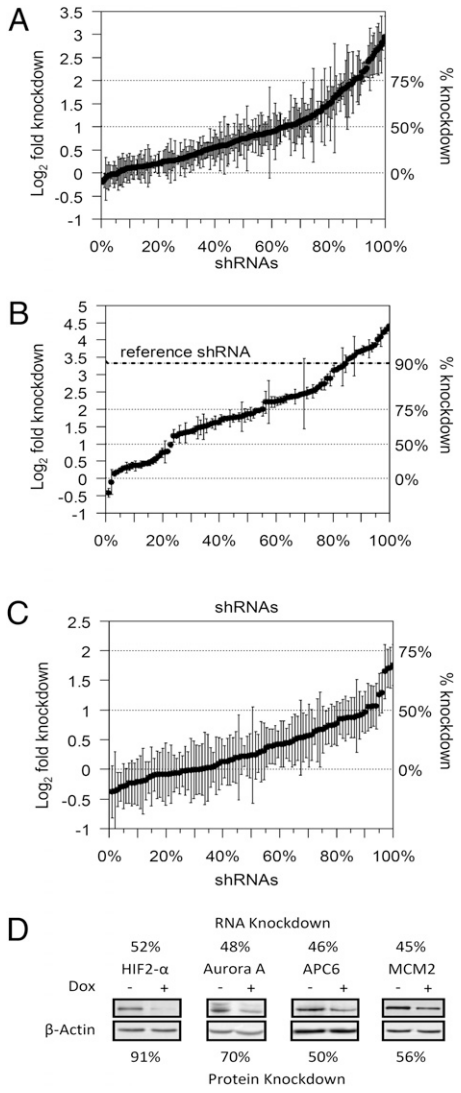


Fig. 2. shRNA activity testing. (*A*) RNAi activity of 230 luciferase-derived shRNAs in luciferase-expressing HT1080 cells, transduced and analyzed in 96-well plates. Target knockdown is measured by reduction in normalized luciferase activity in doxycycline-treated relative to untreated cells. shRNA sequences arranged in the order of increasing activity. Mean and standard deviation shown for triplicate experiments. (*B*) RNAi activity of 96 luciferase-derived shRNAs from the set in *A*, retransduced in 24-well plates. Target knockdown is measured by reduction in normalized luciferase activity in doxycycline-treated cells relative to control vector-transduced cells. Knockdown by reference luciferase shRNA (from Clontech) is marked with a dashed line. Mean and standard deviation are shown for duplicate experiments. (*C*) RNAi activity of 101 cDNA-derived shRNAs in MCF7 cells, transduced and analyzed in 96-well plates. Target knockdown is measured by reduction in RNA levels for the corresponding genes, as measured by QPCR, in doxycycline-treated relative to untreated cells. Mean and standard deviation shown for triplicate QPCR assays. (*D*) Immunoblotting analysis of the indicated proteins in MCF7 cells transduced with shRNAs targeting the corresponding genes and grown in the absence (-) or in the presence of doxycycline (+). Beta-actin was used as a loading control. RNA knockdown was determined by QPCR; protein knockdown was determined by image quantitation.

and >75% knockdown, respectively. Although none of the clones showed >90% knockdown in this assay, the knockdown efficiency was lowered in this assay by the relatively low lentiviral transduction rate in 96-well plates [low transduction leads to lower transgene expression in drug-selected populations (24)]. In the experiment shown in Fig. 2*B*, 96 clones representing a range of RNAi activities were retested by more efficient transduction in 24-well plates, along with the vector expressing a commercially available highly efficient reference shRNA for GL2 luciferase. The knockdown rates in Fig. 2*B* were calculated for each shRNA in the presence of doxycycline relative to cells transduced with insert-free vector. This analysis showed higher knockdown rates and an overall concordance with the data in Fig. 2*A* (Pearson correlation coefficient 0.81, $P = 2 \times 10^{-16}$) (Fig. S2*B*). The reference shRNA produced 90% knockdown, and this efficiency was exceeded by 15 of 96 clones generated from random fragments (Fig. 2*B*). Hence, random fragment conversion can produce shRNA inhibitors that are at least as efficient as rationally designed shRNA.

To evaluate RNAi activity of shRNA clones derived from normalized cDNA, we analyzed 101 cell populations transduced with clones corresponding to 100 different UniGene entries for a decrease in the RNA level of genes targeted by each shRNA. This analysis was performed by quantitative RT-PCR (QPCR) with gene-specific primers using RNA from cell populations grown with or without doxycycline. The distribution of RNAi activities (Fig. 2*C*) shows that 10% of the tested clones produced ≥50% knockdown by this criterion. Immunoblotting analysis for four of the protein targets of shRNAs that produced ≈50% knockdown in mRNA levels showed a decrease in the corresponding protein levels, which was in some cases much stronger than the mRNA decrease measured by QPCR (Fig. 2*D*). The lower apparent frequency of effective inhibitors among cDNA-derived shRNAs than among luciferase-derived shRNAs could reflect higher sensitivity of the artifical GL2 luciferase transcript to RNAi relative to human mRNAs. Alternatively, the difference could be due to lower sensitivity of the QPCR assay relative to the protein activity assay or to different RNAi activities between the cell lines (HT1080 and MCF7) in which these clone sets were tested. Indeed, we were able to achieve only ≈75% luciferase knockdown with the reference shRNA in MCF7 cells, as opposed to >90% knockdown in HT1080 cells. Assuming 10% frequency of active shRNAs and ≈150 shRNAs per gene, an average gene in the library should be represented by ≈15 active shRNAs.

Structure/Activity Analysis of shRNA Sequences. We investigated whether commonly used criteria for shRNA design, namely the choice of SA over AS orientation of shRNA and the Dharmacon score [a cumulative score for several criteria based on the sequence of the 19-nt siRNA guide strand (25)], show activity correlations with our datasets of 201 luciferase-derived and 101 cDNA-derived shRNA activities. The distributions of shRNA activities of both datasets according to the shRNA orientation and the Dharmacon score are shown in Fig. 3*A*. SA orientation was associated with higher activity in both datasets, but this difference was significant only for the luciferase dataset ($P = 0.0002$ for luciferase and $P = 0.1137$ for cDNA dataset, Welch's *t* test). Dharmacon score showed significant correlation with activity in both datasets [Spearman correlation 0.26 ($P = 0.0002$) for luciferase and 0.27 ($P = 0.0062$) for cDNA dataset]. Nevertheless, the majority of active shRNAs in both datasets showed relatively low Dharmacon scores (≤7), and surprisingly, many AS-oriented shRNAs were highly active (Fig. 3*A*). Therefore, we have conducted a structure-activity analysis (*SI Methods*) by including the following parameters (Table S1): the effects of the shRNA length and orientation, the levels of gene expression, the role of individual bases at each position in shRNA (Fig. S2*B*), the presence of runs of identical nucleotides, and five energy parameters based on the structures of both the processed shRNA and its mRNA target calculated using the Sfold algorithm (26). The goal of the analysis was to identify the parameters and cutoff values (filtering criteria) associated with RNAi activity.

Most of the tested parameters, including secondary structure formation by siRNA guide strand, siRNA target binding energy, average internal stability at the cleavage site, and levels of gene expression (for the cDNA-derived dataset), showed no correlations with shRNA activity. The filtering criteria that produced

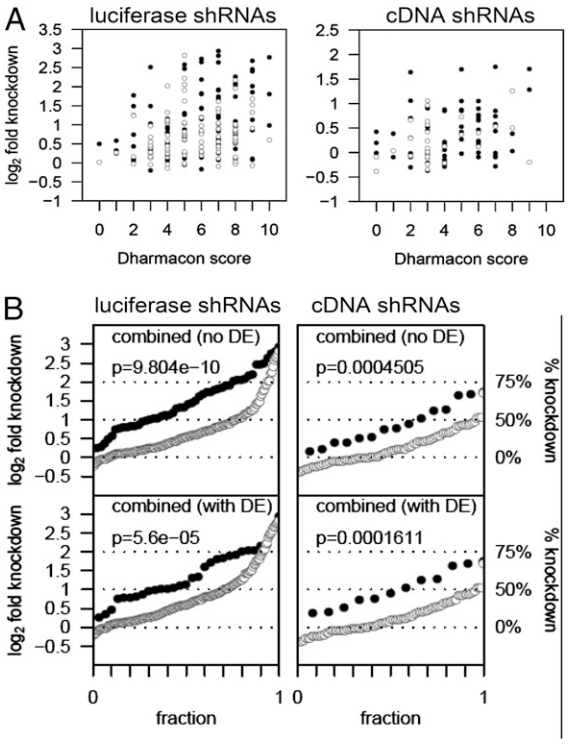


Fig. 3. shRNA structure-activity analysis. (*A*) Distribution of shRNA activities according to the Dharmacon score and shRNA orientation. Closed circles, SA shRNAs. Open circles, AS shRNAs. (*B*) Activity of shRNAs that passed (closed circles) or failed (open circles) the combination of five filtering criteria but not target disruption energy (*Top*) or all six filters combined (*Bottom*). *P* values were determined by Welch's *t* test (two sided, unequal variance). *Left:* luciferase-derived shRNA. *Right:* cDNA-derived shRNA.

significant activity discrimination in both datasets are presented in Figs. S3, S4, S5, and S6. The preferences identified in both cDNA and luciferase sets include (*i*) the SA orientation in those shRNAs that contain G as the first base; (*ii*) absence of runs of three G, three C, or four A nucleotides anywhere within the shRNA transcript, or three U nucleotides preceding the first U of the termination signal; (*iii*) GC content between 35% and 60%; (*iv*) The free-energy difference between the 5′ and the 3′ ends of the processed siRNA guide strand characterized by DSSE (differential stability in siRNA duplex ends) values of ≤−1 kcal/mol; and (*v*) the absence of C in the second position of the guide strand (Fig. S2*B*). As presented elsewhere (27), we have also found that the activity in the cDNA dataset was significantly correlated with target disruption energy, a measure of accessibility of the target mRNA. cDNA-derived shRNAs showed lower activity when their disruption energy was <−15 kcal/mol, which we have used as the sixth filtering criterion. Target disruption energy was also found to be significant in independent siRNA datasets from three human genes (27) but did not correlate with RNAi activity in the luciferase dataset, suggesting that luciferase inhibition by RNAi does not depend on the accessibility of individual sequences within this artificial target.

Fig 3*B* compares the activities of shRNAs that either passed or failed the combination of the first five filtering criteria (excluding target disruption energy) or all six filters (the comparisons for individual filters are shown in Fig. S3). After applying the first five filters, the fraction of luciferase-derived shRNAs that inhibit luciferase activity >2-fold increased from 34% in the unfiltered set to 71%. The active cDNA-derived shRNAs, defined as those that decrease target mRNA >2-fold by QPCR, increased from 10% to 40%. The addition of the target disruption energy filter produced an additional improvement in the cDNA dataset, increasing the active fraction to 50% (6 of 12) and allowing us to identify five of six most active shRNAs. Future analysis of additional clones from the cDNA-derived library should yield other significant correlations that will allow for even more rigorous selection of active shRNAs.

Identification of Genes Required for Breast Carcinoma Cell Growth Through Growth-Inhibitory shRNA Selection and Massive Parallel Sequencing. Genes required for the cell growth are expected to give rise to shRNAs that would inhibit cell proliferation. Such inhibitors can be isolated through negative selection techniques, such as BrdU suicide selection, previously used to identify growth-inhibitory GSEs (22, 28). We have now used our normalized cDNA library in the same selection (Fig. 4*A*), taking advantage of the massive parallel sequencing technology for identifying sequences enriched in selection.

The scheme of this analysis strategy is presented in Fig. 4*B*. The shRNA library was transduced into 2.5×10^7 MDA-MB-231 breast carcinoma cells expressing tTR-KRAB. 25% of the cells were used for DNA extraction, and the rest were selected for doxycycline-dependent resistance to BrdU suicide. Cells surviving the selection were used for DNA extraction. The integrated shRNA templates were amplified by PCR from genomic DNA extracted from the unselected and the BrdU-selected library-transduced cells using vector-specific primers, and the PCR product was subjected to 454 pyrosequencing. BLAST analysis of the sequence data yielded 53,201 sequences with homology to Unigene database entries before selection and 53,803 sequences after selection. These sequences matched 14,699 and 3,316 Unigene clusters respectively, indicating that selection has occurred. Sequences of 741 genes in the selected subset were enriched at least 4-fold after selection, and one of the most enriched genes was *KRAS*, an oncogene that has undergone an activating mutation in MDA-MB-231 cells (29). Table 1 shows the top 10 Kyoto Encyclopedia of Genes and Genomes pathways that were significantly enriched in this group of genes, as determined using the Pathway-Express program (30). Strikingly, 9 of the 10 pathways were associated with various cancers, cell cycle progression, or apoptosis. This analysis validates our selection system as capable of identifying oncogenes.

Of the identified genes, 119 were targeted by two to six different enriched shRNA sequences and therefore represent the most likely targets. These genes, among which *KRAS* showed the strongest enrichment, are listed in Table S2. To verify the role of such genes in cell growth, we have picked 22 of the most enriched genes represented by at least two selected shRNA sequences and 12 genes that showed no change in shRNA representation after selection. Instead of the laborious process of individually assaying specific shRNAs enriched by selection, we tested the role of candidate genes in cell growth by an independent procedure, based on transfection with synthetic siRNAs (designed by Qiagen), to determine whether such siRNAs will inhibit cell growth. Individual siRNAs were transfected into MDA-MB-231 cells, at four siRNAs per gene. A cytotoxic mixture of siRNAs derived from several essential genes was used as a positive control, and siRNAs targeting either no known genes or GFP were used as negative controls. Relative cell number was determined 6 days after siRNA transfection. As shown in Fig. 4*C*, one to four siRNAs per gene, targeting 19 of 22 tested genes (86%), inhibited cell growth relative to the negative controls (*P* values <0.05 for 17 genes and 0.053 and 0.07 for two other genes); *KRAS* targeting siRNAs showed the strongest effect. In contrast, none of the siRNAs in the control group of 12 genes with unchanged shRNA representation inhibited cell growth (Fig. 4*D*). Hence, the use of an enzymatically generated shRNA library coupled with BrdU suicide selection, massive parallel sequencing, and target verification by synthetic siRNAs, provides for efficient identification of genes required for cell growth. The analysis strategy of the present study should be generally applicable to high-throughput identification of genes involved in many different phenotypes.

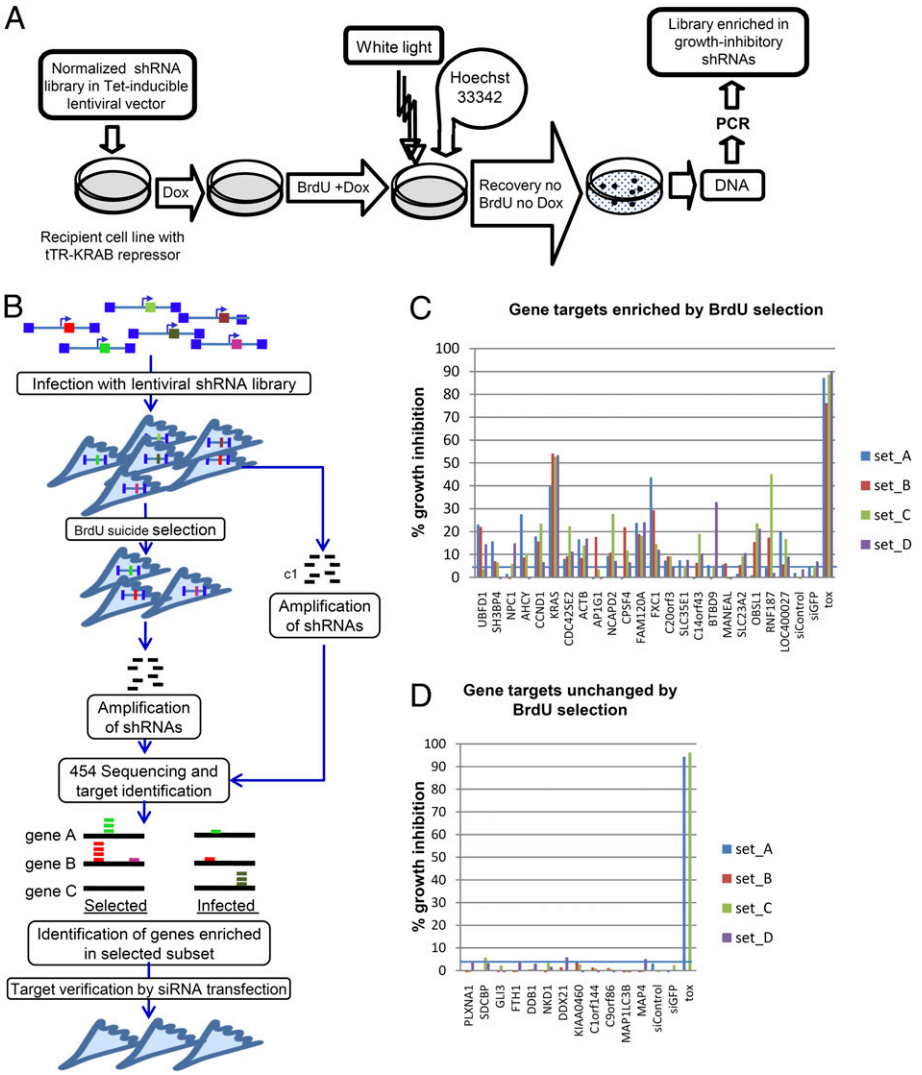


Fig. 4. Selection and validation of growth-inhibitory shRNAs. (*A*) Scheme of BrdU suicide selection for doxycycline-dependent growth inhibition. (*B*) Scheme of selection and analysis of growth-inhibitory shRNAs. (*C*) Testing 22 gene targets enriched by shRNA selection. Four synthetic siRNAs per target (sets A–D) from Qiagen were tested for the ability to inhibit MDA-MB-231 cell growth. A cytotoxic siRNA mixture (*tox*) was used as a positive control; siRNAs targeting GFP (*siGFP*) or targeting no human gene (*siControl*) were used as negative controls. (*D*) The same analysis conducted on 12 gene targets with unaltered shRNA representation after BrdU suicide selection.

Methods

All of the procedures are described in detail in *SI Methods*. The coding sequence of GL2 firefly luciferase from pGEM-luc vector (Promega) was used to generate the luciferase-based library. The starting material for the cDNA-based library was a GSE library of normalized cDNA fragments from MCF7 breast carcinoma cells, prepared as described previously (22). The lentiviral vector LLCEP TU6LX is an optimized version of LLCEP TU6X (20).

The scheme of the high-throughput activity assays of individual shRNA clones is shown in Fig. S2. Luciferase activity assays were carried out in human HT1080 fibrosarcoma cells expressing GL2 luciferase and tTR-KRAB repressor with DsRed fluorescent protein. Human gene knockdown assays were carried out in MCF7 breast carcinoma cells expressing tTR-KRAB, by QPCR. To select growth-inhibitory shRNAs, the cDNA-based library was transduced into MDA-MB-231 breast carcinoma cells expressing tTR-KRAB repressor and subjected to selection for doxycycline-dependent resistance to BrdU suicide. The shRNA sequences, amplified by PCR from the DNA of the unselected and BrdU-selected cells using vector-specific primers, were subjected to ultra-high throughput sequencing using the 454 Sequencing System. For siRNA verification assays, four siRNAs per gene (Qiagen), along with control siRNAs, were transfected into MDA-MB-231 cells in 96-well plates, in triplicates, at 5 nM of siRNA per well. Relative cell number was

Table 1. Pathways enriched among genes targeted by shRNAs enriched by BrdU suicide selection (ABC, ATP-binding cassette)

Rank	Pathway name	Impact factor	Genes in pathway (*n*)	Input genes in pathway (*n*)	Input genes in pathway (%)	Pathway genes in input (%)	*P* value
1	Prostate cancer	8.207	90	11	1.488	12.222	2.73E-04
2	Pathways in cancer	7.663	330	24	3.248	7.273	4.70E-04
3	Cell cycle	7.598	118	12	1.624	10.169	5.01E-04
4	Small cell lung cancer	5.928	86	9	1.218	10.465	0.002664
5	Pancreatic cancer	5.698	72	8	1.083	11.111	0.003354
6	Apoptosis	5.691	89	9	1.218	10.112	0.003375
7	ABC transporters	5.557	44	6	0.812	13.636	0.003862
8	Chronic myeloid leukemia	5.527	75	8	1.083	10.667	0.003977
9	Colorectal cancer	4.835	84	8	1.083	9.524	0.007948
10	Endometrial cancer	4.823	52	6	0.812	11.538	0.008042

determined 6 days after siRNA transfection by staining cellular DNA with Hoechst 33342. The Student *t* test (one-tailed, unequal variance) was used for assessing statistical difference between the inhibitory effects of tested and control siRNA.

Sequencing of plasmid DNA from randomly picked colonies was carried out on ABI 3730. Attribution of shRNA sequences was performed using National Center for Biotechnology Information BLAST and Java and Perl programs written for this analysis. Clone representation was correlated with the data from Affymetrix U133 Plus 2.0 microarray analysis of gene expression in exponentially growing MCF7 cells (23), using GeneSpring (Agilent). Different parameters used for shRNA structure-activity analysis were computed using the Sfold program (26). Plotting of the calculated parameters and nucleotides at each position in the guide strand, as well as statistical analysis, was carried out using R 2.3.1 software.

ACKNOWLEDGMENTS. We thank Dr. Yongzhi Xuan, who constructed the GSE library; George Kampo and Gregory Hurteau of Ordway Research Institute's Functional Genomics Facility for assistance with high-throughput activity assays; Drs. Pascal Bouffard and Michael Egholm for sequencing support; Dr. Inder Verma for lentiviral packaging constructs; Dr. Didier Trono for tTR-KRAB-expressing vector; and Drs. Brian Davis and Clarence Chan for helpful discussions. The Computational Molecular Biology and Statistics Core at the Wadsworth Center provided computing resources for this work. This work was supported by National Institutes of Health (NIH) Grants R33 CA95996, R01 AG028687, R01 CA62099, and R01 AG17921 (to I.B.R.), Department of Defense Grant W81XWH-08-1-0070 (to M.S.), and National Science Foundation Grant DBI-0650991 and NIH Grant R01 GM068726 (to Y.D.).

1. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev* 20:515–524.
2. Pei Y, Tuschl T (2006) On the art of identifying effective and specific siRNAs. *Nat Methods* 3:670–676.
3. Taxman DJ, et al. (2006) Criteria for effective design, construction, and gene knockdown by shRNA vectors. *BMC Biotechnol* 6:7.
4. Berns K, et al. (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 428:431–437.
5. Paddison PJ, et al. (2004) A resource for large-scale RNA-interference-based screens in mammals. *Nature* 428:427–431.
6. Silva JM, et al. (2005) Second-generation shRNA libraries covering the mouse and human genomes. *Nat Genet* 37:1281–1288.
7. Ngo VN, et al. (2006) A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* 441:106–110.
8. Moffat J, et al. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* 124:1283–1298.
9. Mullenders J, Fabius AW, Madiredjo M, Bernards R, Beijersbergen RL (2009) A large scale shRNA barcode screen identifies the circadian clock component ARNTL as putative regulator of the p53 tumor suppressor pathway. *PLoS One* 4:e4798.
10. Schlabach MR, et al. (2008) Cancer proliferation gene discovery through functional genomics. *Science* 319:620–624.
11. Sen G, Wehrman TS, Myers JW, Blau HM (2004) Restriction enzyme-generated siRNA (REGS) vectors and libraries. *Nat Genet* 36:183–189.
12. Shirane D, et al. (2004) Enzymatic production of RNAi libraries from cDNAs. *Nat Genet* 36:190–196.
13. Luo B, Heard AD, Lodish HF (2004) Small interfering RNA production by enzymatic engineering of DNA (SPEED). *Proc Natl Acad Sci USA* 101:5494–5499.
14. Dinh A, Mo YY (2005) Alternative approach to generate shRNA from cDNA. *Biotechniques* 38:629–632.
15. Du C, et al. (2006) PCR-based generation of shRNA libraries from cDNAs. *BMC Biotechnol* 6:28.
16. Fukano H, Hayatsu N, Goto R, Suzuki Y (2006) A technique to enzymatically construct libraries which express short hairpin RNA of arbitrary stem length. *Biochem Biophys Res Commun* 347:543–550.
17. Xu L, et al. (2007) Construction of equalized short hairpin RNA library from human brain cDNA. *J Biotechnol* 128:477–485.
18. Siolas D, et al. (2005) Synthetic shRNAs as potent RNAi triggers. *Nat Biotechnol* 23: 227–231.
19. Goomer RS, Kunkel GR (1992) The transcriptional start site for a human U6 small nuclear RNA gene is dictated by a compound promoter element consisting of the PSE and the TATA box. *Nucleic Acids Res* 20:4903–4912.
20. Maliyekkel A, Davis BM, Roninson IB (2006) Cell cycle arrest drastically extends the duration of gene silencing after transient expression of short hairpin RNA. *Cell Cycle* 5:2390–2395.
21. Patanjali SR, Parimoo S, Weissman SM (1991) Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci USA* 88:1943–1947.
22. Primiano T, et al. (2003) Identification of potential anticancer drug targets through the selection of growth-inhibitory genetic suppressor elements. *Cancer Cell* 4:41–53.
23. Chen Y, Dokmanovic M, Stein WD, Ardecky RJ, Roninson IB (2006) Agonist and antagonist of retinoic acid receptors cause similar changes in gene expression and induce senescence-like growth arrest in MCF-7 breast carcinoma cells. *Cancer Res* 66: 8749–8761.
24. Schott B, Iraj ES, Roninson IB (1996) Effects of infection rate and selection pressure on gene expression from an internal promoter of a double gene retroviral vector. *Somat Cell Mol Genet* 22:291–309.
25. Reynolds A, et al. (2004) Rational siRNA design for RNA interference. *Nat Biotechnol* 22:326–330.
26. Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 32 (Web Server issue):W135–41.
27. Shao Y, et al. (2007) Effect of target secondary structure on RNAi efficiency. *RNA* 13: 1631–1640.
28. Pestov DG, Lau LF (1994) Genetic selection of growth-inhibitory sequences in mammalian cells. *Proc Natl Acad Sci USA* 91:12549–12553.
29. Kozma SC, et al. (1987) The human c-Kirsten ras gene is activated by a novel mutation in codon 13 in the breast carcinoma cell line MDA-MB231. *Nucleic Acids Res* 15: 5963–5971.
30. Draghici S, et al. (2007) A systems biology approach for pathway level analysis. *Genome Res* 17:1537–1545.

Supporting Information

Shtutman et al. 10.1073/pnas.1003055107

SI Methods

Vectors. Doxycycline-regulated lentiviral vector LLCEP TU6LX (Fig. 1*B* in main text) regulated by doxycycline via tTR-KRAB repressor (1), was derived from LLCEP TU6X (2), which was modified by mutagenizing the 3′ end of human U6 promoter to create an SrfI site. The SrfI site and the downstream XbaI site were then used to clone a modified cassette comprising *CAT* gene (chloramphenicol resistance) and bacterial toxin ccdB (a negative selection marker), derived from Gateway Reading Frame Cassette A (Invitrogen). The vector was propagated in *Escherichia coli* strain DB3.1 (Invitrogen) resistant to ccdB, in the presence of ampicillin and chloramphenicol. As a positive control for luciferase-derived shRNA, we inserted the GL2 luciferase targeting shRNA from pSIREN Control Vector Set (Clontech) into LLCEP TU6X. tTR-KRAB repressor was expressed either from dsRed-expressing lentiviral vector pLV-tTR-KRAB-red (1), or from a vector termed LFC-tTR-KRAB, which was derived by replacing luciferase with tTR-KRAB in the lentiviral vector pLFC-GL2 (2) that carries a blasticidin resistance marker.

Cell Lines. HT1080 cells were cultured in DMEM with 10% FC2 serum (Invitrogen) supplemented with penicillin, streptomycin, and glutamine. 293FT cells (Invitrogen) were cultured in DMEM with 10% FC2 supplemented with penicillin, streptomycin, glutamine, and nonessential amino acids. MCF7 cells were cultured in DMEM with 10% FBS (Invitrogen) supplemented with penicillin, streptomycin, glutamine, sodium pyruvate, nonessential amino acids, and insulin. To generate recipient cell line for luciferase assays, HT1080 fibrosarcoma subline E14 (3) was transduced with retroviral vector LNCLuc (4) expressing GL2 luciferase and selected with 800 μg/mL G418. The cells were then transduced with pLV-tTR-KRAB-red and selected for DsRed fluorescence by flow sorting using FACSAria (Becton Dickinson). To generate recipient cells for regulated human gene knockdown, MCF7 and MDA-MB-231 breast carcinoma cell lines were obtained from American Type Culture Collection and then infected with LFC-tTR-KRAB and selected with 6 μg/mL blasticidin (MCF7) or infected with pLV-tTR-KRAB-red followed by two rounds of FACS selection for DsRed positive cells (MDA-MB-231). To test for tetracycline/doxycycline-dependent regulation, tTR-KRAB-transduced cells were infected with an enhanced green fluorescent protein (EGFP)-expressing tetracycline/doxycycline-inducible pLLCEm-GFP lentiviral vector. The level of activation of EGFP expression by treatment with 100 ng/mL doxycycline was approximately 150-fold.

shRNA Library Construction from GL2 Luciferase. Coding sequence of GL2 firefly luciferase was amplified by PCR from pGEM-luc vector (Promega) and digested with DNase I (Amersham/GE Healthcare) as described previously (5) to produce 50–400-bp fragments, the ends of which were then ligated to hairpin adaptor (HA) (Fig. S1*A*). HA contains MmeI restriction site and the loop of mir23 miRNA. After digestion with MmeI restriction enzyme (New England Biolabs), HA-containing fragments migrating at ≈32–34 bp were purified by electrophoresis in a 10% TBE-polyacrylamide gel, and their MmeI-generated 3′ NN overhangs were ligated to termination adaptor (TA) (Fig. S1*A*). TA contains a single-stranded nick that primes the extension with Klenow fragment (Fig. 1*A* in main text, step 5), without the need to denature the hairpin and anneal an external primer. TA also provides a Pol III termination signal and a 3′ (G/A)N overhang, which improves Pol III transcription by placing a purine at +1 position from the promoter (6). Primer extension from the primer within TA was performed with Klenow fragment of DNA polymerase I (Fermentas). Extended fragments (139–143 bp long) were purified on an 8% Tris/borate/EDTA (TBE)-polyacrylamide gel and digested with MlyI and XbaI restriction enzymes. The ≈78–80-bp digestion product was purified on an 8% TBE-polyacrylamide gel and then ligated into the LLCEP TU6LX backbone, which had been prepared by gel purification of plasmid digested with SrfI and XbaI to remove the CAT-ccdB cassette. The resulting library was transformed into ccdB-sensitive *E. coli* 10G Supreme (Lucigen), which selects for ccdB-free insert-containing clones.

We have also generated another luciferase-derived shRNA library using a modified TA that was designed to shorten the shRNA stem by removing a nucleotide at the 5′ end of the shRNA transcript and leaving an unpaired cDNA-derived nucleotide at the 3′ end of the shRNA. Subsequent analysis showed no difference in shRNA activity between the two adaptors.

Library Construction from Normalized Human cDNA Fragments. A genetic suppressor elements (GSE) library of normalized cDNA fragments from MCF7 breast carcinoma cells was prepared as described previously (7) and cloned in retroviral vector LmGCX (3). cDNA inserts with their flanking 5′ and 3′ adaptors were amplified from the GSE library by PCR using adaptor-derived primers. The primer corresponding to the 5′ adaptor was biotinylated, and the primer corresponding to the 3′ adaptor was sequence-modified to create an MmeI site at a position that allows for MmeI digestion within the cDNA sequence after random octanucleotide reverse transcription priming site. At the next step, MmeI digestion was used to remove the adaptor and the octanucleotide-derived sequence, generating a two-nucleotide NN overhang at the 3′ end. The MmeI-digested 100–500-bp fragments were gel-purified and ligated with HA, which was the same as in Fig. S1*A* except for the addition of a NN overhang at the 3′ end. The ligated material was bound to Dynabeads M-270 Streptavidin magnetic beads (Invitrogen/Dynal) and digested at the MmeI site in the HA, so that fragments containing the HA and 19–21 bp of cDNA sequences could be separated from fragments containing the 5′ adaptor, which remained bound to the streptavidin beads. The purified fragments were then used for ligation with TA and subsequent steps of shRNA template generation, as described for the luciferase-derived library.

Sequence Analysis. For sequence analysis of individual library clones, individual bacterial colonies were picked at random and grown in 2-mL cultures in 96-well deep culture blocks (Whatman). Plasmid DNA was isolated using either Wizard SV 96 Plasmid DNA Purification System (Promega) or QIAprep 96 Miniprep Kit (Qiagen) and sequenced on an ABI 3730 DNA Analyzer using a 1–5 ratio of BigDye 3.1 dGTP dye terminator and BigDye 3.1 Dye terminator mixtures. For massive parallel sequencing, PCR-amplified shRNA templates were subjected to ultra-high-throughput sequencing by 454 Life Science.

Sequence attribution of shRNA sequences was performed using National Center for Biotechnology Information (NCBI) BLAST (standard default "blastn" settings) against NCBI's human genome database located at ftp.ncbi/nih/gov/respository/Unigene/Homo_sapiens/. The first program utilizes the BioJava version 1.4 framework (http://www.biojava.org) and NCBI BLAST software to process chromatogram and sequence data generated by DNA sequencing and base calling software, and it allows exploration of sequencing and analysis results in 96-well format using a graphical user interface. The program uses BLAST to identify correctly structured library shRNA clones and to identify target mRNAs in

NCBI BLAST-formatted sequence databases, with a special emphasis on identification of targets using the NCBI UniGene cluster database. For the attribution of batch sequence data from ultra-high-throughput sequencing, the best "high-scoring segment pairs" or HSPs (i.e., those with the lowest E-value) were extracted using a custom Perl script using BioPerl modules. Only those HSPs that fell within 4 nt of the estimated start of the nonadaptor portion of the sequence were scored as legitimate hits.

To characterize the library representation of genes with different levels of expression in MCF7 cells, we have used the data from Affymetrix U133 Plus 2.0 microarray analysis of gene expression in exponentially growing MCF7 cells (8) with Microsoft Excel and GeneSpring software (Agilent). The probe sets in the Affymetrix array were matched with their corresponding UniGene ID numbers, and one probe set per UniGene entry was selected on the basis of the highest raw hybridization signal in MCF7 cells. The same selection was carried out for UniGene entries identified by sequence analysis in cDNA-derived shRNA library. The resulting lists of probe sets were then plotted against the raw hybridization data and signal significance values using GeneSpring.

shRNA Activity Analysis. After sequence analysis, plasmid clones were selected and arrayed into 96-well plates using a SciClone ALH 3000 Workstation (Caliper) and adjusted by addition of water to a fixed concentration. Ninety-six-well transfection of lentiviral vectors into 293FT packaging cells was performed using TransFectin Lipid Reagent (Bio-Rad). Two hundred nanograms of plasmid DNA for each library clone or control clone was transferred to a sterile 96-well round-bottom plate and combined with 200 ng of Δ8.91 lentiviral packaging plasmid and 66 ng VSV-G plasmid (gifts of Dr. I. Verma, Salk Institute) in serum-free DMEM. TransFectin reagent diluted in serum-free DMEM was added to allow formation of lipid complexes. After complex formation, 50,000 293FT cells in DMEM with 10% FC2 were added to each well. After overnight incubation, media were removed, and 125 μL fresh media were added. On the next day, plates were centrifuged, and 60 μL of virus-containing media were removed from each well. The media were added to the recipient cells, plated the day before at 2,500–5,000 cells per well in 96-well flat-bottom plates, and treated with 50 μM chloroquine before transduction. Transduction was repeated the following day. Two days after last transduction, transduced cells were treated with 100 ng/mL doxycycline for 2 days to induce EGFP-Puro expression, and selected with 2 μg/mL puromycin for 2 to 3 days. After selection, cells were passaged either in 96-well or in 24-well plates.

For luciferase assays, cells were plated in opaque black Nunclon Delta 96-well plates (Nalge/Nunc). Cell numbers were equalized using DsRed fluorescence as measured by FluoStar Optima multifunction plate reader (BMG Labtech). Doxycycline (100 ng/mL) or media without doxycycline were added for 2 days, then cells were washed three times with PBS and lysed in 25 μL of Passive Lysis Buffer (Promega). Firefly luciferase activity was measured using Promega Luciferase Assay System and normalized by DsRed fluorescence, measured in the same wells with the plate reader before luciferase measurement. Each plate was assayed three times on separate days.

For quantitative RT-PCR (QPCR) analysis, microarray data for MCF7 cells was examined to determine which targets were likely to be expressed at a detectable level. In addition, we selected shRNAs that target mRNAs for which QPCR primer sequences are available at PrimerBank database (http://pga.mgh.harvard.edu/primerbank/). Cells of each shRNA-transduced population were plated in two wells of 96-well plates, one of which contained 100 ng/mL doxycycline. After 3 days, poly(A)+ RNA was isolated using TurboCapture 8 mRNA strips (Qiagen) and converted into cDNA by reverse transcription using random hexanucleotide primers and SuperScript III reverse transcriptase (Invitrogen). QPCR was carried out in triplicate using an ABI 7900HT real-time PCR instrument, with primers for corresponding to each target gene selected from PrimerBank database (9). RNA levels normalized relative to β-actin were calculated using a modified $\Delta\Delta C_T$ method that took into account the amplification efficiency for each primer set, using QPCR SDS software (ABI) and the LinRegPCR program (10).

For immunoblotting analysis, MCF7 cells carrying individual shRNAs were grown for 3 days in the presence or absence of 100 ng/mL doxycycline in P100 plates. Immunoblotting analysis was carried out by standard procedures. Two-fold serial dilutions of protein extracts were used to assure the excess of the antibody. Primary antibodies against the corresponding shRNA targets were as follows: HIF2α (mouse monoclonal, Abcam ab8365), MCM2 (mouse monoclonal, Abcam ab6153); APC6 (rabbit polyclonal, Abcam ab18299); Aurora A (mouse monoclonal, EMD Biosciences, PK11). Immunoblots were developed using the corresponding HRP-conjugated secondary antibodies and ECL chemiluminescence agent. Image quantitation was carried out using the Bio-Rad VersaDoc imaging system.

Structure–Activity Analysis. We have evaluated the shRNA length and orientation, the role of individual bases at each position in shRNA, the presence of runs of identical nucleotides, and five energy parameters based on the structures of both the processed shRNA and its mRNA target (Table S1). We took into account the fact that each shRNA can be cut by Dicer into two different siRNA sequences, with either a 19-nt or a 20-nt guide strand, at comparable yields (11) (Fig. S1*B*), and carried out our calculations for both lengths. Dharmacon score, GC composition, differential stability in siRNA duplex ends (DSSE), minimum free energy (MFE) of the siRNA guide strand, siRNA binding energy, average internal stability (AIS) at the cleavage site (the average of internal stability values for positions 9–14 of the antisense strand (12)), and target disruption energy were computed with the Sfold algorithm (13). DSSE is an Sfold implementation of the rule for asymmetry of duplex ends (12, 14). MFE is the energy of the lowest-energy structure for the siRNA guide as computed by mfold (15) and is available from Sfold, and thus it measures the potential of siRNA self-folding. Dharmacon score is the sum of eight component scores for various sequence features of the siRNA duplex, with a range between −2 and 10 (16). Target disruption energy is a quantitative measure of target accessibility for the siRNA binding site and is based on mRNA secondary structures predicted by the Sfold program.

Correlations between shRNA activities and the above parameters, as well as effects of target sequence length, shRNA orientation, and individual nucleotides at each position in shRNA were plotted and analyzed using R software (http://cran.r-project.org/). Differences in shRNA activity of datasets that passed or failed filtering criteria were evaluated by statistical testing for two-group comparison. The Welch's *t* test (two tailed) is used because the sizes of our datasets are sufficiently large, and this test does not rely on the assumption of equal variance.

Library Transduction and Selection for Resistance to BrdU Suicide. The cDNA-derived shRNA library in pLLCE-TU6-LX vector was transduced into MDA-MB-231 breast carcinoma cells expressing tTR-KRAB. The infection rate (as determined by QPCR analysis of integrated provirus) was 95%. Twenty-five percent of the infected cells were subjected to DNA purification, and the rest were plated at a density of 1×10^6 cells per P150, to a total of 10^8 cells. These cells were subjected to selection for doxycycline-dependent resistance to BrdU suicide, as follows. Cells were treated with 0.1 μg/mL of doxycycline for 18 h, then with 0.1 μg/mL of doxycycline and 50 μM BrdU for 48 h. Cells were then incubated with 10 μM Hoechst 33258 for 3 h and

illuminated with fluorescent white light for 15 min on a light box, to destroy the cells that replicated their DNA and incorporated BrdU in the presence of doxycycline. Cells were then washed twice with PBS and allowed to recover in normal medium (DMEM, 10% FBS) for 7–10 days. The surviving cells were collected, followed by DNA purification. The shRNA inserts were amplified by two rounds of PCR from genomic DNA preparations of the infected cells, unselected and BrdU-selected, using the following primers: shRNA-S (AAGGAATTCAAGGTCGGGCAGGAAGAG), shRNA-AS (GGATCTAGACACGGCCGCTCTAGACAAGTC), shRNA-454-seq-A (GCCTCCCTCGCGCCATCAGTCAAGGTCGGGCAGGAAGAG), and shRNA-454-seq-B (GCCTTGCCAGCCCGCTCAGACGGCCGCTCTAGACAAGTC).

Vector-specific primers shRNA-S and shRNA-AS were used for the first round of amplification, followed by the second round of amplification with primers shRNA-454-seq-A and shRNA-454-seq-B containing overhangs compatible with 454 A and B primers. Amplified products were subjected to 454 ultra-high-throughput sequencing with 454 B primer.

Testing the Role of Target Genes in Cell Growth by siRNA Knockdown. Four siRNAs per gene, obtained from Qiagen Human Whole Genome siRNA set 1.0, were transfected into MDA-MB231 cells in 96-well plates, in triplicates, at 5 nM of siRNA per well using Silentfect transfection reagent (Biorad) and the manufacturer's (Qiagen) instructions. A cytotoxic mixture of siRNA derived from several essential genes (Qiagen, All star cell death Hs siRNA, #1027298), was used as a positive control, and siRNAs targeting no known genes (Qiagen, Negative Control siRNA #1022076) or targeting GFP (Qiagen, GFP-22 siRNA, #1022064) were used as negative controls. Cells were cultured in DMEM media with 10% FBS serum, and the relative cell number was determined 6 days after siRNA transfection by staining cellular DNA with Hoechst 33342 (Polysciences, #23491-52-3).

1. Wiznerowicz M, Trono D (2003) Conditional suppression of cellular genes: Lentivirus vector-mediated drug-inducible RNA interference. *J Virol* 77:8957–8961.
2. Maliyekkel A, Davis BM, Roninson IB (2006) Cell cycle arrest drastically extends the duration of gene silencing after transient expression of short hairpin RNA. *Cell Cycle* 5:2390–2395.
3. Kandel ES, et al. (1997) Applications of green fluorescent protein as a marker of retroviral vectors. *Somat Cell Mol Genet* 23:325–340.
4. Schott B, Iraj ES, Roninson IB (1996) Effects of infection rate and selection pressure on gene expression from an internal promoter of a double gene retroviral vector. *Somat Cell Mol Genet* 22:291–309.
5. Gudkov A, Roninson IB (1997) *Methods in Molecular Biology: cDNA library protocols*, eds Cowell IG, Austin CA (Humana Press, Totowa, NJ), pp 221–240.
6. Goomer RS, Kunkel GR (1992) The transcriptional start site for a human U6 small nuclear RNA gene is dictated by a compound promoter element consisting of the PSE and the TATA box. *Nucleic Acids Res* 20:4903–4912.
7. Primiano T, et al. (2003) Identification of potential anticancer drug targets through the selection of growth-inhibitory genetic suppressor elements. *Cancer Cell* 4:41–53.
8. Chen Y, Dokmanovic M, Stein WD, Ardecky RJ, Roninson IB (2006) Agonist and antagonist of retinoic acid receptors cause similar changes in gene expression and induce senescence-like growth arrest in MCF-7 breast carcinoma cells. *Cancer Res* 66: 8749–8761.
9. Wang X, Seed B (2003) A PCR primer bank for quantitative gene expression analysis. *Nucleic Acids Res* 31:e154.
10. Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* 339: 62–66.
11. Rose SD, et al. (2005) Functional polarity is introduced by Dicer processing of short substrate RNAs. *Nucleic Acids Res* 33:4140–4156.
12. Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* 115:209–216.
13. Ding Y, Chan CY, Lawrence CE (2004) Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* 32 (Web Server issue):W135–41.
14. Schwarz DS, et al. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell* 115:199–208.
15. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.
16. Reynolds A, et al. (2004) Rational siRNA design for RNA interference. *Nat Biotechnol* 22:326–330.

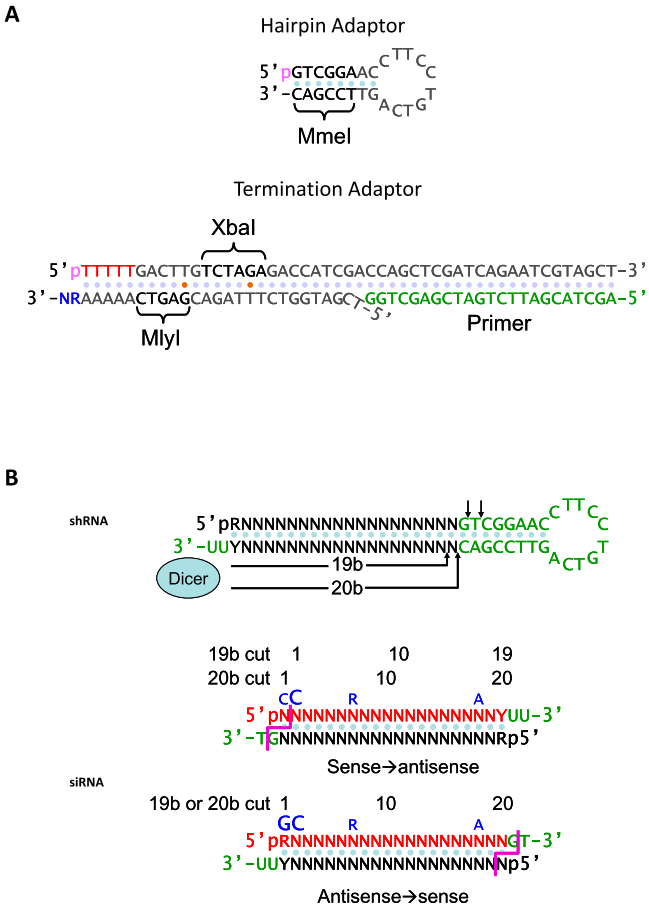


Fig. S1. (*A*) Structure of adaptors used in shRNA library construction. (*B*) Structure of library-derived shRNA and of siRNA processed from SA- and AS-oriented shRNAs. The positions of alternative Dicer cuts are marked. Red letters, guide strand; black letters, passenger strand; green letters, adaptor-derived sequences. Blue letters, nucleotides associated with reduced shRNA activity; larger letters indicate nucleotides that show a significant difference in both datasets, and smaller letters indicate nucleotides that show a significant difference only in the cDNA-derived dataset.

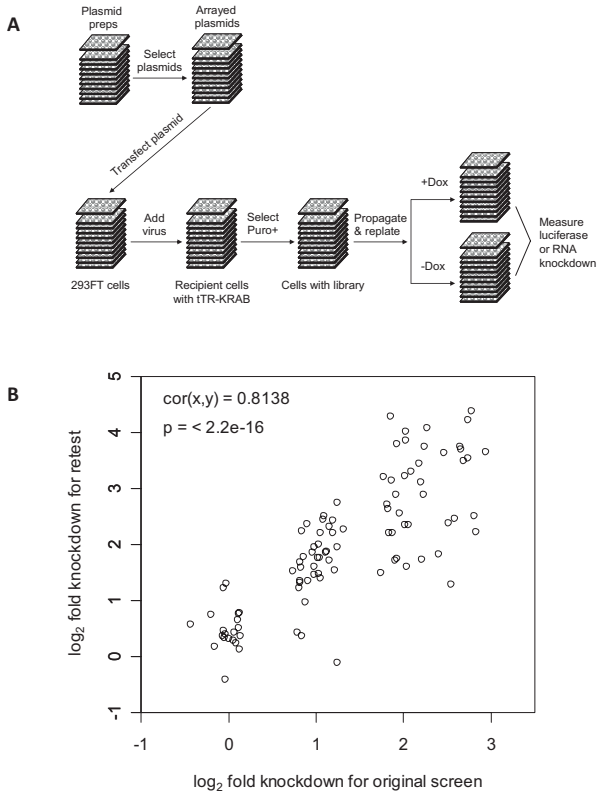


Fig. S2. (*A*) Scheme of functional assays for RNAi activity. (*B*) Comparison of shRNA activities (luciferase knockdown) for 96 luciferase-derived shRNAs, as measured in the original assays (Fig. 2*A*, main text) after transduction in 96-well plates (*x* axis) and when retested after transduction in 24-well plates (Fig. 2*B*, main text) (*y* axis). The correlation coefficient and the corresponding *P* value are shown.

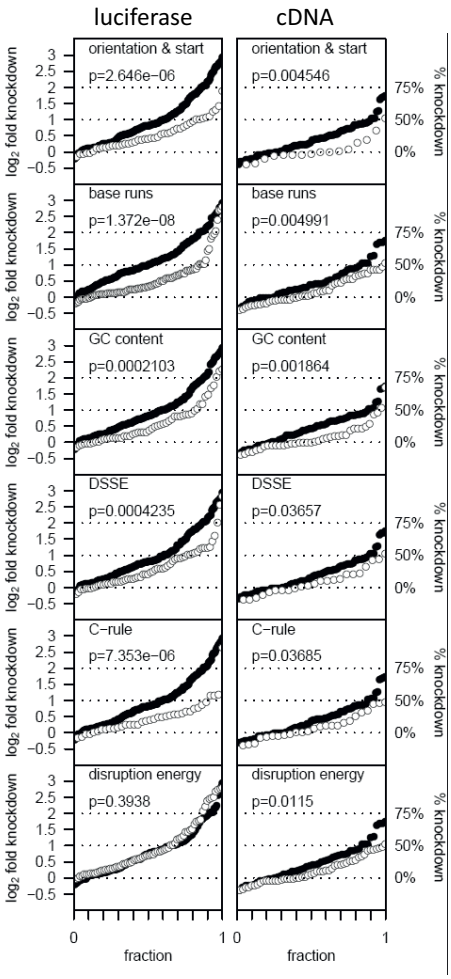
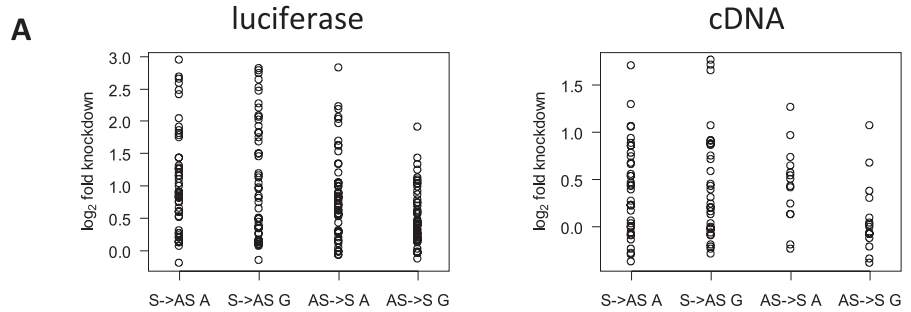


Fig. S3. Activities of shRNAs that passed (closed circles) or failed (open circles) the indicated filtering criteria, as shown in Figs. S4–S6 and described in the text. *P* values were determined by Welch's *t* test (two sided, unequal variance). *Left:* Luciferase-derived dataset. *Right:* cDNA-derived dataset.



Orientation and base comparison	p-value	
	luciferase	cDNA
AS->S G vs AS->S A	0.012482	0.017987
AS->S G vs S->AS A	0.000102	0.014523
AS->S G vs S->AS G	0.000211	0.014101
AS->S A vs S->AS A	0.106484	0.707337
S->AS G vs S->AS A	0.786669	0.777805
S->AS G vs AS->S A	0.082995	0.917521

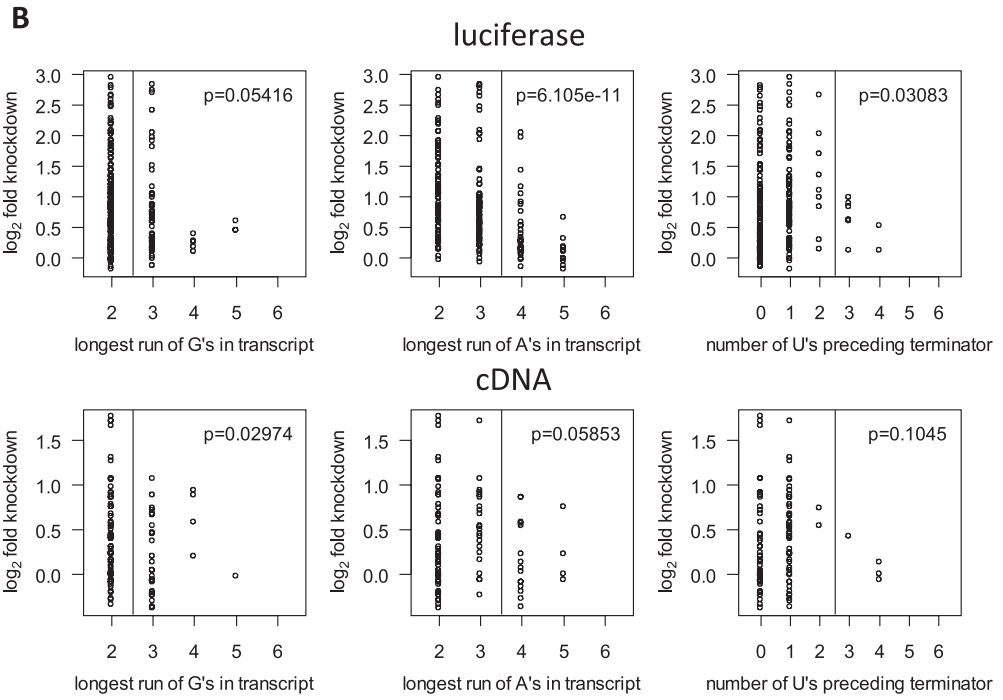


Fig. S4. (*A*) Distribution of shRNA activities according to orientation (SA, sense-strand-first; AS, antisense-strand first) and the starting nucleotide of the shRNA transcript (A or G). The table shows *P* values for all of the two-group combinations (Welch's *t* test). (*B*) Distribution of shRNA activities according to the presence of the longest runs of the indicated nucleotides in the shRNA transcripts. Vertical dotted lines show the chosen cutoffs. *P* values are shown for two-group comparisons of shRNA activities below and above the cutoff (Welch's *t* test). The distribution of the runs of Cs is identical to the corresponding distribution for Gs.

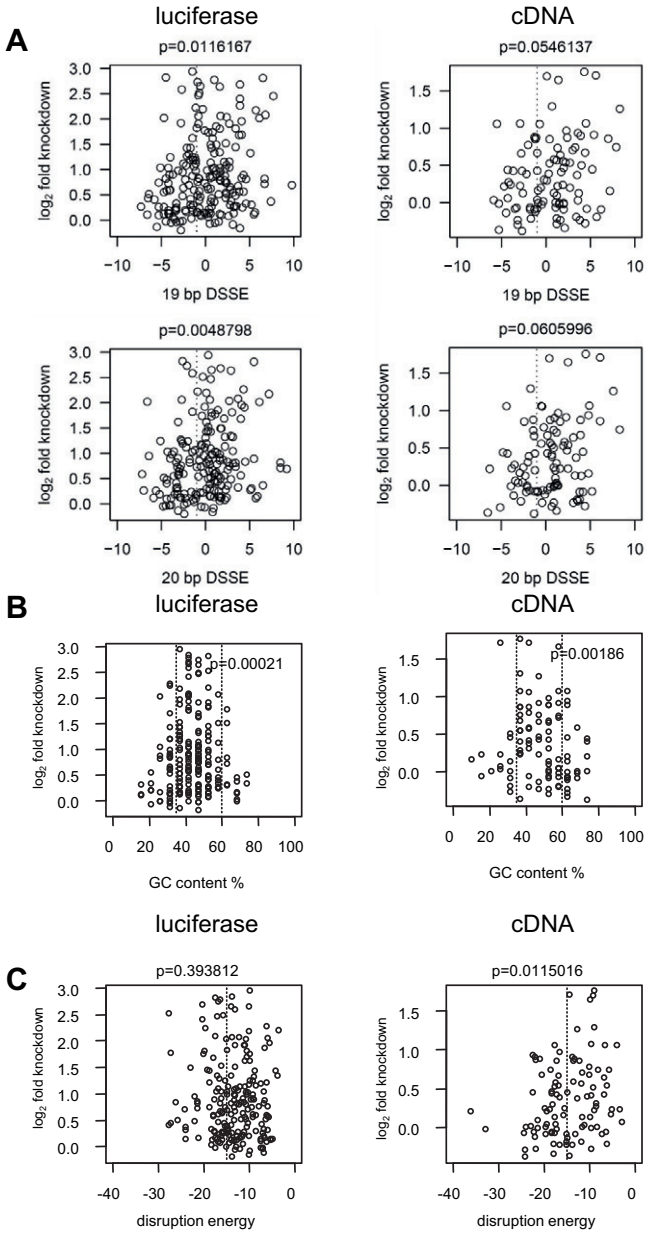


Fig. S5. (*A*) Distribution of shRNA activities according to the DSSE values, as calculated for 19-nt length (*Top*) or 20-nt length of the guide strand (*Bottom*). Vertical dotted lines show the chosen cutoffs. *P* values are shown for two-group comparisons of shRNA activities below and above the cutoff (Welch's *t* test). (*B*) Distribution of shRNA activities according to the GC content (as calculated for 19-nt length of the guide strand). Vertical dotted lines show the chosen cutoffs. *P* values are shown for two-group comparisons of shRNA activities within and outside the cutoffs (Welch's *t* test). (*C*) Distribution of shRNA activities according to target disruption energy values, as calculated for 19-nt length of the guide strand; 20-nt based calculations produced very similar results. Vertical dotted lines show the chosen cutoffs. *P* values are shown for two-group comparisons of shRNA activities below and above the cutoff (Welch's *t* test).

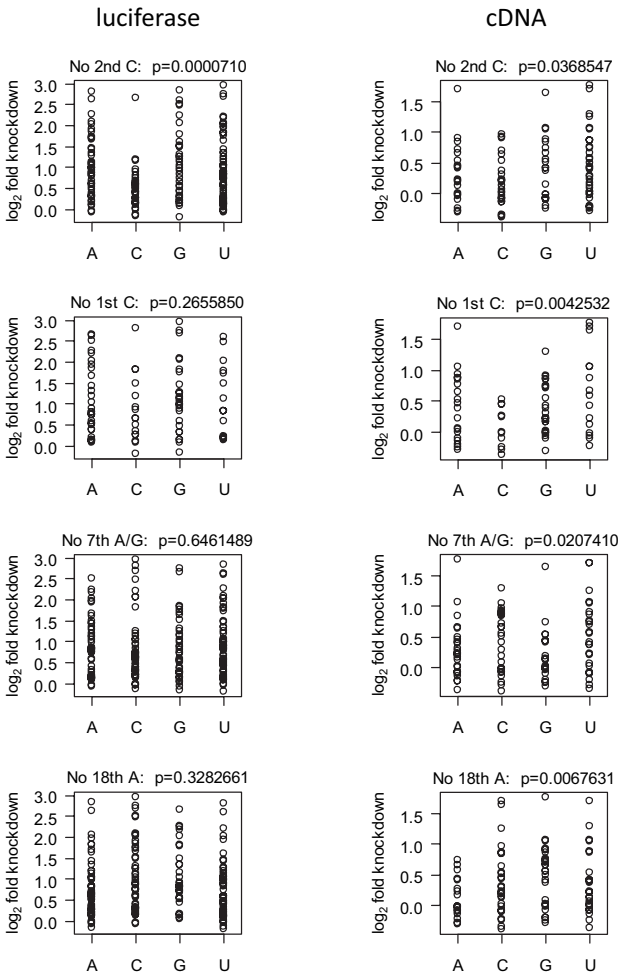


Fig. S6. Distribution of shRNA activities according to the presence of the indicated nucleotides at the indicated positions in the guide strand, as calculated for 20-nt length. The "No 1st C" distribution is shown for SA-oriented shRNAs only, because the corresponding nucleotide is absent from AS-oriented shRNAs; all of the other distributions combine SA- and AS-oriented shRNAs. *P* values are shown for two-group comparisons of shRNA activities for the indicated nucleotide (s) relative to all of the other nucleotides (Welch's *t* test).

Table S1. Correlations of energy parameters calculated on the basis of shRNA and target mRNA structures with RNAi activity

	Luciferase dataset				cDNA dataset			
	19-nt		20-nt		19-nt		20-nt	
Parameters	PCC	*P* value	PCC	*P* value	PCC	*P* value	PCC	*P* value
Binding energy	−0.09	0.2210	−0.08	0.3009	0.13	0.1911	0.15	0.1278
DSSE	0.21	0.0027*	0.20	0.0074*	0.25	0.0115*	0.27	0.0064*
AIS	−0.01	0.8792	0.00	0.9902	0.10	0.3287	0.12	0.2430
MFE	0.01	0.8468	0.04	0.5450	−0.15	0.1392	−0.15	0.1407
Target disruption energy	−0.06	0.3909	−0.03	0.7238	0.24	0.0166*	0.23	0.0193*

The calculations and interpretations of siRNA template binding energy, DSSE, AIS at the cleavage site, MFE of the siRNA guide strand, and the disruption energy of the target sequence are described in *SI Methods*. Pearson's correlation coefficients (PCC) and their corresponding *P* values are shown; an alternative nonparametric Spearman correlation analysis produced essentially the same results.
*Significant correlation ($P < 0.05$).

Table S2. Genes giving rise to at least two shRNA sequences enriched by BrdU selection in MDA-MB-231 cells

Unigene_ID	Gene name	Annotation	Selection to infection ratio	No. of different shRNA sequences	Enrichment factor
Hs#S24527772	*KRAS*	*Homo sapiens* v-K_i-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS)	43.22394	2	86.44787
Hs#S1824400	*FXC1*	Fracture callus 1 homolog (rat)	34.19699	2	68.39397
Hs#S5978514	*C20orf3*	Chromosome 20 ORF 3	11.11896	6	66.71377
Hs#S14802153	*LOC400027*	Hypothetical gene supported by BC047417	20.19062	3	60.57187
Hs#S4084609	*ICT1*	*Homo sapiens* immature colon carcinoma transcript 1 (ICT1)	28.77849	2	57.55698
Hs#S1728579	*AGPS*	Alkylglycerone phosphate synthase	25.10415	2	50.20829
Hs#S26336067	*AP1G1*	Adaptor-related protein complex 1, gamma 1 subunit	24.11579	2	48.23159
Hs#S19132849	*SLC23A2*	Solute carrier family 23 (nucleobase transporters), member 2	16.06072	3	48.18217
Hs#S17512792	*CDC42SE2*	CDC42 small effector 2	23.72045	2	47.44091
Hs#S16906232	*SH3BP4*	SH3-domain binding protein 4	15.45423	3	46.3627
Hs#S2366249	*ACTB*	*Homo sapiens* actin, beta (ACTB), mRNA	15.31946	3	45.95838
Hs#S37583284	*OBSL1*	Obscurin-like 1	11.48009	4	45.92036
Hs#S5931131	*MANEAL*	*Homo sapiens* mannosidase, endo-alpha-like (MANEAL)	22.2924	2	44.5848
Hs#S31785054	*RNF187*	Ring finger protein 187	14.82528	3	44.47585
Hs#S18152367	*C14orf43*	Chromosome 14 ORF 43	14.5429	3	43.62869
Hs#S1731541	*NPC1*	Niemann-Pick disease, type C1	14.49583	3	43.4875
Hs#S4554552	*CSNK2A1*	*Homo sapiens* casein kinase 2, alpha 1 polypeptide (CSNK2A1)	21.74375	2	43.4875
Hs#S39298991	*BTBD9*	BTB (POZ) domain containing 9	20.26122	2	40.52244
Hs#S2294357	*UBFD1*	Ubiquitin family domain containing 1	13.37364	3	40.12092
Hs#S16820105	*SLC35E1*	Solute carrier family 35, member E1	19.60232	2	39.20464
Hs#S226144	*NCAPD2*	Non-SMC condensin I complex, subunit D2	7.742092	5	38.71046
Hs#S1732011	*CPSF4*	Cleavage and polyadenylation specific factor 4, 30kDa	12.77255	3	38.31766
Hs#S1728763	*BAG1*	BCL2-associated athanogene	19.10814	2	38.21629
Hs#S15631764	*MTHFD1*	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolatesynthetase	12.70739	3	38.12216
Hs#S4262094	*RPS15A*	Ribosomal protein S15a	18.577	2	37.154
Hs#S5495022	*IQCE*	IQ motif containing E	18.28452	2	36.56903
Hs#S21286877	*PCMT1*	Protein-L-isoaspartate (D-aspartate) O-methyltransferase	17.7322	2	35.4644
Hs#S1729966	*TERF2*	Telomeric repeat binding factor 2	16.80199	2	33.60398
Hs#S2483479	*PTTG1*	Pituitary tumor-transforming 1	16.5196	2	33.0392
Hs#S4296148	*ATP5I*	*Homo sapiens* ATP synthase, H+ transporting, mitochondrial F0complex, subunit E (ATP5I), nuclear gene encoding mitochondrialprotein	16.01131	2	32.02261
Hs#S17878167	*CNNM4*	Cyclin M4	15.93718	2	31.87436
Hs#S34544410	*HSPD1*	*Homo sapiens* heat shock 60kDa protein 1 (chaperonin) (HSPD1),nuclear gene encoding mitochondrial protein	15.36888	2	30.73775
Hs#S4613863	*AHCY*	S-adenosylhomocysteine hydrolase	7.511477	4	30.04591
Hs#S1615068	*HSP90AB1*	*Homo sapiens* heat shock protein 90kDa alpha (cytosolic), classB member 1 (HSP90AB1)	14.82528	2	29.65057
Hs#S1727203	*NCBP1*	Nuclear cap binding protein subunit 1, 80kDa	14.82528	2	29.65057
Hs#S2801073	*CD9*	*Homo sapiens* CD9 molecule (CD9)	14.49583	2	28.99167
Hs#S24303272	*SSRP1*	Structure specific recognition protein 1	7.225196	4	28.90078
Hs#S3110506	*DNTTIP1*	Deoxynucleotidyltransferase, terminal, interacting protein 1	13.8987	2	27.79741
Hs#S3987156	*CCND1*	Cyclin D1	5.459469	5	27.29735
Hs#S24303058	*THRA*	Thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian)	8.98502	3	26.95506
Hs#S1728947	*EPRS*	Glutamyl-prolyl-tRNA synthetase	12.65091	2	25.30182
Hs#S2333982	*HIGD2A*	*Homo sapiens* HIG1 domain family, member 2A (HIGD2A)	12.56619	2	25.13239

Table S2. Cont.

Unigene_ID	Gene name	Annotation	Selection to infection ratio	No. of different shRNA sequences	Enrichment factor
Hs#S20091302	*APTX*	Aprataxin	12.45324	2	24.90648
Hs#S1731389	*MAPK6*	Mitogen-activated protein kinase 6	12.3544	2	24.70881
Hs#S1367624	*M-RIP*	Myosin phosphatase-Rho interacting protein	8.153906	3	24.46172
Hs#S36352307	*ZBTB38*	Zinc finger and BTB domain containing 38	6.071306	4	24.28523
Hs#S3220225	*VPS11*	Vacuolar protein sorting 11 homolog (*S. cerevisiae*)	12.07986	2	24.15972
Hs#S4300247	*TRIM29*	*Homo sapiens* tripartite motif-containing 29 (TRIM29)	7.906818	3	23.72045
Hs#S4618683	*CRIP1*	Cysteine-rich protein 1 (intestinal)	7.836221	3	23.50866
Hs#S1368546	*ABCC3*	ATP-binding cassette, subfamily C (CFTR/MRP), member 3	7.820874	3	23.46262
Hs#S2138915	*AFF4*	AF4/FMR2 family, member 4	11.67491	2	23.34982
Hs#S38872165	*ATXN7L3*	Ataxin 7-like 3	5.581283	4	22.32513
Hs#S16818069	*SPRY2*	Sprouty homolog 2 (Drosophila)	11.1462	2	22.2924
Hs#S2935335	*CAPN2*	*Homo sapiens* calpain 2, (m/II) large subunit (CAPN2)	11.1462	2	22.2924
Hs#S19132577	*PLEKHG2*	Pleckstrin homology domain containing, family G (with RhoGef domain) member 2	11.0366	2	22.0732
Hs#S26153400	*TNPO2*	Transportin 2 (importin 3, karyopherin beta 2b)	7.298601	3	21.8958
Hs#S1263959	*FZD1*	Frizzled homolog 1 (Drosophila)	10.87187	2	21.74375
Hs#S23881883	*BCL2L1*	BCL2-like 1	7.238227	3	21.71468
Hs#S21106926	*FAM120A*	Family with sequence similarity 120A	4.28286	5	21.4143
Hs#S24273099	*FXR1*	Fragile X mental retardation, autosomal homolog 1	10.31956	2	20.63912
Hs#S32813027	*ASB1*	Ankyrin repeat and SOCS box-containing 1	9.703822	2	19.40764
Hs#S1731625	*CTPS*	CTP synthase	9.663888	2	19.32778
Hs#S3619291	*LMNB2*	Lamin B2	4.726902	4	18.90761
Hs#S2654982	*DUSP14*	Dual specificity phosphatase 14	9.356401	2	18.7128
Hs#S14273019	*TAF15*	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa	5.963058	3	17.88918
Hs#S15116077	*CDA*	Cytidine deaminase	5.930113	3	17.79034
Hs#S34547277	*ARPC3*	*Homo sapiens* actin related protein 2/3 complex, subunit 3, 21kDa (ARPC3), mRNA.	8.89517	2	17.79034
Hs#S2138748	*REPIN1*	Replication initiator 1	5.831278	3	17.49383
Hs#S5978605	*KIAA0664*	*Homo sapiens* TSC22 domain family, member 2, mRNA	5.637268	3	16.9118
Hs#S18928610	*PTRF*	Polymerase I and transcript release factor	5.633608	3	16.90082
Hs#S16056802	*POLR2J2*	*Homo sapiens* DNA directed RNA polymerase II polypeptide J-related(POLR2J2), mRNA	8.445919	2	16.89184
Hs#S4074937	*PFDN5*	Prefoldin subunit 5	8.445919	2	16.89184
Hs#S2653589	*YPEL5*	Yippee-like 5 (Drosophila)	8.400994	2	16.80199
Hs#S2293860	*C20orf11*	Chromosome 20 ORF 11	8.400994	2	16.80199
Hs#S4616288	*AES*	Amino-terminal enhancer of split	8.400994	2	16.80199
Hs#S4618103	*PCYT2*	Phosphate cytidylyltransferase 2, ethanolamine	8.346085	2	16.69217
Hs#S38795021	*USP19*	Ubiquitin specific peptidase 19	5.435937	3	16.30781
Hs#S16391281	*NF2*	Neurofibromin 2 (bilateral acoustic neuroma)	8.104488	2	16.20898
Hs#S4833432	*LOC51035*	SAPK substrate protein 1	8.071543	2	16.14309
Hs#S15640600	*RBM20*	RNA binding motif protein 20	7.906818	2	15.81364
Hs#S16819411	*LOC283970*	probably fused seq	7.906818	2	15.81364
Hs#S17873428	*FDPS*	*Homo sapiens* farnesyl diphosphate synthase (farnesyl pyrophosphate synthetase, dimethylallyltranstransferase, geranyltranstransferase) (FDPS)	7.906818	2	15.81364
Hs#S4026358	*CFL1*	Cofilin 1 (nonmuscle)	5.188849	3	15.56655
Hs#S21310805	*KRT13*	Keratin 13	7.542688	2	15.08538
Hs#S3438578	*C13orf23*	Chromosome 13 ORF 23	7.511477	2	15.02295
Hs#S4029085	*EIF4B*	EIF4B eukaryotic translation initiation factor 4B	7.46755	2	14.9351
Hs#S16056656	*USP37*	Ubiquitin specific peptidase 37	7.430801	2	14.8616
Hs#S3307627	*0*	Target clone is not clearly identified (homology with himeric products)	4.889743	3	14.66923
Hs#S1824507	*PRDM4*	PR domain containing 4	7.116136	2	14.23227
Hs#S1726632	*DDX3X*	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	7.072896	2	14.14579
Hs#S16819337	*TH1L*	TH1-like (Drosophila)	4.67221	3	14.01663
Hs#S3219661	*UBE2O*	Ubiquitin-conjugating enzyme E2O	6.918466	2	13.83693

Table S2. Cont.

Unigene_ID	Gene name	Annotation	Selection to infection ratio	No. of different shRNA sequences	Enrichment factor
Hs#S1970972	*ANKMY2*	Ankyrin repeat and MYND domain containing 2	6.918466	2	13.83693
Hs#S19656836	*MAPKAPK3*	Mitogen-activated protein kinase-activated protein kinase 3	6.808649	2	13.6173
Hs#S1727033	*ITGB1*	*Homo sapiens* integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) (ITGB1)	6.766411	2	13.53282
Hs#S3219600	*ST14*	Suppression of tumorigenicity 14 (colon carcinoma)	6.720795	2	13.44159
Hs#S3940065	*LARP1*	La ribonucleoprotein domain family, member 1	6.606354	2	13.21271
Hs#S17083357	*RNF34*	Ring finger protein 34	6.424289	2	12.84858
Hs#S1177138	*NF1*	*Homo sapiens* neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease) (NF1)	6.424289	2	12.84858
Hs#S1730187	*WDR57*	WD repeat domain 57 (U5 snRNP specific)	6.149747	2	12.29949
Hs#S15967295	*RPL23*	*Homo sapiens* ribosomal protein L23 (RPL23)	6.071306	2	12.14261
Hs#S34542802	*ARHGDIA*	*Homo sapiens* Rho GDP dissociation inhibitor (GDI) alpha (ARHGDIA)	5.930113	2	11.86023
Hs#S3619207	*SLC35B4*	Solute carrier family 35, member B4	5.930113	2	11.86023
Hs#S3993862	*GIT2*	G protein-coupled receptor kinase interactor 2	5.930113	2	11.86023
Hs#S3782070	*SLC38A5*	Solute carrier family 38, member 5	5.930113	2	11.86023
Hs#S16885581	*TOLLIP*	Toll interacting protein	5.683025	2	11.36605
Hs#S875916	*ADAR*	Adenosine deaminase, RNA-specific	5.660563	2	11.32113
Hs#S15515243	*SHKBP1*	*Homo sapiens* SH3KBP1 binding protein 1 (SHKBP1)	5.600663	2	11.20133
Hs#S34542796	*MAFG*	*Homo sapiens* v-maf musculoaponeurotic fibrosarcoma oncogene homologG (avian) (MAFG)	5.600663	2	11.20133
Hs#S1729506	*OXSR1*	Oxidative-stress responsive 1	5.559481	2	11.11896
Hs#S1728450	*IDUA*	*Homo sapiens* iduronidase, alpha-L- (IDUA),	5.435937	2	10.87187
Hs#S4284382	*TMEM43*	Transmembrane protein 43	5.051578	2	10.10316
Hs#S1732446	*POLS*	Polymerase (DNA directed) sigma	4.941761	2	9.883522
Hs#S3508192	*PSG4*	*Homo sapiens* pregnancy specific beta-1-glycoprotein 4 (PSG4)	4.941761	2	9.883522
Hs#S542952	*BRI3*	*Homo sapiens* brain protein I3 (BRI3)	4.941761	2	9.883522
Hs#S3782069	*VPS13A*	Vacuolar protein sorting 13 homolog A (*S. cerevisiae*)	4.941761	2	9.883522
Hs#S24303175	*MVD*	Mevalonate (diphospho) decarboxylase	4.941761	2	9.883522
Hs#S1727134	*MADD*	MAP-kinase activating death domain	4.851911	2	9.703822
Hs#S1729605	*CBL*	Cas-Br-M (murine) ecotropic retroviral transforming sequence	4.61231	2	9.224621
Hs#S1731172	*IQGAP1*	IQ motif containing GTPase activating protein 1	4.534793	2	9.069585
Hs#S21504143	*TRMT5*	TRM5 tRNA methyltransferase 5 homolog (*S. cerevisiae*)	4.447585	2	8.89517
Hs#S34541381	*KRT8*	*Homo sapiens* keratin 8 (KRT8), mRNA	4.447585	2	8.89517
Hs#S24303340	*EEF1A2*	Eukaryotic translation elongation factor 1 alpha 2	4.34875	2	8.6975
Hs#S2651797	*TREML2*	Triggering receptor expressed on myeloid cells-like 2	4.324041	2	8.648082
Hs#S15644477	*TUSC4*	Tumor suppressor candidate 4	4.28286	2	8.565719
Hs#S2140695	*MRPL48*	Mitochondrial ribosomal protein L48	4.28286	2	8.565719
Hs#S2653639	*SNX12*	Sorting nexin 12	4.28286	2	8.565719
Hs#S24639346	*PKN1*	Protein kinase N1	4.151079	2	8.302159
Hs#S4807559	*SDK1*	Sidekick homolog 1 (chicken)	4.105463	2	8.210926
Hs#S1729352	*USP11*	Ubiquitin specific peptidase 11	4.094602	2	8.189204
Hs#S17668848	*PDE8A*	Phosphodiesterase 8A	4.063226	2	8.126452

"Selection to infection ratio" is the number of sequence reads for the corresponding gene in the samples from BrdU-selected cells relative to unselected lentiviral library-infected cells. The "enrichment factor" is the "selection to infection ratio" multiplied by the number of different shRNA sequences for a given gene found in the BrdU-selected sample.