(54) Title: SYSTEMS AND METHODS FOR SELECTION AND DESIGN OF SHORT INTERFERING RNA

(57) Abstract: The present invention provides systems and methods for selecting and designing siRNAs to effectively inhibit or reduce levels and/or expression of target transcripts. In particular, the invention provides computer based systems and methods for implementing rules useful in selecting and designing siRNAs. The invention provides siRNAs selected and/or designed in accordance with the inventive methods. The invention further provides methods for systematically testing siRNAs, thereby permitting identification of hypersensitive sites and preferred siRNAs.
Published:
— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.
SYSTEMS AND METHODS FOR SELECTION AND DESIGN OF SHORT INTERFERING RNA

Priority Claim

[0001] The present application claims priority to USSN 60/415,235 filed October 1, 2002, the entire contents of which are incorporated herein by reference.

Government Support

[0002] The United States Government has provided grant support utilized in the development of the present invention. In particular, National Cancer Institute contract number P01-CA42063, National Institutes of Health contract numbers R37-GM34277, R01-AI32486, and R21-AI45306 have supported development of this invention. The United States Government may have certain rights in the invention.

Background of the Invention

[0003] Since the earliest days of genetics there has been tremendous interest in determining the function of individual genes. Classical studies of gene and protein function relied on the isolation of mutants and the performance of genetic screens designed to enrich for such mutants. Although such forward genetic approaches are appealing in that the phenotype of the mutant frequently sheds light on the function of the mutated gene, they suffer from a number of limitations. For example, after identifying a mutant using classical genetic approaches further steps are necessary to identify the mutation and thus the gene responsible for the phenotype. In addition, many gene functions are likely to be redundant, and many phenotypes are multifactorial, involving the interplay of numerous genes. These drawbacks, among others, have become more acute with the advent of large scale cDNA and genomic sequencing.

[0004] Reverse genetic approaches aim to allow the researcher to determine the function of a gene and/or its encoded product(s) given only a complete or partial sequence, a goal that is frequently unattainable using the techniques described above. Reverse genetic approaches include the use of homologous recombination, in which a
gene of interest is specifically deleted or mutated within a cell, organism, etc. Although methods for performing homologous recombination are well developed, the technique can be time-consuming and costly and is thus not well suited for studies comparing gene function among a large number of different cell types, or under a variety of different experimental conditions, etc. Antisense technology is another reverse genetic approach that has been extensively explored as a means of reducing or eliminating expression of a gene of interest, but this technique also suffers from a number of limitations. For example single-stranded RNA can be extremely susceptible to degradation by any of a the wide variety of ribonucleases either before or after introduction into a cell or organism. Large amounts of antisense RNA must frequently be delivered. Selection of effective antisense sequences represents another challenge.

[0005] The discovery in *C. elegans* of a new technique involving specific gene silencing by double-stranded RNA, referred to as RNA interference (RNAi), was followed by the realization that similar processes operate in a wide variety of eukaryotes including both animals and plants. Accordingly, there has been interest in utilizing RNAi for the purpose of reducing or inhibiting expression of genes in various cell types and organisms. At this point in time there remains a need for improved understanding of the phenomenon of RNAi and of how to harness it, e.g., for research and/or therapeutic purposes. The present invention provides systems and methods for selecting RNA sequences useful in mediating RNAi, thereby addressing some of these needs.

**Summary of the Invention**

[0006] The present invention provides methods and accompanying computer-based systems and computer-executable code stored on a computer-readable medium for selecting and designing preferred siRNA sequences for mediating RNA interference. In certain embodiments of the invention the siRNA comprises two RNA strands having a region of complementarity approximately 19 nucleotides in length and optionally further comprises one or two single-stranded overhangs or loops. In certain embodiments of the invention the siRNA comprises a single RNA strand having a region of self-complementarity. The single RNA strand may form a hairpin
structure with a stem and loop and, optionally, one or more unpaired portions at the 5' and/or 3' portion of the RNA.

[0007] In one aspect, the invention provides a method for selecting an siRNA targeted to a target transcript comprising (i) applying a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3' overhang; and (ii) applying a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs. The Target Portion Selection Rules may select target portions according to a tiling approach or a constrained overhang approach. In various embodiments of the invention the Complexity Rules include Composition Rules and/or Cluster Avoidance Rules. The method may further include applying one or more Suboptimal Element Positioning Rules, base pair optimization rules, Overhang Refinement Rules, Global Positioning Rules, or Specificity Rules.

[0008] In another aspect, the present invention provides a computer system for selecting an siRNA targeted to a target transcript, the computer system comprising memory means which stores a program comprising computer-executable process steps and a processor which executes the process steps so as (i) to apply a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3' overhang; and (ii) to apply a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs. According to certain embodiments of the invention the computer system receives input from a user specifying a target transcript and provides preferred sequences to the user. In certain embodiments of the invention interaction with the user occurs via the Internet using World Wide Web pages. Certain embodiments of the invention provide the user with various options that can be used to choose rules and parameters to guide the selection process. In certain embodiments of the invention the computer system includes an
online ordering module that allows the user to order a preferred siRNA. According to
the invention the siRNA is then shipped to the user.

[0009] In another aspect, the invention provides computer-executable process
steps stored on a computer-readable medium, the computer-executable process steps
to select an siRNA targeted to a target transcript, the computer-executable process
steps comprising: code to apply a target portion selection rule to select portions of the
target transcript, thereby identifying a set of core regions, wherein each core region
corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and
wherein each duplex portion of a candidate siRNA thus specified comprises a sense
strand and an antisense strand, each of which optionally includes a 3’ overhang; and
code to apply a plurality of Complexity Rules to the candidate siRNAs, thereby
selecting preferred siRNAs. According to certain embodiments of the invention the
computer-executable process steps include code to receive input from a user
specifying a target transcript and code to provide preferred sequences to the user. In
certain embodiments of the invention the computer-executable process steps include
code that allows the user to order a preferred siRNA online. The invention thus
includes a method for providing an siRNA comprising steps of: receiving information
identifying a target transcript from a user; selecting one or more preferred siRNA
targeted to the transcript by applying a set of siRNA selection rules to the target
portion, thereby selecting one or more preferred siRNAs; and providing at least one
siRNA to the user.

[0010] In another aspect, the invention provides methods for identifying an
siRNA hypersensitive site on a target transcript by systematically testing siRNAs.
According to certain embodiments of the invention the method comprises steps of
providing at least ten siRNAs, wherein each siRNA includes an antisense strand
having a sequence that is perfectly complementary to a target portion of the target
transcript, wherein the siRNAs span at least fifty percent of the target transcript;
delivering each siRNA to a different population of cells, which cells each contain a
level of the target transcript; determining to what extent each siRNA reduces the level
of the target transcript or reduces expression of the target transcript in the population
of cells to which it was delivered; and identifying a site on the target transcript as a
hypersensitive site if an siRNA whose sense strand sequence either includes the site’s
sequence or is included by the site's sequence reduces the level of the target transcript or reduces expression of the target transcript to a greater extent than the other siRNAs. Those of ordinary skill in the art will readily appreciate that different cells may contain different levels of the target transcript; so long as a detectable change in level occurs, the inventive methods and systems apply. The invention further comprises hypersensitive sites identified according to the inventive methods, siRNAs targeted to such sites, and methods of treating disease by delivering siRNAs targeted to hypersensitive sites to a subject.

[0011] In another aspect, the invention provides methods of identifying a preferred siRNA to inhibit a target transcript comprising steps of: providing at least ten siRNAs, wherein each siRNA includes an antisense strand having a sequence that is perfectly complementary to a target portion of the target transcript, and wherein the siRNAs span at least fifty percent of the target transcript; delivering each siRNA to a different population of cells, which cells each contain a level of the target transcript; determining to what extent each siRNA reduces the level of the target transcript or expression of the target transcript in the population of cells to which it was delivered; and identifying an siRNA that reduces the level of the target transcript or reduces expression of the target transcript to a greater extent than the other siRNAs as a preferred siRNA. The invention also encompasses methods of treating a disease or clinical condition comprising administering preferred siRNAs identified according to the inventive methods. It is noted that although for purposes of description the present application refers to siRNA, this term is used herein to refer to RNA structures that encompass siRNA precursors such as shRNA. In addition, it will be evident that the methods for selecting duplex portions of siRNA are useful in the design of vectors for use in effecting intracellular synthesis of siRNA and shRNA.


5 Brief Description of the Drawing

[0013] Figure 1 shows the structure of siRNAs observed in the Drosophila system.
[0014] Figure 2 presents a schematic representation of the steps involved in RNA interference in Drosophila.
[0015] Figure 3 shows a variety of exemplary siRNA structures useful in accordance with the present invention.
[0016] Figure 4 shows an siRNA structure comprising a stem and loop.
[0017] Figure 5 shows an siRNA structure comprising a stem and two loops.
[0018] Figure 6 presents a representation of an alternative inhibitory pathway, in which the DICER enzyme cleaves a substrate having a base mismatch in the stem to generate an inhibitory product that binds to the 3' UTR of a target transcript and inhibits translation.
[0019] Figure 7 presents one example of a construct that may be used to direct transcription of both strands of an inventive siRNA.
[0020] Figure 8 depicts one example of a construct that may be used to direct transcript of a single-stranded siRNA according to the present invention.
[0021] Figure 9 shows the results of experiments indicating that CD4-siRNA selected in accordance with the methods described herein inhibits HIV entry and infection in Magi-CCR5 cells. Panel A shows flow cytometric analysis of CD4 expression (CD4-PE) 60 hours after Magi-CCR5 cells were either mock transfected or transfected with CD4-siRNA, antisense strand of CD4-siRNA only (CD4-asRNA) or HPRT-siRNA (control siRNA). Cell numbers in each panel represent the percent of gated CD4 positive cells. Panel B shows a Northern blot for CD4 expression in control (CD4-negative) HeLa cells (lane 1), mock (lane 2), CD4-siRNA (lane 3, CD4-asRNA (lane 4) and control siRNA (lane 5) transfected cells. ß-actin expression was used as a loading control. Panel C shows ß-gal expression in CD4-siRNA (lane 1),
CD4-asRNA (lane 2) and control siRNA (lane 3) transfected cells, 2 days after infection with HIV-1 NL43 (left) or BAL (right). A reduction in the number of β-gal positive cells in CD4-siRNA transfected cells compared with control siRNA transfected cells indicates decreased transactivation of endogenous LTR-β-gal expression by HIV-1 Tat. Error bars are the average of 2 experiments. Panel D shows a photomicrograph of β-gal stained Magi-CCR5 cells either uninfected or infected with HIV-1 NL43 after mock, CD4-siRNA, CD4-asRNA, or control siRNA transfection. Syncytia formation and LTR activation are reduced in the CD4-siRNA transfected cells compared to controls. Panel E presents levels of viral p24 antigen of cell free HIV production from the samples described in C as measured by ELISA 2 days after transfected Magi-CCR5 cells were infected with HIV-1 strains NL43 (left) or BAL (right). Error bars are the average of 2 experiments. Panel F shows alternate washes of the Northern blot shown in Panel B. The upper portion of the panel shows a lower stringency wash used for quantification of transcription after gene silencing.

The middle panel is a higher stringency wash of the same blot used to demonstrate that the smudge near the CD4 silenced lane was non-specific.

[0022] Figure 10 presents results of experiments demonstrating that p24-siRNA selected in accordance with the methods described herein inhibits viral replication in HeLa-CD4 cells. Panel A shows flow cytometric analysis of p24-siRNA-directed inhibition of viral gene expression (p24RD1) in uninfected, control and mock-, p24-siRNA-, p24-siRNA-antisense strand- and GFP-siRNA (control siRNA) transfected HeLa-CD4 cells 2 d after infection with HIVIIIb, demonstrating specificity of the inhibitory effect. Panel B shows a Northern blot for p24 expression in uninfected (lane 1), mock (lane 2), p24-siRNA (lane 3), p24-siRNA-antisense strand (lane 4), and control siRNA (lane 5) transfected cells. β-actin expression was used as a loading control. Panel C shows flow cytometric analysis of p24 expression (p24RD1) in uninfected control and mock, p24-siRNA and GFP-siRNA (control siRNA) transfected HeLa-CD4 cells 5 days post infection with HIVIIIb. Cell numbers in each panel represent the percent of gated p24 cells. Panel D is a Northern blot for p24, Nef and β-actin expression in stably infected control (lane 1), uninfected (lane 2), mock (lane 3), p24-siRNA (lane 4), and control siRNA (lane 5) transfected cells. Compared to mock or control siRNA transfected cells, p24-siRNA transfected cells showed
decreased expression of the full length, 9.2 Kb HIV transcripts and/or genomic RNA as well as the 4.3 and 2.0 Kb Nef-containing transcripts. β-actin expression was used as a loading control. Panel E gives levels of viral p24 antigen measured by ELISA in uninfected control (lane 1) and mock (lane 2), p24-siRNA (lane 3) and control siRNA (lane 4) transfected cells infected with HIVIII and demonstrates that reduction of cell free virus production only in p24-siRNA transfected HeLa-CD4 cells. Error bars represent the average of three experiments.

Figure 11 demonstrates siRNA-directed knockdown of viral gene expression in HeLa-CD4 cells within established HIV infection using siRNAs designed in accordance with the methods described herein. Four days after infection with HIVIII, HeLa-CD4 cells were either mock transfected or transfected with p24-siRNA or GFP-siRNA (control siRNA) and analyzed 2 days later for p24 expression (p24-RD1) by flow cytometry. Overlay histogram depicts the uninfected control shown in panel 1. Cell numbers in each panel depicts mean fluorescent intensity of the cells expressing p24.

Figure 12 presents a schematic of part of an mRNA target transcript including a target portion and a corresponding siRNA targeted to the target transcript in accordance with the present invention.

Figure 13 presents a complete cDNA sequence corresponding to a target transcript (the vaccinia virus genome, Genbank accession number NC_001559) with candidate siRNAs indicated in boxes and preferred siRNAs selected in accordance with the invention indicated with asterisks. The file was created using the DNA Strider™ program.

The upper strand of the cDNA is the sense strand, i.e., the strand identical to the corresponding mRNA.

Figure 14 depicts microRNAs hybridized to a target site.

Figure 15 depicts a representative embodiment of a computer system of the present invention.

Definitions

The term hybridize, as used herein, refers to the interaction between two complementary nucleic acid sequences. The phrase hybridizes under high stringency conditions describes an interaction that is sufficiently stable that it is maintained under
art-recognized high stringency conditions. Guidance for performing hybridization reactions can be found, for example, in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y., 6.3.1-6.3.6, 1989, and more recent updated editions, all of which are incorporated by reference. See also Sambrook, Russell, and Sambrook, *Molecular Cloning: A Laboratory Manual*, 3rd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2001. Aqueous and nonaqueous methods are described in that reference and either can be used. Typically, for nucleic acid sequences over approximately 50-100 nucleotides in length, various levels of stringency are defined, such as low stringency (e.g., 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by two washes in 0.2X SSC, 0.1% SDS at least at 50°C (the temperature of the washes can be increased to 55°C for medium-low stringency conditions)); medium stringency (e.g., 6X SSC at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 60°C; high stringency hybridization (e.g., 6X SSC at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 65°C; and very high stringency hybridization conditions (e.g., 0.5M sodium phosphate, 0.1% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C.).

Hybridization under high stringency conditions only occurs between sequences with a very high degree of complementarity. One of ordinary skill in the art will recognize that the parameters for different degrees of stringency will generally differ based upon various factors such as the length of the hybridizing sequences, whether they contain RNA or DNA, etc. For example, appropriate temperatures for high, medium, or low stringency hybridization will generally be lower for shorter sequences such as oligonucleotides than for longer sequences.

[0029] The term *human immunodeficiency virus (HIV)*, is used here to refer to any strain of HIV-1 or HIV-2 that is capable of causing disease in a human subject, or that is an interesting candidate for experimental analysis. Furthermore, as will be clear from context, the term *HIV* is often used to refer to a virus (e.g., FIV, SIV) that is highly related to HIV but infects a different host. A huge number of HIV and SIV isolates have been partially or completely sequenced. Sequences of HIV genes are therefore readily available to, or determinable by, those of ordinary skill in the art.
[0030]  *Isolated*, as used herein, means 1) separated from at least some of the components with which it is usually associated in nature; and/or 2) not occurring in nature.

[0031]  *Purified*, as used herein, means separated from many other compounds or entities. A compound or entity may be *partially purified, substantially purified, or pure*, where it is *pure* when it is removed from substantially all other compounds or entities, i.e., is preferably at least about 90%, more preferably at least about 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or greater than 99% pure.

[0032]  The term *regulatory sequence* is used herein to describe a region of nucleic acid sequence that directs, enhances, or inhibits the expression (particularly transcription, but in some cases other events such as splicing or other processing) of sequence(s) with which it is operatively linked. The term includes promoters, enhancers and other transcriptional control elements. In some embodiments of the invention, regulatory sequences may direct constitutive expression of a nucleotide sequence; in other embodiments, regulatory sequences may direct tissue-specific and/or inducible expression. For instance, non-limiting examples of tissue-specific promoters appropriate for use in mammalian cells include lymphoid-specific promoters (see, for example, Calame et al., *Adv. Immunol.* 43:235, 1988) such as promoters of T cell receptors (see, e.g., Winoto et al., *EMBO J.* 8:729, 1989) and immunoglobulins (see, for example, Banerji et al., *Cell* 33:729, 1983; Queen et al., *Cell* 33:741, 1983), and neuron-specific promoters (e.g., the neurofilament promoter; Byrne et al., *Proc. Natl. Acad. Sci. USA* 86:5473, 1989). Developmentally-regulated promoters are also encompassed, including, for example, the murine hox promoters (Kessel et al., *Science* 249:374, 1990) and the α-fetoprotein promoter (Campes et al., *Genes Dev.* 3:537, 1989). In some embodiments of the invention regulatory sequences may direct expression of a nucleotide sequence only in cells that have been infected with an infectious agent. For example, the regulatory sequence may comprise a promoter and/or enhancer such as a virus-specific promoter or enhancer that is recognized by a viral protein, e.g., a viral polymerase, transcription factor, etc.

[0033]  *A short, interfering RNA (siRNA)* comprises an RNA duplex that is preferably approximately 19 basepairs long and optionally further comprises one or two single-stranded overhangs or loops. An inventive siRNA may comprise two
RNA strands hybridized together, or may alternatively comprise a single RNA strand that includes a self-hybridizing portion. siRNAs may include one or more free strand ends, which may include phosphate and/or hydroxyl groups. Inventive siRNAs include a portion that hybridizes under stringent conditions with a target transcript. In certain preferred embodiments of the invention, one strand of the siRNA (or, the self-hybridizing portion of the siRNA) is precisely complementary with a region of the target transcript, meaning that the siRNA hybridizes to the target transcript without a single mismatch. In most embodiments of the invention in which perfect complementarity is not achieved, it is generally preferred that any mismatches be located at or near the siRNA termini as described in more detail below. For the purposes of the present invention, any RNA comprising a duplex structure, one strand or portion of which binds to a target transcript and reduces its expression, whether by triggering degradation, by inhibiting translation, or by other means, is considered to be an siRNA, and any structure that generates such an siRNA is useful in the practice of the present invention.

[0034] The term subject, as used herein, refers to an individual to whom as siRNA is to be delivered, e.g., for experimental and/or therapeutic purposes. Preferred subjects are mammals, particularly domesticated mammals (e.g., dogs, cats, etc.), primates, or humans.

[0035] An siRNA is considered to be targeted for the purposes described herein if 1) the stability of the target gene transcript is reduced in the presence of the siRNA as compared with its absence; and/or 2) the siRNA shows at least about 90%, more preferably at least about 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, or 100% precise sequence complementarity with the target transcript for a stretch of at least about 17, more preferably at least about 18 or 19 to about 21-23 nucleotides; and/or 3) the siRNA hybridizes to the target transcript under stringent conditions.

[0036] The term vector is used herein to refer to a nucleic acid molecule capable of mediating entry of, e.g., transferring, transporting, etc., another nucleic acid molecule into a cell. The transferred nucleic acid is generally linked to, e.g., inserted into, the vector nucleic acid molecule. A vector may include sequences that direct autonomous replication, or may include sequences sufficient to allow integration into host cell DNA. Useful vectors include, for example, plasmids, cosmids, and viral
vectors. Viral vectors include, *e.g.*, replication defective retroviruses, adenoviruses, adeno-associated viruses, and lentiviruses. As will be evident to one of ordinary skill in the art, viral vectors may include various viral components in addition to nucleic acid(s) that mediate entry nucleic acid(s). The present invention provides vectors from which siRNAs designed according to the inventive methods may be expressed in relevant expression systems, *e.g.*, cells. Preferably, such expression vectors include one or more regulatory sequences operatively linked to the nucleic acid sequence(s) to be expressed.

Detailed Description of Certain Preferred Embodiments of the Invention

**RNA Interference and siRNA**

[0037] RNA interference (RNAi) is a phenomenon involving sequence-specific, post-transcriptional gene silencing, which can be initiated by double-stranded RNA (dsRNA) that is homologous in sequence to the silenced gene. Naturally occurring gene silencing is believed to occur at least in part via sequence-specific messenger RNA degradation mediated by 21- and 22-nucleotide small interfering RNAs (siRNAs) generated by ribonuclease III cleavage from longer dsRNAs.

[0038] Small inhibitory RNAs were first discovered in studies of RNAi in *Drosophila*, as described in WO 01/75164. In particular, it was found that, in *Drosophila*, long double-stranded RNAs are processed by an RNase III-like enzyme called DICER (Bernstein et al., *Nature* 409:363, 2001) into smaller dsRNAs comprised of two 21 nt strands, each of which has a 5’ phosphate group and a 3’ hydroxyl, and includes a 19 nt region precisely complementary with the other strand, so that there is a 19 nt duplex region flanked by 2 nt-3’ overhangs (see Figure 1). These small dsRNAs (siRNAs) act to silence expression of any gene that includes a region complementary to one of the dsRNA strands, presumably because a helicase activity unwinds the 19 bp duplex in the siRNA, allowing an alternative duplex to form between one strand of the siRNA (the antisense strand) and the target transcript. This new duplex then guides an endonuclease complex, RISC, to the target RNA, which it cleaves ("slices") at a single location corresponding to a position near the
middle of the siRNA duplex, producing unprotected RNA ends that are promptly degraded by cellular machinery (see Figure 2).

[0039] Relatives of the DICER enzyme, e.g., enzymes having RNase III-like activity, have now been found in diverse species ranging from *E. coli* to humans (Sharp, *Genes Dev.* 15;485, 2001; Zamore, *Nat. Struct. Biol.* 8:746, 2001), raising the possibility that an RNAi-like mechanism might be able to silence gene expression in a variety of different cell types including mammalian, or even human, cells. Unfortunately, however, long dsRNAs (e.g., dsRNAs having a double-stranded region longer than about 30 nucleotides) are known to activate the interferon response in mammalian cells. Thus, rather than achieving the specific gene silencing observed with the *Drosophila* RNAi mechanism depicted in Figure 2, introduction of long dsRNAs into mammalian cells would be expected to lead to interferon-mediated non-specific suppression of translation, potentially resulting in cell death. Long dsRNAs are therefore not thought to be useful for inhibiting expression of particular genes in mammalian cells.

[0040] However, it has been found that siRNAs, when introduced into mammalian cells, can effectively reduce the expression of host cell genes and/or heterologous genes such as those present on plasmids or in the genome of infectious agents such as viruses. For example, an siRNA targeted to human CD4, and selected in accordance with the rules disclosed herein, reduces the amount of CD4 mRNA and protein produced in human cells (Example 1). In addition, an siRNA targeted to the HIV p24 gene, selected in accordance with the rules disclosed herein, reduces the levels of p24 protein, and also reduces the levels of a variety of viral transcripts (Example 3). Similar results have been obtained with respect to a diverse array of genes, indicating the general applicability of the approach.

[0041] An siRNA may comprise two RNA strands hybridized together, or may alternatively comprise a single RNA strand that includes a self-hybridizing portion. siRNAs generally include a base-paired region approximately 19 nt long, and may optionally have one or more single-stranded overhangs or looped ends. Certain siRNAs may include bulges within the duplex region, representing less than perfect complementarity. In particular, RNAs known as microRNAs may typically display less than perfect complementarity within the duplex region. Figures 3, 4, and 5
present various structures that can be utilized as siRNAs. For example, Figure 3 shows the structure found to be active in the Drosophila system described above, and may represent the species that is active in mammalian cells.

Figures 4 and 5 present two alternative structures that may be used as siRNAs. Figure 4 shows an agent comprising an RNA strand containing two complementary elements that hybridize to one another to form a stem (element B), a loop (element C), and an overhang (element A). Preferably, the stem is approximately 19 bp long, the loop is about 1-20, more preferably about 4-10, and most preferably about 6-8 nt long and/or the overhang is about 1-20, and more preferably about 2-15 nt long. In general, the stem is minimally 19 nucleotides in length and may be up to approximately 29 nucleotides in length. One of ordinary skill in the art will appreciate that loops of 4 nucleotides or greater are less likely subject to steric constraints than are shorter loops and therefore may be preferred. In some embodiments, the overhang includes a 5' phosphate and a 3' hydroxyl. An agent having the structure depicted in Figure 4 can readily be generated by in vivo or in vitro transcription, in which case the transcript tail may be included in the overhang, so that often the overhang will comprise a plurality of U residues, e.g., between 1 and 5 U residues. Figure 5 shows an agent comprising an RNA circle that includes complementary elements sufficient to form a stem approximately 19 bp long (element B).

It will be appreciated by those of ordinary skill in the art that in general agents having any of the structures depicted in Figures 3, 4, and 5, or any other effective structure as described herein, may be comprised entirely of natural RNA nucleotides, or may instead include one or more nucleotide analogs. A wide variety of such analogs is known in the art; the most commonly-employed in studies of therapeutic nucleic acids being the phosphorothioate (for some discussion of considerations involved when utilizing phosphorothioates, see, for example, Agarwal, Biochim. Biophys. Acta 1489:53, 1999). In particular, in certain embodiments of the invention it may be desirable to stabilize the siRNA structure, for example by including nucleotide analogs at one or more free strand ends in order to reduce digestion, e.g., by exonucleases. The inclusion of deoxynucleotides, e.g., pyrimidines such as deoxythymidines at one or more free ends may serve this purpose.
Alternatively or additionally, it may be desirable to include one or more nucleotide analogs in order to increase or reduce stability of the 19 bp stem, in particular as compared with any hybrid that will be formed by interaction of one strand of the siRNA with a target transcript. According to certain embodiments of the invention various nucleotide modifications are used selectively in either the sense or antisense strand. For example, it may be preferable to utilize unmodified ribonucleotides in the antisense strand while employing modified ribonucleotides and/or modified or unmodified deoxyribonucleotides at some or all positions in the sense strand. According to certain embodiments of the invention only unmodified ribonucleotides are used in the duplex portion of the antisense and/or the sense strand of the siRNA while the overhang(s) of the antisense and/or sense strand may include modified ribonucleotides and/or deoxyribonucleotides.

Numerous nucleotide analogs and nucleotide modifications are known in the art, and their effect on properties such as hybridization and nuclease resistance has been explored. For example, various modifications to the base, sugar and internucleoside linkage have been introduced into oligonucleotides at selected positions, and the resultant effect relative to the unmodified oligonucleotide compared. A number of modifications have been shown to alter one or more aspects of the oligonucleotide such as its ability to hybridize to a complementary nucleic acid, its stability, etc. For example, useful 2'-modifications include halo, alkoxy and allyloxy groups. US patent numbers 6,403,779; 6,399,754; 6,225,460; 6,127,533; 6,031,086; 6,005,087; 5,977,089, and references therein disclose a wide variety of nucleotide analogs and modifications that may be of use in the practice of the present invention. See also Crooke, S. (ed.) Antisense Drug Technology: Principles, Strategies, and Application (1st ed), Marcel Dekker; ISBN: 0824705661; 1st edition (2001) and references therein. Certain analogs and/or modifications may result in an siRNA with increased oral absorbability, increased stability in the blood stream, increased ability to cross cell membranes, etc. As will be appreciated by one of ordinary skill in the art, analogs or modifications may result in altered Tm, which may result in increased tolerance of mismatches between the siRNA sequence and the target while still resulting in effective suppression.
[0045] As will be appreciated by one of ordinary skill in the art, analogs and modifications may be tested using, e.g., the assays described herein or other appropriate assays, in order to select those that effectively reduce expression of a gene of interest. The extent to which the presence of such analogs and modifications affects the efficiency of siRNA mediated gene silencing remains to be determined, but the methods of the invention may in any event be applied to select preferred siRNA sequences regardless of the particular analogs and/or modifications that are employed. However, according to certain embodiments of the invention the systems and methods described herein are to be understood as applying to siRNA comprised of ribonucleotides such as are found within naturally occurring RNA, with the proviso that deoxyribonucleotides may be employed in the free 3' ends.

**siRNA Selection and Design**

[0046] The findings described above and in the Examples demonstrate that different siRNAs can effectively reduce or inhibit expression of a variety of endogenous and heterologous genes, resulting in decreased levels of the corresponding mRNA and its encoded protein(s). The potential for exploiting this phenomenon for research and for therapeutic purposes is immense. During the course of the experiments a variety of siRNAs with different sequences were designed and tested, and it became evident that not all siRNAs were equally effective in reducing or inhibiting expression of any particular target gene. Other workers have also reported variability in the efficacy of different siRNAs (Holen, T., et al., *Nucleic Acids Res.*, 30(8):1757-1766).

[0047] Initially, the selection of siRNA sequences proceeded via a process with few or no constraints imposed on the siRNA sequence other than that it display complementarity to the mRNA target transcript. For example, initially siRNAs were selected by (i) identifying 23 nt regions in the target transcript consisting of 19 nt regions flanked by two AA residues at the 5' end and two TT residues at the 3' end and then (ii) selecting siRNAs having an antisense strand perfectly complementary to nucleotides 1 – 21 of the 23 nt region and a sense strand perfectly identical to nt 3 - 23 of the 23 nt region. However, as experience accumulated, the inventors developed methods that resulted in an improved ability to select and/or design effective siRNAs.
It will be recognized that the process of choosing an siRNA sequence may involve elements of both selection and design and that these terms overlap. Their use either separately or in combination herein is not intended to imply a clear distinction between them, though different aspects of the process of choosing an siRNA sequence may more closely approximate a selection process while other aspects more closely approximate a design process. In general, the use of the term “selection” is intended to encompass elements of design and vice versa.

[0048] The present invention therefore provides systems and methods for selecting and/or designing preferred siRNA sequences for forming the duplex portion of the siRNA based on a target sequence. Certain embodiments of the invention include systems and methods for selecting and/or designing additional portions of an siRNA, which may include free 3’ ends and/or unpaired loop regions.

[0049] In general, in accordance with findings of the inventors and others, inventive siRNAs will preferably include a region (the “inhibitory region” or “duplex region”) that is substantially complementary to that found in a portion of the target transcript (the “target portion”), so that a precise hybrid can form in vivo between the antisense strand of the siRNA and the target transcript. This duplex region, also referred to as the “core region” is understood not to contain the overhangs, although the overhangs may also be complementary to the target transcript. Preferably, this substantially complementary region includes most or all of the stem structure depicted in Figures 3, 4, and 5. The relevant inhibitor region of the siRNA is preferably perfectly complementary with the target transcript although siRNAs including one or more non-complementary residues have been shown to mediate silencing in some experiments and may be designed in accordance with the teachings herein.

Preferably, fewer than three residues or alternatively less than about 15% of residues, in the inhibitory region are mismatched with the target.

[0050] Rules for selection and design of optimal siRNA sequences are described below. For purposes of description it will be assumed that the rules below operate on a target mRNA transcript, which is equivalent in sequence to a sense strand of a cDNA. Of course the rules can operate on sequences provided in any format (e.g., as double-stranded DNA or cDNA), provided that the identity of the sense transcript is clear. One of ordinary skill in the art will readily be able to adapt these rules to
operate on target transcript sequences presented in any format. In addition, these rules may be used to select siRNAs formed by hybridization of two independent RNA strands and/or to select siRNAs that are formed by intramolecular hybridization between complementary portions of a single RNA strand (e.g., stem-loop structures).

I. Target Portion Selection Rules.

Scan the target transcript, preferably including 5' and 3' untranslated regions if available, and identify potential target portions having an appropriate length. This step identifies at least a core region that will be a duplex in the resulting siRNA and may also identify overhangs. Typically the length of the core region is 19 nucleotides (nt), which may be set as the default parameter in computer-based embodiments of the method. The length may also be a parameter that may be selected by the user. For purposes of description herein the length will be assumed to be 19 nucleotides, and a 19 nucleotide sequence is referred to as N19. However, the core region may range in length from 15 to 29 nucleotides. In addition, it is assumed that the siRNA N19 inhibitory region will be chosen so that the antisense strand of the siRNA is perfectly complementary to the mRNA target, though as mentioned above one or more mismatches may be tolerated. In general it is desirable to avoid mismatches in the duplex region if an siRNA having maximal ability to reduce expression of the target transcript is desired. However, as described below, it may be desirable to select an siRNA that exhibits less than maximal ability to reduce expression of the target transcript. In such a situation it may be desirable to incorporate one or more mismatches in the duplex portion of the siRNA.

Figure 12 presents a schematic of an mRNA target 100 with nucleotides indicated in accordance with the discussion below. Only part of the target transcript is depicted. In Figure 12, an “X” opposite an “N” represents a nucleotide complementary to N. Nucleotides indicated with an “N” and no subscript represent target portion 110. As indicated on the figure, target portion 110 is an N19 portion, 19 nucleotides in length. The siRNA 120 comprises a sense strand 130 and an antisense strand 140. One or more of nucleotides 150 and 155 indicated with “N1” and “N2” located immediately 5' of the target portion may be complementary to the 3' overhang 160 of the antisense strand of a corresponding siRNA. In other words, one or more of N1 and N2 may be complementary to X1 and X2 respectively. One or
more of the nucleotides located immediately 3' of the target portion (indicated with "N22" and "N23") may be identical to the 3' overhang 170 of the sense strand of a corresponding siRNA. In other words, one or more of N22 and N23 in the target transcript may be identical to N20 and N21 in the sense strand of the siRNA respectively. As described below, the siRNA overhangs need not correspond to the target transcript. Selection of target portions may be performed, in general, according to two different approaches:

(A) Tiling Approach. According to this approach each stretch of nucleotides of appropriate length (e.g., 19 nucleotides) is a potential target portion. For example, if it is desired to select target portions for the sequence below

5' - AGTTAACTGCTTAGCTCATCAGTGCTTACAAAA-3' (SEQ ID NO: 15), then the first four target portions would be:

5'-AGTTAACTGCTTAGCTCAT-3' (SEQ ID NO: 16)
5'-GTTHAACTGCTTAGCTCATT-3' (SEQ ID NO: 17)

5'-TTAACTGCTTAGCTCATTC-3' (SEQ ID NO: 18)
5'-TACCTGCTTAGCTCATTTCA-3' (SEQ ID NO: 19)

Overhangs may be added in accordance with rule V. Alternately, 3' overhangs may be selected to match the two nucleotides immediately 3' of the target portion on each strand, so that the N21 sense strand of the siRNA has the same sequence as the target and the N21 antisense strand of the siRNA is complementary to the target.

Constrained Overhang Approach. According to this approach portions of the target transcript in which the 5' end of the core N19 sequence is flanked by one or two purines (A or G) and/or the 3' end of the core N19 siRNA sequence is flanked by one or two pyrimidines (C or T) are selected. Sequences of 23 nucleotides (N23) are thus identified in the target transcript, which conform to the pattern 5'-N1N2(N19)N22N23-3', where either or both of N1 and N2 are purines (A or G) and/or either or both of N22 and N23 are pyrimidines (C or U). The corresponding siRNA includes a sense strand identical in sequence to the (N19)N22N23 sequence in the target and an antisense strand complementary to the N1N2(N19) sequence in the target. Thus selection of the N23 portion of a target transcript is sufficient to specify both strands of the siRNA. Note that if a double-stranded cDNA or genomic DNA
sequence is used, this process is equivalent to selecting target portions in which the core N19 sequence is flanked by 2 nucleotide 3’ overhangs on each strand, where at least one of the nucleotides in one of the overhangs is a pyrimidine (C or T), i.e., an N21 sense strand and an N21 antisense strand.

[0061] The following rules and guidelines may be applied to compare, score, and/or rank sequences selected according to the constrained overhang approach. These rules are expressed in terms of the N21 siRNA sense and antisense sequences (or, equivalently except for the use of U rather than T, in terms of the corresponding N21 cDNA sense and antisense sequences) rather than the corresponding N23 sequence in the target transcript.

[0062] (1) In general, U is preferred to C, and C is preferred to G and A.

[0063] (2) The presence of pyrimidines in the 3’ overhang of the antisense strand of the siRNA is more important than the presence of pyrimidines in the 3’ overhang of the sense strand of the siRNA. Therefore, siRNA sequences having antisense strands conforming to the pattern 5’-(N19)PyPy-3’, 5’-(N19)PyPu-3’, or 5’-(N19)PuPy-3’ are preferred to siRNA sequences containing one or more pyrimidines in the 3’ overhang of the sense strand only.

[0064] (3) In general, for 3’ overhangs on either the sense or antisense siRNA strand, the presence of a pyrimidine in the ultimate (last) position is more important than the presence of a pyrimidine in the penultimate (second to last) position of the overhang. For example, an siRNA sequence having an antisense strand conforming to the pattern 5’-(N19)PuPy-3’ is preferred to an siRNA sequence having an antisense strand conforming to the pattern 5’-(N19)PyPu-3’.

[0065] According to certain embodiments of the invention a numerical score is assigned to each siRNA based on the identities of the nucleotides at each position in the 3’ overhangs of the sense and antisense strands. Each overhang nucleotide contributes to the score depending on its identity. For example, according to such a system, a pyrimidine in the last position in the 3’ overhang of the antisense strand contributes a higher value to the score than a pyrimidine in the penultimate position of that overhang. Similarly, a pyrimidine in either position in the 3’ overhang of the antisense strand contributes a higher value to the score than a pyrimidine in either position of the 3’ overhang of the sense strand of the siRNA. Such a scoring system
permits the ranking of siRNA sequences based on the identity of the nucleotides in the 3’ overhangs.

[0066] Note that in general use of the constrained overhang approach can rapidly eliminate many potential siRNA sequences. For example, if the selection is constrained by the requirement that both overhangs in the siRNA are UU (and thus that the corresponding portion of the transcript has the configuration 5’-AA(N19)UU-3’, or equivalently that the sense strand of the cDNA has the configuration 5’-(N19)TT-3’ and the antisense strand of the cDNA has the configuration 5’-(N19)TT, then it is likely that only a small number of candidate siRNAs will remain after application of rule I(B). This greatly reduces the amount of processing required to evaluate the sequences in accordance with the remaining rules described below.

[0067] II. Complexity Rules. Both the overall composition of the siRNA sequence (i.e. the relative percentage of G/C vs A/T) and also the presence of strings of repeated nucleotides are significant.

[0068] (A) Composition. The composition of the sequence is evaluated by calculating the percentages of A, T, G, and C on either strand. Note that if overhangs are present the ratio of nucleotides may differ depending on whether calculations are based on the sense strand or the antisense strand. In this case it is preferable to make the calculation based on the antisense strand, including the overhang. The ratio should be approximately equimolar (1:1:1:1). Sequences having ratios progressively closer to 1:1:1:1 are ranked progressively higher as the ratio approaches 1:1:1:1. Sequences in which the percentage of GC pairs (G/C content) is greater than approximately 70% or less than approximately 30% should be avoided. Sequences having GC to AU pair ratios closer to 1:1 are ranked progressively higher as the ratio approaches 1:1. For example, it is preferable to select a target site so that the ratio of GC to AU basepairs in the siRNA is within the range of approximately 0.75:1 to approximately 1.25:1, preferably within the range of approximately 0.9:1 to approximately 1.1:1, more preferably closer to approximately or exactly 1:1. The desired range of ratios may be a parameter that is set by the user. Note that the ratio of GC to AU pairs alone does not adequately describe the complexity of the sequence. For example, the sequence GAGAGAGAGAG (SEQ ID NO: 20) has a GC:AU ratio of 1:1 but has lower complexity than the sequence GACTGACTGACT (SEQ ID

Page 21 of 86
NO:21). This may be seen by comparing the nucleotide ratios (1:1:0:0) for SEQ ID NO: 21 as compared to (1:1:1:1) for SEQ ID NO: 22. Thus both overall GC content and nucleotide ratios are significant in evaluating sequence complexity.

[0069] (B) Cluster avoidance. Complexity involves more than simply the relative proportions of the four nucleotides. The presence of strings of repeated nucleotides reduces overall complexity independent of the particular nucleotide composition. For example, even though the sequences GGGCCCAATTT (SEQ ID NO: 23) and GTCACTGCTAGA (SEQ ID NO: 24) both contain 3 G residues, 3 C residues, 3 A residues, and 3 T residues, the second sequence exhibits greater complexity than the first since it lacks contiguous blocks of G, C, A, or T. Generally it is desirable to utilize high complexity target sites, e.g., sites that include most or all residues, preferably in a stochastic pattern, avoiding stretches in which a single residue is repeated multiple times. Methods, including computer based methods, for performing pattern analysis and determining the complexity of any particular sequence are well known in the art. For purposes of description, a doublet is defined as a string of two consecutive identical nucleotides; a triplet is defined as a string of three consecutive identical nucleotides; a doublet repeat is a string in which the same set of two nonidentical nucleotides is repeated, e.g., GTGTGTGT (SEQ ID NO: 25); and a triplet repeat is a string in which the same set of three nucleotides (at least two of which are different) is repeated, e.g., GACGACGAC (SEQ ID NO: 26) or GGAGGAGGA (SEQ ID NO: 27). According to certain embodiments of the invention ranking of sequences based on complexity follows one or more of the following rules:

[0070] (1) Sequences having one or more stretches of four consecutive identical nucleotides should be avoided.

[0071] (2) Sequences having one or more stretches of three consecutive identical nucleotides are less preferred than sequences lacking such stretches.

[0072] (3) Sequences having a stretch of three consecutive identical nucleotides are preferred to sequences having a row of three or more doublets (e.g., a sequence containing CCC is preferred to a sequence containing CCGGAA (SEQ ID NO: 28).

[0073] (4) Sequences containing doublet or triplet repeats are less preferred to sequences lacking such repeats.
Sequences may be compared, scored, and/or ranked according to the degree to which they conform to or violate these criteria. In sum, the invention provides a method for selecting an siRNA targeted to a target transcript comprising: applying a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3’ overhang; and applying a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs. Note that identifying a target portion does not require that the portion be explicitly identified as such. Identifying a sequence that includes a target portion, e.g., a sequence that includes overhangs, implicitly identifies a target portion.

III. Suboptimal element positioning. In many situations it will not be possible to select an siRNA sequence that satisfies all the criteria outlined above. Such siRNAs are defined as containing suboptimal elements. For example, a string of three or four consecutive identical nucleotides is a suboptimal element. A string of repeated doublet or triplet repeats, e.g., GTGTGT (SEQ ID NO: 29) or GACGACGAC (SEQ ID NO: 30) is a suboptimal element. When comparing, scoring, or ranking siRNA sequences containing suboptimal elements, sequences in which the suboptimal element(s) are located toward the ends (either 5’ or 3’) of the siRNA are preferred to sequences in which the suboptimal element(s) are closer to the middle. Thus the distance of any suboptimal element from the end can be incorporated into the score assigned to the sequence.

IV. Base Pair Optimization.

(A) siRNA sequences that contain clusters of Gs (e.g., groups of three or more consecutive Gs) on one strand near clusters of three or more As or Us on the opposite strand are less preferred than siRNA sequences lacking such clusters. Avoidance of such clusters reduces the likelihood of out of register pairing, i.e., pairing in which nucleotides in one strand fail to pair with the nucleotide in the appropriate position in the opposite strand. For example, if both strands are N21, then each nucleotide at position X in the N19 core of the sense strand should pair with the nucleotide at position 20 – X in the antisense strand, where nucleotides are counted...
from the 5' end of the strand, starting with 1. By "near" is meant that the cluster of As or Ts on the opposite strand is either adjacent to the cluster of Cs that would be base paired with the Gs in a correctly base paired duplex, or is separated from that cluster by 1, 2, or 3 nucleotides.

5 (B) siRNA sequences that minimize possibilities for non Watson-Crick base pairing are preferred to sequences that offer such possibilities. Watson-Crick base pairing refers to G-C and A-U pairs. Non Watson-Crick base pairing possibilities include G-U wobble (e.g., G-A, G-U), Hoogsteen base pairing (e.g., A-U, A-A, U-U), and inosine base pairing (e.g., I-U, I-A, I-C). Inosine is an intermediate in purine synthesis, and adenosine may be deaminated to inosine by cellular adenosine deaminase. Thus possibilities for non Watson-Crick base pairing involving inosine may exist in vivo.

10 V. Overhang Refinement. Rules listed under I outlined two general approaches to the selection of the core N19 sequence of the siRNA. The rules in this section address selection of the overhangs in the case that they are not completely determined by I.

15 (A) If the candidate siRNA sequences were selected according to the tiling approach described in I(A), then the 3' overhangs may be chosen so that the 3' overhang of the sense siRNA strand is partly or completely identical to the nucleotides immediately 3' of the N19 portion of the target mRNA transcript and the 3' overhang of the antisense siRNA strand is partly or completely complementary to the nucleotides immediately 5' of the N19 portion of the target mRNA transcript. Alternately, the sequences of the 3' overhangs may be selected freely. To the extent that the selection process does not determine the sequences of the overhangs based on the sequence of the target transcript, they may be selected in accordance with the principles in V(B). In particular, if the duplex portion of the siRNA has a preferred sequence in accordance with the rules described in II, III, IV, and/or VI, i.e., but contains purines in the overhangs, it may be desirable to replace them as described below.

20 (B) If the candidate siRNA sequences were selected according to the constrained overhang approach described in I(B), then the 3' overhangs were initially determined during the selection process. These initially selected overhangs may be
modified in any of a number of ways. For example, if there are purines in the 3’ overhangs, they may be replaced by pyrimidines (U, T, C, dU, dT, or dC) may advantageously selected. In particular, dT has been shown to be effective in a variety of experiments in mammalian cells) and may be preferred. Pyrimidines may offer increased stability compared with purines due to decreased nuclease sensitivity. Deoxythymidine, in particular, offers the possibility of purifying the siRNA using an oligo-dA column, which is less expensive than an oligo-A column, and may also increase the stability of the siRNA to exonucleolytic attack.

VI. Specificity. It is preferred that the siRNA does not display significant sequence identity or homology with other known sequences, particularly for genes that are important to the biological process under study and/or essential for cell viability.

(A) To avoid selection of sequences having significant identity or homology with genes that are important to the biological process under study and/or essential for cell viability, a database search using programs such as BLAST, BLASTNR, or CLUSTALW (or variations thereof) in a comprehensive database such as GenBank, Unigene, etc., can be performed using, e.g., default parameters and matrices (e.g., BLOSUM substitution matrix). (BLAST is described in Altschul, SF, et al., Basic local alignment search tool, J. Mol. Biol., 215(3): 403-410, 1990, Altschul, SF and Gish, W, Methods in Enzymology. Additional discussion and references to appropriate computer programs are found in Baxevanis, A., and Ouellette, B.F.F., Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Wiley, 1998; and Misener, S. and Krawetz, S. (eds.), Bioinformatics Methods and Protocols (Methods in Molecular Biology, Vol. 132), Humana Press, 1999.) Preferably both strands of the siRNA are searched, as is done automatically by the BLAST program. In general, it is preferable to avoid sequences that display significant identity or homology to known genes even if the products of such genes are not known to be important in the biological process under study or essential for cell viability.

(B) If identity or homology does exist over part of the siRNA, it is preferable that the region of identity or homology is located towards either end of the siRNA rather than near the middle. For example, the length of any stretch of identity or significant homology and the distance between the middle of such a stretch and the
middle of the siRNA may be used when assigning a score or rank to the siRNA or comparing it with other siRNAs. Generally identity over the first and last thirds of the sequence is acceptable provided that the middle third, particularly the central 4 to 7 nucleotides differ. Identity of as many as 17 or 18 out of 19 nucleotides in the duplex region may be acceptable, provided that the 1 or 2 nonidentical nucleotides occur(s) in the central third or, more preferably, the central 4 to 7 nucleotides of the siRNA. However, such a high degree of identity is not preferred when designing siRNAs for maximum inhibition of the target. Thus in the context of these rules, two 19 nt sequences do not display significant identity or homology if any of the following conditions are met: (i) any areas of identity are confined to nucleotides between positions 1 to 6 or 7 and/or positions 14 or 15 to 19; (ii) any areas of identity do not include nucleotides 9 to 12 of the sequences; (iii) any areas of identity do not include nucleotides 10 to 13 of the sequences; (iv) the sequences contain at least one nonidentical nucleotide at a position between nucleotides 8 to 12 of the sequence; (v) the sequences contain at least one nonidentical nucleotide at a position between nucleotides 9 to 13 of the sequence; (vi) the sequences contain at least two nonidentical nucleotides at a position between nucleotides 8 to 12 of the sequence; (vii) the sequences contain at least two nonidentical nucleotides at a position between nucleotides 9 to 13 of the sequence.

(C) As described further below, there is evidence to suggest that certain siRNAs that bind to the 3' UTR of a template transcript may inhibit expression of a protein encoded by the template transcript by a mechanism related to but distinct from classic RNA interference, e.g., by reducing translation of the transcript rather than decreasing its stability. Such RNAs are referred to as microRNAs (miRNAs) and are typically approximately 22 nt in length. It is believed that they are derived from larger precursors known as small temporal RNAs (stRNAs) approximately 70 nt long. siRNAs such as miRNAs that bind within the 3' UTR (or elsewhere in a target transcript) and inhibit translation may tolerate a larger number of mismatches in the siRNA/template duplex, and particularly may tolerate mismatches within the central region of the duplex. In fact, there is evidence that some mismatches may be desirable or required as naturally occurring stRNAs frequently exhibit such mismatches as do miRNAs that have been shown to inhibit translation in vitro (Zeng,
et al., referenced above). For example, when hybridized with the target transcript such siRNAs frequently include two stretches of perfect complementarity separated by a region of mismatch as shown schematically in Figure 14A, which depicts a microRNA 280 hybridized to a target site 285. As shown therein, the hybridized complex includes two regions of perfect complementarity (duplex portions) 290 indicated as nucleotide pairs and a two nucleotide area of mismatch (bulge) 295 separating the two duplex portions. A variety of structures are possible. For example, the miRNA may include multiple areas of nonidentity (mismatch). The areas of nonidentity (mismatch) need not be symmetrical in the sense that both the target and the miRNA include nonpaired nucleotides. For example, Figure 14B shows a structure in which only one strand includes nonpaired nucleotides. Typically the stretches of perfect complementarity are at least 5 nucleotides in length, e.g., 6, 7, or more nucleotides in length, while the regions of mismatch may be, for example, 1, 2, 3, or 4 nucleotides in length.

[0087] The inventors have recognized that any particular siRNA may function to inhibit gene expression both via (i) the "classical" siRNA pathway, in which stability of a target transcript is reduced and in which perfect complementarity between the siRNA and the target is frequently preferred and also by (ii) the "alternative" pathway in which translation of a target transcript is inhibited. Generally the transcripts targeted by a particular siRNA via mechanism (i) would be distinct from the transcript targeted via mechanism (ii) although it is possible that a single transcript might contain regions that could serve as targets for both the classical and alternative pathways. (Note that the terms "classical" and "alternative" are used merely for convenience and do not reflect the importance, effectiveness, or other features of either mechanism.)

[0088] Given the possibilities outlined above, the inventors have recognized that when intending to target a particular transcript via mechanism (i) it is generally desirable to avoid inadvertently targeting a transcript via mechanism (ii). Therefore, according to certain embodiments of the invention it is desirable to avoid siRNA sequences that meet the criteria for inhibiting target transcripts via mechanism (ii). For example, when designing an siRNA to inhibit a first target transcript, it is desirable to avoid sequences that display two regions of identity (or complementarity)
with any known gene separated by a region of nonidentity (or mismatch). According to certain embodiments of the invention it is desirable to avoid siRNAs wherein the two regions of identity (or complementarity) are between 5 and 10 nucleotides in length, e.g., 5, 6, 7, 8, 9, or 10 nucleotides in length, and wherein the region(s) of nonidentity (or mismatch) is/are between 1 and 6 nucleotides in length, e.g., 1, 2, 3, 4, 5, or 6. According to certain embodiments of the invention, for example, it is preferred to avoid siRNAs that include two regions of identity (or complementarity), each 6 or 7 nucleotides in length, separated by a region of nonidentity (or mismatch) of 2 to 4 nucleotides in length, with any known gene, mRNA, etc. Database searches using programs such as BLAST, BLASTNR, or CLUSTALW (or variations thereof) may be used to identify siRNAs that have the potential to inhibit genes other than the intended target via mechanism (ii).

VII. Global Positioning. Any portion of a target transcript may be searched to identify candidate siRNA sequences, including the 5' and 3' UTR. In general, it is preferable to avoid selecting sequences within introns (which might occur if the selection is based on genomic DNA or unspliced pre-mRNA). Information about the location of a candidate siRNA relative to the 5' and/or 3' end of the target transcript and information about the position of the candidate siRNA relative to any exon/intron boundaries that might exist is helpful in selecting preferred siRNAs.

(1) Position relative to transcript ends. In general, sequences located closer to the 3' end of the target transcript are preferred to sequences nearer to the middle or 5' end of the mRNA target. In general, sequences closer to the 5' end of the target are preferred to sequences closer to the middle with the proviso that the first 50 to 75 nucleotides following the AUG may be less preferred since they may be protected by the translational machinery. While not wishing to be bound by any theory, the inventors suggest that the 3' portion of target transcripts may be less likely to exhibit secondary structure that may inhibit or interfere with siRNA activity, e.g., by reducing accessibility.

The distance of a candidate siRNA sequence from the 3' and 5' ends of the transcript may be used when assigning a score to the siRNA. For example, with respect to the global positioning parameter, an N19 sequence located at positions
1000-1018 of the target would receive a higher score than a sequence located at positions 100-118, which would in turn receive a higher score than a sequence located at positions 500-518 (assuming a scoring system in which a higher score indicates a preferred sequence).

5 [0092] (2) Position relative to exon/exon boundaries. In general, siRNA sequences located within a single exon are preferred to siRNA sequences that span an exon/exon boundary. While not wishing to be bound by any theory, the inventors suggest that here may be sequences, bound by trans-acting RNA binding factors that obscure sites on the target mRNA. Some factors that may be present at or occupy a site on the mRNA might include nonsense-mediated decay factors.

10 [0093] VIII. Accessibility. Any information available regarding mRNA accessibility may be used to guide selection of preferred siRNAs. For example, the accessibility of various portions of a target transcript may be assessed using RNase H protection techniques, taking advantage of the ability of RNase H to selectively cleave the RNA portion of RNA/DNA hybrids. In one such assay, oligonucleotides having the sequence of either strand of a candidate siRNA are allowed to hybridize to target RNA transcripts. The target transcript is exposed to RNase H under conditions compatible with RNase H activity. If the oligonucleotide is able to anneal to the complementary sequence of the RNA, RNase H will cleave the RNA within the double-stranded DNA/RNA region. However, regions of the target RNA that are capable of forming secondary structures, e.g., self-complementary regions, are more likely to be resistant to RNase H digestion than regions that do not form such structures. Portions of the RNA that survive such exposure are isolated and sequenced. These portions represent sequence that may be less accessible and thus not preferred for the design of siRNAs. RNA to be tested may be chemically synthesized, synthesized using in vitro transcription, or purified from cells. The latter approach may also reveal regions of the RNA that may be prevented from binding to oligonucleotides, e.g., by proteins, and may thus be less likely to be preferred regions to use in designing siRNAs. (See, e.g., Scherr, M. and Rossi, J.J. Nucl. Acids Res. 26, 5079-5085 (1998); Scherr, M., et al., Mol. Ther. 4, 454-460 (2001); Lee, N.S., et al., Nat. Biotech. 20, 500-505 (2002); Gunzl, A., et al, Methods, 26(2):162-9, Feb., 2002)).
[0094] Of course the general approach embodied in the foregoing method is not limited to RNase H but may employ any other nuclease that preferentially digests the RNA portion of a DNA/RNA hybrid. Enzymes that preferentially degrade or cleave double-stranded RNA while leaving single-stranded RNA intact (or vice versa), may be used in a similar fashion to identify preferred portions of the target (e.g., portions with a lesser propensity to assume secondary structures relative to other portions) for use in designing siRNAs.

[0095] Alternatively or additionally, RNA accessibility may be evaluated using a variety of computational approaches, e.g., such as those utilized in RNA folding programs such as Mfold. See, e.g., programs and information available at the Website having URL http://bioinfo.math.rpi.edu/~zukerm/rna/. See also Zuker, M., et al., “Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide” in Barciszewski, J. and Clark, B.F.C. (eds.), RNA Biochemistry and Biotechnology, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999 and Mathews, D., et al., J. Mol. Biol., 288, 911-950, 1999. The existence and affects of RNA secondary structure on the ability of an siRNA antisense strand to hybridize with its target transcript may be evaluated as described in Vickers, TA, et al., Nucleic Acids Research, 28(6), 2000. According to certain embodiments of the invention information regarding RNA secondary structure and/or accessibility is provided by the user. According to other embodiments of the invention a program such as Mfold is automatically invoked and used to evaluate potential siRNA sequences. Of course it is not necessary to utilize information about RNA secondary structure and/or accessibility to select effective siRNA sequences.

[0096] IX. Polymorphism Avoidance. If possible it is preferred to avoid sequences that include known polymorphisms since polymorphisms may result in differences in sequence between the siRNA and the target transcript in the particular system under study.

[0097] X. Additional Considerations. The methods and systems described herein can readily be applied to select and/or design potential new siRNAs, targeted to any particular gene of interest. As described above, the selection and design methods are generally based on the sequence of the target transcript without considering the biological context. However, as will be evident to one of ordinary skill in the art, in
certain situations information specific to the particular biological context in which the siRNA is to be used is helpful in selecting and/or designing a preferred siRNA sequence and may even supersede certain of the considerations described above. For example, if certain regions of the mRNA are known to be associated with RNA binding proteins it may be desirable to avoid selecting sequences corresponding to such regions.

[0098] The rules described above may generally be grouped into a number of categories. Rules listed under I are referred to as Target Portion Selection Rules because they identify a portion of the target transcript whose sequence will supply the N19 duplex core of the siRNA. Rules listed under II are referred to as Complexity Rules because they relate to the complexity of the siRNA sequence. Rule II(A) is referred to as a Composition Rule. Rules listed under II(B) are referred to as Cluster Avoidance Rules. Rules listed under III are referred to as Suboptimal Element Positioning Rules. Rules listed under IV are referred to as Base-Pair Optimization Rules. Rules listed under V are referred to as Overhang Refinement Rules. Rules listed under VI are referred to as Specificity Rules. Rules listed in VII are referred to as Global Positioning Rules. In general, rules in groups II through IV, VI, and VII can be applied to siRNA sequences selected in accordance with either approach outlined in Rule I and/or after modification of the overhangs in accordance with Rules listed under V.

[0099] To use the rules above in selection of siRNA sequences it will typically be necessary to compare candidate sequences with each other, which may involve assigning a score or rank to candidate sequences in accordance with the degree to which they conform or fail to conform to the preferences embodied in the rules. The comparison, scoring, and/or ranking steps can be performed in any of a number of different ways. For example, each potential siRNA could be assigned a certain number of “points” initially, and points could be subtracted for any deviation from perfect conformity with the preferences described above, in which case a lower score would indicate a preferred sequence. Alternately, siRNA sequences may be given points if they conform to the preferences. Sequences may be assigned an overall score and/or they may be assigned subscores for each set of rules. The different sets of rules may function as “filters”, which may be applied sequentially to a set of
candidate sequences to eliminate less preferred sequences. The rules may be applied in successive iterations, refining an initially selected set of preferred siRNAs. According to certain embodiments of the invention the rules are implemented as a set of "if-then" statements (or the equivalent, depending on the programming language). Such rules may add or subtract appropriate values to a score for an siRNA depending on whether or not it meets the condition embodied in the "if" statement. By assigning different scores, the rules may be assigned different weights.

[00100] It will be appreciated that certain of the rules may be more significant than others in terms of selecting an optimum siRNA sequence. Thus the rules may be given different weights in accordance with their importance. Such weights may be assigned automatically (e.g., as default values). Alternately, some or all of the weights may be selected by a user. In general, rules listed under II, III, IV, and VII are listed in the order of importance. For example, a sequence that conforms with Rule II(B)(1) but does not conform with Rule II (B)(2) would be preferred to a sequence that conforms with Rule II(B)(2) but does not conform with Rule II(B)(1) (assuming equal conformance with the other rules). However, weights may be selected so that an siRNA sequence that would be preferred based on its characteristics with respect to a single more important rule but that violates several rules of lesser importance would not necessarily be preferred over a sequence that violates the more important rule but that conforms with the rules of lesser importance.

In general, unless otherwise stated, when a rule indicates that a sequence having or lacking a particular feature is preferred (or not preferred) relative to a sequence lacking or having the feature respectively, the comparison assumes that the sequences are otherwise identical with respect to features not considered by that rule. As will be evident to one of ordinary skill in the art, the degree to which preferred siRNA sequences can conform to the rules is constrained to a certain extent by the sequence of the target transcript, and it may not always be possible to select a sequence that conforms to one or more of the rules. The methods are therefore flexible in order to handle such situations.

[00101] Comparisons between different siRNAs can be performed at any step during application of the rules. The siRNAs can be compared pairwise, or any particular siRNA can be compared with all other potential candidates. As selection
progresses, a list of candidates may be maintained. This list may be specified to remain a certain size, so that once a certain number of candidates is identified addition of another candidate to the list requires removing a less preferred sequence. It is not necessary to rank all possible candidates. In certain embodiments of the invention it is preferred to divide the candidates into different groups, e.g., an “optimum” group, an “acceptable” group, a “nonpreferred group”, etc. According to certain embodiments of the invention all or a substantial proportion (e.g., greater than approximately 50%, greater than approximately 75%, greater than approximately 90%, or more, up to 100% of the candidate siRNAs are ranked in order of their predicted efficacy in inhibiting expression of the target transcript.

[00102] One aspect of the invention is the recognition that siRNAs of differing effectiveness in terms of the degree to which they inhibit expression of the target transcript may be useful for different purposes. For example, it may be of interest to reduce expression of the target transcript by a factor of 2 but not eliminate its expression entirely. Such a reduction in expression may, for example, mimic the effects of a recessive mutation and would be useful for studying phenotypes resulting from such mutations. In the case of essential genes, eliminating expression entirely would be lethal, making it difficult or impossible to study the function of the gene. In addition, in the case of genes that have multiple activities not all of which are essential, it may be difficult to discover the nonessential functions since cells die in the absence of the essential function(s). However, reducing rather than eliminating expression allows the identification of nonessential functions and facilitates analysis of the essential function(s). In general, the ability to predict siRNAs capable of causing intermediate knockdown phenotypes would be an invaluable tool in the setting of analysis of any gene whose function (or lethality) varies with the level of gene expression. Thus the rules find use not only in selecting and designing siRNAs having maximum efficacy but also in selecting and designing siRNAs having a range of different efficacies. In general, the ability to predict siRNAs capable of causing intermediate knockdown phenotypes would also be an invaluable tool in the setting of analysis of any gene whose function (or lethality) varies with the level of gene expression.
The invention therefore encompasses the use of the rules to generate libraries of siRNAs having a range of efficacies in terms of their ability to inhibit expression of a target transcript. According to certain embodiments of the invention a library includes at least 2 siRNAs, of which one reduces expression to less than 50% of the level of expression that exists in the absence of the siRNA and one of which reduces expression to more than 50% of the level of expression that exists in the absence of the siRNA. According to certain embodiments of the invention a library includes at least 3 siRNAs, of which one reduces expression to less than 75% of the level of expression that exists in the absence of the siRNA, one of which reduces expression to less than 50% of the level of expression that exists in the absence of the siRNA, and one of which inhibits expression to less than 25% of the level of expression that exists in the absence of the siRNA. According to certain embodiments of the invention a library includes at least 5 siRNAs that inhibit expression to differing degrees. According to certain embodiments of the invention a library includes at least 10 siRNAs that inhibit expression to differing degrees. According to certain embodiments of the invention two siRNAs (siRNA1 and siRNA2) inhibit expression “to differing degrees” if the level of expression in the presence of siRNA1 and the level of expression in the presence of siRNA2 differ by at least 5% of the level of expression that exists in the absence of either siRNA. For example, if the level of expression in the absence of either siRNA is given by X, the level of expression in the presence of siRNA1 is given by Y, and the level of expression in the presence of siRNA2 is given by Z, then siRNA1 and siRNA2 inhibit expression to varying degrees if the absolute value of (Y - Z) is greater than or equal to 0.5 * (X) where the asterisk indicates multiplication. Any of the libraries may include an siRNA that has minimal effects on expression, e.g., reduces expression by less than approximately 3%, less than approximately 2%, less than approximately 1%, or not at all (i.e., no detectable difference in expression is caused by the siRNA). Any of the libraries may include an siRNA that has maximum effects on expression, e.g., reduces expression by more than approximately 90%, 95%, 98%, 99%, or more, e.g., so that expression is undetectable.

The operation of the rules may be illustrated in reference to Figure 13 (SEQ ID NO: 31), which presents the 2031 nucleotide cDNA sequence 200 of the
vaccinia virus genome (as presented by the program DNA Strider™). The sequence as shown includes restriction sites. The identity and location of these sites is not significant for purposes of the present invention. According to one of many possible approaches, Rule I(B) (the constrained overhang method) is used to identify all potential siRNA sequences in which both strands have 3' pyrimidine overhangs. The 15 candidate siRNA sequences 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, and 275 so identified are enclosed in boxes with staggered ends.

Applying the Complexity Rules, the next step is to determine the nucleotide composition of each candidate, which may be done by counting the number of A, G, C, and T nucleotides on the antisense (lower) strand. These numbers are indicated on Figure 13 adjacent to the row in which the sequence appears. Application of Rule II(A) (the Composition Rule) eliminates sequences 205, 210, 215, 225, 235, 240, 250, 260, 265, 270, and 275 from consideration because either (i) the percentage of GC pairs (G/C content) is greater than approximately 70% or less than approximately 30% or (ii) the A:G:C:T ratio deviates too far from 1:1:1:1, leaving sequences 220, 230, 245, and 255 as candidates.

[00105] Next, the Cluster Avoidance Rules in II(B) may be applied. None of the remaining candidates includes a stretch of four or more identical nucleotides. However, 220 and 255 each contains a triplet and at least one doublet whereas the other two candidates each contain three doublets. Application of Rule II(B)(2) indicates that sequences 230 and 245 are preferred to sequences 220 and 255 contains a triplet and a doublet whereas the other two contain only doublets. Thus 230 and 245 are recommended as preferred siRNAs sequences. The choice between these two sequences could be further refined by considering, e.g., (i) the positions of the doublets relative to the middle of the siRNA; (ii) the position of the siRNA sequence relative to the 5' and 3' ends of the transcript; (iii) the position of the siRNA sequence relative to any exon/intron boundaries that might exist. However, as mentioned above, it is not necessary to identify a "best" sequence or to exhaustively apply all the rules.

[00106] The discussion above is exemplary and is not intended to be limiting. A wide variety of ranking, scoring, and comparison methods may be employed in the utilization of the above rules.
[00107] Although the rules above have been described in reference to an siRNA comprised of two separate, complementary strands as depicted in Figure 5, it will be appreciated that they are equally applicable to alternative structures, e.g., those involving hairpin loops. Additional rules, e.g., rules to minimize the likelihood of base pairing between loop sequences and sequences in the stem portion of a hairpin may be added. For example, the loop nucleotides may be selected so as to avoid complementarity with the stems.

[00108] The rules described above were designed primarily for selection of siRNAs that inhibit expression by reducing target transcript levels. However, there is evidence to suggest that certain siRNAs that bind to the 3' UTR of a template transcript may inhibit expression of a protein encoded by the template transcript by a mechanism related to but distinct from classic RNA interference, e.g., by reducing translation of the transcript rather than decreasing its stability. Specifically, as shown in Figure 6, the DICER enzyme that generates siRNAs in the Drosophila system discussed above and also in a variety of organisms, is known to also be able to process a small, temporal RNA substrate into an inhibitory agent referred to as a microRNA that, when bound within the 3’ UTR of a target transcript, blocks translation of the transcript (see Figure 6; Grishok, A., et al., Cell 106, 23-24, 2001; Hutvagner, G., et al., Science, 293, 834-838, 2001; Ketting, R., et al., Genes Dev., 15, 2654-2659. Similar RNAs have been identified in a number of organisms including mammals, suggesting that this mechanism of post-transcriptional gene silencing may be widespread (Lagos-Quintana, M. et al., Science, 294, 853-858, 2001; Pasquinelli, A., Trends in Genetics, 18(4), 171-173, 2002, and references in the foregoing two articles). MicroRNAs have been shown to block translation of target transcripts containing target sites in mammalian cells (Zeng, Y., et al., Molecular Cell, 9, 1-20, 2002).

[00109] Rules appropriate for selecting preferred siRNAs that function in a similar fashion to naturally occurring stRNAs or miRNAs may differ from those appropriate for siRNAs that function via the classical siRNA mechanism. For example, as noted above, siRNAs that bind within the 3’ UTR and inhibit translation may tolerate a larger number of mismatches in the siRNA/template duplex, and particularly may tolerate mismatches within the central region of the duplex. In fact, there is evidence
that some mismatches may be desirable or required as naturally occurring siRNAs frequently exhibit such mismatches. Notwithstanding the possibility that such differences exist, the present invention encompasses the application of the rules to selection of siRNAs that function by inhibiting translation in addition to those that function by causing decreased transcript stability. In particular, the present invention encompasses the application of the rules to selection of miRNAs that include regions of identity (or complementarity) to a target transcript wherein the regions of identity (or complementarity) are chosen in accordance with certain of the rules such as the Complexity Rules, base pair optimization rules, and Suboptimal Element Positioning Rules.

Systematic Testing of siRNAs

[00110] The present invention provides methods for systematically testing siRNAs on a large scale. These methods have a number of applications. For example, systematic testing on a large scale is useful for further development, refinement, and testing of the inventive methods described above. In addition, systematic testing of siRNAs allows the identification of “hypsensitivs sites” within a target transcript. In the context of RNAi, a hypersensitive site is a region within an mRNA that is particularly susceptible to the effects of siRNA mediated gene silencing. The efficacy of any particular siRNA in reducing gene expression may be defined in terms of the fold reduction in mRNA level or protein level that results following delivery of the siRNA to the cell, organism, etc., that contains the siRNA and in any of a variety of other ways. For purposes of description it will be assumed that the siRNA is to be delivered to a cell, either exogenously or by expression of a vector that directs synthesis of the siRNA within the cells.

[00111] According to the methods, a plurality of siRNAs (an siRNA set) are selected and/or designed. Preferably the siRNA set includes siRNAs that are located across the length of the target transcript, e.g., siRNAs corresponding to target portions that span at least 50% to 75% of the target transcript. By “span at least X% of the transcript” is meant that the distance between the middle nucleotide (or the more 3’ of the middle two nucleotides if the target portion consists of an even number of nucleotides) of the most 5’ target portion of the transcript and the middle nucleotide
(or the more 3' of the middle two nucleotides if the target portion consists of an even number of nucleotides) of the most 3' target portion of the transcript is at least X% of the length of the transcript. For example, if the target transcript is 1000 nucleotides in length, then two siRNAs corresponding to target portions located at nucleotides 200-218 and 700-718 would span 50% of the transcript. (When considering whether a set of nucleotides spans X% of a transcript, deviations of one, two, or several nucleotides are acceptable.) According to certain embodiments of the invention at least 10 siRNAs are tested, preferably relatively evenly distributed along the region of the target transcript that is spanned. For example, if the region of the target transcript that is spanned is divided in half, then preferably between 30% and 70% of the target portions corresponding to siRNAs to be tested are located in each half of the region.

According to other embodiments of the invention at least 20, at least 30, at least 50, or greater than 50 siRNAs are tested, preferably relatively evenly distributed along the region of the target transcript that is spanned. According to certain embodiments of the invention, if Y represents the target transcript length, then at least Y/19, at least 2Y/19, at least 3Y/19, or at least 5Y/19 transcripts are tested, preferably distributed across the transcript so that all or most of the transcript is "covered" by the siRNA sequences.

[00112] According to certain embodiments of the invention rather than being located across the length of the target transcript, the siRNAs span only a portion of the target transcript, which may be referred to as a window. For example, according to certain embodiments of the invention the siRNAs span a window at least 200 nucleotides in length. According to certain other embodiments of the invention the siRNAs span a window at least 300 nucleotides, at least 400 nucleotides, at least 500 nucleotides, at least 600 nucleotides, at least 700 nucleotides, at least 800 nucleotides, at least 900 nucleotides, at least 1000 nucleotides in length, etc. Regardless of the exact length, the goal is to systematically explore siRNAs located across the length of the window rather than selectively focusing on particular regions thereof. The inventive approach is therefore distinguished from ad hoc methods in which, for example, the region around particular siRNAs is explored. Instead, by tiling across an entire window (though not necessarily in one nucleotide increments) the inventive method explores the window in an unbiased manner.
[00113] In one approach, siRNAs targeted to different window portions of multiple target transcripts are tested. For example, siRNAs to a plurality of target transcripts (e.g. corresponding to 8 to 10 genes) are synthesized. For each transcript, a window region of a particular length is chosen as described above. For some of the transcripts the window will be in the 5’ region of the transcript while for others the window will be toward the 3’ end of the coding region. Other window may either include the 3’end of the test gene extending into the 3’UTR or be located entirely within the 3’UTR of the transcript. siRNAs will be synthesized by “tiling” across the entire window region of the target transcript (though not necessarily one nucleotide at a time). Table 1 presents a representative list of a set of genes that may be tested according to the inventive approach. These genes include genes having a range of different lengths and subcellular localizations and include members of a variety of gene families. Note that the designations in the table are intended to refer to genes and target transcripts rather than their encoded proteins although the nomenclature employed may apply to the protein. The accession numbers correspond to the cDNAs for the genes.

<table>
<thead>
<tr>
<th>Gene/Accession No.</th>
<th>Localization</th>
<th>Splicing Variants</th>
<th>mRNA Distribution</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 c-myc/</td>
<td>cyto/nuclear</td>
<td>yes</td>
<td>likely polysomes</td>
<td>2121 bp</td>
</tr>
<tr>
<td>V00568</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 RAPTOR/</td>
<td>cyto/membrane</td>
<td>none known</td>
<td>free ribosomes/polysomes</td>
<td>6431 bp</td>
</tr>
<tr>
<td>NM 020761</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 SMN1/</td>
<td>cyto/nuclear</td>
<td>yes</td>
<td>unknown</td>
<td>1621 bp</td>
</tr>
<tr>
<td>NM 000344</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 HPRT/</td>
<td>cytoplasmic</td>
<td>none known</td>
<td>free ribosomes/polysomes</td>
<td>1331 bp</td>
</tr>
<tr>
<td>NM 000194,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M31642,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J00205,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V00530</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 hTERT/</td>
<td>cyto/nuclear</td>
<td>yes</td>
<td>unknown</td>
<td>4015 bp</td>
</tr>
<tr>
<td>NM 003219</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 mTOR/</td>
<td>cyto/membrane</td>
<td>none known</td>
<td>free ribosomes/polysomes</td>
<td>8587 bp</td>
</tr>
<tr>
<td>NM 020009</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 lamin A/X03444</td>
<td>nuclear</td>
<td>yes</td>
<td>unknown</td>
<td>2404 bp</td>
</tr>
<tr>
<td>8 TK1/BC007872</td>
<td>cytoplasmic</td>
<td>none known</td>
<td>polysomes</td>
<td>1430 bp</td>
</tr>
<tr>
<td>9 p53/NM_000546</td>
<td>cyto/nuclear</td>
<td>none known</td>
<td>unknown</td>
<td>2629 bp</td>
</tr>
</tbody>
</table>
According to the inventive approach, each siRNA targeted to a particular transcript has a different sequence specifying endonucleolytic attack, thus creating a plurality of siRNAs with different specificities. Each of these siRNAs is most comparable to its nearest neighbors and will be constrained by similar secondary structures in the mRNA (if such constraints exist). This method has the advantage that each siRNA will be related to its nearest neighbors and therefore will share many of the parameters of the optimal siRNA and will likely deviate from its nearest neighbors in only one or two parameters for each nucleotide tiled. The resulting siRNAs can thus be compared better and differences may be most readily discernible as opposed to comparing siRNAs that do not share significant sequence overlap.

Preferably the siRNAs include a 19 bp region that is perfectly complementary to a portion of the target transcript. Such a portion of a target transcript is referred to herein as an “inhibitory region”, a “target region”, etc. The siRNAs may include one or more 3’ overhangs. While it is not required that such overhangs be included, preferably all siRNAs tested in relation to any particular target transcript are designed consistently with regard to the presence or absence of overhangs. According to certain embodiments of the invention overhangs are added to a 19 bp duplex chosen to be complementary to a target region, without regard for the particular sequences present in the target transcript immediately 5’ of the target region. For example, thymidine (TT) or deoxythymidine overhangs (dTdT) may be added to the 3’ ends of the siRNA regardless of whether As appear in the target transcript in the positions 5’ of the target region. Alternatively, overhangs may be selected based on nucleotides immediately 5’ and 3’ of the N19 target sequence, etc. Each siRNAs is delivered to a population of cells, and the ability of the siRNA to reduce expression of the target transcript is evaluated. Preferably cells to which the various siRNAs are delivered are maintained under similar or identical environmental conditions and methods used for measuring expression of the target transcript are consistent. Although any target transcript can be tested according to the approach described herein, in certain embodiments of the invention it is preferred to test endogenous transcripts since results obtained using endogenous transcripts rather than exogenous transcripts may be more relevant to a typical experimental or therapeutic context in which siRNAs will be employed.
Any of a wide variety of methods may be used to determine the extent to which an siRNA reduces gene expression. Such methods include measurement of the level of the target transcript (e.g., by reverse transcription (RT)-PCR which may be followed by densitometry of ethidium stained gels, microarray analysis, Northern blot, etc.) and measurement of the level of a polypeptide encoded by the target transcript (e.g., by FACS, protein microarray, mass spectrometry, Western blot, immunoassay, immunofluorescence, etc.) Use of real-time, quantitative PCR after reverse transcription (e.g., Taqman™ assays, Roche Molecular Biosystems) may be preferred.

Depending on the identity of the target transcript and the biological setting in which it is expressed, various other methods of assessing siRNA efficacy may be employed. For example, in the case of an siRNA that targets a viral transcript, the ability of the candidate siRNA(s) to reduce inhibit and/or suppress one or more aspects or features of the viral life cycle such as viral replication, pathogenicity, and/or infectivity is then assessed. For example, cell lysis, syncytia formation, production of viral particles, etc., can be assessed either directly or indirectly using methods well known in the art. Cells to which inventive siRNA compositions have been delivered (test cells) may be compared with similar or comparable cells that have not received the inventive composition (control cells). The susceptibility of the test cells to viral infection can be compared with the susceptibility of control cells to infection. Production of viral protein(s) and/or progeny virus may be compared in the test cells relative to the control cells. Other indicia of viral infectivity, replication, pathogenicity, etc., can be similarly compared. Generally, test cells and control cells would be from the same species and of similar or identical cell type (e.g., T cell, macrophage, dendritic cell, etc.). For example, cells from the same cell line could be compared. When the test cell is a primary cell, typically the control cell would also be a primary cell.

While such methods of determining efficacy of a candidate siRNA may be preferred in certain situations, generally for purpose of systematically testing a large number of candidate siRNAs, more straightforward methods of evaluating target transcript levels will be appropriate. For example, according to one possible approach, a set of siRNAs targeted to a transcript of interest is synthesized.
Preferably the set spans most or all of the transcript, e.g., by tiling across the
transcript as described above, whereby each nucleotide serves as the first nucleotide
in a different N19 target portion, thus shifting the target site by one nucleotide at a
time. The siRNA sequences are selected according to the constrained overhang
approach, in which the sequence of the 3’ overhangs is determined by the sequence of
the target transcript. Alternatively or in addition to the foregoing, a set of siRNA
sequences can be selected according to the tiling approach, and 3’ overhangs
consisting of TT, dTdT, etc., are added to each strand of the N19 core. It is not
necessary to test every potential siRNA, and it may be preferable to tile across by
selecting every other nucleotide, every third nucleotide, every fifth nucleotide, etc., as
the first nucleotide in a different N19 target region in order to reduce the number of
siRNAs to be tested, i.e., shifting the siRNA target site by two, three, five nucleotides
at a time, etc.

[00119] Sense and antisense primers for performing RT-PCR, e.g., quantitative
PCR, real-time PCR, etc., are synthesized for the target transcript and also for a
transcript that will serve as an internal and loading control. Preferably the transcript
is one whose levels are not affected by levels of the target transcript and that lacks
regions of identity or significant homology to the target transcript. Either or both of
the target and control transcripts may be endogenous transcripts or may be
heterologous transcripts, e.g., transcripts derived from genes introduced into the cell
by transfection, infection, etc. Cells are plated, e.g., in multiwell dishes, and
candidate siRNAs are transfected into populations of cells, e.g., using a lipophilic
transfection reagent such as Oligofectamine™ (See Example 1). The experiment can
be performed using a single cell type, e.g., a well characterized, easily transfectable
cell type such as HeLa cells. siRNA transfection conditions are very well established
for HeLa cells growing in early log phase in 6 well dishes. For example, transfection
of 100 pmoles of GFP siRNA complexed with 3 μL oligofectamine in 100 μl DMEM
and added to 1x10^5 logarithmically growing cells washed and resuspended in 900 μL
serum-free DMEM per well of a 6 well tissue culture dish, will approach 100%
transfection efficiency. These conditions can readily be modified to 24, 48 or 96 well
dishes. The experiment may also be performed in different cell types, which may
reveal cell-type specific differences in silencing, e.g., due to different regulatory
proteins. However, it is expected that both the rules and the results obtained when testing specific siRNAs in HeLa cells will be applicable in a wide variety of organisms, and the rules are applicable to siRNAs targeted to transcripts found in any organism as well as to artificial (non-naturally occurring) transcripts.

[00120] The effectiveness of an siRNA in inhibiting expression of the target transcript may be measured at either the RNA or protein level in different embodiments of the method. If RNA is to be measured, RNA is harvested from the cells after an appropriate time period, e.g., 18-72 hours after transfection of the siRNAs. According to one approach, the levels of both target transcript and control transcript are measured using a Taqman™ assay according to the directions of the manufacturer, and levels of target transcript are normalized based on the level of the control transcript. The degree to which each siRNA reduces the target transcript level is determined. Measurement of absolute RNA levels is not required and therefore standardizing the plasmid control is not necessary, though this step may be added.

[00121] As an alternative to the quantification of RNA, the level of a polypeptide encoded by the transcript may be measured, e.g., using FACS with an antibody to the polypeptide. In general, proteins with a short half life and membrane localization may be preferred for such detection. If the transcript encodes a readily detectable marker such as GFP, luciferase, etc., detection of this marker may be used to quantify the polypeptide. The control transcript or polypeptide may be, for example, a transcript or polypeptide whose levels generally do not vary significantly over a wide variety of physiological states of cells. Thus the level of such transcripts or polypeptides would not be expected to vary as a result of inhibition of the target transcript. Transcripts and polypeptides encoded by “housekeeping genes”, e.g., β-actin are useful in this regard. In general, the experiments can be performed using a high throughput format, automated plate handling devices, robotic liquid handling machines, etc., which are well known in the art. See, e.g., Web sites having URLs www.zymark.com (Zymark Corporation, Hopkinton, MA) and www.robsci.com (Robbins Scientific Corporation, Sunnyvale, CA) describing laboratory automation equipment, etc., that may be used in conjunction with the present invention.

[00122] The effectiveness of a given siRNA may be expressed in terms of the extent to which levels of the target transcript are reduced in the presence of that
siRNA relative to the level of the target transcript in the absence of the siRNA, e.g., as a fraction of the transcript level that exists in the absence of the siRNA, % reduction, fold-reduction, % transcript in presence of siRNA vs in its absence, etc. The siRNAs can then be ranked according to the degree to which they reduce expression of the target transcript. Once the effectiveness of the siRNAs is determined, the siRNA sequences and the positions of the siRNA sense sequences within the target transcript may be analyzed in relation to the rules described above in order to refine and further develop the rules. For example, siRNAs that are predicted to be relatively effective in reducing target transcript levels based on their sequences and positions but that in fact fail to do so, and siRNAs that are predicted to be relatively ineffective in reducing target transcript levels based on their sequences and positions but that in fact do reduce target transcript levels substantially may be particularly informative. According to certain embodiments of the invention the contribution of each factor independently as well as its covariance with other factors is determined. Thus, a multi-dimensional matrix is established with each axis corresponding to an optimal sequence characteristic.

[00123] One aspect of the invention is the inventors’ recognition that the systematic approach to testing siRNAs for their ability to inhibit expression of a first target transcript as described above may result in identification of siRNAs that act as miRNAs and inhibit a second target transcript distinct from the first target transcript. For example, an siRNA that is identical (or complementary) to a first transcript may include two regions of identity (or complementarity) to a second transcript, separated by a region of nonidentity (or mismatch), allowing the siRNA to function as an miRNA and inhibit translation of the second transcript. Such an effect may be detected, for example, by observing an unexpected phenotype that would not typically be attributed to knockdown or knockout of the target (either because the knockdown/knockout phenotype is known or because measurements show that expression of the target is not inhibited). In such a case the gene inhibited via the miRNA pathway may be identified by doing database searches to locate genes with two regions of identity (or complementarity) to the siRNA separated by short regions of nonidentity (or mismatch). (Note that the terms identity or complementarity are both used here because the appropriate term will differ depending upon whether one
is considering the sense or antisense strand of the siRNA. The terms nonidentity or mismatch are both used for the same reason.

[00124] According to certain embodiments of the invention miRNAs having two regions of identity (or complementarity) to any known gene, transcript, etc., wherein the regions are between 5 and 10 nucleotides in length, e.g., 5, 6, 7, 8, 9, or 10 nucleotides in length, and one or two regions of nonidentity (or mismatch) wherein the regions of nonidentity (or mismatch) are between 1 and 6 nucleotides in length, e.g., 1, 2, 3, 4, 5, or 6 are identified. According to certain embodiments of the invention, miRNAs that include two regions of identity (or complementarity), each 6 or 7 nucleotides in length, separated by a region of nonidentity (or mismatch) of 2 to 4 nucleotides in length, with any known gene, transcript, etc., are identified. The invention encompasses miRNAs that target a second transcript identified according to the systematic approach for identifying siRNAs that target a first transcript described herein.

[00125] According to certain embodiments of the invention the computer program analyzes the results and uses them to refine the rules. The experiment described above allows evaluation of the effects of nucleotide composition and complexity, position within target transcript, influence of identity of nucleotides in overhang, Tm of the siRNA, etc., to be determined. In order to facilitate use of the results of systematic testing to refine the rules and/or to derive additional rules, programs that permit immediate input of the level of silencing based upon the results of RT (reverse transcription)-PCR, of Real-Time PCR (after Reverse Transcription), or from the movement of cells out of specified gates in FACS analysis may be employed. For example, the inventive computer program embodying the rules may input results directly from CellQuest™ (if FACS is the method of testing) or from the ABI-Prism™ software (if Real-Time PCR is the method of testing) or from comparable software if automated RT-PCR is the method of testing.

[00126] As mentioned above, information from the systematic testing of siRNAs may allow the derivation of additional rules. For example, while the rules above may incorporate factors that reflect the Tm of the siRNA, they do not explicitly utilize Tm in the selection process. However, information from the systematic testing of siRNAs may allow the derivation of a rule involving Tm. The Tm is defined as the
temperature at which 50% of a nucleic acid and its perfect complement are in duplex in solution while the Td, defined as the temperature at a particular salt concentration, and total strand concentration at which 50% of an oligonucleotide and its perfect filter-bound complement are in duplex, relates to situations in which one molecule is immobilized on a filter.

[00127] One common way to determine the actual Tm is to use a thermostatted cell in a UV spectrophotometer. If temperature is plotted vs. absorbance, an S-shaped curve with two plateaus will be observed. The absorbance reading halfway between the plateaus corresponds to Tm. The simplest equation for Td is the Wallace rule: Td = 2(A+T) + 4(G+C) Wallace, R.B.; Shaffer, J.; Murphy, R.F.; Bonner, J.; Hirose, T.; Itakura, K., *Nucleic Acids Res.* 6, 3543 (1979). The nature of the immobilized target strand provides a net decrease in the Tm observed relative to the value when both target and probe are free in solution. The magnitude of the decrease is approximately 7-8°C. Another useful equation for DNA which is valid for sequences longer than 50 nucleotides from pH 5 to 9 within appropriate values for concentration of monovalent cations, is: Tm = 81.5 + 16.6 log M + 41(XG+XC) - 500/L - 0.62F, where M is the molar concentration of monovalent cations, XG and XC are the mole fractions of G and C in the sequence, L is the length of the shortest strand in the duplex, and F is the molar concentration of formamide (Howley, P.M; Israel, M.F.; Law, M-F.; Martin, M.A., *J. Biol. Chem.* 254, 4876). Similar equations for RNA are: Tm = 79.8 + 18.5 log M + 58.4 (XG+XC) + 11.8(XG+XC)2 - 820/L - 0.35F and for DNA-RNA hybrids: Tm = 79.8 + 18.5 log M + 58.4 (XG+XC) + 11.8(XG+XC)2 - 820/L - 0.50F. These equations are derived for immobilized target hybrids.

[00128] Several studies have derived accurate equations for Tm using thermodynamic basis sets for nearest neighbor interactions. The equation for DNA and RNA is: Tm = (1000ΔH)/A + ΔS + Rln(Ct/4) - 273.15 + 16.6 ln[Na⁺], where ΔH (Kcal/mol) is the sum of the nearest neighbor enthalpy changes for hybrids, A (eu) is a constant containing corrections for helix initiation, ΔS (eu) is the sum of the nearest neighbor entropy changes, R is the Gas Constant (1.987 cal deg⁻¹ mol⁻¹) and Ct is the total molar concentration of strands. If the strand is self complementary, Ct/4 is replaced by Ct. Values for thermodynamic parameters are available in the literature. For DNA see Breslauer, et al., *Proc. Natl. Acad. Sci. USA* 83, 3746-3750, 1986. For

[00129] Since the sequence of any given siRNA contains an N19 core that corresponds in sequence to a portion of the target transcript, identification of siRNAs that are particularly effective in reducing transcript expression also identifies portion of the target transcript that are preferred target portions for the design of siRNAs. Regions of the target transcript that correspond to particularly effective siRNAs may be referred to as hypersensitive sites. While not wishing to be bound by any theory, such a hypersensitive site may reflect a region of greater RNA accessibility relative to other regions of the transcript or may reflect the a property of the particular sequence in that region, such as its complexity, Tm, etc.

[00130] A hypersensitive site may be defined in any of a number of ways and may be of any length. For example, a hypersensitive site may be a region shorter in length than an siRNA, such that if the region is present within a target portion of the transcript then siRNA sequences encompassing that site display increased effectiveness in reducing target transcript levels and/or levels of a polypeptide encoded by the transcript. Alternatively, a hypersensitive site may be a region equal in length to an siRNA (e.g. 19 nucleotides or 23 nucleotides if overhangs are included) such that the siRNA whose sequence corresponds to that region displays increased effectiveness in reducing target transcript levels and/or levels of a polypeptide encoded by the transcript. Alternatively, a hypersensitive site or region may be one greater than 23 nt in length such that siRNAs having a core sequence (e.g., an N19 core sequence) that is found within the site or region display increased effectiveness in reducing target transcript levels and/or levels of a polypeptide encoded by the transcript. In one embodiment of the invention a hypersensitive site is defined as a site between approximately 19 and 23 nucleotides in length such that an
siRNA containing a core (e.g., an N19 core) whose sequence corresponds to that site is more effective in reducing transcript levels and/or levels of a polypeptide encoded by the transcript than any siRNA containing a core (e.g., an N19 core) corresponding to a different target portion of the transcript. The site may be somewhat shorter in length than 19 nt, e.g., 17 or 18 nt, or somewhat longer in length than 23 nt, e.g., 24 or 25 nt.

[00131] An siRNA displaying increased effectiveness may be defined in relation to the effectiveness of other siRNAs. For example, assuming that an adequate set of siRNAs are tested, an siRNA displaying increased effectiveness may be defined as being the most effective siRNA, or as being among the 10% to 25% of siRNAs that exhibit the most effectiveness. An adequate set of siRNAs would vary in number depending on the length of the target transcript but would typically be at least 50 siRNAs and/or at least Y/19 siRNAs where Y = length of target transcript.

[00132] The invention thus provides methods for the identification of siRNA hypersensitive sites. The invention additionally encompasses any such sites identified according to the inventive technique. In addition, the invention provides methods for effecting siRNA mediated gene silencing comprising delivering an siRNA targeted to a hypersensitive site to a cell, organism, etc. The methods may be used for any purposes for which gene silencing is useful, including research purposes and therapeutic purposes.

Preparation and Delivery of siRNAs

[00133] Those of ordinary skill in the art will readily appreciate that siRNA agents selected and/or designed according to the methods described herein may be prepared according to any available technique including, but not limited to chemical synthesis, enzymatic or chemical cleavage in vivo or in vitro, or template transcription in vivo or in vitro. As noted above, inventive siRNAs may be delivered as a single RNA strand including self-complementary portions, or as two (or possibly more) strands hybridized to one another. For instance, two separate 21 nt RNA strands may be generated, each of which contains a 19 nt region complementary to the other, and the individual strands may be hybridized together to generate a structure such as that depicted in Figure 5A.
Alternatively, each strand may be generated by transcription from a promoter, either in vitro or in vivo. For instance, a construct may be provided containing two separate transcribable regions, each of which generates a 21 nt transcript containing a 19 nt region complementary with the other. Alternatively, a single construct may be utilized that contains opposing promoters and terminators positioned so that two different transcripts, each of which is at least partly complementary to the other, are generated is indicated in Figure 7.

In another embodiment, an inventive siRNA agent is generated as a single transcript, for example by transcription of a single transcription unit encoding self complementary regions. Figure 8 depicts one such embodiment of the present invention. As indicated, a template is employed that includes first and second complementary regions, and optionally includes a loop region. Such a template may be utilized for in vitro or in vivo transcription, with appropriate selection of promoter (and optionally other regulatory elements). The present invention encompasses gene constructs encoding one or more siRNA strands.

In vitro transcription may be performed using a variety of available systems including the T7, SP6, and T3 promoter/polymerase systems (e.g., those available commercially from Promega, Clontech, New England Biolabs, etc.). As will be appreciated by one of ordinary skill in the art, use of the T7 or T3 promoters typically requires an siRNA sequence having two G residues at the 5’ end while use of the SP6 promoter typically requires an siRNA sequence having a GA sequence at its 5’ end. Vectors including the T7, SP6, or T3 promoter are well known in the art and can readily be modified to direct transcription of siRNAs. When siRNAs are synthesized in vitro they may be allowed to hybridize before transfection or delivery to a subject. It is to be understood that inventive siRNA compositions need not consist entirely of double-stranded (hybridized) molecules. For example, siRNA compositions may include a small proportion of single-stranded RNA. This may occur, for example, as a result of the equilibrium between hybridized and unhybridized molecules, because of unequal ratios of sense and antisense RNA strands, because of transcriptional termination prior to synthesis of both portions of a self-complementary RNA, etc. Generally, preferred compositions comprise at least approximately 80% double-stranded RNA, at least approximately 90% double-
stranded RNA, at least approximately 95% double-stranded RNA, or even at least approximately 99-100% double-stranded RNA.

[00137] Those of ordinary skill in the art will appreciate that, where inventive siRNA agents are to be generated in vivo, it is generally preferable that they be produced via transcription of one or more transcription units. The primary transcript may optionally be processed (e.g., by one or more cellular enzymes) in order to generate the final agent that accomplishes gene inhibition. It will further be appreciated that appropriate promoter and/or regulatory elements can readily be selected to allow expression of the relevant transcription units in mammalian cells. In some embodiments of the invention, it may be desirable to utilize a regulatable promoter; in other embodiments, constitutive expression may be desired.

[00138] In certain preferred embodiments of the invention, the promoter utilized to direct in vivo expression of one or more siRNA transcription units is a promoter for RNA polymerase III (Pol III). Pol III directs synthesis of small transcripts that terminate within a stretch of 4-5 T residues. Certain Pol III promoters such as the U6 or H1 promoters do not require cis-acting regulatory elements (other than the first transcribed nucleotide) within the transcribed region and thus are preferred according to certain embodiments of the invention since they readily permit the selection of desired siRNA sequences. In the case of naturally occurring U6 promoters the first transcribed nucleotide is guanosine, while in the case of naturally occurring H1 promoters the first transcribed nucleotide is adenine. (See, e.g., Yu, J., et al., Proc. Natl. Acad. Sci., 99(9), 6047-6052 (2002); Sui, G., et al., Proc. Natl. Acad. Sci., 99(8), 5515-5520 (2002); paddison, P., et al., Genes and Dev., 16, 948-958 (2002); Brummelkamp, T., et al., Science, 296, 550-553 (2002); Miyagashi, M. and Taira, K., Nat. Biotech., 20, 497-500 (2002); Paul, C., et al., Nat. Biotech., 20, 505-508 (2002); Tuschl, T., et al., Nat. Biotech., 20, 446-448 (2002). Thus in certain embodiments of the invention, e.g., where transcription is driven by a U6 promoter, the 5'-nucleotide of preferred siRNA sequences is G. In certain other embodiments of the invention, e.g., where transcription is driven by an H1 promoter, the 5' nucleotide may be A.

[00139] It will be appreciated that in vivo expression of constructs such as those depicted in Figures 7 and 8 can desirably be accomplished by introducing the constructs into a vector, such as, for example, a viral vector, and introducing the
vector into mammalian cells. Any of a variety of vectors may be selected, though in
certain embodiments it may be desirable to select a vector that can deliver the siRNA-
encoding construct(s) to one or more cells that are susceptible to HIV infection. The
present invention encompasses vectors containing siRNA transcription units having
sequences selected and/or designed in accordance with the inventive methods
described herein, as well as cells containing such vectors or otherwise engineered to
contain expressable transcription units encoding one or more siRNA strands. In
certain preferred embodiments of the invention, inventive vectors are gene therapy
vectors appropriate for the delivery of an siRNA-expressing construct to mammalian
cells, preferably domesticated mammal cells, and most preferably human cells.
Preferred gene therapy vectors include, for example, adenovirus vectors, adeno-associated virus vectors, retroviral vectors and lentiviral vectors. In certain instances
(e.g., gene therapy applications for HIV), lentiviruses will often be particularly
preferred, due to their ability to infect resting T cells, dendritic cells, and
macrophages. Lentiviral vectors can also transfer genes to hematopoietic stem cells
with a superior gene transfer efficiency and without affecting the repopulating
capacity of these cells. See, e.g., Mautino and Morgan, AIDS Patient Care STDS
describing the FUGW lentiviral vector; Somia, N., et al. J. Virol. 74(9): 4420-4424,

[00140] It will be appreciated that two separate, complementary siRNA strands can
be transcribed using a single vector containing two promoters, each of which directs
transcription of a single siRNA strand. Alternatively, a vector containing a promoter
that drives transcription of a single siRNA strand comprising two complementary
regions (e.g., a hairpin) may be employed, or a vector containing multiple promoters,
each of which drives transcription of a single siRNA strand comprising two
complementary regions is used. Alternately, the vector may direct transcription of
multiple different siRNAs, either from a single promoter or from multiple promoters.
A variety of configurations are possible. For example, a single promoter may direct
synthesis of a single RNA transcript containing multiple self-complementary regions,
each of which may hybridize to generate a plurality of stem-loop structures. These
structures may be cleaved in vivo, e.g., by DICER, to generate multiple different
siRNAs. It will be appreciated that such transcripts preferably contain a termination signal at the 3' end of the transcript but not between the individual siRNA units. Single RNAs from which multiple siRNAs can be generated need not be produced in vivo but may instead be chemically synthesized or produced using in vitro transcription and provided exogenously.

[00141] Vectors for synthesis of siRNA may include multiple promoters, each of which directs synthesis of a self-complementary RNA that hybridizes to form an siRNA. The multiple siRNAs may all target the same transcript, or they may target different transcripts.

Those of ordinary skill in the art will further appreciate that in vivo expression of siRNAs may allow the production of cells that produce the siRNA over long periods of time (e.g., greater than a few days, preferably at least several months, more preferably at least a year or longer, possibly a lifetime).

[00142] Inventive siRNAs selected and/or designed according to the methods described herein may be introduced into cells by any available method. For instance, siRNAs or vectors encoding them can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms "transformation" and "transfection" are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA or RNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, DEAE-dextran-mediated transfection, lipofection, injection, or electroporation.

[00143] The present invention encompasses any cell manipulated to contain an inventive siRNA selected according to the methods described herein. Preferably, the cell is a mammalian cell. In some embodiments of the invention, the cells are non-human cells within an organism. For example, the present invention encompasses transgenic animals engineered to contain or express inventive siRNAs. Such animals are useful for studying the function and/or activity of inventive siRNAs, and/or of the genes whose expression is affected by their presences. As used herein, a "transgenic animal" is a non-human animal, preferably a mammal, more preferably a rodent such as a rat or mouse, in which one or more of the cells of the animal includes a transgene. Other examples of transgenic animals include non-human primates, sheep, dogs, cows, goats, chickens, amphibians, and the like. A transgene is exogenous DNA or a
rearrangement, e.g., a deletion of endogenous chromosomal DNA, which preferably is integrated into or occurs in the genome of the cells of a transgenic animal. A transgene can direct the expression of an encoded siRNA product in one or more cell types or tissues of the transgenic animal. According to certain embodiments of the invention the transgenic animal is of a variety used as an animal model (e.g., murine or primate) for testing potential therapeutics.

Implementation Systems and Methods

[00144] The methods described above may advantageously be implemented using a computer-based approach, and the present invention therefore includes a computer system for practicing the methods. Figure 15 depicts a representative embodiment of a computer system that may be used for this purpose. Computer system 300 comprises a number of internal components and is also linked to external components. The internal components include processor element 310 interconnected with main memory 320. For example, computer system 310 can be an Intel Pentium™-based processor such as are typically found in modern personal computer systems. The external components include mass storage 330, which can be, e.g., one or more hard disks (typically of 1 GB or greater storage capacity). Additional external components include user interface device 335, which can be a keyboard and a monitor including a display screen, together with pointing device 340, such as a "mouse", or other graphic input device. The interface allows the user to interact with the computer system, e.g., to cause the execution of particular application programs, to enter inputs such as data and instructions, to receive output, etc. The computer system may further include disk drive 350, CD drive 355, and zip disk drive 360 for reading and/or writing information from or to floppy disk, CD, or zip disk respectively. Additional components such as DVD drives, etc., may also be included.

[00145] The computer system is typically connected to one or more network lines or connections 370, which can be part of an Ethernet link to other local computer systems, remote computer systems, or wide area communication networks, such as the Internet. This network link allows computer system 300 to share data and processing tasks with other computer systems and to communicate with remotely located users.
The computer system may also include components such as a display screen, printer, etc., for presenting information, e.g., for displaying preferred siRNA sequences.

[00146] A variety of software components, which are typically stored on mass storage 330, will generally be loaded into memory during operation of the inventive system. These components function in concert to implement the methods described herein. The software components include operating system 400, which manages the operation of computer system 300 and its network connections. This operating system can be, e.g., a Microsoft Windows™ operating system such as Windows 98, Windows 2000, or Windows NT, a Macintosh operating system, a Unix or Linux operating system, an OS/2 or MS/DOS operating system, etc. Software component 410 is intended to embody various languages and functions present on the system to enable execution of application programs that implement the inventive methods. Such components, include, for example, language-specific compilers, interpreters, and the like. Any of a wide variety of programming languages may be used to code the methods of the invention. Such languages include, but are not limited to, C, C++, JAVATM, various languages suitable for development of rule-based expert systems such as are well known in the field of artificial intelligence, etc. According to certain embodiments of the invention the software components include Web browser 420, e.g., Internet Explorer™ or Netscape Navigator™ for interacting with the World Wide Web.

[00147] Software component 430 represents the siRNA selection and design methods of the present invention as embodied in a programming language of choice. Different sets of rules may, but need not be, coded as individual rule modules 440, 450, ranking module 460, etc. In certain embodiments of the invention the software includes online ordering module 470. In an exemplary implementation, to practice the methods of the invention using such a computer system, a user provides the sequence of a target transcript to the computer, which sequence may then be loaded into computer memory 320. The sequence can be directly entered by the user from monitor and keyboard 335, or from other computer systems linked by network connection 370 (see below), or on removable storage media, etc. In certain embodiments of the invention the user may enter the sequence in any of a variety of formats, e.g., as a genomic DNA sequence, as a cDNA sequence, as a mRNA
sequence, etc. The user may be allowed to select an appropriate sequence option
and/or may be prompted to select such an option. According to certain embodiments
of the invention the user enters the sequence by cutting and pasting from a file such as
a DNA Strider™ file or similar file. According to certain embodiments of the
invention the user need not enter the sequence itself but may instead enter information
sufficient to identify the sequence such as, for example, a database accession number,
and the system accesses the database and retrieves the sequence.

[00148] According to certain embodiments of the invention, software component
430 offers various options to the user either before or after receiving the sequence.

For example, the user may have the option of choosing whether candidate siRNAs are
to be selected according to the tiling approach or the constrained overhang approach.
The user may be allowed to select parameters such as GC content, length of duplex,
number of mismatches, etc. The user may be allowed to determine the relative
weights given to different rules or sets of rules or may be allowed to specify that
certain rules will or will not be applied. The user may make these selections using
any of a number of methods, e.g., pull-down or pop-up menus, check boxes, radio
buttons, fill in the blank, etc.

[00149] The description above has generally envisioned a system in which the user
interacts directly with the computer that executes the application program encoding
the methods of the invention. However, according to certain embodiments of the
invention the system is implemented as a client/server system in which users enter
information at a client computer, which information is then transmitted to a server
computer that executes the application program. The client computer system can
comprise any available computer but is typically a personal computer equipped with a
processor, memory, display, keyboard, mouse, storage devices, appropriate interfaces
for these components, and one or more network connections.

[00150] Both the server and client computers are provided with software to support
World Wide Web interactions. Server systems are typically equipped with a Web
server application program, i.e., a Web server engine. A number of such server
engines are available, e.g., Microsoft’s Internet Information Server (IIS) software
running under Microsoft’s NT operating system, Apache HTTP Server software
running under the Unix, Linux, or other operating systems, BEA Systems Weblogic™
server, etc. Client computers are typically equipped with a Web browser, i.e., an application program that facilitates the requesting and displaying of World Wide Web pages.

[00151] According to certain embodiments of the invention, interaction between the client and server computers takes place via Web pages, which allow the user to enter data such as sequence information and to choose between various options as described above. Output such as preferred siRNA sequences can also be transmitted to the user via Web pages although other formats (e.g., e-mail, fax, or non-electronic formats may also be used). Web pages can be coded using methods and languages well known in the art (e.g., HTML, XML, etc.). Scripts to be executed in response to user input may be coded using methods and languages known in the art (e.g., Javascript™, CGI/Perl, etc.) Creation of Web pages may be facilitated by use of an application environment such as Microsoft’s Active Server Pages.

[00152] According to certain embodiments of the invention, software component 420 is implemented in conjunction with a Web based online ordering system, whereby a user may enter a target transcript, select desired siRNA sequences using the rules, and order the siRNA without leaving the Web site. The system may offer the user a variety of options depending upon the extent to which the user wishes to exert control over the selection process. For example, the user may be offered a “standard” option in which the user merely enters the target sequence (e.g., by pasting the sequence into a window, by entering an accession number, etc.) and the system automatically applies the rules and present the user with a list (optionally ranked) of preferred siRNAs, from which the user may select the siRNA(s) he/she wishes to order. A standard format, e.g., incorporating 3’ dTdT overhangs on both strands, may be employed. The siRNAs would be synthesized and shipped to the user.

Alternatively, the user may select a “custom” option, in which he or she would be able to explicitly control the selection process in various ways as described above. Of course any number of options may be offered, designed to satisfy users ranging from those who want to obtain an effective siRNA with minimal effort to those who desire an siRNA meeting particular requirements or specifications.

[00153] The foregoing description is to be understood as being representative only and is not intended to be limiting. Alternative systems and techniques for
implementing the methods of the invention will be apparent to one of skill in the art and are intended to be included within the accompanying claims. In particular, the accompanying claims are intended to include alternative program structures for implementing the methods of this invention that will be readily apparent to one of skill in the art.

[00154] The ability to select or design effective siRNAs based on the sequence of the target transcript without the necessity of testing multiple candidates in order to identify an effective siRNA opens a variety of possibilities for analysis of gene activity and function on a large scale. In particular, the sequences of multiple organisms ranging from retroviruses to humans are now available in public databases such as GenBank. See, e.g., the Web site having URL www.ncbi.nlm.nih.gov/Genomes/index.html. The invention encompasses application of the rules to generate libraries of effective siRNAs targeted to all or a substantial fraction of genes (e.g., greater than 70%, greater than 80%, greater than 90%, greater than 95%, greater than 98%, greater than 99% of the genes or essentially all) expressed in any organism including humans. Such libraries allow a genome-wide analysis of gene function.

[00155] Many genes are members of multigene families that may share various sequence motifs, and one of the challenges of modern biology is to dissect the functions and activities of individual family members. Members of such families include, e.g., G protein coupled receptors, tyrosine kinase receptors and nonreceptor proteins, cytochrome p450 enzymes, ion channels, as well as numerous transcription factor families. Traditional forward and reverse genetic approaches are limited for a variety of reasons. For example, genes that are highly similar in sequence often have overlapping and/or redundant functions. Thus mutation or knockout of one family member may cause little or no phenotypic change, making such functions extremely difficult to identify.

[00156] While knocking out a sufficient number of family members in a single cell or organism using traditional techniques such as homologous recombination might shed light on the role of the remaining family member(s), this method is unwieldy and impractical for families with many members. However, such an approach is feasible using siRNA. Members of a multigene family may be identified, e.g., by searching
publicly available databases. Effective siRNAs targeted against each family member may be designed in accordance with the methods and rules described herein. Collections of siRNAs (either as oligonucleotides or vectors, as appropriate depending upon the experimental setting) designed to inhibit a large proportion of family members may be delivered to a cell or organism. For example, siRNAs targeted to all or a substantial fraction of (e.g., greater than 70%, greater than 80%, greater than 90%, greater than 95%, greater than 98%, greater than 99% of the genes or essentially all) genes belonging to a particular gene family may be delivered. As a consequence, the cell or organism either does not express any genes in the gene family or expresses only one or a few genes in the gene family, allowing identification and/or study of the function and activity of that gene or genes. The invention encompasses application of the rules to generate libraries of effective siRNAs targeted to all or a substantial fraction of genes (e.g., greater than 70%, greater than 80%, greater than 90%, greater than 95%, greater than 98%, greater than 99% of the genes or essentially all) in a multigene family expressed in any organism including humans and also encompasses libraries generated according to the inventive rules and methods.

Therapeutic Applications

[00157] Compositions containing inventive siRNAs of the present invention may be used to inhibit or reduce expression of genes involved in any of a wide variety of diseases and conditions. In such applications, an effective amount of an inventive siRNA composition is delivered to a cell or organism prior to, simultaneously with, or after development of the disease or condition. Preferably, the amount of siRNA is sufficient to reduce or delay one or more symptoms of the disease or condition.

[00158] Inventive siRNA-containing compositions may contain a single siRNA species, targeted to a single site in a single target transcript, or alternatively may contain a plurality of different siRNA species, targeted to one or more sites in one or more target transcripts. In some embodiments of the invention, it will be desirable to utilize compositions containing collections of different siRNA species targeted to different genes. Some embodiments will contain more than one siRNA species targeted to a single transcript. To give but one example, it may be desirable to include at least one siRNA targeted to coding regions of a target transcript and at least one
siRNA targeted to the 3' UTR. This strategy may provide extra assurance that products encoded by the relevant transcript will not be generated because at least one siRNA in the composition will target the transcript for degradation while at least one other inhibits the translation of any transcripts that avoid degradation. As described above, the invention encompasses “therapeutic cocktails”, including approaches in which a single vector directs synthesis of siRNAs that inhibit multiple targets (which may, but need not be, multiple targets that affect the same disease process) and/or directs synthesis of multiple siRNAs that inhibit a single target and/or directs synthesis of RNAs that may be processed to yield a plurality of siRNAs.

It will often be desirable to combine the administration of inventive siRNAs with one or more other therapeutic agents in order to inhibit, reduce, or prevent one or more symptoms or characteristics of the disease or condition. In certain preferred embodiments of the invention, the inventive siRNAs are combined with FDA approved agents. In some embodiments of the invention, it may be desirable to target administration of inventive siRNA compositions to particular cells, e.g., tumor cells, cells infected with an infectious agent, etc.

siRNAs may be delivered using gene therapy. Gene therapy protocols may involve administering an effective amount of a gene therapy vector capable of directing expression of an inhibitory siRNA to a subject either before, substantially contemporaneously, with, or after development of a disease or clinical condition. Another approach that may be used alternatively or in combination with the foregoing is to isolate a population of cells, e.g., stem cells or immune system cells from a subject, optionally expand the cells in tissue culture, and administer a gene therapy vector capable of directing expression of an inhibitory siRNA to the cells in vitro.

The cells may then be returned to the subject. Optionally, cells expressing the siRNA (and thus having reduced levels of the target transcript) can be selected in vitro prior to introducing them into the subject. In some embodiments of the invention a population of cells, which may be cells from a cell line or from an individual who is not the subject, can be used. Methods of isolating stem cells, immune system cells, etc., from a subject and returning them to the subject are well known in the art. Such methods are used, e.g., for bone marrow transplant, peripheral blood stem cell transplant, etc., in patients undergoing chemotherapy.
[00161] In yet another approach, oral gene therapy may be used. For example, US 6,248,720 describes methods and compositions whereby genes under the control of promoters are protectively contained in microparticles and delivered to cells in operative form, thereby achieving noninvasive gene delivery. Following oral administration of the microparticles, the genes are taken up into the epithelial cells, including absorptive intestinal epithelial cells, taken up into gut associated lymphoid tissue, and even transported to cells remote from the mucosal epithelium. As described therein, the microparticles can deliver the genes to sites remote from the mucosal epithelium, i.e. can cross the epithelial barrier and enter into general circulation, thereby transfecting cells at other locations.

Pharmaceutical Formulations

[00162] Compositions containing siRNAs selected and/or designed utilizing the systems and methods described herein may be formulated for delivery by any available route including, but not limited to parenteral (e.g., intravenous), intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, rectal, and vaginal. Preferred routes of delivery include parenteral, transmucosal, rectal, and vaginal. Inventive pharmaceutical compositions typically include an siRNA or other agent(s) such as vectors that will result in production of an siRNA after delivery, in combination with a pharmaceutically acceptable carrier. As used herein the language “pharmaceutically acceptable carrier” includes solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. Supplementary active compounds can also be incorporated into the compositions.

[00163] A pharmaceutical composition is formulated to be compatible with its intended route of administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerine, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such
as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple dose vials made of glass or plastic.

[00164] Pharmaceutical compositions suitable for injectable use typically include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition should be sterile and should be fluid to the extent that easy syringability exists. Preferred pharmaceutical formulations are stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. In general, the relevant carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, it will be preferable to include isotonic agents, for example, sugars, polyalcohols such as manitol, sorbitol, sodium chloride in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

[00165] Sterile injectable solutions can be prepared by incorporating the active compound in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, the preferred methods of preparation are
vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

[00166] Oral compositions generally include an inert diluent or an edible carrier. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules, e.g., gelatin capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash. Pharmacologically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

[00167] Formulations for oral delivery may advantageously incorporate agents to improve stability within the gastrointestinal tract and/or to enhance absorption.

[00168] For administration by inhalation, the inventive siRNA agents are preferably delivered in the form of an aerosol spray from pressured container or dispenser which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

[00169] Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art.

[00169] The compounds can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.
In one embodiment, the active compounds are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. The materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

It is advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier.

Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD$_{50}$ (the dose lethal to 50% of the population) and the ED$_{50}$ (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD$_{50}$/ED$_{50}$. Compounds which exhibit high therapeutic indices are preferred.

While compounds that exhibit toxic side effects can be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

The data obtained from cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds lies preferably within a range of circulating concentrations that include the ED$_{50}$ with little or no toxicity. The dosage can vary within this range depending upon the
dosage form employed and the route of administration utilized. For any compound
used in the method of the invention, the therapeutically effective dose can be
estimated initially from cell culture assays. A dose can be formulated in animal
models to achieve a circulating plasma concentration range that includes the IC$_{50}$ (i.e.,
the concentration of the test compound which achieves a half-maximal inhibition of
symptoms) as determined in cell culture. Such information can be used to more
accurately determine useful doses in humans. Levels in plasma can be measured, for
example, by high performance liquid chromatography.

[00174] A therapeutically effective amount of a pharmaceutical composition
typically ranges from about 0.001 to 30 mg/kg body weight, preferably about 0.01 to
25 mg/kg body weight, more preferably about 0.1 to 20 mg/kg body weight, and even
more preferably about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5
to 6 mg/kg body weight. The pharmaceutical composition can be administered at
various intervals and over different periods of time as required, e.g., one time per
week for between about 1 to 10 weeks, between 2 to 8 weeks, between about 3 to 7
weeks, about 4, 5, or 6 weeks, etc. For certain conditions it may be necessary to
administer the therapeutic composition on an indefinite basis to keep the disease
under control. The skilled artisan will appreciate that certain factors can influence the
dosage and timing required to effectively treat a subject, including but not limited to
the severity of the disease or disorder, previous treatments, the general health and/or
age of the subject, and other diseases present. Generally, treatment of a subject with
an siRNA as described herein, can include a single treatment or, in many cases, can
include a series of treatments.

[00175] Exemplary doses include milligram or microgram amounts of the
inventive siRNA per kilogram of subject or sample weight (e.g., about 1 microgram
per kilogram to about 500 milligrams per kilogram, about 100 micrograms per
kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to
about 50 micrograms per kilogram.) It is furthermore understood that appropriate
doses of an siRNA depend upon the potency of the siRNA, and may optionally be
tailored to the particular recipient, for example, through administration of increasing
doses until a preselected desired response is achieved. It is understood that the
specific dose level for any particular animal subject may depend upon a variety of
factors including the activity of the specific compound employed, the age, body
weight, general health, gender, and diet of the subject, the time of administration, the
route of administration, the rate of excretion, any drug combination, and the degree of
expression or activity to be modulated.

[00176] The nucleic acid molecules of the invention can be inserted into vectors
and used as gene therapy vectors as described herein. Gene therapy vectors can be
delivered to a subject by, for example, intravenous injection, local administration, or
by stereotactic injection (see e.g., Chen et al. (1994) Proc. Natl. Acad. Sci. USA
91:3054-3057). In certain embodiments of the invention gene therapy vectors may be
delivered orally or inhalationally and may be encapsulated or otherwise manipulated
to protect them from degradation, enhance uptake into tissues or cells, etc. The
pharmaceutical preparation of the gene therapy vector can include the gene therapy
vector in an acceptable diluent, or can comprise a slow release matrix in which the
gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery
vector can be produced intact from recombinant cells, e.g., retroviral or lentiviral
vectors, the pharmaceutical preparation can include one or more cells which produce
the gene delivery system.

[00177] Inventive pharmaceutical compositions can be included in a container,
pack, or dispenser together with instructions for administration.

Exemplification

Example 1: Transfection with CD4-siRNA Reduces CD4 Transcript Levels

[00178] This example presents data showing that siRNAs selected in accordance
with the inventive rules described herein effectively inhibit a variety of target
transcripts including both HIV transcripts and host cell transcripts. The following
Materials and Methods were employed in this and following Examples.

[00179] Cell Culture. Magi-CCR5 cells were grown in DMEM containing 200
ug/ml neomycin, 100 ug/ml hygromycin, and 10% heat-inactivated fetal calf serum
(FCS). HeLa-CD4 cells were grown in DMEM containing 200 ug/ml neomycin and
10% heat-inactivated FCS.
Preparation of siRNAs. siRNAs with the following sense and antisense sequences were used (where the presence of a phosphate at the 5’ end of the RNA is indicated with a P):

CD4 (sense): 5’-GAUCAAGAGACUCCUCAGUdGdA-3’ (SEQ ID NO:1)
CD4 (antisense): 5’-ACUGAGGAAGUCUCUUGAUdCdGdG-3’ (SEQ ID NO:2)
p24 (sense): 5’-P.GAUUGACUGAGAGACAGGCU-3’ (SEQ ID NO:3)
p24 (antisense): 5’-P.CCUGUCUCAGUACAUCU-3’ (SEQ ID NO:4)
GFP (sense): 5’-P.GGCUCAGUCGAGAGCCACC-3’ (SEQ ID NO:5)
GFP (antisense): 5’-P.UGCGUCUCCUGAGCUAGGCCU-3’ (SEQ ID NO:6)

HPRT (sense): 5’-P.GGUCAUAUGUGAAACUAGGAA-3’ (SEQ ID NO:7)
HPRT (antisense): 5’-P.CCAGUUUCACUAUAGACACA-3’ (SEQ ID NO:8)

All siRNAs were synthesized by Dharmacon Research (Lafayette, CO) using 2’ACE protection chemistry. The siRNA strands were deprotected according to the manufacturer’s instructions, mixed in equimolar ratios and annealed by heating to 95°C and slowly reducing the temperature by 1°C every 30 s until 35°C and 1°C every min until 5°C.

siRNA transfection. Magi-CCR5 and HeLa cells were trypsinized and plated in 6 cm wells at 1 x 10^5 cells per well for 12-16 h before transfection. Cationic lipid complexes, prepared by incubating 100 pmol of indicated siRNA with 3 ul oligofectamine (Gibco-Invitrogen, Rockville, MD) in 100 ul DMEM (Gibco-Invitrogen) for 20 min, were added to the wells in a final volume of 1 ml. After overnight incubation, cells were washed and used for infection with HIV-1. For transfection of suspension cells, cationic lipid complexes were prepared by 20 min incubation with 100 pmol of indicated siRNA and 0.5 ul oligofectamine (Gibco-Invitrogen) in 50 ul AIM V T-cell medium (Gibco-Invitrogen). Log phase cultures of H9 cells were resuspended at 1 x 10^5 cells per well in 50 ul AIM V media and combined with the cationic lipid complexes in 96 well flat bottom plates. Cells were transfected overnight, washed and resuspended in RPMI medium containing serum and were used for infection of HIV-1.

Flow cytometry. Phycoerythrin (PE)-conjugated αHIV-1 p24 monoclonal antibodies were used for staining (Shankar, P., et al., Blood 94, 3084-3093 (1999)). Data were acquired and analyzed on FACSscalibur with CellQuest software (Becton Dickinson, Franklin Lakes, NJ).
[00192] **Northern Analysis.** Northern blot analysis was performed with 5-10 ug total RNA (RNAEasy, Qiagen, Valencia, CA) and blotting was performed using the Northern Max protocol (Ambion, Austin, TX).

[00193] CD4 probe was PCR amplified from the T4pMV7 plasmid (Madden, P.J., et al., *Cell* 47, 333-348 (1986)) using the following primers:

- **CD4-forward** 5'-'TGAAGTGAGGAGACCAGAAGG-3' (SEQ ID NO:9)
- **CD4-reverse** 5'-CTTGCCCATCTGGAGGCCTTAG-3' (SEQ ID NO:10)

[00196] p24 and nef probes were PCR amplified from the HXB2 plasmid (Ratner, et al., *AIDS Res. Hum. Retroviruses* 3, 57-69 (1987) using the following primers:

- **p24-forward** 5'-CCAGGGCAGATGGCATCAAGGCACTATA-3' (SEQ ID NO:11)
- **p24-reverse** 5'-CCTCTCTGTGAGGCCTGCTGCTTTA-3' (SEQ ID NO:12)
- **nef-forward** 5'-ATAGGGTGGCAGTGGTCAAAAATGAGG-3' (SEQ ID NO:13)
- **nef-reverse** 5'-GTTGGGAGATCTACAGGCTGCTGTAAGG-3' (SEQ ID NO:14)

[00201] β-actin probe (Ambion) was used as an internal standard. PCR products (25-30 ng) were labeled with α-[32P]dATP (DECAprimeII, Ambion), purified by NucAway spin columns (Ambion), heated to 95°C and used as probes in Northern blots.

[00202] **HIV infection.** Magi-CCR5 cells were infected with R5 BAL and X4 NL43 strains of HIV-1 using 10 ng of p24 gag antigen per well. HeLa-CD4 cells were infected with 10-20 ng of p24 antigen per well of X4 HIVIIIB virus. At indicated times, cells were trypsinized and evaluated for HIV-1 p24 expression. H9 cells were infected with viral supernatants from pR7-GFP (Liu, R., et al., *Cell* 86, 367-377 (1996)) transfected 293 T cells at an MOI of 0.1.

[00203] β-gal staining. Magi-CCR5 cells were infected in the presence of DEAE-dextran (20 ug/ml) and then fixed and stained 2 d later (Chackerian, B., et al., *J. Virol.*, 71, 3932-3939 (1997)). Cell counts represent number of blue cells per 10 high power fields. Cell-free p24 antigen was measured by ELISA in supernatants at indicated times (Beckman-Coulter, Brea, CA).

[00204] **Results.** To investigate the feasibility of using siRNA selected in accordance with the rules presented herein to suppress HIV replication, the CD4 molecule, the principal receptor for the virus (Klatzmann et al., *Nature* 312:767, 1984; Madden et al., *Cell* 47:333, 1986) was targeted. Specifically, the HeLa-derived cell line Magi-CCR5, which expresses human CD4, as well as CXCR4, the co-receptor
for T-cell-tropic HIV, and CCR5, the co-receptor for macrophage-tropic virus (Chackerian et al., *J. Virol.* 71:3932, 1997) was used. In addition, Magi-CCR5 cells have an integrated HIV-LTR-β-galactosidase gene that reflects Tat-mediated transactivation and can be used to score for viral entry and early gene expression.

[00205] Magi-CCR5 cells were transfected either with siRNA directed against human CD4 or with control siRNA, and were analyzed for CD4 expression by flow cytometry. As shown in Figure 9A, CD4-siRNA selected in accordance with the rules disclosed herein specifically reduced CD4 expression eight-fold in about 75% of the cells. Northern analysis, shown in Figure 9B, revealed approximately an eight-fold reduction in CD4 mRNA, confirming that the CD4 silencing occurred at the level of mRNA stability. The exposure of the blot used for quantitation is shown in Figure 9F.

Example 2: CD4-siRNA Suppresses HIV Entry and Infection

[00206] To assess the effect of CD4 silencing on viral entry, Magi-CCR5 cells were first transfected with CD4-siRNA. Sixty hours later, the time of maximal gene silencing, the cells were infected with both R5 (BAL) macrophage tropic and X4 (NL43) T cell tropic strains of HIV. Figure 9C shows the level of β-galactosidase activity observed 48 hours post-infection, which is an indicator of viral entry; Figure 9C shows the extent of syncytia formation, an indicator of viral infection. As can be seen, β-galactosidase levels were reduced 4-fold, and syncytia formation was almost abolished. Furthermore, early production of cell free virus, measured by p24 ELISA 48 hours post-infection, was reduced four-fold compared to cells treated with either antisense or control siRNA (see Figure 9E). These findings, when taken together with those reported in Example 1, demonstrate that siRNA selected in accordance with the rules described herein effected silencing of CD4, which specifically inhibited HIV entry into cells and therefore blocked viral replication.

Example 3: p24-siRNA Reduces Levels of p24 and of Viral Transcripts

[00207] The HIV capsid is expressed from the intact viral RNA as a gag polyprotein that is proteolytically cleaved into p24, p17 and p15 polypeptides to form the major structural core of the virus. The p24 polypeptide also functions in
uncoating and packaging virions. To score for the ability of siRNA selected in accordance with the inventive rules to effect silencing of viral genes, the gag gene was targeted because cleavage in this region could inhibit both viral RNA accumulation and production of p24. HeLa cells expressing human CD4 (HeLa-CD4; Madden et al., Cell 47:333, 1986) were transfected with p24-siRNA 24 hours prior to infection with HIVIIIb. Two days after infection, p24-siRNA transfected cells showed a greater than four-fold decrease in viral protein, compared with controls (Figure 10A). Furthermore, silencing of full-length viral mRNA levels (as assessed by Northern blotting for p24 expression) was observed only in the p24-siRNA transfected HeLa-CD4 cells (Figure 10B). Only 14.5% of p24-siRNA-transfected cells expressed p24 antigen above background levels 5 days after infection, while 92% of cells transfected with control siRNA had detectable p24 expression by flow cytometry (see Figure 10C). When production of viral particles was measured by p24 ELISA 5 days after infection, p24 titers in culture supernatants were reduced 25-fold compared to mock transfected cells or cells transfected with control siRNA (see Figure 10D). Northern blots of cellular RNA harvested 5 days after infection showed that after transfection with p24-siRNA, the amount of 9.2 Kd viral transcript containing gag p24 mRNA was reduced ten fold as compared with its level in control transfected cells (see Figure 10E).

[00208] The level of various HIV transcripts in the presence (or absence) of p24-siRNA was also assessed. There are at least ten HIV transcripts (Pavlakis et al. in Annu. Rev. AIDS Res. (Kennedy et al., Eds) Marcel Dekker, New York: pp. 41-63, 1991), and multiple messenger RNAs—including several singly or multiply spliced messages, that are expressed from the integrated HIV provirus at various stages of the viral life cycle (Kim et al., J. Virol. 63:3708, 1989). The full-length HIV transcript is expressed only from the integrated provirus and serves as both the mRNA for the gag-pol genes and the genomic RNA of progeny virus. By contrast, some genes, including Tat, Rev, and Nef, may be expressed from the provirus prior to integration into the host genome (Wu et al., Science 293:1503, 2001).

[00209] Since Nef is the 3'-most gene and is contained in many virally-derived transcripts, a probe against Nef was used to test the effect of siRNA-directed knockdown on different viral transcripts. As shown in Figure 10C, the 4.3 and 2.0 Kb
Nef-containing transcripts were reduced approximately ten-fold, comparably to the knockdown of full-length transcript detected with p24 or Nef gene probes. [00210] Mechanistically, these data suggest at least three possibilities: 1) the siRNA may target the viral genomic RNA directly when the virus first enters the cell, thereby affecting all subsequently-expressed HIV transcripts; 2) the siRNA may inhibit the pre-spliced mRNA in the nucleus; and/or 3) the siRNA may inhibit gag gene expression late in the viral life cycle either by targeting progeny viral genomes directly and/or by inhibiting viral capsid assembly, thereby blocking amplification and re-infection of the virus. Without wishing to be bound by any particular theory, it is proposed that the second possibility is least likely. In particular, it is noted that intronic sequences have not been reported to be good targets for siRNA.

[00211] The effects of p24-siRNA were further characterized by asking whether this siRNA were able to suppress viral production post-integration. Specifically, HeLa-CD4 cells were infected with HIV four days prior to transfection with p24-siRNA. Two days after transfection, the mean fluorescent intensity of p24 expression on a per-cell basis was assessed. As shown in Figure 11, it was found that, in the setting of 80-90% HIV infection, mean fluorescent intensity of p24 expression was reduced 50% as compared with mock or control transfections. These results suggest that siRNA-directed silencing can reduce the steady-state levels of virus even in the setting of an established infection.

[00212] To further eliminate any potential effect of transfected siRNA on parental virus genomes before integration into the host genome, a latently infected T-cell clone (ACH2), which can be induced to produce high levels of infectious HIV-1 by phorbol myristate acetate (PMA) stimulation was assayed. ACH2 cells were grown in RPMI containing 10% heat-inactivated fetal calf serum. ACH2 cells were transfected with p24-siRNA and then induced by treating with PMA at 1 μg/ml. Two days after induction, 70% of control cells expressed p24 compared with 23% of the p24-siRNA-transfected cells.

Equivalents

[00213] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of
the invention described herein. The scope of the present invention is not intended to be limited to the above Description, but rather is as set forth in the appended claims.
We claim:

1. A method for selecting an siRNA targeted to a target transcript comprising:
   applying a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3’ overhang; and
   applying a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs.

2. The method of claim 1, wherein a core region is 19 nucleotides in length.

3. The method of claim 1, wherein a core region is perfectly complementary to the antisense strand of a corresponding siRNA.

4. The method of claim 1, wherein a core region contains one mismatch relative to the antisense strand of a corresponding siRNA.

5. The method of claim 1, wherein the target portion selection rule selects target portions according to a tiling approach, wherein target portions of a specified length are selected without regard to the identity of nucleotides 5’ or 3’ from the portion having the specified length.

6. The method of claim 1, wherein the target portion selection rule selects target portions according to a constrained overhang approach, wherein the target transcript includes at least two nucleotides immediately 5’ of the target portion and at least two nucleotides immediately 3’ of the target portion, wherein target portions of a specified length are selected, and wherein target portions are selected so that either or both of the two nucleotides immediately 5’ of the target portion are purines, or wherein either or both of the two nucleotides immediately 3’ of the target portion are pyrimidines.
7. The method of claim 6, wherein the target portions are selected so that both of the two nucleotides immediately 5’ of the target portion are purines, thereby resulting in selection of an siRNA sequence in which the antisense strand includes a 3’ overhang comprising two pyrimidines.

8. The method of claim 7, wherein the pyrimidines are thymidines.

9. The method of claim 7, wherein the pyrimidines are cytosines.

10. The method of claim 7, wherein the pyrimidines are uridines.

11. The method of claim 7, wherein the pyrimidines are deoxythymidines.

12. The method of claim 1, wherein the target portion selection rule selects target portions according to a constrained overhang approach, wherein the target transcript includes at least two nucleotides immediately 5’ of the target portion and at least two nucleotides immediately 3’ of the target portion, wherein target portions of a specified length are selected, and wherein target portions are selected so that either or both of the two nucleotides immediately 5’ of the target portion are pyrimidines, or wherein either or both of the two nucleotides immediately 3’ of the target portion are purines.

13. The method of claim 1, wherein the target portion selection rule selects target portions according to a constrained overhang approach, wherein the target transcript includes at least two nucleotides immediately 5’ of the target portion and at least two nucleotides immediately 3’ of the target portion, wherein target portions of a specified length are selected, and wherein target portions are selected so that either or both of the two nucleotides immediately 5’ of the target portion are deoxythymidines, or wherein either or both of the two nucleotides immediately 3’ of the target portion are deoxythymidines.

14. The method of claim 1, wherein the Complexity Rules include two Composition Rules, one of which selects preferred siRNAs based on the relative molar ratios of all four nucleotides in each candidate siRNA and one
of which selects preferred siRNAs based on the relative GC versus AT content of each candidate siRNA.

15. The method of claim 1, wherein the Complexity Rules include one or more Composition Rules and one or more Cluster Avoidance Rules.

16. The method of claim 15, wherein the Cluster Avoidance Rules include a rule specifying that sequences having one or more stretches of four consecutive identical nucleotides should be avoided.

17. The method of claim 15, wherein the Cluster Avoidance Rules include a rule specifying that sequences having one or more stretches of three consecutive identical nucleotides are less preferred than sequences lacking such stretches.

18. The method of claim 15, wherein the Cluster Avoidance Rules include a rule specifying that sequences having a stretch of three consecutive identical nucleotides are preferred to sequences having a row of three or more doublets.

19. The method of claim 15, wherein the Cluster Avoidance Rules include a rule specifying that sequences containing doublet or triplet repeats are less preferred than sequences lacking such repeats.

20. The method of claim 1, further comprising applying a suboptimal element positioning rule.

21. The method of claim 20, wherein suboptimal elements include elements selected from the group consisting of: a triplet, a doublet or triplet repeat, and a string of three or more doublets, and wherein the rule specifies that siRNA sequences in which the suboptimal element(s) are located toward the ends of the siRNA are preferred to sequences in which the suboptimal element(s) are closer to the middle, thereby selecting preferred siRNAs.

22. The method of claim 1, further comprising applying a base pair optimization rule.
23. The method of claim 22, wherein the base pair optimization rule specifies that siRNA sequences that contain clusters of Gs on one strand near clusters of three or more As or Ts on the opposite strand are less preferred than siRNA sequences lacking such clusters siRNA.

24. The method of claim 22, wherein the base pair optimization rule specifies that siRNA sequences that minimize possibilities for non Watson-Crick base pairing are preferred to sequences that offer such possibilities, wherein non Watson-Crick base pairing possibilities include, but are not limited to, G-U wobble, Hoogsteen, and inosine basepairing.

25. The method of claim 1, further comprising applying an overhang refinement rule.

26. The method of claim 25, wherein an overhang selected based on the sequence of the target transcript comprises at least one purine, and wherein the overhang refinement rule specifies that the purine is to be replaced by a pyrimidine.

27. The method of claim 1, further comprising applying a specificity rule.

28. The method of claim 27, wherein the specificity rule specifies that siRNA sequences that lack significant sequence identity or homology to other known sequences are preferred to sequences displaying such identity or homology.

29. The method of claim 28, wherein the specificity rule further specifies that if such identity or homology does exist over part of the siRNA, it is preferable that the region of identity or homology is located towards either end of the siRNA rather than near the middle.

30. The method of claim 27, wherein the specificity rule specifies that it is preferable to avoid siRNA sequences that include two regions of identity or complementarity with any known gene separated by a region of nonidentity or mismatch with the gene.
31. The method of claim 30, wherein the regions of identity or complementarity are between 5 and 10 nucleotides in length, and wherein the region of nonidentity or mismatch is between 1 and 6 nucleotides in length.

32. The method of claim 30, wherein the regions of identity or complementarity are either 6 or 7 nucleotides in length.

33. The method of claim 1, further comprising applying a global positioning rule.

34. The method of claim 33, wherein the global positioning rule specifies that the target portion of a preferred siRNA sequence is located within one or more exons.

35. The method of claim 33, wherein the global positioning rule is selected from the group consisting of a rule specifying that siRNA sequences located closer to the 3' end of the target transcript are preferred to sequences nearer to the middle or 5' end of the mRNA target and sequences closer to the 5' end of the target are preferred to sequences closer to the middle, and a rule specifying that siRNA sequences located within a single exon of the target transcript are preferred to siRNA sequences that span an exon/exon boundary.

36. The method of claim 1, further comprising applying an accessibility rule.

37. The method of claim 36, wherein the accessibility rule employs results of an RNase H protection assay or an RNA folding program to select preferred siRNAs.

38. The method of claim 1, further comprising: augmenting selected siRNA sequences with additional nucleotides to create a loop connecting the 3' end of the sense strand with the 5' end of the antisense strand of the siRNA.

39. A library of effective siRNAs selected according to the method of claim 1, wherein the library includes siRNAs targeted to all or a substantial fraction of genes expressed in an organism.
40. A library of effective siRNAs selected according to the method of claim 1, wherein the library includes siRNAs targeted to all or a substantial fraction of genes in a gene family expressed in an organism.

41. A method of identifying an siRNA hypersensitive site on a target transcript comprising steps of:
   providing at least ten siRNAs, wherein each siRNA includes an antisense strand having a sequence that is perfectly complementary to a target portion of the target transcript, wherein the siRNAs span at least fifty percent of the target transcript;
   delivering each siRNA to a different population of cells containing a level of the target transcript;
   determining to what extent each siRNA reduces the level of the target transcript or reduces expression of the target transcript in the population of cells to which it was delivered; and
   identifying a site on the target transcript as a hypersensitive site if an siRNA whose sense strand sequence either includes the site’s sequence or is included by the site’s sequence reduces the level of the target transcript or reduces expression of the target transcript to a greater extent than the other siRNAs.

42. The method of claim 41, wherein the antisense strand of each siRNA further includes a 3’ overhang that is perfectly complementary to the nucleotides immediately 5’ of the target portion in the target transcript.

43. The method of claim 41, wherein the antisense strand of each siRNA further includes a 3’ overhang comprising two pyrimidines.

44. The method of claim 41, wherein the target portions are between 15 and 29 nucleotides in length.

45. The method of claim 41, wherein the target portions are 19 nucleotides in length.

46. The method of claim 41, wherein the cells are HeLa cells.
47. The method of claim 41, wherein each siRNA includes an antisense strand having a sequence that contains at most four mismatches relative to a target portion of the target transcript

48. The method of claim 41, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least two fold relative to the reduction resulting from any of the other siRNAs.

49. The method of claim 41, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least five fold relative to the reduction resulting from any of the other siRNAs.

50. The method of claim 41, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least ten fold relative to the reduction resulting from any of the other siRNAs.

51. The method of claim 41, wherein the providing step comprises providing at least 25, at least 50, or at least 100 siRNAs.

52. The method of claim 41, wherein the target transcript contains Y nucleotides where Y is an integer, and wherein the providing step comprises providing a number of siRNAs greater than or equal to Y divided by 19.

53. The method of claim 41, wherein the providing step comprises providing a plurality of siRNAs that span at least 75%, at least 90%, or at least 95% of the target transcript.

54. An siRNA hypersensitive site identified according to the method of claim 41.

55. The siRNA hypersensitive site of claim 54, wherein the target transcript is selected from the group consisting of: c-myc, RAPTOR, SMN1, HPRT, hTERT, mTOR, lamin A/C TK1, and p53.
56. The siRNA hypersensitive site of claim 54, wherein the site occurs within a target transcript whose presence or activity within a cell is associated with a disease or clinical condition.

57. The siRNA hypersensitive site of claim 54, wherein the target transcript encodes an oncoprotein.

58. A method of treating or preventing a disease or clinical condition in an individual comprising:
   delivering an siRNA composition to the individual, wherein the composition comprises an siRNA whose antisense strand includes or is included by an siRNA hypersensitive site identified according to the method of claim 41.

59. A method of identifying a preferred siRNA to inhibit a target transcript comprising steps of:
   providing at least ten siRNAs, wherein each siRNA includes an antisense strand having a sequence that is perfectly complementary to a target portion of the target transcript, and wherein the siRNAs span at least fifty percent of the target transcript;
   delivering each siRNA to a different population of cells containing a level of the target transcript;
   determining to what extent each siRNA reduces the level of the target transcript or expression of the target transcript in the population of cells to which it was delivered; and
   identifying an siRNA that reduces the level of the target transcript or reduces expression of the target transcript to a greater extent than the other siRNAs as a preferred siRNA.

60. The method of claim 59, wherein the antisense strand of each siRNA further includes a 3' overhang that is perfectly complementary to the nucleotides immediately 5' of the target portion in the target transcript.
61. The method of claim 59, wherein the antisense strand of each siRNA further includes a 3' overhang comprising two pyrimidines.

62. The method of claim 59, wherein the target portions are between 15 and 29 nucleotides in length.

63. The method of claim 59, wherein the target portions are 19 nucleotides in length.

64. The method of claim 59, wherein the cells are HeLa cells.

65. The method of claim 59, wherein each siRNA includes an antisense strand having a sequence that contains at most four mismatches relative to a target portion of the target transcript.

66. The method of claim 59, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least two fold relative to the reduction resulting from any of the other siRNAs.

67. The method of claim 59, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least five fold relative to the reduction resulting from any of the other siRNAs.

68. The method of claim 59, wherein an siRNA whose sense strand either includes the site sequence or is included by the site sequence reduces the level of the target transcript or reduces expression of the target transcript by at least ten fold relative to the reduction resulting from any of the other siRNAs.

69. The method of claim 59, wherein the providing step comprises providing at least 25, at least 50, or at least 100 siRNAs.

70. The method of claim 59, wherein the target transcript contains $Y$ nucleotides where $Y$ is an integer, and wherein the providing step comprises providing a number of siRNAs greater than or equal to $Y$ divided by 19.
71. The method of claim 59, wherein the providing step comprises providing a plurality of siRNAs that span at least 75%, at least 90%, or at least 95% of the target transcript.

72. An siRNA identified according to the method of claim 59.

73. A vector comprising: a nucleic acid construct, the construct characterized in that when present in a cell, the construct directs transcription of the siRNA of claim 72.

74. A cell comprising: the siRNA of claim 72.

75. A transgenic animal engineered to contain or express the siRNA of claim 72.

76. The siRNA of claim 72, wherein the target transcript is selected from the group consisting of: c-myc, RAPTOR, SMN1, HPRT, hTERT, mTOR, lamin A/C, TK1, and p53.

77. A method of identifying a microRNA, which miRNA is capable of inhibiting expression of a gene or transcript, comprising steps of:

   testing an siRNA for its ability to inhibit a first target transcript;
   observing an unexpected phenotype, wherein the unexpected phenotype is a phenotype that would not be expected to occur solely due to inhibition of the first target transcript;
   searching in a database to identify a genes or a second transcript that includes two regions of identity or complementarity with the siRNA separated by a region of nonidentity or mismatch with the siRNA; and
   determining whether expression of the gene or second transcript is inhibited by the siRNA.

78. The method of claim 77, wherein the regions of identity or complementarity are between 5 and 10 nucleotides in length, and wherein the region of nonidentity or mismatch is between 1 and 6 nucleotides in length.

79. The method of claim 77, wherein the regions of identity or complementarity are either 6 or 7 nucleotides in length.
80. The method of claim 77, wherein determining whether expression of the gene or transcript is inhibited by the siRNA comprises determining whether the siRNA inhibits translation of the transcript or of an mRNA transcribed from the gene.

81. The method of claim 77, wherein a plurality of siRNAs are tested.

82. An miRNA identified according to the method of claim 77.

83. A pharmaceutical composition comprising:
   the siRNA of claim 72; and
   a pharmaceutically acceptable carrier.

84. A method of treating or preventing a disease or clinical condition in an individual comprising:
   delivering an siRNA composition to the individual, wherein the composition comprises the siRNA of claim 72.

85. The siRNA of claim 72, wherein the siRNA targets a target transcript whose presence or activity within a cell is associated with a disease or clinical condition.

86. The siRNA of claim 85, wherein the target transcript encodes an oncoprotein.

87. A method for providing an siRNA comprising steps of:
   receiving information identifying a target transcript from a user;
   selecting one or more preferred siRNA targeted to the transcript by applying a set of siRNA selection rules to the target portion, thereby selecting one or more preferred siRNAs; and
   providing at least one siRNA to the user.

88. The method of claim 87, wherein the rules include:
   a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a
sense strand and an antisense strand, each of which optionally includes a 3' overhang; and

a plurality of Complexity Rules.

89. The method of claim 87, wherein the information is received via the Internet.

90. The method of claim 87, further comprising:

providing a plurality of options to the user, the options allowing the user to specify one or more rules to be applied to select one or more preferred siRNAs.

91. The method of claim 87, further comprising:

providing a plurality of options to the user, the options allowing the user to specify one or more parameters to be applied in conjunction with a rule used to select one or more preferred siRNAs.

92. A computer system for selecting an siRNA targeted to a target transcript, the computer system comprising:

memory means which stores a program comprising computer-executable process steps; and

a processor which executes the process steps so as (i) to apply a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3' overhang; and (ii) to apply a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs.

93. The computer system of claim 92, wherein the processor further executes process steps so as to receive information identifying the target transcript from a user.

94. The computer system of claim 92, wherein the processor further executes process steps to as to provide a plurality of options to the user, the options
allowing the user to specify one or more rules to be applied to select one or more preferred siRNAs.

95. The computer system of claim 92, wherein the processor further executes process steps so as to provide a plurality of options to the user, the options allowing the user to specify one or more parameters to be applied in conjunction with a rule used to select one or more preferred siRNAs.

96. The computer system of claim 92, wherein the processor further executes process steps so as to provide the sequence of one or more preferred siRNAs to the user.

97. The computer system of claim 96, wherein the computer system comprises an Internet connection enabling World Wide Web access, and wherein the steps of receiving information identifying a target transcript and providing the sequence of one or more preferred siRNAs to the user take place via the World Wide Web.

98. Computer-executable process steps stored on a computer-readable medium, the computer-executable process steps to select an siRNA targeted to a target transcript, the computer-executable process steps comprising:

   code to apply a target portion selection rule to select portions of the target transcript, thereby identifying a set of core regions, wherein each core region corresponds to and thus specifies at least a duplex portion of a candidate siRNA, and wherein each duplex portion of a candidate siRNA thus specified comprises a sense strand and an antisense strand, each of which optionally includes a 3’ overhang; and

   code to apply a plurality of Complexity Rules to the candidate siRNAs, thereby selecting preferred siRNAs.

99. Computer-executable process steps according to claim 98, further comprising:

   code to receive information identifying a target transcript.

100. Computer-executable process steps according to claim 98, further comprising:
code to provide a plurality of options to the user, the options allowing
the user to specify one or more rules to be applied to select one or more
preferred siRNAs.

101. Computer-executable process steps according to claim 98, further comprising:
    code to provide a plurality of options to the user, the options allowing
the user to specify one or more parameters to be applied in conjunction with a
rule used to select one or more preferred siRNAs.

102. Computer-executable process steps according to claim 98, further comprising:
    code to provide the sequence of one or more preferred siRNAs to the
user.

103. Computer-executable process steps according to claim 102, wherein the code
to receive information identifying the target transcript and the code to provide
the sequence of one or more preferred siRNAs to the user comprise code to
perform these operations via World Wide Web interactions.
2 nt 3' overhang

19 bp ds portion

2 nt 3' overhang

FIGURE 1
RNA INTERFERENCE IN DROSOPHILA

long dsRNA

DICER-mediated

cleavage of ds RNA

HO

p

p

OH

siRNA

HELICASE?

m7G(5')ppp(5')G

AAAAA AAAA mRNA

Slicing of mRNA by RISC complex

m7G(5')ppp(5')G

TRANSCRIPT DEGRADATION

FIGURE 2
Figure 6

DICER

target transcript

CAP

3' UTR

Translation
Fig. 2. p24-siRNA inhibits viral replication in HeLa-CD4 cells. a. Specificity of p24-siRNA-directed inhibition of viral gene expression demonstrated by flow-cytometric analysis of p24 expression (p24RD1) in uninfected control and mock-, p24-siRNA-, p24-siRNA-antisense strand- and GFP-siRNA- (control siRNA) transfected HeLa-CD4 cells 2 days after infection with HIVΔv. Percentage of gated p24+ cells shown in each panel. b. Northern blot for p24 expression in uninfected (lane 1) mock- (lane 2), p24-siRNA- (lane 3), p24-siRNA-antisense strand- (lane 4) and control siRNA- (lane 5) transfected cells. β-actin expression was used as a loading control. c. Flow-cytometric analysis of p24 expression (p24RD1) in uninfected control and mock-, p24-siRNA- and GFP-siRNA- (control siRNA) transfected HeLa-CD4 cells 5 d after infection with HIVΔv. Percentage of gated p24+ cells are shown in each panel. d. Northern blot for p24, Nef and β-actin expression in stably infected control (lane 1), uninfected (lane 2) mock- (lane 3), p24-siRNA- (lane 4), and control siRNA- (lane 5) transfected cells. Compared with mock- or control siRNA- transfected cells, p24-siRNA-transfected cells showed decreased expression of the full-length, 9.2-kb HIV-1 transcripts and/or genomic RNA as well as the 4.3- and 2.0-kb nef-containing transcripts. β-actin expression was used as a loading control. e. Viral p24 antigen measured by ELISA in uninfected control (lane 1) and mock- (lane 2), p24-siRNA- (lane 3) and control siRNA- (lane 4) transfected cells infected with HIVΔv demonstrates reduction of cell-free virus production only in p24-siRNA-transfected HeLa-CD4 cells. Error bars represent the average of 3 experiments.

FIGURE 10
Figure 11
Figure 12
entitled Sequence #1 → Restriction Map

DNA sequence 231 b.p. atggaaaaagatt atataataacta linear

**Figure 13**