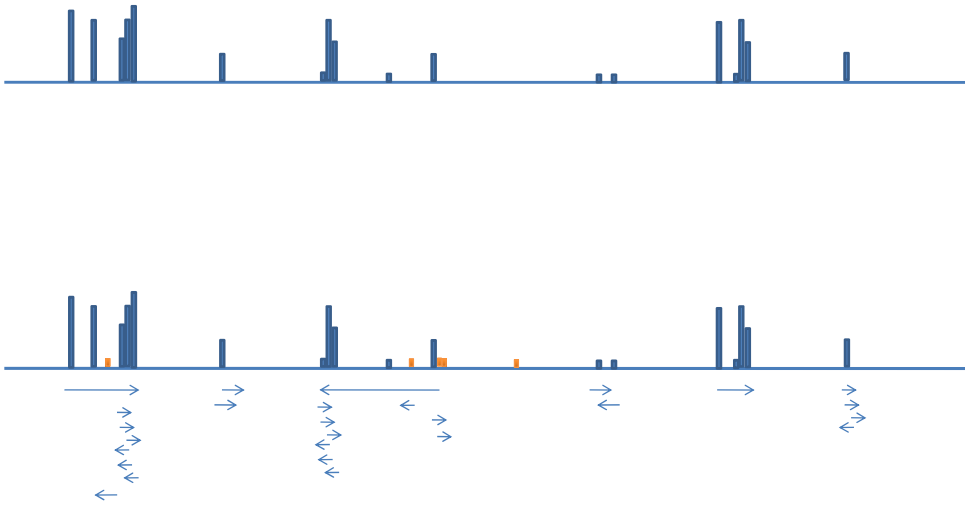
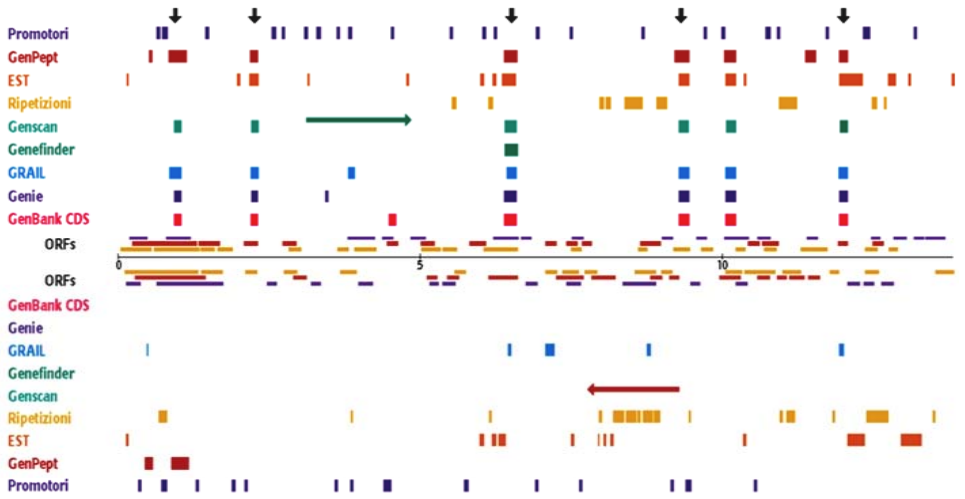
Genome annotation





Annotazione bioinformatica di un segmento di 15 kb del genoma umano, contenente un gene, utilizzando il browser Genotator.

Tutto il genoma viene continuamente ri-annotato considerando le nuove evidenze sperimentali, se liberamente disponibili.

Of course, results of tiling array hybridization, CAGE/SAGE and –notably- RNA Seq will enormously improve functional genomic annotation

but

this requires a lot of time and work....

Public projects (such as NCBI, ENSEMBL , EBI) are absolutely needed, nobody will do it privately → too expensive, no advantage

except

microbiology – related genomes (virus, bacteria,...)
and plants related to human food production

RNA transcript annotation is much but not all...

RNA transcripts (even capped) may arise from unexpected phenomena, such as cleavage of longer transcript, *trans*-splicing followed by processing, editing, and even by RNA-dependent RNA polymerization (even though such a polymerase, found in lower eukaryotes (e.g *c. elegans*), has never been observed in Vertebrates).

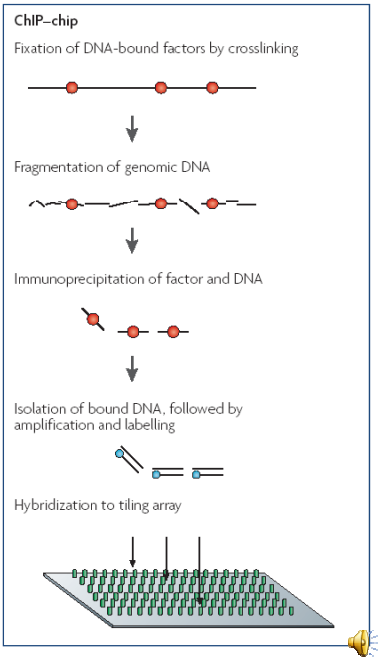
Other functional annotations may help, such as:

- the presence of RNA polymerase(s)
- the presence of other proteins necessary for transcription (and RNA processing)
- histone modifications that detect the local functional status of the chromatin
- DNA covalent modifications (e.g. CpG methylation)
- ....

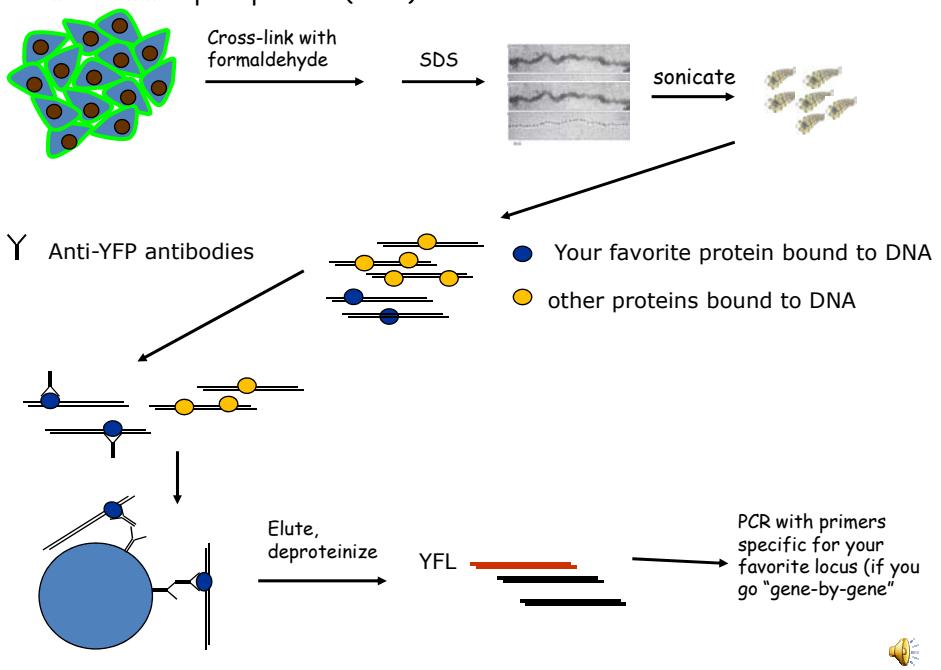
Other methods to identify genome-wide promoters have been used.

One strategy is to identify all the sequences that are bound by RNA Polymerase II and by basal transcription factors, such as TAFs, by using ChIP-on-chip.

As a part of the ENCODE project, for example, 1%-human genome coverage tiling arrays were used to identify sequences bound by RNA PolII and TAF1 (see TAF nomenclature).



Chromatin immunoprecipitation (ChIP)



Direct isolation and identification of promoters in the human genome

Tae Hoon Kim,[1] Leah O. Barrera,[1] Chunxu Qu,[1] Sara Van Calcar,[1] Nathan D. Trinklein,[4] Sara J. Cooper,[4] Rosa M. Luna,[2] Christopher K. Glass,[2] Michael G. Rosenfeld,[3] Richard M. Myers,[4] and Bing Ren[1,2,5]

[1]Ludwig Institute for Cancer Research, [2]Department of Cellular and Molecular Medicine, and [3]Howard Hughes Medical Institute, University of California, San Diego, La Jolla, California 92093, USA; [4]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

Transcriptional regulatory elements play essential roles in gene expression during animal development and cellular response to environmental signals, but our knowledge of these regions in the human genome is limited despite the availability of the complete genome sequence. Promoters mark the start of every transcript and are an important class of regulatory elements. A large, complex protein structure known as the pre-initiation complex (PIC) is assembled on all active promoters, and the presence of these proteins distinguishes promoters from other sequences in the genome. Using components of the PIC as tags, we isolated promoters directly from human cells as protein–DNA complexes and identified the resulting DNA sequences using genomic tiling microarrays. Our experiments in four human cell lines uncovered 252 PIC-binding sites in 44 semirandomly selected human genomic regions comprising 1% (30 megabase pairs) of the human genome. Nearly 72% of the identified fragments overlap or immediately flank 5′ ends of known cDNA sequences, while the remainder is found in other genomic regions that likely harbor putative promoters of unannotated transcripts. Indeed, molecular analysis of the RNA isolated from one cell line uncovered transcripts initiated from over half of the putative promoter fragments, and transient transfection assays revealed promoter activity for a significant proportion of fragments when they were fused to a luciferase reporter gene. These results demonstrate the specificity of a genome-wide analysis method for mapping transcriptional regulatory elements and also indicate that a small, yet significant number of human genes remains to be discovered.

[Supplemental material is available online at www.genome.org.]

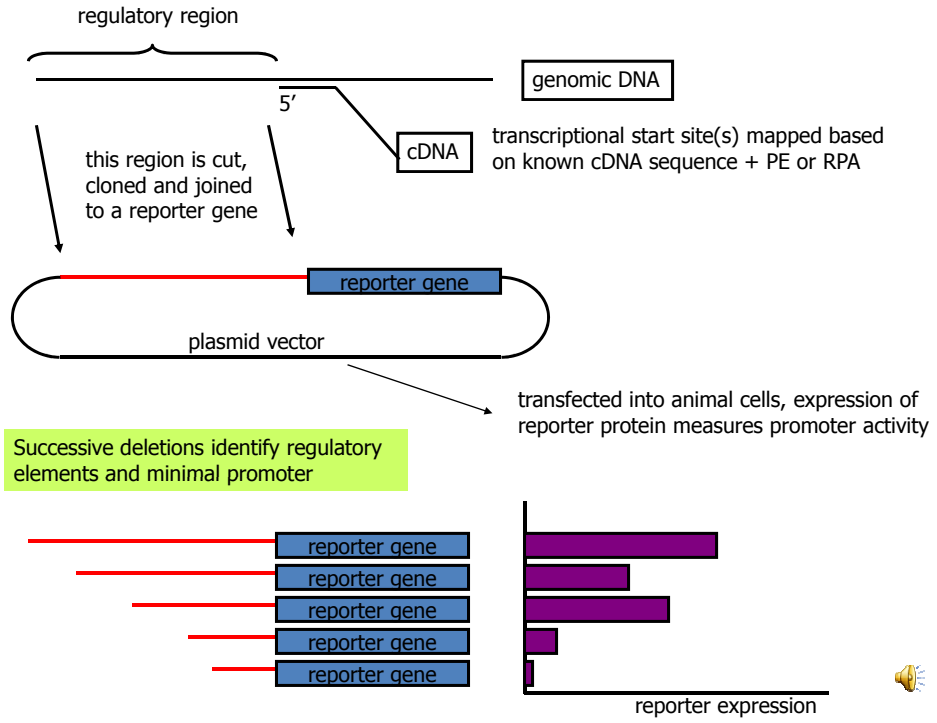
15:830–839 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

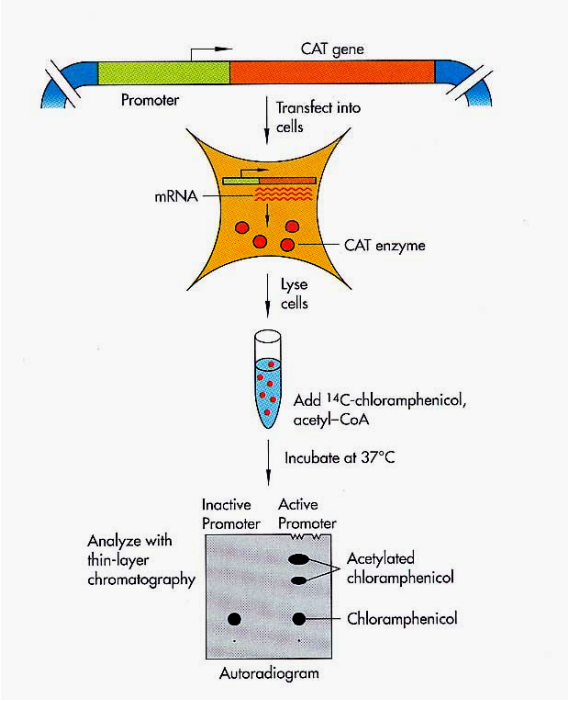
When a sequence is identified as a putative promoter, the functional validation consists in the classical **reporter assay**

It this assay, a minigene is constructed fusing the putative promoter sequence to a **reporter** gene, i.e. a gene whose product can be easily measured.

This construct is then transfected into cultured cells and the product measured after a period of time necessary for transgnene import and expression.

Mammalian **reporter genes** are usually nonmammalian genes, such as
- CAT cloramphenicol acetyltransferase
- Luc firefly luciferase
- GFP Jellyfish green fluorescent protein
- β-gal beta-galactosidase





The reporter gene assay for promoter and enhancers

CAT reporter was very used in the past, but biochemical evaluation is quite labor-intensive

Nevertheless, it provides a paradigmatic example of the reporter vector techninque

Some of the TU identified are:
micro RNA genes

Revealing the world of RNA interference

Craig C. Mello[1,2] & Darryl Conte Jr[2]

[1]Howard Hughes Medical Institute and [2]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA (e-mail: craig.mello@umassmed.edu)

The recent discoveries of RNA interference and related RNA silencing pathways have revolutionized our understanding of gene regulation. RNA interference has been used as a research tool to control the expression of specific genes in numerous experimental organisms and has potential as a therapeutic strategy to reduce the expression of problem genes. At the heart of RNA interference lies a remarkable RNA processing mechanism that is now known to underlie many distinct biological phenomena.

REVIEW

see Ghildiyal & Zamore 2009, Nature Rev Genetics - moodle pdf

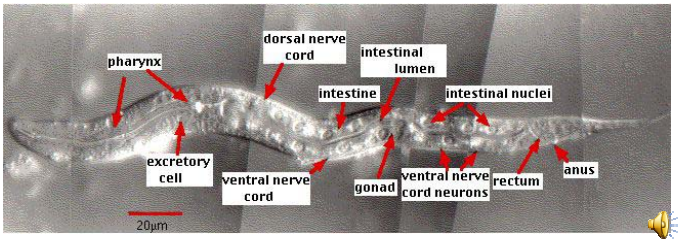
Before… It was only known that, when a double-strad RNA enters a mammalian cell, an interferon response is induced and general protein synthesis is turned down

In 1995.. Guo & Kemphues were attempting to knock-out *par1* mRNA in *C. elegans*, and were transfecting large amounts of in vitro transcribed antisense RNA, using as control "sense" RNA. Surprise: the *par1* mRNA was downregulated by either sense or antisense RNA.

In 1998.. Fire et al., transfect both sense and antisense RNA and find that PTGS (post-transcriptional gene silencing) is 10- to 100-fold stronger ! They call this phenomenon RNA interference.

With more surprise, they find that silencing effect can be transmitted in the germ line and passed up through the sperm or the egg for up to several generations

Even more surprising, silencing can also spread from cell to cell and from tissue to tissue.



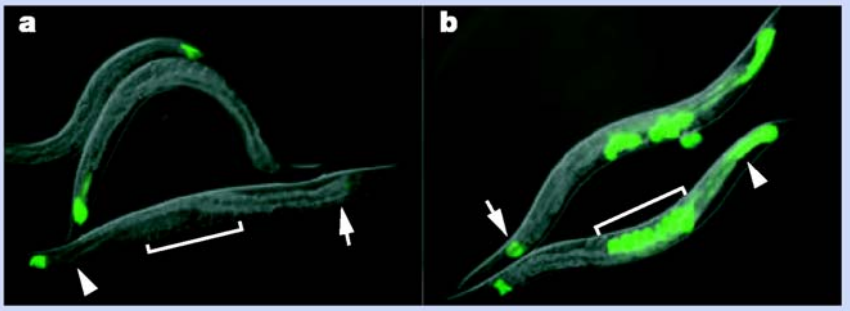
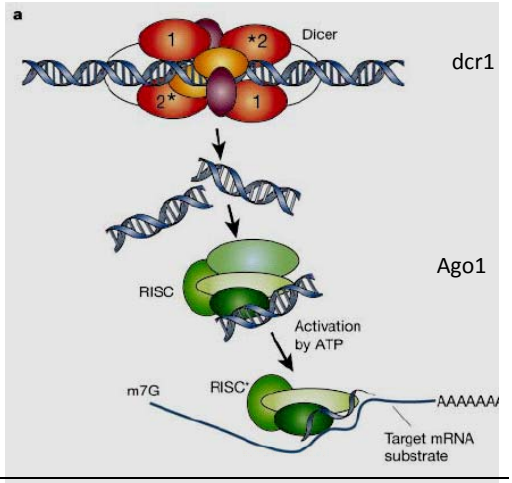


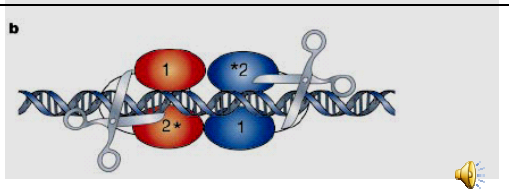
Figure 1 RNAi in *C. elegans*. Silencing of a green fluorescent protein (GFP) reporter in *C. elegans* occurs when animals feed on bacteria expressing GFP dsRNA (**a**) but not in animals that are defective for RNAi (**b**). Note that silencing occurs throughout the body of the animal, with the exception of a few cells in the tail that express some residual GFP. The signal is lost in intestinal cells near the tail (arrowhead) as well as near the head (arrow). The lack of GFP-positive embryos in **a** (bracketed region) demonstrates the systemic spread and inheritance of silencing.

Dicer and RISC (RNA-induced silencing complex).

a, RNAi is initiated by the Dicer enzyme (two Dicer molecules with five domains each are shown), which processes double-stranded RNA into ~22-nucleotide small interfering RNAs. Based upon the known mechanisms for the RNase III family of enzymes, Dicer is thought to work as a dimeric enzyme. Cleavage into precisely sized fragments is determined by the fact that one of the active sites in each Dicer protein is defective (indicated by an asterisk), shifting the periodicity of cleavage from ~9–11 nucleotides for bacterial RNase III to ~22 nucleotides for Dicer family members40. The siRNAs are incorporated into a multicomponent nuclease, RISC (green). Recent reports suggest that RISC must be activated from a latent form, containing a double-stranded siRNA to an active form, RISC*, by unwinding of siRNAs41. RISC* then uses the unwound siRNA as a guide to substrate selection31.



b, diagrammatic representation of Dicer binding and cleaving dsRNA (for clarity, not all the Dicer domains are shown, and the two separate Dicer molecules are coloured differently). Deviations from the consensus RNase III active site in the second RNase III domain inactivate the central catalytic sites, resulting in cleavage at 22-nucleotide intervals.



What is RNA interference ?

From the **cognitive** point of view, a fundamental and completely unexpected mechanism that demonstrates a primary role of RNA for controlling genome activity

and

revalutates the RNA world hypothesis

From the **applicative** point of view, one of the most important and impacting discoveries in the last 15 years.

Tools for experimentally knocking-down (downregulating) genes have been looked for since decades, especially for mammalian cells, where there was only the transgenic mouse alternative to low-efficiency and cumbersome antisense oligos or ribozymes.

RNAi allows knocking-out expression of the gene you need in virtually all model systems

In some organisms (not all, for example not in higher animals)

an additional component required:

RNA-dependent RNA Polymerase (C.elegans, plants, S. pombe: rpd1)

amplifies the RNA to be silenced by constructing complemetary copies

movie

Micro RNA are a family of small RNA that are transcribed from several locations in genomes. The human genome may contain (estimated) up to 1,000 miRNA genes.

They have a typical structure, making a stem-loop structure with some mismatches in the stem

They are processed by Dicer and target usually the 3'-UTR of several mRNAs, leading to cleavage and degradation or inhibiting translation.

Complementary target sequence limited to few nucleotides: one miRNA targets multiple mRNA (co-regulons).

Roles in growth control, development, cancer have been found for miRNAs

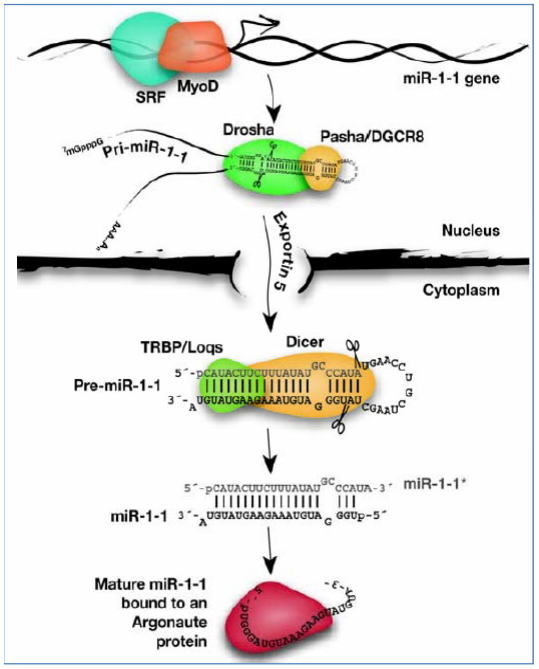


Fig. 2. A day in the life of the miRNA miR-1. In developing cardiac tissue, the trans-cription factors SRF (serum response factor) and MyoD promote RNA Pol II–directed transcription of pri-miR-1. In the nucleus, the RNase III endonuclease Drosha, together with its dsRNA-binding partner, Pasha/ /DGCR8, excises pre-miR-1 from pri-miR-1, breaking the RNA chain on both the 5′ and 3′ sides of the pre-miR-1 stem, leaving a 2-nt, single-stranded 3′ overhang end.

Exportin 5 recognizes this characteristic pre-miRNA end structure, transporting pre-miR-1 from the nucleus to the cytoplasm. In the cytoplasm, a second RNase III endo-nuclease, Dicer, together with its dsRNA-binding partner protein, Loqs/TRBP, makes a second pair of cuts, liberatingmiR-1 as a ''miRNA/miRNA*'' duplex. Mature, 21-nt long miR-1 is then loaded from the duplex into an Argonaute family member and miR-1* is destroyed. miR-1 guides the Argonaute protein to its target RNAs, such as the 3′ untranslated region of the hand2 mRNA. Binding of the miR-1–programmed Argonaute protein represses production of Hand2 protein, halting cardiac cell proliferation.

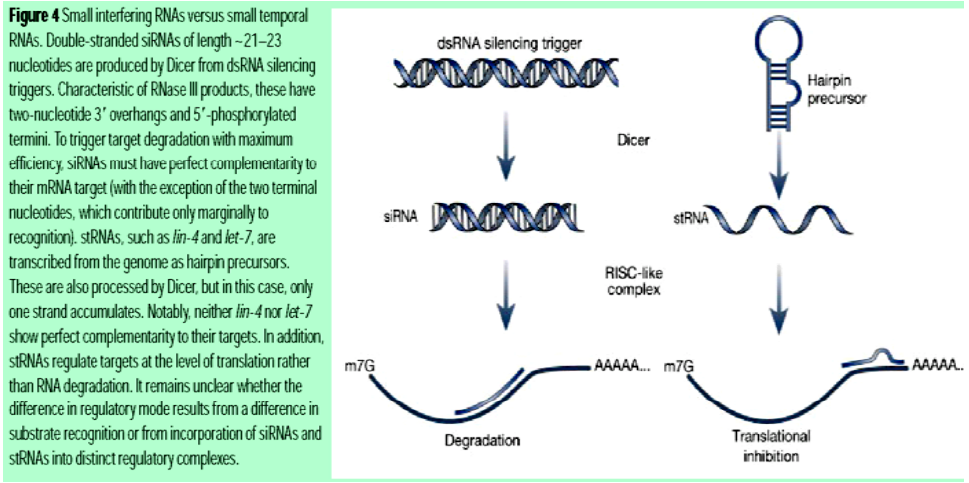


Figure 4 Small interfering RNAs versus small temporal RNAs. Double-stranded siRNAs of length ~21–23 nucleotides are produced by Dicer from dsRNA silencing triggers. Characteristic of RNase III products, these have two-nucleotide 3′ overhangs and 5′-phosphorylated termini. To trigger target degradation with maximum efficiency, siRNAs must have perfect complementarity to their mRNA target (with the exception of the two terminal nucleotides, which contribute only marginally to recognition). stRNAs, such as *lin-4* and *let-7*, are transcribed from the genome as hairpin precursors. These are also processed by Dicer, but in this case, only one strand accumulates. Notably, neither *lin-4* nor *let-7* show perfect complementarity to their targets. In addition, stRNAs regulate targets at the level of translation rather than RNA degradation. It remains unclear whether the difference in regulatory mode results from a difference in substrate recognition or from incorporation of siRNAs and stRNAs into distinct regulatory complexes.

stRNA now called microRNA = miRNA

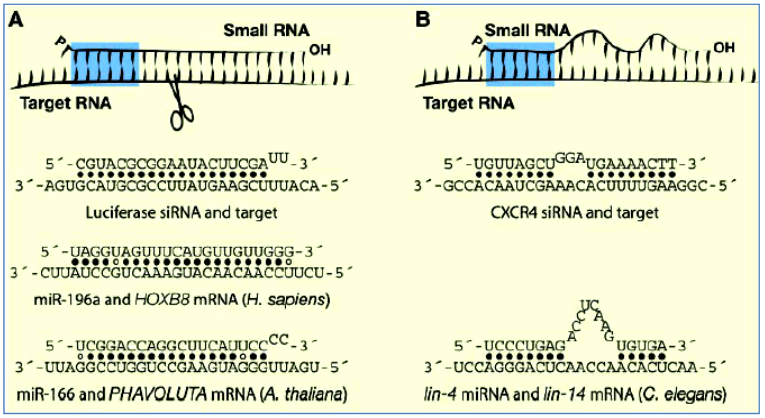


Fig. 3. Small RNA binding modes.
(A) **Extensive pairing** of a small RNA to an mRNA allows the Piwi domain of a catalytically active Argonaute protein (e.g., Ago2 in humans or flies) to cut a single phosphodiester bond in the mRNA, triggering its destruction. Synthetic siRNAs typically exploit this mechanism, but some mammalian miRNAs (such as miR-196a) and most, if not all, plant miRNAs direct an Argonaute protein to cut their mRNA targets.
(B) **Partial pairing** between the target RNA and the small RNA, especially through the ''seed'' sequence —roughly nucleotides 2 to 7 of the small RNA—tethers an Argonaute protein to its mRNA target. Binding of the miRNA and Argonaute protein prevents translation of the mRNA into protein. siRNAs can be designed to trigger such ''translational repression'' by including central mismatches with their target mRNAs; animal miRNAs such as lin-4, the first miRNA discovered, typically act by this mode because they are only partially complementary to their mRNA targets. The seed sequence of the small RNA guide is highlighted in blue.

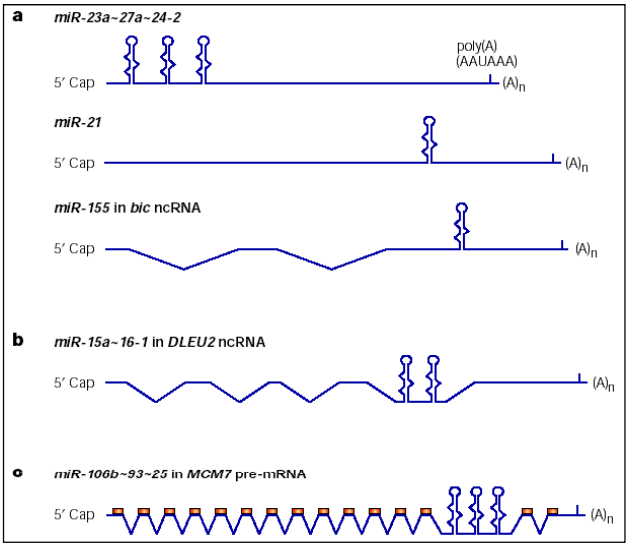


Figure 1 | **The structure of five pri-miRNAs.**

Primary transcripts that encode miRNAs, primiRNAs, contain 5′ cap structures as well as 3′ poly(A) tails. miRNAs can be categorized into three groups according to their genomic locations relative to their positions in an exon or intron.

a | Exonic miRNAs in non-coding transcripts such as an *miR-23a~27a~24-2* cluster, *miR-21* and *miR-155*. *miR-155* was found in a previously defined non-coding RNA (ncRNA) gene, *bic*17.

b | Intronic miRNAs in non-coding transcripts. For example, an *miR-15a~16-1* cluster was found in the fourth intron of a previously defined non-coding RNA gene, *DLEU2* (REF. 126). **c** | Intronic miRNAs in protein-coding transcripts. For example, an *miR-106b~93~25* cluster is embedded in the thirteenth intron of DNA replication licensing factor *MCM7* transcript (variant 1, which encodes isoform 1). The mouse *miR-06b~93~25* homologue is also found in the thirteenth intron of the mouse *MCM7* homologue gene15. The hairpins indicate the miRNA stem-loops. Orange boxes indicate the protein-coding region. This figure is not to scale.

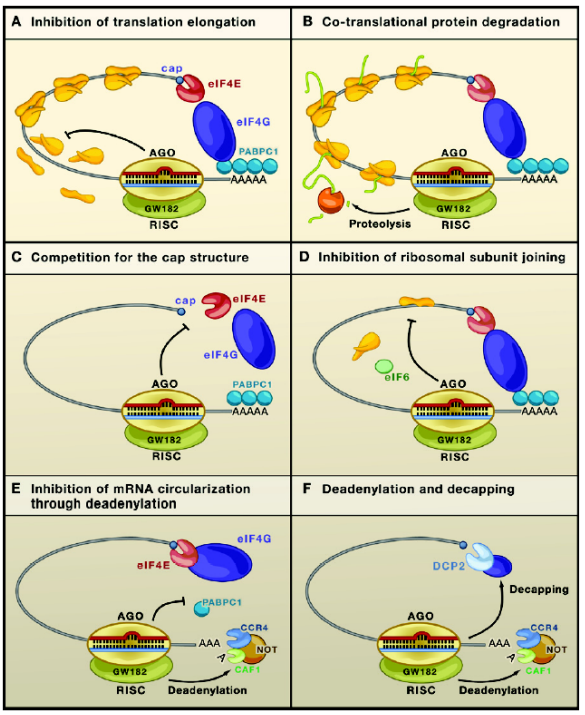


Figure 1. Mechanisms of miRNA-Mediated Gene Silencing

(A) Postinitiation mechanisms. MicroRNAs (miRNAs; red) repress translation of target mRNAs by blocking translation elongation or by promoting premature dissociation of ribosomes (ribosome drop-off).

(B) Cotranslational protein degradation. This model proposes that translation is not inhibited, but rather the nascent polypeptide chain is degraded cotranslationally. The putative protease is unknown.

(C–E) Initiation mechanisms. MicroRNAs interfere with a very early step of translation, prior to elongation. (C) Argonaute proteins compete with eIF4E for binding to the cap structure (cyan dot).

(D) Argonaute proteins recruit eIF6, which prevents the large ribosomal subunit from joining the small subunit.

(E) Argonaute proteins prevent the formation of the closed loop mRNA configuration by an ill-defined mechanism that includes deadenylation.

(F) MicroRNA-mediated mRNA decay. MicroRNAs trigger deadenylation and subsequent decapping of the mRNA target. Proteins required for this process are shown including components of the major deadenylase complex (CAF1, CCR4, and the NOT complex), the decapping enzyme DCP2, and several decapping activators (dark blue circles). (Note that mRNA decay could be an independent mechanism of silencing, or a consequence of translational repression, irrespective of whether repression occurs at the initiation or postinitiation levels of translation.) RISC is shown as a minimal complex including an Argonaute protein (yellow) and GW182 (green). The mRNA is represented in a closed loop configuration achieved through interactions between the cytoplasmic poly(A) binding protein (PABPC1; bound to the 3′ poly(A) tail) and eIF4G (bound to the cytoplasmic cap-binding protein eIF4E).