RNA types

Progressive knowledge of the transcriptome has shown that most of (nonrepetitive) genome sequences are transcribed.

Most of noncoding transcripts (either antisense, intronic, promoter, termination and intergenic) show indeed a very low level of expression.

As illustrated in the next two slides by Ponting & Belgard, from the quantitative point of view 88% of the transcript are mapped to known exons (good news!);
but
from the qualitative point of view, expressed exon sequences are only 22% of the total.

Conclusion: what are all these low-level RNAs ? Do they represent only artifacts or “leaky” transcription or do they have a functional role?
Take into accout that most of this noncoding RNA-encoding DNA is much more conserved than expected if no functional role was thought.

Human Molecular Genetics, 2010, Vol. 19, Review Issue 2 **R162–R168**
doi:10.1093/hmg/ddq362
Advance Access published on August 25, 2010

Transcribed dark matter: meaning or myth?

Chris P. Ponting* and T. Grant Belgard

MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Received July 13, 2010; Revised July 13, 2010; Accepted August 21, 2010

Genomic tiling arrays, cDNA sequencing and, more recently, RNA-Seq have provided initial insights into the extent and depth of transcribed sequence across human and other genomes. These methods have led to greatly improved annotations of protein-coding genes, but have also identified transcription outside of annotated exons. One resultant issue that has aroused dispute is the balance of transcription of known exons against transcription outside of known exons. While non-genic 'dark matter' transcription was found by tiling arrays to be pervasive, it was seen to contribute only a small percentage of the polyadenylated transcriptome in some RNA-Seq experiments. This apparent contradiction has been compounded by a lack of clarity about what exactly constitutes a protein-coding gene. It remains unclear, for example, whether or not all transcripts that overlap on either strand within a genomic locus should be assigned to a single gene locus, including those that fail to share promoters, exons and splice junctions. The inability of tiling arrays and RNA-Seq to count transcripts, rather than exons or exon pairs, adds to these difficulties. While there is agreement that thousands of apparently non-coding loci are present outside of protein-coding genes in the human genome, there is vigorous debate of what constitutes evidence for their functionality. These issues will only be resolved upon the demonstration, or otherwise, that organismal or cellular phenotypes frequently result when non-coding RNA loci are disrupted.

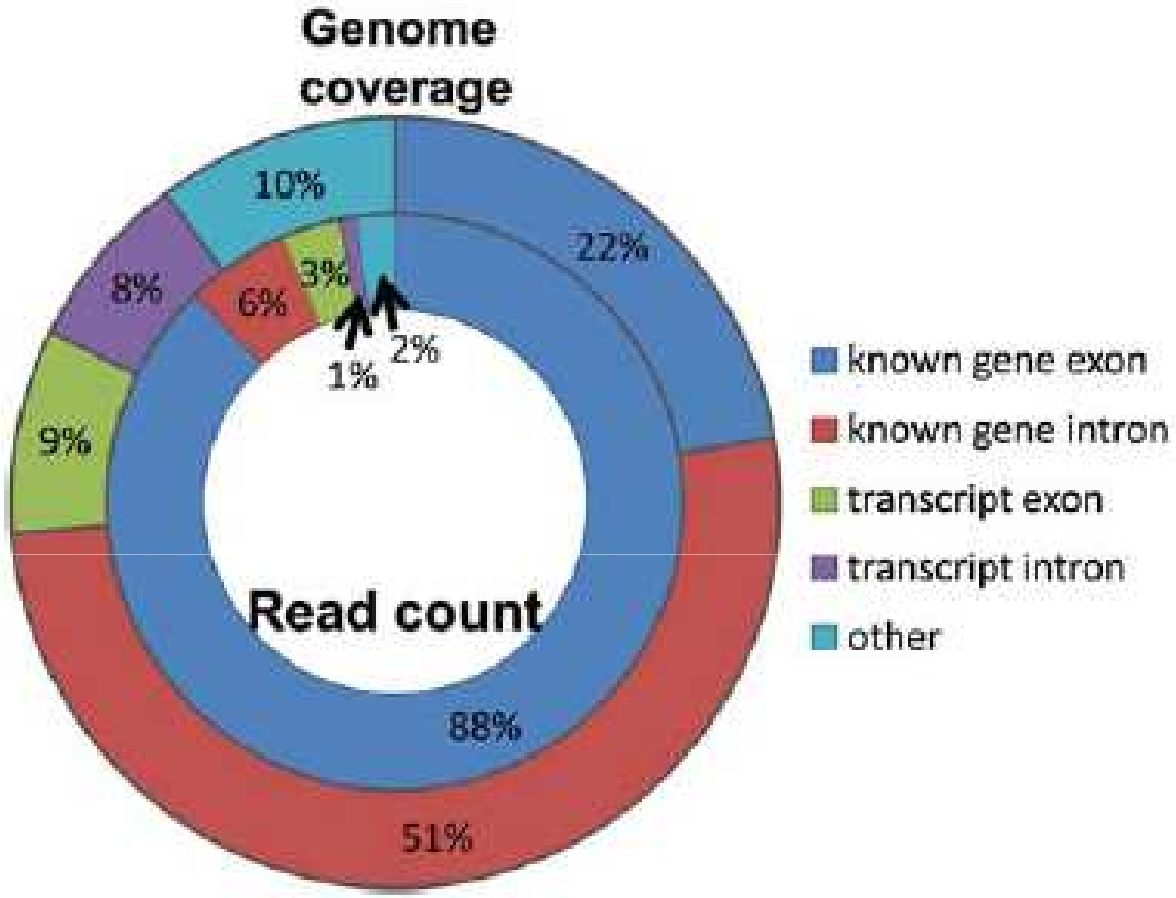


Figure 1. Exons from known genes are associated with 88% of uniquely mapping short reads, but provide 22% of genomic sequence that is transcribed [human data from van Bakel *et al*. (19)]. On the other hand, only 6% of uniquely mapping reads are in intergenic sequence, but these lowly expressing regions cover about one quarter of all transcribed genomic sequence.

We have an increasingly complex anthology of different noncoding RNA classes.

see for example:

RNA list (a Wikipedia vision of transcriptome)

For some classes, we have now good knowledge of function

For example, small RNA that are linked to the RNA interference pathway have been quite well characterized

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.

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.

Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*

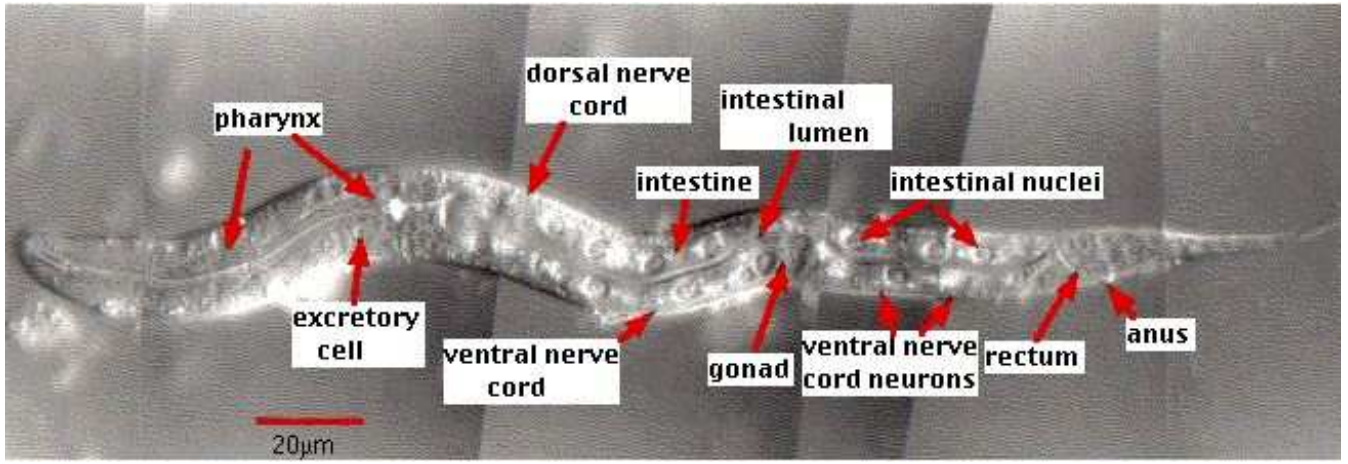
Andrew Fire*, SiQun Xu*, Mary K. Montgomery*, Steven A. Kostas*†, Samuel E. Driver‡ & Craig C. Mello‡

* *Carnegie Institution of Washington, Department of Embryology, 115 West University Parkway, Baltimore, Maryland 21210, USA*
† *Biology Graduate Program, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA*
‡ *Program in Molecular Medicine, Department of Cell Biology, University of Massachusetts Cancer Center, Two Biotech Suite 213, 373 Plantation Street, Worcester, Massachusetts 01605, USA*

Experimental introduction of RNA into cells can be used in certain biological systems to interfere with the function of an endogenous gene[1,2]. Such effects have been proposed to result from a simple antisense mechanism that depends on hybridization between the injected RNA and endogenous messenger RNA transcripts. RNA interference has been used in the nematode *Caenorhabditis elegans* to manipulate gene expression[3,4]. Here we investigate the requirements for structure and delivery of the interfering RNA. To our surprise, we found that double-stranded RNA was substantially more effective at producing interference than was either strand individually. After injection into adult animals, purified single strands had at most a modest effect, whereas double-stranded mixtures caused potent and specific interference. The effects of this interference were evident in both the injected animals and their progeny. Only a few molecules of injected double-stranded RNA were required per affected cell, arguing against stochiometric interference with endogenous mRNA and suggesting that there could be a catalytic or amplification component in the interference process.

Andrew Fire and Craig Mello had the Nobel prize in 2006.

http://en.wikipedia.org/wiki/Caenorhabditis_elegans



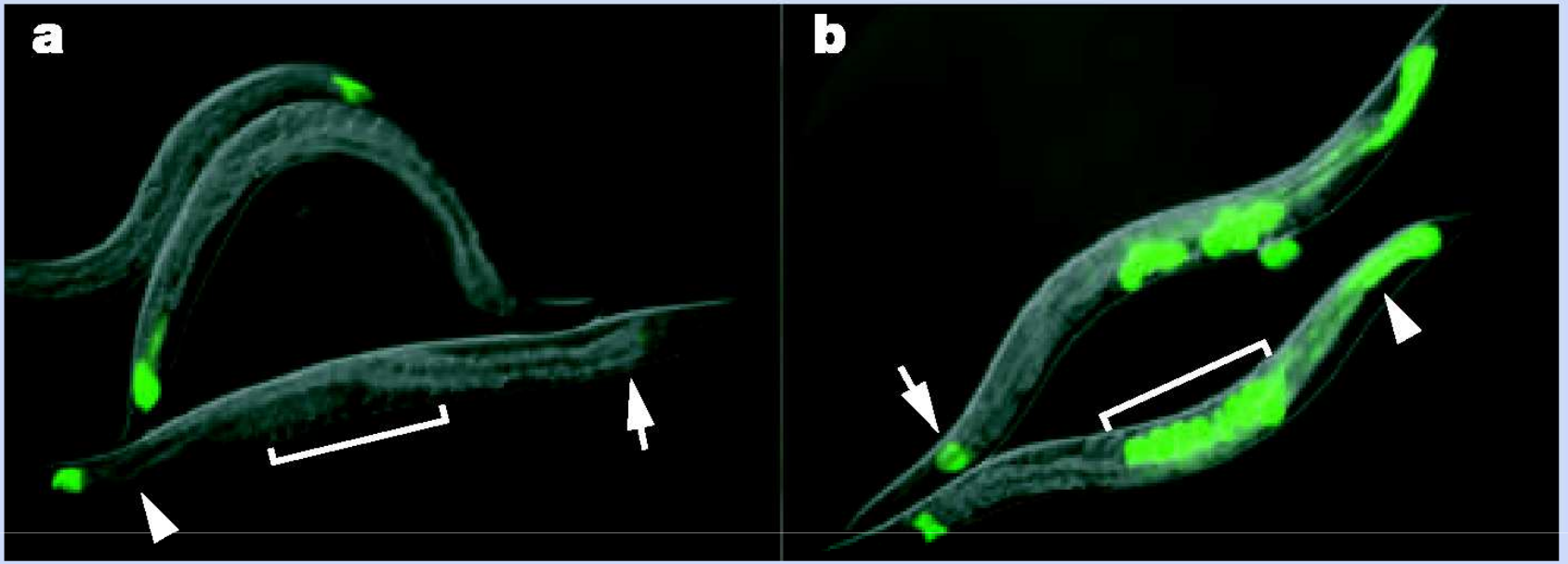
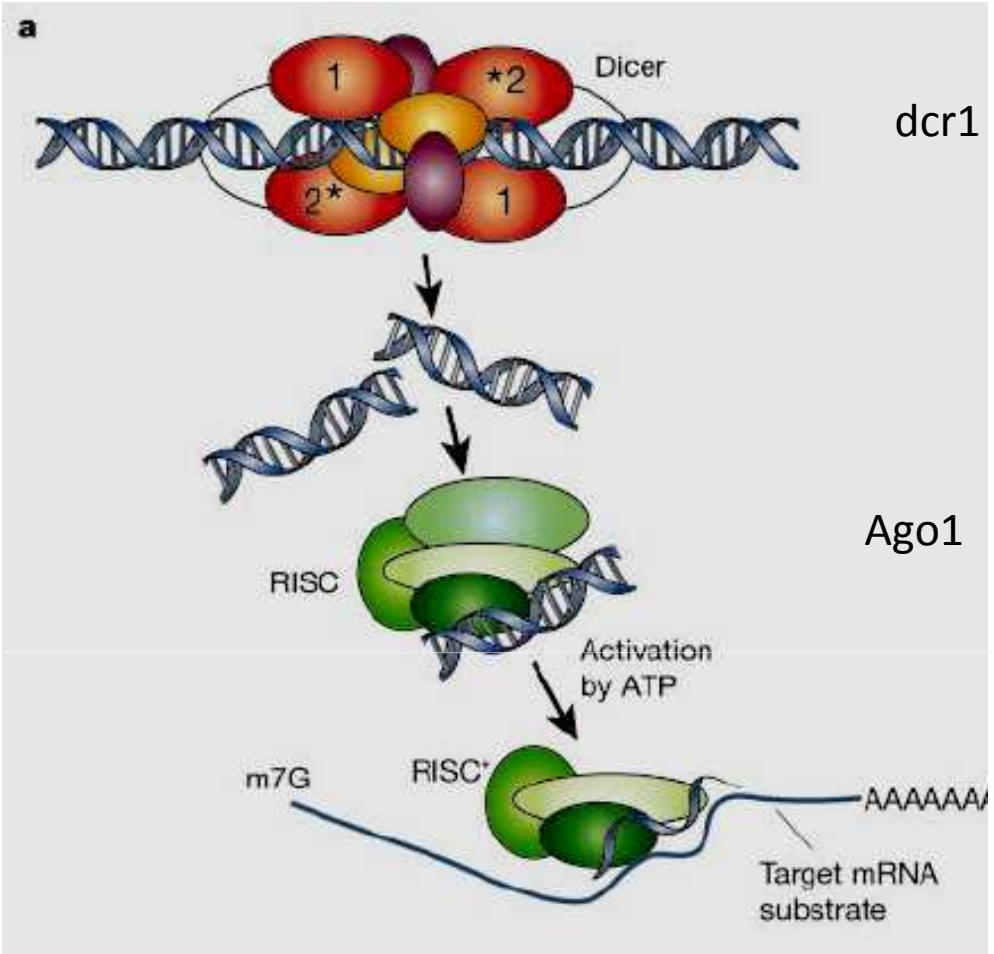


Figure 1 RNAi in *C. elegans*. Silencing of a green fluorescent protein (GFP) reporter in *C. elegans* occurs when animals feed on bacteria expressing GFP dsRNA (**a**) but not in animals that are defective for RNAi (**b**). Note that silencing occurs throughout the body of the animal, with the exception 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).

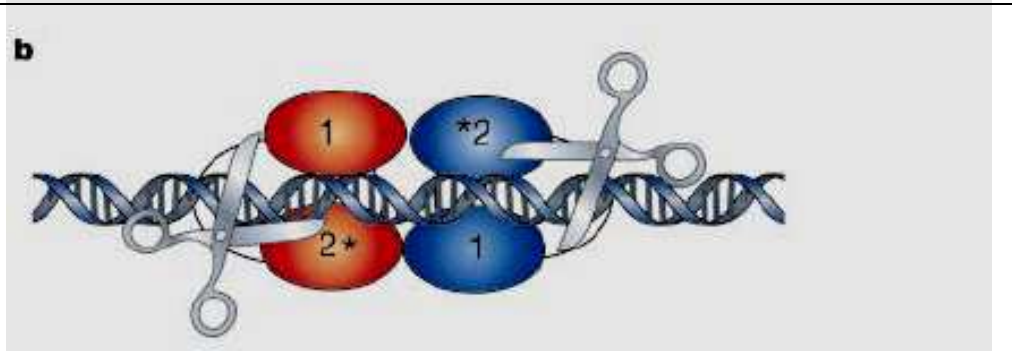
a, RNAi is initiated by the Dicer enzyme (two Dicer molecules with five domains each are shown), which processes double-stranded RNA into ~22-nucleotide small interfering RNAs. Based upon the known mechanisms for the RNase III family of enzymes, Dicer is thought to work as a dimeric enzyme. Cleavage 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.



dcr1

Ago1

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

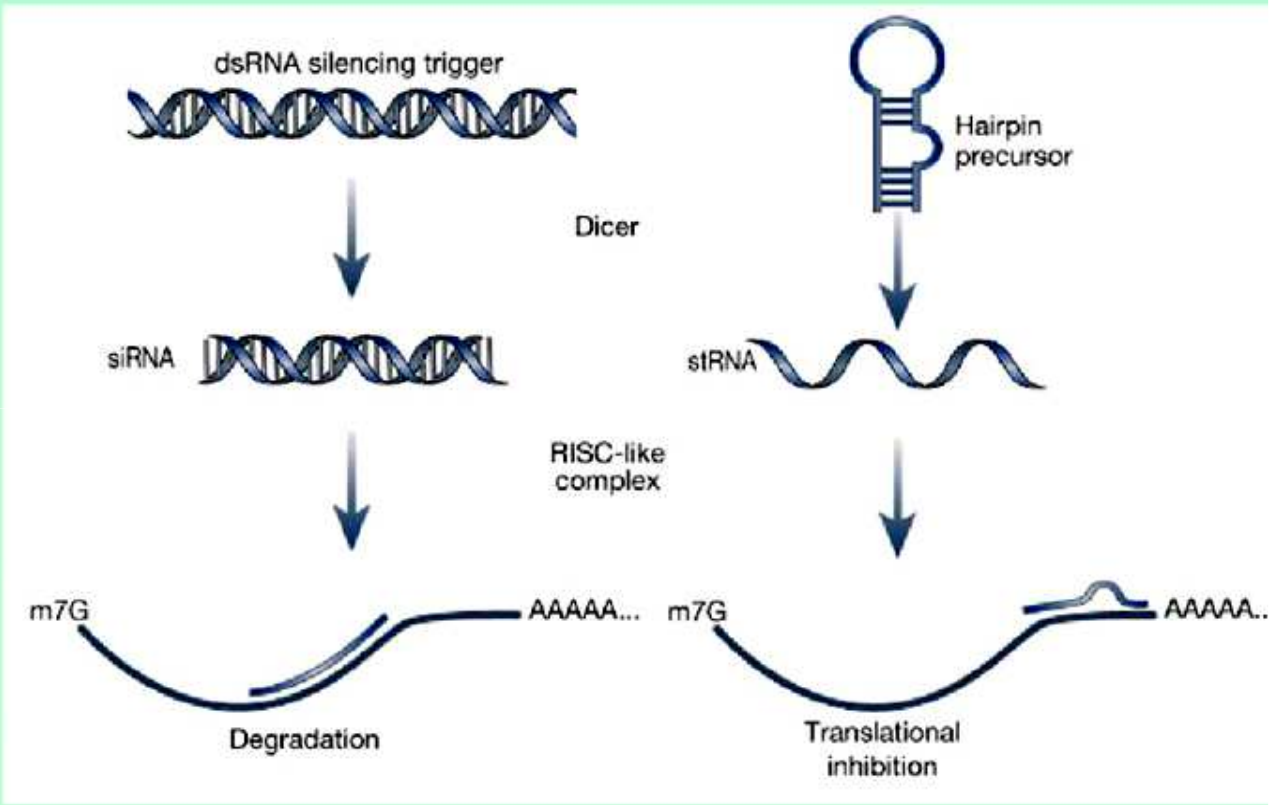
Micro RNA are a family of small RNA that are transcribed from several locations in genomes. The human genome may contain (estimated) up to 1,000 miRNA genes.

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).

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

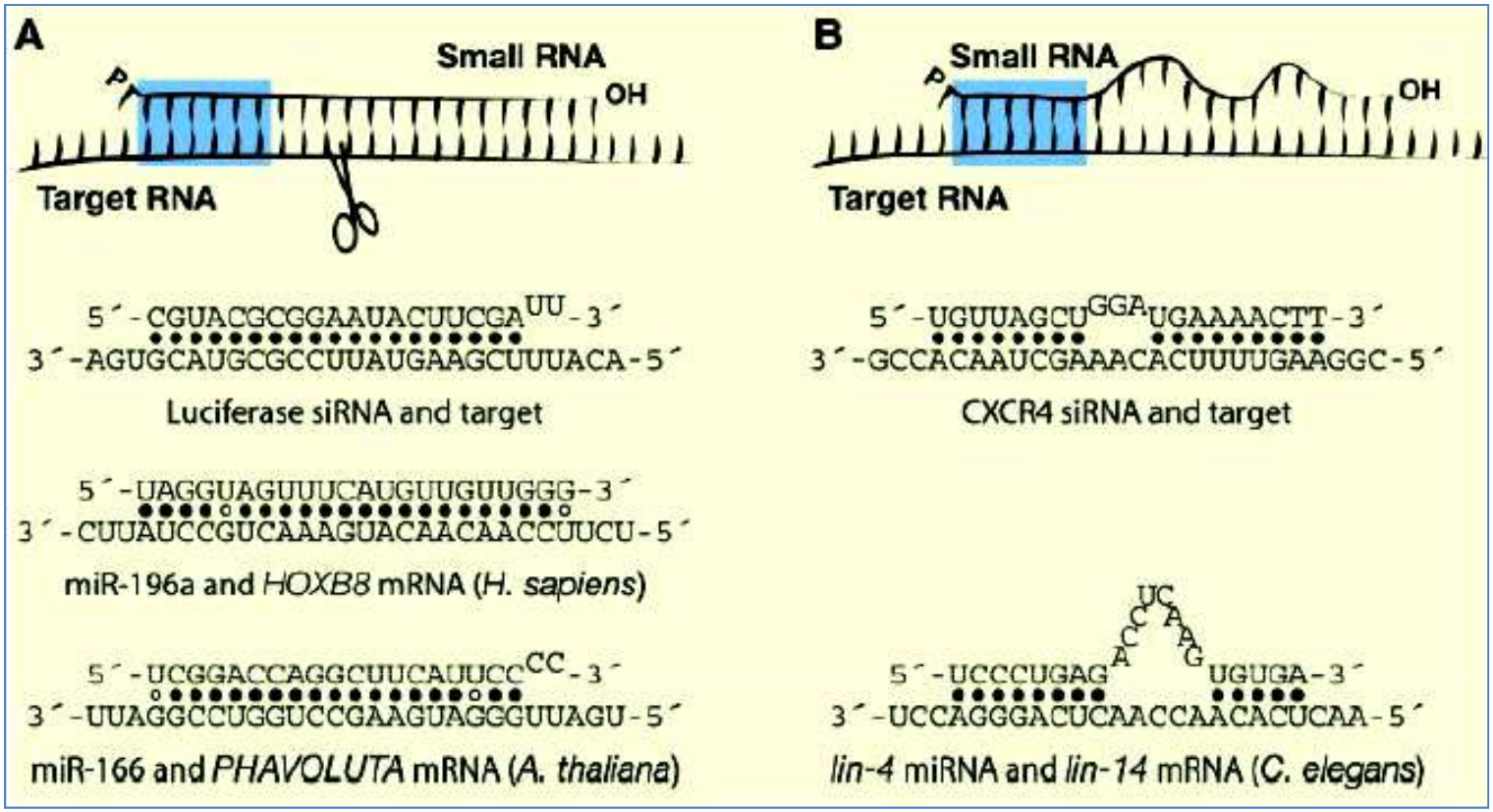


Fig. 3. Small RNA binding modes.
(A) **Extensive pairing** of a small RNA to an mRNA allows the Piwi domain of a catalytically active Argonaute protein (e.g., Ago2 in humans or flies) to cut a single phosphodiester bond in the mRNA, triggering its destruction. Synthetic siRNAs typically exploit this 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.

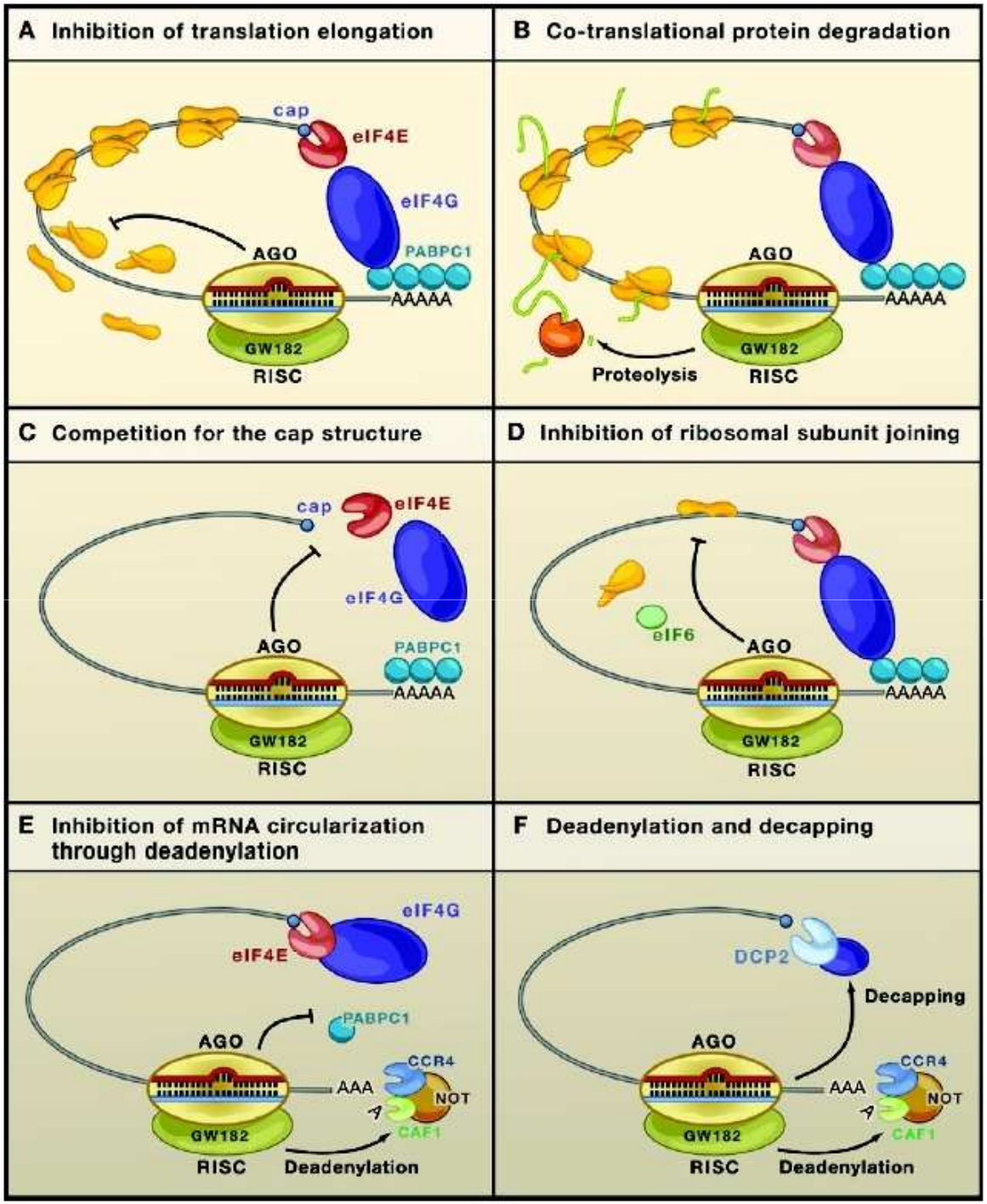


Figure 1. Mechanisms of miRNA-Mediated Gene Silencing

(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).

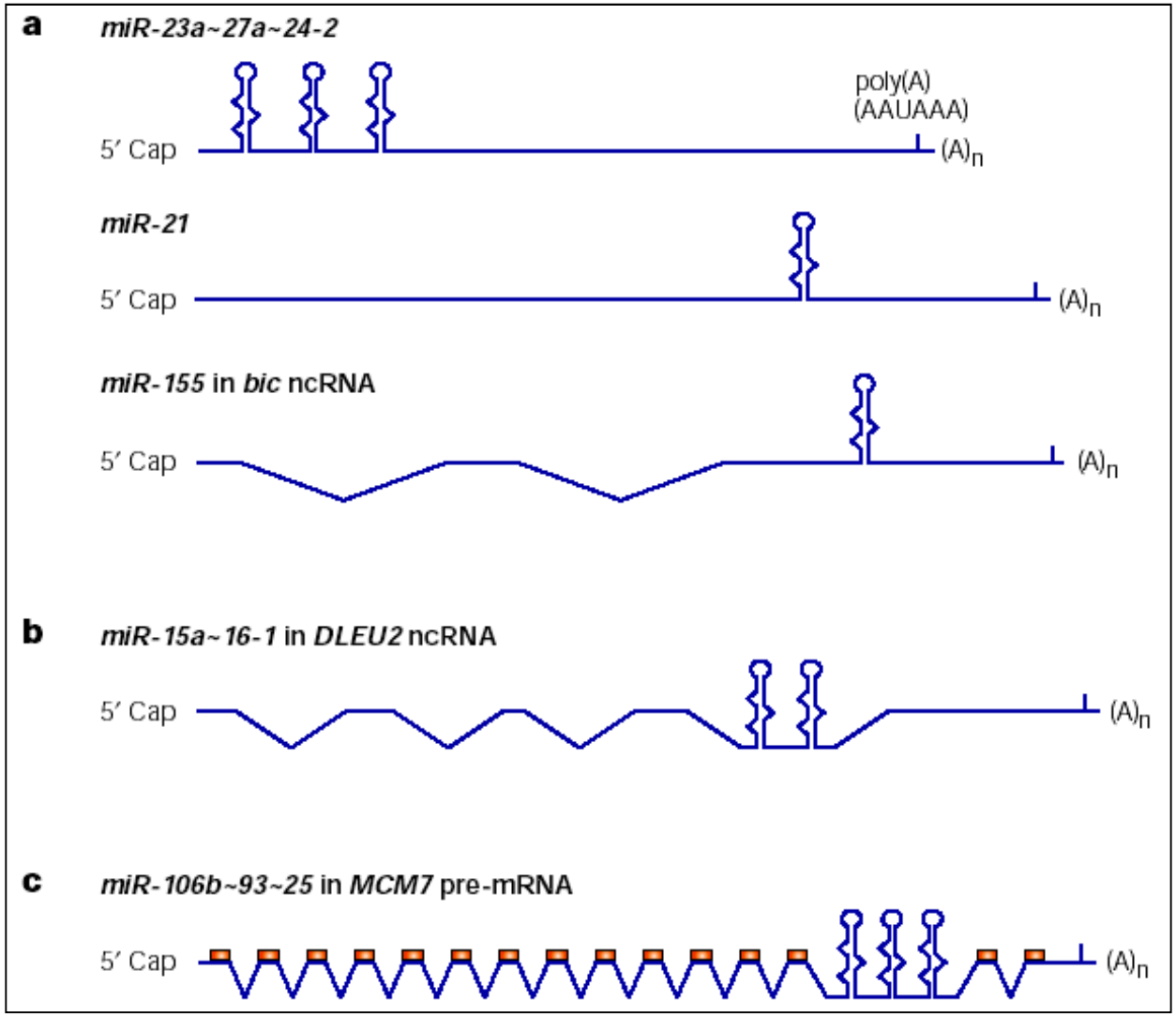


Figure 1 | **The structure of five pri-miRNAs.**

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.

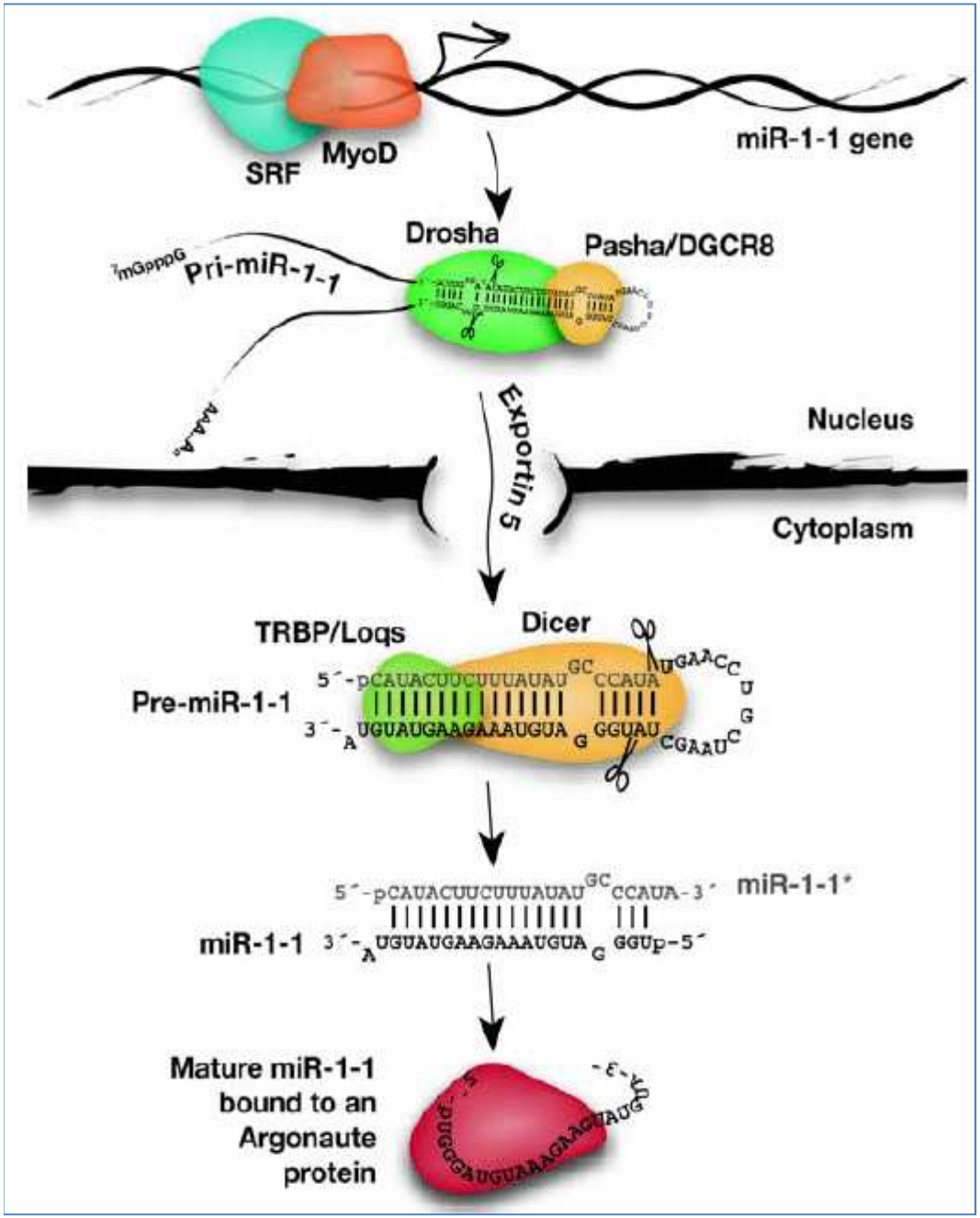


Fig. 2. A day in the life of the miRNA miR-1. In developing cardiac tissue, the trans-cription factors SRF (serum response factor) and MyoD promote RNA Pol II–directed transcription of pri-miR-1. In the nucleus, the RNase III endonuclease Drosha, together with its dsRNA-binding partner, Pasha/ /DGCR8, 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.

Small silencing RNAs: an expanding universe

Review

Megha Ghildiyal and Phillip D. Zamore

Abstract | Since the discovery in 1993 of the first small silencing RNA, a dizzying number of small RNA classes have been identified, including microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). These classes differ in their biogenesis, their modes of target regulation and in the biological pathways they regulate. There is a growing realization that, despite their differences, these distinct small RNA pathways are interconnected, and that small RNA pathways compete and collaborate as they regulate genes and protect the genome from external and internal threats.

Table 1 | **Types of small silencing RNAs**

Name	Organism	Length (nt)	Proteins	Source of trigger	Function	Refs
miRNA	Plants, algae, animals, viruses, protists	20–25	Drosha (animals only) and Dicer	Pol II transcription (pri-miRNAs)	Regulation of mRNA stability, translation	93–95, 200–202,226
casiRNA	Plants	24	DCL3	Transposons, repeats	Chromatin modification	38,44,51, 52,61–63
tasiRNA	Plants	21	DCL4	miRNA-cleaved RNAs from the *TAS* loci	Post-transcriptional regulation	64–68
natsiRNA	Plants	22	DCL1	Bidirectional transcripts induced by stress	Regulation of stress-response genes	71,72
		24	DCL2			
		21	DCL1 and DCL2			
Exo-siRNA	Animals, fungi, protists	~21	Dicer	Transgenic, viral or other exogenous dsRNA	Post-transcriptional regulation, antiviral defense	4,5,8,227
	Plants	21 and 24				
Endo-siRNA	Plants, algae, animals, fungi, protists	~21	Dicer (except secondary siRNAs in *C. elegans*, which are products of RdRP transcription, and are therefore not technically siRNAs)	Structured loci, convergent and bidirectional transcription, mRNAs paired to antisense pseudogene transcripts	Post-transcriptional regulation of transcripts and transposons; transcriptional gene silencing	75–79,82, 83,86,87, 200,201, 228
piRNA	Metazoans excluding *Trichoplax adhaerens*	24–30	Dicer-independent	Long, primary transcripts?	Transposon regulation, unknown functions	157, 163–169, 177,202
piRNA-like (soma)	*Drosophila melanogaster*	24–30	Dicer-independent	In *ago2* mutants in *Drosophila*	Unknown	76
21U-RNA piRNAs	*Caenorhabditis elegans*	21	Dicer-independent	Individual transcription of each piRNA?	Transposon regulation, unknown functions	114, 173–175
26G RNA	*Caenorhabditis elegans*	26	RdRP?	Enriched in sperm	Unknown	114

ago2, Argonaute2; casiRNA, *cis*-acting siRNA; DCL, Dicer-like; endo-siRNA, endogenous small interfering RNA; exo-siRNA, exogenous small interfering RNA; miRNA, microRNA; natsiRNA, natural antisense transcript-derived siRNA; piRNA, Piwi-interacting RNA; Pol II, RNA polymerase II; pri-miRNA, primary microRNA; RdRP, RNA-dependant RNA polymerase; tasiRNA, *trans*-acting siRNA.

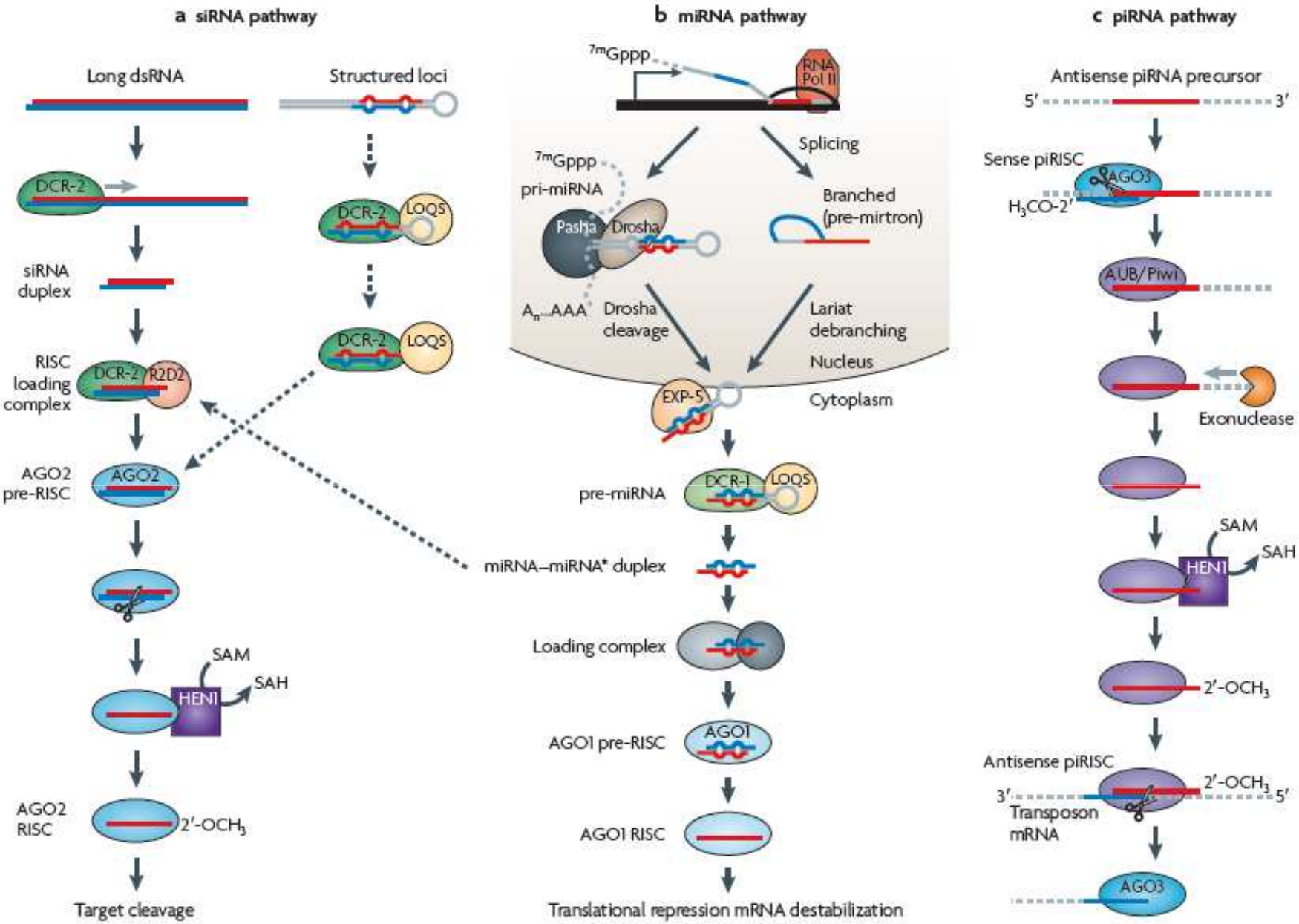
Both exo-siRNA and endo-siRNA

Variety and sources vary among different organisms

siRNA pathway: Ago2 is the main actor

Diversity among organisms: C. elegans has 27 different Argonaute proteins, D. melanogaster has 5, A. thaliana has 10.
Numbers of Dicer also vary, mammals have a single Dicer.

Choice between **guide** and **passenger** strains of siRNA are selected based on thermodynamic stability of 5' end (the higher → guide)
The cut is between nucleotides paired to positions 10 and 11 of the guide (AGO2).



Small RNA silencing pathways in Drosophila.

Plants exhibit a surprising diversity of small RNA types and the proteins that generate them.

In plants, inverted-repeat transgenes or coexpressed sense and antisense transcripts produce two sizes of siRNAs: 21 and 24 nucleotides. The DCl4-produced 21-mers typically associate with AGo1 and guide mRNA cleavage. The 24-mers associate with AGo4 (in the major pathway) and AGo6 (in the surrogate pathway), and promote the formation of repressive chromatin.

In plants, single-stranded sense transcripts from tandemly repeated or highly expressed single-copy transgenes are converted to dsRNA by RDR6, a member of the RNA-dependent RNA polymerase (RdRP) family that transcribe ssRNAs from an RNA template. RDR6 and RDR1 also convert viral ssRNA into dsRNA, initiating an antiviral RNAi response.

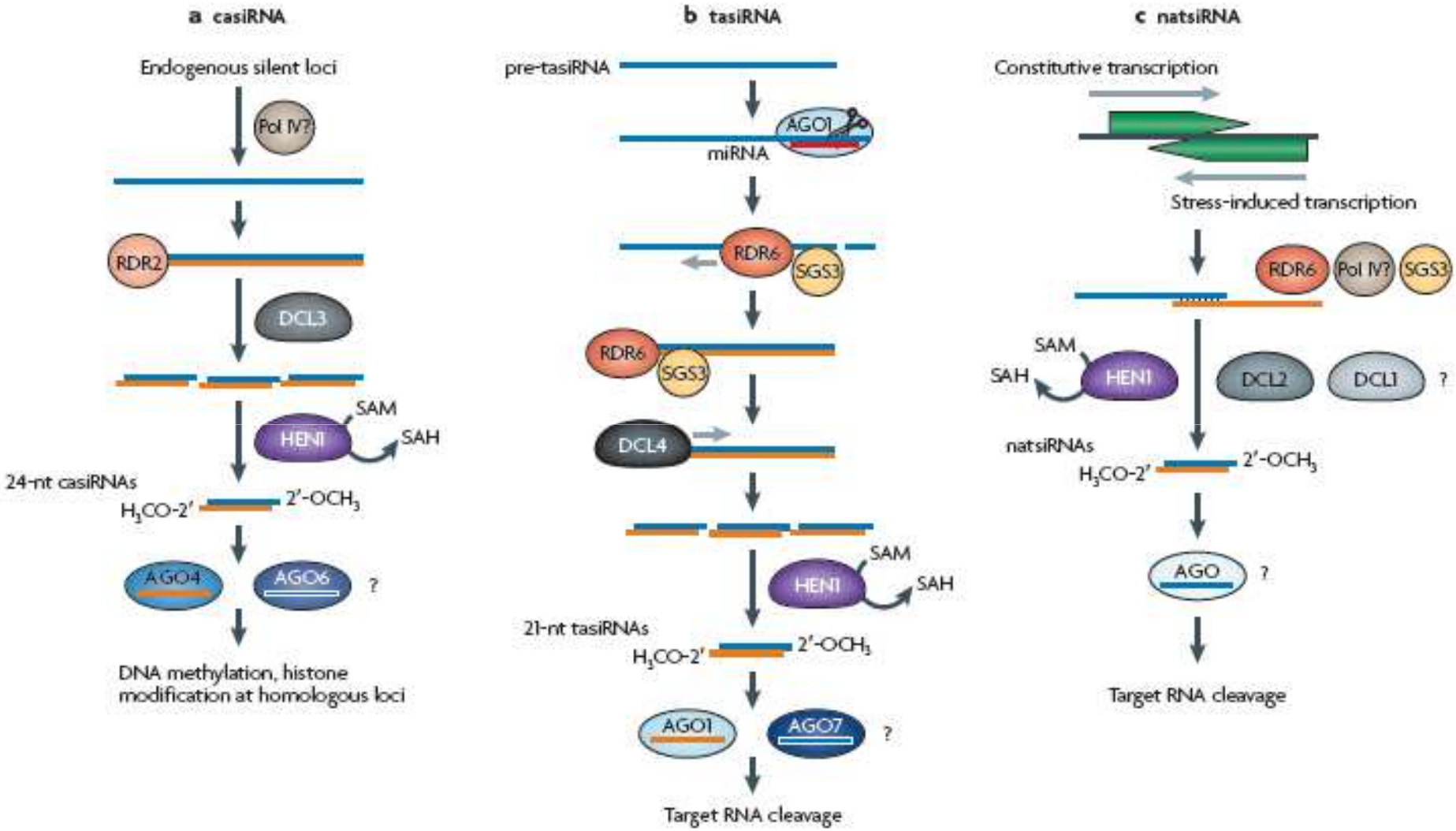


Figure 2 | **Plant endogenous small interfering RNA (endo-siRNA) biogenesis.** *Cis*-acting siRNAs (casiRNAs), *trans*-acting siRNAs (tasiRNAs) and natural antisense transcript-derived siRNAs (natsiRNAs) are derived from distinct loci. Several of the proteins involved in their biogenesis are genetically redundant, whereas others have specialized roles.

endo-siRNA

The first endo-siRNAs were detected in plants and *C. elegans, and the recent discovery of endo-siRNAs* in flies and mammals suggests that endo-siRNAs are ubiquitous among higher eukaryotes.

In many cases plant and worm endo-siRNA dependt upon RDRP activity.

The genomes of flies and mammals do not seem to encode such RdRP proteins, so the recent discovery of endo-siRNAs in flies and mice was unexpected.

The first mammalian endo-siRNAs to be reported corresponded to the long interspersed nuclear element (L1) retrotransposon and were detected in cultured human cells (2006).

More recently, endo-siRNAs have been detected in somatic and germ cells of *Drosophila species and in* mouse oocytes.
(most done by AGO2 immunoprecipitation followed by RNA-Seq)

Fly endo-siRNAs derive from transposons, heterochromatic sequences, intergenic regions, long RNA transcripts with extensive structure and, most interestingly, from mRNAs.

A subset of fly endo-siRNAs derives from ‘structured loci’, RNA transcripts of which can fold into long intramolecularly paired hairpins (intramolecular information) others from pseudogenes (trans).

endo-siRNAs have also been identified in mouse oocytes
(Tam et al., Nature 453: 534-538, 2008; Watanabe et al., Nature 453: 539-543, 2008).

As in flies, mouse endo-siRNAs are 21 nucleotides, Dicer-dependent and derived from a variety of genomic sources

A subset of mouse oocyte endo-siRNAs maps to regions of protein-coding genes that are capable of pairing to their cognate pseudogenes, and to regions of pseudogenes that are capable of forming inverted repeat structures.
(Interestingly, pseudogenes can no longer encode proteins, but they drift from their ancestral sequence more slowly than would be expected if they were simply junk DNA).

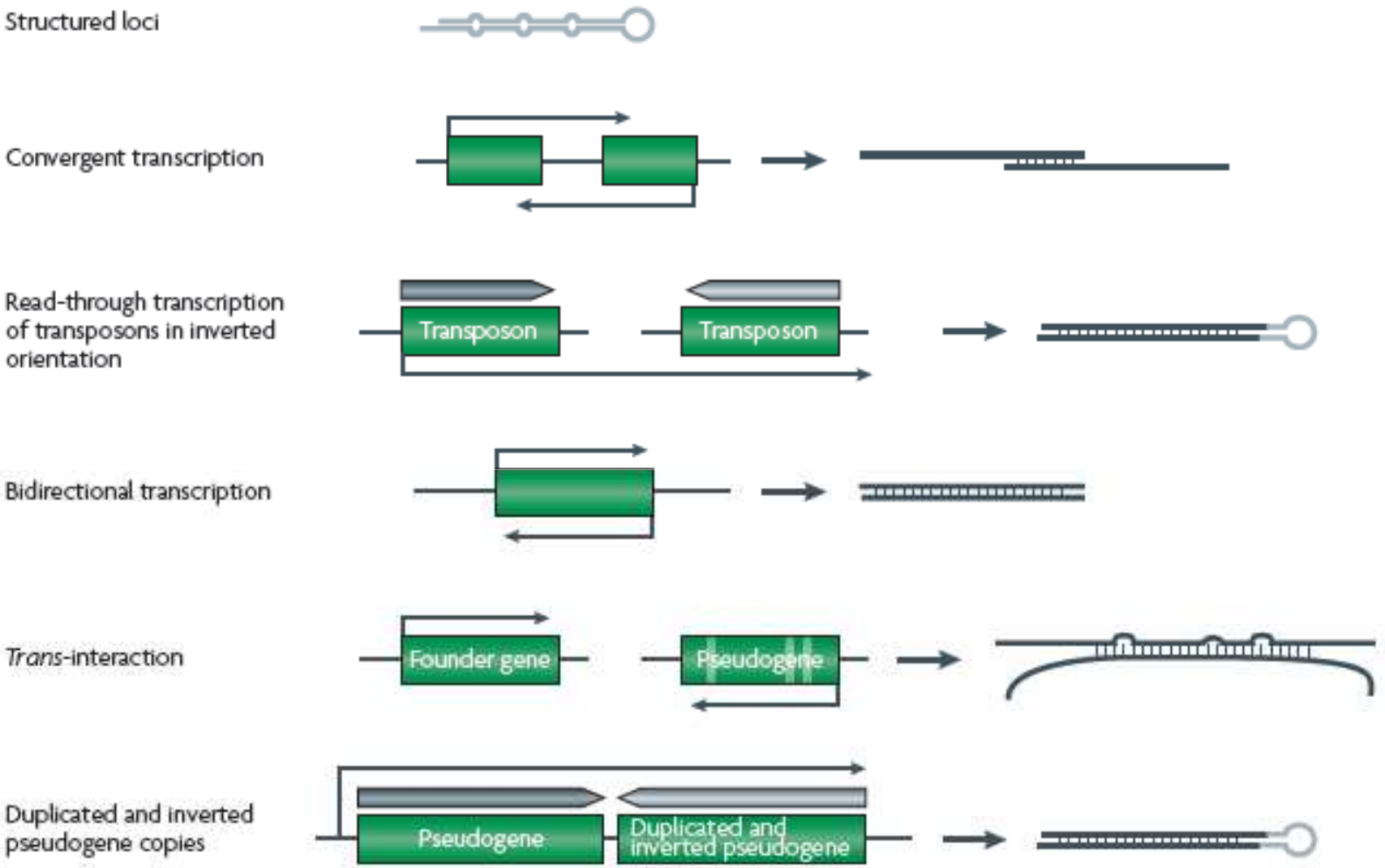


Figure 3 | **Genomic sources of dsRNA triggers for endogenous small interfering RNAs (endo-siRNAs) in flies and mammals**

piRNA are small RNAs associated to the Piwi-subfamily of Argonaute proteins.

piRNAs were first proposed to ensure germline stability by repressing transposons when Aravin and colleagues discovered in flies a class of longer small RNAs (~25–30 nucleotides) associated with silencing of repetitive elements

Mammalian piRNAs can be divided into pre-pachytene and pachytene piRNAs, according to the stage of meiosis at which they are expressed in developing spermatocytes. like piRNAs in flies, pre-pachytene piRNAs predominantly correspond to repetitive sequences and are implicated in silencing transposons, such as L1 and intracisternal A-particle

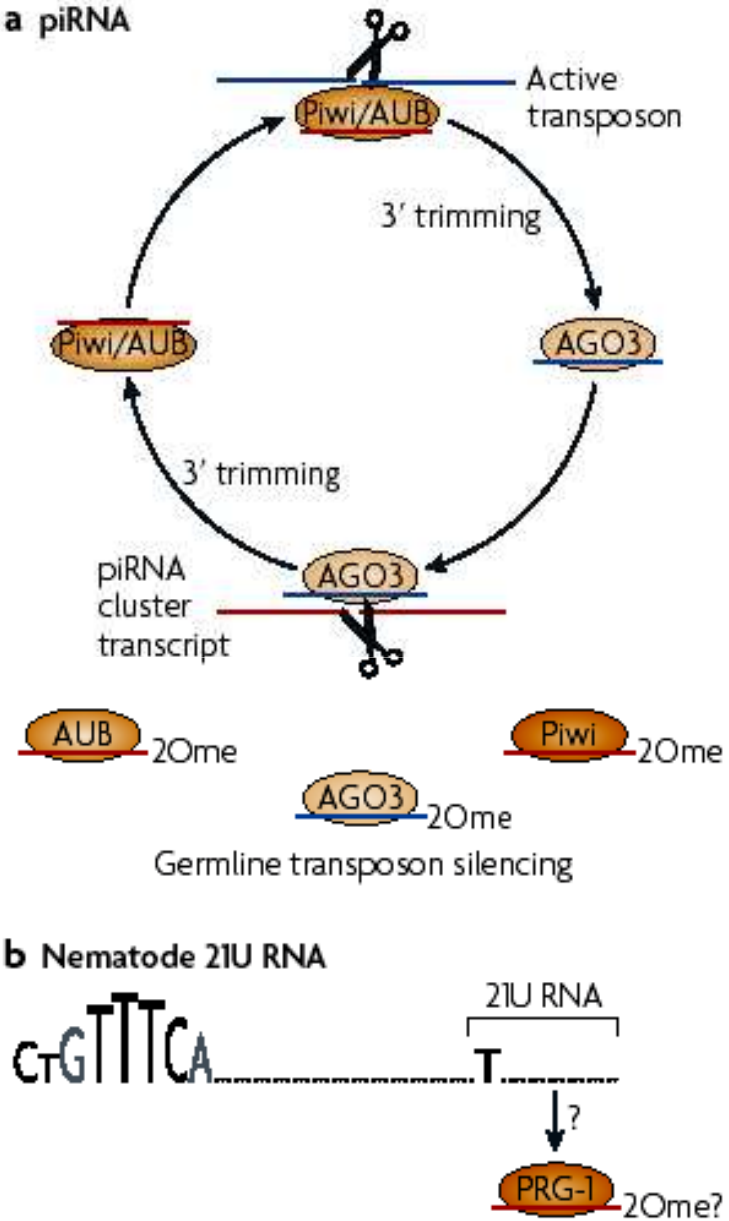


Figure 2 | **Specialized small-RNA regulatory pathways in the animal germ line.** These are mediated by Piwi-class Argonaute proteins (ovals). **a** | The Piwi-interacting (pi)RNA pathway operates in the *Drosophila melanogaster* and vertebrate germ line. A 'ping-pong' strategy amplifies piRNAs from complementary transcripts, in which the slicer activity of Piwi proteins (Piwi, Aubergine (AUB) and AGO3 in *D. melanogaster*) reciprocally define piRNA 5′ ends. The mechanism that defines the 3′ ends of piRNAs is not known. A conserved role of the piRNA pathway is to restrict transposon activity in the germ line; however, there might be other roles for abundant non-transposon-derived piRNAs that are found in mammals. **b** | Nematode 21U RNAs might be a functional analogue of piRNAs. These 21-nucleotide RNAs begin with U and are produced from genomic loci with a characteristic upstream motif (CTGTTTCA), and they are bound by the Piwi protein PRG-1. The details of 21U biogenesis and function are unclear, but 21Us are linked to spermatogenesis and control of Tc3 transposition. 2Ome, 2′-*O*-methyl group.

Since the “seeding” sequence is very short (6-8nt), potential targets for known miRNA can be identified in the 3’-UTR of hundreds of genes each.

However, experiment using either trasfection of miRNA in cultured cells or knock-down of endogenous miRNA function by “antagomir” , followed by gene expression profiling with microarrays, demonstrated that a limited number of targets exist for each miRNA, and that (in fewer cases) new unidentified targets may exist.

Many studies have shown that several context-dependent factors are important :

1) the number of miRNA targets / 3’UTR
2) cooperativity with different miRNA
3) position of the targets
4) RNA-binding sites

Transcription of miRNA encoding genes is made primarily by RNA Pol II and is controlled in a very similar way as protein-encoding genes, i.e. it depends on the same transcription factors and co-regulators.

This is evident, of course, also for miRNA that are embedded in introns of protein-encoding genes, that are co-regulated with the host gene.

This means that we may evidence “circuits” of control, where a specific Transcription factors controls a5t the same time transcription of a miRNA and of its target (targets) mRNA(s).

From a bioinformatic point of view, this is very evident.

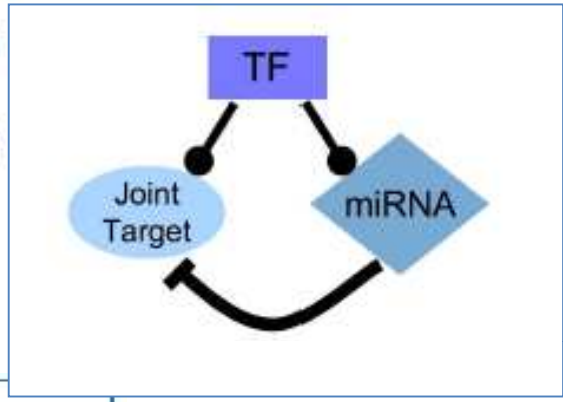
See the paper that follows:

Friard et al. BMC Bioinformatics 2010, 11:435
http://www.biomedcentral.com/1471-2105/11/435

BMC Bioinformatics

DATABASE Open Access

CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse



Olivier Friard[1], Angela Re[2], Daniela Taverna[1,3,4], Michele De Bortoli[1,3], Davide Corá[1,5*]

Abstract

Background: Transcription Factors (TFs) and microRNAs (miRNAs) are key players for gene expression regulation in higher eukaryotes. In the last years, a large amount of bioinformatic studies were devoted to the elucidation of transcriptional and post-transcriptional (mostly miRNA-mediated) regulatory interactions, but little is known about the interplay between them.

Description: Here we describe a dynamic web-accessible database, `CircuitsDB`, supporting a genome-wide transcriptional and post-transcriptional regulatory network integration, for the human and mouse genomes, based on a bioinformatic sequence-analysis approach. In particular, `CircuitsDB` is currently focused on the study of mixed miRNA/TF Feed-Forward regulatory Loops (FFLs), i.e. elementary circuits in which a master TF regulates an miRNA and together with it a set of Joint Target protein-coding genes. The database was constructed using an ab-initio oligo analysis procedure for the identification of the transcriptional and post-transcriptional interactions. Several external sources of information were then pooled together to obtain the functional annotation of the proposed interactions. Results for human and mouse genomes are presented in an integrated web tool, that allows users to explore the circuits, investigate their sequence and functional properties and thus suggest possible biological experiments.

Conclusions: We present `CircuitsDB`, a web-server devoted to the study of human and mouse mixed miRNA/TF Feed-Forward regulatory circuits, freely available at: http://biocluster.di.unito.it/circuits/

Besides the quite well charcterized miRNA, siRNA and piRNA (and related) pathways, there are several other **uncharacterized transcripts**

They belong both to “intragenic” and “intergenic” cathegory

and are either short RNA (<200nt) or long RNA (up to several megabases !!!)

Of paricular interests are all the transcripts that tiling microarrays and RNA-Seq experiments have revealed around the promoter and at the end of known protein coding genes, that are illustrated and discussed in the review that follows and in the previous one, and depicted in the next figures.

DNA Research **17**, 51–59, (2010)
Advance Access Publication: 7 March 2010
doi:10.1093/dnares/dsq006

RNA Dust: Where are the Genes?

Review

Piero Carninci*

Omics Science Center, RIKEN Yokohama Institute, Kanagawa, Japan

*To whom correspondence should be addressed. Tel. +81 45-503-9331. Fax. +81 45-503-9216.
Email: carninci@riken.jp

Edited by Osamu Ohara
(Received 24 December 2009; accepted 5 February 2010)

Abstract

Initial gene discovery efforts through analysis of genome sequences and identification and characterization of expressed RNAs have revealed that only a relatively small portion of the genome is transcribed into protein coding mRNAs in vertebrates. However, in contrast with this paucity of protein coding 'genes', there is an enormous complexity in transcription and the protein coding mRNAs contribute to a very small fraction of transcripts in comparison with the different varieties of non-coding RNAs (ncRNAs). This transcriptome complexity may be hypothesized to have a regulatory role that is required for the development and function of organisms as complex as vertebrates. At the same time, it raises the fundamental question of the unequivocal definition of a gene. It is intriguing to postulate that many ncRNAs might finely modulate gene activity by acting as regulatory elements. The emerging hypotheses suggest that the gene regulatory machinery may be deeply interconnected with the world of short RNAs. These RNAs may generally act for fine-tuning of the protein-coding transcriptome.

Key words: transcriptome; non-coding RNAs; sense–antisense transcription; cDNA annotation; RNA

Table 1. Definition of RNAs classes discussed in this review

Short name of RNA classes	Full name of RNA classes	Notes
PALRs	Promoter-associated long RNAs	Hundreds nt long RNAs spanning regions on proximal promoters to the first exon
PASRs	Promoter-associated short RNAs	20–70 nt long RNAs spanning regions around core promoters
TASRs	Termini-associated short RNAs	20–70 nt long RNAs spanning regions around transcription termination sites
PROMPTs	Promoter upstream transcripts	Unstable transcripts mapping 0.5–2 kb upstream the transcription starting sites
TSSa-RNAs	Transcription start sites antisense RNAs	RNAs, generally short and non-coding, generated from bidirectional activity of mammalian RNA Polymerase II
NRO-RNAs	Nuclear run-on assay derived RNAs	Short RNA detected by nuclear run-on assays, mapping 20 to 50 downstream to transcriptions starting sites of mRNAs
RE RNAs	Retrotransposon-derived RNAs	A heterogeneous class of RNAs which starting sites overlap retrotransposon elements
tiRNAs	Tiny transcription initiation RNAs	RNAs about 18 nt long, positioned about 20 bp after the transcription starting sites of highly expressed mRNAs
snoRNAs	Small nucleolar RNAs	Small ncRNAs that guide chemical modifications of other non-coding RNAs
siRNAs	Small interfering RNAs	Double-stranded RNA molecules, 20–25 nucleotides in length, that act in various silencing pathways
miRNAs	microRNAs	Single-stranded RNA molecules of 21–24 nucleotides in length, which regulate gene expression

continued

Table 1. Definition of RNAs classes discussed in this review

Short name of RNA classes	Full name of RNA classes	Notes
LincRNAs	Large intervening non-coding RNAs	Large non-coding RNAs that map in intergenic locations
ncRNAs	Non-coding RNAs	Generic definition for non-protein coding RNAs
sRNA	Short RNAs	Generic definition for short RNAs
snRNA	Small nuclear RNAs	Nuclear small non-coding RNAs involved in various functions including splicing
piRNA	Piwi interacting RNAs	26–31 nt long RNAs involved in transcriptional gene silencing, including retrotransposons

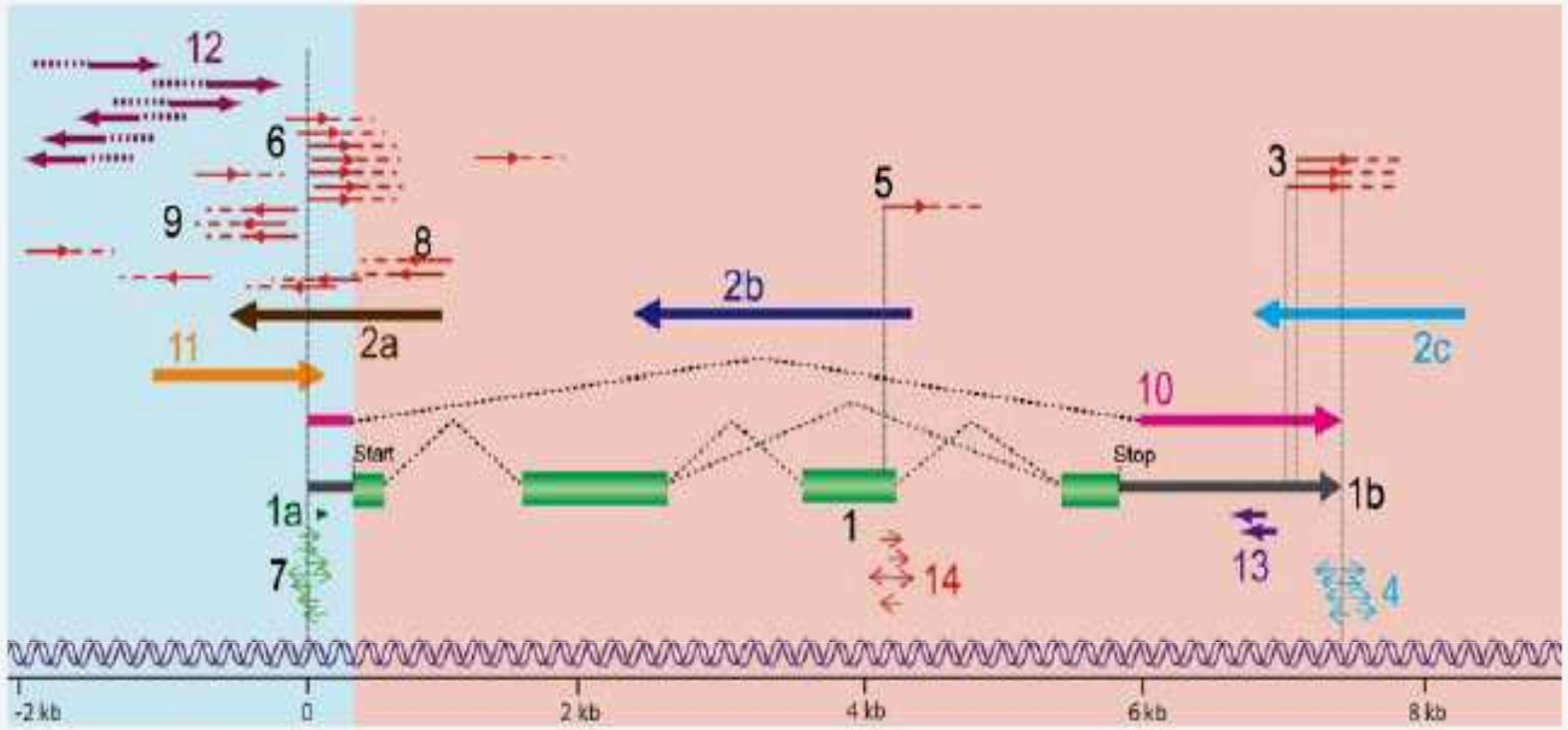
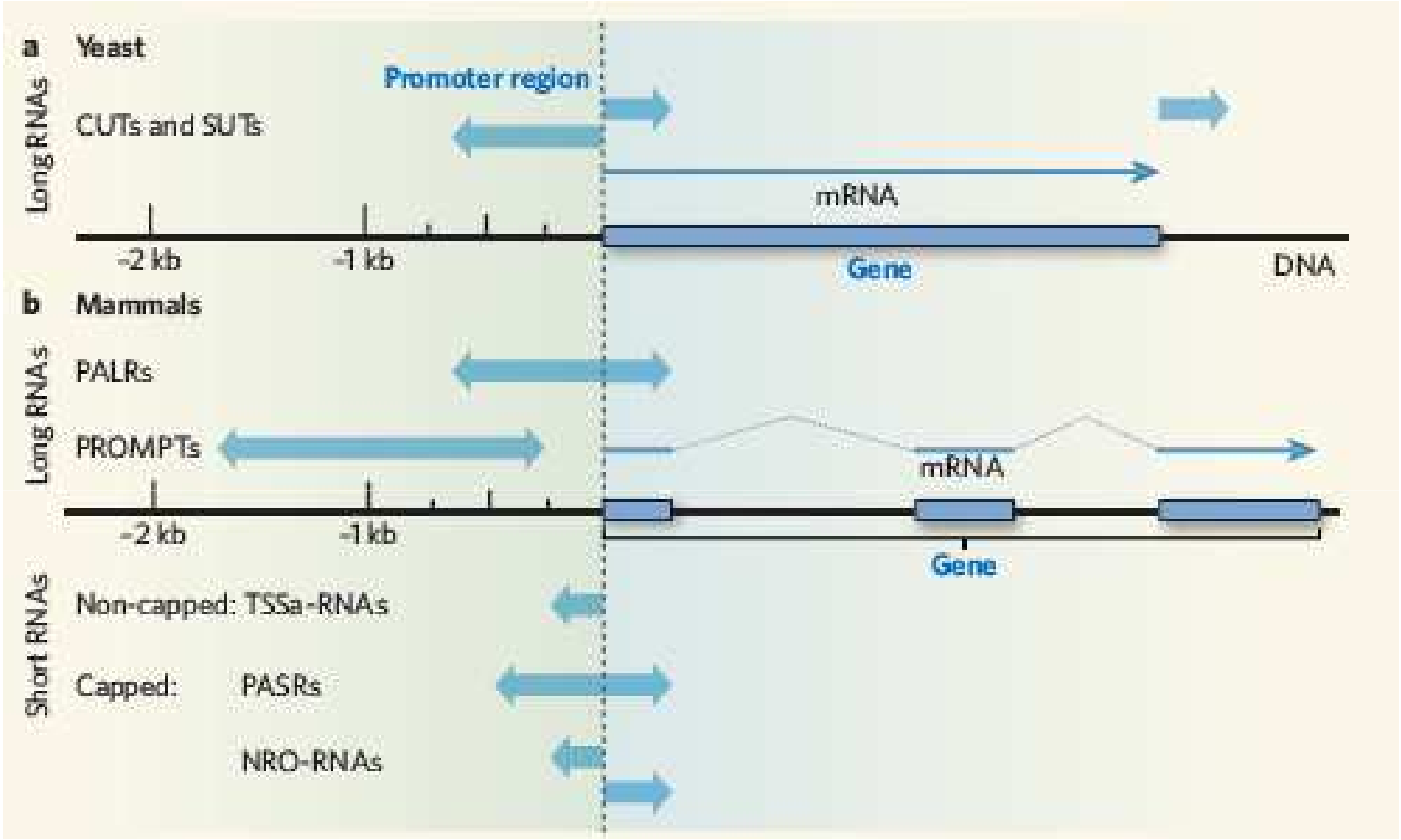


Figure 1. Complexity of transcriptome around a hypothetical gene. Green boxes represent exons of typical protein coding gene. Not all the sRNAs and genomic elements are in scale. CAGE tags (red lines; dots beyond the arrowheads indicate than the 3′ ends are unknown) identify TSSs or other capped molecules; the dashed lines on CAGE tags indicate that the 3′ end is not determined. Various types of transcripts are indicated by numbers: (1) protein coding mRNA transcript gene (green boxes: coding exons, gray lines, 5′ and 3′ untranslated regions); (2, a–c), antisense RNAs in various relation with the transcripts (3′–3′ overlap, full overlap, 5′–5′ overlap); (3) CAGE tags identify transcript in the 3′-UTRs, likely polyadenylated; (4) termination-associated sRNAs (TASRs); (5) exonic long-capped transcripts; (6) CAGE tags identifying TSS (exact location can vary) and may overlap PALRs; (7) PASRs (green) and tiny 18 nt long RNAs (tiRNAs, arrowhead only); (8) antisense transcription events detected by CAGE, often (but not limited to) the first exons and introns; (9) bidirectionally transcribed RNAs from core promoters; (10) ncRNA splicing isoforms only partially overlapping to coding mRNA sequences; (11) PALRs; (12) PROMPTs, unstable transcripts on upstream regulatory regions; (13) miRNAs and endogenous siRNAs (deriving mostly from other loci, not perfectly matching in most cases in animal cells); (14) other sRNAs associated to exonic-capped long transcripts. The list of different types of RNAs is continuously growing and subject to revisions and further classifications.


a
Yeast
Promoter region
Long RNAs
CUTs and SUTs
mRNA
-2 kb
-1 kb
Gene
DNA
b
Mammals
Long RNAs
PALRs
PROMPTs
mRNA
-2 kb
-1 kb
Gene
Short RNAs
Non-capped: TSSa-RNAs
Capped:
PASRs
NRO-RNAs

What about functions?
Especially for RNA “around genes” a regulatory role on trascription of the protein-coding gene is suggested.
Several papers have evidenced that these RNA may guide in proximity of the gene proteins that organize heterochromatin, resulting in silencing of the gene.
Other, on the contrary, have shown that a noncoding RNA can function as co-activator of transcription of the nearest gene, by binding to, and activate, a specific coactivating protein.

Long non-coding RNAs: insights into functions

Tim R. Mercer, Marcel E. Dinger and John S. Mattick

Abstract | In mammals and other eukaryotes most of the genome is transcribed in a developmentally regulated manner to produce large numbers of long non-coding RNAs (ncRNAs). Here we review the rapidly advancing field of long ncRNAs, describing their conservation, their organization in the genome and their roles in gene regulation. We also consider the medical implications, and the emerging recognition that any transcript, regardless of coding potential, can have an intrinsic function as an RNA.

regulating the expression of neighbouring protein-coding genes. The importance of this localized regulation was foreshadowed by the phenomenon of 'transvection', in which non-coding loci affect the expression of nearby protein-coding genes in *trans*[19].

Chromatin modification. Long ncRNAs can mediate epigenetic changes by recruiting chromatin remodelling complexes to specific genomic loci. For example, hundreds of long ncRNAs are sequentially expressed along the temporal and spatial developmental axes of the human homeobox (Hox) loci, where they define chromatin domains of differential histone methylation and RNA polymerase accessibility[21]. One of these ncRNAs, Hox transcript antisense RNA (*HOTAIR*), originates from the *HOXC* locus and silences transcription across 40 kb of the *HOXD* locus in *trans* by inducing a repressive chromatin state, which is proposed to occur by recruitment of the Polycomb chromatin remodelling complex PRC2 by *HOTAIR*[21]

A short list of recent papers suggesting functions for long noncoding RNAs.

36. Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.
37. Mondal, T., Rasmussen, M., Pandey, G.K., Isaksson, A. and Kanduri, C. (2010) Characterization of the RNA content of chromatin. *Genome Res.*, **20**, 899–907.
38. Kanhere, A., Viiri, K., Araujo, C.C., Rasaiyaah, J., Bouwman, R.D., Whyte, W.A., Pereira, C.F., Brookes, E., Walker, K., Bell, G.W. *et al.* (2010) Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol. Cell*, **38**, 675–688.
39. Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E. and Chang, H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.
40. Szutorisz, H., Dillon, N. and Tora, L. (2005) The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem. Sci.*, **30**, 593–599.
41. Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.

One of the main problems in assigning a function to noncoding RNA is that they are not well defined by the method used.
Both tiling arrays and RNA –Seq, indeed, give millions of “short reads”: sometimes they cover entirely and continuously a genome region: in this case it is easy to understand where the RNA strats and ends and how it is composed.
In other cases (frequently with those low-level noncoding) the coverage is discontinuous, so that the reseacher can only “guess” how the RNA is organized.

Sequencing techniques must evolve !!! Weren’t they spectacular enough, were they ?

As a matter of fact, NGS are spectacular in terms of “throughput” (millions of reads per experiment) but are very limited since they sequence only shortly after the primer.

A method capable of sequencing long to very long stretches of DNA is eagerly awaited.....

See this paper, for example:

Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach,† Stephen Turner†

We present single-molecule, real-time sequencing data obtained from a DNA polymerase performing uninterrupted template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleoside triphosphates (dNTPs). We detected the temporal order of their enzymatic incorporation into a growing DNA strand with zero-mode waveguide nanostructure arrays, which provide optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions. Conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows continuous observation of DNA synthesis over thousands of bases without steric hindrance. The data report directly on polymerase dynamics, revealing distinct polymerization states and pause sites corresponding to DNA secondary structure. Sequence data were aligned with the known reference sequence to assay biophysical parameters of polymerization for each template position. Consensus sequences were generated from the single-molecule reads at 15-fold coverage, showing a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates.

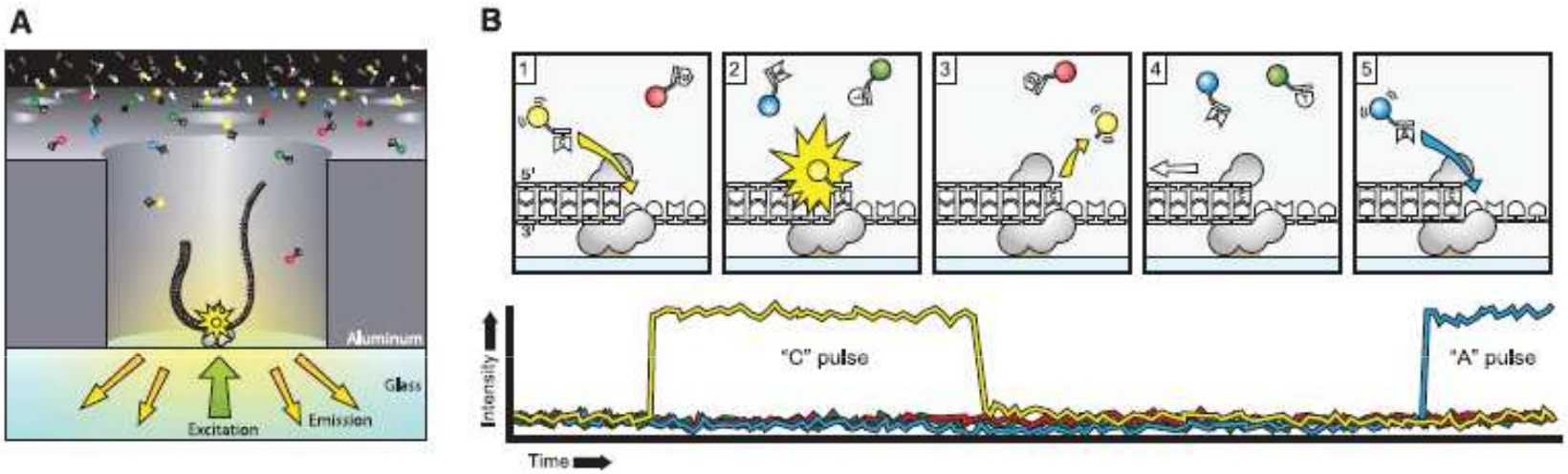


Fig. 1. Principle of single-molecule, real-time DNA sequencing. (**A**) Experimental geometry. A single molecule of DNA template-bound Φ29 DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10^{-21} liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. (**B**) Schematic event sequence of the phospholinked dNTP incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.

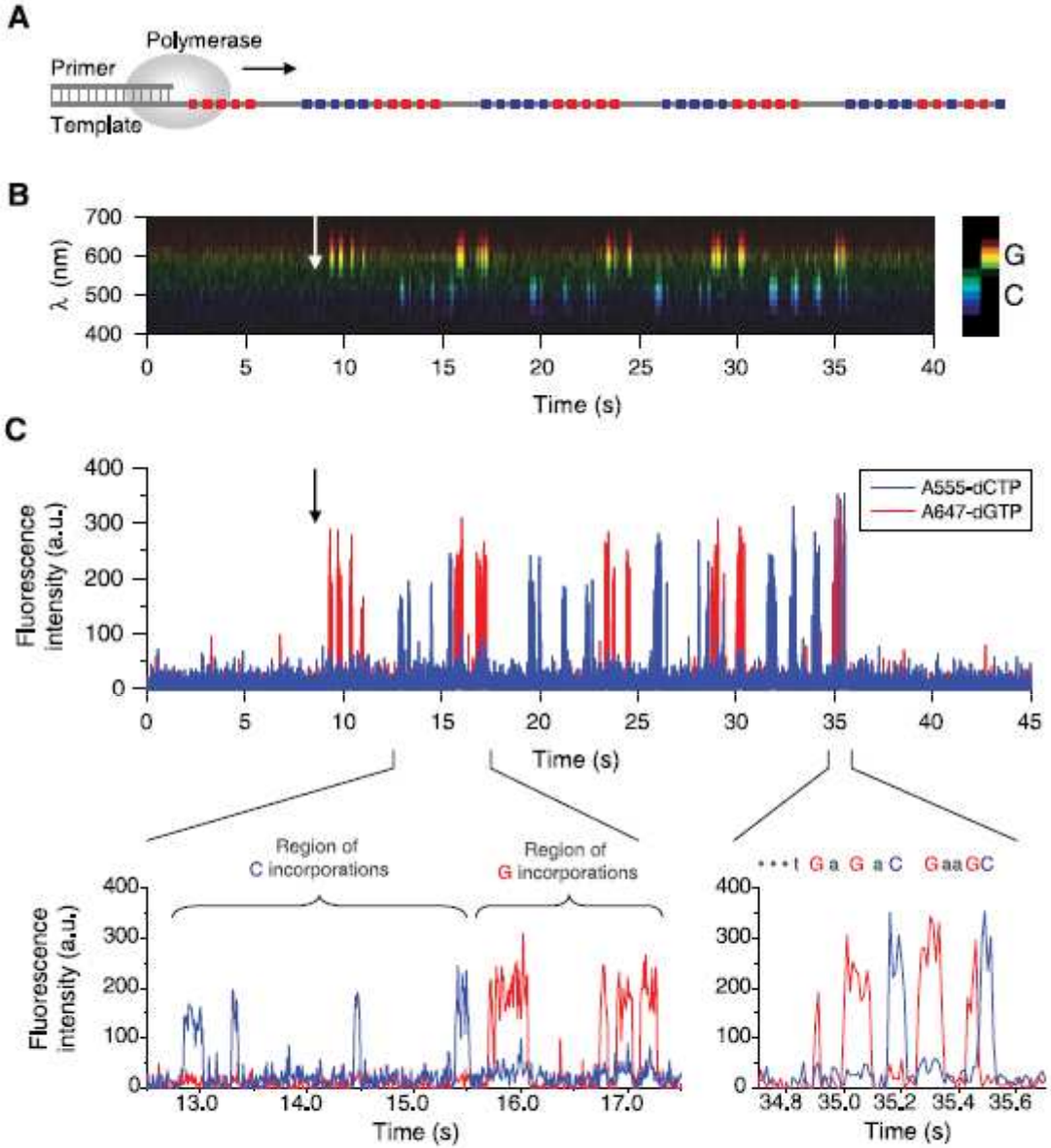


Fig. 2. Real-time detection of single-molecule DNA polymerase activity. (**A**) DNA template design for two-base sequence pattern detection. The sequence of a linear, single-stranded DNA template was designed to yield incorporation of alternating blocks of two phospholinked nucleotides (A555-dCTP and A647-dGTP), interspersed with the other two, unmodified dNTPs. (**B**) Time-resolved fluorescence intensity spectrum from a ZMW. Data from a 15 × 5 pixel area from each movie frame were spatially collapsed to a 15-pixel spectrum, which is shown as a function of time. The expected fluorescence emission profiles for the two labeled nucleotides are shown at the right. The arrow denotes addition of the catalytic metal ion that initiated the polymerization reaction. The complete data set from which the time trace was extracted

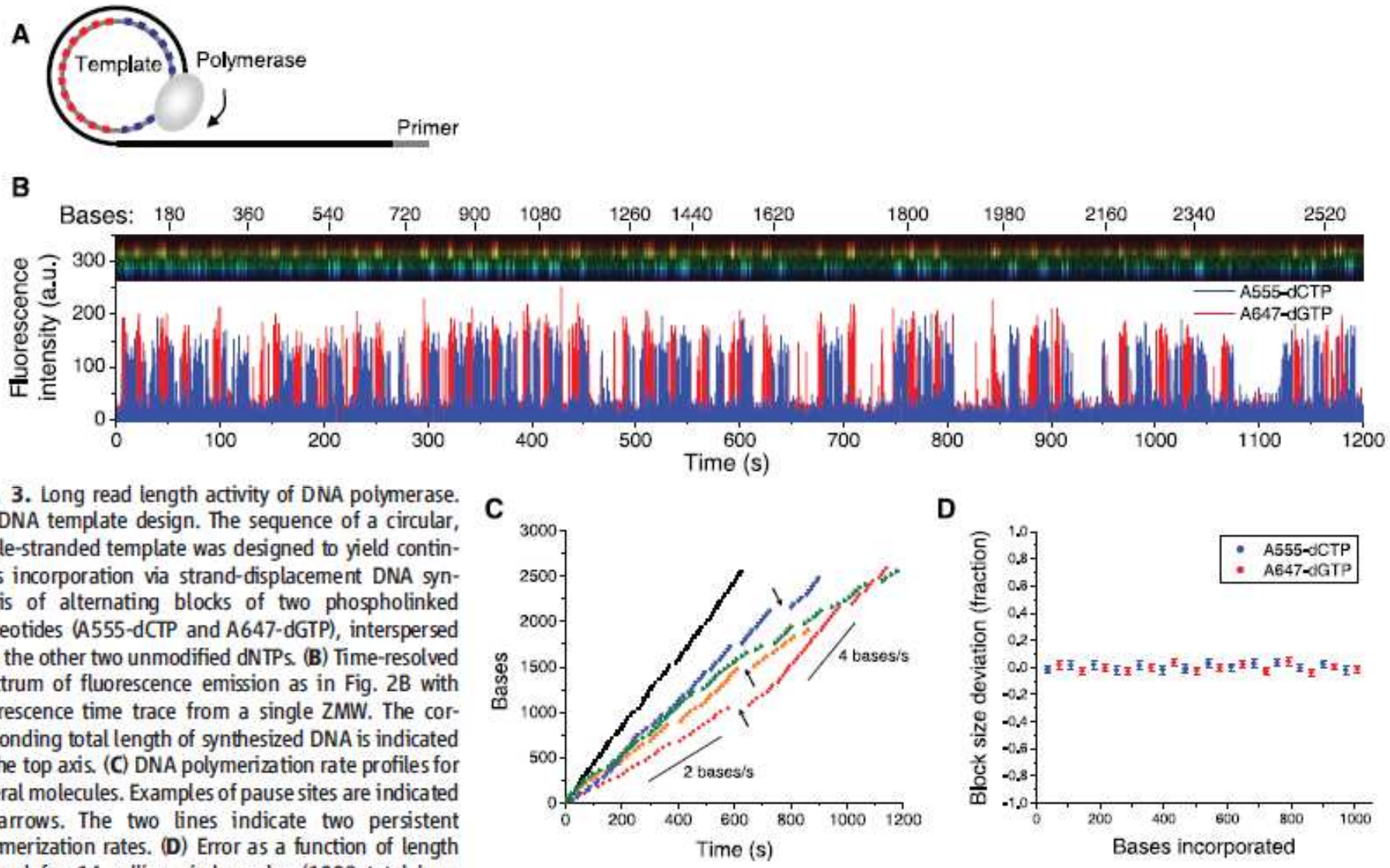


Fig. 3. Long read length activity of DNA polymerase. (**A**) DNA template design. The sequence of a circular, single-stranded template was designed to yield continuous incorporation via strand-displacement DNA synthesis of alternating blocks of two phospholinked nucleotides (A555-dCTP and A647-dGTP), interspersed with the other two unmodified dNTPs. (**B**) Time-resolved spectrum of fluorescence emission as in Fig. 2B with fluorescence time trace from a single ZMW. The corresponding total length of synthesized DNA is indicated by the top axis. (**C**) DNA polymerization rate profiles for several molecules. Examples of pause sites are indicated by arrows. The two lines indicate two persistent polymerization rates. (**D**) Error as a function of length of read for 14 rolling circle cycles (1008 total base incorporations; n = 186 reads). The fractional deviation from the average number of pulses per block (12 A555-dCTP and 12 A647-dGTP observed phospholinked dNTP pulses per cycle, respectively), mean ± SE, is plotted as a function of template position. The 95% confidence interval for the slope is −0.027 to +0.036 blocks per 1008 bases of incorporation.