NEWS & VIEWS

GENOMICS

Protein fossils live on as RNA

Rajkumar Sasidharan and Mark Gerstein

Pseudogenes constitute many of the non-coding DNA sequences that make up large parts of genomes. Once considered merely protein fossils, it now emerges that some of them have active regulatory roles.

A central challenge in genome annotation is determining the function of sequences that do not encode proteins, but make up the overwhelming bulk of large genomes — some 99% in humans. A significant fraction of these sequences are pseudogenes, or fossils of ancient proteins, and although many of them are transcribed into RNA, they have hitherto been deemed 'junk'. However, given the abundance of pseudogenes, it is unlikely that they are useless. One function suggested for them is gene regulation, and RNA interference (RNAi) has been proposed as the mechanism for carrying this out. Six papers[1–6], including three in this issue (pages 793, 798 and 803), significantly expand the known scope of RNAi by describing the discovery of natural small interfering RNA (siRNA) sequences in mice and fruitflies, some of which are potentially transcribed from pseudogenes.

The textbook definition of a pseudogene is an inheritable genetic element that is similar to a functioning gene, yet is non-functional. But what is meant by non-functional is debatable — not transcribed, not translated, or not under control of a promoter sequence? Pseudogenes are similar to protein-coding genes because they are usually copied from a parent gene, either through unsuccessful duplication or by retrotransposition (whereby a gene is transcribed into RNA, which is then 'reverse-transcribed' back into DNA and inserted somewhere different in the genome). Because all this copying does not yield a normal, functioning protein, pseudogenes are usually identified by obvious 'disablements' in their sequence, such as frameshifts or premature stops. They have been of interest because they provide records of ancient molecules encoded by the genome.

Although pseudogenes have generally been considered as evolutionary 'dead-ends', one of the surprises of genome sequencing has been how abundant they are: tens of thousands of pseudogenes are found in mammalian genomes (roughly the same number as protein-coding genes in all mammals sequenced so far)[7]. In addition, a large proportion of these sequences seem to be under some form of purifying selection[8] — whereby natural selection eliminates deleterious mutations from the population — and genetic elements under selection have some use. Finally, several large-scale genomic studies probing non-gene parts of the genome for biochemical activity have found many pseudogenes being transcribed and regulatory factors binding upstream of them. One such investigation, the ENCODE pilot project[9], which looked at a representative 1% of the sequence of the human genome, found strong evidence for at least one-fifth of pseudogenes being actively transcribed.

These observations indicate that pseudogenes might not be purely dead relics of past genes but could be resurrected for new biochemical activities. Indeed, functioning pseudogenes have been reported previously. For instance, in snails, a pseudogene is involved in translational control of the gene that codes for nitric oxide synthase[10]. And transcripts of the mouse pseudogene *makorin1-p1* have been proposed to inhibit degradation of their parent gene's mRNA, effectively enhancing its expression[11], although this observation has been debated. Nevertheless, a clear mechanism for the functioning of pseudogenes has been lacking. The six studies — four in flies[1–4] and two in mice[5,6] — provide such a direct pathway,

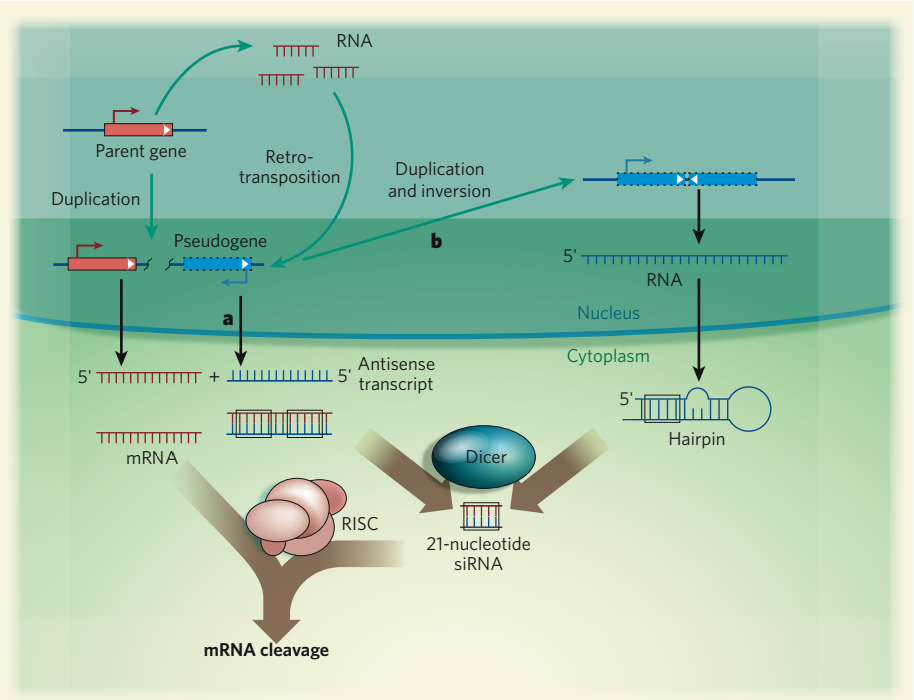


Figure 1 | Pseudogene-mediated production of endogenous small interfering RNAs (endo-siRNAs). Pseudogenes can arise through the copying of a parent gene (by duplication or by retrotransposition). **a**, An antisense transcript of the pseudogene and an mRNA transcript of its parent gene can then form a double-stranded RNA. **b**, Pseudogenic endo-siRNAs can also arise through copying of the parent gene as in **a** and then nearby duplication and inversion of this copy. The subsequent transcription of both copies results in a long RNA, which folds into a hairpin, as one half of it is complementary to its other half. In both **a** and **b**, the double-stranded RNA is cut by Dicer into 21-nucleotide endo-siRNAs, which are guided by the RISC complex to interact with, and degrade, the parent gene's remaining mRNA transcripts. The mRNA from genes is in red and that from pseudogenes is in blue. Green arrows indicate DNA rearrangements.

showing that pseudogene transcripts can act as natural siRNAs.

Broadly speaking, RNAi involves various types of small 'guide' RNA sequence regulating protein levels by targeting mRNA for degradation. Pseudogenic siRNAs provide two of the four categories posited by the six studies to organize the natural, or 'endo', siRNAs (Box 1).

Endo-siRNAs in the first category mediate transposon silencing, which is typically a feature of Piwi-interacting RNAs (piRNAs). The studies were therefore careful to distinguish between endo-siRNAs associated with transposons and piRNAs on the basis of size (21–22 nucleotides versus 24–30) and Argonaute effector-protein partner (Ago2 versus Piwi). The second category of endo-siRNAs arise from bidirectional transcription of partially overlapping loci on opposite DNA strands[1,12]. Studies in mice[5,6] identify a few examples of these, and around 1,000 have been reported in flies[1], with their target genes consisting mainly of those with nucleic-acid functions, such as nuclease activity and transcription-factor binding[12].

The third category of siRNAs, which have been identified only in mice[5,6], are products of the interaction between a spliced mRNA transcript from a protein-coding parent gene and an antisense transcript from its pseudogene, which can be located far away from its parent gene, on the same or a different chromosome (Fig. 1a). Endo-siRNAs of the fourth category are closely related to those in the third. They arise from hairpin-shaped sequences, which in mice[5,6] can come from inverted-repeat structures of pseudogenes (Fig. 1b). Here, the pseudogene also regulates its parent gene, but the double-stranded RNA precursor of the endo-siRNA comes from transcription of an inverted-repeat sequence, producing a hairpin. The reports show that mouse proteins affected by the third and fourth categories of endo-siRNAs are disproportionately involved in particular functions — such as regulating cytoskeletal dynamics — which indicates that their underlying pseudogene-mediated regulation has been explicitly selected for and is not simply caused by random pairing of transcribed genes and pseudogenes.

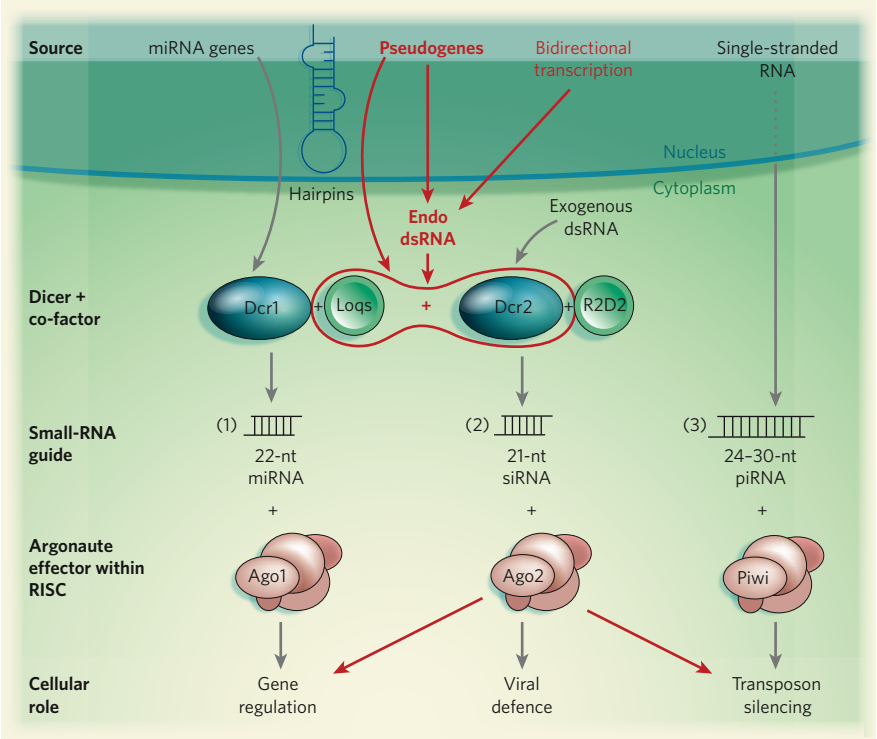
Hairpin precursors of endo-siRNAs have also been found in flies, but the evidence links them only weakly with inverted repeats of pseudogenes. Thus, most of the new data for pseudogenic siRNAs come from mouse rather than fly studies. One possible reason for this is that the mouse genome contains many more pseudogenes than the fly genome[13]. In fact, even compared with other metazoan organisms such as worms, flies are particularly poor in pseudogenes, possibly owing to pronounced genomic deletion processes known to occur in this organism[14].

The scarcity of pseudogenes in flies makes their detection particularly difficult. Nevertheless, there is suggestive evidence for fly pseudogenes functioning as endo-siRNAs. First, an appreciable number (~30) have an inverted-repeat structure, associated with the formation of hairpins. Second, many of the sequences obtained by ultra-high-throughput sequencing of small RNAs in the fly coincide with DNA regions containing pseudogenes. In particular, a small but significant number of the 'reads' found using the Solexa sequencing technology[1,4] can be intersected with some 70 pseudogenes, for an average of roughly 12 reads each. Finally, there is strong evidence that for several genes — particularly the β-esterase gene and its pseudogene — a duplicated pseudogene forms a functional complex with its parent gene, with regulatory consequences[15].

Of course, to demonstrate the activity of pseudogenes conclusively, further experiments are needed. Deleting a pseudogene and demonstrating an effect on its potentially regulated parent gene would be most definitive. Also of great value would be studying the expression patterns of a potential endo-siRNA-producing pseudogene and its regulated parent gene across various tissues — data which should be generated by the ENCODE and modENCODE projects.

In addition to connecting RNAi with pseudogenes, the new studies[1–6] also blur the distinctions between the three 'traditional' classes of small RNA — siRNAs, piRNAs and

Box 1 | Small but significant



There are three main classes of small RNA, which generally differ in biogenesis, sorting and function[19].

(1) MicroRNAs (miRNAs) mainly regulate genes involved in developmental processes. Specific miRNA genes encode mRNA-like primary transcripts that form hairpin structures, which are in turn excised by the enzyme Drosha (not shown) to form precursor miRNAs. In flies, further cleavage of these sequences by the Dicer enzyme Dcr1 and its specific co-factor Loqs yields mature miRNAs of ~22 nucleotides (nt). To carry out their function, miRNAs are incorporated into the RISC protein complex, which contains the effector protein Ago1, a member of the Argonaute family.

(2) Conventional small-interfering RNAs (siRNAs) of ~21 nucleotides are produced through cleavage of double-stranded RNA (dsRNA) — in flies, by the Dicer enzyme Dcr2 and its co-factor R2D2 (refs 17, 19). These small RNAs bind to the Argonaute-family effector Ago2 and function in defence against external nucleic acids, such as synthetic dsRNAs or intermediates of viral replication.

(3) Discrete genomic loci give rise to single-stranded RNA sequences (ssRNA), which are then processed to ~27 nt Piwi-interacting RNAs (piRNAs). piRNA biosynthesis remains somewhat ambiguous, but is known not to require Dicer. piRNAs bind to Piwi, another member of the Argonaute family that seems to be expressed only in germline cells. It is believed that these small RNAs function as master controllers of mobile genetic sequences called transposable elements[20].

In the figure, grey lines indicate known relationships, whereas red lines indicate new ones reported in the six papers[1–6]. Clearly, the boundaries between the three small-RNA classes have been somewhat blurred by these reports. For details of how endo-siRNAs arise from pseudogenes, see Figure 1. **R.S. & M.G.**

microRNAs (miRNAs) — which are distinct in their biogenesis and cellular roles (Box 1). The studies[1–6] find that endo-siRNAs regulate transposons as piRNAs do; that, like miRNAs, they can arise from hairpins; and that, in flies, their processing involves a similar co-factor to the processing of miRNAs (Box 1).

This blurring of boundaries among different types of small RNA, together with the newly established links between siRNAs and pseudogenes, has interesting evolutionary implications. In plants, inverted duplications containing a protein-coding gene have been proposed[16] as a mechanism to create new miRNAs. Thus, one can imagine a gene being copied (either by duplication or retrotranscription) and this copy then being duplicated (again) in inverted fashion. Given the ubiquitous nature of genomic transcription, the copy and its inverted duplicate could potentially be transcribed to a hairpin precursor of endo-siRNAs to regulate the parent gene.

As the function of the hairpin no longer has anything to do with encoding protein, its sequence, still under selection, can acquire frameshifts and stop codons, making it seem pseudogenic. One could even imagine its sequence drifting further and becoming gradually transformed into a miRNA gene, the sequence of which is much less similar to the gene encoding its target mRNA. So pseudogenes encoding endo-siRNAs might provide a crucial intermediate link to understanding the evolution of miRNA-mediated regulation[17]. Although speculative, the plausibility of this theory is bolstered by a recent survey[18] of the genomic context of more than 300 human miRNA loci, which identified two that lie within pseudogenes. ■

Rajkumar Sasidharan and Mark Gerstein are in the Departments of Molecular Biophysics and Biochemistry, and Computer Science, Yale University, New Haven, Connecticut 06520, USA.
e-mail: mark.gerstein@yale.edu

1. Czech, B. *et al. Nature* **453,** 798–802 (2008).
2. Ghildiyal, M. *et al. Science* **320,** 1077–1081 (2008).
3. Kawamura, Y. *et al. Nature* **453,** 793–797 (2008).
4. Okamura, K. *et al. Nature* **453,** 803–806 (2008).
5. Tam, O. H. *et al. Nature* **453,** 534–538 (2008).
6. Watanabe, T. *et al. Nature* **453,** 539–543 (2008).
7. Zhang, Z., Carriero, N. & Gerstein, M. *Trends Genet.* **20,** 62–67 (2004).
8. Zheng, D. *et al. Genome Res.* **17,** 839–851 (2007).
9. The ENCODE Project Consortium *Nature* **447,** 799–816 (2007).
10. Korneev, S. A., Park, J.-H. & O'Shea, M. *J. Neurosci.* **19,** 7711–7720 (1999).
11. Hirotsune, S. *et al. Nature* **423,** 91–96 (2003).
12. Okamura, K., Balla, S., Martin, R., Liu, N. & Lai, E. C. *Nature Struct. Mol. Biol.* doi:10.1038/nsmb.1438 (2008).
13. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. *Nucleic Acids Res.* **31,** 1033–1037 (2003).
14. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. *Nature* **384,** 346–349 (1996).
15. Balakirev, E. S., Anisimova, M. & Ayala, F. J. *J. Mol. Evol.* **62,** 496–510 (2006).
16. Allen, E. *et al. Nature Genet.* **36,** 1282–1290 (2004).
17. Chapman, E. J. & Carrington, J. C. *Nature Rev. Genet.* **8,** 884–896 (2007).
18. Devor, E. J. *J. Hered.* **97,** 186–190 (2006).
19. Matranga, C. & Zamore, P. D. *Curr. Biol.* **17,** R789–R793 (2007).
20. Brennecke, J. *et al. Cell* **128,** 1089–1103 (2007).

ATTOSECOND PHYSICS

An easier route to high harmony

Mark I. Stockman

The generation of ultrashort light pulses by atomic ionization and recombination doesn't come cheap. But by niftily exploiting the play of light on a nanostructured surface, it can be done on a table-top.

Extreme ultraviolet (EUV) radiation has great potential to be extreme not just in name, but in usefulness. It is the band of ultraviolet light with the shortest wavelength — around 5–50 nanometres, between 100 and 10 times shorter than that of visible light. In applications such as microscopy and lithography it can thus be used to probe and etch at tiny scales. What's more, this wavelength regime is that of many atomic resonances, making EUV light ideally suited for spectroscopic applications. On page 757 of this issue[1], Kim *et al.* detail a deft new way to produce EUV radiation — one that could be considerably more economical that previous approaches.

The way in which EUV radiation is currently generated is extremely fiddly. It starts with the amplification of light pulses from an oscillator, a source of laser light. These are used to drive the repeated ionization of noble-gas atoms. The electrons freed during this process are accelerated in the light field and, because the sign of the field reverses after half a cycle, re-collide with their parent atoms[2,3], releasing the electrons' surplus energy as light. The result is a sequence of ultrashort (attosecond) pulses that are themselves useful tools for high-time-resolution metrology[4,5]. More detailed consideration of the process reveals that the spectrum of these pulses consists of a comb of 'high harmonics' — spectral lines at wavelengths equal to the wavelength of the driving field divided by some integer. (Owing to symmetries of the particular situation, only light corresponding to odd-integer divisors is generated in this case.) The highest-harmonic (shortest-wavelength) component of this spectrum can be selected by filtering to produce a single attosecond pulse at an EUV wavelength.

Things would be much simpler if EUV radiation could be produced directly from an oscillator — an ultrashort pulsed laser of relatively low intensity — without the need for sophisticated, complex and expensive amplifiers to produce a high-intensity optical field. Kim *et al.*[1] provide a distinct glimmer of an indication that such an approach could be viable. They illuminate an intricate, nanoscale gold antenna structure with light from a standard titanium–sapphire laser, with a wavelength of 800 nm. The interaction of the light with the antennas produces high harmonics right up to the 17th harmonic — whose wavelength of 47 nm lies within the EUV range. The optical intensity required to generate this light is, at 10^{11} W cm^{-2}, about 100 times less than in the traditional approaches.

The secret of the authors' success is the nanoscale behaviour of 'quasiparticles' known as surface plasmons. These packets of optical energy represent rapid oscillations of electron density that spring up in the surface regions of metal nanoparticles when bathed in an incident light field. If this incident light is of the right frequency, the surface plasmons can enter resonance, greatly increasing the local field intensity over that of the excitation wave. This phenomenon has a central role in, for example, surface-enhanced Raman scattering[6], a spectroscopy and imaging technique that is sensitive enough to detect the presence of individual molecules adsorbed on a metal surface.

The extent of this field enhancement is determined by the nanoparticles' plasmonic resonance properties, which in turn depend mostly on the resistivity of the metal at the frequency of the optical light. Additional magnitude comes from geometric effects[7–9], both in narrow gaps between particles where there is a significant localization of optical energy, leading to the formation of 'gap plasmons', and also similarly around sharp tips, a phenomenon known as the lightning-rod effect. Nanoparticles have been specifically engineered in

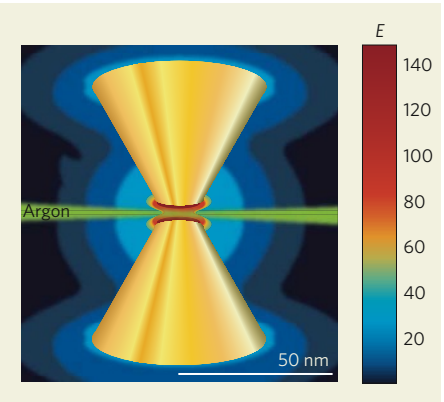


Figure 1 | Stripping on the table-top. The bow-tie-shaped gold nanoantennas used by Kim *et al.*[1] develop electric-field strengths in the gap through interactions with quasiparticles known as surface plasmons. (The field strength *E* is colour-coded; note that intensity is proportional to the square of the field strength.) When a beam of argon atoms (green) is directed towards the gap, the field strips them of electrons, which subsequently recombine — a process that results in the generation of high harmonics of the original light, including the sought-after extreme ultraviolet radiation. (Data and calculations courtesy of J. Aizpurua.)