Noncoding RNA transcription beyond annotated genes

Piero Carninci[1,2] and Yoshihide Hayashizaki[1,2]

Recent analyses based on high-throughput transcriptome data have revealed that the fraction of the genome that is transcribed largely exceeds the fraction encoding protein. Transcription of unconventional genes into noncoding RNAs is widespread and, in mammals, these RNAs comprise at least half the total number of RNAs transcribed by RNA polymerase II. Although the function of the majority of noncoding RNAs has yet to be discovered, many of them are transcribed from both strands of the genome, and evidence points towards a regulatory function for many noncoding RNAs in mammalian cells.

Addresses
[1] Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan
[2] Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan

Corresponding author: Carninci, Piero (rgscerg@gsc.riken.jp) and Hayashizaki, Yoshihide (rgscerg@gsc.riken.jp)

Current Opinion in Genetics & Development 2007, **17**:139–144

This review comes from a themed issue on
Chromosomes and expression mechanisms
Edited by Tom Misteli and Abby Dernburg

Available online 20th February 2007

0959-437X/$ – see front matter

DOI 10.1016/j.gde.2007.02.008

Introduction

In recent years, the science community has witnessed the sequencing of the human genome and an increasingly large number of other animal and plant genomes. Despite the apparent different organism complexity, vertebrate and invertebrate genomes have similar numbers of protein-coding genes. Mice and humans have as few as 20 000–25 000 protein-coding genes [1,2], slightly more than in the *Drosophila* (13 000) [3] and *C. elegans* (19 000) genomes [4] and, surprisingly, even less than in some plants, such as rice (more than 46 000) [5]. Yet mammals have evolved very complex structures. For instance, in the human brain, more than 100 billion neurons need instruction to properly connect during development. The low number of protein-coding genes has been a surprise: before the completion of these genome sequencing projects, scientists assumed that a large number of 'genes' would be required to control many complex biological phenomena. Although part of this task can be performed by alternative splicing, novel datasets are pointing at new directions.

In this review, we discuss recent works that suggest that the transcriptional output of the genome is much more complex than estimates based on the number of protein-coding genes; these studies point at non-coding RNA as a source of regulatory element.

There are more transcripts than 'genes'

In the past couple of years, it has become evident that the genome produces many more RNAs than there are protein-coding genes and genes encoding structural RNAs (e.g. rRNAs and tRNAs), and that most of these RNAs have still unknown functions. Novel non-protein-coding RNAs (from here on referred to as noncoding RNAs) have been identified in plants (e.g. sense–antisense RNAs in *Arabidopsis* [6] and rice [7]), invertebrates (e.g. *Drosophila* [8] and honey bee [9]) and vertebrates, as described below. Novel datasets consisting of whole-genome tiling arrays and sequences tags have demonstrated that, in mammals, the number of the noncoding RNAs is very large, in agreement with theories suggesting that the increase in complexity of body plans, and its regulation, is associated with the expansion in the number of noncoding RNAs as controlling elements [10,11]. Indeed, the whole-genome tiling arrays [12,13••] demonstrate that the human genome is widely transcribed, extending beyond the conventionally annotated protein-coding genes. In particular, Cheng and colleagues [13••] separated the cytosolic and nuclear cellular fractions, and extracted from each of them polyA$^+$ and polyA$^-$ RNAs, to hybridize whole-genome tiling arrays. They identified transcripts that, to date, have been hidden from other experimental approaches for gene discovery. In fact, 51.3% of the species of RNAs are detected exclusively in the nucleus, and 43% of all the RNAs do not have a polyA tail; these fractions are not targeted during cDNA cloning for expressed sequence tag (EST) sequencing determination, the strategy of which is based on oligo-dT priming. Indeed, in the mouse, the sequences of large noncoding RNAs, which probably have no 3′ polyA tail, were reconstructed from the fragments of truncated cDNAs. These cDNAs were generated from oligo-dT molecules that were non-specifically hybridized onto A-rich stretches of RNAs, leading to the production of internally primed cDNAs with incomplete 3′ ends. In fact, these fragments are difficult to clone as full-length cDNAs, because the original RNAs are too long and lack a 3′ polyA tail [14]. These 66 RNAs, identified accidentally, represent the tip of the iceberg of a large class of RNAs, members of which are difficult to

clone and which are preferentially localized in the nucleus and devoid of polyA tails. A large fraction of these transcripts are encoded from genomic regions that do not contain known protein-coding genes, suggesting that many of them are noncoding RNAs.

In addition to humans, the mouse transcriptome has been deeply analyzed. However, for the mouse, the transcriptome was studied with an extensive collection of full-length cDNAs and other datasets called 5′–3′ 'ditags', which identify the start and termination sites of RNAs. [15••,16,17]. The RIKEN mouse full-length cDNA collection differs from the human one. Besides tissue sampling, one of the large human cDNA contributors, the Mammalian Gene Collection project [18], uses for full-insert sequencing only those cDNAs that have coding potential. However, the RIKEN mouse cDNAs have been randomly sequenced and therefore the noncoding RNA content is different. By analyzing all of the available mouse cDNA sequences, it was found that the mouse genome encodes at least 44 000 distinct transcriptional units; a transcriptional unit comprises all the RNAs that share a common exonic sequence transcribed from the same genomic strand. Analysis of transcriptional units shows that more than half (~23 000) do not have protein-coding potential in any of the identified variants [15••]. Given that the technique used to isolate relatively large full-length cDNA makes use of cap selection [19], most of these noncoding RNAs are likely to be transcribed by RNA pol II. To ascertain that the novel RNAs identified across high-throughput platforms are not artefacts [17], experimental validation confirms that the majority of noncoding RNAs are genuine transcripts. Tiling arrays and tags partially validate each other when applied on the same sample. Analysis of a human liver cell line (Hep-G2) with both CAGE (cap-analysis gene expression) and whole-genome tiling arrays shows that CAGE tags are generally validated by tiling array, and generally extend their edges, but a large number of tiling array transfrags (transcribed fragments) cannot be validated by tags [17]. Besides difficulties in mapping 5′ tags to internal exons, it is also possible that a large number of tiling arrays identify non-capped RNAs. Rapid amplification of cDNA ends (RACE) [20] largely validates the existence and structure of the unknown noncoding RNA transcripts, whereas microarrays and reverse transcriptase PCR (RT-PCR) show that noncoding RNAs are dynamically expressed in different tissues and upon induction [21]. Additionally, we have recently identified novel dimensions of complexity: there are noncoding RNA variants of protein-coding genes, including transcripts that originate from the exonic sequence of mRNAs transcribed from promoters enriched in TATA-boxes, and transcripts originating from a novel type of promoter in the 3′ untranslated regions (UTRs) [16]. Interpretations that the former could be associated with chromatin remodeling and transcriptional control, and that the latter might be involved in regulating sense-antisense transcription and other information that is encoded in the 3′ UTRs are awaiting functional validation.

Lack of conservation: why we are not monkeys?

Conservation has been assumed to be synonymous to functionality. This is intuitively true for protein domains that have to perform under structural constrains, such as when forming a membrane channel, constructing a catalytic site of an enzyme or, in the case of transcription factors, binding to double-stranded DNA. However, if noncoding RNAs have regulatory functions, some of their parts can act by pairing to other nucleic acids. Alternatively, other noncoding RNAs can work *in cis*, by altering the chromatin and other binding proteins through the act of transcription itself. Clearly, these types of interaction do seem to be dramatically constrained against base substitutions. Indeed, a very large fraction of noncoding RNAs are poorly conserved, which has led to the speculation that they are not functional [22]. However, this observation is not necessarily associated to lack of functionality, because known functional noncoding RNAs, such as *Xist* (responsible for X chromosome inactivation) and *Air* (involved in imprinted gene-silencing at the *IGFR2* locus) are poorly conserved [23]. Indeed, putative promoter regions located upstream of a validated set of noncoding RNAs show clear patterns of conservation that extend broadly upstream to the starting site but which are reduced downstream of the transcription starting site [15••]. Intriguingly, an independent study on expression of intergenic sequences in human and chimpanzee shows conservation of expression of noncoding RNAs in equivalent genomic positions, but not conservation of the RNA sequences, demonstrating that that noncoding RNAs are under expression constraints, particularly in brain [24]. Instead of looking for similarity of expression, Pollard and colleagues [25] looked for differences between human and chimpanzee. They analyzed genomic regions that evolve faster than neutral regions (evidence of positive selection) and identified about 200 of these regions. Among them, one expresses an RNA during the human cortical development in neocortex from 7 to 19 gestational weeks, a period crucial for cortical neuron specification and migration [26•]. This noncoding RNA, called *HAR1F*, is co-localized with Reelin, a protein with a fundamental role in specifying the six-layer structure of the human cortex. Although we await the final proof of functionality, such as interaction of Reelin and the *HAR1F* RNA, this work establishes the importance of looking for lack of conservation to explain evolutionary differences.

Sense–antisense transcription

Validation of tiling arrays transfrags (which roughly correlate with exonic regions [17]) has identified overlapping transcripts for about half of the expressed genomic

regions [20]. A consistent fraction of the antisense RNAs validated from transfrags includes transcripts whose splicing sites seem to be the mirror image of the normally spliced RNAs [13••], leading the authors to hypothesize the existence of a putative RNA-dependent RNA polymerase that would produce such antisense RNAs in mammals. This hypothesis awaits further confirmation, because such transcripts were not identified from other cDNA collections, and only conventionally spliced antisense transcripts have been widely identified.

The 5′ end CAGE tags have revealed antisense transcription in 72% of the transcriptional unit, when considering transcripts antisense to introns [27]. Other technologies have also revealed a large extent of transcription overlap: large-scale serial analysis gene expression (SAGE; 8.5 millions of tags in the mouse genome) [28] suggests the existence of a large number of unannotated transcripts. Besides the prediction of 23 000 novel transcriptional units, they identified a large extent of sense–antisense transcripts (36%), which confirms the majority of RIKEN sense–antisense transcripts analyzed RNAs [29].

Sense–antisense relationships can extend for more than a pair of transcriptional units, by involving as much as 11 transcriptional units, which overlap as sense–antisense or share a bidirectional promoter. The transcriptional units in such relationships are said to form 'chains' [30]. Because the expression of transcriptional units within chains tends to be positively correlated, they are likely to be located in genomic regions that are under similar transcriptional control, and indeed they tend to be co-expressed. About 1000 of such chain pairs are conserved between mouse and human, despite the generally poor sequence conservation of the RNAs involved in chains [30]. The failure to identify positional conservation for other chains might be due to lack of appropriate cDNA coverage resulting from different sampling and gene discovery strategies (P Carninci, unpublished).

It is widely accepted that mammalian genomes frequently transcribe overlapping RNAs, and that sense and antisense RNAs influence each other's expression levels [27]. Furthermore our knowledge of the molecular mechanisms of *in trans* sense–antisense relationships (such as miRNAs [31]) is growing. Despite all this, there is not yet consensus on how sense–antisense transcription exerts its function in mammals. Extrapolating from the lessons of the siRNA and miRNA machinery, it is tempting to hypothesize that sense–antisense pairs produce short siRNAs (silencing RNAs). These siRNAs would then be used by the RISC complex to cleave larger RNAs, thereby promoting their degradation. However, such short RNAs could not be identified [32] for two gene pairs mapping on the genome in a tail to tail fashion, leading to the interpretation that the action of this type of sense–antisense transcript does not depend upon Dicer and does not involve formation of short siRNA, at least in the cytosolic fraction. In the cytoplasm, large sense–antisense transcripts do not seem to interact [32], but the interaction might be nuclear. Accordingly, noncoding RNAs that are involved in sense–antisense relationships tend to be poorly polyadenylated and enriched in the nuclear fraction [27,29]. Mechanisms can differ for the large fraction of bidirectionally transcribed regions spanning promoter regions [27]. Indeed, Dicer acts in the nucleus of mammalian cells, where it controls transcription at the β-globin locus, from which bidirectional transcription has been detected. Knocking down Dicer expression in mammalian cells promotes intergenic transcription at the β-globin gene cluster and has general effects on the whole locus [33]. Given that this is also associated with histone modifications that facilitate transcription, these experiments clearly point at the nucleus as the compartment where bidirectional transcription has an important role in mammals, and where transcription is required for subsequent silencing of specific chromosomal regions, although the mechanisms are much less clear than in plants [34].

Not every noncoding RNA needs to be diced

It would be an oversimplification to assume a single mode of action for noncoding RNAs, and various mechanisms have been extensively reviewed elsewhere [10,11,17,35]. Apart from their potential involvement in *cis* and *trans* sense–antisense transcripts, the action of noncoding RNA transcription involves other mechanisms, including the formation of complexes with proteins, such as NRON (noncoding repressor of NFAT), which modifies nuclear transportation [36], coating of chromosomal regions (e.g. X inactivation), and transcriptional interference [17]. A model for transcriptional interference was proposed for *Air*, a large (108 kb) pol II-transcribed noncoding RNA that is mainly retained in the nucleus. Its instability and the fact that it isn't exported to the cytoplasm lead to the hypothesis that its actions are based on transcriptional interference: transcription of *Air* would proceed through hypothetical domain regulatory element(s), displacing the factor bound there, which would otherwise activate transcription of *Igf2r*, *Slc22a2* and *Slc22a3* — the latter two do not show overlapping expression, and yet are controlled by *Air* [37]. This mechanism is opposite to transcription of HOX antisense noncoding RNAs, the expression of which seems to prevent silencing [30].

Although *Air* was considered to be an unusual transcript, analysis of tiling arrays [13••] and cDNA [14] suggests that such nuclear RNAs are very abundant, and transcription through controlled regions might represent a common mechanism of action. For most of the noncoding large RNAs, however, novel high-throughput screening methods will probably reveal more mechanisms of action.

But more diversity among short RNAs

Discovery of novel short RNAs has been recently reported, increasing the dimensions of the transcriptome complexity. In fact, a series of almost simultaneous papers has recently described a new large family of short RNAs (25–31 nucleotides), members of which are slightly larger than the miRNAs and siRNAs, and which bind to Piwi and its orthologous proteins in mouse (Miwi and Mili) and rat (Riwi) [38•–42•]. These short RNAs are produced by relatively few regions in each respective genome and are essential for the maturation of the sperms [42•]. They are produced from less than 100 genomic regions, which map in orthologous regions of the mouse, rat and human genomes. These RNAs, renamed piRNAs (Piwi-associated RNAs) are hypothesized to work at the nuclear level, by associating with DNA, histone or RNAs [43], and form a complex with RecQ that has helicase activity. A remarkable feature of these short transcripts is that, unlike other short RNAs, they do not seem to originate from the cleavage of double-stranded RNAs but from direct processing of single-stranded RNAs.

There are still additional novel mechanisms of action of short RNAs: by analyzing the ovary, Watanabe *et al.* [42•] identified yet another class of short RNA. Members of this group are 20–24 nucleotides long and derive from retro-elements (e.g. SINEs [short interspersed elements], LINEs [long interspersed elements] and LTRs [long terminal repeats]), and they were shown to destabilize transcripts containing such repeat elements, suggesting that they work through the RNAi pathway. It is uncertain, however, where these short repeat-associated RNAi molecules exercise their action. Besides working like the repeat-associated RNAs (rasiRNAs) in invertebrates [44], it is intriguing that 17% of mRNA isoforms encoding proteins that are specifically expressed in mouse early embryos make use of LTR elements as promoter elements in a short window of the early development in mice [45•]. The short repeat-associated RNAs might provide a link to regulate those transcripts specifically promoted by repeat elements. Intriguingly, repeat elements are also present in *Drosophila* ovary, where such repeats are associated with Piwi but have a larger size (25–29 nucleotides) [46]. It is very likely that novel dimensions of complexity will be added by analyzing other tissues and biological phenomena: for instance, noncoding repetitive RNAs are produced from cells under stress in nuclear stress bodies [47]; and, in plants, high-throughput methods have identified 30 000 different short transcripts [48], promising that more short RNAs will soon be found.

Conclusions

Noncoding RNAs constitute a very heterogeneous group of RNAs, for which we are progressively deciphering the basic mechanisms of action. High-throughput screening methods might reveal their functions and help to admit them into the category of non-protein coding RNA 'genes'. However, owing to the difficulties in measuring phenotypes, in particular those having mild effects, we might not be able to assess the function of many of them, nor of a fraction of protein-coding genes. If the non-genic RNAs have regulatory functions, regulatory network redundancy might further complicate the functional phenotype assessment in laboratory animals. Until that time, we scientists are facing a dilemma: should we start considering (and classifying) all of these RNAs as non-protein coding genes or keep them to one side until their function will be determined? This is causing a growing dichotomy in genomics and biology. While gene annotation and nomenclature is very conservative and tends to be limited to protein-coding mRNA with solid evidence of function, a parallel world of noncoding transcripts clearly exists, which it is not properly represented in the maps.

Acknowledgements

We thank the members of the GERG and GSL for support. In particular, Hanna Daub for editing the manuscript, and Miki Nishikawa and Harumi Uruma for handling the manuscript and secretarial assistance. This work was supported by a research grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to YH, a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and a grant from the RIKEN Frontier Research System.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al.*: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.

3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.

4. consortium Ces: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science* 1998, **282**:2012-2018.

5. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: **A draft sequence of the rice genome (Oryza sativa L. ssp. indica)**. *Science* 2002, **296**:79-92.

6. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M *et al.*: **Empirical analysis of transcriptional activity in the *Arabidopsis* genome**. *Science* 2003, **302**:842-846.

7. Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K *et al.*: **Antisense transcripts with rice full-length cDNAs**. *Genome Biol* 2003, **5**:R5.

8. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, Celniker SE, Rubin GM: **Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster***. *Proc Natl Acad Sci USA* 2005, **102**:5495-5500.

9. Honeybee Genome Sequencing Consortium: **Insights into social insects from the genome of the honeybee *Apis mellifera***. *Nature* 2006, **443**:931-949.

10. Mattick JS: **Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms**. *Bioessays* 2003, **25**:930-939.

11. Mattick JS: **RNA regulation: a new genetics?** *Nat Rev Genet* 2004, **5**:316-323.

12. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S *et al.*: **Global identification of human transcribed sequences with genome tiling arrays**. *Science* 2004, **306**:2242-2246.

13. •• Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G *et al.*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution**. *Science* 2005, **308**:1149-1154.

This paper describes the whole-genome tiling arrays for 10 chromosomes at 5-bp resolution — one oligonucleotide for every 5 base pairs of non-repeated genome. Remarkably, the authors have used for the first time various cellular fractions (e.g. nuclear and cytoplasmic RNAs) and various RNA fractions (e.g. RNA with and without polyA tails, from both compartments). This has proven that the unknown part of the transcriptome is several-fold larger than expected.

14. Furuno M, Pang KC, Ninomiya N, Fukuda S, Frith MC, Bult C, Kai C, Kawai J, Carninci P, Hayashizaki Y *et al.*: **Clusters of internally primed transcripts reveal novel long noncoding RNAs**. *PLoS Genet* 2006, **2**:e37.

15. •• Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C *et al.*: **The transcriptional landscape of the mammalian genome**. *Science* 2005, **309**:1559-1563.

We have released a very large collection of full-length mouse cDNAs and tags that identify the starting and termination sites, demonstrating that there are at least 181 000 different transcripts in a mammalian genome.

16. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution**. *Nat Genet* 2006, **38**:626-635.

17. Carninci P: **Tagging mammalian transcription complexity**. *Trends Genet* 2006, **22**:501-510.

18. Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF *et al.*: **Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences**. *Proc Natl Acad Sci USA* 2002, **99**:16899-16903.

19. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D *et al.*: **Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia**. *Genome Res* 2003, **13**:1273-1289.

20. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays**. *Genome Res* 2005, **15**:987-997.

21. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM *et al.*: **Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome**. *Genome Res* 2006, **16**:11-19.

22. Babak T, Blencowe BJ, Hughes TR: **A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription**. *BMC Genomics* 2005, **6**:104.

23. Pang KC, Frith MC, Mattick JS: **Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function**. *Trends Genet* 2006, **22**:1-5.

24. Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, Paabo S: **Functionality of intergenic transcription: an evolutionary comparison**. *PLoS Genet* 2006, **2**:e171.

25. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R *et al.*: **Forces shaping the fastest evolving regions in the human genome**. *PLoS Genet* 2006, **2**:e168.

26. • Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A *et al.*: **An RNA gene expressed during cortical development evolved rapidly in humans**. *Nature* 2006, **443**:167-172.

This study establishes two principles: that looking for differences is the key to understanding recent human evolution; and that expressed, diverging noncoding RNAs are the other key to understanding differences. This goes against the common sense to look for similarities in search of function.

27. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al.*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**:1564-1566.

28. Siddiqui AS, Khattra J, Delaney AD, Zhao Y, Astell C, Asano J, Babakaiff R, Barber S, Beland J, Bohacec S *et al.*: **A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells**. *Proc Natl Acad Sci USA* 2005, **102**:18485-18490.

29. Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K: **Disclosing hidden transcripts: mouse natural sense–antisense transcripts tend to be poly(A) negative and nuclear localized**. *Genome Res* 2005, **15**:463-474.

30. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, Lavorgna G, Brozzi A, Luzi L, Tan SL, Yang L *et al.*: **Complex loci in human and mouse genomes**. *PLoS Genet* 2006, **2**:e47.

31. Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**:350-355.

32. Faghihi MA, Wahlestedt C: **RNA interference is not involved in natural antisense mediated regulation of gene expression in mammals**. *Genome Biol* 2006, **7**:R38.

33. Haussecker D, Proudfoot NJ: **Dicer-dependent turnover of intergenic transcripts from the human β-globin gene cluster**. *Mol Cell Biol* 2005, **25**:9724-9733.

34. Henderson IR, Zhang X, Lu C, Johnson L, Meyers BC, Green PJ, Jacobsen SE: **Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning**. *Nat Genet* 2006, **38**:721-725.

35. Mattick JS: **The functional genomics of noncoding RNA**. *Science* 2005, **309**:1527-1528.

36. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG: **A strategy for probing the function of noncoding RNAs finds a repressor of NFAT**. *Science* 2005, **309**:1570-1573.

37. Seidl CI, Stricker SH, Barlow DP: **The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export**. *EMBO J* 2006, **25**:3565-3575.

38. • Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE: **Characterization of the piRNA complex from rat testes**. *Science* 2006, **313**:363-367 See annotation [42•]..

39. • Grivna ST, Beyret E, Wang Z, Lin H: **A novel class of small RNAs in mouse spermatogenic cells**. *Genes Dev* 2006, **20**:1709-1714.
See annotation [42•].

40. • Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T *et al.*: **A novel class of small RNAs bind to MILI protein in mouse testes**. *Nature* 2006, **442**:203-207.
See annotation [42•].

41. • Girard A, Sachidanandam R, Hannon GJ, Carmell MA: **A germline-specific class of small RNAs binds mammalian Piwi proteins**. *Nature* 2006, **442**:199-202.
See annotation [42•].

42. • Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H: **Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes**. *Genes Dev* 2006, **20**:1732-1743.

This series of studies [38•–42•] identifies a novel class of short RNAs that bind to Piwi and related mammalian proteins (Riwi, Miwi and Mili) essential for spermatogenesis.

43. Carthew RW: **Molecular biology. A new RNA dimension to genome control**. *Science* 2006, **313**:305-306.

44. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD: **A distinct small RNA pathway silences selfish genetic elements in the germline**. *Science* 2006, **313**:320-324.

45. • Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB: **Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos**. *Dev Cell* 2004, **7**:597-606.
This paper reports that 17% of transcripts in full-grown oocytes are driven by specific promoters composed specifically of LTR retrotransposons. These LTR sequences are used as alternative promoters of otherwise normal protein-coding genes. This is the only known example in which mammalian transcription is widely driven by retro-elements.

46. Saito K, Nishida KM, Mori T, Kawamura Y, Miyoshi K, Nagami T, Siomi H, Siomi MC: **Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome**. *Genes Dev* 2006, **20**:2214-2222.

47. Valgardsdottir R, Chiodi I, Giordano M, Cobianchi F, Riva S, Biamonti G: **Structural and functional characterization of noncoding repetitive RNAs transcribed in stressed human cells**. *Mol Biol Cell* 2005, **16**:2597-2604.

48. Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ: **Elucidation of the small RNA component of the transcriptome**. *Science* 2005, **309**:1567-1569.