EMBO reports

concept

Non-coding RNAs: the architects of eukaryotic complexity

John S. Mattick+

ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

Received July 31, 2001; revised September 10, 2001; accepted September 11, 2001

Around 98% of all transcriptional output in humans is non-coding RNA. RNA-mediated gene regulation is widespread in higher eukaryotes and complex genetic phenomena like RNA interference, co-suppression, transgene silencing, imprinting, methylation, and possibly position-effect variegation and transvection, all involve intersecting pathways based on or connected to RNA signaling. I suggest that the central dogma is incomplete, and that intronic and other non-coding RNAs have evolved to comprise a second tier of gene expression in eukaryotes, which enables the integration and networking of complex suites of gene activity. Although proteins are the fundamental effectors of cellular function, the basis of eukaryotic complexity and phenotypic variation may lie primarily in a control architecture composed of a highly parallel system of *trans*-acting RNAs that relay state information required for the coordination and modulation of gene expression, via chromatin remodeling, RNA–DNA, RNA–RNA and RNA–protein interactions. This system has interesting and perhaps informative analogies with small world networks and dataflow computing.

The genome sequencing projects have revealed an unexpected problem in our understanding of the molecular basis of developmental complexity in the higher organisms: complex organisms have lower numbers of protein coding genes than anticipated. The fruitfly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans* appear to have only about twice as many protein coding genes (~12–14 000) as microorganisms such as *Saccharomyces cerevisiae* (~6200) and *Pseudomonas aeruginosa* (~5500) (Rubin *et al.*, 2000; Stover *et al.*, 2000). Humans appear to have only twice as many again (~30 000) (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), although there is some debate about this (Wright *et al.*, 2001; see also below). While the repertoire of protein isoforms expressed in the higher organisms is greatly increased by alternative splicing (Graveley, 2001), the other striking feature of the evolution of the higher organisms, which has been largely overlooked to date, is the huge increase in the amount of non-protein-coding RNA, which in humans accounts for ~98% of all genomic output (see below).

Have we missed something fundamental? Are these RNAs functional, and if so might they represent an important development in the genetic operating system of the higher organisms, as opposed to the mainly protein-based systems of microbes?

Phenotypic diversity in eukaryotes

The proteomes of the higher organisms are relatively stable. Humans and mice share 99% of their protein coding genes in common (J.C. Venter, personal communication), and differentiation in these and other complex eukaryotes appears to be achieved primarily by modular re-use and multitasking of different subsets of the proteome (Pawson, 1995; Duboule and Wilkins, 1998). Moreover, of the ~3 000 000 sequence differences per haploid genome between individual humans, only ~10 000 (0.3%) occur in protein-coding sequences, mostly as silent (third base) changes (Venter *et al.*, 2001).

Thus, phenotypic variation between both individuals and species may be based largely on differences in non-protein-coding sequences and be mainly a matter of variation in gene expression, i.e. due to the control architecture of the system. This further implies that, although protein variation will also contribute, the primary source of complex traits and of quantitative trait variation is embedded in this control architecture. If so, this has significant implications for understanding the basis of differentiation and development and the regulatory networks that underlie neural function, disease susceptibility and cancer.

While the control architecture is assumed to be primarily located in *cis*-acting gene promoters and enhancers, which are subject to combinatorial inputs from transcription factors modulated by signaling pathways, this may be only part of the answer. This view ignores the possible role(s) of non-coding RNAs,

+Corresponding author. Tel: +61 7 3365 4446; Fax: +61 7 3365 8813; E-mail: j.mattick@imb.uq.edu.au

which represent the vast majority of genomic output in higher organisms. The failure to recognize the possible significance of these RNAs is based on the central dogma, as determined from bacterial molecular genetics, that genes are synonymous with proteins, and that RNAs are just temporary reflections of this information. This view is reinforced by the prevailing biochemical perspective that proteins comprise the primary regulators of cell and organismal biology, which is essentially the case in prokaryotes (although non-coding RNAs are occasionally used), but may not be true for higher eukaryotes.

Genomic output in the higher eukaryotes

Non-protein-coding RNA transcription in the eukaryotes falls into two classes: introns and other non-protein-coding RNAs. In humans, introns account for ~95% of the pre-mRNA transcripts of protein coding genes, and are generally of high sequence complexity. As far as can be judged from *ad hoc* reports and from hybridization kinetic analysis of the relative complexity of heterogeneous nuclear (hn) RNA versus mRNA, other non-coding RNAs represent half to three quarters of all transcription from the genomes of the higher organisms (Davidson *et al.*, 1977; Mattick and Gagen, 2001; Shabalina *et al.*, 2001). These RNAs include a plethora of antisense transcripts and 'intergenic' transcripts (Ashe *et al.*, 1997; Askew and Xu, 1999; Eddy, 1999; Erdmann *et al.*, 2001; Mattick and Gagen, 2001) and may include many of the estimated 65 000–75 000 transcriptional units in the human genome (Wright *et al.*, 2001). If we assume that approximately two thirds of all transcripts do not contain protein-coding sequences, the real number of 'genes' (defined as those that produce separate primary transcripts and are separately regulated) in mammalian genomes may in fact be of the order of 100 000. Where these non-coding RNAs have been examined, they are developmentally regulated and have genetic effects. A good example is the bithorax-abdominal A/B complex of *Drosophila* which spans ~200 kb and expresses seven major transcripts that cover almost the entire region. Only three of these contain protein-coding sequences, but all are spatially and temporally regulated and the interruption or deletion of the DNA that encodes them has known phenotypic consequences (Lipshitz *et al.*, 1987; Sanchez-Herrero and Akam, 1989).

Potential *trans*-acting mediators of cellular networking and regulation

If these non-protein-coding RNAs are functional, their most obvious role would be in networking, i.e. the production of parallel *trans*-acting signals that allow activity at one locus to be connected with others in real time. This further implies that suites of gene activity and other levels of systems control may be directly coordinated and integrated in a programmed manner via efference RNA signals (eRNAs) (Figure 1) and that this may be fundamental to the operation of the system. These eRNAs could act as a cellular memory of recent transcription events (Mattick, 1994), as a kind of soft wiring (Herbert and Rich, 1999a). At face value this would represent an enormous increase in network connectivity and functionality over the situation where system activity is solely regulated through protein-based feedback loops that relay metabolic and environmental state information (Mattick, 1994; Mattick and Gagen, 2001). Moreover, if a system utilizing an RNA communication network has evolved, it would not be surprising if many loci had evolved solely to express RNA.

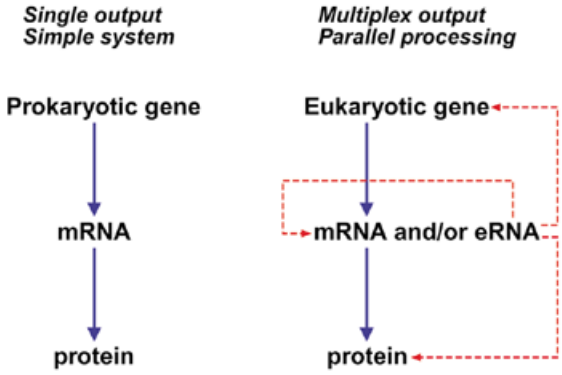

Fig. 1. Comparison of the prokaryotic and proposed eukaryotic genetic operating systems. The left panel shows the central dogma in which genes code, via mRNA, for proteins, which carry out the catalytic, structural, signal transduction and regulatory functions of the cell. The right panel shows the proposed operating system in eukaryotes wherein genes may express two levels of information: mRNA for proteins, and eRNAs that carry out concomitant networking and other functions within the organism. Thus there are three types of genes in eukaryotes: those that encode only protein (which are rare), those that encode only eRNA, and those that encode both.

The origin and evolution of eukaryotic nuclear introns

When nuclear introns were first discovered they were assumed to be non-functional and were postulated to be remnants of the prebiotic assembly of genes from exonic cassettes of protein-coding information (Gilbert *et al.*, 1986). However, it is now clear that modern nuclear introns invaded eukaryotic genes late in evolution, after the separation of transcription and translation (Mattick, 1994; Cho and Doolittle, 1997; Logsdon, 1998; Wolf *et al.*, 2000). The fragmentation of protein-coding genes by introns may have conferred an advantage by facilitating the modular shuffling of eukaryotic protein domains in evolutionary time and in real time via alternative splicing, but this is not necessarily the prime reason for their dominance. Alternative splicing signals are usually short and located near intron–exon boundaries (Lopez, 1998), and cannot account for the vast tracts of intronic sequences that populate most protein-coding genes in the higher organisms.

Nuclear introns are clearly derived from self-splicing group II introns of prokaryotes, which have the same splicing mechanism and which have expanded in eukaryotes by retrotransposition and other mutational, recombinational and insertional processes (Lambowitz and Belfort, 1993; Jacquier, 1996; Tarrio *et al.*, 1998; Cousineau *et al.*, 2000; Eickbush, 2000). The evolution of the spliceosome by the devolution of *cis*-acting catalytic RNAs into *trans*-acting general factors (spliceosomal RNAs) and the recruitment of accessory proteins would have reduced the internal sequence constraints on these introns, and allowed them considerable freedom to drift, expand and evolve. Any sequences that acquired a useful function, for example as *trans*-acting signals capable of transmitting other

information in parallel with their associated protein coding sequences, would have had a certain selective value and formed the genesis of a networking system in eukaryotic cells (Mattick, 1994). This does not imply that all introns will have evolved function, as each will be evolving largely independently, but rather that an increasing number may well have done so. In the pufferfish *Fugu rupripes* for example, which has a highly compact genome (Elgar, 1996), about three-quarters of the introns are very small, and probably represent vestigial remnants of past insertions, whereas the remainder are considerably larger and probably contain functional information.

Intronic RNA and other non-protein-coding RNAs now constitute the majority of genomic output in complex eukaryotes. Moreover, after accounting for variable amounts of repetitive DNA, there is a good correlation between intron density and developmental complexity (Mattick and Gagen, 2001). Introns and other noncoding RNAs have high sequence complexity and, in some cases, show interesting patterns of conservation across large evolutionary distances. Conservation is often found in large blocks that are indicative of selective constraints (Jareborg *et al.*, 1999; Mattick and Gagen, 2001; Shabalina *et al.*, 2001). The fact that most introns are less conserved than their associated protein-coding exons does not mean that they lack function, but rather that they are subject to less severe constraints.

Evidence that introns and other non-coding RNAs have function

Examples of intronic and other non-protein-coding RNAs that contain functional information are increasingly coming to light (Askew and Xu, 1999; Eddy, 1999) (see also below). One interesting subclass of these are small nucleolar RNAs (snoRNAs), which are produced from intronic RNAs derived from genes encoding ribosomal proteins and nucleolar proteins, as well as from other genes whose exons no longer have any protein coding capacity (Maxwell and Fournier, 1995; Tycowski *et al.*, 1996; Filipowicz, 2000). These introns are processed through pathways involving endonucleolytic cleavage by double-stranded RNase III-related enzymes, exonucleolytic trimming and possibly RNA-mediated cleavage, which occur in large complexes called exosomes (Allmang *et al.*, 1999; van Hoof and Parker, 1999).

Other interesting examples of non-coding RNAs with functional activity are the small temporal RNAs *lin-4* and *let-7*, which control developmental timing in *C. elegans* via RNA–RNA interactions that affect the translation and stability of other transcripts (Moss, 2000). *let-7* is conserved among vertebrates and invertebrates (Pasquinelli *et al.*, 2000). These small RNAs are derived from larger precursors and are around 22 nucleotides in length, similar to the size of RNAs produced by RNA interference (RNAi)-mediated RNA processing (see below). Indeed, it has been shown recently that the production of these RNAs is dependent on homologs of the Dicer and RDE-1 families of proteins that are also involved in RNAi (Grishok *et al.*, 2001). It is quite conceivable that such pathways are involved in the downstream processing of a wide range of intronic and other non-coding RNAs, whose products may number in the tens or hundreds of thousands and which may act as guide RNAs to regulate many different processes. There are many other examples of non-coding RNAs that have a role during development in both animals and plants, including Xist and roX1/roX2 which are involved in dosage compensation, as well as H19, Pgc, NTT, bic, BORG, BC200, his-1, Bsr, hsr-omega, ENOD40, CR20, among many others (Nakamura *et al.*, 1996; Teramoto *et al.*, 1996; Liu *et al.*, 1997; Tam *et al.*, 1997; Takeda *et al.*, 1998; Eddy, 1999; Komine *et al.*, 1999; Erdmann *et al.*, 2001). Some of these RNAs are alternatively spliced or have alternative polyadenylation sites and are probably derived from genes that have lost their protein coding capacity.

It seems safe to predict that the vast majority of non-coding RNAs have not yet been catalogued (see e.g. Ashe *et al.*, 1997), as most genomic screens have been intrinsically biased against their discovery (Eddy, 1999). It is only recently that some attempts to do this more systematically have been initiated (Olivas *et al.*, 1997; Hüttenhofer *et al.*, 2001). In addition, it is likely that single-base mutations in non-coding RNAs will be hard to detect phenotypically. As is the case for promoters, such sequences may be somewhat more flexible than are protein-coding sequences, especially if the affected RNAs are part of a scale-free network that is resistant to damage (Albert *et al.*, 2000). On the other hand, it is relatively easy to find mutations in genomic sequences encoding non-coding RNAs by insertional and deletional mutagenesis, as in the case for the *Drosophila* bithorax locus referred to above.

Complex genetic phenomena involving RNA

A central role for RNA signaling and RNA metabolism in eukaryotic biology is becoming more obvious. There are a number of poorly understood genetic phenomena in higher eukaryotes which include RNAi, co-suppression, transgene silencing, position effect variegation, imprinting, DNA methylation, X-chromosome dosage compensation and transvection, all of which share features in common (Judd, 1995; Fire, 1999; Jones *et al.*, 2000; Kelley and Kuroda, 2000; Mette *et al.*, 2000; Morel *et al.*, 2000; Sleutels *et al.*, 2000; Wassenegger, 2000; Sharp, 2001). Without going into detail, RNA signals have been shown to be central to, or at least implicated in, all of these phenomena, which involve RNA–RNA and RNA–DNA interactions as well as chromatin remodeling (see Mattick and Gagen, 2001; Sharp, 2001; and references therein). RNAi and post-transcriptional gene silencing in animals and plants is mediated by 21–22 nucleotide RNAs generated by RNase III cleavage from longer double-stranded RNAs (Hammond *et al.*, 2001; Sharp, 2001), a length similar to that required for RNA-directed DNA methylation (Wassenegger, 2000), and which is probably close to the optimal minimum required for stable base-pairing and sequence-specific interactions within complex genomes. While some of these pathways may be utilized in defense against viruses and transposon mobilization (Baulcombe, 2001), it is also clear that they are an integral part of normal cell and developmental biology (see Grishok *et al.*, 2001).

Large families of proteins are involved in RNA metabolism and signaling

It has also become obvious that there are many large gene families which encode proteins involved in RNA metabolism, some of which have come to light by the genetic analysis of RNAi, and

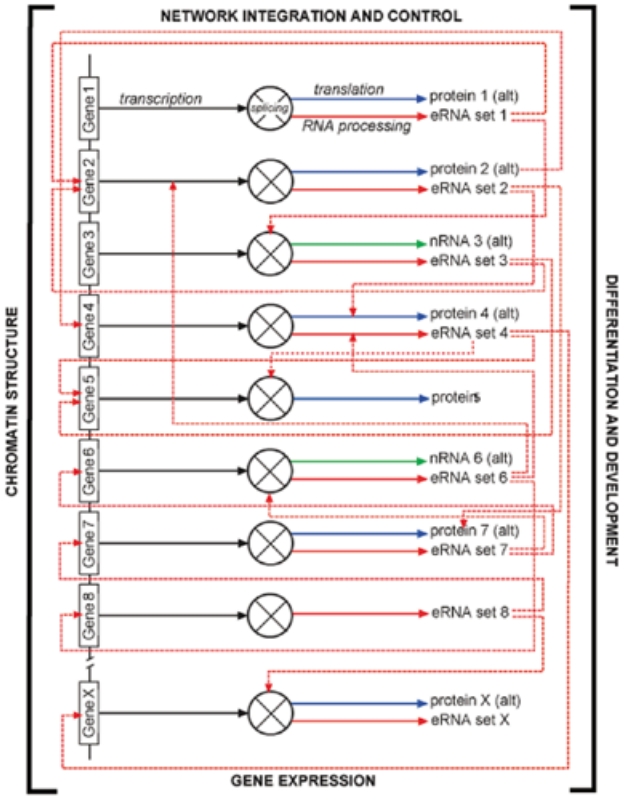

Fig. 2. A more detailed schematic of the proposed role of eRNAs in eukaryotic system networking and control. Genes, packaged in chromatin, express primary transcripts which are then (alternatively) spliced to yield an mRNA and/or *n* introns, which may be further processed to form multiple smaller species, such as *let-7*. Some noncoding RNA genes may yield functional RNAs from both introns and exons (nRNA). These RNAs may then act as signaling or guide molecules to integrate activity at this locus with that of related parts of the network, via effects on chromatin structure, transcription, splicing, other levels of RNA processing, mRNA translation, mRNA stability and other levels of RNA-mediated signal transduction within the cell. The evidence indicates that many if not most of these interactions will be homology (primary sequence) dependent, and involve RNA–DNA, RNA–RNA and RNA–protein interactions, but others may involve secondary or tertiary RNA structures and RNA-mediated catalysis. This scheme is not comprehensive, but is intended to give a sense of the complexity and potential of such networks for programmed control and system integration of complex suites of gene activity in differentiation and development.

which also affect co-suppression and transgene silencing. Apart from RNaseD-type 3′-5′ exonucleases and double-stranded RNase IIIs, of which there are many homologs in metazoan genomes, these include: the Dicer family of proteins that contain similar domains (RNase type III domains and dsRNA-binding domains) together with an RNA helicase domain and a PAZ domain; adenosine deaminases that act on dsRNAs (ADARS); RNA-dependent RNA polymerases; RNA helicases and DExH/D box proteins; the RDE-1 (Argonaute/piwi/zwille) family of proteins found in plants, fungi, invertebrates and mammals (which also contain a PAZ domain), with at least 20 homologs in *C. elegans*; and others identified in genetic screens but yet to be defined biochemically (Cerutti *et al.*, 2000; Fagard *et al.*, 2000; Baulcombe, 2001; Grishok *et al.*, 2001; Schwer, 2001).

Other families of RNA-binding proteins include those with one or more RRM (RNA recognition motif) domains, KH domains and RG domains, among others (Perez-Canadillas and Varani, 2001), and it seems likely that RNA-binding proteins of one sort or another constitute the largest group of proteins in the genomes of the higher eukaryotes. In addition many proteins that are considered to be 'transcription factors', such as Y-box (cold shock) proteins, winged-helix-turn-helix proteins, and zinc finger proteins such as Sp1 and WT1, appear to bind RNA or RNA–DNA hybrids, and may well be recognizing not DNA *per se* but higher order structures involving RNA, as well as associating in complexes with other proteins such as DNA methyltransferase, histone H5 and hnRNP K (Shi and Berg, 1995; Ladomery, 1997; Herbert and Rich, 1999b; Fierro-Monti and Mathews, 2000; Shnyreva *et al.*, 2000).

RNA regulates chromatin architecture

There is also good evidence that RNA regulates chromatin architecture. DNA methylation is RNA-directed, at least in plants and probably in all higher eukaryotes (Wassenegger, 2000). The phenomenon of transvection, or allelic cross-talk, which has been largely described in *Drosophila* but which also occurs in other higher eukaryotes (Wu and Morris, 1999), has been implicated in genomic imprinting and X chromosome inactivation and almost certainly involves *trans*-acting RNA signals (see Mattick and Gagen, 2001). Transvection, co-suppression and transgene silencing have all been shown to involve Polycomb-group proteins (Birchler *et al.*, 2000), which are involved in chromatin remodeling via histone deacetylation (van der Vlag and Otte, 1999; Gebuhr *et al.*, 2000), leading to the suggestion that *trans*-acting RNAs may direct the gene-specific binding of Polycomb complexes (Sharp, 2001).

Importantly, it has recently been shown that a conserved domain called a chromodomain, which occurs in Polycomb-group proteins, as well as in other proteins involved in chromatin remodeling such as the HP1 and CHD families (Jones *et al.*, 2000) and the histone acetyltransferase MOF, is an RNA-binding module (Akhtar *et al.*, 2000). The chromodomain controls sequence and target specificity (Jones *et al.*, 2000) and different Polycomb-group protein complexes function at different genomic sites (Strutt and Paro, 1997). Chromodomain-containing proteins are also involved in position effect variegation (Kennison, 1995). In addition, a non-coding RNA has been shown to act as a transcriptional co-activator for steroid receptors (Lanz *et al.*, 1999), whose action also requires chromatin remodeling and the recruitment of histone acetyltransferases (Zhang and Lazar, 2000). Thus chromatin structure and hence gene expression in higher eukaryotes appears to be controlled not just by protein factors but also by *trans*-acting RNA signals.

RNA networks have parallels with other complex information processing systems

Taken together, these observations suggest that a complex network of RNA signaling with a sophisticated infrastructure operates in higher eukaryotes, which enables direct gene–gene communication and the integration and regulation of gene activity at many different levels, including chromatin structure, DNA methylation, transcription, RNA splicing and processing, RNA translation, RNA stability, and RNA signaling in other pathways (Figure 2). This is reminiscent of network control in other

information processing systems, such as computers and the brain, where control codes (which are mainly internally sourced) are used to integrate and multitask complex patterns of activity (Mattick and Gagen, 2001). Such systems require multiple inputs and outputs, which in neurobiology are referred to as 'efference' signals (Bridgeman, 1995), and it has been suggested that *trans*-acting RNAs may play a central role in regulating gene expression in the brain (see Smalheiser *et al.*, 2001).

A more detailed presentation of the evidence for this hypothesis and its relationship to information processing in other domains is presented in Mattick and Gagen (2001). Such a system has interesting and perhaps instructive analogies with small world networks and dataflow computing. Experimental approaches to testing this hypothesis will include examination of the effects of ectopic production of introns and other non-coding RNAs on gene expression patterns and phenotypic indices, aided by bioinformatic analysis to identify conserved sequences in RNA and DNA that may act as transmitters or receivers in the network, as most of these RNA-dependent effects would appear to be homology-dependent. Comparison of the human, mouse and other mammalian genomes shows a surprisingly large degree of sequence homology outside of protein-coding regions (V. Bonazzi, personal communication; Mayor *et al.*, 2000). If correct, understanding the biology of higher organisms will not simply require understanding of the proteome, which is the focus of so much research at present, but also the identification of all non-coding RNAs, their expression patterns, processing, and signaling pathways. It also suggests that, far from being evolutionary junk, introns and other non-coding RNAs form the primary control architecture that underpins eukaryotic differentiation and development.

Acknowledgements

I would like to thank Michael Gagen (Physics Department, University of Queensland) for many stimulating and informative discussions on the intersection between genetics and information processing systems. I would also like to thank Philip LoCascio (Oak Ridge National Laboratory, TN) for pointing out the similarities between this hypothesis and dataflow computing. Apologies are extended to authors whose work was not cited directly due to space limitations.

References

Akhtar, A., Zink, D. and Becker, P.B. (2000) Chromodomains are protein–RNA interaction modules. *Nature*, **407**, 405–409.

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E. and Tollervey, D. (1999) Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J.*, **18**, 5399–5410.

Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P. and Proudfoot, N.J. (1997) Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.*, **11**, 2494–2509.

Askew, D.S. and Xu, F. (1999) New insights into the function of noncoding RNA and its potential role in disease pathogenesis. *Histol. Histopathol.*, **14**, 235–241.

Baulcombe, D. (2001) Diced defence. *Nature*, **409**, 295–296.

Birchler, J.A., Bhadra, M.P. and Bhadra, U. (2000) Making noise about silence: repression of repeated genes in animals. *Curr. Opin. Genet. Dev.*, **10**, 211–216.

Bridgeman, B. (1995) A review of the role of efference copy in sensory and oculomotor control systems. *Ann. Biomed. Eng.*, **23**, 409–422.

Cerutti, L., Mian, N. and Bateman, A. (2000) Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the piwi domain. *Trends Biochem. Sci.*, **25**, 481–482.

Cho, G. and Doolittle, R.F. (1997) Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.*, **44**, 573–584.

Cousineau, B., Lawrence, S., Smith, D. and Belfort, M. (2000) Retrotransposition of a bacterial group II intron. *Nature*, **404**, 1018–1021.

Davidson, E.H., Klein, W.H. and Britten, R.J. (1977) Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript. *Dev. Biol.*, **55**, 69–84.

Duboule, D. and Wilkins, A.S. (1998) The evolution of 'bricolage'. *Trends Genet.*, **14**, 54–59.

Eddy, S.R. (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev.*, **9**, 695–699.

Eickbush, T.H. (2000) Molecular biology: Introns gain ground. *Nature*, **404**, 940–941.

Elgar, G. (1996) Quality not quantity: the pufferfish genome. *Hum. Mol. Genet.*, **5**, 1437–1442.

Erdmann, V.A., Barciszewska, M.Z., Szymanski, M., Hochberg, A., de Groot, N. and Barciszewski, J. (2001) The non-coding RNAs as riboregulators. *Nucleic Acids Res.*, **29**, 189–193.

Fagard, M., Boutet, S., Morel, J.B., Bellini, C. and Vaucheret, H. (2000) AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl Acad. Sci. USA*, **97**, 11650–11654.

Fierro-Monti, I. and Mathews, M.B. (2000) Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem. Sci.*, **25**, 241–246.

Filipowicz, W. (2000) Imprinted expression of small nucleolar RNAs in brain: Time for RNomics. *Proc. Natl Acad. Sci. USA*, **97**, 14035–14037.

Fire, A. (1999) RNA-triggered gene silencing. *Trends Genet.*, **15**, 358–363.

Gebuhr, T.C., Bultman, S.J. and Magnuson, T. (2000) Pc-G/trx-G and the SWI/SNF connection: developmental gene regulation through chromatin remodeling. *Genesis*, **26**, 189–197.

Gilbert, W., Marchionni, M. and McKnight, G. (1986) On the antiquity of introns. *Cell*, **46**, 151–153.

Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

Grishok, A. *et al.* (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell*, **106**, 23–34.

Hammond, S.M., Caudy, A.A. and Hannon, G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Genet.*, **2**, 1110–1119.

Herbert, A. and Rich, A. (1999a) RNA processing in evolution: The logic of soft-wired genomes. *Ann. N. Y. Acad. Sci.*, **870**, 119–132.

Herbert, A. and Rich, A. (1999b) Left-handed Z-DNA: structure and function. *Genetica*, **106**, 37–47.

Hüttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J.-P. and Brosius, J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Jacquier, A. (1996) Group II introns: elaborate ribozymes. *Biochimie*, **78**, 474–487.

Jareborg, N., Birney, E. and Durbin, R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9**, 815–824.

Jones, D.O., Cowell, I.G. and Singh, P.B. (2000) Mammalian chromodomain proteins: their role in genome organisation and expression. *Bioessays*, **22**, 124–137.

Judd, B.H. (1995) Mutations of zeste that mediate transvection are recessive enhancers of position-effect variegation in *Drosophila melanogaster*. *Genetics*, **141**, 254–253.

Kelley, R.L. and Kuroda, M.I. (2000) Noncoding RNA genes in dosage compensation and imprinting. *Cell*, **103**, 9–12.

Kennison, J.A. (1995) The Polycomb and trithorax group proteins of *Drosophila*: trans-regulators of homeotic gene function. *Annu. Rev. Genet.*, **29**, 289–303.

Komine, Y., Tanaka, N.K., Yano, R., Takai, S., Yuasa, S., Shiroishi, T., Tsuchiya, K. and Yamamori, T. (1999) A novel type of non-coding RNA expressed in the rat brain. *Brain Res. Mol. Brain Res.*, **66**, 1–13.

Ladomery, M. (1997) Multifunctional proteins suggest connections between transcriptional and post-transcriptional processes. *Bioessays*, **19**, 903–909.

Lambowitz, A.M. and Belfort, M. (1993) Introns as mobile genetic elements. *Annu. Rev. Biochem.*, **62**, 587–622.

Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U., Wong, J., Tsai, S.Y., Tsai, M.J. and O'Malley, B.W. (1999) A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell*, **97**, 17–27.

Lipshitz, H.D., Peattie, D.A. and Hogness, D.S. (1987) Novel transcripts from the Ultrabithorax domain of the bithorax complex. *Genes Dev.*, **1**, 307–322.

Liu, A.Y., Torchia, B.S., Migeon, B.R. and Siliciano, R.F. (1997) The human NTT gene: identification of a novel 17-kb noncoding nuclear RNA expressed in activated $CD4^+$ T cells. *Genomics*, **39**, 171–184.

Logsdon, J.M.J. (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.*, **8**, 637–648.

Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.*, **32**, 279–305.

Mattick, J.S. (1994) Introns: evolution and function. *Curr. Opin. Genet. Dev.*, **4**, 823–831.

Mattick, J.S. and Gagen, M.J. (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.*, **18**, 1611–1630.

Maxwell, E.S. and Fournier, M.J. (1995) The small nucleolar RNAs. *Annu. Rev. Biochem.*, **64**, 897–934.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S. and Dubchak, I. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16**, 1046–1047.

Mette, M.F., Aufsatz, W., van Der Winden, J., Matzke, M.A. and Matzke, A.J. (2000) Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J.*, **19**, 5194–5201.

Morel, J., Mourrain, P., Beclin, C. and Vaucheret, H. (2000) DNA methylation and chromatin structure affect transcriptional and post-transcriptional transgene silencing in *Arabidopsis*. *Curr. Biol.*, **10**, 1591–1594.

Moss, E.G. (2000) Non-coding RNA's: lightning strikes twice. *Curr. Biol.*, **10**, R436–R439.

Nakamura, A., Amikura, R., Mukai, M., Kobayashi, S. and Lasko, P.F. (1996) Requirement for a noncoding RNA in *Drosophila* polar granules for germ cell establishment. *Science*, **274**, 2075–2079.

Olivas, W.M., Muhlrad, D. and Parker, R. (1997) Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res.*, **25**, 4619–4625.

Pasquinelli, A.E. *et al*. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.

Pawson, T. (1995) Protein modules and signalling networks. *Nature*, **373**, 573–580.

Perez-Canadillas, J.M. and Varani, G. (2001) Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.*, **11**, 53–58.

Rubin, G.M. *et al*. (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Sanchez-Herrero, E. and Akam, M. (1989) Spatially ordered transcription of regulatory DNA in the bithorax complex of *Drosophila*. *Development*, **107**, 321–329.

Schwer, B. (2001) A new twist on RNA helicases: DExH/D box proteins as RNPases. *Nature Struct. Biol.*, **8**, 113–116.

Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A. and Kondrashov, A.S. (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.*, **17**, 373–376.

Sharp, P.A. (2001) RNA interference-2001. *Genes Dev.*, **15**, 485–490.

Shi, Y. and Berg, J.M. (1995) Specific DNA–RNA hybrid binding by zinc finger proteins. *Science*, **268**, 282–284.

Shnyreva, M., Schullery, D.S., Suzuki, H., Higaki, Y. and Bomsztyk, K. (2000) Interaction of two multifunctional proteins. Heterogeneous nuclear ribonucleoprotein K and Y-box-binding protein. *J. Biol. Chem.*, **275**, 15498–15503.

Sleutels, F., Barlow, D.P. and Lyle, R. (2000) The uniqueness of the imprinting mechanism. *Curr. Opin. Genet. Dev.*, **10**, 229–233.

Smalheiser, N.R., Manev, H. and Costa, E. (2001) RNAi and brain function: was McConnell on the right track? *Trends Neurosci.*, **24**, 216–218.

Stover, C.K. *et al*. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.

Strutt, H. and Paro, R. (1997) The polycomb group protein complex of *Drosophila melanogaster* has different compositions at different target genes. *Mol. Cell Biol.*, **17**, 6773–6783.

Takeda, K., Ichijo, H., Fujii, M., Mochida, Y., Saitoh, M., Nishitoh, H., Sampath, T.K. and Miyazono, K. (1998) Identification of a novel bone morphogenetic protein-responsive gene that may function as a noncoding RNA. *J. Biol. Chem.*, **273**, 17079–17085.

Tam, W., Ben-Yehuda, D. and Hayward, W.S. (1997) *bic*, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. *Mol. Cell Biol.*, **17**, 1490–1502.

Tarrio, R., Rodriguez-Trelles, F. and Ayala, F.J. (1998) New *Drosophila* introns originate by duplication. *Proc. Natl Acad. Sci. USA*, **95**, 1658–1662.

Teramoto, H., Toyama, T., Takeba, G. and Tsuji, H. (1996) Noncoding RNA for CR20, a cytokinin-repressed gene of cucumber. *Plant Mol. Biol.*, **32**, 797–808.

Tycowski, K.T., Shu, M.D. and Steitz, J.A. (1996) A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, **379**, 464–466.

van der Vlag, J. and Otte, A.P. (1999) Transcriptional repression mediated by the human polycomb-group protein EED involves histone deacetylation. *Nature Genet.*, **23**, 474–478.

van Hoof, A. and Parker, R. (1999) The exosome: a proteasome for RNA? *Cell*, **99**, 347–350.

Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wassenegger, M. (2000) RNA-directed DNA methylation. *Plant Mol. Biol.*, **43**, 203–220.

Wolf, Y.I., Kondrashov, F.A. and Koonin, E.V. (2000) No footprints of primordial introns in a eukaryotic genome. *Trends Genet.*, **16**, 333–334.

Wright, F.A. *et al*. (2001) A draft annotation and overview of the human genome. *Genome Biol.*, **2**, 0025.1–0025.18.

Wu, C.T. and Morris, J.R. (1999) Transvection and other homology effects. *Curr. Opin. Genet. Dev.*, **9**, 237–246.

Zhang, J. and Lazar, M.A. (2000) The mechanism of action of thyroid hormones. *Annu. Rev. Physiol.*, **62**, 439–466.

John S. Mattick

DOI: 10.1093/embo-reports/kve230