The Journal of Experimental Biology 210, 1526-1547
Published by The Company of Biologists 2007
doi:10.1242/jeb.005017

A new paradigm for developmental biology

John S. Mattick

ARC Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia

e-mail: j.mattick@imb.uq.edu.au

Accepted 19 February 2007

Summary

It is usually thought that the development of complex organisms is controlled by protein regulatory factors and morphogenetic signals exchanged between cells and differentiating tissues during ontogeny. However, it is now evident that the majority of all animal genomes is transcribed, apparently in a developmentally regulated manner, suggesting that these genomes largely encode RNA machines and that there may be a vast hidden layer of RNA regulatory transactions in the background. I propose that the epigenetic trajectories of differentiation and development are primarily programmed by feed-forward RNA regulatory networks and that most of the information required for multicellular development is embedded in these networks, with cell–cell signalling required to provide important positional information and to correct stochastic errors in the endogenous RNA-directed program.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1526/DC1

Key words: non-coding RNA, intron, regulation.

Introduction

The developmental ontogeny of a human from an embryo to a fully formed adult involves the construction of an organism of approximately 100 trillion cells, with an extremely precise architecture and many differentiated tissues. These include intricately sculpted bones, organs and muscles, such as the dozens of fine muscles in the face (Gray, 1918), as well as a brain that evolves *in situ* in response to experience (Edelman, 1993). This is an extraordinary feat of genetic programming, which in all likelihood, requires enormous amounts of information. This information directs not just a human developmental program, or that of another species, but the idiosyncrasies of the particular program that was inherited by the individual from their parents and their ancestors, as exemplified by the shape of our nose, mouth and ears and other identifying familial features.

How is this feat achieved, and where is this information embedded? In the only well-studied case, the nematode worm *Caenorhabditis elegans*, it is known that developmental ontogeny is precise and invariant, with each cell in the adult being the result of a spatially and temporally ordered progression of cell division, selected apoptosis (programmed cell death) and, ultimately, differentiation into nerve, muscle, gut, germ and other specialized cells (Ambros, 2001; Sternberg and Felix, 1997). Similar processes are observed in the development of insects and mammals (Baehrecke, 2002; McCarthy, 2003), for example in the apoptosis that sculpts the eye ommatidia in the former (Clark et al., 2002) and separates the digits of the fore- and hindlimbs in the latter (Zuzarte-Luis and Hurle, 2005). Thus, it is likely that the ontogeny of higher animals, while vastly more complex and likely to be subject to individual (genomic) variation, is also precisely programmed (Clarke and Tickle, 1999). Indeed, the almost exact identity of monozygotic twins in their physical characteristics and idiosyncrasies, as well as a high degree of concordance in their psychological characteristics (independent of environment), is clear testimony to the precision and reproducibility of the genetic instructions they share.

The genetic programming of development is usually considered to be directed by proteins involved in morphogenetic signalling and various aspects of gene regulation. These include homeodomain-containing proteins, chromatin-modifying proteins, and transcription factors acting on *cis*-regulatory elements, informed by those involved in cell surface receptor and signal transduction systems. Together they form elaborate modular regulatory networks (Arnone and Davidson, 1997; Bantignies and Cavalli, 2006; Levine and Davidson, 2005; Levine and Tjian, 2003) – notwithstanding the recent discovery of microRNAs (see below) that are regarded as an interesting extension of the current paradigm (Davidson, 2006) rather than the vanguard of another entire layer of regulation. This protein-centric perspective underpins most

conceptions of the control of development, as exemplified by elegant studies on sea urchin embryogenesis and fruitfly development (Ben-Tabou de-Leon and Davidson, 2006; Davidson, 2006; Levine and Davidson, 2005; Stathopoulos and Levine, 2005). On the other hand, many proteins are shared in common throughout the metazoa (Duboule and Wilkins, 1998). Moreover, the genomes of *C. elegans* (Stein et al., 2003), which only has 1000 cells, and sea urchins (Sodergren et al., 2006) have essentially the same number of annotated protein-coding genes as those of vertebrates, including humans (Aparicio et al., 2002; International Human Genome Sequencing Consortium, 2004a; International Human Genome Sequencing Consortium, 2004b; Goodstadt and Ponting, 2006; Taft et al., 2007).

All of these observations suggest that significant amounts of relevant information must lie beyond protein-coding sequences, presumably in expanded regulatory regions that control the expression of these proteins (Kleinjan and van Heyningen, 2005; Taft et al., 2007). It also seems likely, although firm conclusions are limited by the poor cDNA library coverage in many species, that the proteome is expanded in more developmentally complex species by the increased use of alternative splicing (Graveley, 2001; Smith and Valcarcel, 2000; Stamm et al., 2005). This in turn, however, mandates an increase in regulation, assuming that cell- or tissue-specific alternative splicing is not random. Thus evolutionary innovation and phenotypic divergence is achieved not only by variations in the structure and function of proteins, but also and probably more so, by those in the regulatory circuitry that controls their deployment (Davidson, 2006; Duboule and Wilkins, 1998; Jacob, 1977; Zuckerkandl and Cavalli, 2007).

Analogue components and digital information transfer in complex systems

Proteins are extraordinarily versatile macromolecules that perform the vast bulk of the catalytic, structural and (to a greater or lesser extent; see below) regulatory functions in biology. As such, proteins (and their derived products such as carbohydrates, lipids and infrastructural RNAs) may be thought of as the analogue components of cells, in the same way that windows, chairs, wheels, gears, sensors and signalling systems comprise the analogue components of bicycles and aircraft. Damage to components usually has severe consequences for the function of the system and is therefore likely to be very evident, although there will be exceptions.

In addition to sophisticated operational controls, complex entities (whether aircraft or organisms) require extensive and detailed design plans for their construction, information about which has also to be stored in the system, along with the specifications of the components themselves. Random changes to assembly plans may have more subtle effects than those that alter component structure (particularly those that compromise component function), creating design variations that often have less severe consequences, although there will be exceptions in both directions. In biology these changes will therefore often result in minor defects, quantitative trait variation or alterations in disease susceptibility. Altered regulatory information has been shown to underlie such variation in a number of cases where it has been possible to map the causative nucleotide changes to completion in well-structured pedigrees (Clark et al., 2006; Clop et al., 2006; Ishii et al., 2006; Smit et al., 2003; Van Laere et al., 2003).

While it has long been recognized that genetic information is encoded digitally in DNA, it has also been widely assumed that the cellular outputs of this information, expressed *via* the intermediate of messenger RNA (mRNA), are almost exclusively analogue components. That is, it has been assumed that most genes are synonymous with proteins and that most genetic information is transacted by proteins. This is essentially true for the prokaryotes, whose genomes comprise densely packed protein-coding sequences, although these genomes clearly also encode a limited number of small regulatory RNAs that function in part by sequence-specific interactions with other RNAs and DNA (Gottesman, 2005; Mattick and Makunin, 2006; Vogel and Sharma, 2005; Winkler, 2005). The situation is similar in unicellular eukaryotes such as the yeasts *Saccharomyces cerevisiae* (David et al., 2006; Olivas et al., 1997) and *Schizosaccharomyces pombe* (Watanabe et al., 2002). Interestingly, although of similar complexity, the former has more protein-coding sequences than the latter, whereas the latter has many more introns (Goffeau et al., 1996; Wood et al., 2002) and a more elaborate RNA signalling infrastructure, which includes the basic components of the RNA interference (RNAi) pathway (Martienssen et al., 2005). This suggests that there may be some trade-off between protein- and RNA-based forms of gene regulation in simple eukaryotes. In any case, at first approximation it is reasonable to say that micro-organisms, particularly the prokaryotes, are in fact largely analogue devices (the 'bicycles' of biology) and that proteins not only comprise the primary structural and catalytic components of these cells but are also the main agents by which they are regulated.

For the past 50 years it has been assumed that the same applies in more complex organisms, i.e. that regulation, particularly developmental regulation, is also largely analogue (protein-based) in multicellular organisms (Davidson, 2006), despite the fact that genome sequence analysis has shown that the numbers of protein-coding genes do not scale strongly or consistently with morphological complexity (Taft et al., 2007) (Fig. 1). This apparently quite reasonable assumption (at least initially) led logically to two subsidiary assumptions: (i) that the increased regulatory sophistication of more complex organisms is achieved through combinatoric interactions of regulatory proteins intersecting with more complex regulatory sequences in promoters and untranslated regions of mRNAs (etc.) (Buchler et al., 2003; Levine and Tjian, 2003); and (ii) that the vast amounts of non-protein-coding sequences in more complex organisms are, apart from a limited amount of *cis*-acting regulatory sequences, evolutionary debris. The latter view has been reinforced by the fact that many of these non-coding sequences are derived from transposons (DNA

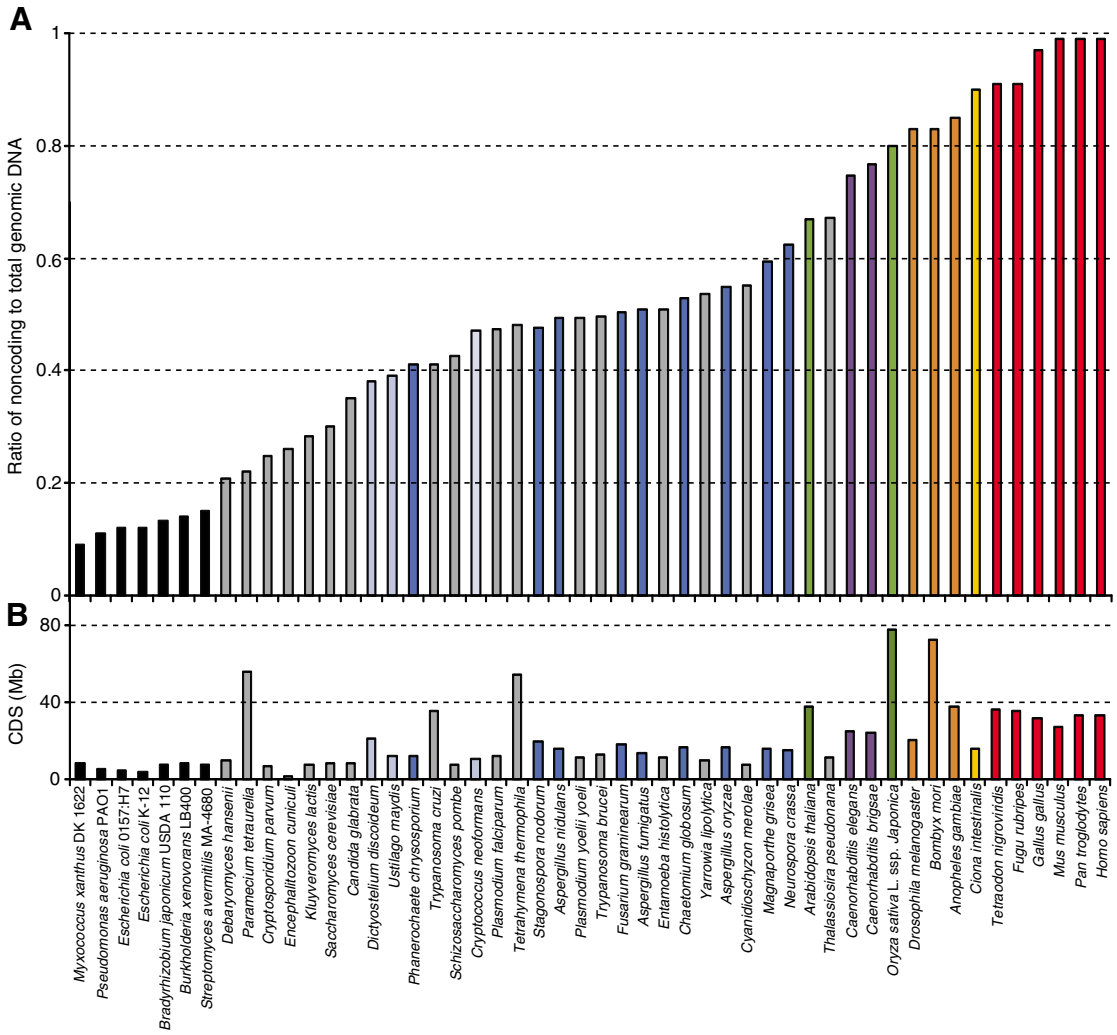

Fig. 1. The fraction of non-protein-coding DNA and megabases of protein coding sequence (CDS) per haploid genome in different species. (A) The ratio of the total bases of non-protein-coding to the total bases of genomic DNA per sequenced genome across phyla (i.e. the fraction of non-protein-coding DNA). The four largest prokaryote genomes and two well-known bacterial species are depicted in black. Single-celled organisms are shown in gray, organisms known to be both single and multicellular depending on lifecycle are light blue, basal multicellular organisms are blue, plants are green, nematodes are purple, arthropods are orange, ascidians are yellow, and vertebrates are red. Species names are listed below B. (B) The amount (in megabases) of CDS per genome for species ranked by fraction of non-protein-coding DNA. Figure adapted from Taft et al. (Taft et al., 2007) with permission from *BioEssays*.

sequences that can move within the genome to new positions), themselves widely assumed to be non-functional, selfish DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980) and to be evolving 'neutrally' (Waterston et al., 2002). These assumptions have remained largely unquestioned for many years and have become articles of faith, but they are not necessarily correct.

Non-linear scaling of regulatory information in integrated systems

In earlier papers it was shown that the requirement for endogenous communication and regulatory information in integrated complex systems, whether cells or computers, scales faster than linearly with function and thus must hit a limit (Gagen and Mattick, 2005; Mattick and Gagen, 2005). This limit can only be relaxed and raised by changing the physical basis and efficiency of the control architecture. In other domains, this limit has been raised by superimposition of digital communication and control systems, using symbolic or sequence-specific strings to store and transmit information within the system. This allows both higher information density and improved transmission accuracy, the latter to overcome the problem of amplified noise (unintended crosstalk) inherent in analogue computation, thereby achieving higher operational sophistication and complexity (see e.g. Collen, 1994). Good examples are the transition from analogue to digital computing (Weinstein and Keim, 1965) and the evolution of aircraft from purely mechanical devices to modern passenger or military jets, wherein a large proportion of the information and cost is

entailed in the computing and software systems, including hundreds of kilometers of optical fiber (Csete and Doyle, 2002). Imagine what a bicycle engineer, or even an aeronautical engineer, might have made of the latter when unexpectedly confronted with it, a situation akin to the discovery of introns in the late 1970s (see below).

It should be noted that the power and precision of digital communication and control systems has only been broadly established in the human intellectual and technological experience during the past 20–30 years, well after the central tenets of molecular biology were developed and after introns had been discovered. The latter was undoubtedly the biggest surprise (Williamson, 1977), and its misinterpretation possibly the biggest mistake, in the history of molecular biology. Although introns are transcribed, since they did not encode proteins and it was inconceivable that so much non-coding RNA could be functional, especially in an unexpected way, it was immediately and almost universally assumed that introns are non-functional and that the intronic RNA is degraded (rather than further processed) after splicing. The presence of introns in eukaryotic genomes was then rationalized as the residue of the early assembly of genes that had not yet been removed and that had utility in the evolution of proteins by facilitating domain shuffling and alternative splicing (Crick, 1979; Gilbert, 1978; Padgett et al., 1986). Interestingly, while it has been widely appreciated for many years that DNA itself is a digital storage medium, it was not generally considered that some of its outputs may themselves be digital signals, communicated *via* RNA[1].

Analysis of prokaryotic genomes has shown that, as predicted (Croft et al., 2003), the numbers of genes encoding regulatory proteins scale almost quadratically with gene number or genome size (Croft et al., 2003; Gagen and Mattick, 2005; Mattick, 2004; Mattick and Gagen, 2005; van Nimwegen, 2003). In addition, extrapolation of these relationships show that the point where the number of new regulatory genes is predicted to exceed the number of new (non-regulatory) functional genes is close to the observed upper size limit of bacterial genomes (Gagen and Mattick, 2004; Gagen and Mattick, 2005). This implies (albeit does not prove) that bacteria have reached a complexity ceiling imposed by the accelerating cost of protein-based regulation, possibly early in evolution. It also implies (i) that the more complex eukaryotes must have solved the problem some other way, most likely by the co-option of RNA as a sequence-specific regulatory molecule [microRNAs (miRNAs) being a good example] and, more subtly, (ii) that the combinatorics of regulatory factors *per se* cannot be used to enlarge the regulatory space to get past this ceiling, as there is no *a priori* reason to expect that prokaryotes could not have easily evolved more complex promoters and recruited additional transcription factors, etc. This in turn suggests that the complex gene regulatory regimes in the higher organisms may operate through multiple layers of regulation and regulatory decisions, rather than multiple (combinatoric) inputs at any given point.

In any case, and consistent with the non-linear scaling of regulatory information, there is a strong relationship between the extent of non-protein-coding DNA sequences in the genomes of higher organisms and their relative complexity. Indeed this appears to be the only consistent relationship between genome information content and complexity (Taft et al., 2007) (Fig. 1). These non-protein-coding sequences occupy almost 99% of the human genome (Frith et al., 2005), and it has been inconceivable to many that they might all be functional as *cis*-acting regulatory elements (although these have clearly expanded in complex organisms). Again this view is implicitly predicated on the assumption that most genetic information is transacted by proteins.

The major output of metazoan genomes is non-coding RNA

In apparent opposition to the above assumption, it is now evident that most of the non-protein-coding sequences in genomes are in fact expressed (i.e. transcribed), either as introns in the primary transcripts of protein-coding genes (which occupy ~40% of the human genome) or as intergenic or antisense transcripts (Frith et al., 2005; Mattick and Makunin, 2006). Indeed it appears that the vast majority of all genomes, from yeast to insects and mammals (wherein most studies have been done), are transcribed, much on both strands (Carninci et al., 2005; Cheng et al., 2005; David et al., 2006; Manak et al., 2006). Both cDNA (Carninci et al., 2005; Katayama et al., 2005; Okazaki et al., 2002) and genome tiling array studies (Cheng et al., 2005; Kampa et al., 2004; Kapranov et al., 2002; Kapranov et al., 2005) of the transcriptome have revealed an extraordinarily complex landscape of interleaved and overlapping transcripts, with distal exons, elaborate splicing patterns and alternative polyadenylation sites, many of which appear to have no protein-coding capacity (Mattick and Makunin, 2006). The most recent data show that at least 85% of the *Drosophila* genome (Manak et al., 2006), 70% of the mouse genome (Carninci et al., 2005) and 93% of the ENCODE regions of the human genome (The ENCODE Project Consortium, manuscript submitted for publication) have experimentally documented transcripts. Moreover, there also appears to be a large and mostly distinct population of non-polyadenylated transcripts located in the nucleus and the cytoplasm, which (despite indications from some very early

[1]On some early occasions it was suggested that RNA may act as a regulatory molecule. The possibility was first mooted briefly by Jacob and Monod in 1961 (Jacob and Monod, 1961) but lapsed when the archetypal gene regulatory factor, the *lac* repressor, was subsequently shown to be a protein (Gilbert and Muller-Hill, 1966). The existence of RNA regulatory networks was first postulated by Britten and Davidson in 1969 (Britten and Davidson, 1969; Davidson et al., 1977), in an attempt to explain the vastly greater complexity of the RNA in the nucleus (then called 'heterogenous nuclear RNA' or hnRNA) compared to the cytoplasm where mRNA is located. Although this paper is of historical importance for first proposing a major role for regulatory mechanisms in the evolution of higher eukaryotes, the idea of RNA regulation itself was not pursued, even following the discovery of introns, despite the fact that this discovery provided an explanation (at least in part) of the origin of hnRNA and an obvious potential source of the co-production of gene regulatory signals from the excised intronic RNA (Mattick, 1994; Mattick and Gagen, 2001).

studies) it was not appreciated existed, because of the widespread use of oligo dT to purify mRNA and to construct cDNA libraries (Cheng et al., 2005).

There are literally tens of thousands of long non-coding RNAs (ncRNAs) that have been identified in mammals (Carninci et al., 2005; Kampa et al., 2004; Okazaki et al., 2002), including many antisense transcripts (Alfano et al., 2005; Cocquet et al., 2005; Katayama et al., 2005; Korneev and O'Shea, 2005; Pandorf et al., 2006; Reis et al., 2004; Tufarelli et al., 2003; Werner, 2005; Werner and Berdal, 2005) and large numbers of smaller RNAs such as miRNAs (Berezikov et al., 2006a; Berezikov et al., 2006b) and piRNAs (Aravin et al., 2006; Girard et al., 2006; Lau et al., 2006). Many of these ncRNAs are expressed in a cell- or tissue-specific manner, suggesting that they are developmentally regulated. Characterized long ncRNAs include *H19* (Barsyte-Lovejoy et al., 2006; Brannan et al., 1990; Wrana, 1994), *7H4* (Velleca et al., 1994), *bic* (Tam et al., 1997), *NTT* (Liu et al., 1997), *BORG* (Takeda et al., 1998), *Xist* (Brockdorff, 1998), *Tsix* (Lee et al., 1999), *DD3* (Bussemakers et al., 1999), *Msx1* (Blin-Wakkach et al., 2001), *Air* (Sleutels et al., 2002), *MALAT-1* (Ji et al., 2003), *adapt33* (Wang et al., 2003), *SCA8* (Mutsuddi et al., 2004), *MIAT* (Ishii et al., 2006), *CTN* (Prasanth et al., 2005), *NFAT* (Willingham et al., 2005), *PRINS* (Sonkoly et al., 2005), *TUG1* (Young et al., 2005), *PINC* (Ginger et al., 2006), *SAF* (Yan et al., 2005), *Evf-2* (Feng et al., 2006), *HSR1* (Shamovsky et al., 2006) and *HAR1* (Pollard et al., 2006), most of which have been associated with specific cellular or developmental functions and/or disease. However, most of the ncRNAs discovered in genome-wide transcriptomic analyses or expressed from particular genomic regions have not been studied in any detail, although high-throughput cell-based and other screening strategies are beginning to be deployed to ascertain their function (Mattick, 2005; Reis et al., 2004; Willingham et al., 2005). Moreover, the documented numbers of these RNAs are conservative estimates: more are being regularly discovered as genomic analyses of one sort or another delve deeper into the transcriptome. Recent evidence suggests that deep sequencing has not remotely exhausted the repertoire of either long ncRNAs (Carninci et al., 2005) or short ncRNAs (Berezikov et al., 2006a; Berezikov et al., 2006b; Cummins et al., 2006; Ruby et al., 2006) and that there may be hundreds of thousands of small RNAs expressed in humans (T. R. Gingeras, personal communication; L. Croft, R. J. Taft and J.S.M., unpublished data).

These observations confront and very largely contradict the traditional protein-centric view of genetic information and genome organization (Mattick and Makunin, 2006). Either the bulk of the transcriptional output from the human genome and those of other complex organisms is random 'noise' (or, in the case of introns, the residue of evolutionary baggage retained and accumulated within genes, as widely assumed) or this transcription comprises a massive but hitherto hidden layer of expression of systemic genetic information that is transacted by RNA (Mattick, 1994; Mattick, 2001; Mattick, 2003; Mattick, 2004). The former has been described as a rather nihilistic view (Werner, 2005), but is one that is comfortable for the prevailing orthodoxy. On the other hand, the latter is strongly supported by the observations that: (i) all well-studied loci in insects and mammals express a large number of non-protein-coding transcripts (e.g. Ashe et al., 1997; Bae et al., 2002; Holmes et al., 2003; Jones and Flavell, 2005; Lemons and McGinnis, 2006; Lipshitz et al., 1987; Sanchez-Herrero and Akam, 1989; Sessa et al., 2007); (ii) many of the experimentally detected ncRNAs are differentially expressed (Carninci et al., 2005; Cheng et al., 2005; Ravasi et al., 2006), apparently under the control of common transcription factors (Barsyte-Lovejoy et al., 2006; Cawley et al., 2004); (iii) at least some have specific subcellular locations (Ginger et al., 2006; Prasanth et al., 2005); and (iv) at least some have been shown to be functional (Brannan et al., 1990; Brockdorff, 1998; Feng et al., 2006; Ginger et al., 2006; Prasanth et al., 2005; Velleca et al., 1994; Willingham et al., 2005; Wrana, 1994; Young et al., 2005).

Microarray analyses have shown that large numbers of ncRNAs are dynamically regulated during the differentiation of embryonal stem cells, myoblasts, neuronal cells and the gonadal ridge, as well as during T-cell and macrophage activation (M. E. Dinger, K. C. Pang, I. Qureshi, M. Crowe, A. C. Perkins, S. M. Grimmond, D. A. Hume, P. A. Koopman, G. E. O Muscat, S. Bruce, M. F. Mehler and J.S.M., manuscript in preparation) and in cancer (Lu et al., 2005; Reis et al., 2004). In addition, *in situ* hybridization analyses are revealing large numbers of ncRNAs that are expressed in particular regions of the brain and in particular subcellular locations (T. R. Mercer, M. E. Dinger, S. Sunkin, M. F. Mehler and J.S.M., in preparation). Many of these ncRNAs are antisense or intronic to genes encoding proteins important in neural development, function and disease. It is also now evident that many of the complex genetic phenomena in complex organisms, including transcriptional and post-transcriptional gene silencing (Cogoni and Macino, 2000; Matzke et al., 2001; Zamore and Haley, 2005), imprinting (Kelley and Kuroda, 2000; Morison et al., 2005; Nikaido et al., 2003) and probably also transvection (Mattick and Gagen, 2001) and transinduction (Ashe et al., 1997), are linked to RNA signalling (Mattick, 2003; Mattick and Gagen, 2001).

Digital–analogue conversion of RNA signals

A key advantage of RNA is its sequence specificity, in that it can direct a precise interaction with its target by base pairing, over short stretches of nucleotides, far more efficiently than can be achieved by proteins. This allows large numbers of regulatory controls to be encoded compactly in genomes, especially as those genomes come under pressure to contain exponentially greater amounts of regulatory information as complexity increases. These regulatory controls can also be flexibly altered and re-configured by evolution to achieve phenotypic variation without altering the underlying components of the system, a concept that is well established in engineering (Mattick and Gagen, 2001). A good case in point

is that of miRNAs, some of which are widely distributed among species and highly conserved while others are species-specific (Berezikov et al., 2006a; Berezikov et al., 2006b), with two documented cases of mutations in miRNA target sites underpinning disease (Abelson et al., 2005) or quantitative trait variation (Clop et al., 2006). RNAs also intrinsically possess much more precise specificity of interactions with other RNAs and DNA than is usually possible by and between proteins, thus potentially improving the precision of the control system and minimizing noise from crosstalk, especially in complex regulatory networks. (The problem of noise was a primary limitation of analogue computers and a primary driving force in the transition to digital computing.) Thus it appears that evolution may have discovered the power of digital communication and control systems a billion years before we did (see below).

However, the sequence-specific interaction of a regulatory RNA with its target is relatively sterile unless this interaction can be converted into a meaningful analogue action. At its simplest level, this may comprise antisense binding to block another interaction, and this primitive mechanism seems to be a common feature of regulatory RNAs in prokaryotes. However, a more sophisticated strategy is to embed secondary signals either in the RNA itself or in the structure of the resulting RNA:RNA or RNA:DNA complex, to recruit different types of complexes, which then undertake the type of analogue action required upon receipt of the signal. Good examples are (i) the complexes of RNA-modifying enzymes that act at a site adjacent to and determined by the position of the sense:antisense interaction between small nucleolar RNAs (snoRNAs) and their targets (Bachellerie et al., 2002; Meier, 2005), and (ii) the RNA-induced silencing (RISC) complexes that act on RNAs bound to small interfering RNAs (siRNAs) and miRNAs (Tang, 2005). Thus, there are two components to RNA signals: a sequence-specific interaction with the intended target(s) and a secondary or tertiary structural component that acts as a transducer to recruit generic infrastructural proteins to impart different types of actions. Indeed, this two-stage principle also applies to other classes of functional RNAs including snRNAs and tRNAs, which recognize splice junctions in pre-mRNAs or codons in mRNAs and recruit the spliceosome or ribosome, respectively. That is, RNAs function as adaptors, with a target sequence-specific address code and separate structural motifs that specify the type of consequent function and bind the appropriate proteins.

Such considerations suggest that a receptive infrastructure for RNA signalling must have co-evolved with the RNA signals themselves and become progressively more sophisticated as RNA regulatory and transport networks gained currency during the evolution of the eukaryotes. Examples include the proteins of the argonaute family and others associated with RNA interference (Carmell et al., 2002), and those containing RRM domains, KH domains, SR domains, SET domains, pumilio-homology domains and double-stranded RNA-binding domains, which occur in a wide range of developmental regulators with global functions (Anantharaman et al., 2002; Bernstein and Allis, 2005; Saunders and Barber, 2003; Wang et al., 2002). Indeed many of the so-called nucleic acid binding proteins and chromatin-binding proteins whose target specificity is uncertain or unknown may in fact recognize different types of RNA signals. This possibility is supported by evidence suggesting that regulatory proteins containing C2H2 zinc fingers (Shi and Berg, 1995), Y-boxes (Ladomery, 1997), chromodomains (Akhtar et al., 2000; Bernstein and Allis, 2005), tudor domains (Maurer-Stroh et al., 2003) and SET domains (Krajewski et al., 2005), and others such as DNA methyl transferases (Jeffery and Nakielny, 2004), may recognize such RNA signals in one form or another.

The origin and evolution of RNA-based regulatory networks in complex organisms

I suggest that the transition from a largely analogue protein-based regulatory control to digitally based RNA regulation was a fundamental rate-limiting step in the emergence of complex organisms (Mattick, 1994; Mattick and Gagen, 2001), together with other factors such as the level of atmospheric oxygen (Canfield et al., 2007). It follows that the RNA-based regulatory systems underpinning the ability to control more complex developmental trajectories must have been largely in place prior to the metazoan radiation and have been a critical factor enabling this evolutionary event (Mattick, 1994; Mattick, 2001; Mattick, 2004; Mattick and Gagen, 2001). Following the emergence of all modern animal phyla at that time, often referred to as the Cambrian explosion (Fig. 2), these new dynasties of multicellular organisms settled down to 'battle it' out in evolutionary competition. This was achieved, firstly, by refining and introducing new adaptations to body plans to improve their competitiveness for survival and reproduction, and to enable the colonization of new ecological niches and new domains such as the land and the air. The latter presented new physical and physiological challenges, which required significant innovations in proteins as well as in the regulatory architecture controlling developmental ontogeny (Bejerano et al., 2004; Kleinjan and van Heyningen, 2005; Mattick and Gagen, 2001). Recent data indicate that many regulatory RNAs, such as miRNAs, emerged in the ancestors of the Bilateria (Hertel et al., 2006; Prochnik et al., 2007) and in major transitions of metazoan evolution, including the advent of the vertebrates and eutherian mammals (Hertel et al., 2006). Secondly, there would have been considerable evolutionary advantage, and therefore pressure, to enhance sensory and cognitive capacities to recognize and respond to opportunities and threats and to alter the environment in favour of better survival and reproduction. This led to the evolution of learning and memory, an even greater mechanistic challenge that almost certainly involved RNA editing as a means of dynamically intersecting the environment with otherwise hardwired genetic information, ultimately leading to the emergence of higher-order cognition (Mehler and Mattick, 2007).

Although RNA is an ancient molecule and may well have been the progenitor of both DNA and proteins (Gesteland et al., 2006), its evolution as a regulatory molecule with associated infrastructure and networks probably had its genesis in the invasion of eukaryotic protein-coding genes by mobile self-splicing group II introns (Cavalier-Smith, 1991; Cousineau et al., 2000; Lambowitz and Zimmerly, 2004; Mattick, 1994; Palmer and Logsdon, Jr, 1991). These sequences occur in prokaryotes (Ferat and Michel, 1993; Martinez-Abarca and Toro, 2000) but are restricted to non-protein-coding sequences by the intimate coupling between transcription and translation (Cavalier-Smith, 1991; Mattick, 1994), thereby restricting the target area for evolutionary experimentation. While RNA regulation occurs in prokaryotes, it is not well developed, just as there is little need for digital control systems in a bicycle. The need to find solutions to the accelerating problem of increasing regulatory sophistication required to underpin multicellular development – ultimately through the co-option of RNA as a compact signalling molecule and later connecting these signals to different types of actions through the co-evolution of different types of RNA binding and effector proteins – might have been felt by both prokaryotes and eukaryotes, but the latter may have had more opportunity to do so, especially given the compartmentalization of their cells. This latter feature probably arose due to the lifestyle of early eukaryotes as phagocytic cellular predators, such as amoebae or macrophages (Cavalier-Smith, 1991). Importantly, the separation of transcription from translation by the introduction of a nuclear membrane allowed introns to invade protein-coding sequences, as their negative effects could be minimized as long as they were (self) spliced out before export to the cytoplasm. In so doing, it also created the raw material for a new round of molecular evolution of RNA signals produced in parallel with protein-coding sequences (Mattick, 1994) (Fig. 2).

The subsequent evolution of the spliceosome occurred by the devolution of the originally *cis*-acting catalytic sequences within introns to *trans*-acting generic co-factors (spliceosomal RNAs) and the recruitment of ancillary proteins. This reduced the internal sequence constraints on the introns, allowing them more freedom to evolve and flexibly explore new functional space (as RNA molecules). It also made their excision from primary transcripts more efficient, perversely providing them with even greater facility to expand and invade other genes (Mattick, 1994). As these RNA networks began to be established, proteins capable of recognizing subsets of signals in these networks would have been selected for, increasing the sophistication of the system. Moreover, it would be expected that increasing numbers of genes would have evolved solely to express RNA as higher-order regulators in this increasingly complex system. This will have occurred at least in part by gene duplication followed by loss of protein-coding capacity, as appears to have happened in *Xist* (the ncRNA controlling X chromosome inactivation in female mammals) (Duret et al., 2006) and in many of the non-protein-coding genes that encode snoRNAs or miRNAs in their introns (Cavaille et al., 2001; Mattick and Makunin, 2005; Rodriguez et al., 2004; Tycowski et al., 1996; Ying and Lin, 2005). Interestingly, many ncRNAs are alternatively spliced (Cocquet et al., 2005; Pang et al., 2005), suggesting that there is an operational distinction between RNA sourced from exons and introns. The other major source of functional RNAs has almost certainly been various other types of mobile (transposable) elements, many of which are derived from small RNAs and have been a potent force in genome evolution and genetic innovation (Brosius, 1999; Brosius, 2005; Waterston et al., 2002).

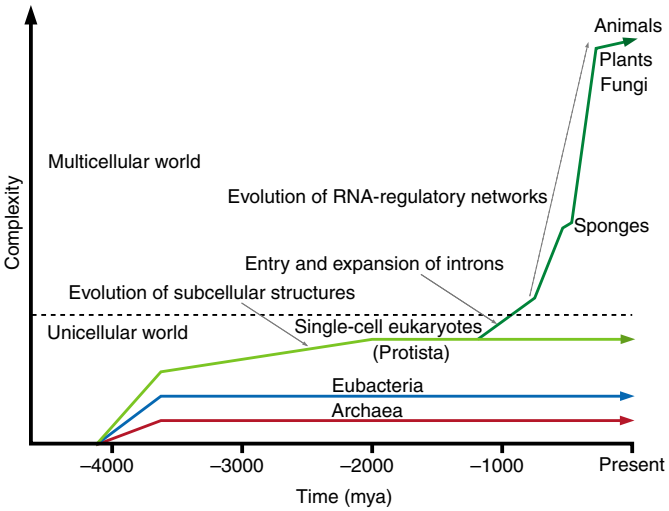

Fig. 2. A simplified view of the biological history of the Earth.

The extent of the genome under evolutionary selection

This raises the question of the composition, rate of evolution and functionality of the genome as a whole, especially as it is now known that most of the genome is transcribed. A large percentage of the mammalian genome (~46% in humans) is composed of transposon-derived sequences (Lander et al., 2001; Waterston et al., 2002), often pejoratively referred to as repeats, and assumed to be non-functional and therefore evolving 'neutrally' (Waterston et al., 2002). The same assumption has often been made about introns, although it is now evident that there are significant amounts of conserved sequences within them (Dermitzakis et al., 2003; Hare and Palumbi, 2003; Sironi et al., 2005), presumably reflecting either functional RNA products or important *cis*-acting regulatory sequences. In any case, on the assumption that ancient repeats (ARs) can be used as a yardstick of the background neutral evolutionary rate, it has been estimated that ~5% of the human genome is under purifying selection in mammals (Waterston et al., 2002), and therefore functional, with the remainder largely considered to comprise genetically inert, neutrally evolving evolutionary debris.

This is in direct contradiction to the suggestion that much of the genome-wide transcription, which is developmentally regulated, is functional. However, it is questionable whether the ARs that are used as yardsticks for these estimations are

really evolving neutrally. First, if ARs have no functional relevance to the organism, they would be expected to evolve freely and eventually to either acquire function or be deleted (M. Pheasant and J.S.M., manuscript submitted for publication), as appears to have occurred with a large fraction of ARs (Waterston et al., 2002). That is, the more ancient the extant sequence, the more likely it is to have acquired function. Second, in agreement with this logic, there are increasing numbers of transposon-derived sequences of all classes, both ancient and modern, including lineage-specific repeats such as *Alu* elements that have been shown to have undergone functional exaptation as gene promoters, regulatory elements, exons and microRNA precursors (Bejerano et al., 2006; Britten, 2006; Brosius, 1999; Dagan et al., 2004; Ferrigno et al., 2001; Hasler and Strub, 2006; Krull et al., 2005; Landry et al., 2001; Lev-Maor et al., 2003; Lippman et al., 2004; Matlik et al., 2006; Nigumann et al., 2002; Smalheiser and Torvik, 2005; Smalheiser and Torvik, 2006; Volff, 2006; Zhou et al., 2002).

These observations throw increasing doubt on the widespread assumption that such sequences are mostly parasitic, and remain as inert genomic passengers. Transposable elements have also been found to underlie the birth of new genes and regulatory networks (Brandt et al., 2005; Cordaux et al., 2006; Landry et al., 2001; Zhou et al., 2002) and to influence early development (Peaston et al., 2004) and phenotypic variation (Whitelaw and Martin, 2001). It is also possible to identify AR sequences that are clearly conserved, some of which are very ancient (Nishihara et al., 2006), such as recently discovered classes of ARs in humans sharing common ancestors with those in marsupials (Kamal et al., 2006) and fish (Ogiwara et al., 2002; Xie et al., 2006), including an example of the slowest evolving regions of the human genome (Bejerano et al., 2006). Moreover, some major classes of ARs show variable rates of sequence conservation within them. One example is the class of so-called ‘mammalian interspersed repeats’ (MIRs), of which there are ~300 000 copies in the human genome (Smit and Riggs, 1995). These MIRs date back ~130 million years and are tRNA-derived SINEs (short interspersed elements) with a consensus length of ~260 nt including a 70 nt central region and 15–25 nt more highly conserved core (Silva et al., 2003; Smit and Riggs, 1995). The fact that hundreds of thousands of such elements have an internal sequence that is conserved more highly than the rest of the element is *prima facie* evidence that this class of ARs (or at least the conserved core within them) is not neutrally evolving and is likely under selection, presumably for function and possibly as regulatory RNAs.

It is also clear that there are widely different rates of evolution of different types of genomic sequences, particularly of gene regulatory sequences, some of which are extraordinarily highly conserved blocks (Bejerano et al., 2004), while many others cover extended genomic regions and exhibit rapid turnover (Fisher et al., 2006; Frith et al., 2006; Smith et al., 2004; Taylor et al., 2006). The latter includes the remarkable functional conservation of regulatory sequences controlling *ret* gene expression in zebrafish and humans, although there is little recognizable primary sequence conservation (Fisher et al., 2006). The *cis*-regulatory elements of the HoxA cluster have also been shown to undergo accelerated evolution, presumably under positive selection during the origin of amniotes and mammals (Wagner et al., 2004). Moreover, it is evident that phenotypic diversification may be due as much, if not more, to changes in regulatory architecture than to the protein components (Duboule and Wilkins, 1998; Levine and Tjian, 2003; Mattick and Gagen, 2001). Indeed, regulatory sequences often exhibit considerable evolutionary plasticity (depending on the number of their interacting targets; see below) and relatively low conservation (Pang et al., 2006) compared with proteins whose evolutionary flexibility is limited by both analogue structure–function relationships and multitasking, i.e. the differential use of the same components in multiple contexts (Duboule and Wilkins, 1998; Mattick and Gagen, 2001).

There are also other regions of the genome under evolutionary constraints that are not evident at the primary sequence level, including shuffled *cis*-regulatory elements (Sanges et al., 2006), gene deserts (Ovcharenko et al., 2005), transposon-free regions (Simons et al., 2006), chromatin domains (Bernstein et al., 2005; Bernstein, B. E. et al., 2006), regions under indel-purifying selection (Lunter et al., 2006), the distances between ultra-conserved elements (Sun et al., 2006) and regions predicted to contain common RNA secondary or tertiary structures (Lescoute et al., 2005; Washietl et al., 2005). Thus, the proportion of functionally meaningful DNA in the human genome is substantially greater than estimated from sequence conservation alone (Smith et al., 2004).

Different rates of evolution also occur within and between different classes of functional gene products, both RNAs and proteins. While most protein-coding sequences are highly constrained and hence highly conserved, some are much more flexible and others have diverged under positive selection (Bustamante et al., 2005). The estimated 5% of the human genome that is conserved with mouse does not include 35% of annotated protein-coding sequences and 17% of RefSeq annotated genes (M. Pheasant and J.S.M., manuscript submitted for publication). Many miRNAs are highly conserved (Pang et al., 2006) but many are not, being lineage- or even species-specific (Berezikov et al., 2006a; Berezikov et al., 2006b). There are also thousands of recently discovered small RNAs (piRNAs) expressed in testis that are not conserved between rodents and humans, although similar RNAs are produced from syntenically orthologous loci (Aravin et al., 2006; Girard et al., 2006; Lau et al., 2006). SnoRNAs have very divergent sequences and many are identifiable only by the loose consensus and positioning of the C/D (RUGAUGA/CUGA) (Shanab and Maxwell, 1992) or H(ANANNA)/ACA boxes (Meier, 2005). It is also clear that many longer functional non-protein-coding RNAs (ncRNAs), such as the *Xist* and *Tsix* transcripts involved in X-chromosome dosage compensation, are evolving quickly (Chureau et al., 2002; Migeon et al., 2001; Nesterova et al., 2001; Pang et al.,

2006). In other cases, there is evidence of recent positive selection in ncRNAs, such as the *HAR1* transcript expressed in particular regions of the brain (Pollard et al., 2006). While functionally validated RNAs do not presently add up to a large fraction of the genome, they do illustrate that lack of conservation does not necessarily equate to lack of function (Pang et al., 2006; Smith et al., 2004). They also point to the likelihood that many functional transcripts, particularly regulatory ncRNAs, are not highly conserved over significant evolutionary distances.

Most of the mammalian genome appears to be evolving more quickly than protein-coding sequences, and at a (regionally adjusted) rate similar to ancient transposon-derived sequences. However, this is evidence simply that the majority of the genome is under similar average selection pressures (M. Pheasant and J. S. Mattick, manuscript submitted for publication), rather than being non-functional and evolving neutrally, although the latter is the favored explanation (Waterston et al., 2002) being consistent with the orthodox view. Moreover, it has been known for some time that the nucleotide substitution frequency varies across the genome. This has often been interpreted as the result of regional variation in the background mutation or fixation (related to recombination) frequencies, rather than selection, as it was (again) inconceivable that the vast intronic and intergenic sequences could be under selection, since that in turn would impute function. Variation in substitution frequencies beyond that which might be expected from random events is also observed at close range within genomic regions, and the data are more consistent with the genome comprising different types of genetic information that are evolving at different rates under different selection pressures and different structure–function constraints (M. Pheasant and J.S.M., manuscript submitted for publication).

Functional constraints on the evolution of regulatory RNAs

Structure–function constraints are different for different types of molecules. As noted already, proteins are analogue components that have quite strict structural specifications. There are only so many ways to construct a wheel, a catalytic site, or an oxygen-binding pocket that is responsive to O_2 and CO_2 partial pressures, and it is hard to vary a successful design. On the other hand, sequence-specific regulatory signals like miRNAs are purely informational and only need to address the right targets; thus at first glance it seems a mystery why many of the known miRNAs have been so fiercely conserved – more so than most protein-coding sequences (Pang et al., 2006) – over 500 million years of evolution from worms to mammals. The exact sequence of these small RNAs does not seem to matter that much: it is easy to design them artificially against almost any sequence, and such siRNAs are now commonly employed as experimental tools (Chalk et al., 2005; Truss et al., 2005). So why have some been so frozen in evolution? The answer appears to be that those miRNAs that were first cloned are common central regulators that have multiple targets (John et al., 2004; Lewis et al., 2005; Lim et al., 2005), which makes co-variation almost impossible in evolutionary terms. If the odds of a miRNA and a target co-varying by compensatory mutations in the same generation are 10^{-5}, the odds of co-variation of an miRNA with 20 targets are 10^{-100}. Most miRNAs that have been subsequently identified through bioinformatics means have also invoked evolutionary conservation as a filter (Berezikov et al., 2005; Jones-Rhoades and Bartel, 2004), thereby likely also restricting their discovery to those that have multiple targets.

Clearly, the level of selection pressure on such sequences will be a function of the number of interactions that must be maintained, rather than the precise sequence itself. Those with one or few interacting partners will be able to evolve relatively freely and also explore new connections in regulatory networks, which themselves can evolve to explore new developmental space, which (given a relatively stable proteome) may be the major route to higher complexity and phenotypic variation. Thus, logic would suggest that there may be many miRNAs that are not highly conserved over significant evolutionary distances, for which there is some supporting evidence (Berezikov et al., 2006a; Berezikov et al., 2006b; Lindow and Krogh, 2005). There is also good reason to expect that some, and perhaps many, miRNAs will have very restricted expression, as exemplified by the miRNA *lsy-6*, which controls left/right neuronal asymmetry in *C. elegans* and is expressed in only a few neurons (Johnston and Hobert, 2003). Indeed, recent deep sequencing shows that the rate of new miRNA discovery continues unabated, albeit with a logarithmic drop as deeper sequencing finds those that are not so highly expressed or are only expressed in a limited subset of cells. Many of these rarer miRNAs are less conserved, being order- or species-specific (Berezikov et al., 2006a; Berezikov et al., 2006b; Cummins et al., 2006). Moreover, if conservation is dropped as a requirement for the bioinformatics prediction, there are well over 1 million plausible miRNA precursor (stem–loop) structures in the mammalian genome, with a large fraction showing evidence of producing small RNA products in array-based assays (L. Croft, R. Taft and J.S.M., unpublished observations).

Other newly discovered classes of putative small regulatory RNAs, such as the 26–31 nt piRNAs (Aravin et al., 2006; Girard et al., 2006; Lau et al., 2006) and 21 nt 21U-RNAs (Ruby et al., 2006), show little long range evolutionary conservation. Many longer ncRNAs exhibit short-range sequence conservation only in small patches (Pang et al., 2006), as exemplified by the case of *Xist* in mammals, even though mutational studies have suggested that most of the molecule is functional (Nesterova et al., 2001). Thus, it seems safe to predict that the sequence of many, if not most, regulatory RNAs will not be highly conserved over significant evolutionary distances, even in cases of conserved function, due to more relaxed structure–function constraints (allowing rapid drift) and to selection pressures for adaptive radiation by altering the endogenous regulatory circuitry (network structure) underpinning developmental processes.

Endogenous feed-forward control of development by RNA networks

The simple logic is that if all of the transcribed and processed ncRNAs are functional, these ncRNAs must in the main be regulatory, because catalytic versatility is not the *forte* of RNA, notwithstanding its central role in splicing and translation and the identification of catalytic RNAs in other contexts (Gesteland et al., 2006). This is not to deny that some RNAs may have interesting (and as yet unappreciated) catalytic functions (Salehi-Ashtiani et al., 2006), or that secondary structural motifs or domains in RNA may be important mediators of interactions with proteins. Nonetheless, if the major function of the massive numbers of ncRNAs transcribed from animal, and particularly mammalian, genomes is regulation, as is likely, the logical extension is that the main (but not exclusive) role of such regulation is to control differentiation and development, rather than (simply) the short-term physiological responses of terminally differentiated cells. In summary, if functional, these RNAs must be mainly regulatory and, if so, their major regulatory function must be to direct development.

This conclusion is well supported by what we currently know or suspect of RNA regulation at many different levels of gene control (see below), but has one very profound implication: that the enormous amount of information required to program development is endogenously embedded in these RNA networks and that most regulatory transactions during development are directed by RNA, albeit mediated by proteins and supplemented by external cues that are conveyed by proteins (see below).

These RNA (and protein) networks, initially laid down by transcription in the female (and also possibly the male) gamete, create an epigenetic state that is asymmetric in the fertilized embryo and that is asymmetrically inherited by daughter cells. Thus each of the daughter cells has a defined subsequent state and is on a pre-programmed pathway of division and differentiation controlled by internal and external cues, the latter of which probably becomes operative at the time of syncytial formation in insects and morula formation in mammals. Thus, every cell in the developing organism contains an epigenetic memory[2] of what its pathway has been, and where it is headed. In computer science, this is akin to what is termed a dynamical recurrent neural network (Aussem et al., 1995; Sudharsanan and Sundareshan, 1994), in which the current state of the network (in this case the gene regulatory and expression network) is defined as the combination of part history and current (external) inputs.

This information about the state of the network (and the embedded trajectories) is enclosed in the structure of the chromatin (almost certainly itself controlled by RNA signalling; see below), the protein repertoire (also directed and regulated by RNA; see below) and, ultimately, the RNA networks that are current in individual cells. These RNA networks have been described as the cellular 'soft wiring' or 'ribotype' (Herbert and Rich, 1999a; Herbert and Rich, 1999b). Thus, RNA transcription and processing may be thought of as a series of steps, one or more of which have two mutually exclusive outcomes: a default outcome and an alternative outcome that is controlled by appropriate regulatory signals. These outcomes can be used either to regulate cellular responses directly or to control other RNA processing events, the latter forming networks wherein the processing of one RNA (either to produce more regulatory RNAs or alternative splice variants of mRNAs) is sequentially contingent on another (Herbert and Rich, 1999a; Herbert and Rich, 1999b).

While such networks would be clearly subject to natural selection (Herbert and Rich, 1999a; Herbert and Rich, 1999b; Mattick, 1994), I suggest that they now dominate the genomic programming of complex organisms and are the primary drivers of development in an unfolding cascade of regulatory interactions that gives each cell a unique identity and vectorial place in the developmental trajectory. This therefore constitutes an endogenous feed-forward regulation of differentiation and development, which is largely predetermined by embedded unfolding RNA networks. Thus, the current behaviour and trajectory of each cell are determined by the networks operative in the preceding cell or state, until the terminal state is reached, at which point the cell cycle is suspended and differentiation completed. This also suggests that there are, in fact, ~10^{14} different cells (i.e. cells with a specific history and identity) in humans, leaving aside those that may have clonally expanded during (e.g.) fat storage or immune responses.

Parallel expression of exonic sequences and efferent RNA signals

An important feature of the proposed exaptation of introns as a source of *trans*-acting RNAs is the potential to produce regulatory signals in parallel with mRNA sequences (and other non-coding RNAs) that may then make contacts to alter settings at multiple loci or targets (Fig. 3). This is akin to what is described by neurobiologists as 'efferent signals' (which are essential to motor coordination, cognition and memory) (Andersen et al., 1997; Bridgeman, 1995; Elman, 1998; Plunkett et al., 1997) and would in theory, and possibly practice, permit much more complex communication and control networks to operate in different cells and states during ontogeny (Mattick, 1994; Mattick, 2001; Mattick and Gagen, 2001). Almost all snoRNAs and a large proportion of miRNAs in animals are encoded within introns (Baskerville and Bartel, 2005; Cai et al., 2004; Mattick and Makunin, 2005; Rodriguez et al., 2004; Ying and Lin, 2005). Moreover, many snoRNA and miRNA gene loci appear to be polycistronic (Cavaille et al., 2002; Huang et al., 2004; Lau et al., 2001; Runte et al., 2001; Seitz et al., 2004). Although introns are thought to be degraded after excision from primary transcripts (Padgett et al., 1986), there is good evidence that intronic RNAs may

[2]This memory can be quite plastic and can be modulated by contextual cues (cell signalling), set in a new direction by artificial translocation of the cell to a new context, or (in some cases) recapitulated when required, such as during the regeneration of fingertips, tails, limbs or rays in mammals, lizards, axolotls and starfish.

actually be processed to smaller RNAs with significantly long half-lives and specific subcellular locations (Clement et al., 2001; Clement et al., 1999). Recently, it was shown that ectopic expression of intronic sequences derived from the *CFTR* gene causes specific changes in transcription of various genes in HeLa cells, with different intron sequences resulting in a distinctive pattern of effects on specific subsets of genes (Hill et al., 2006). There is also evidence that coding and noncoding regions contain sequences that match others in the genome in functionally congruent networks (Rigoutsos et al., 2006).

Layers of RNA-directed control of gene expression in development

RNA is known or strongly implicated to be involved in the regulation of gene expression (both protein-coding and non-coding) at all levels in animals, creating extraordinarily complex hierarchies of interacting controls. This includes chromatin modification and associated epigenetic memory, transcription, alternative splicing, RNA modification, RNA editing, mRNA translation, RNA stability, and cellular signal transduction and trafficking pathways.

Chromatin structure and epigenetic memory

The fine control of chromatin structure is one of the major hallmarks of eukaryotes and of gene regulation in multicellular development (Margueron et al., 2005). Chromatin architecture is altered by DNA modification (methylation) and histone modifications of various types (including compound patterns of methylation, acetylation and phosphorylation at various residues) (Lam et al., 2005; Peterson and Laniel, 2004) in different ways at many different loci in different cell lineages.

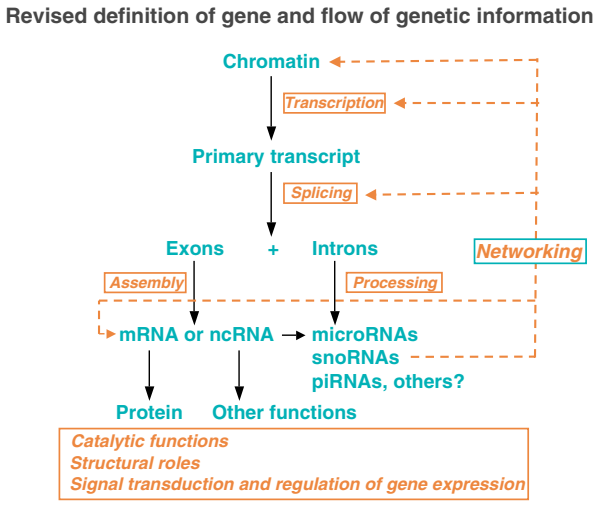

Fig. 3. The flow of genetic information in higher eukaryotes. Primary transcripts may be (alternatively) spliced and further processed to produce a range of protein isoforms and/or ncRNAs of various types, which are involved in complex networks of structural, functional and regulatory interactions.

This involves proteins such as the polycomb group and trithorax group, which mediate repressive and permissive effects, respectively (epigenetic memories), on gene expression in development (Bantignies and Cavalli, 2006; Cernilogar and Orlando, 2005; Lund and van Lohuizen, 2004). As there are only a limited number of enzymes (DNA methyltransferases, histone acetylases and deacetylases, etc.) that perform these modifications, there must be some other signal that specifically directs these modifications to the myriad of target loci around the genome. Indeed, in the absence of an army of DNA sequence-specific binding proteins, the only logical alternative is RNA signals.

While the details of this putative RNA signalling are unknown, there is a great deal of evidence to support its existence (Andersen and Panning, 2003; Bernstein and Allis, 2005; Lippman and Martienssen, 2004; Schmitt and Paro, 2006). This includes the observations that (i) DNA methytransferase and some domains in chromatin remodelling enzymes and binding effector proteins, such as SET, tudor domains and chromodomains, appear to interact with RNA (Bernstein and Allis, 2005; Jeffery and Nakielny, 2004; Sanchez-Elsner et al., 2006), (ii) many regulatory regions affecting chromatin structure and the expression of adjacent protein-coding genes are themselves transcribed in spatially and temporally regulated ways (Bae et al., 2002; Lipshitz et al., 1987; Sanchez-Elsner et al., 2006), and (iii) such non-coding transcripts play important roles in activation of gene expression by targeting global protein regulators such as HP1, Ash1 and the chromatin insulator protein CP190 to the cognate sequences in *cis*-regulatory response elements, including polycomb- and trithorax-response elements (PREs and TREs) (Grimaud et al., 2006; Lei and Corces, 2006b; Maison et al., 2002; Sanchez-Elsner et al., 2006; Schmitt et al., 2005) (see also below). It also includes the well-characterized roles of RNAs in DNA methylation and transcriptional silencing in plants (Aufsatz et al., 2002; Mette et al., 2000; Wassenegger, 2000) and in animals (Bayne and Allshire, 2005; Imamura et al., 2004; Jeffery and Nakielny, 2004; Morris et al., 2004; Ting et al., 2005; Tufarelli et al., 2003; Weiss et al., 1996), imprinting in mammals (Sleutels et al., 2002), heterochromatin formation in *Drosophila* (Birchler et al., 2004; Pal-Bhadra et al., 2004), global activation or repression of sex chromosomes for dosage compensation in insects and mammals (Andersen and Panning, 2003), RNA interference-mediated heterochromatin assembly and chromosome dynamics in fission yeast (Martienssen et al., 2005; Verdel and Moazed, 2005), meiosis (Cho et al., 2005; Watanabe et al., 2001), and programmed DNA elimination in *Tetrahymena* (Mochizuki and Gorovsky, 2004). More recently it has been shown that a specialized set of RNAi components, including members of the argonaute family, are required for DNA methylation in plants (Qi et al., 2006) and yeast (Irvine et al., 2006), as well as transcriptional gene silencing and associated alterations to chromatin structure involving polycomb recruitment in *Drosophila* (Grimaud et al., 2006) and in human cells (Kim et al., 2006). This indicates that the RNAi machinery may regulate higher-order nuclear organization to

orchestrate gene expression during development (Lei and Corces, 2006a). The nuclear organization of chromatin insulators is also affected by the RNAi machinery (Lei and Corces, 2006b).

The proteins of the polycomb group (PcG) and trithorax group (TrxG) are important global regulators of transcriptional silencing and activation and mediators of epigenetic memory in development, best characterized in homeotic loci (Boyer et al., 2006; Guenther et al., 2005; Negre et al., 2006; Ringrose and Paro, 2004; Schwartz et al., 2006; Schwartz and Pirrotta, 2007; Squazzo et al., 2006). Both PcG and TrxG are recruited to genomic elements (termed PREs and TREs, respectively) that encompass hundreds of base pairs. These elements have a very weak consensus in *Drosophila* and none have been identified yet in mammals (Ringrose and Paro, 2004; Ringrose and Paro, 2007). Although five proteins associated with PcG or TrxG complexes (GATA, PSQ, Zeste, PHO and PHO-like) have DNA-binding properties, they bind to rather degenerate sequences and have not been demonstrated to have a role in target recognition *in vivo* (Ringrose and Paro, 2004). Moreover many PREs/TREs are transcribed as ncRNAs (Schmitt et al., 2005), and *Hox* gene loci exhibit complex patterns of non-coding transcripts on both strands (Carninci et al., 2005; Engstrom et al., 2006). The activation of the *HoxA* genes is also accompanied by intergenic antisense ncRNA transcription (Sessa et al., 2007). These observations, together with recent data suggesting that such transcripts and the RNAi pathway play a central role in PcG- and TrxG-mediated epigenetic regulation (Bernstein, E. et al., 2006; Grimaud et al., 2006; Lei and Corces, 2006a; Sanchez-Elsner et al., 2006; Schmitt et al., 2005), suggest that the specificity of this process is controlled by RNA. Thus, the locus- and stage-specific epigenetic modification of chromatin by proteins with global functions may be viewed as the first derivative of a genomically encoded developmental program that is elaborated *via* unfolding RNA regulatory networks, informed by contextual cues and modulated by environmental inputs.

Transcription

There is also increasing evidence that transcription itself is influenced, directly or indirectly, by RNA signalling (Goodrich and Kugel, 2006; Kim et al., 2006). Not only do certain classes of transcription factors either bind RNA or have high affinity for nucleic acid structures involving RNA (Ladomery, 1997; Shi and Berg, 1995), but also transcription has been shown to be both inhibited by single-stranded RNA directed at transcription start sites (Janowski et al., 2005) and activated by double-stranded RNAs (dsRNAs) directed at promoter sequences (Li et al., 2006). The latter requires the Argonaute 2 (Ago2) protein and is associated with a loss of lysine-9 methylation on histone 3 at dsRNA-target sites (Li et al., 2006). The β-globin LCR (‘locus control region’), which is considered to be the archetypal long-distance transcriptional ‘enhancer’, is itself specifically transcribed in erythroid cells (Ashe et al., 1997). Enhancers controlling expression of homeotic and other genes are also specifically transcribed (Feng et al., 2006; Jones and Flavell, 2005; Ronshaugen and Levine, 2004). It has also been shown that transactivation of the steroid receptor, as well as MyoD (which regulates skeletal myogenesis), requires the ncRNA called SRA (Caretti et al., 2006; Hube et al., 2006; Lanz et al., 1999; Lanz et al., 2002). The ncRNA 7SK is involved in the transcriptional activation of the proto-oncogene *c-myc* (Krause, 1996), among other examples (Goodrich and Kugel, 2006).

Splicing

There is an enormous amount of post-transcriptional processing of RNA, both protein-coding and non-coding, much of which involves and is probably regulated by other RNAs. The mechanism of control of alternative splicing, the other major hallmark of developmentally complex organisms, is not known, but a range of circumstantial evidence suggests that this process too is controlled by RNA signals. This evidence includes: (i) that alternative splicing choices are not well explained by what is known about proteins involved in splicing or splicing regulation, despite speculations about combinatorial interactions (Blencowe, 2006; Caceres and Kornblihtt, 2002; Pozzoli and Sironi, 2005; Soller, 2006); (ii) that alternative splice sites are generally more highly conserved than constitutive splice sites (suggesting that a sequence-specific *trans*-acting sequence is required to address the former) (Sorek and Ast, 2003; Sugnet et al., 2004; Sugnet et al., 2006); and, most convincingly, (iii) the well-established observation that synthetic RNA derivatives directed against splice sites can easily alter splicing patterns both *in vivo* and *in vitro* (Garcia-Blanco, 2005; Gendron et al., 2006; Kole and Sazani, 2001; Roberts et al., 2006; Wilton and Fletcher, 2005). If this can easily be achieved by artificial means, then it is not unlikely that nature will employ a similar mechanism. It is not immediately obvious where such regulatory RNAs may be sourced, as conserved splice sites do not have obvious orthologous sequences elsewhere in the genome. However, it is possible that antisense transcripts are the source of these signals (Yan et al., 2005). It is also possible that, given the high affinity of RNA:RNA interactions, the antisense elements in putative *trans*-acting RNAs are short and difficult to identify, as they are in reverse when trying to identify the possible targets of orphan (non-rRNA directed) small nucleolar RNAs (Cavaille et al., 2000) (see below).

RNA modification and RNA editing

There is also considerable post-transcriptional modification and editing of RNA in eukaryotes, especially complex eukaryotes. SnoRNAs range from 60 to 300 nucleotides in length and guide the site-specific modification of target RNAs *via* short regions of base pairing. There are two major classes: (i) the box C/D snoRNAs, which guide 2′-O-ribose-methylation, and (ii) the box H/ACA snoRNAs, which guide pseudo-uridylation of target RNAs (Bachellerie et al., 2002; Henras et al., 2004; Kiss et al., 2004; Meier, 2005). The action of snoRNAs was initially thought to be restricted to rRNA modification in the nucleolus during ribosome biogenesis, but

it is now evident that they can target other RNAs, including small nuclear (spliceosomal) RNAs and mRNAs (Bachellerie et al., 2002; Henras et al., 2004; Kishore and Stamm, 2006; Kiss et al., 2004; Meier, 2005). A subset of box H/ACA snoRNAs is located in Cajal bodies (a class of small nuclear organelle), and are sometimes called scaRNAs (small Cajal body RNAs) (Meier, 2005), where they modify telomerase RNA in a cell-cycle dependent manner (Jady et al., 2004; Jady et al., 2003). At least some snoRNAs exhibit tissue-specific and developmental regulation and/or imprinting (Cavaille et al., 2000; Cavaille et al., 2002; Cavaille et al., 2001; Rogelj and Giese, 2004), which is indicative of a regulatory function. There are also a number of so-called 'orphan' snoRNAs without known targets (Cavaille et al., 2000; Cavaille et al., 2002; Cavaille et al., 2001; Huttenhofer et al., 2001; Kiss et al., 2004; Vitali et al., 2003), one of which has recently been shown to be involved in the aberrant splicing of the serotonin receptor 5-HT(2C)R gene in Prader–Willi syndrome patients (Cavaille et al., 2000; Kishore and Stamm, 2006).

RNAs may also be edited by enzymes termed ADARs (*A*denosine *D*eaminases *A*cting on *R*NAs), which catalyze the deamination of adenosine to inosine to alter coding capacity, splicing patterns or regulatory functions, and also by the APOBEC family of cytidine deaminases, which catalyze C-U/C-T editing of both RNA and DNA (Navaratnam and Sarwar, 2006). The targets of RNA editing include not only mRNAs but also miRNAs and other ncRNAs whose functions are as yet unknown (Athanasiadis et al., 2004; Blow et al., 2004; Blow et al., 2006; Levanon et al., 2004; Yang et al., 2006). RNA editing appears to be the major mechanism by which environmental signals overwrite encoded genetic information to modify gene function and regulation, particularly in the nervous system, where it is well documented to modify transcripts encoding proteins involved in fast neural transmission. These include ion channels and ligand-gated receptors (Bass, 2002; Valente and Nishikura, 2005) such as the serotonin receptor, which is regulated in the same region by snoRNA-mediated RNA modification (Kishore and Stamm, 2006). In humans, where RNA editing is considerably more prevalent than in mouse (Athanasiadis et al., 2004; Blow et al., 2004; Levanon et al., 2004), RNA editing alters many transcripts from genomic loci encoding proteins involved in neural cell identity, maturation and function. This implies a role for RNA editing not only in the regulation of neural transmission but also of brain development (Mehler and Mattick, 2007).

mRNA translation and stability

It is now well established that mRNA translation and mRNA stability are controlled by miRNAs, primarily directed at sequences in the 3′ untranslated region (UTR). 3′ UTRs have expanded greatly during metazoan evolution and in humans occupy over 1% of the genome, accounting for almost as much of the mRNA sequences as the protein-coding sequences themselves (Frith et al., 2005), and suggesting that extremely complex regulatory controls are embedded within them. miRNAs have been shown to be centrally involved in gene regulation in both plants and animals (Bartel, 2004; Carrington and Ambros, 2003; Mattick and Makunin, 2005; Pasquinelli et al., 2005), including flowering in plants (Chen, 2004) and many aspects of development (Bernstein et al., 2003; Giraldez et al., 2005; Hornstein et al., 2005; Kanellopoulou et al., 2005; Ronshaugen et al., 2005), cell growth and differentiation (Baehrecke, 2003; Brennecke et al., 2003; Chen et al., 2004; Hatfield et al., 2005; Johnston and Hobert, 2003; Kuwabara et al., 2004; Naguibneva et al., 2006; Wienholds et al., 2005) in animals. miRNA regulation has also been shown to be perturbed in developmental abnormalities including cancer (Croce and Calin, 2005; Esquela-Kerscher and Slack, 2006; Hammond, 2006) and possibly other diseases (Abelson et al., 2005), as well as in quantitative trait variation (Clop et al., 2006). Some miRNAs have also been shown to regulate *Hox* gene expression (Hornstein et al., 2005; Mansfield et al., 2004; Naguibneva et al., 2006; Yekta et al., 2004) and to exhibit expression patterns reminiscent of *hox* genes in embryonic development (Mansfield et al., 2004). Moreover, as noted above, while there are ~10^3 known miRNAs, there may be far more expressed in mammals. It is also worth noting that neither specific nor general biological functions have yet been ascribed to the thousands of piRNAs that are expressed in mammalian testis, although they are known to interact with the Piwi subfamily of Argonaute proteins, which are required for germ cell maintenance and meiosis (Aravin et al., 2006; Girard et al., 2006; Lau et al., 2006). The same is true of the class of 21U-RNAs recently discovered in *C. elegans* (Ruby et al., 2006).

RNA intersection in signalling cascades and other aspects of cell biology

While it is already clear that various proteins involved in gene regulation have RNA binding domains or domains that intersect with complexes involving RNA, there is also evidence that proteins involved in cellular signal transduction cascades also bind RNA. This is exemplified by the RasGAP-binding protein G3BP/rasputin, which contains both an RNA recognition motif (RRM) and SH3 binding domains (Irvine et al., 2004; Pazman et al., 2000; Zekri et al., 2005). There is also evidence that ncRNAs may be involved in regulating nuclear factor trafficking (Willingham et al., 2005), and the large numbers of ncRNAs that appear to have a cytoplasmic location (Cheng et al., 2005) suggest that many other cellular functions are also regulated by such RNAs.

The role of proteins in development

It is clear that proteins, many of which (such as homeotic proteins, signalling proteins and transcription factors) are referred to as regulatory and are differentially expressed in different cells and tissues, are intimately intertwined with regulatory RNAs in the control of development, and that the boundaries between them may often be blurred. However, without putting too fine a point on it, I suggest that there are

two general classes of proteins involved in developmental regulation.

The first class encompasses those whose role is to transmit contextual signals from the cell and the external environment (other cells and circulating signals) into the gene regulatory networks of the cell. It includes secreted proteins (as well as other ligands such as steroid hormones) that act locally or systemically and their receptors, for example the patched-hedgehog (Murone et al., 1999) and Wnt-frizzled systems (Gordon and Nusse, 2006), which may be positioned asymmetrically on different parts of the cell surface. This class also includes internal protein kinase-mediated signal transduction cascades. These signals are critical to the fidelity of the developmental process, both as feedback controls to correct (inevitable) stochastic errors in the endogenous RNA-directed program, and as important additional positional information to supplement the endogenously specified developmental program. For example, imagine a robot that has been given full instructions for the specification and assembly of a motor vehicle but is denied any environmental reference information (through vision, touch, etc.). It would be impossible to design a program whose execution would be sufficiently precise as to preclude the necessity for feedback controls or (particularly in the case of self assembling multicellular systems) remove the enormous advantages of positional information and cell–cell communication during growth and development. However, as noted already, simply because environmental signalling is critical to the process of ontogeny, this does not mean that this is where the majority of the relevant information is embedded.

The second class of 'regulatory' proteins important for development encompasses those that effect analogue functions to control gene expression at various levels. These proteins are directed to the appropriate site of action in many, if not most, cases by RNA signals, albeit also influenced (activated or repressed) by intersections with the protein-based signal transduction systems that usually operate *via* phosphorylation. These effector proteins include homeotic and other types of chromatin-modifying proteins, transcription factors, splicing factors, RNA editing enzymes, RNA modification enzymes, and Argonaute proteins and others in RISC complexes. Many of these proteins are themselves developmentally regulated at the transcriptional and post-transcriptional level, and are contributory variables in the complex matrix of RNA:DNA:protein interactions and the resultant regulatory networks.

Genetic signatures of RNA regulatory networks

If RNA regulatory networks pervade the cell and developmental biology of complex organisms in such a profound manner, why have they not been recognized sooner, especially in genetic screens? Apart from the fact that the sheer complexity of the ncRNA population has only recently been revealed by sophisticated transcriptomic analysis (genome-scale tiling arrays, extensive cDNA libraries and, most recently, deep sequencing of small RNA fractions) and the possibility that regulatory networks may be intrinsically robust, most genetic screens have suffered a strong expectational, perceptual and technical bias towards mutations in protein-coding sequences. The expectational bias derives from the long-held orthodoxy that most genes encode proteins and their *cis*-acting regulatory elements. This has been reinforced by the perceptual bias that mutations in proteins (as key analogue components of the system) will, in most cases, produce a strongly impaired and often visibly affected phenotype. In contrast, those in regulatory circuits will, in many if not most cases, produce more subtle effects that may not be noticed at all, except in sensitive genetic screens such as that which identified the miRNA *lsy-6* in *C. elegans* (Johnston and Hobert, 2003). Indeed, the entire world of miRNAs and their central role in regulating differentiation and development lay hidden for many years despite intense genetic scrutiny of fruitflies and mammals. It was only revealed by the characterization of the small RNA products of the *let-7* and *lin-4* loci, which control developmental timing in *C. elegans* (Lee et al., 1993; Reinhart et al., 2000), and the intersection of these findings with the characterization of the similar-sized small RNAs produced by RNAi (Hammond et al., 2000; Zamore et al., 2000), also discovered in *C. elegans* (Fire et al., 1998). This suggested that a similar mechanism may produce (other) short regulatory RNAs (Grishok et al., 2001; Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Ruvkun, 2001).

There may also be a difference between protein-coding sequences and those encoding regulatory sequences (whether acting in *cis* or in *trans via* RNA) in terms of their functional sensitivity to point mutations, which comprise the bulk of the natural and induced mutations in mammals. On the other hand, many non-coding regulatory mutations have been known for a long time in *Drosophila*, where many mutations have been obtained by deletion or insertion. However, these have almost inevitably been interpreted as affecting *cis*-regulatory DNA elements (Duncan, 2002), despite the fact that many of the regions concerned, such as *bxd* in the *bithorax* complex (Lipshitz et al., 1987; Petruk et al., 2006), which includes PRE/TRE response elements (Tillib et al., 1999), are known to be transcribed into separately regulated ncRNAs (i.e. may represent separate genetic units) and to be involved in the complex and still poorly understood genetic phenomena of transvection and polycomb-mediated developmental memory (Mattick and Gagen, 2001). Apart from a few cases of regulatory mutations affecting quantitative traits that have been mapped to completion in well-structured animal pedigrees, most screens in mammals (especially in humans) progress from positional mapping to mutation screening of exons, with little prospect (due to the enormous technical and statistical difficulties) of identifying mutations that lie outside these limited regions in large intronic or intergenic sequences. However, some such mutations are being identified, including one in a novel ncRNA called MIAT, which appears to increase the risk for myocardial infarction (Ishii et al., 2006). I predict that as re-sequencing of target regions in affected populations

becomes feasible with new sequencing technologies, more of these mutations/variations will be discovered and that many will affect regulatory RNAs sourced from such regions. Indeed, apart from revealing the true extent of the involvement of RNA in the developmental programming of humans and other complex organisms, these discoveries will go to the heart of what is perhaps the most interesting aspect of our biology, the genetic factors controlling or influencing our individual physical, physiological and psychological variation, including disease susceptibility.

Conclusion

I suggest that we have fundamentally misunderstood the nature of genetic programming of complex organisms for the past 50 years, because of the presumption – largely true in the prokaryotes but not in the complex eukaryotes – that most genetic information is transacted by proteins. This view was derived from studying simple organisms in an analogue age before the power and use of digital information systems were appreciated. However, it now seems increasingly likely that most of the human genome, and those of other complex organisms, encodes a vast and hitherto hidden layer of regulatory RNAs (Mattick and Makunin, 2005; Mattick and Makunin, 2006). This evolved to breach the operational limits imposed by solely protein-based regulatory systems, in the face of the nonlinear scaling of regulatory requirements as living organisms explored higher organizational and macro-functional complexity (Mattick, 2004). Indeed, it may well be that most of the human genome is functional (M. Pheasant and J.S.M., manuscript submitted for publication), including many sequences such as introns and other mobile element-derived sequences that have been long considered as parasitic evolutionary debris rather than the historic raw material for genetic innovation and the current embodiment of higher levels of regulatory sophistication. Thus it appears that the genome is largely composed of sequences encoding components of RNA regulatory networks that co-evolved with a sophisticated protein infrastructure to interact with RNAs and act on their instructions.

The advantages of RNA over protein as a regulatory molecule are its genomic compactness, its high sequence specificity, and its mutability and associated ease of re-configuration of interaction networks to explore phenotypic and functional diversity. This leads to a new conception of how multicellular development is regulated and where the relevant information is embedded, i.e. that development is primarily driven by endogenous RNA regulatory networks, which are contextually informed and whose instructions are functionally executed by proteins. There is clearly a long way to go to understand and parse these networks, with many surprises yet in store, including the likely discovery of new classes and subclasses of small and large regulatory RNAs, and many biological and mechanistic aspects to decipher. Whatever the details may be, the irony is that what was dismissed as junk because it was not understood may well comprise the majority of the information that underpinned the emergence and now directs the development of complex organisms (Mattick, 1994; Mattick, 2004; Mattick and Gagen, 2001), including ultimately the brain (Mehler and Mattick, 2007). It probably also contains a large fraction of the information that determines the phenotypic differences between individuals and the diversity of species.

This article draws on, and to some extent integrates, ideas elaborated in others that I have co-authored with Michael Gagen, Igor Makunin, Mark Mehler, Michael Pheasant and Ryan Taft, which are cited in the appropriate places in the text. I am grateful to them, as well as all members of my laboratory, for many stimulating discussions and research contributions over many years. I am also grateful to collaborators, particularly Yoshihide Hayashizaki, Piero Carninci and Harukazu Suzuki from RIKEN, and other senior scientists in my own institute and elsewhere. I particularly thank Paulo Amaral, Andy Cossins, Larry Croft, Martin Feder, Michaela Handel, Ian Holmes, Igor Makunin, Michael Pheasant and Cas Simons for comments and suggestions on the manuscript. This work was supported by an Australian Research Council Federation Fellowship.

References

Abelson, J. F., Kwan, K. Y., O'Roak, B. J., Baek, D. Y., Stillman, A. A., Morgan, T. M., Mathews, C. A., Pauls, D. L., Rasin, M. R., Gunel, M. et al. (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**, 317-320.

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

Alfano, G., Vitiello, C., Caccioppoli, C., Caramico, T., Carola, A., Szego, M. J., McInnes, R. R., Auricchio, A. and Banfi, S. (2005). Natural antisense transcripts associated with genes involved in eye development. *Hum. Mol. Genet.* **14**, 913-923.

Ambros, V. (2001). The temporal control of cell cycle and cell fate in *Caenorhabditis elegans*. *Novartis Found. Symp.* **237**, 203-214; discussion 214-220.

Anantharaman, V., Koonin, E. V. and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30**, 1427-1464.

Andersen, A. A. and Panning, B. (2003). Epigenetic gene regulation by noncoding RNAs. *Curr. Opin. Cell Biol.* **15**, 281-289.

Andersen, R. A., Snyder, L. H., Bradley, D. C. and Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annu. Rev. Neurosci.* **20**, 303-330.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. et al. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-1310.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T. et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203-207.

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851-1864.

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

Athanasiadis, A., Rich, A. and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2**, e391.

Aufsatz, W., Mette, M. F., van der Winden, J., Matzke, A. J. and Matzke, M. (2002). RNA-directed DNA methylation in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **99**, 16499-16506.

Aussem, A., Murtagh, F. and Sarazin, M. (1995). Dynamical recurrent neural networks – towards environmental time series prediction. *Int. J. Neural Syst.* **6**, 145-170.

Bachellerie, J. P., Cavaille, J. and Huttenhofer, A. (2002). The expanding snoRNA world. *Biochimie* **84**, 775-790.

Bae, E., Calhoun, V. C., Levine, M., Lewis, E. B. and Drewell, R. A. (2002). Characterization of the intergenic RNA profile at abdominal-A and Abdominal-B in the *Drosophila* bithorax complex. *Proc. Natl. Acad. Sci. USA* **99**, 16847-16852.

Baehrecke, E. H. (2002). How death shapes life during development. *Nat. Rev. Mol. Cell Biol.* **3**, 779-787.

Baehrecke, E. H. (2003). miRNAs: micro managers of programmed cell death. *Curr. Biol.* **13**, R473-R475.

Bantignies, F. and Cavalli, G. (2006). Cellular memory and dynamic regulation of polycomb group proteins. *Curr. Opin. Cell Biol.* **18**, 275-283.

Barsyte-Lovejoy, D., Lau, S. K., Boutros, P. C., Khosravi, F., Jurisica, I., Andrulis, I. L., Tsao, M. S. and Penn, L. Z. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* **66**, 5330-5337.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297.

Baskerville, S. and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**, 241-247.

Bass, B. L. (2002). RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817-846.

Bayne, E. H. and Allshire, R. C. (2005). RNA-directed transcriptional gene silencing in mammals. *Trends Genet.* **21**, 370-373.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* **304**, 1321-1325.

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J. and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87-90.

Ben-Tabou de-Leon, S. and Davidson, E. H. (2006). Deciphering the underlying mechanism of specification and differentiation: the sea urchin gene regulatory network. *Sci. STKE* **2006**, pe47.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21-24.

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E. and Plasterk, R. H. (2006a). Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38**, 1375-1377.

Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S. et al. (2006b). Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* **16**, 1289-1298.

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., 3rd, Gingeras, T. R. et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169-181.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K. et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326.

Bernstein, E. and Allis, C. D. (2005). RNA meets chromatin. *Genes Dev.* **19**, 1635-1655.

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V. and Hannon, G. J. (2003). Dicer is essential for mouse development. *Nat. Genet.* **35**, 215-217.

Bernstein, E., Duncan, E. M., Masui, O., Gil, J., Heard, E. and Allis, C. D. (2006). Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. *Mol. Cell. Biol.* **26**, 2560-2569.

Birchler, J. A., Kavi, H. H. and Fernandez, H. R. (2004). Heterochromatin: RNA points the way. *Curr. Biol.* **14**, R759-R761.

Blencowe, B. J. (2006). Alternative splicing: new insights from global analyses. *Cell* **126**, 37-47.

Blin-Wakkach, C., Lezot, F., Ghoul-Mazgar, S., Hotton, D., Monteiro, S., Teillaud, C., Pibouin, L., Orestes-Cardoso, S., Papagerakis, P., Macdougall, M. et al. (2001). Endogenous Msx1 antisense transcript: in vivo and in vitro evidences, structure, and potential involvement in skeleton development in mammals. *Proc. Natl. Acad. Sci. USA* **98**, 7336-7341.

Blow, M., Futreal, P. A., Wooster, R. and Stratton, M. R. (2004). A survey of RNA editing in human brain. *Genome Res.* **14**, 2379-2387.

Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R. and Stratton, M. R. (2006). RNA editing of human microRNAs. *Genome Biol.* **7**, R27.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K. et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353.

Brandt, J., Schrauth, S., Veith, A. M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C. and Volff, J. N. (2005). Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* **345**, 101-111.

Brannan, C. I., Dees, E. C., Ingram, R. S. and Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.* **10**, 28-36.

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. and Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell* **113**, 25-36.

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

Britten, R. (2006). Transposable elements have contributed to thousands of human proteins. *Proc. Natl. Acad. Sci. USA* **103**, 1798-1803.

Britten, R. J. and Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science* **165**, 349-357.

Brockdorff, N. (1998). The role of Xist in X-inactivation. *Curr. Opin. Genet. Dev.* **8**, 328-333.

Brosius, J. (1999). RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238**, 115-134.

Brosius, J. (2005). Disparity, adaptation, exaptation, bookkeeping, and contingency at the genome level. *Paleobiology* **31**, 1-16.

Buchler, N. E., Gerland, U. and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proc. Natl. Acad. Sci. USA* **100**, 5136-5141.

Bussemakers, M. J., van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F., Schalken, J. A., Debruyne, F. M., Ru, N. and Isaacs, W. B. (1999). DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **59**, 5975-5979.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D. et al. (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157.

Caceres, J. F. and Kornblihtt, A. R. (2002). Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186-193.

Cai, X., Hagedorn, C. H. and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**, 1957-1966.

Canfield, D. E., Poulton, S. W. and Narbonne, G. M. (2007). Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. *Science* **315**, 92-95.

Caretti, G., Schiltz, R. L., Dilworth, F. J., Di Padova, M., Zhao, P., Ogryzko, V., Fuller-Pace, F. V., Hoffman, E. P., Tapscott, S. J. and Sartorelli, V. (2006). The RNA helicases p68/p72 and the noncoding RNA SRA are coregulators of MyoD and skeletal muscle differentiation. *Dev. Cell* **11**, 547-560.

Carmell, M. A., Xuan, Z., Zhang, M. Q. and Hannon, G. J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.* **16**, 2733-2742.

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

Carrington, J. C. and Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science* **301**, 336-338.

Cavaille, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C. I., Horsthemke, B., Bachellerie, J. P., Brosius, J. and Huttenhofer, A. (2000). Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl. Acad. Sci. USA* **97**, 14311-14316.

Cavaille, J., Vitali, P., Basyuk, E., Huttenhofer, A. and Bachellerie, J. P. (2001). A novel brain-specific box C/D small nucleolar RNA processed from

tandemly repeated introns of a noncoding RNA gene in rats. *J. Biol. Chem.* **276**, 26374-26383.

Cavaille, J., Seitz, H., Paulsen, M., Ferguson-Smith, A. C. and Bachellerie, J. P. (2002). Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum. Mol. Genet.* **11**, 1527-1538.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. *Trends Genet.* **7**, 145-148.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J. et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509.

Cernilogar, F. M. and Orlando, V. (2005). Epigenome programming by Polycomb and Trithorax proteins. *Biochem. Cell Biol.* **83**, 322-331.

Chalk, A. M., Warfinge, R. E., Georgii-Hemming, P. and Sonnhammer, E. L. (2005). siRNAdb: a database of siRNA sequences. *Nucleic Acids Res.* **33**, D131-D134.

Chen, C. Z., Li, L., Lodish, H. F. and Bartel, D. P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303**, 83-86.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science* **303**, 2022-2025.

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

Cho, Y. S., Iguchi, N., Yang, J., Handel, M. A. and Hecht, N. B. (2005). Meiotic messenger RNA and noncoding RNA targets of the RNA-binding protein Translin (TSN) in mouse testis. *Biol. Reprod.* **73**, 840-847.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P. and Duret, L. (2002). Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.* **12**, 894-908.

Clark, R. M., Wagler, T. N., Quijada, P. and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* **38**, 594-597.

Clark, S. W., Fee, B. E. and Cleveland, J. L. (2002). Misexpression of the eyes absent family triggers the apoptotic program. *J. Biol. Chem.* **277**, 3560-3567.

Clarke, J. D. and Tickle, C. (1999). Fate maps old and new. *Nat. Cell Biol.* **1**, E103-E109.

Clement, J. Q., Qian, L., Kaplinsky, N. and Wilkinson, M. F. (1999). The stability and fate of a spliced intron from vertebrate cells. *RNA* **5**, 206-220.

Clement, J. Q., Maiti, S. and Wilkinson, M. F. (2001). Localization and stability of introns spliced from the Pem homeobox gene. *J. Biol. Chem.* **276**, 16919-16930.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J. M., Eychenne, F. et al. (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.* **38**, 813-818.

Cocquet, J., Pannetier, M., Fellous, M. and Veitia, R. A. (2005). Sense and antisense Foxl2 transcripts in mouse. *Genomics* **85**, 531-541.

Cogoni, C. and Macino, G. (2000). Post-transcriptional gene silencing across kingdoms. *Curr. Opin. Genet. Dev.* **10**, 638-643.

Collen, M. F. (1994). The origins of informatics. *J. Am. Med. Inform. Assoc.* **1**, 91-107.

Cordaux, R., Udit, S., Batzer, M. A. and Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. USA* **103**, 8101-8106.

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

Crick, F. (1979). Split genes and RNA splicing. *Science* **204**, 264-271.

Croce, C. M. and Calin, G. A. (2005). miRNAs, cancer, and stem cell division. *Cell* **122**, 6-7.

Croft, L. J., Lercher, M. J., Gagen, M. J. and Mattick, J. S. (2003). Is prokaryotic complexity limited by accelerated growth in regulatory overhead? *Genome Biology Preprint Depository* http://genomebiology.com/qc/2003/5/1/p2.

Csete, M. E. and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science* **295**, 1664-1669.

Cummins, J. M., He, Y., Leary, R. J., Pagliarini, R., Diaz, L. A., Jr, Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A. E., Labourier, E. et al. (2006). The colorectal microRNAome. *Proc. Natl. Acad. Sci. USA* **103**, 3687-3692.

Dagan, T., Sorek, R., Sharon, E., Ast, G. and Graur, D. (2004). AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res.* **32**, D489-D492.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W. and Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. *Proc. Natl. Acad. Sci. USA* **103**, 5320-5325.

Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. New York: Academic Press.

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.

Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C. and Antonarakis, S. E. (2003). Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033-1035.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601-603.

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

Duncan, I. W. (2002). Transvection effects in *Drosophila*. *Annu. Rev. Genet.* **36**, 521-556.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. and Avner, P. (2006). The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653-1655.

Edelman, G. M. (1993). Neural Darwinism: selection and reentrant signalling in higher brain function. *Neuron* **10**, 115-125.

Elman, J. L. (1998). Connectionism, artificial life, and dynamical systems: new approaches to old questions. In *A Companion to Cognitive Science* (ed. W. Bechtel and G. Graham), pp. 488-505. Oxford: Basil Blackwood.

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

Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs – microRNAs with a role in cancer. *Nat. Rev. Cancer* **6**, 259-269.

Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P. and Kohtz, J. D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20**, 1470-1484.

Ferat, J. L. and Michel, F. (1993). Group II self-splicing introns in bacteria. *Nature* **364**, 358-361.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J. P., White, R. J. and Aberdam, D. (2001). Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* **28**, 77-81.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E. and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806-811.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. and McCallion, A. S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276-279.

Frith, M. C., Pheasant, M. and Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet.* **13**, 894-897.

Frith, M. C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayshizaki, Y. and Sandelin, A. (2006). Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16**, 713-722.

Gagen, M. J. and Mattick, J. S. (2004). Inherent size constraints on prokaryote gene networks due to 'accelerating' growth. *arXiv Preprint Archive* http://arXiv.org/abs/q-bio.MN/0312021.

Gagen, M. J. and Mattick, J. S. (2005). Inherent size constraints on prokaryote gene networks due to 'accelerating' growth. *Theory Biosci.* **123**, 381-411.

Garcia-Blanco, M. A. (2005). Making antisense of splicing. *Curr. Opin. Mol. Ther.* **7**, 476-482.

Gendron, D., Carriero, S., Garneau, D., Villemaire, J., Klinck, R., Elela, S. A., Damha, M. J. and Chabot, B. (2006). Modulation of 5′ splice site selection using tailed oligonucleotides carrying splicing signals. *BMC Biotechnol.* **6**, 5.

Gesteland, R. F., Cech, T. R. and Atkins, J. F. (2006). *The RNA World*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Gilbert, W. (1978). Why genes in pieces? *Nature* **271**, 501.

Gilbert, W. and Muller-Hill, B. (1966). Isolation of the *lac* repressor. *Proc. Natl. Acad. Sci. USA* **56**, 1891-1898.

Ginger, M. R., Shore, A. N., Contreras, A., Rijnkels, M., Miller, J., Gonzalez-Rimbau, M. F. and Rosen, J. M. (2006). A noncoding RNA is

a potential marker of cell fate during mammary gland development. *Proc. Natl. Acad. Sci. USA* **103**, 5781-5786.

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., Hammond, S. M., Bartel, D. P. and Schier, A. F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833-838.

Girard, A., Sachidanandam, R., Hannon, G. J. and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199-202.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M. et al. (1996). Life with 6000 genes. *Science* **274**, 563-567.

Goodrich, J. A. and Kugel, J. F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* **7**, 612-616.

Goodstadt, L. and Ponting, C. P. (2006). Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* **2**, e133.

Gordon, M. D. and Nusse, R. (2006). Wnt signalling: multiple pathways, multiple receptors, and multiple transcription factors. *J. Biol. Chem.* **281**, 22429-22433.

Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.* **21**, 399-404.

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

Gray, H. (1918). *Anatomy of the Human Body*. Philadelphia: Lea & Febiger.

Grimaud, C., Bantignies, F., Pal-Bhadra, M., Ghana, P., Bhadra, U. and Cavalli, G. (2006). RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* **124**, 957-971.

Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G. and Mello, C. C. (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.

Guenther, M. G., Jenner, R. G., Chevalier, B., Nakamura, T., Croce, C. M., Canaani, E. and Young, R. A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. *Proc. Natl. Acad. Sci. USA* **102**, 8603-8608.

Hammond, S. M. (2006). MicroRNAs as oncogenes. *Curr. Opin. Genet. Dev.* **16**, 4-9.

Hammond, S. M., Bernstein, E., Beach, D. and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293-296.

Hare, M. P. and Palumbi, S. R. (2003). High intron sequence conservation across three mammalian orders suggests functional constraints. *Mol. Biol. Evol.* **20**, 969-978.

Hasler, J. and Strub, K. (2006). Alu elements as regulators of gene expression. *Nucleic Acids Res.* **34**, 5491-5497.

Hatfield, S. D., Shcherbata, H. R., Fischer, K. A., Nakahara, K., Carthew, R. W. and Ruohola-Baker, H. (2005). Stem cell division is regulated by the microRNA pathway. *Nature* **435**, 974-978.

Henras, A. K., Dez, C. and Henry, Y. (2004). RNA structure and function in C/D and H/ACA s(no)RNPs. *Curr. Opin. Struct. Biol.* **14**, 335-343.

Herbert, A. and Rich, A. (1999a). RNA processing and the evolution of eukaryotes. *Nat. Genet.* **21**, 265-269.

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

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I. L. and Stadler, P. F. (2006). The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**, 25.

Hill, A. E., Hong, J. S., Wen, H., Teng, L., McPherson, D. T., McPherson, S. A., Levasseur, D. N. and Sorscher, E. J. (2006). Micro-RNA-like effects of complete intronic sequences. *Front. Biosci.* **11**, 1998-2006.

Holmes, R., Williamson, C., Peters, J., Denny, P. and Wells, C. (2003). A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Res.* **13**, 1410-1415.

Hornstein, E., Mansfield, J. H., Yekta, S., Hu, J. K., Harfe, B. D., McManus, M. T., Baskerville, S., Bartel, D. P. and Tabin, C. J. (2005). The microRNA miR-196 acts upstream of Hoxb8 and Shh in limb development. *Nature* **438**, 671-674.

Huang, Z. P., Zhou, H., Liang, D. and Qu, L. H. (2004). Different expression strategy: multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. *J. Mol. Biol.* **341**, 669-683.

Hube, F., Guo, J., Chooniedass-Kothari, S., Cooper, C., Hamedani, M. K., Dibrov, A. A., Blanchard, A. A., Wang, X., Deng, G., Myal, Y. et al. (2006). Alternative splicing of the first intron of the steroid receptor RNA activator (SRA) participates in the generation of coding and noncoding RNA isoforms in breast cancer cell lines. *DNA Cell Biol.* **25**, 418-428.

Huttenhofer, 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.

Imamura, T., Yamamoto, S., Ohgane, J., Hattori, N., Tanaka, S. and Shiota, K. (2004). Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochem. Biophys. Res. Commun.* **322**, 593-600.

International Human Genome Sequencing Consortium (2004a). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945.

International Human Genome Sequencing Consortium (2004b). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716.

Irvine, D. V., Zaratiegui, M., Tolia, N. H., Goto, D. B., Chitwood, D. H., Vaughn, M. W., Joshua-Tor, L. and Martienssen, R. A. (2006). Argonaute slicing is required for heterochromatic silencing and spreading. *Science* **313**, 1134-1137.

Irvine, K., Stirling, R., Hume, D. and Kennedy, D. (2004). Rasputin, more promiscuous than ever: a review of G3BP. *Int. J. Dev. Biol.* **48**, 1065-1077.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M. et al. (2006). Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *J. Hum. Genet.* **51**, 1087-1099.

Jacob, F. (1977). Evolution and tinkering. *Science* **196**, 1161-1166.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318-356.

Jady, B. E., Darzacq, X., Tucker, K. E., Matera, A. G., Bertrand, E. and Kiss, T. (2003). Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm. *EMBO J.* **22**, 1878-1888.

Jady, B. E., Bertrand, E. and Kiss, T. (2004). Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body-specific localization signal. *J. Cell Biol.* **164**, 647-652.

Janowski, B. A., Huffman, K. E., Schwartz, J. C., Ram, R., Hardy, D., Shames, D. S., Minna, J. D. and Corey, D. R. (2005). Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. *Nat. Chem. Biol.* **1**, 216-222.

Jeffery, L. and Nakielny, S. (2004). Components of the DNA methylation system of chromatin control are RNA-binding proteins. *J. Biol. Chem.* **279**, 49479-49487.

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., Bulk, E. et al. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 6087-6097.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S. (2004). Human microRNA targets. *PLoS Biol.* **2**, e363.

Johnston, R. J. and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845-849.

Jones, E. A. and Flavell, R. A. (2005). Distal enhancer elements transcribe intergenic RNA in the IL-10 family gene cluster. *J. Immunol.* **175**, 7437-7446.

Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* **14**, 787-799.

Kamal, M., Xie, X. and Lander, E. S. (2006). A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci. USA* **103**, 2740-2745.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G. et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331-342.

Kanellopoulou, C., Muljo, S. A., Kung, A. L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D. M. and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19**, 489-501.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916-919.

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

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

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

Kim, D. H., Villeneuve, L. M., Morris, K. V. and Rossi, J. J. (2006). Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nat. Struct. Mol. Biol.* **13**, 793-797.

Kishore, S. and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311**, 230-232.

Kiss, A. M., Jady, B. E., Bertrand, E. and Kiss, T. (2004). Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol.* **24**, 5797-5807.

Kleinjan, D. A. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8-32.

Kole, R. and Sazani, P. (2001). Antisense effects in the cell nucleus: modification of splicing. *Curr. Opin. Mol. Ther.* **3**, 229-234.

Korneev, S. and O'Shea, M. (2005). Natural antisense RNAs in the nervous system. *Rev. Neurosci.* **16**, 213-222.

Krajewski, W. A., Nakamura, T., Mazo, A. and Canaani, E. (2005). A motif within SET-domain proteins binds single-stranded nucleic acids and transcribed and supercoiled DNAs and can interfere with assembly of nucleosomes. *Mol. Cell. Biol.* **25**, 1891-1899.

Krause, M. O. (1996). Chromatin structure and function: the heretical path to an RNA transcription factor. *Biochem. Cell Biol.* **74**, 623-632.

Krull, M., Brosius, J. and Schmitz, J. (2005). Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* **22**, 1702-1711.

Kuwabara, T., Hsieh, J., Nakashima, K., Taira, K. and Gage, F. H. (2004). A small modulatory dsRNA specifies the fate of adult neural stem cells. *Cell* **116**, 779-793.

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

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858.

Lam, A. L., Pazin, D. E. and Sullivan, B. A. (2005). Control of gene expression and assembly of chromosomal subdomains by chromatin regulators with antagonistic functions. *Chromosoma* **114**, 242-251.

Lambowitz, A. M. and Zimmerly, S. (2004). Mobile group II introns. *Annu. Rev. Genet.* **38**, 1-35.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Landry, J. R., Medstrand, P. and Mager, D. L. (2001). Repetitive elements in the 5′ untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* **76**, 110-116.

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.

Lanz, R. B., Razani, B., Goldberg, A. D. and O'Malley, B. W. (2002). Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). *Proc. Natl. Acad. Sci. USA* **99**, 16081-16086.

Lau, N. C., Lim, L. P., Weinstein, E. G. and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858-862.

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P. and Kingston, R. E. (2006). Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367.

Lee, J. T., Davidow, L. S. and Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.* **21**, 400-404.

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294**, 862-864.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854.

Lei, E. P. and Corces, V. G. (2006a). A long-distance relationship between RNAi and Polycomb. *Cell* **124**, 886-888.

Lei, E. P. and Corces, V. G. (2006b). RNA interference machinery influences the nuclear organization of a chromatin insulator. *Nat. Genet.* **38**, 936-941.

Lemons, D. and McGinnis, W. (2006). Genomic evolution of Hox gene clusters. *Science* **313**, 1918-1922.

Lescoute, A., Leontis, N. B., Massire, C. and Westhof, E. (2005). Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res.* **33**, 2395-2409.

Lev-Maor, G., Sorek, R., Shomron, N. and Ast, G. (2003). The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* **300**, 1288-1291.

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D. et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22**, 1001-1005.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* **102**, 4936-4942.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* **424**, 147-151.

Lewis, B. P., Burge, C. B. and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20.

Li, L. C., Okino, S. T., Zhao, H., Pookot, D., Place, R. F., Urakami, S., Enokida, H. and Dahiya, R. (2006). Small dsRNAs induce transcriptional activation in human cells. *Proc. Natl. Acad. Sci. USA* **103**, 17337-17342.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. and Johnson, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-773.

Lindow, M. and Krogh, A. (2005). Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics* **6**, 119.

Lippman, Z. and Martienssen, R. (2004). The role of RNA interference in heterochromatic silencing. *Nature* **431**, 364-370.

Lippman, Z., Gendrel, A. V., Black, M., Vaughn, M. W., Dedhia, N., McCombie, W. R., Lavine, K., Mittal, V., May, B., Kasschau, K. D. et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471-476.

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.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A. et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-838.

Lund, A. H. and van Lohuizen, M. (2004). Polycomb complexes and silencing mechanisms. *Curr. Opin. Cell Biol.* **16**, 239-246.

Lunter, G., Ponting, C. P. and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* **2**, e5.

Maison, C., Bailly, D., Peters, A. H., Quivy, J. P., Roche, D., Taddei, A., Lachner, M., Jenuwein, T. and Almouzni, G. (2002). Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat. Genet.* **30**, 329-334.

Manak, J. R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A. et al. (2006). Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat. Genet.* **38**, 1151-1158.

Mansfield, J. H., Harfe, B. D., Nissen, R., Obenauer, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A. E., Ruvkun, G. et al. (2004). MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat. Genet.* **36**, 1079-1083.

Margueron, R., Trojer, P. and Reinberg, D. (2005). The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* **15**, 163-176.

Martienssen, R. A., Zaratiegui, M. and Goto, D. B. (2005). RNA interference and heterochromatin in the fission yeast *Schizosaccharomyces pombe*. *Trends Genet.* **21**, 450-456.

Martinez-Abarca, F. and Toro, N. (2000). Group II introns in the bacterial world. *Mol. Microbiol.* **38**, 917-926.

Matlik, K., Redik, K. and Speek, M. (2006). L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.* **2006**, 71753.

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

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986-991.
Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* **25**, 930-939.
Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat. Rev. Genet.* **5**, 316-323.
Mattick, J. S. (2005). The functional genomics of noncoding RNA. *Science* **309**, 1527-1528.
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.
Mattick, J. S. and Gagen, M. J. (2005). Accelerating networks. *Science* **307**, 856-858.
Mattick, J. S. and Makunin, I. V. (2005). Small regulatory RNAs in mammals. *Hum. Mol. Genet.* **14**, R121-R132.
Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* **15**, R17-R29.
Matzke, M., Matzke, A. J. and Kooter, J. M. (2001). RNA: guiding gene silencing. *Science* **293**, 1080-1083.
Maurer-Stroh, S., Dickens, N. J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F. and Ponting, C. P. (2003). The Tudor domain 'Royal Family': tudor, plant Agenet, chromo, PWWP and MBT domains. *Trends Biochem. Sci.* **28**, 69-74.
McCarthy, J. V. (2003). Apoptosis and development. *Essays Biochem.* **39**, 11-24.
Mehler, M. F. and Mattick, J. S. (2007). Non-protein-coding RNAs and RNA editing in brain development, functional diversification and neurological disease. *Physiol. Rev*. **87** (in press).
Meier, U. T. (2005). The many facets of H/ACA ribonucleoproteins. *Chromosoma* **114**, 1-14.
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.
Migeon, B. R., Chowdhury, A. K., Dunston, J. A. and McIntosh, I. (2001). Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *Am. J. Hum. Genet.* **69**, 951-960.
Mochizuki, K. and Gorovsky, M. A. (2004). Small RNAs in genome rearrangement in Tetrahymena. *Curr. Opin. Genet. Dev.* **14**, 181-187.
Morison, I. M., Ramsay, J. P. and Spencer, H. G. (2005). A census of mammalian imprinting. *Trends Genet.* **21**, 457-465.
Morris, K. V., Chan, S. W., Jacobsen, S. E. and Looney, D. J. (2004). Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305**, 1289-1292.
Murone, M., Rosenthal, A. and de Sauvage, F. J. (1999). Hedgehog signal transduction: from flies to vertebrates. *Exp. Cell Res.* **253**, 25-33.
Mutsuddi, M., Marshall, C. M., Benzow, K. A., Koob, M. D. and Rebay, I. (2004). The spinocerebellar ataxia 8 noncoding RNA causes neurodegeneration and associates with staufen in Drosophila. *Curr. Biol.* **14**, 302-308.
Naguibneva, I., Ameyar-Zazoua, M., Polesskaya, A., Ait-Si-Ali, S., Groisman, R., Souidi, M., Cuvellier, S. and Harel-Bellan, A. (2006). The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation. *Nat. Cell Biol.* **8**, 278-284.
Navaratnam, N. and Sarwar, R. (2006). An overview of cytidine deaminases. *Int. J. Hematol.* **83**, 195-200.
Negre, N., Hennetin, J., Sun, L. V., Lavrov, S., Bellis, M., White, K. P. and Cavalli, G. (2006). Chromosomal distribution of PcG proteins during Drosophila development. *PLoS Biol*. **4**, e170.
Nesterova, T. B., Slobodyanyuk, S. Y., Elisaphenko, E. A., Shevchenko, A. I., Johnston, C., Pavlova, M. E., Rogozin, I. B., Kolesnikov, N. N., Brockdorff, N. and Zakian, S. M. (2001). Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11**, 833-849.
Nigumann, P., Redik, K., Matlik, K. and Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628-634.
Nikaido, I., Saito, C., Mizuno, Y., Meguro, M., Bono, H., Kadomura, M., Kono, T., Morris, G. A., Lyons, P. A., Oshimura, M. et al. (2003). Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling. *Genome Res.* **13**, 1402-1409.
Nishihara, H., Smit, A. F. and Okada, N. (2006). Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res*. **16**, 864-874.
Ogiwara, I., Miya, M., Ohshima, K. and Okada, N. (2002). V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. *Genome Res.* **12**, 316-324.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573.
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.
Orgel, L. E. and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature* **284**, 604-607.
Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W. and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137-145.
Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. and Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55**, 1119-1150.
Pal-Bhadra, M., Leibovitch, B. A., Gandhi, S. G., Rao, M., Bhadra, U., Birchler, J. A. and Elgin, S. C. (2004). Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. *Science* **303**, 669-672.
Palmer, J. D. and Logsdon, J. M., Jr (1991). The recent origins of introns. *Curr. Opin. Genet. Dev.* **1**, 470-477.
Pandorf, C. E., Haddad, F., Roy, R. R., Qin, A. X., Edgerton, V. R. and Baldwin, K. M. (2006). Dynamics of myosin heavy chain gene regulation in slow skeletal muscle: role of natural antisense RNA. *J. Biol. Chem.* **281**, 38330-38342.
Pang, K. C., Stephen, S., Engstrom, P. G., Tajul-Arifin, K., Chen, W., Wahlestedt, C., Lenhard, B., Hayashizaki, Y. and Mattick, J. S. (2005). RNAdb – a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res.* **33**, D125-D130.
Pang, K. C., Frith, M. C. and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1-5.
Pasquinelli, A. E., Hunter, S. and Bracht, J. (2005). MicroRNAs: a developing story. *Curr. Opin. Genet. Dev.* **15**, 200-205.
Pazman, C., Mayes, C. A., Fanto, M., Haynes, S. R. and Mlodzik, M. (2000). Rasputin, the Drosophila homologue of the RasGAP SH3 binding protein, functions in ras- and Rho-mediated signalling. *Development* **127**, 1715-1725.
Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D. and Knowles, B. B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597-606.
Peterson, C. L. and Laniel, M. A. (2004). Histones and histone modifications. *Curr. Biol.* **14**, R546-R551.
Petruk, S., Sedkov, Y., Riley, K. M., Hodgson, J., Schweisguth, F., Hirose, S., Jaynes, J. B., Brock, H. W. and Mazo, A. (2006). Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. *Cell* **127**, 1209-1221.
Plunkett, K., Karmiloff-Smith, A., Bates, E., Elman, J. L. and Johnson, M. H. (1997). Connectionism and developmental psychology. *J. Child Psychol. Psychiatry* **38**, 53-80.
Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M. A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A. et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167-172.
Pozzoli, U. and Sironi, M. (2005). Silencers regulate both constitutive and alternative splicing events in mammals. *Cell. Mol. Life Sci.* **62**, 1579-1604.
Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., Zhang, M. Q. and Spector, D. L. (2005). Regulating gene expression through RNA nuclear retention. *Cell* **123**, 249-263.
Prochnik, S. E., Rokhsar, D. S. and Aboobaker, A. A. (2007). Evidence for a microRNA expansion in the bilaterian ancestor. *Dev. Genes Evol.* **217**, 73-77.
Qi, Y., He, X., Wang, X. J., Kohany, O., Jurka, J. and Hannon, G. J. (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature* **443**, 1008-1012.
Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M. et al. (2006).

Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16**, 11-19.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901-906.

Reis, E. M., Nakaya, H. I., Louro, R., Canavez, F. C., Flatschart, A. V., Almeida, G. T., Egidio, C. M., Paquola, A. C., Machado, A. A., Festa, F. et al. (2004). Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* **23**, 6684-6692.

Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A. and Platt, D. (2006). Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc. Natl. Acad. Sci. USA* **103**, 6605-6610.

Ringrose, L. and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38**, 413-443.

Ringrose, L. and Paro, R. (2007). Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development* **134**, 223-232.

Roberts, J., Palma, E., Sazani, P., Orum, H., Cho, M. and Kole, R. (2006). Efficient and persistent splice switching by systemically delivered LNA oligonucleotides in mice. *Mol. Ther.* **14**, 471-475.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14**, 1902-1910.

Rogelj, B. and Giese, K. P. (2004). Expression and function of brain specific small RNAs. *Rev. Neurosci.* **15**, 185-198.

Ronshaugen, M. and Levine, M. (2004). Visualization of trans-homolog enhancer-promoter interactions at the Abd-B Hox locus in the Drosophila embryo. *Dev. Cell* **7**, 925-932.

Ronshaugen, M., Biemar, F., Piel, J., Levine, M. and Lai, E. C. (2005). The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes Dev.* **19**, 2947-2952.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**, 1193-1207.

Runte, M., Huttenhofer, A., Gross, S., Kiefmann, M., Horsthemke, B. and Buiting, K. (2001). The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genet.* **10**, 2687-2700.

Ruvkun, G. (2001). Glimpses of a tiny RNA world. *Science* **294**, 797-799.

Salehi-Ashtiani, K., Luptak, A., Litovchick, A. and Szostak, J. W. (2006). A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* **313**, 1788-1792.

Sanchez-Elsner, T., Gou, D., Kremmer, E. and Sauer, F. (2006). Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. *Science* **311**, 1118-1123.

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

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F. and Stupka, E. (2006). Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* **7**, R56.

Saunders, L. R. and Barber, G. N. (2003). The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J.* **17**, 961-983.

Schmitt, S. and Paro, R. (2006). RNA at the steering wheel. *Genome Biol.* **7**, 218.

Schmitt, S., Prestel, M. and Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. *Genes Dev.* **19**, 697-708.

Schwartz, Y. B. and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. *Nat. Rev. Genet.* **8**, 9-22.

Schwartz, Y. B., Kahn, T. G., Nix, D. A., Li, X. Y., Bourgon, R., Biggin, M. and Pirrotta, V. (2006). Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*. *Nat. Genet.* **38**, 700-705.

Seitz, H., Royo, H., Bortolin, M. L., Lin, S. P., Ferguson-Smith, A. C. and Cavaille, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res.* **14**, 1741-1748.

Sessa, L., Breiling, A., Lavorgna, G., Silvestri, L., Casari, G. and Orlando, V. (2007). Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster. *RNA* **13**, 223-239.

Shamovsky, I., Ivannikov, M., Kandel, E. S., Gershon, D. and Nudler, E. (2006). RNA-mediated response to heat shock in mammalian cells. *Nature* **440**, 556-560.

Shanab, G. M. and Maxwell, E. S. (1992). Determination of the nucleotide sequences in mouse U14 small nuclear RNA and 18S ribosomal RNA responsible for in vitro intermolecular base-pairing. *Eur. J. Biochem.* **206**, 391-400.

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

Silva, J. C., Shabalina, S. A., Harris, D. G., Spouge, J. L. and Kondrashovi, A. S. (2003). Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82**, 1-18.

Simons, C., Pheasant, M., Makunin, I. V. and Mattick, J. S. (2006). Transposon-free regions in mammalian genomes. *Genome Res.* **16**, 164-172.

Sironi, M., Menozzi, G., Comi, G. P., Cagliani, R., Bresolin, N. and Pozzoli, U. (2005). Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet.* **14**, 2533-2546.

Sleutels, F., Zwart, R. and Barlow, D. P. (2002). The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810-813.

Smalheiser, N. R. and Torvik, V. I. (2005). Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21**, 322-326.

Smalheiser, N. R. and Torvik, V. I. (2006). Alu elements within human mRNAs are probable microRNA targets. *Trends Genet.* **22**, 532-536.

Smit, A. F. and Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23**, 98-102.

Smit, M., Segers, K., Carrascosa, L. G., Shay, T., Baraldi, F., Gyapay, G., Snowder, G., Georges, M., Cockett, N. and Charlier, C. (2003). Mosaicism of Solid Gold supports the causality of a noncoding A-to-G transition in the determinism of the callipyge phenotype. *Genetics* **163**, 453-456.

Smith, C. W. and Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381-388.

Smith, N. G., Brandstrom, M. and Ellegren, H. (2004). Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84**, 806-813.

Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C., Angerer, L. M., Arnone, M. I., Burgess, D. R., Burke, R. D. et al. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941-952.

Soller, M. (2006). Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell. Mol. Life Sci.* **63**, 796-819.

Sonkoly, E., Bata-Csorgo, Z., Pivarcsi, A., Polyanka, H., Kenderessy-Szabo, A., Molnar, G., Szentpali, K., Bari, L., Megyeri, K., Mandi, Y. et al. (2005). Identification and characterization of a novel, psoriasis susceptibility-related noncoding RNA gene, PRINS. *J. Biol. Chem.* **280**, 24159-24167.

Sorek, R. and Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631-1637.

Squazzo, S. L., O'Geen, H., Komashko, V. M., Krig, S. R., Jin, V. X., Jang, S. W., Margueron, R., Reinberg, D., Green, R. and Farnham, P. J. (2006). Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* **16**, 890-900.

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A. and Soreq, H. (2005). Function of alternative splicing. *Gene* **344**, 1-20.

Stathopoulos, A. and Levine, M. (2005). Genomic regulatory networks and animal development. *Dev. Cell* **9**, 449-462.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A. et al. (2003). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol.* **1**, E45.

Sternberg, P. W. and Felix, M. A. (1997). Evolution of cell lineage. *Curr. Opin. Genet. Dev.* **7**, 543-550.

Sudharsanan, S. I. and Sundareshan, M. K. (1994). Supervised training of dynamical neural networks for associative memory design and identification of nonlinear maps. *Int. J. Neural Syst.* **5**, 165-180.

Sugnet, C. W., Kent, W. J., Ares, M., Jr and Haussler, D. (2004). Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* **2004**, 66-77.

Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams, A., Kulp, D., Blume, J. E., Haussler, D. et al. (2006).

Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol.* **2**, e4.

Sun, H., Skogerbo, G. and Chen, R. (2006). Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet.* **15**, 2911-2922.

Taft, R. J., Pheasant, M. and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* **29**, 288-299.

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.

Tang, G. (2005). siRNA and miRNA: an insight into RISCs. *Trends Biochem. Sci.* **30**, 106-114.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Semple, C. A. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2**, e30.

Tillib, S., Petruk, S., Sedkov, Y., Kuzin, A., Fujioka, M., Goto, T. and Mazo, A. (1999). Trithorax- and Polycomb-group response elements within an Ultrabithorax transcription maintenance unit consist of closely situated but separable sequences. *Mol. Cell. Biol.* **19**, 5189-5202.

Ting, A. H., Schuebel, K. E., Herman, J. G. and Baylin, S. B. (2005). Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. *Nat. Genet.* **37**, 906-910.

Truss, M., Swat, M., Kielbasa, S. M., Schafer, R., Herzel, H. and Hagemeier, C. (2005). HuSiDa–the human siRNA database: an open-access database for published functional siRNA sequences and technical details of efficient transfer into recipient cells. *Nucleic Acids Res.* **33**, D108-D111.

Tufarelli, C., Stanley, J. A., Garrick, D., Sharpe, J. A., Ayyub, H., Wood, W. G. and Higgs, D. R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* **34**, 157-165.

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.

Valente, L. and Nishikura, K. (2005). ADAR gene family and A-to-I RNA editing: diverse roles in posttranscriptional gene regulation. *Prog. Nucleic Acid Res. Mol. Biol.* **79**, 299-338.

Van Laere, A. S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A. L., Haley, C. S., Buys, N., Tally, M. et al. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832-836.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet.* **19**, 479-484.

Velleca, M. A., Wallace, M. C. and Merlie, J. P. (1994). A novel synapse-associated noncoding RNA. *Mol. Cell. Biol.* **14**, 7095-7104.

Verdel, A. and Moazed, D. (2005). RNAi-directed assembly of heterochromatin in fission yeast. *FEBS Lett.* **579**, 5872-5878.

Vitali, P., Royo, H., Seitz, H., Bachellerie, J. P., Huttenhofer, A. and Cavaille, J. (2003). Identification of 13 novel human modification guide RNAs. *Nucleic Acids Res.* **31**, 6543-6551.

Vogel, J. and Sharma, C. M. (2005). How to find small non-coding RNAs in bacteria. *Biol. Chem.* **386**, 1219-1238.

Volff, J. N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* **28**, 913-922.

Wagner, G. P., Fried, C., Prohaska, S. J. and Stadler, P. F. (2004). Divergence of conserved non-coding sequences: rate estimates and relative rate tests. *Mol. Biol. Evol.* **21**, 2116-2121.

Wang, X., McLachlan, J., Zamore, P. D. and Hall, T. M. (2002). Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110**, 501-512.

Wang, Y., Davies, K. J., Melendez, J. A. and Crawford, D. R. (2003). Characterization of adapt33, a stress-inducible riboregulator. *Gene Expr.* **11**, 85-94.

Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A. and Stadler, P. F. (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23**, 1383-1390.

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

Watanabe, T., Miyashita, K., Saito, T. T., Yoneki, T., Kakihara, Y., Nabeshima, K., Kishi, Y. A., Shimoda, C. and Nojima, H. (2001). Comprehensive isolation of meiosis-specific genes identifies novel proteins and unusual non-coding transcripts in *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **29**, 2327-2337.

Watanabe, T., Miyashita, K., Saito, T. T., Nabeshima, K. and Nojima, H. (2002). Abundant poly(A)-bearing RNAs that lack open reading frames in *Schizosaccharomyces pombe*. *DNA Res.* **9**, 209-215.

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

Weinstein, S. M. and Keim, A. (1965). *Fundamentals of Digital Computers*. New York: Holt, Rinehart and Winston.

Weiss, A., Keshet, I., Razin, A. and Cedar, H. (1996). DNA demethylation in vitro: involvement of RNA. *Cell* **86**, 709-718.

Werner, A. (2005). Natural antisense transcripts. *RNA Biol.* **2**, 53-62.

Werner, A. and Berdal, A. (2005). Natural antisense transcripts: sound or silence? *Physiol. Genomics* **23**, 125-131.

Whitelaw, E. and Martin, D. I. (2001). Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat. Genet.* **27**, 361-365.

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H. R., Kauppinen, S. and Plasterk, R. H. (2005). MicroRNA expression in zebrafish embryonic development. *Science* **309**, 310-311.

Williamson, B. (1977). DNA insertions and gene structure. *Nature* **270**, 295-297.

Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B. and Schultz, P. G. (2005). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**, 1570-1573.

Wilton, S. D. and Fletcher, S. (2005). RNA splicing manipulation: strategies to modify gene expression for a variety of therapeutic outcomes. *Curr. Gene Ther.* **5**, 467-483.

Winkler, W. C. (2005). Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.* **9**, 594-602.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S. et al. (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871-880.

Wrana, J. L. (1994). H19, a tumour suppressing RNA? *BioEssays* **16**, 89-90.

Xie, X., Kamal, M. and Lander, E. S. (2006). A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl. Acad. Sci. USA* **103**, 11659-11664.

Yan, M. D., Hong, C. C., Lai, G. M., Cheng, A. L., Lin, Y. W. and Chuang, S. E. (2005). Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum. Mol. Genet.* **14**, 1465-1474.

Yang, W., Chendrimada, T. P., Wang, Q., Higuchi, M., Seeburg, P. H., Shiekhattar, R. and Nishikura, K. (2006). Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat. Struct. Mol. Biol.* **13**, 13-21.

Yekta, S., Shih, I. H. and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304**, 594-596.

Ying, S. Y. and Lin, S. L. (2005). Intronic microRNAs. *Biochem. Biophys. Res. Commun.* **326**, 515-520.

Young, T. L., Matsuda, T. and Cepko, C. L. (2005). The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Curr. Biol.* **15**, 501-512.

Zamore, P. D. and Haley, B. (2005). Ribo-gnome: the big world of small RNAs. *Science* **309**, 1519-1524.

Zamore, P. D., Tuschl, T., Sharp, P. A. and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25-33.

Zekri, L., Chebli, K., Tourriere, H., Nielsen, F. C., Hansen, T. V., Rami, A. and Tazi, J. (2005). Control of fetal growth and neonatal survival by the RasGAP-associated endoribonuclease G3BP. *Mol. Cell. Biol.* **25**, 8703-8716.

Zhou, Y. H., Zheng, J. B., Gu, X., Saunders, G. F. and Yung, W. K. (2002). Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res.* **12**, 1716-1722.

Zuckerkandl, E. and Cavalli, G. (2007). Combinatorial epigenetics, 'junk DNA', and the evolution of complex organisms. *Gene* **390**, 232-242.

Zuzarte-Luis, V. and Hurle, J. M. (2005). Programmed cell death in the embryonic vertebrate limb. *Semin. Cell Dev. Biol.* **16**, 261-269.

The Journal of Experimental Biology 210, 1492-1496
Published by The Company of Biologists 2007
doi:10.1242/jeb.02783

Glossary of terms

This section is designed to help readers adapt to the complex terminology associated with contemporary molecular genetics, genomics and systems biology. Fuller descriptions of these terms are available at http://www.wikipedia.org/

Ab initio prediction	methods used to predict the potential genes encoded in the genome, which are trained on datasets made of known genes, and used computationally to predict coding regions out of genome without the aid of cDNA sequence. Although their performance is improving, these algorithms perform very poorly on non-protein coding genes.
Annotation	as applied to proteins, DNA sequences or genes. The storage of data describing these entities (protein/gene identities, DNA motifs, gene ontology categorisation, etc.) within a biological database. Active projects include FlyBase and WormBase. See Gene ontology.
Assembly	the process of aligning sequenced fragments of DNA into their correct positions within the chromosome or transcript.
cDNA	complementary DNA. This is DNA synthesised from a mature mRNA template by the enzyme reverse transcriptase. cDNA is frequently used as an early part of gene cloning procedures, since it is more robust and less subject to degradation than the mRNA itself.
ChIP	chromatin immunoprecipitation assay used to determine which segments of genomic DNA are bound to chromatin proteins, mainly including transcription factors.
Chip	see Microarray.
ChIP-on-chip	use of a DNA microarray to analyse the DNA generated from chromatin immunoprecipitation experiments (see ChIP).
cis-acting	a molecule is described as *cis*-acting when it affects other genes that are physically adjacent, on the same chromosome, or are genetically linked or in close proximity (for mRNA expression, typically a promoter).
Collision-induced dissociation	a mechanism by which molecules (e.g. proteins) are fragmented to form molecular ions in the gas phase. These fragments are then analysed within a mass spectrometer to provide mass determination.
Connectivity	a term from graph theory, which indicates the number of connections between nodes or vertices in a network. Greater connectedness between nodes is generally used as a measure of robustness of a network.
CpG islands	regions that show high density of ‘C followed by G’ dinucleotides and are generally associated with promoter elements; in particular, stretches of DNA of at least 200 bp with a C–G content of 50% and an observed CpG/expected CpG in excess of 0.6. The cytosine residues can be methylated, generally to repress transcription, while demethylated CpGs are a hallmark of transcription. CpG dinucleotides are under-represented outside regulatory regions, such as promoters, because methylated C mutates into T by deamination.
Edge	as in networks. Connects two nodes (or vertices) within a system. These concepts arise from graph theory.
Enhancer	a short segment of genomic DNA that may be located remotely and that, on binding particular proteins (*trans*-acting factors), increases the rate of transcription of a specific gene or gene cluster.
Epistasis	a phenomenon when the properties of one gene are modified by one or more genes at other loci. Otherwise known as a genetic interaction, but epistasis refers to the statistical properties of the phenomenon.

eQTL	the combination of conventional QTL analysis with gene expression profiling, typically using microarrays. eQTLs describe regulatory elements controlling the expression of genes involved in specific traits.
EST	expressed sequence tag. A short DNA sequence determined for a cloned cDNA representing portions of an expressed gene. The sequence is generally several hundred base pairs from one or both ends of the cloned insert.
Exaptation	a biological adaptation where the current function is not that which was originally evolved. Thus, the defining (derived) function might replace or persist with the earlier, evolved adaptation.
Exon	any region of DNA that is transcribed to the final (spliced) mRNA molecule. Exons interleave with segments of non-coding DNA (introns) that are removed (spliced out) during processing after transcription.
Gene forests	genomic regions for which RNA transcripts, produced from either DNA strand, have been identified without gaps (non-transcribed genomic regions). Conversely, regions in which no transcripts have ever been detected are called ‘gene deserts’.
Gene interaction network	a network of functional interactions between genes. Functional interactions can be inferred from many different data types, including protein–protein interactions, genetic interactions, co-expression relationships, the co-inheritance of genes across genomes and the arrangement of genes in bacterial genomes. The interactions can be represented using network diagrams, with lines connecting the interacting elements, and can be modelled using differential equations.
Gene ontology (GO)	an ontology is a controlled vocabulary of terms that have logical relationships with each other and that are amenable to computerised manipulation. The Gene Ontology project has devised terms in three domains: biological process, molecular function and cell compartment. Each gene or DNA sequence can be associated with these annotation terms from each domain, and this enables analysis of microarray data on groups of genes based on descriptive terms so provided. See http://www.geneontology.org
Gene set enrichment analysis	a computational method that determines whether a defined set of genes, usually based on their common involvement in a biological process, shows statistically significant differences in transcript expression between two biological states.
Gene silencing	the switching-off of a gene by an epigenetic mechanism at the transcriptional or post-transcriptional levels. Includes the mechanism of RNAi.
Genetic interaction (network)	a genetic interaction between two genes occurs when the phenotypic consequences of a mutation in one gene are modified by the mutational status at a second locus. Genetic interactions can be aggravating (enhancing) or alleviating (suppressing). To date, most high-throughput studies have focussed on systematically identifying synthetic lethal or sick (aggravating) interactions, which can then be visualised as a network of functional interactions (edges) between genes (nodes).
Genome	a portmanteau of gene and chromosome, the entire hereditary information for an organism that is embedded in the DNA (or, for some viruses, in RNA). Includes protein-coding and non-coding sequences.
Heritability	phenotypic variation within a population is attributable to the genetic variation between individuals and to environmental factors. Heritability is the proportion due to genetic variation usually expressed as a percentage.
Heterologous hybridization	the use of a cDNA or oligonucleotide microarray of probes designed for one species with target cRNA/cDNAs from a different species.
Homeotic	the transformation of one body part to another due to mutation of specific developmentally related genes, notably the *Hox* genes in animals and *MADS-box* genes in plants.
Hub	as in networks. A node with high connectivity, and thus which interacts with many other nodes in the network. A hub protein interacts with many other proteins in a cell.

Hybridisation	the process of joining (annealing) two complementary single-stranded DNAs into a single double-stranded molecule. In microarray analysis, the target RNA/DNA from the subject under investigation is denatured and hybridised to probes that are immobilised on a solid phase (i.e. glass microscope slide).
Hypomorph	in genetics, a loss-of-function mutation in a gene, but which shows only a partial reduction in the activity it influences rather than a complete loss (cf. hypermorph, antimorph, neomorph, etc).
Imprinting	a phenomenon where two inherited copies of a gene are regulated in opposite ways, one being expressed and the other being repressed.
Indel	insertion and deletion of DNA, referring to two types of genetic mutation. To be distinguished from a 'point mutation', which refers to the substitution of a single base.
Interactome	a more or less comprehensive set of interactions between elements within cells. Usually applied to genes or proteins as defined by transcriptomic, proteomic or protein–protein interaction data.
Intron	see Exon.
KEGG	The Kyoto Encyclopedia of Genes and Genomes is a database of metabolic and other pathways collected from a variety of organisms. See http://www.genome.jp/kegg
Metabolomics	the systematic qualitative and quantitative analysis of small chemical metabolite profiles. The metabolome represents the collection of metabolites within a biological sample.
Metagenomics	the application of genomic techniques to characterise complex communities of microbial organisms obtained directly from environmental samples. Typically, genomic tags are sequence characterised as markers of each species to inform on the range and abundance of species in the community.
Microarray	an arrayed set of probes for detecting molecularly specific analytes or targets. Typically, the probes are composed of DNA segments that are immobilised onto the solid surface, each of which can hybridise with a specific DNA present in the target preparation. DNA microarrays are used for profiling of gene transcripts.
Model species	a species used to study particular biological phenomena, the outcome offering insights into the workings of other species. Usually, the selection is based on experimental tractability, particularly ease of genetic manipulation. For the geneticist, it is an organism with inbred lines where sibs will be >98% identical (i.e. *Drosophila*, *Caenorhabditis elegans* and mice). For genomic science, it refers to a species for which the genomic DNA has been sequenced.
miRNA	a category of novel, very short, non-coding RNAs, generated by the cleavage of larger precursors (pri-miRNA). These short RNAs are included in the RNA-induced silencing complex (RISC) and pair to the 3′ ends of target RNA, blocking its translation into proteins (in animals) or promoting RNA cleavage and degradation (in plants).
mRNA	a protein-coding mRNA containing a protein-coding region (CDS), preceded by a 5′ and followed by a 3′ untranslated region (5′ UTR and 3′ UTR). The UTRs contain regulatory elements. A full-length cDNA contains the complete sequence of the original mRNA, including both UTRs. However, it is often difficult to assign the starting–termination positions for protein synthesis unambiguously. A cDNA containing the entire CDS is often considered acceptable for bioinformatic and experimental studies requiring full-length cDNAs.
ncRNA	non-coding RNA is any RNA molecule with no obvious protein-coding potential for at least 80 or 100 amino acids, as determined by scanning full-length cDNA sequences. It includes ribosomal (rRNA) and transfer RNAs (tRNA) and is now known to include various sub-classes of RNA, including snoRNA, siRNA and piRNA. Just like the coding mRNAs, a large proportion of ncRNAs are transcribed by RNA polymerase II and are large transcripts. A description of the many forms of ncRNA can be found at http://en.wikipedia.org/wiki/Non-coding_RNA.

Node	as in networks. Objects linked by edges to create a network.
PCR	polymerase chain reaction. A molecular biology technique for replicating DNA *in vitro*. The DNA is thus amplified, sometimes from very small amounts. PCR can be adapted to perform a wide variety of genetic manipulations.
piRNA	Piwi-interacting RNA. A class of RNA molecules (29–30 nt long) that complex with Piwi proteins (a class of the Argonaute family of proteins) and are involved in transcriptional gene silencing.
PMF	peptide mass fingerprinting. An analytical technique for protein identification in which a protein is fragmented using proteases. The resulting peptides are analysed by mass spectrometry and these masses compared against a database of predicted or measured masses to generate a protein identity.
Polyadenylation	the covalent addition of multiple A bases to the 3′ tail of an mRNA molecule. This occurs during the processing of transcripts to form the mature, spliced molecule and is important for regulation of turnover, trafficking and translation.
Post-source decay	in mass spectrometry. The fragmentation of precursor molecular ions as they accelerate away from the ionisation source of the mass spectrometer. All precursor ions leaving the ion source have approximately the same kinetic energy, but fragmentation results in smaller product ions that can be distinguished from precursor ions using a ‘reflectron’ by virtue of their lower kinetic energies.
Post-translational modification	the chemical modification of a protein after synthesis through translation. Some modifications, notably phosphorylation, affect the properties of the protein, offering a means of regulating function.
Principal component analysis (PCA)	a technique for simplifying complex, multi-dimensional datasets to a reduced number of dimensions, the principal components. This procedure retains those characteristics of the data that relate to its variance.
Promoter	a regulatory DNA sequence, generally lying upstream of an expressed gene, which in concert with other often distant regulatory elements directs the transcription of a given gene.
Proteome	the entire protein complement of an organism, tissue or cell culture at a given time.
Quantitative trait	inheritance of a phenotypic property or characteristic that varies continuously between extreme states and can be attributed to interactions between multiple genes and their environment.
qPCR	quantitative real-time PCR, sometimes called real-time PCR. A more quantitative form of RT-PCR in which the quantity of amplified product is estimated after each round of amplification.
QTL	quantitative trait loci. A region of DNA that contains those genes contributing to the trait under study.
RISC	RNA-induced silencing complex. A protein complex that mediates the double-stranded RNA-induced destruction of homologous mRNA.
RNAi	RNA interference or RNA-mediated interference. The process by which double-stranded RNA triggers the destruction of homologous mRNA in eukaryotic cells by the RISC.
RT-PCR	reverse transcription–polymerase chain reaction. A technique for amplifying a defined piece of RNA that has been converted to its complementary DNA form by the enzyme reverse transcriptase. See qPCR.
siRNA	small interfering RNA, or silencing RNA. A class of short (20–25 nt), double-stranded RNA molecules. It is involved in the RNA interference pathway, which alters RNA stability and thus affects RNA concentration and thereby suppresses the normal expression of specific genes. Widely used in biomedical research to ablate specific genes.

snoRNA	small nucleolar RNA. A sub-class of RNA molecules involved in guiding chemical modification of ribosomal RNA and other RNA genes as part of the regulation of gene expression.
SNP	single nucleotide polymorphism. A single base-pair mutation at a specific locus, usually consisting of two alleles. Because SNPs are conserved over evolution, they are frequently used in QTL analysis and in association studies in place of microsatellites, and in genetic fingerprinting analyses.
SSH	suppressive subtractive hybridisation. A powerful protocol for enriching cDNA libraries for genes that differ in representation between two or more conditions. It combines normalisation and subtraction in a single procedure and allows the detection of low-abundance, differentially expressed transcripts, such as those involved in signalling and signal transduction.
Structural RNAs	a class of non-coding RNA, long known to have a structural role (for instance, the ribosomal RNAs), transcribed by RNA polymerase I or III.
Systems biology	treatment of biological entities as systems composed of defined elements interacting in defined ways to enable the observed function and behaviour of that system. The properties of the systems are embedded in a quantitative model that guides further tests of systems behaviour.
TATA-boxes	sequences in promoter regions constituted by TATAAA, or similar variants, which were considered the hallmark of Promoters. Recent data show that they are present only in the minority of promoters, where they direct transcription at a single well-defined location some 30 bp downstream of this element.
trans-acting	a factor or gene that acts on another unlinked gene, a gene on a separate chromosome or genetically unlinked usually through some diffusible protein product (for mRNA expression, typically a transcription factor).
Transcript	an RNA product produced by the action of RNA polymerase reading the sequence of bases in the genomic DNA. Originally limited to protein-coding sequences with flanking UTRs but now known to include large numbers of products that do not code for a protein product.
Transcriptome	the full set of mRNA molecules (transcripts) produced by the system under observation. Whilst the genome is fixed for a given organism, the transcriptome varies with context (i.e. tissue source, ontogeny, external conditions or experimental treatment).
Transgene	a gene or genetic material that has been transferred between species or between organisms using one of several genetic engineering techniques.
Transinduction	generation of transcripts from intergenic regions. At least some such products do not relate to a definable promoter or transcriptional start site.
Transposon	sequences of DNA able to move to new positions within the genome of a single cell. This event might cause mutation at the site of insertion. Also called ‘mobile genetic elements’ or ‘jumping genes’.
Transvection	an epigenetic phenomenon arising from the interaction between one allele and the corresponding allele on the homologous chromosome, leading to gene regulation.
TUs	transcriptional units. Used to group all of the overlapping RNA transcripts that are transcribed from the same genomic strand and share exonic sequences.
UTR	untranslated region. Regions of the mRNA that lie at either the 3′ or 5′ flanking ends of the molecule (i.e. 3′ UTR and 5′ UTR). They bracket the protein-coding region and contain signals and binding sites that are important for the regulation of both protein translation and RNA degradation.