47. Ferber, D. GM crops in the cross hairs. *Science* **286**, 1662–1666 (1999).
48. Bosch, X. USA fights Europe's ban on genetically modified food. *The Lancet* **361**, 1798 (2003).
49. Bosch, X. GM foods in new dispute. *The Lancet* **362**, 714 (2003).
50. Mitchell, P. Europe angers US with strict GM labeling. *Nature Biotechnol.* **21**, 6 (2003).
51. World Trade Organization. *European Communities — Measures Affecting the Approval and Marketing of Biotech Products* [online], <http://www.wto.org/english/tratop_e/dispu_e/dispu_subjects_index_e.htm#gmos> (2003).
52. Kolata, G. *Clone: The Road to Dolly and the Path Ahead* (William Morrow and Company, New York, 1998).
53. Anderson, N. Human cloning success startles lawmakers. *Los Angeles Times* A13 (27 Nov 2001).
54. Isasi, R. *Database of Global Policies on Human Cloning and Germ-line Engineering* [online], <http://www.glphr.org/genetic/genetic.Htm> (2003).
55. Lenoir, N. Universal declaration on the human genome and human rights: the first legal and ethical framework at the global level. *Columbia Human Rights Law Review* **30**, 537–561 (1999).
56. Carson, R. *Silent Spring* (Houghton Mifflin, Boston, 1962).
57. Whorton, J. *Before Silent Spring: Pesticides and Public Health in Pre-DDT America* (Princeton Univ. Press, Princeton, 1974).
58. Graham, F. *Since Silent Spring* (Houghton-Mifflin, Boston, 1970).
59. Ashworth, W. *The Carson Factor* (Hawthorn Books, New York, 1979).
60. Hilgartner, S. *Science on Stage: Expert Advice as Public Drama* (Stanford Univ. Press, Stanford, California, 2000).
61. Clayton, E. W. Ethical, legal, and social implications of genomic medicine. *N. Engl. J. Med.* **349**, 562–569 (2003).
62. Hellman, D. What makes genetic discrimination exceptional? *Am. J. Law Med.* **29**, 77–116 (2003).
63. Zitner, A. Senate blocks genetic discrimination. *Los Angeles Times* 16 (15 Oct 2003).
64. Nys, H. *et al. Genetic Testing: Patient's Rights, Insurance and Employment — A Survey of Regulations in the European Union* (Office for Official Publications of the European Communities, Luxemborg, 2002).
65. Carnegie Commission on Science, Technology and Government. *Science, Technology, and Congress: Organizational and Procedural Reforms: A Report of Carnegie Commission on Science, Technology, and Government* (The Commission, New York, 1994).
66. European Commission. *Life Science and Biotechnology: A Strategic Vision* [online], <http://europa.eu.int/comm/biotechnology/introduction_en.html > (2004).
67. Shapiro, H. T. Reflections on the interface of bioethics, public policy and science. *Kennedy Inst. Ethics J.* **9**, 209–224 (1999).
68. Yudell, M. A. Accounting for the fear factor. *Genome Technol.* 56 (2001).
69. Terry, S. F. & Davidson, M. E. Empowering the public to be informed consumers of genetic technologies and services. *Community Genet.* **3**, 148–150 (2000).
70. Ard, C. F. & Natowicz, M. R. A seat at the table: membership in federal advisory committees evaluating public policy in genetics. *Am. J. Public Health* **91**, 787–790 (2001).
71. Caulfield, T., Gold, E. R. & Cho, M. K. Patenting human genetic material: refocusing the debate. *Nature Rev. Genet.* **1**, 227–231 (2000).
72. Knoppers, B. M. Status, sale and patenting of human genetic material: an international survey. *Nature Genet.* **22**, 23–26 (1999).
73. *Health Effects Institute* [online], <http://www.healtheffects.org> (2004).
74. Anderson, F. R. Science advocacy and scientific due process. *Issues Sci. Technol.* **16**, 71–76 (2001).
75. Schulte, P. A. & Lomax, G. Assessment of the scientific basis for genetic testing of railroad workers with carpal tunnel syndrome. *J. Occup. Environ. Med.* **45**, 592–600 (2003).
76. Weiss, R. Ignorance undercuts gene tests' potential. *Washington Post* A1 (2 Dec 2000).
77. Greely, H. T. Human genomics research. New challenges for research ethics. *Perspect. Biol. Med.* **44**, 221–229 (2001).
78. Wendler, D., Prasad, K. & Wilfond, B. Does the current consent process minimize the risks of genetics research? *Am. J. Med. Genet.* **113**, 258–262 (2002).
79. Rothstein, M. A. & Epps, P. G. Pharmacogenomics and the (ir)relevance of race. *Pharmacogenomics J.* **1**, 104–108 (2001).
80. Weijer, C. & Miller, P. B. Protecting communities in pharmacogenetic and pharmacogenomic research. *Pharmacogenomics J.* **4**, 9–16 (2004).
81. Foster, M. W., Sharp, R. R. & Mulvihill, J. J. Pharmacogenetics, race, and ethnicity: social identities and individualized medical care. *Ther. Drug Monit.* **23**, 232–238 (2001).
82. Silver, L. *Remaking Eden: Cloning and Beyond in a Brave New World* (Avon Books, New York, 1999).
83. Walters, L. & Palmer, J. G. *The Ethics of Human Gene Therapy* (Oxford Univ. Press, New York, 1997).
84. Rothman, D. & Rothman, S. *The Pursuit of Perfection: The Promise and Perils of Medical Enhancement* (Panthoen Books, New York, 2003).
85. Andrews, L. B. *The Clone Age: Adventures in the New World of Reproductive Technology* (Henry Holt, New York, 1999).
86. Leary, W. E. Panel urges caution in producing gene-altered animals. *New York Times* A12 (21 Aug 2002).

Acknowledgements

The work reported in this publication was supported in part by the National Institute of Environmental Health Sciences and the National Human Genome Research Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Environmental Health Sciences, the National Human Genome Research Institute or the National Institutes of Health. The authors wish to thank R. DeSalle and D. Rosner for their thoughtful comments on earlier versions of the paper, and M. Sander for her editorial assistance.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

Database of Global Policies on Human Cloning and Germline Engineering: http://www.glphr.org/genetic/genetic.htm

FURTHER INFORMATION

European Initiative for Biotechnology Education: http://www.eibe.lnfo

National Council of State Legislators. Genetic Laws and Legislative Activity: http://www.ncsl.org/programs/health/genetics/charts.htm

Wellcome Trust Biomedical Ethics Programme: http://www.wellcome.ac.uk/en/1/pinbiorev.html

Access to this interactive links box is free online.

OPINION

RNA regulation: a new genetics?

John S. Mattick

Do non-coding RNAs that are derived from the introns and exons of protein-coding and non-protein-coding genes represent a fundamental advance in the genetic operating system of higher organisms? Recent evidence from comparative genomics and molecular genetics indicates that this might be the case. If so, there will be profound consequences for our understanding of the genetics of these organisms, and in particular how the trajectories of differentiation and development and the differences among individuals and species are genomically programmed. But how might this hypothesis be tested?

Perhaps the most fundamental belief in molecular biology is that genes are generally protein-coding — an extension of the central dogma and the fundamental ethos of biochemistry. The central dogma holds that genetic information flows from DNA to RNA to proteins. However, this has usually been interpreted to mean that genetic information flows from DNA to proteins via mRNA — that is, that genes are generally synonymous with proteins, and that genetic output is entirely or almost entirely transacted by proteins.

This conclusion is essentially correct for prokaryotes, in which the early experiments that defined our understanding of genes and gene expression were carried out. It has since been confirmed by the complete sequencing of many bacterial and archaeal genomes, which predominantly comprise protein-coding sequences that are flanked by 5′ and 3′ *cis*-regulatory elements that operate to control the expression of these sequences at the transcriptional or translational level. The only exceptions are genes that encode infrastructural RNAs (rRNAs, tRNAs) that are required for protein synthesis, and a small number of genes that express non-translated RNAs with regulatory functions[1–3], which occupy no more than 1% of the genome sequence. So, in prokaryotes at least, proteins comprise not only the primary functional and structural components of cells, but are also the main agents by which cellular dynamics are controlled, in conjunction with *cis*-regulatory elements and environmental signals.

It has long been assumed that the same is true in multicellular organisms, despite the fact that the proportion of protein-coding sequences declines as a function of complexity and is only a small minority of the genomic

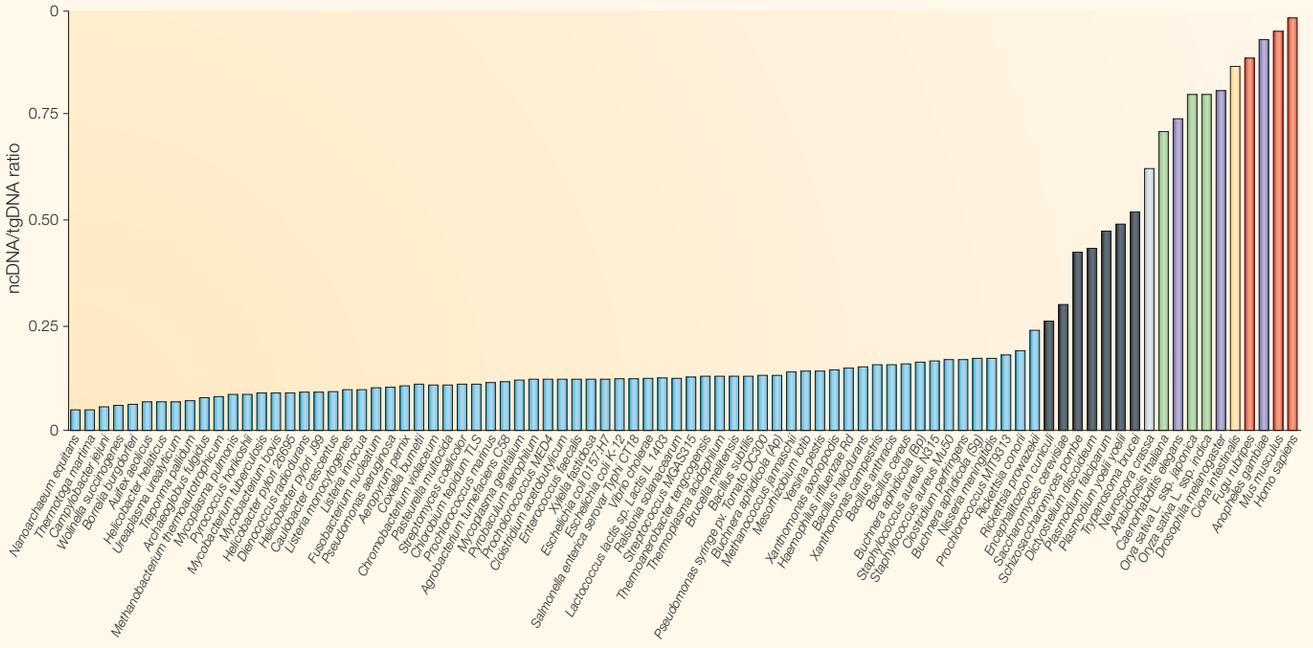

Figure 1 | **The ratio of non-coding to protein-coding DNA rises as a function of developmental complexity.** Prokaryotes have less than 25% non-coding DNA, simple eukaryotes have between 25 and 50% non-coding DNA and more complex fungi, plants and animals have more than 50%, rising to approximately 98.5% non-coding DNA in humans — which also have a genome size that is three orders of magnitude larger than prokaryotes. Note that this analysis corrects for ploidy, whereas pre-genomic estimations of the amount of DNA in different organisms did not. The different colours represent prokaryotes (bacteria and archaea) (blue), simple eucharyotes (black), *Neurospora crassa* (grey), plants (green), non-chordate invertebrates (nematodes, insects) (purple), *Ciona intestinalis* (urochordate) (yellow) and vertebrates (red). ncDNA, non-coding DNA; tgDNA, total genomic DNA. Reproduced with permission from REF. 77 © (2003) BioMed Central Ltd.

programming of complex organisms such as mammals (FIG. 1). This assumption has led to several logical extensions and subsidiary assumptions. In particular, it is assumed that the increased complexity of eukaryotes is explained by the combinatorics of regulatory factors that intersect with more complex promoters[4,5], with the corollary that most non-protein-coding sequences in eukaryotic genomes (98.5% in humans) are either *cis*-regulatory and structural elements or molecular hobos and evolutionary debris[6–10].

This article challenges these assumptions and suggests that our understanding of the amount and structure of genetic information in higher organisms is vastly incomplete. I examine what is required to programme complex objects and explain the logic behind my previously published hypothesis that the main output of the genomes of complex organisms is genetically active but non-coding RNA (ncRNA)[11–14]. I also review the new evidence that further strengthens this hypothesis and, most importantly, suggest how it might be tested. I argue that the principal advance in complex organisms was the development of a digital programming system based on ncRNA signalling, which bypassed the complexity limits that are imposed by accelerating regulatory networks that operate with proteins alone.

If this hypothesis is correct, the current conceptions of how genetic information is encoded and transmitted in higher organisms will need to be re-assessed, and a new framework developed for the analysis of genomic sequence data. This framework might allow us to understand the true basis of the evolution and developmental programming of complex organisms, and the basis of individual and species diversity.

Programming complex organisms

Complex organisms require two interrelated levels of programming. The first involves specifying their structural and functional components (proteins and their derived products). The second involves specifying how these components are arrayed and assembled into higher levels of organization (cells and organs), together with the control systems that manage their function, which include components that act as environmental sensors and relays. All of this information must ultimately be encoded in the genome.

Combinatorics and complexity. In humans, there are trillions of precisely patterned and positionally distinct cell types (BOX 1). Can this degree of positional and functional identity, and detailed four-dimensional architecture, be specified solely by the combinatorics of protein regulators that intersect with compound control sequences? The validity of this assumption is rarely examined, but is bound up, at least in part, with the question of how many regulatory inputs can sensibly be integrated, for example, at different promoters or splicing complexes, to produce different outcomes[4,5]. It is also bound up with a consideration of how the regulatory overhead must scale with the increased complexity of organized systems (see below).

It is true that complexity is an emergent property of interactions[15]. However, although necessary, this is not sufficient to explain organized complexity. As elegantly articulated by Dennett[16], combinatorics can generate vastly expanding universes of possibilities, but most of these are chaotic and meaningless, and both evolution and development have to navigate a course through these possibilities to find those that are sensible and competitive. Evolution does this by trial and error, with selected outcomes in the derived genomes that not only specify the structural and functional components of cells, but also the higher-order architectural programmes for growth and development.

The problem is not how to generate complexity — that is easy — but rather how to control it to specify ordered trajectories that lead to highly organized and complex organisms.

Box 1 | How many different cells are there in complex organisms?

The nematode worm *Caenorhabditis elegans*, the cellular ontogeny of which has been precisely mapped, has 1,179 and 1,090 distinct somatic cells (including those that undergo programmed cell death) in the male and female, respectively, each with a defined history and fate. Therefore, if we take the developmental trajectories and cell position into account, *C. elegans* has 10^3 different cell identities, even if many of these cells are functionally similar. By this reasoning, although the number of different cell types in mammals is often considered to lie in the order of hundreds, it is actually in the order of 10^{12} if their positional identity and specific ontogeny are considered. Humans have an estimated 10^{14} cells, mostly positioned in precise ways and with precise organization, shape and function, in skeletal architecture, musculature and organ type, many of which (such as the nose) show inherited idiosyncrasies. Even if the actual number of cells with distinct identities is discounted by a factor of 100 (on the basis that 99% of the cells are simply clonal expansions of a particular cell type in a particular location or under particular conditions (for example, fat, muscle or immune cells)), there are still 10^{12} positionally different cell types.

This requires an enormous amount of information, particularly regulatory information. Indeed, the best (albeit abstract) definition of relative complexity is the minimum amount of information that is required to specify the ontogeny and operation of the object or system[17]. On this basis, the minimum amount of DNA sequence information that is required to specify a vertebrate, at least according to our present knowledge, is 365 Mb (the genome size of the pufferfish *Fugu rubripes*), of which only approximately 10% encodes protein[9]. The rest cannot be easily dismissed as junk, as it largely comprises DNA sequences of high complexity (approximately 22% in introns and the rest intergenic), and is therefore apparently information rich.

How does regulation scale with complexity? Both intuitive and mathematical considerations indicate that the amount of regulation must increase as a nonlinear (probably quadratic) function of the number of genes in the network[18–20]. First, unless they are constitutively expressed, new genes (or splice variants) with different functions will need to be specifically regulated, which gives a linear increase in the number of regulators or combinations thereof. This is then compounded by the fact that a proportion of these new regulators will also require regulation, and that the impact of the activity of the new genes will have to be integrated into the existing regulatory circuitry of the organism as a whole, if the system is not to become disconnected. So, as the system becomes more complex, an increasing proportion has to be devoted to regulation. This nonlinear relationship between regulation and function is a feature of all integrally organized systems. Therefore, all such systems have an intrinsic complexity limit that is imposed by their accelerating control architecture (that is, if the fractional cost of additional regulation exceeds the benefit of new functions), until or unless the physical nature of the regulatory system undergoes a state transition to a more powerful system, for example, by the use of digital instead of analogue controls (J.S.M. and M. J. Gagen, manuscript in preparation).

In agreement with this prediction, the number of protein regulators in prokaryotes has been found to increase quadratically with genome size[18] (FIG. 2). Moreover, extrapolation indicates that the point at which the number of new regulators will exceed the number of new functional modules (operons) is close to the observed upper limit of bacterial genome sizes[19]. That is, the system seems to have become saturated, with further genomic and functional complexity constrained by the accelerating regulatory cost. This indicates that the complexity of prokaryotes, for which only simple developmental structures and transitions are possible, has had a ceiling imposed throughout evolution by regulatory overhead[11], rather than by environmental, structural or biochemical factors as has been commonly assumed. This is consistent with the limitation of life on Earth to microbial systems for most of its evolutionary history (FIG. 3).

This also indicates that protein-based regulation has reached its effective limit in prokaryotes, and that combinatoric controls cannot overcome this limit — there is no *a priori* reason why prokaryotes could not easily have evolved more complex promoters and combinatoric regulatory control if this was a viable option. Reciprocally, eukaryotes must have found a solution to this problem as a precondition of their exploration of more complex space.

RNA: a digital solution?

Genome sequencing projects have largely been reported in terms of the number of identifiable protein-coding genes. However, until relatively recently, the enormous increase in the transcription of ncRNA in these organisms — which accounts for approximately 98% of all genomic output in humans[13] — had gone unnoticed. This ncRNA comprises introns in protein-coding genes and other transcripts that do not seem to encode proteins. So, either the genomes of complex organisms are replete with useless transcription, or these ncRNAs are fulfilling some unexpected functions. If the latter is true, these functions must be transmitted through RNA, which indicates that RNA has evolved a new significance in the genetic programming of higher organisms.

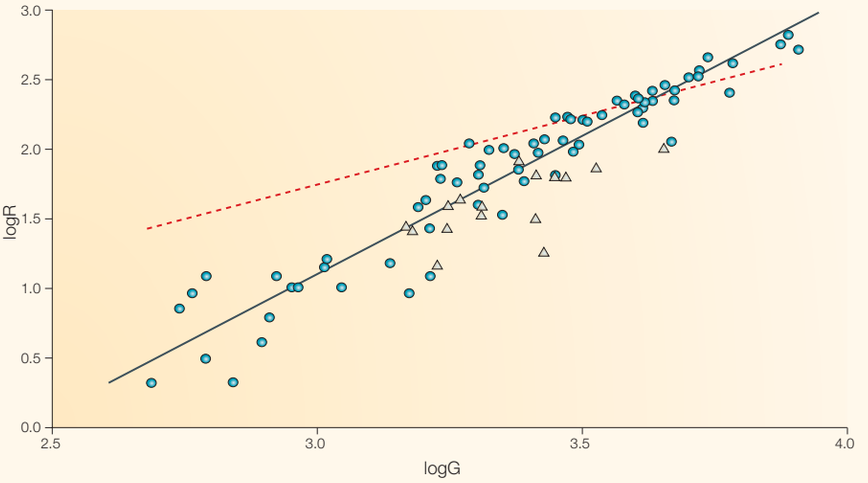

Figure 2 | **Double-logarithmic plot of the number of genes that encode regulatory proteins (R) against the total number of genes (G) for bacteria (circles) and archaea (triangles).** The log–log distribution is well described by a straight line with slope 1.96 ± 0.15 (95% confidence interval), corresponding to a quadratic relationship between regulator number and total gene number (note that if $R = AN^x$, a log transformation yields $\log R = x\log N + \log A$, in which the slope of the resulting line is equal to the exponent x). Dashed lines show the best linear fit to the data. Modified from REFS 18,19.

Introns. The key to understanding the transition to a predominantly RNA-based regulatory system in higher organisms is to first consider introns (BOX 2). Introns account for approximately 95–97% of the average protein-coding gene in humans[6,7,21], which means that although protein-coding sequences occupy only about 1.5% of the human genome, at least one-third of the genome must actually be transcribed. Furthermore, if the numerous other genes that express ncRNAs are taken into account, then at least half of the human genome is transcribed[14].

Although it is widely believed that intronic RNA is non-functional (simply being degraded and recycled after excision by splicing), there is another equally, if not more, plausible possibility — that introns are genetically active and that intronic RNA feeds genetic information into the regulatory network of the cell[11,12]. Given the long history of the presence of these sequences in eukaryotic genes, it would be surprising if evolution had not explored this possibility.

ncRNAs: a parallel digital regulatory system. If the possibility is entertained that introns are functional (actively transmitting genetic information through RNA molecules), then an entirely different type of regulation becomes possible, with an entirely different set of logical extensions and interesting predictions.

First, it would mean that the genetic operating system of complex eukaryotes is fundamentally different and much more sophisticated than that of simple prokaryotes. Eukaryotic genes would express two types of information in parallel — proteins and (to borrow a term from neurobiology) EFFERENCE RNA SIGNALS that can communicate with other genes or gene products independently of the biochemical function of the encoded protein in the host transcript[11–13]. This leads to the deeper prediction that the emergence of a true parallel processing system was, in all likelihood, fundamental to the evolution and development of complex organisms[11,12].

Second, it would be predicted that these ncRNA sequences have been under selective encouragement to expand in complex organisms, with the further prediction that some, perhaps many, genes have evolved to only produce RNA signals as higher-order regulators in this network[11,12]. Both predictions are consistent with the known data. Complex eukaryotes have much more extensive introns than simpler ones. There are also increasing numbers of ncRNA transcripts being identified, which might account for half or more of all transcripts in mammals[14]. Furthermore,

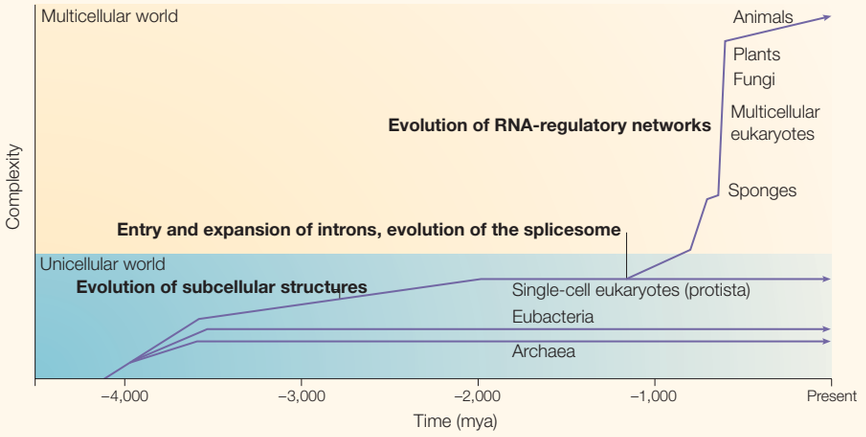

Figure 3 | **A simplified biological history of the Earth.** This graph is intended to present an overview. Some dates are still being debated and the abscissa ('complexity') has an arbitrary scale. Life appeared on Earth approximately 4,200 million years ago (mya), either arising as, or quickly streaming into, three main kingdoms — the eukarya, the bacteria and the archaea. Life remained unicellular, or at best colonial, for at least 3,000 million years. The common ancestor of the animals, plants and fungi is thought to have arisen approximately 1,200 mya, around the time at which the mitochondria entered the lineage through a rickettsial-like endosymbiont, an event that is postulated to have also brought with it type II self-splicing introns[75]. Whether or not these events were coincidental, the incidence of introns (and other non-coding sequences) correlates with the complexity of the organism after that point. In the Cambrian period (~520 mya), complex animal life exploded in an event known as the metazoan radiation, in which recognizable ancestors of all modern phyla appeared only in a single strata of rock[78]. What restrained the appearance of organized multicellular organisms for so long? Was it environmental or biochemical factors (such as oxygen tension and oxidative energy metabolism), or a primitive genetic operating system?

some genes, notably those that encode small nucleolar RNAs (although there are probably many others), are known to transmit information from introns, as their exons do not contain any open reading frames and seem to be degenerate[22–24]. Other genes express ncRNAs that are assembled from multiple exons, and in at least some cases are alternatively spliced[25–29].

Third, it would also be predicted that many of these RNAs will be processed, after transcription and splicing, into numerous smaller signals that can address different targets in the network, to influence chromatin architecture, transcription, alternative splicing, translational efficiency and RNA stability, and so on, at other loci[14]. The discovery of microRNAs, which are derived from both the introns and exons of longer precursors[30–32], and the involvement of RNA interference (RNAi) in chromosome dynamics and developmental pathways (see below), is entirely consistent with this prediction.

Fourth, as most of these RNA molecules themselves are unlikely to be catalytic (although some edit other RNAs)[23,24], these signals must be largely regulatory, sending epigenetic signals downstream into the system. For example, if the intronic RNAs that are derived from transcription of the β-globin gene are functional, it seems unlikely that they would be involved in oxygen transport, but rather in aspects of the developmental regulation and coordination of gene expression in the erythroid lineage, for which there is, in fact, good circumstantial evidence[33]. This is essentially a feed-forward system of endogenous control, a programme that in theory could set developmental trajectories, guided by environmental signals to provide contextual cues and to correct stochastic noise in the endogenous programme.

Finally, these RNAs must be generally conveying sequence-specific signals to their targets, presumably (in the main) to other RNAs and DNA. These targets must also then be acted on by a receptive infrastructure — that is, proteins that can recognize the secondary or tertiary structure of these signalled complexes and take appropriate action — for example, by modification of chromatin[34,35] or target degradation by RNAi[36]. This is, therefore, a digital system in which the signals and the consequent actions are separated.

Such a digital system would allow a quantum shift in regulatory sophistication, efficiency and versatility. A sequence-specific RNA signal in animals and plants can be just 22 nucleotides, almost 2 orders of magnitude smaller than that required to encode an average protein. It would also be an ideal way of embedding a forward-control system that can specify the complex suites of gene activity that underpin the ontogeny of complex organisms, particularly if increases in functional

Box 2 | **The history of nuclear introns**

Undoubtedly, the greatest surprise in the history of molecular biology was the discovery in the late 1970s that many genes in eukaryotes, especially in the higher eukaryotes, were fragmented into mosaics of protein-coding mRNA sequences (exons) that were interspersed with non-protein-coding sequences (intervening sequences or introns), which were excised before translation by splicing. These introns, because they did not encode protein, were generally assumed to be genetically inert (apart from possibly containing *cis*-regulatory signals). Introns were consequently rationalized as the stigmata of the prebiotic assembly of genes from cassettes of protein-coding information, albeit with a role in enabling protein-domain shuffling, a view that is essentially presented as fact in most molecular biology textbooks. Subsequent work has established that, in all likelihood, modern nuclear introns descended from self-splicing group II introns and expanded in eukaryotic genes relatively late in evolution. This expansion was aided by the separation of transcription from translation, which, conversely, is a strong counter-selective force in prokaryotes[11,75,76]. Whatever the precise origins of introns, the subsequent evolution of the SPLICEOSOME in the eukaryotes led to relaxation of their internal sequence constraints and an increase in the efficiency of their excision from primary transcripts[11]. This in turn provided the opportunity for these sequences to both expand and to drift and to explore new evolutionary and functional space, based on RNA rather than on proteins, although this is not to suggest that all introns will have acquired such capacity in any given lineage.

and architectural complexity require exponential increases in endogenous regulatory information. Indeed, as pointed out by Csete and Doyle[37], explosions in complexity in virtually all systems occur as a result of advanced controls and embedded networking, most of which is invisible to the observer.

Emerging evidence

There is now considerable evidence that RNA-mediated regulation is widespread in higher organisms, much of which has been summarized in previous articles[12–14]. However, recently there have been several surprising new observations that further strengthen the case.

Comparative genomics. Comparative analysis of the increasing number of sequenced animal genomes has uncovered patterns of conservation in intronic and intergenic sequences that collectively account for a much higher proportion of the observed conservation between genomes than protein-coding sequences[8,38–40]. For example, analysis of the *CFTR* and *SIM2* loci in several vertebrate species has identified many conserved segments[40,41]. Most of these segments are in introns and intergenic regions, and many cannot be detected by pair-wise sequence comparisons alone[40,41], which indicates that they are evolving under selective constraints (both positive and negative) in lineage-specific ways. Comparison of the dog, mouse and human genomes (which are relatively distant in terms of mammalian evolution) shows that there is significant conservation outside protein-coding sequences (estimated to be between three and ten times the amount of conservation that is observed within protein-coding sequences)[8,39]. This conservation occurs in blocks, the size and distribution of which is not consistent with neutral drift from a common ancestral sequence; there are some sequence blocks of several hundred nucleotides in which there is hardly a single nucleotide change between different vertebrate species (M. Pheasant, I. Makunin and J.S.M., manuscript in preparation). The selective pressures on these particular sequences are unknown, but one explanation is that they are part of networks with multiple interacting partners, making the odds of obtaining compensatory changes in all components effectively zero. In other cases, the level of sequence divergence is higher than would be expected, which indicates that there is a positive selection for changes to these sequences (related to phenotypic divergence) or that the underlying rate of neutral substitution is much higher than previously thought, which would make the blocks of conservation in non-protein-coding sequences even more impressive.

Non-coding transcripts. An increasing number of ncRNA genes are being identified, several of which have links to human diseases such as B-cell lymphoma, lung cancer, prostate cancer, cartilage-hair hypoplasia, spinocerebellar ataxia type 8, DiGeorge syndrome, autism and schizophrenia, among others[14,25,26,29,42–44]. Reliable estimates indicate that at least 7% of all transcripts do not encode protein[45]. This is likely to be just the tip of an iceberg, the full dimensions of which might take time to assess, particularly given the difficulties of establishing the functional relevance of non-coding transcripts[29,46].

Full-length cDNA analysis of the mouse has identified thousands of transcripts that do not contain any significant open reading frame[47]. Although the problems of incomplete reverse transcription and genomic contamination cannot be entirely discounted, many of these transcripts are distant from protein-coding sequences and most seem to be developmentally regulated[45]. Some of these transcripts are antisense to known or predicted genes, and it has been estimated that as many as 20% of all human genes have associated antisense transcripts[48]. Examination of EST collections indicates that the real figure might be much higher[49]. Antisense regulation has been shown to cause human genetic disease[50] and is clearly important at IMPRINTED LOCI[51], but it could be a more general mechanism for inter-allelic communication and dosage compensation at non-imprinted loci that involves local RNA regulatory loops[52]. This suggestion is consistent with the lower number of antisense transcripts on the mammalian X chromosome[53].

Related to this is the recent discovery of sense regulation by ncRNA. It has recently been reported that a non-coding pseudogene transcript regulates the expression of its homologous protein-coding gene[54]. There are around 20,000 pseudogenes in the human genome, which had been presumed to be non-functional[10]. This might have been a premature assumption, and it is now a moot point as to what fraction of these pseudogenes might be genetically active as RNA.

RNAi, disease and development. The phenomenon of RNAi, which is unique to eukaryotes, has now been shown to be central to plant and animal development[55,56], as well as to meiosis, mitosis and other aspects of chromosome dynamics[57–59].

MicroRNAs are believed to be involved in human disease[60,61], and at least some are derived from introns[32,62,63]. Many disease association studies are now finding no correlated mutations in exons, which indicates that the causative mutations are in the adjacent *cis*- or *trans*-acting regulatory sequences. Perhaps the best recent example of this is the elegant dissection of the callipyge ('beautiful buttocks') locus in sheep — an imprinted region with several protein-coding and ncRNA genes — in which a single nucleotide change in an intergenic region (the transcriptional status of which is unknown) is responsible for a changed musculature of the buttock[64]. A similar story has emerged with the genetic variation in the muscle mass of the domestic versus wild pig, which involves a single nucleotide change in the intron of the *IGF2* gene[65].

Molecular genetic analyses of the bithorax locus in *Drosophila melanogaster* — which, like all other well-studied loci including the

globin locus[33,64,66], contains a predominance of developmentally regulated ncRNA genes — have shown that the segment-specific transcription of 'intergenic' regulatory regions is required to establish an epigenetically inheritable activation of the expression of adjacent homeotic protein-coding genes[66,67]. This again indicates that local RNA-mediated regulatory loops are important in setting the subsequent epigenetic and transcriptional profiles of cells during development in complex organisms, and might finally provide an explanation for complex genetic phenomena such as TRANSVECTION[12] and TRANSINDUCTION[33].

Testing the new genetics

Analysis of the functions of ncRNA genes. The dissection of RNA-mediated genetic signalling will not be easy. The allocation of function to the increasing number of ncRNA genes that are discovered will be arduous[46]. It will involve examining sequence homologies, developmental expression patterns, subcellular localization, and both knockout and ectopic expression studies in transgenic animals, and will at least establish the general importance of these previously overlooked genetic outputs. Such experiments are underway, initially targeting a selection of the most highly conserved non-coding sequences in vertebrates and insects, at least some of which are expressed as stable RNAs (I. Makunin, E. Glazov, M. Pheasant and J.S.M., unpublished observations).

Molecular genetic analysis of intron-encoded signals. The key proof-of-principle experiments will be to show that intronic RNAs are genetically active, as this will validate the concept of a parallel output of both protein- and efference RNA regulatory signals from eukaryotic genes. These experiments are underway in several model organisms. In yeast, despite the limited number of introns, bioinformatic analysis has already — surprisingly — uncovered patterns of networks of sequence conservation between introns and other sequences within the genome, which cluster with high statistical significance within congruent GENE ONTOLOGY groups (S. Stanley and J.S.M., manuscript in preparation). Several of these introns are being targeted for site-specific deletion, using perturbation of microarray transcriptional profiles as a convenient phenotypic end point, followed by complementation studies to distinguish between conventional *cis*-regulatory elements and the possibility of *trans*-acting RNA sequences.

Bioinformatic analysis of RNA-mediated regulatory networks. The main advantage of RNA as a regulatory molecule is its compact size and sequence specificity. As noted already, the likelihood is that most RNA signals will be transmitted through primary sequence-specific interactions with other RNAs and with DNA, forming complexes that are recognized by proteins that contain particular types of domain. This provides an opportunity to identify both the potential transmitters and receivers (targets) in such networks, as well as the types of interacting protein (see below). Importantly, most of these interactions would be expected to involve RNA–RNA and RNA–DNA interactions (potentially including triplexes and other higher-order structures) that do not obey canonical base-pairing rules[68–70]. So, new algorithms will need to be developed that allow the search for these different types of interaction in genomic sequences. Moreover, conventional homology search algorithms (such as BLAST) are poor at finding short sequence homologies, especially if these involve mismatches (see below). More complete search algorithms, such as those based on SUFFIX ARRAYS and SUFFIX TREES[71], will be needed to analyse this properly.

Identification of RNA signalling complexes recognized by different classes of protein. Given that many types of RNA signal would be predicted to function at many levels (chromatin modification, transcriptional control, regulation of alternative splicing, and so on), how might such digital signals result in specific functional consequences? The ability of RNA to form strong interactions with other RNAs provides a clue — RNA–RNA and (to a lesser extent) RNA–DNA base pairing is stronger than DNA–DNA base pairing, and can allow for stable mismatches and the formation of particular secondary structures, such as bulges, stems and loops, which, rather than being seen as mismatch errors (as in DNA repair), might in fact contain embedded structural motifs that can be recognized by particular proteins. For example, perfect versus imperfect matching of microRNAs to their targets determines whether the mRNA target is actively degraded by the RNAi pathway or is translationally repressed[36].

So, the prediction can be made that if there are different types of RNA signal the different structures of the resulting complexes will be recognized and acted on by particular classes of nucleic-acid-binding protein, many of which have at present unknown or poorly understood target specificity. If this is generally correct, understanding these secondary structural and mismatch rules will in turn enrich the bioinformatic approaches to dissecting these networks at the genomic level. It will also allow better prediction of the regulatory consequences of different types of RNA signal by the development of specific algorithms to identify particular subsets that obey different sets of rules for the combination of sequence specificity and the type of secondary structure that is created by the interaction, bearing in mind that parts of the network will be silent in any given cell or lineage because an RNA transmitter or target is not expressed, or because a DNA target has been made inaccessible by chromatin modification.

Conclusions

If RNA-mediated regulation is real, then why has this system gone unrecognized for so long? First, we were unprepared for the possibility of an extensive RNA control network, despite early predictions that RNA might have important regulatory roles[72,73], because of the general assumption that regulatory information is transacted primarily by proteins.

Glossary

EFFERENCE RNA SIGNALS
Regulatory RNA signals that are produced in parallel with the primary gene product that allow forward control and coordination of networks of gene activity.

GENE ONTOLOGY
A hierarchical organization of concepts (ontology) with three organizing principles: molecular function, the tasks done by individual gene products; biological process that are accomplished by ordered assemblies of molecular functions; and cellular components, subcellular structures, locations and macromolecular complexes.

IMPRINTED LOCI
Loci at which the expression of an allele is different depending on whether it is inherited from the mother or the father.

SPLICEOSOME
A ribonucleoprotein complex that is involved in splicing nuclear pre-mRNA. It consists of 5 small nuclear ribonucleoproteins (snRNPs) and more than 50 non-snRNPs, which recognize and assemble on exon–intron boundaries to catalyse the excision of introns from the pre-mRNA.

SUFFIX ARRAYS
An array of all terminal substrings of a sequence string in lexicographical order, which allows a binary search.

SUFFIX TREES
A compact representation of a tree that corresponds to the suffixes of a given string in which all nodes with one child are merged with their parents in a branching structure.

TRANSINDUCTION
The induction of intergenic transcription of the β-globin cluster in non-erythroid cells by the expression of transiently transfected β-globin genes, which is not dependent on protein expression.

TRANSVECTION
Apparent cross-talk between alleles, in which complementation is observed between promoter mutations in one allele and structural mutations in the other, although in many cases the promoter region itself might produce a separate transcript.

Second, the system has largely been biochemically invisible — RNA is labile, and many, if not most, of the RNA signals are ephemeral and small, literally and metaphorically off the radar screen at the bottom of the gel. If not for the genetic discovery of *lin-4* and *let-7* in *C. elegans*, and their link with the RNAi pathway [74], itself an accidental discovery, it is doubtful whether we would be aware that microRNAs were present in eukaryotic cells.

Third, this regulatory system is genetically subtle, with different phenotypic signatures to those of protein-coding genes. Damage to proteins by mutation is usually obvious, and so tends to dominate the visible landscape of genetic screens. This is particularly true for those mutations that mainly involve single base changes, which in protein-coding sequences can be catastrophic, but in regulatory sequences might have much more subtle consequences. Known mutations in human promoters that give recognized phenotypes are rare, but that does not mean that promoters are non-functional, simply that they have different constraints and a degree of greater plasticity. So it will be for any regulatory sequence, particularly if they are participating in networks that might be intrinsically robust. It is important to remember that these networks are involved in programming the architecture, rather than the functional components, of complex systems. Embedded in these networks is the genetic specification of the body plan, and therefore both the principal source of species differences (given a relatively common component set or proteome) and the individual differences that underpin quantitative trait variation and susceptibility to disease.

The RNA regulatory system might have been the essential prerequisite to both the evolution of developmentally sophisticated multicellular organisms and the rapid expansion of phenotypic complexity into uncontested environments[11,12]. This also indicates that the principal source of the evolutionary diversity of complex organisms and their ability to colonize different ecological niches is the regulatory architecture, which is primarily encoded in ncRNA genes and introns, and therefore also indicates that most of their genomes are devoted to developmental programming and are under selective influences. This includes transposable elements that might have entered the system from elsewhere, but subsequently evolved *in situ* as part of the (genomic) community. Those sequences that are conserved among particular lineages are presumably those that are common to, for example, vertebrate or mammalian biology in general, whereas those sequences that are different might be equally functional but involved in specifying the differences among species. The best route to understanding this will be the intersection of molecular genetics and comparative genomics.

Note added in proof

Mapping of transcripts and transcription factors along human chromosomes 21 and 22 has indicated that the human genome contains approximately equal numbers of protein-coding and non-coding genes, consistent with my earlier predictions[14], that are bound by common transcription factors and regulated by common environmental signals[79,80].

John Mattick is at the ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia. e-mail: j.mattick@imb.uq.edu.au

doi:10.1038/nrg1321

1. Wassarman, K. M., Zhang, A. & Storz, G. Small RNAs in *Escherichia coli*. *Trends Microbiol.* **7**, 37–45 (1999).
2. Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* **11**, 941–950 (2001).
3. Klein, R. J., Misulovin, Z. & Eddy, S. R. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl Acad. Sci. USA* **99**, 7542–7547 (2002).
4. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
5. Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA* **100**, 5136–5141 (2003).
6. Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
7. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
8. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
9. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
10. Harrison, P. M. *et al.* Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280 (2002).
11. Mattick, J. S. Introns: evolution and function. *Curr. Opin. Genet. Dev.* **4**, 823–831 (1994).
12. Mattick, J. S. & Gagen, M. J. 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 (2001).
13. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986–991 (2001).
14. Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930–939 (2003).
15. Weng, G., Bhalla, U. S. & Iyengar, R. Complexity in biological signaling systems. *Science* **284**, 92–96 (1999).
16. Dennett, D. *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon Schuster, New York, 1995).
17. Li, M. & Vitanyi, P. M. B. *An Introduction to Kolmogorov Complexity and its Applications* 2nd edn (Springer, New York, 1997).
18. Croft, L. J., Lercher, M. J., Gagen, M. J. & Mattick, J. S. Is prokaryotic complexity limited by accelerated growth in regulatory overhead? *Genome Biol. Preprint Depository* [online], <http://genomebiology.com/qc/2003/5/1/p2> (2003).
19. Gagen, M. J. & Mattick, J. S. Inherent size constraints on prokaryote gene networks due to 'accelerating' growth. *arXiv Preprint Archive* [online], <http://arXiv.org/abs/q-bio.MN/0312021> (2004).
20. Gagen, M. J. & Mattick, J. S. Failed 'nonaccelerating' models of prokaryote gene regulatory networks. *arXiv Preprint Archive* [online], <http://arXiv.org/abs/q-bio.MN/0312022> (2004).
21. Scherer, S. W. *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).
22. Tycowski, K. T., Shu, M. D. & Steitz, J. A. A mammalian gene with introns instead of exons generating stable RNA products. *Nature* **379**, 464–466 (1996).
23. Kiss, T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**, 145–148 (2002).
24. Bachellerie, J. P., Cavaille, J. & Huttenhofer, A. The expanding snoRNA world. *Biochimie* **84**, 775–790 (2002).
25. Sutherland, H. F. *et al.* Identification of a novel transcript disrupted by a balanced translocation associated with DiGeorge syndrome. *Am. J. Hum. Genet.* **59**, 23–31 (1996).
26. Bussemakers, M. J. *et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **59**, 5975–5979 (1999).
27. Raho, G., Barone, V., Rossi, D., Philipson, L. & Sorrentino, V. The gas 5 gene shows four alternative splicing patterns without coding for a protein. *Gene* **256**, 13–17 (2000).
28. Charlier, C. *et al.* Human-ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (clpg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. *Genome Res.* **11**, 850–862 (2001).
29. Wolf, S. *et al.* B-cell neoplasia associated gene with multiple splicing (BCMS): the candidate *B-CLL* gene on 13q14 comprises more than 560 kb covering all critical regions. *Hum. Mol. Genet.* **10**, 1275–1285 (2001).
30. Lee, Y., Jeon, K., Lee, J. T., Kim, S. & Kim, V. N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* **21**, 4663–4670 (2002).
31. Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
32. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
33. Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N. J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).
34. Akhtar, A., Zink, D. & Becker, P. B. Chromodomains are protein–RNA interaction modules. *Nature* **407**, 405–409 (2000).
35. Hall, I. M. *et al.* Establishment and maintenance of a heterochromatin domain. *Science* **297**, 2232–2237 (2002).
36. Hutvagner, G. & Zamore, P. D. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**, 2056–2060 (2002).
37. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
38. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* **17**, 373–376 (2001).
39. Kirkness, E. F. *et al.* The dog genome: survey sequencing and comparative analysis. *Science* **301**, 1898–1903 (2003).
40. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
41. Frazer, K. A. *et al.* Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14**, 367–372 (2004).
42. Ridanpaa, M. *et al.* Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell* **104**, 195–203 (2001).
43. Vulliamy, T. *et al.* The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature* **413**, 432–435 (2001).
44. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 6087–6097 (2003).
45. Numata, K. *et al.* Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.* **13**, 1301–1306 (2003).
46. Storz, G. An expanding universe of noncoding RNAs. *Science* **296**, 1260–1263 (2002).
47. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
48. Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. & Hayashizaki, Y. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13**, 1324–1334 (2003).
49. Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnol.* **21**, 379–386 (2003).
50. Tufarelli, C. *et al.* Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nature Genet.* **34**, 157–165 (2003).

51. Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813 (2002).
52. Andersen, A. A. & Panning, B. Epigenetic gene regulation by noncoding RNAs. *Curr. Opin. Cell Biol.* **15**, 281–289 (2003).
53. Kiyosawa, H. & Abe, K. Speculations on the role of natural antisense transcripts in mammalian X chromosome evolution. *Cytogenet. Genome Res.* **99**, 151–156 (2002).
54. Hirotsune, S. *et al.* An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **423**, 91–96 (2003).
55. Carrington, J. C. & Ambros, V. Role of microRNAs in plant and animal development. *Science* **301**, 336–338 (2003).
56. Palatnik, J. F. *et al.* Control of leaf morphogenesis by microRNAs. *Nature* **425**, 257–263 (2003).
57. Hall, I. M., Noma, K. & Grewal, S. I. RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast. *Proc. Natl Acad. Sci. USA* **100**, 193–198 (2003).
58. Volpe, T. *et al.* RNA interference is required for normal centromere function in fission yeast. *Chromosome Res.* **11**, 137–146 (2003).
59. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell* **110**, 689–699 (2002).
60. McManus, M. T. MicroRNAs and cancer. *Semin. Cancer Biol.* **13**, 253–258 (2003).
61. Metzler, M., Wilda, M., Busch, K., Viehmann, S. & Borkhardt, A. High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes Chromosom. Cancer* **39**, 167–169 (2004).
62. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
63. Llave, C., Kasschau, K. D., Rector, M. A. & Carrington, J. C. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**, 1605–1619 (2002).
64. Georges, M., Charlier, C. & Cockett, N. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. *Trends Genet.* **19**, 248–252 (2003).
65. Van Laere, A. S. *et al.* A regulatory mutation in *IGF2* causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832–836 (2003).
66. Drewell, R. A., Bae, E., Burr, J. & Lewis, E. B. Transcription defines the embryonic domains of *cis*-regulatory activity at the *Drosophila* bithorax complex. *Proc. Natl Acad. Sci. USA* **99**, 16853–16858 (2002).
67. Rank, G., Prestel, M. & Paro, R. Transcription through intergenic chromosomal memory elements of the *Drosophila* bithorax complex correlates with an epigenetic switch. *Mol. Cell Biol.* **22**, 8026–8034 (2002).
68. Toulme, J. J., Di Primo, C. & Moreau, S. Modulation of RNA function by oligonucleotides recognizing RNA structure. *Prog. Nucleic Acid Res. Mol. Biol.* **69**, 1–46 (2001).
69. Vasquez, K. M. & Glazer, P. M. Triplex-forming oligonucleotides: principles and applications. *Q. Rev. Biophys.* **35**, 89–107 (2002).
70. Sczyrba, A., Kruger, J., Mersch, H., Kurtz, S. & Giegerich, R. RNA-related tools on the Bielefeld Bioinformatics Server. *Nucleic Acids Res.* **31**, 3767–3770 (2003).
71. Sadakane, K. & Shibuya, T. Indexing huge genome sequences for solving various problems. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 175–183 (2001).
72. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).
73. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357(1969).
74. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
75. Cavalier-Smith, T. Intron phylogeny: a new hypothesis. *Trends Genet.* **7**, 145–148 (1991).
76. Lynch, M. & Richardson, A. O. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **12**, 701–710 (2002).
77. Taft, R. J. & Mattick, J. S. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *arXiv Preprint Archive* [online], <http://www.arxiv.org/abs/q-bio.GN/0401020> (2003).
78. Conway Morris, S. The Cambrian 'explosion': slow-fuse or megatonnage? *Proc. Natl Acad. Sci. USA* **97**, 4426–4429 (2000).
79. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
80. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).

Acknowledgements

I thank the members of my research group, collaborators and colleagues for stimulating discussions and for their input into different aspects of this work.

Competing interests statement

The author declines to provide information about competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
CFTR | *IGF2* | *lin-4* | *let-7* | *SIM2*

OMIM: http://www.ncbi.nlm.nih.gov/Omim
autism | B-cell lymphoma | cartilage-hair hypoplasia | DiGeorge syndrome | lung cancer | prostate cancer | schizophrenia | spinocerebellar ataxia type 8

FURTHER INFORMATION

University of Queensland Institute for Molecular Bioscience: http://www.imb.uq.edu.au/index.html?id=11681

Access to this interactive links box is free online.

ONLINE CORRESPONDENCE

Nature Reviews Genetics publishes items of correspondence online. Such contributions are published at the discretion of the Editors and are subject to peer review. Correspondence should be a scholarly comment on a specific Review or Perspective article that has been published in the journal. To view correspondence, please go to our home page at http://www.nature.com/reviews/genetics and select the link to New correspondence, or, alternatively, go to the archived correspondence at http://www.nature.com/nrg/correspondence.

The following correspondence has recently been published:

Introgression from genetically modified crops to wild populations: getting the details right
By Norman C. Ellstrand

Reply:
By C. Neal Stewart Jr, Matthew D. Halfhill and Suzanne I. Warwick
This correspondence relates to the article:

TRANSGENE INTROGRESSION FROM GENETICALLY MODIFIED CROPS TO THEIR WILD RELATIVES

C. Neal Stewart Jr, Matthew D. Halfhill and Suzanne I. Warwick

Nature Reviews Genetics **4**, 806–817 (2003)

T-loop formation and abrupt telomere shortening
By Ivica Rubelj

Reply:
By Arthur J. Lustig
This correspondence relates to the article:

CLUES TO CATASTROPHIC TELOMERE LOSS IN MAMMALS FROM YEAST TELOMERE RAPID DELETION

Arthur J. Lustig

Nature Reviews Genetics **4**, 916–923 (2003)