The central role of RNA in human development and cognition

John S. Mattick

Institute for Molecular Bioscience, University of Queensland, Brisbane 4072, Australia

A R T I C L E   I N F O

Article history:
Received 29 April 2011
Accepted 3 May 2011
Available online 6 May 2011

Edited by Sergio Papa, Gianfranco Gilardi and Wilhelm Just

Keywords:
Non-coding RNA
Evolution
Gene regulation
Epigenesis
RNA editing
Retrotransposition

A B S T R A C T

It appears that the genetic programming of humans and other complex organisms has been misunderstood for the past 50 years, due to the assumption that most genetic information is transacted by proteins. However, the human genome contains only about 20,000 protein-coding genes, similar in number and with largely orthologous functions as those in nematodes that have only 1000 somatic cells. By contrast, the extent of non-protein-coding DNA increases with increasing complexity, reaching 98.8% in humans. The majority of these sequences are dynamically transcribed, mainly into non-protein-coding RNAs, with tens if not hundreds of thousands that show specific expression patterns and subcellular locations, as well as many classes of small regulatory RNAs. The emerging evidence indicates that these RNAs control the epigenetic states that underpin development, and that many are dysregulated in cancer and other complex diseases. Moreover, it appears that animals, particularly primates, have evolved plasticity in these RNA regulatory systems, especially in the brain. Thus, it appears that what was dismissed as ‘junk’ because it was not understood holds the key to understanding human evolution, development, and cognition.

@ 2011 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

The purpose of this article is to make the case that the genomic programming of humans and other complex organisms has been misunderstood, because of the assumption that most genetic information is transacted by proteins. This assumption derived from the early studies of the lac operon in Escherichia coli, and from the ensuing common interpretation of the central dogma, i.e., that information mostly flows from DNA through the temporary intermediate of RNA, which is then translated into proteins that effect all of the major structural, catalytic and (notably, for the purposes of this discussion) regulatory functions of the cell. While it has always been clear that some RNAs are end-point gene products in themselves, and Crick recognized this in the central dogma [1], this has traditionally been thought to be limited to infrastructural RNAs such as ribosomal, transfer, spliceosomal and small nucleolar RNAs, involved directly or indirectly in protein expression and other core cellular functions.

The protein-centric view of molecular genetics and cell biology, extended later into developmental biology, has deep roots, dating back to the early biochemical studies, even prior to the elucidation of the double helical structure of DNA in 1953. It was fashioned in a mechanical age that had little appreciation that genetic information may be transacted (as opposed to inherited) in other ways, although the use of codes and the code-breaking successes of World War II, from Morse code through to the Enigma machine, led to the ready acceptance of the concept of the ‘genetic code’, at least insofar as it applied to protein-coding sequences, and later to cis-acting sites in DNA and RNA recognized by regulatory proteins.

This protein-centric view became entrenched within the first few years of molecular biology, but was not without its contemporary challengers. The Nobel Laureate Barbara McClintock, who was celebrated for her insight that transposons are ‘controlling elements’ in corn, and was possibly the most original thinker in the early years of molecular biology, wrote in 1950 [2]: “Are we letting a philosophy of the [protein-coding] gene control [our] reasoning? What, then, is the philosophy of the gene? Is it a valid philosophy? . . . When one starts to question the reasoning behind the origin of the present notion of the gene (held by most geneticists), the opportunity for questioning its validity becomes apparent.” It seems few heeded her admonition, especially as her insights were apparently discredited in the eyes of others by her promotion of the idea that ‘controlling elements’ are the key to understanding development.

A few others also kept an open mind. François Jacob and Jacques Monod, who received the Nobel Prize for their work on the lac operon, mooted the possibility that the lac repressor, which they had identified genetically but not biochemically, might be an RNA [3]. However, the idea was lost when it was subsequently shown that the lac repressor was a protein [4], which undoubtedly was an important factor in the entrenchment of the protein-centric orthodoxy, especially as it pertained to gene regulation and the subsequent rise of the concept of ‘transcription factors’.

E-mail address: j.mattick@imb.uq.edu.au

0014-5793/$36.00 © 2011 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2011.05.001
Occasionally related ideas emerged. Most notably, in the late 1960s, Roy Britten and Eric Davidson, using renaturation hybridization kinetics to study the sequence complexity of DNA and RNA, noted that the extent of genomic DNA broadly increased with developmental complexity, and made two unexpected observations, specifically (i) that the population of ‘heterogenous nuclear RNA’ (hnRNA) is far more complex than messenger RNA (mRNA), and (ii) that a substantial proportion of the genome is comprised of low complexity/high copy number ‘repetitive’ sequences, some of which at least are differentially expressed at different developmental stages. This led them to propose that there may be considerable regulatory RNA in the nucleus of eukaryotic cells, and that repetitive sequences later found to be transposon-derived may comprise parts of regulatory networks [5,6]. They also predicted that many of these putative regulatory RNA would be chromatin-associated, which has now been shown to be the case (see below). Unfortunately, these ideas could not be tested at the time and, although their papers had been highly cited, they appeared to have been not well received, or ignored, by the mainstream, similar to McClintock’s experience. In any case, and somewhat surprisingly, these ideas were not re-visited later, when the discovery of introns explained, at least in part, the nature of hnRNA, and the genome projects later revealed the full repertoire of transposon-derived sequences.

2. The great surprises

This protein-centric conceptual framework imbues almost every aspect of the ‘philosophy’ (as McClintock put it) of molecular biology, and has persisted until the present, despite a number of subsequent surprises that, like Britten and Davidson’s observations, should have given serious pause for thought, and that collectively paint an entirely different picture.

The first of these was the discovery in late 1977 that most protein-coding genes in mammals and other complex organisms are not co-linear with their encoded products, but are mosaics of small segments of protein-coding sequences (‘exons’) interspersed with often vast tracts of non-protein-coding sequences (‘intervening sequences’ or ‘introns’) [7,8]. Without question, this was then and still remains the biggest surprise in the history of molecular biology [9]. However, within a very short period it was universally concluded that, because they did not code for protein, and despite the fact that they are transcribed into RNA, these sequences are mostly non-functional evolutionary debris [10,11], which was then rationalized as the likely hangover of the early evolution/assembly of genes [12,13] and/or retrotransposition of ‘selfish DNA’ (see below). The fact that these sequences were excised (and apparently discarded) and a contiguous mRNA assembled by splicing allowed the community to breathe easy, as the core flow of information from DNA to protein was left operationally undisturbed. There is, of course, another far more interesting interpretation, with potentially profound consequences – that the excised RNAs are also transmitting information, which would mean that the system, at least in the higher eukaryotes, is far more complex and sophisticated than conventionally thought, and by logical extension that there is a hidden layer of RNA regulatory networks that might be important, even central, to developmental ontogeny [14].

The second surprise, extending from the earlier work of Britten and Davidson, was that the genomes of humans and other complex eukaryotes are full of transposon-derived sequences of various classes, pejoratively referred to as ‘repetitive elements’ or more simply ‘repeats’. Again, on the same logic (that they do not encode proteins, except occasionally for their own mobilization) and despite McClintock’s insights, these sequences were quickly assumed to be mostly non-functional and their presence rationalized as ‘selfish DNA’ [15,16], part of the emerging view of most of the human genome as an evolutionary graveyard. The vast tracts of intergenic and intronic sequences in the human genome, riddled with transposon-derived sequences, have been described, among other things, as ‘junk . . . in the attic’, on the proposition that some have been exapted for function but many or most only have potential future value [17], although the proportion that may have been exapted for function was and is still unknown. Again, of course, there is a more interesting and potentially more profound explanation.

The raw material for evolution is duplication and transposition, with the latter providing far more flexibility in terms of dissemination of functional cassettes and re-structuring of regulatory networks for phenotypic divergence. This appears to be the major place that evolution has played in relation to the radiation of animals, which generally have a similar proteome (see next). Nonetheless, and again despite McClintock’s insights, on the dubious and entirely circular assumption of the non-functionality of the transposon-derived ‘ancient repeats’ (ARs) that are common to the genomes of humans and other mammals (which have persisted with us for tens if not hundreds of millions of years), and the independent but equally dubious assumption that the recognizable ex- tant population of ARs is representative of the original distribution, not simply its more conserved end [18], it has been estimated that only ~5% of the human genome is under purifying selection [19,20] (although other analyses, which do not rely on this assumption, have indicated that evolutionary selection is widespread across the genome [21]). That is, the assumed non-functionality of a small proportion of the genome has been used to impute the non-functionality of most of the remainder, because they are evolving at similar rates. If either this or the other assumption mentioned above is wrong, then so is the conclusion [18].

The third great surprise, which was entirely contrary to expectations, is that the number and repertoire of protein-coding genes remains relatively static across the metazoan lineage, despite enormous increases in developmental and cognitive complexity [22]. The simple nematode Caenorhabditis elegans, which only has ~1000 somatic cells, has almost 20,000 protein-coding genes, similar to that in the human [23–25]. However, humans have approximately 10^{14} cells, sculpted into a myriad of different muscles, bones and organs that have complex architectures, as well as a brain with approximately 10^{11} neurons [26]. Moreover, despite some interesting expansions and innovations, such as RNA editing in vertebrates (see below), the majority of these genes are orthologous (i.e., have similar functions), even in sponge, the most primitive metazoan [27], including most of those involved in the cell signaling and homeotic pathways that underpin multicellular development. That is, all animals have a similar protein toolkit [28], and therefore the relevant information that programs progressively more complex organisms must lie elsewhere in the genome, presumably in an expanded regulatory architecture.

3. Scaling of regulatory architecture

The response to this lack of gene scaling and limited proteomic diversification has been to assert the power of combinatoric control – specifically that the power of combinatorial control by transcription factors (and other regulatory proteins, such as those involved in alternative splicing), will lead to “a dramatic expansion in regulatory complexity” [29]. As a logical extension, it is then simply proposed that the developmental programming of more complex organisms has been enabled by an expansion in the numbers and complexity of the cis-acting sequences recognized by these regulatory proteins, alteration of which, along with alterations in the expression of the regulatory proteins and the networks
in which they participate, lies at the heart of phenotypic diversity [29,30].

While there is little doubt that regulatory alterations and expansions underpin the emergence and divergence of more developmentally complex organisms, the essential argument here (although not spelled out explicitly in mathematical terms) is that the range of regulatory options scales factorially with the numbers of regulatory proteins, and that in animals this number is so great (>1000 regulatory proteins in human or C. elegans [29] = potentially >1000! combinations = a number far bigger than the estimated number of atoms in the universe) as to be superficially more than capable, even if heavily discounted, of providing sufficient regulatory complexity to program human development, so there is no need for further concern or further justification. The necessary power is implicit in the assumption. However, while it is clear that many factors can influence a decision to transcribe a gene, so there is some sort of ‘combinatorial’ control, it is by no means clear that this scales factorially. Indeed, the implied assertion has not only never been clearly articulated, but it has also (consequently) never been justified, mathematically, by reference to decision theory, or mechanistically; nor has it been subjected to critical scrutiny, as the lack of articulation obscures the issue, and the assertion fits comfortably with mainstream preconceptions.

How do regulatory factors really scale with gene number? This is a difficult question to answer, but the available evidence suggests that it is quite the opposite to that which is assumed. Prokaryotic genomes are predominantly composed of protein-coding sequences, and therefore it is possible to do a first approximation analysis of the relationship between the numbers of genes encoding regulatory factors and the total numbers of genes in cells of different genetic complexity (despite the existence of a limited set of regulatory RNAs in these organisms). This is not possible with eukaryotes, which contain indeterminate but apparently large numbers of regulatory RNAs, both large and small (see below). In prokaryotes, however, it is clear that the number of regulatory genes R (those encoding proteins with characteristic motifs such as DNA-binding domains) scales quadratically with gene number N (R ∝ N^2, not as some sort of inverse factorial R ∝ 1/N!), over the entire range of bacterial genome sizes [31–33]. This empirical fact has many implications. Since regulatory genes are scaling twice as fast as total genes, and there is no hint of any deviation at the top end of the range, there must be a limit (as the number of regulatory genes cannot exceed the total), which is ostensibly the observed upper limit of bacterial genome sizes (about 10 Mb, ~9000 genes), where over 20% of the genes are regulatory. This limit was likely reached early in evolution.

The quadratic relationship and inferred limit also implies that higher complexity eukaryotes must have solved this problem, in all likelihood by moving to a more genomically efficient (and evolutionarily flexible) RNA-based regulatory system (that separates regulatory signal from consequent action, as exemplified by RNA interference pathways [34,35]), together with the introduction of other levels of control and compartmentalization, all hallmarks of eukaryotes, to mitigate the global scaling problem. Finally, this relationship implies that combinatorial action of regulatory factors, as envisaged to allow dramatic expansions in regulatory capacity, does not hold, at least in prokaryotes (as the scaling of regulatory factors would be entirely different) and therefore also probably not in eukaryotes, unless one can conceive of chemical space (specifically protein–protein interaction space) that is permitted in eukaryotes but not prokaryotes. If so, the clear implication is that in all cells, and indeed in all functionally integrated complex systems, the proportion of regulatory information increases with increasing complexity, i.e., occupies a progressively greater amount of the total information required to program the ontogeny and operation of the system [33].

Few would disagree that increased developmental complexity requires an expansion of regulatory information and that this information resides mainly in the non-protein-coding portion of the genome. As noted already, most molecular biologists have assumed that this information is (largely) restricted to cis-acting regulatory sequences that interact with ‘regulatory’ proteins to control gene expression, and, since it is inconceivable that such sequences could cover a huge fraction of the human genome, have been generally comfortable with the proposition that the majority of the genome is non-functional. However, while the number of protein-coding genes and the extent of protein-coding sequences remains surprisingly static, it is clear that the extent of non-protein-coding intronic and intergenic sequences does scale with increased developmental complexity [22,36] (Fig. 1), and indeed is the only variable yet demonstrated to do so (along with the complement of regulatory RNAs, see below). This does not prove anything, but is at least consistent with the proposition that high developmental complexity requires a vastly expanded regulatory architecture, which can only be falsified by a downward exception, i.e., a complex organism that has much less non-coding sequence that demonstrably simpler ones (as opposed to an upward outlier that has more non-coding sequence than organisms of equivalent complexity), which so far has not been observed. Here it is important to note that Fugu rubripes and Arabidopsis thaliana, which possess the smallest known vertebrate and angiosperm genomes, respectively, do not deviate substantially from the broad relationship described above [22].

4. Pervasive transcription of the genome

The fourth great surprise of genomic analyses has been that, irrespective of the extent of non-protein-coding sequences in different genomes, the vast majority of these sequences are transcribed, apparently in a developmentally regulated manner, mainly into non-protein-coding RNAs (ncRNAs). A variety of transcriptomic studies, primarily using high throughput sequence analyses of normalized cDNA libraries [37–42] and cDNA interrogation of genome tiling arrays [43–48], have shown that the mammalian transcriptome is amazingly complex and consists of a myriad of overlapping and interlacing transcripts from both strands, including intronic, antisense and intergenic transcripts that exhibit different start sites, termination points, splicing and expression patterns in different cells [49–52].

A frequent response to these observations, especially as many of these unexpected ncRNAs appear to be expressed at relatively low levels, has been to suspect or assert that they are ‘transcriptional noise’, an alternative hypothesis with the appeal of not disturbing the prevailing orthodoxy of gene regulation. Despite the substantial evidence that genomic sequences specifying these ncRNAs have all of the hallmarks of conventional genes and the rapidly accumulating numbers of functionally validated examples ([53], see below), the debate about the relevance of these transcripts continues. A recent article, for example, that compared the signals derived from transcriptomic interrogation of genome tiling arrays with next generation short read sequencing data, concluded that the former had significant problems with false positives, that many of the short sequence ‘singleton’ tags in the latter were so scattered as to appear “of random character” (i.e., consistent with noise) and that much of the genome is “not pervasively transcribed” [54].

However, apart from problems with the methodology employed, the data are equally consistent with low level sampling of rare transcripts (that may, e.g., be cell type-specific), considering that such deep sequencing suffers from diminishing sensitivity for uncommon transcripts by the dominance of common highly expressed (usually protein-coding) RNAs [55], especially in the brain.
where much of the analysis was done [54]. Indeed some ncRNAs are easily detectable by the relatively insensitive technique of in situ hybridization in particular subregions of the brain, such as the dentate gyrus [56] (Fig. 2), which occupies a tiny fraction of the brain and would be difficult to detect in whole brain transcriptomic analysis, even by ‘deep’ sequencing. Moreover, a new technique called RNA Capture-Seq, which combines array capture of cDNAs with deep sequencing to increase the sensitivity of transcriptomic interrogation of targeted loci, has revealed that so-called ‘gene deserts’ actually express a wide range of spliced transcripts, many of which were not detected by a single tag in pre-capture libraries, as well as previously unknown isoforms of intensively studied protein-coding loci such as p53 [57].

5. A world of long non-coding RNAs

There are tens if not hundreds of thousands of long antisense, intergenic and intronic ncRNAs (lncRNAs) expressed from mammalian genomes [41,57], with abundant evidence of their involvement in eukaryotic cell and developmental biology (for recent review see [58]). A large fraction of lncRNAs is expressed in the brain [56,59–61]. Many lncRNAs are also derived from enhancers [62–65], enigmatic non-coding regulatory elements that act at a distance to control gene expression during development, and which are thought to act by recruiting transcription factors and inducing chromosomal looping to bring these factors into contact with target promoters [66]. Reciprocally some lncRNAs emulate the properties of enhancers [67], which suggests that enhancer action may integrally involve a derived RNA [68], possibly to mediate higher order chromatin interactions and/or epigenetic changes [69,70]. These effects may explain the equally enigmatic genetic phenomena of transvection [68,69] and transinduction [63], the latter (along with clear evidence of selection on synonymous codon sites [71]) suggesting that even mRNAs may have embedded regulatory functions in addition to their protein-coding capacity.

It has also recently become evident that conventional protein-coding loci may also produce both small and large non-coding RNAs, by regulated post-transcriptional cleavage of mRNAs [72,73], including within 3’ untranslated regions (3’UTRs) [74,75]. The latter can be expressed in a highly cell-specific manner (e.g., in the cortex and hippocampus in the brain, or Sertoli cells in the testis) [75] (Fig. 3), with genetic evidence dating back many years showing that these sequences can transmit information in trans separately from their normally associated protein-coding sequences and independently of their normal cis-acting functions in the regulation of mRNA translation and stability. This includes the restoration of the oogenesis defect in Drosophila oskar mutants [76] and the inhibition of cell division, suppression of malignancy and induction of differentiation by ectopically expressed 3’UTRs from a variety of genes in mammalian cells [77–81].

In addition, genome tiling array-based transcriptomic studies, which do not require ribosomal RNA depletion by oligo(dT) purification, showed that almost half of the transcripts found in human cells are not polyadenylated, and are largely of a different sequence composition from the RNA polymerase II-derived polyA+ fraction [46]. These transcripts are possibly synthesized by RNA polymerase III [82], indicating that for a generation a large portion of the transcriptome has been hidden from view for technical reasons. The same applies to repetitive sequences, which are frequently masked out of such analyses, but for which there is increasing evidence of differential expression [83–86], dating back many years [87,88].

6. Functions of lncRNAs

Although the vast majority of lncRNAs await characterization, there are compelling genome-wide indices of their functionality [53], as indicated by:
(i) Conservation of their promoters, splice junctions, exons, predicted structures, genomic position and expression patterns [41, 89–100].

(ii) Dynamic expression and alternative splicing during differentiation [94, 95, 101–103].

(iii) Altered expression or splicing patterns in cancer and other diseases [103–117].

(iv) Association with particular chromatin signatures that are indicative of actively transcribed genes [94, 95].

(v) Regulation of their expression by key morphogens, transcription factors and hormones [94, 95, 116, 118, 119] and

(vi) Tissue- and cell-specific expression patterns and subcellular localization [59, 102, 116, 120–130].
Indeed, different but substantial subsets of all of polled IncRNAs are differentially and dynamically expressed in all differentiation or disease systems that have been examined. These include the differentiation of embryonic stem cells [94], neuronal cells [131], muscle cells [102], T cells [132], breast epithelial cells [103] and cancer [117]. A survey of the expression of over 1300 IncRNAs in mouse brain showed that over 600 were expressed in highly specific locations, such as different regions of the hippocampus, different layers of the cortex, or different parts of the cerebellum (Fig. 2), with most (where the resolution was sufficient to determine) showing specific subcellular locations [56]. Since lack of conservation is uninformative [18,89], the only reliable index of function, although by no means proof of function, is differential expression [18]. By this criterion, most of the genome is imputed to be functional.

Many IncRNAs are associated with particular subcellular structures [133], including novel subnuclear domains in a subset of neurons [125] or Purkinje cells [56], and paraspeckles, which are as yet not well understood mammal-specific subnuclear domains involved in the retention of edited transcripts (containing Alu elements, see below) and induced in differentiated cells [102, 126–129,134–136]. What feature of mammalian differentiation and development that involves these structures (and is not required in other vertebrates) is not known, but is an intriguing question, especially in view of the recent report that mice lacking paraspeckles appear superficially normal [137]. Preliminary evidence suggests that paraspeckles may have a role in regulating the nuclear-cytoplasmic shuttling of RNAs that are subject to RNA editing [134,138], which appears to be associated with the rise of cognition (see below), and may be a feature not only of mammalian brain function, but also other aspects of mammalian reproduction, development and physiology associated with the exchange of high reproductive rates for extended nurturing, presumably with a net advantage for survival and reproductive fitness.

Although very few have yet been studied, there are rapidly increasing numbers of reports showing that IncRNAs have biological function [139], usually using siRNA-mediated knockdown and/or ectopic expression in cell-based assays (for recent compilations see [53,140]). These include functions, for example, in the modulation of the reprogramming of induced pluripotent stem cells [141], regulation of homeotic gene expression [101], cancer metastasis [117,142], breast development [103,143], retinal development [144,145], parental imprinting [146–151], X-chromosome dosage compensation [152–157], paraspeckle assembly [102,126–129] and p53 regulation [158–160].

In some cases, mechanistic insights into the mode of action of IncRNAs are emerging. For example, the nuclear IncRNA Malat1 (‘metastasis-associated lung adenocarcinoma transcript 1’) is expressed in numerous tissues and is highly abundant in neurons, where it has been shown to play a role in synaptogenesis [161] and to regulate alternative splicing by interacting with SR family pre-mRNA-splicing factors [161,162] and other nuclear RNA binding proteins [163]. It also yields a tRNA-like cytoplasmic RNA by 3’ end processing, although the function of this small RNA is unknown [164].

Most functionally characterized IncRNAs appear to play a role in differentiation and development, which is the broad prediction from the presence and progressive expansion of these transcripts in developmentally complex organisms. Consistent with this, a major, if not the major, function of these IncRNAs appears to be the regulation of epigenetic processes.

7. RNA regulation of epigenetic processes

Epigenetic processes are central to differentiation and development [165–170], long-term responses to environmental variables [171,172], and brain function [173–179]. Epigenetic information is encoded by the methylation [180] and hydroxymethylation [181,182] of cytosines in DNA, and in a wide range of modifications of the histones that package DNA into nucleosomes [183]. These are catalyzed by a suite of ~60 generic enzymes/chromatin modifying complexes that impose a myriad of different chemical marks at hundreds of thousands of genomic locations in different cells at different stages of differentiation [183].

What determines the site-selectivity of these chromatin remodeling complexes, how is the position of nucleosomes regulated, and what is the molecular basis of environment–epigenome interactions? The likely answer to all of these questions is RNA [184–186]. It had been thought that these processes were directed by transcription factors, which clearly can exert powerful effects on cell state, capable of (re-)programming stem cells [187] or forcing differentiation into e.g., myoblasts [188]. However, the enormous and underappreciated challenge for genetic programming is not simply to define the phenotypic state of a cell, or to wind it back or forward, but rather to organize the 4-dimensional growth and differentiation of cells into a myriad of precisely 3-and 4-dimensionally sculpted organs and tissues [189] – different vertebrae or muscles, for example, have a unique architecture tuned to their functional position. While many proteins, including transcription factors and homeotic proteins, which are often differentially expressed and
state-specific (see e.g., [190,191]), have a role in this process, these clearly also require additional information for their site-specificity (as only a subset of potential sites are actually addressed in any given context). This additional information increasingly appears to be embedded in regulatory RNAs that control chromatin architecture by directing effector ‘regulatory’ proteins to their sites of action [184] (Fig. 4).

The evidence implicating RNAs in the epigenetic control of chromatin architecture and transcription dates back many years [69,184,192–194] and includes the physical association of RNA with chromatin [195–198] and chromatin regulatory proteins such as DNA methyltransferases and REST [199,200], the presence of RNA binding domains in proteins in chromatin remodeling complexes [201–206] and the observations that major classes of transcription factors, such as zinc finger and Y-box proteins, have RNA or RNA:DNA binding activity [207–209]. The direct evidence of IncRNA involvement in epigenetic processes includes the association of subsets of IncRNAs with Trithorax-group proteins and forms of modified histones in active chromatin [94,210], or with Polycomb-group proteins and histone modifications associated with repressed chromatin [101,150,151,198,200,211–217], which also appears to involve elements of the RNA interference pathway [218–220].

Apart from IncRNAs that regulate allele-specific expression at imprinted loci and sex chromosomes, a number of interesting examples have emerged over the past few years. These include the epigenetic regulation of the tumour suppressor gene p15 by its antisense RNA [217,221], called ANRIL, independently shown to be associated with susceptibility to range of complex diseases, including coronary disease, intracranial aneurysm, type 2 diabetes, gliomas and basal cell carcinomas, among others, with preliminary evidence suggesting that the common factor is its role in the regulation of cell proliferation and senescence [222]. Indeed, most genetic variants associated with complex diseases are non-coding [223], presumably regulatory, many if not most of which may affect intergenic, intronic and antisense IncRNAs [53]. Another that has been well studied (by IncRNA standards) is HOTAIR, a relatively rapidly evolving mammal-specific transcript [224], which is derived from the HOXC locus and was first identified among 231 HOX-associated IncRNAs that are differentially expressed along spatiotemporal axes during development [101]. It represses transcription at the HOXD locus in trans [101], is up-regulated in breast cancer and increases cancer invasiveness and metastasis [142], which has also been observed with another IncRNA SPRY4-IN1, which is up-regulated in melanoma [117].

These examples are undoubtedly the tip of a very big iceberg, especially when one considers that, in all likelihood, most of the 10^14 cells in a human, like the 10^5 in C. elegans, will have a unique ontogeny and positional identity (a proposition that is supported by the phenotypic similarity of monozygotic twins) that is epigenetically controlled by such regulatory RNAs.

8. A world of small RNAs

There are also many classes of small RNAs that have been discovered in recent years. The best characterized is microRNAs (miRNAs), ~23 nt small RNAs derived from short hairpin precursors that generally appear to control mRNA translation and stability via their (usually imperfect) recognition of target sites usually in the 3’UTR of mRNAs and interaction with Argonaute-containing RNA-induced silencing complexes (RISCs) [225]. Approximately 1000 miRNAs have been identified in human, but there are likely to be many more, especially if significant numbers are lineage- or species-restricted [226] and/or, like lsy6 in C. elegans [227], are cell-specific.

The repertoire of miRNAs, like IncRNAs, has expanded during animal evolution [228–231]. They regulate most mRNAs [232,233] (and possibly IncRNAs), appear to influence almost every facet of animal and plant development [234–242], as well as many aspects of brain function (see e.g., [243–246]), and are often dysregulated in cancer and other complex diseases [247,248]. There is much more to be learned about their biology and functional mechanisms, including the parameters that determine target recognition [249–252], especially in view of the observations that miRNA isoforms are developmentally regulated [253] and that at least some miRNAs are localized in the nucleus [254,255].

Similar small RNAs, termed siRNAs (small interfering RNAs), derived from longer hairpin precursors or duplex RNAs, also act through the RNA interference pathway, usually by perfect matches with the target sequence, resulting in mRNA degradation [35] as well as transcriptional silencing via epigenetic effects on target sequences [184,219,256–258], with as yet not well understood interplay between different facets of the RNA interference system [35]. There appear to be a number of pathways as well as intercellular transport of small RNA signals involved, at least in plants [259], and possibly also in animals [260].

Related animal-specific small (~24–30 nt) RNAs that also involve the RNA interference system, piRNAs (piwi-interacting RNAs) [261–263], are involved in the silencing of transposons, primarily in the germine [264–267]. PiRNAs are 2’-O-methylated at their 3’ ends, which permits their specific recognition by the PAZ domain of specific Argonautes that are expressed in the germine [268,269]. They appear to be involved in genome defense [267,270,271], although another more interesting possibility is that they have evolved to regulate retrotransposon expression in more subtle ways during early development [256,272–276]. These RNAs are also linked to epigenetic changes (DNA methylation and histone modifications) [277–279], assembly of the telomere protection complex [273], and complex genetic phenomena such as hybrid dysgenesis [280], position effect variegation [256] and trans-silencing [281].

Another class of small RNAs, termed small nucleolar RNAs (snoRNAs), which may be spread by retrotransposition [282–284], are derived from the introns of protein-coding and non-coding host transcripts [285–288], with at least two cases of the latter having independent functions as IncRNAs [103,289]. SnoRNAs guide specialized protein complexes to impart sequence-specific 2’-O-methylation (box C/D snoRNAs) or the isomerization of specific uridines to pseudouridines (box H/ACA snoRNAs) in target RNAs, and are usually localized in the nucleolus, with a related class of box H/ACA snoRNAs found in (nuclear) Cajal-bodies [290–293] using a specific localization signal that is also found in telomerase RNA [294]. SnRNs were initially thought simply to modify ribosomal RNAs to tune their chemical properties in translation [295], but some show imprinted, tissue-specific and/or context-dependent expression, especially in the brain [296–300] and target other cellular RNAs, including small nuclear (spliceosomal) RNAs (snRNAs), transfer RNAs (tRNAs) and possibly even mRNAs, in ways and for purposes that are not yet well-defined or understood [291]. RNA can also be modified by cytosine methylation. Dnmt2, named because of its homology to DNA methyltransferases, is in fact an RNA methyltransferase [301,302], which, although its target spectrum is not known, plays a role in the development of the brain and other organs [303] and is required for retrotransposon silencing in somatic cells of Drosophila [304]. Recently, it has been shown that snoRNAs, snRNAs and tRNAs are cleaved at specific positions to produce smaller RNAs [305–307]. Most, if not all, small nucleolar RNAs in eukaryotes, from fission yeast to human, are processed to three different sub-species of small RNAs of three different canonical sizes (~17–19 nt and ~30 nt from box/C/D snoRNAs, and ~20–24 nt from box
H/ACA snoRNAs), which show altered expression patterns in mutants affecting the RNA interference pathway [306]. The size of the box H/ACA-derived small RNAs is similar to miRNAs, and some have been shown to function as miRNAs [308,309]. A number of annotated miRNA precursors have box H/ACA snoRNA features [310] and some miRNAs have a nucleolar location [254]. Small RNAs derived from box C/D snoRNAs have been implicated in the regulation of splicing [311], whereas those derived from tRNAs have been found to be preferentially associated with different Argonauts, and to affect the global regulation of the RNA silencing system [307]. These findings all point to a complex evolutionary and functional interplay between different classes of small RNAs in the regulation of gene expression, whose dimensions have barely begun to be explored.

Analysis of deep sequencing small RNA datasets identified a new class of animal-specific nuclear-localized tiny (~18 nt) RNAs derived just downstream from transcription initiation sites (tiRNAs) [255,312]. Their size and position suggest that tiRNAs may be produced by RNA polymerase II interaction with the first downstream nucleosome and backtracking [312,313], a well-established phenomenon that involves TFIIIS-mediated cleavage of a small RNA from the 3’ end of the nascent transcript prior to resumption of synthesis [314]. This possibility was given strong support by the finding that the position of tiRNAs is different in human and Drosophila, reflecting the different position of the first nucleosome downstream of the transcription start site [313] (Fig. 5), with the implication that tiRNAs may be marking, and perhaps regulating, the position of the nucleosome [313].

While it has been known for some time that there is a nucleosome-free region around transcription start sites, with periodicity elsewhere in nucleosome spacing [315–319], it came as a significant surprise to discover recently that nucleosomes are in fact preferentially positioned at exons in somatic cells [320–324] and germ cells [322], the latter implying trans-generational epigenetic inheritance, for which there is increasing evidence [325–327] and which may provide selective advantage to cope with environmental challenges [328]. Presumably there is also some logic to the positioning of nucleosomes in intronic and intergenic regions, which also produce many functional (non-coding) RNAs, although this has yet to be established. In any case, this finding shows that chromatin is far more organized than previously suspected, and that the position of nucleosomes must be regulated by some active or passive mechanism(s). It also provides an explanation for the observed coupling of chromatin structure, transcription and splicing [329], and a potential basis for exon selection through various histone modifications within these nucleosomes that report the status of particular exons during differentiation and development [323] – i.e., that alternative splicing may be controlled by histone modifications [320–324], a prediction that has since been confirmed, at least in part [330,331].

Subsequent deep sequencing of nuclear small RNAs identified another class of small RNAs positioned at splice sites (spliRNAs) [255] (Fig. 6). These spliRNAs are similar in size (17–18 nt) to tiRNAs and, like tiRNAs, are only found in animals. This suggests that they may have a similar ontogeny and function, but spliRNAs are derived from the 3’ end of exons [255], whereas tiRNAs are positioned on the 5’ side (upstream) of the first nucleosome [313]. However, there is (at least) one form of modified histone (H3K36me3) that has a different position, which is associated with actively expressed genes [332] and which appears to be positioned at the exon–intron boundary [333], or just downstream of it [323]. This would (in principle) place spliRNAs on the 5’ side of these nucleosomes, like tiRNAs, a possibility that we are currently testing by deep sequencing on small RNAs associated with different forms of modified nucleosomes. It is worth noting that individual tiRNAs and spliRNAs are only present at very low levels in deep sequenc-

ing datasets [255,312]; neither would have been identified without the orthogonal intersection of small RNA datasets with specific genomic features (transcription start sites and exon–intron junctions, respectively), which also suggests that there may be many more locus-specific regulatory RNAs to be discovered.

9. Transcriptomic, epigenetic and genomic plasticity in gene-environment interactions and brain function

RNA also appears to be the substrate for environmental–epigenome interactions. There is emerging evidence that RNA is subject to a great deal of context-dependent editing, especially in the brain. RNA editing involves base deamination (as distinct from snoRNA-mediated modification) and is catalyzed by two classes of enzymes in animals: ADARs (adenosine deaminases that act on RNA) change adenosine to inosine (A > I) [334,335], which behaves similarly to guanosine (e.g., in sequencing protocols), but has different base pairing qualities; and APOBECs (named after ‘ApoB editing complex’, see below), which are vertebrate-specific and change cytosine to uracil (C > U) and may act on RNA or DNA [336,337].

There are three ADAR orthologs in animals. ADAR1 and ADAR2 occur in invertebrates and vertebrates [338], and are expressed in most tissues, but particularly highly expressed in the nervous system [334,335,339]. Loss of these genes in mice is embryonic or postnatally lethal [340,341]. ADAR3 is vertebrate- and brain-specific [338,342], but little is known about its function. Little is known about how RNA editing is regulated [343], but ADARs can be localized in both the nucleus and cytoplasm [344,345], and there is evidence that RNA editing activity is connected to canonical cell signaling pathways [346], implying response to external cues.

A > I editing was first discovered a generation ago by cDNA-genomic comparisons of sequences encoding important neurotransmitters, such as glutamate, GABA and serotonin receptors, where it alters the amino acid sequence, ostensibly to tune the electrophysiological properties of the synapse [334]. It has since been regarded as an interesting but somewhat idiosyncratic subfield of both molecular biology and neuroscience. However four papers published in 2004, which have attracted surprisingly little attention, showed by comparison of large scale cDNA libraries with genomic DNA that A > I editing is far more widespread than previously suspected, and occurs in thousands of transcripts [347–350].

Most of the edited sites occur in non-coding regions, implying that editing is not only modifying the structure-function properties of proteins, but also RNA-based regulatory circuits, and therefore potentially epigenetic processes, which are central to learning and brain function [173–177,179,351].

Moreover, these studies showed that there is a massive increase (~35×) in the intensity of RNA editing in humans compared to mouse. Most (>90%) of this editing occurs in primate-specific Alu sequences [347–350], which evolved from a functional RNA ancestor (the 7SL RNA of the signal recognition particle) [352,353]. Alu sequences invaded the primate lineage in three successive waves, and now comprise ~1.2 million mostly sequence unique copies that collectively occupy ~10.5% of our genome [354,355]. A subsequent study showed that the intensity of A > I editing also increased during primate evolution, and that new editable Alu insertions after the human-chimpanzee split are significantly enriched in genes related to neurological functions and neurological diseases [356]. These SINEs (short interspersed nuclear elements), consistent with the general view of such elements as junk, have long been regarded as the most recent transpositional storm to hit our lineage. However, these observations suggest a radically different and much more interesting interpretation – i.e., that such sequences, while having being recruited for many functions [353],
have flourished as modular substrates for RNA editing, permitting the introduction and spread of the transcriptomic and epigenomic plasticity necessary for epigenome–environment interactions, driven by positive selection for cognitive function [185,328,356,357].

The APOBECs are even more intriguing. They were discovered initially by their action on ApolipoproteinB mRNA where a C > U change introduces a stop codon to generate a truncated isoform of the protein in intestine versus the longer form produced in liver [336]. There are five families of APOBECs, two of which (APOBEC 1 and 3) are mammal-specific [336,337]. The best characterized is AID, which is involved in somatic rearrangements and hypermutation of immunoglobulins in the immune system [337]. AID appears to act on DNA but may be targeted by RNA [358]. Moreover, AID deaminates 5’-methylcytosine (to form thymine) [359] and is required for the reprogramming of cells to pluripotency [360], and APOBEC2 is required for muscle differentiation [361], suggesting a wider role

Fig. 5. A model of tiRNA biogenesis, and evidence that tiRNAs are linked to the position of the first nucleosome downstream of the transcription start site (TSS). (A) RNAPII generates a short nascent RNA, stalls at the first (+1) nucleosome and backtracks, and the 3’ end of the nascent RNA is cleaved by TFIIH to generate an ~18 nt tiRNA. Values next to human or fly silhouettes show the average distances from the TSS to the nucleosome or the putative position of tiRNA biogenesis. (B) Peak densities of the 5’ ends of tiRNAs in human and Drosophila are offset, consistent with positions of phased +1 nucleosomes, which are different between the two species, and the proposed model of biogenesis. Reproduced from [313] with permission.
for such enzymes in developmental processes. Interestingly, there are many parallels between the nervous and adaptive immune systems, including the presence of immunoglobulin domains in many neuronal cell surface receptors [362,363], indicating that the adaptive immune system evolved (in vertebrates) from the nervous system, and that both may use similar mechanisms to tune receptor interactions. Moreover, the existence of many unusual DNA repair enzymes, many of which appear to be linked to reverse transcriptase activity, suggests that RNA-directed DNA recoding may play a role in long-term memory formation [357].

The APOBEC3 family originated after the divergence of the marsupial and placental lineages and has greatly expanded in the primate lineage, with very strong signatures of positive selection [337,364,365]. At least some (as well as APOBEC1 [366]) appear to be involved in the control of exogenous and endogenous retrotransposition, possibly by inhibition of reverse transcription, and are therefore thought to be involved in host/genome defense [337,367–370], but why this should be particularly important in mammals and especially primates is problematic. An alternative possibility is that these enzymes (one of which, APOBEC3G) is expressed in neurons [371]) have evolved to domesticate transposition. This possibility has been given strong support by the recent observations that de novo L1 retrotransposition events occur in neural progenitor cells and may therefore contribute to individual somatic mosaicism in the brain [372,373]. Moreover, this process that appears to be regulated by Wnt signaling pathways and transcription factors known to be important in neural differentiation [374]. The process is also regulated by MeCP2 (methyl-CpG-binding protein 2), which is involved in global DNA methylation and neurodevelopmental diseases [375], and is modulated in the brain itself by environmental influences [376]. That is, transposon mobilization may not simply have played a role in genome evolution but also in real time genome dynamics that enable the extraordinary in situ evolution [377] and functional complexity of the neuronal networks in the human brain [378].

10. Concluding remarks

The emerging evidence suggests that evolution has shaped the human genome in far more sophisticated ways than ever imagined,
and that most of the information it holds is involved in complex regulatory processes that underpin development and brain function. This includes the vast numbers of non-coding RNAs and transposons, which rather than being junk, appear to provide the regulatory power and plasticity required to program our ontogeny and cognition [14, 69, 184, 185, 194, 379–381]. In this scenario, DNA might be viewed as a zip file/hard disc, proteins (including most ‘regulatory’ proteins) as the analog effectors, and RNA as the computational engine of the system [189, 382]. Moreover, it seems that might be viewed as a zip file/hard disc, proteins (including most non-protein-coding DNA and eukaryotic complexity. Bioessays 29, 288–299.


Tait, R.J. and Mattick, J.S. (2003) Differences in cooption and domestication of retrotransposons, to allow the development of higher-order information processing, learning, language and thought, the most powerful selective advantage of all.

Acknowledgements
I am extremely grateful for the inspiration, advice and contributions of the members of my research group, collaborators and many colleagues around the world, as well as the support of the Australian Research Council and National Health and Medical Research Council. I especially thank Paulo Amaral for helpful comments and corrections, and Ryan Taft for bringing the McClintock quote [2] to my attention.

References


Simon, B. et al. (2011) Recognition of 2′-O-ribose methylation of rRNAs.


