Introns: evolution and function

John S Mattick

University of Queensland, Brisbane, Australia

Current Opinion in Genetics and Development 4: 823-831 (1994)

The debate continues on the issue of whether nuclear introns were present in eukaryotic protein-coding genes from the beginning (introns-early) or invaded them later in evolution (introns-late). Recent studies concerning the location of introns with respect to gene and protein structure have been interpreted as providing strong support for both positions, but the weight of argument is clearly moving in favour of the latter. Consistent with this, there is now good evidence that introns can function as transposable elements, and that nuclear introns derived from self-splicing group II introns, which then evolved in partnership with the spliceosome. This was only made possible by the separation of transcription and translation. If introns did colonize eukaryotic genes after their divergence from prokaryotes, the original question as to the evolutionary forces that have seen these sequences flourish in the higher organisms, and their significance in eukaryotic biology, is again thrown open. I suggest that introns, once established in eukaryotic genomes, might have explored new genetic space and acquired functions which provided a positive pressure for their expansion. I further suggest that there are now two types of information produced by eukaryotic genes - mRNA and iRNA - and that this was a critical step in the development of multicellular organisms.

Introduction

The discovery of intervening sequences (introns) in 1977 took the molecular biology world by surprise [1]. The interpretation of these sequences had its roots in the presumption that a gene is synonymous with protein, which had become an article of faith following the classical studies into the *lac* operon and the 'genetic code' over the preceding two or three decades. Because introns did not code for protein, they were assumed to be non-functional, in which case their presence had to be explained by other factors, historically and evolutionarily. In 1978, Gilbert [2] suggested that introns allowed the shuffling of protein-coding sequences (exons), which would both increase genetic complexity by differential splicing and speed evolution by permitting new arrangements by recombination, implying that this was their *raison d'être*. Gilbert's hypothesis was extended by Blake [3] with the sequitur that exons would be predicted to encode relatively discrete protein structural elements such as domains or 'smaller, supersecondary structures' [3,4]. Darnell and Doolittle (see [5]) soon pointed out that there could be little short-term selective advantage in dividing previously contiguous protein-coding sequences and concluded, therefore, that the introns must have been there in the first place, that is, in the genes of the progenote. This view was supported by the developing ideas of prebiotic evolution, and of the assembly of primordial genes from cassettes of nucleotide sequence that specified peptide, domains (and that were separated by non-functional intervening sequences), which has become known as the exon theory of genes. These ideas arose largely before it was appreciated that there are different types of introns, and that RNA can have catalytic functions, but nonetheless have maintained their own momentum. However, they have also been seriously challenged by Cavalier-Smith [6] and Palmer and Logsdon [7], who have mounted compelling arguments that nuclear pre-mRNA introns in fact spread late into eukaryotic lineages (see also [8]).

Since the discovery of introns, there has been considerable study of the biochemistry of the process by which these sequences are removed from primary transcripts in the nucleus (for recent reviews, see [9-13]). It has also become clear that there are at least three other types of intervening sequences: group I introns, group II introns, and archaeal introns. Group I and II introns have significant secondary structure and can self-splice, a process that may be aided by protein factors, such as maturases, that are encoded in the intron. These introns occur in both eubacteria and eukaryotes, but have a restricted distribution, being found primarily in rRNA and tRNA genes, and in a few protein-coding genes of organelles and bacteriophages. Archael introns are quite distinct from group I and II introns and thus far have been found only in archaebacterial tRNA and rRNA genes [14*]. They have no conserved internal structure, are not self-splicing, and require protein but not *trans*-acting RNA factors [14*]. They appear to have originated quite separately, and their relationship to other types of introns remains obscure.

The position of introns with respect to protein and gene structure

A crucial prediction of the introns-early hypothesis is that introns should, by and large, delineate structural or functional units within proteins, which were originally expected to be globular [3]. As there appear to be many exceptions to this, and it is clear that many exons are too small to code for globular domains, those supporting this view have attempted to redefine these units as 'supersecondary structures', 'compact modules' or 'least-extended units'. These notions were consistent with theoretical calculations of the average length of open reading frames in a random (i.e. prebiotic) nucleotide sequence, which was superficially similar to the average length of exons, at least in vertebrates [15,16]. This matter has also been confounded by some confusion between those genes that are clearly ancient, and those which have been assembled more recently. This issue has been clearly articulated by a number of authors (e.g. see [7,17,18**]), who point out that there is a logical distinction between the exon theory of genes and exon shuffling, the latter of which has clearly occurred in the recent evolution of some proteins and is frequently exploited in differential splicing, but does not necessarily explain the origin of introns, nor necessarily imply that ancestral genes had such a structure.

Certain proteins. have been used as models for the analysis of intron position, on the basis that they are clearly ancient and have a well characterized structure. These proteins include globin, triose phosphate isomerase (TPI), pyruvate kinase, alcohol dehydrogenase, and glyceraldehydes-3-phosphate dehydrogenase (GAPDH), some of which have been claimed to show a correlation between intron position and protein structure (e.g. see [19]). The definition of the relevant structural unit to which exons might correspond has become increasingly nebulous, but Gilbert and Glynias [20*] have reported recently that exons in TPI have a statistically significant tendency to form compactly folded domains or 'least-extended polypeptide structures', by quantifying the 'extensivity' (average C_-C_ distances) of exons in this protein relative to random permutations. Stoltzfus *et al.* [18**] have since reported a more comprehensive study of the position of some 62 introns in genes from various species encoding four reference proteins, including TPI. These authors confirmed the finding of Gilbert and Glynias [20*], but found that this did not hold for alcohol dehydrogenase, globins or pyruvate kinase, and that in all cases, there was no correlation between intron position and the secondary or tertiary structure of the protein. They concluded that there is no significant evidence that ancient proteins were assembled from exon-encoded modules of structure and, therefore, that the exon theory of genes is untenable [18**]. Significantly, this study was co-authored by WF Doolittle, one of the original proponents of the introns-early hypothesis, and may represent a turning point in the debate.

A related approach to this issue has been to examine the positions of introns in homologous genes from different phyla and kingdoms in the eukaryotes, on the hypothesis that a common position reflects the situation in the progenote. The fact that most introns are not found in common positions between phylogenetically distant homologs would appear to support the idea that introns were inserted into these genes late in their evolution ([21]; for a recent summary, see [7]), but those favouring the intron-early position argue that this is a misleading consequence of the differential loss of introns, compounded by intron 'slippage', in different lineages from a common ancestral gene that was originally peppered with introns [22,23**]. Thus it has been difficult to distinguish between intron loss and intron gain, although likelihood is shifting in favour of the latter [24]. The most recent contribution to this debate has been provided by Kersanach *et al.* [23**], who reported an analysis of 47 known intron positions in GAPDH genes and concluded that the frequency of identical or near identical positions between (some) introns in plants, animals and fungi could not have occurred by chance (P =2 X 10^{-5}), thereby apparently supporting the exon theory of genes. However, this conclusion has been strongly challenged on the basis, among others, of the following: Firstly, most of the intron positions are not conserved. Secondly, those that are have a very restricted phylogenetic distribution. Thirdly, the average size of exons in the putative ancestral gene containing all 47 introns would have been ridiculously small. Fourthly, and particularly, one would expect that there may be preferential sites for intron insertion (the proto-splice site), which would easily account for the limited number of similar or matching intron positions in these genes [25**]. In reply, Cerff and co-workers [25**] argue that the proponents of the introns-late position must provide a plausible (alternative) hypothesis for the assembly of genes in early evolution, although this may be an entirely different issue. It remains entirely possible that primordial genes were initially assembled from small open reading frames by RNA splicing and/or recombination. This does not necessarily mean, however, that such mosaics persisted throughout the first 3 billion years of cellular evolution (see below), and thus, in this context, the introns-early and introns-late hypotheses may not be incompatible, but merely unlinked.

In interpreting the evidence that a few introns may have a common position between animals, plants and fungi, one must bear in mind their evolutionary history. Comparisons of rRNA sequences support the idea that the three kingdoms of multicellular eukaryotes form monophyletic groups, which seem to have originated almost simultaneously, possibly from a common ancestor [26-291. This would also explain the conservation of the position of some introns among homologous genes within these kingdoms, without having to invoke the presence of introns in these genes from the very beginning.

The alternative scenario: introns-late

The proponents of the introns-early position posit that the lack of introns in prokaryotes and their low abundance in protista is a consequence of the pressure to streamline their genome and minimize replication time for competitive growth advantage [30-32]. They sidestep .the critical issue of how (and why) the prokaryotic and eukaryotic lineages that ultimately gave rise to the intron-rich multicellular eukaryotes might have retained their introns throughout the 3000 million years or so of prior cellular history. The statements of Gilbert and colleagues [32] that "introns were lost in the course of evolution" and that "only genes in slowly replicating cells of complex organisms still retain the full stigmata of their birth" imply that slowly replicating complex eukaryotes diverged early from their simpler relatives, rather than having evolved from them. Moreover, as chloroplasts and mitochondria arose from eubacterial ancestors relatively late in evolution, one has to postulate that introns in those nuclear genes that were post-endosymbiotically transferred from organelle genomes had been preserved in eubacteria for most of their history, but have since been removed [7]. Indeed, the entire introns-early argument rests on the unstated proposition that introns were lost from most unicellular lineages preferentially in the last 500-1000 million years. This seems unlikely, to say the least. Because transcription and translation are intimately coupled in prokaryotic cells, it seems much more reasonable to suggest that prokaryotes could not tolerate the presence of introns in protein genes because of their disruption of protein synthesis, and that this represents the real (and very powerful) selection against them [6,33], irrespective of whether or not ancestral genes were initially assembled from RNA sequence mosaics. Significantly, the only introns found in prokaryotes to date (excluding bacteriophage and organelle genomes) have been in rRNA and tRNA genes (see below), which would not be subject to such strong negative selection, provided that splicing out occurs within a reasonable time-frame relative to the biology of the cell, or if there are multiple copies of such genes.

The discovery of self-splicing introns in chloroplast and mitochondrial genes has provided valuable insights into the likely evolutionary history of introns and their invasion of eukaryotic genes [6,7]. These sequences probably derive from the prebiotic RNA world [34-37] and represent mobile genetic elements (for recent review, see [14*]) that survived in out-of-the-way places in prokaryotic cells, such as tRNA genes. (Interestingly, it has been reported recently that splice-site selection bears similarities to RNA-guided decoding on ribosomes, including inhibition by aminoglycoside antibiotics, suggesting a common origin in early evolution [38,39].) Quite significantly, group II introns are spliced by a lariat reaction mechanism essentially identical to nuclear pre-mRNA introns and have similar 5' and 3' consensus splice-site sequences [14*,40-43]. However, unlike nuclear pre-mRNA introns, group II introns have a complex conserved RNA structure required for splicing [14*]. It has been shown that these RNA domains can be separated and can function in *trans*, both *in vitro* and *in vivo*, leading to the conclusion that they represent the precursors of the small nuclear RNAs (snRNAs) that function in modern spliceosomes [14*,40,41,44**,45].

Recognizable eukaryotic cells (on the basis of their size, other external features, and geochemical signatures) are first detectable around 1.5-2 billion years ago [46,47], although rRNA sequence comparisons suggest that this lineage diverged much earlier in biological history, probably from an archaeal ancestor [48]. Numerous theories have been put forward regarding the origin of eukaryotic cells, one of the most attractive being that they originally evolved as cellular predators with a capacity to ingest solids, including other cells. This would have required the development of a mobile plasma membrane to allow endocytosis, with a cytoskeleton that was capable of controlling this process and which may have led to the evolution of other internal structures of the cell, including lysosomes, the endoplasmic reticulum, and the nuclear membrane [49]. Mitochondria and chloroplasts were acquired subsequently from bacterial endosymbionts [50,51]. The development of a nuclear membrane that isolated the DNA from the rest of the cell may have been to protect the DNA from a more aggressive or complex internal metabolism, or may have been a necessary packaging to keep control of the organization of an increasingly complex cell [49,52]. In any case, the sequestration of DNA into the nucleus resulted in the decoupling of transcription and protein synthesis.

Under these circumstances, there would have been considerably less constraint on the invasion of eukaryotic genes by self-splicing parasitic sequences [6,33]. As noted above, there is very good circumstantial evidence that these sequences originated from group II introns, which were probably introduced via the bacterial ancestor of the mitochondrion [6,53**], and which then co-evolved in partnership with the spliceosome [9,14**,40,41]. The latter has become very sophisticated, containing perhaps as many as 30-100 proteins, and snRNAs [9, 11,12,41,54], and is almost as complex as the ribosome. Moreover, and somewhat perversely, the evolution of the spliceosome would in turn have reduced negative selection against the insertion of such elements by ensuring their efficient excision from transcripts, and have also reduced the sequence requirements (beyond the retention of minimal splicing signals), allowing the internal regions of these elements to degenerate and to drift. In other words, provided they could be conveniently edited, these sequences were able to spread progressively into eukaryotic genomes and remain as relatively stable genetic passengers, which were then free to evolve and explore new genetic space.

Apart from the obvious mechanistic and sequence similarities between self-splicing group II introns and spliceosomal introns, and the discovery of intermediate forms, the evidence for the entry of introns into eukaryotes via bacterial endosymbionts has been strengthened in recent years by the discovery of group I introns in tRNA^{Leu} genes of all cyanobacteria and some other eubacterial phyla (for summaries, see [6,7]). Group I introns had also been found in nuclear rRNA genes of *Physarum* and *Tetrahymena*, the latter apparently having derived from several independent insertions [55]. Critically, however, group II introns (the proposed forerunners of nuclear pre-mRNA introns) had not been found in bacteria. Last year, however, using PCR amplification, Ferat and Michel. [56**] provided evidence for the presence of group II introns in two cyanobacteria and in a _-proteobacterium *(Azotobacter vinelandii)* (also see [53**]), adding strong weight to the hypothesis that these introns entered eukaryotes via the endosymbiotic ancestor of the mitochondrion .[6]. A prediction of this hypothesis is that such introns should be absent in those eukaryotic lineages that diverged earliest and never had mitochondria, such as *Giardia* [7], which has thus far held true. However, even if this were ultimately found not to be true, it would not preclude introns having come through (so to speak) one or more early eukaryotic lineage as an endogenous element, or having been laterally transferred from another source.

The only other issue to consider is the mechanism by which introns might have inserted into new genomic sites. Sharp [34] suggested originally that this could occur by the reversal of the splicing reaction (reverse *trans*-splicing) at proto-splice sites (which have limited sequence requirements), followed by reverse transcription and integration. Reverse self-splicing of group I and group II introns has been demonstrated *in vitro* [14*,57-59], and the recent discovery of introns within introns (‘twintrons’) [44**,60] and in snRNAs in particular lineages of fungi [61] provides strong evidence that this process can and has occurred *in vivo*. Analyses of intron positions in a number of nuclear genes have also indicated that introns have been inserted at different times at particular sites that conform to the proto-splice site [21,62]. It has been suggested that such processes may be one mechanism for the origin of alternative splicing [44**]. The generation of new introns (or alternative splice sites) may also have occurred by mutations at cryptic splice sites, following sequence duplication or transposon insertion [7], or illegitimate recombination.

In summary, the hypothesis that pre-mRNA introns evolved late in eukaryotic history [6,7,63] is strongly supported by recent evidence from a variety of sources. The question now is not so much how introns came to be present within nuclear genomes, but rather why have they become so prevalent? Introns are a major element in the genomes of the higher eukaryotes, in many cases occupying tens of kilobases in length, and accounting for up to 95% of the primary transcriptional output. What were the evolutionary forces at work that have seen them spread (and expand) so successfully into nuclear genes?

Introns: have they evolved function?

Once introns had invaded nuclear genes, and especially once most internal sequence restrictions had been removed by the evolution of the spliceosome, they would have been free to drift and evolve with minimal constraint. Under such conditions, it would not be unreasonable to suggest that any random mutational change that produced a beneficial outcome would have positive selection value and be retained. Upon a moments reflection, this prospect would not simply be regarded as speculative, but be expected. Importantly, such evolution would be able to occur in parallel with protein expression, without directly interfering with it, with the essential difference being that the evolving molecule(s) would be RNA. That is, the entry of introns into eukaryotes may have initiated a new round of molecular evolution, based on RNA rather than protein.

At least one outcome of this process was alternative splicing, which has increased genetic complexity by allowing the production of a set of related proteins with different properties from a single gene. Surprisingly, although the biochemistry of splicing and the structure of the spliceosome per se has been well studied [9,11,12,54], only relatively recently has significant progress been made in understanding the families of proteins that regulate RNA processing and alternative splicing. A number of genes encoding such proteins have now been identified [64-67], and recently it has been estimated, using conserved sequences in the RNA recognition motif to design primers for PCR amplification, that there may be as many as 300 different genes encoding such proteins in the *Drosophila* genome [68*]. Thus, there appears to be a matrix of transcriptional and splicing controls that regulate protein expression in eukaryotes. There is every possibility that this was an important factor in the evolution of multicellular organisms with organized differentiated cells expressing subsets of the overall genetic program. However, the sequences involved in alternative splicing are small; mutational studies have shown that only very few bases, usually located at or near the intron/exon boundary, are required for splice site selection [69,70], which at face value cannot account for the vast tracts of intronic sequences in the higher organisms. These introns bear many of the signatures of information, including high sequence complexity, non-random base distribution and intriguing patterns of conservation (see below).

I suggest that introns have evolved functions of their own and that there are in fact two levels of information produced by gene expression in the higher organisms – mRNA and informational RNA (iRNA). If one accepts this possibility, the significance of nuclear introns and other types of iRNAs (see below) in the eukaryotic cell takes on an entirely different perspective. Immediately, one can envisage that these sequences could create a new dimension of genetic programming that would potentially allow genes to communicate directly with each other, via RNA signals that are implicit in the primary transcript, providing an alternative regulatory network that does not depend on indirect biochemical mechanisms to establish meaningful contacts between different genes. (Interestingly, the possibility of RNA-based regulation and integration of gene activity was mooted by Britten and Davidson [71] almost a decade before the discovery of introns, on the basis of the large differences in complexity between nuclear and cytoplasmic RNA populations in the higher organisms, but has not since been re-visited - cf. [72].) This process would have accelerated as it became more established, leading ultimately to a radical change in the genetic operating system of the cell. This does not mean that all introns contain information. Nevertheless, I suggest that transcription in the higher eukaryotes results in the simultaneous expression of both structural (i.e. protein-coding) and networking information, allowing multiplex contacts between genes and their products. This would potentially include RNA-DNA, RNA-RNA and RNA-protein interactions, and there are known or implied examples of all three. Indeed, one might expect any and all possibilities to be exploited in different circumstances.

It is clear that a close correlation exists between intron density and developmental complexity. Palmer and Logsdon [7] have estimated that there is an average of one intron per kb of coding sequence in simpler eukaryotes such as *Dictyostelium* and *Plasmodium*, rising to 3-4 per kb in plants and fungi, and reaching a zenith in animals (an average 6 per kb of coding sequence in vertebrates), which was interpreted simply in terms of a restricted phylogenetic distribution and therefore evidence of late insertion. Although this is true, it is also consistent with the evolution of function and positive selection in these lineages. Moreover, and perhaps more significantly, average intron size also increases, accounting for no more than 10-20% of the primary transcripts in protista, but as much as 95% in vertebrates [73]. Not all introns in multicellular organisrns are large, and many may simply represent relics of past insertions with no specific function, but the key point is that these organisms have a progressively bigger complement of introns. This is also consistent with positive selection, and although there are other potential sources of such pressures [24], they are not mutually exclusive. There may also be negative selection to reduce intronic sequences in some cases (e.g. see [70*]), but this is not inconsistent with the hypothesis that a significant number of introns have acquired and convey information in the higher organisms.

The interpretation of intron load has been confused by (and with) the so-called C-value paradox, that is, the apparent lack of correlation between genome size and phylogenetic position. This appears to be primarily due to different levels of repetitive sequences (which occur for unknown reasons), rather than intron load, although the former may impose on the latter. In general, it is clear that nuclear introns have high sequence complexity, suggestive of information content, although they may include some simpler elements within them. Firstly, it has been known for some time that the so-called ‘heterogenous nuclear RNA’ (hnRNA), which was widely studied before the discovery of introns, and which presumably represents the nuclear pool of pre-mRNA (and other) transcripts, largely comprises highly complex unique sequences [75]. Secondly, introns do not appear to contain significant amounts of repeated sequences or stretches of cryptic simplicity [76,77]. Introns also exhibit non-random nucleotide distribution; for example, the dinucleotides GG, GC and TC exhibit opposite trends in coding sequences and introns, implying (in both cases) that there are some selective pressures acting on these sequences [78].

Introns may also be more highly conserved than generally acknowledged. The observation that many introns are less conserved in sequence between organisms than their associated exons has frequently been interpreted as evidence of non-functionality. This may be true in some (many) cases, but this preconception is challenged by a number of papers that report unexpected patterns of intron conservation. For example, the single intron in the oligodendrocyte-myelin glycoprotein is highly conserved between mouse and human (75% overall), with one region exhibiting over 90% sequence identity [79]. There are a number of anomalies in the pattern of intron conservation among *Drosophila Adh* genes, including an unexpectedly high conservation of intron 1, which in one case exceeds 96% between different species [80]. A highly conserved sequence cassette from the first intron of a metallothionein gene from sea urchin is found in a number of other RNAs in eggs and embryos, in both orientations [81]. The introns of the *Ubx* gene are highly conserved among insects (ME Akam, personal communication). Sequences in the third intron of the _-actin gene have been highly conserved between human and *Xenopus* [82]. The first intron of the human major histocompatibility complex *DQA1* gene exhibits both evolutionary stability and non-random sequence variation [83,84]. Many other examples exist, but perhaps the most spectacular to date, and which has focussed attention most sharply on this area, is the recent finding that the sequences of the mouse and human T-cell receptor genes show 71% homology over their entire 100kb length, even though less than 6% encodes exons [85].

Whether these examples represent isolated cases, or evidence of a more general phenomenon, remains to be determined. The assumption that introns are non-functional has meant that many introns, especially those that are large, remain unsequenced, which therefore restricts the ability to discern patterns. This may be rectified by genomic sequencing projects that do not suffer such assumptions. Moreover, we do not know what the rules might be - how information might be encoded in introns, how that information might be transduced, what constraints might operate on it, and how quickly it may evolve. Nevertheless, one can say that intronic sequences are complex and that some exhibit striking conservation, both of which are prima facie evidence of information content.

Indications of intron-mediated gene regulation are beginning to appear. It has been reported recently that the product of the *lin-4* gene of *Caenorhabditis elegans* (which regulates the expression of the *lin-14* gene, which in turn encodes a nuclear protein that is involved in the temporal control of post-embryonic development) is a small RNA that originates in the intron of another mRNA [86,87]. Deletion of the conserved third intron of human _-actin affects the morphology of transfected cells, suggesting that these sequences “convey functionally significant information to the cell” [88]. Perhaps even more telling is the discovery of a number of small nucleolar RNAs that are derived from the processed introns of transcripts encoding a number of ribosome-associated proteins (L1, L5, S3, S8 and eIF-4A) as well as the nucleolar protein nucleolin, the hsc70 protein, and the cell-cycle regulated protein RCC1 [89,90]. This seems to be both a clear example of a dual output from these genes, and a potential instance of feedback regulation by intronic sequences on related genes, that is, the expression of rRNAs in the nucleolus. It has also been suggested previously that intronic sequences might operate in *trans* to regulate rRNA expression [91]. The lack of known examples of specific intron-derived RNAs in the nucleoplasm may be simply because this is a more complex milieu and individual species are more difficult to detect. A number of regulatory effects mapped to introns are not explained easily by conventional transcription factor interactions (e.g. see [92]).

A prediction of this hypothesis is that some genes will have evolved to express only iRNA, and there is clear evidence for this in such cases as the *XIST* and *H19* transcripts of mammals and in the *bithorax-infraabdomi*nal region of *Drosophila*. The *H19* transcript appears to act as a tumour suppressor, but does not encode a protein [93,94]. Similarly, *XIST* appears to be involved in X-chromosome inactivation, but also does not appear to encode a protein [95,96]. In these cases, the primary transcripts are spliced to produce relatively large poly(A)$^+$ RNAs, raising the possibility that either the introns or exons of such genes, or both, may transmit information. Five classes of mutations affect expression of the homeotic protein Ubx in *Drosophila,* only one of which affects the protein-coding sequences [97]. The others are located intronically or in the upstream *bxd* region. The latter produces a 27kb transcript that has a number of large introns and is subject to differential splicing to give various small (~1.2kb) poly(A)$^+$ RNAs, none of which contains a significant open reading frame [97-99]. The expression of this transcript is highly regulated during embryogenesis, in a pattern that is partially reflective of the *Ubx* transcript [98,99]. A number of *bxd* insertional mutations have no effect on the amount or the size of the *bxd* poly(A)$^+$ RNAs, suggesting that this species is irrelevant to the observed phenotypes [97]. The real import of the transcription of these sequences may be to produce nuclear intronic RNAs, which, interestingly, exhibit some homology to sequences in the introns of *Ubx* [97]. A similar situation is observed in the 100kb *infraabdominal* region which lies between the homeotic genes *abd-A* and *Abd-B* in the BXC locus. Most or all of the DNA in the intergenic region is transcribed in a distinct spatially restricted pattern during embryogenesis, involving at least three transcriptional domains, whose expression is at least partially regulated by gap genes, and within which are found a number of mutations that affect the parasegmental expression pattern of *abd-A* and *Abd-B* [100]. Indeed, it seems that virtually the entire 200kb BXC locus is transcribed, into at least seven transcription units, only three of which contain protein-coding sequences, but all of which appear to have genetic function.

The evolution of parallel processing

There are a number of puzzling genetic phenomena and molecular genetic observations which may be explained by RNA signals. This and other evidence in support of the hypothesis will be discussed elsewhere (JS Mattick, manuscript in preparation). Nevertheless, taken together, many apparently disconnected observations support the somewhat unexpected notion that eukaryotes have evolved an RNA-based gene regulatory system, with its roots in the colonization of their genomes by introns. This was only made possible by the separation of transcription from translation. Thus, there are three types of informational genes in higher eukaryotes - those that encode protein, those that encode iRNA, and those that encode both. Such a dual mRNA + iRNA system represents a form of parallel processing, which would have vastly expanded the options for regulating and integrating more complex genetic datasets and programs, and which may have been the critical prerequisite for the evolution of multicellular organisms. A corollary of this suggestion is that the developmental and genetic complexity of prokaryotes may not have been restricted by their biochemical complexity or the available diversity of polypeptide structures, but rather by a primitive operating system that relied solely on protein-DNA loops. The fact that the Earth was limited to unicellular or at best colonial life forms for most of its 3500+ million year biological history, and that multicellular organisms arose recently from within a relatively narrow taxonomic group, characterized by high intron content, is consistent with this.

Acknowledgements

Much of the background research for this article was conducted at the Department of Genetics, University of Cambridge, UK. I thank the Australian Research Council and the University of Queensland for their support of the work. I am also grateful to James Castelli-Gair, Mike Akam and many others for helpful discussions, and to Alan Weiner and Laurence Hurst for their helpful comments on the manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
* of special interest
** of outstanding interest

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

2. Gilbert W: **Why genes in pieces?** *Nature* 1978, **271**:501.

3. Blake CCF: **Do genes-in-pieces imply proteins-in-pieces.** *Nature* 1978, **273**:267.

4. Holland SK, Blake CCF: **Proteins, exons, and molecular evolution.** In *Intervening sequences in evolution and development.* Edited by Stone EM, Schwartz RJ. New York: Oxford University Press; 1990:10-42.

5. Darnell JE, Doolittle WF: **Speculations on the early course of evolution.** Proc *Natl Acad Sci USA* 1986, **83**:1271-1275.

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

7. Palmer JD, Logsdon JM: **The recent origins of introns.** *Curr Opin Genet Dev* 1991, **1**:470-477.

8. Rogers JH: **The role of introns in evolution.** *FEBS Lett 1990,* **268**:339-343.

9. Green MR: **Biochemical mechanisms of constitutive and regulated pre-mRNA splicing.** *Annu Rev Cell Biol* 1991, **7**:559-599.

10. Balvay L, Libri D, Fiszman MY: **Pre-mRNA secondary structure and the regulation of splicing.** *Bioessays* 1993, **15**:165-169.

11. Lamond AI: **The spliceosome.** *Bioessays* 1993, **15**:595-603.

12. Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG: **hnRNP proteins and the biogenesis of mRNA.** *Annu Rev Biochem* 1993, **62**:289-321.

13. Horowitz DS, Krainer AR: **Mechanisms for selecting 5' splice sites in mammalian pre-mRNA splicing.** *Trends Genet* 1994, **10**:100-106.

14*. Lambowitz AM, Belfort M: **Introns as mobile genetic elements.** *Annu Rev Biochem* 1993, **62**:587-622.

A comprehensive and well written review of the evidence that group I and group II introns can function as mobile elements, with a strong emphasis on mechanisms and evolutionary relationships.

15. Naora H, Deacon NJ: **Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes.** *Proc Nati Acad Sci USA* 1981, **79**:6196-6200.

16. Senapathy P: **Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications.** *Proc Nati Acad Sci USA* 1986, **83**:2133-2137.

17. Rogers J: **Exon shuffling and intron insertion in serine protease genes.** *Nature* 1985, **315**.458-459.

18**. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM Jr, Doolittle WF: **Testing the exon theory of genes: the evidence from protein structure.** *Science* 1994, **265**:202-207.

An intellectually and experimentally disciplined examination of the hypothesis that early genes were assembled from a mosaic of exons separated by intron spacers and that exons correspond to units of protein structure. The authors point out the distinction between the exon theory of genes and exon shuffling in recent eukaryotic evolution. They also highlight the esoteric nature of attempts to define the relevant ‘unit’ of protein structure, in the face of the lack of any obvious correlation between the latter and exons that would theoretically have comprised the primordial gene when all intron positions are taken into account.

19. Stone EM, Schwartz RJ (Eds): *Intervening sequences in evolution and development.* New York: Oxford University Press; 1990.

20*. Gilbert W, Glynias M: **On the ancient nature of introns.** *Gene* 1993, **135**:137-144.

A recent summary of the exon theory of genes and analysis of intron position with respect to protein structure in triosephosphate isomerase, which exemplifies the increasingly convoluted analyses put forward in support of this hypothesis.

21. Dibb NH, Newman AJ: **Evidence that introns arose at proto-splice sites**. *EMBO J* 1989, **8**:2015-2021.

22. Crabtree GR, Comeau CM, Fowlkes DM, Fornace DJ, Malley JD, Kant JA: **Evolution and structure of the fibrinogen genes: random insertion of introns or selective loss?** *J Mol Biol* 1985,**185**:1-19.

23**. Kersanach R, Brinkmann H, Liaud M-F, Zhang D-X, Martin W, Cerff R: **Five identical intron positions in ancient duplicated genes of eubacterial origin**. *Nature* 1993, **367**:387-389.

The most recent and high profile attempt to use phylogenetic comparisons to show that intron position is conserved and therefore ancient. A different perspective is articulated in [25**].

24. Hurst LD: **The uncertain origin of introns**. *Nature* 1994, **371**:381-382.

25**. Logsdon JM Jr, Palmer JD, Stoltzfus A, Cerff R, Martin W, Brinkmann H: **Origin of introns - early or late?** *Nature* 1994, **369**:526-528.

An illuminating debate between opponents and proponents of the introns-early position, conducted in the Scientific Correspondence section of *Nature*, following publication of [23**].

26. Sogin ML: **The phylogenetic significance of sequence diversity and length variations in eukaryotic small subunit ribosomal RNA coding regions**. In *New perspectives in evolution*. Edited by Warren L, Koprowski H. New York: Wiley Liss; 1991:175-188.

27. van der Peer Y, Neefs J, de Rijk P, de Wachter R: **Evolution of eukaryotes as deduced from small ribosomal subunit RNA sequences**. *Biochem Syst Ecol* 1993, **21**:43-55.

28. Wainwright PO, Hinkle G, Sogin ML, Stickel SK: **Monophyletic origins of the metazoa: an evolutionary link with fungi**. *Science* 1993, **260**:340-341.

29. Kobayashi M, Takahashi M, Wada H, Satoh N: **Molecular phylogeny inferred from sequences of small subunit ribosomal DNA supports the monophyly of the metazoa**. *Zool Sci* 1993, **10**:827-833.

30. Doolittle WF: **Genes in pieces: were they ever together?** *Nature* 1978, **272**:581-582.

31. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution**. *Nature* 1980, **284**:601-603.

32. Gilbert W, Marchionni M, McKnight G: **On the antiquity of introns**. *Cell* 1986, **46**:151-154.

33. Cavalier-Smith T: **The origin of eukaryote and archaebacterial cells.** *Ann NY Acad Sci USA* 1987, **503**:17-54.

34. Sharp PA: **On the origin of RNA splicing and introns.** *Cell* 1985, **42**:397-400.

35. Cech TR: **The generality of self-splicing RNA: relationship to nuclear mRNA splicing**. *Cell* 1986, **44**:207-210.

36. Gilbert W: **The RNA world**. *Nature* 1986, **319**:618.

37. Lamond AI, Gibson TJ: **Catalytic RNA and the origin of genetic systems**. *Trends Genet* 1990, **6**:145-149.

38. Schroeder R, Streicher B, Wank H: **Splice-site selection and decoding: are they related?** *Science* 1993, **260**:1443-1444.

39. Schroeder R: **Dissecting RNA function**. *Nature* 1994, **370**:597-598.

40. Sharp PA: **Five easy pieces**. *Science* 1991, **254**:663.

41. Guthrie C: **Messenger RNA splicing in yeast: clues as to why the spliceosome is a ribonucleoprotein**. *Science* 1991, **253**:157-163.

42. Newman AJ, Norman C: **U5 snRNA interacts with exon sequences at 5' and 3' splice sites**. *Cell* 1992, **68**:743-754.

43. Madhani HD, Guthrie C: **A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic action of the spliceosome**. *Cell* 1992, **71**:803-817.

44**. Copertino DW, Hallick RB: **Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns**. *Trends Biochem Sci* 1993, **18**:467-471.

A recent review summarizing the discovery and analysis of intermediate forms which provide strong support for the view that nuclear introns evolved from group II introns, and that these sequences are capable of invading pre-existing sequences, including pre-existing introns.

45. Bonen L: ***Trans*-splicing of pre-mRNA in plants, animals, and protists**. *FASEB J* 1993, **7**:40-46.

46. Summons RE, Walter MR: **Molecular fossils and microfossils of prokaryotes and protists from Proterozoic sediments**. *Am J Sci* 1990, **290**:212-244.

47. Han TM, Runnegar B: **Megascopic eukaryotic algae from the 1.2-billion-year-old Negaunee iron-formation, Michigan**. *Science* 1992, **257**:232-235.

48. Olsen GJ, Woese CR: **Ribosomal RNA: a key to phylogeny**. *FASEB J* 1993, **7**:113-123.

49. Cavalier-Smith T: **Origin of the cell nucleus**. *Bioessays* 1988, **9**:72-78.

50. Margulis L, Schwartz KV: *Five kingdoms.* San Francisco: Freeman; 1982.

51. Gray MW: **Origin and evolution of organelle genomes**. *Curr Opin Genet Dev* 1993, **3**:884-890.

52. Cavalier-Smith T: *The evolution of genome size*. New York: Wiley; 1985.

53**. Roger AJ, Doolittle F: **Why introns-in-pieces**. *Nature* 1993, **364**:289-290.

A summary of recent data, prompted by [56**], which supports Cavalier-Smith's scheme for the spread of introns into nuclear genes after the origin of the nucleus, via group II introns introduced by the bacterial ancestor of the mitochondrion [6]. However, as the authors correctly point out, this scheme does not explain the expansion of introns within nuclear genomes, especially in multicellular organisms.

54. Ruby SW, Abelson J: **Pre-mRNA splicing in yeast**. *Trends Genet* 1991, **7**:79-85.

55. Sogin ML, Ingold A, Karlock M, Nielsen H, Endberg J: **Phylogenetic evidence for the acquisition of ribosomal introns subsequent to the divergence of some of the major *Tetrahymena* groups**. *EMBO J* 1986, **5**:3625-3630.

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

An important and previously missing link in the hypothesis that group II introns entered eukaryotes via the eubacterial ancestors of organelles.

57. Woodson SA, Cech TR: **Reverse self-splicing of the *Tetrahymena* group I intron: implication for the directionality of splicing and for intron transposition**. *Cell* 1989, **57**:335-345.

58. Augustin S, Muller MW, Schweyen RJ: **Reverse self-splicing of group II intron RNAs *in vitro***. *Nature* 1990, **343**:383-386.

59. Mori M, Schmelzer C: **Integration of group II introns into a foreign RNA by reversal of the self-splicing reaction *in vitro***. *Cell* 1990, **60**:629-636.

60. Copertino DW, Christopher DA, Hallick RB: **A mixed group II/group III twintron in the *Euglena gracilis* chloroplast ribosomal protein S3 gene: evidence for intron insertion during gene evolution**. *Nucleic Acids Res* 1991, **19**:6491-6497.

61. Tani T, Ohshima Y: **mRNA-type introns in U6 small nuclear RNA genes: implications for the catalysis in pre-mRNA splicing**. *Genes Dev* 1991, **5**:1022-1031.

62. Lee VD, Stapleton M, Huang B: **Genomic structure of *Chlamydomonas calcitracin*.. evidence for intron insertion suggests a probable genealogy for the EF-hand superfamily of proteins**. *J Mol Biol* 1988, **221**:175-191.

63. Cavalier-Smith T: **Selfish DNA and the origin of introns**. *Nature* 1985, **315**:283-284.

64. Haynes SR: **The RNP motif protein family**. *New Biol* 1992, **4**:421-429.

65. Zahler AM, Lane WS, Stolk JA, Roth MB: **SR proteins: a conserved family of pre-mRNA splicing factors**. *Genes Dev* 1992, **6**:837-847.

66. Biamonti G, Riva S: **New insights into the auxiliary domains of eukaryotic RNA binding domains**. *FEBS Lett* 1994, **340**:1-8.

67. Burd CG, Dreyfuss G: **Conserved structures and diversity of functions of RNA-binding proteins**. *Science* 1994, **265**:615-621.

68*. Kim Y-J, Baker BS: **Isolation of RRM-type RNA-binding protein genes and the analysis of their relatedness by using a numerical approach**. *Mol Cell Biol* 1993, **13**:174-183.

An interesting attempt to use PCR-homology cloning to explore the family of RNA-binding proteins (in *Drosophila),* especially those which might control alternative splicing, an area which had, until recently, been neglected in studies on the control of gene expression in the higher organisms.

69. McKeown M: **Alternative mRNA splicing**. *Annu Rev Cell Biol* 1992, **8**:133-155.

70. Watakabe A, Tanaka K, Shimura Y: **The role of exon sequences in splice site selection**. *Genes Dev* 1993, **7**:407-418.

71. Britten RJ, Davidson EH: **Gene regulation for higher cells: a theory**. *Science* 1969, **165**:349-357.

72. Davidson EH: **Molecular biology of embryonic development: how far have we come in the last ten years?** *Bioessays* 1994, **16**:603-615.

73. Cavalier-Smith T: **Eukaryotic gene numbers, non-coding DNA and genome size**. In *The evolution of genome size.* Edited by Cavalier-Smith T. Chichester: John Wiley & Sons; 1985:70-103.

74*. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, Aparicio S: **Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome**. *Nature* 1993, **366**:265-268.

The *Fugu* genome (400 Mb) is only about one-seventh the size of the human genome. It contains very little repetitive DNA and predominantly small introns, three-quarters of which are apparently less than 120bp (unpublished results cited in the text). At face value, this would seem to contradict the hypothesis put forward herein, but it is important to note that this hypothesis does not predict that all (or even most) introns will necessarily contain information. This has been, and will undoubtedly be, a confounding factor in any judgement as to their function, in general. However, a clear prediction of the hypothesis is that some introns will contain larger and more complex sequences, indicative of information, and it will be interesting to examine the remaining 25% of *Fugu* introns in this light. Indeed, *Fugu* may offer an uncluttered system for the identification and analysis of those introns important for vertebrate gene regulation and development.

75. Davidson EH: *Gene activity in early development.* New York: Academic Press; 1976.

76. Tautz D, Trick M, Dover GA: **Cryptic simplicity in DNA is a major source of genetic variation**. *Nature* 1986, **322**:652-656.

77. Tautz D, Tautz C, Webb D, Dover GA: **Evolutionary divergence of promoters and spacers in the rDNA family of four *Drosophila* species**. *J Mol Biol* 1987, **195**:525-542.

78. Gutierrez G, Oliver JL, Martin A: **Dinucleotides and G+C content in human genes: opposite behaviour of GpG, GpC, and TpC at II-III codon positions and in introns**. *J Mol Evol* 1993, **37**:131-136.

79. Mikol DD, Rongnoparut P, Allwardt BA, Marton LS, Stefansson K: **The oligodendrocyte-myelin glycoprotein of mouse: primary structure and gene structure**. *Genomics* 1993, **17**:604-610.

80. Sullivan DT, Atkinson PW, Starmer WT: **Molecular evolution of the alcohol dehydrogenase genes in the genus *Drosophila***. In *Evolutionary* biology. Edited by Hecht MK, Wallace B, Macintyre RJ. New York: Plenum Press; 1990:107-147.

81. Nemer M, Bai G, Stuebing EW: **Highly identical cassettes of gene regulatory elements, genomically repetitive and present in RNA**. *Proc Nati Acad Sci USA* 1993, **90**:10851-10855.

82. Erba HP, Eddy R, Shows T, Kedes L, Gunning P: **Structure, chromosomal location, and expression of the human _-actin gene: differential evolution, location, and expression of the cytoskeletal _- and _-actin genes**. *Mol Cell Biol* 1988, **8**:1775-1779.

83. Simons MJ, Ashdown ML, Kibble MJ, Chapman DG, McGinnis MD: **Intron sequence variation in HLA class II genes is non-random, stable, and informative of haplotype evolution**. *Hum Immunol* 1992, **34 (suppl 1, abstract B7.4.172)**:92.

84. McGinnis MD, Labo RV, Quinn DL, Simons MJ: **Ancient, highly polymorphic human major histocompatability complex DQA1 intron sequences**. *Am J Hum Genet* 1994, in press.

85. Koop BF, Hood L: **Striking sequence similarity over almost 100kilobases of human and mouse T-cell receptor DNA**. *Nature Genet* 1994, **7**:48-53.

86. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complimentarity to *lin-14.*** *Cell* 1993, **75**:843-854.

87. Wightman B, Ha I, Ruvkin G: **Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans***. *Cell* 1993, 75:855-862.

88. Lloyd C, Gunning P: **Noncoding regions of the _-actin gene influence the impact of the gene on myoblast morphology**. *J Cell Biol* 1993, **121**:73-82.

89. Soliner-Webb B: **Novel intron-encoded small nucleolar RNAs**. *Cell* 1993, **75**:403-405.

90. Nicoloso M, Caizergues-Ferrer M, Michot B, Azum M,. Bachellerie J: **U20, a novel small nucleolar RNA, is encoded in an intron of the nucleolin gene in mammals**. *Mol Cell Biol* 1994, **14**:5766-5776.

91. Sekeris CE: **The role of hnRNA in the control of ribosomal gene transcription**. *J Theor Biol* 1985, **114**:601-604.

92. Meredith J, Storti RV: **Developmental regulation of the *Drosophila tropomyosin* gene in different muscles is controlled by muscle-type-specific intron enhancer elements and distal and proximal promoter control elements**. *Dev Biol* 1993, **159**:500-512.

93. Brannan CI, Dees EC, Ingram RS, Tilghman SM: **The product of the *H19* gene may function as an RNA**. *Mol Cell Biol* 1990, **10**:28-36.

94. Hao Y, Crenshaw T, Moulton T, Newcomb E, Tycko B: **Tumour. suppressor activity of *H19* RNA**. *Nature* 1993, **365**:764-767.

95. Brown CJ, Hendrich BD, Rupert JL, Lafraniere RG, Xing Y, Lawrence J, Willard HF: **The human *XIST* gene: analysis of a 17kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus**. *Cell* 1992, **71**:527-542.

96. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S: **The product of the mouse *Xist* gene is a 15kb inactive X-specific transcript containing no conserved ORF and located in the nucleus**. *Cell* 1992, **71**:515-526.

97. Hogness DS, Lipschitz HD, Beachy PA, Peattie DA, Saint RB, Goldschmidt-Clermont M, Harte PJ, Gavis ER, Helfand SL: **Regulation and products of the *Ubx* domain of the bithorax complex**. *Cold Spring Harb Symp Quant Biol* 1985, **50**:181-194.

98. Akam ME, Martinez-Arias A, Weinzeirl R, Wilde CD: **Function and expression of Ultrabithorax in the *Drosophila* embryo**. *Cold Spring Harb Symp Quant Biol* 1985, **50**:195-200.

99. Akam ME, Martinez-Arias A: **The distribution of Ultrabithorax transcripts in *Drosophila* embryos**. *EMBO J* 1985, **4**:1689-1700.

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

Abbreviations

GAPDH-glyceraldehyde-3-phosphate dehydrogenase;
iRNA-informational RNA;
snRNA-small nuclear RNA;
TPI-triose phosphate isomerase.