Genetica **118:** 123–131, 2003.

Introns in gene evolution

Larisa Fedorova[1] & Alexei Fedorov[2,*]
[1]*Vision Research Laboratories, New England Medical Center, Tufts University School of Medicine, Boston, MA 02111, USA (Phone: +1-617-636-9020; Fax: +1-617-636-8945; E-mail: lfedorova@lifespan.org);* [2]*Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA (Phone: +1-617-495-0560; Fax: +1-617-496-4313; E-mail: afedorov@fas.harvard.edu);* [*]*Author for correspondence (E-mail: afedorov@fas.harvard.edu)*

Key words: evolution, exons, genes, genome, introns, splicing

Abstract

Introns are integral elements of eukaryotic genomes that perform various important functions and actively participate in gene evolution. We review six distinct roles of spliceosomal introns: (1) sources of non-coding RNA; (2) carriers of transcription regulatory elements; (3) actors in alternative and trans-splicing; (4) enhancers of meiotic crossing over within coding sequences; (5) substrates for exon shuffling; and (6) signals for mRNA export from the nucleus and nonsense-mediated decay. We consider transposable capacities of introns and the current state of the long-lasting debate on the 'early-or-late' origin of introns. Cumulative data on known types of contemporary exon shuffling and the estimation of the size of the underlying exon universe are also discussed. We argue that the processes central to introns-early (exon shuffling) and introns-late (intron insertion) theories are entirely compatible. Each has provided insight: the latter through elucidating the transposon capabilities of introns, and the former through understanding the importance of introns in genomic recombination leading to gene rearrangements and evolution.

Introduction

One and half years have passed since the human genome was published but we still do not know the exact number of the human genes. Recent report by Hogenesch et al. (2001) shows that different computer programs exhibit little agreement in their predictions of novel human genes. The main reason why the unknown genes escape accurate detection lies in their exon-intron gene structure. Introns are those parts of genes and their precise pre-mRNA replicas that are excised during maturation of the transcripts into mRNA molecules. Thus, introns are excluded from the main gene function – coding the protein. About 95% of human genes harbor introns. The average human gene contains 5–6 introns. The average length of introns is 2100 nucleotides (nt), but some have been found of enormous sizes of more than 100,000 nt. Exons look like small islands among these non-coding regions with 80% of exons having sizes between 50 and 300 nt, with a peak of 125 nt. The longest human genes are more than a million nucleotides long due to their numerous and lengthy introns.

Introns are present in all studied eukaryotic organisms. Relatively small number of introns were found in single-cell organisms (e.g., yeast *S. cerevisiae* has about 300 introns) while dozens of thousands were discovered in the completely characterized genomes of plants (*A. thaliana*), invertebrates (*D. melanogaster* and *C. elegans*) and vertebrates (mouse, human). Evolution of intron-exon structure is very controversial. What is the role of introns in the genome? Notwithstanding that introns were discovered 25 years ago, there are still opposite viewpoints on this question. In this paper, we discuss the following long-standing dilemmas concerning introns: (1) Do introns have important cellular functions or are they predominantly selfish elements of genomes? (2) Are introns ancient genomic elements or were they acquired only recently in the evolution of eukaryotes? (3) Have introns been

universally involved in the formation of new genes by exon shuffling or is it an uncommon process in gene evolution?

In our paper we consider only the main type of introns, known as 'spliceosomal'. This type consists of all introns found in nuclear protein-coding genes transcribed by RNA polymerase II. Two other types of introns, group I and group II, are very small in number and mainly restricted to the genomes of cellular organelles. These types were recently described elsewhere (Lambowitz & Belford, 1993; Belford & Perlman, 1995; Martinez-Abarca & Toro, 2000; Bonen & Vogel, 2001).

Intron functions

In 1985 Cavalier-Smith (1985) suggested that introns were selfish DNA with no distinct cellular function. This transposon theory of the origin of introns became very popular and lately transformed into the intron-late theory. However, several important functions of introns have been uncovered gradually since that time. New data has rendered a conception of introns as selfish genomic elements obsolete. In this section, we survey six distinct functions of introns in the cell.

Source of non-coding RNAs

Despite the fact that introns were discovered in 1977 (Berget, Moore & Sharp, 1977; Chow et al., 1977; Jeffreys & Flavell, 1977), for more than a decade investigators paid very little attention to what happens with intronic RNA fragments after their removal from pre-mRNA. Only in 1990 did Liu and Maxwell (1990) show that intronic sequences of the mouse hsc70 heat shock gene are the source for U14 small nucleolar RNA (or snoRNA). Non-coding snoRNAs are guiding molecules for precise chemical modification of different RNAs (reviewed by Maxwell & Fournier, 1995; Weinsein & Steitz, 1999). SnoRNAs guide the process of pseudouridylation and 2′-O-ribose methylation in ribosomal rRNA by complementary pairing of their guide sequences with rRNAs. SnoRNAs are also involved in pseudouridylation of telomerase complex (Mitchell & Collins, 2000; Pogacic, Dragon & Filipowicz, 2000) and spliceosomal snRNAs (Peculis, 2000). About 200 types of snoRNA were found in vertebrates, and a majority of them are formed from intronic RNA segments processed after splicing. In 2000, Cavaille et al. (2000) found snoRNA-like molecules specifically expressed in mammalian brain. Some of these novel snoRNAs are also formed from introns. These snoRNAs do not appear to be involved in the chemical modification of rRNAs or snRNAs since they do not have complementary guide sequences for these molecules. Interestingly, Cavaille et al., showed that one of the human brain-specific snoRNAs, named HBII-52, has a 18-nt long guidance sequence of perfect complementarity to the coding sequence of serotonin 2C receptor mRNA. Moreover, the 5th position of the HBII-52 snoRNA guidance sequence, usually responsible for a chemical modification of target RNAs, corresponds to the adenosine→inosine editing site of serotonin 2C receptor mRNA. This editing site is also very close to the alternative splice site of the 5th exon in the serotonin 2C receptor gene. Cavaille et al., showed that HBII-52 snoRNA is phylogenetically conserved and its counterpart in the mouse genome has 18-nt complementarity to the mouse serotonin 2C receptor gene as well. Therefore, the authors proposed a potential role of HBII-52 snoRNA in the processing of serotonin 2C receptor mRNA. This discovery opens the possibility that snoRNA could have more versatile functions in the nucleus and that there could be many more yet unknown snoRNA-like molecules encoding by introns.

Recently, new types of non-coding RNAs with a diverse range of functions were discovered (review by Eddy, 1999). We are only at the beginning of the investigation of non-coding RNAs (Filipowicz, 2000; Eddy, 1999). Thus, there is a possibility that together with snoRNAs, some other yet unknown non-coding RNAs are present within introns.

Source of regulatory elements

A number of elements regulating gene expression have been found within intronic sequences. For example, the second intron of the human apolipoprotein B gene is required for expression of this gene in liver (Brooks et al., 1994). Similarly, Lothian and Lendahl (1997) showed that the evolutionarily conserved region in the second intron of the human nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. Thereafter, this nestin intronic element was successfully used in transgenic experiments for guiding specific expression of transfected genes within neural stem cells (Akamatsu & Okano, 2001; Aoki et al., 2000). Hural and co-authors demonstrated that a cis-acting element in the second intron of the

murine interleukin IL-4 gene has a dual function: first, it regulates transcription in mast cells; second, through directing methylation of the gene it influences chromatin structure (Hural et al., 2000). The first intron of the human keratin 18 gene contains a 100-bp enhancer element with binding sites of the AP-1 and Ets transcription factors and mediates increased expression by the Ras-mitogen-activated protein kinase signal transduction pathway (Rhodes & Oshima, 1998; Pankov et al., 1994; Oshima et al., 1990). Another enhancer that binds a developmentally regulated factor, PRL-1 intron enhancer complex (PIEC), was found within the first intron of the human nuclear protein tyrosine phosphatase PRL-1 gene (Peng et al., 1998). Howell and Hill revealed that a 107-bp region of the first intron of the *Xenopus* dorsal mesoderm XFKH1 gene acts as an enhancer and confers activin inducibility (Howell & Hill, 1997). Pan and Simpson (1999) showed that 280-nt region of the first intron of the human c-mic gene contains three binding sites for nuclear phosphoproteins whose binding blocks transcription elongation of the gene, while in the first intron of the human N-myc gene there is a 116-nt element that directs tissue-specific expression (Silvak et al., 1999).

Besides navigating gene expression, elements inside introns participate in determination of alternative splicing. For instance, intronic elements regulate alternative splicing in the mammal calcitonin/calcitonin gene-related peptide (CT/CGRP) (Lou, Gagel & Berget, 1996) and alternative splicing of exon 7B in the hnRNP A1 human gene (Simard & Chabot, 2000).

Keeping in mind that introns have transposable properties (see below), it is clear that they have the potential to transfer regulatory elements from one gene to another. Hence, introns suggest themselves as functional genomic carriers of gene regulatory elements.

Actors in alternative splicing and trans-splicing

The presence of introns in a gene provides that gene an opportunity to generate alternative coding messages through alternative splicing of pre-mRNA. Alternative splicing – an expression of different types of mRNA from a single gene – is widespread in multicellular eukaryotes and results from the use of alternative 5′-splice sites, alternative 3′-splice sites, optional exons, mutually exclusive exons, and retained introns (reviewed by Lopez, 1998). According to the low-end estimations of Croft et al. (2000) at least 35% of the human genes are alternatively spliced. Often, a gene produces only a few types of alternatively spliced mRNA molecules, but sometimes the variation in splicing forms can be enormous – up to hundreds or maybe even thousands of different forms from a single gene (Black, 2000) (e.g., DSCAM gene in *Drosophila* (Schmucker et al., 2000) and neuroxin genes in Human (Missler & Sudhof, 1998)). Therefore, alternative splicing is tremendously important in increasing the protein diversity in the organism (reviewed by Black, 2000). The capability for alternative splicing is due to the existence of introns which are 'providers' of the splicing process itself.

Another way in which splicing is used to increase the protein diversity of organisms is trans-splicing, in which two different pre-mRNA molecules recombine to form a single mRNA. The trans-splicing process occurs through recombination within non-coding parts of the molecules very similar to ordinary introns. Trans-splicing is widespread among lower eukaryotes (Nilsen, 2001), and has also been found in mammals (Akopian et al., 1999; Caudevilla et al., 2001; Takahara et al., 2000), *Drosophila* (Dorn, Reuter & Loewendorf, 2001), and plants (Kawasaki et al., 1999). Approximately 70% of mRNAs in *C. elegans* are trans-spliced to conserved 21- to 23-nucleotide leader RNAs (Krause & Hirsh, 1987; Ferguson & Rothman, 1999; Evans & Blumenthal, 2000). In our opinion, at present, we do not have sufficient information about the prevalence of trans-splicing in different species and, hence, it is difficult to estimate its importance of this process for the eukaryotes as a whole.

Enhancers of meiotic crossing over within coding sequences

Complex eukaryotic organisms generate dozens of different tissue-types all of whose cells have the same genome. The formation of these organisms requires the precise and coordinated expression of genes in a tissue-specific and time-specific manner, which can be finely tuned by numerous extracellular signals. This is made possible by compound promoter regions in multicellular eukaryotes which are much larger than those in prokaryotes. In humans, promoter elements regulating transcription of a gene can be found up to 50 kb upstream or 50 kb downstream from the gene transcription initiation site. On average, promoter regions of higher eukaryotes are much larger than the coding sequences whose expression they control. Such promoter regions harbor up to several tens of different regulatory elements for activating or suppressing gene

expression in response to different extracellular or intracellular signals. Many of these promoter elements are binding sites for proteins or other molecules and are very short – usually from 4 to 20 nt. The rest and the largest part of a promoter region is non-functional besides providing a playground for the appearance of new elements regulating gene expression during evolution. For instance, for an appearance of a given 8-nucleotide element at a particular genomic site during evolution, it takes on average $4^8 = 6.5 \times 10^4$ different point mutations. Taking into account that the mutation rate is low (about 10^{-8} per site per generation for humans), the appearance of a particular 8-nucleotide element at a particular genomic site in at least one individual of the whole next generation of 6 billion human population has a probability of 10^{-3} (probability $= 10^{-8} \times 6 \times 10^9/4^8$). At the same time, this particular 8-nucleotide element will appear on average in 10 humans of next-generation inside a promoter region of 10 thousand nucleotides long. Thus, extensive non-coding regions of genomes of multicellular eukaryotes are important for the organisms by providing the space for the evolution of promoter elements. Introns can harbor functional transcription elements and, therefore, provide an additional space for the evolution of gene regulation along with 'classical' promoter regions located upstream of genes.

A most important function of introns as providers of evolutionary experimentation is that they increase the rate of meiotic crossing over within a coding sequence, as shown in Figure 1. In the case of intronless genes with large promoter regions, crossing over will be infrequent within a coding sequence and will occur preferentially outside the gene (Figure 1(a)). On the other hand, in the case of intron-containing genes, introns re-distribute non-coding sequences from upstream promoter regions to the regions inside coding sequences. So, the probability of crossing over between segments of a coding sequence (exons) increases considerably (Figure 1(b)). This meiotic recombination between coding regions of a gene is of great importance for the evolution of the protein. It brings together different mutations and tries them in different combinations, for those which have synergetic selective advantages. So, the presence of introns in multicellular eukaryotes could increase immensely the speed of protein evolution. The idea of the importance of introns for gene recombination was first outlined by Gilbert (1978). Recently, this idea received confirmation in a short paper by Carvalho and Clark (1999) and then in a thorough analysis by Comeron and Kreitman (2000), who show a negative correlation between intron length and recombination rate in *Drosophila* and human genes. This correlation means that in the chromosomal regions where meiotic crossing over is infrequent the introns tend to be larger and by this elongation they increase the rate of recombination between exons.

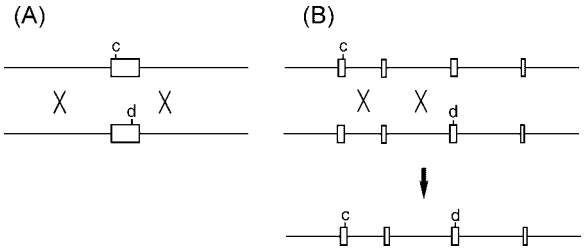

Figure 1. Meiotic crossing over events in the case of (A) an intronless gene, (B) an intron-containing gene. Coding sequences are shown as boxes, non-coding sequences including introns – as lines. Sites of meiotic crossing overs are shown as (X). Letters c and d mark maternal and paternal alleles of the genes. In the case of intron-containing gene (B) crossing overs frequently occur within intronic sequences, which brings c and d mutations to the same gene. On the other hand, in the case of intronless gene (A) crossing overs most frequently occur outside the coding sequences.

Actors in exon-shuffling

Introns-early theory proposes that genes were assembled from exonic 'pieces' and that this assembly was implemented by illegitimate recombination within intronic sequences. In his review of 1999, Patthy showed that exon-shuffling frequently occurred in the genomes of metazoa. Examination of exon duplication, a particular type of exon-shuffling in which an exon is reused within the same gene, demonstrated that at least 6% of all human exons appeared to be the result of duplication, according to the low-end estimates by Fedorov et al. (1998). The majority of the largest human genes like titin, dystrophin, collagen, von Willebrand factor and others have duplicated exons inside them. Thus, exon shuffling has been of major importance for eukaryote evolution and introns are the providers of this process. The details of exon shuffling are discussed in the last section of this paper.

Signals for mRNA export from the nucleus and nonsense-mediated decay

Another intriguing possibility for introns lies in the very tight and intricate interaction of splicing with the processes of mRNA export from the nucleus and regulation of mRNA stability in cytoplasm.

A model of how splicing is coupled with mRNA export nonsense-mediated decay has been excellently described by Reed and Magni (2001) and Maniatis and Reed (2002). Therefore, we will not repeat this theme and will confine ourselves to the references to these reviews.

Any other functions?

Above we have discussed six different functions of introns. Since the investigation of non-coding sequences has not been a major priority of the worldwide scientific community, there is a possibility that we have not yet detailed all functions of introns and that some more remain to be discovered.

Early or late?

Initially in 1978, Doolittle (1978) and Darnel (1978) proposed that introns are very ancient genetic elements which existed at the beginning of life before the divergence of eukaryotes and prokaryotes. The exon theory of the genes developed by Gilbert holds that introns existed at the earliest time of life in the 'RNA World' (Gilbert, 1987). The contrary view on the origin of introns was formulated in 1991 in the papers of Cavalier-Smith (1991) and Palmer and Logsdon (1991). These author claims that introns appeared relatively recently in the genomes of eukaryotes long after their divergence from prokaryotes. An important contribution of introns-late theory to our understanding of the exon/intron gene structure and evolution is that introns are transposable elements capable to be acquired into a gene or be deleted from it. The notion of intron transposition came from the fact that intron positions vary in homologous genes of different organisms (Logsdon et al., 1995; Logsdon, 1998; Logsdon, Stoltzfus & Doolittle, 1998). At the same time, transposition of introns occurs very infrequently during evolution. If we compare, for example, intron positions in genes from mouse and human, species that diverged at least 50 million years ago, intron positions are the same in the majority of orthologous gene pairs. The mechanism of intron transposition is still an enigma, because we know only a few examples of intron transposition *de novo*. One hypothesis is that intron transposition occurs through a process of reverse splicing (Sharp, 1985; Roger & Doolittle, 1993), while another suggests that intron moves like transposons via excision/insertion at the DNA level (Giroux et al., 1994).

By itself, transposable capability of introns does not give us a clue to the time of intron origin. A main argument of supporters of introns-late theory lies in the confined distribution of introns on the phylogenetic tree of eukaryotes. They claim that introns do not exist in the earliest branches of eukaryotes (Logsdon, 1998). However, the latest data on sequenced genomes do not support this phylogenetic view. A thorough genomic investigation showed that the deepest known branches in the eukaryotic line of descent – diplomonads *Giardia liamblia*, trichomonads *Trichomonas vaginalis*, and microsporidian *Nosema locustae* genomes contain several genes for spliceosomal proteins, testifying for intron presence in these species (Hartman & Fedorov, 2002; Fast & Doolittle, 1999; and Fast et al., 1998, respectively). Nixon and co-authors only just have proved the existence of spliceosomal intron in *Giardia* (Nixon et al., 2002). So far no intronless eukaryotic species has been found. Introns have been found even in the microsporidia *Encephalitozoon cuniculi* which has the tiniest eukaryotic genome of 2.9 megabase and whose genome compaction is reflected by reduced intergenic spacers and by the shortness of most proteins (Katinka et al., 2001). Recently 17 spliceosomal introns have been reported in the nucleomorph genome of the cryptomonad *Guillardia theta* (Douglas et al., 2001). This nucleomorph genome is also extremely compact and has a length of 551 kb. Therefore, all recent data support the assumption that introns already existed at the earliest stages of eukaryote evolution.

Can we now answer the long lasting question of whether introns have existed before the divergence of eukaryotes and prokaryotes? The cumulative genomic data show that this divergence was not a simple event – a time-point during evolution at which two cellular domains appeared from the same ancestor. There have been multiple events of horizontal gene transfer between domains of life (Doolittle, 1999). Recently a large-scale comparison of prokaryotic and eukaryotic proteins by Hartman and Fedorov (2002) gave evidence to the Chronocyte hypothesis for the origin of the eukaryotic cell. According to this hypothesis a third ancient cellular domain, named Chronocyte, existed at the earliest stages of evolution. The eukaryotes were formed when the Archaea and Bacteria were engulfed by the Chronocyte. The authors of this hypothesis proposed that the predecessor of eukaryotes – Chronocyte may have existed as far as 3.8 billion years ago and it was RNA-based organism contained introns. The more data we have the more complex the picture of

early evolution becomes. At the present time we do not have a reliable picture of ancient prokaryotes and eukaryotes and, therefore, only unconfirmed speculations about the appearance of introns are available.

What is important to note concerning the early-or-late question is the very tight connection of introns with transcriptional and translational machinery. Introns cannot exist within the present-day prokaryotes simply because translation immediately follows the process of gene transcription. Ribosomes bind partially transcribed mRNA molecules which are still undergoing transcription and, therefore, there is no room and time for introns to be excised. Introns only exist within prokaryotic-like organelles, mitochondria and chloroplasts, where ribosomes and genomic DNA are usually bound to membrane. So, in these organelles, there is some period of time for introns to be spliced out during the transposition of the newly synthesized pre-mRNA to a ribosome. A few group I and II introns are known in prokaryotes, which predominantly located within non-protein-coding genes for rRNAs and tRNAs which do not undergo ribosomal 'attack' and, hence, have a possibility for splicing. Interestingly, group I introns have been detected in eukaryotic nuclear genes, but only in those coding for rRNA and tRNA (Muscarella & Vogt, 1989; Nikoh & Fukatsu, 2001). On the other hand, eukaryotic spliceosomal introns have transposable properties, but they propagate only within genes transcribed by RNA polymerase II. They have never been detected within genes transcribed by RNA polymerase I or III. So each type of introns corresponds to a specific RNA-polymerase. To know more about possible splicing in ancient prokaryotes we need to know the details about transcription and translation in these ancient cells.

As Francis Crick wrote in 1979 "When did introns first arise? The obvious suggestion is that they came in with the eukaryotes. Two investigators (Doolittle, 1978 and Darnel, 1978) have proposed they originated at a much earlier time. This issue may prove difficult to resolve, and I shall not pursue it further here" (Crick, 1979). These words are as true today as when he wrote them 23 years ago.

Exon shuffling

Introns-early theory of the assembly of genes from exon 'pieces' was initially outlined in 1978 by Gilbert (1978), Doolittle (1978) and Blake (1978) and then was elaborated in details as 'the exon theory of genes' by Gilbert (1987). In the original form this theory describes to the earliest steps of evolution starting at the time of 'RNA world'. The theory claims that ancient genes were assembled from mini-exons during the exon-shuffling process. It proposes a simple scheme of molecular events at the genomic level that could dramatically increase the speed of gene evolution. Under this scheme, instead of trying practically unlimited numbers of gene mutations to produce functional proteins, genes were assembled from exonic 'pieces' coding functional protein segments. The exon shuffling is realized via recombination within intronic sequences. Since there are very primitive forms of introns (group I and II introns) with the ability to self-splicing due to simple spatial organization and intrinsic RNA-enzymatic activity (Lambowitz & Belford, 1993; Belford & Perlman, 1995; Bonen & Vogel, 2001), it is likely, that such introns arose at the earliest stages of evolution as far back as the 'RNA world'. After 20 years of rigorous debates on the exon-shuffling origin of genes, it appears that it is practically impossible to proof or refute this theory by establishing what exactly had happened more than two billion years ago on the planet.

It is of interest to consider known cases of exon shuffling detected in the genes of present-day eukaryotes. Two possible mechanisms for this process are reviewed by Long (2001) and many examples of exon shuffling are surveyed by Patthy (1999). Analyzing gene structures in different branches of eukaryote Patthy came to a conclusion that "exon-shuffling acquired major significance at the time of metazoan radiation . . . and the rise of exon-shuffling coincides with a spectacular burst of evolutionary creativity: the Big Bang of metazoan radiation." Interestingly, that present-day exon shuffling differs from the initially proposed one. Many of recent examples show that not a single but a group of exons, representing a protein domain, is participating in a shuffling event (Fedorov et al., 1998). Domon and Steinmetz (1994) demonstrated for anther-specific genes from sunflower that "two genes have originated via exon shuffling during which a copy of a DNA segment including the promoter region as well as a signal peptide coding sequence has been transferred into the upstream region of two different unrelated coding sequences, generating two novel genes which display the same specificity of expression and which both encode an extracellular protein". This example shows that exon shuffling can be useful not only for addition of a new 'piece' of a coding sequence, but for a creation of new expression

pattern for a gene. In many genes the first exon is located in the 5′-non-translated region and shuffling of such 5′-terminal exons could switch gene expression from one type to another. It is appropriate to remind the case of the extremely compact nucleomorph genome of the cryptomonad *Guillardia theta:* "Among 464 genes from this genome only 17 genes contain spliceosomal introns, all located in the 5′region, many immediately after the initiator AUG. Eleven introns are in the ribosomal proteins" (Douglas et al., 2001). Maybe some kind of concerted shuffling of 5′terminus exons in these genes coordinate the expression of ribosomal proteins? Known cases of the appearance of new genes described in Drosophila genome (Long & Langley, 1993; Long, Wang & Zhang, 1999; Nurminsky et al., 1998), demonstrated that in both cases the exon/intron gene structure played an important role in the formation of these genes. However, the formation of both the described genes was not through a 'classical' exon-shuffling process, but through shuffling of much larger gene parts. Therefore, modern exon-shuffling process in the eukaryotic cells is much more complex and diverse than that originally proposed for the beginning of evolution.

There have been several attempts to estimate the 'exon universe' (i.e., the number of different exons from which all genes were assembled.) Some attempts were done on the basis of intron phase distribution (Fedorov et al., 1992; Long, Rosenberg & Gilbert, 1995; Long & Rosenberg, 2000). A direct method of exon universe estimation through the alignment of all exon sequences with each other was conducted by Dorit Schoengach & Gilbert (1990) and the latest comparison by Saxonov and Gilbert (2003) is presented in this issue. Dorit et al., estimated that as small a number as 1000–7000 ancient exons could be involved in the formation of all present genes. However, the detailed recalculation using a much more extensive database of all exonic sequences from the GenBank, release 116, increased this estimation of exon universe to about 15–30 thousands of ancient units (Saxonov & Gilbert, 2002). It is important to note that these calculations were done with the assumption that exon length could be changed only slightly during evolution. Since introns are transposable and can be acquired or lost from a gene, exon length is not an invariable parameter, but one that can be changed dramatically during evolution. These evolutionary changes hide the ancient features of exon–intron structures. That is why the estimations of the rate of exon shuffling during evolution is still only a rough approximation and, also, why correspondence of exons to protein domains is not as strong and obvious as it was expected previously.

Conclusions

The conception of introns as selfish DNA has become out of date. Undoubtedly, all elements of the living world of carbonic molecules, including introns, must have some selfishness to survive under the constant pressure of natural selection. Yet, cumulative data show that, in addition, introns are helpful symbionts of eukaryotic genomes that carry out various gene regulatory functions and actively participate in gene evolution.

The question of the point in time of the origin of introns is still unresolved. The 'early' and 'late' theories represent different aspects of a complex picture of intron/exon structure and evolution and we do not see a fatal confrontation between them. The introns-early theory is a biological idea in which a simple chain of genomic events can dramatically change the course of evolution. It is clear that a process of genomic recombination, leading to gene rearrangements and the assembly of genes from 'pieces' played a crucial role in gene evolution. However, the details of these genomic recombinations at the time of the origin of life are unclear (was it 'classical' exon-shuffling or another type of nucleic acids rearrangements). On the other hand, the introns-late theory has proven its insight through revealing the properties of introns as transposons. Thus, introns have been not only passive providers of exon shuffling, but also active participants in gene evolution.

Acknowledgements

This work was particularly inspired by the ideas of Walter Gilbert concerning exon–intron evolution. We are thankful to Dr Gilbert for laboratory discussions on crucial issues of biology and evolution. We wish to thank Scott Roy for valuable comments and suggestions on our paper.

References

Akamatsu, W. & H. Okano, 2001. No to Hattatsu. Brain Dev. 33: 114–120.

Akopian, A.N., K. Okuse, V. Souslova, S. England, N. Ogata & J.N. Wood, 1999. Trans-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. FEBS Lett. 445: 177–182.

Aoki, Y., Z. Huang, S.S. Thomas, P.G. Bhide, I. Huang, M.A. Moskowitz & S.A. Reeves, 2000. Increased susceptibility to ishemia-induced brain damage in transgenic mice overexpressing a dominant negative form of SHP2. FASEB J. 14: 1965–1973.

Belford, M. & P.S. Perlman, 1995. Mechanisms of intron mobility. J. Biol. Chem. 270: 30237–30240.

Berget, S.M., C. Moore & P.A. Sharp, 1977. Spliced segments at the 5′terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. USA 74: 3171– 3175.

Black, D.L., 2000. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103: 367–370.

Blake, C.C.F., 1978. Do genes-in-pieces imply proteins-in-pieces? Nature 273: 267.

Bonen, L. & J. Vogel, 2001. The ins and outs of group II introns. Trends Genet. 17: 322–331.

Brooks, A.R., B.P. Nagy, S. Taylor, W.S. Simonet, J.M. Taylor & B. Levy-Wilson, 1994. Sequences containing the second-intron enhancer are essential for transcription of the human apolipoprotein B gene in the livers of transgenic mice. Mol. Cell. Biol. 14: 2243–2256.

Carvalho, A.B. & A.G. Clark, 1999. Intron size and natural selection. Nature 401: 344.

Caudevilla, C., C. Codony, D. Serra, G. Plasencia, R. Roman, A. Graessmann, G. Asins, M. Bach-Elias & F.G. Hegardt, 2001, Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing. Nucl. Acids Res. 29: 3108–3115.

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

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

Cavalier-Smith, T., 1991. Intron phylogeny: a new hypothesis. Trend. Genet. 7: 145–148.

Chow, L.T., R.E. Gelinas, J.R. Broker & R.J. Roberts, 1977. An amazing sequence arrangement at the 5′ends of adenovirus 2 messenger RNA. Cell 12: 1–8.

Comeron, J.M. & M. Kreitman, 2000. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics 156: 1175–1190.

Crick, F., 1979. Split genes and RNA splicing. Science 204: 264–271.

Croft, L., S. Schandroff, F. Clark, K. Burrage, P. Arctander & J.S. Mattick, 2000. ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. Nat. Genet. 24: 340–341.

Darnel, J.E., 1978. Implications of RNA. RNA splicing in evolution of eukaryotic cells. Science 202: 1257–1260.

Domon, C. & A. Steinmetz, 1994. Exon shuffling in anther-specific genes from sunflower. Mol. Gen. Genet. 244: 312–317.

Doolittle, W.F., 1978. Genes in pieces: were they ever together? Nature 272: 581–582.

Doolittle, W.F., 1999. Lateral genomics. Trends Cell Biol. 9: M5–M8.

Dorit, R.L., L. Schoengach & W. Gilbert, 1990. How big is the universe of exons? Science 250: 1377–1382.

Dorn, R., G. Reuter & A. Loewendorf, 2001. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. Proc. Natl. Acad. Sci. USA 98: 9724–9729.

Douglas, S., S. Zauner, M. Fraunholz, M. Beaton, S. Penny, L.T. Deng, X. Wu, M. Reith, T. Cavalier-Smith & U.G. Maier, 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410: 1091–1096.

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

Evans, D. & T. Blumenthal, 2000. Trans splicing of polycistronic *Caenorhabditis elegans* pre-mRNAs: analysis of the SL2 RNA. Mol. Cell. Biol. 20: 6659–6667.

Fast, N.M., A.J. Roger, C.A. Richardson & W.F. Doolittle, 1998. U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. Nucl. Acids Res. 26: 3202–3207.

Fast, N.M. & W.F. Doolittle, 1999. Trichomonas vaginalis possesses a gene encoding the essential spliceosomal component, PRP8.

Fedorov, A., G. Suboch, M. Bujakov & L. Fedorova, 1992. Analysis of nonuniformity in intron phase distribution. Nucl. Acids Res. 20: 2553–2557.

Fedorov, A., V. Starshenko, L. Fedorova, V. Filatov & E. Grigor'ev, 1998. Influence of exon duplication and shuffling on intron phase distribution. J. Mol. Evol. 46: 263–271.

Ferguson, K.C. & J.H. Rothman, 1999. Alterations in the conserved SL1 trans-spliced leader of *Caenorhabditis elegans* demonstrate flexibility in length and sequence requirements in vivo. Mol. Cell. Biol. 19: 1892–1900.

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

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

Gilbert, W., 1987. The exon theory of genes. Cold Spring Harbor Symp. Quant. Biol. 52: 901–905.

Giroux, M.J., M. Clancy, J. Baier, L. Ingham, D. McCarty & C. Hannah, 1994. De novo synthesis of an intron by the maize transposable element Dissociation. Proc. Natl. Acad. Sci. USA 91: 12150–12154.

Hartman, H. & A. Fedorov, 2002. The origin of the eukaryotic cell – a genomic investigation. Proc. Natl. Acad. Sci. USA, 99: 1420–1425.

Hogenesch, J.B., K.A. Ching, S. Batalov, A.I. Su, J.R. Walker, Y.S.A. Zhou, Kay, P.G. Schultz & M.P. Cooke, 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 106: 413–415.

Howell, M. & C.S. Hill, 1997. Xsmad2 directly activates the activin-inducible, dorsal mesoderm gene XFKH1 in *Xenopus embryos*. EMBO J. 16: 7411–7421.

Hural, J.A., M. Kwan, G. Henkel, M.B. Hock & M.A. Brown, 2000. An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells. J. Immunol. 165: 3239–3249.

Jeffreys, A.J. & R.A. Flavell, 1977. The rabbit beta-globin gene contains a large insert in the coding sequence. Cell 12: 1097–1108.

Katinka, M.D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach & C.P. Vivares, 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414: 450–453.

Kawasaki, T., S. Okumura, N. Kishimoto, H. Shimada & H. Ichikawa, 1999. RNA maturation of the rice SPK gene may involve trans-splicing. Plant J. 18: 625–632.

Krause, M. & D. Hirsh, 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell 49: 753–761.

Lambowitz, A.M. & M. Belford, 1993. Introns as mobile genetic elements. Annu. Rev. Biochem. 62: 587–622.

Liu, J. & E.S. Maxwell, 1990. Mouse U14 snRNA is encoded in an intron of the mouse cognate hsc70 heat shock gene. Nucl. Acids Res. 18, 6565– 6571.

Logsdon, J.M., M.G. Tyshenko, C. Dixon, J.D. Jafari, V.K. Walker & J.D. Palmer, 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the intronslate theory. Proc. Natl. Acad. Sci. USA 92: 8507–8511.

Logsdon, J.M., 1998. The recent origin of spliceosomal introns revised. Curr. Opin. Genet. Dev. 8: 637–648.

Logsdon, J.M., A. Stoltzfus & W.F. Doolittle, 1998. Molecular evolution: recent cases of spliceosomal intron gain? Curr. Biol. 8: R560–R563.

Long, M. & C.H. Langley, 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260: 91–95.

Long, M., C. Rosenberg & W. Gilbert, 1995. Intron phase correlations and the evolution of the intron–exon structure of genes. Proc. Natl. Acad. Sci. USA 92: 12495–12499.

Long, M., W. Wang & J. Zhang, 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei in Drosophila. Gene 238: 135–141.

Long, M. & C. Rosenberg, 2000. Testing the 'proto-splice sites' model of intron origin: evidence from analysis of intron phase correlations. Mol. Biol. Evol. 17: 1789–1796.

Long, M., 2001. Evolution of novel genes. Curr. Opin. Genet. Dev. 11: 673–680.

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

Lothian, C. & U. Lendahl, 1997. An evolutionary conserved region in the second intron of the human nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. Eur. J. Neurosci. 9: 452–462.

Lou, H., R.F. Gagel & S.M. Berget, 1996. An intron enhancer recognized by splicing factors activates polyadenylation. Genes Dev. 10: 208–219.

Maniatis, T. & R. Reed, 2002. An extensive network of coupling among gene expression machines. Nature 416: 499–506.

Martinez-Abarca, F. & N. Toro, 2000. Group II introns in the bacterial world. Mol. Microbiol. 38: 917–926.

Maxwell, E.S. & M.J. Fournier, 1995. The small nucleolar RNAs. Ann. Rev. Biochem. 35: 897–934.

Missler, M. & T.C. Sudhof, 1998. Neuroxins: three genes and 1001 products. Trends Genet. 14: 20–26.

Mitchell, J.R. & K. Collins, 2000. Human telomerase activation requires two independent interactions between telomerase RNA and telomerase reverse transcriptase. Mol. Cell 6: 361–371.

Muscarella, D.E. & V.M. Vogt, 1989. A mobile group I intron in the nuclear rDNA of *Physarum polycephalum*. Cell 56: 443–454.

Nikoh, N. & T. Fukatsu, 2001. Evolutionary dynamics of multiple group I introns in nuclear ribosomal RNA genes of endoparasitic fungi of the genus *Cordyceps*. Mol. Biol. Evol. 81: 1631–1642.

Nilsen, T.W., 2001. Evolutionary origin of SL-addition trans-splicing: still an enigma. Trends Genet. 17: 678–680.

Nixon, J.E.J., A. Wang, H.G. Morrison, A.G. McArthur, M.L. Sogin, B.J. Loftus & J. Samuelson, 2002. A Spliceosomal intron in *Giardia lamblia*. Proc. Natl. Acad. Sci. USA 99: 3701–3705.

Nurminsky, D.I., M.V. Nurminskaya, D. DeAguiar & D.L. Hartl, 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature 396: 572–575.

Oshima, R.G., L. Abrams & D. Kulesh, 1990. Activation of an intron enhancer within the keratin 18 gene by expression of c-fos and c-jun in undifferentiated F9 embryonal carcinoma cells. Genes Dev. 4: 835–848.

Palmer, J.D. & J.M. Logsdon, 1991. The recent origin of introns. Curr. Opin. Genet. Dev. 1: 470–477.

Pan, Q. & R.U. Simpson, 1999. C-myc intron element-binding proteins are required for 1,25-dihydroxyvitamin D3 regulation of c-myc during HL-60 cell differentiation and the involvement of HOXB4. J. Biol. Chem. 274: 8437–8444.

Pankov, R., A. Umezawa, R. Maki, C.J. Der, C.A. Hauser & R.G. Oshima, 1994. Oncogene activation of human keratin 18 transcription via the Ras signal transduction pathway. Proc. Natl. Acad. Sci. USA 91: 873–877.

Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling – a review. Gene 238: 103–114.

Peculis, B.A., 2000. RNA-binding proteins: if it looks like a sn(o)RNA. Curr. Biol. 10: R916–R918.

Peng,Y., A. Genin, N.B. Spinner, R.H. Diamond & R. Taub, 1998. The gene encoding human nuclear protein tyrosine phosphatase, PRL-1. Cloning, chromosomal localization and identification of an intron enhancer. J. Biol. Chem. 273: 17286–17295.

Pogacic, V., F. Dragon & W. Filipowicz, 2000. Human H/ACA small nucleolar RNPs and telomerase share evolutionarily conserved proteins NHP2 and NOP10. Mol. Cell. Biol. 20: 9028–9040.

Reed, R. & K. Magni, 2001. A new view of mRNA export: separating the wheat from the chaff. Nat. Cell Biol. 3: E201–E204.

Rhodes, K. & R.G. Oshima, 1998. A regulatory element of the human keratin 18 gene with AP-1-dependent promoter activity. J. Biol. Chem. 273: 26534–26542.

Roger, A.J. & W.F. Doolittle, 1993. Why introns-in-pieces? Nature 364: 289–290.

Saxonov, S. & W. Gilbert, 2003. The universe of exons revisited. Genetica 118: 267–278.

Schmucker, D., J. Clemens, J. Shu, C. Worby, J. Xiao, M. Muda, J. Dixon & L. Zipursky, 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101: 671–684.

Sharp, P.A., 1985. On the origin of RNA splicing and introns. Cell 42: 397–400.

Silvak, L.E., G. Pont-Kingdon, K. Le, G. Mayr, K.F. Tai, B.T. Stevens & W.L. Carroll, 1999. A novel intron element operates posttranscriptionally to regulate human N-myc expression. Mol. Cell. Biol. 19: 155–163.

Simard, M.J. & B. Chabot, 2000. Control of hnRNP A1 alternative splicing: an intron element represses use of the common 3′ splice site. Mol. Cell. Biol. 20: 7353–7362.

Takahara, T., S.I. Kanazu, S. Yanagisawa & H. Akanuma, 2000. Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic trans-splicing. J. Biol. Chem. 275: 38067–38072.

Weinsein, L.B. & J.A. Steitz, 1999. Guided tours: from precursor snoRNA to functional snoRNP. Curr. Opin. Cell Biol. 11: 378–384.