

## REVIEW

## The rise and falls of introns

R Belshaw<sup>1</sup> and D Bensasson<sup>2</sup><sup>1</sup>Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK; <sup>2</sup>Department of Biological Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK

There has been a lively debate over the evolution of eukaryote introns: at what point in the tree of life did they appear and from where, and what has been their subsequent pattern of loss and gain? A diverse range of recent research papers is relevant to this debate, and it is timely to bring them together. The absence of introns that are not self-splicing in prokaryotes and several other lines of evidence suggest an ancient eukaryotic origin for these introns, and the subse-

quent gain and loss of introns appears to be an ongoing process in many organisms. Some introns are now functionally important and there have been suggestions that invoke natural selection for the ancient and recent gain of introns, but it is also possible that fixation and loss of introns can occur in the absence of positive selection.

*Heredity* (2006) **96**, 208–213. doi:10.1038/sj.hdy.6800791; published online 1 February 2006

**Keywords:** group II intron; selfish DNA; effective population size; introns early; introns late; spliceosome

## Introduction

Nuclear genes of eukaryotes typically contain multiple regions called introns that are removed from the pre-mRNA. The remaining regions that are translated are called exons, and the process of intron removal and exon fusion is called splicing. Most intron splicing requires a complex of small nuclear RNAs and many proteins called the spliceosome (Jurica and Moore, 2003; Nilsen, 2003); hence these eukaryotic introns are often called spliceosomal introns. No evidence of these introns has been found in prokaryotes despite the sequencing of more than 100 genomes (Lynch and Richardson, 2002).

In eukaryotes, two types of spliceosome are recognised. The common U2-type splices GT-AG introns, so-called because they have GT and AG dinucleotides at the 5' and 3' end, respectively. The second is known as the U12 type, and it splices the very rare – and misleadingly named – AT-AC introns, which have a range of dinucleotides at their ends (confusingly, most often GT-AG) (Lewandowska *et al*, 2004); however, the two types are sufficiently similar for us to treat them as a single entity in this review. Eukaryotes vary considerably in the length and abundance of their introns (Deutsch and Long, 1999), yet most seem to have them and even unicellular eukaryotes have extremely complex spliceosomes for their removal (Jurica and Moore, 2003; Nilsen, 2003). This suggests that their origin is indeed ancient. Two main theories, called *Introns Early* and *Introns Late*, have been proposed to account for the origin of introns, but recently all workers have accepted that there must have been subsequent processes of loss and gain. The question of how introns arose has, therefore, been supplemented by more quantitative ones: how dynamic

are intron movements now, and are there unifying explanations for intron diversity in the different lineages of life? Many recent studies are highly pertinent to these questions, and we review them briefly below.

## Introns early... or introns late?

The discovery of introns and splicing in the 1970s led to two theories of their origin that became known as *Introns Early* and *Introns Late*.

The *Introns Early* theory proposed that introns were present in the common ancestor of prokaryotes and eukaryotes, where they were merely the genomic regions between genes (Darnell, 1978; Doolittle, 1978; Gilbert, 1978). These regions then suffered different fates in the different lineages: they were lost in all prokaryote lineages, while in eukaryotes they were maintained as introns by the appearance of the spliceosome. According to this theory, a modern protein is a concatenation of earlier, smaller proteins achieved by one of these two evolutionary processes.

Since it was first proposed, evidence has built up against the *Introns Early* theory, and Gilbert's related (1987) *Exon Theory of Genes*. Despite some early observations to the contrary, for example in the first vertebrate globin genes to be sequenced (Go, 1981), there appears to be no general match between exons and protein domains such as would be expected if today's exons represent ancient genes. Stoltzfus *et al* (1994) found no correspondence between exons and protein structure within four ancient proteins. Indeed, there is dispute about the existence of any correlation between exons and the domain structure of proteins (de Souza *et al*, 1998; Qiu *et al*, 2004). A second line of evidence proposed for the *Introns Early* theory was that approximately one-half of known introns do not interrupt the gene's reading frame (they are said to be in phase 0), as opposed to one-third that perhaps might be expected if they had been inserted at random at a later date, and that these phase 0 introns

Correspondence: R Belshaw, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

E-mail: robert.belshaw@zoo.ox.ac.uk

Received 15 September 2005; accepted 22 November 2005; published online 1 February 2006

may represent ancient intergenic regions (de Souza *et al*, 1998). However, Wolf *et al* (2001) found the same pattern among introns in genes of *Caenorhabditis elegans* that appear to have been transferred from organelles and so are likely to have acquired their introns later (see below). Also, many phase 0 introns appear to have been acquired recently, according to phylogenetic analysis (Coghlan and Wolfe, 2004; Nielsen *et al*, 2004; Qiu *et al*, 2004; Rogozin *et al*, 2003). An alternative explanation for phase 0 bias is that it can arise as a consequence of codon usage bias, because the exon nucleotides that flank introns are not random. They are commonly Gs, and thus GG pairs may be considered potential insertion sites for new introns; if random sequences are generated using eukaryote codon usage frequencies, such GG pairs are especially common across codon boundaries (Ruvinsky *et al*, 2005; but see Long *et al*, 1998). The above arguments also apply to the tendency for adjacent introns to be in the same phase as each other.

The *Introns Late* theory, in contrast, proposed that spliceosomal introns only appeared in eukaryotes, where they were derived from self-splicing introns that invaded previously undivided genes, and that the spliceosome evolved as a way of removing them (Cavalier-Smith, 1991; Palmer and Logsdon, 1991; Boeke, 2003). Self-splicing introns (sometimes called retrointrons) are a type of genomic parasite: they insert themselves into the host genome and, when transcribed, their RNA catalyses its own excision – although sometimes assisted by a protein translated from within the intron (Lambowitz and Zimmerly, 2004). The argument for self-splicing introns having given rise to spliceosomal introns and their spliceosomes is based on similarities of function and structure, combined with parallels between the present taxonomic distribution of self-splicing introns and the likely origin of the eukaryote cell. More specifically, one type of self-splicing intron, called group II introns, and spliceosomal introns have similar splicing mechanisms: in both, the 5' end of the intron becomes bound to an adenine near its 3' end to form a lariat (lasso) structure that is excised. This similarity is highlighted by replacement experiments: the splicing activity of a group II intron from which a particular stem-loop sequence has been removed can be restored by addition of the RNA molecule that appears to serve the same function in the spliceosome (Hetzer *et al*, 1997). There are also similarities in the structure of group II introns and spliceosomes, although they are insufficient to prove their common ancestry (Lynch and Richardson, 2002; Villa *et al*, 2002).

The taxonomic distribution of group II introns is also consistent with them having given rise to spliceosomal introns. Group II introns were first discovered in yeast mitochondria, although uncharacterised at that time (Bonitz *et al*, 1980). They were later found to parasitise the organelles of many eukaryotes (except for those of animals) as well as many eubacteria (Ferat and Michel, 1993), where typically they are found in plasmids and between rather than within genes. Significantly, they appear to be absent from most archaeobacteria, and their presence in one genus of archaeobacteria in which they have been found (the large-genomed *Methanosarcina*) appears to be a secondary acquisition caused by multiple horizontal transfers from eubacteria (Rest and Mindell, 2003). The eukaryote nuclear genome is thought to share

its most recent common ancestor with the archaeobacteria (Brown and Doolittle, 1995) and the eukaryote organellar genomes are thought to share common ancestors with various eubacteria (Gray, 1999). There was a large-scale transfer of genes from these organelles to the nucleus (Gray *et al*, 1999), which may have led to the introduction of group II intron-like elements in the eukaryote nucleus. The spliceosome could then have arisen and spread through the fragmentation of a group II intron, even in the absence of positive selection, through a series of steps reviewed in Lynch and Richardson (2002). To summarise, the *Introns Late* theory proposes that spliceosomal introns arose in the first eukaryotes from group II intron-like elements present in their endosymbiotic organelles. This is more parsimonious than the *Introns Early* theory's requirement that there has been secondary loss of all ancient introns from both eubacteria and archaeobacteria.

## Colonising the genome

One problem with the origin of spliceosomal introns from group II introns, and in understanding their current dynamics, is to explain how such introns could have spread throughout the eukaryote nuclear genome. Today, group II introns have a homing mechanism that restricts them to the same locus in the host, and which typically is in an organelle or plastid (Dai *et al*, 2003). This problem has recently been overcome. In the bacterium *Lactococcus lactis*, Cousineau *et al* (2000) found evidence for movement to novel chromosomal locations by the group II intron *LtrB*, which usually is found in a plasmid. The introns appear to colonise new loci in the host chromosome by reverse splicing directly into the DNA sequence (Cousineau *et al*, 2001; Ichiyanagi *et al*, 2002). Also, in natural populations of several other taxa, we see group II introns with almost identical sequences occupying multiple locations (in the same or different hosts) that share only low sequence similarity, indicating that some movement to new loci has occurred (Dai and Zimmerly, 2002).

## The problems of phylogenetic reconstruction

As multiple genes from a range of taxa have become available, some researchers have used phylogenetic methods to examine the pattern of intron gains and losses. Rogozin *et al* (2003) analysed eight divergent eukaryote taxa and found major differences between the two rates in the different lineages. By using the same dataset, Roy and Gilbert (2005a, b) suggested that – following an initial rapid gain in their common ancestor – there has been overall intron loss in most lineages. However, Qiu *et al* (2004) estimated that, for 95% of the introns in 10 gene families (a total of 677 gene sequences taken from many taxa), the probability that they were in the most recent common ancestor of the gene was below 0.05.

The taxa in the above studies are so distantly related that shared introns will have lost any sequence similarity. Such studies rely instead on the assumption that introns in the same position in genes are homologous, that is, are descended directly from a common ancestor. However, Sadusky *et al* (2004) found that, if you removed experimentally the splice donor site from the 5' end of the intron in actin genes from Human,

*Arabidopsis* (a herbaceous plant) and *Physarum* (a slime mould), a number of alternate splice sites are created. These 'cryptic splice sites' tend to have the AG|GT motif (corresponding to the typical exon|intron and intron|exon boundaries) and, more importantly, eight out of nine sites corresponded to known intron sites in other taxa (Sadusky *et al*, 2004; Stoltzfus, 2004). Thus, introns in the same site in distantly related taxa may have been acquired in parallel and not be homologous.

There are additional problems with phylogenetic analyses involving distantly related taxa: there is a strong sampling bias due to our tendency to sequence organisms with smaller genomes, and the phylogenies are poorly known – much more poorly than is generally recognised (Philippe *et al*, 2005; Philippe *et al*, 2005).

## Mechanisms of intron loss

Comparisons of genomes typically reveal losses of varying sized regions through evolutionary time, and introns would not be immune from this. Like other noncoding DNA regions, *Drosophila melanogaster* introns appear to lose DNA by the accumulation of small deletions (Parsch, 2003). Llopart *et al* (2002) found an example of intron deletion in *D. teissieri* (leaving a 12 bp fragment that now adds four amino acids to the protein). Also, introns are lost when new genes are created by reverse transcription of (intron-less) mRNA followed by insertion of the cDNA into a new genomic location, although this process appears to be rare – at least in *Drosophila* (Betrán *et al*, 2002). Another mechanism that may have operated is reverse transcription followed by gene conversion. Fink (1987) suggested that, as in the formation of processed pseudogenes, some copies of processed mRNA are reverse transcribed into cDNA by reverse transcriptase derived from retroelements in the cell. Homologous recombination between the original gene and the now intronless cDNA may then bring about loss of the intron. Reverse transcription followed by gene conversion would be expected to remove preferentially introns from the 3' end of the gene. This trend would arise because transcription is affected by length-dependent dissociation starting at the 3' end, which produces cDNAs truncated at their 5' end – the tendency has been used to explain the pattern of truncation observed in processed pseudogenes (Pavliček *et al*, 2002). Such a mechanism may explain why introns tend to be concentrated near the 5' end of genes in intron-sparse genomes (Mourier and Jeffares, 2003), although it remains unclear why they are evenly distributed in intron-rich genomes. This mechanism also predicts a 3' bias in intron loss. Evidence for such a trend is uncertain. In their analysis of intron gain and loss among seven diverse taxa, Roy and Gilbert (2005a) found that introns near the 3' end of genes were indeed more likely to be lost, but Nielsen *et al* (2004) found no discernable 3' bias for intron loss in three filamentous fungal lineages.

## Mechanisms of intron gain

It appears that introns can give rise to other introns. *C. elegans* and *C. briggsae* diverged around 100 million years ago, and over 250 introns are present in one species but absent in the other (Kent and Zahler, 2000). By determining that there was no intron at the same location

in a range of outgroups, Coghlan and Wolfe (2004) were able to identify those introns that had been gained in one of the two lineages. Of 122 such introns, 28 could be identified by sequence similarity to have originated from other introns that were in either the same gene or other genes within the same organism (three and 25, respectively). There are other examples of apparent duplication of existing introns, for example within the xanthine dehydrogenase gene in *Drosophila* (Tarrío *et al*, 1998).

In their analysis of intron gain in *C. elegans* and *C. briggsae*, Coghlan and Wolfe (2004) favoured reverse-splicing as the predominant mechanism. If the spliceosome remains combined with a recently excised intron and attaches itself to an unoccupied but potentially functional splice site of the same or another pre-mRNA, then it might – instead of splicing out the intron – catalyse the reverse reaction. If this pre-mRNA with its novel intron is then reverse transcribed, its cDNA might recombine with its homologous DNA to insert the intron into a novel chromosomal location. The main evidence for such a mechanism from this study is that, of the genes that gained introns, more were expressed in the germline and more were inferred to be involved in mRNA processing functions than would be expected by chance alone.

## The diversity of introns

Evidence is mounting that introns cannot be considered as uniform. Commonly, we see a bimodal distribution of intron length with a high peak of short (termed 'minimal length') introns and a much flatter peak of longer introns, ranging up to thousands of base pairs in length in humans (Yu *et al*, 2002). Furthermore, in some organisms there appears to be a functional link between intron length and gene expression, with introns tending to be smaller in highly expressed genes – possibly as a result of a need for transcriptional efficiency. This appears to be the case in humans, where it is linked to other evidence of selection for transcription efficiency (Urrutia and Hurst, 2003). The relationship is more marked when only one copy of the gene is expressed (and hence intron length variations in individual alleles are more likely to affect fitness), for example in imprinted genes – where the gene from only one parent is expressed (Hurst *et al*, 1996) – and in *Arabidopsis* genes expressed in the (haploid) pollen (Seoighe *et al*, 2005). However, the relationship between intron length and expression is reversed in baker's yeast (*Saccharomyces cerevisiae*), with highly expressed genes having longer introns (Vinogradov, 2001). This may indicate some (unknown) functional role for introns in gene expression in this organism, but the situation is complicated by the fact that *S. cerevisiae* has few introns and these are concentrated in highly expressed ribosomal protein genes (Ares *et al*, 1999).

In addition to the GT and AG motifs at the 5' and 3' end of the intron, respectively, and the adenine for lariat formation, there appear to be other signalling motifs within introns. Bergman and Kreitman (2001) found that long introns in *Drosophila* were subject to a level of sequence constraint that was indistinguishable from that seen in intergenic regions thought to be involved in cis-regulation. Haddrill *et al* (2005) also found in *Drosophila* a strong negative correlation between intron length and sequence divergence, with apparent selective constraint

on longer introns while short introns appear to evolve at a similar rate to synonymous sites. In humans (Majewski and Ott, 2002) and in rat and mouse (Keightley and Gaffney, 2003) the levels of sequence variation are surprisingly low in some intron regions, suggesting that the sequence in these regions is under selective constraint, and the intron nearest to the 5' end of the gene appears to evolve slower than the others.

Given that some introns may have functional roles while others appear to be under selective pressure to reduce their length, it is likely that the dynamics of gains and loss of such different introns will differ.

## The role of natural selection in the rise and falls of introns

From its inception, the *Introns Early* theory stressed the role of selection in creating the diversity and taxonomic distribution of introns. In prokaryotes, the hypothesised intergenic regions became lost when a proposed increase in transcriptional efficiency allowed larger genes to be transcribed – a form of genomic streamlining (Gilbert, 1987). In eukaryotes, however, the evolution of the spliceosome – and the conversion of these intergenic regions into introns – allowed genes to be combined while gaining a proposed evolutionary plasticity from the appearance of alternate splicing, exon shuffling and enhanced within-gene recombination. Another variant of the *Introns Early* theory proposed that introns originally played an important role in detecting copying errors by forming stem-loops during meiosis to facilitate recombination (Barrette *et al*, 2001). Although we argue above that the evidence points to a later origin of introns, many authors have suggested that they are maintained by indirect positive selection on lineages for mechanisms such as alternate splicing, etc; the evolution of such 'evolvability' traits is currently under investigation (Radman *et al*, 1999). The process of splicing has now become closely integrated with the production of mature mRNAs (Maniatis and Reed, 2002), and introns play a role in the detection of prematurely terminated RNAs through a mechanism called Nonsense Mediated Decay (or NMD). Indeed, Lynch and Kewalramani (2003) found that the spatial distribution of introns within genes was consistent with maximising the efficiency of NMD (eg introns were overdispersed), and they suggested that this resulted from positive selection for certain intron insertions once NMD had evolved. There are also examples of gene expression signals migrating to introns, and excised intron RNA molecules being used by the organism (Fedorova and Fedorov, 2003). These current roles for introns are such that for most eukaryotes a secondary evolutionary loss of introns is probably impossible – although this might still be viewed as a co-option of sequences that were parasitic in origin, for example as has occurred in the case of some endogenous retroviruses (Bock and Stoye, 2000).

In the spirit of Occam's razor, Lynch (2002) proposed a nearly neutral alternative to hypotheses invoking positive selection, which explains the taxonomic abundance and distribution of introns in terms of population size. Introns may often impose a cost, for example wastage of resources in transcribing introns and then splicing them out, and disrupting a gene with an insertion that may not

always be spliced accurately. In one of the rare examples of a natural population found to be polymorphic for the presence/absence of an intron, Llopart *et al* (2002) found evidence of selection against the intron-present allele in the *jingwei* gene in *D. teissieri*. If we consequently regard introns as weakly deleterious mutations for the host, some will occasionally drift to fixation provided that the host population is not too large, which may well be the case in eukaryotes but not prokaryotes. The much larger population size in prokaryotes may then explain both the general rarity of their mobile self-splicing introns (such as group II introns) compared to the abundance of spliceosomal introns in eukaryotes, and the failure of anything similar to spliceosomal introns to evolve within prokaryotes. Also, the fact that introns tend to be rarer in unicellular eukaryotes compared to multicellular eukaryotes may reflect the former's larger population sizes.

## Conclusions and prospects

It appears to us most likely that spliceosomal introns are descended from group II-like introns that were present within the organelles of the first eukaryotes, but there have been many gains and losses of introns in all lineages since then. Eukaryotes vary greatly in the distribution, length and structure of their introns and so we can expect that the rates of intron gain and loss will vary depending upon both the lineage and the intron type.

This is an exciting time for the study of introns. Although we believe that at present studies relying on phylogenies of distantly related taxa are problematic, as more closely related genomes are sequenced the gains and losses of introns will be much better resolved (Coghlan and Wolfe, 2004). Well-sequenced taxonomic groups such as *Drosophila*, *Saccharomyces*, and primates are ideal for such work. Advances in biochemistry promise to improve our understanding of how spliceosomes work and should improve our understanding of possible constraints on the genic positions of introns, and this will in turn inform evolutionary studies. The proposed role of reverse transcription in intron gain and loss can be examined by a large-scale comparison of genes that are expressed in the germline with those that are not. As genome sequences gather and our knowledge of intron distributions improves, there will also be increasing scope for the use of estimates of effective population size to examine the relationship between population size and intron density and test hypotheses concerning the role of selection.

## Acknowledgements

We would like to thank Casey Bergman and Aris Katzourakis and two anonymous reviewers for their comments on the manuscript.

## References

- Ares Jr M, Grate L, Pauling MH (1999). A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* 5: 1138–1139.
- Barrette IH, McKenna S, Taylor DR, Forsdyke DR (2001). Introns resolve the conflict between base order-dependent stem-loop

- potential and the encoding of RNA or protein: further evidence from overlapping genes. *Gene* **270**: 181–189.
- Bergman CM, Kreitman M (2001). Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**: 1335–1345.
- Betrán E, Thornton K, Long M (2002). Retroposed new genes out of the X in *Drosophila*. *Genome Res* **12**: 1854–1859.
- Bock M, Stoye JP (2000). Endogenous retroviruses and the human germline. *Curr Opin Genet Dev* **10**: 651–655.
- Boeke JD (2003). The unusual phylogenetic distribution of retrotransposons: a hypothesis. *Genome Res* **13**: 1975–1983.
- Bonitz SG, Coruzzi G, Thalenfeld BE, Tzagoloff A, Macino G (1980). Assembly of the mitochondrial membrane system. Structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrome oxidase. *J Biol Chem* **255**: 11927–11941.
- Brown JR, Doolittle WF (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci USA* **92**: 2441–2445.
- Cavalier-Smith T (1991). Intron phylogeny: a new hypothesis. *Trends Genet* **7**: 145–148.
- Coghlan A, Wolfe KH (2004). Origins of recently gained introns in *Caenorhabditis*. *Proc Natl Acad Sci USA* **101**: 11362–11367.
- Cousineau B, Lawrence S, Smith D, Belfort M (2000). Retrotransposition of a bacterial group II intron. *Nature* **404**: 1018–1021.
- Cousineau B, Lawrence S, Smith D, Belfort M (2001). Correction: Retrotransposition of a bacterial group II intron. *Nature* **414**: 84.
- Dai L, Toor N, Olson R, Keeping A, Zimmerly S (2003). Database for mobile group II introns. *Nucleic Acids Res* **31**: 424–426.
- Dai L, Zimmerly S (2002). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* **30**: 1091–1102.
- Darnell Jr JE (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science* **202**: 1257–1260.
- de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998). Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA* **95**: 5094–5099.
- Deutsch M, Long M (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* **27**: 3219–3228.
- Doolittle WF (1978). Genes in pieces: were they ever together? *Nature* **272**: 581–582.
- Fedorova L, Fedorov A (2003). Introns in gene evolution. *Genetica* **118**: 123–131.
- Ferat JL, Michel F (1993). Group II self-splicing introns in bacteria. *Nature* **364**: 358–361.
- Fink GR (1987). Pseudogenes in yeast? *Cell* **49**: 5–6.
- Gilbert W (1978). Why genes in pieces? *Nature* **271**: 501.
- Gilbert W (1987). The exon theory of genes. *Cold Spring Harb Symp Quant Biol* **52**: 901–905.
- Go M (1981). Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**: 90–92.
- Gray MW (1999). Evolution of organellar genomes. *Curr Opin Genet Dev* **9**: 678–687.
- Gray MW, Burger G, Lang BF (1999). Mitochondrial evolution. *Science* **283**: 1476–1481.
- Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P (2005). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol* **6**: R67.
- Hetzer M, Wurzer G, Schweyen RJ, Mueller MW (1997). Trans-activation of group II intron splicing by nuclear U5 snRNA. *Nature* **386**: 417–420.
- Hurst LD, McVean G, Moore T (1996). Imprinted genes have few and small introns. *Nat Genet* **12**: 234–237.
- Ichihyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M (2002). Retrotransposition of the LL.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. *Mol Microbiol* **46**: 1259–1272.
- Jurica MS, Moore MJ (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell* **12**: 5–14.
- Keightley PD, Gaffney DJ (2003). Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. *Proc Natl Acad Sci USA* **100**: 13402–13406.
- Kent WJ, Zahler AM (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*–*C. elegans* genomic alignment. *Genome Res* **10**: 1115–1125.
- Lambowitz AM, Zimmerly S (2004). Mobile group II introns. *Annu Rev Genet* **38**: 1–35.
- Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin CF *et al* (2004). Determinants of plant U12-dependent intron splicing efficiency. *Plant Cell* **16**: 1340–1352.
- Llopart A, Cameron JM, Brunet FG, Lachaise D, Long M (2002). Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc Natl Acad Sci USA* **99**: 8121–8126.
- Long M, de Souza SJ, Rosenberg C, Gilbert W (1998). Relationship between ‘proto-splice sites’ and intron phases: evidence from dicodon analysis. *Proc Natl Acad Sci USA* **95**: 219–223.
- Lynch M (2002). Intron evolution as a population-genetic process. *Proc Natl Acad Sci USA* **99**: 6118–6123.
- Lynch M, Kewalramani A (2003). Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol Biol Evol* **20**: 563–571.
- Lynch M, Richardson AO (2002). The evolution of spliceosomal introns. *Curr Opin Genet Dev* **12**: 701–710.
- Majewski J, Ott J (2002). Distribution and characterization of regulatory elements in the human genome. *Genome Res* **12**: 1827–1836.
- Maniatis T, Reed R (2002). An extensive network of coupling among gene expression machines. *Nature* **416**: 499–506.
- Mourier T, Jeffares DC (2003). Eukaryotic intron loss. *Science* **300**: 1393.
- Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE (2004). Patterns of intron gain and loss in fungi. *PLoS Biol* **2**: e422.
- Nilsen TW (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* **25**: 1147–1149.
- Palmer JD, Logsdon Jr JM (1991). The recent origins of introns. *Curr Opin Genet Dev* **1**: 470–477.
- Parsch J (2003). Selective constraints on intron evolution in *Drosophila*. *Genetics* **165**: 1843–1851.
- Pavliček A, Pačes J, Zika R, Hejnar J (2002). Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. *Gene* **300**: 189–194.
- Philip GK, Creevey CJ, McInerney JO (2005). The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Mol Biol Evol* **22**: 1175–1184.
- Philippe H, Lartillot N, Brinkmann H (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* **22**: 1246–1253.
- Qiu WG, Schisler N, Stoltzfus A (2004). The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol Biol Evol* **21**: 1252–1263.
- Radman M, Matic I, Taddei F (1999). Evolution of evolvability. *Ann NY Acad Sci* **870**: 146–155.
- Rest JS, Mindell DP (2003). Retroids in archaea: phylogeny and lateral origins. *Mol Biol Evol* **20**: 1134–1142.
- Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003). Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* **13**: 1512–1517.
- Roy SW, Gilbert W (2005a). The pattern of intron loss. *Proc Natl Acad Sci USA* **102**: 713–718.

- Roy SW, Gilbert W (2005b). Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc Natl Acad Sci USA* **102**: 5773–5778.
- Ruvinsky A, Eskesen ST, Eskesen FN, Hurst LD (2005). Can codon usage bias explain intron phase distributions and exon symmetry? *J Mol Evol* **60**: 99–104.
- Sadusky T, Newman AJ, Dibb NJ (2004). Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr Biol* **14**: 505–509.
- Seoighe C, Gehring C, Hurst LD (2005). Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. *PLoS Genet* **1**: e13.
- Stoltzfus A (2004). Molecular evolution: introns fall into place. *Curr Biol* **14**: R351–R352.
- Stoltzfus A, Spencer DF, Zuker M, Logsdon Jr JM, Doolittle WF (1994). Testing the exon theory of genes: the evidence from protein structure. *Science* **265**: 202–207.
- Tarrío R, Rodríguez-Trelles F, Ayala FJ (1998). New *Drosophila* introns originate by duplication. *Proc Natl Acad Sci USA* **95**: 1658–1662.
- Urrutia AO, Hurst LD (2003). The signature of selection mediated by expression on human genes. *Genome Res* **13**: 2260–2264.
- Villa T, Pleiss JA, Guthrie C (2002). Spliceosomal snRNAs: Mg(2+)-dependent chemistry at the catalytic core? *Cell* **109**: 149–152.
- Vinogradov AE (2001). Intron length and codon usage. *J Mol Evol* **52**: 2–5.
- Wolf YI, Kondrashov FA, Koonin EV (2001). Footprints of primordial introns on the eukaryotic genome: still no clear traces. *Trends Genet* **17**: 499–501.
- Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK (2002). Minimal introns are not ‘junk’. *Genome Res* **12**: 1185–1189.