The evolution of spliceosomal introns: patterns, puzzles and progress

Scott William Roy[*] *and Walter Gilbert*[‡]

Abstract | The origins and importance of spliceosomal introns comprise one of the longest-abiding mysteries of molecular evolution. Considerable debate remains over several aspects of the evolution of spliceosomal introns, including the timing of intron origin and proliferation, the mechanisms by which introns are lost and gained, and the forces that have shaped intron evolution. Recent important progress has been made in each of these areas. Patterns of intron-position correspondence between widely diverged eukaryotic species have provided insights into the origins of the vast differences in intron number between eukaryotic species, and studies of specific cases of intron loss and gain have led to progress in understanding the underlying molecular mechanisms and the forces that control intron evolution.

Nonsense-mediated decay
A mechanism by which a stop codon that is encountered by the ribosome upstream of an intron–exon boundary leads to degradation of the transcript.

Introns are genomic sequences that are removed from the corresponding RNA transcripts of genes. Group I and II introns are both found in some bacterial and organellar genomes, and group I introns are also found in ribosomal RNAs (rRNAs) of protist and fungal nuclei[1–3]. These two groups have distinct RNA structures that facilitate their self-splicing activity. They also contain internal ORFs, which facilitate both intron removal from RNA transcripts and intron propagation to intronless sites through reverse transcription. In total, around 1,500 group I and 200 group II introns have been identified[1].

By contrast, a third group of introns — spliceosomal introns — are found in the nuclear genomes of all characterized eukaryotes. They have quasi-random sequences and generally lack ORFs. Their lengths vary widely between species, from just tens of bases in some protists to hundreds of kilobases in mammals. The spliceosome, a complex that comprises five RNAs and hundreds of proteins, removes spliceosomal introns from RNA transcripts, a process that is coupled to several other transcript-processing steps[4]. Despite important differences between spliceosomal and other introns, similarities between the splicing mechanisms of group II and spliceosomal introns indicate a possible evolutionary relationship between the two[5–9].

The timing and causes of spliceosomal intron evolution are matters of great interest in the study of genome evolution as a whole. Spliceosomal introns are absent in prokaryotes and their numbers vary tremendously between eukaryotic species, from fewer than 100 introns per genome in some species to hundreds of thousands per genome in vertebrates and plants (FIG. 1). However, despite the huge numbers of intron gains and/or losses that are implied by these differences, there is less certainty about the mechanisms and forces that underlie intron gain and loss than about any other major class of genetic element. There are millions of known introns in coding regions, but there is only one known intraspecific presence/absence polymorphism[10], and there are only three recently inserted introns for which the origins have been confidently traced[11–13]. Currently, several plausible hypotheses compete to explain the origin of new introns, and no consensus has been reached as to whether introns are positively, negatively or neutrally selected.

The vast interspecific differences in intron number between eukaryotic genomes constitute an important puzzle with which theories about the determinants of genome complexity — for example, selfish genetic elements[14–16], organismal complexity[15–19] or population size[20,21] — must come to terms. Indeed, introns have prominent roles in several ambitious theories of genome evolution. A long-standing theory of intron origin postulates that recombination within introns facilitated the construction of the first full-length genes[17,22–38], and a similar theory has been proposed for the origins of many multidomain metazoan genes[39,40]. Introns have also been suggested to increase fitness by increasing intragenic recombination[22,41–43] or to have boosted transcript fidelity in early eukaryotes through nonsense-mediated decay (NMD)[44]. Differences in intron number have also served as the potential crowning example of the hypothesis that increases in genome complexity are often the results of,

[*]*Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.*
[‡]*15 Gray Gardens West, Cambridge, Massachusetts 01238, USA.*
Correspondence to S.W.R.
e-mail: scottwroy@gmail.com
doi:10.1038/nrg1807

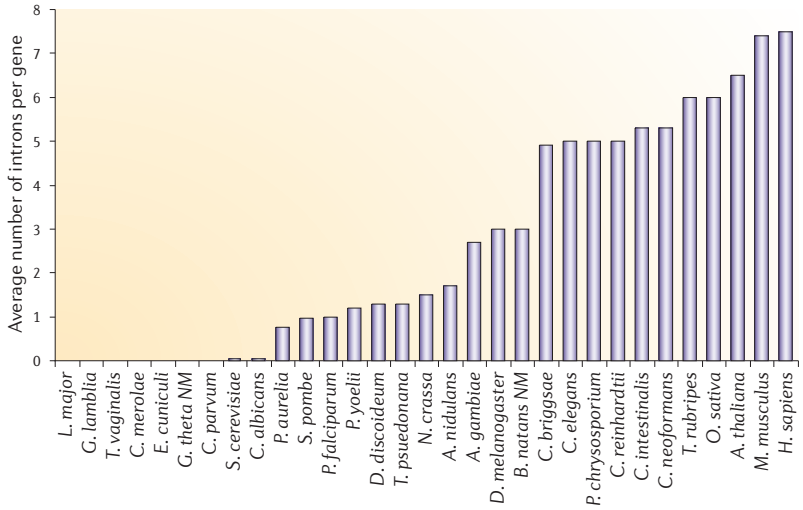


Figure 1 | **Distribution of spliceosomal introns in eukaryotic species.** The number of introns per gene is shown for a selection of well-characterized eukaryotic species, derived from a survey of the literature. The full names of the species shown are as follows: *Anopheles gambiae*; *Arabidopsis thaliana*; *Aspergillus nidulans*; *Bigelowiella natans* Nucleomorph; *Caenorhabditis briggsae*; *Caenorhabditis elegans*; *Candida albicans*; *Chlamydomonas reinhardtii*; *Ciona intestinalis*; *Cryptococcus neoformans*; *Cryptosporidium parvum*; *Cyanidioschyzon merolae*; *Dictyostelium discoideum*; *Drosophila melanogaster*; *Encephalitozoon cuniculi*; *Giardia lamblia*, *Guillardia theta* Nucleomorph; *Homo sapiens*; *Leishmania major*; *Mus musculus*; *Neurospora crassa*; *Oryza sativa*; *Paramecium aurelia*; *Phanerochaete chrysosporium*; *Plasmodium* falciparum; *Plasmodium yoelii*; *Saccharomyces cerevisiae*; *Schizosaccharomyces pombe*; *Takifugu rubripes*; *Thalassiosira pseudonana*; *Trichomonas vaginalis*.

or are themselves, deleterious mutations[20,21]. Therefore, the evolution of spliceosomal introns has broad implications for many fundamental evolutionary questions.

Research into the timing, mechanisms and causes of spliceosomal intron evolution has been extremely active in the past few years, resolving some old controversies and sparking some new ones. Here we discuss recent studies of the rise and fall of intron number through eukaryotic evolution, mechanisms of intron gain and loss, and the evolutionary forces that might be responsible for these changes.

The timing of intron evolution

The introns early–introns late debate. There are two main, long-standing alternative explanations for the origin of introns. The introns-early (IE) model proposes that introns are extremely old, and were numerous in the ancestors of eukaryotes and prokaryotes[17,22–38], with introns allowing the modular assembly of very early full-length genes from shorter exon-encoded fragments through 'exon shuffling'. In this model, introns were then lost from prokaryotic genomes. By contrast, according to the introns-late (IL) model, the phylogenetic restriction of spliceosomal introns to eukaryotes reflects their more recent insertion into originally intronless genes after the divergence of eukaryotes and prokaryotes[7–9,45–56].

With accumulating evidence, the more extreme versions of these two models — which view nearly all introns as being either extremely old or new — have yielded to more nuanced perspectives. Currently, IE adherents postulate that only a minority of modern introns predate the eukaryote–prokaryote split[31–38], whereas most IL supporters believe that introns evolved from type II bacterial introns in relatively early eukaryotes[7–9]. However, vigorous debate continues about both the presence of introns in prokaryote–eukaryote ancestors and the relative importance of intron loss and intron gain in eukaryotic evolution. We begin our discussion with the key issues of intron conservation, loss and gain.

Patterns of intron retention, gain and loss. The massive variation in intron number among eukaryotic species shows no simple phylogenetic pattern, with intron-rich and intron-poor species interspersed in the eukaryotic phylogenetic tree. This pattern implies recurrent episodes of massive intron loss and/or gain. At one extreme, the common ancestors of intron-rich and intron-poor species could have been intron-poor, with intron-rich species having undergone more recent insertions. In this case, intron-position correspondence between distant species should be relatively rare. At the other extreme, nearly all modern introns could be inherited from intron-rich ancestors, with intron-poor species having experienced massive intron loss. In this case, intron-position coincidence between widely diverged species might be expected to be nearly complete.

The actual degree of intron-position correspondence lies between these two extremes[57,58]. Rogozin and co-workers found significant but incomplete correspondence of intron positions in 684 sets of orthologous genes from 8 species with fully sequenced genomes[58]. They found that 25% of human introns are at the exact same position (between the homologous pair of nucleotides in the alignment) as an intron in the orthologous gene from *Arabidopsis thaliana*, and that 40% of *Schizosaccharomyces pombe* intron positions match an intron position from a non-fungus. On the other hand, 20–68% of introns in a species are specific to that species. Together, these results imply considerable intron gain and/or loss over the past hundreds of millions of years.

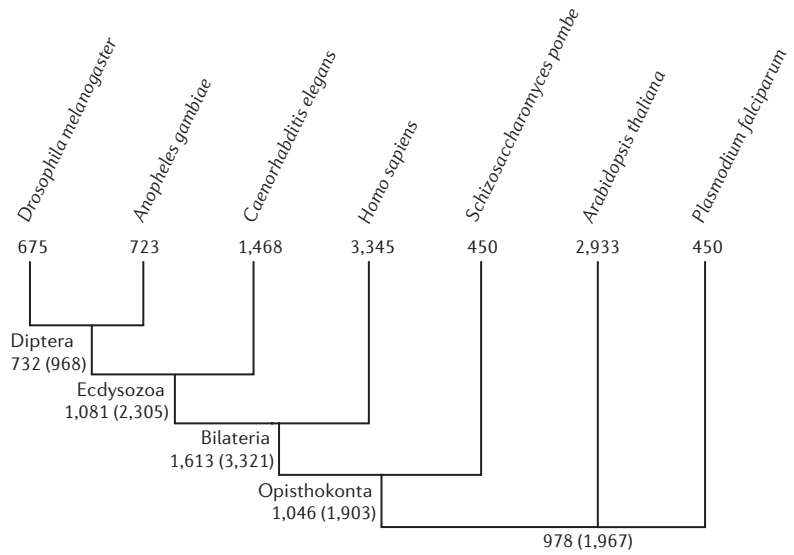
Is it possible that these intron-position correspondences are due to independent insertions into the homologous position along different lineages[47,59–67]? Occasional cases of such 'parallel insertion' have been documented[59], and there is accumulating evidence that intron insertions 'prefer' certain sequences[47,60–67], increasing the possibility of such multiple insertions. However, parallel insertion seems unlikely to explain a significant fraction of observed intron-position correspondences, as simulations of targeted intron insertion for genes in the Rogozin *et al.* data set showed only 5–10% as many correspondences as are actually observed[68]. The number of actual parallel insertions could be even lower: if many introns have been retained from ancestral species, there has been less subsequent insertion than the simulations assumed, and therefore fewer parallel insertions.

However, even if most intron-position correspondences do represent ancestral introns, there is still disagreement about the meaning of the observed patterns of correspondence. Rogozin *et al.* used Dollo parsimony

Exon shuffling
A process by which ectopic recombination within introns leads to the creation of new genetic products.

Dollo parsimony
A method in which a character (in this case an intron position) is inferred to have arisen exactly once on the evolutionary tree in the ancestor of the most distantly related pair of species that share the character. Absence of the character in descendents of this ancestor is then explained by the minimal pattern of losses necessary to explain the observed phylogenetic distribution.

a



b

	Opisthokonta	Bilateria	Ecdysozoa	Diptera
D. melanogaster	27 (37)	64 (70)	68 (75)	80 (87)
A. gambiae	28 (35)	63 (69)	67 (73)	80 (87)
C. elegans	19 (31)	43 (52)	46 (51)	
H. sapiens	27 (41)	45 (75)		
S. pombe	55 (60)			

Figure 2 | **Intron conservation, gain and loss for 684 sets of eukaryotic orthologues.** The extent of intron conservation and change is shown for conserved regions of 684 sets of eukaryotic orthologues[58]. **a** | For each modern species studied, the numbers of introns present in analysed regions are indicated. For each ancestral node, two estimates for the number of introns in analysed regions are given. The first is derived from a Dollo parsimony analysis[58]. The second, in parentheses, is derived from a maximum likelihood analysis, assuming no parallel insertion and equal rates of loss for different introns along the same lineage[69]. The estimates at the base of the tree are for the most recent common ancestor of two of the three unresolved groups (opistokonts (animals and fungi), plants and apicomplexans (including *Plasmodium falciparum*)), which might be either the plant–apicomplexan or plant–opisthokont ancestor. Data are from REFS 58,69. **b** | The age of modern introns. Values indicate the percentage of introns in the genomes of the modern species listed on the left that are estimated to have been present in the ancestor (for example, the *Drosophila melanogaster*–dipteran entry gives the percentage of studied introns in the *D. melanogaster* genome that are estimated to have been present in the dipteran ancestor). As in panel **a**, the first value gives estimates that are derived from a parsimony analysis[58], the second gives estimates from a maximum likelihood analysis, assuming no parallel insertion and equal rates of loss across different introns[69]. In the parsimony case, an intron position in a modern species is assumed to represent an intron that was present in the ancestor if the position is shared with another species from which the first species diverged at or before the time of origin of that ancestor. The ancestors are as indicated in the tree in panel **a**. Data are from REFS 58,69.

Maximum likelihood analysis
A statistical method that finds the maximum of the likelihood function given a set of data, where the likelihood function gives the probability of obtaining the data for a set of unknown variables.

to estimate numbers of introns that were present in the common ancestors of the species that they studied, as well as numbers that were subsequently lost or gained along the respective evolutionary branches (FIG. 2a). They argued for important roles for both ancestral intron retention and massive, lineage-specific episodes of intron loss and gain[58]. However, this approach does not correct for the apparently widespread intron loss in the data set[69]. For example, among the 927 introns that are shared between animals and *A. thaliana* or *Plasmodium falciparum* — which were therefore probably present in the fungus–animal ancestor — only 14% are retained in either *S. pombe* or *Saccharomyces cerevisiae*[69]. This suggests massive intron loss in these fungi. Similarly, only 37% of the 907 introns shared between humans and a non-animal are found in a second animal, suggesting independent massive loss in animals. These losses will lead to the underestimation of ancestral intron numbers using the parsimony method[69].

We carried out a maximum likelihood analysis that incorporated intron loss to analyse the same data set, which suggested that, in general, the common ancestors of the species in the data set contained many introns — nearly as many as are found in the most intron-dense modern organisms[69] (FIG. 2a). These estimates indicate that intron number has decreased along many lineages, leading to more moderate intron densities in modern species. However, before these results are universally accepted, further work will be necessary[68,70] to better understand the reliance of these estimates on two important assumptions: that all shared intron positions reflect ancestral introns and that all introns are lost at equal rates along a given branch.

Both the parsimony and maximum likelihood estimates attest to an important role for intron loss, with some branches undergoing more intron loss than gain. The possibility of recurrent decreases in intron number along diverse lineages is perhaps surprising; however, studies of more closely related species have shown the same pattern. Six intron losses but no gains were found among 10,000 intron positions in rodents and humans[71], and several studies have found more intron losses than gains among species of *Caenorhabditis*[72–74], *Plasmodium* and *Drosophila* (S.W.R. and D.L. Hartl, unpublished observations). Moreover, the apparent conservation of significant numbers of ancestral introns in *S. pombe*, *A. thaliana*, and *P. falciparum* implies massive intron-number reduction in the related, near-intronless species (in corresponding order) *Encephalitozoon cuniculi*, *Cyanidioschyzon merolae* and *Cryptosporidium parvum*. Therefore, decrease in intron number seems to be a common occurrence in diverse eukaryotic species — even those with high or moderate intron densities — whereas the occurrence of similarly dramatic episodes of intron gain remains a matter of debate.

Caveats to intron loss-dominated evolution? Other studies have argued for an important role for intron gain in recent evolution. *Caenorhabditis elegans* genes that might have been laterally transferred from intronless prokaryotes have intron densities that are comparable to other *C. elegans* genes[75]. This implies significant intron gains within nematodes, although the BLAST-based methods that were used might not accurately identify laterally transferred genes[34]. The urochordate *Oikopleura dioica* has also experienced both massive intron gain and loss, leading to numerous unique intron positions[76,77]. So, at least some lineages have experienced significant recent intron gain.

Two large-scale studies of intron positions in gene families (which arise by gene duplication) have also argued for a general excess of intron gain over loss[64,78]. One of these used the presence of introns in widely diverged species to infer intron presence or absence at the time of gene duplication, and concluded that there has been an excess of intron gain over loss[78]. However, the fact that some intron positions that are shared between duplicates (and so were probably present at the time of duplication) are not represented in the widely diverged species used shows that these species are not always good surrogates for intron presence at the time of duplication. This leads to systematic underestimation of ancestral introns, and therefore overestimation or underestimation of recent gains and losses, respectively. An alternative statistical reconstruction of the data from this study that incorporates intron loss indicates the opposite pattern, with more loss than gain having occurred (S.W.R., unpublished observations). The second study of gene families[64], which also argued for an important role of intron gain, unrealistically assumed constant ratios of rates of intron loss to gain across lineages, and that the only intron positions that are possible are those that are observed. Both assumptions bias the method towards parallel insertion over intron loss. More study is necessary to confirm the relative rates of intron loss and gain in these data sets.

Introns and early eukaryotes. The studies described above suggest the presence of at least moderate numbers of introns as long ago as the divergence of the major eukaryotic groups. But what of the earliest eukaryotes? This question is complicated by uncertainty about the eukaryotic phylogeny. The divergence between plants and animals might represent an extremely early branching within eukaryotes[79,80], in which case the numerous observed plant–animal intron-position correspondences indicate significant intron numbers in very early eukaryotes. Alternatively, the plant–animal divergence might be relatively recent, with other eukaryotic groups branching much earlier on[81,82], in which case intron-position correspondence between plants and animals or fungi is not informative about very early eukaryotes. What can be said about introns in early eukaryotes in the latter case?

Both introns and spliceosomal components have been found in many species that could have diverged from other eukaryotes very early in eukaryotic evolution[83–96]. There is also increasing evidence for conservation of intron positions between such 'early diverging' species and distantly related eukaryotic species[58,93,96]. This suggests the presence of at least some introns, and a nascent spliceosome, in the ancestor of all modern eukaryotes. Recently, Collins and Penny found that many known spliceosomal and spliceosome-associated proteins are not only conserved between fungi, plants and animals[97], but also in potentially early diverging eukaryotes[95]. A complex spliceosome was therefore probably present in the ancestor of all extant eukaryotes.

It is tempting to imagine that the massive complexity of the spliceosome arose in the context of reasonable numbers of introns in early eukaryotes. If so, potentially early diverging intron-poor eukaryotic species must have subsequently lost most of their ancestral introns. Moreover, the emergence of such a complex spliceosome probably required significant time, suggesting the presence of at least a basic splicing mechanism well before the eukaryotic radiation. It is particularly notable that, apart from a possible relationship between the spliceosome and the much simpler splicing mechanism of type II introns, no intermediate spliceosomal form has been found. All characterized bacteria and archaeans lack both a spliceosome and spliceosomal introns (either owing to their loss or to their never having been present in prokaryotes); all eukaryotes are probably descended from an ancestor that had a complex spliceosome and perhaps a significant number of introns.

How old are modern introns? Given the mix of stasis and change in intron evolution, what is the distribution of ages of modern introns? Assuming that intron positions that are common to multiple species represent retained ancestral introns, the minimum age of an intron is the deepest divergence between species that share that intron position. By this reasoning, at least 43–63% of introns in studied bilaterans were present in the bilateran ancestor, and at least around 19–55% of introns in studied animals and fungi were present in the animal–fungi ancestor[58] (FIG. 2b). A maximum likelihood method that incorporates intron loss increases these estimates to 52–75% and 35–60%, respectively[69] (FIG. 2b). So, the distribution of ages of modern introns is weighted towards relatively old introns, with only a minority having been inserted in the past hundreds of millions of years. Both relatively intron-rich genomes and many modern introns themselves therefore seem to date at least to major divergences within eukaryotes.

Intron early or introns late? The findings of at least moderate intron density in relatively deep eukaryotic ancestors, and of significant numbers of introns that date back at least hundreds of millions of years, are consistent with the IE model. In addition, the near-intronless states of diverse eukaryotes seem to be due to the massive loss of ancestral introns, bolstering the idea of complete intron loss in prokaryotes. Indeed, our maximum likelihood reconstructions, which suggest large numbers of introns in relatively old ancestors, imply that massive intron loss might be commonplace in eukaryotic evolution, whereas massive intron gain might be very rare. If so, prokaryotic intron loss might be more likely than the massive intron gain that is necessary to support the IL theory.

Several more direct tests have also supported the IE model. A central prediction of the theory is that if early genes were assembled from exon-encoded fragments through exon shuffling, the positions of ancient introns should tend to delineate protein structure in ancient genes — that is, they should fall between sequences that encode discrete elements of protein structure[24,26–28,30–33,35,36]. Recent results have shown that this gene–protein structure correlation is stronger in the subsets of introns that are most likely to be ancient.

These include: so-called 'phase-zero' introns that fall between codons[31–33], the abundance of which throughout eukaryotes might reflect their dominance in ancient genomes[29,31]; introns in genes that are shared between prokaryotes and eukaryotes, but not in more recently arising genes[33]; and introns that are shared between multiple eukaryotic kingdoms, which are likely to be relatively old in terms of the eukaryote phylogeny[30,32,35]. In addition, phase-zero introns, and domains that are flanked by them, are more common in putatively ancient regions of coding genes, which is as expected if ancient exon-shuffling events disproportionately involved phase-zero introns (see REFS 37,38 for a more detailed discussion).

However, other patterns are less clear. The regularity of intron phases has been used as evidence to implicate introns in gene formation, but debate surrounds this issue. Studies have reached opposite conclusions as to whether the observed intron phase pattern could be alternatively explained by insertion into protosplice sites[98–100]. Furthermore, contradictory results have been obtained about whether introns that are shared between multiple eukaryotic kingdoms are more likely to fall in phase zero[34,58]. Further comparative genomic studies will be necessary to fully resolve these issues.

Mechanisms of intron loss and gain

Mechanisms of intron loss. There are two main models for intron loss (FIG. 3a). In the classical model[100–106], the genomic copy of a gene undergoes gene conversion or double recombination with a reverse-transcribed copy of a spliced mRNA transcript (RT-mRNA), deleting one or more adjacent introns. Alternatively, introns could be lost by (near) exact genomic deletion[72,107]. The two models make several distinct predictions. First, recombination with RT-mRNAs should excise introns exactly, whereas genomic deletion should be less tidy, sometimes deleting adjacent coding sequence and/or leaving residual intron sequence (for example, FIG. 3b). Second, RT-mRNA-mediated loss requires an mRNA intermediate. As genomic changes are only heritable if they occur in germline cells, in species with a dedicated germ line intron loss should be limited to germline-transcribed genes. Third, because reverse transcriptase processes from the 3′ end to the 5′ end of RNA molecules and often produces incomplete transcripts, reverse-transcriptase products are biased towards 3′ sequences, predicting 3′ biased intron loss (although such a bias is also expected from an overrepresentation of regulatory elements in 5′ introns[108]). Finally, occasional RT-mRNA gene conversions that span multiple intron positions should lead to concerted loss of adjacent introns, whereas genomic deletion should always delete single introns.

Only the third and fourth of these predictions have been systematically tested, with mixed results. Neither 3′-loss bias nor adjacent intron loss were detected in several nematode genes[72], 2,073 sets of orthologues from 4 species of Euascomycetes fungi[109], or a large set of multidomain metazoan genes[110], which suggests genomic deletion. However, other studies support the RT-mRNA model. Introns in intron-sparse genes[111] and genomes[112] are concentrated towards the 5′ ends of genes, consistent with 3′ biased loss from

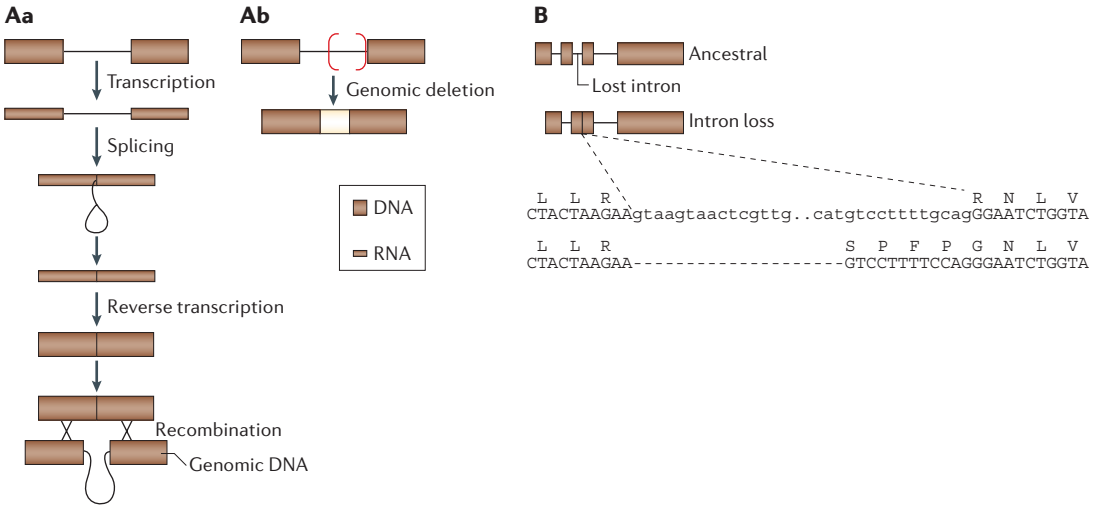


Figure 3 | **Models and examples of intron loss. A** | The two main models of intron loss. The classical model of intron loss is shown in panel **Aa**[101–106]. A gene is transcribed and the intron spliced out. This spliced transcript is then reverse transcribed and the resultant cDNA undergoes recombination with the genomic copy, leading to intron loss. Panel **Ab** shows the genomic deletion model of intron loss[72,107,108], which leads to exact or inexact deletion of the DNA sequence that encodes the intron sequence. **B** | An apparent example of intron loss by genomic deletion. A genomic deletion of most of an intron sequence from the *jingwei* gene in *Drosophila teissieri* left a 12-bp residue[10]. This residual 12 bp is an effective insertion of 4 codons, encoding the amino-acids serine, proline, phenylalanine and proline in the new allele. This case represents the only known case of an intron presence/absence polymorphism within a species. The new intron-loss allele segregates at 77% in the species and is associated with a decrease in expression levels. Population studies show that the new intron-loss allele is evolving under positive selection, although it is not known whether this is due to the coding sequence insertion, the change in mRNA level, or the absence of the intron itself. Data are from REF. 10.

Protosplice sites
A consensus motif into which newly inserted introns seem to insert (or at least in which they are found), which is generally thought to be a variant of MAG|GT, where M denotes an A or C, and the line indicates the point of insertion.

Gene conversion
Any process by which a genomic element changes to the sequence of a paralogous element; this probably takes place mainly by double recombination.

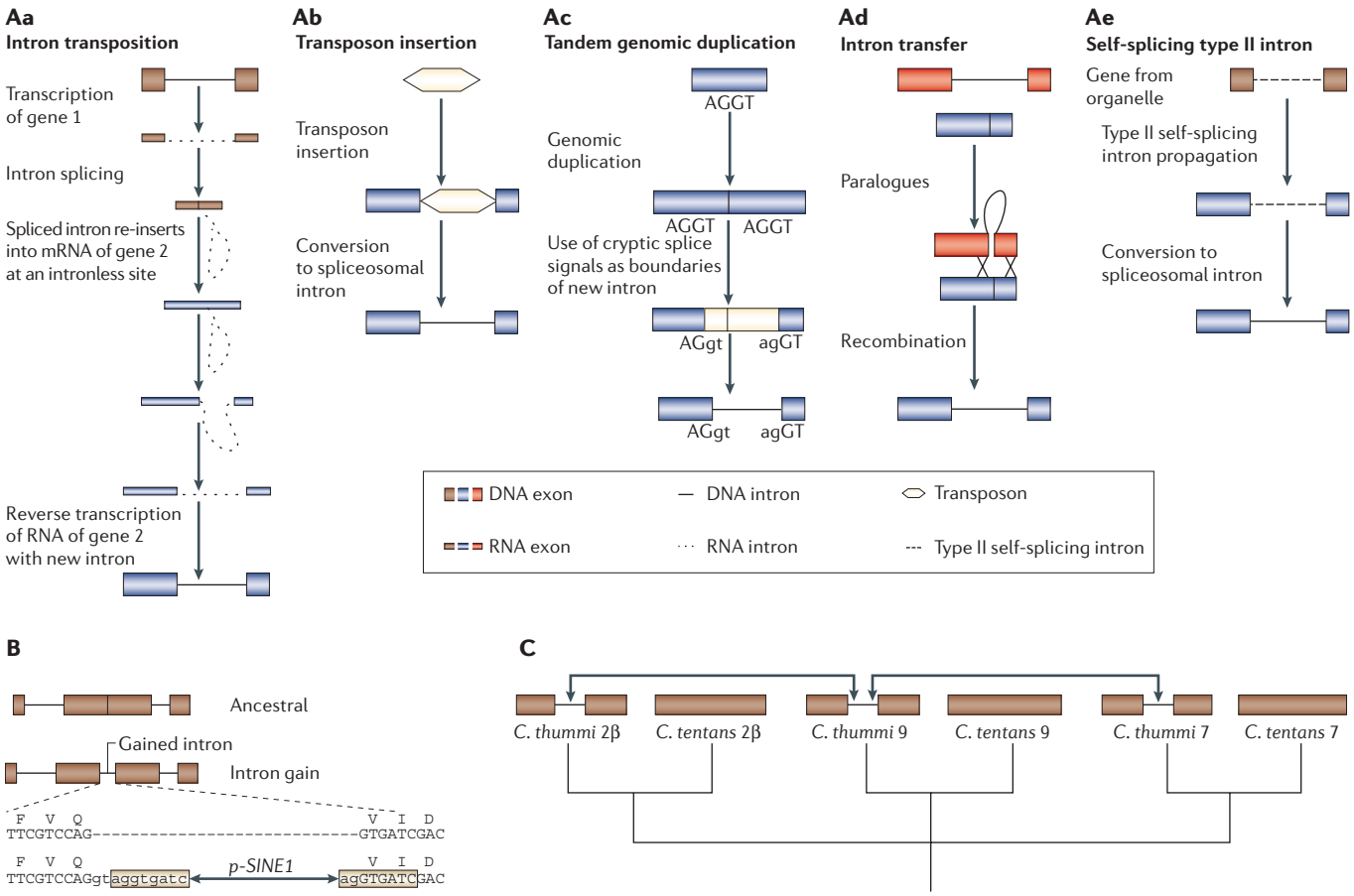


Figure 4 | **Models and examples of intron gain. A** | Five models of intron gain. Panel **Aa** shows intron transposition[46]. An intron from the mRNA transcript of one gene is spliced out, and then reverse splices into a previously intronless site of a transcript of another (or the same) gene. Panel **Ab** shows transposon insertion[6,11,12,117–120,122]. A transposable element inserts at the DNA level into a previously intronless site of the genomic copy of a gene. RNA copies of this intervening transposon sequence are removed from mRNA transcripts of the gene by the spliceosome, leading to the creation of a new intron. Panel **Ac** shows tandem genomic duplication[6]. A stretch of coding sequence with an internal AGGT sequence is duplicated. The two resultant AGGTs are interpreted by the spliceosome as the 5′ and 3′ boundaries of a new intron, effectively leading to the creation of a new intron, with conservation of the original coding sequence. Panel **Ad** shows intron transfer from paralogues[13]. Homologous recombination between paralogues leads to transfer of an intron from the intron-containing paralogue to the other paralogue, which previously lacked an intron at this site. Panel **Ae** shows conversion of a type II intron to a spliceosomal intron[7–9]. A type II intron from an organellar gene is transferred to the eukaryotic nucleus and inserts into a previously intronless site. It is subsequently removed from the transcripts by the spliceosome, leading to the creation of a new spliceosomal intron. Modified with permission from REF. 122 © (2004) BioMed Central Ltd. **b** | An example of intron insertion by transposition[11,12] that shows intron gain through the transposon *p-SINE1* insertion in the rice catalase A gene. Apart from a duplicated sequence of 8 nucleotides, the intron, which is absent in maize and other plants, is completely identical to a *p-SINE1* element. Modified with permission from REF. 11 © (1998) Springer-Verlag. **c** | Example of intron insertion by intron transfer from a paralogue[13]. All three globin paralogues in the midge *Chironomus thummi* contain an intron at the homologous position, whereas all three copies are intronless in several related species (for example, *Chironomus tentans*), indicating insertion into one copy and subsequent transfer to the others. Modified with permission from REF. 13 © (1997) Elsevier Science.

more intron-rich ancestral structures. In addition, individual convincing cases of concordant loss of adjacent introns have been found[113,114], introns shared between widely diverged species are also 5′ biased[115], and intron loss in 684 groups of orthologous genes showed 3′ intron-loss bias and concerted loss of adjacent introns[116]. The increasing focus of genome sequencing projects on closely related species should clarify the relative importance of these two mechanisms.

Mechanisms of intron gain. There are five main models for the origin of new introns[6,8,11–13,45,46,117–121] (FIG. 4). Intron transposition — the perennial favourite, despite lack of direct evidence — postulates that RNA intron sequences that have been recently spliced out of transcripts reinsert into new positions of another (or the same) mRNA. The intron-acquiring transcript is then reverse transcribed and the reverse-transcribed copy transfers the new intron to the genomic copy by gene conversion. Indirect

evidence for this model includes analogy to type II introns, which self-propagate by similar mechanisms, as well as the resemblance of insertion sites of new introns to sites that flank older introns, which might indicate the involvement of the spliceosome in intron insertion[46,47,66]. However, such correspondences could also reflect post-insertion selection on the splicing efficiency of new introns, based on spliceosome requirements for flanking exonic sequences.

Recently, Coghlan and Wolfe studied 122 introns in *C. elegans* and *Caenorhabditis briggsae* that were apparently inserted since their divergence around 100 million years ago (REF. 63). The sequences of some of these introns resembled other *Caenorhabditis* introns (both ancestral and recently gained), possibly suggesting intron transposition. However, the regions of inter-intron homology largely comprise multicopy palindromic repetitive elements that are also found in intergenic regions (REF. 122; and S.W.R., unpublished observations), which indicates that these similarities might instead derive from the genomic insertion of transposons with palindromic sequences. Indeed, the preponderance of palindromic repetitive elements in the new introns indicates that the insertions themselves might have created them[6,11,12,117,118,120]. The fact that the transposons involved have palindromic sequences is particularly intriguing, as the tendency of such elements to form hairpin structures would juxtapose the ends of new introns, possibly facilitating their splicing[122].

Another study examined introns from the Rogozin *et al.* data set that are specific to a single species, which the authors interpreted as recently gained introns[115]. These introns are 3′ biased, which was taken as evidence for reverse-transcriptase-mediated gain. However, the species involved are only distantly related, so some species-specific introns are probably ancestral introns that have been lost from other lineages. Therefore, 3′-biased intron loss could also explain this pattern. The increasing availability of genomic sequence from various *Caenorhabditis* species should allow for even clearer identification of recent intron gains to resolve these issues.

A two-tiered system of intron origin. Importantly, no single model offers clear explanations for both the initial origins of the spliceosomal system and more recent intron gains. Each plausible model for the recently inserted nematode introns — intron transposition, transposon insertion and genomic duplication — requires a pre-existing, efficient spliceosome. The only proposal that includes a mechanism for spliceosomal origin is the conversion of type II endosymbiont introns to spliceosomal introns[6,8,9], which cannot explain the recently inserted *Caenorhabditis* introns as animal mitochondria lack type II introns. According to this model, the spliceosome and the first spliceosomal introns were descended from type II introns. It is possible that the spliceosome and the first spliceosomal introns would be descended from type II introns, whereas some recently gained introns (for example, the recent *Caenorhabditis* gains) would be gained by another mechanism. Modern introns would not then be truly homologous, but would be unified only by serving as substrates for the spliceosome. Alternatively, some future model of intron origin might offer a more complete explanation.

The causes of intron evolution

Debate also continues about the evolutionary forces that are responsible for modern intron distributions. Several proposed advantages of introns do not seem to have contributed greatly to their initial spread (BOX 1). In addition, previously proposed explanations for intron-number differences between species do not predict recent findings that show relatively intron-rich structures in relatively early eukaryotes.

Cell number, generation time and population size. Some evidence indicates that multicellular species with long generation times and small population sizes tend to be intron-dense, whereas unicellular species with short generation times and large population sizes tend to be intron-sparse. The classical interpretation[15–17,23] is that introns and other non-essential DNA are disfavoured in unicellular species that are under strong selective pressure for short genome replication time, but are nearly neutral in multicellular species.

A more recent proposal from Lynch sees population size as the driving variable[21]. Sequence requirements for intron splicing presumably impose constraints at sites that are otherwise free to vary, and this extra constraint might impose a weak selective disadvantage on intron-containing alleles. Slightly deleterious alleles are more likely to become fixed in small populations, so that species with small populations should be more intron-rich.

Neither proposal predicts the recent findings of relatively intron-rich early eukaryotes and subsequent recurrent intron loss that we discussed above. The classical model directly predicts low intron number in unicellular early eukaryotes. To reconcile relatively intron-rich early eukaryotes with the population-size model would require low population sizes among early eukaryotes and continuing subsequent global population expansions, which is an improbable scenario. Moreover, other preliminary data are not consistent with the population-size proposal. The proposed constraint cannot explain the magnitude of the difference in intron number between multicellular and unicellular species, and there is no apparent relationship between population size and intron number among either multicellular species or unicellular species (see the supplemental material from REF. 20; and S.W.R., unpublished observations).

A new proposal: differences in recombination rate. The proposals above implicitly seek to explain differences in intron number as differences in intron loss–gain equilibria that are due to ecology. In species with long generation times or small population sizes, intron number will mainly reflect a balance between the rates of spontaneous intron gain and loss mutations. Equilibrium intron numbers in fast-replicating species or species with large population sizes will be lower, owing to a large number of successful intron-loss mutations and fewer successful

Fixation
With respect to a given mutant, the condition in which all alleles in the population are descendents of that mutant.

Box 1 | **Selective forces that might favour introns**

Introns have been shown or proposed to carry out many functions. Selection for some of these might increase the evolutionary success of introns, although there is no convincing evidence that these possible advantages are important in driving intron gain or loss.

Nonsense-mediated decay. In nonsense-mediated decay (NMD), if a transcriptional error leads to a stop codon upstream of a nearby intron–exon boundary, that transcript is targeted for degradation[127]. The presence of introns could therefore be important for transcript fidelity. However, selection for intron gain on the basis of NMD predicts moderately numerous introns at some optimum number (reflecting a balance between positive NMD-based selection and other proposed negative selection) in species with large population sizes (and therefore more effective selection). Smaller populations should be less optimized, with either fewer or more introns, owing to differences in loss and gain rates[128]. Instead, the opposite is true — species with small populations show consistently large numbers of introns, whereas numbers in species with large populations vary radically. In addition, if NMD is an important factor, 3′ introns should be favoured relative to 5′ introns, as the former will tend to recognize more upstream transcription errors. Therefore, 3′ bias should increase with population size — instead, introns are 5′ biased, particularly in species with large population sizes[112].

Alternative splicing. Introns provide the possibility of generating new gene products by alternative splicing, which has doubtlessly had an important role in the evolution of at least plant, animal, and a minority of characterized fungal genomes. However, given that alternative splicing seems to have a less important role in many other groups, the results discussed here, which show that large fractions of modern introns predate the main eukaryotic divergences, suggest that alternative splicing has not been a major force in the colonization of eukaryotic genomes by introns.

Exon shuffling. Patthy and others have shown that exon shuffling was vital in the creation of large numbers of the multidomain genes of metazoans, and the introns-early theory proposes that exon shuffling dates back much further to the ancestors of eukaryotes and prokaryotes[51,67,68]. The results discussed here indicate that many of the introns involved in exon shuffling in early metazoans dated back much further into eukaryotic history, making it unlikely that a marked rise in intron number was associated with metazoan exon shuffling. By contrast, according to the introns-early model, introns were themselves vital for the formation of the earliest genes, in which case the presence of introns would have been of extreme importance, leading not only to the success of intron-containing genes, but to life as we know it.

Increased recombination. The presence of introns in genes increases the recombination rate between parts of the coding region, allowing the creation of new products and increasing the efficiency of selection[22,41–43]. That this ability might be important was originally borne out by the finding of an inverse correlation between recombination rate and intron length in *Drosophila*[42], which suggests that introns in low-recombination areas might be longer owing to selection for increased recombination. However, further analysis has shown that the pattern is not so simple, suggesting that increased recombination is not a major factor in intron evolution[129–131].

intron gains, owing to a greater intensity or efficiency of selection against introns.

However, the studies of ancestral intron number discussed above suggest widespread non-equilibrium, with diverse groups experiencing ongoing changes in intron number. If, as our maximum likelihood estimates suggest, eukaryotic evolution has been characterized by more intron loss than gain, differences in intron number might largely reflect differential levels of retention of ancestral introns. If so, the most important determinant of intron number might simply be rate of evolution: slow-evolving species will retain more ancestral introns than fast-evolving species. If intron loss proceeds through gene conversion by RT-mRNAs, rates of intron loss might depend on overall rates of paralogous recombination. In species where most recombination occurs during meiosis, intron-loss rate should correspond to the number of meioses per unit time: species that have undergone more sexual generations will have lost more introns than longer-generation species.

At first glance, the data seem supportive. Since the mouse–human divergence, mice have experienced both more generations and more intron loss[71], and *O. dioica*, which has an extremely fast life cycle, has lost far more ancestral introns than other chordates[76,77]. Furthermore, vertebrates retain more ancestral introns than dipterans or *C. elegans*[110,123], and the filarial worm *Brugia malayi* has a much longer life cycle and roughly twice as many introns[124] as *C. elegans*. Humans might be more intron-rich than other well-studied species as the latter are mostly model organisms, chosen in part for their short generation times, and/or parasites, for which short replication times are presumably an advantage. The high paralogous gene-conversion rate in *S. cerevisiae*[125] might explain the loss of most of its ancestral introns[58].

However, this proposal cannot explain the observed tenfold difference in intron-gain rates between lineages[23]. We found an inverse correspondence between intron loss and gain rates for six widely diverged lineages[123], although this correspondence might be exaggerated. Over long timescales, many introns that are gained along a lineage will subsequently be lost, leading to an underestimation of gains. In correcting for such multiple events, we assumed a constant loss rate for all introns. However, rates of loss for new introns are probably faster than for introns that have had hundred of millions of years to co-evolve with the surrounding genes, perhaps leading us to underestimate recent gains in rapid-loss lineages (leading to an artefactual inverse gain–loss rate correspondence). Species of *Caenorhabditis*[123] and *Oikopleura*[76,77] show high rates of both gain and loss, whereas vertebrates and species of *Cryptococcus* show very low rates of both processes (REF. 109; J.E. Stajich, S.W.R. and F.S. Dietrich, unpublished observations). So,

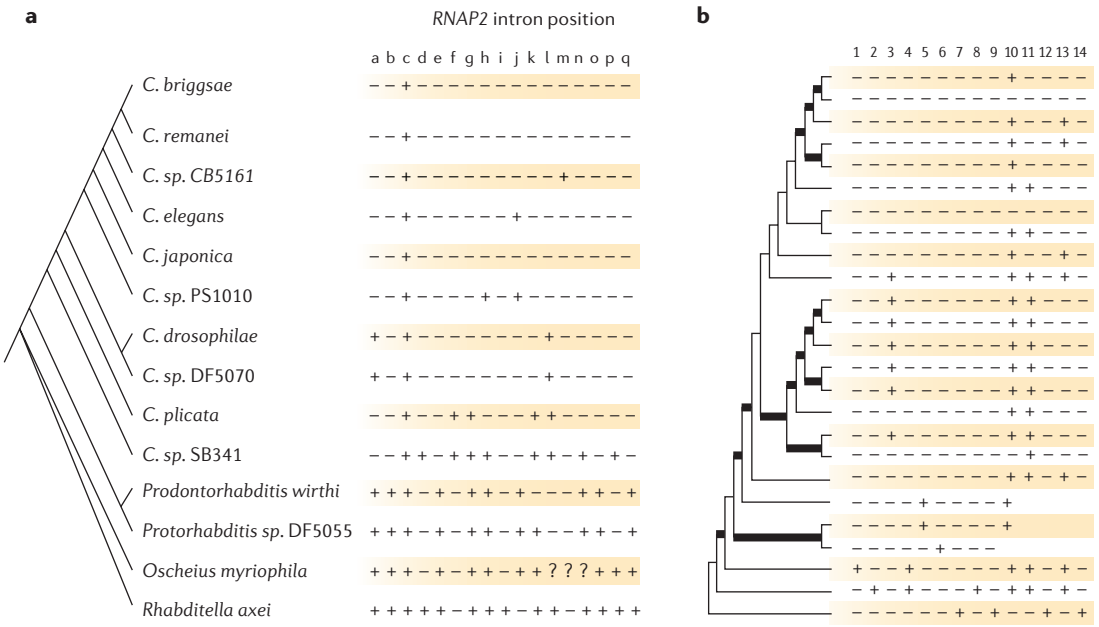


Figure 5 | **Strange patterns of intron loss and gain. a** | Strange patterns of intron conservation in the RNA binding protein 2 gene (*RNAP2*; also known as *CUGBP2*) of nematodes[73]. The pattern of intron conservation at position j requires that either the intron was present in the ancestor and was subsequently lost five times along different branches of the tree, or that the ancestor had no intron at this position but that introns have subsequently been gained at the site independently at least three times along different branches. Modified with permission from REF. 73 © (2003) National Academy of Sciences. **b** | Recurrent loss of intron 13 in the *white* gene of arthropods[126]. The observed pattern of intron conservation of intron 13 requires at least five independent intron losses, although few losses of the other introns are observed. Modified with permission from REF. 126 © (2002) Oxford University Press.

it is possible that differences in intron-gain rates could be largely due to coincidental differences in molecular biology between species. This could give rise to variations in the rates of various types of moderately long spontaneous insertions, and in the chances of their serendipitous recognition by the spliceosome.

Of course, even if a key determinant of intron number is the rate of intron-loss mutations, selection against introns might still be important, particularly in shaping the success of new insertions. Whereas the main determinant of intron number might now be the different rates of approaching equilibrium through intron loss, in the distant future differences in equilibrium frequencies — which are set by a combination of factors, including selection — could dominate. Even now, these factors are likely to contribute. Indeed, it would be surprising if introns were not disfavoured in species under strong genomic streamlining pressure, and higher population-wide rates of mutation to intron-lacking alleles in large populations will lead to higher rates of loss if introns are disadvantageous.

Strange patterns of intron conservation. A few examples illustrate how much we still have to learn about the forces that drive intron loss and gain. The phylogenetic pattern at an intron position of the gene that encodes RNA binding protein 2 (*RNAP2*; also known as *CUGBP2*) in nematodes requires either five independent losses of the same intron (often while leaving adjacent introns intact), or three independent intron gains in the same position[73] (FIG. 5a). Similarly, the pattern at an intron position in arthropod *white* genes (FIG. 5b) requires at least five independent gains or losses[126]. These examples are not outliers that have been culled from large-scale studies, but are results from small-scale studies. Inferring loss in these cases implies massive differences in loss rates between introns — in FIG. 5a one intron has been retained in all studied species. However, inferring more than two parallel insertions in such closely related groups seems incredible. Another study of over 200 sets of orthologues from 4 species of Euascomycete fungi found an average of around one-third of an intron gain per gene; however 1 gene showed more than 20 (REF. 116). Clearly, powerful forces at least occasionally drive introns in and out of eukaryotic genes, although what these forces might be remains mysterious.

Conclusions

Despite exciting recent advances, several important questions remain to be answered in the field of intron evolution. First, how are new introns created, and how homogeneous is the insertion mechanism (or mechanisms) across species? Coghlan and Wolfe[63] have taken an important step by identifying recent gains in nematodes, but their results are not straightforward to interpret and are of unknown generality across taxa. Although many current genome projects that focus on clusters of closely related species unfortunately address species with low rates of intron gain (vertebrates, fungi and apicomplexans), the gradual accretion of genome sequences will hopefully allow us to answer these questions in the future. The sequencing

of *O. dioica*, with its wildly divergent intron–exon structures, is particularly promising.

Another important question is why introns first arose. Despite 30 years of intense study, we are barely closer to answering this riddle. Perhaps introns were present before the prokaryote–eukaryote ancestor and were vital to the creation of early genes. Perhaps introns first arose in massive numbers through the insertion of transposable elements, an insult that provoked the evolution of a host defence mechanism against further massive insertion. Perhaps the spliceosomal organization of genes facilitated other important transformations of early eukaryotic evolution, from the coordination of transcription and translation to processing and nuclear export of transcripts.

Finally, why do different species show such radically different intron numbers? To the previous explanations of organismal complexity and population size we add a third: that of differences in recombination rates. Comparisons of closely related species should help to finally lift the shroud on this very old question.

1. Cannone, J. J. *et al.* The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *MBC Bioinformatics* **3**, 2 (2002).
2. Bonen, L. & Vogel, J. The ins and outs of group II introns. *Trends Genet.* **17**, 322–331 (2001).
3. Lambowitz, A. M. & Zimmerly, S. Mobile group II introns. *Annu. Rev. Genet.* **38**, 1–35 (2004).
4. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell* **12**, 5–14 (2003).
5. Cech, T. R. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* **44**, 207–210. (1986).
6. Rogers, J. H. How were introns inserted into nuclear genes? *Trends Genet.* **5**, 213–216 (1989).
Two of the five current models of intron creation are proposed in this speculative piece.
7. Sharp, P. A. Five easy pieces. *Science* **254**, 663 (1991).
8. Cavalier-Smith, T. Intron phylogeny: a new hypothesis. *Trends Genet.* **7**, 145–148 (1991).
An important statement of the idea that introns might be descended from type II introns that are transferred from early eukaryotic organelles.
9. Stoltzfus, A. On the possibility of constructive neutral evolution. *J. Mol. Evol.* **49**, 169–181 (1999).
10. Llopart, A., Comeron, J. M., Brunet, F. G., Lachaise, D. & Long, M. Intron presence–absence polymorphism in *Drosophila* driven by positive Darwinian selection. *Proc. Natl Acad. Sci. USA* **99**, 8121–8126 (2002).
The sole known cases of polymorphic intron absence–presence within a species, notably in a gene with a fascinating evolutionary history.
11. Iwamoto, M., Maekawa, M., Saito, A., Higo, H. & Higo, K. Evolutionary relationship of plant catalase genes inferred from intron–exon structures: isozyme divergence after the separation of monocots and dicots. *Theor. Appl. Genet.* **97**, 9–19 (1998).
The first convincing case of intron gain in which the source of the intron, an inserted SINE element, is clear.
12. Iwamoto, M., Nagashima, H., Nagamine, T., Higo, H. & Higo, K. p-SINE1-like intron of the CatA catalase homologs and phylogenetic relationships among AA-genome *Oryza* and related species. *Theor. Appl. Genet.* **98**, 853–861 (1999).
13. Hankeln, T., Friedl, H., Ebersberger, I., Martin, J. & Schmidt, E. R. A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. *Gene* **205**, 151–160 (1997).
14. Dawkins, R. *The Selfish Gene* (Oxford Univ. Press, 1976).
15. Orgel, L. E. & Crick, F. H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
16. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1981).
Along with reference 15, this article contains early statements of the idea of genome evolution by insertion of selfish elements and differential selection on such elements between species of different complexity.
17. Gilbert, W. The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901–905 (1987).
18. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
19. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46**, 111–138 (1971).
20. Lynch, M. Intron evolution as a population-genetic process. *Proc. Natl Acad. Sci. USA* **99**, 6118–6123 (2002).
21. Lynch, M. & Conery, J. The origins of genome complexity. *Science* **302**, 1401–1404 (2002).
22. Gilbert, W. Why genes in pieces? *Nature* **271**, 501 (1978).
23. Doolittle, W. F. Genes in pieces – were they ever together? *Nature* **272**, 581–582 (1978).
24. Blake, C. C. F. Do genes in pieces imply proteins in pieces? *Nature* **273**, 267 (1978).
References 22–24 provide the backbone of the IE theory.
25. Perler, F. *et al.* The evolution of genes — the chicken preproinsulin gene. *Cell* **20**, 555–566 (1980).
26. Go, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **291**, 90–92 (1981).
27. Stone, E. M., Rothblum, K. N. & Schwartz, R. J. Intron-dependent evolution of chicken glyceraldehyde phosphate dehydrogenase gene. *Nature* **313**, 498–500 (1985).
28. Straus, D. & Gilbert, W. Genetic engineering in the Precambrian: structure of the chicken triosephosphate isomerase gene. *Mol. Cell Biol.* **5**, 3497–3506 (1985).
29. Long, M., Rosenberg, C. & Gilbert, W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl Acad. Sci. USA* **92**, 12495–12499 (1995).
30. De Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert., W. Introns correlate with module boundaries in ancient proteins. *Proc. Natl Acad. Sci. USA* **93**, 14632–14636 (1996).
31. De Souza, S. J. *et al.* Towards 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 (1998).
An important early statement of the 'synthetic' or 'mixed' variant of the IE theory.
32. Roy, S. W., Nosaka, M., de Souza, S. J. & Gilbert, W. Centripetal modules and ancient introns. *Gene* **238**, 85–91 (1999).
33. Fedorov, A. *et al.* Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proc. Natl Acad. Sci. USA* **98**, 13177–13182 (2001).
34. Roy, S. W., Lewis, B. P., Fedorov, A. & Gilbert, W. Footprints of primordial introns on the eukaryotic genome. *Trends Genet.* **17**, 496–498 (2001).
35. Fedorov, A., Roy, S., Cao, X. & Gilbert, W. Phylogenetically older introns strongly correlate with module boundaries in ancient proteins. *Genome Res.* **13**, 1155–1157 (2003).
36. Roy, S. W., Fedorov, A. & Gilbert, W. The signal of ancient introns is obscured by intron density and homolog number. *Proc. Natl Acad. Sci. USA* **99**, 15513–15517 (2002).
37. De Souza, S. J. The emergence of a synthetic theory of intron evolution. *Genetica* **118**, 117–121 (2003).
38. Roy, S. W. Recent evidence for the exon theory of genes. *Genetica* **118**, 251–266 (2003).
39. Patthy, L. Genome evolution and the evolution of exon-shuffling — a review. *Gene* **238**, 103–114 (1999).
A comprehensive review of the known cases of exon shuffling.
40. Patthy, L. Modular assembly of genes and the evolution of new functions. *Genetica* **118**, 217–231 (2003).
41. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. *Proc. Natl Acad. Sci. USA* **75**, 1485–1489 (1978).
42. Comeron, J. M. & Kreitman, M. The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* **156**, 1175–1190 (2000).
43. Duret, L. Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends Genet.* **17**, 172–175 (2001).
44. Lynch, M. & Kewalramani, A. Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol. Biol. Evol.* **20**, 563–571 (2003).
45. Cavalier-Smith, T. Selfish DNA and the origin of introns. *Nature* **315**, 283–284 (1985).
An important early statement of the IL hypothesis.
46. Sharp, P. A. On the origin of RNA splicing and introns. *Cell* **42**, 397–400 (1985).
47. Dibb, N. J. & Newman, A. J. Evidence that introns arose at proto-splice sites. *EMBO J.* **8**, 2015–2021 (1989).
48. Palmer, J. D. & Logsdon, J. M. Jr. The recent origin of introns. *Curr. Opin. Genet. Dev.* **1**, 470–477 (1991).
49. Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. Jr & Doolittle, W. F. Testing the exon theory of genes: the evidence from protein structure. *Science* **265**, 202–207 (1994).
50. Stoltzfus, A. Origin of introns — early or late? *Nature* **369**, 526–527 (1994).
51. Logsdon, J. M. Jr *et al.* Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl Acad. Sci. USA* **92**, 8507–8511 (1995).
52. Kwaitowski, J., Krawczyk, M., Kornacki, M., Bailey, K. & Ayala, F. J. Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene. *Proc. Natl Acad. Sci. USA* **92**, 8503–8506 (1995).
53. Cho, G. & Doolittle, R. F. Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.* **44**, 573–584 (1997).
54. Rzhetsky, A., Ayala, F. J., Hsu, L. C., Chang, C. & Yoshida, A. Exon/intron structure of aldehyde dehydrogenase genes supports the 'introns-late' theory. *Proc. Natl Acad. Sci. USA* **94**, 6820–6825 (1997).
55. Logsdon, J. M. Jr. The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **8**, 637–648 (1998).
56. Logsdon, J. M. Jr, Stoltzfus, A. & Doolittle, W. F. Molecular evolution: recent cases of spliceosomal intron gain? *Curr. Biol.* **8**, R560–R563 (1998).
57. Fedoro v, A. F., Merican, A. F. & Gilbert, W. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl Acad. Sci. USA* **99**, 16128–16133 (2002).
58. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13**, 1512–1517 (2003).
59. Tarrio, R., Rodriguez-Trelles, F. & Ayala, F. J. A new *Drosophila* spliceosomal intron position is common in plants. *Proc. Natl Acad. Sci. USA* **100**, 6580–6583 (2003).
60. Dibb, N. J. Proto-splice site model of intron origin. *J. Theor. Biol.* **151**, 405–416 (1991).
61. Paquette, S. M., Bak, S & Feyereisen, R. Intron–exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* **19**, 307–317 (2000).
62. Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. & Koonin, E. V. Reconstruction of ancestral protosplice sites. *Curr. Biol.* **14**, 1505–1508 (2004).
63. Coghlan, A. & Wolfe, K. H. Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl Acad. Sci. USA* **101**, 11362–11367 (2004).
The first sequence analysis of a large number of putative recently gained introns. The results are interpreted by the authors as evidence for intron transposition, although the answer might not be so straightforward.

64. Qiu, W. G., Schisler, N. & Stoltzfus, A. The evolutionary gain of spliceosomal introns: sequence and phase preferences. *Mol. Biol. Evol.* **21**, 1252–1263 (2004).
65. Tordai, H. & Patthy, L. Insertion of spliceosomal introns in proto-splice sites: the case of secretory signal peptides. *FEBS Lett.* **575**, 109–111 (2004).
66. Sadusky, T., Newman, A. J. & Dibb, N. J. Exon junction sequences as cryptic splice sites: implications for intron origin. *Curr. Biol.* **14**, 505–509 (2004).
67. Stoltzfus, A. Molecular evolution: introns fall into place. *Curr. Biol.* **14**, R351–352 (2004).
68. Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. & Koonin, E. V. Conservation versus parallel gains in intron evolution. *Nucleic Acids Res.* **33**, 1741–1748 (2005).
69. Roy, S. W. & Gilbert, W. Complex early genes. *Proc. Natl Acad. Sci. USA* **102**, 1986–1991 (2005).
The reanalysis of data from reference 58, which indicates that intron loss, not gain, has dominated intron evolution.
70. Rogozin, I. B., Sverdlov, A. V., Babenko, V. N. & Koovin, E. V. Analysis of evolution of exon–intron structure in eukaryotic genes. *Brief Bioinform.* **6**, 118–134.
71. Roy, S. W., Fedorov, A. & Gilbert, W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl Acad. Sci. USA* **100**, 7158–7162 (2003).
72. Cho, S., Jin, S. W., Cohen, A. & Ellis, R. E. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14**, 1207–1220 (2004).
73. Kiontke, K. *et al. Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl Acad. Sci. USA* **101**, 9003–9008 (2004).
74. Robertson, H. M. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* **8**, 449–463 (1998).
75. Wolf, Y. I., Kondrashov, F. A. & Koonin, E. V. Footprints of primordial introns on the eukaryotic genome: still no clear traces. *Trends Genet.* **17**, 499–501 (2001).
76. Seo, H. C. *et al.* Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506 (2001).
77. Edvardsen, R. B. *et al.* Hypervariable and highly divergent intron–exon organizations in the chordate *Oikopleura dioica*. *J. Mol. Evol.* **59**, 448–457 (2004).
78. Babenko, V. N., Rogozin, I. B., Mekhedov, S. L. & Koonin, E. V. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* **32**, 3724–3733 (2004).
79. Embley, T. M. & Hirt, R. P. Early branching eukaryotes? *Curr. Opin. Genet. Dev.* **8**, 624–629 (1998).
80. Simpson, A. G. & Roger, A. J. Eukaryotic evolution: getting to the root of the problem. *Curr. Biol.* **12**, R691–R693 (2002).
81. Sogin, M. L. Early evolution and the origin of eukaryotes. *Curr. Opin. Genet. Dev.* **1**, 457–463 (1991).
82. Hashimoto, T. & Hasegawa, M. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. *Adv. Biophys.* **32**, 73–120 (1996).
83. Stiller, J. W., Duffield, E. C. & Hall, B. D. Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. *Proc. Natl Acad. Sci. USA* **95**, 11769–11774 (1998).
84. Biderre, C., Metenier, G. & Vivares, C. P. A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi*. *Mol. Biochem. Parasitol.* **94**, 283–286 (1998).
85. Fast, N. M., Roger, A. J., Richardson, C. A. & Doolittle, W. F. U2 and U6 snRNA genes in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. *Nucleic Acids Res.* **26**, 3202–3207 (1998).
86. Fast, N. M. & Doolittle, W. F. *Trichomonas vaginalis* possesses a gene encoding the essential spliceosomal component, PRP8. *Mol. Biochem. Parasitol.* **99**, 275–278 (1999).
87. Breckenridge, D. G, Watanabe, Y., Greenwood, S. J., Gray, M. W. & Schnare, M. N. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc. Natl Acad. Sci. USA* **96**, 852–856 (1999).
88. Ismaili, N. *et al.* Characterization of a SR protein from *Trypanosoma brucei* with homology to RNA-binding *cis*-splicing proteins. *Mol. Biochem. Parasitol.* **102**, 103–105 (1999).
89. Schnare, M. N. & Gray, M. W. Structural conservation and variation among U5 small nuclear RNAs from trypanosomatid protozoa. *Biochim. Biophys. Acta.* **1490**, 362–366 (2000).
90. Dacks, J. B. & Doolittle, W. F. Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* **107**, 419–425 (2001).
91. Edgcomb, V. P., Roger, A. J., Simpson, A. G., Kysela, D. T. & Sogin, M. L. Evolutionary relationships among 'jakobid' flagellates as indicated by α- and β-tubulin phylogenies. *Mol. Biol. Evol.* **18**, 514–522 (2001).
92. Archibald, J. M., O'Kelly, C. J. & Doolittle, W. F. The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. *Mol. Biol. Evol.* **19**, 422–431 (2002).
93. Nixon, J. E. *et al.* A spliceosomal intron in *Giardia lamblia*. *Proc. Natl Acad. Sci. USA* **99**, 3701–3705 (2002).
94. Simpson, A. G., MacQuarrie, E. K & Roger, A. J. Eukaryotic evolution: early origin of canonical introns. *Nature* **419**, 270 (2002).
95. Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**, 1053–1066 (2005).
A demonstration of the presence of a sophisticated spliceosome in the common ancestor of all extant eukaryotes.
96. Vanacova, S., Yan, W., Carlton, J. M. & Johnson, P. J. Spliceosomal introns in the deep-branching eukaryote *Trichomonas vaginalis*. *Proc. Natl Acad. Sci. USA* **102**, 4430–4435 (2005).
97. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids. Res.* **30**, 1427–1464 (2002).
98. Ruvinsky, A., Eskesen, S. T., Eskesen, F. N. & Hurst, L. D. Can codon usage bias explain intron phase distributions and exon symmetry? *J. Mol. Evol.* **60**, 99–104 (2005).
99. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. Relationship between 'proto-splice sites' and intron phases: evidence from dicodon analysis. *Proc. Natl Acad. Sci. USA* **95**, 219–223 (1998).
100. Long, M. & Rosenberg, C. Testing the 'proto-splice sites' model of intron origin: evidence from analysis of intron phase correlations. *Mol. Biol. Evol.* **17**, 1789–1796 (2000).
101. Bernstein, L. B., Mount, S. M. & Weiner, A. M. Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. *Cell* **32**, 461–472 (1983).
102. Lewin, R. How mammalian RNA returns to its genome. *Science* **219**, 1052–1054 (1983).
103. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55**, 631–661 (1986).
104. Fink, G. R. Pseudogenes in yeast? *Cell* **49**, 5–6 (1987).
105. Long, M. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
106. Derr, L. K. The involvement of cellular recombination and repair genes in RNA-mediated recombination in *Saccharomyces cerevisiae*. *Genetics* **148**, 937–945 (1998).
107. Kent, W. J. & Zahler, A. M. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Res.* **10**, 1115–1125 (2000).
108. Fedorova, L. & Fedorov, A. Introns in gene evolution. *Genetica* **118**, 123–131 (2003).
109. Nielsen, C. B., Friedman, B., Birren, B., Burge, C. B. & Galagan, J. E. Patterns of intron gain and loss in fungi. *PLoS Biol.* **2**, e422 (2004).
110. Banyai, L. & Patthy, L. Evidence that human genes of modular proteins have retained significantly more ancestral introns than their fly or worm orthologues. *FEBS Lett.* **565**, 127–132 (2004).
111. Sakurai, A. *et al.* On biased distribution of introns in various eukaryotes. *Gene* **300**, 89–95 (2002).
112. Mourier, T. & Jeffares, D. C. Eukaryotic intron loss. *Science* **300**, 1393 (2003).
113. Frugoli, J. A., McPeek, M. A., Thomas, T. L. & McClung, C. R. Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* **149**, 355–365 (1998).
114. Wada, H. *et al.* Dynamic insertion-deletion of introns in deuterostome EF-1α genes. *J. Mol. Evol.* **54**, 118–128 (2002).
115. Sverdlov, A. V., Babenko, V. N., Rogozin, I. B. & Koonin, E. V. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron loss. *Gene* **338**, 85–91 (2004).
116. Roy, S. W. & Gilbert, W. The pattern of intron loss. *Proc. Natl Acad. Sci. USA* **102**, 713–718 (2005).
117. Crick, F. H. Chromosome structure and function — future prospects. *Eur. J. Biochem.* **83**, 1–3 (1978).
118. Crick, F. Split genes and RNA splicing. *Science* **204**, 264–271 (1979).
119. Tsujimoto, Y. & Suzuki, Y. The DNA sequence of *Bombyx-mori* fibroin gene including the 5′ flanking, mRNA coding, entire intervening and fibroin protein coding regions. *Cell* **18**, 591–600 (1979).
120. Giroux, M. J. *et al. De novo* synthesis of an intron by the maize transposable element Dissociation. *Proc. Natl Acad. Sci. USA* **91**, 12150–12154 (1994).
The authors show that a transposable element inserted into the *Sh2* gene of maize is sometimes exactly spliced out of transcripts, supporting the idea that transposable element insertions could give rise to new introns in some cases.
121. Rogers, J. H. The role of introns in evolution. *FEBS Lett.* **268**, 339–343 (1990).
122. Roy, S. W. The origin of recent introns: transposons? *Genome Biol.* **5**, 251 (2004).
123. Roy, S. W. & Gilbert, W. Rates of intron loss and gain: implications for early eukaryotic evolution. *Proc. Natl Acad. Sci. USA* **102**, 5773–5778 (2005).
124. Guiliano, D. B. *et al.* Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biol.* **3**, research0057 (2002).
125. Gao, L. Z. & Innan, H. Very low gene duplication rate in the yeast genome. *Science* **306**, 1367–1370 (2004).
126. Krzywinski, J. & Besansky, N. J. Frequent intron loss in the white gene: a cautionary tale for phylogeneticists. *Mol. Biol. Evol.* **19**, 362–366 (2002).
127. Hentze, M. W. & Kulozik, A. E. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* **96**, 307–310 (1999).
128. Ast, G. How did alternative splicing evolve? *Nature Rev. Genet.* **5**, 773–782 (2004).
129. Castillo-Davis, C. I., Mekhedov, S. L., Hartl, D. L., Koonin, E. V. & Kondrashov, F. A. Selection for short introns in highly expressed genes. *Nature Genet.* **31**, 415–418 (2002).
130. Ometto, L., Stephan, W. & De Lorenzo, D. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* **169**, 1521–1527 (2005).
131. Prachumwat, A., DeVincentis, L. & Palopoli, M. F. Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. *Genetics* **166**, 1585–1590 (2004).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
FlyBase: http://flybase.bio.indiana.edu
***Saccharomyces* Genome Database:** http://genome-www.stanford.edu/Saccharomyces
***Schizosaccharomyces pombe* GeneDB:** http://www.genedb.org/genedb/pombe/index.jsp
The *Arabidopsis* Information Resource: http://www.arabidopsis.org
WormBase: http://www.wormbase.org
Access to this interactive links box is free online.