Minireview

The origin of recent introns: transposons?

Scott W Roy

Address: Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. E-mail: scottroy@fas.harvard.edu

Published: 29 November 2004

Genome ***Biology*** 2004, **5:**251

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/12/251

Abstract

The long-standing question of how genes acquire introns has provoked much debate. A recent study makes considerable progress by identifying numerous recently gained introns in nematodes - although it remains difficult to distinguish definitively between models of intron gain.

The origin of spliceosomal introns is one of molecular biology's longest-standing unsolved mysteries. Despite 27 years of extensive study, we are confident of the origin of an intron in only two cases: a short interspersed nucleotide element (SINE) insertion that gave rise to a new intron in the coding region of the catalase A gene of rice [1], and two midge globin genes that acquired an intron via gene conversion with an intron-containing paralog [2]. Previous large-scale studies have failed to find a single convincing case of intron gain since the divergence of humans and mice [3] or a single case of convincing sequence homology between introns in the same genome for a range of taxa [4], and although some other cases of recent intron insertion have been discovered, the sources of these introns remain unknown. Yet, all characterized metazoan species and most other eukaryotes harbor multiple introns per gene, requiring hundreds of thousands, if not millions, of individual intron gains to have occurred throughout eukaryotic evolution.

There are five hypotheses for the origin of new introns (Figure 1). The intron transposition hypothesis states that introns propagate at the RNA level via reinsertion of spliced introns into previously intronless sites in a transcript; the new intron-containing RNA is reverse-transcribed and undergoes gene conversion with the original locus, leading to a new intron (Figure 1a) [5-7]. According to the transposon hypothesis, introns originate as transposon insertions (Figure 1b), as in the case of the new rice intron [1]. The new insertion either serendipitously possesses or quickly acquires signals allowing it to be efficiently spliced out of transcripts [1,8]. The tandem duplication hypothesis (Figure 1c) says that introns originate by tandem genomic duplication of a region containing part or all of an exon, followed by use of the two copies of an internal exonic AGGT sequence as the splice sites for a new intron. The new intron encompasses the 3′ end of the 5′ copy and the 5′ end of the 3′ copy of the duplicated region [9,10]. Genes can also acquire introns via gene conversion by intron-containing paralogs, as with the globin genes [2] - the intron transfer hypothesis (Figure 1d). Finally, spliceosomal introns could originate by the insertion of type II self-splicing introns transferred to the nucleus from an organelle [9,11] (Figure 1e).

Coghlan and Wolfe [12] recently studied newly gained introns in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. They identified 122 apparently recent gains by searching for introns that are present in only one of the two species and are absent from the distantly related parasitic nematode *Brugia malayi* as well as from paralogs and orthologs from several other species. These introns are longer than control introns, are more likely to lie in genes expressed in the germline, and contain more palindromic sequences and microsatellites. The absence of type II introns in *Caenorhabditis* mitochondria rules out the self-splicing intron model as an explanation for the origins of these introns; the authors' requirement that the new intron be at a site which is intronless in known paralogs excludes the intron transfer hypothesis. Coghlan and Wolfe [12] then sought to distinguish between the three remaining hypotheses.

They found that 21 of 81 new introns in *C. elegans* and 7 of 41 in *C. briggsae* show significant sequence similarity to

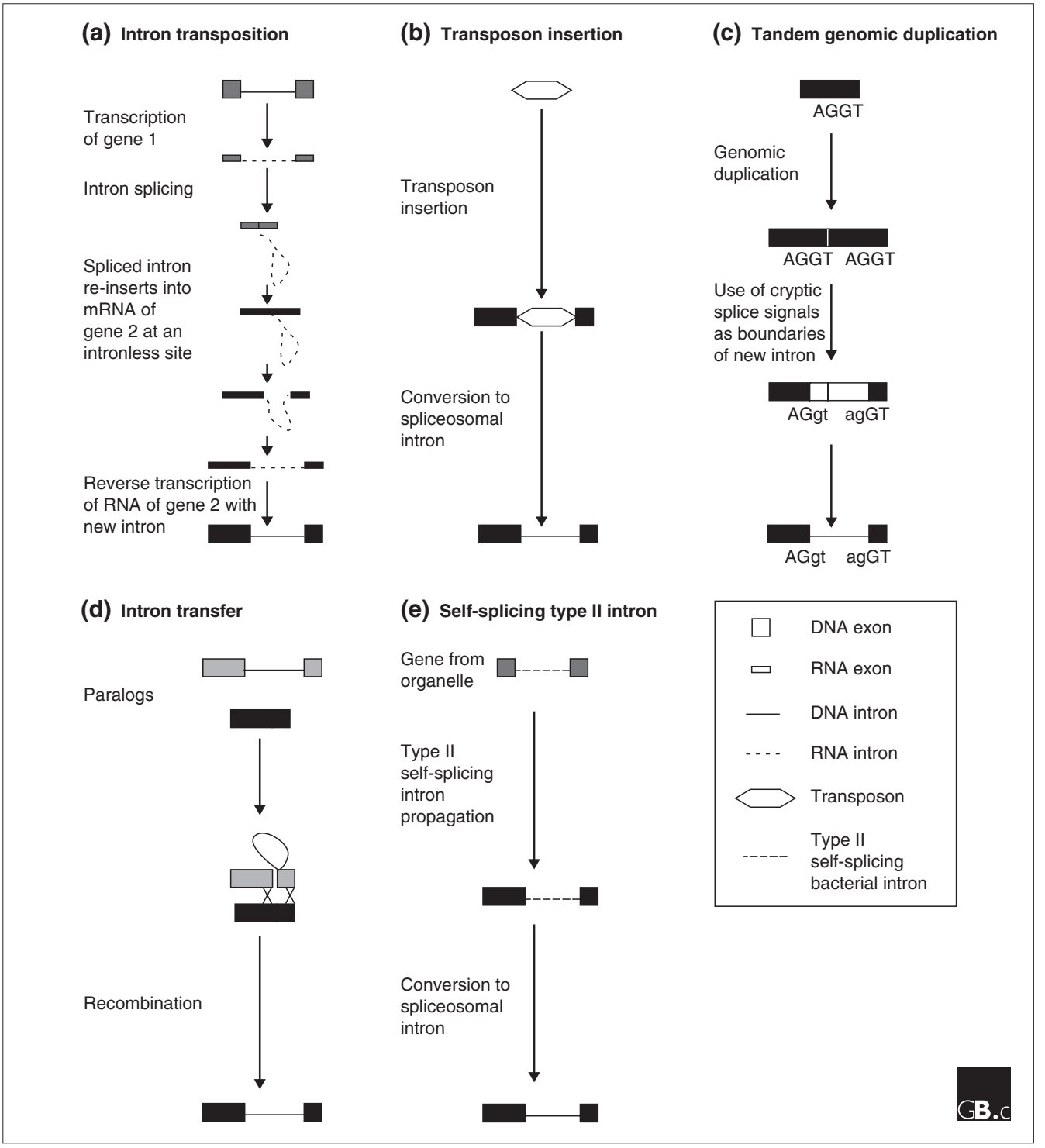

Figure 1
Models of spliceosomal intron gain. **(a)** Intron transposition. An intron from one gene is spliced out of an mRNA transcript. That intronic RNA sequence then reinserts into a previously intronless site of a transcript of the same or a different gene. That structure is then retroposed to give a DNA copy of the gene containing an intron at a new site. The retroposed copy then undergoes gene conversion with a genomic copy (not shown). **(b)** Transposon insertion. A transposon inserts into a contiguous coding region and is transformed into an intron. **(c)** Tandem genomic duplication. A region including part or all of an exon with an internal AGGT is duplicated. The two homologous AGGTs are then used as 5' and 3' splicing boundaries for a new intron comprising the 3' end of the upstream copy and the 5' end of the downstream copy. **(d)** Intron transfer. A gene undergoes a gene conversion or simple double recombination with an intron-containing paralog. **(e)** A self-splicing type II intron, presumably from an organelle of the same organism, inserts into a contiguous region of coding sequence of a nuclear genome and is then converted to a spliceosomal intron.

other introns in the same genome [12]. In three of these 28 cases, two in *C. briggsae* and one in *C. elegans*, the recently gained intron shows homology to another intron in the same gene. In 19 cases, the new intron matches multiple introns in the same genome. Sequence similarity of new introns to other introns is clearly a central expectation of the intron transposition model. Such similarities are also consistent with the transposon model, however, because a second copy of the intron-forming transposon may independently insert into another, previously existing, intron, and with the genomic duplication model as the new intron sequence would be homologous to nearby exonic and intronic sequences. Further analysis showed that the newly gained introns are not enriched for known repetitive elements relative to control introns (apparent evidence against the transposon hypothesis) and that the ends of new introns show no similarity to flanking exonic sequences, apparent evidence against the genomic duplication model. Thus intron transposition seems to be supported by a process of elimination.

The story is not so simple, however. Figure 2 shows the observed intron pattern for one of the *C. briggsae* genes with a newly acquired intron and its orthologs from other species. Newly gained intron 3 shows sequence homology to both introns 1 and 5 of the same gene. The three introns each contain multiple copies of a roughly 170 base-pair palindromic element. *C. briggsae* intron 3 is putatively new, so its sequence similarity to the other introns could be due to intron transposition. Introns 1 and 5 are shared with humans however, and thus date to at least the Cambrian explosion, far too long ago for intronic sequences to maintain the observed high level of sequence similarity. Indeed, the sequences of the corresponding introns in *Caenorhabditis remanei*, *C. elegans*, and *Homo sapiens* do not contain the palindromic sequence. Thus the sequence similarity between introns 1 and 5 appears to be not a vestige of intron birth but due to more recent events, most probably independent insertions of the same palindromic element into both pre-existing introns. In fact, the palindromic elements present in these introns are quite common in the *C. briggsae* genome - a nucleotide similarity search, using BLASTN [13] comparing the third intron to the whole *C. briggsae* genome yields over one hundred hits with Expectation (*e*) values of less than 10^{-10}.

This raises the possibility that intron 3 acquired this palindromic element not by transposition of another intron but by a third transposon insertion, either into a pre-existing intron 3 or into a contiguous coding region, leading to the creation of intron 3 (the transposon model). The finding of Coghlan and Wolfe [12] that new introns are generally enriched in palindromic sequences suggests the latter. The possibility of intron origin by insertion of palindromic transposons is enticing, because the tendency of palindromic elements to form hairpin structures could bring the 5′ and 3′splice sites of the new intron into proximity,

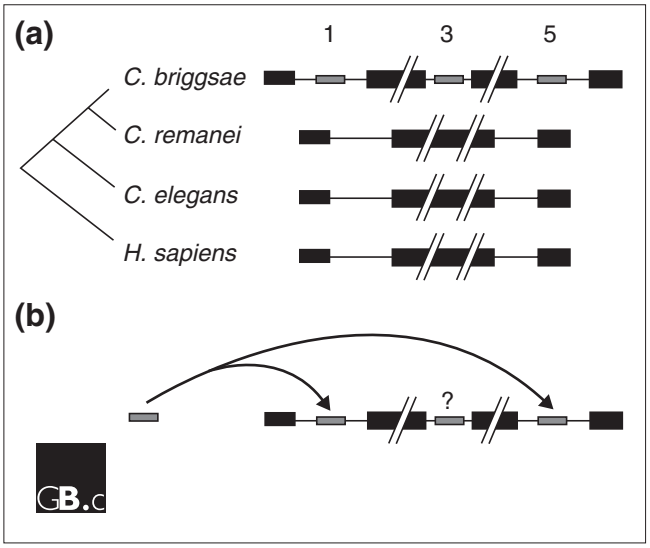

Figure 2
Pattern of conservation between species for one of the *C. briggsae* genes with a newly acquired intron, and a hypothesis for intron acquisition. **(a)** Intron presence and absence, and sequence similarity, for *C. briggsae* gene CBG18597 and orthologs. Introns 1 and 5 are common to all orthologs; intron 3 is unique to *C. briggsae*. Black boxes represent exons and lines represent introns. The gray boxes represent the common palindromic sequence within introns. Other introns in the genes are omitted for simplicity (indicated by breaks in the boxes). Not drawn to scale. **(b)** Probable origin of the sequence similarity between introns 1 and 5. The absence of the common palindromic sequence in introns of orthologs suggests that both *C. briggsae* introns 1 and 5 acquired the sequence through independent recent transposon insertions. Intron 3 may contain the common sequence as a result of transposition of another intron (interpretation of Coghlan and Wolfe [12]), by creation of the intron by a third transposon-insertion event (argued here) or via a third transposon insertion into a previously existing intron 3.

perhaps facilitating splicing. (A shorter hairpin structure is maintained by selection in the first intron of the *Adh* gene in *Drosophila melanogaster* [14].) The intron sequence could then gradually lose its palindromic character as other compensatory local mutations increased the intron's splicing efficiency, leading eventually to the quasi-random sequence characteristic of most introns. Although the authors' [12] finding that recently gained introns are not enriched in known repetitive elements seems to be evidence against transposon origins for these introns, this could be reconciled if the palindromic elements involved are extinct, and their extant copies too diverged (the intron matches in the Coghlan and Wolfe study [12] show around 70% nucleotide identity) to warrant inclusion in libraries of known transposable elements.

Other mechanisms could also account for the excess of palindromic elements in new introns, however. Regions with more stable DNA secondary structures (such as palindromic elements) are expected to experience more replication slippage, leading to higher rates of duplication of short-to-medium stretches of DNA. If such duplications occasionally

lead to the creation of new introns (the tandem duplication hypothesis), these introns would themselves contain the palindromic sequences of adjacent regions. That the authors find no similarity between the terminal 25 base-pair regions of new introns and those of flanking exons could be due to the age of the gains (the levels of observed sequence similarity in the study are around 70%, a level that is not significant over short stretches) and/or to stronger positive selection near the boundaries of the new introns. The higher frequency of intron-acquiring genes in the germline is, however, harder to explain by genomic duplication except by recourse to the generally faster evolution of germline genes. Also, these arguments do not exclude intron transposition as a possibility. As pointed out by the authors [12], the palindromic character of new introns could reflect longer survival times of introns with stable secondary structures, affording more opportunity to be reverse-spliced. In cases where the new intron is homologous both to a transposon and to another intron, however, it seems more parsimonious to postulate a reasonably common single transposon insertion rather than a series of three rarer events (intron reinsertion, transcript retroposition and gene conversion).

What evidence remains for intron transposition? First, germline-expressed genes preferentially acquire introns, as would be expected if intron gain occurs at the RNA level, although this could instead reflect preferential insertion of palindromic elements into actively transcribing regions [15] or generally faster evolution of germline-expressed genes. Second, genes involved in mRNA processing and splicing preferentially gain introns. This is a surprise under any model, though it does intuitively seem to implicate the spliceosome in intron gain. As Coghlan and Wolfe [12] point out, however, it is hard to imagine why a mechanism that inserts introns via a protein complex would tend to favor insertion into the genes coding for these proteins. More attention will be necessary to determine the cause and generality across taxa of this intriguing bias. By identifying clear recent intron gains, Coghlan and Wolfe [12] have taken a large step forward in deciphering the origins of introns. That even this study is subject to interpretation underscores the slipperiness of the problem. The increasing focus of sequencing projects on closely related genomes is promising, and similar comparative studies in other taxa should help to finally unravel this mystery.

References

1. Iwamoto M, Maekawwa M, Saito A, Higo H, Higo K: **Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots.** *Theor Appl Genet* 1998, **97:**9-19.
2. Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER: **A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain.** *Gene* 1997, **205:**151-160.
3. Roy SW, Fedorov A, Gilbert W: **Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain.** *Proc Natl Acad Sci USA* 2003, **100:**7158-7162.
4. Fedorov A, Roy S, Fedorova L, Gilbert W: **Mystery of intron gain.** *Genome Res* 2003, **13:**2236-2241.
5. Cavalier-Smith T: **Selfish DNA and the origin of introns.** *Nature* 1985, **315:**283-284
6. Palmer JD, Logsdon JM Jr.: **The recent origin of introns.** *Curr Opin Genet Dev* 1991, **1:**470-477.
7. Logsdon JM Jr.: **Worm genomes hold the smoking guns of intron gain.** *Proc Natl Acad Sci* 2004, **101:**11195-11196.
8. Crick F: **Split genes and RNA splicing.** *Science* 1979, **204:**264-271.
9. Rogers JH: **How were introns inserted into nuclear genes?** *Trends Genet* 1989, **5:**213-216.
10. Venkatesh B, Ning Y, Brenner S: **Late changes in spliceosomal introns define clades in vertebrate evolution.** *Proc Natl Acad Sci USA* 1999, **96:**10267-10271.
11. Cavalier-Smith T: **Intron phylogeny: a new hypothesis.** *Trends Genet* 1991, **7:**145-148.
12. Coghlan A, Wolfe KH: **Origins of recently gained introns in *Caenorhabditis*.** *Proc Natl Acad Sci USA* 2004, **101:**11362-11367.
13. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/blast/]
14. Chen Y, Stephan W: **Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster Adh* gene.** *Proc Natl Acad Sci USA* 2003, **100:**11499-11504.
15. Timakov B, Liu X, Turgut I, Zhang P: **Timing and targeting of p-element local transposition in the male germline cells of *Drosophila melanogaster*.** *Genetics* 2002, **160:**1011-1022.