

LETTERS

Smoke Without Fire: Most Reported Cases of Intron Gain in Nematodes Instead Reflect Intron Losses

Scott William Roy and David Penny

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Identification of recently gained spliceosomal introns would provide crucial evidence in the continuing debate concerning the age and evolutionary significance of introns. A previously published genomic analysis reported to have identified 122 introns that had been gained since the divergence of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* ~100 MYA. However, using newly available genomic sequence from additional *Caenorhabditis* species, we show that 74% (60/81) of the reported gains in *C. elegans* are present in a *C. briggsae* relative. This pattern indicates that these introns represent losses in *C. briggsae*, not gains in *C. elegans*. In addition, 61% (25/41) of the reported gains in *C. briggsae* are present in the more distant *C. briggsae* relative, in a pattern suggesting that additional reported gains in *C. elegans* and/or *C. briggsae* may in fact represent unrecognized losses. These results underscore the dominance of intron loss over intron gain in recent eukaryotic evolution, the pitfalls associated with parsimony in inferring intron gains, and the importance of genomic sequencing of clusters of closely related species for drawing accurate inferences about genome evolution.

The origin of the spliceosomal introns of eukaryotes constitutes a 30-year-old mystery (de Souza 2003; Jeffares et al. 2006; Martin and Koonin 2006; Rodríguez-Trelles et al. 2006; Roy and Gilbert 2006). In 1998, Logsdon et al. laid out conditions for determining the source of a recently gained intron: 1) strong evidence for the intron's recent gain, derived from "dense phylogenetic sampling"; and 2) the "molecular smoking gun"—an intronic sequence whose clear similarity to another genetic element betrays the intron's origin. The subsequent years have been an extremely active time for the study of intron evolution (Tarrío et al. 1998, 2003; Venkatesh et al. 1999; Sakharkar et al. 2001; Seo et al. 2001; Wolf et al. 2001; Fedorov et al. 2002; Llopert et al. 2002; Wada et al. 2002; Bon et al. 2003; Fedorov et al. 2003; Rogozin et al. 2003; Nielsen et al. 2004; Slamovits and Keeling 2006). However, until 2004 only a single clearly characterized intron gain had been reported (Iwamoto 1998). Then finally, 2 years ago Coghlan and Wolfe (2004) reported the cases of 81 potentially recently gained introns in *Caenorhabditis elegans* and 41 in *Caenorhabditis briggsae*. Each of the 122 introns was not found in the other *Caenorhabditis* species or in various outgroups (apparently fulfilling the first criterion; fig. 1), and 28 of the introns showed sequence similarity to other *Caenorhabditis* introns (apparently fulfilling the second). These results have been widely discussed (e.g., Roy 2004; Rodríguez-Trelles et al. 2006) and widely cited (>35 citations) and were hailed by Logsdon (2004) as the long-awaited "smoking gun."

The first point of contention concerned interpretation of the intron sequences themselves (the second criterion). Coghlan and Wolfe (2004) as well as Logsdon (2004) interpreted observed sequence similarity between the reported apparent gains and other introns as evidence of intron gain by transposition of existing introns into new positions in the same or different genes. However, sequence similarity

between introns often spanned only a fraction of the introns' lengths, evidence against a simple intron transposition event (Roy 2004; Roy SW, unpublished data). Also, regions of similarity were often limited to many-copy repetitive elements, which were also found in intergenic regions, leading one of us to suggest that the reported gains might instead be due to transposable element (TE) insertions (Roy 2004).

Here, we report evidence that many of the reported intron gains are not even true intron gains (criterion 1) but instead reflect intron losses. We examined putatively orthologous sequences from newly available genomic sequences from 2 relatives of *C. briggsae*: *Caenorhabditis remanei* and *Caenorhabditis sp. 4* (fig. 1). If the 81 *C. elegans* introns reported to be recent gains are in fact just that, they should clearly be absent from these species. Instead, 74% (60/81) were found to be shared with one or both species (table 1, see e.g., in fig. 2; a more detailed summary is available as Supplementary Material Online). This implies that these introns' absence in *C. briggsae* is due to intron loss and not due to recent gain in *C. elegans*. The remaining 21 possible *C. elegans* gains may either be actual gains in *C. elegans* or losses in the *C. briggsae*–*C. sp. 4* ancestor (table 1, branch ii). Thus, in most cases (at least for *C. elegans*), there is no smoking gun.

Among the 41 reported *C. briggsae* gains, fully 61% (25) are present in *C. sp. 4* and, thus, could represent gains in the *C. briggsae*–*C. remanei*–*C. sp. 4* ancestor (branch ii) or losses in *C. elegans*. If all 25 of these introns and all 21 *C. elegans*–specific introns were true intron gains, there would be no losses but numerous gains in both *C. elegans* (branch i) and in the *C. briggsae*–*C. sp. 4* ancestor (branch ii). This would be surprising in light of the observation of 3.75 losses per gain (60/16) in the sample occurring in *C. briggsae* since the *C. briggsae*–*C. sp. 4* ancestor (branch iii). This suggests that some of the remaining possible gains may also represent unrecognized losses. Though direct estimation of the loss/gain numbers here is not possible, the ratio of losses to gains in both the *C. briggsae*–*C. sp. 4* ancestor and in *C. elegans* would also equal 3.75, if there were 5.6 gains and 20.9 losses in *C. elegans* and 4.1 gains and 15.4 losses in the

Key words: intron gain, genome complexity, genome annotation, genome sequencing, genome evolution, parsimony.

E-mail: scottwroy@gmail.com.

Mol. Biol. Evol. 23(12):2259–2262. 2006

doi:10.1093/molbev/msl098

Advance Access publication August 30, 2006

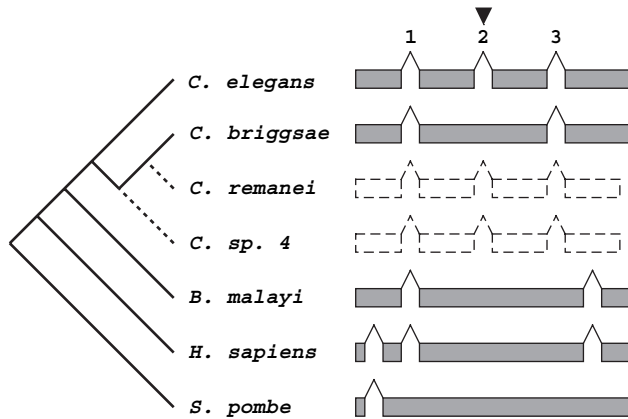


FIG. 1.—Some previously reported intron gains instead represent intron losses. Coghlan and Wolfe (2004) previously aligned homologous sequences from *Caenorhabditis elegans* and *Caenorhabditis briggsae*, the distantly related nematode *Brugia malayi*, and other available outgroups (but not *Caenorhabditis remanei* or *Caenorhabditis sp. 4*). They identified introns specific to one of the two *Caenorhabditis* species and absent in outgroups (e.g., intron 2), as recent intron gains. However, although 85% of introns are shared between *C. elegans* and *C. briggsae* (Kent and Zahler 2000), intron conservation between *Caenorhabditis* and available outgroups is 50% or less (Guiliano et al. 2002; Rogozin et al. 2003), thus some introns present in the *C. elegans*–*C. briggsae* ancestor may not be represented in outgroups (e.g., intron 3). We assessed presence of reportedly gained introns in *C. remanei* and *C. sp. 4*. As shown in the figure, most reported *C. elegans* gains are present in one or both species (e.g. intron 2), suggesting intron presence in the *C. elegans*–*C. briggsae* ancestor and subsequent loss in *C. briggsae*. In addition, we assessed intron presence/absence of reported *C. briggsae* intron gains in the 2 *C. briggsae* relatives (not shown). Species relationships are from Cho et al. 2004 and Kiontke et al. 2004.

C. briggsae–*C. sp. 4* ancestor. In this case, only 21% ($16 + 5.6 + 4.1 = 25.7$ out of 122) of the reported intron gains would represent true intron gains.

Most biases identified by Coghlan and Wolfe (2004) among the 122 reported gains are not strong for the remaining 21 possible *C. elegans* gains and 16 probable *C. remanei* gains. Only 4/37 are found in genes involved in mRNA splicing. Only 27.0% (10/37) show sequence

similarity to other introns, similar to the *C. briggsae* losses (25.0%, 15/60). Only the reported bias toward oocyte expression remains: among genes that contain *C. elegans* gains and/or probable *C. briggsae* gains for which oocyte expression is available (Hill et al. 2000), 70% (16/23) are present in oocytes, more than 42% for all genes assessed ($P \sim 0.01$ by a Fisher's Exact test). This bias could reflect additional undetected oocyte-biased intron losses, a higher frequency of insertion of intron-creating TE insertions into germline-expressed genes (perhaps due to more accessible chromatin structure), or a dependency of intron gain on an mRNA intermediate, as suggested by Coghlan and Wolfe (2004).

The suggestive biases among the introns reported by Coghlan and Wolfe (sequence similarity to other introns/TEs, gene biases toward germline expression and involvement in mRNA processing) are thus apparently a case of smoke without fire (or evidence without a crime), as these introns are primarily derived from cases of intron loss, not gain. Why should introns in one species that are lost in the other show sequence similarity to other introns/TEs? It seems likely that these sequence similarities are due to independent intronic and intergenic insertions of the same TE. If rates of intron loss and intronic TE insertion were correlated due to the dependence of both processes on local recombination rate or general DNA accessibility (chromatin structure) in the germline, the same introns that are lost in one species might tend to experience TE insertion in the other. Preferential intron loss from germline-expressed genes could reflect mRNA-mediated intron loss (Mourier and Jeffares 2003), as could the bias toward mRNA splicing-related genes, though why splicing-associated transcripts, as opposed to proteins, should associate with the spliceosome is unclear (Coghlan and Wolfe 2004). These surprising biases in intron loss deserve further attention.

These results highlight 3 important points. First, along many eukaryotic lineages, recent evolution has been characterized by a dominance of intron loss over intron

Table 1
Observed phylogenetic patterns among 122 *Caenorhabditis* introns previously reported to be recently gained (Coghlan and Wolfe 2004)

				Number	Conclusion
<i>Cele</i>	<i>Csp4</i>	<i>Crem</i>	<i>Cbri</i>		
+	+/?*	+	–	41	<i>Cbri</i> loss
+	+	–/?**	–	12	<i>Cbri</i> loss
+	–	+	–	7	<i>Cbri</i>, <i>Csp4</i> double loss
+	–	–	–	21	<i>Cele</i> gain or <i>Cbri</i> loss
–	+	+	+	21	<i>Cbri</i> gain or <i>Cele</i> loss
–	–	+	+	4	Probable <i>Cbri</i> gain
–	–	–	+	12	Probable <i>Cbri</i> gain
–	+	–	+	4	<i>Crem</i> loss, <i>Cele</i> loss or branch ii gain

NOTE.—*Cele*, *Cbri*, *Crem*, and *Csp4* indicate *Caenorhabditis elegans*, *Caenorhabditis remanei*, *Caenorhabditis briggsae*, and *Caenorhabditis sp. 4*, respectively. Intron presence (+), absence (–), or uncertainty (?) are indicated. The most likely conclusion(s) for each category, as discussed in the main text, are listed under "Conclusion." Bold text under "Conclusion" indicates likely intron loss. The top and bottom 4 categories were previously reported to be intron gains in *C. elegans* and *C. briggsae*, respectively. Branches i, ii, and iii, as referred to in the text, are indicated. *, Intron presence/absence in *C. sp4* was uncertain for 2/41 cases. **, Intron presence/absence in *C. remanei* was uncertain for 4/12 cases. A more detailed summary of our results is available as Supplementary Material Online.

Cele	ATG	GGA	GGG	AGT	GGT	GCC	GGT	AAG	ACG	ACT	CTG	ATG	AAT	ATC	CTG	GCC	CAT	TTG	GAT	ACT	AAC	GGA		
Cbri	ATG	GGA	GGT	AGT	GGA	GCC	GGA	AAA	ACC	ACT	TTA	ATG	AAT	ATT	TTG	GCT	CAT	TTG	GAT	ACG	AAT	GGC		
Crem	ATG	GGC	GGA	AGT	GGA	GCA	GGT	AAA	ACG	ACA	CTG	ATG	AAT	ATA	CTT	GCT	CAT	TTG	GAT	ACC	AAT	GGA		
Csp4	ATG	GGT	GGT	AGT	GGA	GCC	GGC	AAA	ACA	ACT	CTG	ATG	AAC	ATT	CTG	GCT	CAT	TTG	GAT	ACC	AAT	GGA		
	M	G	G	S	G	A	G	K	T	T	L	M	N	I	L	A	H	L	D	T	N	G		
Cele	GTT	GAG	gtgagcccagctccccgagctacctatcattttctcgatttttcag [46bp]																TAC	CTC	GGC			
Cbri	GTT	GAA	[intron has been lost]																			TAT	TAT	GGT
Crem	GTG	GAG	gtgggacatctggacaagacaacccgatgggggttttttttttcaaaatcttgatttcag [60bp]																TAT	TAC	GGA			
Csp4	GTT	GAG	gtatgtgggcgagctttctgaattcacccaaataaataattacag [46bp]																TAC	TAC	GGT			
	V	E	<u>INTRON</u>																Y	L/Y	G			
Cele	GAC	GTC	ACT	GTC	AAT	GGC	AAG	AAG	ATC	ACC	AAG	CAG	AAA	ATG	CGG	CAA	ATG	TGC	GCC	TAC	GTT			
Cbri	GAC	GTC	ACT	GTA	AAT	GGA	AAA	AAG	ATC	ACA	AAA	CAA	AAA	ATG	CGT	CAA	ATG	TGT	GCG	TAT	GTG			
Crem	GAT	GTG	ACG	GTC	AAC	GGG	AAG	AAG	ATA	ACC	AAA	CAG	AGA	ATG	CGT	CAA	ATG	TGT	GCA	TAT	GTT			
Csp4	GAC	GTG	ACT	GTC	AAT	GGA	AAA	AAG	ATC	ACG	AAA	CAA	AAG	ATG	CGT	CAA	ATG	TGT	GCC	TAC	GTT			
	D	V	T	V	N	G	K	K	I	T	K	Q	K/R	M	R	Q	M	C	A	Y	V			

FIG. 2.—Example of a reported intron gain in *Caenorhabditis elegans*, which instead appears to be an intron loss in *Caenorhabditis briggsae*. The first intron in *C. elegans* gene F02E11.1 was reported by Coghlan and Wolfe (2004) to have been recently gained since the *C. elegans*–*C. briggsae* ancestor. The first and second *C. elegans* exons (uppercase) and intervening intron (lowercase) are shown, along with putatively orthologous sequences for *C. briggsae*, *Caenorhabditis remanei*, and *Caenorhabditis sp. 4*. The intron is clearly present in the *C. briggsae* relatives *C. remanei* and *C. sp. 4*, suggesting intron loss in *C. briggsae*. Typically, that the intervening *C. remanei* and *C. sp. 4* sequences are in fact intronic is reinforced by the “gt...ag” boundary structure (underlined), the presence of in-frame stop codons in both sequences (bold), and the fact that the intervening sequence in *C. sp. 4* is not a multiple of 3 bp. The sequences are from *C. elegans* gene F02E11.1, *C. briggsae* gene CBG04332, *C. remanei* contig Cont107.34, and *C. sp. 4* read CPAA-aga86e12.b1, respectively. Cele, Cbri, Crem, and Csp4 indicate *C. elegans*, *C. briggsae*, *C. remanei*, and *C. sp. 4*, respectively.

gain (e.g., Roy et al. 2003; Cho et al. 2004; Kiontke et al. 2004; Lin et al. 2006; Roy and Hartl 2006; Stajich and Dietrich 2006): in this case, even the introns that appeared most likely to represent cases of recent gains are instead mostly due to loss. Second, these results provide an important case study of the utility of parsimony in the face of high degrees of evolutionary change. Third, these results demonstrate the importance of greater taxonomic sampling and the indispensability of sequencing additional genomes for answering even seemingly straightforward questions about genome structure and evolution.

The general dearth of clear recent intron gains continues to frustrate attempts to understand mechanisms and causes of intron creation (Roy et al. 2003; Lin et al. 2006; Roy and Hartl 2006; Roy et al. 2006; Stajich and Dietrich 2006). These observed low intron gain rates are curious as huge numbers of introns in various eukaryotic genomes attest to substantial intron creation at some point in evolution. One possible explanation is improved policing of genome insertions in modern eukaryotes relative to early/pre-eukaryotic evolution. Investigation is ongoing.

Methods

From the text and supplementary materials of Coghlan and Wolfe 2004, we extracted amino acid sequences flanking apparent intron gains, gene functions, and names of introns exhibiting sequence similarity to other introns from the same genome. We performed TblastN searches against the assembled *C. remanei* genome (version 1, downloaded from Wormbase [http://www.wormbase.org]). Intron presence/absence was determined by either 1) the presence of a gain in the resultant alignment, almost always with stop codons in the gapped *C. remanei* sequence or 2) the presence of 2 independent HSPs to the same contig, one upstream and one downstream of the intron position, with the sequence stopping abruptly at the intron position,

and with a significant intervening gap. Both the best hit and other highly significant hits were surveyed to determine intron presence in all possible orthologous or closely related sequences. An analogous search was made of the 2,714,032 available genomic shotgun sequencing reads for *Caenorhabditis sp. 4*, downloaded from TraceDB (www.ncbi.nlm.nih.gov). Following Coghlan and Wolfe, genes that are always or sometimes expressed in oocytes were determined from the oligonucleotide studies of Hill et al. (2000)

Supplementary Material

Summary of results including for each intron, the gene name, intron number, and presence (+), absence (–), or uncertainty (?) in the *C. briggsae* relatives is available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank Manuel Irimia for constructive comments and helpful discussions during the preparation of this manuscript.

Literature Cited

- Bon E, Casaregola S, Blandin G, et al. (11 co-authors). 2003. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res.* 31:1121–1135.
- Cho S, Jin SW, Cohen A, Ellis RE. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* 14:1207–1220.
- Coghlan A, Wolfe KH. 2004. Origins of recently gained introns in *Caenorhabditis*. *Proc Natl Acad Sci USA.* 101:11362–11367.
- de Souza SJ. 2003. The emergence of a synthetic theory of intron evolution. *Genetica.* 118:117–121.

- Fedorov A, Merican AF, Gilbert W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci USA*. 99:16128–16133.
- Fedorov A, Roy S, Fedorova L, Gilbert W. 2003. Mystery of intron gain. *Genome Res*. 13:2236–2241.
- Guiliano DB, Hall N, Jones SJ, Clark LN, Corton CH, Barrell BG, Blaxter ML. 2002. Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biol*. 3:RESEARCH0057.
- Hill AA, Hunter CP, Tsung BT, Tucker-Kellogg G, Brown EL. 2000. Genomic analysis of gene expression in *C. elegans*. *Science*. 290:809–812.
- Iwamoto M, Maekawa M, Saito A, Higo H, Higo K. 1998. Evolutionary relationship of plant catalase genes inferred from exon-intron structures: isozyme divergence after the separation of monocots and dicots. *Theor Appl Genet*. 97:9–19.
- Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet*. 22:16–22.
- 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.
- Kiontke K, Gavin NP, Raynes Y, Rehrig C, Piano F, Fitch DH. 2004. Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc Natl Acad Sci USA*. 101:9003–9008.
- Lin H, Zhu W, Silva JC, Gu X, Buell CR. 2006. Intron gain and loss in segmentally duplicated genes in rice. *Genome Biol*. 7:R41.
- Llopart A, Comeron 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.
- Logsdon JM Jr. 2004. Worm genomes hold the smoking guns of intron gain. *Proc Natl Acad Sci USA*. 101:11195–11196.
- Logsdon JM Jr, Stoltzfus A, Doolittle WF. 1998. Molecular evolution: recent cases of spliceosomal intron gain? *Curr Biol*. 8:R560–R563.
- Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature*. 440:41–45.
- 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.
- Rodríguez-Trelles F, Tarrío R, Ayala FJ. 2006. Origins and evolution of spliceosomal introns. *Annu Rev Genet*. 40:47–76.
- 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. 2004. The origin of recent introns: transposons? *Genome Biol*. 5:251.
- Roy SW, Fedorov A, Gilbert W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci USA*. 100:7158–7162.
- Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*. 7:211–221.
- Roy SW, Hartl DL. 2006. Very little intron loss/gain in *Plasmodium*: intron loss/gain mutation rates and intron number. *Genome Research*. 16:750–756.
- Roy SW, Irimia M, Penny D. 2006. Very little intron gain in *Entamoeba histolytica* genes laterally transferred from prokaryotes. *Mol Biol Evol*. 23:1824–1827.
- Sakharkar MK, Tan TW, de Souza SJ. 2001. Generation of a database containing discordant intron positions in eukaryotic genes (MIDB). *Bioinformatics*. 17:671–675.
- Seo H-C, Kube M, Edvardsen RB, et al. (11 co-authors). 2001. Miniature genome in the marine chordate *Oikopleura dioica*. *Science*. 294:2506.
- Slamovits CH, Keeling PJ. 2006. A high density of ancient spliceosomal introns in oxymonad excavates. *BMC Evol Biol*. 6:34.
- Stajich JE, Dietrich FS. 2006. Evidence of mRNA-mediated intron loss in the human-pathogenic fungus *Cryptococcus neoformans*. *Eukaryotic Cell*. 5:789–793.
- Tarrío R, Rodríguez-Trelles F, Ayala FJ. 1998. New *Drosophila* introns originate by duplication. *Proc Natl Acad Sci USA*. 95:1658–1662.
- Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2003. A new *Drosophila* spliceosomal intron position is common in plants. *Proc Natl Acad Sci USA*. 100:6580–6583.
- Venkatesh B, Ning Y, Brenner S. 1999. Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc Natl Acad Sci USA*. 96:10267–10271.
- Wada H, Kobayashi M, Sato R, Satoh N, Miyasaka H, Shirayama Y. 2002. Dynamic insertion-deletion of introns in deuterostome EF-1 alpha genes. *J Mol Evol*. 54:118–128.
- Wolf YI, Kondrashov FA, Koonin EV. 2001. Footprints of primordial introns on the eukaryotic genome: still no clear traces. *Trends Genet*. 17:499–450.

William Martin, Associate Editor

Accepted August 24, 2006