Origins of recently gained introns in *Caenorhabditis*

Avril Coghlan and Kenneth H. Wolfe*

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin 2, Ireland

Edited by Jeffrey Donald Palmer, Indiana University, Bloomington, IN, and approved May 28, 2004 (received for review December 10, 2003)

The genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* both contain ≈100,000 introns, of which >6,000 are unique to one or the other species. To study the origins of new introns, we used a conservative method involving phylogenetic comparisons to animal orthologs and nematode paralogs to identify cases where an intron content difference between *C. elegans* and *C. briggsae* was caused by intron insertion rather than deletion. We identified 81 recently gained introns in *C. elegans* and 41 in *C. briggsae*. Novel introns have a stronger exon splice site consensus sequence than the general population of introns and show the same preference for phase 0 sites in codons over phases 1 and 2. More of the novel introns are inserted in genes that are expressed in the *C. elegans* germ line than expected by chance. Thirteen of the 122 gained introns are in genes whose protein products function in premRNA processing, including three gains in the gene for spliceosomal protein SF3B1 and two in the nonsense-mediated decay gene *smg-2*. Twenty-eight novel introns have significant DNA sequence identity to other introns, including three that are similar to other introns in the same gene. All of these similarities involve minisatellites or palindromes in the intron sequences. Our results suggest that at least some of the intron gains were caused by reverse splicing of a preexisting intron.

How introns spread within and among genes remains a central but largely unresolved question in evolutionary biology (1–4). Although genome-scale studies have shown that both losses and gains of introns occurred at substantial rates during the evolution of the major eukaryotic lineages (5), studies focused on more recent evolutionary periods have found many examples of losses but few gains. A survey of mammalian genes found six cases of intron losses in rodents relative to human but no intron gains (6). Recent intron losses are also frequently seen in plant genes (7). Fedorov *et al.* (8) did not detect any recently duplicated (i.e., gained) introns within the genome sequences of human, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. Finding recent intron gains and identifying the origin of their DNA is likely to be a key to understanding where new introns come from.

Although few in number, some examples of recent intron gain are supported by strong evidence. Logsdon *et al.* (9) compared triose-phosphate isomerase genes from many different eukaryotes and found that in some cases an intron's phylogenetic distribution could be explained by either a single gain or up to 12 losses. Other convincing gains have been found in the fruit fly xanthine dehydrogenase gene (10), in the globin genes of midges (11), in the rice catalase gene (12, 13), and in *C. elegans* chemoreceptor genes (14). Despite the evidence that intron gains occur, the mechanism is unknown.

Five different mechanisms by which spliceosomal introns could be gained have been proposed and are summarized briefly here. (*i*) Shortly after the discovery of introns, Crick (15) hypothesized that novel introns arise by insertion of a transposon. There is a large body of evidence that some recent insertions of transposable elements in laboratory strains of animals and plants can be spliced out, often with little or no phenotypic consequence (16, 17). However, the sole possible example of this occurring on an evolutionary timescale is the similarity between a short novel intron in the catalase gene of some rice species and a SINE element (13). (*ii*) Rogers (18) suggested that new spliceosomal introns may originate by insertion of a group II intron via reverse self splicing, but there is no evidence to support this. (*iii*) Rogers also proposed that novel introns could be formed by tandem duplication of an internal fragment of an exon containing the sequence AGGT, with activation of the resultant cryptic splice sites (18). Three novel introns in fish may have been formed this way (19). (*iv*) A preexisting spliceosomal intron could be reverse-spliced into a new site in the same or a different mRNA, which is then reverse-transcribed to a cDNA that recombines with the genome (20). Tarrío *et al.* (10) attributed three novel introns in fly xanthine dehydrogenase genes to this mechanism, but their analysis has been questioned (2). (*v*) An unspliced mRNA could be reverse-transcribed and the cDNA recombine with a homologous gene in the genome that previously lacked an intron at that site. There is strong evidence that an intron was gained in a midge globin gene by this mechanism (11).

Nematode genes have a particularly high rate of intron turnover compared to other animals, as first noticed by Logsdon *et al.* (9). By comparing the whole *C. elegans* genome to 8% of that of its sister species *Caenorhabditis briggsae*, Kent and Zahler (21) found evidence of ≈250 introns present in one species but not in the other. Recently, we reported that in 12,155 orthologous gene pairs in the whole genomes of *C. elegans* and *C. briggsae*, there are 4,379 *C. elegans*-specific introns and 2,200 *C. briggsae*-specific introns (22). We estimated that intron gains or losses have occurred at a rate of at least 0.005 per gene per million years in nematodes, which far exceeds the rate in chordates (22). Intron–exon structure seems to be in flux across the entire phylum Nematoda: in 11 orthologs compared between *C. elegans* and its distant relative *Brugia malayi*, only 50% of *C. elegans* introns are conserved in *B. malayi*, and 25% of *B. malayi* introns are conserved in *C. elegans* (23).

Here, we searched for novel introns that can be identified unambiguously as having been gained after the divergence of *C. elegans* and *C. briggsae*. Our results point to reverse splicing of preexisting introns (20) as the main mechanism of intron gain during recent nematode evolution.

Methods

Here we summarize our methods; a more detailed description is included as *Supporting Methods* and Appendix 1, which are published as supporting information on the PNAS web site.

Sequence Data and Homolog Sets. The *C. elegans* data set was Wormpep 104 (19,588 proteins). The *C. briggsae* data set (19,507 proteins) was created as part of its genome project (22). We used Ensembl human release 15.33.1, mouse release 15.30.1, *Drosophila* release 15.3a.1, and *Anopheles* release 15.2.1. For each *C. elegans* or *C. briggsae* gene, we searched for homologs in six animal genomes (*C. elegans*, *C. briggsae*, human, mouse, fruit fly,

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviation: snRNA, small nuclear RNA.

See Commentary on page 11195.

*To whom correspondence should be addressed. E-mail: khwolfe@tcd.ie.

www.pnas.org/cgi/doi/10.1073/pnas.0308192101

and mosquito) by BLASTP (24). We sorted the homologs of a gene in order of significance and took the most significant hits. We found homolog groups for 16,590 *C. elegans* genes and 16,438 *C. briggsae* genes.

Detecting Intron Gains from Alignments. The proteins in each of the 33,028 homolog groups were aligned by using CLUSTALW (25). To detect recently gained introns, we mapped intron positions onto the protein sequence alignment. If *C. briggsae* or *C. elegans* gene *A* has an intron A_i after its *i*th amino acid residue, and residue *i* is at the *j*th position of the alignment, then intron A_i is at the *j*th position of the alignment. We excluded introns that fall in unreliable regions of the alignment, considering intron positions to be reliable only if (*i*) $\geq 5/10$ of the aligned residues $j-9$ to j, and $\geq 5/10$ of those from $j+1$ to $j+10$ are either identical or conserved among all of the sequences in the homolog group from the six animal genomes; and (*ii*) there are no gaps between positions $j-9$ to $j+10$. Taking only those introns whose positions are reliable, an intron is considered as a putative recent gain in gene *A* if there is no intron in any of the homologs of *A* from $j-4$ to $j+5$, to exclude possible intron sliding. This analysis yielded 244 putative novel introns in *C. elegans* and 124 in *C. briggsae*, which were then tested for absence in *B. malayi* and phylogenetic support as described below.

Comparison to *B. malayi*. We checked whether the 368 putative novel introns are absent in the distantly related nematode *B. malayi*, whose genome is being sequenced by the Institute for Genomic Research (26). Gene predictions are not yet available, so we ran TBLASTN (24) with the *Caenorhabditis* protein as query against the *B. malayi* contigs. If a putative novel intron was at residue *i* in the *Caenorhabditis* protein, then we took the intron to be absent in *B. malayi* if the top TBLASTN hit included a large *B. malayi* exon, and residue *i* was internal to the *B. malayi* exon, at least five residues from either end. We found clear evidence that 112 *C. elegans* and 57 *C. briggsae* novel introns are absent from *B. malayi* and retained these for further analysis.

Phylogenetic Support for Intron Gains. For each gene containing a putative novel intron, we constructed a phylogenetic tree for the corresponding protein and its homologs. The outgroup for the tree was a SwissProt (release 41.15) protein that was clearly more distant from the other proteins than they were from each other. Protein sets for each tree were aligned by T-COFFEE (27). Neighbor-joining trees were drawn by using PROTDIST and NEIGHBOR (28) with the Γ correction, and 1,000 bootstrap replicates were made by SEQBOOT (28). A phylogenetic tree was accepted only if there were at least two internal branches with bootstrap values $\geq 70\%$ between the outgroup and the gene containing a putative novel intron. We found phylogenetic support for 41 *C. briggsae* and 81 *C. elegans* putative novel introns.

Control Sets of Introns. To compare the novel introns to the entire *C. elegans* and *C. briggsae* intron populations, we created control sets of introns for each species. We included introns in our control sets only where ± 10 aa adjacent to the intron's position are well conserved among the six animal species. This criterion was the same as we required for novel introns. The control sets consist of 19,942 *C. elegans* introns (20% of all *C. elegans* introns) and 18,516 *C. briggsae* introns (20%).

Repeat Elements and Similarity Among Introns. To find repeat elements in introns, we used FASTA (29) with an *E* value cutoff of 10^{-10} and searched the repeat libraries for *C. elegans* and *C. briggsae* (22). The program PALINDROME (30) was used to find palindromes with a repeating unit of 50–150 bp. Minisatellites of 7–50 bp were detected by using a sliding-window approach (31). The thermodynamic stability of intron RNAs was predicted by using MFOLD (32). Known repeat elements from the repeat libraries were masked before running PALINDROME and MFOLD. To detect sequence similarity among introns and estimate its significance, we used SSEARCH and PRSS (29) after masking repeats from the repeat libraries.

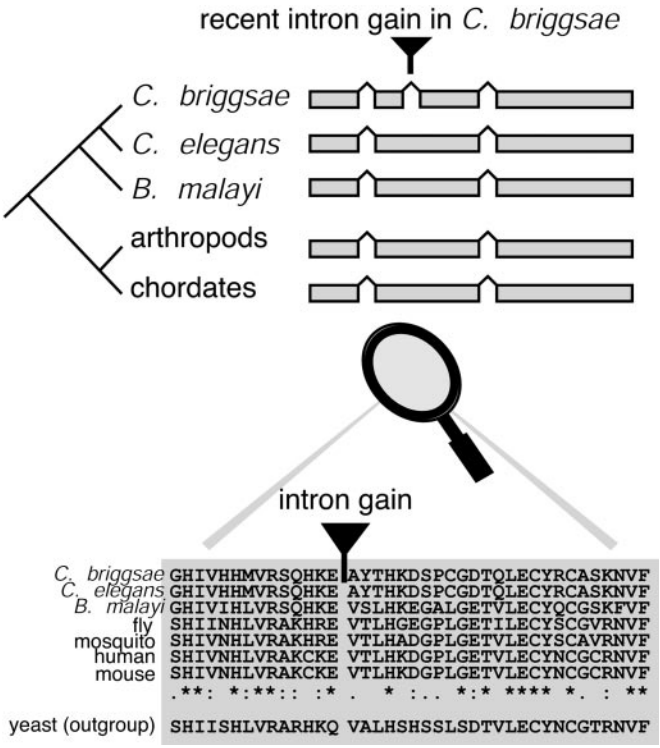
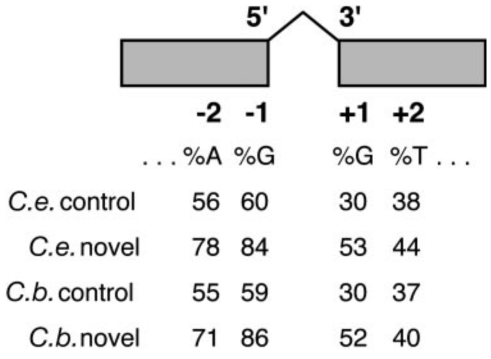

recent intron gain in C. briggsae
C. briggsae
C. elegans
B. malayi
arthropods
chordates
intron gain
C. briggsae GHIVHHMVRSQHKE AYTHKDSPCGDTQLECYRCASKNVF
C. elegans GHIVHHMVRSQHKE AYTHKDSPCGDTQLECYRCASKNVF
B. malayi GHIVIHLVRSQHKE VSLHKEGALGETVLECYQCGSKFVF
fly SHIINHLVRAKHRE VTLHGEGPLGETILECYSCGVRNVF
mosquito SHIVNHLVRAKHRE VTLHADGPLGETVLECYSCAVRNVF
human SHIVNHLVRAKCKE VTLHKDGPLGETVLECYNCGCRNVF
mouse SHIVNHLVRAKCKE VTLHKDGPLGETVLECYNCGCRNVF
.**: *:**:: :* . * :.. *:* **** *. : **
yeast (outgroup) SHIISHLVRARHKQ VALHSHSSLSDTVLECYNCGTRNVF

Fig. 1. Identifying novel introns. To ensure that a putative novel intron was almost certainly caused by insertion rather than by deletion, we drew phylogenetic trees of the gene and its animal and nematode orthologs. We required that there be at least three nodes between the gene and the outgroup. We also required that, in a protein alignment of the gene and its orthologs, ≥5/10-aa residues on either side of the intron be identical or well conserved among the animal genomes.

Results

Identification of Novel Introns and Control Intron Sets. We aimed to find clear examples of introns that are recent gains in one nematode species, rather than to compile an exhaustive list of all possible gains. We considered a *C. elegans* or *C. briggsae* intron to be novel if it is absent from the gene's orthologs in the other *Caenorhabditis* species, the nematode *B. malayi* (26), chordates (human and mouse), and arthropods (fruit fly and mosquito), as well as in any close nematode paralogs. To ensure that a putative novel intron was almost certainly caused by intron insertion rather than by deletion, we drew phylogenetic trees for the gene with its homologs and required that there be at least three nodes between the gene and the outgroup (Fig. 1). Because at least three independent intron losses or one gain could explain the intron distribution, it is more parsimonious to infer intron gain. Furthermore, to ensure that a putative novel intron is very unlikely to be due to intron sliding, the novel intron had to be more than five codons from the nearest intron in any homolog. We also used stringent parameters for both global and local sequence alignment quality (see *Methods*). Using this rigorous approach, we found 41 novel introns in 39 *C. briggsae* genes and 81 novel introns in 74 *C. elegans* genes. There are seven cases where introns have been gained (at different sites) in both a *C. briggsae* gene and its *C. elegans* ortholog, so in total 106 different



	. . . %A	%G	%G	%T . . .
C.e. control	56	60	30	38
C.e. novel	78	84	53	44
C.b. control	55	59	30	37
C.b. novel	71	86	52	40

Fig. 2. Exon splice site consensus of novel introns in *C. elegans* and *C. briggsae*, compared with the consensus for control sets of introns from each genome. Numbers show the percentage of introns in each group that have the indicated base at each position.

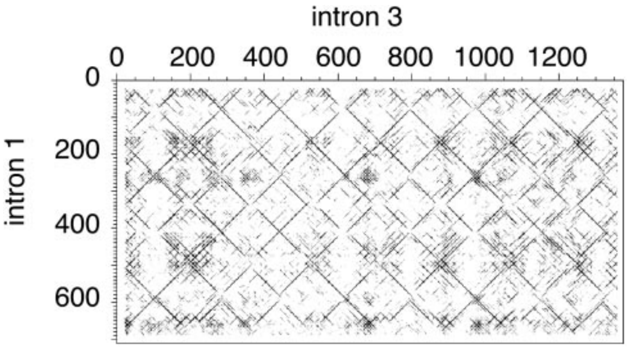


Fig. 3. Dot matrix comparison of introns in *C. briggsae* gene *CBG18597*. Intron 3 is novel and has similarity to intron 1. The plot was made by DOTTER (34), with min = 0 and max = 100 in the GREYRAMP tool.

genes have gained introns. The phylogenetic trees and protein alignments showing the positions of novel introns can be viewed at http://wolfe.gen.tcd.ie/avril/introns.html.

To compare the novel introns to the entire *C. elegans* and *C. briggsae* intron populations, we created control sets of ≈19,000 introns from each species that conformed to the same criteria regarding protein sequence conservation as were applied to the novel introns (see *Methods*). Novel introns are significantly longer (median 60 bp) than the control introns (54 bp; two-sided Wilcoxon test, $P = 0.01$; *C. briggsae* $P = 0.2$; and *C. elegans* $P = 0.2$).

Exon Splice Site Consensus of Novel Introns. Spliceosomal introns tend to be flanked by exon consensus sequences, with AG immediately upstream of the intron and GT immediately downstream of it (18, 33). The novel nematode introns conform more strongly to the exon consensus sequences at all four nucleotide sites in both *C. elegans* and *C. briggsae* than do the control sets of introns (Fig. 2). For the 81 novel *C. elegans* introns, the differences are statistically significant at the A_{-2}, G_{-1}, and G_{+1} positions, all with $P < 10^{-5}$ (one-sided Fisher's test). For the 41 novel *C. briggsae* introns, the differences are again significant at the same sites ($P < 0.04$, $P < 10^{-4}$, and $P < 0.002$, respectively). At the T_{+2} site, the novel introns in both species have a higher frequency of T than the control set (Fig. 2), but the differences are not statistically significant.

Phases of Novel Introns. An intron is described as phase 0 if it lies between two codons in a gene, phase 1 if it is after the first base of a codon, or phase 2 if it is after the second base of a codon. If introns inserted into random positions in genes, novel introns would have an equal probability of being phase 0, 1, or 2. However, of the 41 novel introns in *C. briggsae*, 22 (54%) are phase 0, 12 (29%) are phase 1, and 7 (17%) are phase 2. This is a significant deviation from equal proportions of each phase (χ^2 test; $P = 0.01$). This trend is also seen in novel *C. elegans* introns, which are 52% phase 0, 26% phase 1, and 22% phase 2 ($P = 0.002$). The phase distributions of novel introns are similar to those of the control sets of introns; the frequencies of phases 0, 1, and 2 in the control sets are 51%, 24%, and 25%, respectively, in *C. briggsae* and 53%, 24%, and 22% in *C. elegans.* There is no significant difference between the phase distributions of novel and control introns in either species (χ^2 test; $P \geq 0.4$).

Test for Partial Exon Duplication. If novel introns arise by duplication of an exon region containing AGGT (18, 19), we would expect the region around the 5′ intron–exon boundary to be homologous to that around the 3′ boundary. We found 10 novel introns that have significant similarity in PRSS ($P \leq 0.05$ with -f -7 -g -3 options; ref. 29), taking a region from 25 bp upstream to 25 bp downstream of each boundary. If these 10 introns resulted from exon duplication, we would expect that, if we aligned the 5′ and 3′ boundaries, the nn$\downarrow$GT (5′ end) and AG$\downarrow$nn (3′ end) should line up. However, we did not see this for any of the 10 introns when we aligned the boundary pairs using SSEARCH (29).

Repeat Elements in Novel Introns. We tested the hypothesis that novel introns originate from transposable elements (13, 15) by testing whether novel introns contain more repeat elements than do control introns. We used a recent reannotation of repeat families in both species for this analysis (22). The novel introns in both species include repeat elements from a wide variety of families, but the proportion of novel introns that contain annotated genomic repeats (9%) is not significantly higher than the proportion in the control set (7%; one-sided Fisher's test, $P = 0.1$; *C. briggsae* $P = 0.1$; and *C. elegans* $P = 0.2$).

Sequence Similarity Between Novel and Old Introns. To test the hypothesis that new introns arise by propagation of preexisting introns (20), we compared novel introns to all other introns in the same species. Known repetitive element sequences from the repeat libraries were masked. We found the best alignment on the sense strand between introns using SSEARCH and calculated a *P* value for the alignment by using 500 PRSS iterations (29). This search strategy is computationally intensive but more sensitive than BLAST or FASTA for sequences that are as short as typical nematode introns. We identified 32 novel introns that have significant similarity to other introns in the same genome, using the criteria $P < 0.001$ in PRSS and ≥60% sequence identity over ≥100 bp. We rejected 4 of the 32 because they had large numbers of hits to other introns, presumably due to undescribed repeat elements that were missing from the repeat libraries. We retained the other 28 novel introns (7 *C. briggsae* and 21 *C. elegans*) for further analysis. These introns are listed in the *Supporting Methods* and have 1–11 hits each to other introns.

The similarities among the 28 novel introns and other introns in the same genome seem to be largely due to minisatellite repeats or larger palindromes, or to low-copy-number genomic repeat elements (Fig. 3). That is, in all cases, the PRSS match region (in the query or the hit or both) either contains minisatellite repeats or palindromic repeats or has a weak FASTA hit ($E \leq 0.1$) to a known repeat element in the repeat libraries (too weak to have been detected by the masking algorithm). We used 7–50 bp as a size definition for a minisatellite and 50–150 bp per repeating unit for palindromes (see *Methods*). This distinction is somewhat arbitrary, because large minisatellite arrays often form palindromes.

The DNA in the whole set of novel introns tends to be more internally repetitive than in other introns. In both *C. elegans* and *C. briggsae*, there are palindromes in 27% of novel introns, compared to 12% of 1,000 control introns from the same species (Fisher's test; $P = 0.005$ for *C. elegans* and 0.0005 for *C. briggsae*). Furthermore, the novel *C. briggsae* introns are enriched in minisatellites (for 40% of them, ≥70% of their length is occupied by minisatellite, compared to 16% of control introns; Fisher's test; $P = 0.0004$), although no enrichment is seen in *C. elegans* (11% in novel introns and 13% in controls; $P = 0.8$). Furthermore, more of the 122 novel introns are predicted to fold into stable RNA structures ($\Delta G \leq -100$ kcal/mol) compared to the 2,000 control introns (21% vs. 13%; one-sided Fisher's test, $P = 0.009$; *C. briggsae* $P = 0.04$, *C. elegans* $P = 0.06$). This may be partly because novel introns tend to be longer than control introns, and ΔG decreases with length.

Among the 28 novel introns with similarity to other introns, there are three with similarity to another intron in the same gene. For example, intron 3 of *C. briggsae* gene *CBG18597* is novel and has similarity to intron 1 of the same gene (68% identity over 735 bp). Both introns contain multiple copies of a ≈170-bp palindromic repeat (Fig. 3). Intron 3 also has similarity to intron 5 (65% identity over 1,493 bp). The other same-gene matches are between novel intron 7 and old introns 5 and 6 in *C. briggsae* *CBG21228* (>70% identity over >470-bp alignments) and in *C. elegans Y22D7AL.5* (*hsp-60*), where novel intron 4 matches old intron 5 (63% over 580 bp). There are minisatellites of ≈10 and ≈20 bp in *CBG21228* intron 7 and *Y22D7AL.5* intron 4, respectively. Additional dot-matrix plots showing similarity between novel introns and old introns, and minisatellites or palindromes within novel introns, are included as *Supporting Methods*.

If a novel intron had an equal probability of hitting any other intron in the genome, the probability of hitting another intron from the same gene would be $\approx 4 \times 10^{-5}$, because there are $\approx 10^5$ introns in the genome and five introns per gene. Hence, among the 148 PRSS matches between the 28 novel introns and other introns, we would expect to see no same-gene hits ($148 \times 4 \times 10^{-5} \approx 0$), but we observe five (two in each of *CBG18597* and *CBG21228* and one in *Y22D7AL.5*). Thus, same-gene hits do seem to occur more frequently than we would expect by chance alone. Furthermore, the same-gene matches are the strongest matches had by any novel intron in either species; they are the only ones with ≥63% identity over ≥450 bp. However, the counts are too small to allow statistical testing of whether there are more same-gene than other-gene hits.

Germ-Line Expression of Genes That Have Gained Introns. To become fixed, an intron gain must occur in a germ-line cell or a cell that is going to become one (2). We investigated whether intron gain also requires gene transcription in the germ line. Hill *et al.* (35) used oligonucleotide arrays to identify 5,951 *C. elegans* genes that are always or sometimes expressed in oocytes. Of the 74 genes that have gained introns in *C. elegans*, 57 were studied by Hill *et al.* (35), whereas their data set covers 4,752 of the genes containing control introns. The proportion of the 57 genes that gained introns that are always or sometimes oocyte-expressed (63%) is significantly greater than the proportion of the 4,752 control genes that are always or sometimes oocyte-expressed (42%; one-sided Fisher's test; $P = 0.001$). Thus, genes that are expressed in the germ line are more susceptible to gaining introns than genes not expressed in the germ line.

For novel introns to originate by a reverse-splicing mechanism (20), both the gene containing the novel intron and the gene from which the intron was derived should be expressed in the germ line. The 21 novel *C. elegans* introns with similarity to other introns have been inserted into 19 "recipient" genes, 11 of which were studied by Hill *et al.* (35), who found 9 of 11 (82%) to be germ line expressed. In contrast, of the 87 candidate "source" genes containing introns with PRSS matches to the 21 novel introns, 49 were studied by Hill *et al.* (35), of which 39% are germ line expressed. This is not significantly different from the proportion of control genes that are germ line expressed (42%, two-sided Fisher's test, $P = 0.8$). However, it is obvious that, at most, only 21 of the 87 candidates could actually have been sources of novel introns.

Functions of Genes Containing Novel Introns. The 122 novel introns are inserted into 106 different genes, counting pairs of orthologs as a single gene. Thirteen genes gained two or three introns (Table 1), which is surprising given the low total number of gains, but it should be noted that our ability to detect novel introns depends on gene-specific (rate of sequence evolution, existence of orthologs in other species, and bootstrap support for a phylogenetic tree) as well as intron-specific factors.

It is striking that several genes with novel introns, including one that gained three introns, code for proteins involved in premRNA splicing or surveillance (Table 1). These genes include *smg-2*, which functions in nonsense-mediated decay (36), and *F49D11.1*, which is predicted to catalyze the second step of splicing (homolog of *Saccharomyces cerevisiae* Cdc40; ref. 37). Novel introns were also found in nematode homologs of three well-characterized *S. cerevisiae* spliceosomal proteins (Hsh155/SF3B1, Prp6, and Prp19), two others (Imd2 and Ssa1) that are associated with the spliceosomal penta-snRNP in yeast (38), homologs of *S. cerevisiae* Dis3 (a component of the exosome, which processes the 3′ end of U4 small nuclear RNA (snRNA); refs. 39 and 40), and a homolog of human gene *CPSF5*, coding for a subunit of premRNA cleavage factor I_m (41). Of the 122 novel introns, 13 are in genes with known splicing-related functions, and four more are in putative RNA helicase genes with DEAD-box motifs (Table 1).

As an approximate test of the significance of this observation, we tested whether genes with mRNA processing functions are overrepresented in the novel intron group, compared to their frequency in a control group of germ-line-expressed genes containing control introns. We used Gene Ontology annotations for all of SwissProt instead of *C. elegans* alone, because documentation of some premRNA processing and spliceosome components is more complete in yeasts and vertebrates. We identified nematode genes with BLASTP hits ($E < 10^{-50}$) to proteins in the Gene Ontology category "mRNA processing." This method inferred mRNA processing roles for 5 of the 106 nematode genes with novel introns (4.7%), compared to only 17 of the 1,990 *C. elegans* control genes that are expressed in the germ line (0.9%; one-sided Fisher's test; $P = 0.004$).

It is also notable that genes that have gained introns tend to be part of operons. For *C. elegans*, whose operons have been mapped (42), 26% of genes with novel introns but only 14% of genes in the control set are in operons (Fisher's test; $P = 0.005$). However, this seems to just reflect the tendency of operons to be expressed in the germ line; taking just those control genes expressed in the germ line, 30% are in operons.

Discussion

Of the possible mechanisms of intron gain listed in the Introduction, group II intron insertion is improbable in nematodes, because their mitochondrial genomes do not contain group II introns. Also, intron gain by gene conversion with a homologous intron-containing gene can result only in the novel intron being gained at the same position as the source intron (11). We included only novel introns for which there was no intron at the same position in any close homolog, so the novel introns in our data set could not have arisen by this mechanism. Thus, in the following discussion, we consider whether the remaining three

Table 1. Some of the nematode genes with novel introns

Gene name		Intron gains			
C. elegans	*C. briggsae*	*C. elegans*	*C. briggsae*	Homologs*	Predicted function*
Genes with mRNA-related functions					
T08A11.2	*CBG21228*	0	3	*Sc HSH155*; *Sp prp10*; *Hs SF3B1*	U2 snRNP protein (*Sc*) (ref. 38)
smg-2	*CBG08494*	1	1	*Sc NAM7*; *Hs RENT1*	Nonsense mediated decay *(Ce)* (ref. 36)
dis-3	*CBG13499*	2	0	*Sc DIS3*; *Sp dis3*	Exosome subunit (Sc) (ref. 40)
Y59A8B.6	*CBG05596*	1	0	*Sc PRP6*; *Sp prp1*	U4/U6·U5 snRNP protein (*Sc*) (ref. 38)
T10F2.4	*CBG21324*	1	0	*Sc PRP19*; *Sp cwf8*	Penta-snRNP specific protein *(Sc)* (ref. 38)
F49D11.1	*CBG20408*	0	1	*Sc CDC40*; *Sp prp17*; *Hs hPRP17*	Second step of splicing *(Sc, Hs)* (ref. 37)
F43G9.5	*CBG12500*	1	0	*Hs CPSF5*	Pre-mRNA cleavage factor I_m subunit *(Hs)* (ref. 41)
F32D1.5	*CBG23625*	1	0	*Sc IMD2*†	GMP reductase; *Sc* Imd2 is penta-snRNP associated (ref. 38)
stc-1	*CBG00564*	1	0	*Sc SSA1* family†	Heat shock protein 70, penta-snRNP associated *(Sc)* (ref. 38)
H27M09.1	*CBG12746*	2	0	*Sc DBP2* family† *Hs DDX41*; *Dm Abstrakt*	DEAD-box RNA helicase
T26G10.1	*CBG10097*	1	0	*Sc RRP3*; *Hs DDX47*	DEAD-box RNA helicase
Y65B4A.6	*CBG05145*	1	0	*Sc FAL1*; *Hs DDX48*	DEAD-box RNA helicase
Genes with other functions					
tre-3	*CBG23281*	2	1	*Sc NTH1*	Trehalase
F57C2.5	*CBG20870*	2	0	No clear orthologs	Strictosidine synthase family
K08E3.1	*CBG18278*	2	0	No clear orthologs	Tyrosinase family
hsp-60	*CBG11701*	2	0	*Sc HSP60*; *Sp hsp60*	Mitochondrial chaperonin 60
Y25C1A.5	*CBG19635*	2	0	*Sc SEC26*; *Sp sec26*; *Hs COPB*	Coatomer complex β chain (β-COP)
F02E11.1	*CBG04332*	1	1	*Sc YOL075C*	ABC transporter
Y50D4A.4	*CBG01087*	1	1	*Hs MIC1*	Unknown
klp-20	*CBG15720*	1	1	*Hs KIF3A*	Kinesin family
Y74C10AM.1	*CBG03989*	1	1	*Sc ATM1*; *Hs ABCB7*	ABC transporter

Genes with functions related to mRNA processing and all genes that have gained more than one intron are listed.

*Abbreviations of species names: *Ce*, *C. elegans*; *Dm*, *D. melanogaster*; *Hs*, *Homo sapiens*; *Sc*, *S. cerevisiae*; *Sp*, *Schizosaccharomyces pombe*.

†Not reciprocal best BLASTP hit of the *C. elegans* gene.

mechanisms could explain our data: transposon insertion, partial exon duplication, and reverse splicing of a preexisting intron.

We found that 63% of *C. elegans* genes that gained introns are expressed in the germ line, compared to 42% of control genes. If introns are gained by reverse splicing, one would expect intron gains to occur mainly in germ-line-expressed genes (2). Alternatively, if novel introns arise by transposon insertion, the transposons may have an insertion preference for actively transcribed regions of the genome, as has been observed for the *Drosophila P* element (43). But if intron gains occur by partial exon duplication, we see no reason why there would be a bias for germ-line-expressed genes. We also did not find any cases of obvious partial exon duplication in our data. Thus, we consider that partial exon duplication can be discarded as a possible mechanism in *Caenorhabditis*.

Our novel introns tend to be inserted at AG ↓ G, where ↓ is the insertion site. This is similar to the "proto-splice site" (MAG ↓ R) proposed by Dibb and Newman (33) and agrees with findings that the AG ↓ G consensus is stronger in species-specific introns than in all introns in *Caenorhabditis* (21), that recently gained introns in 10 eukaryotic protein families seem to have inserted into AG ↓ G sites (44), and that introns specific to one animal phylum have a stronger exon consensus than those common to two or more phyla (45). If introns are gained by reverse-splicing, the spliceosome may insert the novel intron into AG ↓ G, because this would be the reverse of its normal role of removing an intron from AG ↓ G. Alternatively, if novel introns arise by transposon insertion with a target site duplication containing AGG, the resultant intron would be found at AG ↓ G (16). We also found, similar to Rogozin *et al.* (5) and Qiu *et al.* (44), that novel introns tend to insert into phase 0 positions in codons. If novel introns insert into AG ↓ G, 51% of insertions will be in phase 0 because of the genetic code (ref. 46; see ref. 3 for discussion). This is close to the fraction of novel introns in phase 0 that we observed: 54% in *C. briggsae* and 52% in *C. elegans*. Thus, the excess of phase 0 introns among the novel introns is likely to be a result of their tendency to insert at AG ↓ G sites. However, an alternative is that the phase bias results from selection subsequent to intron insertion. Lynch (47) pointed out that if intron sliding occurred subsequent to insertion, it would have greater negative consequences for phase 1 and 2 than phase 0 introns.

We found that novel introns are as likely as control introns to contain genomic repetitive elements from the repeat libraries. This result suggests that novel introns probably did not originate by insertion of transposable elements. However, we also found that novel introns are more likely than control introns to contain palindromes. We cannot tell whether the repeats that form palindromes are uncharacterized locally distributed transposable elements that have produced new introns, whether these repeats are somehow formed when the new intron is formed, or whether introns that are repetitive are more likely to give rise to new introns by reverse splicing. It seems unlikely that the palindrome repeats are locally distributed transposable elements, because the proportion of *C. elegans* novel introns with PRSS matches to introns from genes that are within ±500 genes on the same chromosome (8%) is not any greater than expected by chance (5%; one-sided Fisher's test, $P = 0.1$). However, palindromes in RNA molecules often fold into hairpins. The fact that novel introns are predicted to fold into more stable RNA structures than do most introns would fit the expectation that introns with longer half-lives are more likely to be duplicated by reverse splicing (4).

Our finding that several novel introns are inserted into genes coding for proteins with functions related to splicing provides circumstantial support for a reverse-splicing model. When spliceosomal introns were discovered in genes for the U1, U2, U5, and U6 snRNA components of the spliceosome in fungi, it was suggested that they had originated from mishaps during splicing (48, 49). An excised intron from some other transcript became integrated into the snRNA, which was then reverse transcribed into cDNA and recombined with the chromosomal snRNA gene. Brow and Guthrie (48) suggested that this reverse splicing was facilitated by the closeness of the snRNAs to the catalytic center of the spliceosome. A similar argument can be made for the novel introns we found in genes for spliceosomal proteins such as SF3B1 (Table 1), but the argument is complicated by the fact that these genes are protein coding. Conceivably, mRNAs for proteins with splicing-related functions might somehow be more available for reverse-splicing reactions than other mRNAs, perhaps due to autoregulation (50) or occasional aberrant events such as the attempted incorporation into the spliceosome of nascent proteins that are still associated with their mRNAs. Spliceosomal proteins are part of the core cellular machinery that is conserved across eukaryotes, and "core" genes tend to be both germ line expressed and located within operons (35, 51). However, our Gene Ontology analysis indicated that novel introns are unusually frequent in genes with mRNA processing functions, relative to germ-line-expressed genes, which suggests that it is the function of these genes, rather than their mode of transcription, that makes them amenable to gaining introns.

Logsdon *et al.* (2) commented that for an intron gain to be credible, it should have strong phylogenetic support, and the source of the intron DNA should be identifiable. They referred to this second criterion as a "molecular smoking gun." We identified three novel nematode introns with significant sequence similarity to another intron in the same gene, a result that is suggestive of a reverse-splicing model where an excised intron sometimes reintegrates back into a different site in the same mRNA (10). However, interpretation of the sequence similarities is complicated by the repetitive structures of the introns (Fig. 3). Our results indicate a reverse-splicing origin for some novel nematode introns but do not exclude the possibility that other mechanisms were involved in other intron gains. The best way to confirm our proposal that reverse splicing is one of the principal mechanisms of intron gain in nematodes is to identify intron gains that are even more recent than those examined here, because their source of intron DNA would be more obvious, for example by identifying introns that have been gained after the divergence of *C. briggsae* from its closer relative *Caenorhabditis remanei* (22).

Note Added in Proof. Using BLAST searches of unassembled reads from the *C. remanei* genome (http://genome.wustl.edu/blast/client.pl), we found that 15 of the 41 *C. briggsae* novel introns are absent from its sister species *C. remanei* and so must have been gained since speciation. Another 19 introns are shared by the two species, although we could not unambiguously score the remaining 7. The fraction of the 15 younger introns that have PRSS matches to other introns in the same genome (5/15; 33%) is significantly greater than the fraction of the 19 older introns with same-genome matches (0%; one-sided Fisher's test, P = 0.01). This strongly suggests that the same-genome PRSS matches are vestiges of intron birth.

We thank Richard Durbin and Lincoln Stein for access to *C. briggsae* data, Andrew Hill (Department of Genomics, Wyeth Research, Cambridge, MA) for kindly providing *C. elegans* germ-line expression data, and Des Higgins for magnanimously allowing A.C. to complete this work in his laboratory. This work was supported by Science Foundation Ireland. We thank The Institute for Genome Research (TIGR) for allowing prepublication use of data from the International *Brugia* Genome Sequencing Project, which is supported by an award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

1. Gilbert, W. (1978) *Nature* **271,** 501.
2. Logsdon, J. M., Jr., Stoltzfus, A. & Doolittle, W. F. (1998) *Curr. Biol.* **8,** R560–R563.
3. Logsdon, J. M., Jr. (1998) *Curr. Opin. Genet. Dev.* **8,** 637–648.
4. Lynch, M. & Richardson, A. O. (2002) *Curr. Opin. Genet. Dev.* **12,** 701–710.
5. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. (2003) *Curr. Biol.* **13,** 1512–1517.
6. Roy, S. W., Fedorov, A. & Gilbert, W. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 7158–7162.
7. Charlesworth, D., Liu, F. L. & Zhang, L. (1998) *Mol. Biol. Evol.* **15,** 552–559.
8. Fedorov, A., Roy, S., Fedorova, L. & Gilbert, W. (2003) *Genome Res.* **13,** 2236–2241.
9. Logsdon, J. M., Jr., Tyshenko, M. G., Dixon, C., D-Jafari, J., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8507–8511.
10. Tarrío, R., Rodriguez-Trelles, F. & Ayala, F. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1658–1662.
11. Hankeln, T., Friedl, H., Ebersberger, I., Martin, J. & Schmidt, E. R. (1997) *Gene* **205,** 151–160.
12. Frugoli, J. A., McPeek, M. A., Thomas, T. L. & McClung, C. R. (1998) *Genetics* **149,** 355–365.
13. Iwamoto, M., Nagashima, H., Nagamine, T., Higo, H. & Higo, K. (1999) *Theor. Appl. Genet.* **98,** 853–861.
14. Robertson, H. M. (2001) *Chem. Senses* **26,** 151–159.
15. Crick, F. (1979) *Science* **204,** 264–271.
16. Giroux, M. J., Clancy, M., Baier, J., Ingham, L., McCarty, D. & Hannah, L. C. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 12150–12154.
17. Purugganan, M. D. (2002) in *Horizontal Gene Transfer*, eds. Syvanen, M. & Kado, C. I. (Chapman & Hall, London), pp. 187–195.
18. Rogers, J. H. (1989) *Trends Genet.* **5,** 213–216.
19. Venkatesh, B., Ning, Y. & Brenner, S. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 10267–10271.
20. Sharp, P. A. (1985) *Cell* **42,** 397–400.
21. Kent, W. J. & Zahler, A. M. (2000) *Genome Res.* **10,** 1115–1125.
22. Stein, L., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M., Chen, J., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., *et al.* (2003) *PLoS Biol.* **1,** 166–192.
23. Guiliano, D. B., Hall, N., Jones, S. J., Clark, L. N., Corton, C. H., Barrell, B. G. & Blaxter, M. L. (2002) *Genome Biol.* **3,** RESEARCH0057.
24. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
25. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
26. Ghedin, E., Wang, S., Foster, J. M. & Slatko, B. E. (2004) *Trends Parasitol.* **20,** 151–153.
27. Notredame, C., Higgins, D. G. & Heringa, J. (2000) *J. Mol. Biol.* **302,** 205–217.
28. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
29. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
30. Rice, P., Longden, I. & Bleasby, A. (2000) *Trends Genet.* **16,** 276–277.
31. Katti, M. V., Ranjekar, P. K. & Gupta, V. S. (2001) *Mol. Biol. Evol.* **18,** 1161–1167.
32. Zuker, M. (2003) *Nucleic Acids Res.* **31,** 3406–3415.
33. Dibb, N. J. & Newman, A. J. (1989) *EMBO J.* **8,** 2015–2021.
34. Sonnhammer, E. L. & Durbin, R. (1995) *Gene* **167,** GC1–G10.
35. Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. & Brown, E. L. (2000) *Science* **290,** 809–812.
36. Page, M. F., Carr, B., Anders, K. R., Grimson, A. & Anderson, P. (1999) *Mol. Cell. Biol.* **19,** 5943–5951.
37. Ben Yehuda, S., Dix, I., Russell, C. S., Levy, S., Beggs, J. D. & Kupiec, M. (1998) *RNA* **4,** 1304–1312.
38. Stevens, S. W., Ryan, D. E., Ge, H. Y., Moore, R. E., Young, M. K., Lee, T. D. & Abelson, J. (2002) *Mol. Cell* **9,** 31–44.
39. van Hoof, A., Lennertz, P. & Parker, R. (2000) *Mol. Cell. Biol.* **20,** 441–452.
40. Allmang, C., Petfalski, E., Podtelejnikov, A., Mann, M., Tollervey, D. & Mitchell, P. (1999) *Genes Dev.* **13,** 2148–2158.
41. Ruegsegger, U., Blank, D. & Keller, W. (1998) *Mol. Cell* **1,** 243–253.
42. Blumenthal, T., Evans, D., Link, C. D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W. L., Duke, K., Kiraly, M., *et al.* (2002) *Nature* **417,** 851–854.
43. Timakov, B., Liu, X., Turgut, I. & Zhang, P. (2002) *Genetics* **160,** 1011–1022.
44. Qiu, W. G., Schisler, N. & Stoltzfus, A. (2004) *Mol. Biol. Evol.* **21,** 1252–1263.
45. Sverdlov, A. V., Rogozin, I. B., Babenko, V. N. & Koonin, E. V. (2003) *Curr. Biol.* **13,** 2170–2174.
46. Long, M., de Souza, S. J., Rosenberg, C. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 219–223.
47. Lynch, M. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6118–6123.
48. Brow, D. A. & Guthrie, C. (1989) *Nature* **337,** 14–15.
49. Takahashi, Y., Tani, T. & Ohshima, Y. (1996) *J. Biochem. (Tokyo)* **120,** 677–683.
50. Lewis, B. P., Green, R. E. & Brenner, S. E. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 189–192.
51. Blumenthal, T. & Gleason, K. S. (2003) *Nat. Rev. Genet.* **4,** 112–120.