Genome Biology and Evolution

SMBE

GBE

Family Size and Turnover Rates among Several Classes of Small Non–Protein-Coding RNA Genes in Caenorhabditis Nematodes

Paul Po-Shen Wang and Ilya Ruvinsky*

Department of Ecology and Evolution, Institute for Genomics and Systems Biology, The University of Chicago

*Corresponding author: E-mail: ruvinsky@uchicago.edu.

Accepted: 23 March 2012

Abstract

It is important to understand the forces that shape the size and evolutionary histories of gene families. Here, we investigated the evolution of non–protein-coding RNA genes in the genomes of Caenorhabditis nematodes. We specifically focused on nested arrangements, that is, cases in which an RNA gene is entirely contained in an intron of another gene. Comparing these arrangements between species simplifies the inference of orthology and, therefore, of evolutionary fates of nested genes. Two distinct patterns are evident in the data. Genes encoding small nuclear RNAs (snRNAs) and transfer RNAs form large families, which have persisted since before the common ancestor of Metazoa. Yet, individual genes die relatively rapidly, with few orthologs having survived since the divergence of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. In contrast, genes encoding small nucleolar RNAs (snoRNAs) are either single-copy or form small families. Individual snoRNAs turn over at a relatively slow rate—most *C. elegans* genes have clearly identifiable orthologs in *C. briggsae*. We also found that in Drosophila, genes from larger snRNA families die at a faster rate than their counterparts from single-gene families. These results suggest that a relationship between family size and the rate of gene turnover may be a general feature of genome evolution.

Key words: birth-and-death, gene family, evolution, small RNA, nested genes, *C. elegans*.

Introduction

Gene families originate as single genes. Many remain as single-gene families, but some expand into families that contain from two to hundreds of members (Annilo et al. 2006; Prachumwat and Li 2008). Family expansions can be caused by a variety of mechanisms, including whole-genome duplications (Ohno 1970), unequal crossing-overs (Takahashi et al. 1982), and retrotransposition (Betrán et al. 2002; Marques et al. 2005). Acting in the opposite direction, a number of mechanisms restrain the growth of gene families. For example, redundant recently duplicated copies can acquire deleterious mutations (Li 1997; Kondrashov and Kondrashov 2006) or selection may prevent fixation of a duplicated gene in order to maintain appropriate dosage (Papp et al. 2003).

Different families evolve in different regimes under the pressure of a multitude of evolutionary forces. Consequently, they differ in size as well as age—some are ancient, dating to the origin of cellular life (Leipe et al. 2002), whereas others are quite recent (Hahn et al. 2007). A classical view of evolution within families is one that combines gene duplications with relatively low rates of gene death leading to a slow divergence in gene complements (Li 1997). It was noticed, however, that the evolutionary histories of several gene families were inconsistent with this model (Nei et al. 1997). Instead a "birth-and-death" process was proposed to account for the observations (Nei and Rooney 2005). This view combines extensive gene duplications with rampant loss of different genes in different lineages. Over time, birth-and-death processes can lead to complex organization of gene families—some genes persist for long periods of time, whereas others are young. Therefore, gene complements can be quite different even between closely related species.

Whereas the birth-and-death process is now widely appreciated as being an important mode of gene family evolution, the forces that influence the rates of gene birth and death are less well understood. Here, we studied the evolutionary histories of three major classes of small non–protein-coding RNAs (ncRNAs)—small nuclear RNAs

Table 1
Nested ncRNAs in *Caenorhabditis elegans*

	Number of genes in *C. elegans*	Number of nested genes in *C. elegans*	Number of nested gene arrangements[a]	Number of host genes
snRNA	120	47 (39%)	41	35
snoRNA	142	111 (78%)	111	88
tRNA	608	264 (43%)	229	204

[a] A nested gene arrangement may include more than one paralogous ncRNA nested inside the same intron.

(snRNAs), small nucleolar RNAs (snoRNAs), and transfer RNAs (tRNAs)—in Caenorhabditis nematodes. The compact sequenced genomes of these animals (*Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Caenorhabditis brenneri*, and *Caenorhabditis remanei*) permit relatively straightforward gene identification and allow accurate determination of gene loss and gain events. In particular, we investigated the role of family size in shaping evolution of these genes.

Materials and Methods

Genome Sequences and Gene Annotations

Whole-genome sequences for the nematode species were downloaded from Wormbase (www.wormbase.org). Version WS190 was used for *C. elegans* and *C. briggsae* and WS204 was used for *C. remanei*, *C. brenneri*, and *C. japonica*. These databases provided protein-coding gene annotations for all species, whereas RNA annotations were available only for *C. elegans*. Wormbase annotations of microRNAs (miRNAs) were supplemented by miRBase data (www.mirbase.org), whereas snoRNAs were compiled from a previous study (Wang and Ruvinsky 2010). In general, non-*C. elegans* annotations were not as comprehensive, therefore, we carried out additional gene discovery (see below). All Drosophila sequences and annotations were downloaded from Flybase (www.flybase.org), r5.34 for *Drosophila melanogaster*, r1.2 for *Drosophila virilis*, and r2.17 for *Drosophila pseudoobscura*.

Identification of Unannotated RNA Sequences

We carried out discovery of additional RNA sequences (in *C. elegans* and other genomes) using WU-Blast (blast.wustl.edu). A BlastN search (with word-size option $W = 6$) was performed using annotated *C. elegans* sequences as queries. Matches with higher than 60% identity were considered as possible homologs. Multiple sequence alignment via ClustalW (www.clustal.org; Thompson et al. 1994) was used to filter out spurious matches. Specifically, related sequences were expected to have regions of high conservation, even if global conservation was low. Therefore, sequences were not considered further if they aligned poorly in regions where high conservation was expected, even if their overall alignment passed the threshold (60%). The final set of RNA sequences used in this study is shown in supplementary table S1 (Supplementary Material online). To establish paralogy relationships between ncRNAs, we carried out sequence similarity comparisons using the same parameters as described above.

Identification of Orthologous Host Genes and Nested Arrangements

Wormbase annotations for protein-coding genes in non-*C. elegans* species are incomplete. To identify orthologs of *C. elegans* host genes, their sequences were used to search non-*C. elegans* genomes using WU-Blast (TBlastN). Loci with the highest number of matching exons and residues were designated as putative orthologs. We next examined the intron orthologous to the host intron of the *C. elegans* gene to establish whether it contained an RNA homologous to the nested *C. elegans* gene.

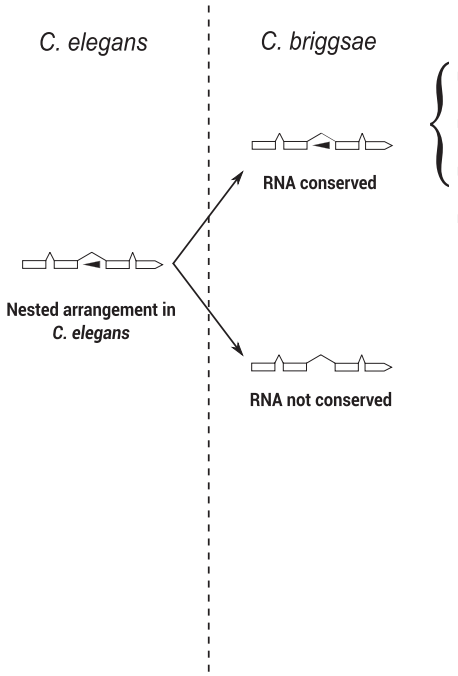
Results

Large Fraction of ncRNA Genes Are Nested in the *C. elegans* Genome

The availability of completely sequenced genomes offers an opportunity to investigate the evolution of entire gene families. Whereas the evolution of protein-coding gene families has been extensively studied over the years, relatively less is known about ncRNAs. Yet, just three major classes of these genes—snRNA (Newman 1993; Staley and Guthrie 1998), snoRNA (Bachellerie et al. 2002), and tRNA (Schmitt et al. 1998; Sprinzl and Vassilenko 2005)—constitute nearly 5% of genes in the *C. elegans* genome (Table 1). Furthermore, their unique structures, functions, and modes of regulation (Sharp et al. 1981; Mattaj et al. 1993; Reynolds 1995; Matera et al. 2007) make them interesting subjects of an evolutionary study.

We first catalogued the entire complements of snRNA, snoRNA, and tRNAs in the genome of *C. elegans*. Notably, substantial fractions of all three classes of these genes are nested, that is, completely contained within introns of protein-coding genes (Table 1; supplementary table S1, Supplementary Material online). Whereas only a small fraction (<0.5%) of the approximately 100,000 introns in the genome are occupied by nested genes, in most instances, there is only one nested gene per host intron.

We restricted the subsequent evolutionary analysis to nested ncRNAs. In addition to comprising nearly half of



Host gene structure	snoRNA	tRNA	snRNA
All exons intact	70	66	5
Immediate flanking exons intact	28	46	2
Missing a flanking exon, but exons from both sides are found	2	4	0
All exons on one side are missing	0	5	0
Cel arrangements conserved in *Cbr*	100	121	7
All exons intact	9	56 (-1)	21
Immediate flanking exons intact	1	34	12
Missing a flanking exon, but exons from both sides are found	1	7 (-1)	0
All exons on one side are missing	0	7	1
No homolog is found for the host gene	0	4 (-4)	0
Cel arrangements not conserved in *Cbr*	11	108 (-6)	34
Number of nested arrangements in *Cel*	**111**	**229 (-6)**	**41**

Fig. 1.—Conservation of *Caenorhabditis elegans* (Cel) host gene structures in *Caenorhabditis briggsae* (Cbr). In the great majority of cases, the exons immediately flanking the host intron were also found in *C. briggsae*, regardless of whether the RNA is conserved, suggesting that the gain or loss of nested RNAs had little impact on the host genes. We regarded as conserved only those arrangements (indicated by the bracket) which had conserved exons surrounding host introns on both sides. Orthologs of six host genes were not found either in *C. briggsae*, *Caenorhabditis remanei*, or *Caenorhabditis brenneri*. These cases are shown in parenthesis and were excluded from the subsequent analyses. Therefore, numbers in figure 2 are the same as in this figure, except for tRNAs for which there were 223 (=229 − 6) conserved host genes.

all genes from their respective classes, these offer an additional advantage. Ascertaining orthology and paralogy relationships can be difficult, particularly for short genes. One solution to this problem is to examine the evolutionary history of closely linked loci, which can be instructive for the understanding the evolution of the gene(s) in question (Bailey et al. 1997; Ruvinsky and Silver 1997).

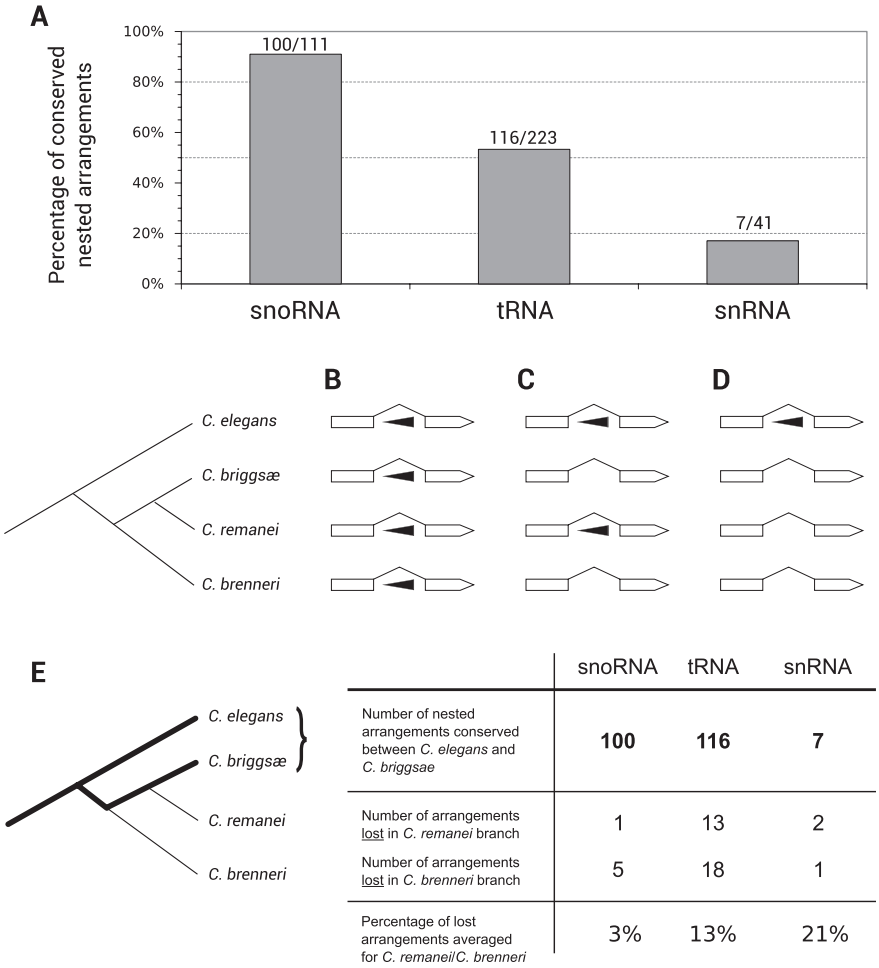
Nested genes are completely contained within introns of their host protein-coding genes (Chen and Stein 2006; Assis et al. 2008). Having to identify single genes, not extended regions with multiple linked genes as would be required for analysis of synteny, overcomes a practical difficulty that not all sequenced genomes have been completely assembled yet. Furthermore, gene order evolves relatively rapidly, and decay of synteny is evident even between closely related species (Hillier et al. 2007; Ranz et al. 2007; Vergara and Chen 2010; von Grotthuss et al. 2010). Therefore, focusing our analysis on nested genes simplified the ascertainment of orthology and permitted confident inferences of trends governing the evolution of these genes.

Conservation of Nested ncRNAs between *C. elegans* and *C. briggsae*

We sought to identify *C. briggsae* orthologs of all nested arrangements involving annotated *C. elegans* snRNAs, snoRNAs, and tRNAs. All *C. elegans* host genes containing snRNAs and snoRNAs and 98% (200/204) of those containing tRNAs have at least one putative *C. briggsae* ortholog (fig. 1; supplementary table S2A, Supplementary Material online). For all host genes with putative orthologs, we were able to establish whether a given nested RNA was conserved, that is, present in the homologous intron. In nearly all cases, there was a single plausible *C. briggsae* ortholog for a host gene containing a nested RNA in the *C. elegans* genome (the few exceptions are shown in supplementary table S2B, Supplementary Material online).

We found that 90% of the *C. elegans* snoRNAs have an ortholog in *C. briggsae* (fig. 2*A*). However, this ratio is considerably lower for tRNAs (52%) and lower still for snRNAs (17%). Importantly, in over 85% of instances (first two lines in "RNA not conserved" portion of fig. 1) when a *C. elegans* nested RNA did not have an ortholog in *C. briggsae*, the host gene had a clear ortholog, implying that losses and gains of nested genes most often proceed without substantial alterations of the host genes.

We note that there is an inherent discovery bias in our search procedure. Nonconserved cases above reflect *C. elegans*-specific gain events as well as loss events along the *C. briggsae* lineage. Because de novo predictions in newly sequenced genomes can be difficult (e.g., Wang



	snoRNA	tRNA	snRNA
Number of nested arrangements conserved between *C. elegans* and *C. briggsae*	**100**	**116**	**7**
Number of arrangements lost in *C. remanei* branch	1	13	2
Number of arrangements lost in *C. brenneri* branch	5	18	1
Percentage of lost arrangements averaged for *C. remanei*/*C. brenneri*	3%	13%	21%

Fig. 2.—Conservation of *Caenorhabditis elegans* nested ncRNA arrangements. (*A*) The fraction of *C. elegans* nested ncRNA arrangements conserved in *Caenorhabditis briggsae*, case counts are given above bars. (*B–D*) Possible scenarios of arrangement conservation: the arrangement is conserved in all species (*B*), the arrangement is conserved in only one species other than *C. elegans* (*C*), the arrangement is found only in *C. elegans* (*D*). (*E*) Lineage-specific losses of nested arrangements. By considering only arrangements found in both *C. elegans* and *C. briggsae* (indicated by thick lines), the lineage-specific losses can be inferred for the *Caenorhabditis remanei* and *Caenorhabditis brenneri* branches.

and Ruvinsky 2010), of necessity we started the search using currently annotated ncRNAs of *C. elegans*. This limited our ability to identify newly arisen genes in the remaining three genomes. It did not, however, compromise our ability to infer lineage-specific gene losses. Altogether, the data so far imply that snoRNA arrangements are somewhat static within the Caenorhabditis nematodes, whereas tRNAs and snRNAs are lost more rapidly.

Different Classes of ncRNAs Have Different Rates of Gene Loss in Caenorhabditis Nematodes

To obtain a more precise estimate of the number of nested RNA genes lost during Caenorhabditis evolution and to avoid counting nonindependent events, we devised the following strategy. We identified orthologs of nested *C. elegans* RNAs in the genomes of *C. brenneri* and *C. remanei* using the same approach as that used for *C. briggsae*. This permitted us to categorize all nested *C. elegans* RNAs as conserved in four species (fig. 2*B*), lost in one or more lineages (fig. 2*C*), or uniquely gained in the *C. elegans* lineage (fig. 2*D*). We consider it more parsimonious to interpret cases like that shown in figure 2*C* as two independent loss events than two independent gain events. This is because the former scenario would require that a particular 1 of the ~100,000 introns in the genome would acquire homologous nested ncRNAs in two different genomes.

The genes conserved between *C. elegans* and *C. briggsae* must have existed in their common ancestor, and the phylogenetic relationship of the nematode species (fig. 2*E*; Kiontke et al. 2004) dictates that they must have also existed in the genomes of the last common ancestors of *C. briggsae*/*C. remanei* and *C. briggsae*/*C. brenneri*. Thus, the losses of any of the genes conserved between *C. elegans* and *C. briggsae* in *C. brenneri* and/or *C. remanei* can be considered as independent events along their respective lineages.

Two strikingly different patterns are evident in the data (fig. 2*E*; supplementary table S1, Supplementary Material online). Only one and five snoRNA genes were lost along *C. remanei* and *C. brenneri* lineages, respectively, of the 100 genes that existed in the last common ancestors of each of these two species and *C. briggsae*. The divergence times of *C. brenneri* and *C. remanei* from *C. briggsae* were relatively close (Cutter 2008). Thus, the approximate fraction of genes that were lost, averaged between the two lineages, is ~3% (i.e., (1/100 + 5/100)/2). Gene loss is, therefore, considerably less common compared with snRNAs ((2/7 + 1/7)/2 = 21%) and tRNAs ((18/116 + 13/116)/2 = 13%). These findings are consistent with previously reported results, which suggested that turnover of individual genes may be slower for snoRNA genes (Hoeppner et al. 2009) than for snRNAs (Marz et al. 2008) or tRNAs (Rogers et al. 2010).

We note that whereas the genomes of *C. brenneri* and *C. remanei* have been sequenced, the coverage is not complete due to considerable residual heterozygosity (Barrière et al. 2009). In every case when sequence data were not available confidently to infer conservation or loss of a gene, we assumed a loss. Our estimates were not substantially altered if instead we assumed that missing sequences covered conserved arrangements—the numbers of snRNAs, snoRNAs, and tRNAs lost along the two lineages combined would be 3, 4, and 29, which are 21%, 2%, and 12%, respectively.

A Hypothesis to Explain Different Rates of Gene Death between Different Classes of ncRNAs

The data presented above suggest that within Caenorhabditis genomes, the rates of gene loss (i.e., the fractions of lost genes within a given class) are considerably higher for tRNAs and particularly snRNAs compared with snoRNA genes. Over 90% of the annotated *C. elegans* snoRNAs have orthologs in *C. briggsae*, whereas comparable fractions for tRNAs and snRNAs are only 52% and 17%, respectively (fig. 2*A*). Of the genes that can be confidently inferred to have been present in the common ancestor of the four examined Caenorhabditis species, only ~3% of snoRNAs have been lost in either *C. remanei* or *C. brenneri*, whereas the fraction was substantially higher for tRNAs (13%) and snRNAs (21%; fig. 2*E*).

To understand how the different rates of gene death manifest over longer timescales, we sought to identify orthologs of individual nested *C. elegans* genes in *D. melanogaster*. The divergence between *C. elegans* and *C. briggsae* is estimated to have occurred ~20 million years ago (Cutter 2008). This is approximately 30- to 50-fold more recently than the pre-Cambrian divergence of nematodes and arthropods (Valentine 1994).

We found no evidence of orthology between any of the three classes of nested ncRNAs in *C. elegans* and *D. melanogaster* (supplementary table S3, Supplementary Material online). That is, we did not identify a single conserved RNA gene in an orthologous intron of an orthologous host gene between the two species.

Gene families, however, appear to have arisen before the divergence of arthropods and nematodes. All but 3 of the 608 *C. elegans* tRNA genes and all of the 69 major spliceosome snRNA genes have clearly identifiable homologs in the genome of *D. melanogaster*. We identified several cases of *C. elegans* snoRNAs sharing extended sequence similarity with *D. melanogaster* snoRNAs (supplementary fig. S1, Supplementary Material online), suggesting that at least some families may have survived since the common ancestor of worms and flies. Extending this inference to other snoRNA families is complicated by the somewhat higher rate of sequence divergence of these genes compared with snRNAs and tRNAs (supplementary fig. S1 and table S4, Supplementary Material online).

A simple explanation can be offered to account for the different rates of gene death between different classes of ncRNAs and for the persistence of families despite death of individual genes. Namely, tRNAs and snRNAs form larger families than those of snoRNA genes. Because genes within individual families are somewhat functionally redundant, this effectively reduces the selective pressure to maintain each individual gene, leading to higher rates of gene turnover, while the families persist. An analogy can be made to the robustness at the system level (gene family carrying out a particular function) being distinct from the robustness at the level of individual components, that is, genes (Frank 2007). Below we offer five lines of evidence consistent with predictions of this model.

Different Classes of ncRNAs Form Gene Families of Dramatically Different Sizes

First, we sought to establish that different classes of ncRNAs form families of different sizes.

For snoRNAs and snRNAs, the assignment of genes into families was straightforward, as genes from the same family showed considerable sequence similarity, whereas genes from different families could not be aligned (supplementary fig. S2, Supplementary Material online). Assignment of tRNA families was more challenging because most genes are related to one another. Rather than using an arbitrary cutoff, we examined family compositions at varying levels of sequence identity. When genes were included in the same family as long as they had at least 80% identity to another gene in the group, we found a remarkable correspondence between gene function and sequence similarity (fig. 3). Nearly, all families encode anticodons for only one amino acid (in many cases a single anticodon), and thus, tRNAs encoding different amino acids rarely share sequence similarity higher than 80%.

While nearly all tRNAs (597/608 = 98%) and snRNAs (118/120 = 98%) had clearly identifiable paralogs within

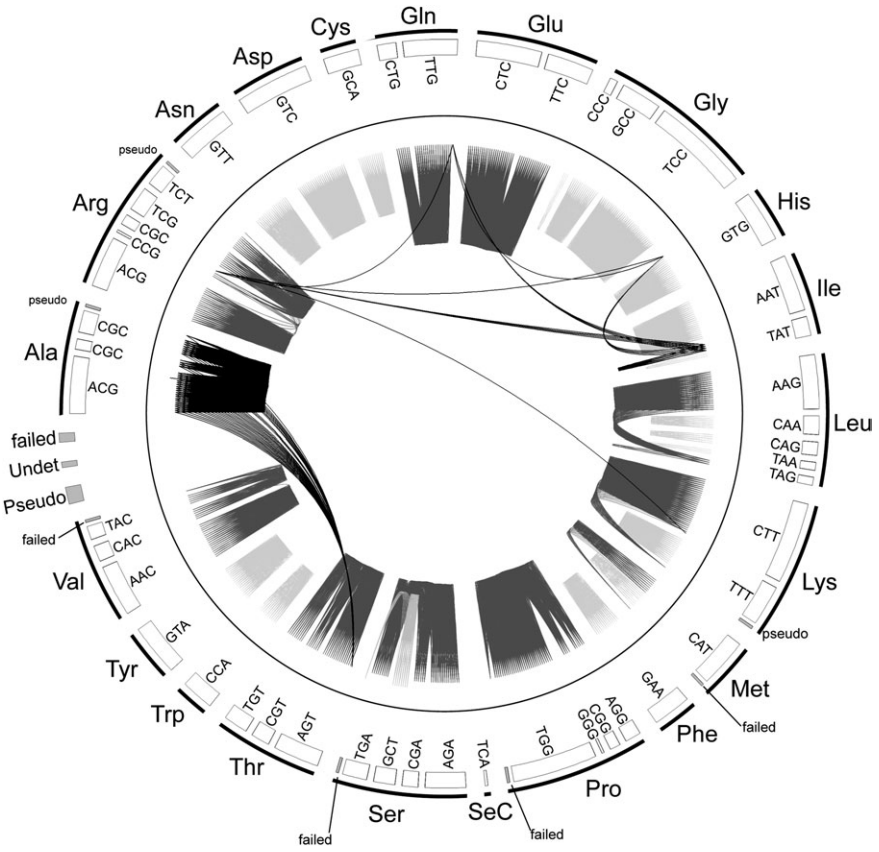


Fig. 3.—Assignment of tRNAs into families using sequence identity corresponds well with anticodon families. The 608 Wormbase-annotated *Caenorhabditis elegans* tRNAs are grouped by their encoded amino acid and corresponding anticodon (outside circle). The anticodon for each sequence was determined by tRNAscan-SE (Lowe and Eddy 1997). This program classified several sequences as pseudogenes (""Pseudo"), undetermined isotype ("Undet"), or not as tRNA ("failed"). Individual sequences that share greater than 80% identity are connected with lines. Vast majority of these connections are between genes with the same anticodon (light gray lines) or encoding for the same amino acid (dark gray lines). The few exceptions (black lines) include, for example, similarity between one Thr tRNA (AGT) sequence and Ala tRNAs.

the *C. elegans* genome, fewer snoRNAs did so (41/142 = 29%), suggesting that snoRNA genes often exist as single-copy loci. Indeed snRNAs and tRNAs formed dramatically larger families than did snoRNAs. Specifically, 82% (501/608) of all *C. elegans* tRNAs belong to subfamilies with 15 or more members, 89% (107/120) of snRNAs belong to groups with 10 or more members, while 71% (101/142) of snoRNAs are single-copy genes (fig. 4). We observed very similar trends when only nested genes were considered (supplementary fig. S3, Supplementary Material online).

Importantly, genes within individual families of snRNAs and tRNAs appear to be functionally redundant. Individual snRNA families (U1, U2, etc.) carry out unique and conserved functions (Newman 1993; Staley and Guthrie 1998). Within each of these families, genes display high degree of sequence identity within *C. elegans* and between *C. elegans* and *D. melanogaster* (supplementary table S4, Supplementary Material online). Functional redundancy can be inferred for genes within individual families of tRNAs (fig. 3). Furthermore, it has been shown that tRNA isotypes encoding the same amino acid are regulated in a coordinated and conserved manner (Kutter et al. 2011). Functional redundancy of paralogous genes is further suggested by the dispensability of individual snRNAs (Parker et al. 1988) and tRNAs (Leung et al. 1991).

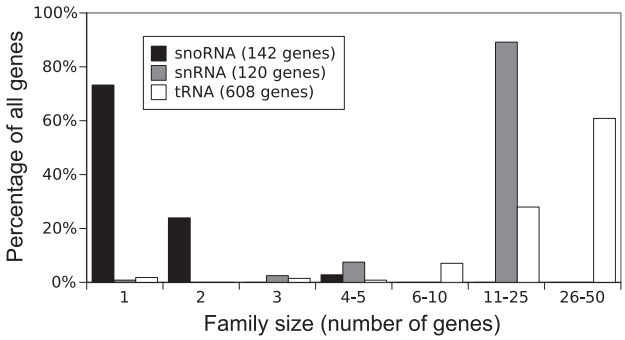


Fig. 4.—Distribution of family sizes of ncRNAs in *Caenorhabditis elegans*. The tRNA, snRNA, and snoRNA genes were classified into families based on sequence similarity.

Single-Gene Families Are Not Static

Second, we sought evidence that single-gene families were undergoing birth-and-death evolution, albeit at a slower rate, as opposed to being static. It is formally possible that single genes are born, persist in the genome without duplicating and eventually die. Alternatively, they may duplicate, but one of the two paralogs may die, leading to an apparently static family size of one gene, that in reality reflects a steady state between gene birth and death. To discriminate between these possibilities, we counted all gene births and deaths for snoRNAs in the genomes of the four Caenorhabditis nematodes (supplementary table S5, Supplementary Material online). We observed the same number of events in which a single gene duplicated (16 cases) as those in which gene families with two members have lost one of the paralogs (16 cases). Single-gene deaths that led to family extinction were considerably less common (2 cases). These results are consistent with the idea that snoRNAs undergo a dynamic albeit slow birth-and-death evolution, while maintaining single-gene families.

Pseudogenes Exist for Large Gene Families

Third, we sought evidence that there exist pseudogenes related to genes comprising large families. During the birth-and-death evolution, some genes accumulate deleterious mutations and subsequently decay (Nei and Rooney 2005). If the death rate is substantial, we would expect to observe partially degraded pseudogenes before sequence divergence has completely erased the evidence of their relationship with functional genes. We looked for the presence of pseudogenes of tRNAs by performing BlastN search against *C. elegans* genome using annotated tRNA sequences. We designated sequences as putative tRNA pseudogenes if they satisfied the following three criteria—1) in the *C. elegans* genome they showed greater sequence similarity to genes of a certain family than to genes from other families, 2) there were no sequences in *C. briggsae* that were more similar to them than the *C. elegans* paralogs described in 1) above, and 3) they were not classified as functional by tRNAscan-SE (Lowe and Eddy 1997). Following these criteria, we found evidence of multiple pseudogenes (supplementary fig. S4A, Supplementary Material online). We used a similar approach to identify pseudogenes for several snRNA families. This was made easier by the high degree of sequence identity within families and no sequence similarity between families. In all instances, putative pseudogenes were the most divergent members of their families and contained mutations that appeared incapacitating (supplementary fig. S4B, Supplementary Material online). We also found evidence of gene death in snRNA families, that is, we documented two instances in which the *C. briggsae* ortholog of an annotated *C. elegans* gene appeared to be decaying (supplementary fig. S4C, Supplementary Material online). We found no evidence of pseudogenes for any of the snoRNAs. It must be noted that in the absence of functional data, it is more challenging to distinguish between divergent snoRNAs and pseudogenes than between divergent tRNAs or snRNAs and their pseudogenes because of the faster sequence divergence of orthologous snoRNAs. Nevertheless, Blast searches revealed no more than five sequences related to previously discovered snoRNAs, indicating that at most a few snoRNA pseudogenes exist.

Evolution of miRNAs in Caenorhabditis Genomes

Fourth, if the relationship between family size and the rate of gene birth and death is general, it should apply to other classes of genes. One major class of ncRNAs that has not been examined above is the miRNAs. Demarcation of miRNA gene families is complicated because they can be grouped by the seed sequence, the mature sequence, or the precursor sequence. We found that if we assign genes into families based on the precursor sequences, then most families contain only a single gene, consistent with results reported for Drosophila miRNAs (average of 1.22 genes per family for all 12 species; Nozawa et al. 2010).

We examined conservation of *C. elegans* miRNAs in *C. briggsae*, *C. remanei*, and *C. brenneri* using the same methodology as described above (supplementary table S6, Supplementary Material online). Unlike snRNAs, snoRNAs, and tRNAs, a large fraction of miRNAs do not have apparent homologs in the non-*C. elegans* species (17/49 = 35% of nested miRNA), consistent with the reports of high de novo birth rates (Lu et al. 2008). However, among the 32 nested *C. elegans* miRNAs that exist in at least one non-*C. elegans* species, there were only four losses (three in *C. briggsae* and one *C. brenneri*). Thus, miRNAs appear to have death rates on par with snoRNAs, consistent with our expectation for single-gene families.

Different Rates of Gene Turnover in Large and Small Families of Drosophila snRNA Genes

Fifth, we tested whether the relationship between family size and the rate of turnover is evident in other genomes. We therefore examined the size of gene families and the rate of gene loss for snRNA genes in Drosophila (Table 2). In the genome of *D. melanogaster*, four types of these genes exist as single-copy loci (all are minor spliceosomal snRNAs, none of which exist in *C. elegans*) and five as multigene families (Marz et al. 2008). Some of these genes are nested, whereas others are located in intergenic regions. Using synteny as a guide, we identified all orthologous loci in the genomes of *D. pseudoobscura* and *D. virilis* (supplementary table S7, Supplementary Material online). These two species are separated from *D. melanogaster* by approximately the same phylogenetic distance as that between *C. elegans* and *C. briggsae* (Kiontke et al. 2004).

Table 2
Conservation of snRNA Loci in *Drosophila*

Family type	Number of loci in *Dmel*	Conserved in *Dpse*	Conserved in *Dvir*
Single gene (U11, U12, U4atac, U6atac)	4	4	3[a]
Multigene (U1, U2, U4, U5, U6)	22	12 (13)[b]	8 (15)[b]

[a] Sequence coverage around the fourth locus is insufficient to determine whether it is conserved.

[b] Because genome assemblies in *Drosophila pseudoobscura* (*Dpse*) and *Drosophila virilis* (*Dvir*) are not as complete as *Drosophila melanogaster* (*Dmel*), orthology relationships for some loci cannot be unequivocally determined. The numbers in parenthesis represent the highest number of conserved loci in these species.

Remarkably, single-copy genes have clear one-to-one orthologs in the genomes of the three examined flies. In contrast, only approximately half of the loci belonging to multigene families were retained during the same evolutionary time. These results confirm our inference of a positive relationship between family size and the rate of gene loss.

Discussion

We conducted an evolutionary analysis of nested ncRNAs in the genomes of Caenorhabditis nematodes. We specifically concentrated on nested gene arrangements because the highly conserved structures of the host loci allow a more reliable identification of orthologous genes. Applied to the sequenced genomes of the four closely related species, this approach revealed a detailed evolutionary history of several gene classes.

Our principal finding is that the evolution of ncRNAs is best described by one of two alternative scenarios. In one, snoRNAs form small (in most cases single gene) families and rarely die—over 97% of the genes conserved between *C. elegans* and *C. briggsae* have also survived in *C. brenneri* and *C. remanei*. In a strikingly different pattern, snRNAs and tRNAs belong to larger families (typically greater than 10 members), but the probability of an individual gene dying along a given nematode lineage is four to seven times higher than that for a snoRNA gene. The existence of snRNA and tRNA pseudogenes is consistent with the idea of higher death rates for these genes. Single-gene families are not static, however, as demonstrated by the lower but detectable rate of birth and death of snoRNAs and miRNAs (here, we refer to birth via duplication as opposed to de novo origin). The relationship between family size and the relative rate of turnover appears to hold for other genomes, as exemplified by snRNA genes in Drosophila. We infer that functional redundancy between genes of a given snRNA or tRNA family contributes to the faster turnover by rendering each individual gene potentially dispensable for the overall function of the family.

Our findings have several implications. First, although our study concentrated on nested ncRNAs, the rules we detailed above may apply to other classes of genes. Previous results appear consistent with this possibility. For example, the ABC transporters constitute a large and ancient gene family of protein-coding genes (Annilo et al. 2006). Whereas these genes are well conserved among Caenorhabditis nematodes (Zhao et al. 2007), the conservation of individual genes between nematodes and humans or Drosophila is poor (Sheps et al. 2004). In general, the rate of birth-and-death evolution appears to be higher in larger gene families (Nam et al. 2004; Thomas 2006, 2007; Nowick et al. 2011). Of course, the relationship between the survivorship of individual genes and the size of the gene families to which they belong is likely contingent on the specifics of individual cases such as gene function, dispensability, etc. Indeed, genes exist that have apparently persisted as single-copy loci for as long as 2 billion years (Fernandes et al. 2008). Yet in cases when other features of genes are similar, it is expected that single-copy genes undergo birth and death at a slower rate than their counterparts from large families (Innan and Kondrashov 2010). Our results with Drosophila snRNAs support this notion. We observed that in a given class of genes, which presumably have similar functions and selective pressures, genes that belonged to larger gene families underwent faster turnover.

Second, selection is commonly invoked to explain why some nested arrangements are conserved. That is, when an arrangement between a nested gene and its host is seen in multiple species, it is inferred to be maintained by natural selection (Chen and Stein 2006; Hudson et al. 2007; Hoeppner et al. 2009). It is assumed that the regulatory elements of a nested gene are dispersed throughout the host locus and that selection is acting to preserve a particular mode of gene regulation (Tsang et al. 2009). It is certainly possible that some nested genes are indeed kept inside their hosts by this type of selection. We found, however, that genes from larger families are lost more rapidly from within nested arrangements than their single-copy counterparts. This suggests that the probability of conservation of a particular arrangement may strongly depend on the size of a family to which a nested gene belongs.

Finally, whereas some classes of genes tend to form multigene families, others persist as single-copy genes. Certainly, a number of forces determine gene family size. One important contributing factor is whether a gene is likely to survive a duplication event. Propensity of a gene to undergo successful duplications may be influenced, among other factors, by its mode of regulation. A comparison of two of the four classes of ncRNAs studied here appears instructive. The majority of eukaryotic snoRNAs are located in introns of host genes and are typically expressed by nuclease processing of the host introns during splicing (Kiss and Filipowicz 1995; Tycowski and Steitz 2001). *Caenorhabditis elegans* snoRNA genes also appear to lack external promoter elements (Deng et al. 2006), and

their expression is somewhat correlated to the expression of their host genes (He et al. 2006). This may limit the ability of snoRNA genes to undergo successful duplications because to maintain proper expression the entire surrounding genomic locus would have to be duplicated. In contrast, many tRNA genes are regulated, at least in part, by internal promoters that are located within the transcribed portion of the gene (Paule and White 2000; Geiduschek and Kassavetis 2001). This may increase the probability that a tRNA gene, once duplicated, would survive. Similar analyses in the future may help to elucidate additional general rules that shape the size and evolutionary dynamics of gene families.

Supplementary Material

Supplementary figures S1–S4 and tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We are grateful to members of the Ruvinsky laboratory for invaluable input throughout the course of this work and to Kevin Bullaughey and Steve Frank for reading the manuscript and offering insightful suggestions. This work was supported in part by a grant from the National Science Foundation (IOS-0843504) to I.R.

Literature Cited

Annilo T, et al. 2006. Evolution of the vertebrate ABC gene family: analysis of gene birth and death. Genomics 88:1–11.

Assis R, Kondrashov AS, Koonin EV, Kondrashov FA. 2008. Nested genes and increasing organizational complexity of metazoan genomes. Trends Genet. 24:475–478.

Bachellerie JP, Cavaillé J, Hüttenhofer A. 2002. The expanding snoRNA world. Biochimie 84:775–790.

Bailey WJ, Kim J, Wagner GP, Ruddle FH. 1997. Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol. 14: 843–853.

Barrière A, et al. 2009. Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. Genome Res. 19:470–480.

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12:1854–1859.

Chen N, Stein LD. 2006. Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. Genome Res. 16:606–617.

Cutter AD. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol. 25:778–786.

Deng W, et al. 2006. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res. 16:20–29.

Fernandes JM, Macqueen DJ, Lee HT, Johnston IA. 2008. Genomic, evolutionary, and expression analyses of cee, an ancient gene involved in normal growth and development. Genomics 91:315–325.

Frank SA. 2007. Maladaptation and the paradox of robustness in evolution. PLoS One 2(10):e1021.

Geiduschek EP, Kassavetis GA. 2001. The RNA polymerase III transcription apparatus. J Mol Biol. 310:1–26.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3:e197.

He H, et al. 2006. Profiling *Caenorhabditis elegans* non-coding RNA expression with a combined microarray. Nucleic Acids Res. 34:2976–2983.

Hillier LW, et al. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol. 5:e167.

Hoeppner MP, White S, Jeffares DC, Poole AM. 2009. Evolutionarily stable association of intronic snoRNAs and microRNAs with their host genes. Genome Biol Evol. 1:420–428.

Hudson SG, et al. 2007. Phylogenetic and genomewide analyses suggest a functional relationship between kayak, the Drosophila fos homolog, and fig, a predicted protein phosphatase 2c nested within a kayak intron. Genetics 177:1349–1361.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet. 11:97–108.

Kiontke K, et al. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A. 101:9003–9008.

Kiss T, Filipowicz W. 1995. Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. Genes Dev. 9:1411–1424.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. J Theor Biol. 239:141–151.

Kutter C, et al. 2011. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nat Genet. 43:948–955.

Leipe DD, Wolf YI, Koonin EV, Aravind L. 2002. Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol. 317:41–72.

Leung J, Sinclair DA, Hayashi S, Tener GM, Grigliatti TA. 1991. Informational redundancy of tRNA(4Ser) and tRNA(7Ser) genes in *Drosophila melanogaster* and evidence for intergenic recombination. J Mol Biol. 219:175–188.

Li WH. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964.

Lu J, et al. 2008. The birth and death of microRNA genes in *Drosophila*. Nat Genet. 40:351–355.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:e357.

Marz M, Kirsten T, Stadler PF. 2008. Evolution of spliceosomal snRNA genes in metazoan animals. J Mol Evol. 67:594–607.

Matera AG, Terns RM, Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol. 8:209–220.

Mattaj IW, Boelens W, Izaurralde E, Jarmolowski A, Kambach C. 1993. Nucleocytoplasmic transport and snRNP assembly. Mol Biol Rep. 18:79–83.

Nam J, et al. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A. 101:1910–1915.

Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A. 94:7799–7806.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.
Newman AJ. 1993. RNA: RNA interactions in the spliceosome. Mol Biol Rep. 18:85–91.
Nowick K, Fields C, Gernat T, Caetano-Anolles D, Kholina N, Stubbs L. 2011. Gain, loss and divergence in primate zinc-finger genes: a rich resource for evolution of gene regulatory differences between species. PLoS One 6:e21553.
Nozawa M, Miura S, Nei M. 2010. Origins and evolution of microRNA genes in *Drosophila* species. Genome Biol Evol. 2:180–189.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.
Parker R, Simmons T, Shuster EO, Siliciano PG, Guthrie C. 1988. Genetic analysis of small nuclear RNAs in *Saccharomyces cerevisiae*: viable sextuple mutant. Mol Cell Biol. 8:3150–3159.
Paule MR, White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res. 28:1283–1298.
Prachumwat A, Li WH. 2008. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. Genome Res. 18:221–232.
Ranz JM, et al. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol. 5:e152.
Reynolds WF. 1995. Developmental stage-specific regulation of Xenopus tRNA genes by an upstream promoter element. J Biol Chem. 270:10703–10710.
Rogers HH, Bergman CM, Griffiths-Jones S. 2010. The evolution of tRNA genes in *Drosophila*. Genome Biol Evol. 2:467–477.
Ruvinsky I, Silver LM. 1997. Newly identified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. Genomics 40:262–266.
Schmitt E, Panvert M, Blanquet S, Mechulam Y. 1998. Crystal structure of methionyl-tRNAfMet transformylase complexed with the initiator formyl-methionyl-tRNAfMet. EMBO J. 17:6819–6826.
Sharp S, DeFranco D, Dingermann T, Farrell P, Söll D. 1981. Internal control regions for transcription of eukaryotic tRNA genes. Proc Natl Acad Sci U S A. 78:6657–6661.
Sheps JA, Ralph S, Zhao Z, Baillie DL, Ling V. 2004. The ABC transporter gene family of *Caenorhabditis elegans* has implications for the evolutionary dynamics of multidrug resistance in eukaryotes. Genome Biol. 5:R15.
Sprinzl M, Vassilenko KS. 2005. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 33(Suppl 1): D139–D140.
Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. Cell 92:315–326.
Takahashi N, et al. 1982. Structure of human immunoglobulin gamma genes: implications for evolution of a gene family. Cell 29:671–679.
Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 16: 1017–1030.
Thomas JH. 2007. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 3:e67.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Tsang WH, Shek KF, Lee TY, Chow KL. 2009. An evolutionarily conserved nested gene pair—Mab21 and Lrba/Nbea in metazoan. Genomics 94:177–187.
Tycowski KT, Steitz JA. 2001. Non-coding snoRNA host genes in Drosophila: expression strategies for modification guide snoRNAs. Eur J Cell Biol. 80:119–125.
Valentine JW. 1994. Late Precambrian bilaterians: grades and clades. Proc Natl Acad Sci U S A. 91:6751–6757.
Vergara IA, Chen N. 2010. Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. BMC Genomics 11:516.
von Grotthuss M, Ashburner M, Ranz JM. 2010. Fragile regions and not functional constraints predominate in shaping gene organization in the genus *Drosophila*. Genome Res. 20:1084–1096.
Wang PPS, Ruvinsky I. 2010. Computational prediction of *Caenorhabditis* box H/ACA snoRNAs using genomic properties of their host genes. RNA 16:290–298.
Zhao Z, Thomas JH, Chen N, Sheps JA, Baillie DL. 2007. Comparative genomics and adaptive selection of the ATP-binding-cassette gene family in Caenorhabditis species. Genetics 175:1407–1418.

Associate editor: Gunter Wagner

Genome Biology and Evolution SMBE