Published online 4 March 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 12* ***5203–5214***
doi:10.1093/nar/gkr102

The *Caenorhabditis elegans* intermediate-size transcriptome shows high degree of stage-specific expression

Yunfei Wang[1,2], Jingjing Chen[1,2], Guifeng Wei[1,2], Housheng He[1,3], Xiaopeng Zhu[1], Tengfei Xiao[1,2], Jiao Yuan[1,2], Bo Dong[1,2], Shunmin He[1,4], Geir Skogerbø[1,]* and Runsheng Chen[1,5,6,]*

[1]Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, [2]Graduate School of the Chinese Academy of Science, Beijing 100080, China, [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02115, USA, [4]Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China, [5]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100080 and [6]Chinese National Human Genome Center, Beijing 100176, China

Received November 19, 2010; Revised February 6, 2011; Accepted February 8, 2011

ABSTRACT

Earlier studies have revealed a substantial amount of transcriptional activity occurring outside annotated protein-coding genes of the *Caenorhabditis elegans* genome. One important fraction of this transcriptional activity relates to intermediate-size (70–500 nt) transcripts (is-ncRNAs) of mostly unknown function. Profiling the expression of this segment of the transcriptome on a tiling array through the *C. elegans* life cycle identified 5866 hitherto unannotated transcripts. The novel loci were distributed across intronic and intergenic space, with some enrichment toward protein-coding gene termini. The majority of the putative is-ncRNAs showed either stage-specific expression, or distinct developmental variation in their expression levels. More than 200 loci showed male-specific expression, and conserved loci were significantly enriched on the X chromosome, both observations strongly suggesting involvement of is-ncRNAs in sex-specific functions. Half of the novel loci were conserved in other nematodes, and numerous loci showed significant conservational correlations to nearby coding genes. Assuming functional roles for most of the novel loci, the data imply a nematode is-ncRNA tool kit of considerable size and variety.

INTRODUCTION

Recent years have seen increasing efforts toward the unraveling of the functional roles of non-protein coding RNAs (ncRNAs) in organismal development. Noncoding RNAs have broadly been divided into small (<200 nt) and long (>200 nt) transcripts (1), and research has been particularly intense on microRNAs and other RNAs ranging between 15 and 40 nt in size. Simultaneously, increasing efforts are being made to investigate the roles of many long and mRNA-like ncRNAs found in mammalian transcriptomes (2). However, eukaryote transcriptomes are also composed of several classes of transcripts whose size range spans the border between small and long RNA. For practical purposes we will, in the following text, refer to transcripts in this size range (70–500 nt) as ‘intermediate-size ncRNAs’ (is-ncRNAs). Such ncRNAs include the well-studied snRNAs and snoRNAs, but it has also been known since early this century that this transcript range also comprises numerous other transcripts with less well-defined roles (3,4). Large-scale transcriptome analyses by tiling array or deep sequencing have recently demonstrated the existence of considerable numbers of transcripts in this size range in all investigated organisms (5–7). Very many of the intermediate-size transcripts in eukaryotes appear to occur in the context of protein coding loci and are being referred to under various denominations, such as PASRs (promoter-associated short RNAs), TASRs (terminator-associated short RNAs), CUTs (cryptic

*To whom correspondence should be addressed. Tel: +86 64888543, Fax: +86 10 64889892; Email: crs@sun5.ibp.ac.cn
Correspondence may also be addressed to Geir Skogerbø. Tel: +86 64888543, Fax: +86 10 64889892; Email: zgb@moon.ibp.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

unstable transcripts), SUTs (stable unannotated transcripts), PROMTs (transcripts upstream of core promoters) and eRNAs (enhancer RNA) (8–12), depending on the organism of origin, size range and specific genomic location. However, even after very stringent filtering of tiling array and RNA-Seq data, there remained several thousand putative is-ncRNAs in mammalian transcriptomes that could not be accounted for in this way (13).

With the exception of snRNAs, snoRNAs and a few others, whose cellular roles were largely established in the final decades of the 20th century, the functional properties of is-ncRNAs are just beginning to be touched upon. Transcripts arising around and in concert with transcription of protein coding loci may be involved in transcription activation of the coding loci, or simply be 'by-products' of such activation (14–16). It is probably premature to explain all coding locus-associated transcription in this way, and is-ncRNA genes located in deep intergenic space can hardly be thus accounted for. On the contrary, there is compelling evidence that the large numbers of identified but yet unstudied non-coding transcripts have intrinsic functionality, as indicated by the conservation of their promoters, structures, genomic position and expression patterns (17–20). Investigations into a number of is-ncRNAs in *C. elegans* suggested that on the one hand, they are fairly recalcitrant to knock-down by RNAi, and on the other hand, their cellular stability largely depends on interactions with proteins or protein complexes (21). This would suggest that transcripts in this size range exert their functions in stable ribonucleoproteins which, in addition to being vehicles of their cellular function, also confer resistance to cellular ribonucleases. In the former respect, is-ncRNAs may resemble miRNAs in that they link the digital information of the nucleotide to the analogue information of protein structure (22). Given their sheer numbers (apparently in the thousands) and the relatively low research effort invested in elucidating their functional roles, is-ncRNAs have the potential to fill a regulatory space of a magnitude similar to that occupied by microRNAs.

The current annotation of the ws190 data of *C. elegans* genome estimated ~20 000 protein coding genes and ~900 intermediated sized (70–500 bp) ncRNA genes (23,24) and computational predictions have suggested the presence of an additional 3000–4000 is-ncRNAs in the genome (4,25). We have previously carried out a tiling array analysis which identified approximately 1200 novel intermediate-size transcripts in a mixed stage culture of *C. elegans* (5). Much of the mammalian tiling array data have not stood up well to scrutiny in light of deep sequencing data (13); however, careful analyses of both methodologies in *C. elegans* demonstrate that the tiling array compares well with deep sequencing when necessary measures are in place (26). We, therefore, applied tiling array analysis to six developmental and two conditional stages of the nematode, detecting 5866 novel intermediate-size transcripts [or transcribed fragments (transfrags) of unknown function; TUFs]. Fifty-two (85%) of 59 tested TUFs were verifiable by reverse transcription-polymerase chain reaction (RT–PCR) and an additional 10 of 10 TUFs were verifiable by Northern blot and RACE experiments. These TUFs exhibited more complex expression patterns across stages, and most showed features different from that of known is-ncRNAs types and coding genes, suggesting the existence of novel functional types of intermediate-size RNAs.

MATERIALS AND METHODS

Preparation of RNA samples and tiling array

RNA samples were prepared from wild-type N2 strain worms at larval stages 1–4 (L1–L4), mature adult (MA) and male (ML) worms, dauer stage worms (DU), and worms subjected to heat-shock (HS). Total RNA was extracted from each of the eight different developmental stages and environmental conditions according to the Trizol (Invitrogen) protocol. Intermediate-size RNAs (70–500nt) were isolated using a QIAGEN tip (Qiagen), and remaining rRNAs were removed by adapting the MicrobExpress kits (Ambion). The enriched is-RNAs were dephosphorylated with CIAP (Fermentas) and then ligated to the 3′-adaptor oligonucleotide by T4 RNA ligase (Fermentas). Each RNA sample was reverse transcribed using random hexamers and a primer complementary to the 3′-adaptor. Double-strand cDNA was fractioned, labeled and hybridized to the Affymetrix GeneChip® *C. elegans* Tiling 1.0R Array according to Affymetrix's GeneChip Whole Transcript (WT) Double-Stranded Target Assay Manual. RNA sample preparation and hybridization was carried out twice for each of the *C. elegans* stages or conditions, except for MLs and MAs, which only were sampled once.

Computational analysis

The genome annotation, sequence and conservation data were downloaded from Wormbase (http://www.wormbase.org, version WS190) (23) and the UCSC genome browser (http://genome.ucsc.edu/, version ce6) (27). The raw tiling array data was pre-processed using the Affymetrix Tiling Analysis Software (TAS, version 1.1.02). Briefly, quantile-normalization was performed on the tiling array replicates and signal intensity values were then adjusted to yield a median intensity of 100. Log_2 [max (PM−MM, 1)] was calculated for each probe as an estimate of the expression level at each genomic position. Probe signal intensities were considered as significant over background if above the threshold associated with a false-positive rate of 0.05. Transcribed fragments (transfrags) were identified using a sliding window method with window size = 100, maxgap = 30 and minrun = 70.

For normalization within arrays, the signal intensity of all stages and conditions were quantile-normalized (R, limma package). The transfrags were filtered with the normalized signal by removing the ones with low signal intensity [threshold = 6, false detection rate (FDR) = 0.05]. The remaining transfrags were annotated by mapping to known is-ncRNAs (Wormbase and other published is-ncRNAs) (4,5,28), or to other annotations from UCSC (SangerGene annotations, pseudogene annotation and repeat annotations) (27), introns or unannotated intergenic regions. The unannotated transfrags dataset was refined by removing TUFs covering probes corresponding to multiple genomic regions or with homology to ESTs (identity >95% and alignment >35 bp).

Chromosome location, GC content, secondary structure (29) and development profile analyses were done for all TUFs. Conservation analysis was implemented using phastCons data (UCSC, goldenPath/ce6/phastCons6wayScores). BLAST and Infernal (30) were performed against several non-coding RNA databases (28,31–34) for sequence and functional homology. Prediction of snoRNA was implemented using snoscan (35), snoReport (36) and snoGPS (37) to both strand of TUFs. Motif analysis was done using the MEME/MAST program (38,39).

The TUF signal intensity profile data were combined with coding gene-expression profile data obtained from the Genome B.C. *Candida elegans* Gene Expression Consortium (http://elegans.bcgsc.bc.ca), and both data sets were quantile normalized.

Validation of TUFs by RT–PCR

Total RNA was digested with DNase I (Fermentas), dephosphorylated and ligated to the 3AD (see Supplementary Data) oligo. Reverse transcription was performed by using a 3RT primer (see Supplementary Data). First-strand cDNA was used as template for PCR with a pair of TUF sequence-specific primers. Total RNA digested with DNase I was used as negative control while genomic DNA was used as positive controls. Total RNA without DNase I digestion was used as control for genomic DNA contamination.

Northern blot and RACE

The RNA probes used for Northern blots were labeled with DIG-UTP (Roche) by *in vitro* transcription, hybridized to 4 μg total RNA at 62°C overnight, and detected with CDP-Star (Roche). For 3′ RACE, total RNA was ligated to the 3AD oligo, reverse transcribed with the 3RT primer and PCR amplified with GSP primers. 5′ RACE was performed using the SMART cDNA synthesis kit (ClonTech). Oligos used in these experiments are listed in Supplementary Document S4.

RESULTS

We have previously reported a survey of the genomic transcriptional activity in *C. elegans* which identified approximately 1200 novel intermediate-size (70–500 nt) transcripts in a mixed stage worm population (5). In order to obtain a more detailed map of the intermediate-size RNAs through the *C. elegans* life cycle, we applied the same tiling microarray approach (5) to worms in eight different developmental stages and environmental conditions. These included the four larval stages L1–L4, the MA stage, ML, worms in the DU stage, and worms exposed to HS. A transcribed fragment [transfrag (40)] was defined as at least four consecutive positive probes each separated by a gap of ≤30 bp. The data were normalized and we applied $\log_2$ (signal intensity) (L2SI) of six as the lower threshold cutoff for transfrag selection (see ‘Methods’ section and Supplementary Data for details). This rendered 32 230 transfrags, covering 3.58 million base pairs of the *C. elegans* genome, which were retained for further analysis. About 56.1% of the expressed base pairs had been annotated as either coding sequences or untranslated regions (UTRs) of coding transcripts, and 3.1% were annotated as either is-ncRNAs or pseudogenes (Figure 1A). The remaining 40.8% of the transcribed nucleotides either locate to introns (17.4%) or intergenic regions (23.4%). The transfrags were distributed almost evenly on the six *C. elegans* chromosomes (Supplementary Figure S1).

The tiling arrays detected 73.3% of all known is-ncRNAs (Table 1). This is a lower fraction than reported in a previous tiling microarray assay of mixed stage worms (5), and owes mainly to a more stringent data-filtering procedure (see ‘Methods’ section and Supplementary Data for details). This effect was most prominent for the relatively short tRNAs, but was also seen for other classes of known is-ncRNAs. The protocol applied for sample preparation was not designed to detect mature miRNAs (or other similarly sized small RNAs like 21U-RNAs), but the tiling array nonetheless produced positive signals for 29% of the 138 known miRNA loci and a small number of 21U-RNAs loci. A comparison to previously published tiling array analysis of mixed stage intermediate size RNAs showed that our TUFs overlapped 36–53% of the mixed stages TUFs (see Supplementary Document S3). The overlapping TUFs tended to be ubiquitously expressed through all developmental stages and had correlated ($r^2 = 0.47$) expression levels in the two datasets (Figure 1B).

Only 364 transfrags were highly expressed (L2SI > 10) at any stage or condition, and the expression level of the majority of the transfrags (∼21 480) were expressed at relatively low levels (6 < L2SI < 7; Figure 1C). This is well below the expression level of well-established is-ncRNA classes, such as snRNAs and snoRNAs, which generally had ×4–6 higher expression (Figure 1D). Other previously verified is-ncRNAs displayed a wider distribution of expression level and a considerable fraction of these fell in the same expression range as the majority of transfrags (Figure 1D). The number of transfrags expressed above the threshold (L2SI > 6) at any given stage or condition varied significantly with nematode development, most transfrags being expressed in the first three larval stages and fewer toward maturity (Figure 1E). The lowest number of transfrags was observed in ML worms.

The 32 230 transfrags were separated into different categories according to their genomic locations. We compared annotation data from Wormbase (ws190) to the locations of our transfrags (Supplementary Figure S2). Approximately 19 000 transfrags wholly or partially overlapped exonic sequences. These could represent independent transcriptional unit overlapping coding genes in either orientation, as recently observed in other species. However, as the percentage of exonic transfrags declined markedly relative to other types of transfrags with higher

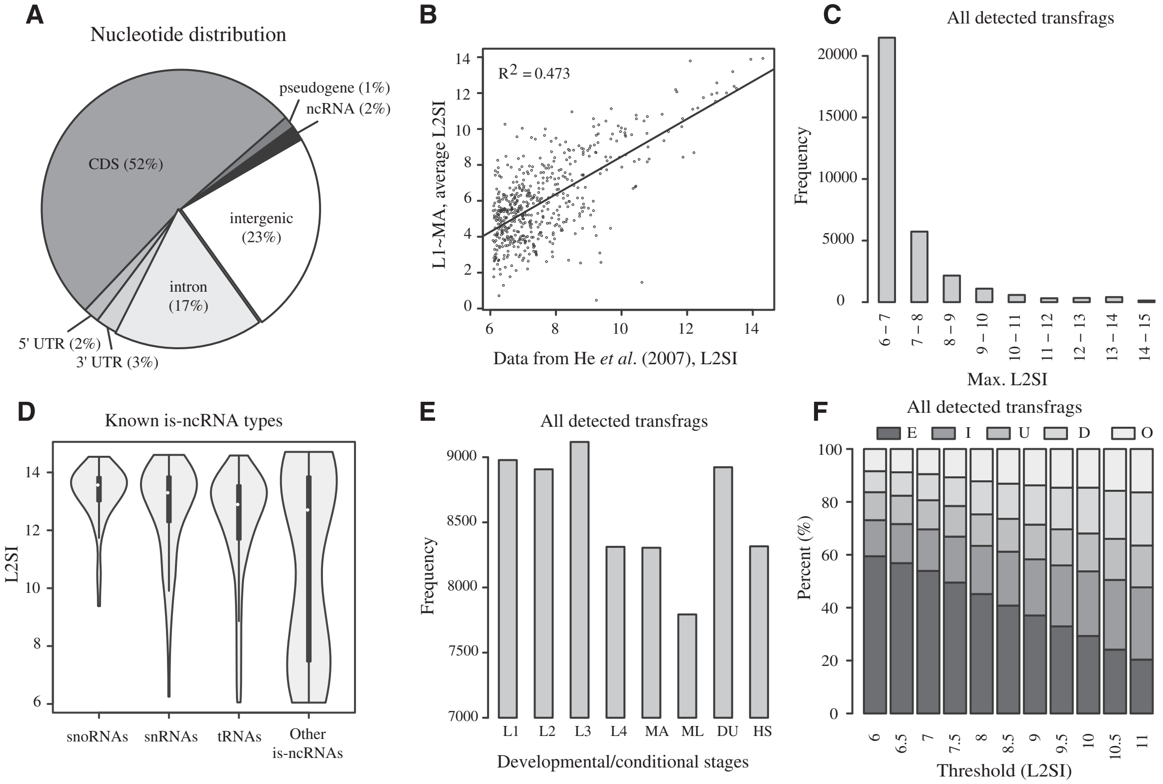


Figure 1. Transcribed fragments. (**A**) Overall nucleotide distribution of the 32 230 tiling array transfrags. (**B**) TUF expression. Comparison to data from He *et al.* (5). (**C**) Transfrag maximum $\log_2$ (signal intensity) (L2SI) distributions. (**D**) Known is-ncRNA maximum L2SI distribution ('Other is-ncRNAs' include all known is-ncRNAs except snoRNAs, snRNAs and tRNAs). (**E**) Number of transfrag expressed in each developmental and conditional stage. (**F**) Genomic distribution of the 32 230 transfrags mapping to exons (E), introns (I) and intergenic regions upstream (U), downstream (D) and distant (>2 kb, O) from the nearest coding gene.

Table 1. Detection rates for known is-ncRNAs loci

ncRNA class	Interrogated	Detected	Detection rate (%)
tRNA	631	440	69.73
rRNA	21	17	80.95
snoRNA	133	119	89.47
snRNA	90	76	84.44
SL2 RNA	8	7	87.50
scRNA	1	1	100.00
sm Y RNA	1	1	100.00
Uncharacterized is-ncRNAs	60	32	53.33
All interrogated RNAs	945	693	73.33
21U-RNA	5356	30	0.56
miRNA	138	40	28.99

signal intensity was increased (Figure 1F), it seems reasonable to assume that at least a fraction of the exonic transfrags represent degradation products of pre-mRNAs and mature mRNAs, or elements that have been spliced from pre-mRNAs.

In subsequent analyses, we focused on transfrags not overlapping with other annotated genomic elements, i.e. intergenic and intronic transfrags. We therefore removed transfrags that had probes corresponding to multiple genome loci and transfrags that overlapped with repeat sequences, exons, known is-ncRNAs and pseudogenes (WS190). The remaining 6552 transfrags were filtered with EST data and WS190 data, leaving 5866 transfrags of unknown function (TUFs) not overlapping any annotated sequences. Recent analyses of mammalian tiling array data (13) have suggested that most short TUFs of low signal intensity frequently represent cross-hybridization and other noise. As the RNA sample used for hybridization to the tiling arrays had been size fractioned and depleted of the most abundant RNA species, the potential for cross-hybridization was greatly reduced compared to tiling arrays hybridized with polyadenylated RNA or total RNA. Furthermore, TUFs consisting of one or more probes matching more than one genomic position were removed (See 'Methods' section). Validation of 59 TUFs randomly sampled in the low-signal intensity range $(6 < \text{max. L2SI} < 8)$ by RT-PCR confirmed this by returning positive amplification for 85% of the selected TUFs (Supplementary Document S2). Northern blot analysis of 10 TUFs all indicated a transcript in the expected size range (Figure 2), and subsequent RACE analysis showed that TUF and transcript size generally differed by 2–34% (Supplementary Document S2). The GC content was lower in TUFs than in known is-ncRNAs, and similar to that of coding exons (Figure 3A). There were no differences between intronic and intergenic TUFs, but in both genomic contexts, the TUF GC content was higher than in the surrounding sequence.

Genomic organization of the TUF loci

The chromosomal distribution of TUFs deviated from that of known is-ncRNA loci. Known is-ncRNAs loci (rRNAs and tRNAs not included) in *C. elegans*, are almost evenly distributed on the autosomes but scarce on the X chromosome. TUFs, on the other hand, were slightly over-represented on the X chromosome (Figure 3B), irrespective of developmental stage or environmental condition under which they were expressed. Intergenic TUFs without nearby coding genes (distant TUFs, see below) showed the strongest tendency to locate on the X chromosome (26%). There were 4025 TUFs (69%) with an intergenic location, amounting to an overall density of about 1 TUF/10 kb of intergenic sequence (98.6 TUFs/Mb). As a number of recent analysis (6,7,41,42) have observed frequent non-coding transcription in the vicinity of active coding loci, we further divided the intergenic TUFs into two ‘proximal’ groups of 1895 and 858 TUFs located within 2 kb upstream or downstream, respectively, of a coding gene, and a group of 1272 distant TUFs located >2 kb away from any gene. The density of proximal upstream (302.4 TUFs/Mb) and downstream (136.9 TUFs/Mb) was on average five times that of distant TUFs (45.0 TUFs/Mb). Closer analysis of 50 bp windows in the gene proximal sequences showed that TUF density peaked within the first hundred base pairs upstream and downstream of the WS190 annotated genes, reaching the highest value at ~150 bp upstream of

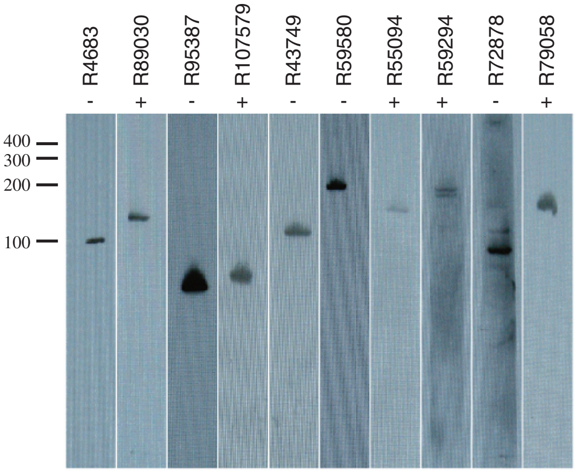


Figure 2. Northern blot analysis of 10 TUFs. All 10 TUFs are within the expected 70–500 nt range.

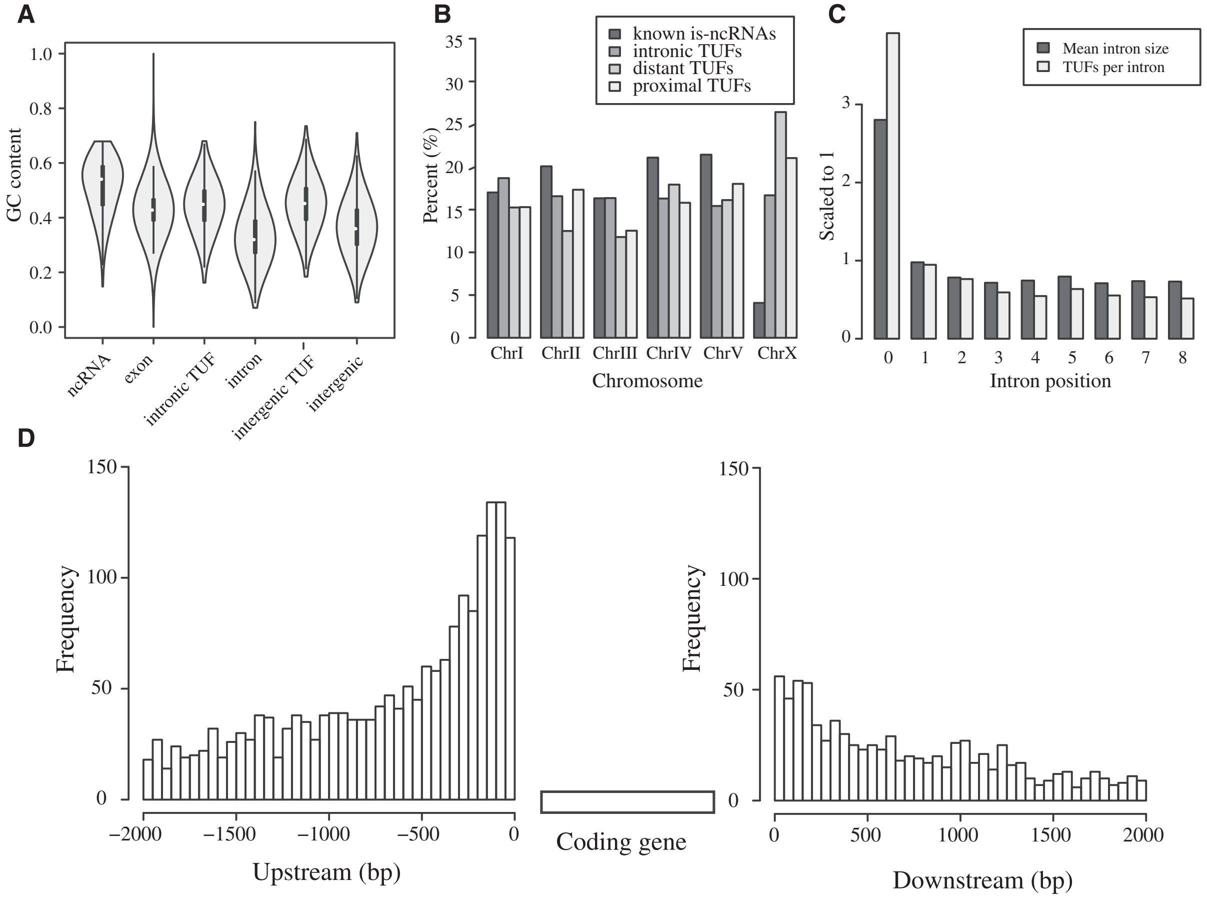


Figure 3. TUF genomic distribution. (**A**) GC content distribution for TUFs and other genomic regions. The GC contents of exons, introns and intergenic regions were calculated using 10 000 randomly selected exons, introns and 100 bp intergenic fragments. (**B**) Chromosomal distribution of known is-ncRNA and TUF loci. (**C**) Intron position, length and TUF density. Average intron size (420.56 nt) and TUF density (0.055 TUFs/intron) values were scaled to 1. Introns located in 5′-UTRs were labelled 0, and intron positons >8 are not shown. (**D**) TUF density in regions flanking coding genes.

the 5′ termini of annotated genes (Figure 3D). Together with the 1841 intronic TUFs, 4594 TUFs (78.3%) are found within or in the vicinity of a protein coding sequence.

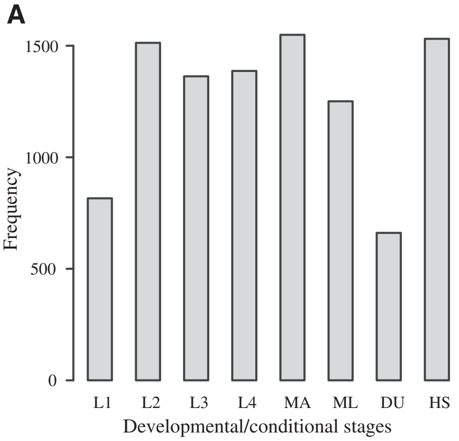
The frequency of TUFs within a given intron varied considerably with intron position and the average number of TUFs in 5′ UTR introns (intron 0) was ~4 times higher than in later introns within the CDS region (Figure 3C). Longer intron 0 is a general property of the eukaryotic gene structure (43), and introns located at 5′ proximal of genes have been shown to have important functional properties, often related to gene expression (44,45). Using reference gene annotation data from UCSC, we calculated intron lengths for 23 665 high-confidence genes. The average length of the intron 0 (Figure 3C) was ~2.5 times longer than other introns ($P < 0.01$, Z-test); however, despite the longer size of intron 0, its TUF density (1 TUF/11.5 kb of intron sequence) was nearly twice that of other introns (1 TUF/18.5 kb of sequence).

Most TUFs show stage-dependent expression

There were no particular differences in average expression levels of TUFs located in the four different genomic regions (Supplementary Figure S3A). However, unlike most of the previously known is-ncRNAs, which show relatively uniform expression through worm development, TUF expression commonly fluctuated considerably across stages (Supplementary Figure S3B). There were marked differences in the number of TUFs that were expressed above cutoff (L2SI = 6) in each developmental stage or condition (Figure 4A), and 3975 (67.8%) of the TUFs were expressed in only one stage or condition. Only 70 TUFs (1.2%) were ubiquitously expressed in all stages and conditions, these, however, were much more strongly expressed (mean L2SI = 11.4) than TUFs expressed in a single stage or condition (mean L2SI = 6.4). The number of TUFs expressed at the first larval stage was markedly lower than in the older larval and the MAs (Figure 4A) and the number expressed in ML worms was ~20% lower than in the general (mostly hermaphrodite) MA populations. The number of expressed TUFs was lowest in the DU stage, which probably reflects the generally reduced physiological activity at this stage. HS worms had a high number of expressed TUFs, and also displayed the highest number of TUFs that were specifically expressed at any stage or condition (427; Figure 4B). However, a high number of specifically expressed TUFs was also seen in the ML stage (229; Figure 4B), despite the relative low total number of expressed TUFs at this stage, suggesting that a disproportionally high number of small transcripts may be required for attaining or maintaining this specific stage. Also, early worm development (L1) was associated with a relatively high number (65) of specifically expressed TUFs, despite the overall low number of TUFs expressed. A small number of TUFs appeared to be present at all but one specific stage, such as the DU (6) and ML (7) stages and the HS (5).

TUFs show dual conservation distributions

In order to estimate the conservation levels of intergenic and intronic TUFs, PhastCons scores from six nematode genomes were downloaded from UCSC (27). To compare the conservation distribution of TUFs to annotated transcriptional units, we calculated phastCons scores for all known is-ncRNAs, and for randomly selected exonic, intronic and intergenic fragments (100 bp for 10 000, times, respectively). Compared with known is-ncRNAs, which are generally well conserved among the nematodes, both intergenic and intronic TUFs displayed a dual distribution, in which ~40% of the TUF are almost completely non-conserved (average phastCons score <0.2), and the majority of the remaining TUF having PhastCons scores in the range 0.4–0.8 (Figure 5A and B). Both intronic and intergenic TUFs have distributions that differ in shape from those of their respective genomic environments (i.e. randomly selected introns and intergenic sequences), but only intronic TUFs had significantly ($P < 0.01$, Wilcoxon's test) higher phastCons scores than introns in



B

Type	Count	L1	L2	L3	L4	MA	ML	DU	HS
HS enriched	427	-	-	-	-	-	-	-	+
HS reduced	5	+	+	+	+	+	+	-	-
Dauer enriched	24	-	-	-	-	-	-	+	-
Dauer reduced	6	+	+	+	+	+	+	-	-
Male enriched	229	-	-	-	-	-	+	-	-
Male reduced	7	+	+	+	+	+	-	+	+
Early stage enriched	65	+	-	-	-	-			
Middle stage enriched	20	-	+	+	-	-			
Late stage enriched	46	-	-	-	+	+			

Figure 4. TUF expression. (**A**) Number of TUFs expressed at each stage and condition. (**B**) Stage-specific TUF expression. The table shows number of TUFs expressed in only one (or one group of) stage (s) ('enriched') or expressed in all (or nearly all) stages but one ('reduced'). The data corresponding to the shaded area were not considered.

general. Only 20% of all TUFs had phastCons scores at the same level as most known is-ncRNAs (average phastCons score > 0.7; Figure 5B); however, these approximately ~1100 TUFs represent more than twice the present number of intermediate-size transcript loci in *C. elegans* with phastCons scores at this level. An additional peculiarity was that TUFs located on the X chromosome had significantly ($P < 2.2 \times 10^{-16}$, Wilcoxon's test) higher phastCons scores than autosomal TUFs (Figure 5C). TUFs on chromosomes I and IV had somewhat lower phastCons scores ($P < 0.01$ and $P < 0.05$, respectively) than TUFs on the remaining autosomes, but the differences were far less pronounced than that between the X chromosome and the autosomes. Except from being conspicuously absent immediately downstream of coding loci, TUFs with high phastCons scores showed no particular distribution on the X chromosome (Figure 5D).

Further analysis of the relationship between TUFs and sequence conservation revealed several interesting correlations. There were 1895 TUFs that were located within 1 kb upstream of 1714 protein coding loci, and phastCons scores of coding loci with at least one TUF located upstream of their annotated 5′-termini had on average significantly ($P < 0.01$, Z-test) higher phastCons scores than other coding genes (Figure 5E). There was also a tendency that immediate up- and down-stream flanking regions (<500 bp) of coding genes displayed higher phastCons scores if a TUF was located within the flanking region (Figure 5E). To further analyze potential interactions between TUFs and neighboring coding genes, coding gene expression profile data from the *C. elegans* Gene Expression Consortium (see 'Methods' section) were compared with expression data from the larval stages (L1–L4). We obtained expression profiles for 1872 coding genes with at least one or more flanking or intronic TUFs, and established 2623 TUF–gene pairs, consisting of one TUF and its nearest gene; however, 94.5% of TUFs showed little or no expressional correlation with their neighboring genes. A recent analysis has found that mouse neuronal enhancers to be commonly transcribed (12). However, a re-analysis of *C. elegans* DNase I hypersensitive sites (46) failed to show any substantial co-location of TUFs and potential *cis*-regulatory elements (Supplementary Figure S4), and transcriptional activation of enhancers may thus not be a prominent feature of the *cis*-regulatory mechanism in the worm.

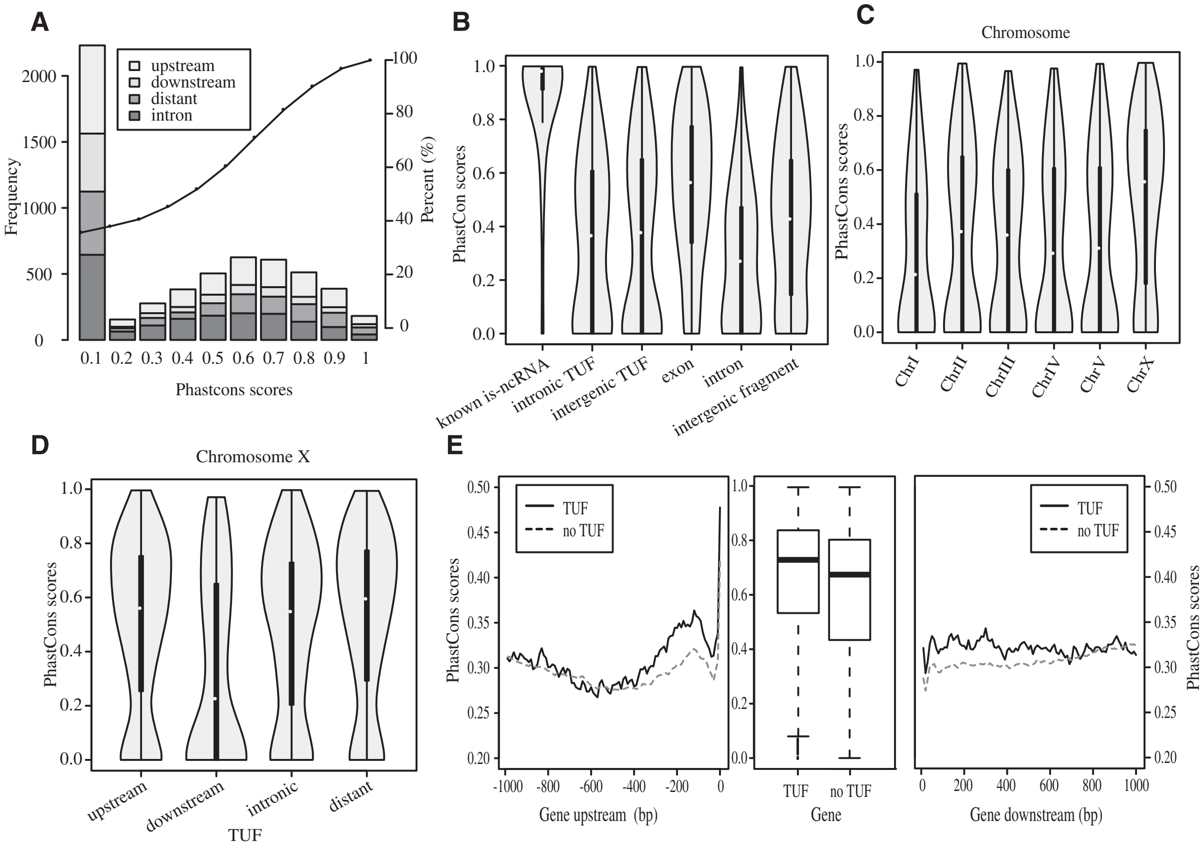


Figure 5. Conservation analyses. (**A**) Distribution of TUF phastCons scores. (**B**) PhastCons score distributions of known is-ncRNAs, TUFs and 100 bp randomly selected intronic, exonic and intergenic sequences. Intronic TUFs have significantly higher phastCons scores than that of introns (Wilcoxon's test, $P < 0.01$). (**C**) Chromosomal distribution of TUF phastCons score. The TUF phastCons scores on chromosome X are significantly higher than that on the autosomes (Wilcoxon's test, $P < 0.01$). (**D**) TUF phastCons scores distribution on chromosome X. (**E**) TUFs are associated with coding gene phastCons scores. A coding gene and its flanking sequences have higher phastCons scores when an upstream TUFs is present (Wilcoxon's test, $P < 0.01$).

Motifs, repeats and conserved structure

A search for conserved structures using the Infernal software (30) against the Rfam database (31) yielded 33 hits. Among these, four were mammalian miRNAs, 26 were sequences that the software assigned as 'HIV-related signal', two were prokaryotic small RNA or mRNA leader sequences, and one was *C. elegans* snoRNA U6-47, which for some reason has not yet been included into Wormbase. Secondary structure analysis predicted 42 and 23 novel C/D box and H/ACA box snoRNAs, respectively (Supplementary Document S1).

Previous analyses of is-ncRNA loci in *C. elegans* have revealed the presence of common upstream motifs with putative or verified promoter activity (4,47). A search for known upstream sequence motifs in 200 bp sequence flanking each was performed with the MAST software (38). The flanking (and most likely upstream) sequence of 15 TUFs displayed one of the three previously reported *C. elegans* is-ncRNA promoter motifs (UM1–UM3, $E < 0.01$) (4,5). Eleven of these were UM1, corresponding to the snRNA proximal sequence element (PSE) previously identified in *C. elegans* (48). The MEME software was also applied to search both strands of the TUFs for internal sequence motifs, yielding one novel internal motif (IM4), which was shared by 8 TUFs and formed part of a predicted stem–loop structure (Figure 6).

About one-third of the TUFs (1865) were flanked within 200 bp by repeats from the UCSC RepeatMasker annotation (2436 repeats in total). Approximately half (1231) were simple repeats, the majority being AT-rich (758), and the other half complex repeats of which CELE14B was the most frequent type. The distance distribution of the repeats showed a conspicuous peak at 10–15 bp away from the TUF (Supplementary Figure S5), which might result from the removal of repetitive region in the design of the chip. A certain fraction of *C. elegans* repeats might act as promoters to initiated the transcription for downstream sequences as that in other organisms (49). The CELE14 MITE repeat family occurs 3020 times in the *C. elegans* genome, mostly clustered near the ends of the autosomes (50) and the CELE14B repeat was observed flanking 80 TUFs. In a few cases, CELE14B was flanking a TUF on both sides, leaving little room for promoter sequences to be located elsewhere than within the CELE14B sequences themselves (Supplementary Figure S6).

DISCUSSION

Using tiling array technology, we have profiled the intermediate-size (70–500 nt) transcriptome of *C. elegans* through eight developmental and conditional stages. After

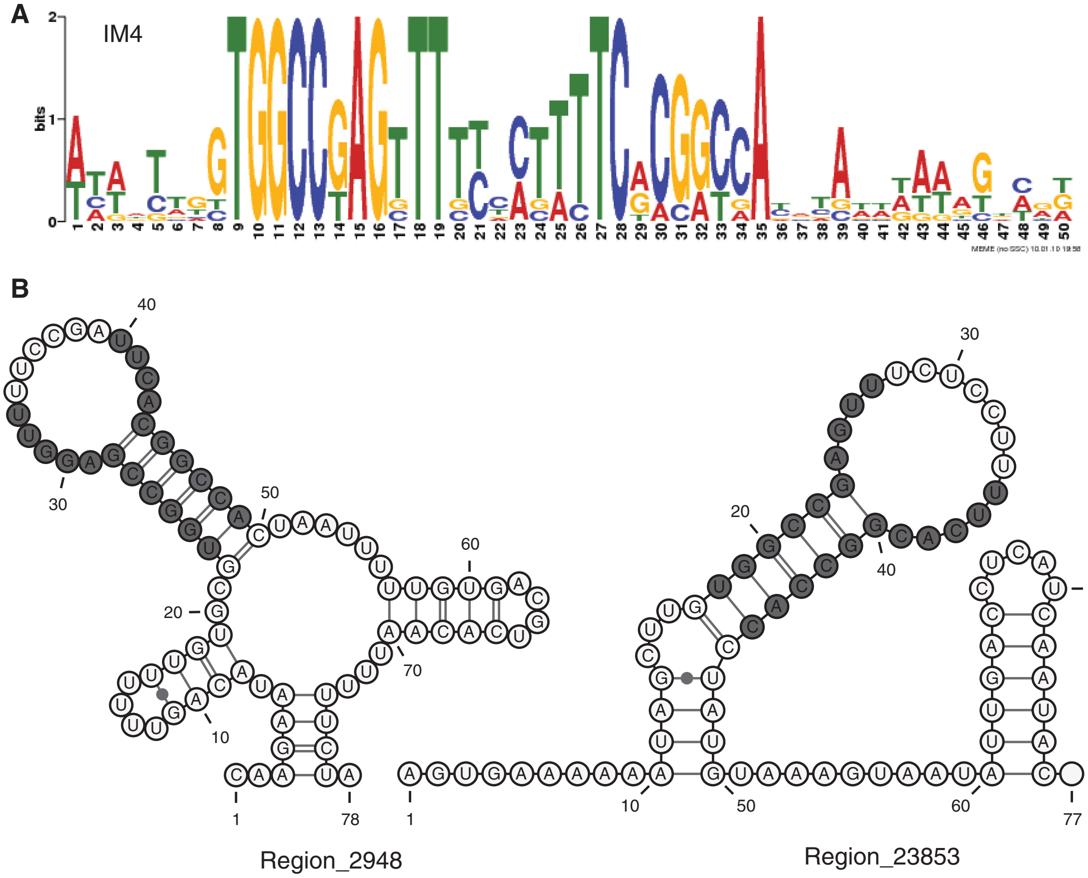


Figure 6. TUF motif. (**A**) Internal motif (IM4, *e* = 1.2e−005) of 8 TUFs. (**B**) Predicted secondary of TUFs containing IM4, which form a stem–loop structure (shaded).

stringent filtering, the analysis altogether included 32 230 transfrags, of which 5866 were located in non-annotated intergenic or intronic regions. Most of these potential RNAs exhibited distinct features common to known classes of is-ncRNAs, and a fraction of them shared novel upstream or internal motifs. The data give an elaborate view of the chromosomal distribution, conservation and expression profiles of this segment of the *C. elegans* transcriptome, and the number and expressional complexity of the novel transcripts suggest roles in nematode development and phenotypic specification (51). We have previously estimated that there may be 3000–4000 is-ncRNA loci in *C. elegans* (4,25) and a previous analysis of mix-stage worm suggested even higher numbers (5). Taking into account the apparent developmental and conditional specificity of TUF expression, analysis of additional developmental stages (e.g. egg, embryo, aging, etc.) and conditions might increase the number to a level of 7000–10 000 intermediate-size transcripts.

The emergence of second generation sequencing (RNA-Seq) data have cast doubt on the quality and correctness of mammalian tiling array data (13). Tiling arrays has generally higher false-positive rates than RNA-Seq data and non-coding transcription appear to be concentrated within and around coding loci rather than pervasive throughout the entire genome (13). Contrary to the human tiling microarrays analyzed by van Bakel *et al.* (13), the *C. elegans* tiling array is designed with both perfect matched and mismatched probes, which facilitates stringent data normalization and filtering. An analysis of *C. elegans* tiling array and RNA-Seq data found good general agreement in expression levels at identical loci in the two data, and 86% of the loci identified by the tiling array as differentially expressed between two developmental stages were confirmed by the RNA-Seq data (26). If anything, the tiling array data underestimated the number of differentially expressed loci (26), and the variation in TUF expression levels across nematode development may actually be even more pronounced than reported here. The risk of high false-positive rates owing to cross-hybridization could mainly be limited to a 'black list' of 2327 regions (26) of which only one remained after our own (independent) filtering of the data. Our validation rates (~85%) were higher than those reported for human tiling array data (25–70%, (6,52), despite the fact that the TUFs used for validation were selected among those with lowest expression levels ($6 < \text{L2SI} < 8$).

We have assumed that the majority of the novel intermediate-size transcripts are not translated into peptides. A recent study in *Drosophila* have shown that some transcripts previously considered to be non-coding RNAs contain short open reading frames encode 11–32 amino acid long bioactive peptides involved in temporal regulation of epidermal morphogenesis (53). Although a number of strategies have been developed to distinguish protein-coding RNAs from ncRNAs, the distinction between the protein-coding and non-coding categories is not entirely clear (1), and the existence of bifunctional RNAs further contribute to this confusion (54,55). However, short (<100 codons) active ORF tend to reside in transcripts that are considerably longer than those identified in this study (53,56).

The higher density of TUFs dwelling within or close to coding genes may reflect some aspect of their biogenesis or function. One distinct feature in *C. elegans* is that nearly half of the gene upstream TUFs were detected upstream of trans-splice acceptor sites or within operons, suggesting the possibility that some TUFs might represent outrons (57) resulting from trans-splicing. In yeast, CUTs are frequently associated with promoters of coding genes (41). In *Arabidopsis*, UNTs are collinear with the 5′ ends of known mRNAs and frequently extend into the first intron of respective overlapping genes. The possibility that UNTs derive from the pre-mRNA is highly improbable, as some UNTs are more abundant than the corresponding gene. Moreover, mapping of human transcriptome has also revealed an abundance of PASRs, possibly produced by pervasive or bidirectional transcription of promoter regions depleted of nucleosomes (6,10,14). The TASRs may be generated by similar mechanism (58). However, one caveat to such an interpretation of the data is provided by a previous analysis which have shown that intronic is-ncRNA loci in *C. elegans* commonly have independent promoter activity and show little or no expressional correlation to the protein-coding loci within which they reside (59).

The peculiar fact that the presence of a proximal TUF correlated with the conservation level of nearby coding gene is suggestive of some sort of functional relationship. Previous studies in yeast have found that ncRNA SRG1 could interfere with the promoter of downstream SER3 stress-responsive gene by blocking the binding of transcriptional factors (9). On the other hand, as the expression levels of most TUFs were not correlated with those of their respective proximal coding genes, and previous analyses have demonstrated a high degree of transcriptional independence even for *C. elegans* intronic is-ncRNAs (4,59), most TUF loci may well be transcriptionally and functionally independent of neighboring coding loci.

The novel TUFs differ from known is-ncRNAs in several aspects. Most of the previously known is-ncRNAs are well conserved across a wide range of organisms, whereas the novel TUFs show a dual conservation distribution, with approximately one-half of the TUFs being conserved within nematodes, and the rest apparently being specific to *C. elegans*. The failure to identify flanking sequence motifs suggests that most TUF loci do not have recognizable promoter and terminator sequences. This may result from most TUFs being by-products of the nearby transcriptional processes; however, a number of previously verified, but functionally uncharacterized is-ncRNA loci show a similar lack of canonical structures (4). Moreover, known is-ncRNAs such as snRNAs and snoRNAs are generally pervasively expressed through most developmental stages, whereas most TUFs showed fluctuant expression across stages. This tendency of novel ncRNAs to display stage-specific expression was also observed in recently published studies of the *C. elegans* transcriptome (60,61), corroborating the data in this study. The observation that a small number of TUFs appear to be present at all but one specific stage is curious. Though absence of evidence is not evidence of absence, it is tempting to speculate that there may exist a small number of transcripts whose presence is required for the 'default' state of the worm (i.e. rapid development toward the mature hermaphrodite), or alternatively, whose removal is required for the entrance into the ML or DU stages, respectively. Also, the average GC content of the TUFs was lower than that of known is-ncRNAs, and was closer to that of protein-coding exons. Thus, the previously known is-ncRNAs may, with some notable exceptions, be an assembly of highly and stably expressed loci that are not very representative of the expressional and functional complexity of intermediate-size transcriptome in *C. elegans*.

A substantial fraction of the TUFs showed no or nearly no conservation in other nematode species. However, lack of conservation does not necessarily imply lack of function, inasmuch as some of the genomic information (coding and non-coding) will be required to specify *C. elegans* as a distinguishable nematode species. With the exception of chromosomal location, the TUF phastCons score distribution was not related to any other TUF characteristic, thus TUFs with high and low phastCons scores were indistinguishable with respect to genomic localization, stage of expression and signal intensity distribution (Supplementary Figure S7). Consequently, the data suggest that there are substantial numbers of transcriptionally active genetic elements in *C. elegans* that are not conserved in other nematodes. These unconstrained TUFs may belong to a large pool of neutral elements that are biologically active but non-orthologous between nematodes (62), and it is possible that the non-conserved TUFs may play important roles in distinguishing *C. elegans* from other nematodes (41).

The functional roles of this large complement of novel loci can at present only be speculated. Many TUFs showed male-specific expression. Combined with the finding that the X chromosome was enriched for TUF loci, this strongly suggests involvement of these loci in sex determination or gender-specific functions. The lack of conservation outside the nematodes, and for a large fraction of the TUF loci, even within the sequenced nematode genomes, might suggest functional roles in specifying the nematode lineage, or even roles in distinguishing *C. elegans* from other nematodes. High numbers of non-conserved transcripts are also identified in several other organisms. In yeast, >80% of the 185 novel CUTs were produced from genomic regions with low conservation scores even among closely related yeast species (41), and in human, most of the detected unannotated transcribed sequences appear not to be strongly conserved in the mouse genome (63,64). On the other hand, since RNA function commonly depends more on secondary structure than primary sequence, it cannot be excluded that conserved (i.e. identical or similar) functions are executed by RNA loci that are no longer alignable. A significant number of human genomic regions not alignable to the mouse genome were found to have signatures of RNA structure and were twice as likely to overlap tiling array detected transfrags (18). A search for rapidly evolving sequences in the human linage identified a non-coding RNA with brain-specific expression (65), and a possible interpretation of the non-conserved TUFs is that they derive from genomic regions that exhibit recent evolutionary change. Mouse eRNAs has been hypothesized as an evolutionary source for new genes (66), and a fraction of the TUFs that are specific to *C. elegans* may well provide a warehouse of neutral elements available for further evolution.

ACCESSION NUMBER

Raw data and processed tables can be accessed from GEO by GSE24023.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dandan He for early discussion on experimental design and Jie Wang and Tao Liu for instructions in tiling array data analysis. The *C. elegans* strains used in this work was provided by the *Caenorhabditis* Genetics Center, which is funded by the National Institutes of Health—National Center for Research Resources.

FUNDING

Funding for open access charge: National Sciences Foundation of China (30630040); National Key Basic Research and Development Program 973 (2009CB825400, 2007CB946901, 2007CB935703).

Conflict of interest statement. None declared.

REFERENCES

1. Dinger,M.E., Pang,K.C., Mercer,T.R. and Mattick,J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, **4**, e1000176.
2. Severin,J., Waterhouse,A.M., Kawaji,H., Lassmann,T., van Nimwegen,E., Balwierz,P.J., de Hoon,M.J., Hume,D.A., Carninci,P., Hayashizaki,Y. *et al.* (2009) FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions. *Genome Biol.*, **10**, R39.
3. Huttenhofer,A., Kiefmann,M., Meier-Ewert,S., O'Brien,J., Lehrach,H., Bachellerie,J.P. and Brosius,J. (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, **20**, 2943–2953.
4. Deng,W., Zhu,X., Skogerbo,G., Zhao,Y., Fu,Z., Wang,Y., He,H., Cai,L., Sun,H., Liu,C. *et al.* (2006) Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis and expression. *Genome Res.*, **16**, 20–29.
5. He,H., Wang,J., Liu,T., Liu,X.S., Li,T., Wang,Y., Qian,Z., Zheng,H., Zhu,X., Wu,T. *et al.* (2007) Mapping the C. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Res.*, **17**, 1471–1477.
6. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes

and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

7. Chekanova,J.A., Gregory,B.D., Reverdatto,S.V., Chen,H., Kumar,R., Hooker,T., Yazaki,J., Li,P., Skiba,N., Peng,Q. *et al.* (2007) Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. *Cell*, **131**, 1340–1353.
8. Goodrich,J.A. and Kugel,J.F. (2006) Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.*, **7**, 612–616.
9. Martens,J.A., Laprade,L. and Winston,F. (2004) Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature*, **429**, 571–574.
10. Preker,P., Nielsen,J., Kammler,S., Lykke-Andersen,S., Christensen,M.S., Mapendano,C.K., Schierup,M.H. and Jensen,T.H. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, **322**, 1851–1854.
11. Yue,X., Schwartz,J.C., Chu,Y., Younger,S.T., Gagnon,K.T., Elbashir,S., Janowski,B.A. and Corey,D.R. (2010) Transcriptional regulation by small RNAs at sequences downstream from 3′ gene termini. *Nat. Chem. Biol.*, **6**, 621–629.
12. Kim,T.K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
13. van Bakel,H., Nislow,C., Blencowe,B.J. and Hughes,T.R. (2010) Most ‘dark matter’ transcripts are associated with known genes. *PLoS Biol.*, **8**, e1000371.
14. Core,L.J., Waterfall,J.J. and Lis,J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.
15. Xu,Z., Wei,W., Gagneur,J., Perocchi,F., Clauder-Munster,S., Camblong,J., Guffanti,E., Stutz,F., Huber,W. and Steinmetz,L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
16. Neil,H., Malabat,C., d’Aubenton-Carafa,Y., Xu,Z., Steinmetz,L.M. and Jacquier,A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, **457**, 1038–1042.
17. Pang,K.C., Frith,M.C. and Mattick,J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.*, **22**, 1–5.
18. Torarinsson,E., Sawera,M., Havgaard,J.H., Fredholm,M. and Gorodkin,J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
19. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
20. Louro,R., El-Jundi,T., Nakaya,H.I., Reis,E.M. and Verjovski-Almeida,S. (2008) Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics*, **92**, 18–25.
21. Aftab,M., He,H., Skogerbo,G. and Chen,R. (2008) Microarray analysis of ncRNA expression patterns in *Caenorhabditis elegans* after RNAi against snoRNA associated proteins. *BMC Genomics*, **9**, 278.
22. St Laurent,G. III and Wahlestedt,C. (2007) Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci.*, **30**, 612–621.
23. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase 2007. *Nucleic Acids Res.*, **36**, D612–D617.
24. Stricklin,S.L., Griffiths-Jones,S. and Eddy,S.R. (2005) *C. elegans* noncoding RNA genes. *WormBook*, 1–7.
25. Missal,K., Zhu,X., Rose,D., Deng,W., Skogerbo,G., Chen,R. and Stadler,P.F. (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *J. Exp. Zool. B Mol. Dev. Evol.*, **306**, 379–392.
26. Agarwal,A., Koppstein,D., Rozowsky,J., Sboner,A., Habegger,L., Hillier,L.W., Sasidharan,R., Reinke,V., Waterston,R.H. and Gerstein,M. (2010) Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, **11**, 383.
27. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
28. He,S., Liu,C., Skogerbo,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
29. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
30. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
31. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
32. Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.
33. Pang,K.C., Stephen,S., Dinger,M.E., Engstrom,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
34. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2007) Noncoding RNAs database (ncRNAdb). *Nucleic Acids Res.*, **35**, D162–D164.
35. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
36. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
37. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M. Jr, Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **32**, 4281–4296.
38. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc./Int. Conf. Intell. Syst. Mol. Biol. ISMB*, **2**, 28–36.
39. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
40. Encode Project Consortium,T. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
41. Lee,A., Hansen,K.D., Bullard,J., Dudoit,S. and Sherlock,G. (2008) Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genet.*, **4**, e1000299.
42. Thiebaut,M., Kisseleva-Romanova,E., Rougemaille,M., Boulay,J. and Libri,D. (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol. Cell*, **23**, 853–864.
43. Bradnam,K.R. and Korf,I. (2008) Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE*, **3**, e3093.
44. Mascarenhas,D., Mettler,I.J., Pierce,D.A. and Lowe,H.W. (1990) Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol. Biol.*, **15**, 913–920.
45. Rose,A.B. (2002) Requirements for intron-mediated enhancement of gene expression in *Arabidopsis*. *RNA*, **8**, 1444–1453.
46. Shi,B., Guo,X., Wu,T., Sheng,S., Wang,J., Skogerbo,G., Zhu,X. and Chen,R. (2009) Genome-scale identification of *Caenorhabditis elegans* regulatory elements by tiling-array mapping of DNase I hypersensitive sites. *BMC Genomics*, **10**, 92.
47. Li,T., He,H., Wang,Y., Zheng,H., Skogerbo,G. and Chen,R. (2008) *In vivo* analysis of *Caenorhabditis elegans* noncoding RNA promoter motifs. *BMC Mol. Biol.*, **9**, 71.
48. Thomas,J., Lea,K., Zucker-Aprison,E. and Blumenthal,T. (1990) The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **18**, 2633–2642.

49. Perez-Stable,C., Ayres,T.M. and Shen,C.K. (1984) Distinctive sequence organization and functional programming of an Alu repeat promoter. *Proc. Natl Acad. Sci. USA*, **81**, 5291–5295.
50. Surzycki,S.A. and Belknap,W.R. (2000) Repetitive-DNA elements are similarly distributed on Caenorhabditis elegans autosomes. *Proc. Natl Acad. Sci. USA*, **97**, 245–249.
51. Gingeras,T.R. (2007) Origin of phenotypes: genes and transcripts. *Genome Res.*, **17**, 682–690.
52. Rinn,J.L., Euskirchen,G., Bertone,P., Martone,R., Luscombe,N.M., Hartman,S., Harrison,P.M., Nelson,F.K., Miller,P., Gerstein,M. *et al.* (2003) The transcriptional activity of human Chromosome 22. *Genes Dev.*, **17**, 529–540.
53. Kondo,T., Plaza,S., Zanet,J., Benrabah,E., Valenti,P., Hashimoto,Y., Kobayashi,S., Payre,F. and Kageyama,Y. (2010) Small peptides switch the transcriptional activity of *Shavenbaby* during Drosophila embryogenesis. *Science*, **329**, 336–339.
54. Chooniedass-Kothari,S., Emberley,E., Hamedani,M.K., Troup,S., Wang,X., Czosnek,A., Hube,F., Mutawe,M., Watson,P.H. and Leygue,E. (2004) The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett.*, **566**, 43–47.
55. Kloc,M., Wilk,K., Vargas,D., Shirato,Y., Bilinski,S. and Etkin,L.D. (2005) Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development*, **132**, 3445–3457.
56. Frith,M.C., Forrest,A.R., Nourbakhsh,E., Pang,K.C., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y., Bailey,T.L. and Grimmond,S.M. (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet.*, **2**, e52.
57. Conrad,R., Thomas,J., Spieth,J. and Blumenthal,T. (1991) Insertion of part of an intron into the 5′ untranslated region of a *Caenorhabditis elegans* gene converts it into a trans-spliced gene. *Mol. Cell. Biol.*, **11**, 1921–1926.
58. van Bakel,H. and Hughes,T.R. (2009) Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genom. Proteom.*, **8**, 424–436.
59. He,H., Cai,L., Skogerbo,G., Deng,W., Liu,T., Zhu,X., Wang,Y., Jia,D., Zhang,Z., Tao,Y. *et al.* (2006) Profiling *Caenorhabditis elegans* non-coding RNA expression with a combined microarray. *Nucleic Acids Res.*, **34**, 2976–2983.
60. Lu,Z.J., Yip,K.Y., Wang,G., Shou,C., Hillier,L.W., Khurana,E., Agarwal,A., Auerbach,R., Rozowsky,J., Cheng,C. *et al.* (2011) Prediction and characterization of non-coding RNAs in *C. elegans* by integrating conservation, secondary structure and high throughput sequencing and array data. *Genome Res.*, doi:10.1101/gr.115758.110.
61. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE Project. *Science*, **330**, 1775–1787.
62. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
63. Kampa,D., Cheng,J., Kapranov,P., Yamanaka,M., Brubaker,S., Cawley,S., Drenkow,J., Piccolboni,A., Bekiranov,S., Helt,G. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, **14**, 331–342.
64. Bertone,P., Stolc,V., Royce,T.E., Rozowsky,J.S., Urban,A.E., Zhu,X., Rinn,J.L., Tongprasit,W., Samanta,M., Weissman,S. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
65. Pollard,K.S., Salama,S.R., Lambert,N., Lambot,M.A., Coppens,S., Pedersen,J.S., Katzman,S., King,B., Onodera,C., Siepel,A. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443**, 167–172.
66. Ren,B. (2010) Transcription: enhancers make non-coding RNA. *Nature*, **465**, 173–174.