Supplementary materials for:

Organisation of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis and expression

Wei Deng, Xiaopeng Zhu, Geir Skogerbø, Yi Zhao, Zhuo Fu, Yudong Wang, Housheng He, Lun Cai, Hong Sun, Changning Liu, Biao Li, Baoyan Bai, Jie Wang, Dong Jia, Shiwei Sun, Hang He, Yan Cui, Yu Wang, Dongbo Bu, Runsheng Chen

Contents

Supplementary figure 1

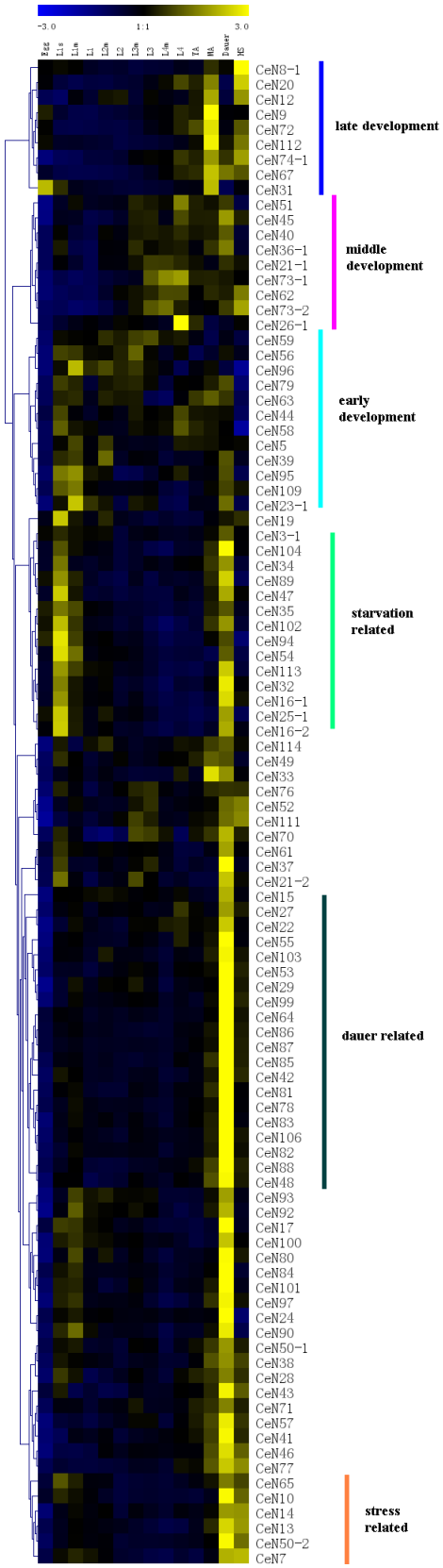


Fig. S-1 Expression patterns for 105 clustered ncRNAs (the figure includes both novel and known ncRNAs). L1s, starved overnight after hatch. L1m, m indicates the developmental midterm between the immediate preceding (L1s) and following (L1) stages. YA, Young Adult. MA, Mature adult. HS, Heat Shock treated (L4 worms at 30℃ 3 hours). See supplementary document for time intervals.

Supplementary figure 2

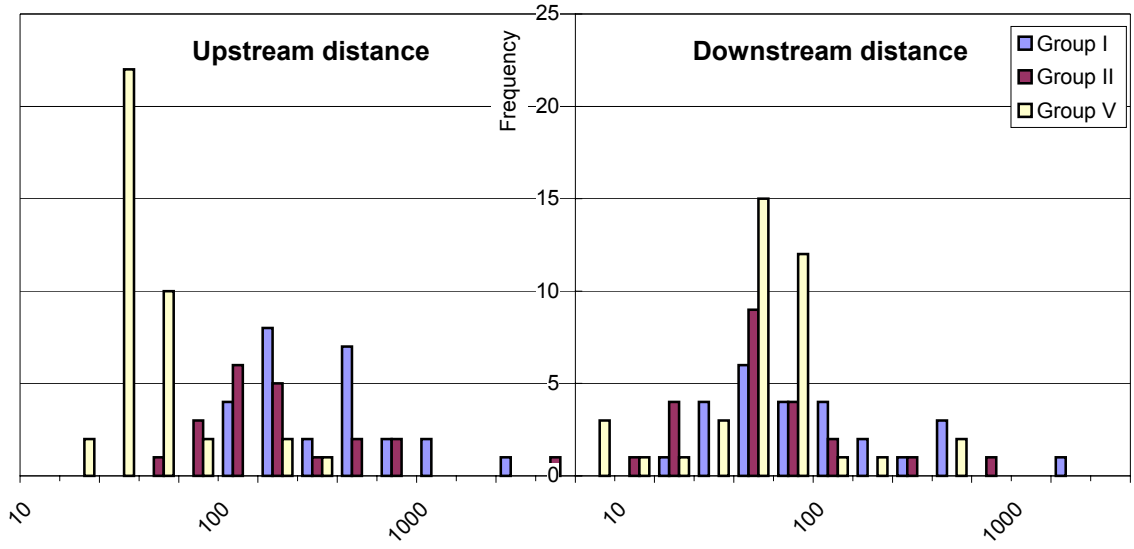


Fig. S-2. Distance (in bp) from intronic ncRNAs to adjacent upstream and downstream exons. The distances to the upstream exon from Group I and II loci ("motif loci") are significantly greater than from Group V loci ("non-motif loci"), suggesting that independently transcribed Group I and Group II loci require a upstream region of a minimum length to accommodate core promoter elements. Group V ncRNAs are transcribed from the host gene promoter, and excised from pre-mRNA or processed from intron lariats, thus having a lower requirement for an upstream sequence length. The size of the region downstream of intronic loci seems not correlated to the mode of biogenesis.

Supplementary figure 3

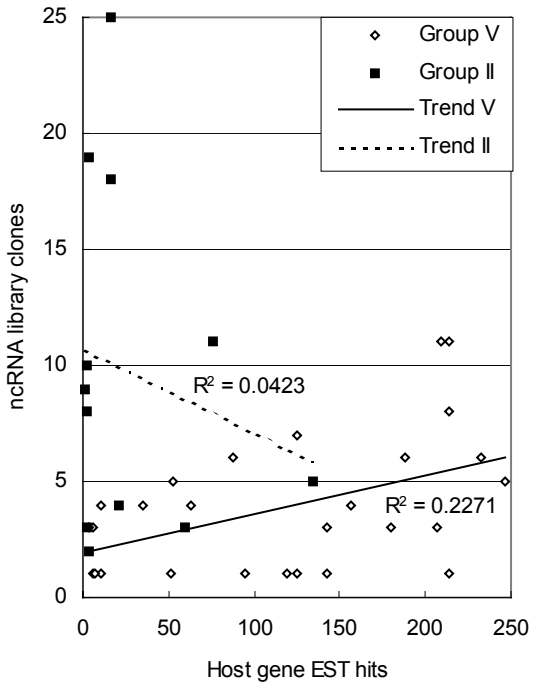


Fig. S-3. Relationship between intronic ncRNAs and host gene expression. The number of clones of each intronic ncRNA species in biogenesis groups II and V (see Supplementary table 1) were plotted against the average number of EST found for their respective host genes. For Group V ncRNAs, assumed to be excised and processed from pre-mRNA or spliced introns, there was a positive relationship between the number of host gene EST hits in WormBase (Harris et al., 2003) and the library clone number (R2 = 0.23). The low R2 value is not surprising given the type and size of data used for the plot, however it may be caused by incompletely intron-processed ncRNAs (thus accounting for samples with high host gene and low ncRNA values). No positive relationship was found for the assumedly independently transcribed Group II ncRNAs. Independently transcribed ncRNAs appear to have higher expression levels than dependently transcribed ncRNAs, whilst the expression of their host genes expression appear be lower than those of co-transcribed ncRNAs.

Supplementary document 1: Capping probability

The library construction procedure included a step to distinguish between 5' capped and non-capped RNAs. Prior to 5' end linker ligation, the ncRNAs were split into to aliquots, one treated with PolyNucleotide Kinase (PNK, Fermentas) to phosphorylate noncapped RNA, the other treated with Tobacco Acid Pyrophosphatase (TAP, Epicentre) to remove 5' end methyl-guanosine caps form capped RNA. This generated two types of libraries, TAP libraries containing mostly 5' capped ncRNAs, and PNK libraries with mostly non-capped ncRNA

For the most known and abundant capped ncRNAs (U1, U2, U4, U5), clone numbers were used to determine the probability that an capped ncRNAs be cloned in TAP or PNK libraries. Among 407 clones, 395 were present/found in the TAP library and 12 reads in the PNK library. For non-capped ncRNA, 5.8 S rRNA clones were used, which had a distribution of 148 and 21 clones in TAP and PNK libraries, respectively.

We also investigated why a number clones occurred in the "opposite" library, and found that almost all incorrectly cloned sequences were 5' truncated, probably brought about the in experimental steps. As we have no information on the exact 5' end structures of the novel ncRNAs other than observed differences between sequence,, we used all sequences reads to determine the probabilities of the cap status.

Table1. Distribution of clones/sequence reads of known capped and non-capped ncRNA in TAP and PNK library

	Including partial ncRNA reads		Ignoring partial ncRNA reads.	
	TAP library	PNK library	TAP library	PNK library
Capped	97% (395)	3% (12)	100% (175)	0
Non-capped	12% (21)	88% (148)	0	100% (114)

The probability of an ncRNA, with *t* clones in the TAP library and *p* clones in the PNK library, to either be capped [P(C|D)] or noncapped [P(N|D)] is determined by the following formula:

$$P(C|D) / P(N|D) = (0.97^{t} \times 0.03^{p}) / (0.12^{t} \times 0.88^{p})$$
$$P(C|D) + P(N|D) = 1$$

According to the calculated probabilities of all ncRNAs, we divided them into five groups, assigned with a value ranging from 1 to -1. These value is used in supplementary tables 2 for capping possibility annotation.

Probability		Assigned value
> 95% probability of being	capped	1
> 80% probability of being	capped	0.5
Information insufficient to determine cap status		0
> 80% probability of being	non-capped	-0.5
> 95% probability of being	non-capped	-1

For ncRNAs with only one sequenced clone and a possibility of greater than 95% to be capped, we assigned them to the >80% group to reduce risk of faulty determinations due to too few samples. We also remove reads from the estimate in cases of affirmed truncations.

Supplementary document 2: Genomic location of the novel ncRNA genes

The genomic locations of the novel small RNAs can be divided into several categories (tab. S-1). A majority of the novel loci (55%) are located in sense direction within a known or putative transcript of another gene, of which intronic loci constitute the larger part (51%). Among loci found within an intron, a considerable fraction (24% of all novel loci) is also located within a *C. elegans* operon. Two loci are found in either introns or UTR regions, depending on which frame is used for translation of the protein, and one locus (CeN42) covers the 3' end of an exon and the 5' end of the following intron.

The approximately 1/3 of loci found outside any known transcribed region are to a large extent located in relative vicinity to a protein coding gene (< 1 kb), whereas no locus is more than 10 kb away from a known or predicted transcribed gene. One locus is found in the antisense direction to a coding exon.

Table S-1. Distribution of novel small RNA genomic locations in C. elegans. It should be noted some RNAs may have several loci. All ncRNAs located in operons are found in introns (or intron + exon) of the operonic genes. The total number of genomic loci is therefore different both from the number of novel small RNAs, and from the sum of loci reported in the table. The "All" column refers to the loci of known plus novel ncRNAs detected in our screen.

	Number of ncRNA loci	
Genomic location	***Novel(%)***	***All (%)***
Within transcribed sequence (sense)	***55 (54.5)***	***90 (45.5)***
Intron	*51 (50.5)*	*85 (42.9)*
In intron or UTR (cen95, cen106)	*2 (2.0)*	*2 (1.0)*
Overlapping exon and intron (cen42)	*1 (1.0)*	*1 (0.5)*
In UTR (cen59)	*1 (1.0)*	*1 (0.5)*
Overlapping 5'UTR and the ORF (cen4.1)	*0 (0.0)*	*1 (0.5)*
In operon	*24 (23.8)*	*29 (14.6)*
Antisense to transcribed sequence	***10 (9.9)***	***25 (12.6)***
- to intron	*10 (9.9)*	*23 (11.6)*
- to exon (cen107.4)	*0 (0.0)*	*1 (0.5)*
- to UTR (cen4.17)	*0 (0.0)*	*1 (0.5)*
Intergenic	***36 (35.6)***	***83 (41.9)***
All	***101***	***198***

1. Intronic ncRNA genes

Intronically located novel ncRNA genes comprise 54 loci, located to 48 different known or predicted protein-coding genes. Altogether 21 different protein functional classes are represented among the genes hosting the intronic small RNAs, however, genes coding for ribosomal and ATP-binding proteins make up 30% and 19% of all functionally annotated genes respectively.

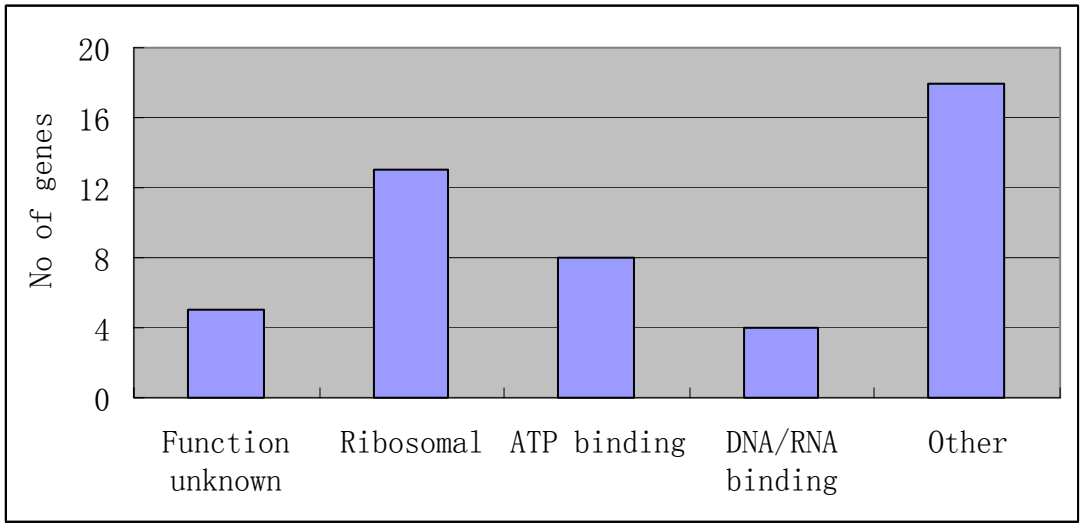


Figure 1. Functional class distribution of coding genes encoding an intronic small RNA transcript. "Other" are various functional classes represented by only one gene each.

1.1 CeN96

The RNA CeN96 represents a peculiar case of an intron located transcript. It is located to introns in two different genes, rpl-7A (Y24D9A.4a) and Y73B3A.18, on chromosomes 4 and X, respectively. The two CeN96 loci differ in 4 bp only, and overlapping parts of their host genes and the surrounding sequence show more than 85% homology (nearest 500 bp), most likely resulting from a chromosomal translocation. Rpl-7A on chr 4 encodes a ribosomal protein, whereas the function of Y73B3A.18 has not been experimentally established.

1.2. CeN42 spans an exon-intron border

CeN42 was first mistaken for a known transcript, as its sequence is partially covered by at least one EST. However, it turned out that this H/ACA snoRNA-like transcript derives from a loci which covers the 61 last basepairs of exon 4 in the putative (partially confirmed) protein kinase gene R166.5 (*1*), and extends 70 bp into the succeeding intron. The CeN42 locus is preceded by a UM2 element entirely embedded in R166.5 exon 4.

The CeN42 transcripts was represent by no less than different clones in our library, and confirmed by Northern blotting, thus it seems unlikely that it should represent a fragment from R166.5 pre-mRNA degradation.

1.3. RNA genes located in either intron or UTR

A few RNA loci fall within a part of a coding gene which is either a UTR or an intron, depending on splicing and/or frame usage of different forms of the protein. RNA CeN106 is located in intron 2 (sense) or in the 3' UTR of of ubl-1 ("Ubiquitin-like family protein 1, isoform a/b"; "H06I04.4a/b";), depending on whether form a or b is translated.

RNA CeN59 could be located either in the 5' UTR or in intron 5 of the hypothetical gene/protein Y71F9AM.4 (or .3). When the reading frame supplied by WormBase (*1*) is used, it would be intronic.

RNA CeN95 is similar to CeN59 in that it is either located in the 5' UTR or in the first intron of two alternative frame usages of the Y37E3.8 gene. WormBase (*1*) only gives the longer frame (.8a), which is the ribosomal protein L27, and in which CeN95 is intronic. RNA CeN88 is located in the (last) intron of the same gene.

2. Small novel ncRNA genes located in operons

The *C. elegans* genome is particular among eukaryotes in that a substantial fractions (15%) of its genes are located in operons, a feature it thus far only found in the *Oikopleura dioica* outside the bacterial realm (*2, 3*). Operon-located genes shared the common exon-intron mosaic found in most eukaryotic coding genes, however, when the operonic pre-mRNA is processed, an additional 22 nt leader sequence is added to the 5' end exon of all downstream genes before the introns are spliced out (*2*).

Of all novel small RNAs detected in this study, 32 have genes located in operons, all within an intron. A few operons contain several small ncRNAs, located in different introns of the same or different genes. Although genes of most functional classes are found in operons, *C. elegans* genes located in introns show a substantial bias towards ribosomal proteins and proteins involved in to RNA degradation (*2*). Out of 22 annotated operonic genes hosting one or more small RNA, 8 codes for ribosomal proteins, whereas the remaining are distributed on various different classes (fig. 2).

As ribosomal genes are highly overrepresented among operonic genes in *C. elegans* (*2*), it seems more likely that the small non-coding RNAs are located to operons due to their preference for ribosomal genes, than due to preference for location in operons *per se.*

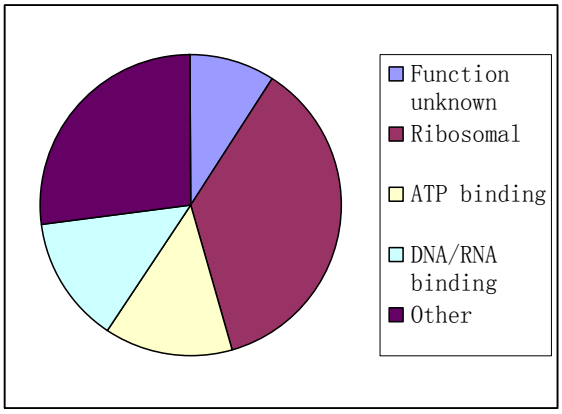


Figure 2. Functional distribution of operonic genes hosting one or more small RNA genes. "Other" are various functional classes represented by only one gene each.

3. RNAs of intergenic location

Thirty-nine small RNAs have loci outside known or predicted transcripts of other genes (tab. S1). Nearly 70% of these are located in relative proximity to (within 1 kb of) known or predicted coding genes, with an equal distribution of upstream and downstream partners. The remaining RNA loci are mostly within 5 kb from a coding gene, and no locus is more than 10 kb from a coding gene.

3.1 CeN53

RNA CeN53 appears to have two adjacent intergenic genomic loci on chromosome I. However, both are located within two approx 8 kb sequences with absolute sequence identity, including the entire or a large part of the nearest upstream coding gene. Most likely this represent an error during assembly of the *C. elegans* genome, if not, it might be a very recent duplication event. CeN53 has a near homologous sequence in the *C. briggsae* genome, however, the surrounding sequences are not conserved.

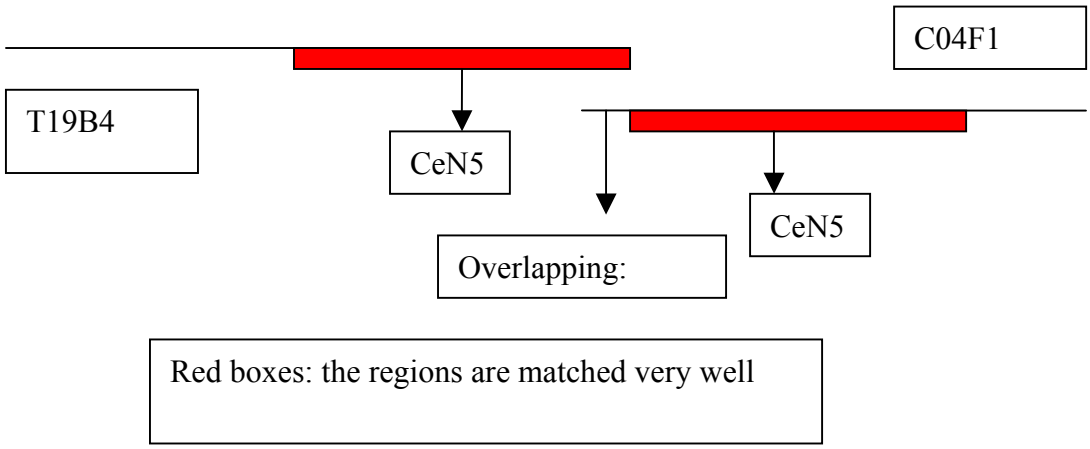


Figure 3. The two intergenic loci of CeN53. The chromosomal regional covering the nc20 loci is assembled from two BACs (T19B4 and C04F1), with an overlapping sequence of 200 bp. b. The sequences rendered in red cover two entirely identical 8.1 kb frements, however, the 200 bp overlapping sequence at the 5' end of the C04F1 BAC does not have a identical sequence at the 5' end of the 8.1 kb (red) sequence in the T19B4 BAC.

4. RNA CeN107-4 is located antisense to ORF

One RNA, CeN107-4, is located antisense to the last exon of the predicted gene B0284.1, overlapping codons 375-447 (out of 481), however the last approx 18 nt of the RNA matches the supposed locus poorly. The RNA sequence is conserved in *C.briggsae*.

1. *T. W. Harris et al., Nucl. Acids Res.* ***31****, 133 (January 1, 2003, 2003).*
2. *T. Blumenthal, K. S. Gleason, Nat Rev Genet* ***4****, 112 (Feb, 2003).*
3. *P. Ganot, T. Kallesoe, R. Reinhardt, D. Chourrout, E. M. Thompson, Mol Cell Biol* ***24****, 7795 (Sep, 2004).*

Supplementary document 3: ncRNA expression analysis

Total RNAs were extracted from worms at 15 different conditions: the mixed stages (freely cultured worms with all developmental representation), Dauer, Heat Shock treated (L4 worms at 30℃ 3 hours) and 12 different developmental stages. These were: Egg, L1s(starved overnight after hatch), L1m (m indicates the developmental midterm between the immediate preceding (L1s) and following stages (L1)), L1, L2m, L2, L3m, L3, L4m, L4, Young Adult (YA) and Mature adult (MA). All development stages were determined by time of culture since feeding of L1s worms at 20℃.

stage	L1s	L1m	L1	L2m	L2	L3m	L3	L4m	L4	YA	MA
time(hour)	0	8	13	18	22	27	32	38	44	56	82

Northern blotting were performed as described in the Method section. Blot signals were collected by an image system *ChemiCapt 3000 (Vilber, France)*, and then anaysized by *Quantity One version 4.3.1(Biorad, USA)*. Normalised intensities were then transformed into relative ncRNA contents (1.0 represents the expression abundance of U5 snRNA in 1ug total RNA sample from mixed-stage worms). Blots with strong background or faint signals or containing dubious alternative bands were removed from further expression analysis. In total, data from 106 ncRNAs, including 20 known ncRNAs were used for further analysis.

Standard Deviation (SD) and Coefficient of Variation (CV) of the expression levels, and mean expression level(E_m) of each ncRNA was calculated, CV_{median} being the median of all CVs.

$$SD = \mathrm{sqrt}(\sum(E_i - E_m)^2)$$

(E_i being the expression level at the *i*th condition (*i* is from 1 to 15).

$$CV = SD / E$$

A Z score was calculated as follows:

$$Z_i = (E_i - E_m)/\mathrm{MAX}\{SD, E_m * CV_{median}\}$$

If $|Z_i \geq 2|$, E_i was considered as significantly higher ($Z_i \geq 2$) or lower ($Z_i \leq 2$) expressed than the mean expression level (E_m).
If $|-1 < Z_i < 1|$ at all conditions, the expression level of the ncRNA was considered unvaried. Unvariedly expressed ncRNAs were not included in the cluster analysis.

To simplify creating of the image (Figure S-1) Z_i was used instead of E_i for hierarchical clustering analysis. This did not change the Pearson correlation values.

To further analyse the two ncRNAs highly expressed at both the egg and MA stages, mRNA expression data from 6 development stages from [1] were combined with ncRNA expression data ($\log_2(E/E_{mix})$)(E_{mix} was the expression levels from the mixed stage). Pearson correlation coefficients for each ncRNA-mRNA pair were calculated, and the top 200 mRNAs were selected, hoping that these 200 pairs might give some hints concerning the functions of the corresponding ncRNAs.

Supplementary document 4: Upstream and internal motifs of the *C. elegans* noncoding RNA loci

165 *C. elegans* noncoding RNAs corresponding to 198 loci in *C. elegans* genome were identified through out our experimental procedure. In order to detect possible conserved internal or upstream sequence features of these ncRNAs, were submitted the transcribed sequences, or their immediate 5' flanking regions, to analysis by the motif discovery software MEME [1].

Discovering the ncRNA Upstream Motifs

The first 100 bp upstreams of the transcription start site of 198 ncRNA genes were extracted from *C. elegans* genome from WormBase [2] and used as the input sequence set for MEME [1]. The parameters entered for MEME were "-dna -nmotifs 10". Some of the 10 motifs are trivial as they belong to only one single RNA family. However, 3 motifs were found upstream of more than 3 different RNAs or RNA families indicating they may be conserved upstream regulatory elements of several different noncoding RNAs. We labelled these UM1, UM2 and UM3. Further analyses provided indicated that these UMs are biological meaningful.
The basic information of UMs are shown in table 1.

Table 1. Statistical data of the upstream motif search. The table gives the motif size in bp (`width'), the total information content ('bits'), number of occurrences in the training set (`sites'), log likelihood ratio (`llr') and E-value of the three upstream motifs. According to the description of the MEME software [1], the statistical significance of a motif is based on its log likelihood ratio, its width and number of occurrences, and the background letter frequencies. The E-value is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and number of occurrences, that one would find in a similarly sized set of random sequences.

Motif	*E value*	*bits*	*width*	*sites*	*llr*
UM1	*4.0×10^{-521}*	*34.0*	*50*	*84*	*1978*
UM2	*7.3×10^{-179}*	*29.5*	*50*	*48*	*982*
UM3	*1.1×10^{-38}*	*53.3*	*50*	*9*	*333*

UM2 and tRNA

A search for additional UM2 motifs in the *C. elegans* genome reveal that a large fraction of the genomic UM2 hits overlapped with annotated tRNA and pseudo-tRNA genes (See SM section on ncRNA estimates). There are no tRNA annotations upstream of any of the 198 loci of our ncRNA, however, 4 pseudo-tRNA annotations (F19H8.t1,

D1046.t1, T08B2.t1, and Y59E1B.t1) upstream of four UM2-containing loci (cen62, cen55, cen43, and cen68, respectively).

In order to determine what may be the relationship between UM2 and tRNA, we extracted *C. elegans* tRNA sequences based on WormBase annotation [2], and applied the MEME motif discovery procedure on the tRNA data set. The tRNA motif profile showed a certain degree of similarity to the UM2 profile, in particular did the box A and box B motifs of the internal tRNA promoter appear to quite similar to the 5' and 3' ends of UM2, respectively.

We therefore asked MEME to find the two most conserved 15 bp motifs within the tRNA dataset of 387 sequences. The two motifs produced largely correspond to the *C. elegans* tRNA box A and box B. We then applied the same procedure first 100 bp upstream of the 47 UM2-containing loci. The search resulted in one motif present at 47 loci, which generally overlaps the first 15-16 bp of UM2 (therefore labelled the Front Box, or FB), and another motif, present at 35 loci, which stretches from bp 41 UM2 and beyond its 3' end (labelled the Tail Box, TB). Statistics concerning the MEME search is found in table 2.

Table 2. Statistics summary for the 15 bp motif MEME searches in the tRNA and UM2-containing loci data sets. (See tab. 1 for explanations).

Motif	*E value*	*bits*	*width*	*sites*	*llr*
tRNA "box A"	3.3×10^{-852}	*14.7*	*15*	*364 out of 387*	*3699*
tRNA "box B"	3.5×10^{-1312}	*18.8*	*15*	*365 out of 387*	*4754*
UM2 FB	6.7×10^{-78}	*14.5*	*15*	*47 out of 47*	*471*
UM2 TB	1.4×10^{-50}	*15.3*	*15*	*35 out of 47*	*371*

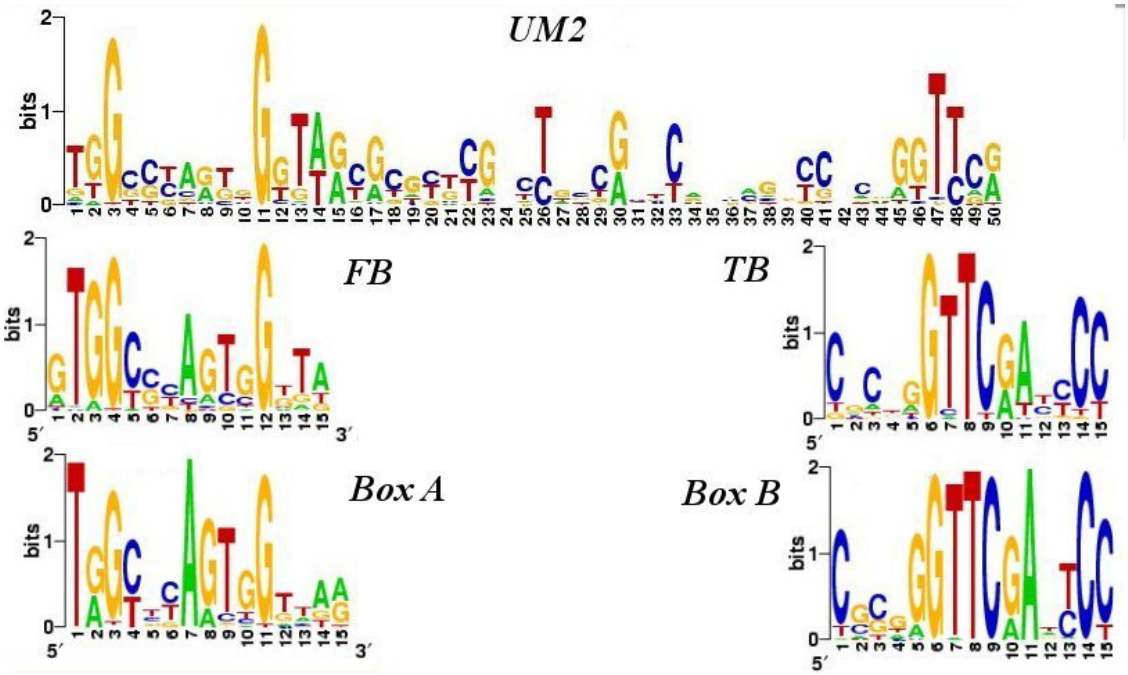


Figure 1. The 15 bp motifs found in tRNAs and upstream of UM2-containing loci. The two short motifs Front Box (FB) and Tail Box (TB) found upstream of UM2-containing loci are compared to the original UM2, and to the two 15 bp motifs located from the tRNA set. The latter two are labelled "box A" and "box B" to denote that they cover the tRNA box A and B promoter motifs. (The actual box A and B motifs correspond to bp 2-11 in "Box A" and bp 4-15 in "Box B" [3].)

The similarity between the tRNA box A and box B and the UM2 FB and TB is striking (fig. 1). It thus seems likely that UM2 corresponds to pseudo-tRNA genes located in front of, and serving as promoters for, snoRNA (and other ncRNAs) genes. A similar arranged of ncRNA genes has been observed in *Arabidopsis*, where a dicistronic primary transcript consisting of an $tRNA^{Gly}$ and a snoRNA is cleaved by the tRNA 3' end-processing enzyme RNA Z, releasing the snoRNA from the tRNA [4]. If a similar arrangement of tRNA and snoRNA genes also exists in *C. elegans*, this could possibly be a very old evolutionary solution to the transcription of snoRNAs genes, dating back to before the divergence of plants and animals. We have, however, no indication of the existence of such a dicistronic primary transcript in our data. Though this may very well be because our experimental protocol was not design to pick up a possibly short-lived dicistronic transcript , an alternative possibility could therefore be that the originally internal tRNA promoter has been transformed to act as a (non-transcribed) upstream core promoter, perhaps utilising a TFIIIC-like transcription factor for recruitment of TFIIIB and polymerase III.

Discovery of ncRNA internal motifs

For the internal motif discovery, the input sequence set was all detected ncRNA transcripts, with the exception that for each family (consisting of nearly identical transcripts), one sequence was

chosen to represent the entire family. The results produced three conserved motifs (tab. 2), of which two (Internal Motif 1 (IM1) and IM2) were found at the 5' and 3' ends of 9 transcripts, and the third (IM3) found in another set of 8 transcripts.

Table 3. Statistics summary for the internal motif search. (See tab. 1 for explanations).

Motif	*E value*	*bits*	*width*	*sites*	*llr*
IM1	*1.2×10^{-19}*	*38.4*	*23*	*7*	*186*
IM2	*1.0×10^{-20}*	*38.9*	*28*	*8*	*216*
IM3	*1.8×10^{-37}*	*40.2*	*37*	*12*	*334*

1. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 1994: p. 28-36.
2. Harris, T.W., et al., *WormBase: a cross-species database for comparative genomics*. Nucl. Acids Res., 2003. **31**(1): p. 133-137. Release W130, Oct. 2004.
3. Ciliberto, G., et al., *Promoter of a Eukaryotic tRNAPro Gene is Composed of Three Noncontiguous Regions.* PNAS, 1982. **79**(4): p. 1195-1199.
4. Kruszka, K., et al., *Plant dicistronic tRNA-snoRNA genes: a new mode of expression of the small nucleolar RNAs processed by RNase Z.* EMBO J., 2003. **22**(3): p. 621-632.

Supplementary document 5: *C. elegans* ncRNA biogenesis groups

Based on their upstream motif, genomic location, 5' end structure and 3' termini, as well as three conserved internal motifs, we suggest the 161 known and novel ncRNAs can be divided into least seven different groups with respect to biogenesis and function (Tab. S-1). The first five groups (I-A, I-B, II, III and IV) are predominantly determined by core promoter structure, and contain ncRNAs with both intronic and intergenic loci. The remaining two groups lack discernible upstream elements, and are grouped according to genomic location and other available information. Group V is made up entirely made up snoRNAs with intronic loci, whereas group VI mainly consists of a smaller number of ncRNAs with intergenic loci, and for which transcription systems is not immediately evident.

Group I-A includes 51 transcripts with a UM1 sequence upstream of their genomic loci. These ncRNAs appear all to carry a 5' methyl-guanosine cap, and include known RNA polymerase II transcripts like the U1, U4 and U5 snRNAs, along with 11 C/D snoRNAs and 17 novel ncRNAs of unknown function. The absence of TATA box (Hernandez 2001; McNamara-Schroeder et al. 2001), and the presences of 5' caps, strongly suggests this group is composed of unprocessed RNA polymerase II transcripts. The majority of UM1 loci were intergenic, however 26 intronic loci also had this upstream motif. A subset of 23 UM1 loci had an additional conserved upstream motif (UM1A) located directly upstream of UM1, at approximately -80 bp (Fig S-3).

Four groups (I-B, II, III and V-B) contain known or likely RNA polymerase III transcripts. These ncRNAs generally have a low frequency of capped transcript, and a high frequency of polymerase III terminator signals. Groups I-B, II and III loci contain upstream motifs UM1, UM2 and UM3, respectively. Loci in groups I-B and III all have a TATA box in addition to their upstream motifs. Group I-B includes U6 snRNA, which in our data showed strong indications of carrying a 5' end cap, probably implying that the γ-monomethyl-GTP cap found on human U6 snRNA (Gupta et al. 1990) is also post-transcriptionally added to *C. elegans* U6 snRNA.

Forty-seven ncRNAs make up group II, most of which (39) have snoRNA-like characteristics. All share the same Upstream Motif 2 (UM2) at their genomic loci. None could be assigned a 5' end cap with any certainty, and 87% of their loci have a canonical RNA polymerase III terminator sequence at their 3' termini, strongly indicating transcription by RNA polymerase III.

Group III ncRNAs are characterised by Upstream Motif 3 (UM3) at their genomic loci, and comprise the 9 sbRNAs. Their loci are, with one exception, all intergenic, and are

frequently found in small clusters of two and three. The loci invariantly contain a TATA box preceded by a conserved G residue. The presence of a TATA box, and the apparent lack of a 5' end cap on all but two of transcripts is indicative of RNA polymerase III transcription. Group IV is made up of four SRP RNAs most likely transcribed from both internal and upstream promoter elements. Their loci show an upstream TATA element and a tRNA type A-block, resembling the *Schizosaccharomyces pombe* type of SRP RNA promoter (Rodicker et al. 1999).

Group V are intronically located snoRNA-like genes with no apparent upstream motifs. These genes are most likely transcribed from the host gene promoter, and either excised from pre-mRNA or processed from intron lariats after splicing of the host mRNA (de Turris et al. 2004). More than 2/3 of group V snoRNAs are potential H/ACA snoRNAs, whereas a similar dominance of C/D snoRNAs are found in other biogenesis groups. The sequence separating the group IV ncRNA from the preceding exon is generally more AT-rich than any of the conserved upstream motifs. The average distance between the ncRNA 5' end and the preceding exon is also only 44 nt, far less than for intronic group A (373 nt) and C loci (204 nt), and hardly sufficient to accommodate for a eukaryotic core promoter (Tab. S-1). More than 60% of the loci reside in ribosomal or other translational related genes, known in vertebrates to have the TOP type promoter, important for regulating the rate of mRNA to snoRNA synthesis in genes hosting intronic snoRNAs (de Turris et al. 2004). None appeared likely to carry a 5' end cap. Relating the expression levels of group V snoRNAs with the frequencies of EST corresponding to exons of their host genes produced a distinct positive correlation not found for intronic ncRNAs in group II (Fig. 4 in text). On the other hand, relating the expression profiles of the group V snoRNAs to the expression profiles of their host genes, produced strong negative correlations the developmental timing of expression of most snoRNAs relative to their respective host gene mRNAs, suggesting that the balance between processing of the host pre-mRNA transcripts to mature, spliced mRNAs or direct excision of the intronic snoRNAs may also be under regulation (de Turris et al. 2004). A further indication in this direction is that many intronic group V snoRNA-like transcript showed substantial variation in expression through development, with particularly high expression in the dauer state. Whereas we cannot exclude the possibility that the high dauer expression levels could be an artefact of the experimental procedure (e.g. due to a reduced rRNA fraction in RNA extracts from dauer worms), the fact the none of the five constitutively expressed spliceosomal snRNAs displayed altered expression in dauer worms indicates that the data reflects physiologically relevant levels of these ncRNAs. Thus, the elevated expression levels of the ncRNAs in the dauer state may represent altered activation of their host genes, or differential processing of pre-mRNAs to snoRNAs rather than to mRNAs (de Turris et al. 2004).

Group VI is composed of ncRNAs with intergenic loci for which no discernible upstream motifs could be found. The group comprises six SL2 RNAs and two snlRNAs

(probably transcribed by RNA polymerase II), and two C/D snoRNA-like RNAs for which no transcription system is suggested.

Supplementary table 1

Tab. S-1. ncRNA biogenesis groups.

Group	No. of ncRNAs	Upstream motif(s) (position)	Capped ncRNAs* (%)	Pol-III terminator$ (%)	Intronic/ intergenic loci	Distance to 5 exon! (bp)	TR host genes& (%)	Suggested RNA polymerase	Remarks /Known ncRNA classes
I-A	51	UM1 (-81)	84	13	20 / 51	287	26	II	U1/2/4/5 snRNAs, SL RNAs, C/D box snoRNAs
I-B	4	UM1 (-89) +TATA	25	100	8 / 12	672	20	III (II?)	RNase P RNA, Y RNA, U6 snRNA
II	47	UM2	2	87	22 / 26	204	15	III	SnoRNAs
III	9	UM3 +TATA	11	78	1 / 8	-	-	III	Novel ncRNAs with IM1 and IM2
IV	4	TATA + int. box A	0	100	1/3	304	0	III	All SRP RNAs
V	36	None	0	31	36 / 0	39	69	P#	All are snoRNAs
VI	10	None	80	10	0 / 10	-	-	Undecided	SL RNAs, 2 snoRNAs

* ncRNAs with more than 95% possibility of being capped.

& TR host genes – Translationally related host genes.

P – Processed from pre-mRNA or intron lariats rather than independantly transcribed.

$ A stretch of at least four consequtive Ts at +/- 10 bp of the 3' terminus of ncRNA transcript.

! Apply only to intronic ncRNA members of each group.

The group I-A, I-B, II and III ncRNAs were clustered according to their respective upstream motifs, and group I-A and I-B were further subdivided according to UM1 position and TATA box. Group IV are SRP RNAs which transcribed from SRP specific transcription promotor. Group V are transcripts with intronic loci without distinct upstream motifs. Group VI all have intergenic loci with no distinct upstream motif.

Supplementary document 6: Estimates of the *C. elegans* small non-coding RNA number

1. Estimate based on intron conservation

Based on the observation that the conservation level of ncRNAs is different from (non-protein coding) non-ncRNA sequences, we used the WABA algorithm[1] to compare the sequence conservation of ncRNA containing introns in *C. elegans* and *C. briggsae* to that of all introns (fig 1). The conservation level of each individual intron was calculated as "waba-strong sequence/total intron sequence", introns with no waba-strong sequence being set to zero.

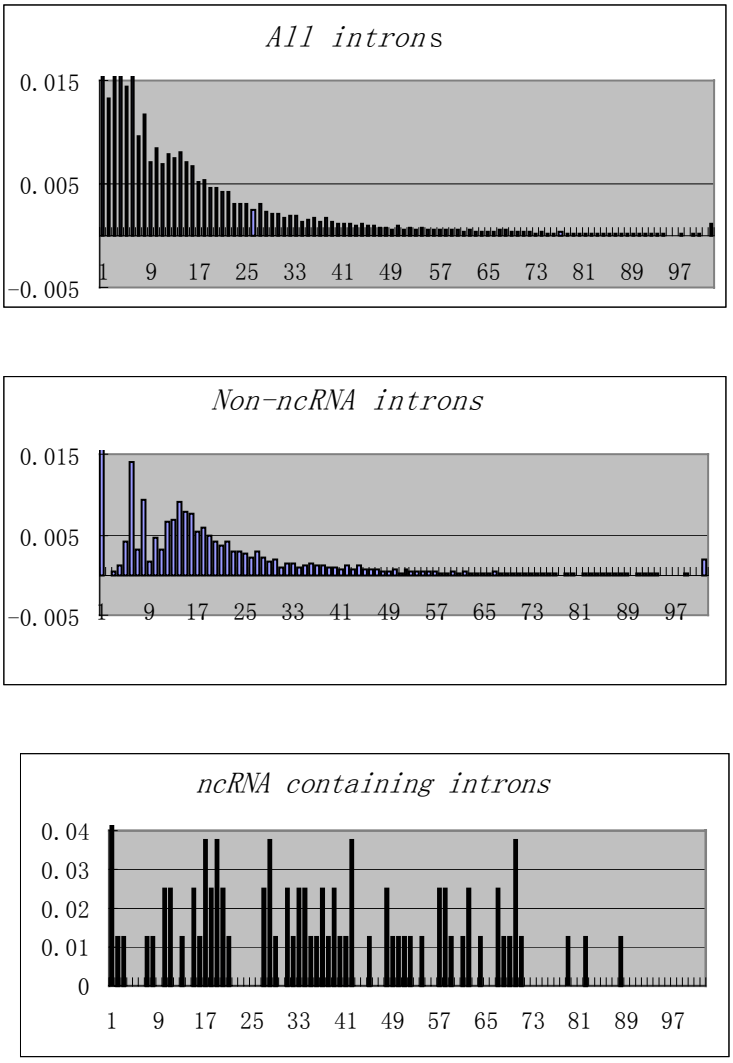


Figure 1. Distribution of C.elegans introns conserved in C. briggsae. Upper: All introns. Middle: Introns not containing ncRNA loci. Lower: Introns containing ncRNA loci.

X-axis: % waba strong sequence, Y-axis: Frequency

As most ncRNAs are longer than 50 bp, we initially took introns shorter than 50 bp to represent non-ncRNA containing introns. Assuming that the distribution of all (known and unknown) ncRNA-containing introns is similar to that of known ncRNA containing introns, we used linear regression analysis to estimate the (total) percentage of ncRNA containing introns in *C. elegans*, yielding a preliminary estimate of 4.1%. However, not knowing to what extent these short (< 60 bp) introns were representative for the longer non-ncRNA containing intron population, we repeated the procedure, taking introns within each 10 bp interval from 30 to 130 bp (no ncRNA-containing intron is shorter than 130 bp) as estimates of the non-ncRNAs intron population (fig. 2). As is evident from the figure, most intron intervals below 140 bp give estimates of 1.5-2.5%. A notable exception is the interval 40-50 bp, which gives a percentage twice that of the average of the remaining <140 bp intervals. Introns in this size interval are exceptionally many, close to 35.000, as opposed to 1-2000 for most other intervals (tab. 1). To avoid undue influence from the shortest fraction of the intron population, we therefore carried out separate calculations with introns below 50 bp either included or excluded.

Table 1. Intron size distribution

Intron size (bp)	***No of introns***
<30	*42*
30 - 40	*300*
40 - 50	*34698*
50 - 60	*14213*
60 - 70	*3410*
70 - 80	*2397*
80 - 90	*1785*
90 - 100	*1493*
100 - 110	*1405*
110 - 120	*1295*
120 - 130	*1113*
130 - 1000	*33046*
1000 - 2000	*4643*
>2000	*1482*
All introns	*101322*

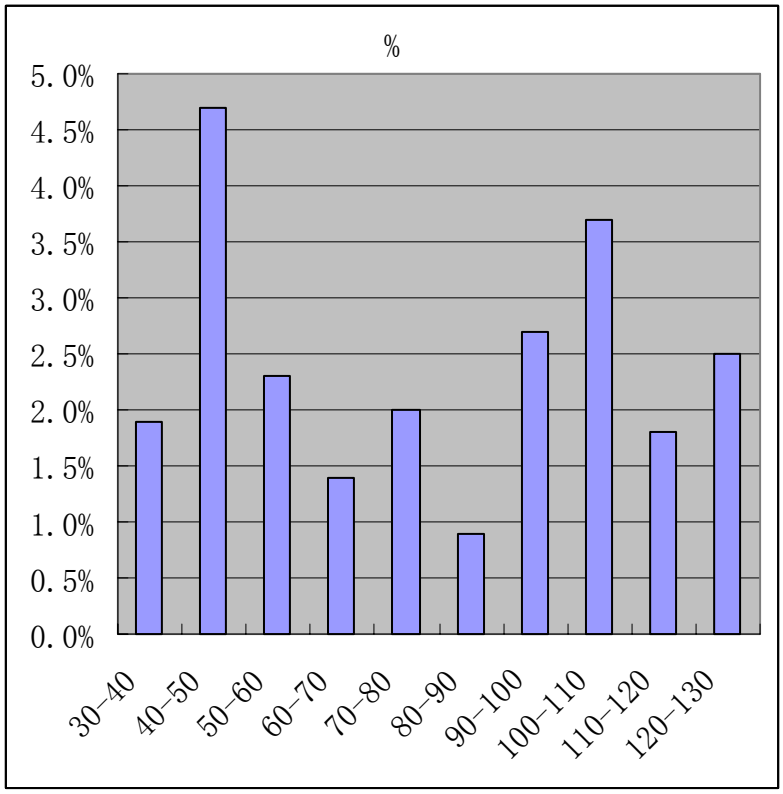


Figure 2. Estimates of the percentage of ncRNA-containing intron when using introns within different length intervals as representative of non-ncRNA containing introns.

Of the 99 introns containing an ncRNA, 89 are shorter than 1000 bp. Moreover, our definition of conservation level is not appropriate for large introns (since the large introns may have a considerable amount of conserved sequence, but the calculated conservation level would still be low. We therefore let the 130 – 1000 bp intron set represent the total set (henceforth TOTAL set). Binning of the conservation level had strong influence on the result, however, the variations leveld out with increasingly large bins, and we therefore used 1/3 (33.33%) as bin size.

Using the linear regression method [2] to deduce the percentage of ncRNA containing introns in the TOTAL set. Let x be the conservation level, $f(x)$ be the function of the conservation distribution of TOTAL set, $f1(x)$ of ncRNA containing intron set, and $f2(x)$ of the non-ncRNA containing intron set. b_0 is the percentage of ncRNA containing introns in the TOTAL set. We then get

$$f(x) = f1(x)b + f2(x)(1 - b)$$

$$b_0 = \arg\min_b \left(\int [f(x) - f1(x)b + f2(x)(1 - b)]^2 dx \right)$$

We used three different intron set to represent non-ncRNA containing introns, SHORT (40-50 bp), LONG (50-130 bp) and ALL (SHORT + LONG), to calculate frequencies of ncRNA containing introns and estimates of the entire *C. elegans* small ncRNA population. Adjusting for the fraction of ncRNA loci in introns larger than 1000 bp (10 out of 99), the ncRNA/intronic loci-ratio (90/99), and the fraction of of non-intronic (eg. intergenic) ncRNAs, we obtained three estimates of the *C. elegans* small non-coding transcriptome (tab 2).

Table 2. Estimates of small ncRNA number in the C. elegans genome. For definitions of SHORT, LONG, ALL, and TOTAL, see text.

Non-ncRNA intron set	*% of TOTAL*	*No of ncRNA introns in TOTAL*	*Total no of ncRNA introns*	*No of ncRNAs with intronic loci*	*% of all introns*	*No ncRNAs in genome*
SHORT	*3.92%*	*2358*	*2623*	*2385*	*2.4%*	*4062*
LONG	*2.74%*	*905*	*1007*	*916*	*0.9%*	*1560*
ALL	*4.08%*	*1348*	*1500*	*1363*	*1.3%*	*2322*
All introns					*101322*	
TOTAL (130- 1000 bp)					*33046*	
Fraction of ncRNA containing introns in TOTAL					*0.90*	
ncRNAs/loci					*0.91*	
Fraction of intergenic ncRNAs					*0.59*	

2. Estimate based on conserved upstream motifs

In order to obtain an estimate the ncRNA numbers in *C. elegans*, we attempted to count the number of occurrences of UM1-3 in the genome. To accomplish this we used the weight matrices computed by MEME [3] as input to the program mhmm (part of the Meta-MEME [4] software) in order to generate a Hidden Markov Model for each upstream motif. We then searched these HMM profiles in a *C. elegans* masked genome [5] using mhmmscan (also part of Meta-MEME). We chose the E-value threshold as 0.1 because most (over 90%, or 129 out of 139) of our proved ncRNA upstream motifs could be identified under this threshold. 1404 genomic sites were reported as UM1 candidates, 527 as UM2 candidates, and 65 as UM3 candidates. Among these, we identified 75 out of 82 of our previously detected UM1 sequences, 39 of out 48 UM2, and 8 out of 9 UM3. (The parameters of Meta-MEME were as follows: mhmm -motif x -type star MemeResultFile; mhmmscan motifx.mhmm worm.masked.dna).

<u>UM1 candidates and a possible transcript repeat</u>

We used a repeat masked genome for the search since mhmmscan tend to give tandem

repeat regions a high score in genome scale scans. We nevertheless found that a *C. elegans* repeat (Ce000293) contained UM1. During our cloning work, we once identified one clone which appeared to be a combination repeat Ce000293 and another repeat. However, as we only found this one clone, and were unable to determine transcript 5' and 3' with any certainty, it was not included in our set of verified novel ncRNAs. There are 314 Ce000293 repeats in *C. elegans* genome but these have not been included in our UM1 count, as the search was done on a repeat masked genome.

Intronic and exonic UM hits

A fraction of each of the ncRNA loci having either of the three upstream motifs are found in introns of protein coding genes, whereas as only three are found overlapping exons in sense directions. Of these three, one (Cen59/UM2) is located entirely in the non-translated part of an UTR, whereas the two others (Cen4.2[snRNA U6]/UM1 and Cen42/UM2) covers parts of translated exons. Thus, the data suggests (as do common sense) that few genuine UM hits should be found covering exons, whereas as considerable fraction could be expected to co-located with introns. We therefore checked to what extent the genomic hits overlapped introns and exons of protein coding genes (tab 3). As seen from the table, "intronic hits" are within the limits of what would be expected with a reasonable error rate, but both UM2 and UM3 have unreasonably high percentages of hits in exons.

Table 3. Distribution of verified loci and genomic hits for upstream motifs 1-3 (UM1-3).

	Verified loci (%)		*Genomic hits*		
Motif	*In intron*	*In exon*	*In intron (%)*	*In exon (%)*	*Total no of hits*
UM1	*31.7*	*1.2*	*54.1*	*7.1*	*1404*
UM2	*45.8*	*4.2*	*40.4*	*40.4*	*527*
UM3	*11.1*	*0.0*	*30.8*	*33.4*	*65*

UM2 and tRNA

Among the 527 UM2 candidates, 254 are tRNA primary transcript according to the WormBase annotation. Additional searches with alternative software (in progress) also indicate a relatively high number (around 2000) of UM2-like sequences, of which about 50% overlap with either tRNA genes or pseudo-tRNA genes (see SM document on "**Upstream Motifs at *C. elegans* noncoding RNA loci**" for details).

Based on the above we conclude that the sets of UM2 and UM3 candidates are unlikely to yield a reliable estimate of the total ncRNA population in *C. elegans*. UM2 is found to overlap considerably with both tRNA and pseudo-tRNA loci (the fraction of

pseudo-tRNA loci being difficult to estimate), and the genomic hits have an unreasonably high tendency of falling within coding exons. UM3 candidates also show a high percentage of exonic locations, and the number of UM3 loci (verified and candidate) is so low that an estimate of the total ncRNA number would be heavily influence by any error in the UM3 candidate figure. Therefore, whereas we do believe that a number of the UM2 and UM3 candidate loci will eventually turn out be actual ncRNA loci, we do not think that the UM2 and UM3 hit numbers are suited for an estimate of the total number of ncRNAs in *C. elegans*.

An estimate based on UM1 seems more likely to give a reasonable reflection of the size of *C. elegans* ncRNA population. UM1 is well conserved and includes a very likely promoter element (the snRNA PSE). It is occurs rather frequently (at 82 of 198 verified ncRNA loci), and apart from one UM1 containing repeat, which is easily removed by repeat masking, we have not found any other confounding factors that might strongly bias an estimate. The genomic search also gave a low number of hits overlapping coding exons, indicating that the candidate UM1 hits mostly occur at probable genomic locations. We have therefore based an estimate of the size of the total ncRNA population in *C. elegans* on the frequency of UM1 hits in genome (tab 4).

.

Table 4. Estimates of ncRNA loci and number of different ncRNAs in the C. elegans *genome based on the occurrence of upstream motifs UM1, UM2 and UM3. (Loci/ncRNA is calculated as all ncRNA loci corresponding to all ncRNAs detected by our screen, i.e. 198/161).*

	No of ncRNAs	*No of ncRNA loci*	*% of all loci*	*Loci/ ncRNA*	*No of genomic hits (E < 0.1)*	*Estimated no of ncRNA loci*	*Estimated no of different ncRNAs*
UM1	*54*	*82*	*41.4%*	*1.23*	*1404*	*3390*	*2757*
Total no of loci		*198*					

3. Estimate based on clone number and quantitative Northern blot data

The model relates to two kinds of data. One is the *number of library clones* found for a single ncRNA species The other is the *concentration* of each ncRNA as determined by quantitative Northern blots. The estimates are based on two assumptions:

1. For an ncRNA, the clone number is related to its concentration in the total RNA sample (i.e. the concentration of the ncRNA in *C. elegans*).
2. For ncRNAs of a given concentration (i.e. within a given concentration interval as estimated by Northern data), the clones numbers as obtained by multiple samplings, are normally distributed.

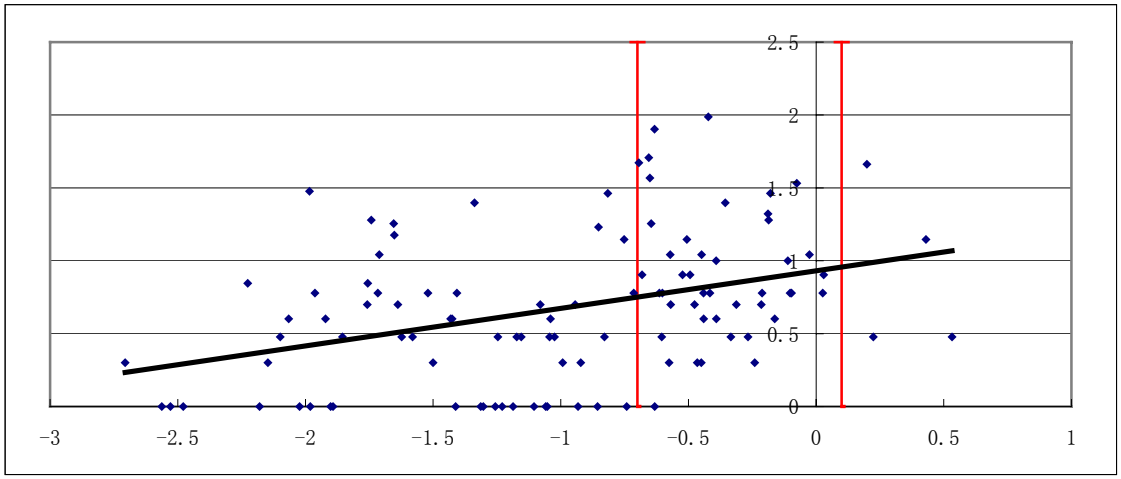
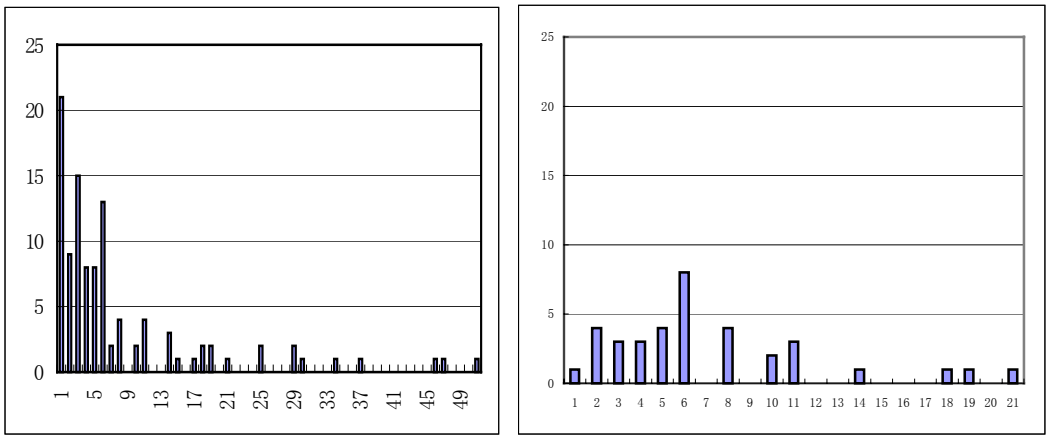


Figure 3. Dots distributions of ncRNA clones in relation to concentration (Y-axis: log Clone number, X-axis: log ncRNA concentration (Northern))

Concerning assumption 1: As shown in Figure 3, ncRNA showing higher concentrations in Northern tend to occur as several clones. At lower concentrations, ncRNAs are most likely to be picked randomly, and just one or very few clones have been sequenced.
Concerning assumption 2: Histograms displaying the distribution of total ncRNA clones are shown in figure 4 A. At the highest (log) concentration range [-0.7 to +0.1] (marked by red lines in figure 3), the distribution of ncRNA reads are similar to a normal distribution (Fig. 4B).



A. All ncRNAs B. Conc. range -0.7 to 0.1

Figure 4 A: Histogram distribution of all ncRNA clones. B: Histogram distribution of the ncRNA clone within the (log) concentration range -0.7 to 0.1. (Y-axis: Frequency; X-axis: Clone number)

If we take a log concentration of –0.5 as a "standard" level, we can, according to assumption 1, calculate the expected number of clones for all ncRNA at this standard level. Figure 5 shows the distribution of expected ncRNAs clones after this

normalisation. The normalisation yields a total of 5781 expected clones, with a mean frequency of 2 clones per ncRNA (fig. 5). As this represents a multinomial distribution [6], the number of different ncRNA species equals the number of samples (i.e. expected clones) divided by the mean frequency. Thus, the total number of ncRNA species in the *C. elegans* ncRNA population equals 5781 x 1/2 = **2936.**

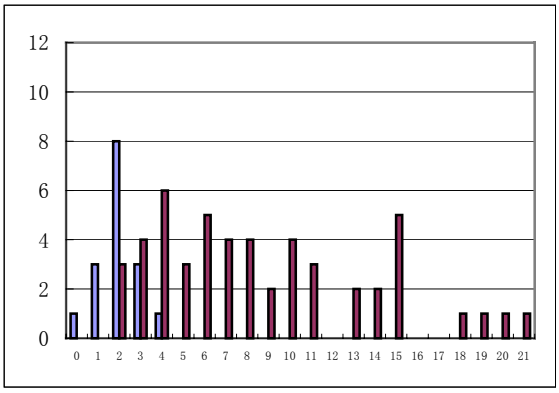


Figure 5. Distribution of -0.5 normalized ncRNA clones. Blue and purple bars represent reads with original concentrations lower and higher than the "standard" level, respectively.

4. Summary of the three estimates

The three different estimates all arrive at a number of different ncRNA in the lower thousands, with an average around 2700 (tab. 5). Most spread is found within model 1 (Intron conservation), however, even if all three sub-estimates of this model are taken into account, the average based on all three models falls between 2400 and 3300. Despite obvious inherent weaknesses of each individual model, we think that the fact that they all arrive at such similar figures suggest at an estimate of 2700 ncRNA species in *C. elegans* may not be too far off the mark.

Table 5. Estimates of the ncRNA transcriptome.

Model		*Estimated no of ncRNAs*
1 Intron conservation		
SHORT non-ncRNA set	*4100*	
ALL non-ncRNA set	*2385*	*2385*
LONG non-ncRNA set	*1600*	
2. Conserved upstream motif (UM1)		*2757*
3. Clone no. vs expression level		*2936*
Average	*(2431-3263)*	*2693*

1. Kent, W.J. and A.M. Zahler, *Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment.* Genome Res, 2000. **10**(8): p. 1115-25.
2. Venables, W.N. and D.M. Smith, *An Introduction to R*. 2002: Network Theory Ltd. 156.
3. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 1994: p. 28-36.
4. Grundy, W.N., et al., *Meta-MEME: motif-based hidden Markov models of protein families.* Comput Appl Biosci, 1997. **13**(4): p. 397-406.
5. Harris, T.W., et al., *WormBase: a cross-species database for comparative genomics.* Nucl. Acids Res., 2003. **31**(1): p. 133-137. Release W130, Oct. 2004.
6. Billingsley, P., *Probability and Measure*. 3rd ed. 1995, New York: Wiley-Interscience. 608.

Supplementary document 7: Oligo sequences used in this work

(Underlineds: RNA; *Italisc*: Restriction Endonuclease site; p.: 5' phosphate; -x: 3'-DMT.)

>5AD; 5' adaptor

GGAGUA*GCAUGC*GUGACGAAAA

>3AD; 3' adaptor

p.UUUUGACCAC*GAGCTC*ACAGGG-x

>5CD; 5' PCR primer

GGAGTA*GCATGC*GTGACGAAA

>3RT; 3' reverse transcption & PCR primer

CCCTGT*GAGCTC*GTGGTCAA

>5S-1pA; probe used for removal of 5S rRNA

AAAAAAAAAAAAAAAAAAAACGTCTCCGATCCAAGTACTAA

>5S-2pA; probe used for removal of 5S rRNA

AAAAAAAAAAAAAAAAAAAATTACAACATCCAGGATTCCC

>5.8S-1pA; probe used for removal of 5.8S rRNA

AAAAAAAAAAAAAAAAAAAAAATTTCACCACTCTAAGCGTCTG

>18S-1pA5; probe used for removal of 18S rRNA

AAAAAAAAAAAAAAAAAAAAGACCTGTTATCGCTCAATCTC

>18S-2pA; probe used for removal of 18S rRNA

AAAAAAAAAAAAAAAAAAAACTACCTTGTTACGACTTTTACCC

>26S-1pA; probe used for removal of 26S rRNA

AAAAAAAAAAAAAAAAAAAATCCCTATTAGTGGGTGAACAA

>CeN1-1 AY948555
aaacttacctggctgggggttatttcgtgatcatgaagacggaatccccatggtgaggcc
tacccattgcacttttgggcgggctgacccgtgtggcagtctcgagttgagattcgccaa
cagcttaatttttgcgtatcggggctgcgtgcgcgcggccctga
>CeN1-2 AY948556
aaacttacctggctgggggtatctcgtgatcatgaagacgggatccccatggtgaggcct
acccattgcacttttgggcgggctgacctgtgtggcagtctcgagttgagattcgccaac
agcttaatttttgcgtatcggggctgcgtgcgcgcggccctg
>CeN1-3 AY948557
aaacttacctggctgggggttatttcgcgatcaagaaggcggaatccccatggtgaggcc
tacccattgcacttttgggcgggctgacctgtgtggcagtctcgagttgagattcgccaa
cagcttaatttttgcgtatcggggctgcgtgcgcgcggccctga
>CeN1-4 AY948558
aaacttacctggctgggggttatctcgcgatcatgaaggcgggatccccatggtgaggcc
tatccattgcacttttggatgggctgacctgtgtggcagtctcgagttgagattcgccaa
cagcttaatttttgcgtatcggggctgcgtgcgcgcggccctga
>CeN1-5 AY948559
aaacttacctggctgggggttatttcgcgatcaagaaggcggaatccccatggtgaggcc
tacccattgcacttttgggcgggctgacctatgtggcagtctcgagttgagattcgccaa
cagcttaatttttgcgtatcggggctgcgtgcgcgcggccctga
>CeN1-6 AY948560
aaacttacctggctgggggttatttcgcgatcaagaaggcggaatccccatggtgaggcc
tacccattgcacttttgggcgggctgacctgtgtggcagtctcgagttgagattcgccaa
cagctttatttttgcgtatcggggctgcgtgcgcgcggccctga
>CeN1-7 AY948591
aaacttacctggctgggggctattttgcgatcaagaaggcagaatccccatggtgaggcc
tacccattgcacttttgggcgggctgacctgtgtggcagtctcgagttgagattcgccaa
cagcttaatttttgcgtatcggggctgcgtacgcgcggccctga
>CeN2-1 AY948570
agctttgcgctggggcgataacgtgaccaatgaggctttgccgaggtgcgtttattgctg
gttgaaaacttttcccaattgcccgcgatgtcccctgaaacatgggtggcatacgcaatt
tttgaacgcctctaggaggcag
>CeN2-2 AY948571
agctttgcgctggggcgataacgtgaccaatgaggctttgccgaggtgcgtttattgctg
gttgaaaacttttcccaattgcccgcgatgacctctgaaacatgggtgccatacgcaatt
tttgaacgcctctaggaggcag
>CeN3-1 AY948573
caactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccg
tgcaaaggagcatacattgagtattacttagaattttggagccttctcgaaagagcaag
gca
>CeN3-2 AY948574
aactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccgt
gcaaaggagcatacattgagtattatatacaatttttggagtccccttgagaaagcggga
ca

>CeN3-3 AY948575
caactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccg
tgcaaaggagcatttactgagtattacatacaatttttggagactccttgagaaagcggg
tca
>CeN3-4 AY948576
aactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccgt
gcaaaggagcatacattgagtattatatataatttttggagtccccttgagaaagcggga
ca
>CeN3-5 AY948577
aactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccgt
gcaaaggagcatacattgagcattatatacaatttttggagtcccctcgagagagcggga
ca
>CeN3-6 AY948596
aactctggttcctctgcatttaaccgtgaaaatctttcgccttttactaaagatttccgt
gcaaaggagcatacattgagtattatacacaaattttggagtcccctcggaagagcggga
ca
>CeN4 AY948578
gttcttccgagaacatatactaaaattggaacaatacagagaagattagcatggcccctg
cgcaaggatgacacgcaaattcgtgaagcgttccaaattttt
>CeN5 AY948686
ggcagtgatgatcacaaatccgtgtttctgacaagcgattgacgatagaaaaccggctga
gccaa
>CeN6 AY948579
ggttttaacccagttaaccaaggttagcatgtattccgaccattcgtaagagtgtgttga
ataacaataatttttggaacagcttcttcggggttatccgtcgaagc
>CeN7 AY948580
ggttttaacccagtttaaccaaggttagctgtcgtttcgatctctcgagagagtgtgtcg
aataaaaaataatttttggaatcgcttcatcggggaatccgttgaagcaa
>CeN8-1 AY948581
ggtttatacccagttaaccaaggttagcattaagtttcgacctttccaagaatgtgttga
aatgcaaattaatttttggaaccgcttcttcggggaatccgttgaagcaa
>CeN8-2 AY948582
ggtttttacccagttaaccaaggttagcattaaatttcgacctttcgcaagaacgcgttg
aaatgcaaatcaatttttggaaccgcttcttcggggaatccgttgaggcaa
>CeN9 AY948642
ggggctcggtccgagtttcatggtctccaatgtgtgtgtgtgtgtgttttctttaggaac
ctcggttccaacctcatcttgaccttgaaactactttgaccgctcct
>CeN10 AY948604
gaggttggccggaagaagacggttgggagagacacaaaggcctgaaacatggcctacaca
actcccgggaaggtctgagagtaggcctttgatgagatctaggagatcttccattatcct
tatagaggagaggctgtagaagaagacgtttcccctgcgagggtttggaaacggcctcgg
ccagcaattctcgtgtaaatgctgagtgcgatcgtcataccaacacaattcagactagtc
ttcggccaacctt
>CeN11 AY948561

ggttttaaccagttaactaaggttaacattaatttcgaccattcgaaagattgtgttgaa
taacaataatttttggaacagtttcttcggggatatccgatgaa
>CeN12 AY948563
acggttttaacccagttaaccaaggttagcatggaattcgatcattcgcaagaatgtgtc
gaaacacaaaatttttggacaagcttcctcggggtatccgtgggagca
>CeN13 AY948665
agtcaatgatgtttttttcaagacgggaccgactggtgaatgatgcataaatgaaatgctg
agact
>CeN14 AY948666
ctgcggtgacgatcaactcttacctactatgacaaaaacaatggttagacgttactcgta
ctgtctgagcagt
>CeN15 AY948667
gtccgttgatgacaacatacatacaccattacgatctctgaagacttcgtgctgatcatg
tatccatgcaacaccaactgaggaca
>CeN16-1 AY948564
ggttttaacccagttactcaaggtacgctggagttctgacctttcgaaagagagtgtcaa
acaactttaacttttggaaccgctctgctggggttatccggtagagca
>CeN16-2 AY948565
ggttttaacccagttactcaaggtacgctggagttctgacctttcgaaagagagtgtcaa
acaactttaatttttggaaaagcttcgctggggttatccggcgaagca
>CeN16-3 AY948566
ggttttaacccagttactcaaggtacgctggagttctgacctttcgaaagagagtgtcaa
acaactttaatttttggaactgctctactggggttatccggtagagca
>CeN16-4 AY948567
ggttttaacccagttactcaaggtacgctggagttctgacctttcgaaagaaagtgtcaa
acgactttaatttttggaaccgctctgctggggtcatccggtagagca
>CeN17 AY948668
caccaatgatgcaatggttaaatcagacgagtctattttggctatctattcgagttcttc
gaagaaattgccgctaagcggggtgaaattgaggcatttgtctgaggtga
>CeN18 AY948568
atcgcttcttcggcttattagctaagatcaaagtgtagtatctgttcttatcgtattaac
ctacggtatacactcgaatgagtgtaataaaggttatatgatttttggaacctagggaag
actcggggcttgctccgacttcccaagggtcgtcctggcgttgcactgctgccgggctcg
gcccag
>CeN19 AY948569
ggttttaacccagttaccaaggtaattcggagtttcgatctttcgaaagagagtgtcgat
tgtgaacaatttttggaatagctcttccggggaatccggtcgggcaa
>CeN20 AY948572
tggtttaaaacccagttaccaaggtaattcggagttctgacctttcgaaagaaagcgtct
tttacaataaatttttggattagttcagtcggggtttccggctgaacaa
>CeN21-1 AY948613
cagtcttcagtatgggtcaatctctgatctgcaactgaatatgatgagttcgggcgatga
tcttctgtgattacatcgcacggcgaggtgggaacgcaatacccgcctgccagcccgatt
ctgaac

>CeN21-2 AY948614
cagtcttcagtatgggtcaatctctgatctgcaactgaatatgatgagttcgggcgatga
tcttttgtgattaaatcgcacggcgagatgggaacgcaatacccgtctggcagcccgatt
ctgaac
>CeN22 AY948615
attgcagaccggtgatgaaactgttctaggaagtgccgtcttagaaacaatgattatgaa
ttggacgctgaggtca
>CeN23-1 AY948616
atctaccttactcgaaaacccgcatatgaagaccactaaatgacgaatcctaataaccca
atgggtttcattgcggatatgaggcatttgtctgagcgga
>CeN23-2 AY948617
catctaccttactcgaaataacccgtcgatgaagaccactaaatgacgaatcctaatagc
ccaatgggtttcattgcggatatgaggcatttgtctgagcggg
>CeN24 AY948669
aagcaaatgacgaaatcgcacctcggcccgactccaaccctgggggcgaaatgagctttt
taactcagatgctt
>CeN25-1 AY948618
cgatttcttataatacttcagggccttactaaagacccagtgtacaatttttgcagccct
gtcttctgaggcagggg
>CeN25-2 AY948619
gctatttataaacgctctcagggccttgcaaaagacccagtgaacaatttttggagaacc
ctgtcctttcgaggtcagggta
>CeN25-3 AY948620
ggtatgatataacactcagggccttacaaaagacccagtgaacaatttttgcaaaccctg
tcctttcgaggtcagggt
>CeN25-4 AY948621
atataatatatcactgtcagggccttacaaaagacccagtgaacaatttttgcaaaccct
gtcttcttgaaggcgggg
>CeN26-1 AY948590
cgagaccatactatacagaatcatttctgcagtatgtatctcgtaattcccatcaaattg
gtagagatgccaactgcgatgaaagggagacgggcaggccgcgagcgtgaagcagtcctt
tagggcttgatgaccgcacggagccaatgaattttggagccgtagtggcctgctcttctg
agcagt
>CeN26-2 AY948588
cgagaccatactatacagaatcatttctgcagtatgtatctcgtaattcccatcaaattg
gtagagatgccatccgcgatgagtgggagacgggcgagccgcgagcgtgaaacagttctt
cggaacttgatgaccgcacggagccgatgaattttggagccgtagtggcttgctcttctg
agcgga
>CeN26-3 AY948622
cgagaccatactatacagaatcatttctgcagtatgtatttcgtaattcccatcaaattg
gtagagattccatccgcaatgagtgggagacgggcgagccgcgagcgtgaaacagttctt
cggaacttgatgaccgcacggagccgatgaattttggagccgtagtggcttgctcttctg
agcgga
>CeN26-4 AY948586

cgagaccatactatacagaatcatttctgcagtatgtatctcgtaattcccatcaaattg
gtagagatgccaactgcgatgaaagggagacgggcaggtcgcgagcgtgaagcagtcctc
taatgacttgatgaccgcacggagccaatgaattttggagccgtagtggcctgctcttct
gagcagt
>CeN26-5 AY948589
cgagaccatactatacagaatcatttctgcagtatgtatctcgtaattcccatcatattg
gtagagatgccaactgcgatgaaagggagacgggcaggccgcgagcgtgaaacagtcctt
tgaggacttgatgaccgcatggagccaatgaattttggagccgtagtggcctgctcttct
gagcagt
>CeN27 AY948670
cacttgttacatccaatgatgagagtttgcgactagggcggtcttacacaatcatggtga
ttctagtcattctgatggta
>CeN28 AY948623
cacgtctggtgaggatgaaaacaggacaggtttcgctaaaatattaccgaatgccaatat
gtcgagacaccttggctgtctgaggacg
>CeN29 AY948624
acccgtgatgaagaaattagatccaactcccaggccagttgatacgtcttctggttcatg
cagataaaggcgaacgaacggg
>CeN30 AY948671
ggctgtgacgattactattcccaacgcttggaatgaaccaaagtgattattaaccaatcc
ttttctgagccaa
>CeN31 AY948625
gttaattcattaacttccaaggccttacacaagacctagtggacaatttttggagaccct
atctagccgtgatagggtg
>CeN32 AY948626
gttaaaccataactaactcagggccttgttaaagacccgaagtacaatttttgaagccct
gtctctcgaggcagggg
>CeN33 AY948672
caagcctcctggcgctgctattagccgggattcataggctggcgatgattgagattgttc
ccacacgcaatttctcctgatccacatgaaggctaaacttccttggacgtctgagccgc
>CeN34 AY948627
gcatgacggcaaggcgagtactcgcagccacgaaacgttccctgttgagcgcgctaactg
tgagcgaaagtccctggaccactggcagaaagtgtcaccctctaggttgggtactctttg
gagttttagtcatagcacagagacgctccttagaacagcatagggcctactctgatcgtc
atgctt
>CeN35 AY948628
aaggcccgtgaagacacgaattaccgtctgataactaatgacgctaccatggctgtaaac
cagaggccga
>CeN36-1 AY948673
catcgctctccagctccatgccgatgtaaaaaagtcagtgtggcgttttcatgagcggaa
attatcactgttccaaaaaacaattgctagtctcctgtgagctaatgatcacctgatggt
tcagacactt
>CeN36-2 AY948674
atcgctctccagctccatgcagacgtgaaaaagtcggtgtggcgttttcatgagcggaaa

ttaacattgttccaaaaaacaattgctagtctcctgtgagctaatgatcacctgatggtt
caaacactt
>CeN37 AY948629
ctgcatctatgtgtactcgctcgtcgtaaatcgacaagtgtagaagggaaatgatgcgaa
attggccgtgaactccgcaggcggacgaggactttatggttcctgtcctagtcaattgga
gttcctcattt
>CeN38 AY948675
acgcgtcattttcatccaattggcaacgtgattctaatgttggcgattcagcgtatttct
gacgcaaaattgataacttctccattgacgtctagtcagactaaactggctcggatacaa
ttagggagtttacactt
>CeN39 AY948676
atggcggataggaaccaggtcatgtttgtacgtgatttgggccgacatccccgccgaaat
agtcctgcgaagatctaaggtcgcttctggatggatttgccaccgggtgcacaccatctt
cgtgcacaatt
>CeN40 AY948677
cagcggtgatcgatgacttgtgcagtggccgaggcgatcggattgtgatgtcgcctgaaa
aggcgggacccaacgtcgcgcctttcgccagaagatggaaatatgcgcaacgtctgagct
ga
>CeN41 AY948678
atgcaccttgcactctacgcctttcctttcaatgggttggtatgatttaataaggatgca
agagaatagtacgggaagttcgaagacttgcggatttgcttccacggctccgcgagttca
taactgtcaccacaatt
>CeN42 AY948679
attccttcttatcctgcactcagtttgttcattgactgtggagcctggaaatggaggaga
aagtaaatttgatgattactaatacttcgctgtctttagaggacgcggagattgtgagac
ttgaaacattt
>CeN43 AY948680
tgcccggtctttgccccggatcaactttaatagttatgacggtttccgactcgggaaaat
agattcttgcataacgacggaaatttcagagtttgttctgaatgtctccagtcgatgtgc
aatacaatt
>CeN44 AY948681
cgacaatgataggataacctagagtctctgaaccatttcgtgtttgcaaaaacaatgctc
cttgtctgagtcag
>CeN45 AY948682
gtggggcacagagttgcagttgattgaaactaaatcatgtgcgctaagttgtcccgaaag
tatctttggatttctacaaaacagtgttccttctttaggatgctgatccgtagagatctg
aacaaat
>CeN46 AY948683
cttgcacgactgcaacctgctttcggaactttaccgtctgcggtacttgtagtgctatat
gagcgtgggccctccgtgtgcgtgtaatttaattgaactagcactccgttggtcaccgga
taatt
>CeN47 AY948630
acgtaaatgaagaaataccatccttgctctgcgagtcgttgagcaatcacttgagaactc
tgatgaggtattgtagcatcgggttacggtagccgagtcagttgattctcatcatgtggc

acttcgacggtgaacagttatggcctctgatacgt
>CeN48 AY948684
atgcacacagaaagtcgagattgccggcttcaaaggctttatcgttgattacgtgtgcta
taacaacgccataggtatcattcttcgtcttcttaggcgcggaacatctaaatatggcta
cattt
>CeN49 AY948685
accgcctctctaccaaacacttgcaagttctgtttattgcaagtgcatggaagaggcgaa
tagaaaacggaatgatgtccactccttcgattatcgttcattggattgcgggcactttac
ctcgttcgacattt
>CeN50-1 AY948631
ctgtgtatgacgacaacgtgttagggacatctgcaccaaccgtgaagatttaacgaaagt
agtactgacacag
>CeN50-2 AY948632
gtgtatgacgacaacatgttagggacatctgcacaaaccgtgaagatttaacgaaagtag
tactgacacag
>CeN51 AY948687
caccgcgtgagtactggctgtttgtgtctccgccattgccaatcagctgtcattaaccac
gcaataccgtacagaggtgctaaattaccgtgaaaaccactttaaaaattggtcactcgg
aggaggcacctcaacattt
>CeN52 AY948633
ccgcctccatgttccttaatttgtgagagcctgtcgttcactgacactttctccatggaa
cgggaaggcatatcacccaaagtcaccattcatagttcacatcattgcctaacgagtgat
gtgtctccgccaggtgtctcattt
>CeN53 AY948688
tggcaatgatcgaattatcattgagccaatccttttctgaattctgtgaggatgtaaatg
ataggtctgagcca
>CeN54 AY948689
gtttgtgatgactgcatacggatcactgggctctgaatctctatgaaccgataatatccg
ttctgata
>CeN55 AY948585 AY948594
cccgcgcctagaaactcagcggtgttttgtctttatgatcatgagctggtagcaggtgca
atattgattcggagtcttaccggcgtccaccttctgaagggtgtgcgttaaaagttcatc
cgaacaatt
>CeN56 AY948634
ttccccgtgattacaaacattgctgaaacctgcccacccgaaagcttcgagattggcgtg
ggggccgatcgaggaaaagctgacaacggtgggaaccatgatacagattatgagagatga
gggat
>CeN57 AY948690
tccacatgatgatacaaccatagcatgagctggcagcagtgatcgctaaatgtcatagtt
acacagatgggt
>CeN58 AY948691
ctgcgtctctccaacccggacaactcgcaacttgttgcaatggtcagttcgactgttaga
gacgctaaataagaaaattggcggattcgctaacgtcgcgtcccatgaccggagtttgaa
agtaggccgctgaacaatt

>CeN59 AY948692
cacggccagtttgagttgattcgctctttcgcaatagagctttgagtcaaattactgtcc
ggaaattatagagatgaagctcatttggagcaataacaattgttgaagctccgaaaatga
tgctcttcccacaatt
>CeN60 AY948693
acgccacgtgatttaggtttatttgctactcttgattaactctcatgatgacaagaaagt
atgatggc
>CeN61 AY948694
ccgtcgatgacgaccaagagttatccctgtctgaatgattgtgaggacaaaagactatgg
taacactccgagacta
>CeN62 AY948695
gtgcgatgaaggttaatgataagtttcggctgactcaaattgatgacacctttaatatgc
tgagcact
>CeN63 AY948696
gttgtcagtgacgatattacttaccgccccaggcatagtgtttgtgatgattggtttatt
ccgagactt
>CeN64 AY948599
ccccgcttcatggtgcattacactttgtaacttgcattgtgtgcgcgactaatgaagcaa
taccaatatctgcagtatctgctcattcgccgtgttgtagttgatgagctcgtagtccga
taacaatt
>CeN65 AY948697
aagcgatgacgattgatatctgctctaatgagtctgaattaccatgttgagatcttgtct
gagct
>CeN66 AY948584
acgcacgcttatttcgcgccgcaagttttgcaatgacgatgtggctaaagtgtagtgcaa
tatgagctcgctcacggcgttgcggaccatagtctgaacacggttcgcgtttatgtgagt
gaaacaatt
>CeN67 AY948698
gcccccattatctttccatcaaatgatttaggacgtcattgatggcttagaatggggaaa
attgaatcggtagaatgtgatttgtgagttgttcactgacacgtggcaactcgtattcgc
tacttcctacattt
>CeN68 AY948699
gtgccggatttaaacacttctgagttgctcacatgctcagaagaacaaggttcggaaaat
tagtgatcatttgatgtgctgaactccaaagagtcaaaactgagttggtgatcgtacatt
t
>CeN69 AY948700
gacaggatgatgagtcactcgctgagtgacaataagccgagtgttagcggtttttatgta
atcgtgatgatcattccctcaaaagcgataattgtgaactgactgtc
>CeN70 AY948701
gcgcgatgaactctttaccatctttcggggcataacactcttgatgataacatacccatt
tgctgagcgct
>CeN71 AY948635
gaattcctgcggtccggatcgtatgggttatcaattctcaaccaccccatacgaactaac
ttgactaccggaatt

>CeN72 AY948636
catcatcggtccggtgttgatgggttattatcctgtggtgcttgtcgtcgctgatcacat
tcaccgtcctctacacatcatcacaaatttgaccgatgtt
>CeN73-1 AY948637
cacatcggtccggagttgatgggttaccagattaattcttctgcttgcaggagagcccgg
tgtccttgtgatgccaaacccgtgttcctaacagaatacaaccccttcccatcgacacca
acttgaccgttgtt
>CeN73-2 AY948638
cacatcggtccggagttgatgggttacccagtcattcttctgcttgcaggagagcccggt
gtccttgtgatgccaaacccgtgttcctaacagaatacaaacccttcccatcgacaccaa
cttgaccgttgt
>CeN74-1 AY948639
gtctcggtccggcgtcagtgggttatcgtatttctctcccttcggggaatttcccatcgg
catcaacttgaccgttgcgt
>CeN74-2 AY948640
gtatcggtccggcgtcagtgggttatcaagtttgcctcccttcgggaatttctcatcggc
accaacttgaccgttgct
>CeN75 AY948593
aatacggtccggagtcggtgggttatctgagaagccccccatcgacaccaacttgaccga
tgaaaattttagttttaaa
>CeN76 AY948641
cagacaggcgtggtccggagtcggtgggttacctttgaaaccccccctcccatcggcacc
aacttgaccgttcctgtt
>CeN77 AY948602
caattcggtccggagtcaatgggttatctttcaaaacccccccattgacaacaacttgacc
ggcgt
>CeN78 AY948598
ttagcatgctgttagagcttgtaaggtatatgtgattttacgagtgttgaagtattgcaa
aagcaaaggacgggcacaattgccatgtgttggtattattgcttcaagttatttgaagct
gtaatatcaataagcatgtctcgtgtgaagtccgacaatt
>CeN79 AY948702
ttgcattgaaaagggacgggctatctttatggattgtttcgtataattttgatgcgagat
aatagagaggcgcactgttactcttcatcacttttctgtgaacggagagtgtcaagcgc
ttcccacaatt
>CeN80 AY948703
cagcatcgaaaatggacggactttcccgatggatcgtttcgtataattttggtgcaaaat
agtagagagacgcagtgttactcttccttacgttacctgtatctggagagtgtcaagcgt
ttcccacattt
>CeN81 AY948704
gtcgaagagtagtcatttgtatacgtgatacacttatacagagttacactttcgtatat
aagtgcttcttctcgttggagttgtttatttaatgagcaattacctctaaactgaaagca
acaatt
>CeN82 AY948705
acgctcttcaaaagcactggttatcggactcagacttgtccatgccagccgtcaaaatga

gcaatatgaaatatcctgtttttgggtgaggtgtaactgtatttagattagatctcaata
acacgatgacagtt
>CeN83 AY948706
ttccacactccttaagctagtttgactgtgattgtgcattttagatcgcttcatgagttc
gctgcgttttgctgtgaattcattgttgccctaatttgtatagtcatgggtgtcaaattg
gcgagttacgtgtggtaaatggatgacgatcaaattttgttctgtgaactttgccgtaga
tgtgatgctgtttgtttatagattgataatgattgtcgaacaatt
>CeN84 AY948707
ttgcctcagtgtctgggcaaagcagactgtttttaatagtactgacttaatccactgatg
gctatagaaaaacattgtgtaatccccgaacaccttgtggtgtattggattcatgacatt
gtgacaatt
>CeN85 AY948597 AY948600
acgcacaacttaagaacttgcgaaatctacacattcttggctcttcttagctgcgttttg
cagttcagattctgccaagttttgtgttcatggttacgcgagtgtcatatttgtgctata
taaaaaatcgtctcacagtaaattataatagttagagtatctaaagtcgcgctaaggccc
cgatattctatttttcctggtgtgaacgatgacattt
>CeN86 AY948708
ctgcaactattcaagagatttgcctcccgtggggcacgccttttttcgttgcgaaataaa
acatgtcctttattcccgctcctggattgtgtttcttgcgcgatgatatggctatgacaa
tt
>CeN87 AY948709
atgctcttcaaaagcactggttttaggatccactattatccaagccagccgtcaaaactg
agctataagaattatcttgtttttgggtgaggtgtattcaattcagaatgcgtctcaata
acacgatgacaatt
>CeN88 AY948710
tttctctcgtcttggcgcttccactggtgaatgtggttgtgtatgttacgagagacaaag
tagcgcctaacggctttcggatctccttcggtgtttgtcttgaatttcgacaaatgtcat
cctctgggttcgagacaatt
>CeN89 AY948711
gcgcgatgaggattgataacacatacacacactctgaagttatgtgaagagataattgaa
gaacggatatctgagcgc
>CeN90 AY948712
atgcagatgtccattacgaaaaggctctttacctttgacgtttagttaaatttgcgaaa
taaaattgatgtctcgaagacatgtgcttcatattttgatgctcatgttcaagatcagca
aacaaac
>CeN91 AY948713
caaccactattcaagttaatcttatgaaaattattgataagatgattatagaactgtgga
atataaacgcgaagcgaattataaaagcagagtgttactttccaaaaatgtttttttaat
gaaatttttggtcttagacactcagacaac
>CeN92 AY948714
ttgcacctaccaattatatcggaagattttgaagtcatagttttcctatcgtagtgtgca
ataaaaaccgttctactgccacacgggtactggccttgaagctgtatcggttccaaagtt
gttcgaacaatt
>CeN93 AY948715

tttccattaagtagtcccgtacgttgaaattttcacgtacggtttggacaatatggagaa
atgaatgttactctttttgcaatgacgctgaaacatattttcgctttgttagccataagt
ttaaaacaatt
>CeN94 AY948716
caagcagttctccaacagaaatatgtgcatgaagtgatggtttgggtaaccatttcatct
tttcatgtcatatatctcctgcctctcacggtttcatctgtgagaagcttcaaagaactg
caatagaagtctagctcgagatgcgtgtgccggtttgcttttaatgcaatatggttcctt
ttctcctcgggacagtt
>CeN95 AY948717
cgctcctatcttggcgcttccattggtgaatatggctgtgtatgttataggagagatagt
agtgccttttggctttcggatctccttcggtgtttgtcttgcatctcgacaaatgtcatc
ctctgggttcgagacaatt
>CeN96 AY948718
ctcgacatgtgactagcgatcctcttcggggttgattgcttttcttagtggagcatcggg
ggctttctgtaaagttgactccgatccaccttgtcgtataccaacagtctgtgtggccct
ctaccgctatttgagatgcccgggctggattgcatattccttatccttccgtagatccta
tgggtccgtgacggttaggacaaaccatgcaaattacaaat
>CeN97 AY948592 AY948603
ttccactgcgaagtcagcagactagtcaatcgattagtaggcgttacagtggtggaatat
caagatttatgtacaccgtctttgcttgaaacagtttttgtttttgctttgttcggaacg
tttttcatgggactgagagttggaaatttagcactggtcgtttacagggttgtcctgcag
tcaattctaagttaccttagtttatgtcccacacattt
>CeN98 AY948643
gtgctgtgaagagaacgtgccactgtactttgcccatcggaagggcattgaaatggagat
atacctggcacaggggccatctgagcacttttttttt
>CeN99 AY948719
gttccacttatttttcaacgtcgccggttctagaatcgatgtgaccccaataagctggta
gataagttgttactctgccacacgtgtactggtatcatgctgtatgtgttccaaagattt
tacaaacaatt
>CeN100 AY948644
aaacgaggtccagagtcacctttcagcaaaatgcagttagggtgctaggatttctcgaa
aatgaaattggtcgtattatgcaagtagggcttgcttttgcgctccccgacttcacctca
ttcccaaacatt
>CeN101 AY948645
aacctctcttataagccgggggactagcattttgttaagttcactagtaaaatgagaggt
aaagcatagagacaaccagacaccgagaatgttttgatgttttcggtcactttgtgtgtc
ccacaaat
>CeN102 AY948646
tatccatgtttactactactttcatctcttttcactgaaatgtcgagtatttttgcga
tgcatggtaaagaaatcaaggtgaccaggtttcttttcaattttccccataattgaaga
gttacaatgccatctgacaagg
>CeN103 AY948605
taccccatgatgtatcaatttagctaattgtgagctacttcgctctcgagattttgtctt
gaagtgtgagtcgtattgaacgaactttgtcgaagctgatctgagggag

>CeN104 AY948647
ctgcaatgttctcagtgcatcctttggttctccgagccgaaagaagtcactagctgaaaa
agagtttatactctgatcacacagtaacattgcaataaaagcttagctcgagatgcgtgt
gccgttctgcttttaatgcgatctggttccttttctcctcgggacaatt
>CeN105 AY948648
gattgtcgtgaatcttgatcggcgtgtttatttacccgcgttgattgtaaaggccgaca
tatatcaaaatttgacattaatgctggaagtttctgaaatagttccatttttcttccaat
tatttcaaatgtcaaacaaac
>CeN106 AY948649
cgatgatgatgaagaatttttgatatggtgtcaggacctctgagagttccgtgatgatgt
ttagagttcctgaatctt
>CeN107-1 AY948606
caccgagcgtcgtggcgggcgcttgtgagtcagcttcttgacggtagataagtgtggatg
gagtgagaggaggagtcctgtgtatgtcgttgtctacgtcgaccgagcgtccgtgccaag
cgctacgtcaccaggggataactgtcggaaggcggttagtcccgggtgcataaggagtcg
tggatggttcaggaccgaaaggtagcagacaaaagcgaccgcgtggtgcagtggccggac
cgcgcttgtgagttgacacacatacgcagccttctcgataccttcagaccacttatcatt
>CeN107-2 AY948607
caccgagcgtcgtggcgggcgcttgtgagtcagcttcttgacggtagataagtgtggatg
gagtgagaggaggagtcctgtgtatgtcgtcgtctacgtcgaccgagcgtccgtgccaag
cgctacgtcaccaggggataactgtcggaaggcggttagtcccgggtgcataaggagtcg
tggatggttcaggaccgaaaggtagcagacaaaagcgaccgcgtggtgcagtggccggac
cgcgcttgtaagttgacacacatacacagccttctcgataccttcagaccacttatcatt
>CeN107-3 AY948608
caccgagcgtcgtggcgggcgcttgtgagtcagcttcttgacggtagatgagtgtggatg
gagtgagaggaggagtcctgtgtatgtcgttgtcaacgtcgaccgagcgtccgtgccaag
cgctgcgtcaccaggggatgactgtcggaagatggtcagtcccgggtgcataaggagtgg
tggatggttcaggaccgaaaggtagcagacaaaagccaccgcgcggtgcagtggccggac
agcgcttgtgagttgacacgtatacacagccttctcaataccttcagaccacttgtcac
>CeN107-4 AY948609
caccgagcgtcgtggcgggcgcttgtgagtcagcttcttgacggtagataagtgtggatg
gagtgagaggaggagtcctgtgtatgtcgttgtcaacgtcgaccgagcgtccgtgccaag
cgctgcgtcaccaggggatgactgtcggaagatggtcagtcccgggtgcataaggagtgg
tggatggttcaggaccgaaaggtagcagacaaaagccaccgcgcggtgcagtggccggac
tgcgcttgtgagttgacacgtatacacagccttctcaataccttcagaccttcagaccac
ttatcatt
>CeN108 AY948650
gtgcgatgaatgacttggataagtttcggctgaaacttggtgatgccaacttttaaaac
tgctgagcac
>CeN109 AY948651
gcgcaatgaatgtttaaccatctttcggctgaatccatgatgccaattttcaaaatactg
agcgcat
>CeN110 AY948652
gagcagatgtccattacgcgaatgcctgtgcttttcgacgtttagtttagtctgcaagat

agaattgatgtctcgaagacaggtacatcagcttttagagttcctgttcaagatcagcaa
acaaac
>CeN111 AY948653
aagcagtgatgattttatagttcagcttatcttcggatttgatgagaaatttcgccccta
tcagagcttt
>CeN112 AY948610
caatcctaaacttaaataacaaaaaacccaaagcctaactcaggacttggtaacaatttt
tggagaccctaacttttattagttagggtg
>CeN113 AY948654
tggctaatgatgttctctgcgaaatacacaacttactacaaactgatcttatttgaattg
agggttactgtagctactactgtagctaccgtaatcctcacagtgtgaacttaaattaaa
gactgaagctt
>CeN114 AY948655
gtgcaaggatgaaaaagaactctctcactgatagatgatgtcttcctacattatcagagc
act
>CeN115 AY948611
cagttctccaagacggacggagctttcgagtttcgtgctacggcacgtctctcggagagc
caagacaaatccaagtcaaaaaccaacaaaccaaataaatgacttaacaaaatgctcccc
gtctttaaattttggaaaagggtcccaatagggacccgg
>CeN116 AY948562
cggtttaattacccaagtttgaggtaaacattgaaactgacccaaagaaatttggcgtta
gctataaattttggaacgtctcctctcggggagacaaa
>CeN117 AY948656
agacagaggagttgatgagaactctaatccattctctgagcgagaaggatggccgaagcg
ggttcgcatttgaggcattaaggtagacgacagagttcttctggaaactactgcctcgcg
ctgacgtcatgccttcgcgggctgaattttgggtctgatcctc
>CeN118 AY948657
aatcggtgatgtgatatccagttctgctactgagttattgtgaagattaactttccccgt
ctgagatt
>CeN119 AY948658
catgtcaatgatgtctaaaaaattactacgatttaattcgaattgctgtgagatcaatct
tatacaattctgagacac
>CeN120 AY948659
aaagccgatgattaccaaaaataaccaaatgtttgagtgattgtttgtgatcgaattttg
tcactatcgctgaggctt
>CeN121 AY948660
cggctgtgatgatttcctattgccgtttacccgtctgaggaaaaccgtgcttgatacaac
ttggaaaaggctgagccga
>CeN122 AY948661
cgggaatgatgaccttctgtgtaggaatctcaatgagtgactgtgacataaaaatgcagt
aaattcactgacccc
>CeN123 AY948662
gatctatgatgagactttcacgacggtcttccgatgtaatacatacctgtggatatcttt
acgtgaagctgagatc

>CeN124 AY948663
cgagcggtgatgaatgcacgtattgctctgacacctcttatgttagcggtaaatttccgt
gccgcgatgagtccactaggatctctgagctc
>CeN125 AY948601
ctcacccacggcaacaaattccaattgtgtgtaacattcaatgcaattgagccgacgccg
tgggaaatcactttcgtggaaccatttgatcccacgctcgttactgttaatgattgtggt
ttgcacgtggtttacaat
>CeN126 AY948664
aagcagtcgttcagtggcgcaagcgattttcccatgcgcttgtgcctaatctttgactgc
gagataaaaatgtcagagtcaagcggtccgctatctgtgtagcattctgttcaagatttg
actgatattt
>CeN127 AY948595
ccgaagtgcgatatccagacagaactttaagagtactgcttggactgagtttactaactt
cgaatatgaaatagttgatcgataaccggaatatactcaaaaaaaaaggtgtgctccgga
caaaaatcaactaacaaaa
>CeN128 AY948583 AY948587
ccaatgatgactcaaaatagctatatgaactctttggtgactcgataataagagaaatct
gaaaattttcaaggtttttctagagatttcaccgattcttttgaagtatcggtgcaccaa
gtgaaaaataaaattgcgacgctttagtgacgctggagaaagtttttatccagtgtctaa
aggcgcggcgcaacatttttaaactttccgatgatttgtgattactaaaaggcgaatctg
aggcga
>CeN129 AY948612
cctcgatgacgattcacctagctcactcagacatacaactggtgataaaaaaatttcgtg
tcttagagac

sort	CeN	Class	State	comment	size	reads	cap	Group	UM	Conserv	Intronic
1.1	CeN1-1	snRNA U1	Known		164	97	1	I-A	1	92.68	Y
1.2	CeN1-2	snRNA U1	Known		162	31	1	I-A	1	93.21	Y
1.3	CeN1-3	snRNA U1	Known		164	18	1	I-A	1	91.46	
1.4	CeN1-4	snRNA U1	Known		164	14	1	I-A	1	93.29	
1.5	CeN1-5	snRNA U1	Known		164	13	1	I-A	1	90.85	
1.6	CeN1-6	snRNA U1	Known		164	3	1	I-A	1	90.85	
1.7	CeN1-7	snRNA U1	Known		164	3	1	I-A	1	90.85	
2.1	CeN2-1	snRNA U4	Known		142	51	1	I-A	1	99.3	
2.2	CeN2-2	snRNA U4	Known		142	4	1	I-A	1	97.89	Y
3.1	CeN3-1	snRNA U5	Known		123	39	1	I-A	1	92.68	
3.2	CeN3-2	snRNA U5	Known		122	17	1	I-A	1	92.62	P
3.3	CeN3-3	snRNA U5	Known		123	12	1	I-A	1	82.11	
3.4	CeN3-4	snRNA U5	Known		122	4	1	I-A	1	91.8	Y
3.5	CeN3-5	snRNA U5	Known		122	2	1	I-A	1	93.44	
3.6	CeN3-6	snRNA U5	Known		122	2	1	I-A	1	93.44	Y
4	CeN4	snRNA U6	Known		102	16	1	I-B	1	100	P
5	CeN5	snoRNA C/D	Known	U18	65	6	1	I-A	1	98.46	
6	CeN6	snRNA sls-2	Puta		107	4	1	I-A	1	87.85	
7	CeN7	snRNA sls-2	Puta	Y75B8A.38	110	3	1	I-A	1	93.64	
8.1	CeN8-1	snRNA sls-2	Puta		110	3	1	I-A	1	34.55	Y
8.2	CeN8-2	snRNA sls-2	Puta		111	2	0.5	I-A	1	85.59	
9	CeN9	scRNA yrn-1	Known		107	2	0	I-B	1	22.43	
10	CeN10	RNAase P RNA	Known		253	2	-0.5	I-B	1	90.12	
11	CeN11	snRNA sls-2	Puta		104	1	0.5	I-A	1	83.65	
12	CeN12	snRNA sls-2	Puta		107	1	0.5	I-A	1	50.47	Y
13	CeN13	snoRNA C/D	Known	Y71D11A.7	65	7	-1	II	2	92.31	Y
14	CeN14	snoRNA C/D	Known	H09I01.2	73	6	-1	II	2	75.34	
15	CeN15	snoRNA C/D	Known	F30H5.4	86	4	1	II	2	89.53	
16	CeN16-1	snRNA sls-2	Puta		108	7	1	VI		85.19	
16	CeN16-2	snRNA sls-2	Puta		108	6	1	VI		87.04	
16	CeN16-3	snRNA sls-2	Puta		108	6	1	VI		85.19	
16	CeN16-4	snRNA sls-2	Puta		108	4	1	VI		85.19	
17	CeN17	snoRNA C/D	Known	U15	110	3	-1	V		82.73	Y
18	CeN18	snRNA U2	Known		186	22	1	I-A	1	98.92	P
19	CeN19	snRNA sls-2	Puta		107	3	1	VI		47.66	
20	CeN20	snRNA sls-2	Puta		108	1	0.5	VI		16.67	
21	CeN21-1		Novel		126	80	1	I-A	1	92.86	
21	CeN21-2		Novel		126	47	1	I-A	1	90.48	
22	CeN22	snoRNA C/D	Novel		76	30	1	I-A	1	23.68	
23	CeN23-1		Novel		100	21	0.5	I-A	1	93	
23	CeN23-2		Novel		103	4	0	I-A	1	96.12	
24	CeN24	snoRNA C/D	Novel		74	19	1	I-A	1	89.19	
25	CeN25-1	*snlRNA*	Novel		77	17	1	I-A	1	57.14	Y
25	CeN25-2	*snlRNA*	Novel		82	4	1	I-A	1	53.66	
25	CeN25-3	*snlRNA*	Novel		78	1	0.5	I-A	1	41.03	
25	CeN25-4	*snlRNA*	Novel		78	1	0.5	I-A	1	41.03	Y
26	CeN26-1	*U3(snlRNA)*	Puta		186	14	1	I-A	1	94.09	Y
26	CeN26-2	*U3(snlRNA)*	Puta		186	10	1	I-A	1	97.85	P
26	CeN26-3	*U3(snlRNA)*	Puta		186	8	1	I-A	1	96.24	
26	CeN26-4	*U3(snlRNA)*	Puta		187	5	0	I-A	1	93.58	
27	CeN26-5	*U3(snlRNA)*	Puta		187	4	1	I-A	1	93.05	
27	CeN27	snoRNA C/D	Novel		80	12	1	I-A	1	96.25	Y

28	CeN28	snoRNA C/D	Novel		88	11	1	I-A	1	90.91	Y
29	CeN29		Novel		82	6	1	I-A	1	93.9	
30	CeN30	snoRNA C/D	Novel		73	6	1	I-A	1	87.67	
31	CeN31	*snlRNA*	Novel		79	5	1	I-A	1	45.57	
32	CeN32	*snlRNA*	Novel		77	4	1	I-A	1	64.94	Y
33	CeN33	snoRNA C/D	Novel		119	3	1	I-A	1	88.24	Y
34	CeN34		Novel		186	2	-0.5	I-B	1	95.7	Y
35	CeN35		Novel		70	2	1	I-A	1	24.29	Y
36	CeN36-1	snoRNA H/ACA	Novel		130	50	-1	II	2	62.31	
36	CeN36-2	snoRNA H/ACA	Novel		129	7	-1	II	2	61.24	
37	CeN37		Novel		131	34	-1	II	2	58.78	
38	CeN38	snoRNA H/ACA	Novel		137	29	-1	II	2	81.75	
39	CeN39	snoRNA H/ACA	Novel		131	25	-1	II	2	90.84	Y
40	CeN40	snoRNA C/D	Novel		122	25	-1	II	2	24.59	
41	CeN41	snoRNA H/ACA	Known	CeR-9	137	19	-1	II	2	58.39	Y
42	CeN42	snoRNA H/ACA	Novel		131	19	-1	II	2	38.17	Y
43	CeN43	snoRNA H/ACA	Novel		129	18	-1	II	2	50.39	Y
44	CeN44	snoRNA C/D poss	Novel		74	18	-1	II	2	90.54	
45	CeN45	snoRNA H/ACA	Novel		127	11	-1	II	2	42.52	Y
46	CeN46	snoRNA H/ACA p	Novel		125	10	-1	II	2	84	Y
47	CeN47	snoRNA C/D	Novel		155	9	-1	II	2	23.23	Y
48	CeN48	snoRNA H/ACA	Novel		125	8	-1	II	2	82.4	Y
49	CeN49	snoRNA H/ACA	Novel		134	8	-1	II	2	87.31	
50	CeN50-1		Novel		73	7	-1	II	2	87.67	
50	CeN50-2		Novel		71	1	-0.5	II	2	87.32	
51	CeN51	snoRNA H/ACA	Novel		139	6	-1	II	2	79.14	
52	CeN52		Novel		144	6	-1	II	2	54.17	
53	CeN53	snoRNA C/D	Novel		74	5	-1	II	2	93.24	
54	CeN54	snoRNA C/D	Novel		68	5	-1	II	2	94.12	Y
55	CeN55	snoRNA H/ACA p	Novel		129	5	0	II	2	82.95	Y
56	CeN56		Known	CeR-5	125	5	-1	II	2	33.6	Y
57	CeN57	snoRNA C/D	Novel		72	5	-1	II	2	86.11	
58	CeN58	snoRNA H/ACA	Novel		139	4	-1	II	2	90.65	Y
59	CeN59	snoRNA H/ACA	Novel		136	3	-1	II	2	50	
60	CeN60	snoRNA C/D	Novel		68	3	-1	II	2	77.94	
61	CeN61	snoRNA C/D	Novel		76	3	-0.5	II	2	69.74	
62	CeN62	snoRNA C/D	Novel		68	3	-1	II	2	50	Y
63	CeN63	snoRNA C/D	Novel		69	2	-1	II	2	95.65	
64	CeN64		Novel		128	2	-0.5	II	2	89.84	Y
65	CeN65	snoRNA C/D	Novel		65	2	-1	II	2	93.85	
66	CeN66		Novel		129	2	0	II	2	37.21	
67	CeN67	snoRNA H/ACA	Novel		134	1	-0.5	II	2	34.33	
68	CeN68	snoRNA H/ACA	Novel		121	1	-0.5	II	2	15.7	
69	CeN69	snoRNA C/D	Known	CeR-19	107	1	-0.5	II	2	91.59	Y
70	CeN70	snoRNA C/D	Novel		71	1	-0.5	II	2	87.32	
71	CeN71	*sbRNA*	Novel		75	10	-1	III	3	25.33	
72	CeN72	*sbRNA*	Novel		100	6	-1	III	3	22	
73	CeN73-1	*sbRNA*	Novel		134	4	1	III	3	32.09	
73	CeN73-2	*sbRNA*	Novel		132	1	0.5	III	3	33.33	
74	CeN74-1	*sbRNA*	Novel		80	2	-1	III	3	87.5	
74	CeN74-2	*sbRNA*	Novel		78	1	-0.5	III	3	80.77	
75	CeN75	*sbRNA*	Novel		80	1	-0.5	III	3	28.75	
76	CeN76	*sbRNA*	Novel		78	1	-0.5	III	3	37.18	

77	CeN77	*sbRNA*	Novel		65	1	-0.5	III	3	78.46	Y
78	CeN78	snoRNA H/ACA	Novel		160	13	-1	V		40.62	Y
79	CeN79	snoRNA H/ACA	Novel		131	11	-1	V		83.97	Y
80	CeN80	snoRNA H/ACA	Novel		131	8	-1	V		54.2	Y
81	CeN81	snoRNA H/ACA	Novel		126	7	-1	V		65.08	Y
82	CeN82	snoRNA H/ACA	Known	CeR-8 NCB	134	6	-1	V		90.3	Y
83	CeN83	snoRNA H/ACA	Novel		225	6	-1	V		8.44	Y
84	CeN84	snoRNA H/ACA	Novel		129	6	-1	V		93.8	Y
85	CeN85	snoRNA H/ACA	Novel		217	6	-1	V		9.22	Y
86	CeN86	snoRNA H/ACA p	Novel		122	5	-1	V		54.1	Y
87	CeN87	snoRNA H/ACA	Known	CeR-3 NCB	134	5	-1	V		93.28	Y
88	CeN88	snoRNA H/ACA	Novel		140	4	-1	V		91.43	Y
89	CeN89	snoRNA C/D	Novel		78	4	-1	V		30.77	Y
90	CeN90	snoRNA H/ACA	Known	CeR-6	127	4	-1	V		16.54	Y
91	CeN91	snoRNA H/ACA	Novel		150	4	-1	V		17.33	Y
92	CeN92	snoRNA H/ACA	Novel		132	3	-1	V		62.12	Y
93	CeN93	snoRNA H/ACA	Novel		131	3	-1	V		58.78	Y
94	CeN94	snoRNA H/ACA	Novel		197	3	-1	V		36.55	Y
95	CeN95	snoRNA H/ACA	Novel		139	3	-1	V		92.09	Y
96	CeN96	snoRNA H/ACA	Novel	snR30	221	3	0	V		85.07	Y
97	CeN97	snoRNA H/ACA	Novel		218	2	-0.5	V		82.57	Y
98	CeN98	snoRNA C/D	Novel		97	1	-0.5	V		37.11	Y
99	CeN99	snoRNA H/ACA	Novel		131	1	-0.5	V		35.88	Y
100	CeN100	snoRNA H/ACA	Known	CeR-4	132	1	0	V		63.64	Y
101	CeN101	snoRNA H/ACA	Novel		128	1	-0.5	V		89.84	Y
102	CeN102	snoRNA H/ACA	Novel		142	1	0.5	V		16.9	Y
103	CeN103	snoRNA C/D	Novel		109	1	-0.5	V		25.69	Y
104	CeN104	snoRNA H/ACA	Novel		169	1	-0.5	V		37.87	Y
105	CeN105	snoRNA H/ACA	Novel		141	1	-0.5	V		40.43	Y
106	CeN106	snoRNA C/D	Novel		78	15	-1	V		83.33	Y
107	CeN107-1	SRP RNA	Puta		300	18	-1	IV		84	
107	CeN107-2	SRP RNA	Puta		300	15	-1	IV		83.67	
107	CeN107-3	SRP RNA	Puta		299	8	-1	IV		85.28	Y
107	CeN107-4	SRP RNA	Puta		308	1	-0.5	IV		80.19	
108	CeN108	snoRNA C/D	Novel		70	2	-0.5	V		48.57	Y
109	CeN109	snoRNA C/D	Novel		67	1	0.5	V		44.78	Y
110	CeN110	snoRNA H/ACA p	Novel		126	1	-0.5	V		53.97	Y
111	CeN111	snoRNA C/D	Novel		70	6	-1	VI		87.14	
112	CeN112	*snlRNA*	Novel		90	5	1	VI		27.78	
113	CeN113	snoRNA C/D poss	Novel		131	3	-1	VI		24.43	
114	CeN114	snoRNA C/D	Novel		63	2	-0.5	II	2	25.4	
115	CeN115	*snlRNA*	Novel		159	1	0.5	VI		17.61	
116	CeN116	snRNA SL1	Puta		98	112	1	I-A	1	98.98	
117	CeN117	snoRNA C/D	Puta	sn2417	163	28	1	I-A	1	93.25	
118	CeN118	snoRNA C/D	Puta	sn1185 con	68	4	-1	II	2	92.65	Y
119	CeN119	snoRNA C/D	Puta	sn2317	78	6	-1	II	2	41.03	Y
120	CeN120	snoRNA C/D	Puta	sn3159	78	9	1	I-A	1	85.9	Y
121	CeN121	snoRNA C/D	Puta	sn2343	79	5	1	I-A	1	68.35	
122	CeN122	snoRNA C/D	Puta	sn2429	75	9	1	I-A	1	92	
123	CeN123	snoRNA C/D	Puta	sn3071	76	1	-0.5	II	2	94.74	Y
124	CeN124	snoRNA C/D	Puta	sn2903	92	9	-1	V		96.74	Y
125	CeN125	snoRNA H/ACA	Novel		138	3	-0.5	II	2	52.9	Y
126	CeN126	snoRNA H/ACA	Novel		130	3	-1	II	2	87.69	Y

127	CeN127	snoRNA H/ACA	Novel	139	1	-0.5	V		23.6	Y	
128	CeN128	snoRNA H/ACA	Novel	246	1	-0.5	V		31.33	Y	
129	CeN129		Novel	70	1	-0.5	II	2	65.71	Y	

gene	Chr.	Start	Stop	Group	UM	Host Gene	Intron Size	Start	Stop	Rest	Host gene molecular_function
cen11	IV	5317260	5317363	I-A	1						
cen1-1	II	12944698	12944861	I-A	1	F58G1.7	431	220	383	48	structural constituent of ribosor
cen116.1	V	17131684	17131781	I-A	1						
cen116.10	V	17133642	17133739	I-A	1						
cen116.2	V	17124248	17124345	I-A	1						
cen116.3	V	17123272	17123369	I-A	1						
cen116.4	V	17125228	17125325	I-A	1						
cen116.5	V	17127176	17127273	I-A	1						
cen116.6	V	17132666	17132763	I-A	1						
cen116.7	V	17126204	17126301	I-A	1						
cen116.8	V	17120732	17120829	I-A	1						
cen116.9	V	17128152	17128249	I-A	1						
cen117	I	3747	3909	I-A	1						
cen12	III	7140371	7140475	I-A	1	B0280.12a	467	258	362	105	AMPA (non-NMDA)-type ionotr
cen1-2	II	6968323	6968484	I-A	1	C15F1.5	359	150	311	48	ATP binding;ATPase activity;A
cen120	II	11224084	11224161	I-A	1	T06D8.3	1253	1111	1188	65	structural constituent of ribosor
cen121	IV	867061	867139	I-A	1						
cen122	V	8250479	8250552	I-A	1						
cen1-3	V	14270131	14270294	I-A	1						
cen1-4	V	9093464	9093627	I-A	1						
cen1-5	V	14465778	14465941	I-A	1						
cen1-6	V	14031060	14031223	I-A	1						
cen1-7	II	7171769	7171932	I-A	1						
cen18.1	II	13835827	13836012	I-A	1						
cen18.2	II	13948862	13949047	I-A	1						
cen18.3	II	13951311	13951496	I-A	1	W07G1.2	334	92	277	57	oxidoreductase activity;
cen18.4	II	13852678	13852863	I-A	1	F08G2.8	609	219	404	205	Aegilops tauschii Gamma-gliad
cen18.5	II	13829200	13829385	I-A	1						
cen18.6	I	12181779	12181964	I-A	1						
cen18.7	I	12293799	12293984	I-A	1						
cen18.8	I	12332812	12332997	I-A	1						
cen18.9	I	12294641	12294826	I-A	1						
cen2-1.1	V	11982706	11982847	I-A	1						
cen2-1.2	V	11480736	11480877	I-A	1						
cen21-1	IV	8428602	8428727	I-A	1						
cen21-2	IV	4880732	4880856	I-A	1						
cen22	I	8253603	8253678	I-A	1						
cen2-2	V	7145310	7145451	I-A	1	F10D2.6	745	134	275	470	UDP-glucuronosyl and UDP-gl
cen23-1	II	15025700	15025799	I-A	1						
cen23-2	III	5653815	5653917	I-A	1						
cen24	II	8091627	8091700	I-A	1						
cen25-1	IV	9511376	9511452	I-A	1	C33A12.3	831	615	691	140	
cen25-2	V	14095157	14095238	I-A	1						
cen25-3	V	14098977	14099053	I-A	1						
cen25-4	IV	14863955	14864032	I-A	1	Y57G11C.38	2139	416	493	1646	electron transporter activity;hel
cen26-1	I	10525030	10525215	I-A	1	F59C6.6	949	221	406	543	pyrophosphatase activity;
cen26-2.1	V	11178117	11178302	I-A	1	T19C4.6	534	224	409	125	signal transducer activity;
cen26-2.2	V	11179523	11179708	I-A	1						
cen26-3	V	11587780	11587964	I-A	1						
cen26-4	I	10573309	10573495	I-A	1						
cen26-5	IV	14326921	14327107	I-A	1						
cen27	V	8351699	8351778	I-A	1	C08D8.1	303	150	229	74	
cen28	II	14018243	14018330	I-A	1	F01D5.10	642	498	585	57	Chondroitin 6-sulfotransferase
cen29	II	5599723	5599804	I-A	1						
cen30	III	9717276	9717348	I-A	1						
cen31	IV	13600822	13600900	I-A	1						
cen3-1	I	2688948	2689070	I-A	1						
cen32	IV	9511746	9511822	I-A	1	C33A12.3	831	245	321	510	

cen3-2.1	IV	9464500	9464621	I-A	1	F38E11.6a	422	128	249	173 structural constituent of riboso
cen3-2.2	IV	9444625	9444746	I-A	1					
cen33	I	2085039	2085157	I-A	1	Y37E3.11	700	487	605	95 nucleotidyltransferase activity;
cen3-3	IV	7316722	7316844	I-A	1					
cen3-4	IV	9001287	9001408	I-A	1	C53D6.8	417	247	368	49
cen35	V	8231492	8231561	I-A	1	C16D9.2a	392	286	355	37 ATP binding;protein kinase act
cen3-5	IV	12787114	12787235	I-A	1					
cen3-6	IV	2655067	2655188	I-A	1	Y69A2AR.31	492	222	343	149 electron transporter activity;he
cen5	II	5701626	5701690	I-A	1					
cen6	I	9065948	9066054	I-A	1					
cen7	III	12250336	12250445	I-A	1					
cen8-1	III	11090883	11090992	I-A	1	W09D6.5	754	609	718	36 Serine/arginine repetitive matri
cen8-2	III	11090286	11090396	I-A	1					
cen10	I	13472005	13472257	I-B	1					
cen34	II	7200160	7200345	I-B	1	E04F6.2	996	508	693	303 Molecular chaperone/chaperor
cen4.1	V	858921	859022	I-B	1					
cen4.10	IV	4885544	4885645	I-B	1					
cen4.11	IV	13443720	13443821	I-B	1					
cen4.12	III	4414154	4414255	I-B	1	R07E5.13	718	588	689	29 ATP binding;DNA binding;DNA
cen4.13	III	9447650	9447751	I-B	1					
cen4.14	III	10989552	10989653	I-B	1	W05B2.7	360	205	306	54 oxidoreductase activity;
cen4.15	III	9445803	9445904	I-B	1					
cen4.16	III	4426823	4426924	I-B	1	C28A5.6	843	720	821	22 ATP binding;protein kinase act
cen4.17	III	5080755	5080856	I-B	1					
cen4.2	IV	13443370	13443471	I-B	1	K09B11.10	962	782	883	79
cen4.3	IV	13435760	13435861	I-B	1	K09B11.10	3947	3816	3917	30
cen4.4	IV	4930950	4931051	I-B	1	Y9C9A.8	375	221	322	53 structural constituent of riboso
cen4.5	IV	4866807	4866908	I-B	1					
cen4.6	IV	13440527	13440628	I-B	1	K09B11.10	1770	1431	1532	238
cen4.7	IV	13439072	13439173	I-B	1					
cen4.8	IV	13440877	13440978	I-B	1					
cen4.9	IV	4864630	4864731	I-B	1					
cen9	IV	7499498	7499603	I-B	1					
cen114	I	8267492	8267554	II	2					
cen118	II	11484586	11484653	II	2	B0334.2	161	52	119	42 TWK (two-P domain K+) potas
cen119	II	10840222	10840299	II	2	M106.1	224	73	150	74 ATP binding;DNA binding;DNA
cen123	V	12306296	12306371	II	2	C51F7.1	267	109	184	83 Membrane-associated protein
cen125	V	19645623	19645758	II	2	Y43F8C.7	1466	418	553	913 hydrolase activity;
cen126	V	12599436	12599565	II	2	C14C10.3	240	86	215	25 ATP binding;ATPase activity;A
cen129	V	6891670	6891739	II	2	K11C4.3a	270	160	229	41 Beta-spectrin
cen13	III	1122497	1122561	II	2	Y71D11A.3a	5251	5114	5178	73 electron transporter activity;he
cen14	IV	8253081	8253153	II	2					
cen15	III	502802	502887	II	2					
cen36-1	III	8937670	8937798	II	2					
cen36-2	X	15619411	15619539	II	2					
cen37	IV	8375402	8375532	II	2					
cen38	III	4689596	4689732	II	2					
cen39	IV	11474036	11474166	II	2	C10C6.6	334	179	309	25 ATP binding;ATPase activity;A
cen40	V	15193561	15193682	II	2					
cen41	V	16372802	16372938	II	2	W08G11.3	284	97	233	51 Early endosome antigen
cen42	II	10545106	10545236	II	2	R166.5a	126	-61	69	57 structural constituent of riboso
cen43	I	6220760	6220888	II	2	T08B2.9a	350	172	300	50 ATP binding;phenylalanine-tRI
cen44	II	9774670	9774743	II	2					
cen45	I	9184382	9184508	II	2	F25H5.3a	395	206	332	63 pyruvate kinase activity;
cen46	IV	12086680	12086804	II	2	F11A10.7	284	144	268	16 RNA recognition motif
cen47	I	5611666	5611820	II	2	F46F11.9a	891	646	800	91 predicted involvement in meios
cen48	II	4942378	4942502	II	2	ZK546.13	283	117	241	42 Vitamin-D-receptor interacting
cen49	II	15165781	15165914	II	2					
cen50-1	IV	9191246	9191318	II	2					
cen50-2	X	6867890	6867960	II	2					

cen51	II	8075423	8075561	II	2						
cen52	I	2514169	2514312	II	2						
cen53.1	I	5715182	5715255	II	2						
cen53.2	I	5707049	5707122	II	2						
cen54	I	4175998	4176065	II	2	R12E2.3	467	110	177	290	26S proteasome regulatory co
cen55	IV	8927253	8927381	II	2	D1046.1	358	168	296	62	mRNA cleavage factor I subun
cen56	IV	2408514	2408638	II	2	Y38F2AR.12	1229	1000	1124	105	hydrolase activity;
cen57	I	8232530	8232601	II	2						
cen58	II	6248663	6248801	II	2	T24H7.2	284	132	270	14	oxidoreductase activity;
cen59	I	2935786	2935921	II	2						
cen60	I	6090873	6090940	II	2						
cen61	I	6113814	6113889	II	2						
cen62	II	14617031	14617098	II	2	F19H8.4	220	103	170	50	
cen63	II	7183937	7184005	II	2						
cen64	III	5602005	5602132	II	2	C05D2.6	569	420	547	22	
cen65	I	2410191	2410255	II	2						
cen66	II	8547129	8547257	II	2						
cen67	III	10432221	10432354	II	2						
cen68	X	1891503	1891623	II	2						
cen69	V	10774710	10774816	II	2	D1054.3	594	376	482	112	Suppressor of G2 allele of skp
cen70	V	15876438	15876508	II	2						
cen71	II	13834343	13834417	III	3						
cen72	V	5590674	5590772	III	3						
cen73-1	V	5589880	5590013	III	3						
cen73-2	V	5589550	5589681	III	3						
cen74-1	X	14476786	14476865	III	3						
cen74-2	X	14477464	14477541	III	3						
cen75	III	7278372	7278451	III	3						
cen76	II	14801237	14801314	III	3						
cen77	III	7278633	7278697	III	3	B0361.11	774	655	719	55	transporter activity;
cen107-1	III	5211175	5211474	IV							
cen107-2	III	5031371	5031670	IV							
cen107-3	III	4369436	4369734	IV		B0285.9	702	304	602	100	choline kinase
cen107-4	III	4378887	4379176	IV							
cen100	III	7208534	7208665	V		R151.3	227	29	160	67	structural constituent of ribosor
cen101	V	10354315	10354442	V		K07C5.4	179	35	162	17	Ribosome biogenesis protein -
cen102	II	12737895	12738036	V		Y46G5A.5	252	50	191	61	Phosphatidylinositol synthase
cen103	I	9797298	9797406	V		F26E4.9	305	52	160	145	cytochrome-c oxidase activity;
cen104	V	8224873	8225041	V		K07C11.2	275	43	211	64	ATP binding;protein kinase act
cen105	III	4757909	4758049	V		B0393.3	229	47	187	42	Glucose-induced repressor
cen106	III	3068796	3068873	V		H06I04.4a	229	63	140	89	ribosomal protein S27a
cen108	II	8560616	8560685	V		F54C9.1	586	61	130	456	nucleic acid binding;translation
cen109	V	8503169	8503235	V		ZK994.3	341	232	298	43	peroxidase activity;
cen110	III	794736	794861	V		B0412.4	268	73	198	70	structural constituent of ribosor
cen124	V	11785579	11785670	V		F55A11.6	174	55	146	28	structural constituent of ribosor
cen127	III	9226749	9226887	V		T23G5.1	222	34	172	50	structural constituent of ribosor
cen128	II	11876359	11876604	V		K12D12.1	345	32	277	68	ATP binding;DNA binding;DNA
cen17	III	7337813	7337922	V		F56C9.1	229	56	165	64	structural constituent of ribosor
cen78	I	4585831	4585990	V		D1007.12	252	31	190	62	structural constituent of ribosor
cen79	III	7209325	7209455	V		R151.3	222	23	153	69	structural constituent of ribosor
cen80	III	7208944	7209074	V		R151.3	198	28	158	40	structural constituent of ribosor
cen81	V	10970078	10970203	V		F17C11.9a	191	27	152	39	translation elongation factor ac
cen82	III	6374351	6374484	V		C16A3.9	222	28	161	61	structural constituent of ribosor
cen83	I	10577240	10577464	V		F25H2.11	255	27	251	4	molecular_function unknown;
cen84	IV	653908	654036	V		K11H12.2	222	36	164	58	structural constituent of ribosor
cen85	II	10560624	10560840	V		C06A1.1	255	35	251	4	ATP binding;hydrolase activity
cen86	IV	8378808	8378929	V		D2096.8	185	32	153	32	DNA binding;
cen87	I	111292	111425	V		F53G12.10	249	32	165	84	structural constituent of ribosor
cen88	I	2069886	2070025	V		Y37E3.8a	214	29	168	46	structural constituent of ribosor
cen89	III	7496494	7496571	V		C07H6.5	148	22	99	49	ATP binding;ATP dependent h

cen90	III	794394	794520	V	B0412.4	199	30	156	43 structural constituent of riboso
cen91	III	7706783	7706932	V	C03B8.4	255	38	187	68 nuclear protein with multiple zi
cen92	III	5679326	5679457	V	F54E7.2	229	32	163	66 structural constituent of riboso
cen93	II	8602892	8603022	V	F28C6.7a	173	32	162	11 structural constituent of riboso
cen94	V	8225371	8225566	V	K07C11.2	296	38	233	63 ATP binding;protein kinase act
cen95	I	2070258	2070396	V	Y37E3.8a	232	46	184	48 structural constituent of riboso
cen96	IV	4390087	4390307	V	Y24D9A.4a	296	29	249	47 structural constituent of riboso
cen97	V	10969285	10969502	V	F17C11.9a	326	36	253	73 translation elongation factor ac
cen98	IV	17117888	17117978	V	Y116A8C.35	748	49	139	609 RNA binding;nucleic acid bindi
cen99	III	5679693	5679823	V	F54E7.2	167	32	162	5 structural constituent of riboso
cen111	IV	7574112	7574181	VI					
cen112	IV	13601611	13601700	VI					
cen113	I	11617368	11617498	VI					
cen115	IV	13709245	13709403	VI					
cen16-1	II	5543566	5543673	VI					
cen16-2	II	5844827	5844934	VI					
cen16-3	II	5544134	5544241	VI					
cen16-4	I	4172207	4172314	VI					
cen19	I	13305927	13306033	VI					
cen20	IV	9130710	9130817	VI					