C. elegans noncoding RNA genes[*]

Shawn L. Stricklin,[1] Sam Griffiths-Jones,[2] Sean R. Eddy[1,§]

[1]Howard Hughes Medical Institute and Department of Genetics, Washington University, St. Louis, MO 63108 USA

[2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Table of Contents

Abstract

The *C. elegans* genome contains approximately 1300 genes that produce functional noncoding RNA (ncRNA) transcripts. Here we describe what is currently known about these ncRNA genes, from the perspective of the annotation of the finished genome sequence. We have collated a reference set of *C. elegans* ncRNA gene annotation relative to the WS130 version of the genome assembly, and made these data available in several formats.

[§]To whom correspondence should be addressed. E-mail: eddy@genetics.wustl.edu

1. Introduction

The *C. elegans* genome contains approximately 1300 genes that are known to produce functional noncoding RNA (ncRNA) transcripts, as opposed to mRNAs that encode proteins. These known ncRNA genes include about 590 transfer RNA (tRNA) genes, 275 ribosomal RNA (rRNA) genes, 140 trans-spliced leader RNA genes, 120 microRNA (miRNA) genes, 70 spliceosomal RNA genes, and 30 snoRNA genes.

Based on what is known about ncRNA-directed functions in other animals, there are additional ncRNA genes performing known biochemical functions that have not yet been identified in the worm genome. These include the telomerase RNA and on the order of 100-200 small nucleolar RNA (snoRNA) genes that direct site-specific 2'-O-ribose methylations and pseudouridylations of ribosomal RNAs and other target RNAs.

It also seems likely that novel ncRNA genes remain to be discovered. The belated realization that the *lin-4* and *let-7* regulatory RNA genes are not just worm-specific anecdotes, but instead are members of a huge gene family of microRNAs with important roles in posttranscriptional gene regulation in many eukaryotes (Lee et al., 1993; Lim et al., 2003; Pasquinelli et al., 2000; Reinhart et al., 2000) was a spectacular demonstration that important genes (indeed, whole gene families) can easily escape standard computational and experimental gene discovery methods. There is a tantalizing possibility that the miRNAs foreshadow the discovery of even more RNA-directed functions.

Here we describe what is currently known about the ncRNA genes of *C. elegans*, from the perspective of the annotation of the finished genome sequence (*C. elegans* Sequencing Consortium, 1998). Based on the literature, Genbank, and on computational searches for homologs of known RNAs and members of known RNA gene families (Benson et al., 2004; Griffiths-Jones, 2004; Griffiths-Jones et al., 2003; Harris et al., 2004; Lowe and Eddy, 1997), we have collated a stable, curated reference set of *C. elegans* noncoding RNA genes, and their chromosomal coordinates relative to the WS130 version of the genome sequence assembly in Table 1. We have made these data available as annotation tracks for WormBase, and downloadable as HTML tables, GFF coordinate files, or FASTA sequence files. We describe how the reference set has been produced, and summarize what it contains.

2. Ribosomal RNAs

The 18S, 5.8S, and 26S subunits of rRNA are transcribed by RNA polymerase I from a 7.2 kb rDNA unit that is tandemly repeated ~55 times at the end of chromosome I (*C. elegans* Sequencing Consortium, 1998; Ellis et al., 1986; Sulston and Brenner, 1974). The 5S rRNAs are transcribed separately by RNA polymerase III from ~110 copies of a ~1 kb tandem repeat unit on chromosome V (Nelson and Honda, 1985; Sulston and Brenner, 1974). The 5S rRNA repeat unit also includes the gene for the SL1 trans-spliced leader; see below. The rRNA genes are systematically underrepresented in the current genome sequence assembly, because tandem arrays are problematic for physical mapping and sequencing. According to WUBLAST searches using the published sequences of the 7.2 kb and 1 kb rDNA repeat units as queries, one copy of the 18S/5.8S/26S rRNA repeat unit is represented in the chromosome I sequence assembly, and fifteen copies of the 5S rRNA gene are included in the chromosome V sequence. Additionally, the mitochondrial DNA contains one 18S rRNA gene and one 23S rRNA gene.

3. Transfer RNAs

We have annotated genes for 569 nuclear tRNAs, 22 mitochondrial tRNAs, and 1072 probable tRNA pseudogenes. The mitochondrial tRNAs were curated from the literature (Okimoto et al., 1992; Wolstenholme et al., 1987). Nuclear tRNAs can be reliably identified by computational methods. We used the programs tRNAscan-SE (Lowe and Eddy, 1997) and ARAGORN (Laslett and Canback, 2004) to identify a combined candidate list of 612 putative tRNA genes and 214 candidate tRNA pseudogenes. These candidate tRNA genes were manually curated to remove an additional 40 putative pseudogenes and 3 false positives, leaving the final set of 569 annotated genes. This gene set is essentially in agreement with the independent analysis of Marck and Grosjean, who identified 529 putative tRNA genes (Marck and Grosjean, 2002). Differences appear to be due to variation in what is called a putative true gene versus a putative pseudogene, and differences in the version of the genome assembly used. As is the case in many eukaryotes, tRNA pseudogenes are numerous in *C. elegans*. Current tRNA scanning programs only detect pseudogenes that are closely related to true tRNAs. Using WUBLAST, we identified an additional 818 sequences with significant similarity to one or more of the 569 tRNA genes and/or 254 tRNA pseudogenes, and added them to annotate a total of 1072 putative pseudogenes. Many of these overlay four previously identified repetitive sequences (Tc4, CEREP3, CELE45, and NDNAX3_CE) defined by RepBase and RepeatMasker searches.

4. Spliced leader RNAs

Approximately 70% of *C. elegans* mRNAs are covalently modified at their 5' end by the addition of 22-nt trans-spliced leader RNA sequences (Blumenthal and Gleason, 2003; Ross et al., 1995; Zorio et al., 1994). Trans-spliced leaders are donated by independently transcribed ~100-110 nt SL RNAs, which come in two forms, SL1 RNA and SL2 RNA (see Trans-splicing and operons). The most abundant form, SL1, is predominantly trans-spliced to the 5' end of pre-mRNAs, including the first cistron in polycistronic (operon) pre-mRNAs; the rarer form, SL2, is generally trans-spliced to downstream cistrons in polycistronic operons (Blumenthal and Gleason, 2003). The genes for SL1 RNA are part of the same tandem repeat unit that encodes 5S rRNA, occurring in ~110 copies on chromosome V (Krause and Hirsh, 1987; Nelson and Honda, 1985). Ten SL1 RNA genes are represented in the current genome sequence assembly. The genes for the ~110 nt SL2 RNAs are dispersed, and show more sequence variation than SL1 RNAs (indeed, some have been named SL3 RNA, SL4 RNA, etc.; we follow the WormBase convention of annotating all as SL2 variants; Huang and Hirsh, 1989; Ross et al., 1995; Zorio et al., 1994). 20 SL2 RNA gene variants are found in the genome sequence, roughly in agreement with the copy number of ~30 predicted from genomic Southerns (Ross et al., 1995; Zorio et al., 1994). SL RNAs are thought to be transcribed by pol II (Krause and Hirsh, 1987).

5. Small nucleolar RNAs (snoRNAs)

In Eukarya and Archaea, two classes of snoRNAs direct site-specific base modifications of ribosomal RNA and other ncRNAs. C/D box snoRNAs direct 2'-O-ribose methylations, and H/ACA snoRNAs direct pseudouridylations. The modifications are catalyzed by snoRNA-associated proteins (fibrillarin in the case of 2'-O-ribose methylation, Cbf5/dyskerin in the case of pseudouridylation); the snoRNAs function to guide a snoRNP complex to a modification site by complementary base pairing between the snoRNA and the target RNA. One snoRNA usually targets one or two modifications, and any given modification site can be targeted by more than one redundant snoRNA. The best-studied targets for snoRNA-directed base modification, ribosomal RNAs, have about 40-100 2'-O-ribose methylations and a comparable number of pseudouridylations in eukaryotes such as yeast and human. We expect a similar number of modifications in *C. elegans* rRNAs, and therefore expect to find on the order of 100-200 rRNA modification guide snoRNAs in the genome (and probably additional snoRNAs that direct modifications of other ncRNAs).

However, snoRNAs are difficult to identify from sequence analysis alone, and we have only a partial map of *C. elegans* rRNA modifications (Higa et al., 2002), so the catalog of *C. elegans* snoRNA genes is currently incomplete. We can annotate 19 C/D and 6 H/ACA snoRNA genes based on Genbank and the current literature (Higa et al., 2002; Wachi et al., 2004).

In vertebrates, almost all known C/D and H/ACA snoRNAs are processed out of introns of host genes, whereas in yeast and plants, snoRNAs are usually independently transcribed (Tycowski et al., 2004). *C. elegans* likely uses a mix of the two expression strategies. Many of the 2'-O-ribose methylation-guide C/D box snoRNAs are preceded by conserved proximal sequence elements (PSEs) typical of snRNA promoters (Thomas et al., 1990). The H/ACA snoRNAs identified by Wachi et al. were found in the introns of protein-coding genes (Wachi et al., 2004).

A small number of additional snoRNAs do not direct nucleotide modifications. An important one is the U3 snoRNA, which is involved in endonucleolytic processing of the 18S/5.8S/26S rRNA precursor. Six U3 snoRNA genes have been identified by sequence similarity searches (TA Jones, SR Eddy, unpublished).

6. microRNAs (miRNAs)

We annotated both the putative Drosha-produced pre-miRNA (~60nt) and the Dicer-produced miRNA (~21nt) forms of each of 117 microRNAs in the Rfam miRNA Registry database (Griffiths-Jones, 2004). We also annotated 37 of the ~21-nt "tiny noncoding RNA" (tncRNA) sequences identified by Ambros et al., which appear to be produced in a Dicer-dependent fashion and thus are related to the miRNA/siRNA pathways (Ambros et al., 2003). See *C. elegans* microRNAs for a thorough description of *C. elegans* microRNA biology.

We have not annotated the ~700 endogenous small interfering RNAs (siRNAs) found by Ambros et al. (Ambros et al., 2003). These ~21 nt siRNAs exhibit perfect complementarity to coding regions, and thus may represent post-transcriptional gene regulation by endogenous RNAi. For a description of the RNAi phenomenon in *C. elegans*, see RNAi mechanisms.

7. Other known RNAs

The spliceosome contains five different small nuclear RNAs (snRNAs), called U1, U2, U4, U5, and U6. Genes for the abundant spliceosomal RNAs typically occur in multiple copies in higher eukaryotes. The *C. elegans* spliceosomal RNA genes were studied by Thomas et al., who identified 18 spliceosomal genes (1 U1, 10 U2, 3 U4, 2 U5, and 2 U6) dispersed in the genome, and who estimated copy numbers based on genomic Southerns of 11 U1, 12 U2, 6 U4, 9 U5, and 10 U6 snRNA genes (Thomas et al., 1990). Based on WUBLASTN searches using the Thomas et al. sequences as queries, we identify a total of 72 spliceosomal RNA loci (12 U1, 19 U2, 5 U4, 13 U5, and 23 U6), at least five of which appear to be pseudogenes.

Unlike plants and some other animals, *C. elegans* does not appear to utilize the alternative U12 spliceosome components U11, U12, U4atac, and U6atac; their homologs are not detectable in the *C. elegans* genome and no ESTs containing characteristic AT/AC splice-junctions have been found.

RNase P RNA, which catalyzes the maturation of tRNA 5' ends, is present in a single copy in the *C. elegans* genome; it was identified by computational homology searches (Klein, 2003; S.M. Marquez and N.R. Pace, unpublished).

The ceY RNA is also present in a single genomic locus. The function of its RNP complex with Ro protein is unclear, but may be to act as a 5S rRNA quality control agent (Van Horn et al., 1995).

SRP RNA is a component of the signal recognition particle, involved in translocation of nascent proteins across the endoplasmic reticulum. Five SRP-RNA genes have been identified in the *C. elegans* genome by homology search methods (Regalia et al., 2002).

In other metazoans, the 3' ends of histone genes are processed via the interaction of a downstream stemloop with the U7 ncRNA. To date, no *C. elegans* U7 RNA homologs have been found. The 7SK RNA, a negative regulator of pol II transcription in higher mammals, also has no detectable homolog in the *C. elegans* genome.

The most important ncRNA that has yet to be identified in the *C. elegans* genome is the telomerase RNA. Similar to almost all other eukaryotes, *C. elegans* telomeres consist of 1-2 kilobases of six-nucleotide (TAAGGC) repeats (Cheung et al., 2004), which in ciliates, vertebrates, and fungi are known to be synthesized by a telomerase RNP that contains a reverse transcriptase and a telomerase RNA template, in addition to other proteins (Blackburn, 2001). Telomerase RNAs are highly diverged, and we do not expect to be able to identify a homologous RNA in *C. elegans* by BLAST searches using known telomerase RNAs as queries (and indeed, we do not). More powerful structure-based searches might be productively applied now that a satisfactory consensus secondary structure for ciliate, vertebrates, and fungi has been worked out (Lin et al., 2004). Using 3' SAGE tagging, Jones et al. detected an abundant, apparently noncoding transcript *tts-1* which was predicted to contain a plausible telomeric template sequence, leading *tts-1* to be suggested as a possible telomerase RNA (Jones et al., 2001). However, this enigmatic transcript seems unlikely to be the true telomerase RNA, as cDNA sequencing has shown that the putative telomeric template region is not part of the major *tts-1* transcript (SL Stricklin and SR Eddy, unpublished).

8. ncRNA conservation

Most known *C. elegans* ncRNAs are well conserved with their *C. briggsae* homologs: similarity ranges from complete identity (for 5S rRNA, U6 RNA, and SL1 RNA) to 89% identical for RNase P and 75% identical for the ceY RNA. About two-thirds of *C. elegans* miRNAs are clearly conserved in *C. briggsae*, but none of the 37 annotated *C. elegans* tncRNAs (or the immediate genomic context of these ~21mers) appear to be conserved in *C. briggsae*.

9. Prospects for novel ncRNAs

It must be noted that it is not currently possible to systematically identify novel noncoding RNA genes that have no homology to known genes. We do not yet have reliable "genefinding" programs for ncRNAs; current computational ncRNA genefinding approaches are suitable as screens, but not for reliable automated annotation. The best current computational methods use comparative sequence analysis to identify conserved RNA secondary structure (Coventry et al., 2004; Rivas and Eddy, 2001; Washietl et al., 2005). Genome sequencing of several related *Caenorhabditis* species is beginning to make such comparative ncRNA genefinding approaches increasingly powerful in *C. elegans*.

Systematic gene discovery by cDNA sequencing (and related transcript discovery methodologies such as SAGE and MPSS) is usually done on poly-A+, mRNA-enriched, ncRNA-depleted populations. It is possible to use these transcript discovery approaches on total RNA or ncRNA-enriched RNA populations (Huttenhofer et al., 2004), but to our knowledge such experiments have not yet been published for *C. elegans*. We also expect that whole-genome tiled expression arrays should soon make it possible to systematically catalog the entire *C. elegans* transcriptome, including both coding and noncoding transcripts (Kampa et al., 2004; Kapranov et al., 2003).

10. Acknowledgments

11. References

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr. Biol. *13*, 807–818. Abstract Article

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res 32 Database issue, D23–D26. Abstract

Blackburn, E.H. (2001). Switching and signaling at the telomere. Cell *106*, 661–673. Abstract Article

Blumenthal, T., and Gleason, K.S. (2003). *Caenorhabditis elegans* operons: form and function. Nat. Rev. Genet. *4*, 112–120. Abstract Article

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium. Science *282*, 2012–2018. Abstract Article

Cheung, I., Schertzer, M., Baross, A., Rose, A.M., Lansdorp, P.M., and Baird, D.M. (2004). Strain-specific telomere length revealed by single telomere length analysis in *Caenorhabditis elegans*. Nucleic Acids Res. *32*, 3383–3391. Abstract Article

Coventry, A., Kleitman, D.J., and Berger, B. (2004). MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. Proc. Natl. Acad. Sci. USA *101*, 12102–12107. Abstract Article

Ellis, R.E., Sulston, J.E., and Coulson, A.R. (1986). The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res. *14*, 2345–2364.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32* Database issue, D109–D111. Abstract Article

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic Acids Res. *31*, 439–441. Abstract Article

Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., et al. (2004). WormBase: a multi-species resource for nematode biology and genomics. Nucleic Acids Res. 32 Database issue, D411–D417. Abstract Article

Higa, S., Maeda, N., Kenmochi, N., and Tanaka, T. (2002). Location of 2(')-O-methyl nucleotides in 26S rRNA and methylation guide snoRNAs in *Caenorhabditis elegans*. Biochem Biophys Res. Commun. *297*, 1344–1349. Abstract Article

Huang, X.Y., and Hirsh, D. (1989). A second trans-spliced RNA leader sequence in the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA *86*, 8640–8644. Abstract

Huttenhofer, A., Cavaille, J., and Bachellerie, J.P. (2004). Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms. Methods Mol. Biol. *265*, 409–428. Abstract

Jones, S.J., Riddle, D.L., Pouzyrev, A.T., Velculescu, V.E., Hillier, L., Eddy, S.R., Stricklin, S.L., Baillie, D.L., Waterston, R., and Marra, M.A. (2001). Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. Genome Res *11*, 1346–1352. Abstract Article

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. Abstract Article

Kapranov, P., Sementchenko, V.I., and Gingeras, T.R. (2003). Beyond expression profiling: next generation uses of high density oligonucleotide arrays. Brief Funct Genomic Proteomic *2*, 47–56. Abstract

Klein, R.J. (2003). Finding noncoding RNA genes in genomic sequences. Ph.D. Thesis, Washington University in St. Louis.

Krause, M., and Hirsh, D. (1987). A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell *49*, 753–761. Abstract Article

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16. Abstract Article

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4*encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843–854. Abstract Article

Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C.B., and Bartel, D.P. (2003). The microRNAs of *Caenorhabditis elegans*. Genes Dev. *17*, 991–1008. Abstract Article

Lin, J., Ly, H., Hussain, A., Abraham, M., Pearl, S., Tzfati, Y., Parslow, T. G., and Blackburn, E.H. (2004). A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein. Proc. Natl. Acad. Sci. USA *101*, 14713–14718. Abstract Article

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964. Abstract Article

Marck, C., and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA *8*, 1189–1232. Abstract Article

Nelson, D.W., and Honda, B.M. (1985). Genes coding for 5S ribosomal RNA of the nematode *Caenorhabditis elegans*. Gene *38*, 245–251. Abstract Article

Okimoto, R., Macfarlane, J.L., Clary, D.O., and Wolstenholme, D.R. (1992). The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. Genetics *130*, 471–498. Abstract

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M. I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., et al. (2000). Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature *408*, 86–89. Abstract Article

Regalia, M., Rosenblad, M.A., and Samuelsson, T. (2002). Prediction of signal recognition particle RNA genes. Nucleic Acids Res. *30*, 3368–3377. Abstract Article

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901–906. Abstract Article

Rivas, E., and Eddy, S.R. (2001). Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics *2*, 8. Abstract Article

Ross, L.H., Freedman, J.H., and Rubin, C.S. (1995). Structure and expression of novel spliced leader RNA genes in *Caenorhabditis elegans*. J. Biol. Chem. *270*, 22066–22075. Abstract Article

Sulston, J.E., and Brenner, S. (1974). The DNA of *Caenorhabditis elegans*. Genetics *77*, 95–104. Abstract

Thomas, J., Lea, K., Zucker-Aprison, E., and Blumenthal, T. (1990). The spliceosomal snRNAs of *Caenorhabditis elegans*. Nucleic Acids Res. *18*, 2633–2642. Abstract

Tycowski, K.T., Aab, A., and Steitz, J.A. (2004). Guide RNAs with 5' Caps and Novel Box C/D snoRNA-like Domains for Modification of snRNAs in Metazoa. Curr. Biol. *14*, 1985–1995. Abstract Article

Van Horn, D.J., Eisenberg, D., O'Brien, C.A., and Wolin, S.L. (1995). *Caenorhabditis elegans* embryos contain only one major species of Ro RNP. RNA *1*, 293–303. Abstract

Wachi, M., Ogawa, T., Yokoyama, K., Hokii, Y., Shimoyama, M., Muto, A., and Ushida, C. (2004). Isolation of eight novel *Caenorhabditis elegans* small RNAs. Gene *335*, 47–56. Abstract Article

Washietl, S., Hofacker, I.L., and Stadler, P.F. (2005). Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. USA *102*, 2454–2459. Abstract Article

Wolstenholme, D.R., Macfarlane, J.L., Okimoto, R., Clary, D.O., and Wahleithner, J.A. (1987). Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. Proc. Natl. Acad. Sci. USA *84*, 1324–1328. Abstract

Zorio, D.A., Cheng, N.N., Blumenthal, T., and Spieth, J. (1994). Operons as a common form of chromosomal organization in *C. elegans*. Nature *372*, 270–272. Abstract Article