Genome Sequence of Additional *Caenorhabditis* species: Enhancing the Utility of *C. elegans* as a Model Organism

Paul W. Sternberg – HHMI/California Institute of Technology/WormBase
Robert H. Waterston – University of Washington
John Spieth – Washington University /WormBase
Sean Eddy – HHMI/Washington University
Richard K. Wilson – Genome Sequencing Center/Washington University

October 10, 2003

Overview

The value of *Caenorhabditis elegans* as a model organism for biomedical research is unquestioned. The interpretation of its 100.2 Mb complete genome sequence has been enhanced by a high quality, 98% complete draft genome sequence for *Caenorhabditis briggsae*. However, comparison of any two species is not sufficient to define many sequence features, since features evolve at different rates, are of differing sizes and are detected with varying ease by a variety of tools. To allow better genome alignment, gene interpretation, promoter analysis, identification of non-coding RNAs (ncRNAs) and other functional features, as well as to explore the forces that mold these genomes, we propose sequencing the genomes of three additional *Caenorhabditis* species: *Caenorhabditis remanei, Caenorhabditis* n. sp. CB5161 and *Caenorhabditis japonica*. The first two are about the same evolutionary distance from *C. elegans* as is *C. briggsae*, while *C. japonica* is the closest outgroup to these four. Pilot studies that compare the sequence of known transcriptional enhancers from four species demonstrate that addition of *C. remanei* and CB5161 add great value when combined with *C. briggsae* data in the analysis of the *C. elegans* genome. Marginally conserved sequences between *C. elegans* and *C. briggsae* can be given more attention if also conserved in additional *Caenorhabditis* species or less attention if not conserved. Other functional elements may emerge from statistical noise in a four or five-way comparison and additional orthology will be determined, all of which will enhance and expand the value of *C. elegans* as a model for understanding human health and disease, and basic biological processes. In addition, these genomes will enable a variety of evolutionary genetic analyses in part because they encode male-female sexual systems while *C. briggsae* and *C. elegans* have hermaphrodite-male reproduction.

Background

C. elegans has been a major model system for basic biological and biomedical research. It is the first animal for which a complete description of its anatomy, development and neural wiring diagram exists (reviewed by Wood, 1988; Riddle, 1997). It was the first multi-cellular organism to have its genome sequenced (The *C. elegans* Sequencing Consortium, 1998; www.wormbase.org), an achievement that has stimulated development of powerful downstream genomic resources such as genome-wide gene inactivation by RNAi (Kamath et al., 2003) and gene expression mapping (Kim et al., 2001). *C. elegans* has a superbly annotated genome, which along with our comprehensive knowledge of its proteome, anatomy, development and behavior make it a useful testbed for methodological developments in genomics and functional genomics. What we learn from analyzing *C. elegans* will help in similar analyses of the human genome.

The relatively small size of *Caenorhabditis* genomes (~100 Mb total genome sequence) make this organism and its close relatives attractive for genomics. Moreover, *Caenorhabditis* species have short generation time (3-5 days) and strains can be frozen for decades and recovered, thereby facilitating intense genetic analysis. The worm's transparency and its small, invariant anatomy are assets for analysis of gene expression and function at the level of individual,

defined cells. The expression pattern of >1600 genes is known (www.wormbase.org), and gene functions have been assigned to individual cells, e.g., the ODR-10 olfactory receptor is expressed and acts in the two AWA sensory neurons (Sengupta et al., 1996).

The value of *C. elegans* as a model organism for understanding human health and disease has long been recognized (reviewed by Ahringer, 1997). Over half of *C. elegans* genes have human orthologs, while ~42% of human disease genes have a homolog in *C. elegans* (Culetto and Sattelle, 2000). In spite of its relatively simple anatomy, many of the cell types associated with complex mammalian functions such as intestine, neurons, muscle and excretory cells can be recognized in *C. elegans*. Furthermore, *C. elegans* has contributed to, and is often at the forefront of, our understanding of fundamental biological processes such as programmed cell death, RNAi, aging, signaling pathways, axon guidance, cell polarity, sex determination and synaptic signaling. The normal functions of disease genes involved in cancer, polycystic kidney disease, torsion dystonia, and mucolipidosis, among others, were elucidated in part based on studies in *C. elegans*.

The *C. elegans* community comprises over 2000 researchers in 463 laboratories worldwide. Many other researchers take advantage of *C. elegans* in computational analyses and occasional experiments. There are annual meetings (an international meeting in odd years; regional East Coast, West Coast, Midwest, European and Asian in even years). Over 1500 researchers attended the 2003 International *C. elegans* Meeting. Over 700 *C. elegans* papers were published last year (2002); the annual output has been increasing steadily. In addition, many more papers use *C. elegans* sequence for comparative purposes. The community is well organized with a genome database, WormBase (www.wormbase.org/), which will store and display the data from this project (P.S. and J.S. are two of the four WormBase PIs); the *Caenorhabditis* Genetics Center (http://biosci.umn.edu/CGC/ CGChomepage.htm), which freezes, stores and distributes strains (including *C. remanei,* CB5161 and *C. japonica*); the ORFeome project (http://worfcb.dfci.harvard.edu/), which has generated open reading frame clones for most *C. elegans* genes; the Gene Knockout Consortium, which generates deletion alleles of genes by user request, the WormAtlas (www.wormatlas.org), which displays information about anatomy; and the Structural Genomics Initiative, which seeks to determine the crystal structure of *C. elegans* proteins. There have been three community-generated books describing the state of knowledge of *C. elegans* (Wood, 1988; Epstein and Shakes, 1995; Riddle, 1997), and a new, open access electronic book tied to WormBase is planned (M. Chalfie, pers. comm.). Finally, there exists tight control over nomenclature such as gene, allele, strain and cell names, greatly facilitating database management and utility.

Utility of *C. briggsae* sequence for improving *C. elegans* gene structures

Accurate knowledge of the *C. elegans* gene set is crucial since every known *C. elegans* gene is being studied at least by RNAi, microarrays, and yeast-two-hybrid studies and for computational studies (e.g., Reboul et al., 2003). The *C. briggsae* sequence is being used to refine the 19,936 *C. elegans* genes plus 2232 alternately spliced variants (WormBase version WS109) and to predict new ones. An extrapolation from careful manual curation of several regions of colinearity between *C. elegans* and *C. briggsae* suggests that there may be 1300 missing *C. elegans* genes, and over 2800 existing exons in need of extension or truncation in the predicted gene set in WormBase version WS77 (Stein *et al*, 2003). Taking a different approach, preliminary results using TWINSCAN (Korf *et al*, 2001), which predicts genes based on two genomic sequences, identified ~3000 potential new genes, of which 20% have recognizable Pfam (pfam.wustl.edu) domains suggesting that >600 of these may be truly new genes (J. Spieth, unpublished observations). Many of the remaining 80% may as well be real genes, but clearly, conservation in additional species would be a strong filter for these predictions.

All of these new and discrepant genes will have to be manually inspected, but sequence

conservation with *C. briggsae* clearly has become a persuasive predictor for *C. elegans* gene structures. This will be especially so for genes expressed at low levels in single cells, or for rare alternatively-spliced transcripts where EST and mRNA data are not available. *C. briggsae* has already proven useful in identifying alternative exons including previously unrecognized 5' exons. For example, D. Sherwood and P. Sternberg (unpublished) identified a nucleotide substitution associated with a mutant allele of the *evl-5* locus that was not in an exon of the existing gene structure. A *C. briggsae-C. elegans* comparison narrowed nine potential exons to three, one of which was confirmed by RNA analysis. (Additional species likely would have narrowed this to the true one.) Similarly, comparison of *C. elegans* and *C. briggsae* supported a potential alternative 5' exon (and hence an alternative promoter) in the *lin-3* EGF-like growth factor gene that was then verified by RNA analysis (Liu et al., 1999).

Another example of the utility of the *C. briggsae* sequence can be found in G-protein coupled receptor (GPCR) gene families, which are very poorly represented in EST collections. Detection of GPCRs is usually achieved by TBLASTN searches, but in this rapidly evolving family, precise gene structure is often hard to define. Sequence conservation between *C. elegans* and *C. briggsae* for orthologous genes was often the only way to identify exon/intron boundaries (H. Robertson, personal communication). Nonetheless, considerable species-specific evolution between *C. elegans* and *C. briggsae* is found in these families, such that orthologous relationships are difficult to determine with just two species. Robertson estimates that up to 25% of some GPCR families have undergone duplication since the divergence of *C. elegans* and *C. briggsae.* However, without an appropriate outgroup for such comparisons, it is not possible to make orthology determinations (see Eisen, 1998).

Utility of *C. briggsae* to define regulatory sequences

One major use of *C. briggsae* sequence has been for the identification of regulatory sequences. The use of *C. briggsae* for cis-regulatory sequence comparison is now standard (e.g., Zucker-Aprison and Blumenthal, 1989; Prasad and Baillie, 1989; Kennedy et al., 1993; Thacker et al., 1999; Aamodt et al., 2000; Marshall and McGhee, 2001; Cui and Han, 2003). A first step in defining important regulatory regions is a comparison of *C. elegans* and *C. briggsae* between orthologous genes. For example, prior to the release of the *C. briggsae* sequence, Kirouac and Sternberg (2003) used a deletion analysis to identify regions of three *C. elegans* genes sufficient to direct expression in particular cells. When the *C. briggsae* sequence became available, a retrospective analysis indicated that the *C. elegans-C. briggsae* comparison correctly identifies at a resolution of ~100 nucleotides all the regions identified by deletion analysis, thus greatly accelerating the identification of functional regulatory regions. A more systematic attempt at identifying candidate regulatory elements in conserved, non-coding sequences is underway as a part of the more global analysis of the two genomes. Nonetheless, these comparisons will be hampered by the relatively weak signals combined with the considerable noise, resulting in large numbers of both false positives and false negatives.

Utility of *C. briggsae* for finding non-coding RNAs

Micro-RNAs were discovered in *C. elegans* (Lee et al., 1993; Lee and Ambros, 2001; Lau et al., 2001), and the availability of *C. briggsae* sequence has been very helpful in identifying the now large family of micro-RNAs in *Caenorhabditis* (Lim et al, 2003).

Regulatory elements and non-coding RNAs: *C. remanei* and CB5161

The central challenge faced in exploiting the *C. elegans* sequence today is the complete and accurate identification of gene regulatory elements and non-coding RNA genes. Protein coding genes are actually reasonably well described today based on a combination of *ab initio* methods, experimental data, particularly EST and cDNA sequences (for about half the genes) and the refinement provided by the *C. briggsae* sequence in defining exon boundaries, in detecting alternative exons and even in overall gene structure. The identification of transcriptional regulatory elements and non-coding RNA genes is, by contrast, in its infancy.

The *C. briggsae* sequence reveals many conserved features that are not protein coding (only about half the aligned sequences between the two genomes fall into coding exons). With only two genomes, current computational methods only partially help unravel their function. Several recent studies have highlighted the utility of additional genomes for finding regulatory regions and non-coding RNAs (Boffelli et al., 2003; Cliften et al., 2003; Kellis et al., 2003; McCutcheon and Eddy, 2003). Ideally one would want sequences that are as divergent as possible yet readily alignable.

Any comparative genome sequencing proposal for *C. elegans* should then have as a goal the elucidation of these two poorly recognized features (in addition to the more general goal of understanding the evolutionary mechanisms that drove the formation of these species) and should exploit the available species. Figure 1 illustrates the phylogenetic relationship of the known *Caenorhabditis* species. In view of this tree and the two principal goals outlined above, we think that *C. remanei* and CB5161 will have the most immediately useful genomes, with *C. japonica* acting as an important outgroup. We illustrate in next paragraphs the basis for these choices and what might be achieved with some examples.

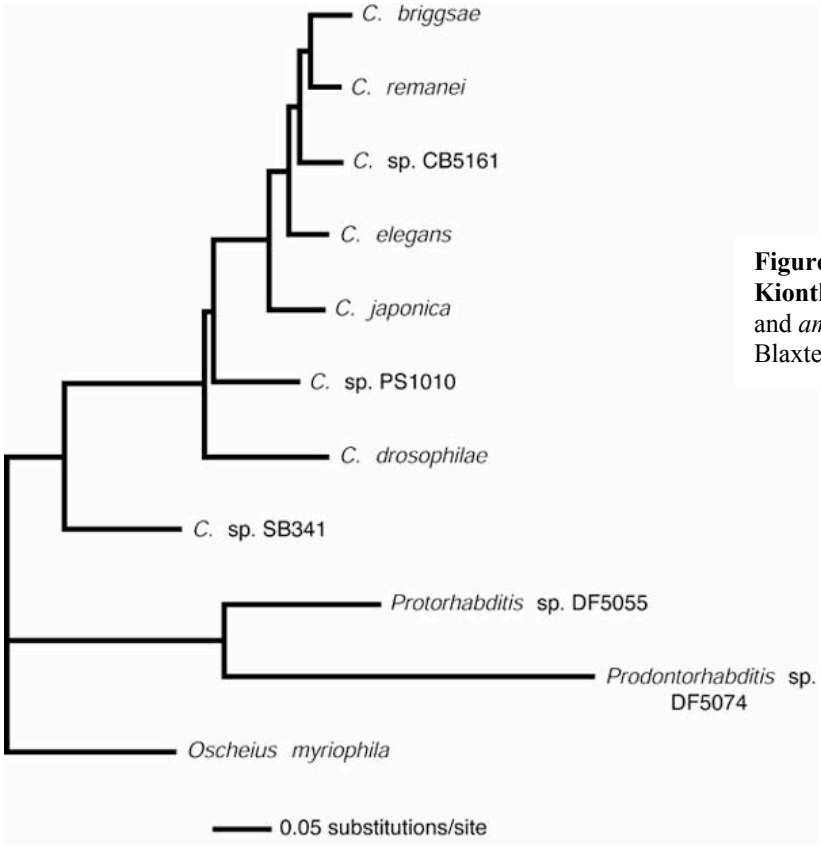

Figure 1. Phylogeny from D. Fitch and K. Kiontke (pers. comm.). Based on 18S, 28S and *ama-1* (RNA Pol II) sequences. See also Blaxter et al., (1998).

Orthologs for eight genes have been cloned and sequenced for unnamed species represented by strains CB5161 and PS1010 (E. Schwarz, J. DeModena, E. Moon, H. Shizuya, B. Wold and P. Sternberg, unpublished observations). For non-coding sequences it is straightforward to align CB5161 but not PS1010 with *C. elegans* and *C. briggsae*. *C. remanei* has been shown to align well to *C. elegans* (e.g., Xue et al., 1992). In the future, better algorithms may allow the analysis of species as divergent from *C. elegans* as PS1010 but as this is not now the case, we reject PS1010 as a candidate for sequencing at this time.

There is some intergenic sequence available for both CB5161 and *C. remanei*. For the anchor

cell regulatory element of *lin-3* (Figure 2), the four way *C. elegans-C. briggsae-C. remanei*-CB5161 indicates the added utility of *C. remanei* and CB5161. A striking example is that one element (B) conserved between *C. elegans* and *C. briggsae* has been shown to not be important in enhancer function based on gel shift and site-directed mutagenesis experiments, while elements A, C and D are functionally important (Hwang and Sternberg, 2003). Site B drops from consideration with the three- and four-way comparisons. Addition of either *C. remanei* or CB5161 helps in this case, and together make the case compelling. For other short sequences, there is divergence in either CB5161 or *C. remanei* but not both. Thus, the availability of both additional sequences will be of immense value.

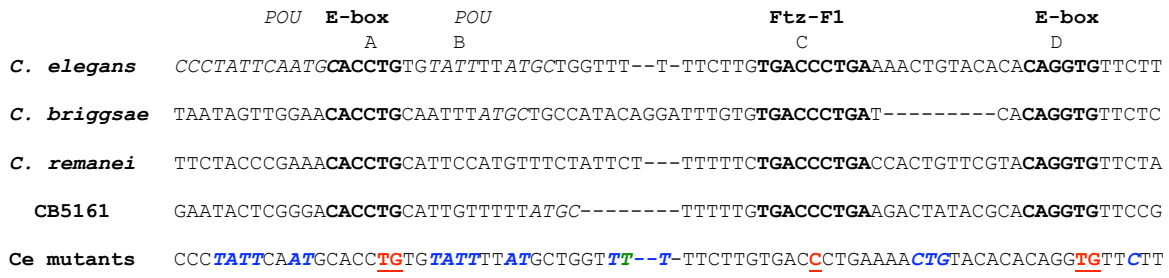

Figure 2. Four-way alignment of the anchor cell enhancer of *lin-3*. A 59 bp enhancer element from the *lin-3* EGF-like ligand encoding gene of *C. elegans* directs expression to a basal promoter specifically in the anchor cell (Hwang & Sternberg, 2003). Three functional elements (two E-boxes and one Ftz-F1 site) within this enhancer have been defined by site-directed mutagenesis experiments ("Ce mutants"): blue italics indicates mutations with no effect; red underlined indicates the mutation eliminated expression. Each species contributes information. The *C. elegans-C. briggsae* comparison indicated a potential POU domain binding site; this is not important for enhancer function, and addition of the other two species does not support this site. *C. remanei* and CB5161 sequence is from Schwarz et al. (unpublished).

An exemplary quantitative analysis of the utility of the additional sequences can be seen in Table 1. Hexamers corresponding to the two known functional motifs in the *lin-3* anchor cell enhancer and one functional motif in the *lin-11* uterine enhancer (Gupta and Sternberg, 2002) were correctly identified *ab initio* by YMF/Explanators (Blanchette and Sinha, 2001; Sinha and Tompa, 2002) from *C. elegans*, *C. briggsae*, *C. remanei* and CB5161. Both *C. remanei* and CB5161 contribute to the identification of or confidence in a site; *C. remanei* was more important for the E-box while CB5161 was more important for the Ftz-F1 and Su(H)/LAG-1 binding sites. This conclusion matches that drawn from inspection of sequence alignments in which a given position might be divergent in only one of the four species. By contrast, addition of PS1010 to the *lin-3* analysis improved neither the 4-way or the 2-way predictions

Motif	**z-scores for hexamer with YMF/Explanators**			
	4-way (+CB5161+*C. remanei*)	**3-way (+ *C. remanei*)**	**3-way (+CB5161)**	**2-way**
CACCTR (~E-box)	10.5	8.5	Not Found	Not Found
AGGGTY (~Ftz-F1)	8.9	7.0	7.7	6.2
ATGGGA (~LAG-1)	11.8	9.6	11.1	7.4

Table 1. Quantitative analysis of overrepresented hexamers from 2-4 species. Z-scores represent number of standard deviations from the mean genomic background frequency. For the first two motifs ~650 nucleotides from each species corresponding to the *C. elegans lin-3* intron containing the anchor cell enhancer shown in Figure 2. The third line is from ~450 nucleotides of the *C. elegans lin-11* 5' flanking region containing the uterine enhancer. Software is described by Blanchette and Sinha (2001) and Sinha and Tompa (2002). Analysis by E. Schwarz.

Similarly, an analysis using very different type of software (MUSSA, an n-way version of the Seq Comp/Family Relations program of Brown et al. (2002) which searches for blocks of conserved sequence in sliding windows) also indicates the strong marginal utility of the fourth species over three species (E. Schwarz, et al., pers. comm.). Thus the value of a fourth *Caenorhabditis* species is not software-dependent.

Given that there are some biological differences among the *Caenorhabditis* species, comparing two species can fail to identify conserved sequence elements. Furthermore, multi-species sequence comparisons can point to sequences that have evolved in one species but not others. For example, *C. elegans* differs from the other species in certain features of the expression and function of *lin-48* (Wang and Chamberlin, 2002; and pers. comm.). Alignment of *C. elegans* and *C. briggsae* sequence from a regulatory region that contributes to this difference identifies little sequence that is conserved. By contrast, addition of *C. remanei* and CB5161 sequences aligns several blocks of conserved sequence shared among *C. briggsae*, *C. remanei*, and CB5161, but not *C. elegans*. The four-way comparison identifies potential conserved regulatory elements shared among *C. briggsae*, *C. remanei*, and CB5161, and the potential regulatory divergence of *C. elegans*. These features are missed by any two-way alignment. One important comparison still missing from these studies is that to a gene from an outgroup species. In the *lin-48* comparisons, *C. elegans* is the "outgroup"; because of this, we cannot determine if the evolutionary change in regulation occurred in the lineage to *C. elegans* or in the lineage shared by the other three (more closely related) species.

These genomes will also enable identification of noncoding RNA genes by comparative sequence analysis; for instance, using the Rivas and Eddy QRNA program (Rivas and Eddy, 2001). Experience with ncRNA identification in *S. cerevisiae* and *E. coli* has shown, unsurprisingly, that using multiple comparative genomes instead of one comparative genome increases the power of the analysis; furthermore, that genomes selected for "optimal" distance for coding gene annotation and phylogenetic footprinting of cis-regulatory sites are also optimal for comparative analysis of ncRNA genes (Rivas et al., 2001; McCutcheon and Eddy, 2003).

We conclude that either *C. remanei* or CB5161 will be of high value for identification of sequence motifs, and that availability of both will increase the value of all sequences. The availability of highly inbred strains and background biological knowledge of *C. remanei* (see below) make it of slightly higher priority than CB5161.

A close outgroup: *C. japonica*

An outgroup to the *elegans* species group (*C. briggsae, C. elegans, C. remanei* and. CB5161) is crucial. The closest possible outgroup will allow the best rooting of all gene families or regulatory elements that have evolved recently in *Caenorhabditis*. Most importantly, a closely related outgroup is required to ensure the highest success in distinguishing orthologs from paralogs. *C. japonica* (Kiontke et al., 2003) is the closest known outgroup (see Figure 1). Evolutionary correlations between variations in behavior, ecology, the composition of gene families (e.g. hormone receptors and GPCRs), and in genes themselves should ultimately provide additional understanding with regard to genome structure and function.

For example, one 34,261-nucleotide region from CB5161 encodes a cluster of seven genes whose predicted protein products have moderate similarity (24-37% identity; E value $\leq 10^{-9}$) to roughly 50 seven-transmembrane proteins in *C. elegans* (E. Schwarz, pers. comm.). These similarities are statistically significant but do not fall into uniquely strong pairs of sequences, as would be normally seen for orthologous genes of CB5161 and *C. elegans.* The *C. elegans* proteins do not currently have assigned locus names, or any known motif; they instead share a Pfam domain (DUF_216; PF02695) of unknown function. At the same time, the CB5161 proteins have 22-37% identity to one another. Having the genome sequence for an outgroup species in the context of a known species phylogeny would allow us to determine which

proteins are related through speciation events (i.e., are orthologous) and which are related through gene duplications (paralogous). Comparisons of such families of receptor and signaling proteins between multiple sibling *Caenorhabditis* species should additionally identify candidates for genes involved in microevolution, ecological adaptation to species-specific niches, and ethological tuning of species-specific behaviors. The genome sequence of an outgroup representative with well-defined ecology (*C. japonica*; see below) will greatly aid the investigation of these and related issues.

Understanding the organization of splicing will also be enhanced by the genome sequence of a closely related outgroup species (e.g., Lee et al. 1992). For example, we know that there are many introns in *C. briggsae* that are not in *C. elegans* and vice versa (Stein et al., 2003). Because an outgroup allows us to infer the direction of evolutionary changes, we can determine which introns were gained and which were lost (which might also help us figure out where new introns come from if they are ever gained). Also, since there exist differences in operon structure between *C. elegans* and *C. briggsae* (Blumenthal and Gleason, 2003; Stein et al., 2003), having an outgroup will shed light on evolution of trans-spliced operons.

Besides functional conservation during evolution, we also predict that there has been functional divergence. Merely comparing pairs or even quadruples of sequences may not provide information about functionality that is not shared. With an outgroup, we have a higher chance of identifying shared changes (apomorphies) that result in new or changed functions. Extrapolating from the *lin-48* example above, it is possible that *C. japonica* shares regulatory sequences with *C. elegans*. If so, the regulation seen in *C. elegans* is ancestral and that in *C. briggsae, C. remanei,* and CB5161 would be derived. Such an evolutionary reconstruction of changes in regulatory sequence correlated with changes in regulation provides considerable predictive power regarding regulatory sequence function.

Experimental use of the genome sequences

The examples above make clear that genome sequences of *C. remanei*, CB5161 and *C. japonica* will be of value for computational studies. The ability to use the sequence experimentally in the context of the organism adds further value. *C. briggsae* sequence is enabling studies in *C. briggsae*, and we expect that the other sequences will similarly lead to experimentation in those species. All these species share transparency, essentially invariant anatomy and small cell number, the ability to freeze and thaw strains, among other features that make them suitable for laboratory culture and experimentation.

There is already a small *C. briggsae* community (a >12 lab subset of the *C. elegans* community; http://wormlab.caltech.edu/briggsae/) that has started genetic and molecular genetic experiments with *C. briggsae*. There are over 500 mutant lines including those with defects in vulval development, dauer pathway and sex determination pathways. Several of these mutants (e.g., *unc-4*, *lin-11*, *sma-6, daf-4*) have been cloned by a candidate gene approach involving transformation rescue and sequencing, but others display novel phenotypes (B. Gupta, T. Inoue, S. Phan and P.S., unpublished observations). A SNP map is being constructed based on the genome sequence (R. Miller and R. Waterston, unpublished), and thus will facilitate the positional cloning of novel loci as well as ordering of the contigs on each chromosome. A *C. remanei* SNP map is being generated (S. Baird, personal communication).

In *C. remanei*, RNAi inactivation of several genes has allowed the probing of functional conservation and divergence (Haag and Kimble, 2000; Rudel and Kimble, 2001, 2002; Ashcroft et al 1999; Stothard et al 2002). *C. briggsae* is resistant to RNAi by feeding (M. Montgomery, personal communication) but transgenics expressing *C. elegans sid-2* are sensitive to bacterial expressed dsRNA (C. Hunter, personal communication). RNAi has not yet been successful with CB5161 (M. Montgomery, personal communication; Stothard et al., 2002). Morpholinos work in *Oscheius tipulae* (Louvet-Vallee et al. Genetics 2003; Dichtel- Danjoy and Félix, 2003)

and the more distant *Pristionchus pacificus* (R. Sommer, personal communication) and thus this technique is a reasonable fallback to RNAi.

DNA-mediated transformation of *C. briggsae* and *C. elegans* is routine. Transformation works with CB5161 and PS1010 (M.-A. Félix, pers.comm.) but has not been tested with *C. japonica* since it has only recently been described and available. *C. remanei* can be transformed but the extrachromosomal arrays typical in *C. elegans* transformation are not common (E. Haag, M.-A. Félix, personal communications).

C. elegans and *C. briggsae* both have hermaphrodite-male sexual systems. Most of the remaining *Caneorhabditis* species are male-female. The evolution of sex determination is thus a well-defined problem in this genus, and is being intensively studied (reviewed by Goodwin and Ellis, 2002; Stothard and Pilgrim, 2003). For example, T. Schedl (pers. comm.) has found that *fog-2*, necessary for spermatogenesis in the *C. elegans* hermaphrodite, is a member of a rapidly evolving gene family in *Caenorhabditis*, and that the mechanism of hermphroditism is probably different between *C. elegans* and *C. briggsae.* Having access to male-female species such as *C. remanei*, which is a sibling to *C. briggsae*, CB5161, an outgroup to *C. remanei* and *C. briggsae*, and C. *japonica*, an outgroup to *C. elegans* and the others, should enable this exciting research area to reach fruition.

C. *remanei* has been proposed as a good species for analysis of ecological and evolutionary genetics because natural populations of *C. remanei* are readily available (Baird, 1999; Jovelin et al., 2003). *C. japonica* is tightly associated with an insect host (Kiontke et al., 2003), and thus might serve as a good model for some of the ecological adaptations of *Caenorhabditis* species.

Impact on the Research Community

Two major communities will be served by these additional genomes. One is the bioinformatics community that has used *C. elegans* since 1998 when it was the first animal genome sequence available, and *C. briggsae* as data accumulated (e.g., Kent and Zahler, 2000; Webb et al., 2002). The second is the 463-laboratory core *C. elegans* community; this is a highly productive community, and is poised to use the additional sequence data. The *C. briggsae* sequence was used as soon as it entered the trace repository, and it is similarly expected that these data will be used as soon as the reads are generated.

There is now an intense focus on transcriptional regulation in *C. elegans*; this research focus will be greatly facilitated by the availability of additional closely related genomes. This focus is based on analysis of signaling pathways and the need to identify their transcriptional targets; from large scale projects that have identified co-expressed, co-regulated and co-functional genes such that searches for *cis*-regulatory elements have become one logical next step in their analysis; from analysis of specific genes that control development, behavior or physiology that has led to identification of single cell type foci of action; and from a transition of analysis of one pathway to pathway integration, with an initial focus on transcriptional regulation.

There is also an intense, but much smaller community focusing on micro-RNAs (reviewed by Pasquinelli, 2002; Ambros et al., 2003). These sequences will be of great value to this emerging sub-field as researchers try to find targets for the >100 micro RNAs. Other non-coding RNAs will also undoubtedly emerge, since extensive (although scattered) evidence for other types of ncRNA already exists both in *C. elegans* (Morse et al., 2002) and in metazoa generally (Szymanski et al., 2003).

A majority of *C. elegans* biologists rely on information about protein structure and function, in part because they interpret the consequences of mutations affecting protein structure, and in part because *C. elegans* is a facile system for structure-function analysis. For rapidly evolving proteins, having four or five species to identify conserved amino acids will provide powerful

insights into structure function relationships. The additional genomes will also help predict alternatively spliced exons. Since alternatively spliced mRNAs can give rise to proteins with different and even antagonistic functions, knowledge of potential variants is important for functional analyses.

In addition to these communities, there is growing use of *Caenorhabditis* for studies of evolutionary biology (e.g., Delattre and Felix, 2001); having additional genomes of the *elegans* group of this genus will stimulate this research area.

Genome Readiness

The size of the *C. remanei* genome is not known with any certainty. A preliminary analysis of *C. remanei* random reads showed that the fraction of reads matching the *C. elegans* genome sequence or falling in the rDNA repeat was roughly that expected from a genome size similar to *C. elegans* and *C. briggsae* and with a similar rDNA repeat number of about 100 copies (R. Waterston, unpublished observations). Random sequencing of genome fragments to identify SNPs indicates that approximately 20-25% of the genome is coding (S. Baird, personal communication). We have no reason to believe at this time that the size of the *C. remanei* genome is much different, but will make a more refined estimate during data collection and adjust the read number accordingly. *C. briggsae* is now estimated at 104 Mb, slightly larger than *C. elegans* (100.2 Mb) (Stein et al., 2003).

Library screening of CB5161 with probes for 12 genes and prelimnary analysis of ~300 kb of genomic sequence indicates a genome size similar to *C. elegans*. In addition, as is the case with *C. remanei*, the size of the coding and total regions of 48 genes has been compared to those of *C. elegans* and found to be similar (E. Schwarz, pers. comm.).

There is no information about the genome size of *C. japonica*. However, library screening and preliminary genomic sequencing of PS1010, outgroup to *C. japonica* and *C. elegans*, indicates a similar size to CB5161, and by extension both to *C. elegans* and *C. briggsae*. We have no reason to suspect that *C. japonica* is significantly larger than *C. elegans* and *C. briggsae*, but a genome size estimate will be developed prior to whole genome shotgun sequencing. Sequencing from 5 genes (18S and 28S rRNA, *ama-1*, *par-6*, and *pkc-3*) from all nine *Caenorhabditis* species in culture (see Fig. 1) shows that intron numbers and total gene sizes in *C. japonica* is most similar to those of *C. elegans*, *C. briggsae*, and *C. remanei* (K. Kiontke and D. Fitch, in prep.).

Choice of strains for sequencing

An open workshop to discuss community priorities for *Caenorhabditis* genome sequencing, organized by Bob Waterston and Paul Sternberg, was held at the 2003 International *C. elegans* Meeting. The 100 attendees appeared uniformly enthusiastic about having access to additional sequences. For close relatives of *C. elegans*, the focus of this white paper, most attention was focused on *C. remanei* because of the biologists interest in sex determination. Sequence analysts were excited about the combination of CB5161 and *C. remanei*. All were enthusiastic about an outgroup, and the consensus was in favor of *C. japonica* over more distant species such as PS1010. In addition to the desire for close relative sequences, which was the focus of the workshop, there was some discussion of nematodes used for studies of the evolution of development and to understand parasitic forms in addition to the ongoing *Brugia* sequence. These species included *Strongyloides* and plant parasitic nematodes. *Pristionchus pacificus* was raised as a promising choice for evolution of development studies.

C. remanei. The *Caenorhabditis* Genetics Center permanently stocks several different isolates of *C. remanei* (Sudhaus, 1974), including the EM464, SB146, and VT733 isolates that have been most often used in the *C. elegans* community. These three *C. remanei* isolates are closely

related, and essentially at the same evolutionary distance from *C. elegans*. A choice between SB146 and EM464 is arbitrary as both have been used for gene cloning and functional studies (e.g., Haag and Kimble [2000] used SB146 while Chen et al. [2001] used EM464). Scott Baird has inbred the EM464-derivative PB4641 20 generations, and we will thus use this strain.

CB5161. The strain CB5161 defines the new species called here *Caenorhabditis n. sp.* CB5161. To ensure homozygosity, we are inbreeding CB5161 20 generations, which is expected to be completed this fall. The inbred strain will be permanently archived as a frozen culture at the Caenorhabditis Genetics Center. There is only a single known isolate of this species (although another new isolate may be conspecific); a formal species description is being discussed (W. Sudhaus and D. Fitch, personal communication).

C. japonica. Karin Kiontke and D. Fitch (pers. comm.) have two healthy, 17-generation inbred strains of *C. japonica* that have been sent to the *Caenorhabditis* Genetics Center. One will be chosen for sequencing.

Sequencing Strategy

The strategy that we propose to sequence the genomes of *C. remanei,* CB5161, and *C. japonica* is similar to the strategy we used to sequence the genome of *C. briggsae*. The *C. briggsae* genome was sequenced using a combination of large insert clone mapping and whole genome shotgun (WGS) assembly. The map currently contains 188 contigs, made up of a mixture of fosmid (16,414) and BAC (17,855) clones. The WGS used some 2.3 million sequencing reads, representing about 11-fold sequence coverage. Included in this set were 20,000 BAC ends to link the sequence assembly with the map. The assembly yielded 105.8 Mbp of sequence 5341 contigs with an N_{50} contig size statistic of 41 kbps. Scaffolding of these contigs, using read pair information, results in 107.5 Mbp of scaffold length in 899 scaffolds with an N_{50} scaffold size of 474 kbps. Comparison with finished sequence shows there are undetected overlaps between contigs and some redundancy. Using the position and orientation of the BAC end reads and the FPC map, these scaffolds were positioned onto the FPC map contigs, resulting in 142 ultracontigs spanning 102 Mbp and 436 unplaced scaffolds containing 6 Mbp (mostly highly repetitive). Comparison with finished sequences indicates the assembly covers 98% of the *C. briggsae* genome. An automated process is now being used to close many of the remaining gaps cheaply and efficiently.

For the new species, we would propose to generate 1.6 million sequence read-pairs from plasmid and fosmid clones to produce with the improved success rates and greater readlengths of current technology approximately 9-fold coverage of the genome (1.6 M reads*0.88 success*650 bases per read = 915 M bases). The ongoing improvements in WGS assembly and the almost exclusive use of these new sequences to inform the interpretation of the *C. elegans* sequence removes the need for a clone-based physical map. The relatively high coverage will yield almost as high a continuity as was achieved for *C. briggsae*. We will then use the automated directed approach mentioned above to close about two-thirds of the remaining gaps. Given the high degree of divergence of *C. remanei,* CB5161 and *C. japonica* from either of the other two genomes, relatively high continuity will be needed to align adequately many regions of the genomes. One improvement that we would make to the approach used for *C. briggsae* is to increase the proportion of fosmid-based sequence reads from 1% to 2% of the total. This increase should provide a better assembly framework with an increased overall long-range continuity of supercontigs, which again will be important for maximally aligning these genomes.

Display of information

WormBase is already storing and displaying *C. briggsae* sequence and annotations, as well as any available nematode ESTs. Lincoln Stein, John Spieth and Richard Durbin (three of the four WormBase PIs) were involved in the annotation and display of the *C. briggsae* sequence and its

relationship to that of *C. elegans*. Thus, the sequence and annotations from *C. remanei*, CB5161 and *C. japonica* will readily be incorporated into WormBase.

References

Aamodt E, Shen L, Marra M, Schein J, Rose B, and McDermott JB. (2000). Conservation of sequence and function of the *pag-3* genes from *C. elegans* and *C. briggsae*. Gene 243:67-74.

Ahringer J. (1997). Turn to the Worm! Curr. Opin. Genet. Dev 7: 410-415.

Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D. (2003). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr Biol 13(10):807-18.

Ashcroft NR, Srayko MA, Kosinski ME, Mains PE, Golden A. (1999). RNA-mediated interference of a *cdc25* homolog in *Caenorhabditis elegans* results in defects in the embryonic cortical membrane, meiosis, and mitosis. Develop. Biol. 206, 15-32.

Baird, S. E. (1999). Natural and experimental associations of *Caenorhabditis remanei* with *Trachelipus rathkii* and other terrestrial isopods. Nematology 1:471-475

Blanchette M, Sinha S. (2001). Separating real motifs from their artifacts. Bioinformatics 17: S30-S38.

Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, Vida JT, and Thomas WK. (1998). A molecular evolutionary framework for the phylum Nematoda. Nature 392: 71-76.

Blumenthal T, Gleason KS (2003). *Caenorhabditis elegans* operons: forms and function. Nat Rev Genet 4:110-118.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, and Rubin EM. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299(5611):1391-1394.

Brown CT, Rust AG, Clarke PJ, Pan Z, Schilstra MJ, De Buysscher T, Griffin G, Wold BJ, Cameron RA, Davidson EH, Bolouri H. (2002). New computational approaches for analysis of cis-regulatory networks. Dev Biol 246:86-102.

Chen PJ, Cho S, Jin SW, and Ellis RE. (2001). Specification of germ cell fates by FOG-3 has been conserved during nematode evolution. Genetics 158:1513-25.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 301(5629):71-6.

Cui M and Han M. (2003). Cis regulatory requirements for vulval cell-specific expression of the *Caenorhabditis elegans* fibroblast growth factor gene *egl-17*. Dev Biol ;257(1):104-16.

Culetto E and Sattelle DB. (2000). A role for *Caenorhabditis elegans* in understanding the function and interactions of human disease genes. Human Molecular Genetics 9:869-877.

Delattre M and Felix MA. (2001). Polymorphism and evolution of vulval precursor cell lineages within two nematode genera, *Caenorhabditis* and *Oscheius*. Curr Biol. 11:631-43.

Dichtel-Danjoy, M.-L. and Félix, M.-A. (2003). The two steps of vulval induction in *Oscheius tipulae* CEW1 recruit common regulators including a MEK kinase. Devel. Biol., in press.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8:163-167.

Goodwin EB and Ellis RE. (2002). Turning clustering loops: sex determination in *Caenorhabditis* elegans. Curr Biol. 12(3):R111-20.

Haag ES and Kimble J. (2000). Regulatory elements required for development of *Caenorhabditis elegans* hermaphrodites are conserved in the *tra-2* homologue of *C. remanei*, a male/female sister species. Genetics 155(1):105-16. Erratum in: Genetics 2000 Jul;155(3):1485.

Hwang, B. J. and Sternberg, P. W. (2003). A cell specific enhancer that specifies *lin-3* expression in the *C. elegans* anchor cell for vulval development. Development, in press.

Jan E, Yoon JW. Walterhouse D. Iannaccone P, and Goodwin EB (1997). Conservation of the

C. elegans tra-2 3' UTR translational control. EMBO J. 16, 6301-6313.
Jovelin R, Ajie BC, Phillips PC. (2003). Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus *Caenorhabditis*. Mol Ecol. 12(5):1325-37.
Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrman M, Welchman DP, Zipperfen P, Ahringer J., (2003). Systematic functional analysis of the *Caenorhabditis elegans*genome using RNAi. Nature 421: 231-237.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-54.
Kennedy BP, Aamodt EJ, Allen FL, Chung MA, Heschl MF, McGhee JD (1993). The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J Mol Biol 229:890-908.
Kent WJ, Zahler AM. (2000). Conservation, regulation. synteny and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. Genome Research 10: 1115-1125.
Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. (2001). A gene expression map for *Caenorhabditis elegans*. Science 293: 2087-2092.
Kiontke, K., Hironaka, M. & Sudhaus, W. (2003). Description of *Caenorhabditis japonica* n. sp. (Nematoda: Rhabditida) associated with the burrower bug *Parastrachia japonensis* (Heteroptera: Cydnidae) in Japan. Nematology 4, 933-941.
Kiroauc, M, Sternberg, P. W. (2003). cis-regulatory control of three cell-fate genes in vulval organogenesis of *C. elegans* and *C. briggsae*. Developmental Biology, in press.
Korf I, Flicek P, Duan D, Brent MR. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17, S140-S148.
Lau NC, Lim LP, Weinstein EG, Bartel DP. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294(5543):858-62.
Lee RC, Ambros V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294(5543):862-4.
Lee RC, Feinbaum RL, Ambros V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75:843-854.
Lee YH, Huang XY, Hirsh D, Fox GE, Hecht RM (1992). Conservation of gene organization and trans-splicing in the glyceraldehyde-3-phosphate dehydrogenase-encoding genes of *Caenorhabditis briggsae*. Gene 121:227-235.
Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. (2003). The microRNAs of *Caenorhabditis elegans*. Genes Dev. 17:991-1008.
Liu J, Tzou P, Hill RJ, Sternberg PW. (1999). Structural requirements for the tissue-specific and tissue-general functions of the *C. elegans* epidermal growth factor LIN-3. Genetics 153: 1257-1269.
Louvet-Vallee S, Kolotuev I, Podbilewicz B, Felix MA. (2003) Control of vulval competence and centering in the nematode *Oscheius* sp. CEW1. Genetics 163(1):133-46.
Marshall SD, McGhee JD. (2001). Coordination of *ges-1* expression between the *Caenorhabditis* pharynx and intestine. Dev Biol. 239(2):350-63.
McCutcheon, J.P. and Eddy, S.R. (2003). Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. Nucleic Acids Res. 31, 4119-4128.
Morse, D.P., Aruscavage, P.J., and Bass B.L. (2002). RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA. Proc. Natl. Acad. Sci. USA 99, 7906-7911.
Pasquinelli AE. (2002). MicroRNAs: deviants no longer. Trends Genet. 18:171-173.
Prasad SS, Baillie DL (1989). Evolutionarily conserved coding sequences in the *dpy-20-unc-22* region of *Caenorhabditis elegans*. Genomics 5:185-198.
Reboul J, Vaglio P, Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, et al. (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. Nat Genet 34:35-41.
Riddle, D. L., T. Blumenthal, et al., Eds. (1997). C. elegans II. Cold Spring Harbor, NY, Cold

Spring Harbor Laboratory Press.
Rivas E and Eddy, SR. (2001). Noncoding RNA gene detection using comparative sequence analysis. *Bioinformatics*, 2:8.
Rivas E, Klein RJ, Jones TA, Eddy SR. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Current Biology, 11:1369-1373.
Rudel D and Kimble J. (2002). Evolution of discrete Notch-like receptors from a distant gene duplication in *Caenorhabditis*. Evol Dev. 4(5):319-33.
Rudel D and Kimble JE. (2001). Conservation of *glp-1* regulation and function in nematodes. Genetics 157: 639-654.
Ruvinsky I, Ruvkun G (2003). Functional tests of enhancer conservation between distantly related species. Development 130:5133-5142.
Sengupta P. Chou JH ,Bargmann CI (1996). *odr-10* encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl. Cell 84 :899-909
Shakes DC Epstein HF (1995). *Caenorhabditis elegans*: Modern biological analysis of an organism. Methods in Cell Biology 48. Academic Press.
Sinha, S. and Tompa M. (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res. 30, 5549-5560.
Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent M, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fulton L, Fulton R, Griffiths-Jones S, Harris T, Hillier L, Kamath R, Kuwabara P, Marra M, Mardis E, Miner T, Minx P, Mullikin JC, Plumb R, Rogers J, Schein J, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson R, Durbin R,Waterston R. (2003). The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PloS-Biology (November, in press).
Stothard P, Hansen D, Pilgrim D. (2002). Evolution of the PP2C family in *Caenorhabditis* : rapid divergence of the sex-determining protein FEM-2. J. Molec. Evol. 54, 267-282.
Stothard P, Pilgrim D (2003). Sex-determination gene and pathway evolution in nematodes. BioEssay 25:221-231.
Sudhaus W, Fitch D. (2002). Comparative studies on the phylogeny and systematics of the Rhabditidae (Nematoda). Journal of Nematology 33: 1-70.
Sudhaus W. (1974). Zur Systematik, Verbreitung, Okologie und Biologie neuer und wenig bekannter Rhabditiden (Nematoda) 2. Teil. Zool. Jb. Syst. Bd. 101: 417-465.
Szymanski, M., Erdmann, V.A., and Barciszewski, J. (2003). Noncoding regulatory RNAs database. Nucleic Acids Res. 31, 429-431.
Thacker C, Marra MA, Jones A, Baillie DL, Rose AM. (1999). Functional genomics in *Caenorhabditis elegans*: An approach involving comparisons of sequences from related nematodes. Genome Res. 9:348-359.
The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012-2018.
Wang X, Chamberlin HM (2002). Multiple regulatory changes contribute to the evolution of the *Caenorhabditis lin-48 ovo* gene. Genes Dev 16:2345-2349.
Webb CT, Shabalina SA, Ogurtsov AY, Kondrashov AS. (2002). Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. Nucleic Acids Res. 30:1233-9.
Wood, W. B. (1988). The Nematode *Caenorhabditis elegans.* Cold Spring Harbor Laboratory Press.
Xue D, Finney M, Ruvkun G, Chalfie M. (1992). Regulation of the *mec-3* gene by the *C. elegans* homeoproteins UNC-86 and MEC-3. EMBO J. 11, 4969-4979.
Zucker-Aprison E, Blumenthal T. (1989). Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. J Mol Evol. 28:487-96.