2005

Genomics in C. elegans: So many genes, such a little worm

LaDeana W. Hillier
Washington University School of Medicine in St. Louis

Alan Coulson
MRC Laboratory of Molecular Biology

John I. Murray
University of Washington - Seattle Campus

Zhirong Bao
University of Washington - Seattle Campus

John E. Sulston
The Wellcome Trust Sanger Institute

See next page for additional authors

Genomics in *C. elegans*: So many genes, such a little worm

LaDeana W. Hillier, Alan Coulson, John I. Murray, et al.

Genome Res. 2005 15: 1651-1660
Access the most recent version at doi:10.1101/gr.3729105

Perspective

Genomics in *C. elegans*: So many genes, such a little worm

LaDeana W. Hillier,[1] Alan Coulson,[2,3] John I. Murray,[4] Zhirong Bao,[4] John E. Sulston,[3] and Robert H. Waterston[4,5]

[1]*Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108, USA;* [2]*MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, United Kingdom;* [3]*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom;* [4]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA*

The *Caenorhabditis elegans* genome sequence is now complete, fully contiguous telomere to telomere and totaling 100,291,840 bp. The sequence has catalyzed the collection of systematic data sets and analyses, including a curated set of 19,735 protein-coding genes—with >90% directly supported by experimental evidence—and >1300 noncoding RNA genes. High-throughput efforts are under way to complete the gene sets, along with studies to characterize gene expression, function, and regulation on a genome-wide scale. The success of the worm project has had a profound effect on genome sequencing and on genomics more broadly. We now have a solid platform on which to build toward the lofty goal of a true molecular understanding of worm biology with all its implications including those for human health.

[Supplemental material is available online at www.genome.org.]

In 1965 Sydney Brenner selected *Caenorhabditis elegans* for his studies of development and the nervous system because of its simple anatomy, its stereotyped behavior, and the ease of genetic manipulation. Even at inception, the goal of studying the worm was an understanding of how genes dictated form and behavior. This holistic view of the organism (now dubbed "systems biology") stimulated the collection of comprehensive data sets. The anatomy was described through serial electron microscopic reconstruction with the nervous system defined at the level of the synapse (White et al. 1986). The complete cell lineage of the 959 adult somatic cells was determined (Sulston and Horvitz 1977; Kimble and Hirsh 1979; Sulston et al. 1983) and found to be remarkably consistent animal to animal. Investigators commonly sought to collect all genes affecting a certain trait through mutations (however illusory that completeness might be in retrospect).

The construction of a clone-based physical map (Coulson et al. 1986, 1995; Sulston et al. 1988), one of the earliest genome projects, was undertaken in the early 1980s in the same spirit. The map of overlapping cosmids and later Yeast Artificial Chromosomes (YACs) (Coulson et al. 1988, 1991), along with efficient means of transformation, provided the community with the wherewithal to recover the DNA for any well-mapped mutant readily and rapidly. But perhaps more importantly, the existence of a nearly complete physical map in 1989 helped convince James D. Watson, head of the National Center for Human Genome Research at the time, that the worm should be included in the select set of model organisms to be targeted by the Human Genome Project (HGP), the so-called Security Council of the HGP (Sulston and Ferry 2002). We, in turn, were drawn to the project by the vision of a complete genome sequence, whose catalytic effect would drive research on the worm forward.

This review begins with an update on the genome sequence since our last report in 1998 (The *C. elegans* Genome Sequencing Consortium 1998). We describe the current state of the genome annotation of the sequence and then consider the collection of systematic data sets and analyses that the genome sequence has enabled and stimulated. All of these data and more are collected in WormBase (Chen et al. 2005a), which is briefly summarized (see Table 1 for Web sites). In conclusion, we discuss the challenges ahead as we strive for a molecular explanation of how the genome sequence produces a worm.

Genome sequences

The C. elegans *genome sequence is complete*

When the sequence of the 100-Mb genome of *C. elegans* was published in 1998 (The *C. elegans* Genome Sequencing Consortium 1998), very little important information was believed to be missing. Nonetheless, several recalcitrant gaps remained, and we had aimed from the start for a complete description of the content and structure of this benchmark genome. With persistence, we have now accumulated, by a variety of methods, the mapping and sequence information that completes the genome. The work behind this achievement is summarized in Text Box 1 and described in more detail in the Supplemental material.

As a result, the *C. elegans* sequence is fully contiguous telomere to telomere and with the mitochondrial genome totals 100,291,840 bp. A few problems may remain, such as undetected deletions within the clones or minor misassemblies. Some long multicopy tandem repeats, where not completely sequenced, have been characterized with respect to sequence content and tagged as such in sequence entries. But because of the hierarchical (clone) based shotgun methods used, all larger genomic duplications should be resolved (including one tandem repeat of

[5]Corresponding author.
E-mail waterston@gs.washington.edu; fax (206) 685-7301.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3729105.

Table 1. *C. elegans* online repositories

Web address	Description
www.wormbase.org	Biology and genome database
elegans.swmed.edu	*C. elegans* WWW server
www.wormatlas.org	Behavioral and structural anatomy
www.wormbook.org	Online review of *C. elegans* biology
www.wormclassroom.org	Education and online learning community
www.rnai.org	Phenotypic data from RNAi studies

Additional Web sites are available in the Supplemental material.

108 kb with only 10 sequence differences between the two copies). The per base error rate has been estimated at $<10^{-5}$. Reports from the community of problems with the sequence are now exceedingly infrequent, suggesting that remaining problems are rare, indeed. The genome seems in good shape!

Other Caenorhabditis *genomes*

The comparison of related genomes provides a powerful tool for genome interpretation. In support of this objective, a draft sequence of the *Caenorhabditis briggsae* genome was produced (Stein et al. 2003). This whole-genome shotgun project produced a sequence with just 899 supercontigs (ordered and oriented contiguous sequence segments) spanning 106 Mb of DNA sequence with ~3 Mb of undetected overlaps and another ~2Mb of inferred gaps. When combined with the physical map, 102 Mb was placed in 142 ultracontigs ("supercontigs" ordered and oriented by their position within the physical map). More recently, the construction of a genetic map using single nucleotide polymorphic (SNP) markers has positioned 100 Mb along the six chromosomes and refined the sequence map (R.H. Waterston, S. Baird, L. Hillier, and R. Miller, unpubl.).

The *C. briggsae* sequence has proven useful in gene prediction (Wei et al. 2005), definition of regulatory elements (Luersen et al. 2004; Teng et al. 2004), and recognition of microRNAs (miRNAs; see below). But with only two species to compare, the signals of selection are often difficult to tease out from the noise of neutral change. To add power to the analysis, additional nematode genomes are currently under way (http://www.genome.gov/11007952), including the three closest of the known *Caenorhabditis* genomes, *Caenorhabditis remanei*, *Caenorhabditis japonica*, and *Caenorhabditis n. sp. PB2801*, and the more distantly related species *Pristionchus pacificus* and *Brugia malayi* (http://www.genome.gov/10002154). All are based on whole-genome shotgun assemblies. The three additional *Caenorhabditis* sequences should refine the definition of conserved features and may reveal sequences that have changed more rapidly in one lineage but not in others. The sequence of multiple species may be particularly critical in defining regulatory elements and noncoding RNA genes. The multiple *Caenorhabditis* species combined with the more distantly related nematodes should also provide insights into structure–function relationships at the protein level. As sequencing costs continue to drop, complete sequencing of other *C. elegans* isolates will undoubtedly be undertaken and add to our knowledge of the functional elements and their evolution.

Gene annotation

Protein–coding genes

The identification of the full set of *C. elegans* protein-coding genes is approaching completion. WormBase (release WS140) (Chen et al. 2005a) currently lists 19,735 genes with 2685 alternative splice forms, bringing the predicted protein count to 22,420 (producing 22,269 unique peptide sequences). More than 90% of the alternatively spliced genes have only one or two alternative spliced forms (Spieth and Lawson 2005). *Trans*-splicing is common in the worm, with more than half of *C. elegans* premRNAs receiving an SL1 leader sequence and 20% an SL2 (Blumenthal 2005). More than 90% of the genes are directly supported by experimental evidence.

Nematodes are unusual among animals in having operons, polycistronic gene clusters containing two or more genes (Blumenthal and Gleason 2003; Blumenthal 2005). Currently, there are >1000 operons identified, each containing between two and

Text Box 1. Completing the *C. elegans* genome sequence

At publication in 1998, there were tens of unfinished YACs and three unfinished cosmids and fosmids. These clones were all completed over the next year or two using the array of methods available for clone finishing (International Human Genome Sequencing Consortium 2004). In addition, we corrected ~20 misassembled, ambiguous, or deleted regions along with ~200 single base corrections (mostly in early projects) stemming from detailed analysis of Expressed Sequence Tags (ESTs) (McCombie et al. 1992; Waterston et al. 1992; Kohara 1996; The *C. elegans* Genome Sequencing Consortium 1998) and other data including community feedback.

More significantly, there remained two internal map gaps on Chromosomes III and IV, respectively, where no spanning clones were available, and three telomeric (Chromosome II right, where left and right are with reference to the genetic map) or subtelomeric (Chromosome I left and Chromosome X left) gaps. The telomere clone cTel33B (one from a set of eleven isolated by Wicky et al. 1996) eventually overlapped Y74C9 as its sequence was completed, capping the left end of Chromosome I. Plasmid cTel7X was linked to Y35H6 on the left end of Chromosome X through three PCR fragments, capping that chromosome end.

The internal gaps persisted despite the high redundancy of the initially mapped clones (some 30-fold from YAC, cosmid, and fosmid clones) and after screening a new BAC library (Exelixis, http://www.exelixis.com, pers. comm.). Given the rarity of these regions in large insert clone libraries, we turned to a strategy of directly subcloning and shotgun-sequencing a restriction fragment from whole genomic DNA for these internal gaps and the uncloned telomere from Chromosome II right.

The regions containing the internal gaps and the remaining telomere were mapped by macrorestriction Southern-blot analysis, using probes derived from the known flanking sequence. To obtain useful purity of the fragments, we adopted a successive digest scheme, using pulsed field gel electrophoresis (PFGE) to isolate the product of the first digest, digesting this in situ with a second enzyme, and subcloning the isolated DNA from a second PFGE purification. Inevitably these libraries were contaminated with copurifying DNA (50%–95% contaminated), but the dominant contig was easily identified in each case and the rest accounted for with known sequence.

The spanning sequence for the internal gaps was in each case a small fraction of the size predicted by Southern blots (6 kb vs. the predicted 250 kb and 20 kb vs. 70 kb for Chromosomes III and IV, respectively). Perhaps the fragment mobility in PFGE can be anomalous at high concentrations (Doggett et al. 1992) (we used 50–100 μ/mL) or result from unusual sequence features, which might also account for the poor representation of the regions in libraries. The telomere segment was in better agreement (82 kb vs. 90 kb predicted), with the difference accounted for at least in part by exclusion of the telomere repeat from the assembled sequence.

eight genes, and accounting for ~15% of all *C. elegans* genes. Those genes that encode the basic machinery of gene expression are more frequently included in operons, while tissue-specific genes tend not to be part of operons (Blumenthal and Gleason 2003).

The protein-coding gene set was based initially on predictions by GeneFinder (P. Green, unpubl.), a gene prediction program developed in conjunction with the *C. elegans* genome project (The *C. elegans* Genome Sequencing Consortium 1998). The accuracy of individual exon prediction was high, but the prediction of complete genes was less reliable because of the combinatorics of multiexon genes and the challenges in detecting the start and stop of genes, especially in an organism with operons. Nonetheless, the GeneFinder predictions have been an excellent point of departure and have served the worm community well.

The computer predictions have been validated and modified by experimental data. Expressed sequence tags (ESTs) aligned with the genome now number more than a quarter of a million (McCombie et al. 1992; Waterston et al. 1992; Kohara 1996). Most ESTs come from the Kohara lab, which used methods to reduce the prevalence of abundant messages. In most cases, ESTs were derived from both 5′- and 3′-ends of cDNA clones, with the 3′-end establishing the 3′-UTR and the polyadenylation site and the 5′-end sampling the coding region or establishing the 5′-UTR for full-length clones. In turn, these clones provided representatives for full-length cDNA sequencing, with >2800 full-length sequences currently in the database. SAGE (Serial Analysis of Gene Expression) (Velculescu et al. 1995) of more than 30 libraries (http://elegans.bcgsc.ca/home/ge_consortium.html) from worms of a variety of stages, growth conditions, tissues, and cell types has yielded >2.5 million high-quality tags (McKay et al. 2003). These tags provide additional support for 16,212 genes, of which 2682 only have SAGE support. In addition, SAGE tags reveal ~500 open reading frames (ORFs) with *C. briggsae* homology that are not in the present gene predictions (G. Vatcher and D. Moerman, pers. comm.). More recently, a method was developed to obtain 5′-end SAGE-like tags for messages with SL1 or SL2 transpliced leaders (Hwang et al. 2004). An initial set of 13,525 tags identified the 5′-end of 2012 genes, confirming the 5′-end of 1512 known or predicted genes and modifying the end of another 401 genes. The 5′-ends of 99 previously unknown genes were also found. A larger sampling of 5′-end tags, now under way, identifies some 6500 5′-ends with 330 not associated with known or predicted genes (B.J. Hwang, H. Muller, S. McKay, P. Huang, S. Gharib, S. Jones, M. Marra, D. Moerman, D. Baillie and P.W. Sternberg, pers. comm.).

As these random-sampling-based methods become less efficient at gene confirmation/discovery, directed methods that begin with the predicted gene models became more useful. As part of an effort to obtain full-length cDNA clones for all *C. elegans* genes (the ORFeome Project) (Lamesch et al. 2004), >12,500 ORFs have been cloned in Gateway vectors, using RT-PCR starting from the gene models. Beyond confirming the transcription of these models, the data also modify the predicted gene models. Together with the EST libraries, OSTs (ORFeome sequence tags) (Lamesch et al. 2004) define 46,830 exon/intron boundaries. Green and colleagues have also been using RT-PCR to test systematically all unconfirmed intron–exon boundaries (see below) (P. Green, pers. comm).

Many of the remaining unsupported gene models and any as-yet-undetected genes in the genome are likely to be poorly expressed, may have weaker statistical signals, and may be less well conserved across species, making their identification by either computational or experimental means more difficult. Improvements in gene prediction programs may help tease out these signals. Twinscan (Korf et al. 2001), an HMM-based program derived from GenScan (Burge and Karlin 1997) that can use comparative sequence in predictions, has used a more realistic model of intron length, added a minor splice variant to splice tables and the *C. briggsae* sequence to produce an improved gene set over current WormBase predictions (Wei et al. 2005). While most Twinscan predictions overlap at least in part with existing predictions, >2000 are unique to Twinscan. RT-PCR experiments suggest that more than half of these may be transcribed (Wei et al. 2005). In a broad assault on the remaining unconfirmed exons and genes, P. Green (unpubl.) has used a substantially improved GeneFinder with relaxed constraints in order to capture most real genes at the cost of false positives. All the unconfirmed exon–intron boundaries are being tested by RT-PCR across the genome. In addition, SL1 and SL2 primers are being used in combination with internal primers to identify the 5′-ends of transpliced messages. Preliminary analysis of the data indicates that the gene set may rise to >21,000 confirmed protein-coding genes. The drive to complete the gene set will undoubtedly begin to challenge our notions of a gene.

Noncoding RNA genes

Many transcripts function at the RNA level, including rRNAs, tRNAs, snRNAs, and snoRNAs. *C. elegans* contains all the major types of eukaryotic RNA genes: >1300 (Stricklin et al. 2005) of these genes have been identified, including 630 tRNAs, 78 snRNAs, and 17 snoRNAs. Of the rRNA genes, the 18S, 28S, and 5.8S are transcribed separately by RNA polymerase I in the ~55 copies of the 7.2-kb rDNA repeat on I (Sulston and Brenner 1974; The *C. elegans* Sequencing Consortium 1998). The 5S gene along with the SL1 spliced leader gene lies in a 1-kb tandem repeat with ~110 copies on V (Sulston and Brenner 1974; Nelson and Honda 1985). (With uncertainty about the exact copy number of these large tandem repeats, only representative members of each are included in the sequence.) There are also 20 copies of the SL2 repeat dispersed in the genome. In addition to these well-known genes, the *lin-4* and *let-7* genes provided the first examples of functional miRNAs (Lee et al. 1993; Wightman et al. 1993; Reinhart et al. 2000), which are now recognized to be common features of eukaryotic genomes, including human. Indeed, many worm miRNA genes have clear homologs in mammalian genomes. Methods are now being developed for large-scale in vivo validation of predicted miRNA targets in *C. elegans*; for example, a dozen novel predicted targets of *let-7* have been tested using comparative expression analyses in transgenic worms (N. Rajewsky, S. Lall, and F. Piano, unpubl.). Computational and experimental methods have identified at least 114 miRNA genes (Ambros et al. 2003; Griffiths-Jones 2004; http://microrna.sanger.ac.uk/sequences/), and intriguing new work is providing evidence about the roles of these RNAs in cell and developmental processes.

Other novel RNA genes and gene families may well exist in the worm genome. Current computational methods to identify such genes and families use conservation of secondary structure across species but are subject to high false-positive rates (Rivas and Eddy 2001; Lim et al. 2003), obscuring real genes. With the sequencing of additional related species (Rivas and Eddy 2001; Coventry et al. 2004; Washietl et al. 2005) the false-positive rate may drop sufficiently to allow the emergence of additional RNA

genes. SAGE can provide evidence for some RNA genes (Jones et al. 2001), and the development of tiling microarrays covering essentially all of the genome may well point to additional possible genes for more detailed study.

Global studies enabled by the genome sequences

The genome sequence, by providing a comprehensive view of the information needed to specify the animal and its behavior, has stimulated a variety of systematic studies to define the functional elements of the genome and to capture functional information about those elements more effectively. Occasionally these data sets provide direct insight into biological mechanism; more often they provide resources that enable investigators focused on specific mechanisms to speed their work. Increasingly these more systematic approaches are being integrated into the more specific studies. We provide examples of these data sets and their use below.

Gene expression

In a multicellular organism a major insight into gene function comes from when, where, and under what conditions a gene is expressed. Approaches that yield expression data on many genes in parallel and other systematic efforts have been enabled by the genome sequence. Many of these approaches are shared with other organisms; others exploit the comprehensive knowledge of the worm's simple anatomy and the cell lineage to provide high temporal and anatomic resolution.

Large data sets measuring RNA levels in specific worm populations are available for both microarray analysis and SAGE. Microarrays provide data on many genes at once but depend on the current state of gene models, while SAGE and related approaches give a potentially unbiased sampling but are more expensive. Microarray data have been acquired from hundreds of experiments using populations of worms, including various stages, different sexes and mutants, and various growth conditions. Early on, much of the data were generated using spotted DNA arrays, and these continue to be widely used (http://www.genome.wustl.edu/genome/celegans/microarray/ma_gen_info.cgi). These resources have been augmented by arrays from commercial suppliers. For example, Affymetrix offers a chip representing an estimated 22,500 transcripts from almost 19,000 gene models (http://www.affymetrix.com/products/arrays/specific/celegans.affx), and NimbleGen offers a chip with 390,000 probes covering 21,121 genes with a minimum of 17 probes per gene (http://www.nimblegen.com/products). Clustering the resultant expression data reveals sets of genes that respond similarly within the populations examined, and based on the presence of previously characterized genes within those clusters, inferences can be drawn about the role of the genes in the group. For example, in a pioneering study, Kim and colleagues (Kim et al. 2001) found 44 different clusters and were able to associate 30 of these with possible functions. Early SAGE analysis targeted differences in gene expression patterns between dauer and non-dauer worms, highlighting the substantial transcriptional differences in the specialized dauer stage and identifying noncoding transcripts with sequence related to the telomere repeat (Jones et al. 2001). In another application, SAGE was used to compare long-lived mutants with control populations to reveal genes and pathways potentially involved in life-span extension (Holt and Riddle 2003). These experiments also demonstrated the potential of SAGE to reveal previously unknown genes and alternative splice and polyadenylation variants.

Using amplification, with the caveats this introduces, gene expression has been measured in small populations of purified cell types and in carefully staged embryos. Specific cell types can be labeled using tissue-specific promoters driving GFP (Green Fluorescent Protein), and until about the 400-min stage, embryonic cells can be dissociated with the labeled cells recovered by FACS (Fluorescence Activated Cell Sorting). Cells can be harvested immediately or placed in culture to allow further differentiation (Christensen et al. 2002; Zhang et al. 2004; Blacque et al. 2005; Fox et al. 2005) and analyzed for mRNA content by either microarray analysis or SAGE. In a variant of this, a tagged poly(A) binding protein (PABP) has been expressed in specific cell types, and mRNAs from these cells have been recovered by immunoprecipitation (Roy et al. 2002; Kunitomo et al. 2005). To obtain information about the temporal progression of gene expression in early embryogenesis, Baugh and colleagues (Baugh et al. 2003, 2005) staged small cohorts of embryos by visual selection of embryos at the four-cell stage, which were then allowed to develop. Samples were taken at intervals approximating the successive rounds of cell division of the embryo. Quantitative analysis of the resultant data showed successive sets of gene expression, suggesting a causal relationship. This relationship was confirmed for a few examples, revealing several potential regulatory networks.

In contrast to methods that extract RNAs, gene products (mRNA or protein) can be assayed directly in the animal to determine the site and time of gene expression. Both RNA hybridization and antibody have been used traditionally for this purpose. RNA in situ methods are more readily carried out systematically, and Kohara (http://www.nig.ac.jp/section/kohara/kohara-e.html; nematode.lab.nig.ac.jp/db2/index.php;) currently displays whole-mount in situ images of 11,237 cDNA clones with various stages available for inspection. Certain tissue patterns are readily recognized, but individual cell identity is difficult to determine. Antibody methods have been more difficult to scale up, but new methods for generating high-affinity reagents may change this.

The advent of in vivo GFP labeling methods allows gene expression patterns to be visualized in living worms. Promoter::GFP fusions are being generated on a genome scale in conjunction with the Promoterome project (Dupuy et al. 2004)—the effort has already released promoter fusions (up to 2 kb) from ~6500 *C. elegans* genes, and plans are under way for a more comprehensive set. Two groups are systematically transforming these constructs or related ones using PCR and imaging the resultant worms with fluorescent microscopy. The Hope Lab Web site (http://129.11.204.86:591/default.htm) provides descriptions and images for >300 genes, and the BC Genome Center site (http://www.bcgsc.ca/gc/celegans/) provides information on some 1750 genes, with images available on a subset of these. The former group has focused on transcription factors, while the latter has targeted *C. elegans* genes with human homologs. The fidelity of the transgene patterns to native genes is, of course, a central issue with such approaches. Transgenes introduced by injection typically are incorporated into large extrachromosomal arrays and are subject to somatic loss and germ-line silencing; nevertheless, the observed expression patterns have been generally reliable. In addition, promoters and other regulatory sequences are not defined for most genes in *C. elegans*, so that as an expedient both projects use the upstream region of arbitrary length to drive expression. Since intergenic regions in *C. elegans* are usually small, often these constructs extend to the adjacent gene. Despite these ob-

vious limitations, the available gene expression patterns are highly valuable.

A challenge in using the in vivo expression data is the need for an expert to interpret the patterns. To circumvent this, our laboratory (Z. Bao, J. Murray, T. Boyle, and R. Waterston, unpubl.) has embarked on a project that will automate the assignment of gene expression to individual cells throughout early development. The method uses four-dimensional images of worms with nuclei labeled with GFP-histone fusions to follow cell divisions throughout embryogenesis, thereby automating the determination of the cell lineage. Because the lineage in wild type is highly reproducible and the fate of every daughter cell is known, knowledge of the lineage history of an animal is tantamount to knowledge of its anatomy. Introduction of a second reporter gene driven by a promoter sequence of interest into this background thus holds the promise of providing expression data with single-cell resolution and high temporal fidelity automatically. Introduction of the constructs via bombardment also may yield single-copy integrants and circumvent germ-line silencing in many cases. The current implementation traces the lineage through 250 cells with only minor editing and thus is already useful for early embryonic events.

Gene disruption

A second powerful insight into gene function comes from analysis of the phenotype of animals carrying mutant forms of a gene. Traditional methods, including chemical mutagenesis, irradiation, and transposon insertion, have produced mutant alleles in fewer than 1000 genes. Furthermore, homologous recombination, so powerful in yeast and mammals, is relatively ineffective in *C. elegans*.

Fortunately other methods have emerged that allow systematic disruption of gene function. Since its discovery in the worm (Fire et al. 1998; Piano et al. 2000; Sonnichsen et al. 2005), RNA interference (RNAi), where double-stranded RNA induces sequence-specific degradation of homologous mRNAs, has become the most widely used means of inhibiting gene function. The double-stranded RNA can be introduced by injection, soaking, and even by feeding worms bacteria expressing the dsRNA (Timmons and Fire 1998). Inhibition is rarely complete, and neuronally expressed genes are particularly resistant to RNAi effects. Nonetheless, the ease of use of feeding libraries and other modes of delivery has facilitated systematic genome-wide RNAi screens by several groups (Fraser et al. 2000; Maeda et al. 2001; Kamath and Ahringer 2003; Sonnichsen et al. 2005), and currently >18,000 *Escherichia coli* strains have been constructed and have been widely distributed. Initial screens were for easily scored phenotypes such as viability, slow growth, or altered movement and body shape. These screens and others have produced phenotypes for >3300 genes (the *E. coli* RNAi library covers 86% of all *C. elegans* genes) (http://www.gurdon.cam.ac.uk/~ahringerlab/pages/rnai.html), including 721 genes required for embryogenesis (Vidalain et al. 2004). To examine genetic robustness at a functional level, a double RNAi feeding screen is being carried out to test 2000 putative duplicate gene pairs for redundant function (S. Woods and J. Ahringer, unpubl.). The RNAi library is being increasingly used to screen for more specific phenotypes or in certain mutant backgrounds, including backgrounds that appear to enhance RNAi effects (Wang et al. 2005). The success of these has, in turn, stimulated efforts to automate various aspects of phenotype analysis.

To complement RNAi and to provide permanent lines with transmissible defects, projects are under way to knock out genes, using either chemically induced deletions or transposons. Both methods use PCR to detect length differences in populations of treated animals. At present, the Gene Knockout Consortium (http://www.celeganskoconsortium.omrf.org/) and the National Bioresource Center (http://shigen.lab.nig.ac.jp/c.elegans/index.jsp) have each generated gene deletions. The former has generated deletions in some 1800 genes, with >1300 of these stabilized and archived, while the latter lists ~1600 gene deletions. The NemaGENETAG Consortium (http://elegans.imbb.forth.gr/nemagenetag/home.html) has produced >150 Mos1-tagged strains and plans to do more (P. Kuwabara, unpubl.). The TILLING (Targeting Induced Local Lesions IN Genomes) approach (McCallum et al. 2000), because it is adaptable to any organism that can be chemically mutagenized, has been used in *C. elegans* and proven to be successful at generating point mutations including stop codons (R. Plasterk, unpubl.). TILLING has the potential advantage of producing an allelic series (mutations of varying severity). As sequencing costs fall, direct sequencing of mutagenized lines may become the method of choice (R. Plasterk, pers. comm.).

Gene regulation

The signals that control gene activity in time and space are also embedded in the genome. They act at the DNA level as promoters and other *cis*-regulatory elements; at the RNA level as elements that govern translation and stability; and at the protein level through post-translational modification and turnover. In contrast to protein-coding regions, no algorithms currently exist that can effectively recognize these signals ab initio in genome sequence. Early work in the area focused on individual genes and through traditional methods established the precise sequences driving gene expression (Okkema and Fire 1994; Fukushige et al. 1996; Okkema et al. 1997). But with the genome sequence, a combination of gene expression data, comparative sequence analysis, and improving computer programs, there is progress in the recognition of the DNA elements and to some extent the RNA elements.

At the DNA level the gene expression sets described above have been critical, allowing genes to be grouped or stratified by time and tissue. Candidate elements have been identified associated with genes expressed in heat shock (Nikolaidis and Nei 2004), muscle (GuhaThakurta et al. 2004), and the gut (Gaudet et al. 2004), particularly the pharynx. For example, Mango and her colleagues (Gaudet et al. 2004) identified genes expressed in the pharynx by comparing mutant embryos enriched and depleted of pharyngeal cells using microarrays. They grouped the genes by early or late expression and then looking between species and across genes, they identified nine candidate regulatory motifs, two of which were previously known. They confirmed several of these for activity in vivo and, in turn, used the motifs to search for additional genes with the motifs. The resultant sets were significantly enriched for genes expressed in the pharynx. This strategy should become more powerful as additional *Caenorhabditis* genome sequences become available and as gene expression data are refined.

Parallel to expression data and comparative genome analysis, investigators have attempted to identify the target sequence for known transcription factors. Using the yeast one-hybrid system, the motifs recognized by the DNA-binding domains of the

worm's ~600 transcription factors are being systematically dissected (Deplancke et al. 2004). Others are exploring ways to apply chromatin precipitation to discover the in vivo sites of protein–DNA interaction and to use DNAse I hypersensitivity to find regions of open chromatin. SELEX (sytematic evolution of ligands by exponential amplification) offers another approach to identify binding motifs that might be applied at scale (Roulet et al. 2002). Combining knowledge of transcription-factor-binding sites and the identification of functional sites associated with genes could provide powerful insights into the networks of gene regulation that underlie development.

Many motifs encoded in DNA within genes act at the RNA level to regulate splicing, localization, translation, RNA editing, or other processes. These RNA regulatory elements can be studied in largely the same fashion as the transcriptional regulatory elements: sequence conservation can be used to identify candidate elements; pull-down experiments can link RNA-binding proteins to their target genes and candidate motifs; function can be assayed by fusions with reporter proteins. *C. elegans* has ~500 RNA-binding proteins, and genetic, biochemical, and computational analyses have revealed critical roles of protein–RNA complexes, 3′-untranslated regions, RNA-binding proteins (and their targets), and RNA–RNA interactions in development.

A complication for defining the RNA regulatory elements is that the regulatory information often resides inside the three-dimensional RNA secondary structure rather than be encoded directly in the primary sequence. This makes it more difficult to predict regulatory elements computationally. The computational prediction of RNA regulatory elements must proceed hand in hand with the structural analysis of the RNA genes that regulate them.

Proteomics

With the well-annotated *C. elegans* genome in hand, both the study of individual proteins and the study of interactions among those proteins can proceed. In a high-throughput proteomic effort to confirm protein-coding genes, G. Merrihew, J.H. Thomas, and M.J. MacCoss (unpubl.) are using mass spectrometry to validate experimentally even small predicted ORFs. They currently have identified 3363 proteins, 121 of which previously had no experimental support (39 of these were identified based on a translated intergenic ORF set, and the remainder from GeneFinder predictions) (P. Green, unpubl.). Others are finding success using mass spectrometry to quantify relative protein levels in *C. elegans* embryos and adults (Venable et al. 2004). Mass spectrometry approaches should also reveal post-translational modifications that may alter activity.

The study of individual protein structures is also well under way. A *C. elegans* structural genomics group has formed a high-throughput protein-to-structure pipeline (Liu et al. 2005b). They have determined the crystal structure of 78 proteins or protein fragments (http://sgce.cbse.uab.edu/index.php) and solved 19 structures (e.g., Symersky et al. 2003; Lu et al. 2004). Another structural genomics effort (http://www.nesg.org; Wunderlich et al. 2004) identified seven structures.

Identifying protein–protein interactions and the effects of any modifications on those interactions will be key to any molecular understanding of the worm. Computational-aided methods, some using comparative data (Liu et al. 2005a; Sharan et al. 2005), have the potential for revealing these, but large-scale studies depend on experimental data. Armed with the set of 11,000 cloned ORFs (*C. elegans* ORFeome project) (Lamesch et al. 2004), researchers have generated a *C. elegans* interactome network map that contains >5500 potential interactions (Li et al. 2004) and are moving toward defining the entire set. Along with another map for *Drosophila melanogaster* (Sanchez et al. 1999), these data sets, although containing high proportions of false positives and negatives, nevertheless represent the first of their kind for metazoan organisms. Critically, the interactome map serves as a foundation for integration of studies of development and disease, both for individual proteins and at the level of networks of interactions.

Population biology and evolution

Beyond aiding in a molecular understanding of the form and behavior of the worm, the genome sequence has also facilitated studies of the evolutionary processes acting on the worm genome. While we cannot access *C. elegans* ancestors, comparative analysis allows inferences about that ancestral state and the events that have occurred since the divergence of two species.

With the sequence of the laboratory strain N2 in hand, the study of variation in different isolates of *C. elegans* from around the world became straightforward. Variation could be readily determined either through using PCR to recover specific areas or from random whole-genome sequence reads from these different isolates aligned with the N2 sequence. A patchwork pattern of variation within most isolates suggested that most isolates had resulted from an interbreeding event followed by isolation, perhaps facilitated by hermaphroditic reproduction (Koch et al. 2000). Surprisingly, there is high population diversity at the local level—on the scale of centimeters—but the diversity levels off very quickly so that there is about the same amount of diversity among isolates from different countries as among isolates from the same compost heap (Fitch 2005).

Among the different isolates, the Hawaiian strain, CB4856, however, proved to have widely and more uniformly dispersed sequence differences (Wicks et al. 2001). A difference was observed once every 850 bases, with transitions outnumbering transversions (57% vs. 43%) and indels (one or more bases added or removed) accounting for more than one-quarter of the differences. Somewhat surprisingly, this rate of difference suggests an effective population size not much different from that of humans. Recent comparison of the genomes using comparative genome hybridizations with microarrays reveals a surprising number of larger deletions in the Hawaiian strain (D. Moerman, pers. comm.). The single nucleotide polymorphisms (SNPs) have also provided the basis for an effective genetic mapping strategy (Wicks et al. 2001; Swan et al. 2002).

C. elegans autosomes have an unusual organization, with recombination significantly elevated on the terminal thirds compared to the centers. Essential genes are more frequently located in the centers in contrast to gene families, which are overrepresented on the arms. This has led to speculation that the arms are sites of high gene death and birth. Consistent with this notion, SNP density appears to be elevated on the arms (Koch et al. 2000). Comparison of the *C. elegans* and *C. briggsae* genomes has shown dramatic differences in expansion of chemosensory genes on the arms in the two species (Chen et al. 2005b) and for positive selection of members of the srz family of G-protein-coupled receptors (Thomas et al. 2005) also clustered on the arms. Furthermore, protein and regulatory evolution is weakly coupled in orthologs but not paralogs, and duplicates of both species show acceleration of both regulatory and protein evolution compared to orthologs (Castillo-Davis et al. 2004). Strikingly, the *C. briggsae*

genome shows the same pattern of high recombination on the autosome arms, showing that this is a well-established feature of genome architecture (R.H. Waterston, L.W. Hillier, S. Baird, and R. Miller, unpubl.).

Comparative studies of the five *Caenorhabditis* genomes may also shed light on the evolution of the hermaphrodite–male mode of reproduction, which is believed to have evolved independently in *C. elegans* and *C. briggsae*. The other three *Caenorhabditis* species have female–male sexual systems. Just comparison of the genomes of the two self-fertilizing species have yielded insights into the dynamics of sex and gamete-specific gene evolution (Kiontke et al. 2004; Cutter and Ward 2005; Nayak et al. 2005) and the genomic organization of reproductive genes (Miller et al. 2004). Intriguingly, the genomes of both *Caenorhabditis remanei* and *Caenorhabditis n. sp. PB2801* are significantly larger than the genomes of the self-fertilizing species (J.S. Johnston, pers. comm.).

WormBase

Central to making all this information available to the community has been the ongoing development of WormBase (Chen et al. 2005a; http://www.wormbase.org), an outgrowth of ACeDB (A *C. elegans* database; http://www.acedb.org). ACeDB was developed in conjunction with the genome project to coordinate the effort to integrate the sequence with the genetic and physical maps and to provide public access to the project and its data.

WormBase contains a wide range of information about the biology and genomics of the worm. It acts as the repository of all the genome annotation for *C. elegans* as well as *C. briggsae* and related nematodes. It curates gene models, reconciling the predictions and the various experimental data sets. It acquires associated functional information from high-throughput experiments and more traditional experiments reported in the literature. WormBase also contains an extensive bibliography of papers published on *C. elegans* along with unpublished abstracts from regional meetings and the biennial International Worm Meetings and the brief reports in the *Worm Breeder's Gazette*.

WormBase supports five different methods of access through its interactive Web interface, with each adapted to specific purposes. These are

1. Web browsing for the casual user, with simple queries and navigation through a variety of displays;
2. batch retrieval for gene and sequence fields;
3. query language searching allowing ad hoc queries for more sophisticated users;
4. bulk downloads of gene sets, other data sets, or even the entire database to provide local access; and
5. scripting to allow formatting and processing of query results for those with some programming skills.

WormBase also supports the Distributed Annotation System (DAS, also developed in conjunction with the worm genome project) (Dowell et al. 2001) allowing users to add their own data tracks to browser displays.

WormBase continues to evolve, improving user interfaces and adding new data sets, such as movies, protein structures, and new genome sequences.

Conclusions

The *C. elegans* genome sequence, now complete, has spurred research on the worm to an extent only dimly foreseen by the early advocates of the genome project. The impact extends beyond the large data sets, the sequencing of additional nematode genomes, and the development of WormBase. The sequence and the associated resources have empowered individual worm labs to investigate central biological issues, rather than the process of cloning and sequencing. It also places their work in a larger context. The abundance of resources has also drawn very talented new investigators into the field. The worm leads the field in studies of apoptosis, aging, development, neurobiology, and other areas.

But the impact extends well beyond the worm field itself. Stimulated by successes in *C. elegans*, ESTs have been generated for almost every major class of nematode parasites of humans (Mitreva et al. 2005), and with the *C. elegans* genome as a point of reference, these data sets are opening new avenues to conquer these insidious diseases. Nematode-specific genes provide potential drug targets, with *C. elegans* able to serve as an initial testbed for evaluating candidate compounds.

More broadly, the worm sequence, through GenBank and the browsers (UCSC ENSEMBL, NCBI), provides a portal to the worm for investigators of other organisms. Either through direct homology searches or through established orthology tables, scientists can rapidly learn that *C. elegans* has a gene related to their gene of interest and then from WormBase and the literature learn what is known about that gene. They may well be drawn into the field to study the gene in worms, because of the ease of experimentation and wealth of resources. Many a worm researcher has had colleagues appear in their office asking about how to do experiments with the worm. New collaborations result, with the worm field enormously enriched by these "outsiders" perspectives, opening up possibilities for impact on human health and well being that otherwise might have been missed.

The impact of the *C. elegans* genome project extends in other directions. The early success of the *C. elegans* EST project was the direct forerunner of the large-scale public domain human and mouse EST projects, without which mammalian microarray and proteomic investigations in the 1990s would have been extremely limited. In genome sequencing, the worm project demonstrated the feasibility of using Sanger-based sequencing methods and a hierarchical (clone-based) shotgun strategy for the Human Genome Project. Significantly, the worm project also provided the model for the data release policies of the Human Genome Project. The worm genome project had adopted from the start a policy of rapid and open data release, extending the practice of early data sharing of the worm community. This policy drew the worm labs into the project, led to a clear delineation of tasks (the genome centers provided the sequence and the individual labs gave it biological meaning), and accelerated the impact of the sequence.

But the task of understanding the worm at a molecular level has just begun. Having captured the large but finite information of the genome, we can now begin to see the enormity of the task before us. We need a full parts list, not just the protein-coding genes, but the RNA genes, the regulatory elements, and any other functional elements of the genome. We need to know the motifs that transcription factors bind in vivo, and that has to be coupled with a precise knowledge of when, where, and at what level each gene is expressed. *C. elegans* is probably the only experimental animal in which the resolution can be at the single-cell level throughout development; we should exploit this. With this information, the regulatory networks that control development should emerge, yielding circuit diagram models of development. Success with this lowly nematode will again

have profound impact on the efforts to extend this knowledge to human biology, with all its implications for human health and well-being.

But we need to move beyond this network view to achieve a true molecular understanding of worm biology. Protein function will have to be defined in detail. RNAi knock-downs, gene knock-outs, and protein–protein interaction networks will be a start, but our knowledge of function will have to go much deeper. Undoubtedly we will need to understand the small molecule component of cells and their flux as well.

These will be challenging studies as we delve deeper and deeper into the molecular description. It will take common resources, new methods, and perseverance. But the synergy of hypothesis-driven and data-driven science of the past decade combined with the spirit of the community are major assets. These, with the inherent advantages of the worm so presciently recognized by Brenner more than 40 years ago, make *C. elegans* the prime candidate for achieving such a grandiose goal. We can't let the opportunity pass.

Acknowledgments

The authors gratefully acknowledge the members of the *C. elegans* Sequencing Consortium as well as the team members of WormBase. The authors also thank Heidi Browning, Cindi Madej, and Susan Strome for their advice and gifts of macro-restriction Southern blots. We also thank Tim Schedl, Don Moerman, Lincoln Stein, and Paul Sternberg for their comments on the manuscript. This work has been funded by the UK Medical Research Council, the Wellcome Trust, and the National Human Genome Research Institute and the National Institute of General Medical Sciences at the National Institutes of Health.

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.

Baugh, L.R., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. 2003. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* **130:** 889–900.

Baugh, L.R., Wen, J.C., Hill, A.A., Slonim, D.K., Brown, E.L., and Hunter, C.P. 2005. Synthetic lethal analysis of *Caenorhabditis elegans* posterior embryonic patterning genes identifies conserved genetic interactions. *Genome Biol.* **6:** R45.

Blacque, O.E., Perens, E.A., Boroevich, K.A., Inglis, P.N., Li, C., Warner, A., Khattra, J., Holt, R.A., Ou, G., Mah, A.K., et al. 2005. Functional genomics of the cilium, a sensory organelle. *Curr. Biol.* **15:** 935–941.

Blumenthal, T. 2005. *Trans*-splicing and operons. In *WormBook* (ed. The *C. elegans* Research Community). doi/10.1895/wormbook.1.5.1, http://www.wormbook.org.

Blumenthal, T. and Gleason, K.S. 2003. *Caenorhabditis elegans* operons: Form and function. *Nat. Rev. Genet.* **4:** 112–120.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Castillo-Davis, C.I., Hartl, D.L., and Achaz, G. 2004. *cis*-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* **14:** 1530–1536.

The *C. elegans* Genome Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.K., et al. 2005a. WormBase: A comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.* **33:** D383–D389.

Chen, N., Pai, S., Zhao, Z., Mah, A., Newbury, R., Johnsen, R.C., Altun, Z., Moerman, D.G., Baillie, D.L., and Stein, L.D. 2005b. Identification of a nematode chemosensory gene family. *Proc. Natl. Acad. Sci.* **102:** 146–151.

Christensen, M., Estevez, A., Yin, X., Fox, R., Morrison, R., McDonnell, M., Gleason, C., Miller III, D.M., and Strange, K. 2002. A primary culture system for functional analysis of *C. elegans* neurons and muscle cells. *Neuron* **33:** 503–514.

Coulson, A., Sulston, J., Brenner, F.R.S., and Karn, J. 1986. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **83:** 7821–7825.

Coulson, A., Waterston, R., Kiff, J., Sulston, J., and Kohara, Y. 1988. Genome linking with yeast artificial chromosomes. *Nature* **335:** 184–186.

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R. 1991. YACs and the *C. elegans* genome. *Bioessays* **13:** 413–417.

Coulson, A., Huynh, C., Kozono, Y., and Shownkeen, R. 1995. The physical map of the *Caenorhabditis elegans* genome. *Methods Cell Biol.* **48:** 533–550.

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

Cutter, A.D. and Ward, S. 2005. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. *Mol. Biol. Evol.* **22:** 178–188.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J. 2004. A gateway-compatible yeast one-hybrid system. *Genome Res.* **14:** 2093–2101.

Doggett, N.A., Smith, C.L., and Cantor, C.R. 1992. The effect of DNA concentration on mobility in pulsed field gel electrophoresis. *Nucleic Acids Res.* **20:** 859–864.

Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., and Stein, L. 2001. The distributed annotation system. *BMC Bioinformatics* **2:** 7.

Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., et al. 2004. A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* **14:** 2169–2175.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391:** 806–811.

Fitch, D.H. 2005. Evolution: An ecological context for *C. elegans*. *Curr. Biol.* **15:** R655–R658.

Fox, R.M., Von Stetina, S.E., Barlow, S.J., Shaffer, C., Olszewski, K.L., Moore, J.H., Dupuy, D., Vidal, M., and Miller III, D.M. 2005. A gene expression fingerprint of *C. elegans* embryonic motor neurons. *BMC Genomics* **6:** 42.

Fraser, A.G., Kamath, R.S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer, J. 2000. Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408:** 325–330.

Fukushige, T., Schroeder, D.F., Allen, F.L., Goszczynski, B., and McGhee, J.D. 1996. Modulation of gene expression in the embryonic digestive tract of *C. elegans*. *Dev. Biol.* **178:** 276–288.

Gaudet, J., Muttumu, S., Horner, M., and Mango, S.E. 2004. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol.* **2:** e352.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res.* **32:** D109–D111.

GuhaThakurta, D., Schriefer, L.A., Waterston, R.H., and Stormo, G.D. 2004. Novel transcription regulatory elements in *Caenorhabditis elegans* muscle genes. *Genome Res.* **14:** 2457–2468.

Holt, S.J. and Riddle, D.L. 2003. SAGE surveys *C. elegans* carbohydrate metabolism: Evidence for an anaerobic shift in the long-lived dauer larva. *Mech. Ageing Dev.* **124:** 779–800.

Hwang, B.J., Muller, H.M., and Sternberg, P.W. 2004. Genome annotation by high-throughput 5′ RNA end determination. *Proc. Natl. Acad. Sci.* **101:** 1650–1655.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

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.

Kamath, R.S. and Ahringer, J. 2003. Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods* **30:** 313–321.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.

Kimble, J. and Hirsh, D. 1979. The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Dev. Biol.* **70:** 396–417.

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.*

101: 9003–9008.
Koch, R., van Luenen, H.G., van der Horst, M., Thijssen, K.L., and Plasterk, R.H. 2000. Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. *Genome Res.* **10:** 1690–1696.
Kohara, Y. 1996. [Large scale analysis of *C. elegans* cDNA]. *Tanpakushitsu Kakusan Koso* **41:** 715–720.
Korf, I., Flicek, P., Duan, D., and Brent, M.R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1:** S140–S148.
Kunitomo, H., Uesugi, H., Kohara, Y., and Iino, Y. 2005. Identification of ciliated sensory neuron-expressed genes in *Caenorhabditis elegans* using targeted pull-down of poly(A) tails. *Genome Biol.* **6:** R17.
Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhaute, J., Hill, D.E., et al. 2004. *C. elegans* ORFeome version 3.1: Increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14:** 2064–2069.
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.
Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.
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.
Liu, Y., Liu, N., and Zhao, H. 2005a. Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21:** 3279–3285.
Liu, Z.J., Tempel, W., Ng, J.D., Lin, D., Shah, A.K., Chen, L., Horanyi, P.S., Habel, J.E., Kataeva, I.A., Xu, H., et al. 2005b. The high-throughput protein-to-structure pipeline at SECSG. *Acta Crystallogr. D Biol. Crystallogr.* **61:** 679–684.
Lu, S., Symersky, J., Li, S., Carson, M., Chen, L., Meehan, E., and Luo, M. 2004. Structural genomics of *Caenorhabditis elegans*: Crystal structure of the tropomodulin C-terminal domain. *Proteins* **56:** 384–386.
Luersen, K., Eschbach, M.L., Liebau, E., and Walter, R.D. 2004. Functional GATA- and initiator-like-elements exhibit a similar arrangement in the promoters of *Caenorhabditis elegans* polyamine synthesis enzymes. *Biol. Chem.* **385:** 711–721.
Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. 2001. Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* **11:** 171–176.
McCallum, C.M., Comai, L., Greene, E.A., and Henikoff, S. 2000. Targeted screening for induced mutations. *Nat. Biotechnol.* **18:** 455–457.
McCombie, W.R., Adams, M.D., Kelley, J.M., FitzGerald, M.G., Utterback, T.R., Khan, M., Dubnick, M., Kerlavage, A.R., Venter, J.C., and Fields, C. 1992. *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nat. Genet.* **1:** 124–131.
McKay, S.J., Johnsen, R., Khattra, J., Asano, J., Baillie, D.L., Chan, S., Dube, N., Fang, L., Goszczynski, B., Ha, E., et al. 2003. Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68:** 159–169.
Miller, M.A., Cutter, A.D., Yamamoto, I., Ward, S., and Greenstein, D. 2004. Clustered organization of reproductive genes in the *C. elegans* genome. *Curr. Biol.* **14:** 1284–1290.
Mitreva, M., Blaxter, M.L., Bird, D.M., and McCarter, J.P. 2005. Comparative genomics of nematodes. *Trends Genet.* **21:** 573–581.
Nayak, S., Goree, J., and Schedl, T. 2005. fog-2 and the evolution of self-fertile hermaphroditism in *Caenorhabditis*. *PLoS Biol.* **3:** e6.
Nelson, D.W. and Honda, B.M. 1985. Genes coding for 5S ribosomal RNA of the nematode *Caenorhabditis elegans*. *Gene* **38:** 245–251.
Nikolaidis, N. and Nei, M. 2004. Concerted and nonconcerted evolution of the Hsp70 gene superfamily in two sibling species of nematodes. *Mol. Biol. Evol.* **21:** 498–505.
Okkema, P.G. and Fire, A. 1994. The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120:** 2175–2186.
Okkema, P.G., Ha, E., Haun, C., Chen, W., and Fire, A. 1997. The *Caenorhabditis elegans* NK-2 homeobox gene ceh-22 activates pharyngeal muscle gene expression in combination with pha-1 and is required for normal pharyngeal development. *Development* **124:** 3965–3973.
Piano, F., Schetter, A.J., Mangone, M., Stein, L., and Kemphues, K.J. 2000. RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr. Biol.* **10:** 1619–1622.
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.
Rivas, E. and Eddy, S.R. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2:** 8.
Roulet, E., Busso, S., Camargo, A.A., Simpson, A.J., Mermod, N., and Bucher, P. 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.* **20:** 831–835.
Roy, P.J., Stuart, J.M., Lund, J., and Kim, S.K. 2002. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418:** 975–979.
Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roder, L., Euzenat, J., Rechenmann, F., and Jacq, B. 1999. Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.* **27:** 89–94.
Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., and Ideker, T. 2005. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci.* **102:** 1974–1979.
Sonnichsen, B., Koski, L.B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A.M., Artelt, J., Bettencourt, P., Cassin, E., et al. 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434:** 462–469.
Spieth, J. and Lawson, D. 2005. Overview of gene structure. In *WormBook* (ed. The *C. elegans* Research Community). doi/10.1895/wormbook.1.5.1, http://www.wormbook.org.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1:** E45.
Stricklin, S.L., Griffiths-Jones, S., and Eddy, S.R. 2005. *C. elegans* noncoding RNA genes. In *WormBook* (eds. J. Hodgkin and P. Anderson). doi/10.1895/wormbook.1.1.1, http://www.wormbook.org.
Sulston, J.E. and Brenner, S. 1974. The DNA of *Caenorhabditis elegans*. *Genetics* **77:** 95–104.
Sulston, J. and Ferry, G. 2002. *The common thread: A story of science, politics, ethics and the human genome*. Bantam Press, London.
Sulston, J.E. and Horvitz, H.R. 1977. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56:** 110–156.
Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100:** 64–119.
Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T., and Coulson, A. 1988. Software for genome mapping by fingerprinting techniques. *Comput. Appl. Biosci.* **4:** 125–132.
Swan, K.A., Curtis, D.E., McKusick, K.B., Voinov, A.V., Mapa, F.A., and Cancilla, M.R. 2002. High-throughput gene mapping in *Caenorhabditis elegans*. *Genome Res.* **12:** 1100–1105.
Symersky, J., Lin, G., Li, S., Qiu, S., Carson, M., Schormann, N., and Luo, M. 2003. Structural genomics of *Caenorhabditis elegans*: Crystal structure of calmodulin. *Proteins* **53:** 947–949.
Teng, Y., Girard, L., Ferreira, H.B., Sternberg, P.W., and Emmons, S.W. 2004. Dissection of *cis*-regulatory elements in the *C. elegans* Hox gene egl-5 promoter. *Dev. Biol.* **276:** 476–492.
Thomas, J.H., Kelley, J.L., Robertson, H.M., Ly, K., and Swanson, W.J. 2005. Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc. Natl. Acad. Sci.* **102:** 4476–4481.
Timmons, L. and Fire, A. 1998. Specific interference by ingested dsRNA. *Nature* **395:** 854.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. 2004. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1:** 39–45.
Vidalain, P.O., Boxem, M., Ge, H., Li, S., and Vidal, M. 2004. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32:** 363–370.
Wang, D., Kennedy, S., Conte Jr., D., Kim, J.K., Gabel, H.W., Kamath, R.S., Mello, C.C., and Ruvkun, G. 2005. Somatic misexpression of germline P granules and enhanced RNA interference in retinoblastoma pathway mutants. *Nature* **436:** 593–597.
Washietl, S., Hofacker, I.L., and Stadler, P.F. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci.* **102:** 2454–2459.
Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., et al. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nat. Genet.* **1:** 114–123.
Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M.,

and Brent, M.R. 2005. Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.* **15:** 577–582.
White, J.G., Southgate, E., Thomson, J.N., and Brenner, F.R.S. 1986. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. Roy. Soc. London Ser. B Biol. Sci.* **314:** 1–340.
Wicks, S.R., Yeh, R.T., Gish, W.R., Waterston, R.H., and Plasterk, R.H. 2001. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat. Genet.* **28:** 160–164.
Wicky, C., Villeneuve, A.M., Lauper, N., Codourey, L., Tobler, H., and Muller, F. 1996. Telomeric repeats (TTAGGC)n are sufficient for chromosome capping function in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci.* **93:** 8983–8988.
Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell* **75:** 855–862.
Wunderlich, Z., Acton, T.B., Liu, J., Kornhaber, G., Everett, J., Carter, P., Lan, N., Echols, N., Gerstein, M., Rost, B., et al. 2004. The protein target list of the Northeast Structural Genomics Consortium. *Proteins* **56:** 181–187.
Zhang, S., Ma, C., and Chalfie, M. 2004. Combinatorial marking of cells and organelles with reconstituted fluorescent proteins. *Cell* **119:** 137–144.

Web site references

http://www.gurdon.cam.ac.uk/~ahringerlab/pages/rnai.html; *C. elegans* RNAi methods and resources.
http://129.11.204.86:591/default.htm; *C. elegans* Expression Pattern Database, Hope Laboratory.
http://elegans.bcgsc.ca/home/ge_consortium.html; British Columbia *C. elegans* Gene Expression Consortium.
http://elegans.imbb.forth.gr/nemagenetag/home.html; Nematode Gene-Tagging Tools and Resources, NemaGENETAG Consortium.
http://microrna.sanger.ac.uk/sequences/; microRNA database.
http://nematode.lab.nig.ac.jp/db2/index.php; Nematode Expression Pattern Database (NEXTDB).
http://sgce.cbse.uab.edu/index.php; Southeast Collabortory for Structural Genomics of *C. elegans*.
http://shigen.lab.nig.ac.jp/c.elegans/index.jsp; National Bioresource Center–*C. elegans* (Japan).
http://www.acedb.org; AceDB database.
http://www.affymetrix.com/products/arrays/specific/celegans.affx; Affymetrix *C. elegans* genome array.
http://www.bcgsc.ca/gc/celegans/; *C. elegans* Gene Expression Studies, BC Genome Sciences Centre.
http://www.celeganskoconsortium.omrf.org/; *C. elegans* Gene Knockout Consortium.
http://www.genome.gov/10002154; NHGRI Genome Sequencing Proposals.
http://www.genome.gov/11007952; National Human Genome Research Institute (NHGRI) Roundworm Genome Sequencing Program.
http://www.genome.wustl.edu/genome/celegans/microarray/ma_gen_info.cgi; Long Oligomer-based Microarrays for the *C. elegans* genome.
http://www.nesg.org; Northeast Structural Genomics Consortium.
http://www.nig.ac.jp/section/kohara/kohara-e.html; Genome Biology of *C. elegans* Development, Kohara Laboratory.
http://www.nimblegen.com/products/; NimbleGen Systems, Inc. products and services.
http://www.wormbase.org; WormBase database.
http://www.wormbook.org; WormBook.
http://www.exelixis.com; Exelixis, an integrated drug discovery and development company.