2 Genome Mapping and Genomics of *Caenorhabditis elegans*

Jonathan Hodgkin, Michael Paulini, and Mary Ann Tuli

2.1 Introduction to *Caenorhabditis elegans*: Key Experimental Advantages

The small nematode worm *Caenorhabditis elegans* was chosen as a subject for intensive study in the 1960s by Sydney Brenner, and since that time it has become one of the major model organisms for laboratory investigation of a great variety of biological problems. Currently more than 500 laboratories around the world make use of *C. elegans* as a research tool, and the bibliography on this organism now exceeds 10,000 papers. In 1998, it became the first multicellular organism for which a complete genome sequence was determined (*C. elegans* Sequencing Consortium 1998). As a consequence, genomic, and post-genomic studies of *C. elegans* have been very extensive, and the past 14 years have seen an enormous increase in the analysis and understanding of this key genome.

Information on the biology and genome of *C. elegans* is stored on the interactive database WormBase (http://www.wormbase.org/), which can be used to explore all properties of this organism. Anatomical features can be examined in greater detail in WormAtlas (http://www.wormatlas.org/). A companion on-line narrative set of reviews and methods is provided by WormBook (http://www.wormbook.org/). Most of the material in this chapter is covered in greater detail in one or another of the many chapters in WormBook. The genomic data accessible on WormBase are regularly updated, on a 2–4 week cycle; most numbers cited in this chapter are derived from WormBase release WS228.

In nature, *C. elegans* occurs as a free-living, non-parasitic worm, which can be found most readily in decaying plant material such as compost heaps and rotting fruit, where it grows by eating bacteria. It has a global distribution, and isolates of the species have been obtained from many different countries in north and south temperate zones. In the laboratory, it is usually cultured by growth on lawns of *E. coli* bacteria, spread on agar plates (Brenner 1974).

Major experimental advantages of the worm include the ease and cheapness of culture because no special media are required, and because *C. elegans* grows well at room temperature (viable range ~12–25 °C). Its generation time is short, 3 days from egg to egg at 25 °C, which enables rapid experimentation and genetic manipulation. The worms can be stored on starved plates for many weeks at room temperature, or as frozen stocks in liquid nitrogen. Such frozen stocks remain viable indefinitely, with no chance of genetic change, unlike stocks passaged during laboratory culture.

J. Hodgkin (✉)
Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK
e-mail: jonathan.hodgkin@bioch.ox.ac.uk

M. Paulini • M.A. Tuli
Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
e-mail: mt3@sanger.ac.uk; mh6@sanger.ac.uk

P. Denny and C. Kole (eds.), *Genome Mapping and Genomics in Laboratory Animals*,
Genome Mapping and Genomics in Animals 4, DOI 10.1007/978-3-642-31316-5_2,

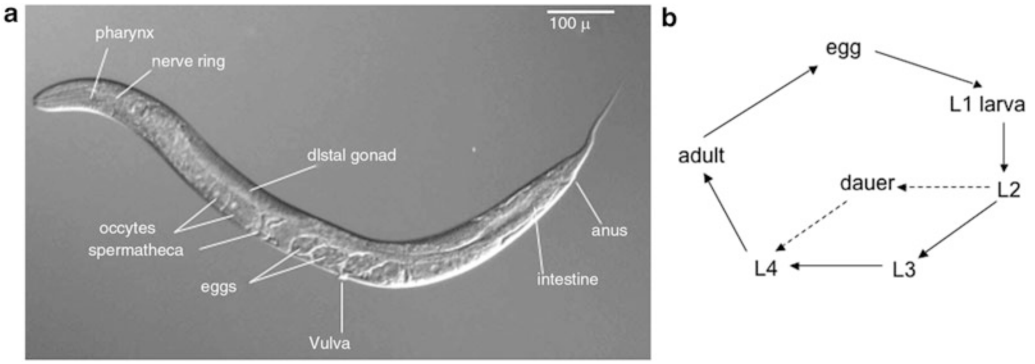

Fig. 2.1 (**a**) Differential interference contrast (Nomarski) image of an adult hermaphrodite of *C. elegans*, illustrating major anatomical features. Adults and larvae have the same general body plan, having a muscular body wall surrounding an internal digestive tract, which runs from the mouth and pharynx (used for grinding up bacterial food) to the anus. The adult has, in addition, a twin-armed gonad that occupies much of the body cavity. Within this gonad, first sperm, and then oocytes differentiate. At the adult stage, oocytes mature and are fertilized by sperm, and then begin development as eggs in the uterus for a short time, before being laid through the centrally located vulva. Photograph provided by Maria Gravato-Nobre. (**b**) Life cycle of *C. elegans*. A self-fertile hermaphrodite lays about 300 eggs, each of which hatches into a small first stage larva (L1). The larvae go through three further larval stages (L2, L3, and L4), separated by molts, before maturing into the fertile adult stage. Under poor nutritional conditions, L2 larvae molt to give an alternative larval form, the dauer larva, which can survive adverse conditions. After provision of food, dauers resume development and molt to give normal L4 larvae. Most (>99 %) progeny produced by an XX hermaphrodite are also XX, but rare progeny lack an X chromosome and mature into XO males, which can cross-fertilize hermaphrodites

Under normal growth conditions, populations consist almost entirely of hermaphrodite worms. Each worm hatches from an egg as a first-stage larva, which grows and molts four times to mature into a self-fertile hermaphrodite adult (Fig. 2.1). Each hermaphrodite produces first sperm and then oocytes from a common pool of germ-line cells. Fertilization occurs internally, followed by an initial phase of embryonic development in the uterus, and eggs are then laid from the centrally located vulva. Each hermaphrodite lays about 300 eggs (self-progeny) during its lifetime. The ability to reproduce by self-fertilization allows more rapid population growth and simplifies propagation of mutant lines, because no mating between individuals is required. It also greatly facilitates genetic screens, because recessive mutations will automatically segregate as homozygotes at each successive generation. Thus, a hermaphrodite that is heterozygous for a recessive mutation causing uncoordinated movement (written *unc*/+) will produce self-progeny in the Mendelian ratio of 25 % *unc*/*unc* homozygotes, which will express the uncoordinated phenotype.

Conveniently, the animal can reproduce by cross-fertilization as well as self-fertilization. Hermaphrodites have a diploid karyotype of 12 chromosomes: five pairs of autosomes and two X chromosomes (abbreviated XX). Individuals with five pairs of autosomes and a single X chromosome (abbreviated XO) arise at low frequency as a result of rare meiotic loss of an X chromosome, and these individuals are males. Their germlines produce only sperm, and they exhibit extensive anatomical and behavioral differences from hermaphrodites, which they can mate with and cross-fertilize. After mating, sperm from the male are used preferentially over the hermaphrodites own sperm. This mating between hermaphrodites and males allows conventional genetic crosses and cross-breeding.

The animal is transparent throughout its lifecycle, which has proved to be of great importance in its exploitation for experimental purposes. Developmental and cellular events can be examined directly and non-invasively in real time, in the living animal. Moreover, the advent of fluorescent protein technology has led to the generation of thousands of different transgenic strains expressing particular proteins tagged with green fluorescent protein (GFP) or its derivatives, permitting further in vivo observation and manipulation.

The small size of the animal means it contains relatively few cells—fewer than one thousand somatic cells in the mature adult—but these cells are well differentiated into distinct tissue types (muscle, gut, skin, nerves) as in more complex animals. The cell lineages and resulting anatomical structures are highly invariant from animal to animal. This invariance has several advantageous consequences. One is that it made it possible to describe the complete cell lineage, from egg to adult. Another is that the entire nervous system could be reconstructed from serial section electron micrographs, leading to a complete "wiring diagram" for the 302 neurons and ~8,000 synapses in this animal. Both the complete cell lineage and the complete wiring diagram are feats of description that are unlikely to be replicated in any more complicated organisms. The high degree of invariance is also useful from an experimental standpoint, because deviations from normal patterns of development or behavior are readily detected and can be reliably ascribed to mutation or manipulation rather than to environmental variability.

The technical advantages mentioned above fuelled much of the initial exploitation of *C. elegans* as an experimental system, but for the past decade the availability of a complete genome sequence, and the development of associated technologies, have driven much greater expansion of its use. The most important of these recent developments has been the use of RNA interference (RNAi). This phenomenon was first discovered in *C. elegans* (Fire et al. 1998), and can be applied in uniquely powerful ways to manipulate and explore this system, as discussed in a later section.

2.2 Genome Mapping

Historically, the genome was first mapped at a recombinational level by means of classical genetic crosses (Brenner 1974). Many hundreds of mutations affecting body morphology, locomotion, or other easily scored features were first generated by means of chemical mutagenesis, and assigned to specific genes by complementation tests. Linkage tests between different genes then allowed assignment of genes to particular linkage groups, and to construction of a recombinational map for each of the six linkage groups that could be inferred. These corresponded to the six pairs of cytologically visible chromosomes. The chromosomes of *C. elegans* are very small, owing to their low DNA content (all <20 Mb) and have almost no distinctive visible features (Fig. 2.2). Each has a length, in standard recombinational units, of 45–50 centiMorgans, which means that each chromosome usually experiences a single crossover event at meiosis.

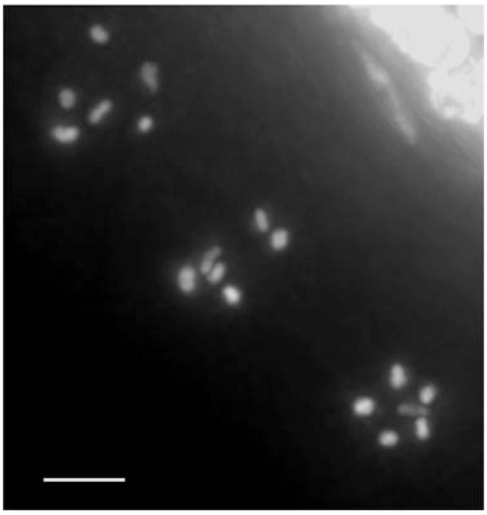

Fig. 2.2 Fluorescence micrograph of three oocyte nuclei stained with DAPI to reveal meiotic chromosomes. At this stage the oocytes are arrested in meiosis I, with the 12 chromosomes paired as six bivalents. The extremely small size of the chromosomes is evident, as is the lack of discernible features such as chromosome bands or constrictions. Scale bar = ~10 μm. Photograph provided by Theresa Zucchero and Shawn Ahmed

Early (pre-molecular) genetic maps were based on data for several hundred genes with visible mutant phenotypes. These were nonrandomly distributed on the autosomes, with conspicuous clustering of most genes in a central cluster flanked by arms with fewer visible markers. In the absence of sequence data, it was not clear whether these clusters represented regions of higher gene density, or higher density of important genes, or lower frequency of recombination. Genomic analysis, discussed below, indicates that all three of these factors contribute to the arm–cluster–arm topology of the autosomes. Clustering is much less apparent on the X chromosome.

With the advent of recombinant DNA technologies, it became possible to clone genes that had been initially defined by mutation and studied for interesting biological properties such as specific effects on differentiation, or neuronal or muscular function. At first, gene cloning was pursued on a gene-by-gene basis, achieved most often by means of transposon-tagging, but these approaches were then powerfully supplemented by a project aimed at generating a complete physical map of the *C. elegans* genome (Coulson et al. 1986). This physical map could then be aligned with the existing genetic map, enormously aiding the positional cloning of mutationally defined genes. This physical map was initially generated by creating and analyzing large libraries of cosmid clones (average size about 40 kb), which were later supplemented by sets of larger yeast artificial chromosome (YAC) clones in order to cover regions of the genome that were missing from the initial libraries (Coulson et al. 1988). Overlaps between clones were identified by means of an efficient high-throughput partial restriction mapping approach, which allowed assembly of larger and larger "contigs" (sets of continuously overlapping clones), eventually covering most of each chromosome.

The physical mapping project consequently resulted in coverage of almost all of the *C. elegans* genome in an ordered array of clones, with much redundancy. This resource, together with the small size of the genome, was a major factor in justifying the complete sequencing of *C. elegans*, at a time when large-scale genomic sequencing was only just becoming feasible.

In addition to their role in enabling whole genome sequencing, the ordered libraries of cosmid and YAC clones remained useful as agents for gene analysis and discovery, for example in their use for transgenic rescue of mutant strains. They have now been largely superseded in these respects by a fosmid library of the *C. elegans* genome, which has major advantages of stability, ease of manipulation and good (close to 100 %) coverage.

2.3 Genomics

2.3.1 General Organization

The genome of *C. elegans* is organized into six nuclear chromosomes, ranging in size from 16 to 20 Mb, plus the small 15 kb mitochondrial genome. In contrast to the genome sequences so far generated for all other multicellular organisms, that of *C. elegans* has been fully determined, telomere to telomere, for each chromosome, allowing an exact statement for the wild-type haploid nuclear genome size: 100,281,426 base pairs.

Some qualifications should be attached to this statement. First, annotation and re-sequencing continue to provide minor corrections to the reference wild-type sequence, at a low rate (less than one correction per month, and usually involving fewer than five nucleotides). Second, some of the sequencing was carried out using DNA from the standard laboratory strain, Bristol N2, and some from an N2-derived nuclease-deficient strain, which may carry additional mutations (but probably fewer than 1/100 kb). Third, individual animals in a wild-type population will all be homozygous at almost all loci, but there will still be some level of variability between individuals, particularly in the copy number of repeats in regions of tandemly repeated sequence, such as the ribosomal RNA gene clusters. Fourth, comparisons between different natural isolates of *C. elegans* carried out by whole genome hybridization suggest that these may differ by substantial (>100 kb) deletions and insertions (Maydan et al. 2007), so the Bristol N2 strain has retained a significant number of genes that are missing in some natural races. Conversely, Bristol N2 has almost certainly lost some genes that are present in other races of this species. Thus,

Table 2.1 Genome organization and gene distribution

Zone	Size (Mb)	Protein genes	Coding %	tRNA genes
LGI	15.07	3,470	26.54	66
L	3.68	622	19.12	6
C	7.13	1,936	31.90	32
R	4.26	912	23.98	28
LGII	15.28	4,090	27.95	56
L	5.17	1,507	28.22	20
C	6.64	1,893	30.98	25
R	3.47	690	21.73	11
LGIII	13.78	3,265	26.40	64
L	3.84	780	20.69	13
C	6.49	1,841	32.34	37
R	3.45	644	21.59	14
LGIV	17.49	3,871	23.13	70
L	4.37	852	19.54	18
C	8.51	1,066	28.60	10
R	4.61	1,953	16.42	42
LGV	20.91	5,570	27.75	78
L	5.67	1,485	26.50	10
C	9.89	1,182	30.44	17
R	5.35	2,903	24.10	51
LGX	17.71	3,578	20.41	274
L	4.90	871	19.32	38
C	7.20	1,400	22.05	125
R	5.61	1,307	19.28	111

Numbers in this table differ somewhat from those originally provided by the *C. elegans* Sequencing Consortium (1998), because of improved gene predictions and revised assessment of boundaries between arm (*L* left, *R* right) and cluster (C) regions. We thank Gary Williams for assistance in preparing this table

"the wild-type sequence" is a slightly idealized concept, even in the case of *C. elegans*, but it is still extremely useful as a completely defined standard for reference.

Molecular sizes and distinctive features for the six chromosomes are provided in Table 2.1, which is an updated version of the equivalent table first assembled by the *C. elegans* Sequencing Consortium (1998). The three chromosomal zones of each autosome, apparent on the genetic map, are also discernible at a molecular level. The central, cluster regions contain a higher density of genes, and there is also a roughly fivefold lower rate of meiotic recombination in these clusters. The arm regions contain a lower density of genes, with (on average) larger introns, and higher recombination frequencies. Consequently, Marey maps (which plot genetic distance versus molecular distance along a chromosome) exhibit a strongly sigmoid shape for the five autosomes and a weaker sigmoid for the X chromosome (Fig. 2.3).

These long-range features of the genome were evident in the first descriptions of the whole genome (*C. elegans* Sequencing Consortium 1998). Functional and evolutionary analyses, carried out on a global scale, revealed additional properties that differentiate between the arms and clusters. There is a higher density of essential genes in the clusters than on the arms. Conversely, the arms contain relatively more genes belonging to large gene families, and the arm genes appear to be evolving more rapidly, particularly in terms of recent gene duplications and deletions. In addition, some general functional differences between genes on the autosomes and genes on the X chromosome have been detected.

Convincing evolutionary or functional explanations of these striking global patterns have yet

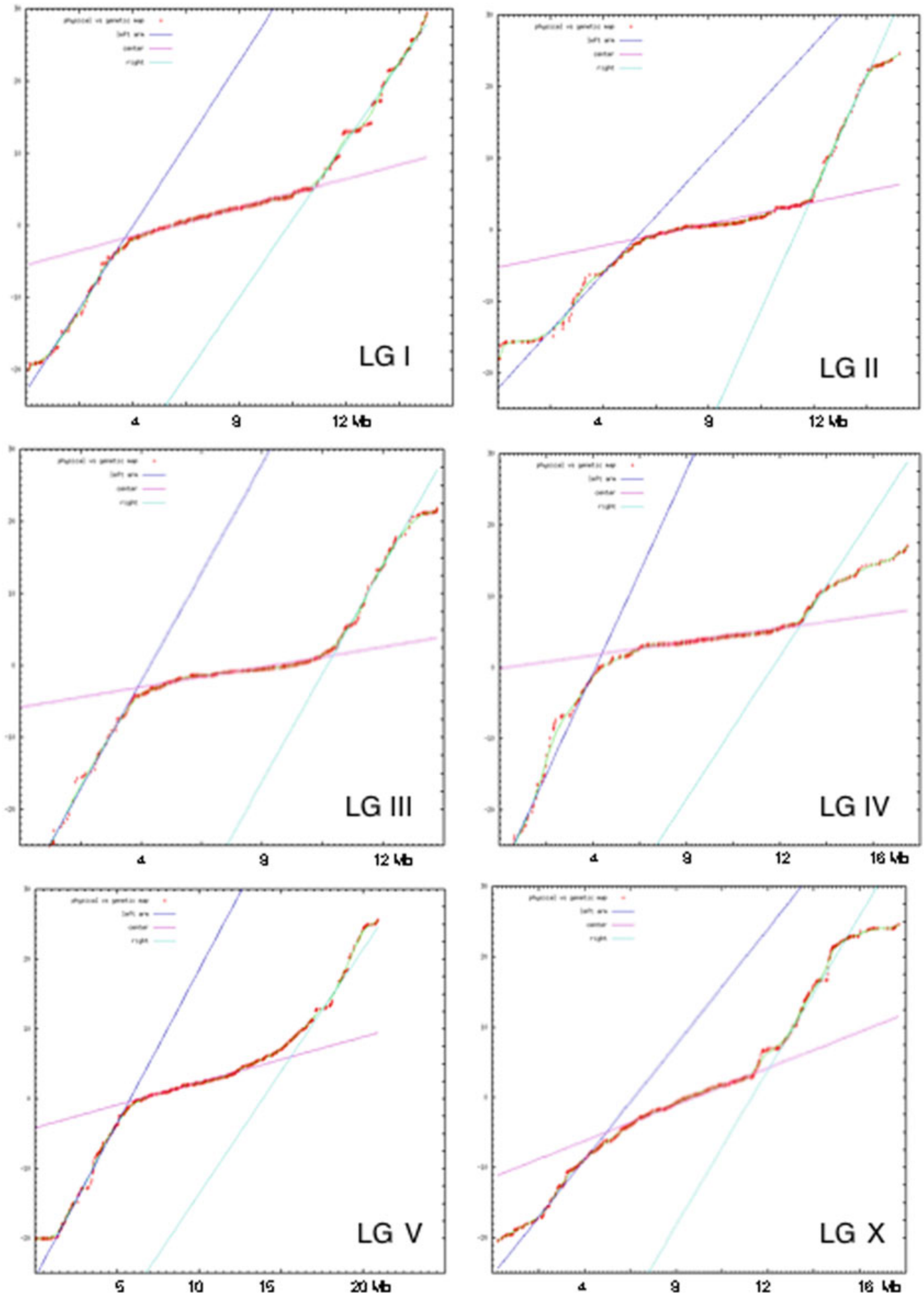

Fig. 2.3 Marey maps plotting genetic map position (centiMorgan scale) against sequence coordinates (Megabase scale), for all recombinationally mapped and cloned genes on each of the six chromosomes of *C. elegans*. The genetic maps of each chromosome are organized around an arbitrarily defined zero point, with negative

to be proposed, but the overall chromosomal organization appears to have long-term stability, because the genome of *Caenorhabditis briggsae* has similar features, and there is a high degree of synteny between *C. elegans* and *C. briggsae*, despite their divergence more than 50 million years ago (Hillier et al. 2007)

The chromosomes of *C. elegans*, like those of other nematode species, are holocentric: that is, they do not have a single mitotic centromere. Instead, at mitosis spindle microtubules appear to attach all along the chromosomes, and thereby mediate segregation. Consistent with this cytologically observable absence of single centromeres, there are no obvious centromeric regions in the genome sequence. It is possible that one or more of the various families of repeated sequence that are widely distributed across the worm genome may act as attachment sites for spindle microtubules, but as yet there is no evidence for this. Furthermore, foreign DNA injected into the germline is able to form extrachromosomal arrays that behave as pseudochromosomes and are transmitted through mitosis with reasonable fidelity, suggesting that there is no sequence specificity in the attachment of mitotic spindle microtubules.

Holocentric chromosomes require some kind of special mechanism in order to allow segregation at the first meiotic division, because otherwise recombination would result in crossover chromosomes being pulled towards both poles at once, leading to chromosome breakage or loss. The problem is solved by chromosome ends acting as centromeres at meiosis I, with either end being usable for this purpose. Pairing of chromosomes at meiosis appears to be initiated by a dedicated pairing site close to one end or the other, and candidate sequences have been identified that may assist in the initial pairing (Sanford and Perry 2001). These are six short (11–16 nt) sequences, each of which is greatly enriched on one chromosome and usually in a cluster of >50 tandem repeats, located in a position consistent with the genetically predicted pairing site.

Telomeres are similar to those of most eukaryotes, consisting of long repeats of a hexanucleotide sequence (TTAGGC, similar to the vertebrate TTAGGG), but there appear to be no specialized subtelomeric regions.

2.3.2 Protein Coding Genes

The WS228 release of WormBase lists 20,389 protein coding genes, about 15 % of which are known to generate more than one mRNA as a result of alternative splicing, to give 23,645 predicted proteins. 35 % of these are fully confirmed at the RNA level, as a result of experimental evidence such as expressed sequence tag (EST) clones. Forty six percent are partly confirmed by these criteria, and 19 % currently lack transcriptional evidence and are based on informatic criteria alone. Some of the genes in the last class may therefore be pseudogenes, discussed further below.

The predicted proteins of *C. elegans* range in size from small peptides (ca. 30 aa) to giant molecules such as the mesocentin DIG-1, with 13,100 aa, for which the gene extends across 60 kb of genomic sequence. Codon usage in *C. elegans* shows some characteristic biases, mostly consistent with the overall GC content of the genome (36 %). Introns are found in most protein coding genes, but are usually small (the commonest size is 47 bases) and there are few of the very large introns found in vertebrate genomes. 5′ and 3′ untranslated regions are

Fig. 2.3 (Continued) coordinates for genes on the *left arm* and positive coordinates for genes on the *right arm* (*Y* axis). The *X* axis plots the sequence from coordinate zero, which is defined as the left end of each chromosome. *Oblique lines* indicate tangents to the main chromosomal regions (left arm, cluster, and right arm). Previous plots of this type (Barnes et al. 1995) were based on the physical map rather than the complete genome sequence and showed less detail, such as the tendency for gene clustering at the ends of chromosomes as well as in the main central clusters. The existence of a central cluster on the X chromosome is also more evident than in earlier maps, as are the demarcations between arm and cluster regions. The regional boundaries used in Table 2.1 were based on this figure generally short, as are regulatory regions, usually not more than 2 kb. This compactness may be a consequence of the generally compressed state of the genome, because the average intergenic distance is also small.

2.3.3 Trans-splicing and Operons

An unusual feature of *C. elegans* is that about 55 % of its genes are trans-spliced to a short leader sequence called SL1, which is transcribed from a repeated set of SL1 genes located at one site in the genome (reviewed by Blumenthal 2005). It is not clear what functional difference there may be between mRNAs with and without SL1; possibly the leader sequence results in more efficient translation. The presence of SL1 means that it is difficult to define the transcriptional start site for the genes concerned, but the 5′ sequences lost by trans-splicing (referred to as "out-tron") are not long, in those cases that have been analyzed in detail.

A further and even more unusual feature of the worm's genome is that a significant fraction of its genes (15 %) are organized into operons (Blumenthal et al. 2002; Blumenthal 2005). At these loci, between two and eight distinct protein coding regions are situated close together and are transcribed from a single promoter. The long primary transcript is then broken up into separate molecules by trans-splicing to a different short leader sequence called SL2, or to an SL2-related leader. The SL2 leader and related leader sequences, which all appear to be functionally equivalent, are transcribed from 18 loci scattered around the genome.

The proteins encoded by any particular operon sometimes share functional properties, copying the pattern seen in bacterial operons, but for many operons this is not the case, and for these the operonic organization looks more like an accident of genomic proximity. Some operons may also contain internal promoters, so presumably under some circumstances the genes within such an operon can behave as conventional genes and do not depend on the single operon promoter.

2.3.4 Pseudogenes

The fact that 19 % of predicted coding sequences are currently unconfirmed by any transcriptional support raises the possibility that these are computational artifacts or pseudogenes. Further evidence of possible pseudogene status for some of the predicted genes comes from systematic screens for patterned expression of transgenes driven by predicted promoter regions. A significant fraction of such tests result in no detectable expression, and examination of these apparently silent genes suggests that some are indeed nonfunctional, frequently containing stop codons in the most probable set of exons (Mounsey et al. 2002). However, it may be that these are simply transcribed at low level, or in very restricted tissues or time windows, and the in-frame stop codons could be avoided by unusual RNA splicing or editing. Moreover, even if they are genuinely non-functional pseudogenes in the Bristol race of *C. elegans*, they may have retained functionality in other races of the species. Clear examples of this effect have been found in various gene families (Stewart et al. 2006). Comparative genomics can be expected to provide powerful evidence on this question: the current availability of a complete sequence for *C. briggsae*, and drafts for *C. japonica*, *C. remanei,* and *C. brenneri* (further discussed below) means that many candidate pseudogenes can be examined for features that will test their functionality.

Processed pseudogenes, which have apparently arisen by reverse transcription and reintegration of a mature mRNA sequence, and therefore lack introns, are much rarer in the *C. elegans* genome than in vertebrate genomes.

The large number of protein coding genes, which was surprising when the worm genome sequence was first established, now looks less anomalous. Various other invertebrates, even apparently simple animals such as the sea anemone *Nematostella vectensis* (Putnam et al. 2007), also have large gene numbers. Also, the amount of alternative splicing in the nematode transcriptome seems to be lower than that in vertebrates, so the total proteomic complexity of *C. elegans* is likely to be much lower than the proteomic complexity of vertebrates, in line with its greater cellular and developmental simplicity.

2.3.5 Major Protein Coding Gene Families

In common with all other higher eukaryotes, certain taxon-specific gene families have been conspicuously expanded at some point in the evolutionary past of this species. For *C. elegans*, one example of such expansion is provided by genes encoding G-protein coupled receptors, of which there are more than a thousand. Most of these are probably chemosensory receptors of one kind or another, because chemoreception provides the major sensory modality for this organism. It has only a rudimentary light sense, but gets most of its information about the environment through a surprisingly sophisticated olfactory sense. Other large families are kinase genes and two classes of transcription factor genes, those encoding NHR (nuclear hormone receptor) proteins and those encoding zinc-finger factors. Over 150 collagen genes can be identified, most of which are involved in construction of the collagenous exoskeleton of the worm. Genes encoding proteins with a C-type lectin motif (clec genes) are also numerous (>250) and may contribute to innate immunity in this organism (O'Rourke et al. 2006). For more extensive review of major protein families, see Schwarz (2005).

2.3.6 RNA Genes: Structural, Translational, and Trafficking

C. elegans has a full complement of the standard translational RNAs. The 18S, 5.8S, and 28S ribosomal RNA genes are arranged in a set of 55 tandem copies on one end of chromosome I, which therefore behaves as the nucleolus organizer. The 5S ribosomal genes are encoded at a different locus on chromosome V, containing about 110 tandem copies of the 5S sequence alternating with the SL1 trans-spliced leader sequence.

tRNA genes are distributed across the genome, with a conspicuous concentration on the X chromosome. A convenient catalogue of the tRNA complement can be found at: http://lowelab.ucsc.edu/GtRNAdb/Celeg/

About 25 % of the 820 tRNA genes appear to be pseudogenes. Of the functional set listed in Table 2.1, most families contain between 4 and 20 members, with abundances approximately matching codon usage in this organism. There is a single selenocysteine tRNA gene.

Genes encoding trans-spliced leaders are located either in the 5S/SL1 cluster, or at dispersed sites for SL2 and related leaders.

The usual sets of snRNAs, scRNA, and other small functional RNA genes are present and have been identified, though some, such as the snoRNAs, are still hard to annotate completely or reliably. The telomerase RNA has not yet been identified, because such RNAs are difficult to recognize and considerably divergent in evolution. At least one abundant but enigmatic RNA species with telomere similarity, the *tts-1* transcript, has been identified as a result of serial analysis of gene expression (SAGE) analysis (Jones et al. 2001).

2.3.7 Small RNA Genes: Regulatory MicroRNAs and Other Species

MicroRNAs, now believed to play major regulatory roles in most multicellular organisms, were first discovered in *C. elegans*, as a result of analysis of the developmental mutants *lin-4* and *let-7*, which when cloned proved to encode small RNAs rather than proteins This finding provided the stimulus for the discovery of substantial miRNA families in other organisms, and also to detailed characterization of the miRNA complement in *C. elegans* itself (reviewed by Vella and Slack 2005). Currently, over 130 *mir*-genes have been recognized in the worm genome, but it is likely that more remain to be discovered.

A second large class of small noncoding RNA genes is the 21U-RNAs, or dasRNAs (diverse autonomously expressed small RNAs), which are all exactly 21 nucleotides long, begin with 5′ UMP, and share an upstream sequence motif (Ruby et al. 2006). They are located primarily in two extended regions on LGIV, which contain thousands of such genes. They appear to be conserved in related nematode species, but their function is currently uncertain.

Many endogenous siRNA species can also be detected in *C. elegans*, but these are produced by the action of RNA-dependent RNA polymerases on the transcripts from protein coding genes and transposons, so they are not encoded by distinct genes.

2.3.8 Transposons

Transposon biology in *C. elegans* has been reviewed in detail by Bessereau (2006). About 12 % of the Bristol genome is taken up with transposons or transposon-derived sequence. Transposons that are currently capable of mobilization fall into eight identified families. Most members of each family are present in 10–50 copies, dispersed through the genome. Tc1, which is one of the founding members of the Tc1/mariner transposon family, has been studied in the most detail. Transposons are quiescent in the genomes of most natural races of *C. elegans*, but can be activated in various mutant backgrounds and they are also actively transposing in the Bergerac race of *C. elegans* and its derivatives. As a result, the Tc1 copy number in the Bergerac race has increased from the normal 30-odd copies to several hundred copies, with some concomitant deleterious effects on the viability and fertility of the worms in this race. The transposon-active strains have been historically useful in providing many polymorphic molecular markers, and in allowing transposon-tagging and cloning of important genes.

2.3.9 Repetitive Sequences

Some 7 % of the genome is taken up with repetitive sequences, belonging to approximately 50 different repeat families. As noted above, some of these are non-randomly distributed in the genome, being either concentrated or depleted in identifiable genomic regions such as the chromosome arms, the central autosomal clusters, the X chromosome, or the meiotic pairing regions.

2.4 Post-genomic Analysis

2.4.1 Continued Annotation

Post-genomic approaches to analyze the expression, function, organization, and evolution of the *C. elegans* genome can be considered briefly, under ten headings. The first of these is a continuing process of annotation of the reference genomic sequence: information from many sources continues to improve gene predictions, identify previously unpredicted genes (especially those producing noncoding RNAs), define new gene families, and reveal candidate transcription factor binding sites and other regulatory regions.

2.4.2 Resequencing

The reference genome sequence is that for the Bristol race of *C. elegans*, which is the standard laboratory strain. Some re-sequencing of the Bristol genome has been carried out, in order to detect any residual errors in the sequence. In addition, many other natural isolates of the species have been obtained, most recently from Africa (Dolgin et al. 2007), and extensive sequencing has been carried out on some of these races. This is in order both to obtain single nucleotide polymorphism (SNP) markers, which are essential for fine-structure mapping and positional cloning, and to examine natural variability in *C. elegans.* A Hawaiian race (strain CB4856) has been examined in most detail, because it appears to be among the most divergent of *C. elegans* races, as compared to the Bristol strain. Significant behavioral and biological differences between the Bristol and Hawaii strains have been studied.

2.4.3 Transcriptome

Numerous cDNA libraries have been generated for *C. elegans* and many thousands of expressed sequence tags (ESTs) have been defined. In addition, a number of SAGE libraries have been generated (Jones et al. 2001).

Initial cDNA collection was limited to whole animal samples, but the ability to sort embryonic cells of particular tissue types or neuronal classes means that transcriptional profiling of specific cell types has become possible (Zhang et al. 2002).

2.4.4 Microarray Analysis

Extensive microarray analysis of *C. elegans* has been carried out, using initially spotted cDNA arrays and more recently oligonucleotide arrays aimed at covering most of the predicted exons in the genome. Many different conditions and mutant backgrounds have been explored. A meta-analysis of early microarray experiments was carried out by Kim et al. (2001), which allowed visualization of correlated gene expression profiles as a three-dimensional "topomap."

2.4.5 Expression Analysis: Spatial and Temporal Patterns

Systematic analyses of expression patterns have been carried out by in situ hybridization (Motohashi et al. 2006) and by constructing transgenic animals expressing β-galactosidase or GFP driven by particular gene promoters. The transparency of the animal means that fluorescent reporters such as GFP and related proteins can be visualized readily in any cell of the animal, throughout development. The complete anatomical description means that all cells can be reliably identified, allowing exact description of anatomical expression profiles (Hunt-Newbury et al. 2007)

High-throughput automated description of temporal expression profiles has also become possible, by coupling a flow-cytometer adapted for nematode profiling together with detection of fluorescent transgenes (Dupuy et al. 2007). Worms are automatically sorted on the basis of length, which corresponds to developmental stage, and fluorescence is recorded along the one dimension of the body axis. This allows generation of "chronograms" which display gene expression patterns in time as well as space.

2.4.6 Functional Analysis: Gene Deletions

Efficient homologous recombination is not currently feasible in *C. elegans*, but a variety of effective methods for isolating gene deletion mutants in genes of interest have been developed (Barstead and Moerman 2006). As a result, putative knockout mutations are now available for thousands of identified genes, and the prospect of achieving complete coverage for all predicted genes seems real. Methods aimed at efficient gene replacement have recently been developed in the worm. One of these uses a Mos transposon from *Drosophila melanogaster* to make targeted gene deletions (Frokjaer-Jensen et al. 2010). The NemaGENETAG consortium (http://elegans.gr/nemagenetag/) has generated a resource of Mos insertions in 14,000 known sites distributed throughout the *C. elegans* genome.

2.4.7 Functional Analysis: RNAi Knockdowns

A distinct method of reducing or blocking gene function was discovered for *C. elegans* in the form of RNAi (Fire et al. 1998), which subsequently proved to be widely applicable to most eukaryotic organisms. *C. elegans* is particularly amenable to RNAi experiments, because it has the capacity to take up double-stranded RNA from the environment and even from the bacteria on which it is fed, which then results in RNAi knockdown of any corresponding endogenous gene (Timmons et al. 2001). Consequently, "feeding libraries" containing many thousands of *E. coli* strains, each expressing a different *C. elegans* dsRNA, have been constructed and used to carry out whole genome screens, efficiently and economically. The first such surveys (Kamath et al. 2002) allowed preliminary assignment of function to about 23 % of genes. The initial RNAi tests on the remaining 77 % revealed no obvious function, however, for a variety of possible reasons, such as subtle or redundant activities, or incomplete knockdown by RNAi. More recent whole-genome screens have used sensitized genetic backgrounds, or have concentrated on particular aspects of the phenotype, and both of these approaches are steadily increasing the number of genes for which some kind of biological function can be identified by means of RNAi.

RNAi has both advantages and disadvantages as compared to stable gene deletion knockouts. Advantages include the extreme convenience of the feeding technique, and the ability to apply it at different times in development. Moreover, since the process acts at the RNA level, it can eliminate both maternal and zygotic contributions to gene expression, which can be important when studying events in early embryogenesis. Disadvantages include variability in effect, incomplete knockdowns (because it is hard to eliminate 100 % of gene activity by RNAi), genes refractory to RNAi, and off-target effects.

2.4.8 Interactome and Gene Networks

Large-scale high-throughput explorations of protein–protein interaction in *C. elegans* have been executed using the yeast 2-hybrid technique (Li et al. 2004). While fallible, this technique is a powerful discovery tool for identifying possible interacting partners for any given protein, which can then be assessed on the basis of other data and subjected to experimental tests. Combination of interactome data together with information about gene expression and function is leading to increasingly sophisticated network biology for *C. elegans* (Piano et al. 2006; Zhong and Sternberg 2006; Lee et al. 2008).

2.4.9 Proteomics and Structural Genomics

Proteomic investigations of *C. elegans* are less advanced than those of mammalian cells, but becoming increasingly effective and important, especially for the characterization of multiprotein complexes such as sperm chromatin (Chu et al. 2006). Mass-spectrometric analyses can also be expected to reveal the full repertoire of posttranslational modification of *C. elegans* proteins.

A different aspect of the proteome is protein structure, acquired either by X-ray crystallography or NMR. Programs have been set up with the goal of acquiring three-dimensional structures for many *C. elegans* proteins, on a high-throughput, genome-driven basis (Luan et al. 2004).

2.4.10 Comparative Genomics

Last but not least of the methods that can be applied to understand the genome of *C. elegans* is comparative genomics, making use of the increasing amount of genomic information for related nematode species and other eukaryotes. In particular, other species from the genus *Caenorhabditis* itself provide invaluable resources for investigating *C. elegans*. There are more than ten known species of Caenorhabditis currently available as laboratory strains. Most of these are conventional gonochoristic species, with female and male sexes, but one of them, *C. briggsae*, is a species like *C. elegans* with hermaphrodite and male sexes. Despite its extreme morphological and biological similarity to *C. elegans*, it appears to have diverged from *C. elegans* at least 50 million years ago, and has evolved a hermaphrodite sex independently (Fig. 2.4). A nearly complete genome sequence has been generated for *C. briggsae*, which has revealed both extensive conservation and extensive divergence (Stein et al. 2003). Remarkably, it appears that synteny between these two species is extreme: although there have been some rearrangements (mostly inversions) within chromosomes, there have been very few exchanges of material between chromosomes (Hillier et al. 2007)

Extensive genomic data are also available for three other species, *C. remanei*, *C. brenneri,* and *C. japonica* with others likely to be sequenced in future (see http://genome.ucsc.edu/cgi-bin/hgGateway). Sequence data for *C. remanei* already reveal a striking genomic difference from the

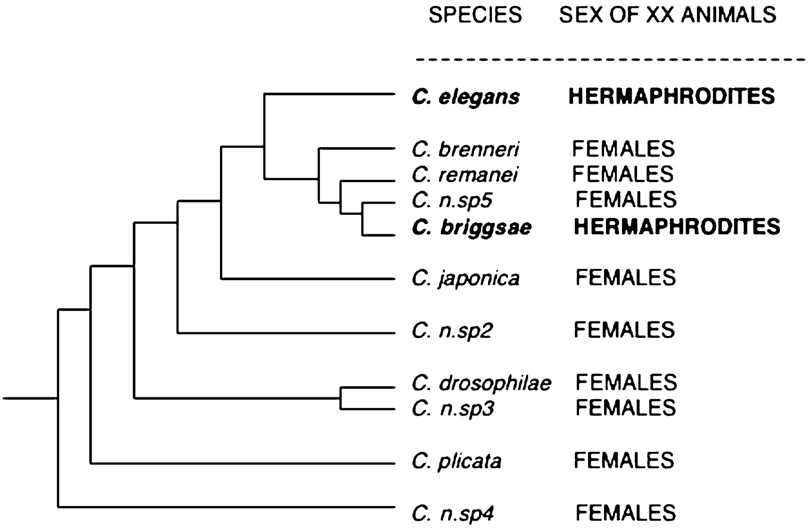

Fig. 2.4 Phylogeny for the genus *Caenorhabditis*, modified from Kiontke and Fitch (2005). Most species in the genus have conventional female and male sexes and are assumed to have XX and XO karyotype, respectively. *C. elegans* and *C. briggsae* have XO male and XX hermaphrodite sexes

two hermaphroditic species, which is that its genome is significantly larger (ca. 150 Mb).

2.5 Conclusion

Thirteen years of exploration and exploitation of the *C. elegans* genome have opened up many new areas for research on this organism. Knowledge about the worm at all levels, from the nucleotide to the whole global population, continues to accumulate and to become ever more accessible and amenable to sophisticated analysis. Integrating different kinds of biological information, and the availability of well-curated and near-exhaustive datasets, can be expected to lead to new kinds of experimental investigation, as well as to new levels of understanding and ultimately to realistic in silico modeling and simulation.

References

Barnes TM, Kohara Y, Coulson A, Hekimi S (1995) Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics 141:159–179

Barstead RJ, Moerman DG (2006) *C. elegans* deletion mutant screening. Methods Mol Biol 351:51–58

Bessereau JL (2006) Transposons in *C. elegans*. In: WormBook. http://www.wormbook.org

Blumenthal T (2005) Trans-splicing and operons. In: WormBook. http://www.wormbook.org

Blumenthal T et al (2002) A global analysis *of Caenorhabditis elegans* operons. Nature 417:851–854

Brenner S (1974) The genetics of *Caenorhabditis elegans*. Genetics 77:71–94

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

Chu DS, Liu H, Nix P, Wu TF, Ralston EJ, Yates JR 3rd, Meyer BJ (2006) Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. Nature 443:101–115

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

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

Dolgin ES, Félix MA, Cutter AD (2007) Hakuna Nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C. briggsae*. Heredity 100:304–315

Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A, Carvunis AR, Pulak R, Shingles J, Reece-Hoyes J, Hunt-Newbury R, Viveiros R, Mohler WA, Tasan M, Roth FP, Le Peuch C, Hope IA, Johnsen R, Moerman DG, Barabási AL, Baillie D, Vidal M (2007) Genome-scale analysis of in vivo spatiotemporal promoter activity in *Caenorhabditis elegans*. Nat Biotechnol 25:663–668

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391:806–811

Frokjaer-Jensen C, Davis MW, Hollopeter G, Taylor J, Harris TW, Nix P, Lofgren R, Prestgard-Duke M, Bastiani M, Moerman DG, Jorgensen EM (2010) Targeted gene deletions in *C. elegans* using transposon excision. Nat Methods 7:451–453

Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA, Koboldt DC, Waterston RH (2007) Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. PLoS Biol 5:e167

Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D, Fang L, Halfnight E, Lee D, Lin J, Lorch A, McKay S, Okada HM, Pan J, Schulz AK, Tu D, Wong K, Zhao Z, Alexeyenko A, Burglin T, Sonnhammer E, Schnabel R, Jones SJ, Marra MA, Baillie DL, Moerman DG (2007) High-throughput in vivo analysis of gene expression in *Caenorhabditis elegans*. PLoS Biol 5:e237

Jones SJM, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA (2001) Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. Genome Res 11:1346–1352

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2002) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421:231–237

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, Fitch DH (2005) The phylogenetic relationships of Caenorhabditis and other rhabditids. In: WormBook. http://www.wormbook.org

Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. Nat Genet 40:181–188

Li S et al (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303:540–543

Luan CH, Qiu S, Finley JB, Carson M, Gray RJ, Huang W, Johnson D, Tsao J, Reboul J, Vaglio P, Hill DE, Vidal M, DeLucas LJ, Luo M (2004) High-throughput expression of *C. elegans* proteins. Genome Res 14:2102–2110

Maydan JS, Flibotte S, Edgley ML, Lau J, Selzer RR, Richmond TA, Pofahl NJ, Thomas JH, Moerman DG (2007) Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization. Genome Res 17:337–347

Motohashi T, Tabara H, Kohara Y (2006) Protocols for large scale in situ hybridization on *C. elegans* larvae. In: WormBook. http://www.wormbook.org

Mounsey A, Bauer P, Hope IA (2002) Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Genome Res 12:770–775

O'Rourke D, Baban D, Demidova M, Mott R, Hodgkin J (2006) Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of *C. elegans* with *M. nematophilum*. Genome Res 16:1005–1016

Piano F, Gunsalus KC, Hill DE, Vidal M (2006) *C. elegans* network biology: a beginning. In: WormBook. http://www.wormbook.org

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell 127:1193–1207

Sanford C, Perry MD (2001) Asymmetrically distributed oligonucleotide repeats in the *Caenorhabditis elegans* genome sequence that map to regions important for meiotic chromosome segregation. Nucleic Acids Res 29:2920–29266

Schwarz EM (2005) Genomic classification of protein-coding gene families. In: WormBook. http://www.wormbook.org

Stein LD et al (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol 1:166–192

Stewart MK, Clark NL, Merrihew G, Galloway EM, Thomas JH (2006) High genetic diversity in the chemoreceptor superfamily of *Caenorhabditis elegans*. Genetics 169:1985–1996

Timmons L, Court DL, Fire A (2001) Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. Gene 263:103–112

Vella MC, Slack FJ (2005) *C. elegans* microRNAs. In: WormBook. http://www.wormbook.org

Zhang Y, Ma C, Delohery T, Nasipak B, Foat BC, Bounoutas A, Bussemaker HJ, Kim SK, Chalfie M (2002) Identification of genes expressed in *C. elegans* touch receptor neurons. Nature 418:331–335

Zhong W, Sternberg PW (2006) Genome-wide prediction of *C. elegans* genetic interactions. Science 311:1481–1484