REVIEW

Around the Genomes

The *C. elegans* Genome Sequencing Project

Mary Berks[1,3] and the *C. elegans* Genome Mapping and Sequencing Consortium[1,2]

[1]The Sanger Centre, Hinxton, Cambridgeshire, CB10 1RQ, UK; [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri USA

Caenorhabditis elegans, a free-living nematode worm, has proved a particularly useful model organism for studying the anatomy, behavior, genetics, and development of a metazoan. It also has one of the smallest genomes of the higher eukaryotes (100 Mb distributed over six chromosomes), making it an ideal candidate for detailed molecular analysis. The *C. elegans* genome project began over 10 years ago and is a collaborative effort between two laboratories (St. Louis, MO, USA, and Cambridge, UK), with the ultimate aim of mapping and sequencing the whole of the 100-Mb genome. The consortium has now completed the sequence of approximately one-fifth of the genome and plans to have sequenced more than half the genome before the end of next year.

More Than 95% of the Genome Covered by Cosmids and YACs

The rapid progress of the sequencing project has been facilitated by the quality and extent of the physical map (Coulson et al. 1986, 1988, 1991). More than 95% of the genome is covered by a combination of 17,500 cosmids and 3,500 YACs with only a few gaps remaining. Mapping of cDNAs (Waterston et al. 1992; Y. Kohara, unpubl.) has indicated that more than 99% of transcripts fall within mapped regions. The mapped cosmids, which represent ~80% of the genome, are the primary resource for the sequencing project: The remaining 20% of the genome, which could not be cloned in multicopy bacterial vectors, is represented by YAC "bridges" that will be sequenced toward the end of the project.

Sequencing Concentrated on Gene-rich Regions

Since the beginning of the project, there has been a close collaboration and exchange of information and resources between the two main mapping and sequencing laboratories and the whole of the *C. elegans* research community. A large number of genetically defined loci are now linked to the physical map and this linkage between the physical and genetic maps is continually improving. Examination of the distribution of known genetic loci on the physical map, or of the distibution of mapped cDNAs, suggests that there is a physical clustering of genes toward the centers of the five autosomes, whereas genes appear more evenly distributed throughout the X chromosome (Fig. 1). This is independent of the extreme clustering of loci in the autosomes on the genetic map, caused by variation in rates of recombination (Greenwald et al. 1987). Sequencing efforts are being concentrated on these relatively gene-rich regions, which together represent a total of ~60% of the genome. Fortunately, these are also the regions best covered by cosmids. Already sequenced are most of the central gene-rich regions of chromosomes II, III, and IV and almost the whole of the X chromosome, with only the gene-rich regions of chromosomes I and V and the gene-sparse regions of the chromosome arms remaining.

Shotgun Sequencing Strategy

Cosmids are chosen for sequencing so that, on average, they overlap by ~5–10 kb. A random shotgun approach has been used, initially utilizing fluorescently labeled dye-primer chemistry to tag sequence a large number of M13 subclones (with an average insert size of 1.3–2 kb) from each cosmid. This is followed by a more directed "finishing" stage in which extra-long dye-primer sequences or reverse-primer sequences from the amplified M13 insert are generated to obtain a contiguous piece of DNA. Any ambiguities are resolved using fluorescently labeled dye-terminator

[3]E-MAIL mb1@sanger.ac.uk; FAX 44-1223-494919.

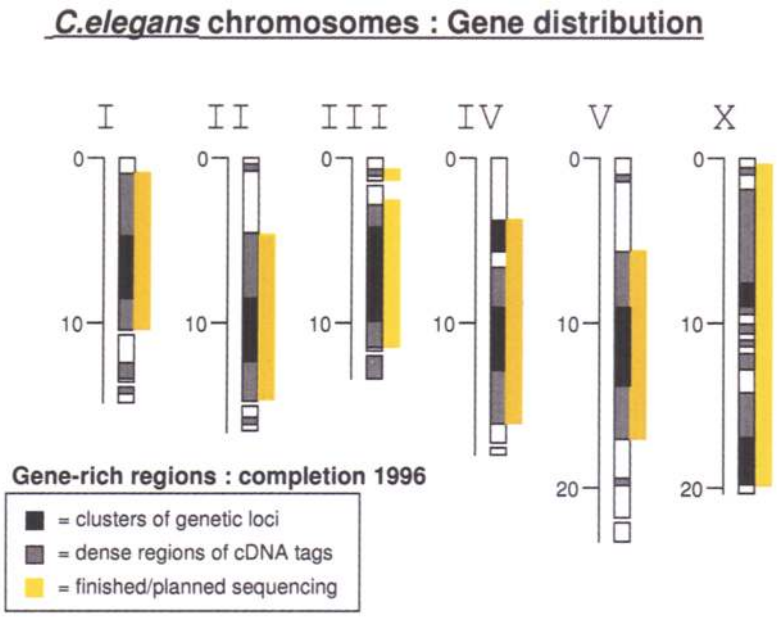

Figure 1 The number and approximate size of *C. elegans* chromosomes is shown schematically. Clusters of genetic loci and dense regions of mapped cDNA tags are superimposed on the chromosomes to indicate the particularly gene-rich regions (representing ~60% of the genome), which we plan to sequence first.

chemistry, and dye-terminator chemistry is also applied to those regions where sequence on both strands is not available. Analyzed data from ABI automated DNA sequencers are screened for contaminating *Escherichia coli*, for sequencing vectors, and for cosmid cloning vectors before assembly into a cosmid project data base. Recently, the assembly algorithm PHRAP (P. Green, unpubl.) has been used, which uses the full length of the sequenced reads to perform the initial assembly. This is then converted to Rodger Staden's XGAP (Staden 1994) format, which is used as an interactive tool to aid sequence editing and "finishing." The final coverage of reads across the cosmid is around fivefold, with an estimated accuracy of ~99.99% (Sulston et al. 1992).

Problems

One problem with this strategy is that there are certain DNA structures that do not subclone efficiently into single-stranded M13 sequencing vectors. In many cases this is due to inverted repeat sequences (see below), with the result that certain regions of the cosmid may not be covered by any M13 subclones. It has been found that screening small-insert (2- to 3-kb) double-stranded phagemids for particular clones or tag-sequencing a number (96) of single-stranded, large-insert (6- to 9-kb) phagemids usually yields suitable templates that span the gap. Alternatively, fragments for sequencing across the gap can be generated from cosmid DNA by PCR.

Costs and Efficiency

The consortium is continually seeking to streamline existing procedures by testing and implementing new developments, both in the biochemistry and in the assembly and finishing software. As a result we have experienced considerable improvements in efficiency over the past few years: Throughput and overall finishing rates have increased while costs have steadily decreased. (Direct costs per completed base, currently ~45 cents, should fall to nearer 20 cents per base over the next year). In particular, this year has seen the development of programs that allow a degree of automatic editing and that choose suitable templates for long reads and reverse reads. This reduces skilled user interaction and helps to increase the output of completed sequence.

Analyzing the Data

Completed sequence data are subjected to a number of routine analysis procedures. The program GENEFINDER (P. Green and L. Hillier, unpubl.) is used to look for particular coding features such as open reading frames, codon biases, and potential donor and acceptor splice sites, and uses this information to make gene predictions. A search is made against the public data bases to find BLAST similarities [BLASTX (Altschul et al. 1990) for protein and BLASTN for nucleotide similarities], particular protein motifs (PROSITE), tRNA genes, and EST matches. The cosmid sequence is analyzed for repeats within itself and for matches to known *C. elegans* repeat families. In addition, new sequence is compared to all other *C. elegans* sequences to highlight long-range repeats indicative of gene families. Completely finished data are stored within the *C. elegans* data base ACEDB (R. Durbin and J. Thierry-Mieg, unpubl.; ftp://ncbi.nih.gov/repository/acedb/), where they are

annotated and then submitted to the public data bases (EMBL or GenBank). In addition, raw data from sequenced and assembled, but unfinished, cosmids are also made available immediately via anonymous FTP (ftp.sanger.ac.uk; http://www.sanger.ac.uk). This availability ensures that experts in particular areas, studying the function either of individual *C. elegans* genes or of protein families in other organisms, have free and instant access to potentially useful sequence data, and has led to numerous useful collaborations and interactions.

Around 30% of Predicted Genes Found So Far

A summary of the analysis of data from > 20 Mb of completed sequence is shown in Table 1. There is now ~7 Mb completed on each of chromosomes III, II, and X, and 0.1 Mb completed on IV, with a total of just under 4000 predicted genes. Using these figures to calculate gene densities suggests an average gene density on the autosomes of one gene every 4.8 kb over the gene-rich regions and on the X chromosome, of one gene every 6.6 kb. Approximately 45% of these predicted genes show some match to data in the public data bases. Several of the remaining novel genes appear to be related to each other and presumably define new gene families, many of which may be nematode-specific.

Table 1. Summary analysis of data from ~20 Mb of completed *C. elegans* sequence

C. elegans genome: (5 autosomes/X chromosome)	100 Mb
Finished sequence: (Aug '95)	21.14 Mb
(chromosome II)	6.56 Mb
(chromosome III)	7.12 Mb
(chromosome IV)	0.14 Mb
(X chromosome)	7.44 Mb
Identified genes (GENEFINDER)	~3980
Gene density:	1 per 4.8 kb (autosomes) 1 per 6.6 kb (X chromosome)
Data base matches:	~45%
Coding sequence:	~28% (exons) ~50% (introns + exons)
Total gene count:	~13,000 (±500)

By comparing the total number of predicted genes with the number of exact EST matches to those genes and dividing by the total number of *C. elegans* ESTs (Waterston et al. 1992; McCombie et al. 1992; Y. Kohara, unpubl.), a figure of ~13,000 (±500) genes can be calculated for the entire genome. This figure assumes only that the expression of genes in the sequenced region is representative of the genome as a whole and is independent of any assumptions regarding the gene density. Comparison of the total number of genes calculated for the whole genome (~13,000) with the number of genes estimated from the completed sequence (~4000) indicates that ~30% of the total number of genes have been found in just over 20% of the genome. This illustrates further the gene-rich distribution over the region sequenced so far.

Interesting Features of the Genome

Now that a relatively large proportion of the genome has been completely sequenced, the search for particularly interesting features that are characteristic of *C. elegans* genome structure can begin. There are several different types of repeat—tandem, inverted, and repeat families. The most common, appearing about every 5 kb on average, are inverted repeats, where a segment of genomic sequence is separated from an inverted copy of itself by a number of unique bases (ranging from a few to several hundred base pairs). Interestingly, a relatively high proportion of these repeats are found in intronic rather than intergenic regions, although the source of the bias and its functional consequences are unknown. Tandem repeats, where multiple copies of the same DNA sequence lie adjacent to one another, are less common (about every 7.5 kb). The most common contain triplet repeats, although dinucleotide repeats are also found. Occasionally, tandem repeats can be complex, with many different tandem repeats arranged in a conserved order in several different parts of the genome (Naclerio et al. 1992). Several *C. elegans* repeats have now been classified into families and at least one of these appears to represent a remnant of an inactive type of transposable element (S. Eddy, unpubl.).

There are a number of examples of gene duplication and in some cases this gives rise to gene clusters showing different levels of divergence within the genes. Figure 2 shows a cosmid that

has nine copies of a cytochrome P450-like gene with identities to each other ranging from 55% to 89% at the protein level and with different intron sizes.

There are several examples of genes, including both tRNA and protein-coding genes, occurring within introns of other genes. In one particular case, a cluster of four related genes on one

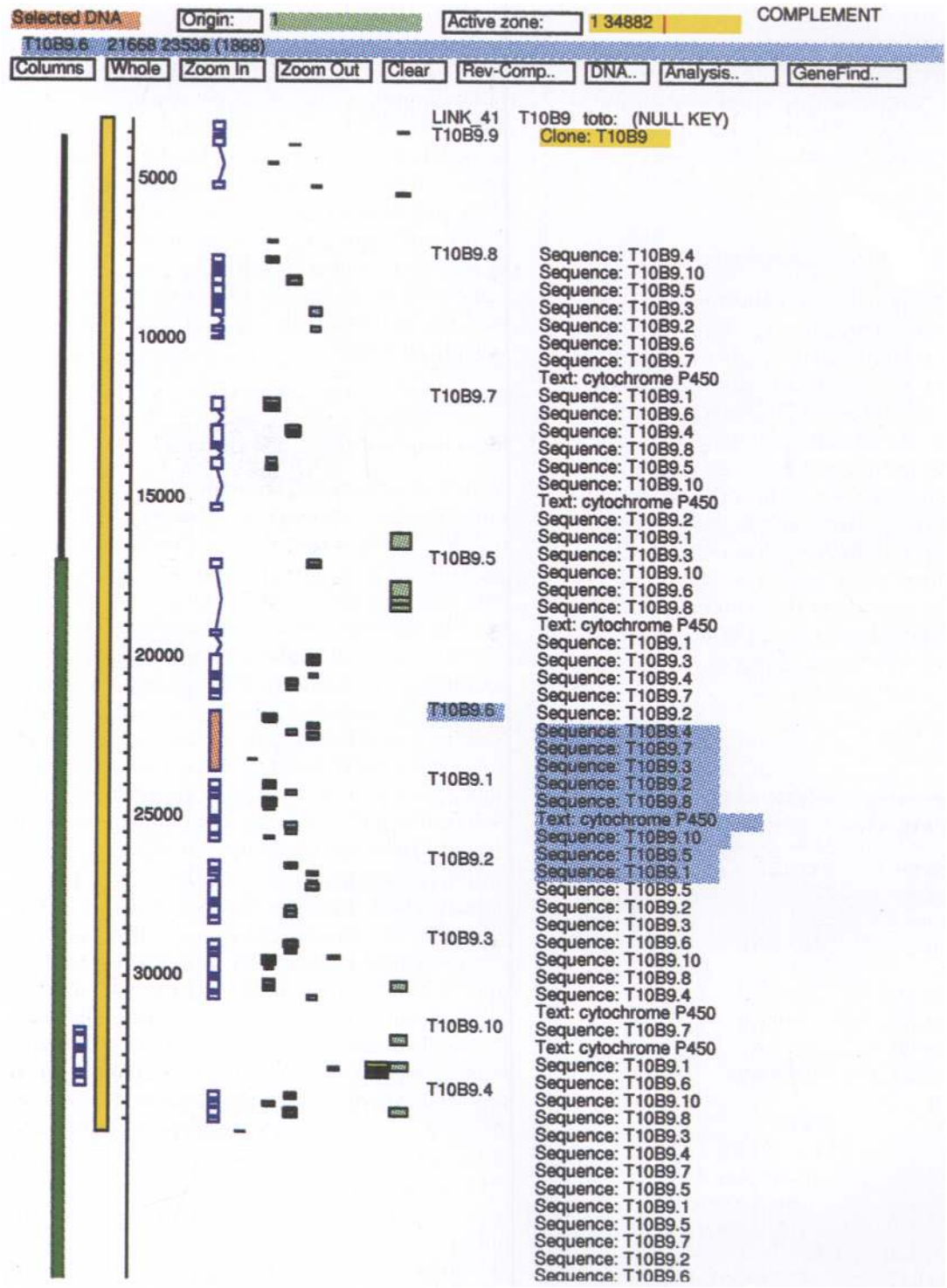

Figure 2 A typical ACEDB output from an analyzed cosmid is shown. Predicted genes (exons and introns) on both strands are indicated on either side of the scale bar. Cosmid T10B9 has a family of nine cytochrome-P450-like genes, varying in the size and number of introns and with varying degrees of divergence in exon sequence.

strand is within a large intron of a predicted gene on the other strand. The head-to-tail arrangement of this small cluster of genes has led to the suggestion that they might be transcribed as a single transcriptional unit (Zorio et al. 1994).

Figure 3 provides a good illustration of the integration of the genomic and transcriptional maps. Although the majority of GENEFINDER-predicted genes do not have matches to *C. elegans* ESTs (only ~25% do), this particular example

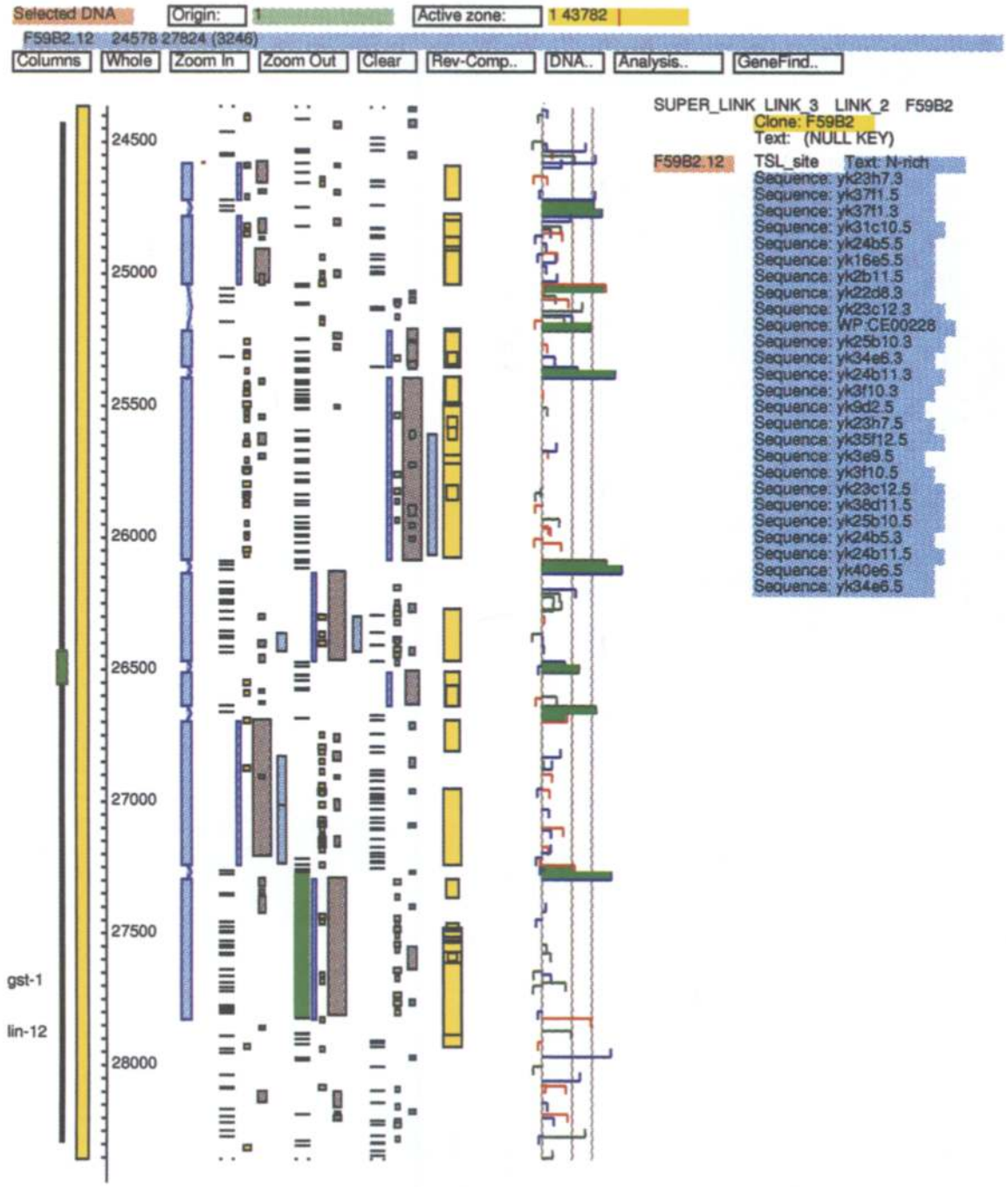

Figure 3 The intron/exon structure of a predicted gene, shown immediately to the *right* of the scale bar, is confirmed by a number of different *C. elegans* ESTs indicated by the yellow boxes toward the *right*. The lines on the extreme *right* represent potential donor and acceptor splice sites, and the green highlighting represents those splice sites that are used by this gene.

shows a gene where each predicted exon is covered by at least one EST, confirming the GENEFINDER exon predictions. In some cases, the EST matches have illustrated alternative splicing patterns for *C. elegans* genes.

Gene-rich Regions to be Completed by the End of 1996

With the current rate of sequencing, the sequence of most of the X chromosome and of the gene-rich regions of the autosomes should be completed by the end of 1996. Assuming that the gene density in the region sequenced so far is typical of the gene-rich region as a whole, then we are likely to have the majority (perhaps as many as 12,000) of the total number of genes. The remaining cosmids in the chromosome arms are expected to yield a significant proportion of the remaining genes, because the regions that cannot be cloned in cosmids tend to be relatively poor in genes. These regions, covered by YAC "bridges," are also likely to be rich in repetitive sequences. Our strategy for dealing with these gaps in the cosmid contigs is still evolving at present: Possible approaches will be to subclone and sequence directly from spanning YACs (Vaudin et al. 1995), to subclone from long-range PCR products, or, possibly, to clone into other vectors such as BACS or PACS.

During these final stages of the project, our strategy toward completing the sequence will change to allow for the larger number of repeat elements and the lower number of genes to be found in these regions. Increasingly, long tandem repeats will be described by their consensus sequence and their number of copies (from restriction digest data).

In summary, the consortium is well on the way to obtaining the complete DNA sequence of a multicellular organism. It is hoped that the large amount of genetic and biological data amassed by the worm research community combined with the detailed molecular information and biochemical resources provided by the consortium will result initially in a greater understanding of *C. elegans* and lead the way to a fuller understanding of other, more complex, organisms.

REFERENCES

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Coulson A., J. Sulston, S. Brenner, and J. Karn. 1986. Toward a physical map the genome of the nematode Caenorhabditis elegans. *Proc. Natl. Acad. Sci.* **83:** 7281–7825.

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

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

Greenwald, I., A. Coulson, J. Sulston, and J. Priess. 1987. Correlation of the physical and genetic maps in the lin-12 region of Caenorhabditis elegans. *Nucleic Acids Res.* **15:** 2295–2307.

McCombie, W., M. Adams, J. Kelley, M. Fitzgerald, T. Utterback, M. Kahn, M. Dubnick, A. Kerlavage, C. Venter, and C. Fields. 1992. Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1:** 124–131.

Naclerio, G., G. Cangiano, A. Coulson, A. Levitt, V. Ruvulo, and A. La Volpe. 1992. Molecular and genomic organisation of clusters of repetitive DNA sequences in Caenorhabditis elegans. *J. Mol. Biol.* **226:** 159–168.

Staden, R. 1994. The Staden package. In *Methods in molecular biology* (ed. A.M. Griffin and H.G. Griffin), vol 25. pp. 9–170. Humana Press Inc., Totawa, NJ.

Sulston, J., Z. Dhu, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qui, S. Dear, A. Coulson, M. Craxton, R. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough, and R. Waterston. 1992. The C. elegans genome sequencing project: A beginning. *Nature* **356:** 37–41.

Vaudin, M., A. Roopra, L. Hillier, R. Brinkman, J. Sulston, R. Wilson, and R. Waterston. 1995. The construction and analysis of M13 libraries prepared from YAC DNA. *Nucleic Acids Res.* **23:** 670–674.

Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R. Durbin, P. Green, R. Shownkeen, N. Halloran, M. Metzstein, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston. 1992. A survey of expressed genes in Caenorhabditis elegans. *Nature Genet.* **1:** 114–123.

Zorio, D., N. Cheng, T. Blumenthal, and J. Spieth. 1994. Operons as a common form of chromosomal organisation in C. elegans. *Nature* **372:** 270–272.