stitutes of Health, Suite 323, 6006 Executive Boulevard, MSC 7052, Bethesda, MD 20892–7052, USA. Human data in this review are derived from published articles and abstracts and from the December 1994 and June 1995 RAC investigator reports. Because the RAC reports are mandated, frequently updated, and public, they are an accurate gauge of the status of the field, although they are not peer-reviewed. Since the first human trial was begun in 1989, there has been an explosion of interest in human gene transfer. In the United States alone, more than 100 human gene transfer protocols have been approved by the RAC, and 697 individuals have participated in human gene transfer trials under RAC-approved protocols (summary data, *RAC Report*, June 1995).
5. H. M. Temin, *Hum. Gene Ther.* **1**, 111 (1990).
6. M. Ali *et al.*, *Gene Ther.* **1**, 367 (1994).
7. T. J. Wickham *et al.*, *Cell* **73**, 309 (1993).
8. M. Rosenfeld *et al.*, *Science* **252**, 431 (1991); P. Lemarchand *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 6482 (1992); G. Bajocchi, S. H. Feldman, R. G. Crystal, A. Mastrangeli, *Nature Genet.* **3**, 229 (1993); H. A. Jaffe *et al.*, *ibid.* **1**, 372 (1992).
9. Y. Yang, Q. Li, H. C. J. Erte, J. M. Wilson, *J. Virol.* **69**, 2004 (1995); Y. Yang *et al.*, *Proc. Nati. Acad. Sci. U.S.A.* **91**, 4407 (1994); Y. Setoguchi, H. A. Jaffe, C.-S. Chu, R. G. Crystal, *Am. J. Resp. Cell Mol. Biol.* **10**, 369 (1994); T. A. Smith *et al.*, *Nature Genet.* **5**, 397 (1993).
10. H. Farhood *et al.*, *Ann N.Y. Acad. Sci.* **716**, 23 (1994); P. L. Felgner, *Liposome Res.* **3**, 3 (1993).
11. J. S. Economou and A. Belldegrun, *RAC Report 9202-015*.
12. R. E. Walker, *ibid. 9209-026*.
13. P. D. Greenberg and S. Riddell, *ibid. 9202-017*.
14. H. E. Heslop *et al.*, *ibid. 9303-038*.
15. C. M. Rooney *et al.*, *Lancet* **345**, 9 (1995).
16. K. Brenner *et al.*, *RAC Report 9102-004*.
17. M. K. Brenner *et al.*, *Lancet* **342**, 1134 (1993).
18. M. K. Brenner, J. Mirro, V. Santana, J. Ihle, *RAC Report 9105-005*; D. R. Rill *et al.*, *Blood* **79**, 2694 (1992).
19. M. K. Brenner, J. Mirro, V. Santana, J. Ihle, *RAC Report 9105-006*.
20. A. B. Deisseroth, *ibid. 9105-007*; A. B. Deisseroth *et al.*, *Blood* **83**, 3068 (1994).
21. K. Cornetta, *RAC Report 9202-014*; K. Cornetta *et al.*, *Blood* **84**, 401A (1994).
22. C. E. Dunbar, *RAC Report 9206-023*.
23. C. E. Dunbar *et al.*, *Blood* **85**, 1306 (1995).
24. C. E. Dunbar, *RAC Report 9206-024*.
25. M. K. Brenner, R. Krance, H. E. Heslop, V. Santana, J. Ihle, *ibid. 9303-039*.
26. A. B. Deisseroth, *ibid. 9206-020*.
27. F. G. Schuening, *ibid. 9209-027*.
28. S. A. Rosenberg, *ibid. 8810-001*; S. A. Rosenberg, *et al.*, *N. Engl. J. Med.* **323**, 570 (1990).
29. M. T. Lotze, *RAC Report 9105-009*; Q. Cai, J. T. Rubin, M. T. Lotze, *Cancer Gene Ther.* **2**, 125 (1995).
30. R. M. Blaese, *RAC Report 9007-002*; R. M. Blaese *et al.*, *Science* **270**, 475 (1995).
31. C. A. Mullen *et al.*, *J. Cell. Biochem.* **18a**, 240 (1994).
32. D. B. Kohn *et al.*, *ibid.*, p. 38; D. B. Kohn *et al.*, *Blood* **82**, 1245 (1993); R. Parkman, K. I. Weinberg, L. Heiss, D. B. Kohn, *J. Cell. Biochem.* **21**, C6 (1995); D. B. Kohn, *J. Cell. Biochem.* **18a**, 57 (1994).
33. S. A. Rosenberg, *RAC Report 9007-003*.
34. M. T. Lotze and J. T. Rubin, *ibid. 9209-033*.
35. M. K. Brenner *et al.*, *ibid.*,*9206-018*.
36. S. A. Rosenberg, *ibid. 9110-011*.
37. G. Dranoff, *ibid. 9411-093*.
38. J. M. Wilson, *ibid. 9110-012*.
39. M. Grossman *et al.*, *Nature Genet.* **6**, 325 (1994); S. E. Raper *et al.*, *Ann. Surg.*, in press.
40. R. E. Walker, *RAC Report 9403-069*.
41. A. B. Deisseroth *et al.*, *ibid. 9406-077*.
42. J. A. Roth, *ibid. 9403-031*.
43. E. H. Oldfield, *ibid. 9206-019*.
44. R. G. Crystal, *ibid. 9212-034*.
45. R. G. Crystal *et al.*, *Nature Genet* **8**, 42 (1994); N. G. McElvaney and R. G. Crystal, *Nature Med.* **1**, 182 (1995).
46. M. J. Welsh, *RAC Report 9212-036*.
47. J. Zabner *et a.l*, *Cell* **75**, 207 (1993).
48. M. J. Welsh, *RAC Report 9312-067*.
49. J. M. Wilson, *ibid. 9212-035*.
50. R. C. Boucher and M. R. Knowles, *ibid. 9303-042*.
51. N. J. Caplen *et al.*, *Nature Med.* **1**, 39 (1995).
52. G .J. Nabel, *RAC Report 9202-013*; G. J. Nabel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11307 (1993); E. G. Nabel *et al.*, *Hum. Gene Ther.* **5**, 1089 (1994); G. J. Nabel, *RAC Report 9306-045*.
53. J. T. Rubin, *RAC Report 9312-064*.
54. E. T. Akporiaye *et al.*, *Annual Meeting of the American Association for Cancer Research*, AACR, Toronto, Canada, 18 to 22 March 1995 (AACR, Philadelphia, 1995); D. Harris *et al.*, *ibid.*
55. N. Vogelzang, *RAC Report 9403-071*.
56. E. Hersh *et al.*, *ibid. 9403-072*.
57. N. M. Kredich and M. S. Hershfield, *The Metabolic Basis of Inherited Disease* (McGraw-Hill, New York, 1989), pp. 1045–1075.
58. M. S. Brown *et al.*, *Nature Genet.* **7**, 349 (1994).
59. J. R. Riordan *et al.*, *Science* **245**, 1066 (1989); B. Kerem *et al.*, *ibid.*, p. 1073; F. S. Collins, *ibid.* **256**, 774 (1992).
60. M. A. Rosenfeld *et al.*, *Cell* **68**, 143 (1992).
61. P. G. Middleton, D. M. Geddes, E. W. Alton, *Eur. Respir. J.* **7**, 2050 (1994); M. R. Knowles, A. M. Paradisco, R. C. Boucher, *Hum. Gene Ther.* **6**, 445 (1995).
62. J. G. Hay *et al.*, *Hum. Gene Ther.*, in press.
63. J. M. Wilson, personal communication.
64. R. G. Crystal *et al.*, *Hum. Gene Ther.* **6**, 667 (1995).
65. R. W. Wilmott, J. A. Whitsett, B. C. Trapnell, *RAC Report 9303-041*.
66. E. H. Oldfield and Z. Ram, *ibid. 9312-059*.
67. R. G. Crystal, unpublished observations.
68. Through the cooperation of investigators, the FDA, and the RAC, production quality control criteria have been agreed on for each vector system. However, because vector design is constantly improving, these criteria continue to evolve.
69. H. Ginsberg and T. Shenk, Chairmen, Adenovirus Breakout Group Report, Cystic Fibrosis Foundation Gene Therapy Meeting, Williamsburg, VA, 4 to 7 June 1995.
70. R. G. Crystal holds equity in GenVec, Inc. (12111 Parklawn Drive, Rockville, MD 20852), a biotechnology company focused on in vivo gene therapy using adenovirus and herpesvirus vectors. I thank N. Wivel and D. Wilson (Office of Recombinant DNA Activities, NIH) for helpful discussions and access to data compiled from human gene transfer trials; E. Falck-Pedersen, A. Mastrangeli, and E. Hirschowitz, Cornell University Medical College, for helpful discussions; and N. Mohamed and J. Macaluso for help in preparing the manuscript.

15 June 1995; accepted 25 September 1995

The Nematode *Caenorhabditis elegans* and Its Genome

Jonathan Hodgkin, Ronald H. A. Plasterk, Robert H. Waterston

Over the past two decades, the small soil nematode *Caenorhabditis elegans* has become established as a major model system for the study of a great variety of problems in biology and medicine. One of its most significant advantages is its simplicity, both in anatomy and in genomic organization. The entire haploid genetic content amounts to 100 million base pairs of DNA, about 1/30 the size of the human value. As a result, *C. elegans* has also provided a pilot system for the construction of physical maps of larger animal and plant genomes, and subsequently for the complete sequencing of those genomes. By mid-1995, approximately one-fifth of the complete DNA sequence of this animal had been determined. *Caenorhabditis elegans* provides a test bed not only for the development and application of mapping and sequencing technologies, but also for the interpretation and use of complete sequence information. This article reviews the progress so far toward a realizable goal—the total description of the genome of a simple animal.

Caenorhabditis elegans has many attractive features as an experimental system (*1*). The life cycle is simple and rapid, with a 3-day generation time, and populations can be grown with ease on agar plates or in liquid, usually by using *Escherichia coli* as a food source. These populations normally consist of only self-fertilizing hermaphrodites, but cross-fertilization is also possible, with the male sexual form. The option of reproduction by either selfing or crossing leads to very convenient genetics so that mutants can readily be generated, propagated, and analyzed (*2*). A simple freezing protocol permits stable storage of all strains, which retain viability indefinitely in the frozen state.

The animal, about 1 mm long when fully grown, is completely transparent at all stages of development. Both development and anatomy are essentially invariant among wild-type individuals. At maturity, all adult hermaphrodites contain 959 somatic nuclei and fewer than 2000 germ cell nuclei. Despite its low cell number, *C. elegans* has fully differentiated tissues corresponding to those of more complicated animals. The transparency and rapid development allow direct examination of cell division and differentiation in living animals with Nomarski microscopy. The small size of the animal also permits reconstruction of the entire anatomy at the ultrastructural level with serial section electron microscopy. However, the

J. Hodgkin is in the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, UK. R. H. A. Plasterk is in the Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121, 1066 CX Amsterdam, Netherlands. R. H. Waterston is in the Department of Genetics and Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63110, USA.

small size does lead to some concomitant disadvantages, as it largely precludes certain experimental approaches, such as tissue transplantation and electrophysiology.

Throughout research on *C. elegans* there has been an emphasis on complete description, both as an end in itself and also to provide the groundwork for experimental studies. The combination of simplicity, lineage tracing, cell identification, and ultrastructural reconstruction has led to a number of landmark accomplishments. A total description of the invariant cell lineage, from the single cell of the fertilized egg to the thousand-odd differentiated cells of the adult, was completed in 1983 (*3*). Accompanying this was a "parts list," enumerating all the identified epithelia, muscles, nerves, intestinal cells, and other cell types that make up the entire anatomy of the adult. All of these parts have been described and reconstructed from serial electron micrographs (EMs). The EM technique was also used for the complete description of the nervous system: The anatomy and connectivity of all 302 adult neurons, as well as the 7600 synaptic junctions they make, were reported in 1986 (*4*).

In parallel, experimental approaches to the *C. elegans* system have been directed at comprehensive analysis. Systematic studies of cell ablation, using a laser microbeam, have been pursued to test functions and regulation within the nervous system and the embryo (*3*). Large-scale hunts for mutants have been carried out, and this mutational approach has often been directed at saturating for certain mutant classes, with the intention of identifying all genes involved in particular developmental or behavioral processes. Other screens have been directed at identifying all essential genes located in defined genetic intervals. The number of genes defined by mutations has expanded from the initial set of 103 reported by Brenner (*2*) in 1974, to approximately 1400 in 1995. Recombinational mapping of these genes defined six linkage groups (five autosomes and an X chromosome), which are roughly equal in size and correspond to the six cytologically visible chromosomes.

Early on, measurements of DNA content and complexity indicated that the genome of *C. elegans* is very small, approaching the apparent limit for a respectably differentiated metazoan animal. This property has loomed ever larger in the investigation of *C. elegans*, first in the construction of a complete physical map, and now in the project to determine the complete DNA sequence.

The Physical Map

Systematic assembly of a physical map of the genome was begun in 1984 (*5*). The initial program made use of a rapid "fingerprinting" technique to identify matches between cloned pieces of DNA in the 10- to 50-kb size range. Several thousand randomly chosen cosmid clones were fingerprinted and assembled in contigs, which are overlapped sets of cosmids that extend from 50 to several hundred kilobases in length. A variety of cosmid vectors and cloning methods were explored at later stages, in the hope of maximizing coverage. Ultimately, more than 17,000 cosmids were analyzed. This phase of the project, known as the thousand islands plan, led to coverage of ~80% of the genome, with cosmids assembled into about 700 island contigs. Many of these could be assigned to precise genetic map locations as a result of the independent cloning of specific genes and also by in situ hybridization.

With time, the cosmid approach began to yield diminishing returns in terms of improving coverage and linking up contigs. The advent of yeast artificial chromosomes (YACs) provided an ideal solution to this problem (*6*). The YAC clones of *C. elegans* DNA have been largely reliable, exhibiting less instability and distinctly less chimerism than those derived from mammalian genomes. Systematic hybridization between YACs and selected cosmids led to the assembly of an almost complete YAC map and to the assignment of almost all cosmids to specific YACs. Ordered arrays of YAC clones on filters, known familiarly as polytene filters, are now generally available. Hybridization of any clone to these filters permits its localization to a resolution of about 100 kb. A set of cosmids corresponding to this location can then be tested, yielding a precise genomic position. All clones generated by the physical mapping program are freely available on request.

The present state of the physical map can be assessed at two levels: the YAC map and the cosmid map. In terms of YACs, the genome is extremely well covered in that there remain only seven gaps in the map, and there is reason to believe that some of these gaps are small. The X chromosome map is now complete, apart from telomeric regions, and consists of a single YAC contig of 18,000 kb. Complete coverage of one of the five autosomes has also been achieved. The reasons for the residual gaps in the autosomal map are not known at present.

A different picture is seen at the level of cosmid coverage, which is dense but far from continuous. On each chromosome there are 50 to 120 regions that are spanned by YACs but not by cosmids; in all, there are a total of about 520 such gaps. The average size of each cosmid contig is therefore ~200 kb, and the longest stretch covered by overlapping cosmids is less than 2000 kb. It is clear, both from statistical arguments and from work with specific regions, that these gaps represent parts of the genome that are difficult to clone in prokaryotic vectors, although they can readily be propagated in YACs. In some cases, the cloning problems can be overcome by using vectors with smaller inserts, or bacterial hosts with different recombinational properties: These may permit isolation of clones that cannot be obtained as cosmids, and consequently lead to gap closure (*7*).

Fortunately, it appears likely that many of the difficult regions of the genome, where cosmid coverage is poor, are also regions containing few genes. For example, one stretch of over 1100 kb on chromosome IV is defined only by YAC clones. No cosmids or known genes are located in this domain, which may contain only heterochromatin, or the *C. elegans* equivalent. These gene-poor regions will not be ignored, but they will be among the last parts of the genome to be subjected to systematic sequencing.

Linking Physical and Genetic Maps

The initial motive for generating such an extensive physical map was to facilitate the positional cloning of known genes. The physical map has indeed been enormously useful in this regard. Work from many laboratories, using a variety of techniques, has led to the tight correlation between the genetic and physical maps of this organism. Over most of the genome, recombinational mapping can now be used to assign a gene to a physical interval of 200 kb or less.

The *C. elegans* transposable element Tc1, a 1.6-kb element belonging to the Tc1-mariner class, has played several important roles in linking the physical and genetic maps (*8*). First, transposon tagging of many genes led to their direct cloning and location on the physical map. Second, restriction fragment length polymorphisms (RFLPs) generated by Tc1 have provided numerous useful landmarks. The standard laboratory strain of *C. elegans* (Bristol) has about 30 copies of Tc1, but there exist other cross-fertile strains with up to 500 copies, distributed apparently at random across the entire genome. Many of these have been cloned, and an increasing number of these have also been used to generate sequence-tagged sites, which provide additional mapping tools for polymerase chain reaction (PCR)–based mapping strategies (*9*). In addition to the polymorphisms generated by Tc1, additional RFLPs between the Bristol strain and other natural isolates are readily identified, at a frequency of one per cosmid or better. These can be used to provide very tight mapping resolution, should this be necessary.

Improving the correspondence between the maps has been significantly aided by the

development of the *C. elegans* database ACeDB, which provides a convenient interactive display of genetic and physical maps, plus underlying genomic and complementary DNA (cDNA) sequences, and much else besides, acting as a universal repository for all information pertaining to this organism (*10*). The correlation between the physical and genetic maps has also produced valuable information about the general organization of the genome, such as demonstrations of long-range variation in gene density and recombination (*11*). The variations in gene density can also be seen in the mapping of cDNA clones, illustrated in the accompanying wall chart.

Sequencing

The current strategy for sequencing the *C. elegans* genome is mainly based on cosmids, because this is the most cost-effective way to proceed during this phase of the project (*12*). The dense physical map permits an optimal choice of minimally overlapped cosmids. Random subclones are generated from each selected cosmid (a "shotgun" approach) and sequenced automatically, giving initial coverage of six- to eightfold for each 35- to 45-kb cosmid. Sequencing reads on each subclone generally yield more than 400 base pairs (bp) of useful data. The shotgun phase is followed by assembly and more dedicated finishing with directed reads, using oligonucleotide primers synthesized for the purpose, in order to fill gaps, to achieve double-stranded coverage, and to complete the linking of contigs. Finished sequence is currently being generated at a total cost of less than 50 cents per base, with an accuracy of ~99.99%.

This phase of the project, which will extend over most of the next 2 years, should provide most of the coding and regulatory information in the genome. However, the cosmid gaps discussed above will present an increasing problem as the project approaches completion. It is probable that further progress will depend on sequencing subclones prepared directly from YACs. This involves additional technical problems but is already feasible. The main difficulty is the unavoidable contamination of purified YACs by substantial amounts of DNA from the yeast host, leading to much wasted time in sequencing and assembling irrelevant yeast sequences. However, this difficulty should be eliminated at the end of 1995, which is the target date for the completion of the complete sequence of *Saccharomyces cerevisiae* (*13*). It will then become possible to discard instantly all sequencing reads that are recognizable as yeast DNA and focus exclusively on the *C. elegans* DNA.

The combination of efficient cosmid and YAC sequencing should yield at least 98% of the *C. elegans* genome by the end of 1998. No radical changes in technology are needed.

Interpretation of Sequence

A completed sequence is subjected to a variety of analyses to detect genes and other features. The most important of these analyses is the GENEFINDER program, which uses properties such as codon usage and splice recognition sequences to predict the coding portions of genes (*14*). Many *C. elegans* introns are small, and splice site consensuses are usually well conserved, so the prediction of exons and gene boundaries seems to be easier and more successful than in larger genomes. However, GENEFINDER is not infallible, nor can it predict alternatively spliced transcripts. An essential source of additional information has been the sequencing of numerous cDNA clones, which confirm exon boundaries predicted by GENEFINDER and also reveal splice variants. Systematic end-sequencing of cDNA clones has been pursued on a large scale (*15*), so by mid-1995 approximately 3500 genes were represented, corresponding to about one-quarter of all the transcribed sequences in the animal. Many of these expressed sequence tags (ESTs) can be immediately assigned to the parts of the genome that have been completely sequenced. Many others have been assigned to specific locations on the physical map, by hybridization to the gridded polytene filters of YAC clones.

Exploration of the sequence of *C. elegans* is an open-ended enterprise at this point, and investigations have hardly begun. Many questions that could never have been answered previously can now be directly addressed. For example, what is the total number of genes in this organism? What are the major gene families? How many genes contain recognizable similarities?

The question of gene number can already be answered fairly precisely with both the predictions from genomic sequence and the distribution of cDNAs: currently the best estimate is 13,100 genes. This is a surprisingly large number, given earlier estimates of fewer than 3000 essential genes (*2*). It suggests that the majority of genes are to some extent redundant or else have subtle functions, perhaps such as those that are significant in the natural ecology of *C. elegans* but not apparent under laboratory conditions.

It seems that we are ignorant about what most of the genome is doing, even in this relatively simple organism. This ignorance is underlined by the fact that more than half of the 3000 genes so far sequenced have no significant similarity to proteins in current databases. We have few clues as to the functions of such genes, a situation that is both challenging and exciting. However, some improvement in this state of affairs is likely to occur as more and more sequence is generated, because genes that currently appear unique may become identifiable as diverged members of known gene families. Information from many organisms can increasingly be used to identify families and ancient conserved regions (*16*).

Already much that is intriguing or exotic has emerged from the millions of base pairs of sequence, such as giant predicted proteins, genes within introns, and clustered genes of unknown function. Some of these features are illustrated in the wall chart: for example, on cosmid F10F2, four related unidentified genes nestle within the largest intron of a recognizable enzyme gene. As shown in Fig. 1, most of cosmid K07E12 seems to be used for encoding a single vast protein that has some similarity to cell adhesion molecules. The expected transcript size for K07E12.1 is 39 kb, corresponding to a predicted protein of 1400 kD.

One notable aspect of the *C. elegans* genome is that ~25% of the genes are organized in polycistronic units of two or more members (*17*). One such example, shown on the wall chart, is cosmid ZK637 which has three adjacent genes that are known to be cotranscribed. The primary transcripts from these units are processed into single-gene transcripts by transplicing to the short RNA splice leaders SL2 or SL1. In some cases, the genes within such "operons" have related functions. In others (like ZK637) there is no evidence of such a connection, but the constituent genes are presumably coexpressed.

Exploitation of Sequence: *C. elegans* as a Test Tube

There are many ways in which the basic data provided by the sequencing consortium can be used and extended. Systematic analyses of the thousands of predicted genes are already in progress. These investigations can be divided into expression studies and functional studies.

Expression patterns for cloned genes can be determined in a number of ways, as illustrated in the wall chart. Ideally, complete cellular and developmental profiles for both transcripts (by in situ hybridization) and protein products (by immunofluorescence) can be obtained. More realistically, much information can be obtained with little effort by carrying out in situ hybridization experiments with many different cDNA clones, leading to the identification of transcripts with distinctive abundance or tissue distribution. A more focused, gene-by-gene transgenic approach is also feasible by using the sequence information to con-

struct fusions between promoter regions and reporter genes such as *lacZ* or the gene for green fluorescent protein (GFP). These fusions can then be tested in transgenic animals, providing data on spatial and temporal expression for each tested gene (*18*). GFP constructs are especially useful in C. *elegans* because the animal is transparent and the patterns of expression can be examined directly in living animals (*19*). Such expression studies are informative and often suggestive as to what each gene is doing, but there is no substitute for more direct functional assays. For this, some kind of genetic approach is essential.

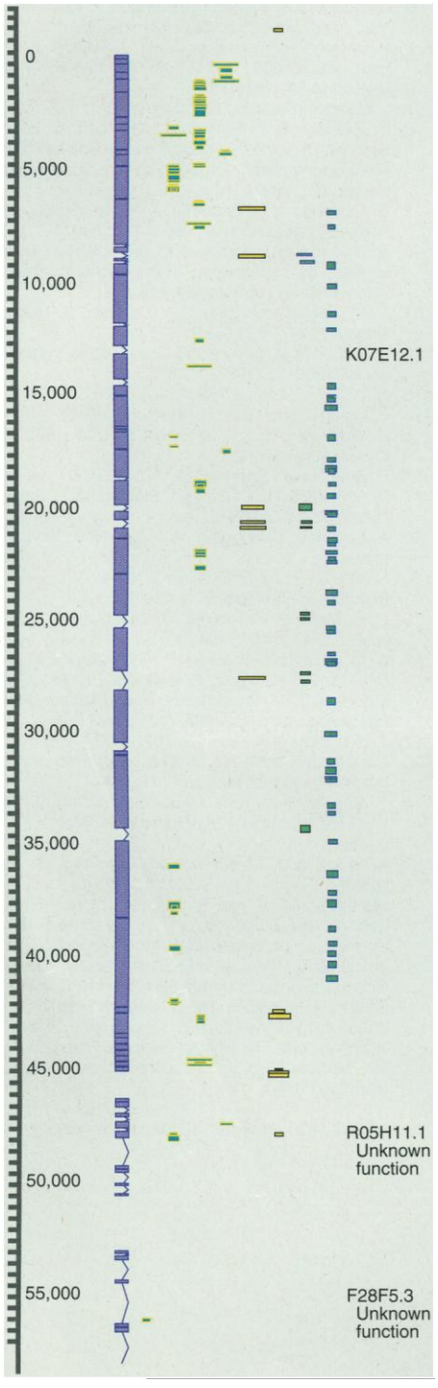

Fig. 1. The genomic region including cosmid K07E12, illustrated with a modified form of the standard ACeDB display (*10*). The scale bar at the left indicates the number of base pairs; next to this is the predicted exon organization (in purple) for three genes (K07E12.1, R05H11.1, and F28F5.3), all transcribed in the same direction. Further to the right are shown significant database matches in different reading frames (three columns), followed by columns indicating the location of repeated sequence families (wide yellow bars) and matches to *C. elegans* cDNA clones (narrower yellow bars). The two rightmost columns (blue-green bars) mark inverted and tandem repeats in the DNA sequence. The giant protein (1400 kD) predicted for K07E12.1 is extensively repetitious and has some sequence similarity to mammalian cell adhesion molecules, but is otherwise novel.

Many biologists working on vertebrate systems are becoming aware of C. *elegans* for the first time because of the identification of a nematode homolog for the gene they happen to be studying. Sometimes the homology will first be detected as partial sequence from a cDNA clone, but increasingly there will be complete genomic information available. What next? First, it is an immediate and automatic bonus to be provided with the entire gene sequence of a C. *elegans* homolog for a vertebrate gene. Sequence comparisons will pinpoint conserved and therefore important protein domains, which may not otherwise be obvious. These conserved regions can then be used to search for other related genes in a more systematic way. Other features, such as intron organization and promoter structure, may also be informative. Second, it will be possible to explore the function of the gene of interest in C. *elegans* with both expression studies and functional tests and by using the power of genetic approaches.

In some cases, mutations may have already been identified in the nematode homolog, as a result of the extensive hunts for mutants carried out in the past. As the correlation of the genetic and physical maps is refined, it is increasingly feasible to associate candidate genes with known mutants. The range of mutant phenotypes that are being studied in C. *elegans* continues to expand, and the characterization of these phenotypes has become steadily more sophisticated. In particular, there is a very large available repertoire of mutations affecting neuronal functions and behavior, a few of which are illustrated in the wall chart. About 100 genes affecting the operation of the locomotory nervous system have been defined by mutation (these are called *unc* genes, for "uncoordinated"), as well as dozens of others affecting sensory functions (chemotaxis, thermotaxis, olfaction, and mechanoreception), specific behaviors (eating, defecation, copulation, and egg-laying), neuronal plasticity, resistance to neuroactive drugs, and others, to a total of about 350 genes at present. As in other animals, a large part of the genome seems to be devoted to neurobiology, and behavior often provides a sensitive assay for subtle alterations in the normal functions of an animal.

Even if no mutations currently exist, a variety of techniques can be used for functional experimentation on an identified gene. These include overexpression, interference by antisense methods, and above all, gene disruption. Both overexpression and antisense methods entail construction of transgenic animals, which are generated by microinjection of cloned DNA (*20*). Such DNA does not normally integrate into the chromosomes but instead forms a multiple copy extrachromosomal array, which can be transmitted fairly stably to offspring. Sometimes the increased copy number of genes in the array is enough to cause a phenotypic effect by itself. Alternatively, expression levels can be boosted by fusing the gene of interest to an inducible promoter, such as a heat-shock promoter.

A more desirable goal is to disrupt the C. *elegans* homolog with procedures that are becoming steadily more streamlined. Targeted gene replacement is not currently feasible; instead, disruption is usually achieved by a two-step process of transposon insertion followed by imprecise excision of the transposon (*21*). Inserts of Tc1 in a gene of interest can be detected by a PCR method, screening DNA samples from a frozen mutant bank. Once such an insertion has been detected, the corresponding population can be thawed, subdivided, and retested. The combined process of PCR testing and sib-selection usually leads rapidly to the isolation of a single worm carrying the desired Tc1 insertion. Sometimes a Tc1 insertion is sufficient to eliminate or greatly reduce gene function [for example, in the case of the gene *goa-1* which encodes one subunit of a heterotrimeric guanosine triphosphate–binding protein (G protein) (*22*)], but often the transposon is located in an intron and has little effect on gene expression. It is then necessary to isolate a derivative mutant line in which the transposon has imprecisely excised, by means of a second round of PCR testing and sib-selection. Imprecise excision usually deletes several kilobases of flanking DNA and thereby creates a null mutation. Some examples of successful isolation of Tc1 inserts are listed in Table 1, to illustrate the range of genes to which this technique has been applied (*23*).

Once mutation or disruption of the nematode gene has been achieved, the resulting phenotype can be examined in detail, and a whole panoply of genetic tech-

Table 1. Reverse genetics: some *C. elegans* genes interrupted by Tc1 insertion.

Gene	Description
apl-1	Amyloid precursor related
cah-1	Adenylate cyclase–associated protein
cct-1	Cytoplasmic chaperone family
cdc-42	Cell cycle protein
cdh-3	Cadherin (cell adhesion molecule)
cdk-5	Cell cycle kinase
ceh-13	Homeobox (labial)
cey-1	Y-box (DNA or RNA binding)
cpr-6	Cathepsin protease
elt-2	GATA transcription factor
fkh-1	Fork-head–related transcription factor
flp-1	Neuropeptide precursor
ges-2	Intestinal carboxylesterase
goa-1	G protein (G_o, alpha subunit)
gpa-2	G protein (G, alpha subunit)
gpb-1	G protein (G, beta subunit)
hlh-3	Helix-loop-helix transcription factor
nhr-1	Nuclear hormone receptor family
odc-1	Ornithine decarboxylase
pes-9	Patterned expression site
pgp-3	P-glycoprotein (multiple-drug resistance)
prk-1	Pim-1 oncoprotein homolog
sod-2	Superoxide dismutase
ssb-1	Single-stranded DNA binding factor
tkr-1	Tachykinin receptor family

niques can be brought into play. Vertebrate homologs and engineered variants of the gene can be tested for function, permitting rapid structure-function correlations. Interactions with mutations in other genes can be explored by construction of double mutants. Screens for genetic modifiers (suppressors or enhancers) can also be carried out. A particular advantage of the *C. elegans* system is that the animal is a diploid which normally reproduces by self-fertilization, so both dominant and recessive modifiers can be selected or screened for. Once modifiers have been identified, they in turn can be mapped and analyzed at the sequence level. In this way, one can expect to identify other interacting genes and elucidate whole pathways. The power of this kind of approach has become well known; for example, in the dissection of *ras*-dependent signaling (*24*).

Prospects

Much current work in molecular biology is constrained by the substantial amount of work needed to clone and sequence a DNA segment. For *C. elegans*, we will soon enter a period in which knowledge of the complete genome sequence can be assumed. This will alter the choice of experimental strategies. Many experiments that are presently very laborious to carry out in vivo or in vitro will become simple and easy to perform "in silico" (by computer), by analyzing the many megabases of stored sequence.

It is reasonable to expect that the complete genome sequence of *C. elegans* will provide, in some sense, the basic formula for constructing a multicellular animal, in much the same way as the complete sequence of *S. cerevisiae* will reveal the basic ingredients for making and maintaining a eukaryotic cell. The phylogenetic position of *C. elegans* is convenient in this regard, because it appears that nematodes diverged at an early point from the rest of the metazoan radiation. Consequently, they provide a universal out-group for the rest of the animal kingdom (*25*). What this means is that if a gene can be identified both in *C. elegans* and in any other kind of animal, whether it be vertebrate, insect, or mollusk, then it must also have been present in the common ancestor of all animals.

With time, it should therefore become clear how much of the *C. elegans* genome is devoted to this basic animal construction kit and how much is associated with specializations that are unique to the phylum Nematoda, or to *C. elegans* itself. For example, the compactness of this genome may be correlated with idiosyncratic features such as the operons, which seem to be absent from larger genomes. Other molecular or biological properties may turn out to be unique to nematodes in general, but absent from other animal groups. Such properties will also be valuable to discern, because nematodes are a large and important group of animals in their own right. Many nematode species have considerable medical or economic significance, as agents of disease and major agricultural pests.

However, it is already clear that a large part of the genome is doing universal things, so much of what is learned from *C. elegans* will also apply to all multicellular organisms. Finally, there is an appealing novelty to the prospect of a complete sequence for such a thoroughly studied animal. How much of the genome will be comprehensible, by arguments from homology or functional experimentation, and how much will remain mysterious? Only by arriving at a complete sequence will we be able to test the limits of our understanding, and perhaps see what questions to ask next.

REFERENCES AND NOTES

1. W. B. Wood, Ed., *The Nematode* Caenorhabditis elegans (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1988); C. Kenyon, *Science* **240**, 1448 (1988); Special issue on *Caenorhabditis elegans*, H. F. Epstein and D. C. Shakes, Eds., *Methods Cell Biol.* **48** (1995).
2. S. Brenner, *Genetics* **77**, 91 (1988).
3. J. E. Sulston and H. R. Horvitz, *Dev. Biol.* **56**, 110 (1977); J. Kimble and D. Hirsh, *ibid.* **70**, 396 (1977); J. E. Sulston, E. Schierenberg, J. N. Thomson, *ibid.* **100**, 64 (1983).
4. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, *Philos. Trans. R. Soc. London* **314**, 1 (1986).
5. A. Coulson, J. Sulston, S. Brenner, J. Karn, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7821 (1986).
6. A. Coulson, R. Waterston, J. Kiff, J. Sulston, Y. Kohara, *Nature* **335**, 184 (1988); A. Coulson *et al.*, *BioEssays* **13**, 413 (1991).
7. J. Hodgkin, *Genetics* **133**, 543 (1993).
8. P. Anderson, S. W. Emmons, D. G. Moerman, in *The Dynamic Genome*, N. Fedoroff and D. Botstein, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992).
9. B. D. Williams, B. Schrank, C. Huynh, R. Shownkeen, R. H. Waterson, *Genetics* **131**, 609 (1992).
10. R. Durbin and J. Thierry-Mieg, unpublished results; available by anonymous FTP from ncbi.nlm.nih.gov in the directory repository/acedb.
11. T. M. Barnes, Y. Kohara, A. Coulson, S. Hekimi, *Genetics* **141**, 159 (1995).
12. J. Sulston *et al.*, *Nature* **356**, 37 (1992); R. Wilson *et al.*, *ibid.* **368**, 32 (1994).
13. N. Williams, *Science* **268**, 1560 (1995).
14. P. Green and L. Hillier, unpublished results.
15. R. Waterston *et al.*, *Nat. Genet.* **1**, 114 (1992); Y. Kohara *et al.*, unpublished results.
16. P. Green *et al.*, *Science* **259**, 1711 (1993).
17. D. A. R. Zorio, N. N. Cheng, T. Blumenthal, J. Spieth, *Nature* **372**, 270 (1994).
18. A. S. Lynch, D. Briggs, I. A. Hope, *Nat. Genet.*, in press.
19. M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, D. C. Prasher, *Science* **263**, 802 (1994).
20. C. C. Mello, J. M. Kramer, D. Stinchcomb, V. Ambros, *EMBO J.* **10**, 3959 (1991).
21. A. M. Rushforth, B. Saari, P. Anderson, *Mol. Cell. Biol.* **13**, 902 (1993); R. R. Zwaal, A. Broeks, J. van Meurs, J. T. M. Groenen, R. H. A. Plasterk, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7431 (1993).
22. J. E. Mendel *et al.*, *Science* **267**, 1652 (1995).
23. Data for this table are derived partly from unpublished work in the laboratory of R.H.A.P.
24. Y. Wu, M. Han, K.-L. Guan, *Genes Dev.* **9**, 742 (1995); K. Kornfeld, K.-L. Guan, H. R. Horvitz, *ibid.*, p. 756.
25. A. Sidow and W. K. Thomas, *Curr. Biol.* **4**, 596 (1994).

26. We thank A. Coulson, R. Durbin, S. Jones, P. Kuwabara, and J. Sulston for valuable comments on the manuscript. Genetic data for *C. elegans* are coordinated by the Caenorhabditis Genetics Center, which is supported by the NIH National Center for Research Resources. The *C. elegans* Mapping and Sequencing Consortium is supported by the NIH National Center for Human Genome Research and the Medical Research Council Human Genome Mapping Project.

14 July 1995; accepted 13 September 1995