Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 10836–10840, November 1995

Review

The genome of *Caenorhabditis elegans*

R. Waterston and J. Sulston†*

**Department of Genetics and Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63110; and †Sanger Centre, Hinxton Hall, Cambs CB10 1RQ, United Kingdom*

ABSTRACT **The physical map of the 100-Mb *Caenorhabditis elegans* genome consists of 17,500 cosmids and 3500 yeast artificial chromosomes (YACs). A total of 22.5 Mb has been sequenced, with the remainder expected by 1998. A further 15.5 Mb of unfinished sequence is freely available online: because the areas sequenced so far are relatively gene rich, about half the 13,000 genes can now be scanned. More than a quarter of the genes are represented by expressed sequence tags (ESTs). All information pertaining to the genome is publicly available in the ACeDB data base.**

Since its inception in the early 1980s, the *Caenorhabditis elegans* genome project has pioneered approaches to physical mapping and genomic sequencing. Today the physical map of the 100-Mb genome is among the largest and most complete yet constructed. More genomic sequence and more sequenced genes are now available from the worm than from any other organism, and we are on target for completion of the full genomic sequence by the end of 1998. The utility of these resources can be judged by the many *C. elegans* laboratories now using the map and sequence to study mutationally defined genes and by the use of sequence homologies by many more laboratories. The genome sequence is critical to gaining a thorough understanding of this important model organism and will aid in studies of human disease.

In this review, we shall attempt to describe the underlying philosophy and the general approaches that we feel have been important for the success of the project. These points are applicable not only to other small genome projects but also to the much larger and more challenging human genome project.

The Genome Map

The physical map consists largely of overlapping cosmid and yeast artificial chromosome (YAC) clones (1–3). Both components are essential: the YACs, by virtue of their large inserts and propagation in yeast, provide long-range continuity and can hold DNA that is unclonable in bacterial cosmid clones, while the cosmids provide high resolution locally and a more convenient substrate for biochemistry. Regions that are unclonable in cosmids are often rich in repetitive sequences and relatively poor in genes.

The principal techniques for physical mapping have been restriction enzyme-based fingerprinting for construction of cosmid contigs; YAC–cosmid hybridizations to gridded arrays; sequence-tagged site (STS) assays for direct detection of YAC–YAC overlaps; and hybridization to *C. elegans* chromosomes for long-range ordering (20). The topological constraints imposed by the ordered cosmids were important for the interpretation of YAC–cosmid hybridizations in that they helped to distinguish genuine matches from spurious matches due to repetitive sequences.

This physical array of cloned DNAs is made into a genome map by the wealth of genetic markers that have been attached to specific clones. This was facilitated by the early and unrestricted distribution of the clone resources and by the research community's readiness to share information prior to publication. In fact, the genetic matches were also critical mapping tools in that they, along with *in situ* hybridization, provide the longest range linkage. Unlike the physical mapping, which was carried out mainly by two central laboratories, the genetic linkage has been achieved by the cooperative effort of the entire *C. elegans* research community. The communal approach is important in two ways. From the point of view of the map, it ensures that the specialized knowledge of all individuals and groups is brought to bear on the project. From the point of view of effort and funding, it means that everyone is involved and allows the central resources to be as lean and focused as possible.

The map as a whole gains from being a multilevel construct; no single technique is sufficient by itself to provide full linkage, and strength arises from partial redundancy between the levels. This is important because all mapping information is to some degree stochastic.

The fingerprinting method is readily scalable and is being applied increasingly to the human genome. Since the original nematode work, the procedure has been automated, and new generations of assembly software are appearing. In contrast to the worm, long-range order in the human genome is being achieved at an earlier stage by STS analysis of YACs and radiation hybrids and by *in situ* hybridization. However, just as in the worm, bacterial clones [whether cosmids, P1s, P1 artificial chromosomes (PACs), or bacterial artificial chromosomes (BACs) (21, 22)] will provide the preferred substrates for biochemistry.

Genome Sequence

In sequencing the genome, we have as far as possible adopted the same philosophy of collective endeavor. The task of the two central laboratories is restricted to collecting the data as efficiently as possible. They strictly refrain from exploiting the sequence for their own research purposes before its release. As soon as the raw data has been assembled into contigs, it is available for screening by anyone [by being placed in an anonymous file transfer protocol (ftp) server]. When each cosmid sequence is finished, it is analyzed by computer with supplemental human interpretation; the annotated sequence is then immediately submitted to GenBank or EMBL, as well as being placed in the *C. elegans* data base ACeDB (4). In this way, the expertise not only of the worm community but of the whole world is brought to bear at the earliest possible stage.

Sequencing at present is concentrated on the central regions of the autosomes and the whole of the X chromosome totaling roughly 60% of the genome. Genetic and cDNA mapping data indicate that these areas contain the majority of the genes (perhaps as many as 90% of the total) (refs. 5 and 6; Y. Kohara, personal communication), and, therefore, the project will deliver high biological value as rapidly as possible. However, this does not weaken the plan to sequence the whole genome; there are many important aspects beyond the acquisition of protein coding sequence.

Starting in this way has the added benefit for us that we are initially sequencing cosmids. Later, we shall have to deal with the "YAC bridges"—i.e., the regions that are cloned in YACs but not in cosmids—and we are experimenting with ways of doing this. Starting from whole YACs,

Abbreviations: YAC, yeast artificial chromosome; STS, sequence-tagged site.

though successfully done (7), is difficult at present because of the limited amounts and purity of material and the greater problems from repeated sequences. However, with improved software and the availability of the yeast sequence, this method may become more practical. Smaller bridges have been successfully rescued by recombination from YACs in yeast, others by long-range PCR, and many regions are susceptible to cloning in λ vectors. During the next year or so, we shall continue to experiment with these and other methods to determine the most efficient approach.

Our principal sequencing strategy is an initial shotgun followed by directed finishing (described in detail in ref. 8). It is worth emphasizing here that the shotgun sequences provide extremely detailed map information, albeit only from a fraction of the subclone insert length. This map information allows the relative positions of the random subclones to be established, while at the same time producing the bulk of the sequence information. Like all mapping methods, it is vulnerable to repeated sequences. The density and accuracy of sequence information, however, compensate for the relatively short length of individual reads. Improved assembly programs (see below) are taking greater advantage of this information, and the sequence itself provides powerful means of evaluating alternative maps. In uncertain areas, additional sequence, for example from the opposite end of the insert, can provide additional map information. In addition, restriction enzyme digests of the parent clone provide a simple and direct means of testing overall map accuracy.

Cost and accuracy are key considerations in evaluating the effectiveness of any strategy. Current direct and indirect costs for the production of the final annotated sequence are approaching $0.45 a base, and total costs of all activities (including development and related research efforts) are below $0.70 a base. In general, the accuracy of the sequence appears to be better than 99.99% judging from comparisons with previously sequenced genes, although these estimates are of limited reliability in the absence of a truly independent means of checking the sequence.

Throughput and the ability to scale the effort are equally important. The combined sequencing rate of the two laboratories should exceed 20 Mb for the current calendar year. This is set to double over the next 18 months, leading to completion in 1996 of the gene-rich regions and in 1997 of the entire set of available cosmids—i.e., 80% of the genome. The throughput is likely to drop during 1997 and 1998, as the remaining gaps are tackled. Realistically, a small mopping up operation may be needed in 1998 and 1999, as the last 1–2% of the genome (involving highly repetitive sequences) is mapped out and representative data are obtained.

It must be made clear that our objective is to extract the information from the genome rather than to exhaustively sequence every last base. For example, even now we sometimes describe long tandem repeats in terms of the consensus and number of copies. The occurrence of such instances is likely to be more frequent as we move into repetitive regions, and therefore, this method of reporting will increase. Conversely, we sequence all other regions as accurately as possible.

The shotgun/directed approach can be applied equally well to the human genome, provided that the extensive repeat families are allowed for in the assembly algorithm. At first, we adapted R. Staden's (23) XGAP by screening the input so that *ALU* sequences were excluded from the initial assembly process. We now begin with P. Green's (University of Washington, Seattle) PHRAP, which makes positive but selective use of repeat sequences in assembly, and then feed the results to XGAP or other editing programs. Given the lower gene density of the human, it is worthwhile to cut costs by deliberately curtailing the finishing process. Many of the problem areas that take the most time to complete are not of immediate biological interest, and therefore time and money can be saved by accurately mapping around them. Should a particular area prove of importance in the future, it can be retrieved by PCR or as a restriction fragment for further investigation.

Status and Applications

The current status of the sequencing effort is summarized in Table 1. All of the sequence has been subjected to a series of programs to provide an initial interpretation of its features. Comparison with expressed sequence tags (ESTs; 11,522 at present from more than 3000 genes) (ref. 6; Y. Kohara, personal communication) allows the construction of a confirmed transcription map for a quarter of the predicted genes.

We have also used the EST information to estimate the number of genes in the entire genome. This can be done by dividing the number of predicted genes in a given sequenced region (1131) by the fraction of the cDNA library for which exact matches have been found in that region (1000/11522 = 0.087). It makes no difference to the calculation that only about a quarter of the genes are represented in the cDNA library. The result, 13,500 genes (±500 at the 95% confidence level), rests on the assumption that the expression of genes in the sequenced region is typical of the genome as a whole; this is likely to be closer to the truth than assuming a uniform gene density.

The genome map is being strengthened by several other systematic studies. These projects are entirely independent, but their findings are united through the map, and some of them draw on its resources. At the University of Leeds, England, Ian Hope is collecting expression data by using transgenic reporter constructs of the predicted genes (refs. 10 and 11; I. Hope, personal communication). Targeted gene disruption by transposon insertion and excision was pioneered by Ronald Plasterk (Netherlands Cancer Institute, Amsterdam) and is now carried out in a number of laboratories; in this way, functionality for the predicted genes can be determined (12). At the National Genetics Institute (Mishima, Japan), Yuji Kohara is continually adding to his set of sequence-tagged cDNAs and is determining their expression patterns by *in situ* hybridization. In

Table 1. Current state of *C. elegans* 100-Mb genome sequencing project

Physical map	17,500 cosmids
	3,500 YACs
	Five autosomes: total of eight gaps (all in gene poor regions)
	X chromosome: single contig of ≈18 Mb
DNA sequence	22.5 Mb completed as of 9/11/95
	Chromosome II: 6.7 Mb
	Chromosome III: 7.2 Mb
	Chromosome X: 8.3 Mb
	Chromosome IV: 0.3 Mb
	4255 putative protein coding genes (≈1 per each 5 kb); approximately 45% have significant similarity to non-*C. elegans* genes.

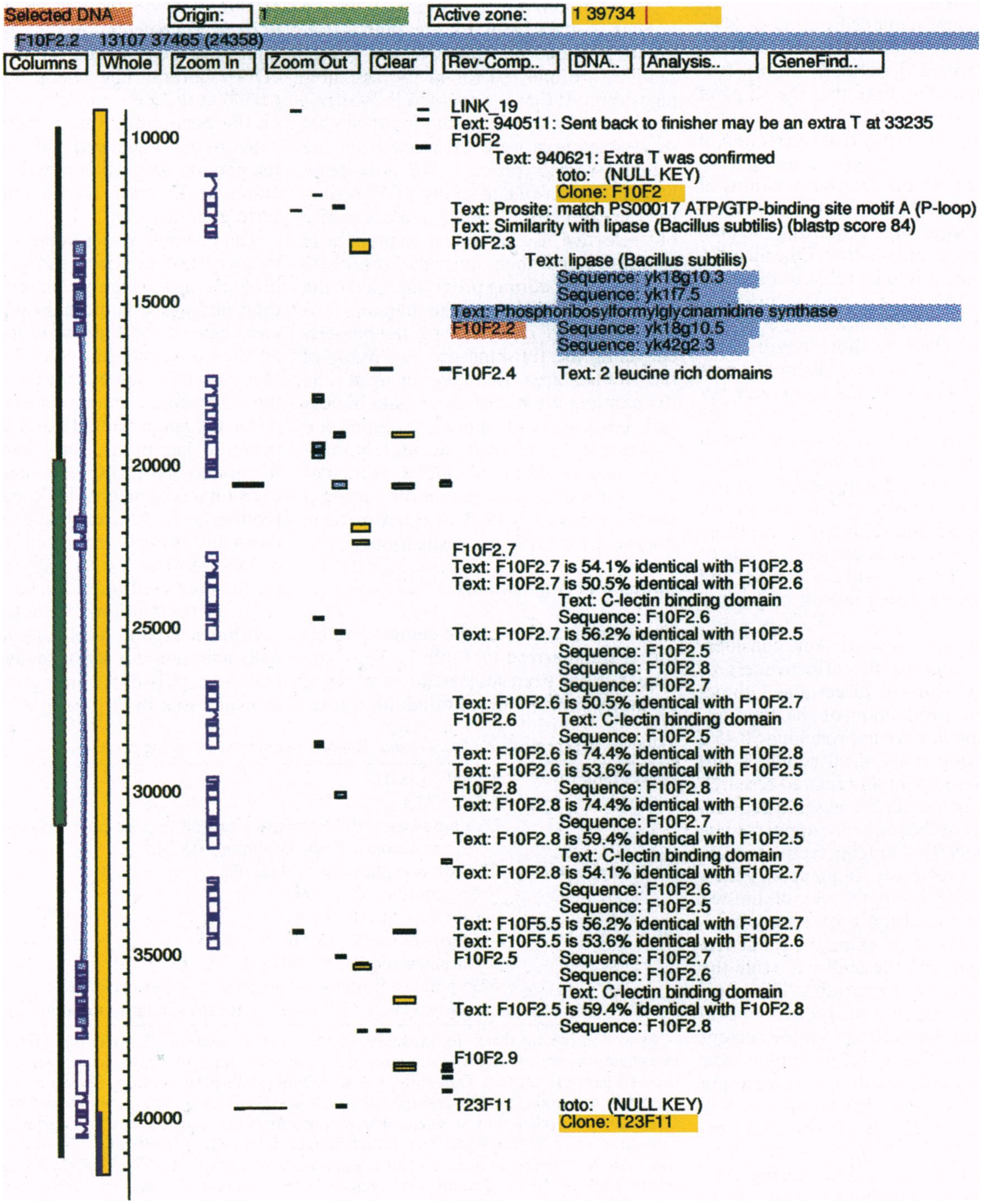
As well as having access to finished sequences in the GenBank, EMBL, and DDBJ data bases, investigators can search these sequence data and also more preliminary unfinished sequences at the two genome centers. Currently, the searchable data base contains 38 Mb of sequence data (22.5 Mb finished, 15.5 Mb unfinished) which is estimated to contain about half of the genes in *C. elegans.* Searches with a nucleotide or protein query use the BLAST programs (9) and are submitted via a World Wide Web interface (see URL http://www.sanger.ac.uk/). This service will soon be available at St Louis (http://genome.wustl.edu/). The Web pages provide additional information about the genome and include help addresses. All data pertaining to the genome (including genetic and physical maps and the sequence) are combined in the data base ACeDB. The latest release can be obtained by anonymous ftp from the following: U.S.A., ncbi.nlm.nih.gov (130.14.20.1) in repository/acedb; U.K., ftp.sanger.ac.uk (193.60.84.11) in pub/acedb; or France, lirmm.lirmm.fr (193.49.104.10) in genome/acedb.

Vancouver, Canada, David Baillie (Simon Fraser University) and Ann Rose (University of British Columbia) are generating transgenic strains incorporating sequenced cosmids; rescue of lethal and visible mutants by these strains will allow precise correlation of the genetic map with the sequence (13–15).

By far the largest use of the genome map and sequence is for the study of specific *C. elegans* genes. Virtually all the 150 *C. elegans* laboratories make use of the map, and most of them request clones from it. Increasingly, laboratories working on other organisms are also using these resources.

Applications of the sequence information are on a variety of levels. At the most mundane level, the prior determination of the sequence by an efficient large-scale operation simply saves subsequent effort and resources. Importantly, the sequence provides ready-made tools, such as a restriction map and information for making primers, that facilitate experimental design.

More creatively, the sequence provides new entry points to the genome. Homologs of known genes or parts of genes can be sought by computer. Not only is this style of searching faster than physical probing but also it is more thorough. Weak similarities, beyond the detection limit of hybridization, can be picked up and evaluated. At present, investigators can search a total of about 38 Mb (22.5 Mb finished, 15.5 Mb unfinished), containing about half of the genes. Searches will become steadily more complete as the



```
LINK_19
Text: 940511: Sent back to finisher may be an extra T at 33235
F10F2
      Text: 940621: Extra T was confirmed
                  toto:  (NULL KEY)
                  Clone: F10F2
Text: Prosite: match PS00017 ATP/GTP-binding site motif A (P-loop)
Text: Similarity with lipase (Bacillus subtilis) (blastp score 84)
F10F2.3
            Text: lipase (Bacillus subtilis)
                  Sequence: yk18g10.3
                  Sequence: yk1f7.5
Text: Phosphoribosylformylglycinamidine synthase
F10F2.2           Sequence: yk18g10.5
                  Sequence: yk42g2.3
F10F2.4           Text: 2 leucine rich domains

F10F2.7
Text: F10F2.7 is 54.1% identical with F10F2.8
Text: F10F2.7 is 50.5% identical with F10F2.6
                  Text: C-lectin binding domain
                  Sequence: F10F2.6
Text: F10F2.7 is 56.2% identical with F10F2.5
                  Sequence: F10F2.5
                  Sequence: F10F2.8
                  Sequence: F10F2.7
Text: F10F2.6 is 50.5% identical with F10F2.7
F10F2.6           Text: C-lectin binding domain
                  Sequence: F10F2.5
Text: F10F2.6 is 74.4% identical with F10F2.8
Text: F10F2.6 is 53.6% identical with F10F2.5
F10F2.8           Sequence: F10F2.8
Text: F10F2.8 is 74.4% identical with F10F2.6
                  Sequence: F10F2.7
Text: F10F2.8 is 59.4% identical with F10F2.5
                  Text: C-lectin binding domain
Text: F10F2.8 is 54.1% identical with F10F2.7
                  Sequence: F10F2.6
                  Sequence: F10F2.5
Text: F10F5.5 is 56.2% identical with F10F2.7
Text: F10F5.5 is 53.6% identical with F10F2.6
F10F2.5           Sequence: F10F2.7
                  Sequence: F10F2.6
                  Text: C-lectin binding domain
Text: F10F2.5 is 59.4% identical with F10F2.8
                  Sequence: F10F2.8

F10F2.9

T23F11            toto:  (NULL KEY)
                  Clone: T23F11
```

FIG. 1. To the left of the kb scale—i.e., on the negative strand—a gene with two large introns is shown. To the right of the scale (positive strand) a family of five genes is seen to lie within the introns. Figs. 1 and 2 are taken from the feature window of ACeDB.

project proceeds and will be greatly enhanced as the emerging families of genes and domains are subjected to cluster analysis.

As the sequenced regions extend along the chromosomes, the large-scale structure of the genome starts to emerge. We are only just beginning to explore this level, but some of the early findings are illustrated in Figs. 1 and 2. Additionally, direct analysis of the sequence yields many interesting features, with applications to gene function, evolution, and medicine. A few selected examples follow.

(*i*) Genes are found within the introns of other genes. In one case, five such genes were found within a single gene (Fig. 1).

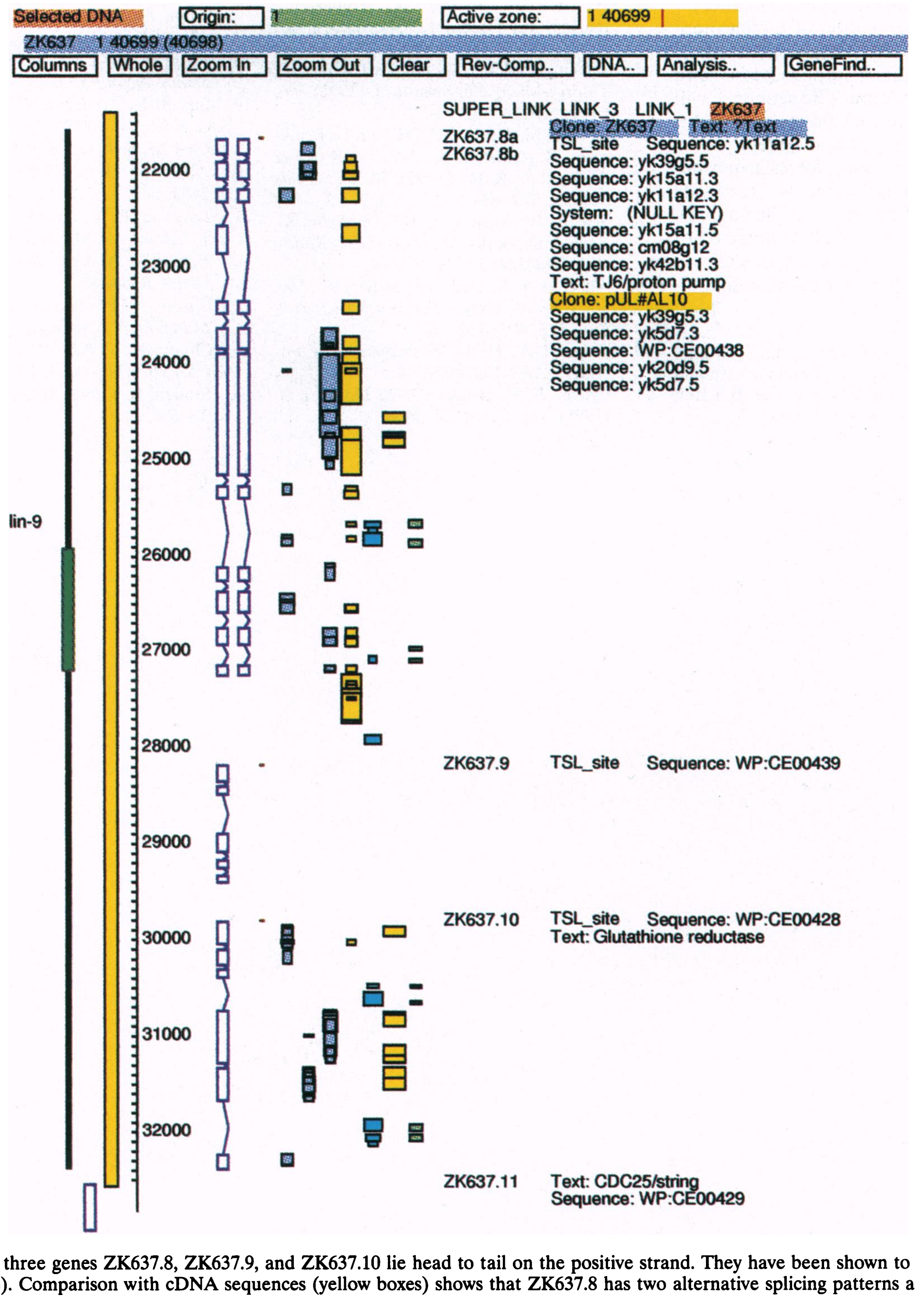
(*ii*) There are clusters of tRNA genes, containing five members in one case and six in another.

(*iii*) Different types of gene families are found: some where the family members are dispersed and others where they are close together in tandem arrays. We can begin to look at the evolution of the individual members.

(*iv*) A relatively high incidence of inverted repeats is found within the introns of predicted genes. The functional significance of this, if any, is unknown.

(*v*) Genes have been found in head-to-tail patterns, which have been shown to be indicative of operons (Fig. 2) (16).

(*vi*) Homologs to genes of medical importance have been found—e.g., the only known invertebrate Lowe syndrome homolog (17).



SUPER_LINK LINK_3 LINK_1 ZK637
Clone: ZK637 Text: ?Text
ZK637.8a TSL_site Sequence: yk11a12.5
ZK637.8b
Sequence: yk39g5.5
Sequence: yk15a11.3
Sequence: yk11a12.3
System: (NULL KEY)
Sequence: yk15a11.5
Sequence: cm08g12
Sequence: yk42b11.3
Text: TJ6/proton pump
Clone: pUL#AL10
Sequence: yk39g5.3
Sequence: yk5d7.3
Sequence: WP:CE00438
Sequence: yk20d9.5
Sequence: yk5d7.5

ZK637.9 TSL_site Sequence: WP:CE00439

ZK637.10 TSL_site Sequence: WP:CE00428
Text: Glutathione reductase

ZK637.11 Text: CDC25/string
Sequence: WP:CE00429

FIG. 2. The three genes ZK637.8, ZK637.9, and ZK637.10 lie head to tail on the positive strand. They have been shown to be transcribed as one operon (16). Comparison with cDNA sequences (yellow boxes) shows that ZK637.8 has two alternative splicing patterns at position 23,000.

(*vii*) The largest known *C. elegans* protein has been predicted from a 45-kb gene in cosmid K07E12. It contains 13,000 amino acid residues and contains multiple copies of the cell adhesion molecule motif (18, 19).

(*viii*) Many repeat families have been identified. Some are thought to be "dead" transposons from an unknown and possibly extinct transposon type.

Apart from patterns of genes, we begin to see the matrix of duplicated, inverted, and transposed pieces of which the genome is composed. Somewhere there are elements that mediate replication, recombination, and segregation of the chromosomes and others that control sex determination, dosage compensation, and global gene expression. The sequence will provide the framework on which hypotheses can be developed and tested.

Finally, the sequence forms a permanent archive whose value we can only begin to tap at the first pass. The analysis, modification, and above all comparison of sequences from different organisms will provide a major route to a full understanding of biology.

This work was supported by the National Institutes of Health National Center for Human Genome Research, the U.K. Medical Research Council, and the Wellcome Trust.

1. Coulson, A. R., Sulston, J. E., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7821–7825.
2. Coulson, A. R., Waterston, R. H., Kiff, J. E., Sulston, J. E. & Kohara, Y. (1988) *Nature (London)* **335,** 184–186.
3. Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J. E. & Waterston, R. H. (1991) *BioEssays* **13,** 413–417.
4. Durbin, R. & Thierry-Mieg, J. (1994) *Computational Methods in Genome Research* (Plenum, New York).
5. Edgley, M. L. & Riddle, D. L. (1990) *Genetic Maps* **5,** 3.
6. Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J. & Sulston, J. (1992) *Nat. Genet.* **1,** 114–123.
7. Vaudin, M., Roopra, A., Hillier, L., Brinkman, R., Sulston, J., Wilson, R. K. & Waterston, R. H. (1995) *Nucleic Acids Res.* **23,** 670–674.
8. Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al.* (1994) *Nature (London)* **368,** 32–38.
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
10. Hope, I. A. (1994) *Development (Cambridge, U.K.)* **120,** 505–514.
11. Lynch, A. S., Briggs, D. & Hope, I. A. (1995) *Nat. Genet.*, in press.
12. Zwaal, R. R., Broeks, A., van Meurs, J. & Plasterk, R. H. A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7431–7435.
13. Howell, A. M. & Rose, A. M. (1990) *Genetics* **126,** 583–592.
14. Rose, A. M., Edgley, M. L. & Baillie, D. L. (1994) *Adv. Mol. Plant Nematol.*, 19–33.
15. Schein, J. E., Marra, M. A., Benian, G. M., Fields, C. A. & Baillie, D. L. (1993) *Genome* **36,** 1148–1156.
16. Zorio, D. A. R., Cheng, N. S. N., Blumenthal, T. E. & Spieth, J. (1994) *Nature (London)* **372,** 270–272.
17. Leahey, A. M., Charnas, L. R. & Nussbaum, R. L. (1993) *Hum. Mol. Genet.* **2,** 461–463.
18. Rathjen, F. G. & Jessell, T. M. (1991) *Semin. Neurosci.* **3,** 297–307.
19. Streuli, M., Krueger, N. X., Tsai, A. Y. M. & Saito, H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 8698–8702.
20. Albertson, D. G. (1985) *EMBO J.* **4,** 2493–2498.
21. Ioannou, P. A., Amemiya, C. T., Garnes, J., Kroisel, P. M., Shizuya, H., Chen, C., Batzer, M. A. & de Jong, P. J. (1994) *Nat. Genet.* **6,** 84–89.
22. Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y. & Simon, M. (1992) *Proc. Acad. Natl. Sci. USA* **89,** 8794–8797.
23. Staden, R. (1994) *Methods Mol. Biol.* **25,** 37–67.