Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

REVIEW

The 97-megabase genomic sequence of the nematode *Caenorhabditis elegans* reveals over 19,000 genes. More than 40 percent of the predicted protein products find significant matches in other organisms. There is a variety of repeated sequences, both local and dispersed. The distinctive distribution of some repeats and highly conserved genes provides evidence for a regional organization of the chromosomes.

The genome sequence of *C. elegans* is essentially complete. The sequence follows those of viruses, several bacteria, and a yeast (*1, 2*) and is the first from a multicellular organism. Some small gaps remain to be closed, but this will be a prolonged process without much biological return. It therefore now makes sense to review the project as a whole.

Here, we describe the origins of the project, the reasons for undertaking it, and the methods that have been used, and we provide a brief overview of the analytical findings. The project began with the development of a clone-based physical map (*3, 4*) to facilitate the molecular analysis of genes, which were being discovered at an ever increasing pace through the study of mutants. This, in turn, initiated a collaboration between the *C. elegans* Sequencing Consortium and the entire community of *C. elegans* researchers (*5*). The resulting free exchange of data and the immediate release of map information (and later sequence) have been hallmarks of the project. The resultant cross correlation between physical and genetic maps is ongoing and is essential for achieving an increasing utility of the sequence.

Along with the genome sequencing project, expressed sequence tag (EST) sequencing has been carried out. Early surveys of expressed sequences were conducted (*6*), but complementary DNA (cDNA) analysis has been carried out primarily by Y. Kohara (*7*). This group has contributed 67,815 ESTs from 40,379 clones, representing an estimated 7432 genes. This extensive information has been invaluable in identifying and annotating genes in the genomic sequence. Others also contributed the 15-kilobase (kb) mitochondrial genome sequence (*8*).

Sequencing

The preexisting physical map, on which sequencing was based, had been initiated by the isolation and assembly of random cosmid clones (with a 40-kb insert, which was the largest insert cloning system available at the time) with a fingerprinting method (*3*). At a sixfold redundant coverage of the genome in cosmids, nonrandom gaps persisted. In most cases, hybridization screening of cosmid libraries failed to yield bridging clones, but the newly developed yeast artificial chromosome (YAC) clones (*9*) rapidly closed most of the cosmid gaps. Incidentally, the YAC clones also covered almost all of the genome, providing a convenient tool for the rapid scanning of the entire genome by hybridization (*4*). About 20% of the genome is represented only in YACs.

By 1989, it became apparent that, with the physical map in hand, complete sequencing of the genome might be both feasible and desirable. Joint funding [from the National Institutes of Health and the UK Medical Research Council (MRC)] for a pilot study was arranged, and in 1990, the first 3-megabase (Mb) sequence was undertaken. Success in this venture (*10, 11*) resulted in full funding and the expansion of the two groups of the consortium in 1993.

Sequencing began in the centers of the chromosomes, where cosmid coverage and the density of genetic markers are high. Cosmids were selected by fingerprint analysis to achieve a tiling path of overlapping clones (in practice, 25% overlap on average). Some sequencing of YACs was explored (*12*), but because of yeast DNA that contaminated preparations of YAC DNA, this approach was deferred in anticipation of the complete sequence of yeast, which enabled contaminating reads to be easily identified. The sequencing process (*13*) can be divided into two major parts: the shotgun phase, which is sequence acquisition from random subclones, and the finishing phase, which is directed sequence acquisition to close any remaining gaps and to resolve ambiguities and low-quality areas. Numerous and ongoing improvements to the shotgun phase have increased sequencing efficiency, improved data quality, and lowered costs. Similarly, finishing tools have improved dramatically. Nonetheless, finishing still requires substantial manual intervention, with a variety of specialized techniques (*14, 15*).

Restriction digests with several enzymes were performed on most cosmids and provided valuable checks on sequence assembly. Where assembly was ambiguous because of repeats, the digests were helpful in resolving the problem. At the start of the project, polymerase chain reaction (PCR) checks were conducted along the length of the sequence to confirm that the assembled sequence of the bacterial clone was an accurate representation of the genome. These checks were abandoned after it became clear that failures in PCR were more common than discrepancies between the clone and the genome.

When available cosmids were exhausted, we screened fosmids (which are similar to cosmids but are maintained at a single copy per cell and thus are potentially more stable) (*16*) and found that a third of the gaps were bridged in the central regions of the chromosomes but very few were bridged in the outer regions. We also used long-range PCR (*17*) to recover some of the central gaps. The remainder of the central gaps and all of the gaps in the outer regions were recovered by sequencing YACs. As for the cosmids, a tiling path of YACs was chosen, and DNA from selected clones was isolated by pulsed-field gel electrophoresis (*18*). Sequencing was performed as for cosmids, with suitable adaptations for the smaller amount of DNA that was available for making libraries. Restriction digests were carried out for assembly checks, but they were not as precisely interpretable as those for bacterial clones. At this stage, the physical map was consolidated and sometimes rearranged as the YAC sequences confirmed or rejected the links made previously by hybridization. The comparison of the assembled YAC sequences with the often extensively overlapping cosmid sequences showed few discrepancies between the two sequences. Generally, further investigation revealed that most discrepancies resulted from a rearrangement in the cosmid. It is interesting (and crucial to the success of the YAC sequencing) that nearly all regions of the YACs can be cloned in

*See genome.wustl.edu/gsc/C_elegans/ and www.sanger.ac.uk/Projects/C_elegans/ for a list of authors. Address correspondence to The Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park Parkway, St. Louis, MO 63108, USA. E-mail: worm@watson.wustl.edu; or The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. E-mail: worm@sanger.ac.uk

bacteria as short fragments, although cosmid and fosmid libraries failed to represent these regions.

The key step in closing sequence assemblies was to obtain subclones that bridged the gaps remaining after the shotgun phase. Often, gaps are spanned by the subclones used in the shotgun phase, because the insert length is deliberately set at two to four times the typical sequence read length. The introduction of plasmid clones halfway through the program greatly improved the coverage of inverted repeats and other unusual structures. In cases where the shotgun phase failed to yield a spanning subclone, plasmid clones that bridged gaps were obtained by isolating and subcloning restriction fragments from cosmids. In YACs, because of their greater size and complexity, screening by hybridization was necessary to recover the desired subclone. In the most difficult cases, we have exploited very short insert plasmid libraries to find gap-bridging clones. PCR was used occasionally, but because of its tendency to yield artifacts in repeat regions, it has recently been used as little as possible. Once isolated, the gap-bridging clone was either sequenced directly or, in cases of a difficult secondary structure, a short insert library (SIL) was made by breaking the insert of the gap-bridging clone into smaller fragments (0.5 kb or even smaller in difficult cases), with break points interrupting the secondary structure (*15*). In some cases, transposon insertion has been used (*19*), although SILs are generally preferred as a first pass because of their ease of throughput.

The 97-Mb sequence is a composite of 2527 cosmids, 257 YACs, 113 fosmids, and 44 PCR products (*20, 21*). For the 12 chromosome ends, nine of the telomere plasmid clones provided by Wicky *et al.* have been linked to the outermost YACs (*22*), either directly by sequence or by long-range PCR and sequencing, where no direct sequence link was found. This probably represents >99% of the genome sequence, on the basis of the representation in the genomic sequence of available EST data and of the sequence from random clones from a whole-genome library.

Much of the remaining DNA likely resides in the three residual gaps between the telomeres and the outermost sequenced YACs and in two internal gaps, where no spanning YAC clone has been identified. One of these is known to be <450 kb, on the basis of Southern (DNA) analysis, but a reliable size estimate is not available for the other gaps. A smaller amount will be recovered from four smaller segments (which are spanned by YACs), where shotgun sequencing has not been completed. Furthermore, very small segments (likely to be <1 kb each) have not been recovered in subclones for 139 segments. Finally, some sequence is likely to be missing from the large tandem repeats, which, in extreme cases, consist of tens of kilobases that are composed of hundreds of copies of a short sequence. Although most have been sized by restriction digestion of the cloned DNA, some segments in the larger YACs are of unknown size. Having established the repeat elements, we cannot usefully work further on them at this stage, because they are likely to be variable and because they do not clone stably; any repeat elements that prove to be important will become the subject of population studies in the future.

As shown by the resolution of discrepancies resulting from matches with sequence data from other sources, the error rate of almost all the product is $<10^{-4}$. In a few regions (predominantly in regions of extensive tandem repeats), the sequence is tagged to indicate that a lower standard of accuracy has been accepted. Accuracy is maintained by a set of criteria (*23*), which is followed by the finisher and by a final checking step that requires specialized software (*24*) and a visual inspection. None of this, however, overcomes errors in the cloning process. A comparison of different clones in overlapping regions and the resolution of discrepancies have indicated a finite error rate associated with cloning. For example, cosmid B0393 (GenBank accession number Z37983) contains a deletion of a large hairpin that was only detected because it overlapped cosmid F17C8 (GenBank accession number Z35719); similarly, we detected a 400–base pair region that had been deleted in all M13 and PCR reads from cosmid F59D12 (GenBank accession number Z81558). The F59D12 deletion was detected by restriction digestion and was recovered in plasmids. However, these instances are rare enough that undetected errors are likely to be few; thus, the advantages of the clone-based sequence, in avoiding long-range confusion in assembly, more than make up for its occasional defects.

Sequence Content

Whereas the sequencing has essentially been completed, analysis and annotation will continue for many years, as more information and better sequence annotation tools become available.

To begin the task, we subjected each completed segment to a series of automatic analyses to reveal possible protein (*25*) and transfer RNA (tRNA) genes (*26*), similarities to ESTs and other proteins (*27–30*), repeat families, and local repeats (*31*). The results were entered in the genome database "a *C. elegans* database" (ACEDB) (*32*), which merges overlapping sequences to provide seamless views across clone boundaries and allows the periodic and automatic updating of entries. To integrate and reconcile the various views of the sequence, we reviewed all data interactively through the ACEDB annotator's graphical workbench (*32*). In particular, the GENEFINDER (*25*) predictions are confirmed or adjusted to account for protein, cDNA, and EST matches, repeats, and so forth, and annotation concerning putative gene function is added.

The interruption of the coding sequence by introns, the generation of alternatively spliced forms, and the relatively low gene density make accurate gene prediction more challenging in multicellular organisms than in microbial genomes. The problem is made more complex in *C. elegans* by transplicing and by the organization of as many as 25% of the genes into operons (*33*). We have used GENEFINDER to identify putative coding regions and to provide an initial overview of gene structure. To quantitate the accuracy of gene identification, we compared introns that were confirmed by ESTs and cDNAs to those that were predicted by GENEFINDER. We found that 92% of the predicted introns had an exact match to the experimentally confirmed ones and that 97% had an overlap. Identification of the start and stop of genes is more difficult, and errors in this process sometimes result in the merging of some neighboring genes and in the splitting of others. To refine the computer-generated gene structure predictions, expert annotators use any available EST and protein similarities, as well as genomic sequence data from the related nematode *C. briggsae*. This information can be especially important in establishing gene boundaries. About 40% of the predicted genes have a confirming EST match, but because ESTs are partial, they presently confirm only ~15% of the total coding sequence. In a number of cases, ESTs have provided direct evidence of alternative splicing; these instances have been annotated in the sequence (*34*).

The genes. The 97-Mb total sequence contains 19,099 predicted protein-coding genes—16,260 of which have been interactively reviewed, for an average density of 1 predicted gene per 5 kb (*35*). Each gene has an average of five introns, and 27% of the genome resides in predicted exons. The number of genes is about three times that found in yeast (*2*) and is about one-fifth to one-third the number predicted for humans. As expected from earlier estimates that were based on much smaller amounts of genome sequence, the number of predicted genes is much higher than the number of essential genes that was estimated from classical genetic studies (*10, 36*).

Similarities to known proteins provide a glimpse of the possible function of the predicted genes. Approximately 42% of predicted protein products have distant matches (outside Nematoda); most of these matches contain functional information (*37*). Another 34% of predicted proteins match only other nematode proteins, but only a few

of these have been functionally characterized. The fraction of genes with informative similarities is far lower than the 70% seen for microbial genomes. This may reflect the smaller proportion of nematode genes that are devoted to core cellular functions (*38*), the comparative lack of knowledge of functions involved in building an animal, and the evolutionary divergence of nematodes from other animals studied extensively at the molecular level.

We compared the available protein sets from *C. elegans*, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens* to highlight qualitative differences in the predicted protein sets (*39*) (Fig. 1). Generally, we found that smaller genomes had matches to a larger fraction of their protein sets and larger genomes had higher numbers of matching proteins. As expected from evolutionary relationships, there were substantially more protein similarities found between *C. elegans* and *H. sapiens* than between any other cross-species pairwise comparison. There were also a substantial number of proteins common to *C. elegans* and *E. coli* that were not found in yeast. Similarly, *C. elegans* lacked proteins that were found in both yeast and *E. coli* (*38*).

Genes encoding proteins with distant matches (outside Nematoda) were more likely to have a matching EST (60%) than those without such matches (20%). This observation suggests that conserved genes are more likely to be highly expressed, perhaps reflecting a bias for "housekeeping" genes among the conserved set. Alternatively, genes lacking confirmatory matches may be more likely to be false predictions, although our analyses do not support this (*40*).

We have also used the Pfam protein family database (*41*) to classify common protein domains in the nematode genome. Of the 20 defined domains that occur most frequently (Table 1), the majority are implicated in intercellular communication or in transcriptional regulation. We find comparatively fewer examples of second messenger proteins (for example, 54 G-beta and 3 Src homology 2 domains). This finding supports models in which the same intracellular signaling pathways are used with variant receptors and transcription factors in different cell states.

In addition to the protein-coding genes, the genome contains at least several hundred genes for noncoding RNAs. There are 659 widely dispersed tRNA genes and at least 29 tRNA-derived pseudogenes (*42*). Forty-four percent of the tRNA genes are found on the X chromosome, which contains only 20% of the total sequence. Several other noncoding RNA genes occur in dispersed multigene families. The U1, U2, U4, U5, and U6 spliceosomal RNA genes occur in 14, 21, 5, 12, and 20 dispersed copies, respectively; there are five dispersed copies of signal recognition particle RNA genes, and there are at least four dispersed copies of splice leader 2 (SL2) RNA genes. A striking feature of these dispersed gene families is their high degree of sequence homogeneity. For example, of the 20 U6 RNA genes, 17 are 100% identical to each other. Either gene conversion or recent gene duplications may account for this homogeneity. Several of these RNA genes occur in the introns of protein-coding genes, which may indicate RNA gene transposition. In general, RNA genes in introns do not appear to occur preferentially in the coding orientation of the encompassing transcript, which indicates that these RNA genes are probably expressed independently.

Other noncoding RNA genes occur in long tandem arrays. The ribosomal RNA genes occur solely in such an array at the end of chromosome I. The 5*S* RNA genes occur in a tandem array on chromosome V, with array members separated by SL1 splice leader RNA genes. A few other known RNA genes, such as the small cytoplasmic Ro-associated Y RNA and the *lin-4* regulatory RNA, are found only once in the genome. Some RNA genes that are expected to be present in the genome have yet to be identified, probably because they are poorly conserved at both the sequence and secondary structure level. These include ribonuclease P RNA, telomerase RNA, and 100 or more small nucleolar RNA genes.

Repetitive sequences. Some of the sequence that does not code for protein or RNA is undoubtedly involved in gene regulation or in the maintenance and movement of chromosomes. A significant fraction of the sequence is repetitive, as in other multicellular organisms. We have classified repeat sequences as either local (that is, tandem, inverted, or simple sequence repeats) or dispersed.

Tandem repeats account for 2.7% of the genome and are found, on average, once per 3.6 kb. Inverted repeats account for 3.6% of the genome and are found, on average, once per 4.9 kb. Many repeat families are distributed nonuniformly with respect to genes and, in particular, are more likely to be found within introns than between genes. For example, although only 26% of the genome sequence is predicted to be intronic, it contains 51% of the tandem repeats and 45% of the inverted repeats. The 47% of the genome sequence that is predicted to be intergenic contains only 49% of the tandem repeats and 55% of the inverted repeats. As expected, only a small percentage

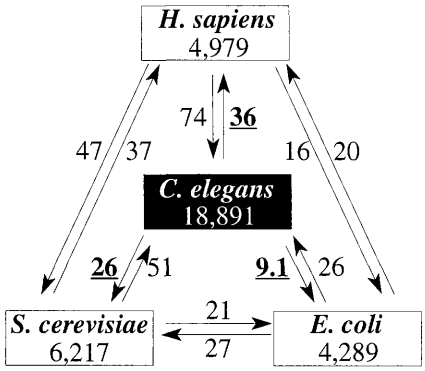

Fig. 1. Percentages of matching proteins resulting from pairwise comparisons (*39*). The organisms and the number of proteins used in the analysis are shown in boxes. For *S. cerevisiae* (a fungus), *C. elegans* (a nematode), and *E. coli* (a bacteria), the numbers reflect proteins that were predicted from an essentially complete genome sequence. The direction of the arrows indicates how the comparison was performed. Numbers that are adjacent to the arrows indicate the percentage of proteins that were found to match. Numbers that are underlined and in bold-faced type indicate the percentage of *C. elegans* proteins that were found to match each of the other organisms.

Table 1. The 20 most common protein domains in *C. elegans* (*41*). RRM, RNA recognition motif; RBD, RNA binding domain; RNP, ribonuclear protein motif; UDP, uridine 5′-diphosphate.

Number	Description
650	7 TM chemoreceptor
410	Eukaryotic protein kinase domain
240	Zinc finger, C4 type (two domains)
170	Collagen
140	7 TM receptor (rhodopsin family)
130	Zinc finger, C2H2 type
120	Lectin C-type domain short and long forms
100	RNA recognition motif (RRM, RBD, or RNP domain)
90	Zinc finger, C3HC4 type (RING finger)
90	Protein-tyrosine phosphatase
90	Ankyrin repeat
90	WD domain, G-beta repeats
80	Homeobox domain
80	Neurotransmitter-gated ion channel
80	Cytochrome P450
80	Helicases conserved C-terminal domain
80	Alcohol/other dehydrogenases, short-chain type
70	UDP-glucoronosyl and UDP-glucosyl transferases
70	EGF-like domain
70	Immunoglobulin superfamily

of the tandem repeats overlaps with the 27% of the genome encoding proteins.

Although local repeat structures are often unique in the genome, others come in families. For example, repeat sequence CeRep26 is the tandemly occurring hexamer repeat TTAGGC, which is seen at multiple sites that are internal to the chromosomes in addition to the telomeres (*22*). CeRep26 and CeRep27 are excluded from introns, whereas other repeat families show a slight positive bias toward introns. The reason for the biased distribution of these repeats is unclear. Furthermore, some repeat families show a chromosome-specific bias in representation. For example, CeRep11, with 711 copies distributed over the autosomes, has only one copy located on the X chromosome.

Altogether, we have recognized 38 dispersed repeat families. Most of these dispersed repeats are associated with transposition in some form (*43*) and include the previously described known transposons of *C. elegans*. However, these repeat elements may not explicitly encode an active transposon (*44*). For example, we have found four new families of the Tc1/mariner type, but these are highly divergent from each other and the other family members; they are probably no longer active in the genome.

In addition to multicopy repeat families, we observe a substantial amount of simple duplication of sequence, that is, segments ranging from hundreds of bases to tens of kilobases that have been copied in the genome. In one case, a segment of 108 kb containing six genes is duplicated tandemly with only 10 sites observed to be different between the two copies. At the left end of chromosome IV, immediately adjacent to the telomere, an inverted repeat is present where each copy of the repeat is 23.5 kb, with only eight different sites found between the two copies. Many cases of shorter duplications are found, which are often separated by tens of kilobases or more that may also contain a coding sequence. These duplications could provide a mechanism for copy divergence and the subsequent formation of new genes. In one example, two 2.5-kb segments, separated by 200 kb, were found to contain genes exhibiting a 98% sequence identity (C38C10.4 and F22B7.5). EST data indicate that both genes are expressed. More commonly, gene duplications are local. In a search for local clusters of duplicated genes, 402 clusters were found distributed throughout the genome (Fig. 2).

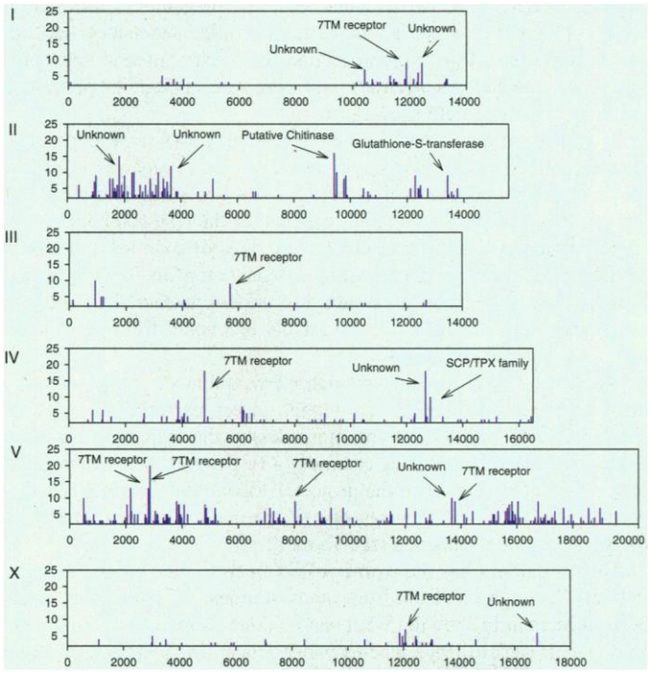

Fig. 2. Locations by chromosome (shown by roman numerals) of local gene clusters. The *x* axis represents the physical distance in kilobases along the chromosomes. The *y* axis represents the size of the clusters. For example, the chitinase cluster on chromosome II contains 17 chitinase-like genes. Local gene clusters were determined by searching for all cases of *N* genes that are similar within a window of 2*N* genes along the chromosomes (for example, three similar genes within a window of six were considered a cluster; clusters were extended until no similar genes could be added). Clusters of $N = 3$ or more were plotted. The criterion for similarity was defined as a BLASTP score of at least 200. ATP, adenosine 5′-triphosphate; TM, transmembrane; Mem. Recep., membrane receptor; SCP/TPX, a family of proteins (SCP, sperm-coating glycoprotein; TPX, Tpx-1, a testis-specific protein).

Chromosome organization. At first sight, the genome looks remarkably uniform; GC content (36%) is essentially unchanged across all the chromosomes, unlike the GC content in vertebrate genomes, such as human, or yeast (*45*). There are no localized centromeres as found in most other metazoa. Instead, the extensive, highly repetitive sequences that are characteristic of centromeres in other organisms may be represented by some of the many tandem repeats found scattered among the genes, particularly on the chromosome arms. Gene density is also fairly constant across the chromosomes, although some differences are apparent, particularly between the centers of the autosomes, the autosome arms, and the X chromosome (Table 2 and Fig. 3).

Striking differences become evident after an examination of other features. Both inverted and tandem repetitive sequences are more frequent on the autosome arms (Fig. 3) than in the central regions of the chromosomes or on the X chromosome. For example, CeRep26 is virtually excluded from the centers of the autosomes (Fig. 3). (The abundance of repeats on the arms is likely to be a contributing factor to the difficulties in cosmid cloning and sequence completion in these regions.) The fraction of genes with

Table 2. Gene density. Autosomes are divided into the genetically defined compartments of the left arm (L), the central cluster region (C), and the right arm (R). The percentage of genes with EST and database matches was determined only from manually inspected genes. Database matches to non-nematode proteins were determined with WUBLASTP ($P \leq 0.001$). Parentheses denote the number of low-scoring predictions thought to be pseudogenes.

Chromosome	Size (Mb)	Protein genes	Density (kb per gene)	tRNA genes	Coding (%)	EST match (%)	Database match (%)
I							
L	3.29	649	5.06	7(2)	21.59	57.0	53.9
C	5.59	1,171	4.77	34(4)	31.65	52.9	52.1
R	4.98	983	5.06	33(2)	25.00	43.4	40.8
II							
L	3.83	1,049	3.65	29(13)	29.00	22.7	26.9
C	7.93	1,719	4.61	38(6)	29.68	49.7	49.8
R	2.96	491	6.03	16(5)	19.89	43.5	39.9
III							
L	3.30	612	5.4	31(14)	20.60	44.2	42.1
C	4.98	1,100	4.52	42(0)	32.21	53.5	53.5
R	4.49	796	5.66	21(3)	23.91	53.1	50.2
IV							
L	5.44	1,050	5.17	38(16)	20.87	39.9	39.7
C	6.51	1,422	4.58	20(3)	29.69	45.7	50.3
R	4.19	622	6.73	26(2)	16.5	36.6	40.7
V							
L	6.19	1,491	4.15	17(4)	27.00	22.0	33.0
C	6.84	1,573	4.34	37(0)	29.40	32.2	43.8
R	7.79	1,018	4.36	152(94)	25.50	19.5	28.8
X	17.22	2,631	6.54	362(33)	19.8	40.9	43.34
Total	95.53	19,141		877(198)			

similarities to organisms other than nematodes tends to be lower on the arms, as does the fraction of genes with EST matches. The difference between autosome arms and central regions is even more obvious in the number of EST matches (*46*). The local gene clusters described above also appear to be more abundant on the arms.

These features, together with the fact that meiotic recombination is much higher on the autosome arms, suggested that the DNA on the arms might be evolving more rapidly than in the central regions of the autosomes. If so, one might expect that the conserved set of eukaryotic genes shared by yeast and *C. elegans* would be largely excluded from the arms. To test this, we identified 1517 proteins in *C. elegans* that are highly similar to yeast genes and plotted their location along the length of the chromosomes (Fig. 3). For four of the five autosomes, the differences in the distribution of core genes are quite striking, with surprisingly sharp boundaries evident. These boundaries appear close to the boundaries in the genetic map that separate regions of high and low rates of recombination (*47*).

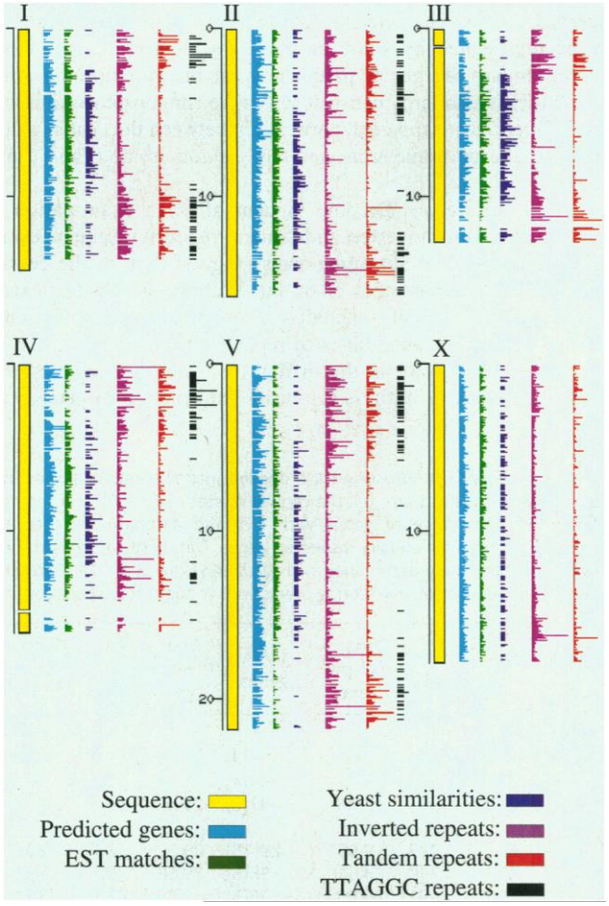

Fig. 3. Distributions of predicted genes; EST matches; yeast protein similarities; and inverted, tandem, and TTAGGC repeats along each chromosome. Gene density varies little along and among the autosomes. On the X chromosome, genes appear at a lower density and are more evenly distributed. In contrast, the frequency of EST matches varies according to their position along the autosomes, indicating a clustering of highly expressed genes. The chromosomal locations of these clusters correlate well with the chromosomal locations of gene products that exhibit significant similarities to yeast proteins (*P* value of 10^{-9}). For the autosomes, repeat density varies dramatically with chromosomal position and is highest on the arms. The density of inverted and tandem repeats on the X chromosome is more uniform, but similar to the autosomes, TTAGGC repeats tend to be located on the arms. Supplemental information regarding the analysis can be found at www.sciencemag.org/feature/data/c-elegans.shl for a general overview.

Conclusions

There are several reasons for completely sequencing a genome. The first and most simple reason is that it provides a basis for the discovery of all the genes. Despite the power of cDNA analysis and its enormous value in interpreting genome sequence, it is now generally recognized that a direct look at the genome is needed to complete the inventory of genes. Second, the sequence shows the long-range relationships between genes and provides the structural and control elements that must lie among them. Third, it provides a set of tools for future experimentation, where any sequence may be valuable and completeness is the key. Fourth, sequencing provides an index to draw in and organize all genetic information about the organism. Fifth, and most important over time, is that the whole is an archive for the future, containing all the genetic information required to make the organism (the greater part of which is not yet understood). As a resource, the sequence will be used indefinitely not only by *C. elegans* biologists, but also by other researchers for the comparison with and the interpretation of other genomes, including the human genome.

As was already known, the genome of a multicellular organism is very different from that of a microbial organism (and even different from that of a eukaryote such as yeast). It is predominantly noncoding, with genes extended (sometimes over many kilobases) by introns. Rather than acting primarily as the source for a set of protein sequences, the genomic sequence itself remains the primary focus of annotation. There are two reasons for this. First, much information about biological function is located in noncoding sequences; second, current methods of gene identification, both experimental and computational, are not yet accurate and complete enough to provide a definitive set of protein sequences.

If we began again now, would we employ the same approach? Almost certainly (*48*). The clone-based physical map was a critical factor in organizing the project between the two sites. The clones of the map have also been valuable reagents for the research community and continue to be so; the discrete assemblies of cosmids and YACs have been essential to disentangling extensive repeats in many areas. For the numerous small areas that are underrepresented in shotgun assemblies, rare subclones can be readily recovered from the cosmid and YAC subclone libraries.

There are two minor changes that we would make in the sequencing approach. We would add longer insert bacterial clones (for example, bacterial artificial chromosomes) to the map, fingerprinting them in the same manner as cosmids (*48*). Second, we would begin YAC sequencing earlier in the project. That we did not do so on this occasion was for historical reasons [in particular, the availability of the yeast genome sequence (see above)].

How important has the worm project been to the Human Genome Project? Through feedback from many sources, we gather that it has been influential in showing what can be done. Certainly, it is remarkable to look back to 1992, when a paper concerning just three cosmids was published as an important milestone (*10*). Undoubtedly, the worm project has contributed to technology and software development; it is not a unique test-bed, but along with the other genome projects, it has explored ways of increasing scale and efficiency.

Where is the finish line? This publication marks more of a beginning than an end and is another milestone in an ongoing process of the analysis of *C. elegans* biology. It is not very meaningful at any particular point to call genomes of this size finished, because of the inevitable imperfections that will only gradually be resolved. This is true no matter what method of sequencing is adopted. The important thing is not a declaration of completion, but rather the provision of the best possible tools to the users at every stage and a commitment to

maintenance and improvement, through interaction with the user community, as long as that is needed.

References and Notes

1. M. S. Chee *et al.*, in *Cytomegaloviruses*, vol. 154 of *Current Topics in Microbiology and Immunology*, J. K. McDougall, Ed. (Springer-Verlag, Berlin, 1990), pp. 125–169; R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996). F. R. Blattner *et al.*, *ibid.* **277**, 1453 (1997); S. T. Cole *et al.*, *Nature* **393**, 537 (1998).
2. H. W. Mewes *et al.*, *Nature* **387** (suppl.), 7 (1997); A. Goffeau *et al.*, *Science* **274**, 546 (1996).
3. A. R. Coulson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7821 (1986).
4. A. Coulson *et al.*, *Bioessays* **13**, 413 (1991); A. Coulson *et al.*, *Nature* **335**, 184 (1988). The current status of the *C. elegans* physical map is accessible on the World Wide Web (*20, 21*).
5. The investigations contributing to the *C. elegans* genome project are too numerous to cite. Two early representative publications are I. Greenwald, A. Coulson, J. Sulston, *Nucleic Acids Res.* **15**, 2295 (1987) and S. Ward *et al.*, *J. Mol. Biol.* **199**, 1 (1988).
6. R. Waterston *et al.*, *Nature Genet.* **1**, 114 (1992); W. R. McCombie *et al.*, *ibid.*, p. 124.
7. Y. Kohara, *PNE Protein Nucleic Acid Enzyme* **41**, 715 (1996).
8. R. Okimoto, J. L. Macfarlane, D. O. Clary, D. R. Wohlstenholme, *Genetics* **130**, 471 (1992).
9. D. T. Burke, G. F. Carle, M. V. Olson, *Science* **236**, 806 (1987).
10. J. Sulston *et al.*, *Nature* **356**, 37 (1992).
11. R. Wilson *et al.*, *ibid.* **368**, 32 (1994).
12. M. Vaudin *et al.*, *Nucleic Acids Res.* **23**, 670 (1995).
13. For details of the sequencing process, see (*49*). The process began with the purification of DNA from selected clones of the tiling path. The DNA was sheared mechanically, and after size selection, the resulting fragments were subcloned into M13 or plasmid vectors. Random subclones were selected for sequence generation (the shotgun sequencing approach). Generally, 900 sequence reads per 40 kb of genomic DNA were generated with fluorescent dye–labeled primers or terminators. Bases were determined with PHRED (*50*). An assembly of these random sequences that was generated with PHRAP (*51*) typically resulted in two to eight contigs. Gap closure and resolution of sequence ambiguities were achieved during finishing [using the editing packages GAP (*52*) and CONSED (*53*) and the collection of additional data] through longer reads, directed sequencing reactions using custom oligonucleotide primers on chosen templates, or additional chemistries as required. High-quality finished sequence was analyzed through the use of a suite of programs (including BLAST and GENEFINDER), and the results were stored in ACEDB and submitted to GenBank. Unfinished and finished sequence data were available to investigators by file transfer protocol (ftp) from both sequencing sites (*20, 21*).
14. C. R. Heiner, K. L. Hunkapiller, S. M. Chen, *Genome Res.* **8**, 557 (1998); L. G. Lee *et al.*, *Nucleic Acids Res.* **20**, 2471 (1992); J. D. Parsons, *Comput. Appl. Biosci.* **11**, 615 (1995).
15. A. A. McMurray, J. E. Sulston, M. A. Quail, *Genome Res.* **8**, 562 (1998).
16. U. J. Kim, H. Shizuya, P. J. de Jong, *Nucleic Acids Res.* **20**, 1083 (1992).
17. S. Cheng, C. Fockler, W. M. Barnes, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5695 (1994).
18. A clean separation of the YAC DNA from the host chromosomal DNA sometimes required the use of yeast strains in which specific yeast chromosomes are altered in size to provide a window around the YAC that is free of the native chromosomes. L. Hamer, M. Johnston, E. D. Green, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 11706 (1995).
19. S. E. Devine, S. L. Chissoe, Y. Eby, *Genome Res.* **7**, 551 (1997).
20. Available at www.sanger.ac.uk.
21. Available at genome.wustl.edu/gsc/gschmpg.html.
22. C. Wicky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8983 (1996).
23. Every region must be sequenced either on each strand or with dye primer and dye terminator chemistry, which extensive comparisons have shown to be at least as reliable as double stranding in revealing and correcting compressions and other base-calling errors. All regions must be represented by reads from two or more independent subclones or from PCR products across the region. If subcloned PCR products are used for a region, three independent clones must be sequenced. Rare exceptions to the general rules of double stranding or alternative chemistry were permitted on the basis of the following. For regions of <50 bases where, despite valid efforts, a finisher is unable to achieve double stranding or double chemistry, the sequence may be submitted (provided the sequence is of high quality and both the finisher and his or her supervisor see no ambiguous bases). When editing, in XGAP, all sequence data must be resolved at the 75% consensus level, either by the collection of additional data or by the editing of poorly called traces. In CONSED, any consensus base with a quality <25% must be manually reviewed to determine if the available data are sufficient to unambiguously support the derived contig sequence. If not, additional data are collected.
24. Each finished sequence is submitted to a series of quality control tests, including verification that all of the finishing rules (*23*) have been followed and a careful verification that the assembly is consistent with all restriction digest information. In addition, every finished sequence undergoes an automatic process of base calling and reassembly with different algorithms than those that were used for the initial assembly and comparison of the resultant consensus by a banded Smith-Waterman analysis [CROSSMATCH (*51*)] against the sequence that was obtained by the finisher. Any discrepancies in assembly or sequence, along with any regions failing to meet finishing criteria, are manually reviewed, and new data are collected as necessary. Only when all discrepancies are accounted for is the sequence passed on for annotation. In turn, if annotation flags any suspicious regions, these are again passed back to the finisher for resolution, either through additional data collection or editing.
25. P. Green and L. Hillier, unpublished software.
26. G. A. Fichant and C. Burks, *J. Mol. Biol.* **220**, 659 (1991); T. M. Lowe and S. Eddy, *Nucleic Acids Res.* **25**, 955 (1997).
27. S. F. Altschul, W. Gish, W. Miller, *J. Mol. Biol.* **215**, 403 (1990); W. Gish, WU-BLAST unpublished software.
28. E. L. L. Sonnhammer and R. Durbin, in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, R. Altman, D. Brutlag, P. Karp, R. Lathrop, D. Searls, Eds. (AAAI Press, Menlo Park, CA, 1994), pp. 363–368.
29. R. Mott, *Comput. Appl. Biosci.* **13**, 477 (1997).
30. E. L. Sonnhammer, S. R. Eddy, E. Birney, *Nucleic Acids Res.* **26**, 320 (1998); S. R. Eddy *Curr. Opin. Struct. Biol.* **6**, 361 (1996).
31. We identified local tandem and inverted repeats with the programs QUICKTANDEM, TANDEM, and INVERTED (*20*), which search for repeats within 1-kb intervals along the genomic sequence. An index of repeat families used by the project is available at www.sanger.ac.uk/Projects/C_elegans/repeats/.
32. R. Durbin and J. Thierry-Mieg, unpublished software. Documentation, code, and data are available from anonymous ftp servers at lirmm.lirmm.fr/pub/acedb/, ftp.sanger.ac.uk/pub/acedb/, and ncbi.nlm.nih.gov/repository/acedb/.
33. In *C. elegans*, two or more genes can be transcribed from the same promoter, with one gene separated by no more than a few hundred nucleotides from another. In genes undergoing transsplicing, the 5′ exon begins with a splice acceptor sequence, making this 5′ exon more difficult to distinguish from internal exons. This combination of factors may result in two genes being merged into one [T. Blumenthal, *Trends Genet.* **11**, 132 (1995)].
34. We have identified 182 genes possessing alternative splice variants, which are predominately from EST data. Of these, 67 genes produce proteins that differ at their amino termini, 57 genes produce proteins that differ at the carboxyl end, and 59 genes produce proteins that display an internal variation. Of the internal variations, seven genes showed complete exon skipping. Thirty-one genes were found where the 5′ end of an exon had changed, 21 of which resulted in a difference of three or fewer codons. In contrast, of the 24 alternative transcripts that changed the 3′ end of an exon, only 4 resulted in a change of three or fewer codons.
35. Available at www.sanger.ac.uk/Projects/C_elegans/Science98/.
36. R. K. Herman, in *The Nematode* Caenorhabditis elegans, W. B. Wood, Ed. (Cold Spring Harbor Laboratory Press, Plainview, NY, 1988), pp. 17–45; R. Waterston and J. Sulston, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10836 (1995).
37. These results were obtained with WU-BLAST (version 2.0a13MP), using default parameters and a threshold P value of 10^{-3}. P. Green *et al.*, *Science* **259**, 1711 (1993).
38. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
39. E. L. Sonnhammer and R. Durbin, *Genomics* **46**, 200 (1997).
40. GENEFINDER systematically uses statistical criteria [primarily log likelihood ratios (LLRs)] to attempt to identify likely genes within a region of genomic sequence. Candidate genes are evaluated on the basis of "scores" that reflect their splice site, translation start site, coding potential LLRs, and intron sizes. These scores are normalized by reference to the distribution of combined scores in a simulated sequence as follows: If a given combined score occurs, on average, once in every 10^s nucleotides in simulated DNA, then the corresponding normalized score is set to s. (For example, exons with a normalized score of 5.0 or greater will be found only once in every 100 kb of simulated DNA. With the current reference simulated sequence, which is 1 Mb in length, 6.0 is the maximum normalized score that can occur.) A dynamic programming algorithm is then used to find the set of nonoverlapping candidate genes (on a given strand) that has the highest total score (among all such sets). About 85% of experimentally verified "exon ORFs" (open reading frames containing true exons) in *C. elegans* genes in GenBank have normalized scores above 5.0 (and many of the remaining 15% are initial or terminal exons, which have a single splice site). The fraction of exons with scores >5.0 may be lower for all *C. elegans* genes because of the bias toward highly expressed genes (which often have very high coding segment scores) in the experimentally verified set. However, even for genes in the current verified set that are expressed at moderate to low levels, a majority of exon ORF scores exceed 5.0; this score should be an effective criterion for identifying at least part of most genes. In theory, high-scoring ORFs could arise in other ways. For example, intergenic or intronic regions having abnormal nucleotide composition might appear to have coding segments and occasionally, by chance, may have high-scoring splice sites. So far, there seem to be relatively few such regions in the *C. elegans* genomic sequence. These regions may account for the anomalous orphan exons that we occasionally find. In addition, there are examples where these GENEFINDER-predicted genes fall into clear gene families that are nematode-specific or have only very distant similarity outside the nematodes, for example, chemoreceptor genes (*54*).
41. Pfam is a collection of protein family alignments that were constructed semiautomatically with hidden Markov models within the HMMER package. The collagen and seven transmembrane chemoreceptor data were obtained with unpublished hidden Markov models. The number of seven transmembrane chemoreceptor genes is lower than that found by Robertson (*54*), which could be due to pseudogenes.
42. Putative tRNA pseudogenes are identified by the search program tRNAscan-SE as sequences that are significantly related to a tRNA sequence consensus but do not appear to be likely to adopt a tRNA's canonical secondary structure (*26*). Many higher eukaryotic genomes have mobile, tRNA-derived short interspersed nuclear elements (SINEs). However, because they are few in number, the nematode tRNA

pseudogenes seem more likely to have arisen by some rare event rather than by the extensive mobility that characterizes mobile SINEs [G. R. Daniels and P. L. Deininger, *Nature* **317**, 819 (1985)].
43. A. F. Smit, *Curr. Opin. Genet. Dev* . **6**, 743 (1996).
44. R. F. Ketting, S. E. J. Fischer, R. H. Plasterk, *Nucleic Acids Res.* **25**, 4041 (1997).
45. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995); B. Dujon *et al.*, *Nature* **369**, 371 (1994).
46. The abundance of *C. elegans* ESTs does not directly reflect expression levels, because they are derived from cDNAs in which more abundantly expressed genes were partially selected against (*6*, *7*).
47. T. M. Barnes, Y. Kohara, A. Coulson, *Genetics* **141**, 159 (1995).
48. This approach is also being used for the human genome (Sanger Centre, Washington University Genome Sequencing Center, *Genome Res.*, in press).
49. For methodological details, see (*20*) or (*21*). For biochemical procedures, see R. K. Wilson and E. R. Mardis, in *Genome Analysis: A Laboratory Manual*, B. Birren, E. D. Green, S. Klapholz, R. M. Myers, J. Roskams, Eds. (Cold Spring Harbor Laboratory Press, Plainview, NY, 1997), vol. 1, pp. 397–454. For software packages, see (*20*) or (*21*) and S. Dear *et al.*, *Genome Res.* **8**, 260 (1998); M. Wendl *et al.*, *ibid.*, p. 975; J. D. Parsons, *Comput. Appl. Biosci.* **11**, 615 (1995); and M. Cooper *et al.*, *Genome Res.* **6**, 1110 (1996).
50. B. Ewing, L. Hillier, M. C. Wendl, *Genome Res.* **8**, 175 (1998); B. Ewing and P. Green, *ibid.*, p. 186.
51. P. Green, personal communication.
52. J. K. Bonfield, K. F. Smith, R. Staden, *Nucleic Acids Res.* **23**, 4992 (1995).
53. D. Gordon, C. Abajian, P. Green, *Genome Res.* **8**, 195 (1998).
54. H. M. Robertson, *Genome Res.* **8**, 449 (1998).
55. This work has been supported by grants from the U.S. National Human Genome Research Institute and the UK MRC. We would also like to thank the many members of the *C. elegans* community who have shared data and provided encouragement in the course of this project.

Zinc Fingers in *Caenorhabditis elegans*: Finding Families and Probing Pathways

Neil D. Clarke and Jeremy M. Berg

REVIEW

More than 3 percent of the protein sequences inferred from the *Caenorhabditis elegans* genome contain sequence motifs characteristic of zinc-binding structural domains, and of these more than half are believed to be sequence-specific DNA-binding proteins. The distribution of these zinc-binding domains among the genomes of various organisms offers insights into the role of zinc-binding proteins in evolution. In addition, the complete genome sequence of *C. elegans* provides an opportunity to analyze, and perhaps predict, pathways of transcriptional regulation.

Less than 15 years ago, it was suggested that repeated sequences found in transcription factor IIIA (TFIIIA) of *Xenopus* might fold into structural domains stabilized by the binding of zinc to conserved cysteine and histidine residues (*1*–*3*). Klug and co-workers further noted that "it would not be surprising if the same 30 residue units were found to occur in varying numbers in other related gene control proteins" (*1*). This proposal proved remarkably prescient: *Caenorhabditis elegans*, for example, turns out to have more than 100 such proteins, and the number of domains per protein varies from one to perhaps as many as fourteen. Unanticipated at the time, though, was the fact that the zinc-binding motif found in TFIIIA is just one of many small zinc-binding domains, a number of which are involved in gene regulation. The properties of a few of these domains have been summarized recently (*4*).

Eukaryotes contain a much greater number of proteins with well-characterized zinc-binding motifs than do bacterial and archaeal organisms (Table 1). The complete genome of *Caenorhabditis elegans* (a metazoan), in conjunction with that of *Saccharomyces cerevisiae* (a yeast), presents a special opportunity to examine the range and diversity of these gene families in eukaryotes. Furthermore, because some of these zinc-binding motifs are sequence-specific DNA-binding proteins, the availability of nearly complete sequence information also permits a preliminary analysis of the distribution of potential binding sites within the entire genome. Such analyses may prove to be of value in deducing development control pathways and in more fully defining the characteristics of eukaryotic promoters.

Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

The Cys_2His_2 Family

The zinc-stabilized domains of TFIIIA are known as "zinc fingers" or Cys_2His_2 domains. The consensus sequence for this family is (Phe, Tyr)-X-Cys-X_{2-4}-Cys-X_3-Phe-X_5-Leu-X_2-His-X_{3-5}-His (*5*–*7*). In both *C. elegans* and the yeast *S. cerevisiae*, roughly 0.7% of all proteins contain one or more Cys_2His_2 zinc finger domains (Table 1). However, the distribution of these domains within proteins is rather different in the two organisms. In yeast, the majority of zinc finger proteins contain exactly two domains, and only a few (~10%) have more than two. In contrast, there are more zinc finger proteins in *C. elegans* that have three or more Cys_2His_2 domains than there are proteins that have exactly two (Fig. 1) (*8*). On the basis of the sequences of mammalian and *Drosophila* zinc finger proteins, it appears that the distribution of Cys_2His_2 domains among *C. elegans* proteins is typical of multicellular organisms.

The GATA, LIM, and Hormone Receptor Families: Implications for Metazoan Evolution

The GATA domain, the LIM domain, and the DNA-binding domains from nuclear hormone receptors each include a four-cysteine zinc-binding domain that can be clustered into the same structural superfamily, and it is possible that they share a common evolutionary origin (Fig. 2) (*9*, *10*). In addition to the Cys_4 superfamily domain, LIM domains contain a similar LIM-specific Cys_2HisCys zinc motif, whereas the hormone receptors have a second and distinct Cys_4 domain. GATA proteins frequently contain a pair of Cys_4 superfamily domains.

Normalized to the number of genes in their respective genomes, the number of GATA and LIM domain homologs is similar in *C. elegans* and *S. cerevisiae*. In striking contrast, the hormone receptor family is completely absent in yeast but is the largest single family of zinc-binding domains in *C. elegans*. In fact, with over 200 family members, the hormone receptors make up nearly 1.5% of the entire coding sequence of *C. elegans*. The differences in the distribution of nuclear hormone receptors in *C. elegans* and *S. cerevisiae* may be relevant to the evolution of multicellular animals. As has been noted before, the evolution of hormone receptors may have been a key event in the development of cell-cell communication and the origins of multicellularity in the metazoa (*11*).

The ligand-binding domains of the hormone receptors have diverged considerably more than the DNA-binding domains. Applying the same criterion for significance to both the DNA- and ligand-binding domains of the hormone receptor family, only about 10% of the open reading frames (ORFs) that have a DNA-binding domain