Caenorhabditis elegans is the first animal to have its genome completely sequenced. To mark this outstanding achievement, we have published a series of articles celebrating worm genetics. This series has examined the contributions that worm genetics has made to fundamental aspects of biology such as cell death, signal transduction and neurobiology. In this final article of the series, the *C. elegans* Genome Consortium reviews the genome project and examines some of the preliminary findings from the near-complete sequence data.

How the worm was won

the *C. elegans* genome sequencing project

The genome sequence of the free-living nematode *Caenorhabditis elegans* is nearly complete, with resolution of the final difficult regions expected over the next few months. This will represent the first genome of a multicellular organism to be sequenced to completion. The genome is approximately 97 Mb in total, and encodes more than 19 099 proteins, considerably more than expected before sequencing began. The sequencing project – a collaboration between the Genome Sequencing Center in St Louis and the Sanger Centre in Hinxton – has lasted eight years, with the majority of the sequence generated in the past four years. Analysis of the genome sequence is just beginning and represents an effort that will undoubtedly last more than another decade. However, some interesting findings are already apparent, indicating that the scope of the project, the approach taken, and the usefulness of having the genetic blueprint for this small organism have been well worth the effort.

The free-living nematode *Caenorhabditis elegans* has been at the forefront of biological discovery for the past two decades. The 'worm', initially proposed as a model organism by Sydney Brenner in 1965, was once viewed as an uninteresting, featureless tube of cells by some early critics. Since then, it has provided a wealth of knowledge in cell biology, development, neurobiology and genetics, and serves as an excellent model system for the study of higher eukaryotes. In the 1970s and 1980s, the complete cell lineage of the worm from fertilized egg to adult was determined by microscopy[1]. Using electron microscopy and serial sectioning, the entire nervous system was reconstructed[2]. Resources such as these, combined with the genetic and genomic data that have been generated during the past decade, have made the worm a powerful tool for the discovery and functional characterization of eukaryotic genes. In this article, we review the beginnings of the *C. elegans* genome project and the trials and successes we have had along the way. We also present some of the preliminary findings from the nearly complete genome sequence and discuss some of the implications of these findings as they relate to the biology of the worm and gene discovery in genomes yet to be sequenced.

The 'worm'… was once viewed as an uninteresting, featureless tube of cells

Building the map

In the mid-1980s, Sulston, Coulson, Waterston and colleagues set out to generate a clone-based physical map of the *C. elegans* genome, then estimated at 100 Mb. Although not given much of a chance for success at the time by others in the field, the project had an immediate impact for the ever-growing community of *C. elegans* biological researchers. The map was initially based on cosmid clones using a fingerprinting approach devised by Sulston and Coulson[3], and later incorporated yeast artificial chromosome (YAC) clones to bridge the gaps between cosmid contigs. The YACs also provided coverage of the approximately 20% of the genome not represented in the cosmid libraries. By 1990, the physical map consisted of fewer than 20 contigs and was useful for rescue experiments that typically could pin down a phenotype of interest to a few kilobases of DNA (Refs 4, 5). Alignment of the existing genetic and physical maps into a genome map for *C. elegans* was greatly facilitated through the cooperation of the entire worm community.

By 1989, with a nearly complete physical map in hand, it became apparent that an effort to sequence the 100 Mb genome might be both feasible and desirable. This was not to be undertaken lightly, being significantly larger than any

The *C. elegans* Genome Consortium

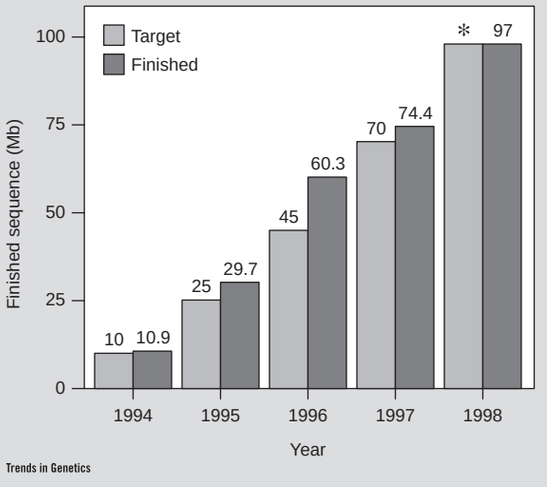
Washington University Genome Sequencing Center, St Louis, MO, USA and The Sanger Centre, Hinxton, Cambridgeshire, UK. (Complete author list available at http://genome.wustl.edu/TIGs and http://www.sanger.ac.uk/TIGs)

Richard K. Wilson (corresponding author)
rwilson@watson.wustl.edu

Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, Box 8501, St Louis, MO 63108, USA.

 PII: S0168-9525(98)01666-7

FIGURE 1. Sequencing project



Sequencing targets and progress in the *C. elegans* genome sequencing project. Asterisk indicates target completion date.

sequencing project ever attempted and nearly two orders of magnitude more expensive than the mapping effort. It was justified in part as a pilot for the Human Genome Project, and the influence of Jim Watson was a key factor in the decision to proceed. Joint funding from the NIH and MRC for a three-year pilot project was arranged, and an effort to sequence 3 Mb of the worm genome was initiated in 1990.

Because it was clearly recognized at the time that the majority of known worm genes were contained in the central regions of the five autosomes, and that these regions were well represented in cosmid clones, the genome sequencing effort initially focused on these gene-rich regions. Our general impression then was that the arms of the autosomes, with sparse cosmid coverage, contained a higher proportion of repetitive elements that could be sequenced later as technology and methods advanced. Armed with such notions, and with a good number of large cosmid contigs in hand, we picked a point near the center of chromosome III and set off in both directions – St Louis to the left and Cambridge to the right – on a journey to build the ultimate map of the worm's genome.

Armed with … a good number of large cosmid contigs in hand, we … set off in both directions – St Louis to the left and Cambridge to the right

The adventure begins

In 1990, only a few whole cosmid clones had ever been sequenced. Before that time, most 'large-scale' sequencing efforts focusing on non-viral genomes had started with purified restriction fragments and utilized either the shotgun strategy of Anderson[6], and of Bankier and Barrell[7], or a more directed approach. The fluorescent sequencing technology developed by Smith *et al.*[8] was still in its infancy and the associated chemistry was inflexible. When it was first proposed to sequence the entire genome of *C. elegans*, the approach that we intended to take was primer-directed sequencing or 'walking' using the cosmid clones as template for sequencing reactions. Automated oligonucleotide synthesis was reasonably inexpensive and robust and we were convinced that walking was the best approach at the time. Both laboratories (Washington University in St Louis and the Laboratory of Molecular Biology in Cambridge) purchased two automated sequencing machines: an Applied Biosystems Model 373 and a Pharmacia ALF. The 373 – an unproved version of the original 370 – had just been introduced the previous year and offered a capacity of 24 samples per run. The Pharmacia instrument, although limited to single-color, four-lane per sample sequencing, seemed to provide the only means of using custom sequencing primers, because dye-labeled terminators were not providing high quality data. With the Pharmacia chemistry, we would be able to synthesize dye-labeled sequencing primers easily on our oligonucleotide synthesizers.

While we worked on improvements to the chemistries for both sequencing platforms, we elected to sequence our first two cosmid clones using fairly standard radioisotopic methods. Very quickly into the project, we became disenchanted with our chosen strategy. The most significant challenges with a walking approach on cosmids were multiple priming events due to repetitive sequences and efficient preparation of sufficient template DNA. To address these problems, we instead decided to fragment the cosmids randomly and subclone the resulting smaller fragments into plasmid and M13 vectors. From this point, we evolved fairly rapidly into a more classic shotgun sequencing strategy, with the majority of sequence data generated from universal priming sites in the subcloning vectors. Performance of the automated sequencing instruments also played a major role in our strategy change. Dealing with autoradiographs, especially for the number of sequencing reactions we were performing, was not efficient. Dear and Staden[9] modified the assembly and contig editing program XBAP to allow direct access to the chromatographic traces produced by both of the automated sequencers. This provided a huge advantage over handling autoradiographs. Because we were able to sequence all of our subclones with a universal dye-labeled primer at a capacity of 24 samples per day per machine, the sequencers quickly replaced radioisotopic methods in both labs. Along with others in the field, we worked to improve fluorescent sequencing methods to the point where they were easily performed in large numbers[10,11]. Before long, both labs purchased additional ABI machines and the ramp up was on! Improvements to the dye-terminator chemistry[12] led to an eventual phasing out of the Pharmacia instrument in favor of four-color, single-lane sequencing with the custom primers that were still utilized for gap closure and the resolution of difficult regions.

Breaking the megabase barrier

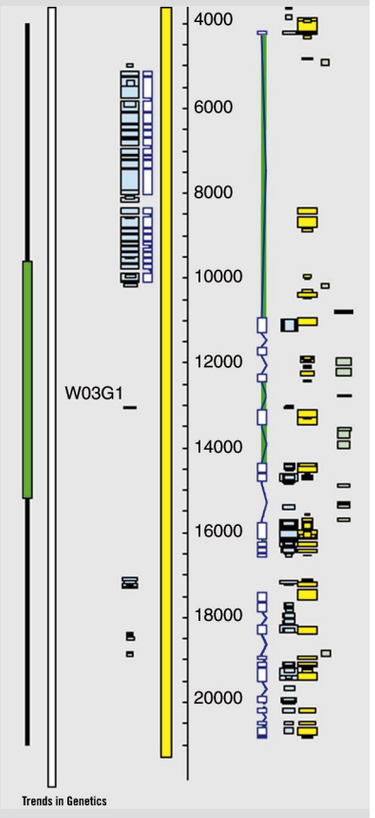
A key goal for the third year of the pilot project was to attain a throughput level of 1 Mb of finished sequence per year. We reasoned that if we could reach this level of sequencing throughput, the additional increases required to complete the *C. elegans* genome in a reasonable time period were possible. Although we had fallen a bit short of our first two years' goals, we were confident that we would more than make up for it during the third year as the methods and technology had become fairly robust. In May 1993, the two groups celebrated the milestone of 1 Mb of finished *C. elegans* genomic sequence. By August 1993, the total had increased to just over 2 Mb (Ref 13). By December 1994, over 10 Mb of the *C. elegans* genome had

been finished. Our success in scaling up the sequencing relied on the implementation of high-throughput devices and semi-automated methods for DNA purification and sequencing reactions, on problem solving or 'finishing', and on software developments that made the processing, analysis and editing of thousands of data files per day a manageable task. Indeed, many of the software tools developed in the *C. elegans* project – ACeDB, PHRED and PHRAP to name a few – have become key components in the current approach to sequencing the human genome.

The most important components of our increasing scale were organization and planning. In both laboratories, we devised an infrastructure that separated the data production and sequence finishing tasks into two distinct, but coupled activities. The production groups focused on template preparation, sequencing reactions and loading of sequencing gels. Once the raw sequence data had been generated, it was processed, assembled and passed off to the finishing groups for editing, gap closure and problem resolution. Other small, distinct groups in each lab focused on library construction, data analysis and annotation, and technology development. In addition, multiple production and finishing groups at each laboratory provided on-site back-up should problems arise, as well as built-in competition to keep each group effectively motivated.

The psychology between the USA and UK sites has always been one of 'together, we can do more', rather than 'one against the other'. Working together made each laboratory all the better and some of the early joint meetings were very productive brainstorming sessions in which methods and plans rapidly came together. It was in these meetings – alternately held in the USA and UK – that the long-term strategic plan for the project was devised. At the joint lab meeting in August 1994, with five years of funding in place for each laboratory, we sketched out the plan for finishing the *C. elegans* genome by the end of 1998. The targets that we set at that meeting (Fig. 1) were daunting at the time, but important to our success as they gave us a clear idea of what we needed to accomplish in each of the next few years.

FIGURE 2. Sequence annotation



Common annotation features are illustrated in an ACeDB sequence display of a portion of the cosmid W03G1. The vertical scale bar, in base pairs, divides the top-strand features to the right of the scale bar from bottom-strand features to the left. Strand-independent features are only displayed on the right. The 5′ position on the top strand is at the top of the figure. GENEFINDER predicted genes, beginning with an initiator-methionine codon and ending with a stop codon, are represented by open blue rectangles (exons) connected by solid blue lines (introns). Introns confirmed by ESTs are indicated by wider straight lines. Solid blue boxes represent BLASTX protein similarities. Solid yellow rectangles represent BLASTN similarities to *C. elegans* ESTs. The width of the blue and yellow rectangles corresponds to the level of similarity: the wider the rectangle, the more similar the two sequences. Local inverted and tandem repeats are depicted with cyan boxes. EST similarities and repeats are displayed as strand-independent features and so are always to the right of the scale bar. Note that the two genes on the bottom strand (one confirmed, in part, by ESTs) fall within the confirmed intron of the gene on the top strand.

Of new buildings, other genomes and continued progress

Our initial sequencing success led to big changes at both sites. We had the funding and a plan to complete the *C. elegans* genome sequence. The next necessary ingredient was room to expand. In 1993, the St Louis group moved into a newly purchased building on the Medical School campus, which provided space for considerable future expansion of operations. In the same year, the Cambridge group relocated to a small estate in the nearby village of Hinxton. The estate was renamed the Sanger Centre in honor of the inventor of the dideoxy method of DNA sequencing. For both laboratories, the move to new and larger facilities was accompanied by additional funding ear-marked for genomes other than that of the worm. Both groups made a commitment to contribute to the international effort to sequence the genome of the yeast *S. cerevisiae* and to begin exploring the adaptation of the methods we were currently using to regional sequencing of the human genome.

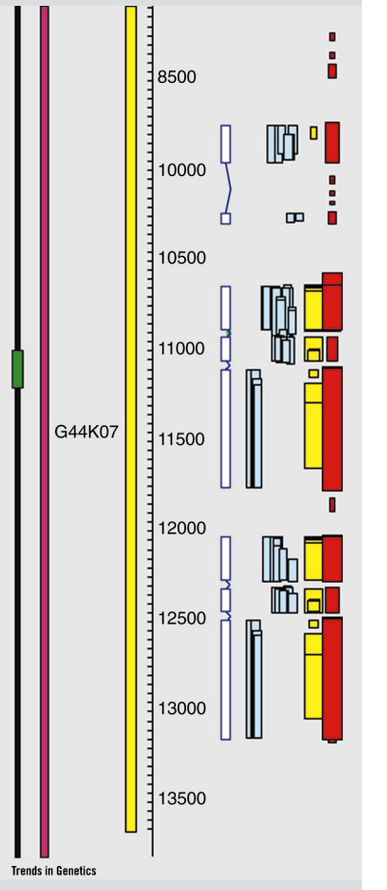
Over the next few years, progress continued at a steady pace. As previously mentioned, we initially focused on cosmid clones, leaving the regions covered by YACs for later. We passed the 50 Mb mark in August of 1996, several months ahead of schedule. At this point, we began to implement a closure strategy for the 20% of the genome not contained in cosmid clones. For gaps in the central regions, we used either long-range PCR or probed a fosmid library in search of a bridging clone. Each of these methods was useful for about a third of the gaps between cosmids. For the remaining gaps in the central regions, and for regions of chromosomes contained only in YACs, the only choice was to use purified YAC DNA as the starting material for shotgun sequencing. We previously had experimented with shotgun sequencing from gel-purified YACs and found that the approach was feasible, although a significant amount of contaminating yeast sequence was unavoidable[14]. By 1996, with improvements in YAC DNA purification, contamination by host DNA in most cases is well below 5%. Because the complete genome sequence of yeast has been determined[15], host DNA sequences can be identified and removed computationally from the raw sequence data before assembly. To complete the worm genome, we calculated that each laboratory would need to sequence about 100 YAC clones. This has progressed well and we have essentially completed all of these final regions with the exception of several particularly nasty repetitive elements. These repeats will be resolved as we continue to develop effective methods for sequencing them to completion. These methods,

which will be of great value for sequencing difficult regions of other genomes, include short insert libraries[16], use of alternative enzymes[17], and 'overgo' probing[18] for specific subclones.

The final genome sequence of the worm is a composite from cosmids, fosmids, YACs and PCR products. The exact genome size is still approximate, mainly because of extensive tandem repeats that cannot be sequenced in their entirety; though most have been accurately sized by restriction digestion. Some tandem repeats in the larger YACs are of unknown size. There is really no point in trying to resolve these further, except possibly in population studies, because they are difficult to clone and likely to be variable. Telomeres were sequenced from plasmid clones provided by Wicky *et al.*[19] Of 12 chromosome ends, 9 have been linked to the outermost YACs on the physical map.

The accuracy of the final sequence product is better than one error per 10 000 bases. Overall accuracy was maintained by adhering to a set of stated criteria followed by the finishers, and by a final checking step using specialized software and visual inspection. None of this, however, overcomes errors in the cloning process. A comparison of discrepancies in the overlap regions of different clones has indicated a finite error rate associated with cloning. For example, the deletion of a large hairpin in one cosmid was detected only because it was present in the overlap with a neighboring clone. Similarly, restriction digestion detected a 400 bp region deleted in all M13 and PCR reads from one particular cosmid clone. The deleted region was subsequently recovered in a plasmid subclone. These instances are rare enough that undetected errors in the worm sequence are likely to be few. However, both of these examples underscore the need for redundant subcloning systems and an independent method for checking sequence assembly. The low error rate of the sequence is further confirmed by the infrequency of problem reports from users, who have been using much of the data for several years. So the product, while undoubtedly flawed in places, is highly functional.

FIGURE 3. Confirming gene annotation



Data from ESTs and the genome of a related nematode confirm gene prediction and annotation. Both are illustrated in an ACeDB sequence display of a portion of the *C. briggsae* fosmid clone. The three *C. briggsae* genes are Cb-mai-1 at the top, followed by Cb-gpd-2 and Cb-gpd-3 at the bottom. The corresponding *C. elegans* gene structures have been confirmed experimentally[33,34]. The display sequence features are described in the legend for Fig. 2. The red rectangular boxes represent BLASTN similarities to *C. elegans* sequence. The small regions of similarity (70–95% identity) found 5′ and 3′ of coding sequences, and in some introns, may represent conserved regulatory sequences.

The beginnings of genome analysis

While sequencing of the worm genome has essentially been completed, analysis and annotation will continue for years to come as more information and better sequence analysis tools become available. However, it is now possible to describe some interesting features of the *C. elegans* genome, based on our analysis of completed segments, along with our first glimpse of the entire sequence.

As individual cosmid clones and larger segments of the genome were finished, a series of computational analysis tools were employed to reveal possible protein and tRNA genes, similarities to ESTs and to other proteins, repeat families and local repeats. The results were entered in the genome database ACeDB, which contiguates overlapping sequence and provides a seamless view across clone boundaries. ACeDB places the sequence and its corresponding annotation in the context of the physical map, genetic markers and other relevant *C. elegans* biological data (Fig. 2), providing a powerful interface between genome sequence and the individual investigator. Browsing through the genome in ACeDB is an enlightening experience as one encounters various unique and interesting features of sequence organization.

Genes

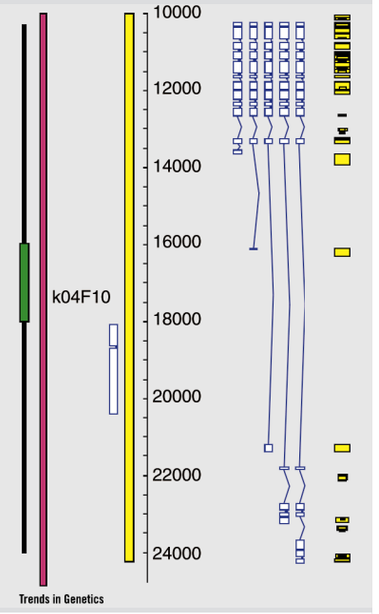
Analysis of 97 Mb of total *C. elegans* genome sequence revealed 19 099 predicted genes (16 260 of which have been manually reviewed) for an average density of one predicted gene per 5 kb. Each gene has an average of five introns and 27% of the genome resides in exons. The gene number is about three times the number found in yeast[20] and is about 1/5 to 1/3 the number predicted for human. As expected from earlier estimates based on much smaller amounts of genome sequence[13,21] this number is much higher than the number of essential genes estimated from classical genetic studies[22,23].

The interruption of the coding sequence by introns and the relatively low gene density make accurate gene prediction more challenging than in microbial genomes. GENEFINDER (P. Green, unpublished) was used to identify putative coding regions and to provide an initial overview of gene structure. To quantitate the accuracy of gene structure prediction, we compared intron–exon junctions confirmed by ESTs and cDNAs to those predicted by GENEFINDER. We found that 92% of predicted introns have an exact match to the experimentally confirmed ones and that 97% have an overlap. To refine the computer-generated gene structure predictions, expert annotators use any available EST and protein similarities, and genomic sequence data from the related nematode *Caenorhabditis briggsae* (Fig. 3). About 40% of predicted genes have a confirming EST match, but as ESTs match only a portion of the gene, only about 15% of the total coding sequence is presently confirmed. In a

number of cases ESTs have provided direct evidence of alternative splicing; these instances have been annotated in the sequence (Fig. 4).

Similarities to known proteins provide an initial glimpse into the possible function of many of the predicted genes. Approximately 42% of predicted protein products have cross-phylum matches, most of which provide putative functional information[24]. 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 less than the 70% observed for microbial genomes. This may reflect the smaller proportion of nematode genes devoted to core cellular functions[20], the comparative lack of knowledge of functions involved in building an animal, and the evolutionary divergence of nematodes from other animals so far studied extensively at the molecular level. Interestingly, genes encoding proteins with cross-phylum matches were more likely to have a matching EST (60%) than those without cross-phylum 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 suggest this.

FIGURE 4. Alternatively spliced genes



An ACeDB sequence display of a portion of the *C. elegans* cosmid K04F10 shows the multiple, alternatively spliced forms of the gene *bli-4*. From left are bli-4E, bli-4A, bli-4B, bli-4C and bli-4D. All the alternatively spliced forms are confirmed either by EST matches or experimentally[35]. The displayed sequence features are described in the legend for Fig. 1.

In addition to the protein-coding genes, the worm genome contains several hundred genes for noncoding RNAs. There are 659 widely dispersed transfer RNA genes, and at least 29 tRNA-derived pseudogenes. Curiously, 44% of the tRNA genes are found on the X chromosome, which contains only 20% of the total sequence. Several other noncoding RNA genes, such as those for spliceosomal RNAs, occur in dispersed multigene families. Several 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, indicating that the 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, and the 5S RNA genes occur in a tandem array on chromosome V, with array members separated by SL1 splice leader RNA genes.

… the most important components of our increasing scale were organization and planning … 'together, we can do more'

Repetitive sequences

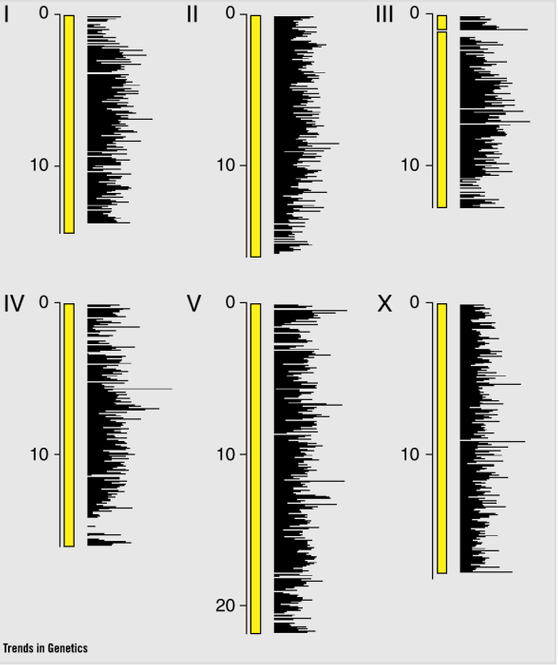
Much of the sequence that does not code for protein or RNA is undoubtedly involved in gene regulation or the replication, maintenance and movement of chromosomes. As in other higher eukaryotes, a significant fraction of the *C. elegans* genome is repetitive and can be classified as either local repeats (e.g. tandem, inverted and simple sequence repeats) or dispersed repeats.

Tandem repeats account for 2.7% of the genome and are found on average every 3.6 kb. Inverted repeats account for 3.6% of the genome and are found on average every 4.9 kb. Throughout the genome, these local repeats are distributed non-uniformly with respect to genes. For example, the 47% of the genome sequence predicted to be intergenic contains 49% of the tandem repeats. However, the 26% of the genome sequence predicted to be intronic contains 51% of the tandem repeats. Not surprisingly, only a small percentage of tandem repeats are found within the 27% of the genome encoding proteins. Conversely, the density of inverted repeats is higher in predicted intergenic regions: 45% of inverted repeats are located within genes, while 55% are located between them.

Although local repeat structures are often unique in the genome, other repeats are members of families. For example, the tandemly occurring hexamer repeat TTAGGC is seen at multiple sites internal to the chromosomes in addition to the telomeres. This repeat family is excluded from introns, while other repeat families show a slight bias toward introns. The reason for the biased distribution of these repeats is unclear. Further, some repeat families show a chromosome-specific bias in representation. Altogether we have recognized 38 dispersed repeat families. Most of these dispersed repeats are associated with transposition in some form[25], and include the previously described transposons of *C. elegans*.

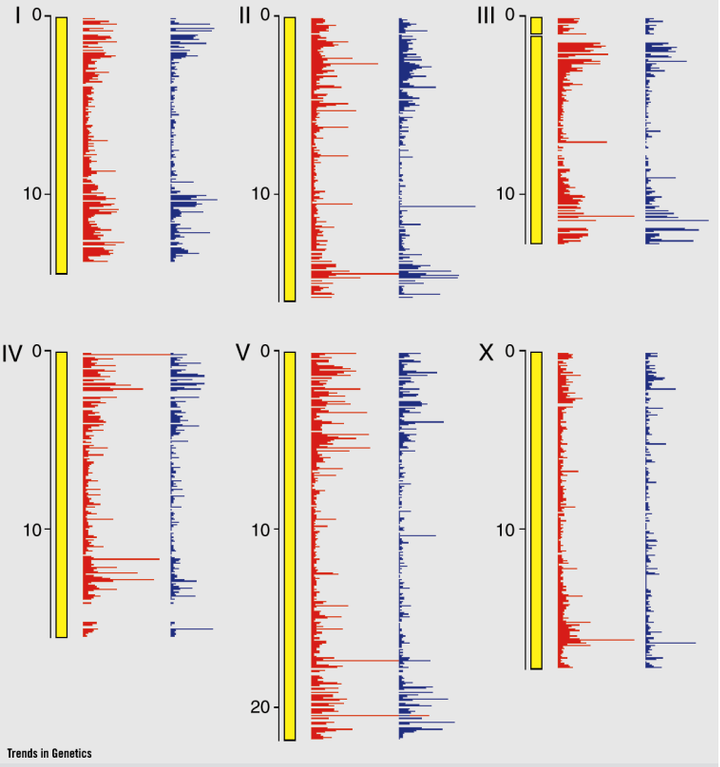
As well as multiple-copy repeat families, we have observed a significant number of simple duplications involving segments that range from hundreds of bases to tens of kilobases that have been copied in the genome. In one case, a segment of 108 kb containing six genes was duplicated tandemly with only ten nucleotide differences observed between the two copies. In another example, immediately adjacent to the telomere at the left end of chromosome IV, an inverted repeat of 23.5 kb was present, with only eight differences found between the two copies. There are many instances of smaller duplications, often separated by tens of kilobases or more that may contain coding sequence. This 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 98% sequence identity (annotated as C38C10.4 and F22B7.5). Based on matches to EST data, both genes are expressed.

FIGURE 5. Gene distribution



The distribution of predicted genes is plotted along each chromsome. The vertical yellow bars represent the clonal physical map of the genome (in Mb).

FIGURE 6. Tandem and inverted repeats



Distribution of local tandem and inverted repeats along each of the chromosomes. Inverted repeats are shown in red while tandem repeats are blue. Both kinds of repeats are more frequent on the arms of the autosomes than in the central gene-rich regions, while they appear more uniformly distributed on the X chromosome.

Chromosome organization

At first glance, the genome looks remarkably uniform: GC content is essentially constant across all chromosomes at 36%, unlike human chromosomes that have different isochores[26]. There are no localized centromeres as are found in most other metazoa. Instead the extensive, highly repetitive sequences that are involved in spindle attachment in other organisms may be represented by some of the many tandem repeats found scattered among the genes, particularly on the chromosome arms (see below). Gene density is also uniform across the chromosomes, although some differences are apparent, particularly between the centers of the autosomes, the autosome arms and the X chromosome (Fig. 5).

More striking differences become evident upon examination of other features. Both inverted and tandem repeat sequences are more frequent on the autosome arms (Fig. 6) than in the central regions or on the X chromosome. This abundance of repeats on the arms is likely the reason for the difficulties in cosmid cloning and sequence completion in these regions. The fraction of genes with cross-phylum similarities 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 looking at the number of EST matches (Fig. 7). Local clusters of genes 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 than elsewhere, suggest that the DNA on the autosome arms might be evolving more rapidly than in the central regions. If this were so, one might expect that the conserved core set of eukaryotic genes shared by yeast and the worm would be largely excluded from the arms. To test this, we identified 1517 genes in *C. elegans* that are highly similar to yeast genes and plotted their location along the length of the chromosomes (Fig. 8). For four of the five autosomes, the differences in distribution of the conserved genes are quite striking, with surprisingly sharp boundaries evident. These boundaries appear close to those seen in the genetic map demarking regions of high and low rates of recombination[27].

Conclusions

The beginnings of analysis of the *C. elegans* genome and the observations that we have introduced in this review provide a preliminary glimpse of the biology of metazoan development. There is much left to be uncovered and understood in the sequence. Of primary interest, all of the genes necessary to build a multicellular organism are now essentially in hand, although their exact boundaries, relationships and functional roles must be more precisely elucidated. The basis for a better understanding of how these genes evolved and are controlled is also now within our grasp. Further, the manner in which the genes are organized represents an additional topic of study. The context in which genes lie undoubtedly contributes to their expression and evolution, although additional genome sequences will be necessary to best understand this. Although EST and genome 'skimming' strategies provide useful data for gene discovery, a comprehensive understanding of the biology of an organism is possible only when the complete genome sequence has been determined. This resource is now available for the worm and it has already fundamentally changed the way individual investigators pose their

queries as to the workings of the animal. Several examples of this have already been described in this journal[28–31].

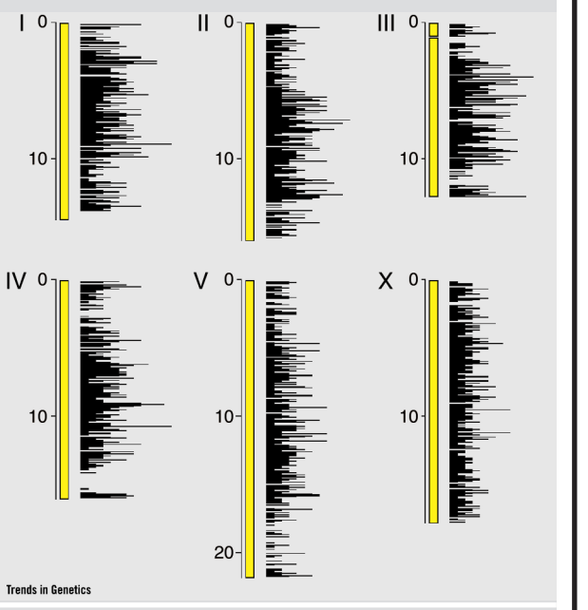
Even though the worm can stand on its own as a model system for biological and genetic experimentation, the *C. elegans* genome effort has always been a part of the Human Genome Project. Besides providing biological data that will facilitate the analysis and understanding of human genes, the practical and technological lessons learned from the worm will be directly applicable to the mapping and sequencing of other complex eukaryote genomes. With that in mind, what lessons has the worm provided as we now anticipate ramping up efforts to sequence the 3000 Mb human genome? As discussed above, organization of the sequencing process is critical. The process must be sufficiently flexible to allow rapid incorporation of new methods and technology. Ongoing incremental improvements in software, instrumentation, and chemistry are at least as important as revolutionary improvements. A close association between the production sequencing operation and the local developers of software and sequencing technology provides for the best possible understanding of current priorities and limitations. This serves to better reduce cost, effort and, ultimately, the time required to complete a project. At least for the next few years, although process automation and computational decision-making have their place in high-throughput DNA sequencing, there are still some tasks and decisions that are best left to humans. For all the discussions of sequencing strategy, the classic shotgun approach with basic Sanger chemistry has best adapted to the ever-changing technology. Furthermore, it is quite clear that a map-based approach is preferred to a 'map-as-you-go' strategy, especially when multiple groups are participating in the sequencing effort.

A high degree of accuracy and continuity are critical for finding the biology contained within the sequence. Similarity searches against existing sequence databases provide a good deal of the information used to annotate the sequence, but to effectively delineate gene boundaries with computational tools such as GENEFINDER, sequence continuity is a must. EST matches provide a good check on gene prediction, as does genomic sequence from other closely related organisms. In the *C. elegans* project, the availability of approximately 6 Mb of sequence from the genome of *Caenorhabditis briggsae* was useful in improving gene prediction. For the human genome, a similar resource will be gained by sequencing regions of the mouse genome.

Rapid public release of all sequencing and mapping data via the Internet has been a hallmark of the worm project. Investigators from the worm and other research communities have constantly monitored our web sites and are able to make use of the data, often when it is still in an unfinished form. In addition to providing access to both finished and unfinished sequence data, we also have offered BLAST servers that give an investigator the option of hunting for their target sequences or for a similarity to a protein of interest, even before our sequence has been annotated. Furthermore, the public often provides useful feedback on the data, which improves both sequence and annotation accuracy.

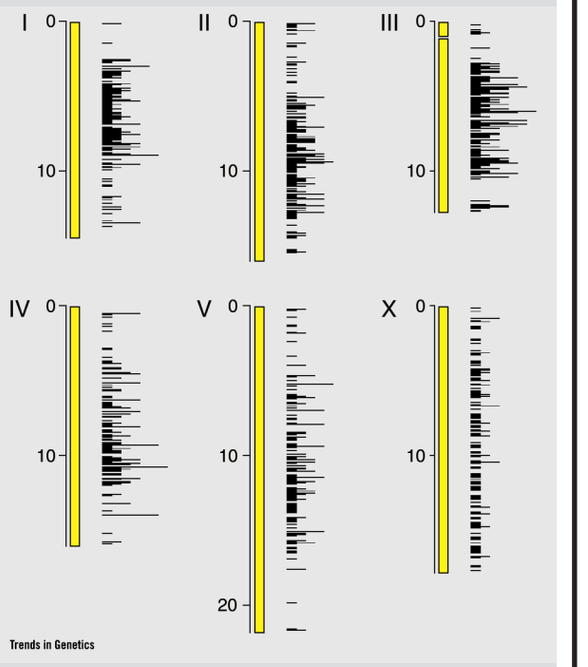
On-going analysis and annotation throughout the life of the project is critical even if the gene predictions are not completely accurate. This provides valuable clues to investigators who may have localized a gene to a specific region (e.g. 40 kb up to 1 Mb). They can browse through a set of gene predictions that may already have some functional information attached. The advantage of this is that the general community often has a difficult time running gene prediction programs and other analysis tools, so that if only the sequence were available, it would be of much less

FIGURE 7. Genes with EST matches



Distribution of predicted genes with EST matches is plotted along each of the chromosomes. The frequency of EST matches indicates clustering of highly expressed genes in the central regions of the autosomes. More uniform distribution is observed on the X chromosome.

FIGURE 8. Comparing *C. elegans* with *S. cerevisiae*



Distribution along each chromosome of the genes that are conserved between *S. cerevisiae* and *C. elegans*. These genes are clustered and coincide with the locations of genes with EST matches.

use. In addition, providing gene predictions on the preliminary data ensures that the sequence is entered into the public protein database. This is important because many investigators typically search only the protein database, rather than the preliminary worm data. Thus, the sequence data is generally more accessible since investigators from all research communities can find matches to *C. elegans* genes.

As was mentioned previously, when the US and UK groups sequencing the worm met in Hinxton in 1994 at a joint lab meeting, we developed the plan for completing the *C. elegans* genome sequence by the end of 1998. It was also during that meeting that we realized that the methods, technology, software and infrastructure that we had developed for the worm could be utilized for systematic sequencing of the human genome. This realization led directly to conversations with both the Wellcome Trust and the NHGRI that resulted in the accelerated program for sequencing human genomic DNA. Now, four years later, the two laboratories combined have contributed over 100 Mb of finished human genomic sequence – about 3% of the human genome – to the public databases[32]. Interestingly, we published our first major paper on the *C. elegans* sequencing project at a similar milestone: with just over 2% of the genome sequenced. As we put the finishing touches on the *C. elegans* genome and fully turn our attention to the genome of *H. sapiens*, it is clear that the worm has led the way.

Acknowledgements

This article is dedicated to the memory of Mary Berks – friend and colleague. The authors wish to thank all members of the consortium, past and present, for their contributions and efforts, and to the members of the *C. elegans* community of researchers for their support, encouragement and help in making the genome project both a reality and a success. This work was supported by funding from the NIH USPHS and the UK MRC.

References

1 Sulston, J. (1988) Cell Lineage, in *The Nematode Caenorhabditis elegans* (Wood, W.B. ed.), pp. 123–155, CSHL Press
2 Chalfie, M. and White, J. (1988) The Nervous System, in *The Nematode Caenorhabditis elegans* (Wood, W.B. ed.), pp. 337–391, CSHL Press
3 Coulson, A. and Sulston, J. (1988) Genome mapping by restriction fingerprinting, in *Genome Analysis: A Practical Approach* (Davies, K., ed.), pp. 19–39, IRL Press
4 Coulson, A.R. *et al.* (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A.* 83, 7821–7825
5 Coulson, A. *et al.* (1988) Genome linking with yeast artificial chromosomes. *Nature* 335, 184–186
6 Anderson, S. (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* 9, 3015–3027
7 Bankier, A.T. and Barrell, B.G. (1983) Shotgun DNA sequencing. *Tech. Nucleic Acid Biochem.* B5, 1–34
8 Smith, L.M. *et al.* (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674–679
9 Dear, S. and Staden, R. (1991) A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* 19, 3907–3911
10 Craxton, M. (1991) Linear amplification sequencing: a powerful method for sequencing DNA. *METHODS: A Companion to Methods in Enzymology* 2, 20–26
11 Fulton, L.L. and Wilson, R.K. (1994) Variations in cycle sequencing. *BioTechniques* 17, 298–301
12 Lee, L.G. *et al.* (1992) DNA sequencing with dye-labeled terminators and T7 DNA polymerase: Effect of dyes and dNTPs on incorporation of dye-terminators, and probability analysis of termination fragments. *Nucleic Acids Res.* 20, 2471–2483
13 Wilson, R. *et al.* (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* 368, 32–38
14 Vaudin, M. *et al.* (1995) The construction and analysis of M13 libraries prepared from YAC DNA. *Nucleic Acids Res.* 23, 670–674
15 Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* 12, 263–270
16 McMurray, A.A. *et al.* (1998) Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res.* 8, 562–566
17 Tabor, S. and Richardson, C.C. (1995) A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxynucleotides. *Proc. Natl. Acad. Sci. U. S. A.* 92, 6339–6343
18 McPherson, J.D. *et al.* Screening large-insert libraries by hybridization in *Current Protocols in Human Genetics* (Dracopoli, N. *et al.*, eds), John Wiley & Sons (in press)
19 Wicky, C. *et al.* (1996) Telomeric repeats (TTAGGC) are sufficient for chromosome capping in *Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A.* 93, 8983–8988
20 Chervitz, S.A. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022–2027
21 Sulston, J. *et al.* (1992) The *C. elegans* genome sequencing project: a beginning. *Nature* 356, 37–41
22 Herman, R.K. (1988) Genetics, in *The Nematode Caenorhabditis elegans* (Wood, W.B., ed.), pp. 17–45, CSHL Press
23 Waterston, R. and Sulston, J. (1995) The genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A.* 92, 10836–10840
24 Green, P. *et al.* (1993) Ancient conserved regions in new gene sequences and the protein database. *Science* 259, 1711–1716
25 Smit, A.F. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* 6, 743–748
26 Bernardi, G. (1995) The human genome: organization and evolutionary history. *Annu. Rev. Genet.* 29, 445–476
27 Barnes, T.M. *et al.* (1995) Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans. Genetics* 141, 159–179
28 Hodgkin, J. and Herman, R.K. (1998) Changing styles in *C. elegans* genetics. *Trends Genet.* 14, 352–357
29 Metzstein, M.M. *et al.* (1998) Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends Genet.* 14, 410–416
30 Sternberg, P.W. and Han, M. (1998) Genetics of RAS signaling in *C. elegans. Trends Genet.* 14, 466–472
31 Jorgensen, E. and Chalfie, M. (1998) *C. elegans* neuroscience: Genetics to genome. *Trends Genet.* 14, 206–212
32 The Sanger Centre and the Washington University Genome Sequencing Center, Toward a complete human genome sequence. *Genome Res.* (in press)
33 Huang, X.Y. *et al.* (1989) Genomic organization of the glyceraldehyde-3-phosphate dehydrogenase gene family of *Caenorhabditis elegans. J. Mol. Biol.* 206, 411–424
34 Spieth, J. *et al.* (1993) Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* 73, 521–532
35 Thacker, C. *et al.* (1995) The *bli-4* locus of *Caenorhabditis elegans* encodes structurally distinct kex2/subtilisin-like endoproteases essential for early development and adult morphology. *Genes Dev.* 9, 956–971