

# A survey of expressed genes in *Caenorhabditis elegans*

R. Waterston<sup>1</sup>, C. Martin<sup>2</sup>, M. Craxton<sup>3</sup>, C. Huynh<sup>1</sup>, A. Coulson<sup>3</sup>, L. Hillier<sup>1</sup>, R. Durbin<sup>3</sup>, P. Green<sup>1</sup>, R. Shownkeen<sup>3</sup>, N. Halloran<sup>1</sup>, M. Metzstein<sup>3</sup>, T. Hawkins<sup>3</sup>, R. Wilson<sup>1</sup>, M. Berks<sup>3</sup>, Z. Du<sup>1</sup>, K. Thomas<sup>3</sup>, J. Thierry-Mieg<sup>4</sup> & J. Sulston<sup>3</sup>

As an adjunct to the genomic sequencing of *Caenorhabditis elegans*, we have investigated a representative cDNA library of 1,517 clones. A single sequence read has been obtained from the 5' end of each clone, allowing its characterization with respect to the public databases, and the clones are being localized on the genome map. The result is the identification of about 1,200 of the estimated 15,000 genes of *C. elegans*. More than 30% of the inferred protein sequences have significant similarity to existing sequences in the databases, providing a route towards *in vivo* analysis of known genes in the nematode. These clones also provide material for assessing the accuracy of predicted exons and splicing patterns and will lead to a more accurate estimate of the total number of genes in the organism than has hitherto been available.

<sup>1</sup>Department of Genetics, Box 8232, Washington University School of Medicine, 4566 Scott Avenue, St Louis, Missouri 63110, USA

<sup>2</sup>Human Genome Center, Building 74, Lawrence Berkeley Laboratory, Berkeley, California 94720, USA

<sup>3</sup>MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

<sup>4</sup>CNRS-CRBM et Physique-Mathématique, Montpellier 34044, France

The systematic sequencing of the 100 megabase (Mb) *C. elegans* genome has begun<sup>1</sup>. This project will eventually reveal all the genes of the organism, regardless of their levels of expression or their family relationships, but will take several years to complete. We have taken advantage of the existence of a sorted cDNA library, prepared three years ago by one of us (C.M.), in order to produce sequence tags for characterizing many of the moderately to abundantly expressed genes. We have gained advance information about the total number of genes, provided the *C. elegans* community with new molecular genetic information, and identified clones against which the predictions of our exon analysis programs (P.G. & L.H. in preparation; ref. 2) can be checked.

The sorted library was produced by a cyclic procedure. In each cycle a few thousand cDNA clones were plated out and screened with probes representing all previously picked clones; only nonhybridizing clones were added to the library. In this way a collection of 1,759 isolates representing the abundant and moderately abundant clones was built up.

A sorted library differs from a random library in two important respects. First, genes expressed at low levels (down to the limit set by the total number of clones examined) are represented nearly as frequently as genes expressed at high levels; genes expressed at still lower levels are picked up with a probability proportional to their abundance. In other words, the library is partially normalized. Second, as far as possible, each selected gene is represented only once. The result is that the operations of sequencing and mapping are much more efficient than would be the case for a random library. A disadvantage of

the simple method for sorting used here (negative screening with all previously selected clones at each round) is that only one or a few members of each gene family are found. However, some of the missing information is later recovered by hybridization mapping to the genome, as we show below.

A single sequence read from the 5' end of each clone provides an identification tag that allows a match to be made against the public protein databases and the genomic sequence. Such a tagging procedure has been applied to a human brain random cDNA library by Venter and coworkers<sup>3</sup>; the same group has also tagged a *C. elegans* random cDNA library as described in the accompanying paper<sup>4</sup>. All the *C. elegans* clones, from both laboratories, are being placed on the genome map<sup>5</sup>, and all the information will be made available through ACEDB (the genome database of *C. elegans*, R.D. and J.T.-M. unpublished; obtainable from NCBI (National Center for Biotechnology Information, Bethesda, Maryland, USA) or via the authors), EMBL and GenBank. The clones themselves are freely available.

## Accuracy

For tag sequencing great accuracy is not required<sup>3,6</sup>. Provided significant similarities in the databases can be detected, the objective of the screen has been achieved. Therefore, although some templates were prepared via plasmid rescue, most were made using a simple and rapid PCR procedure as a means of amplifying the inserts for sequencing. Template quality was variable; when necessary, accuracy was improved by human intervention, checking the automated base-assignment against the

processed trace data produced by the ABI 373A fluorescent sequencing machine<sup>7</sup>. This took as long as the original sequencing, but was worthwhile not only for correcting base assignment errors in regions of adequate traces, but also for improving the interpretation of compressions. Overall, an average read length of 388 bases was used, and the sequence quality was generally good. However, it is not possible to estimate formally the accuracy of an isolated sequence read, because there is no standard of comparison. Previous experience<sup>1</sup> indicates few errors up to 300 bases, other than those due to compressions.

### Representation of the library

The original library contained 1,759 clones, and so far we have succeeded in obtaining sequence from 1,517. Most of the remainder grow poorly or have expired since the library was constructed; the recovered clones have now been stored in a more permanent fashion, at the Cambridge and St Louis laboratories.

The size of the inserts ranged from 1 to 3 kb. In

preliminary experiments, 20/20 clones were shown to have polyA at the end expected to be 3'; subsequently, only 1/1,517 was found to have polyA at the expected 5' end. 209/1,517 (14%) of the inserts carry four or more bases of a *trans*-spliced leader<sup>8</sup> at the 5' end, indicating that at least these are full length.

Table 1 shows the results of the similarity searches of the sequences (translated in all 6 frames) against the NCBI non-redundant protein database. 31% of the cDNAs gave scores above 100. The matched proteins have been arranged alphabetically, with overlapping clones indicated. We have not analysed these findings exhaustively, but comments made below help to indicate their scope.

The list consists largely of genes previously unknown in the nematode: only 15 of the 1194 nonoverlapping sequences match existing *C. elegans* entries in the database. Similarities were found across the whole range of organisms, from yeast to human. Some of the matching genes are novel members of existing gene families: for example, two cDNAs distantly related to actin (cm20d11, cm02a5) and to each other support recent evidence from yeasts that the actin gene family is diverse<sup>9,10</sup>.

Given the bias of the sorted set toward moderately expressed genes, we wanted to know how successful the coverage of moderately expressed genes has been. To make this assessment, we considered classes of moderately expressed genes that were well represented in the databases, and that were likely to show clear similarities to the corresponding nematode genes. A good example of such a group was the set of enzymes involved in the glycolytic pathway and the citric acid cycle. Of the 24 genes or gene families, 17 were found, suggesting single hit coverage of moderately expressed genes. This is not unreasonable, in light of our finding that the total number of cDNA clones screened in construction of the library is roughly comparable to the total number of genes in the organism.

Of the 1,517 processed clones, 1,194 appeared to be unique species of cDNA, as determined by the tag sequencing. This must be an overestimate, as cDNAs shortened by more than 400 bases will appear unique when compared with their full length counterparts. Since processed pseudogenes are rare in *C. elegans*, the additional family members revealed by hybridization (totalling about 300) are likely to represent additional genes. Thus as many as 1,500 genes may have been identified.

### Positions of cDNAs on the genome map

The physical map of *C. elegans* is largely represented by a set of 958 yeast clones containing fragments of worm DNA cloned into YACs. These clones have been gridded onto a nylon membrane and replicated (the 'polytene' filter) to facilitate localization of sequences on the genome map<sup>5</sup>.

Of the 1,517 sequenced cDNAs, 670 have been mapped so far. All except 64 were positioned immediately by probing of the 'polytene' filter. These 64 were used to probe an extended YAC library of 6,000 clones. 13 fell in regions of the genome mapped since the 'polytene' YACs were selected. 6 were positive for probable groups of unmapped YACs, but require further investigation. 37 should have been mapped by 'polytene' probing, but had failed for technical reasons. 5 failed to detect any YACs, and 3 gave paradoxical results; two of the former are encoded by mitochondrial DNA, and technical failure has

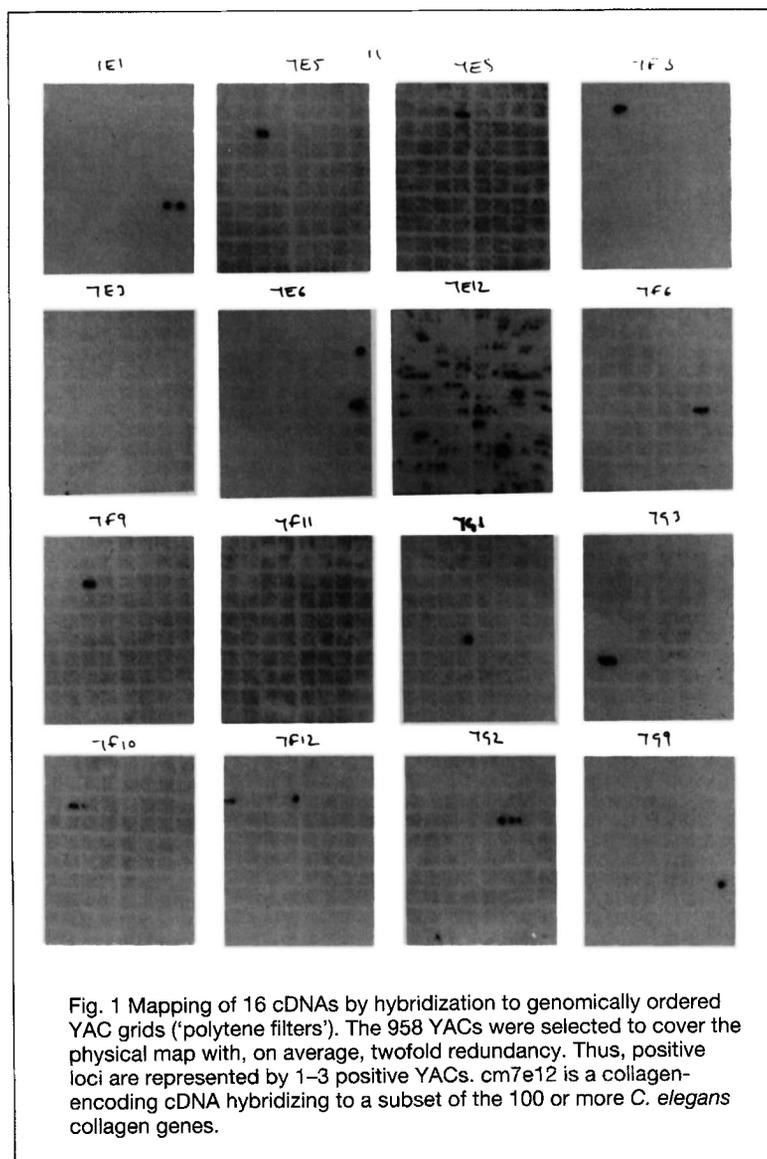


Fig. 1 Mapping of 16 cDNAs by hybridization to genomically ordered YAC grids ('polytene filters'). The 958 YACs were selected to cover the physical map with, on average, twofold redundancy. Thus, positive loci are represented by 1–3 positive YACs. cm7e12 is a collagen-encoding cDNA hybridizing to a subset of the 100 or more *C. elegans* collagen genes.

not been ruled out in the remaining cases.

In 20% of cases, multiple loci were observed. An extreme example is shown by clone cm7e12 (Fig. 1). This collagen-encoding cDNA hybridizes to at least 50 loci, while only seven genes were represented in our sequenced set. More commonly, where 2 or 3 loci were apparent (Fig. 2), a crude measure of relative intensity was entered into the *C. elegans* database ACEDB along with the genomic location. To guard against a possible artefactual origin of multiple loci from chimaeric cDNA clones, we investigated examples of genes that had yielded more than one cDNA clone (indicated by FASTA identities >90% in the set;

compare to Table 1) and that gave multiple loci on probing. In 7/7 such cases, all the cDNAs gave the same set of multiple loci. Thus, chimaerism appears not to be a frequent problem.

In this manner, the cDNAs are located on the physical map to an average precision of 100 kb or better, as about 1,000 YACs represent the 100 Mb genome on the 'polytene' grids. For nearly all of them, further refinement (to 40 kb or better) is possible, because most genes are covered by mapped cosmids as well as YACs.

The distribution of cDNAs on the chromosomes is shown in Fig. 3. All except X show some clustering in the

**Table 1 Similarity searches for *C. elegans* cDNA**

cm2h12	333	1,4-Glucan-6-(1,4-glucano)-transferase	/cm22c9	202	Cathepsin B
/cm08f10	256	3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) - Human	/cm5d10	198	Cathepsin B
/cm01a2	294	3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) alpha chain	/cm04d10	253	Cathepsin B
\cm14f3	306	3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) alpha chain	/cm12f7	114	Cathepsin B
1 cm06e10	301	3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) E2b chain	\cm01a5		
cm04g6	139	4-Nitrophenylphosphatase - <i>S. cerevisiae</i>	/cm14e3	111	Cathepsin B
cm12d10	251	6-Phosphogluconate dehydrogenase	\cm13c10		
cm13c8	248	Acetyl-CoA acetyltransferase	/cm01f3	177	Cathepsin D
cm12g6	262	Acetyl-CoA acetyltransferase	\cm2g11	201	Cathepsin D
cm06b1	138	Acetylcholinesterase precursor	/cm15h12	101	Cathepsin D
cm11e7	171	Acid alpha-glucosidase	/cm15b5		
cm10a4	104	Acid phosphatase	/cm15g10		
cm20d11	172	Actin	\cm15h8		
cm02a5	129	Actin	3 cm11b8	226	Cell cycle protein ; protein kinase
cm12b2	845	Actin - <i>C. elegans</i>	/cm04g2	579	Cell division control protein 2
cm19a6	658	Actin - <i>C. elegans</i>	/cm04g9	230	Cell division control protein 2
/cm01d9	523	Actin - <i>C. elegans</i>	\cm13c2	367	Cell division control protein 2
/cm17d1	353	Actin - <i>C. elegans</i>	cm14b7	106	Cell surface glycoprotein precursor - Mouse
\cm14f4	510	Actin - <i>C. elegans</i>	2 cm10g6	104	Cholinesterase
cm11a7	156	Adenylosuccinate synthetase	cm5h11	129	Chromosome disjoining protein dis2ml - Mouse
cm01a11	118	Adenylosuccinate synthetase	cm10f11	574	Citrate (s1)-synthase
cm7g7	255	Adenylylsulfate kinase	cm18f12	355	Citrate (s1)-synthase
/cm2d9	449	ADP,ATP carrier protein	cm10b11	312	Clathrin coat assembly protein AP50
/cm16d9	427	ADP,ATP carrier protein	cm06a12	414	Clathrin heavy chain
/cm19a3	368	ADP,ATP carrier protein	cm7d1	186	Collagen - <i>C. elegans</i>
/cm19b8	269	ADP,ATP carrier protein	cm10c5	569	Collagen - <i>C. elegans</i>
\cm19h11	190	ADP,ATP carrier protein	/cm04b1	337	Collagen - <i>C. elegans</i>
cm15e6	456	ADP-ribosylation factor	/cm04b12	329	Collagen - <i>C. elegans</i>
cm01d5	242	Alcohol dehydrogenase	/cm06a10	309	Collagen - <i>C. elegans</i>
cm14h3	495	Alcohol dehydrogenase	\cm7c3	336	Collagen - <i>C. elegans</i>
/cm5b1	442	Aldehyde dehydrogenase	cm20b9	161	Collagen alpha 1(IV) chain - Human
\cm5g7	583	Aldehyde dehydrogenase	cm19f3	419	Collagen col-6 - <i>C. elegans</i>
/cm01a9	177	Aldehyde dehydrogenase (NAD+)	cm9g8	587	Collagen dpy-13 precursor - <i>C. elegans</i>
\cm5f10	305	Aldehyde dehydrogenase (NAD+)	cm13c7	450	Collagen sgt-1 precursor - <i>C. elegans</i>
cm18d8	255	Aldehyde reductase	cm14a3	139	Creatine kinase
cm14g9	443	Alpha-actinin	/cm08b11	135	Cytochrome b
cm16b6	139	Alpha-amidating enzyme I precursor	\cm21b10	159	Cytochrome b
cm11b10	201	Amino-peptidase ; BP-1/6C3 antigen, Mouse	cm14f12	179	Cytochrome P450
cm08d9	283	AMP deaminase	cm08b12	126	Cytochrome P450
cm20g8	349	Aspartate aminotransferase	cm18e6	287	Cytochrome C oxidase chain I
/cm04f11	186	Aspartate aminotransferase	/cm11g2	270	Cytochrome C oxidase chain I
\cm06f11	203	Aspartate aminotransferase	/cm19f7	246	Cytochrome C oxidase chain I
cm2g7	122	ATP synthase gamma subunit gene,	/cm01b12	284	Cytochrome C oxidase chain I
cm16h8	505	ATP-dependent DNA helicase	\cm01g5	264	Cytochrome C oxidase chain I
cm7f6	515	ATP-dependent DNA helicase	cm13h8	123	Cytoskeletal protein 4.1
cm15c6	219	Beta-arrestin	cm11b9	360	Cytovillin 2
cm5a6	132	Betaine aldehyde dehydrogenase	cm20f4	149	Debranching protein - <i>S. cerevisiae</i>
cm01a12	154	Bone morphogenetic protein 1 precursor	cm06d3	185	Deoxyctidine kinase
1 cm17f6	300	Branched-chain alpha-keto acid dehydrogenase	cm16b2	273	Dihydroliipoamide dehydrogenase precursor
cm08g12	756	<i>C. elegans</i> cosmid ZK637 gene 3; rat proton pump	cm12b12	104	Dihydroliipoamide acetyltransferase
cm12h4	219	<i>C. elegans</i> cosmid ZK643 gene 6	cm16c12	237	Dimethylaniline monooxygenase
/cm2h2	570	<i>C. elegans</i> cosmid B0303 gene 3; acetyl-CoA acyltransferase	cm9h10	109	Disulfide isomerase-like protein
/cm7h11	587	<i>C. elegans</i> cosmid B0303 gene 3; acetyl-CoA acyltransferase	cm13f2	136	Disulfide isomerase-related protein
\cm19d4	227	<i>C. elegans</i> cosmid B0303 gene 4	4 cm06b9	131	DNA-binding protein A (DBPA) - Human
cm20b1	393	Ca2+-transporting ATPase	/cm10h3	151	DNA-binding protein hXBP-1 - Human
cm7b9	380	Ca2+-transporting ATPase	\cm12c10	142	DNA-binding protein hXBP-1 - Human
cm21d11	253	Calpain heavy chain	cm04c4	110	Dopamine beta-hydroxylase
cm11f6	178	Calreticulin precursor - Rat	cm15b7	129	Drebrins E1 and E2
2 cm7d7	128	cAMP-regulated d2 protein precursor	cm11a9	115	ECA39 protein - Mouse
cm15a6	339	Capping protein alpha-2 isoform (Cap2) mRNA,	cm7b5	245	ECA39 protein - Mouse
/cm18b8	113	Carbonic anhydrase I	cm9g2	127	Ecdysone-induced protein E75A - Drosophila
\cm21f12			/cm14d7	345	Electron transfer flavoprotein alpha
/cm5a2	101	Carnitine octanoyltransferase, hepatic	\cm16c7	400	Electron transfer flavoprotein alpha
\cm13h6	122	Carnitine octanoyltransferase, hepatic - Rat	cm13h2	111	Elongation factor Tu
cm17d2	216	Carnitine palmitoyltransferase II	cm20d1	343	Elongation factor Tu
cm20b12	334	Catalase	/cm22d1	326	Elongation factor Tu
cm12b6	219	Cathepsin B	\cm5c1	264	Elongation factor Tu

Table 1 Continued

/cm01d1	516 Elongation factor Tu alpha	525 H+-ATPase beta subunit isoform II, Bovine
/cm01h2	648 Elongation factor Tu alpha	359 H+-transporting ATP synthase 31K chain
/cm0177	599 Elongation factor 1 alpha	656 H+-transporting ATP synthase alpha chain
/cm1244	766 Elongation factor 1 alpha	405 H+-transporting ATP synthase alpha chain
/cm13d3	448 Elongation factor 1 alpha	311 H+-transporting ATP synthase alpha chain
/cm14b6	531 Elongation factor 1 alpha	341 H+-transporting ATP synthase alpha chain
/cm17b2	332 Elongation factor 1 alpha	578 H+-transporting ATP synthase alpha chain
/cm19c12	495 Elongation factor 1 alpha	246 H+-transporting ATP synthase beta chain
/cm19d3	552 Elongation factor 1 alpha	122 H+-transporting ATP synthase beta chain
/cm19e9	414 Elongation factor 1 alpha	131 H+-transporting ATP synthase beta chain
/cm19f11	432 Elongation factor 1 alpha	273 H+-transporting ATP synthase beta chain
/cm19h5	355 Elongation factor 1 alpha	189 H+-transporting ATP synthase C chain
/cm21a2	204 Elongation factor 1 alpha	193 H+-transporting ATP synthase C chain
/cm11d10	526 Elongation factor 1 alpha	363 Heat shock cognate protein 70
/cm12a8	110 Elongation factor 1 gamma	426 Heat shock cognate protein 79kDa
/cm06a5		689 Heat shock protein 3 gene, C.elegans
/cm7c2	376 Elongation factor 2	567 Heat shock protein 70
/cm01g7	353 Elongation factor 2	590 Heat shock protein 70
/cm01f11	605 Elongation factor 2	571 Heat shock protein 70 C
/cm22e4	115 Elongation factor 3	207 Heat shock protein 70 A - C. elegans
/cm17f8	432 elt-1 gene, C.elegans	689 Heat shock protein 70 A - C. elegans
/cm17f8	364 Enolase alpha chain (2-phospho-D-glycerate hydro-lyase)	333 Heat shock protein 70 A - C. elegans
/cm117	316 Enoyl-CoA hydratase precursor	660 Heat shock protein 70 A - C. elegans
/cm20c3	311 Farnesyl-protein transferase alpha-subunit	360 Heat shock protein B
/cm2e12	103 Fatty acid desaturase	124 Heat shock protein dna7
/cm10e11	102 Fatty acid desaturase	400 Heat shock protein HSP 90-beta
/cm15f4	501 Fibrillarin	142 Heterochromatin-specific chromosomal protein 1 - Fruit fly
/cm08f3	153 Flavlin-containing monooxygenase	305 Heterogeneous nuclear ribonucleoprotein
/cm14c2	127 Flavlin-containing monooxygenase	129 Histidine-rich glycoprotein precursor - Plasmodium lophurae
/cm21c4	355 Fructose-1,6-bisphosphatase	109 Histidine-rich glycoprotein precursor - Plasmodium lophurae
/cm11b6	410 Fructose-1,6-bisphosphatase	329 Histone H1.2 - C. elegans
/cm14f6	380 Fructose-1,6-bisphosphate aldolase	295 Histone H1.2 - C. elegans
/cm12a10	179 Fructose-6-phosphate, 2-kinase	9
/cm1c1	294 Fumarate hydratase	352 Hydratase-dehydrogenase-epimerase
/cm16h5	289 Fumarate hydratase	175 Hydroxacyl-CoA dehydrogenase
/cm01b5	462 Fumarylacetoacetate hydrolase	276 Hydroxymethylglutaryl-CoA reductase
/cm15b10	123 Furin-type protein, Duri, D.melanogaster	333 Hydroxymethylglutaryl-CoA synthase
/cm10b9	246 Gamma-amino-n-butyrate transaminase, Aspergillus nidulans	147 Hypothetical chromosome region maintenance 1 protein - Yeast
/cm13g3	114 GDP dissociation inhibitor for the protein	100 Hypothetical protein (prr122 5' region) - Yeast
/cm19e7	137 Gene frizzled protein precursor - Drosophila	126 Hypothetical protein (PRP4 5' region) - Yeast
/cm08b2	145 Gene suppressor-of-white-apricot protein - Drosophila	240 IMP dehydrogenase
/cm2b4	103 Gene suppressor-of-white-apricot protein - Drosophila	431 Initiation factor eIF-2, alpha chain
/cm2b5	200 Glucose-6-phosphate dehydrogenase	132 Intermediate filament protein B
/cm15d5	216 Glucose-6-phosphate dehydrogenase	432 Intermediate filament protein B
/cm15h5	206 Glucose-6-phosphate dehydrogenase	269 Intermediate filament protein B
/cm18h11	108 Glucose-6-phosphate isomerase	/cm01g8
/cm11e3	332 Glucose-6-phosphate isomerase	231 Intermediate filament protein B
/cm11h2	270 Glucose-6-phosphate isomerase	199 Intermediate filament protein B
/cm1e29	105 Glucose-transporter protein	102 Intermediate filament protein
/cm1e29	290 Glutamate ammonia ligase 2	128 Isocitrate dehydrogenase (NAD+) alpha chain
/cm0e2	409 Glutamate decarboxylase	143 Isocitrate dehydrogenase (NAD+) alpha chain
/cm08h3	196 Glutamate dehydrogenase	292 Isocitrate dehydrogenase (NADP+)
/cm12a9	335 Glutamate dehydrogenase	295 Isocitrate dehydrogenase (NADP+)
/cm5c2	379 Glutamate dehydrogenase	431 Isocitrate dehydrogenase (NADP+)
/cm7h3	167 Glutamine fructose-6-phosphate aminotransferase (isomerizing)	340 Isocitrate dehydrogenase NADPH-specific
/cm2e2	282 Glutathione reductase (NADPH)	419 Isovaleryl-CoA dehydrogenase precursor
/cm3a3	640 Glyceraldehyde-3-phosphate dehydrogenase 1 - C. elegans	135 Isovaleryl-CoA dehydrogenase precursor
/cm01d2	531 Glyceraldehyde-3-phosphate dehydrogenase 3 - C. elegans	148 Kinase-related transforming protein (p1m-1)
/cm9c3	651 Glyceraldehyde-3-phosphate dehydrogenase	264 Kinase-related transforming protein (p1m-1)
/cm15a1	241 Glycyl-3-phosphate dehydrogenase	248 L-Lactate dehydrogenase H chain
/cm5b7	271 Glycine hydroxymethyltransferase	502 L-Lactate dehydrogenase H chain
/cm16g2	107 GSD2 gene - O.sativa (rice)	274 Leukotriene-A4 hydroxylase
/cm01b9	238 GTP-binding protein beta chain	122 Lysosomal membrane sialoglycoprotein
/cm11f9	142 GTP-binding protein beta chain	217 Malate dehydrogenase
/cm21f5	493 H+-ATPase beta subunit isoform I, Bovine	211 Malate dehydrogenase
		268 Malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)

Table 1 Continued

cm12g1	136 Malate dehydrogenase - Methanothermus fervidus	cm17h1	115 Phosphoglucomutase
cm12h2	179 Malate dehydrogenase - Methanothermus fervidus	cm21a12	454 Phosphoprotein phosphatase 1-alpha catalytic chain
/cm15d12	138 Malate dehydrogenase, mitochondrial	cm17a7	110 Phosphoribosyl pyrophosphate synthetase I
cm04e7	274 Malate dehydrogenase, mitochondrial	cm984	531 Phosphoribosyl pyrophosphate synthetase I
cm21a11	218 Mbni protein - Mouse	cm16h11	329 Phosphorylase
cm7h7	316 Methionine adenosyltransferase	/cm14g3	181 Polyadenylate-binding protein
cm08c3	522 Methionine adenosyltransferase	cm16d8	152 Polyadenylate-binding protein
cm08b7	315 Methylenetetrahydrofolate dehydrogenase (NADP+)	cm04c6	151 Polyadenylate-binding protein
cm7b10	304 Methylenetetrahydrofolate dehydrogenase (NADP+)	cm19e11	201 Polyadenylate-binding protein
cm21g3	145 Mevalonate kinase - Rat	/cm17c7	206 Polyadenylate-binding protein
cm01h3	271 MHC H-2K/t-w5-linked open reading frame - Mouse	/cm19f4	134 Polyadenylate-binding protein
cm01a10	120 Microbial aspartic proteinase - Rhizopus chinensis	cm53a11	101 Polyphenyl transferase - S. cerevisiae
cm08h2	156 Mopa box protein - Mouse	cm18e5	344 Polyubiquitin
cm15d8	101 Mopa box protein - Mouse	/cm04h6	216 Porin 31HL - Human
cm598	199 Multidrug resistance protein - Sauroleishmania tarentolae	cm10f10	176 Porin 31HL - Human
cm08b3	169 Myoadenylate deaminase	cm16d5	216 Porin 31HL - Human
cm08b3	174 Myosin heavy chain	cm18g5	215 Porin 31HL - Human
cm20e9	110 Myosin heavy chain	cm19d8	160 Porin 31HL - Human
cm02b4	647 Myosin heavy chain A - C. elegans	cm1798	411 Primary biliary cirrhosis mitochondrial autoantigen - Rat
cm14a12	763 Myosin heavy chain C - C. elegans	/cm18c4	101 Primary biliary cirrhosis mitochondrial autoantigen - Rat
cm17c4	454 Myosin heavy chain C - C. elegans	cm18g6	329 Processing-enhancing protein precursor
cm17g7	225 Myosin heavy chain, smooth muscle	cm21c5	127 Procollagen alpha 1(III) chain precursor - Human
cm5e12	284 Myosin I heavy chain-like protein	cm584	131 Procollagen-proline, 2-oxoglutarate 4-dioxygenase alpha 1 chain
cm20d2	101 Myosin-light-chain kinase	cm01g9	123 Proline dipeptidase
cm04e3	258 Nax- and Cl- dependent betaine transporter, Canis familiaris	7 cm08b1	272 Prolyl 4-hydroxylase beta subunit
cm04e5	132 NADH dehydrogenase	/cm13a10	203 Propionyl-CoA carboxylase
cm12a3	109 NADH dehydrogenase	cm5h2	230 Propionyl-CoA carboxylase
/cm10c7	187 NADH dehydrogenase (ubiquinone) 49k chain	cm9a1	203 Propionyl-CoA carboxylase
cm20f11	121 NADH dehydrogenase (ubiquinone) 49k chain	8 cm22b8	143 Propionyl-CoA carboxylase alpha chain
cm5b3	264 NADH dehydrogenase (ubiquinone) 51k chain	cm17f11	124 Protective protein precursor
cm17a8	205 NADH dehydrogenase (ubiquinone) precursor	/cm02b5	379 Protective protein precursor
cm04h7	124 Nodulin-26 precursor - Soybean	cm17f4	189 Protective protein precursor
cm16b11	542 Nodulin protein p49, Rattus norvegicus liver	cm20c9	161 Protective protein precursor
cm5d4	189 Nuclease sensitive element binding protein-1 - Human	/cm06d9	168 Protective protein precursor
cm01b11	196 Oncoprotein suppressor protein SUP44 - S. cerevisiae	7 cm5b5	290 Protein disulfide isomerase
cm08d12	173 OMP decarboxylase	cm13b3	357 Protein disulfide-isomerase
cm01g6	154 ORF3 protein - Autographa californica nuclear polyhedrosis virus	cm17d8	138 Protein disulfide-isomerase
cm5e9	420 Ornithine aminotransferase	cm14f5	409 Protein disulfide isomerase
cm14h2	397 Ornithine aminotransferase	/cm1e10	194 Protein disulfide-isomerase
cm01d10	128 OSMI protein - S. cerevisiae	cm02c10	123 Protein H precursor 80 kd
cm04f10	113 p48 eggshell protein - Schistosoma mansoni	cm17b4	249 Protein kinase C delta
cm2f8	720 Patamysin - C. elegans	cm2e9	111 Protein kinase homologue B1R gene - Vaccinia
cm7b3	247 PBX2 - Human	cm14h4	129 Protein kinase, cGMP-dependent
/cm21a4	120 Pepsinogen A	cm16a1	280 Protein tyrosine phosphatase
cm15b5	156 Pepsinogen A	cm05a7	280 Proton ATPase proteolipid chain
cm15h8	148 Pepsinogen F	cm01a8	230 Proton pump polypeptide - 116 KD
/cm16b12	147 Pepsinogen F	cm08g6	130 PSE-1 gene - S.cerevisiae
cm01c12	135 Pepsinogen F	cm06e3	254 Pseudoprotease homologue
cm01b7	137 Pepsinogen II-2/3	cm13h4	197 PUP1 gene - S.cerevisiae
5 cm08a9	117 Pepsinogen	cm15b3	214 Pyruvate carboxylase
cm01g7	137 Pepsinogen	cm13e3	286 Pyruvate dehydrogenase
cm15f11	101 Phosphate carrier protein, mitochondrial	/cm03e8	417 Pyruvate dehydrogenase type II alpha subunit
/cm19c2	256 Phosphate transport protein, mitochondrial	cm08e12	370 Pyruvate dehydrogenase type II alpha subunit
cm20a9	220 Phosphate transport protein, mitochondrial	cm2b9	423 Pyruvate dehydrogenase type II alpha subunit
cm5c9	209 Phosphatidylcholine--sterol acyltransferase	407 Pyruvate kinase	
cm16c8	727 Phosphoenolpyruvate carboxylase	cm13a6	244 Pyruvate kinase
cm21f6	232 Phosphoenolpyruvate carboxylase	cm20e6	236 rac protein kinase alpha - Human
/cm06g7	365 Phosphoenolpyruvate carboxylase	cm10h12	142 rac protein kinase beta - Human
cm16b5	277 Phosphoenolpyruvate carboxylase	504 RAD6 - Yeast	
cm29a	352 Phosphoenolpyruvate carboxylase	cm21g4	272 ral-1 - Onchocerca volvulus
/cm12h5	341 Phosphoenolpyruvate carboxylase	cm21z2	260 ras-related protein ; transforming protein (ypt1) - Mouse
cm14g7	185 Phosphoenolpyruvate carboxylase	/cm21g5	427 ras-related protein
cm16d12	506 Phosphoenolpyruvate carboxylase	/cm17e1	110 ref (2)P protein
/cm19c7	250 Phosphoglucomutase	cm11c6	

cm1194	120 Ribonucleoprotein 60 kd - Human	107 tRNA ligase--Lysine
/cm13b11	116 Ribonucleoside-diphosphate reductase chain M2	511 tRNA ligase--Lysine
/cm01b10	245 Ribonucleoside-diphosphate reductase small chain	391 tRNA ligase--Seryl
\cm7b6	262 Ribonucleoside-diphosphate reductase small chain	237 tRNA ligase--Seryl
\cm18f4	409 Ribonucleoside reductase	193 tRNA ligase--Threonine
/cm14h5	248 Ribosomal protein L10e	174 tRNA ligase--Threonine
\cm21d3	255 Ribosomal protein L10e	/cm20c1
\cm15c	405 Ribosomal protein L1a	\cm15e4
/cm01a7	457 Ribosomal protein L3 homologue - Yeast	/cm18h10
/cm01d12	295 Ribosomal protein L3 homologue - Yeast	/cm11g5
/cm01h2	394 Ribosomal protein L3 homologue - Yeast	\cm05c5
/cm01h6	558 Ribosomal protein L3 homologue - Yeast	\cm04h4
/cm14e6	601 Ribosomal protein L3 homologue - Yeast	/cm01d4
/cm19d2	416 Ribosomal protein L3 homologue - Yeast	\cm01f6
\cm18b6	169 Ribosomal protein L3 homologue - Yeast	\cm13g11
/cm12a12	344 Ribosomal protein L3 homologue - Yeast	\cm18d12
\cm17c9	423 Ribosomal protein L5a	/cm01f4
\cm19p3	488 Ribosomal protein L5a	\cm21a7
\cm06c5	264 RNA helicase	\cm14c3
\cm16c2	100 Ryanodine receptor	/cm01h1
\cm20a6	133 Sarcoplasmic reticulum 53K glycoprotein precursor	\cm22e6
\cm12h12	111 SCJ1 protein - S. cerevisiae	\cm10d4
\cm9h4	223 SEC23 protein - S. cerevisiae	\cm2d2
\cm5a10	105 SELB Translation factor	8 cm04c8
\cm7a12	271 Serine-pyruvate aminotransferase	\cm07d12
/cm12d11	562 Signal recognition particle 54K protein	\cm10a2
\cm13c10	343 Signal recognition particle 54K protein	\cm16d1
\cm15f12	284 Signal recognition particle 54K protein	\cm20a12
\cm15g11	286 Signal recognition particle 54K protein	\cm14e12
\cm15e7	269 smg p25A regulatory protein - Bovine	\cm19c3
/cm12h8	324 smg p25A regulatory protein - Bovine	\cm08f5
\cm13f4	231 smg p25A regulatory protein - Bovine	\cm9h7
\cm10e10	107 SNF5 regulatory protein - S. cerevisiae	6 cm21e7
/cm17a2	276 Spectrin alpha-chain	4 cm10e9
\cm9a10	356 Spectrin alpha-chain	4 cm21e9
\cm18c11	144 Spectrin alpha-chain	\cm9b4
/cm5e10	130 Splicing factor mRNA, ST2p33 - Human	-----
\cm5d3	161 Splicing factor mRNA, ST2p33 - Human	Footnotes
\cm22b6	111 Stage specific protein - Trypanosoma brucei	-----
\cm7d11	116 STE7 regulatory protein - S. cerevisiae	1 clone cm06e10 overlaps clone cm17f6
\cm04a9	168 Steroid hormone receptor ERR1 - Human	cm17f6 300 Branched-chain alpha-keto acid dehydrogenase
\cm06b9	161 Sterol carrier protein 2	cm06e10 301 3-Methyl-2-oxobutanate dehydrogenase (lipoamide) E2b chain
\cm12a5	331 Succinate dehydrogenase (ubiquinone)	2 clone cm10g6 overlaps clone cm7d7
\cm12e12	419 Succinate-CoA ligase (GDP-forming) alpha chain	cm10g6 104 Cholinesterase
\cm20b3	283 Succinate-CoA ligase (GDP-forming) alpha chain	\cm7d7 128 CAMP-regulated d2 protein precursor
\cm5f4	377 Succinyl-CoA:alpha-ketoacid coenzyme A transferase	3 clone cm11b7 overlaps clone cm11b8
\cm9h1	337 Succinyl-CoA:alpha-ketoacid coenzyme A transferase	cm11b7 142 rac protein kinase beta - Human
\cm13f9	303 Surf-4 protein - Mouse	cm11b8 226 Cell cycle protein ; protein kinase
/cm13f8	415 T-complex protein 1	4 clone cm06b9 overlaps clone cm10e9
\cm08g10	546 T-complex protein 1	cm06b9 131 DNA-binding protein A (DBPA) - Human
\cm1d11	336 T-complex protein 1	\cm10e9 136 Y-box protein 2 - African clawed frog
/cm12b10	195 T-complex protein 1	5 clone cm08a9 overlaps clone cm01h3
\cm13h1	539 T-complex protein 1	cm08a9 117 Pepsinogen II-2/3
\cm09a1	229 T-complex protein 1	\cm01h3 120 Microbial aspartic proteinase - Rhizopus chinensis
\cm15c5	182 T-protein mRNA - Bovine	6 clone cm7d6 overlaps clone cm21e7
\cm18a3	191 T-protein mRNA - Bovine	cm21e7 249 Nuclease sensitive element binding protein-1 - Human
/cm15e11	331 Triosephosphate isomerase	7 clone cm21e7 187 Y-box protein 1 - African clawed frog
\cm9g9	384 tRNA ligase--Aspartate	cm5b5 290 Protein disulfide isomerase
\cm14h12	191 tRNA ligase--Glutamine	\cm08b1 272 Prolyl 4-hydroxylase beta subunit
\cm20d3	241 tRNA ligase--Lysine	8 clone cm04c8 overlaps clone cm22b8
\cm22a6		\cm04c8 268 Urea amidolyase - Yeast
		cm22b8 143 Propionyl-CoA carboxylase alpha chain

The first column lists the clone name, the second column gives the highest similarity score with BLASTX and the third column gives the generic name of the highest scoring entry. In cases where several entries from various species had similar scores only the name is given; in those cases where the highest scoring entry was distinct from other matches, more detail is given. Comparison of the cDNAs with one another using FASTA revealed several potentially overlapping sequences. Those with a FASTA opt score of at least 200 (equivalent to a contiguous run of 50 identical bases) and 90% or greater identity at the nucleotide level for the region of overlap are noted either by brackets at the left margin where they matched similar database entries or by footnotes where members of a group matched distinct entries. Group members which did not have a BLASTX score > 100 are also included in the list.

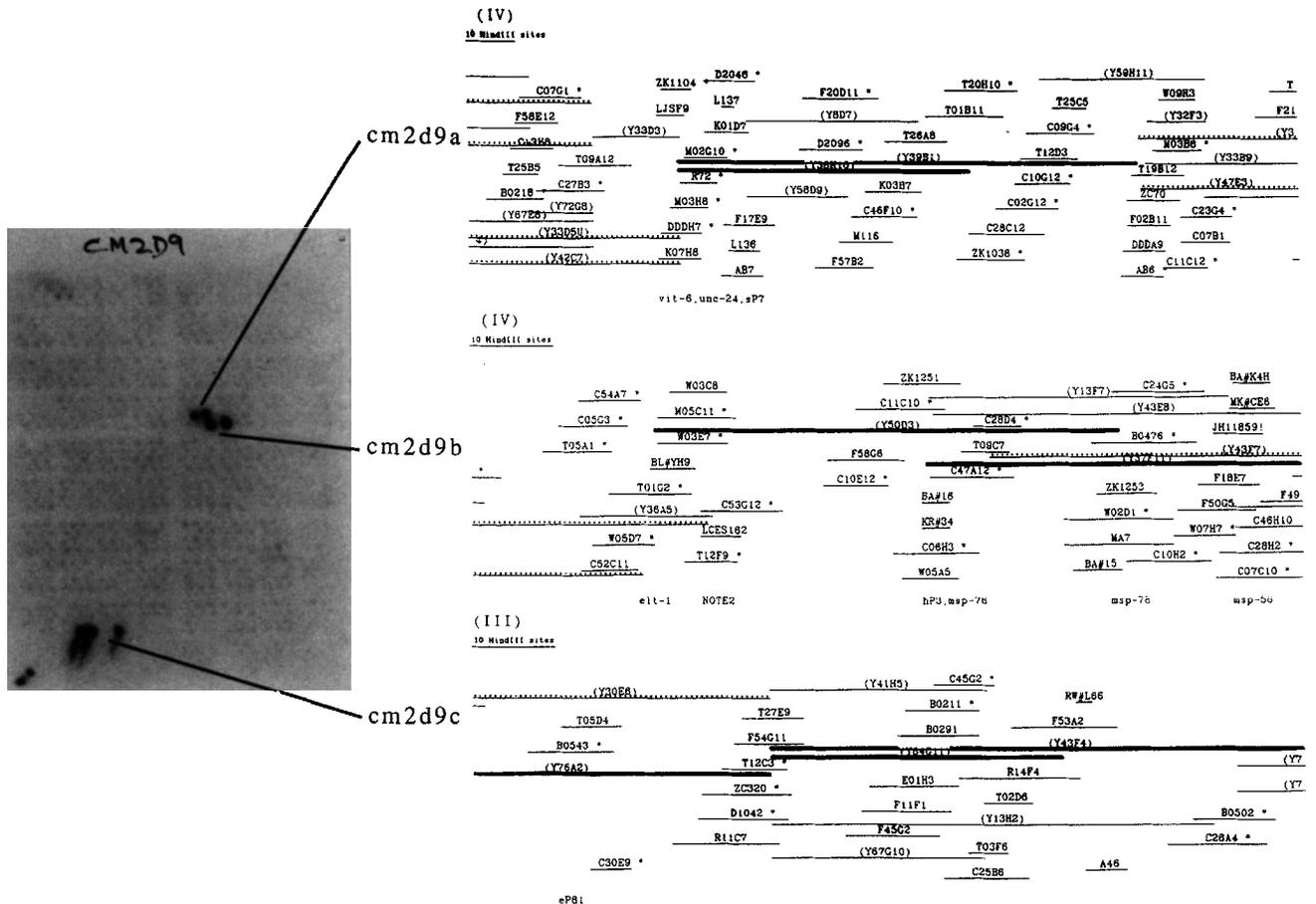


Fig. 2 Detailed physical map locations of clone cm2d9 (see Table 1) and its homologues. On the contig maps: positively hybridizing YACs are shown in bold; YACs that are present on the grid but do not hybridize are dotted. The single positive colony at bottom left is a YAC that had not been mapped when the polytene filters were made, but has subsequently been shown to be associated with Y39B1 and Y38H10. Because the end points of the YACs with respect to the cosmids are not usually precisely known, physical overlap of Y76A2 (but not Y30E6) with Y64G11 and Y43F4 is plausible. Because the apparent intensity of hybridization at each locus is the same, the true location of cm2d9 is not defined.

centre, parallelling the clustering seen among visible mutations (this physical clustering is much less prominent than the genetic clustering<sup>11</sup>, which is exaggerated by reduced recombination in the same areas). The fact that 98% of the cDNAs were located is a clear indication that the gene-containing regions are well represented in the physical map. The data do not address the representation of regions lacking genes; available evidence indicates that such regions may exist on chromosome arms.

cDNAs can serve as ideal STSs for map linkage<sup>12,13</sup>. Because the *C. elegans* map is already well developed, this aspect of the project is not prominent. Nevertheless, a few of the cDNAs have served to confirm pre-existing linkages.

**There are about 15,000 genes in *C. elegans***

A valuable application of the cDNA information will be to estimate the total number of genes in the genome. In five cosmids sequenced and analysed so far, extending over about 200 kb near *unc-32(III)* (Fig. 3), we predicted 50 genes by using the program GENEFINDER<sup>2</sup>. It is possible that some of the putative genes are nonfunctional, or that some real genes have been missed, but the success of the program in predicting exons subsequently detected by other means suggests that the overall count is accurate.

The problem is to extrapolate this figure to the rest of the genome.

If we were to assume that the gene density of the sequenced region is typical of the whole genome, then we would predict a total count of  $50 \times 100,000 / 200 = 25,000$  genes. This is likely to be too high, however, because the distribution of visible mutations along the chromosome (and the clustering of cDNAs found in the present study) suggest that the gene density in the region sequenced is above average. A better assumption is that representation of the sequenced region (4 cDNA clones matched predicted genes) in the sorted cDNA library (1,194 unique clones) is typical of the whole genome - in other words, that moderately to abundantly expressed genes have the same genomic distribution as other genes. Under this assumption, we predict a total of  $50 \times 1,194 / 4 = 15,000$  genes. Although this estimate is highly uncertain at present (binomial 95% confidence limits 6,500–40,000), it will be progressively refined as genome sequencing proceeds. In making this calculation we have considered only exact coincidences between the cDNA library and the region, as we want to use the library merely as a random sampling device.

The estimate of 15,000 is severalfold higher than that

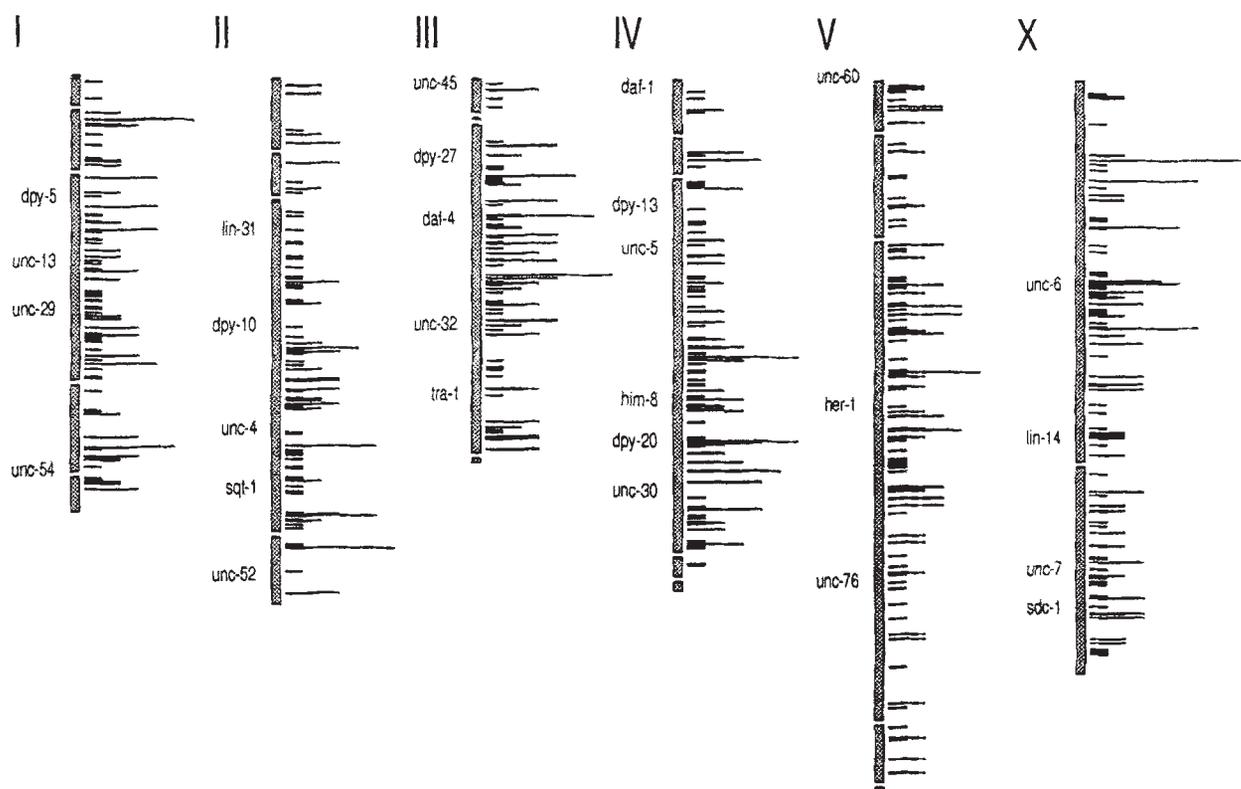


Fig. 3 Summary of clone locations on the genomic map, output directly from ACEDB. The contigs (hatched blocks) representing each of the six chromosomes are shown, divided by arbitrarily small gaps. The size of the gaps between contigs is unknown and is particularly likely to be underestimated towards the ends of the chromosomes; this means that the genes may be somewhat more clustered than is apparent in the Fig. cDNAs in each of various intervals are indicated by the horizontal lines to the right of the contig blocks; where more than one cDNA maps to a given YAC, the lines are allowed to stack horizontally, but on this Fig. the cases of identical cDNAs are not distinguished from the cases of different cDNAs mapping at the same position. For reference the positions of several genes are indicated to the left of the contigs<sup>11</sup>. The region of the genome that is being sequenced lies around *unc-32*.

obtained by counting recessive lethal and visible mutations<sup>14-16</sup>. As for other organisms<sup>17-19</sup>, it is apparent that only a fraction of genes in *C. elegans* can be detected by this means. A twofold shortfall was previously suggested based on the incidence of silent revertants of dominant mutants<sup>20</sup>, and it now seems likely that the ratio is still higher.

#### cDNA versus genomic sequencing

At first sight it may appear that cDNA sequencing can serve as a more efficient substitute for genomic sequencing. According to our current estimates, only about 15% of the genome codes directly for protein. If this is the most useful portion, why should we want to sequence the remainder? Would it not be a better use of resources to sequence all the cDNAs fully? Furthermore, to do so would avoid the difficulties and uncertainties of genomic analysis. Although our routines for predicting exons and splicing patterns are quite successful, they are by no means perfect.

These are significant criticisms and are not to be ignored. Indeed, if it were really possible to collect and sequence, in a simple way, one complete cDNA for each gene in an organism, then a complete cDNA project would be a good way to begin the study of a genome. However, it could

only be a beginning, and even at that level there are problems of representation. Rarely used genes, and rarely used splicing patterns, will always be hard to find as cloned cDNAs, no matter how carefully libraries are normalized and sorted. In contrast, genomic sequencing yields each gene once and once only, and the existence of infrequently used exons (of sufficient size or similarity) is made apparent. These predictions can then be tested by PCR, using uncloned total cDNA as a template.

Even given a complete set of cDNAs, however, their sequences will provide little information about other important components of the genome. Most obvious are the control sequences: at present, we are not able to interpret effectively genome sequence in terms of control elements, but this should improve. Of course, it is always possible to obtain the flanking genomic sequences for a single cDNA clone, but much extra effort is involved.

Finally, noncoding regions can be expected to contain a variety of structural elements and components involved in more global controls, for example in the behaviour of chromosomes during mitosis and meiosis, and both coding and noncoding parts of the genomic sequence are likely to contain clues to its evolution. Much of this information will be difficult or impossible to obtain without a direct examination of the genome itself.

In our opinion, for *C. elegans* research the genomic sequence should be the primary objective; cDNAs, and other approaches to transcript analysis, are useful for its interpretation but are not substitutes for it. In a real sense the genomic sequence is the program that runs the organism. Obtaining that program in its entirety will be invaluable for understanding the present functioning and past evolution of the developmental system.

### Methodology

**Library construction.** mRNA was prepared from a mixed stage, log phase, *C. elegans* liquid culture. First strand cDNA was synthesized in the presence of methyl-dCTP, from a primer containing oligo-dT and an *ApaI* site. Second strand cDNA was synthesized in the presence of dCTP, and *SacI* linkers were added. After cleavage with *ApaI* and *SacI*, fragments >1 kb were size selected and cloned into lambdaSHLX2 (ref. 21). Clones for the library were collected in pools. Isolating a pool began by plating phage at 500 15 cm<sup>-1</sup> plate, using enough plates to give about 10× the number of desired clones for the group. Several filter replicas were prepared from each plate; one was hybridized to a total cDNA probe to identify plaques derived from abundant mRNA species, a second to a vector specific probe to mark all plaques, and the remainder to a series of probes, each representing 250 previously acquired clones. Clones that were positive to the second probe but negative to the rest were picked individually into storage medium. Each clone was subjected to PCR amplification, and those showing inserts >1 kb were plaque purified and added to the collection. Representatives of the abundant mRNA class were collected in a separate pooling operation. In total, two highly abundant and five much larger moderately abundant pools were collected.

**Library manipulation.** Phage were initially stored as lysates prepared from standard cultures either at 4°C or frozen in 8% DMSO. In order to duplicate the library and to reestablish high titre stocks, we adapted an earlier procedure<sup>22</sup>: 5–10 µl of phage stock were mixed with an equal volume of MC1061 host cells (prepared from an overnight growth, concentrated twofold and stored in 10 mM MgCl<sub>2</sub>) in 96 well round-bottom microtitre plates, incubated for 20 min at 37°C, mixed with 25 µl of 0.7% low-melting agarose in NZCYM and pipetted onto 125 µl of NZCYM medium solidified with 1.4% agar in each well. After a brief incubation at 4°C, the plates were incubated overnight (37°C). Phage were eluted with 50 µl of phage diluent. Because the titre of many stocks was too low to provide confluent lysis (and the resultant phage titre was low), the clones were subjected to a second round of amplification. At this round, several copies were prepared, one for shipment, another for making a frozen stock and a third for a working stock. For shipment, the lysed plates were simply capped and shipped. For frozen stocks, DMSO in phage diluent was added to make a final concentration of 8%, mixed with the culture, capped and frozen at -70°C. For working stocks, the phage were eluted with 50 µl water (for PCR) or phage diluent (for plasmid recovery).

**Plasmid templates.** For the initial 200 sequences, the λ clones were converted to plasmids, using the lox sites and the amp<sup>r</sup> gene flanking the inserts. 5 µl of phage lysates were used to infect 50 µl of the 'popout' strain<sup>21</sup>, lysogenic for both phage P1 (to provide recombinase activity) and λ (to reduce lysis of the cells by infection). Because the popout host strain gives low and highly variable yields of plasmid DNA, the plasmids were transferred to a different host, XL-1-blue. For this, DNA was prepared from amp<sup>r</sup> colonies, and 2 µl used to transform 10 µl of XL-1-blue competent cells. Template DNA was

then prepared from amp<sup>r</sup> XL-1-blue colonies. Phage infection, selection of amp<sup>r</sup> popout colonies, transformation, cell growth in liquid and plasmid DNA preps were all done using 96-well microtitre format and, where possible, multichannel pipetors. Selection of the amp<sup>r</sup> XL-1-blue colonies required the use of individual plates, and phenol extraction was done in individual tubes (arrayed in the 96 well format). Plasmid DNA was prepared from 600 µl of a 1 ml culture grown in Micronics tubes covered with parafilm and shaken on a mini-orbital shaker (Bellco), using the alkaline lysis procedure, adapted for microtitre plates<sup>23</sup>. For the plasmid transfer, the procedure was performed up to the first ethanol precipitation, and crude DNA used for transformation. For the sequencing templates, high concentration of RNase was used as described<sup>24</sup>. The final precipitate was resuspended in 25 µl of TE (10 mM Tris, pH 8, 1 mM EDTA) and a total of 6 µl used in sequencing reactions.

**Templates prepared by PCR.** PCR reactions were performed in batches of 96 using microtitre format thermocyclers (Techne MW1 or PHC3, or Cetus 9600). The DNA substrate was obtained either by transferring about 2 µl of top agar from lysed areas of bacterial lawns to 10 µl of water, or by adding about 2 µl of plate lysate to 4 µl of water. The lysates were covered with mineral oil and heated at 95°C for 10–15 min. A suitable mix was added to give: 0.2 µM each amplification primer, 200 µM each deoxynucleotide, 50 µM KCl, 10 µM Tris pH8.5, 1.5 µM MgCl<sub>2</sub>, and 1 U of Taq polymerase per reaction. The reaction was thermally cycled 30–35 times. After cycling, PEG, salt (NaCl or NaOAc) and MgCl<sub>2</sub> were mixed with the aqueous phases in microtubes to give final concentrations of 7%, 0.3M and 1.75mM respectively. After precipitation for 10 min at room temperature, DNA was pelleted by centrifugation, washed twice with 95% ethanol and dispersed in 20–50 µl of water to give a concentration of about 50ng µl<sup>-1</sup>. Amplification primers: 1) 5' GATTTAGGTGACACTATAG (SP6 promoter specific 19mer, adjacent to 5' end of cDNA insert). 2) 5' TAATACGACTCACTATAGGG (T7 promoter specific 20mer, adjacent to 3' end of cDNA insert). 3) 5' TGTAAACGACGGCCAGTGCAGATTTAGGTGACAC (M13-21/SP6 35mer, for use as an alternative to 1).

**Sequencing reactions and analysis.** Sequencing was carried out<sup>24,25</sup> using either SP6 19mer or M13-21 18mer primers (ABI). The trace data was transferred to Sun workstations, where vector sequences were removed from the 5' end and unreliable data removed from the 3' end using the trace editor program ted<sup>26</sup>. Editing of the traces was also done using ted. The sequences were translated in all six frames and compared against the NCBI nonredundant protein database, using the NCBI's GenInfo network Blast<sup>27</sup> server. Scores > 100 (using the PAM120 matrix) were found to exclude reliably matches of dubious significance that result from homopolymeric tracts or other repeated tracts. Comparisons of the cDNA's against themselves to detect overlaps used FASTA<sup>28</sup>, with a score cutoff of 200. Matches of greater than 90% identity were arbitrarily accepted as likely to be overlapping.

**cDNA mapping.** Master grids and 'slave' copies of 958 largely genomically ordered overlapping YACs ('polytene' filters) were prepared essentially as described<sup>29,30</sup>. YACs had been mapped as described<sup>5,30</sup>. Slave copies were printed on Hybond-N+ nylon membranes (Amersham). Radiolabelling of PCR amplified cDNA inserts was by standard procedures using random oligonucleotide hexamer labelling<sup>31,32</sup>. Generally, 48 overnight hybridizations were carried out in parallel in microtitre-plate lids stacked in airtight containers<sup>22</sup>.

Received 20 February; accepted 20 March 1992.

**Acknowledgements**

We thank Andrew Smith, Mohammed Jier and Teresa Copsy for their skilled assistance; Warren Gish and the NCBI for providing access to the GenInfo network BLAST server. C.M. was supported by the Alfred P. Sloan Foundation and acknowledges M. Chalfie for support and helpful discussions. The work was supported by grants from the NIH Human Genome Center and the MRC HGMP, as well as by our respective institutions. All sequences have been entered in genbank and can be obtained directly. The genbank locus name will be CEL + cDNA clone number.

1. Sulston, J. *et al.* *Nature* **356**, 37–41 (1992).
2. Staden, R. *Meth. Enzym.* **183**, 163–180 (1990).
3. Adams, M.D. *et al.* *Science* **252**, 1651–1656 (1991).
4. McCombie, W.R. *et al.* *Nature Genet.* **1**, this issue (1992).
5. Coulson, A. *et al.* *BioEssays* **13**, 413–417 (1991).
6. States, D.J. & Botstein, D. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5518–5522 (1991).
7. Staden, R. & Dear, S. *Nucl. Acids Res.* **19**, 3907–3911 (1991).
8. Krause, M. & Hirsh, D. *Cell* **49**, 753–761 (1987).
9. Leesmiller, J.P., Henry, G. & Helfman, D.M. *Proc. natn. Acad. Sci. U.S.A.* **89**, 80–83 (1992).
10. Schwob, E. & Martin, R.P. *Nature* **355**, 179–182 (1992).
11. Edgley, M.L. & Riddle, D.L. *Genetic Maps* **5**, 3 (1990).
12. Olson, M.V., Hood, L., Cantor, C. & Botstein, D. *Science* **245**, 1434–1435 (1989).
13. Green, E. and Olson, M.V. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1213–1217 (1990).
14. Brenner, S. *Genetics* **77**, 71–94 (1974).
15. Herman, R.K. in *The Nematode Caenorhabditis elegans* (ed. Wood, W.B.) 17–45 (Cold Spring Harbor Laboratory, 1988).
16. Clark, D.V., Rogalski, T., Donati, L.M. and Baillie, D.L. *Genetics* **119**, 345–353 (1988).
17. Hall, L.M.C., Mason, P.J. & Spierer, P. *J. molec. Biol.* **169**, 83–96 (1983).
18. Bossy, B., Hall, L.M.C. & Spierer, P. *EMBO J.* **3**, 2537–2541 (1984).
19. Olson, M. in *Genome Dynamics, Protein Synthesis, and Energetics* (eds Broach, J.R., Pringle, J.R. & Jones, E.W.) 1–41 (Cold Spring Harbor Press, New York, 1991).
20. Park, E.-C. & Horvitz, H.R. *Genetics* **113**, 821–852 (1986).
21. Palazzolo, M.J. *et al.* *Gene* **88**, 25–36 (1990).
22. Kohara, Y., Akiyama, K. & Isono, K. *Cell* **50**, 495–508 (1987).
23. Gibson, T.J. & Sulston, J.E. *Gene Anal. Techn.* **4**, 41–44 (1987).
24. Craxton, M. *Methods: A Companion to Methods in Enzymology* **3**, 20–26 (1991).
25. Halloran, N., Du, Z. & Wilson, R.K. in *Methods in Molecular Biology vol. 10: DNA Sequencing: Laboratory Protocols* (eds Griffin, H.G. & Griffin, A.M.) (Humana Press, Clifton, New Jersey, in the press).
26. Gleeson, T. & Hillier, L. *Nucl. Acids Res.* **19**, 6481–6483 (1991).
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. molec. Biol.* **215**, 403–10 (1990).
28. Pearson, W.R. & Lipman, D.J. *PNAS* **85**, 2444–2448 (1988).
29. Coulson, A. & Sulston, J. in: *Genome analysis: a Practical Approach* (ed. Davies, K.) 19–39 (IRL Press, 1988).
30. Coulson, A., Waterston, R., Kiff, J., Sulston, J. and Kohara, Y. *Nature* **335**, 184–186 (1988).
31. Feinberg, A.P. & Vogelstein, B. *Anal. Biochem.* **132**, 6–12 (1983).
32. Feinberg, A.P. & Vogelstein, B. *Anal. Biochem.* **137**, 266–267 (1984).