Developmental Expression Patterns of Genes Predicted by the *C.elegans* Genome Sequencing Project.

Andrew Sean Lynch

Submitted in accordance with
the requirements for the degree of
Doctor of Philosophy

The University of Leeds
Department of Biology

December 1996

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others

Abstract

The recent progress in the sequencing of entire genomes has identified many novel genes, many of which have no further description. Knowledge of gene expression pattern is one facet of a gene's description which can lead to functional insights. The work presented here aimed to describe the patterns of expression of novel genes identified by genome sequencing of the nematode *Caenorhabditis elegans*.

Gene 5' ends were used to generate translational fusions to the *lacZ* reporter gene, and the resultant *lacZ* fusion expression pattern observed *in situ*. 45 fusions were made, of which 24 were active. The observed patterns of expression, even when the responsible gene was not homologous to a functionally characterised gene, suggested functional possibilities for many genes. The *lacZ* fusion for the gene ZK637.8 exhibited expression specific to the gut lineage in early embryos suggesting that it may be involved in developmental processes. The promoter region of ZK637.8 was subject to further investigation and a region necessary for embryonic expression was identified.

The data, plasmids and nematode strains generated by this project represent a resource which will be valuable to the *C.elegans* research community as a whole. Two means of allowing public access to the data were developed; one utilising the internet, the other incorporating the data into the widely used *C.elegans* database, ACeDB.

Acknowledgements

I would like to thank my supervisors, Ian Hope and Elwyn Isaac, for their great help and guidance during the course of this project. I would also like to thank the many members of the department whose warmth and interest has made it such an enjoyable place to work. Special thanks to Pritti, Jane and Alison for putting up with my sometimes peculiar behaviour in an enclosed space.

I would like to acknowledge the practical support provided by Ian Hope and David Briggs at various times during the project. The majority of the work reported here was supported by a grant by the Human Frontiers Science Project Organisation.

Finally, a huge and heartfelt thanks to Emma and Pritti for allowing me to share their home during the time spent writing this thesis, and for the ultimate will to finish. You have my eternal love and gratitude.

Abbreviations

bp	base pair/s
BSA	Bovine serum albumen
cDNA	complementary DNA
DMF	Dimethyl formamide
DNA	Deoxyribonucleic acid
dNTPS	Deoxyribonucleotide triphosphates
EDTA	Ethylenediaminetetra-acetic acid
EST	expressed sequence tag
kb	kilobase
LB	Luria-Bertani
Mb	megabase
MES	2-[N-morpholino]ethanesulfonic acid
mRNA	messenger RNA
NGM	nematode growth medium
NLS	nuclear localisation signal
nt	nucleotide
ORF	open reading frame
PCR	Polymerase Chain Reaction
PEG	Polyethylene glycol
RNA	Ribonucleic acid
SDS	Sodium dodecyl sulphate
TAE	Tris-acetate EDTA buffer
v/v	volume/volume
WP:"gene name"	worm protein of:"gene name"
w/v	weight/volume
X-gal	5-bromo-4-chloro-3-indoyl-β-D-galactosidase
YAC	Yeast artificial chromosome

List of contents.

List of figures

List of tables

List of plates

Chapter 1

General Introduction

The data produced as a result of the systematic sequencing of whole genomes presents a challenge to modern biology: how best to extract the information encoded in genomic sequence of relevance to understanding the development and functioning of a whole organism. An essential first step must be determining the location and structure of probable genes present in the genomic sequence. The second, and more problematical, step will be functional characterisation of these predicted genes. Mutational techniques have been the primary method for assigning biological function to specific genes. The sheer number of novel genes discovered through the various genome sequencing projects, however, suggests that new techniques will be needed to address the question of their function. Only when this stage has been completed can the question of how the full genetic repertoire of an organism relates to its overall development and physiology be realistically addressed.

The era of whole genome sequencing.

Recent years have seen the inception, and in some cases completion, of projects aimed at sequencing the whole genomes of selected model organisms. The individual aim of each project was the one dimensional, or linear, description of the genetic information encoded in an organism's genome, with the expectation of providing the basic information required to enable a complete biological understanding of that organism (Oliver, 1996). The theoretical basis of this is illustrated in Figure 1.1. Developmental research can be idealised as the attempt to model the four dimensional development and functioning of an organism in terms of the interactions of its component parts. Fundamentally, these interactions are controlled by the one dimensional information encoded in the DNA of the genome. Coordinated gene expression controls the higher order interactions of proteins, organelles and cells/tissues to result in the distinctive biological functions of the organism as a whole. Complete understanding of these processes would not be possible without a complete genetic description.

Specific attributes have favoured the choice of particular organisms for genome sequence projects. In the main, organisms with relatively small and compact genomes have been

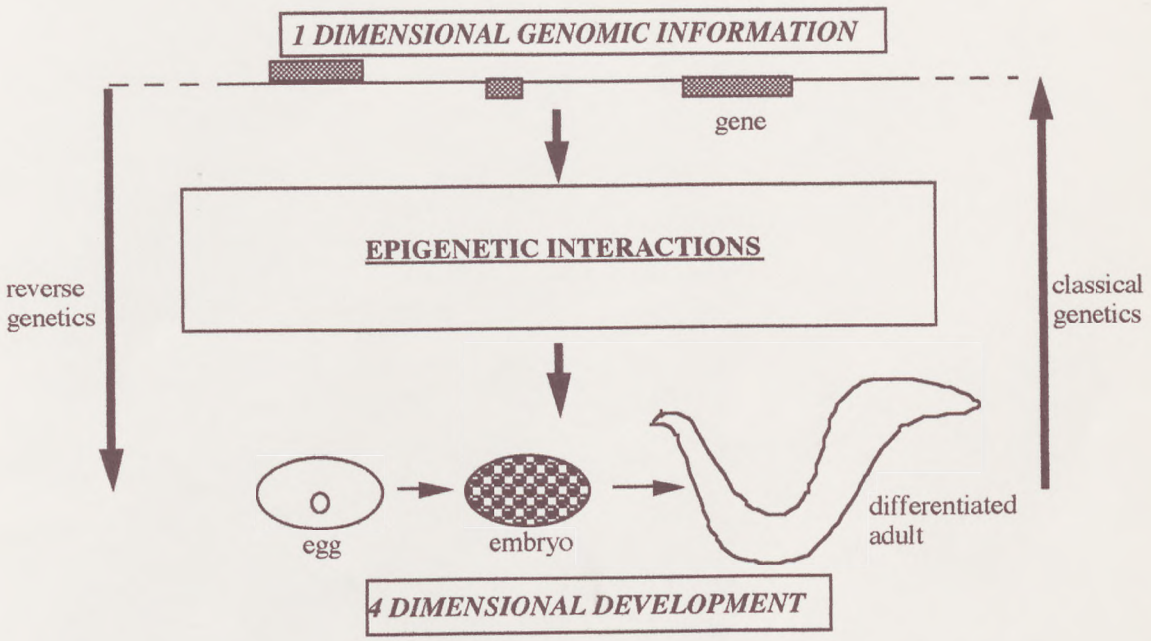


Figure 1.1. Representation of the relationship between the one dimensional genetic information and the four dimensional development of an organism.

The unilateral flow of information from the genome to the developmental process can be studied both by reverse genetics (lefthand arrow) which starts with a gene and aims to find its function in the animal, and classical genetics (righthand arrow) which starts with a phenotype in the animal and aims to identify those genetic elements responsible.

chosen such as *Escherichia coli* (Daniels *et al.*, 1992), *Saccharomyces cerevisiae* (Olson, 1991) and *Caenorhabditis elegans* (Sulston *et al.*, 1992). Such a consideration aids the sequencing effort by reducing the amount of sequencing necessary before a genome is completed. Also of great importance with these three organisms has also been their tractability as genetic systems (Brenner, 1974; Rothstein, 1983). Information from both molecular and mutational analyses is thus most easily gained, allowing studies of the epigenetic interactions underlying biological functions to be approached from different but convergent perspectives (Figure 1.1).

The budding yeast, *S. cerevisiae*, is a unicellular eukaryote and has been historically used to study genetically processes fundamental to eukaryotic life such as cell cycle control (Nurse, 1985), meiotic mechanisms (Simchen and Kassir, 1989) and global chromosome structure (Yanagida, 1990). As explained above, provision of a complete genome description should extend the scope of such studies and provide the means of a general and basic description of eukaryote functions. The bacterium *E. coli* fulfills the same role for prokaryote species and has proved invaluable in the genetic and biochemical description of many processes fundamental to all living organisms such as basic metabolic pathways and DNA-related activities such as the regulation of gene transcription and DNA replication (Blattner *et al.*, 1993).

Caenorhabditis elegans is a simple metazoan animal of approximately a thousand cells (Wood, 1988), and its suitability as a model organism for addressing questions of metazoan development has been long appreciated (Kenyon, 1988). Complete anatomical and cell-lineage descriptions (Sulston *et al.*, 1983; Sulston and Horvitz, 1977), in concert with the ease of genetic manipulations of this species (Brenner, 1974) have enabled detailed analysis of processes essential to multicellular life such as sex determination (Hodgkin, 1988), intercellular signalling (e.g. Greenwald, 1985) and cell fate specification (e.g. Priess and Thomson, 1987). *Arabidopsis thaliana*, with a similar genome size to *C.elegans*, is a more recent model organism selected for whole genome sequencing (Schmidt and Dean, 1993) and is used to study plant specific processes such as flowering (Coupland, 1995) and plant morphogenesis (Lord *et al.*, 1994).

The genetic tractability of a model system need not be a determining factor for all genome sequencing projects, however. Clinical considerations are to the fore in the genome sequencing projects of *Homo sapiens* (Collins and Galas, 1993), and the pathogenic bacteria *Mycoplasma genitalium* (Fraser *et al.*, 1995) and *Haemophilus influenzae* (Fleischmann *et al.*, 1995). New disease loci are being found at the rate of several per month by the Human Genome Project compared to a few per year before the project began (Collins, 1995). The genome of the commercially important rice crop plant, *Oryza sativa*, is also the subject of a genome sequencing project (Havukkala *et al.*, 1995).

Sequencing of eukaryote genomes relies on the prior construction of physical maps representing the genome as overlapping contiguous arrays of genomic clones (Coulson *et al.*, 1986; Olson *et al.*, 1986). Sequences obtained from each individual clone can then be assembled into continuous genomic sequence. Techniques pioneered in one organism have been often adopted in others. For instance, restriction enzyme based fingerprinting was pioneered during constructions of the *C.elegans* physical genome map (Coulson *et al.*, 1986). The technique involves comparing the restriction digest profiles of individual cosmid clone ends to identify those clones which contain the same genomic sequence. Not only are overlapping clones discovered, and the physical map thus made more contiguous and complete, but the number of clones necessary to produce full genome coverage is reduced because only those overlapping at their extremities are included. The fingerprint method is now increasingly used in construction of the human physical map (Waterston and Sulston, 1995). Both mapping and sequencing produce large amounts of data which must be collated and integrated, tasks requiring the processing power of computers. A computer database, ACeDB, was designed to perform these functions for the *C.elegans* genome project (Durbin and Thierry-Mieg, 1991) and now provides a computer platform for many other genome projects, including those of the mouse, *A. thaliana* and *H. sapiens*.

Sequencing of prokaryote genomes, which are generally much smaller then those of eukaryotes, has recently been performed by "whole genome random sequencing"; e.g. the 1.8Mb genome of *H.influenzae* (Fleischmann *et al.*, 1995), the 1.6MB genome of

Methanococcus jannaschii (Bult *et al.*, 1996) and the 0.6Mb genome of *M.genitalium* (Fraser *et al.*, 1995). This technique requires no prior construction of a physical map for sequencing to be accomplished, relying on computer processing power to assemble sequences of randomly generated genomic restriction enzyme fragments into continuous whole genome sequence (Fleischmann *et al.*, 1995). Avoidance of the effort required in physical mapping allows very rapid sequencing of small genomes. *M.jannaschii*, a methanogenic archaebacterium, was discovered in 1982, genome sequencing begun in 1995 and finished in 1996 (Bult *et al.*, 1996). The prospect for prokaryote genome sequencing is obviously much enhanced by this novel approach.

"Expressed Sequence Tag" sequencing projects.

cDNA sequencing projects have also been initiated. Their value has been much appreciated in organisms with larger genomes due to the ability to identify many genes quickly(e.g. Adams *et al.*, 1991). The *C.elegans* genome, for example, contains only 15% coding sequence (Waterston *et al.*, 1992). Much of the effort of genome sequencing will thus be directed against non-coding DNA. This effect is even more exacerbated in the human genome, in which only 3% of sequence is estimated to code for protein (Brenner, 1990). In addition to rapid accumulation of protein coding sequence, cDNA sequence has served to confirm exon predictions from genome sequence in *C.elegans* (Wilson *et al.*, 1994) and human (Adams *et al.*, 1991). Estimation of total gene number can also be done with large scale cDNA data due to the random distribution of the expressed sequence tags (ESTs) obtained. The calculation is done by correlation with predicted genes in a defined genomic interval (Waterston *et al.*, 1992). A reliance on transcript sequencing has drawbacks in terms of study of other aspects of whole genomes, however. First, the representation of genes in the cDNA libraries is biased towards abundantly and moderately expressed genes, less abundantly expressed genes being effectively absent from the cDNA representation. Use of a normalised library - produced by cyclic hybridisation of already picked clones against the whole library and subsequent addition of non-hybridising clones to the normalised library - can partially alleviate this problem but cannot avoid the fact that genes expressed at very low levels are represented with a

probability proportional to their relative abundance *in vivo* (Waterston *et al.*, 1992). Second, information not present in coding sequence is not gained, of particular importance being transcriptional control sequences. Understanding of genetic pathways is essentially done in the context of control elements (Waterston and Sulston, 1995). Finally, many aspects of global control of chromosome function are dependant upon non-coding sequence, e.g. telomeres, centromeres, replication origins (Yanagida, 1990).

Identifying genes in genomic sequence.

The first step after generation of the raw genomic sequence is to identify the genes encoded therein. In *E. coli* and other prokaryotes, the lack of introns means that simple identification of long open reading frames (ORFs) is sufficient to identify many coding sequences. Consensus sequences such as upstream TATA box transcription initiation sites, Kozak sequences identifying translational start codons, and downstream transcriptional termination sites further aid gene identification (Daniels *et al.*, 1992). Homology to cDNA clones and other sequenced genes are also used.

Eukaryotic genomes, which do include introns, present greater difficulty for gene prediction (Fickett, 1982). In *C.elegans* and *S. cerevisiae*, conserved consensus sequences of intron/exon splice sites (e.g. Fields, 1990) along with ORF analysis identify probable exonic sequences (Sulston *et al.*, 1992; Dujon *et al.*, 1994). *C.elegans* has the added advantage of transpliced leader sequence splice sites just upstream of the translation initiation codon of many genes (Spieth *et al.*, 1993). Trans spliced leaders are 22-mer single stranded sequences which are spliced onto the 5' end of many *C.elegans* transcripts, though the biological relevance of this is not fully understood (Krause, 1995). The splice site can be identified in genomic sequence by the presence of a consensus sequence 5 - 15bp upstream of the start codon of a gene (Spieth *et al.*, 1993). Again, cDNA and homology data are used to refine these predictions before the exons are assembled into probable gene structures (Sulston *et al.*, 1992).

Comparative Genomics.

EST and protein homologies are also useful when comparative studies of partial or whole genomes are made. Matches in genomic sequence to foreign genes have been used to identify candidate functional homologues of characterised genes in other organisms (e.g. Tugendreich *et al.*, 1994; Koonin *et al.*, 1995). Such computer based, or *in silico*, screening methods are capable of rapidly identifying homologous sequences, and are much more sensitive than physical procedures (Holm and Sander, 1995).

Whole genome comparisons will be possible as more complete sequences become available. The recent completion of sequencing of the Archaea *M.jannaschii* illustrates the evolutionary significance of this (Bult *et al.*, 1996). It had been hypothesised that the Archaea represented a new order of life distinct from the generally recognised prokaryote and eukaryote divisions, an order sharing the basic cytology of the prokaryotes but being closer molecularly to the eukaryotes (Woese, 1990), a view supported by comparison of individual *Methanococcus* genes to prokaryote and eukaryote homologues (e.g. Brown and Doolittle, 1995). The whole genome comparison performed by Bult *et al.* (1996) with the *M.jannaschii* genome sequence has confirmed its novel Archaeon identity. In particular, metabolic pathways involved in such processes as energy production and nitrogen fixation seem to be closely related to those found in bacterial organisms, but the information processing (e.g. transcription and translation) and secretory mechanisms are more strongly related to those found in eukaryotes. The whole genome approach has thus confirmed, and provided a start point for detailed investigation of, the relative evolutionary histories of Archaea, Bacteria and Eukaryota.

Analysis of the complete set of genes in a genome has also given some idea of the present level of description. In yeast, for example, 65% of all the genes found in the genome sequence have been classified as "functionally characterised", either as a result of direct experimentation in yeast (30%) or by virtue of clear homology with characterised genes in other organisms (35%) (Dujon, 1996). A further 7% (of which only 2-3% are expected to

withstand closer examination) are classed as "partially characterised" on the evidence of less clear-cut homology to functionally characterised genes in other genetic systems. The remaining ~30% either have no significant homology to any other gene or are homologous to genes which have not been functionally characterised. Thus, even in one of the most intensively genetically studied organisms, upto a third of the genes remain completely uncharacterised with respect to function (Dujon, 1996). The number of such orphan genes will, of course, decrease as more gene sequences are added to the databases, but there is likely to be an irreducible fraction which cannot be characterised by sequence homology. Thus, reliance on *in silico* comparative studies of genome sequence to intimate particular gene functions is not likely to address a significant proportion of predicted genes.

Functional analysis of genes.

Phenotype driven analysis.

Historically, primary functional characterisation of genes has been through mutational analysis. Classical genetic procedures start with a phenotype, the genetic mutation responsible then being mapped to the genome and the mutant gene cloned and sequenced (Anderson, 1995). In this way, connections between the linear genome sequence and higher order biological nature are made. Many methods for inducing mutation have been developed, each causing a particular type of genetic lesion. EMS is widely used in *C.elegans* research to induce single base pair changes in genomic sequence, or "point mutations" (Brenner, 1974). Such changes commonly produce missense mutations, in which a wildtype codon is transformed into one encoding a different amino-acid, and nonsense mutations, where a codon is transmuted to a stop codon. Most phenotypes caused by missense mutations are likely to be due to partial loss-of-function of the gene product. More severe loss of function, even to the level of completely removing all functional domains of a gene, is possible with nonsense mutations. Use of other mutagens, such as gamma rays and UV radiation, results in larger scale disruption of the DNA sequence such as chromosomal deletion or translocation events (e.g. Greenwald

and Horvitz, 1980). Such mutants are likely to have more severe phenotypes than the majority of point mutants, as complete loss of function of one or more genes is probable. Such diversity in the types of mutation inducible are important in functional characterisation of a gene (Huang and Sternberg, 1995). Null mutants, i.e. those resulting in complete loss of gene function, reveal overall biological function, whilst less severe loss of function alleles may identify different functions and functional domains in the gene's protein.

Genome sequence driven analysis.

The vast amounts of genome sequence data now being produced make possible the reverse approach, where known gene sequences can be mutated and resultant phenotypes observed (Plasterk, 1995). With this approach, selection is occurring at the level of the genome sequence itself, a gene being targeted for specific inactivation on account of homology with a gene of known function. Techniques exploiting homologous recombination events in mouse (Hasty *et al.*, 1991; Capecchi, 1989) and yeast (Rothstein, 1983) allow the directed replacement of any defined chromosomal segment with DNA supplied *in trans.* Transposon-mediated mutagenesis of specific genes has been developed in the fruit fly (Cooley *et al.*, 1988) and *C.elegans* (Rushforth *et al.*, 1993; Zwaal *et al.*, 1993), gene disruption being accomplished either by simple insertion of a transposon or its subsequent imprecise excision, resulting in deletion of flanking genomic sequence. Both approaches allow many different types of mutation to be induced. Complete removal of gene function is possible by deletion of whole genes. More specific lesions can be engineered *in vitro* such as single base pair changes and deletion of individual functional domains. Homologous recombination allows direct introduction of such mutated sequences in mouse and yeast. The process is more difficult in *Drosophila* and the nematode, but should be achieved by supplying the engineered sequences *in trans* to whole gene deletions generated by transposon excision (Zwaal *et al.*, 1993).

The availability of complete genome sequence provides the possibility of mutating every putative gene to assay for phenotype. Such a systematic project is already underway in

yeast, with a concerted effort to produce strains deleted for each individual gene (Oliver, 1996). The ease of gene disruption by homologous recombination in yeast (Wach *et al.*, 1994; Baudin *et al.*, 1993) means this project is likely to be completed relatively quickly. A correspondingly systematic treatment in *C.elegans* will be more difficult due to the laborious nature of the transposon-mediated mutagenesis method currently used to produce directed genetic lesions in the nematode (Zwaal *et al.*, 1993). A systematic screen must await more facile methods of gene disruption.

Many genes are not mutatable to give an observable phenotype, however. A mutational study in *C.elegans* has highlighted such a genetic population (Park and Horvitz, 1986). Suppressor screens of a range of dominant mutations in the nematode identified many extragenic recessive mutations capable of suppressing the initial phenotypes. When the suppressing loci were outcrossed and assayed for phenotype in a wildtype background, only half were found to cause any obvious phenotype. Reverse genetic approaches have supported this finding. In yeast, complete deletion of each of 55 ORFs in a sequenced portion of chromosome III revealed that only 3 genes were essential for survival under laboratory conditions, and phenotypes for a further 21 were found when the individual deletants were subjected to a series of specific environmental challenges (Oliver *et al.*, 1992). Thus, no observable phenotype could be induced for more than half of the predicted genes. The fact that most of the yeast genome is transcribed into mRNA (Kaback *et al.*, 1979) argues against an explanation suggesting inappropriate prediction of gene coding sequences in genomic sequence. In a P-element insertion saturation screen of 315kb of contiguous *Drosophila* genomic DNA, of 43 identified transcripts only 12 could be induced to cause a mutant phenotype (Bossy *et al.*, 1983). Targeted gene knockouts in the mouse also regularly fail to induce a phenotype (reviewed in Shastry, 1994).

Redundancy of genetic function.

Such results suggest that redundancy of gene function may be a common feature in many organisms. Many clear examples, as defined by genetic test, have been observed (e.g. Lamphier and Ptashne, 1992; Thomas *et al.*, 1993; Bowerman *et al.*, 1992): double

mutants with additional phenotypes to those found in either of the single mutants identify gene pairs which exhibit functional redundancy (Guarente, 1993). Possible mechanisms for generating redundancy have been proposed (Thomas, 1993). In the simplest instance, gene duplication events produce identical copies of a gene, and the duplicates are selected for on the basis of cumulative function. rRNA genes are present at high copy number in all eukaryotes. Multiple copies are required because optimal growth cannot be sustained only with the transcriptional output of one or a few genes (Long and Dawid, 1980). Selection for redundancy on the basis of divergent function will occur when genes encode overlapping functions, which may then provide redundancy for some process, but also have unique functions. The unique functions will be selected for, and the redundant, overlapping functions will be retained as a consequence. Such a case is provided by the *lin-12* and *glp-1* genes of *C.elegans*. They are each a homologue of the Notch gene of *Drosophila*, and encode transmembrane proteins involved in intercellular signalling controlling several inductive developmental processes in the nematode (Yochem and Greenwald, 1989). Single mutants display failure of separate sets of inductive processes: *glp-1* mutants show embryonic phenotypes related to the failure of one of a pair of initially equipotent sister blastomeres to follow an induced cell lineage, and fertility defects due to disruption of mitotic induction in the germline (Priess *et al.*, 1987; Austin and Kimble, 1987); lin-12 mutants have defects in the ventral hypodermis where the structure of the vulva is affected (Greenwald *et al.*, 1983). Double mutants, however, reveal failures in additional interactions leading to an early larval lethality phenotype, which is rarely or never seen in the single mutants (Lambie and Kimble, 1991). These results are consistent with the genes being redundant for action in cells where they are both expressed, but having unique functions in cells where only one is expressed. The two genes lie within a few kilobases of each other and share more than 50% sequence identity (Greenwald *et al.*, 1983), consistent with an evolutionary history of gene duplication followed by sequence diversion and the concurrent acquisition of separate function whilst retaining some overlapping function.

Selection on the basis of process fidelity requires that two genes act together to ensure precise performance of a process. Again, duplication of genes can lead to this situation, as illustrated by the CIN8 and KIP1 yeast kinesin genes (Hoyt *et al.*, 1992; Roof *et al.*, 1992). Mutants with both genes deleted exhibit fatal defects in mitotic spindle formation. CIN8 deletion mutants exhibit reduced fidelity of chromosome segregation at all temperatures, while both CIN8 and KIP1 deletants reduce the maximum viable growth temperature to 26℃, indicating that both genes are redundant for spindle function. The phenotypic character of the set of mutants suggests that the two genes are involved in fidelity of the process of chromosome segregation (Goldstein, 1993).

Redundancy in processes requiring high fidelity can also arise between non-homologous genes and genetic pathways. Such convergent functions act in large metazoan animals to protect against cancer. Multiple pathways are active in restraining cell proliferation, and many cancers only arise by mutation of more than one gene (e.g. Land *et al.*, 1983; Hansen and Cavanee, 1988). Other processes requiring high fidelity of function are widespread through the multiplicity of cellular functions, and include DNA replication and transcription, RNA splicing and translation, and development (Thomas, 1993). It has been proposed that the embryo is privileged in terms of selective pressures on the genetic mechanisms at work during development (Wolpert, 1992). Indeed, there is no evidence that the embryo is subject to the normal pressures on energy efficiency, and the common phenomenon of cell death during development suggests that a certain amount of energetically wasteful activity can be tolerated (e.g. Sulston *et al.*, 1983; White *et al.*, 1991). The only selective pressure will be on the reliability of the developmental process, i.e. production of a reproductively viable adult. In this context, development of many novel alternative mechanisms of differentiation, morphogenesis and cell specification can be readily imagined. Gene duplications, which could acquire new properties by genetic drift, need not be rapidly lost as long as they do not interfere with some essential process, and would provide a pool of material for the evolution of novel functions (Wolpert, 1994; Wagner, 1996). It is interesting to note that the precision of any single developmental event is not likely to be provided through redundancy of homologously similar or

identical pathways or genes. Reliability and fidelity would be best served by multiple non-homologous redundant pathways as such diverse pathways would react differently to environmental variations in the embryo. Fidelity of any process would then be ensured whatever the environmental conditions during embryogenesis.

Such buffering mechanisms do not constitute true genetic redundancy, but on the practical level are difficult to distinguish from it and present the same difficulties in addressing individual gene function. Other genes that do not encode redundant function but have no apparent mutant phenotype will include those which perform a subtle function not easily appreciated in terms of gross phenotype (Oliver, 1996). A gene which enables slightly more rapid mobilisation of stored energy reserves would be positively selected for in a population over evolutionary time scales, but difficult to assign a function to when deletion mutants are under study in a laboratory time scale. Genes may also have more critical functions, but these may only be pertinent under certain environmental conditions. The yeast ORF YCR32w is the longest on chromosome III, but complete deletion resulted in no obvious phenotype (Oliver *et al.*, 1992) Assay of phenotype under many physiological challenges eventually determined that the gene was essential for survival in a low pH glucose-rich environment when the organism was challenged with acetic acid (Chater, 1989). Along with fact that the protein contains several predicted transmembrane domains, these observations have led to the proposal that the gene encodes an acetic acid pump.

Can this approach be pursued for all genes in a genome? In the systematic attempt to do this in yeast, the first aim is to generate strains deleted for each predicted gene. A battery of physiological challenges will be applied to each deletant. This approach is likely to be of limited success with respect to a genome-wide assignation of function for two reasons. First, of the 60-80% of mutants with no obvious phenotype in yeast, a significant fraction are likely to remain uncharacterised even after additional environmental contexts have been tried. Second, many phenotypes may not be very informative with respect to function (discussed in Johnston, 1996). What definitive conclusions can be drawn for a mutant which fails to grow in a high-salt medium, for example? Correlation with other

data, such as which compartment the gene product is expressed in, will undoubtedly serve to refine the functional description of some genes, but it must be expected that many mutant phenotypes will remain uninformative. These reservations will also apply to other gene disruption projects, as evidenced by the discrepancy between genetic units and mutable loci found in yeast, fruit fly and *C.elegans* genomic sequence discussed above.

Mutational techniques can still be used to study those genes which have no apparent phenotype when deleted. Genes encoding redundant function because of multiple gene copies can be investigated through construction of animals mutant for all members of the duplicated gene family. Removal of all copies should remove the collective function of those genes and enable the generation of a synthetic phenotype (e.g. Daignanfornier, 1994). Redundancy due to non-homologous genes and pathways will be less tractable due to difficulty in identifying parallel functions. Enhancer screens (Duncan, 1982; Simon *et al.*, 1991) could be done to identify genes in the same functional pathway as those deletants found to produce relatively uninformative phenotypes. Epistatic studies of genes suspected of being functionally linked, due to experimental results in other organisms for example, would lead to further characterisation (Avery and Wasserman, 1992). Assay for phenotype of single deletants under an increased range of physiological challenges will also define more gene functions (Oliver, 1996). Such an approach requires the experimenters to theorise on all possible functional paradigms: we cannot assume that we will be able to imagine all possible contexts for gene function, a consideration to be taken seriously given the large fraction of the yeast genome remaining uncharacterised after several decades of intense genetic investigation. Thus, the sheer number of uncharacterised genes argues against reliance on just these techniques to provide a systematic survey of genetic function.

<u>Novel genes are likely to encode unappreciated regulatory functions.</u>

It is difficult to imagine that many more basic metabolic functions remain to be discovered. Novel functions are likely to be of unstudied, probably unconstitutive, regulatory pathways such as those involved with integration of metabolic pathways,

response to environmental changes and spatiotemporal control of gene expression during development and the cell cycle. The orphan genes found at high frequency in the genomes of those organisms subject to genome sequencing are interesting in this light (Dujon, 1996; Wilson *et al.*, 1994). Orphan genes are, by definition, not mutatable to give a discernible phenotype and so are likely to be involved in redundant or buffering roles. They are thus excellent candidate genes for any, as yet unappreciated, regulatory functions during development and in specific response to sudden physiological challenges.

It is thus clear that analysis of orphan gene function, as well as genes involved in genetically redundant and buffering pathways, from a purely mutational perspective will not be a facile task. Other techniques will be required to address efficiently the question of what all these currently uncharacterised genes are doing. Even for those genes which can be mutated to give a phenotype, additional types of information will be useful in more fully describing gene function.

What other gene descriptions are of functional significance?

I will relate three other facets of genetic description which have proved useful in defining gene function. Each of these three alternative perspectives relies on knowledge of gene sequence and thus are of great potential for investigation of the huge number of new genes being uncovered by genome sequencing projects. They have previously been regarded as supplying corroborative evidence for function primarily derived from genetic analyses. Given the intractability of certain genes to traditional genetic analysis, such characterisations may need to be viewed as more telling descriptions of gene function. A shift from reliance on phenotype based definitions of biological function will be needed.

The first type of gene description is afforded by sequence homology to genes and proteins in the sequence databases. The remarkable cross-phylum conservation of many genes and genetic pathways has resulted in the increasingly common assumption of function for newly sequenced genes containing high sequence homology to genes already characterised in other organisms (e.g. Bork *et al.*, 1992). Other levels of sequence

homology can also prove useful such as motif homologies including nuclear and other organelle localisation signals, sequence motifs of enzymes such as ATPases, and protein domain matches such as EGF repeats (Oliver, 1996). These types of homology may give some idea of the biochemical function of a protein - e.g. a protein containing a zinc finger motif (Klug and Rhodes, 1987) may be assumed to act as a transcriptional regulator of other genes' expression (e.g. Zarkower and Hodgkin, 1992) - without shedding much light on the biological function of the gene. In as yet functionally uncharacterised genes, they should be regarded as providing a subsidiary part of a gene's description, of potential value when considered in combination with other types of description. Homology in non-coding sequence can also provide information representative of the second class of gene description. Many transcription factors bind to conserved sequences upstream of their target genes. When many binding sites for a specific transcription factor have been identified, a consensus sequence can be defined (Jamieson *et al.*, 1996). Discovery of such consensus binding sites upstream of a gene can be used to identify candidate target genes for specific transcription factors and regulatory pathways (Fondrat and Kalogeropoulos, 1994). In concert with genetic pathways defined through techniques such as epistasis and suppressor screens, specific cascades of directly interacting genes can be identified.

Recently, molecular methods for identifying genes common to particular pathways have been developed. Representational Difference Analysis (RDA) was originally developed to isolate differences between two genomic DNA populations (Lisitsyn *et al.*, 1993), but was subsequently modified for differences between mRNA populations (Hubank and Shatz, 1994). It has been used to isolate and identify transcripts dependant on the presence of specific transcriptional regulators (Braun *et al.*, 1995), and relies on the expectation that the level of expression of genes in a regulatory pathway will depend on the strength of their transcriptional activation. Further approaches pioneered in *E.coli* (Casadaban and Cohen, 1979) and more recently adopted in yeast (Olesen *et al.*, 1987; De Winde and Grivell, 1992) involve expression of a library of *lacZ* fusion genes in two host strains differing only in the presence or absence of a particular transcriptional regulator.

Gene fusions under the direct transcriptional control of the regulator will be differentially expressed in the two strains. The two-hybrid system (Fields and Song, 1989) has isolated directly interacting proteins from many organisms (e.g. Elion *et al.*, 1993; Luban *et al.*, 1993), and requires no knowledge of phenotype although the genes investigated may well have been identified with genetic methods. This approach exploits reconstitution of the GAL4 transcriptional activation system in yeast (Ma and Ptashne, 1987) through the physical association of one hybrid species consisting of a protein "X" fused to a deleted GAL4 protein and another hybrid comprising a protein "Y", which interacts directly with protein "X", and an activating GAL4 protein fragment. The related one-hybrid system has been used to identify direct protein/DNA interactions (Inouye *et al.*, 1994). DNA/protein interactions can also be identified by screening of a cDNA expression library with labelled oligonucleotides (Singh *et al.*, 1988; Vinson *et al.*, 1988). The oligos used represent genomic sequence that has been shown to be necessary and sufficient for tissue-specific expression of a gene, and is thus expected to contain specific transcription factor binding sites. This method has been used in *C.elegans* to isolate the *ceh-22* gene, a transcriptional regulator of the *C.elegans myo-2* gene (Okkema and Fire, 1994).

The third aspect of a gene's description I will consider is its spatio-temporal pattern of expression; that is when and where the gene is expressed during development. Such information can be indicative of function. A gene which is expressed transiently in early embryos, for example, is likely to be involved in activity relevant to embyronic development (Hope, 1994). If that gene also has homology to a transcription factor and its protein is localised to the cell nucleus, a function controlling gene expression in early embryonic cells would be strongly suggested (e.g. Bowerman *et al.*, 1993).

The EST approach to genome analysis often provides expression pattern information in that the cDNA libraries are usually made from mRNA of specific tissues or developmental stages. Indeed, expression profiles may be obtained using EST clones in northern analysis of mRNA prepared from many different tissues and cell lines (Orr *et al.*, 1994; Matsubara and Okubo, 1993). More refined data may be generated by examining transcript expression *in situ* (Tautz and Pfeifle, 1989; Mitani *et al.*, 1993), when precise

cellular and stage specific information can be gathered. The potential of these techniques in genome analysis is being realised with a project to find the mRNA expression pattern of all *C.elegans* genes represented by an accumulation of tag sequenced cDNAs (Yohara, 1996).

An independent and complementary perspective on gene expression is revealed by visualisation of protein accumulation. Differences revealed between protein and transcript distributions for a single gene have provided a starting point for study of postranscriptional control of gene expression (Wightman *et al.*, 1993; Evans *et al.*, 1994). The highest resolution assay of protein expression patterns is provided by immunocytochemistry. The laborious nature of monoclonal antibody production rules out systematic studies of protein expression with antibodies specific to each gene, however. Epitope tagging has overcome this obstacle through incorporation of specific antigen epitope sequences into genes' coding sequences (Munro and Pelham, 1987), allowing a single monoclonal antibody to be used for histochemical staining of each tagged gene. In organisms where chromosomal integration is facile the protein localisations observed are those of the endogenous gene as the epitope sequence is inserted into the true genomic context of the gene (Soldati and Perriard, 1991).

Reporter enzyme genes have also been used to monitor protein expression. Such genes allow more sensitive assay of protein localisation than do immunochemical techniques as the histochemical staining procedures exploit the enzymatic character of the reporter (Lis *et al.*, 1983; Goring *et al.*, 1987). Reporter genes can act as epitope tags in gene fusions, and so can also provide high resolution expression patterns when a monoclonal antibody to the reporter enzyme is used (Fire, 1992). The *E.coli* enzyme beta-Galactosidase (β-Galactosidase), encoded by the *lacZ* gene, has been the most effective reporter gene used in a variety of organisms (reviewed by Silhavy and Beckwith, 1985). Enhancer trapping techniques in *Drosophila* (O'Kane and Gehring, 1987) and the mouse (Allen *et al.*, 1988) utilise a *lacZ* gene with a weak promoter which is randomly integrated into the genome. β-Galactosidase expression is directed by enhancers close to the site of insertion, and accurately mimics the expression pattern of those genes close by (Bellen *et al.*, 1989).

Gene trapping is a similar technique which uses a *lacZ* gene with a 3' splice acceptor sequence just upstream. Random genomic insertion of this construction enables β-Galactosidase expression only when the reporter is translationally fused to a gene coding sequence (Gossler *et al.*, 1989). A promoter trapping approach has been used in *C.elegans* (Hope, 1991). A library of plasmid constructs consisting of a promoterless *lacZ* gene fused to random genomic fragments is constructed *in vitro*. Individual constructs which contain *lacZ* translationally fused to the 5' ends of genes will give an expression pattern when present as transgenes *in vivo* (Hope, 1991; Young and Hope, 1993).

Expression patterns generated by reporter fusions have proved to be of great experimental value in many ways, in addition to suggesting possible functions for specific genes (Hope, 1994). Firstly, they have provided markers of cellular differentiation. In *C.elegans*, this attribute has enabled cell specific evaluation of experimentally induced perturbations: e.g. the phenotypic consequences of mutations (Priess *et al.*, 1987), and the effects of laser ablation of particular cells (Sulston and White, 1980). Secondly, they provide promoters of use in driving cell specific ectopic expression of other gene products (Mello and Fire, 1995; Perry *et al.*, 1993). Finally, they present the opportunity for study of regulation of expression of a specific gene through their ability to report expression of experimentally derived manipulations of gene 5' regions (e.g. Way *et al.*, 1991; Okkema *et al.*, 1993; Okkema and Fire, 1994).

Scope of thesis.

Each of the random approaches detailed above suffers from the need to clone and characterise the gene whose pattern of expression is revealed. The promoter trap approach, for instance, requires initial assay of reporter fusions in pools of 96 as the vast majority will not be active (Hope, 1991). Much subsequent work is required to isolate the active fusion and to clone the gene (Hope, 1991; Young and Hope, 1993; Hope, 1994). The data from the *C.elegans* sequencing project allows a new, directed approach to be envisaged. My PhD project set out to evaluate the applicability of a directed screen of developmental expression patterns for novel genes predicted in *C.elegans* genome sequence. I describe such a pilot screen, and present further analysis of one of the genes identified.

Chapter 2

Methods and Reagents.

All hardware and solutions were sterilised by autoclaving before use, unless noted otherwise. All chemicals were AnalaR grade.

Common recipes.

LB agar

Tryptone 10g, yeast extract 5g, NaCl 5g, agar 15g, in 1l H_2O.

2xYT

Tryptone 10g, yeast extract 5g, NaCl 5g, in 1l H_2O.

TE (pH8)

10mM Tris-Cl (pH8), 1mM EDTA (pH8).

NGM agar

NaCl 3g, Peptone 2.5g, agar 17g, 1ml cholesterol (5mg/ml in ethanol), 975ml H_2O. Autoclave, then add 1ml 1M $CaCl_2$, 1ml 1M $MgSO_4$, 25ml 1M KP_i (pH6).

M9 buffer

NaCl 5g, Na_2HPO_4 6g, KH_2PO_4 3g, in 1l H_2O. Autoclave, then add 1ml 1M $MgSO_4$.

Microbiology.

Cosmid strains.

Cosmids were ordered from Alan Coulson at the Sanger Centre, Cambridge, by email (alan@sanger.ac.uk). On arrival, instant stocks were made by transferring some of the stab culture into 20% glycerol with a sterile loop. This is to ensure that an undeleted source of cosmid is always available. Also on arrival, cosmids were streaked to LB agar plates containing an appropriate antibiotic and incubated overnight at 37°C for 16h. DNA was then prepared as described below.

Frozen bacterial stocks.

Frozen stocks of both cosmids and plasmids were kept in 20 % glycerol at -80°C. Stocks for liquid culture preps were produced by adding 0.4ml of culture medium to 0.1ml of glycerol. The mixture was vortexed and immediately frozen.

Selective antibiotics.

These were used in selective media at the following concentrations:

ampicillin - 100 μg/ml

kanamycin - 50 μg/ml

tetracyclin - 50 μg/ml

DNA preparations and manipulations.

Preparation of cosmid DNA.

Cosmid liquid cultures were produced by inoculating 8ml of 2xYT containing an appropriate selective antibiotic, and growing in an orbital shaker at 37°C for 16h. 0.4ml was used to make a frozen stock. 7ml was added to a 50ml Falcon tube, and centrifuged @4000g for 10min in a benchtop centrifuge. The resulting pellet was dried by decanting the supernatant and inverting the tube on tissue paper for 5min.

Cosmid DNA was produced from the pellet using the Qiagen tip-20 DNA miniprep kit (Qiagen) using the manufacturers instructions.

Preparation of plasmid DNA.

Plasmid liquid cultures were produced by inoculating 2ml of 2xYT containing an appropriate selective antibiotic, and growing in an orbital shaker at 37°C for 16h. 0.4ml was used to make a frozen stock. 1.4ml was added to a 1.5ml eppendorf tube, and centrifuged @13000rpm for 30sec in a benchtop microfuge. The resulting pellet was dried by decanting the supernatant and inverting the tube on tissue paper for 5min.

Plasmid DNA was extracted from the pellet using the Qiaquick kit (Qiagen) using the manufacturers instructions.

Restriction enzyme digests.

All restriction enzymes used were from New England Biologicals (NEB) and were used in the supplied reaction buffers. Restriction enzyme digests were performed in 10μl reactions as follows:

1-2 μl of a standard DNA miniprep

1 μl 10x reaction buffer

1 μl of each restriction enzyme

(1 μl of 10xBSA) - optional

- make upto 10 μl with dd H_2O (double distilled water). Reactions were usually performed @37°C for 1h unless the incubation requirements for one of the enzymes was different, in which case the lower temperature incubation was performed first.

Double restriction digestion of plasmid MCS.

It was sometimes found that double restriction digestion of sites close together in expression plasmid multiple cloning sites (MCS) was difficult to perform. Double cut

plasmids were thus often produced using expression plasmids containing DNA inserts between the two desired sites of the MCS. The insert acted to separate the sites and enable efficient restriction digestion to occur. Double cut vector was separated from the resulting insert fragment by gel extraction.

Agarose gel electrophoresis.

Restriction enzyme digests were run on 0.7% (w/v) agarose gels made with molecular biology grade agarose made up in TAE (40mM Tris-actate, 1mM EDTA pH8) to separate the products of a reaction. DNA samples were mixed with 5x loading buffer (15% (w/v) ficoll, 0.25%(w/v) bromophenol blue) before gel loading. Gels were run at 70V for 2h. Gels were soaked in 0.5 μg/ml ethidium bromide in TAE for 20 mins after running, and then destained in TAE for 10 mins, to visualise DNA bands when the gel was viewed under ultraviolet light.

DNA gel extraction.

DNA was extracted from agarose gels with the Qiaprep kit (Qiagen) using the manufacturers instructions.

DNA ligation reactions.

All ligation reactions were performed in 10 μl aliquots with T4 DNA ligase in its supplied buffer (NEB). Reactions were set up so that there was a 4x molar excess of insert compared to vector, and were performed @14°C for 2h.

DNA transformations.

Ligation mixes were used to transform bacteria that had been made competent for transformation. 1ml of an overnight saturated culture of bacterial cells (TG1-F or XL1-blue) was used to inoculate 50ml of 2xYT, which was returned to the orbital shaker for 1.5h (TG1-F) or 3h (XL1-blue).@ 37°C. The culture was decanted into a 50ml Falcon tube and pelleted in a benchtop centrifuge cooled to 4°C @3500g for 5min. From this point onwards, cells were kept on ice as much as possible. The supernatant was discarded and the pellet resuspended by vortexing in 17ml of ice-cold transformation buffer (TFB: 5x soln: RbCl 6g, $MnCl_2.4H_2O$ 4.5g, hexamine chloride 0.4g, 10ml 1M $CaCl_2$, 10ml 1M MES (pH6.3), in 100ml ddH_2O - sterilise by filter sterilisation). After standing on ice for 30min, the tube was centrifuged as before and the pellet resuspended by gentle shaking in 4ml TFB. 140 μl DMF was added, then 5 min later 140 μl of 5 μl MES (pH6.3), 27 μl 2-mercaptoethanol, 470 μl ddH_2O was added, and a further 10min later 140 μl DMF was added. 100 μl aliquots were added to 100 μl aliquots of ligation mixture in 10mM Tris-Cl, 10mM $CaCl_2$, 10mM $MgCl_2$. This mixture was allowed to stand on ice for 30min, and was then heatshocked @37°C for 5min. The mixtures were then made upto 1ml with 2xYT and incubated in an orbital shaker @37°C for 1h. Cells were pelleted for 30sec

@13000g in a benchtop microfuge and resuspended in a 50 μl residual of 2xYT. Suspensions were spread onto LB agar plates containing suitable selective antibiotic, and the plates incubated overnight @37°C.

Polymerase Chain Reaction (PCR).

Primer design.

Simple rules were followed in the design of primers for PCR. Primers were selected to be 20nt long (not including additional restriction enzyme sites); to have a GC content as close to 50% as possible; to have each of the nucleotides present in as close to equal quantities as possible; and to have a G or C at their 3' end.

PCR reactions.

All PCR reactions were performed with *Taq* polymerase (GIBCO BRL) in its supplied buffer on a thermal cycler (Omnigene). Reaction mixtures were as follows:

1 μl template DNA - 1/100 dilution of a cosmid miniprep

1 μl of each primer - 5pmol each

200mM dNTPs - 50mM each of dCTP, dGTP, dATP and dTTP

5mM $MgCl_2$

2 μl 10x *Taq* polmerase buffer

1 μl *Taq* polymerase

- made upto 20 μl final volume with dd H_2O. The polymerase was always added last. The whole mix was then overlaid with 20 μl of mineral oil, and briefly spun in a benchtop microfuge before loading onto the thermal cycler.

Cycling parameters.

All reactions were performed with the following cycling parameters:

1. 94 °C for 10min (denatures double stranded DNA)
2. 94 °C for 30sec (denatures double stranded DNA)
3. 55 °C for 1min (primer annealing to template DNA)
4. 72 °C for 2min (polymerase extension)
5. repeat steps 2-4 x30
6. 72 °C for 10min
7. Refrigerate at 4 °C.

After cycling, 5 μl of reaction products were visualised on an agarose gel.

Cloning of PCR products.

All PCR products were cloned using the TA cloning kit (Invitrogen) and following the manufacturers instructions.

The pop-out procedure.

This was used to generate plasmid clones of lambda clone inserts sent in lysate form by the *C.elegans* genome project. 10 μl of phage lysate was mixed with 100 μl of a saturated overnight culture of pop-out plasmid and 20 μl of M13K07 helper phage, and allowed to sit at room temperature for 1.25h. 2ml 2xYT (including 100 μg/ml ampicillin) was added and the cultures grown overnight @37°C. 1.2ml of the overnight culture was spun @13000rpm for 2min to pellet the cells. 5 μl of supernatant was added to 100 μl of an overnight XL1-blue culture and the mixture allowed to sit at room temperature for 1.25h. Cells were streaked to an LB agar plate (including 100 μg/ml ampicillin) and colonies picked after overnight incubation @37°C.

DNA sequencing.

All DNA sequencing was performed using the dideoxy sequencing technique on single stranded DNA generated from M13 phagemids mp18 and/or mp19.

Subcloning into M13.

A standard ligation reaction consisting of the insert to be sequenced ligated into the MCS of the M13 phagemid was transformed into competent XL1-blue bacteria (prepared as described above). Immediately after heat shock, 2.5ml of 2xYT top agar (2xYT, 0.75% (w/v) bactoagar) @ 50°C, 1ml 2xYT, 25 μl IPTG (25 mg/ml) and 25 μl X-gal (25 mg/ml in DMF) were added to the transformation mix. This was briefly mixed and immediately poured onto LB plates, and plaques allowed to develop overnight @ 37°C.

Preparation of single stranded DNA.

A 2ml overnight culture of XL1-blue cells was used to freshly inoculate 50 ml of 2xYT. This was incubated in an orbital shaker @ 37°C for 1.5h. 1.5 ml of the resulting culture was then placed into a sterile 5 ml tube, and inoculated with a freshly picked clear plaque of transformed XLI-blue using a sterile pasteur pipette to transfer the plug. Incubation was continued for a further 8h. The culture was then pelleted by centrifugation @13000g for 5min. Avoiding disturbing the pellet, 0.8 ml of the supernatant (containing single strand phagemid DNA) was removed and added to 200 μl 20% (w/v) PEG6000, 2.5M NaCl, mixed, and left to stand for 30min at room temperature. The phagemid DNA was then pelleted by centrifugation @13000g for 5min, and the supernatant discarded. The pellet was then suspended in 110 μl TE (pH8) and extracted with 50 μl phenol. After

another centrifugation, 100 µl of the upper aqueous layer was removed to a fresh tube and precipitated by addition of 10 µl 3M sodium acetate (pH5.2) and absolute alcohol. After incubation @-20°C for 30min, DNA was pelleted by centrifugation @13000g for 10min, and the supernatant removed. The pellet was allowed to dry inverted for 5min, and then resuspended in 100 µl TE (pH8). This solution was reprecipitated as before, and the pellet washed in absolute alcohol. After being allowed to dry, the pellet was finally resuspended in 50 µl TE (pH8) and stored @-20°C.

Dideoxy sequencing.

DNA sequencing was performed with the Sequenase II kit (United States Biochemical) following the manufacturers instructions. Sequence reactions were run on vertical polacrylamide gels.

Preparation of polyacrylamide gels.

Sequencing gels were prepared by pouring a polyacrylamide gel solution between two spaced glass plates (50cm x 21 cm). The plated were cleaned by successive washes in ddH20, absolute alcohol and acetone using lint-free tissues. The inner surface of the luggedplate was then wiped with Sigmacote (Sigma) to enable easy subsequent removal. Wedge shaped spacers were cleaned in the same way and placed at either side of the lugged plate. Onto this was p[laced and clamped the back plate, and the two taped and sealed together along their sides and at the bottom with electrical tape.

Gel solution was prepared by gently mixing 24% (v/v) Sequagel concentrate, 66% (v/v) Sequagel diluent, 10% (v/v) Sequagel buffer (all from Flowgen), 0.8% (w/v) ammonium persulphate and 0.04% (v/v) TEMED in a glass beaker. The mixture was drawn into a 50ml syringe and slowly poured into the space between the two plates. A sharkstooth comb was then inserted to make a large single well when the gel set. The gel was allowed to stand for at least 2h.

Running polyacrylamide gels.

The tape was removed and the plates secured to a Flowgen sequencing rig. The rig was filled with 0.5xTBE (45mM Tris-borate, 1mM EDTA) and pre-run @45V for 30min. The comb was then removed and the well cleared of urea with a pasteur pipette. The comb was replaced teeth-down to form small wells along the top of the gel. The wells were flushed of urea again just before loading of denatured (85°C for 10min) sequence reaction. Gels were run @45V for about 3h, depending on the length of sequence required.

Once run, the lugged plate was prised from the gel, and the gel soaked in 1 litre of 10% (v/v) glacial acetic acid, 15% (v/v) methanol for 45min to wash and fix the gel. The gel

was then transferred to 3MM Whatman paper and dried in a gel dryer. Sequence was viewed after overnight exposure to X-ray film.

Culture and Manipulation of *C.elegans*.

Wildtype strain N2 (Brenner, 1974) was used for all experiments except for observation of male expression when a *him-5* strain was used.

Maintenance of worm strains.

All worm strains were grown on 4.5cm NGM agar petri dishes seeded with the *E.coli* strain OP50 (Sulston and Hodgkin, 1988). Individuals were transferred between plates using a platinum wire pick coated with old "sticky" bacteria. Passaging 4/5 individuals every 3 or 4 days enabled healthy stocks to be kept. Plates were kept at 20°C.

Freezing of worm strains.

5 adult individuals were used to seed a fresh NGM plate. 5 days later, 1 day after the bacterial lawn had been exhausted and the plate was predominantly inhabited by L1 larva, worms were harvested by washing with 1ml M9 buffer. 150 µl aliquots were placed in screw-cap tubes and 150 µl of freezing solution added (NaCl 5.85g, KH_2PO_4 6.8g, glycerol 300 µl, add 995ml water, then 5.6ml 1MNaOH - autoclave, then add 3ml 0.1M $MgSO_4$). The mixture was briefly vortexed and immediately placed in an -80°C freezer.

Samples were thawed by removing from freezer and holding the tube in the palm of the hand until thawed. The tube contents were pipetted onto a fresh NGM plate, and healthy individuals used to found fresh populations after 1-2 days.

Microinjection.

Requirements:

1. Agarose pads. Place 0.06g of agarose into a glass test-tube with 2.5 ml dd H_2O, and place in a beaker of boiling water. As the resulting 2.4% (w/v) solution begins to bubble, withdraw aliquots with a pasteur pipette and place drops onto glass slides. Immediately cover the drops with a coverslip and allow the disc to harden. Withdraw the coverslip with the agarose pad attached by sliding it off sideways and store in an upright position. Its best to do many pads at the same time, but don't re-use the slides on which the pads harden as it can be difficult to remove coverslips from used slides. When you have finished, place the pads in a hot (100°C) oven to desiccate. Store stacked in a coverslip container.

2. Inverted optics microscope (Zeiss Aviovert 10) with attached micromanipulator arm (Narishige). The micromanipulator incorporates a needle holder attached to a high pressure N_2 cylinder (see Mello *et al.*, 1991 for details).

3. A plate of young adult hermaphrodites. Use 4 young hermaphrodites to seed a fresh NGM plate. 3/4 days later, just as the bacterial lawn is cleared, the plate should contain predominantly young adult hermaphrodites.

4. Needles for microinjection. These are made with a dedicated needle-puller (Narishige, PD-5) with borosilicate glass capillaries (GC100TF-10, Clark Electromedical Instruments). Settings were used that resulted in a fine tapering point about 1cm long.

Method:

Recombinant DNA for injection is made by adding 0.5 µl of a typical Qiaquick plasmid miniprep (see above) to 1 µl 500 µg/ml pRF4, 2 µl 5x injection buffer (10% (w/v) PEG 6000, 0.1M KP_i (pH7.5), 15mM potassium citrate (pH7.5)), 6.5 µl ddH_2O. The mixture is vortexed and then centrifuged @13000g for 15min to pellet any particulates in the solution. The injection mix is transferred to the tip of a microinjection needle by means of a mouth pipette holding a drawn out capilliary of a diameter slightly less than that of the needle. Loaded needles are kept horizontally in a humidified chamber until use.

A drop of light mineral oil is placed on an agarose pad, and the pad stuck to the microscope stage with sticky tape. A loaded needle is placed in the micromanipulator and the tip brought into the same focal plane as the surface of the pad. The pressure is turned on and the tip of the needle broken by touching it against debris on the agarose pad. Once a good flow is achieved, the pressure is turned off. The agarose pad is removed and 5 young adult hermaphrodites transferred to it by means of a platinum wire pick. The worms will dry out against the surface of the pad, becoming stuck to it. The pad is returned to the inverted microscope and the needle once again brought into the same focus. The microscope controls are used to centre an immobilised worm in the field of view. The syncitium of the distal arm of one gonad is brought into focus, and the needle point moved into the same focal plane. The needle point is pressed against the worm cuticle close to the gonad syncitium and the stage tapped to force the tip into the gonad. Once the tip is correctly positioned, the pressure is turned on. A wavefront of injected material can be seen traveling away from the tip of the needle and swelling the syncitium. Turn off the pressure when the wavefront reaches the reflex region of the gonad. Withdraw the needle tip and repeat for the other gonad arm. It is advisable to inject all animals on a pad within 5min as longer periods usually result in death.

When injections are completed, remove pad from the microscope stage and place a drop of recovery buffer (Fire, 1986) on top of the worms. The worms should immediately float away from the pad surface. Recovering worms can be left in this state whilst more injections are performed. When all injections are completed, continue recovery of worms by adding a drop of M9 buffer to the pads every 5min for 30min. After this time, the worms should be wriggling strongly in the recovery solution. Transfer individual animals

to seeded NGM plates using a mouth pipette holding a drawn out capillary of a diameter greater than that of a worm.

Generation of stably transformed lines of *C.elegans*.

4 days after injection, pick rolling animals (F1 transformants) from the injected animals' plates. Use 5 rollers to seed new plates. F2 transformants can be picked from the resulting progeny and passaged a further 2 times before being assayed for β-Galactosidase activity.

Embryo preparation.

Embryos could be separated from other life stages by hypochlorite treatment. 3/4 hermaphrodites were used to seed a fresh NGM plate. 3/4 days later, just before the bacterial lawn was exhausted, worms were harvested by washing in 1ml M9 buffer. 0.5ml of hypochlorite solution (300 μl bleach, 80 μl 10M NaOH, 120 μl ddH_2O) was added and the mixture kept at room temperature for 5min with occasional shaking. The released embryos were then pelleted by centrifugation @1000g for 30sec, and washed 3 times in M9 buffer. After the last centrifugation, the pellet was resuspended in 30 μl M9 buffer.

This procedure can also be used to rid populations of bacterial/fungal contamination should it occur. Simply transfer the final egg suspension to a fresh, seeded NGM plate. The resulting population will miraculously be contaminant free.

Histochemical staining for β-Galactosidase activity *in situ*.

The staining procedure is a modification of that developed by Fire (1986). 5 adult hermaphrodites were used to seed a fresh NGM plate. 4/5 days later, just as the bacterial lawn had been cleared, worms were harvested by washing with 1 ml M9 buffer. Worms were centrifuges @1000g for 30sec and 3 μl aliquots pipetted into each well of an 8-well microscope slide. (Embryos prepared by hypochlorite treatment could be used instead.) A coverslip was overlaid and the slide placed on a metal block precooled on dry ice to freeze. The coverslip was flipped off with a razor blade to freeze crack the sample, and the slide immediately placed into a precooled Coplin jar of -20°C methanol to fix for 5min. The slide was then transferred to -20°C acetone for a further 5 min before being air dried on a sheet of tissue paper. 25 μl of staining mix (Fire, 1986) was added to the surface of the slide, and another coverslip overlaid avoiding production of air bubbles and sealed around the edges with nail varnish. Slides were incubated @37°C for stain to develop (upto 24h).

Immunochemical staining for β-Galactosidase activity *in situ*.

Embryos were prepared as described above, and placed in 3 μl aliquots on a BSA coated 8-well microscope slide (BSA coated slides produced by dipping scoured slides in 1mg/ml BSA solution, and air drying). Samples were freeze cracked and fixed as described above. Instead of application of staining mix, however, samples were overlaid with 10 μl per well of a 250x dilution in Tween-TBS (0.5ml Tween 20, 150ml 1MNaCl, 50ml 1mtris (pH7.4), 1.5ml 4M NaOH, 800ml H_2O) of mouse anti- β-Galactosidase IgG antibody (Promega) and incubated overnight @ 4°C in a humidified chamber. Slides were then washed in 3 changes of PBS buffer in a Coplin jar for 10 minutes each change. 10 μl per well of a 50x dilution of FITC-labelled goat anti-mouse-IgG secondary antibody (Promega) with 1mg/ml propidium iodide was then added and the slide incubated a further 2 hours at room temperature in a humidified chamber. The slide was washed in PBS as before, and finally overlaid with 50 μl 10% (v/v) DABCO in 90% glycerol and a coverslip, sealed with nail varnish, and viewed immediately. Slides kept for upto 4 days @4°C.

Observation of stained animals.

Observations of fixed and stained animals were made on a Zeiss Axioplan microscope using DIC optics.

Digital image capture.

Micrograph images were captured onto a Macintosh Quadra 800 HD using a Sony DXC 930P video camera and a Kingfisher NuBus Video Capture Board (Graphics Unlimited). Post-capture image processing was achieved using the Adobe Photoshop 3.0 software.

Chapter 3

A Screen Of Developmental Expression Patterns For Genes Predicted From *C.elegans* Genome Sequence.

Note: some of the work described in this chapter has been previously published (Lynch et al., 1995).

Introduction.

The *C.elegans* Genome Sequencing Project.

The 100Mb genome of *C.elegans* is being sequenced as part of a co-ordinated genome sequencing project (Sulston *et al.*, 1992; Wilson *et al.*, 1994). The project is based on the *C.elegans* genome physical map (Coulson *et al.*, 1986; Coulson *et al.*, 1988; Coulson *et al.*, 1991) which, consisting of over 17000 overlapping and contiguous cosmid clones and 3500 yeast artificial chromosomes (YACs), is among the largest and most comprehensive available for any organism. The value of such a resource was recognised long before the current interest in sequencing genomes (Sulston *et al.*, 1992). This foresight has contributed to the rapid progress of the sequencing effort, such that more genomic sequence and sequenced genes are now available for *C.elegans* than for any other organism and the sequencing is on course for completion in 1998 (Waterston and Sulston, 1995). The first complete genome sequence of a metazoan will therefore belong to *C.elegans*. Publicly accessible ftp sites provide access to all the sequence data, finished and in-progress, for all members of the academic community. At the time of writing 44Mb of finished sequence is available (Figure 3.1).

Predicting Genes in Genome Sequence.

The raw sequence is subjected to several computer analyses before release into the public domain GENBANK and EMBL databases (Wilson *et al.*, 1994). These include several dedicated to identifying potential genes. Finished sequence is first compared to the sequence databases using BLASTX for protein homologies and BLASTN for nucleotide matches (Altschul *et al.*, 1990). The highly conserved consensus sequences for intron/exon splice sites in *C.elegans* (Fields, 1990) allow accurate prediction of gene structures by the GENEFINDER program (Sulston *et al.*, 1992). Results from these analyses, coupled with matches to known *C.elegans* cDNAs (Waterston *et al.*, 1992; McCombie *et al.*, 1992), allow final annotation of the sequence with gene predictions before public release. The greatest inaccuracy in this process concerns prediction of the extreme 5' ends of genes (Jones, pers. comm). Often more than one possible translational start codon can be found for the first predicted exon. Such uncertainty is increased by the presence of transplicing acceptor sites just upstream of approximately 40% (Spieth *et al.*, 1993) of *C.elegans* genes. These sequences (known as 'outrons' (Conrad *et al.*, 1991, 1993)) act as a splice acceptor site in a transplicing reaction unique to *C.elegans* (Krause and Hirsh, 1987; Bektesh *et al.*, 1988; Thomas *et al.*, 1988), but analogous to the conventional intron *cis*-splicing reaction (reviewed by Moore *et al.*, 1993), which places a 22 nucleotide sequence on the 5' end of some *C.elegans* mRNAs (Krause, 1986) The

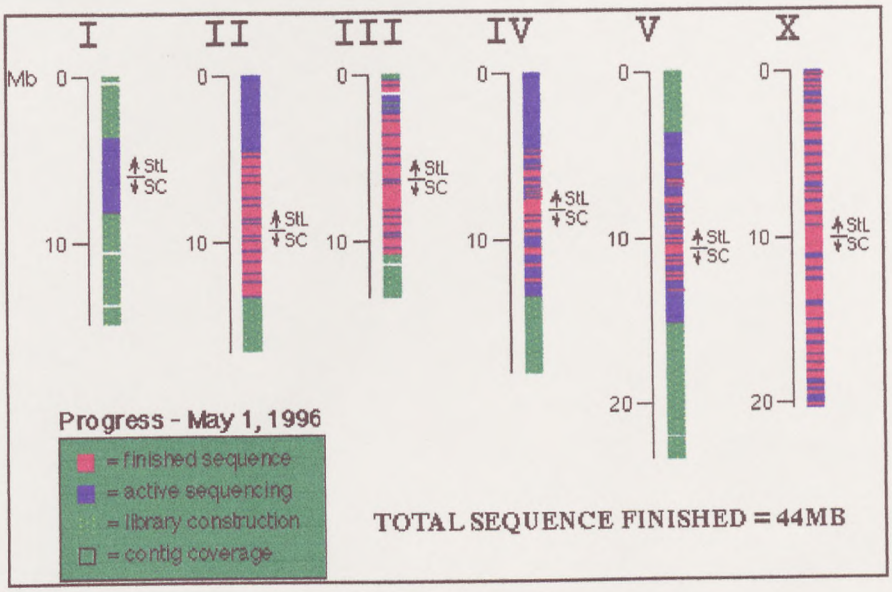


Figure 3.1. Current status of the *C.elegans* Genome Sequencing Project

Sequencing has been started on all chromosomes of the haploid genome. The X chromosome is the closest to completion. The effort of sequencing each chromosome is shared between the two member laboratories involved in the project at St. Louis in the USA (StL) and the Sanger Centre in England (SC). The obvious bias towards sequencing the centres of the chromosomes reflects the relative gene-richness of such regions (Sulston *et al.*, 1992). Concentration initially on these areas means that upto 90% of all C.elegans genes may be characterised in the first 60% of genome sequence completed (Waterston and Sulston, 1995).

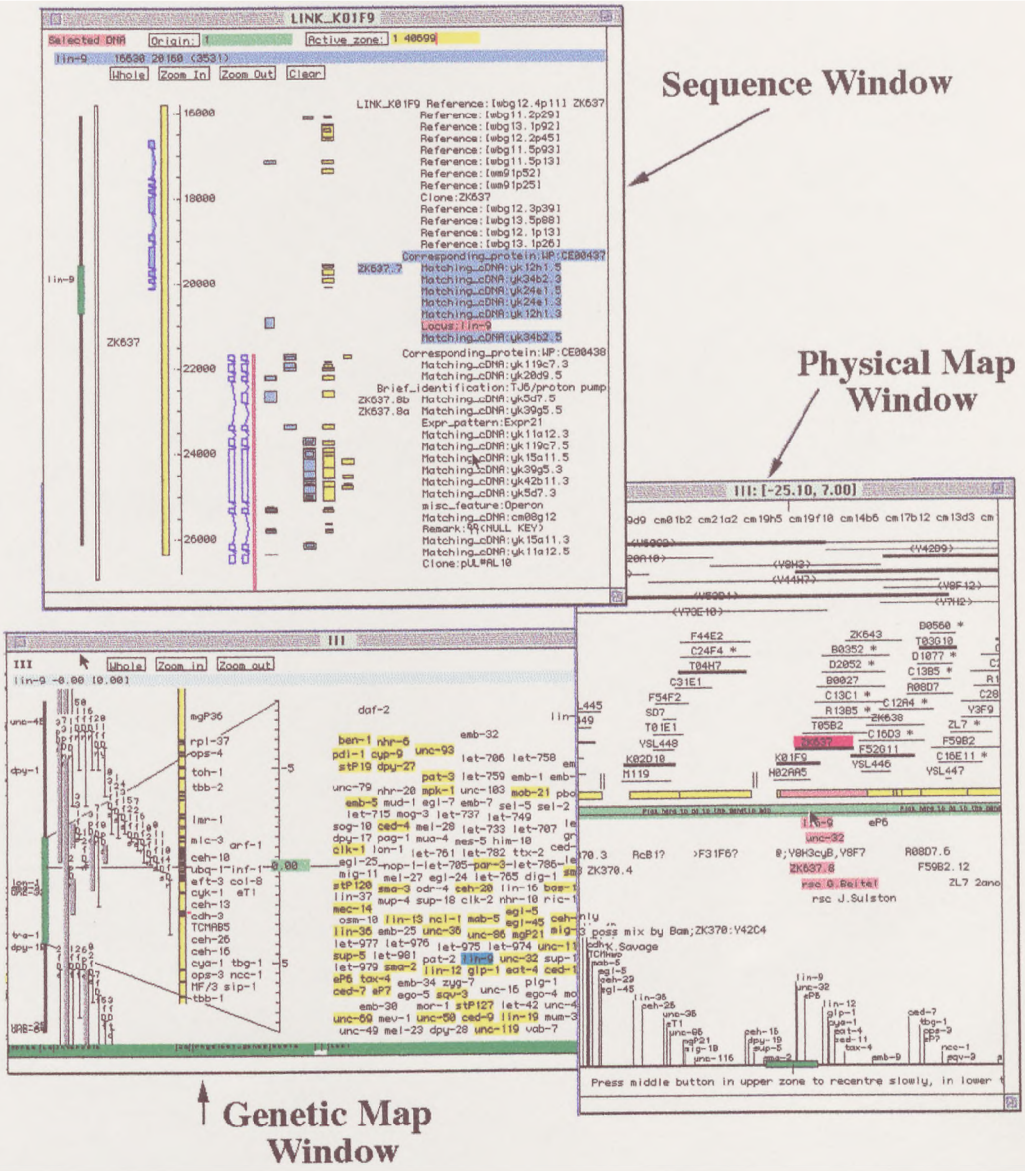


Figure 3.2. The ACeDB database.

Three windows representing the three major facets of information in ACeDB are shown. The sequence window is the format in which the data from the genome sequencing project is displayed. It contains information derived from DNA sequence such as gene predictions (empty blue boxes), cDNA matches (solid yellow boxes) and protein homologies (solid blue boxes). The physical map window contains information on the component cosmid clones of the *C.elegans* physical map (cosmid ZK637 is highlighted in red), and contains links to genome sequence (yellow boxes) and mapped genetic loci (bottom). The genetic map window displays genetic mapping data for *C.elegans* genetic loci (centre and righthand side), as well as other genetic data such as the locations of known deficiencies and duplications (lefthand side).

sequence similarity of outron sequences to intron splice acceptor sequences (Conrad *et al.*, 1991, 1993) may lead to incorrect prediction of the first exon for many genes (Jones, pers. comm.). A graphical and fully integrated representation of all of the genome data (including the physical and genetic maps) is available in a database, ACeDB (Durbin and Mieg, 1991) (Figure 3.2).

A Scheme for Generating Predicted Gene-Reporter Fusions.

The primary aim of my research was to describe expression patterns for genes predicted from the *C.elegans* genome sequence. To screen as many predicted genes as possible, a method involving rapid but specific assay of gene expression must be adopted. Genomic DNA containing predicted gene sequences is readily available from the sequencing labs in the form of the component cosmid clones of the *C.elegans* physical map. Facile construction of *lacZ* reporter fusions for the predicted genes requires the ability to co-locate gene regulatory elements and reporter sequences such that reporter expression depends upon transcriptional or translational fusion with gene specific sequences.

A set of plasmid expression vectors available for analysis of *C.elegans* gene sequences (Fire *et al.*, 1990) allows several approaches to be envisaged. The plasmids comprise a set of interchangeable functional DNA cassettes (Figure 3.3A) which encode features necessary for construction of active reporter fusions. Multiple cloning sites (MCS) 5' and 3' of the *lacZ* gene allow insertion of gene regulatory sequences both upstream and downstream of the reporter gene sequence. Gene specific 5' and 3' flanking sequences can be sequentially inserted into these sites to provide gene specific control of *lacZ* expression (Fire *et al.*, 1990). A similar outcome could be attained by insertion of the *lacZ* coding sequence into an intact copy of the gene. As the genes are available in cosmids which contain approximately 40kb of genomic DNA (Coulson *et al.*, 1991), finding unique sites in common between the *lacZ* MCSs and the gene of interest to allow specific fusion would be unlikely. Smaller genomic DNA fragments would need to be generated to make this approach feasible, such as subcloning of cosmid DNA fragments into smaller plasmid vectors. Both of these approaches would thus involve multiple DNA manipulations and so would be too complex for application in the type of screen intended here, where assay of the maximum number of genes is the main aim. Simple fusion of gene 5' ends to *lacZ* is also capable of producing active reporter fusions (e.g. Hope, 1991). Such constructions require only one DNA subcloning step making them much more desirable in the context of the intended screen.

Other sequences in the expression vectors confer further functionality on potential *lacZ* fusions (Figure 3.3A). Processing and stability of mRNA messages is enhanced in two ways. A synthetic intron is included to exploit the observation that transgene expression in both animals (Brinster *et al.*, 1988) and tissue culture cells (Buchman and Berg, 1988)

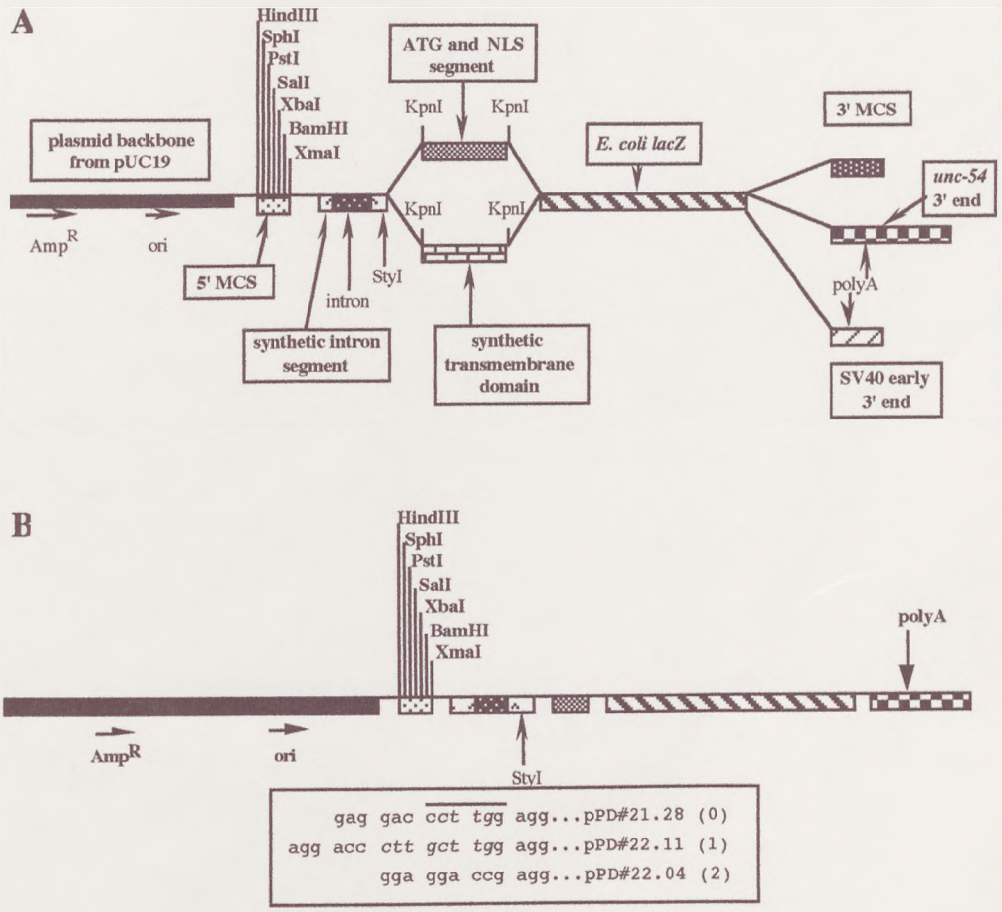


Figure 3.3. Modular structure of C.elegans expression vector plasmids.

A: The vectors made available by Fire *et al.*, 1990, consisted of different combinations of several functional cassettes. The restriction enzyme sites shown are all unique apart from the KpnI sites flanking the segments encoding the nuclear localisation signal and the synthetic transmembrane domain.

B: The form of expression vector used in the expression pattern screen. The StyI site (overlined) is the location of modifications (either duplication or deletion of the 4bp sequence in italics) changing the reading frame between the upstream multiple cloning site and the downstream lacZ reporter gene.

is more efficient for spliced as opposed to unspliced transcripts. Fusions containing no endogenous introns can express more strongly when this DNA segment is present (Fire et al ,1990). For those fusions produced with only 5' gene specific sequences, provision of a 3' gene untranslated region enables maximal signal to be generated. Of particular use has been the 3' end of the *C.elegans* gene *unc-54* (Karn *et al.*, 1983), which promotes *lacZ* fusion expression in a wide variety of *C.elegans* cell types by enabling correct transcript processing (Fire *et al.*, 1990).

Control over the cellular location of *lacZ* fusions is also possible. A transmembrane domain containing cassette can be included to prevent secretion of β-Galactosidase fusion proteins. β-Galactosidase is known to become inactivated on passage through a membrane (Silhavy and Beckwith, 1985) but retains activity when it is located in the cytoplasmic domain of a transmembrane protein (Froshauer *et al.*, 1988). Placement of the synthetic transmembrane domain upstream of *lacZ* in translational fusions to secretory proteins stops cross-membrane transfer of the β-Galactosidase domain, which remains in an active state on the cytoplasmic surface of the membrane (Fire *et al.*,, 1990). Alternatively, a nuclear localisation signal (NLS) cassette containing sequence encoding the NLS of the SV40 T antigen (Kalderon *et al.*, 1984a) can be used. This motif can target proteins to the cell nucleus when present on the N- or C-terminus of a protein or at a variety if internal sites (Kalderon *et al.*, 1984b; Roberts *et al.*, 1987). Nuclear localisation of reporter signal is useful in identifying individual expressing cells (Bonnerot *et al.*, 1987; Hope, 1991).

The NLS cassette also carries an ATG codon which can act as a start codon in transcriptional fusions to *lacZ* (Fire *et al.*, 1990). Transcriptional fusions are produced when regulatory sequences controlling gene specific gene expression are placed upstream of the *lacZ* gene (e.g. Okkema *et al.*, 1993). Tight nuclear localisation of fusion protein is possible with such fusions as the NLS is located right at the N-terminus (Kalderon *et al.*, 1984b), and no gene-encoded protein localisation signals are present. Transcriptional fusions require detailed knowledge of the 5' end of a gene, however, so that gene specific transcriptional control elements can be inserted without any genomic coding sequence. The difficulty in predicting 5' ends of novel genes in *C.elegans* genomic sequence, described above, precludes use of this approach in a screen of their patterns of expression.

Translational fusions contain coding sequence from the 5' end of the gene of interest in addition to transcriptional control regions (e.g. Hope, 1991). Use of such fragments is much more likely to generate functional *lacZ* fusions for the predicted genes in *C.elegans* genomic sequence because, as described above, intragenic coding sequence is easier to identify than gene 5' ends. A critical concern in these situations is that the genomic fragment be in the correct reading frame relative to the *lacZ* gene to ensure production of a functional translational fusion product. 5' MCSs in each of the three possible reading

frames have been generated by modification (Figure 3.3B) of a unique StyI restriction enzyme site flanking the synthetic intron (Fire *et al.*, 1990). Translational fusions have the NLS located at the junction of the fusion peptide, a situation in which the extent of nuclear localisation depends upon the particular molecular context (Roberts *et al.*, 1987). Indeed, previous studies of translational fusions in *C.elegans* have often identified gene fusions in which the NLS is overridden and the fusion protein targeted to other cellular locations (Hope, 1991; Young and Hope, 1993).

Consideration of all of the above points suggested one specific approach for the intended screen. The approach involved production of genomic DNA fusions to a *lacZ* reporter gene such that reporter expression depended upon transcriptional and translational fusion with the 5' end of a gene. The expression vectors to be used (Figure 3.3B) had the following constituent modules: 1) A 5' MCS for insertion of genomic fragments containing gene 5' ends upstream of the *lacZ* gene. 2) A synthetic intron to maximise expression of those fusions containing no endogenous introns. 3) A NLS signal to direct fusion proteins to the cell nucleus, thus concentrating reporter signal and aiding identification of individual expressing cells. 4) A *lacZ* reporter gene to enable histochemical localisation of reporter fusion expression. 5) The 3' end of the *unc-54* gene to provide the necessary signals for correct transcript processing *in vivo*. 6) Three versions of the vector each containing the 5' MCS in a different reading frame relative to *lacZ*. Gene-reporter fusions contained within such expression vectors would be introduced into wildtype *C.elegans* and stable transformants assayed for *lacZ* fusion expression (Figure 3.4).

Selection of Genomic Fragments.

Several criteria were used in choosing the restriction enzyme sites to define the genomic fragment to be inserted into the expression vector. i) For a translational fusion to *lacZ* to result, the 3' site must be in an exon. If possible, sites near the middle of large exons were chosen as, due to the uncertainty inherent in predicting the exact intron/exon splice sites used *in vivo*, these are more likely to lie in actual coding sequence. Sites in regions of high homology to other proteins were preferred for the same reason. ii) The 5' site was chosen to be as far upstream of the 3' site (and the presumed start codon) as was compatible with the approximately 10kb carrying capacity of the expression plasmid. Such sites should maximise the number of putative 5' promoter and enhancer sequences for the endogenous gene included in the fusion construct. iii) Finally, the two restriction enzyme sites chosen for each genomic fragment must enable directional insertion into the vector MCS, i.e. the fragment can only insert in one orientation. This can be ensured by selecting sites defining the 5' and 3' ends of the genomic fragment such that the complementary sites in the MCS are also in 5' and 3' positions relative to *lacZ*. In this

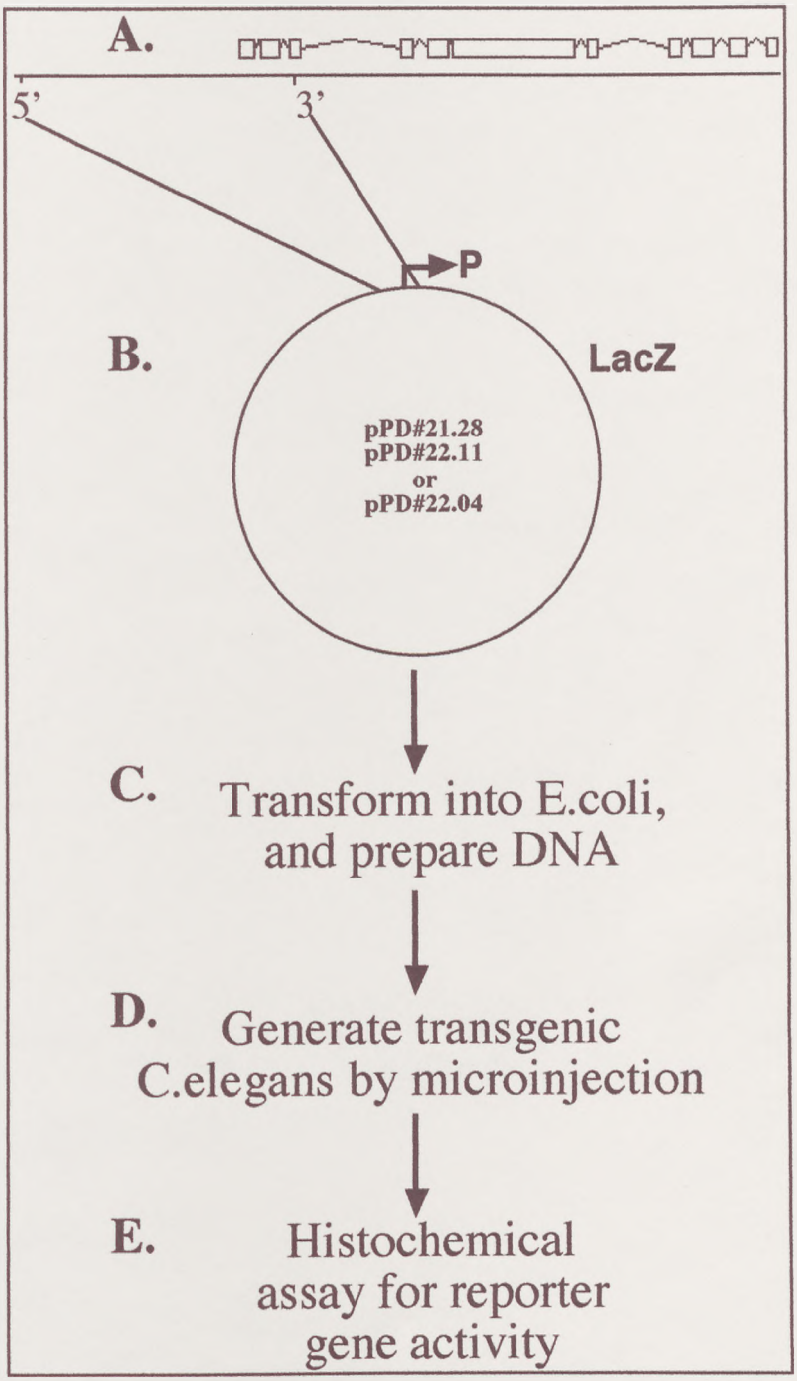


Figure 3.4. Experimental Rationale for assay of expression patterns of predicted genes.

A: Selection of genomic fragments containing gene 5' ends.

B: Translational fusion of gene 5' ends with *lacZ* in an expression vector plasmid.

C: Preparation of recombinant DNA.

D: Generation of worm strains containing lacZ fusion genes.

E: Assay for expression of lacZ fusion genes.

case, all recombinant plasmids should contain the genomic fragment in the correct orientation for transcriptional and translational fusion to result.

Although homology was sometimes used to define the most suitable 3' site, it was not used to bias the selection of genes to be assayed. One of the principal aims of the project was to address the biological relevance of completely novel genes found in genome sequence. Thus, the only requirement in fragment selection was the presence or absence of suitable restriction enzyme sites.

Construction of gene-reporter fusions.

The genomic region covered by this study encompasses 495kb of genome sequence on chromosome III, contained within 14 cosmid clones of the physical map (Figure 3.5). Of the 104 predicted genes in this region, 70 (67%) were found to be suitable for assay (Table 3.1). When the screen was first started, I attempted to generate reporter fusions for all the selected genes on a cosmid before moving on to the next cosmid. This approach was later modified as some genes proved very difficult to sub-clone: to optimise the number of genes examined two attempts only were made for each fusion before work was discontinued on that gene. This resulted in some cosmids being more densely covered than others (e.g. one of the first few cosmids covered, ZK643, had all four of its selected genes processed whilst a later one, R107, had only two of six completed), and only 45 of the 70 selected fusions were eventually made (Table 3.1).

Generation of *C.elegans* strains transgenic for reporter fusion genes.

Reporter fusion plasmids were introduced into *C.elegans* by micro-injection into the syncytial distal arm of the hermaphrodite gonad (Stinchcomb *et al.*, 1985; Fire, 1986; Mello *et al.*, 1991). Recombination events *in vivo* form extrachromosomal arrays of transgenic material which are able to be replicated and distributed stochastically to daughter cells during cell division (Stinchcomb *et al.*, 1985; Mello *et al.*, 1991). Such a segregation mechanism means that transgenic animals are usually mosaic in terms of the cellular distribution of extrachromosomal arrays, leading to mosaicism of patterns (Okkema *et al.*, 1993; Krause et al; 1994). A 50mer oligonucleotide can be co-injected with the reporter plasmid to increase the likelihood of chromosomal integration (Mello *et al.*, 1991). Integrated lines can have much reduced mosaicism of expression as every cell genome will contain copies of the gene fusion (Fire, 1986). Qualitatively, patterns of expression are usually the same whether the genomic context of the reporter gene is chromosomal (i.e. integrated) or extrachromosomal (Mello *et al.*, 1991; Okkema *et al.*, 1993). However, discrepancies have been observed as a result of the chromosomal location of the integration site. For example, a strain bearing an integrated copy of an *unc-54::lacZ* fusion was observed to ectopically express in pharyngeal muscle as well as in bodywall muscle (the gene's normal expression site (Ardizzi and Epstein, 1987;

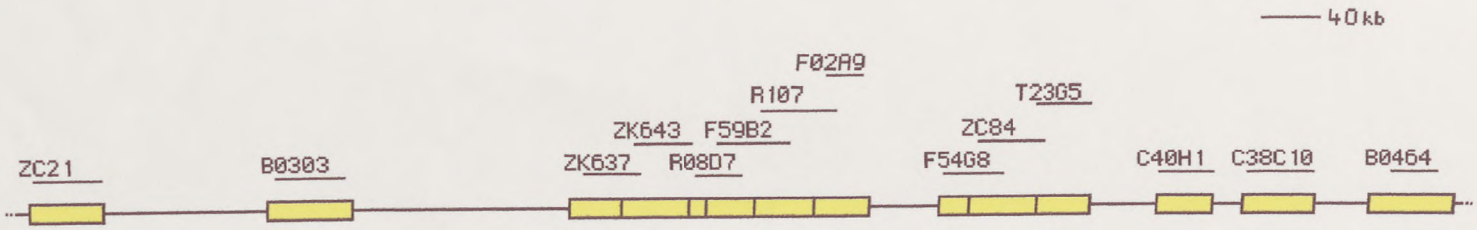


Figure 3.5. The genomic region of chromosome III covered in the screen.

Yellow boxes represent the regions covered on the chromosome by the sequenced cosmid clones which are depicted above the representation of the chromosome. The line under each cosmid clone name is a measure of its actual length in kilobases of DNA (see scale).

Based on output from ACeDB.

COSMID	SIZE (kb)	# genes	# selected	# fusions	# patterns	
B0303	41.074	11	7	5	2	
B0464	40.909	7	6	4	1	
C38C10	34.193	5	4	2	1	
C40H1	27.271	6	4	2	1	
F02A9	26.242	6	5	2	1	
F54G8	31.613	4	4	4	1	
F59B2	43.782	11	6	5	1	
R08D7	27.368	7	5	2	2	
R107	40.97	7	6	2	2	
T23G5	26.926	6	5	3	1	
ZC21	36.087	8	4	4	3	
ZC84	38.955	6	3	2	2	
ZK637	40.699	14	7	4	2	
ZK643	39.534	6	4	4	3	
	495.623	104	70	45	23	Totals
			67.30%		53.33%	

Table 3.1. Cosmids covered in the screen.

Columns:-
COSMID: cosmid name
SIZE(kb): size of genomic insert in kilobases
#genes: number of predicted genes in genomic insert
#selected: number of genes possessing restriction enzyme sites suitable for inclusion in the screen
#fusions: number of genes used to generate *lacZ* fusions
#patterns: number of genes generating an active *lacZ* fusion.

Okkema *et al.*, 1993)). This observation could result from the integration site being close to a chromosomal enhancer of pharynx specific gene expression (Okkema *et al.*, 1993). Integrative transformation was thus not sought during the screen to avoid any similar, misleading events. pRF4 is co-injected with the fusion plasmid to act as a marker of transformation. It carries the *rol-6*(su1006) gene, a dominant effect allele of a collagen gene (Kramer *et al.*, 1990). Worms carrying the *rol-6*(su1006) gene roll to the right as they move, an easily scored phenotype.

Stably transformed strains are established by picking roller progeny of injected animals, followed by replating of rollers through four generations (Stinchcomb *et al.*, 1985). The reporter expression patterns of stable lines can be visualised by histochemical assay for β-Galactosidase enzyme activity (Fire, 1992).

Description of Gene Expression Patterns.

The anatomy of *C.elegans* has been determined at cellular resolution for all stages of development (e.g. Sulston and Horvitz, 1977; Kimble and Hirsh, 1979; Sulston *et al.*, 1983). Comparison of patterns of gene expression obtained in this screen, viewed using Nomarski light microscopy, with published anatomical descriptions permitted the cellular basis of most patterns to be worked out. Exceptions to this occurred most frequently with nerve-specific pattern components. The high concentration of nerve cells in the head makes identification of individual neurones difficult (White *et al.*, 1986). In these cases, however, the ganglion or ganglia in which the nerve nuclei lay could usually be discerned.

At least three independently generated lines for each reporter construct were scored (Table 3.2). This is necessary to allow for the differing structure of each individual extrachromosomal array, expected because of the chance nature of the recombination events by which they are formed (Stinchcomb *et al.*, 1985; Hope, 1991; Mello *et al.*, 1991). Some difference in signal strength was observed between individual lines, but qualitatively identical expression patterns were observed for all transgenic lines of each reporter construct. These observations reflect the expected plasmid copy number differences between extrachromosomal arrays, and suggest that specificity of reporter fusion expression is not compromised by the particular recombinant structure of an extrachromosomal array. Mosaicism of gene expression is expected between individuals of transgenic *C.elegans* strains bearing extrachromosomal arrays (see above). Thus, individual animals of a population need not harbour the complete pattern. The full pattern is constructed from observations of many individuals at all stages of development. Specific expression patterns were observed for 24 (53%) of the 45 reporter constructs assayed (Table 3.2).

gene	homology	plasmid	fragment	frame	5'	strain	# lines	# introns	# cDNA hits
B0303.1	Drosophila Flightless	34	P 10439 -Bs 2434	2	7	UL37	4	3	0
B0303.4	-	35	Xb 7272 -B 11285	0	???	-	3	5	0
B0303.11	-	15	Sa 25443 -Bg 23738	0	0.8	-	7	3	0
B0303.12	-	16	Xb 33443 -A 28265	1	4.4	UL25	8	1	0
B0303.15	ribosomal protein L11	36	Xb 36078 -B 38146	0	1.4	-	6	2	0
B0464.1	aspartyl-tRNA synthetase	41	X 38718 -B 33848	2	3.8	-	3	2	2
B0464.2	-	37	Sa 33378 -Xb 36358	0	1.9	-	4	2	7
B0464.3	-	38	P 25241 -Sa 32399	2	7.2	-	3	0	0
B0464.4	-	42	X 19864 -Se 16882	0	2.9	UL42	3	0	1
C38C10.1	G-protein coupled receptor (tachykinin)	27	H 15518 -N 17754	2	0.4	UL32	6	7	0
C38C10.4	-	28	Xb 26303 -B 21722	1	2.9	-	7	3	3
C40H1.1	ovarian protein (Drosophila)	24	Av 9204 -Bg 14874	0	4.9	-	3	3	0
C40H1.6	-	26	P 18509 -Nh 22793	2	4.2	UL34	4	0	0
F02A9.1	-	44	H 7683 -X 4977	0	2.6	-	3	0	0
F02A9.5	Propionyl-CoA carboxylase	45	H 19835 -Bg 15498	1	3.2	UL43	3	3	6
F54G8.1	-	19	P 3693 -X 6453	2	2.2	-	7	1	0
F54G8.2	Diacylglycerol kinase	14	H 11520 -X 18734	2	2.4	UL24	8	14	0
F54G8.3	integrin alpha chain	21	Se 26400 -Bg 20340	1	2.9	-	7	6	4
F54G8.5	-	17	P 14167 -Bs 9327	0	3.9	-	6	2	0
F59B2.2	-	11	X 6841 -Bg 2063	0	3.1	-	7	4	1
F59B2.5	-	5	N 14051 -X 8397	0	4.5	-	6	1	0
F59B2.9	-	12	P 26869 -X 19224	2	7.6	-	4	0	0
F59B2.12	-	6	X 19224 -B 24635	2	5.4	-	6	0	18
F59B2.13	G-protein coupled receptor (CCK)	7	H 38440 -S 33755	2	2.5	UL20	6	2	0
R08D7.3	-	13	P 10234 -Se 4712	2	4.5	UL23	8	1	2
R08D7.5	caltractin	20	X 601 -Bg 8612	2	7.7	UL27	4	1	0
R107.1	-	32	Sa 15403-Xb 7732	0	5.7	UL35	4	7	0
R107.4	protein kinase	33	N 25377 -Bs 17947	2	6.4	UL36	4	2	1
T23G5.2	SEC14 (yeast golgi transport)	31	Nh 19924 -B 13735	2	5.4	-	6	7	2
T23G5.5	5-HT transporter	29	X 20010 -Xm 26412	2	4.3	UL33	5	9	0
T23G5.6	-	30	Sa 16894 -Bs 21848	2	4.4	-	6	2	0

Table 3.2 contd...

ZC21.1	-	46	N 23006 -A 29325	0	6.2	-	3	0	0
ZC21.2	Transient receptor potential protein	39	N 23006 -B 26920	0	1.7	UL41	4	8	0
ZC21.3	dual bar protein	47	Sa 7316 -A 15370	1	7.6	UL44	6	1	0
ZC21.4	Breakpoint cluster region protein	40	Sa 2406 -Xb 5101	2	1.1	UL60	7	1	0
ZC84.3	-	22	H 12662 -Bg 8713	2	2.8	UL30	5	6	0
ZC84.3a	-	23	H 12662 -Se 8531	2	2.8	UL31	3	6	0
ZK637.3	-	8	P 2163 -X 6327	2	4.1	-	3	0	0
ZK637.5	ArsA - E.coli arsenical pump	9	P 2163 -B 11858	0	8.9	UL21	4	1	1
ZK637.8	proton pump ATPase	10	P 17560 -X 22196	2	4.1	UL22	7	2	8
ZK637.10	glutathione reductase	18	P 23341 -Bs 31362	1	6.5	-	7	4	0
ZK643.1	-	1	Sa 4358 -Se 10132	0	5.2	UL16	5	2	0
ZK643.2	DCMP deaminase	2	Nh 25645 -B 32109	0	6.3	-	4	0	0
ZK643.3	G-protein coupled receptor (calcitonin)	3	N 17883 -B 25209	2	2.3	UL17	6	3	0
ZK643.5	-	4	H 32614 -Se 29841	2	1.1	UL19	3	1	1

Table 3.2. Predicted genes covered in the screen.

Columns:-
gene: name of predicted gene supplying 5' end of *lacZ* fusion. red - B0303.1 is now part of B0523.5. Genes in green had fusions made by David Briggs. Genes in blue had fusions made by Ian Hope.
homology: highest homology of predicted gene
plasmid: plasmid designation (pUL#AL"plasmid desiganation")
fragment: genomic fragment contained in *lacZ* fusion. Coordinates are from Genbank entry for cosmid. A (AgeI), Av (AvrII), B (BamHI), Bg (BglII), Bs (BspEI), H (HindIII), N (NsiI), Nh (NheI), P (PstI), S (SphI), Sa (SalI), Se (SpeI), X (XhoI), Xb (XbaI), Xm (XmaI).
frame: reading frame of expression plasmid vector required to produce a translational fusion with lacZ. 0 - pPD#21.28, 1 - pPD#22.11, 2 - pPD#22.04.
5': amount of genomic sequence in genomic fragment upstream of predicted translationl start (kilobases). ??? - B0303.4 has no predicted start codon.
strain: designation of strains carrying an active *lacZ* fusion. UL22 - an integrated line (UL26) was also generated for this *lacZ* fusion.
#lines: number of transgenic lines assayed for *lacZ* fusion expression.
#introns: number of genomic introns in *lacZ* fusion. 5 - number of introns in known B0303.4 5' end
#cDNA hits: number of ESTs matching gene in ACeDB (September, 1996).

Results.

Specific Gene Expression Patterns Obtained in the Screen.

notes

i) all expression components are nuclear localised unless otherwise stated

ii) all patterns are described as precisely as possible. They are not, however, intended to be definitive and many may be modified in the future as a result of advice from those with more specialised knowledge of particular cells and tissues

B0303.1

This gene has two distinct components of expression. The first begins in a subset of cells of the pharynx during late embryogenesis, endures through the rest of the life-cycle, and is nuclear localised (Plate 3.1A). The subset consists of the m2 muscles and e2 epithelial cells of the procorpus, the m4 muscles in the metacorpus, and the m8 muscle in the most posterior region of the terminal bulb (Albertson and Thomson, 1976).

The second component is restricted to the vm2 vulval muscles and is cytoplasmic (Plate 3.1B). Expression can be detected only in L4 larvae when the vm2 cells are in the final stages of differentiation and become attached to the worm's bodywall in a defined pattern (White *et al.*, 1986). Indeed, expressing cells are seen which appear to be much less elongated than those in adulthood (Plate 3.1C). B0303.1 thus seems to be expressed around the time of the final differentiation of the vm2 cells.

B0303.1 is the *C.elegans* homologue of the *flightless-I* gene of *Drosophila* (Campbell *et al.*, 1993). The *flightless* gene, and its related proteins, has two functional domains: the C-terminal half is homologous to gelsolin, a calcium-regulated actin-binding protein which controls the assembly of actin monomers into filaments, and is also known to bind fibronectin, another component molecule of the cytoskeleton (Lind and Janmey, 1984); the N-terminal portion of the protein is homologous to the leucine-rich-repeat group of proteins known to be involved in protein-protein interactions of the Ras signalling pathway (Claudianos and Campbell, 1995). *Flightless* mutants suggest a role in recruitment of actin to the cytoskeleton, and possibly stabilisation of the cytoskeletal structures so formed (Straub *et al.*, 1996). Weak alleles have abnormal flight muscle morphology, with the myofibrils having a disorganised appearance, indicative of a disruption in the molecular events of myoblast differentiation (Koana and Hotta, 1978; Miklos and de Couet, 1990). Stronger alleles are embryonic lethals, apparently caused by failure of the syncitial embryo to undergo cellularisation (Perrimon *et al.*, 1989).

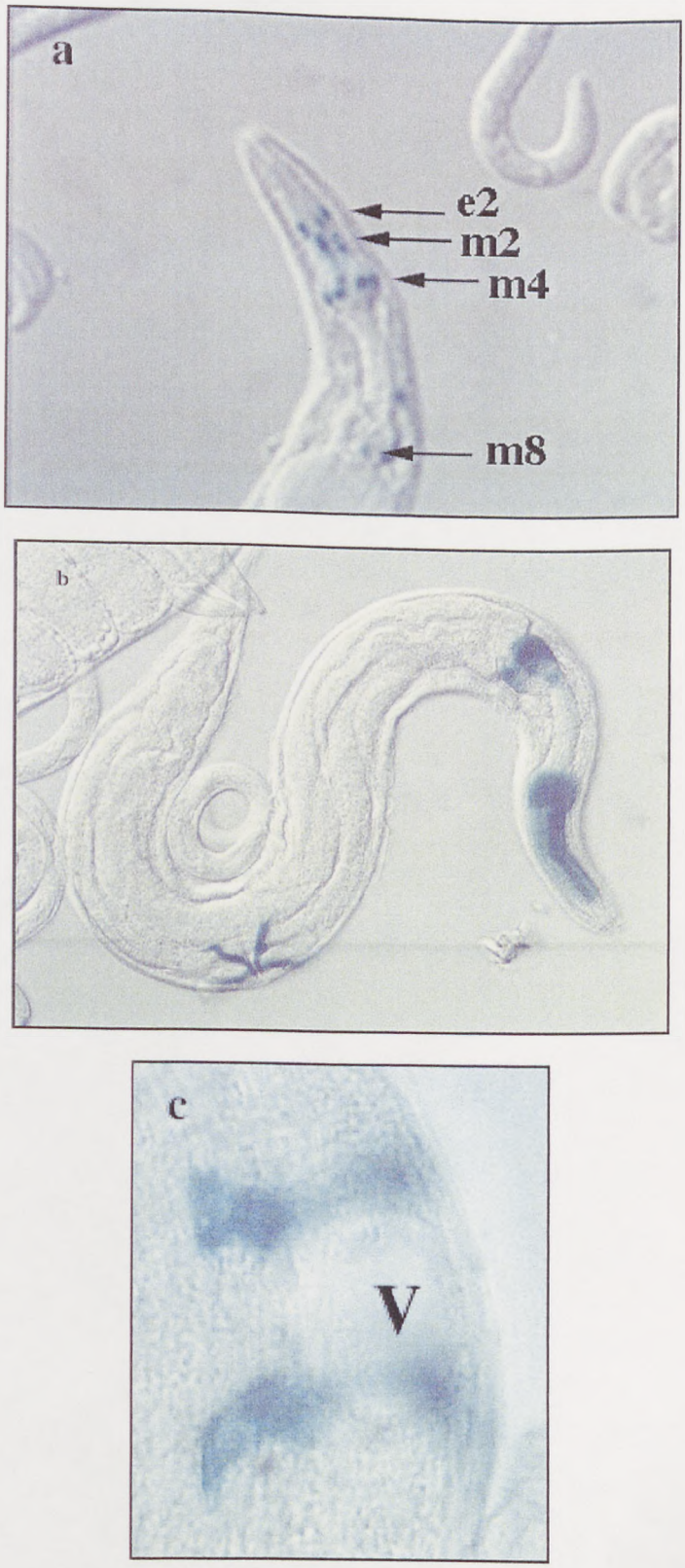


Plate 3.1. LacZ fusion expression in UL37 (B0303.1).
a. Expression in four sets of cells in the pharynx of a young adult hemaphrodite.
b. Expression at the late L4 stage in the pharynx and vm2 vulval muscles.
c. vm2 expression at an earlier time in the L4 stage. The muscle cells are noticeably shorter and wider than those shown in Plate 1.b., as they have not yet become attached to the hypodermis. The open circle of the developing vulva (V) is a good indication that this worm is at the early L4 stage.

Ultrastructural studies of mutant embryos have shown that actin is irregularly associated with the cellularisation membranes (Straub *et al.*, 1996).

Of particular interest in this context is the expression I have observed in the vm2 muscles. Expression is evident when the muscle cells are elongating and attaching to the bodywall, processes which both involve the cytoskeleton. The B0303.1 gene may thus be involved in the cytoskeletal adaptations which underlie these cellular processes.

B0303.12

This gene has two components of expression. In the head, a sinuous line of staining stretches from the hypodermis down into the interior of the worm (Plate 3.2A). The location of the pattern component, close to the terminal bulb of the pharynx, and its singular structure suggest that it identifies the excretory duct (Nelson *et al.*, 1983). This is part of the so-called "excretory system" thought to be responsible for osmotic regulation in the nematode.

Expression from this fusion construct is also evident in all bodywall muscles (Sulston and Horvitz, 1977), and is cytoplasmic. Expression is strongest at the ends of the rows of bodywall muscles (Plate 3.2B).

B0464.4

Five neural nuclei of the anterior ganglion (White *et al.*, 1986) in the head express in transgenic adults (Plate 3.3). Two are already expressing in late embryogenesis, the other three begin to express during the L1 and L2 stages of postembryonic development. This gene has no significant homology to other known genes.

C38C10.1

This gene has an expression pattern restricted to PDEsoL/R and ADEsoL/R (White *et al.*, 1986), and is first visible at the L3 larval stage (Plate 3.4A) when the cells are undergoing terminal differentiation. These are the socket cells of the post-deirid and deirid sensilla respectively, specialised nerve-accessory cells which form a toroid around the sensory process of the sensillum and act as an interface between it and the hypodermis (Ward *et al.*, 1975; Ware *et al.*, 1975). The nuclei of the PDEsoL/R cells are situated in the posterior lateral ganglion in the tail of the worm (Plate 3.4B). Expression ceases during the L4 larval stage.

C38C10.1 has homology to the 7-transmembrane-domain G-protein coupled receptors (GPCRs), and is closest to those receptors for tachykinins (Shigemoto *et al.*, 1990). These neuropeptides are functionally diverse (Nakanishi, 1986). The receptors are

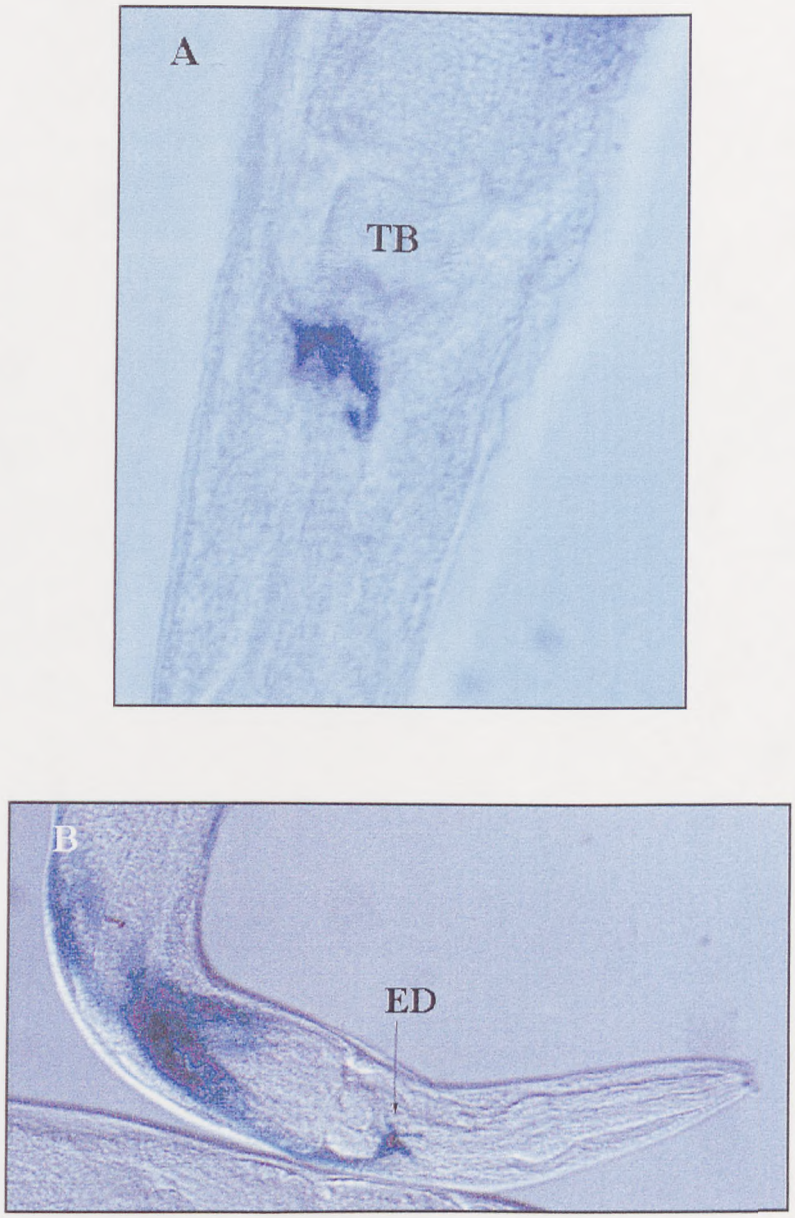


Plate 3.2. *LacZ* fusion expression in UL25 (B0303.12).

A: The excretory duct cell can be seen just posterior to the terminal bulb of the pharynx in the head of an adult hermaphrodite.
B: Another adult hermaphrodite showing expression in the excretory duct cell (ED) and the bodywall muscles. Expression is clearly stronger in the most anterior muscle cells.

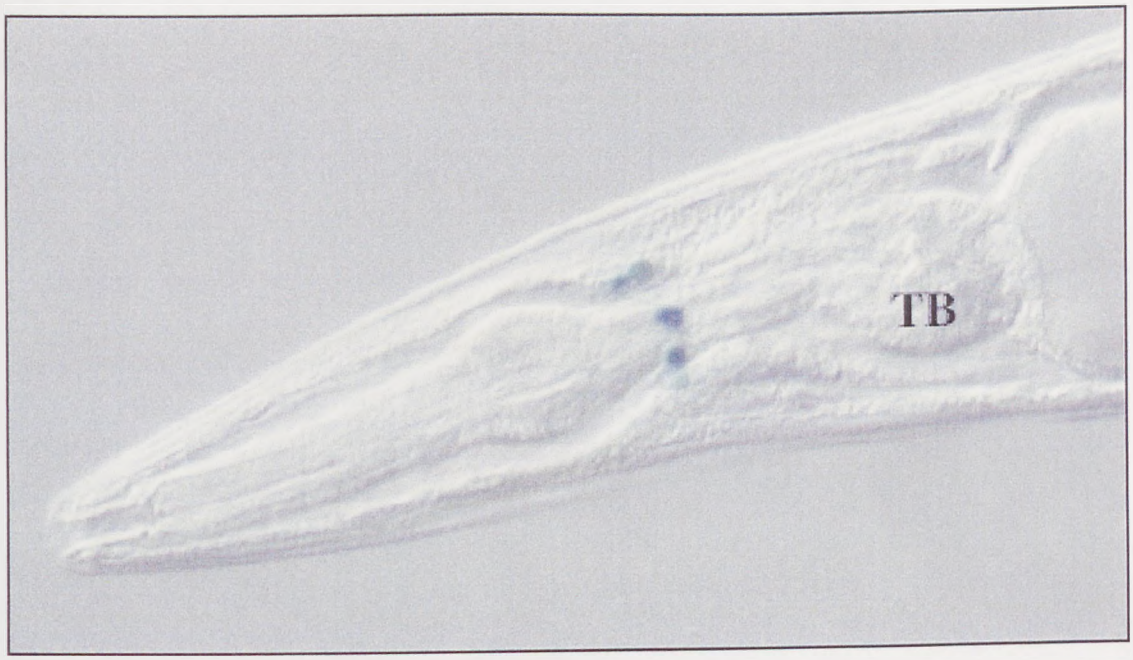


Plate 3.3. *LacZ* fusion expression in UL42 (B0464.4).

5 neural nuclei in the anterior ganglion of an adult hermaphrodite.
TB: terminal bulb of the pharynx.

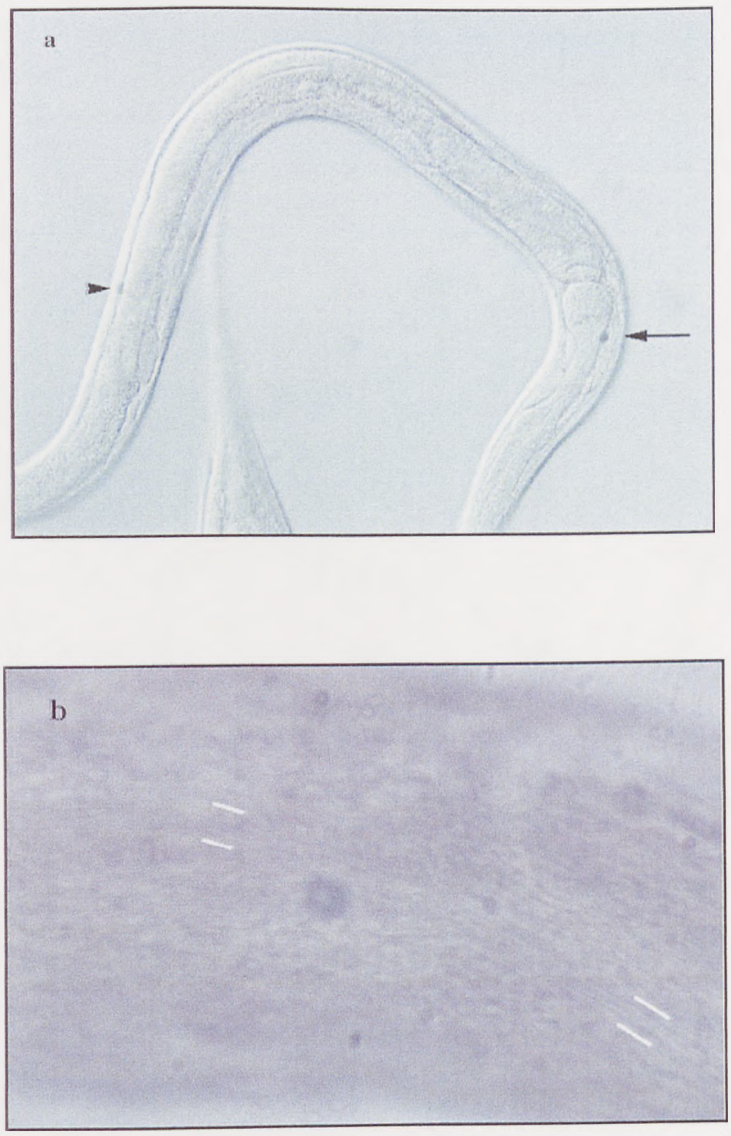


Plate 3.4. *LacZ* fusion expression in UL32 (C38C10.1).

A: Expression can be seen in an L3 larva in the socket cells of the deirid (arrow) and postdeirid (arrowhead) sensilla. In this image, only one member of the pair of each socket cells can be seen.

B: A closer view of expression in a postdeirid socket cell in an L4 larva. The nucleus is located just under the alae of the larval cuticle, a specialised ridge structure running down both lateral sides of the animal. The line of the alae is marked by the sets of white tramlines.

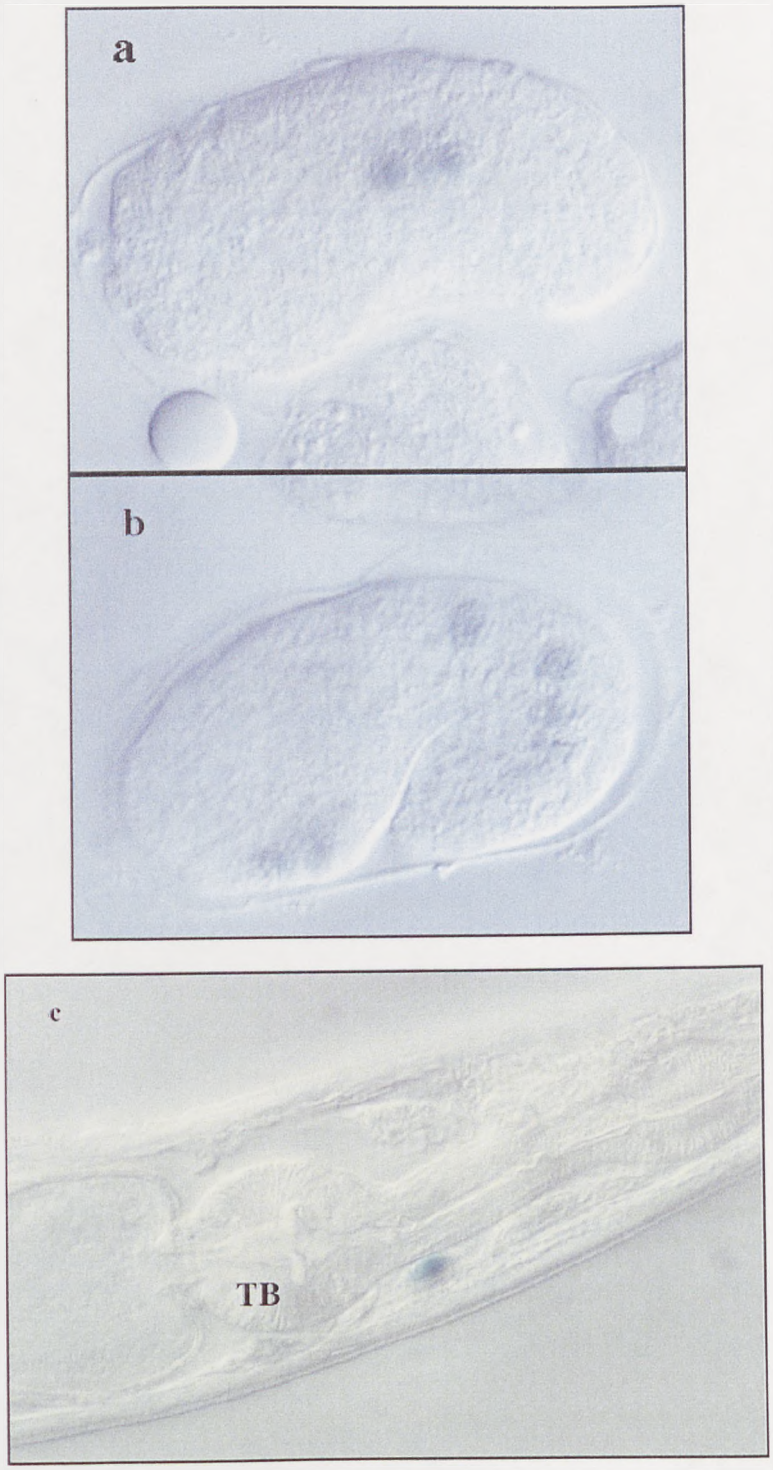


Plate 3.5. *LacZ* fusion expression in UL34 (C40H1.6).

A: 2 nuclei staining in the centre of an embryo just beginning elongation (anterior to the left).
B: An older embryo at the 1.5-fold stage showing expression in ventral nuclei in the anterior (left) and dorsal nuclei at the posterior.
C: An adult hermaphrodite showing expression in the head mesodermal cell. (TB - terminal bulb of the pharynx).

associated with G-proteins in the membrane of receiving cells which activate a phosphotidylinositol-calcium second messenger system (Nakanishi, 1986).

The expression pattern observed for C38C10.1 thus suggests a role in GPCR-mediated signal transduction during final morphogenesis of the deirid and post-deirid sensilla. Specific expression in the socket cells further suggests a role in the formation of a close connection with the hypodermis which is the major cytological specialisation which occurs during late socket cell differentiation in the L3 larval stage (Ward *et al.*, 1975; White *et al.*, 1986).

C40H1.6

Expression begins just before the elongation phase of embryogenesis, when two staining nuclei can be seen in the approximate centre of the embryo (Plate 3.5A). Through the comma stage and up until the 2-fold stage, more nuclei show expression though staining is highly mosaic. The overall pattern at this time consists of two nuclei in each of the dorsal and ventral margins in the head, and four nuclei in the elongating tail region (Plate 3.5B). The location of all these nuclei maps to the approximate positions of hypodermal and bodywall muscle precursors (Sulston *et al.*, 1983). Expression then ceases until late in embryogenesis (3-fold stage) when the large nucleus of the head mesodermal cell near the pharyngeal terminal bulb begins to stain (Sulston *et al.*, 1983). This component lasts through to adulthood (Plate 3.5C).

This gene has no significant homology to other known genes.

F02A9.5

Cytoplasmic expression is evident in both the metacorpus and terminal bulb of the adult pharynx (Plate 3.6). The metacorpal component is restricted to the m4 muscle cells (Albertson and Thomson, 1976). These three cells fill this section of the pharynx and give it its circular character. The expression in the terminal bulb is restricted to the posterior two thirds and marks the location of the m7 muscles.

F02A9.5 is homologous to the family of propionyl-CoA carboxylases, enzymes which act in the catabolic pathway of odd-chain fatty-acids (Kraus *et al.*, 1986). A gene encoding such a vital function for all cells might be assumed to be ubiquitously expressed. The homology score is relatively low for such a highly conserved gene, however, (Nagy *et al.*, 1992) so it is possible the gene codes for a function related but not identical to that of the characterised propionyl-CoA carboxylase gene family. Indeed, a gene with a much higher level of homology (BlastX score 1281), F52E4.1, has been sequenced by the genome sequencing project (data available on ACeDB).

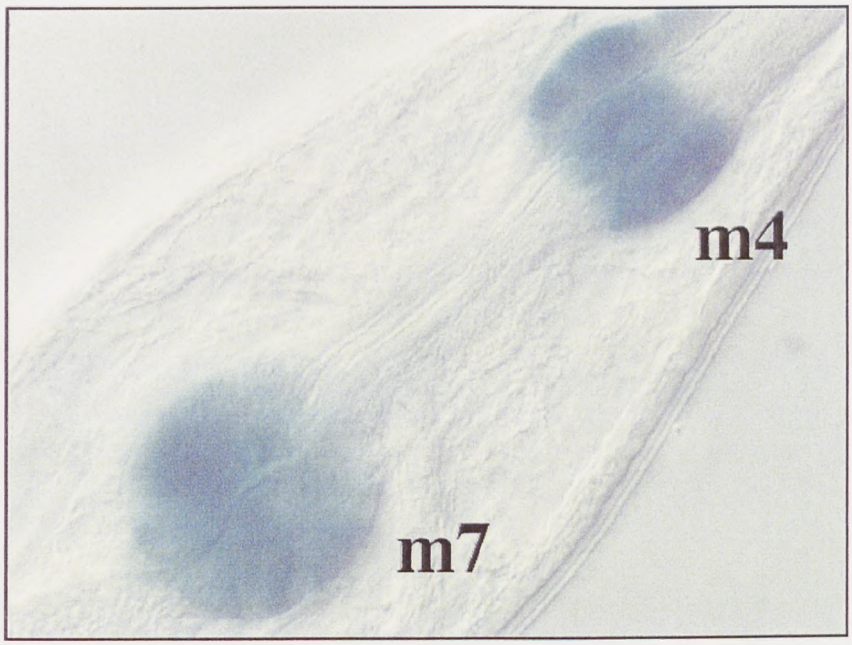


Plate 3.6. *LacZ* fusion expression in UL43 (F02A9.5).

Cytoplasmic expression is seen filling the cell bodies of the m4 and m7 muscle cells in the pharynx of an adult hermaphrodite.

F54G8.2

Adult worms show expression in approximately 10 cells of the head ganglia, in the region of the terminal bulb of the pharynx (Plate 3.7A). Staining seems to outline cell bodies, and can sometimes be seen in neural processes projected from these both into the circumpharyngeal nerve ring and to the extreme anterior of animal (Plate 3.7B). Thus, expression is cytoplasmic. In common with many neural patterns described here, younger worms have fewer positive cells and gradually accumulate more expressing cells, presumably as and when they are born or differentiate (White *et al.*, 1986).

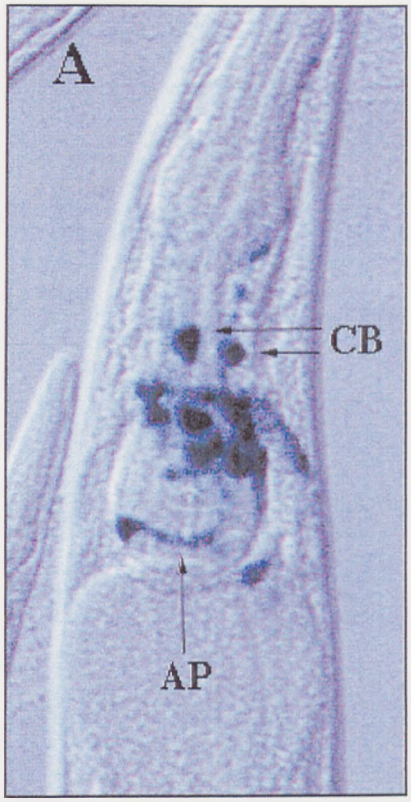
F54G8.2 has good homology to the family of diacylglycerol kinases (e.g. Masai *et al.*, 1992), which convert the second messenger diacylglycerol into phosphotidate, thus attenuating the activity of protein kinase C. They thus act in the control of signal transduction from the cell surface to the cellular interior (Berridge, 1987). The closest homologue to F54G8.2 is in *Drosophila*, where expression has been shown to be restricted to neural cells from late embryos and to continue in these cells through adulthood (Harden *et al.*, 1993).

F54G8.2 may thus encode a signal transduction function in a subset of nerve cells in *C.elegans*.

F59B2.13

Five nuclei of cells of the head ganglia show expression from L1 onwards (Plate 3.8A). A single nucleus in the pharyngeal terminal bulb also expresses from this time (Plate 3.8B). Its position and singularity suggest it may be one of the pharyngeal interneurones I4, I5 or I6 (Albertson and Thomson, 1976). Mid-larval stages mark the emergence of two more spatially distinct pattern components, both of which remain in place through the rest of the life cycle. The cell-bodies of the CANL/R cells can be identified in the lateral mid-body (Plate 3.8C), with their processes running laterally to the anterior and posterior (Plate 3.8D). These two cells run most of the length of the nematode and are of unknown function, though they are required for survival of the worm (White *et al.*, 1986). In the tail, two more nuclei express. One is the anal depressor muscle, the other is possibly one of the intestinal muscles (White *et al.*, 1986).

F59B2.13 has weak homology to the family of G-protein coupled receptors, but lacks two of the transmembrane domains these proteins usually have (data in ACeDB). The neural character of the expression pattern obtained for this gene suggests that it may still encode a neural function involved with GPCRs (see pattern descriptions for C38C10.1 and ZK643.3).



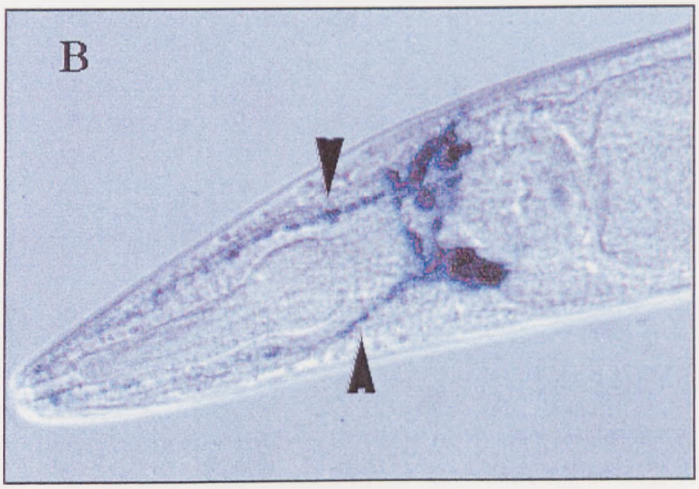


Plate 3.7. *LacZ* fusion expression in UL24 (F54G8.2).

A: Approximately 10 neuronal cell bodies (CB) in the head of an adult hermaphrodite show expression, along with expression in several axon processes (AP).
B: Expression can also be found in processes leading to the extreme anterior (arrowheads).

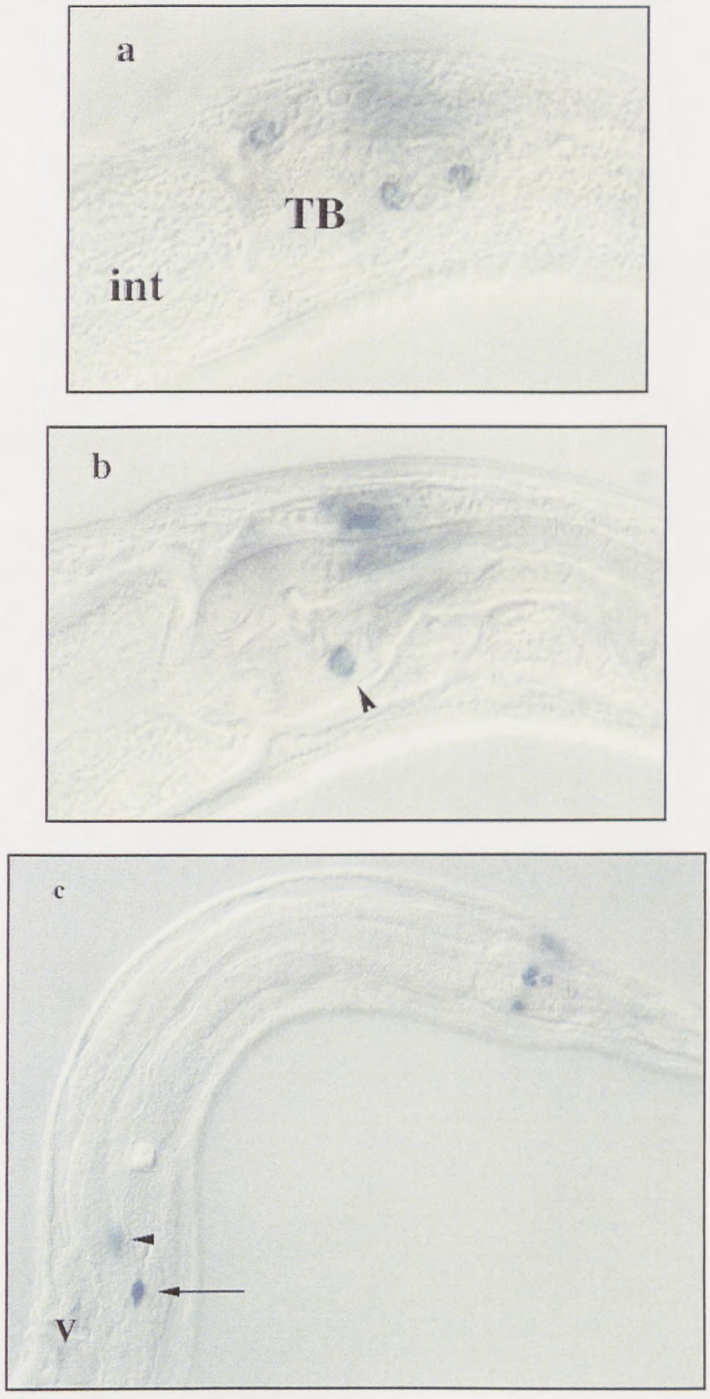
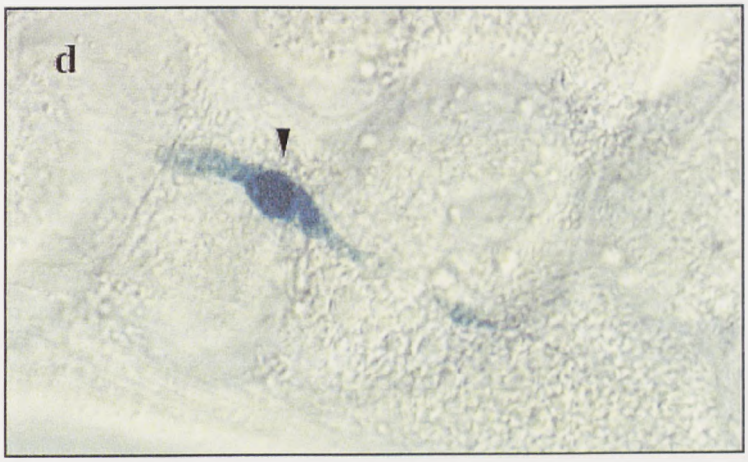


Plate 3.8. *LacZ* fusion expression in UL20 (F59B2.13).

A: Neural expression in the head of a young larva. (TB - terminal bulb of pharynx; int - intestine).
B: A different focal plane in the same animal showing expression in a single cell of the terminal bulb.
C: An L4 larva showing continued expression in head neurones, and expression in CAN neurones in the midbody. (arrow and arrowhead). V marks the position of the vulva.



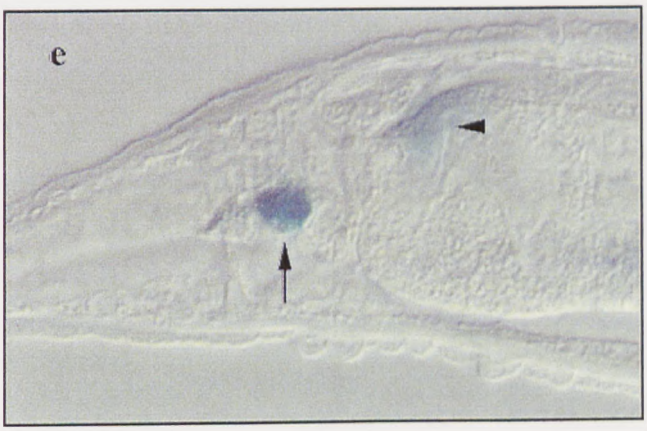


Plate 3.8 contd.

D: Staining in the CAN cells not only fills the cell body (arrowhead), but can be seen in the processes running anteriorly and posteriorly from it.
E: Adult worms show expression in the large nucleus of the anal depressor muscle (arrow) and in the intestinal muscles (aroowhead) at the margin of the posterior intestine.

R08D7.3

This pattern comprises four neural nuclei in the head ganglia (Plate 3.9). The nuclei begin to stain during late embryogenesis, and continue through all subsequent stages.

This gene has no significant homology to other known genes.

R08D7.5

This gene is expressed only in neural cells. At hatching, two cells in the head ganglia and three in the tail ganglia stain for reporter activity (Plate 3.10A). As the worm grows, more neurones become positive for expression-presumably indicating the birth of these cells (White *et al.*, 1986). Adulthood, in which the full pattern can best be seen, has ~15 cells staining in the head and tail ganglia (Plate 3.10B), with a further 6-8 in the most anterior region of the ventral nerve cord (Plate 3.10C).

The most homologous gene to R08D7.5 is calmodulin (Gawienowski *et al.*, 1993). This protein mediates the control of a large number of enzymes such as protein kinases and phosphotases, involved in transducing signals from the cellular membrane to the cell interior (O'Neil and DeGrado, 1990). Calmodulin undergoes a conformational change upon binding calcium ions, facilitating interactions with downstream enzymes (Gawienowski *et al.*, 1993). Thus, R08D7.5 is likely to function in control of calcium ion dependant signal transduction in a subset of nerve cells in *C.elegans*.

R107.1

All expression of this gene occurs during the earlier stages of development. An initial symmetrical set of two pairs of four nuclei at the 300 cell stage are the first cells to express R107.1 (Plate 3.11A). Of these, only Z1 and Z4 continue to show expression for any significant length of time, the other six presumably being their closest relations (Sulston *et al.*, 1983). As elongation and morphogenesis begin in the embryo, Z1 and Z4 can be seen to migrate posteriorly to join Z3 and Z2 in the midbody of the animal (Plates 3.11B/C), there to form the gonad primordium (Plate 3.11D/E). (One large nucleus, which I have identified as that of the head mesodermal cell (Sulston *et al.*, 1983), does continue to express the transgene at the start of these migrations (Plate 3.11B) but ceases by the comma stage (Plate 3.11C).) They remain as such until mid L1 when staining disappears, a time concomitant with their first divisions to form the somatic gonad (Kimble and Hirsh, 1979).

Late embryogenesis and the L1 stage see 2 nuclei in both the head and tail show expression (Plate 3.11E). This lasts only until the end of L1/early L2, suggesting these cells to be neuroblasts (Sulston and Horvitz, 1977). Occasional animals have 3 cells in

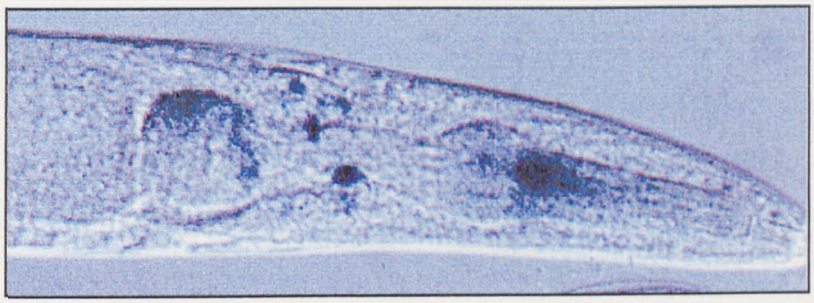


Plate 3.9. *LacZ* fusion expression in UL23 (R08D7.3)

Staining is evident in 4 neural nuclei in the head of an L3 larva. Non-specific staining can also be detected in the pharynx.

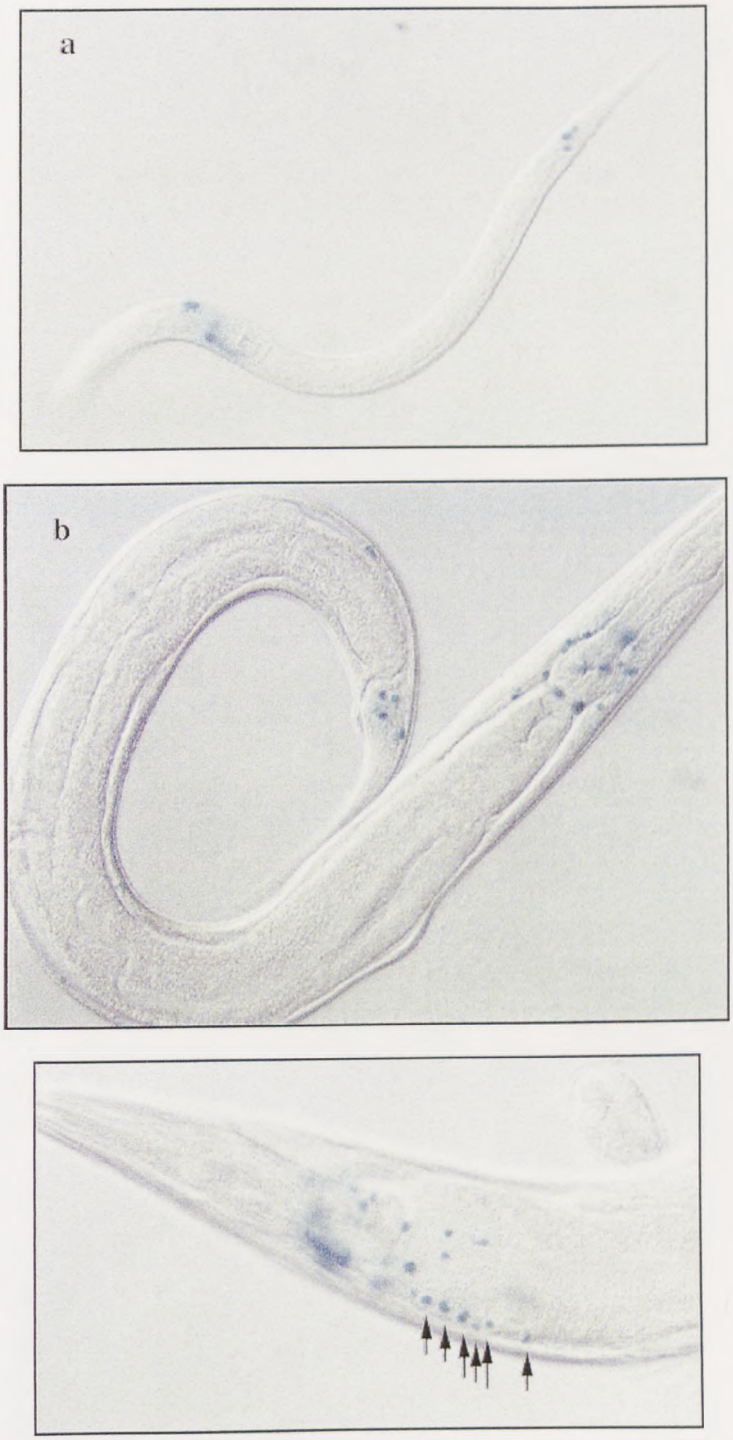


Plate 3.10. *LacZ* fusion expression in UL27 (R08D7.5).

A: Staining in a few neural cells of the head and tail in an L1 larva.
B: An adult animal has many more expressing neurones, including approximately 10 in the head and 5 in the tail.
C: Adults also exhibit staining in several cells in the anterior of the ventral nerve cord (arrows)

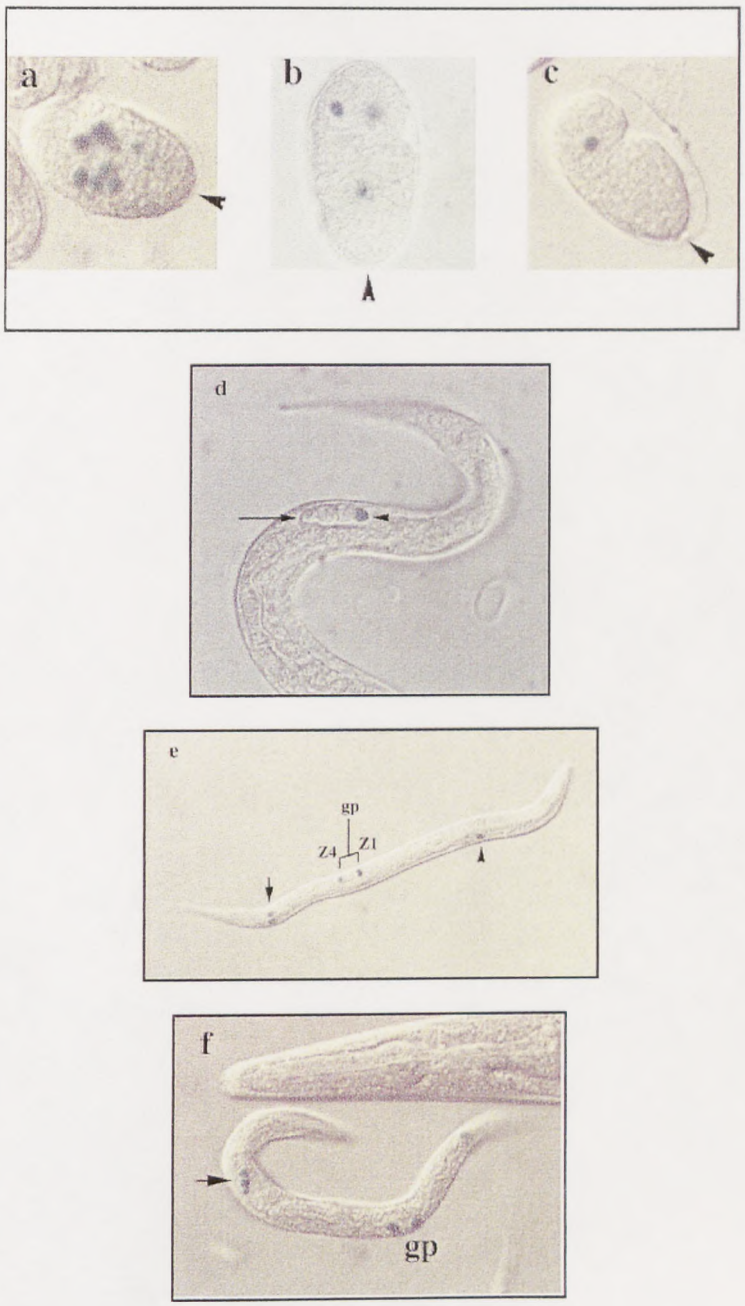


Plate 3.11. *LacZ* fusion expression in UL35 (R107.1).

A, B and C: Three modes of expression during embryogenesis. Anterior is marked by the arrowheads. A: 8 symmetrically placed nuclei at the 300 cell stage. B: Z1 and Z4 showing expression in an embryo just starting elongation, along with one nucleus in the central anterior (head mesodermal cell?). C: A comma stage embryo showing expression only in one of Z1/4.
D: Z1 (arrow) and Z4 (arrowhead) can be clearly seen in the gonad primordium of an L1 larva.
E: Another L1 larva showing exression in neuroblasts of the head (arrowhead) and tail (arrow), as well as in the gonad primordium (gp).
F: An L2 larva reveals expression in 3 nuclei in the head (arrow).

these locations, as opposed to two, so expression may last a short time after the neuroblasts' first division (Plate 3.11F).

This gene has no significant homology to other known genes.

R107.4

All hatched stages exhibit expression in three cells in the anal region of the worm (Plate 3.12). One is positioned directly over the anal duct in the intestino-rectal region, and has thus been identified as the anal sphincter cell (Sulston and Horvitz, 1977). The remaining two are symmetrically placed on the lateral ventral surfaces of the posterior intestine, a position consistent with that of the two intestinal muscles. These cells form part of the machinery necessary for the process of defecation in *C.elegans*, and are closely coupled together by gap junctions (White *et al.*, 1986).

R107.4 is most closely homologous to a serine/threonine kinase from yeast (Feldmann *et al.*, 1994). A characterised example of this gene family in *C.elegans* is the *lin-45* gene which is controlled by a Ras signalling pathway regulating differentiation of the vulva (Han *et al.*, 1993).The signal to defecate is thought to be provided by synaptic output of the single neuron, DVB, which synapses with all the components of the defecatory apparatus (White, 1988). R107.4 may thus not be involved in the primary defecation induction, but may instead act to modulate this behaviour.

The defecatory apparatus has one other component, the anal depressor muscle (White *et al.*, 1986), which is not included in the expression pattern observed for R107.4. One explanation could be that the control elements necessary for expression in this cell are not present in the *lacZ* fusion made for R107.4; however, 6.4kb of upstream sequence is included in the fusion (Table 3.2) so this seems unlikely. Another explanation would assume some functional divergence between the anal depressor cell and the cells represented in the expression pattern.

T23G5.5

Nuclear localised expression in 6 cells of the head ganglia can be observed from the L1 stage onwards (Plate 3.13A). A further component manifests itself during L2, when a bilaterally symmetric pair of nuclei in the posterior lateral ganglia begin expression (Plate 3.13B). This overall pattern is identical to that seen for the dopamine expressing cells in *C.elegans* (Sulston *et al.*, 1975). This congruence positively identifies the expressing cells thus: the 6 cells in the head are the cephalic neurons (CEPDL/R and CEPVL/R) and the deirid neurons (ADEL/R), whilst the 2 cells in the tail are the postdeirid neurons (PDEL/R).

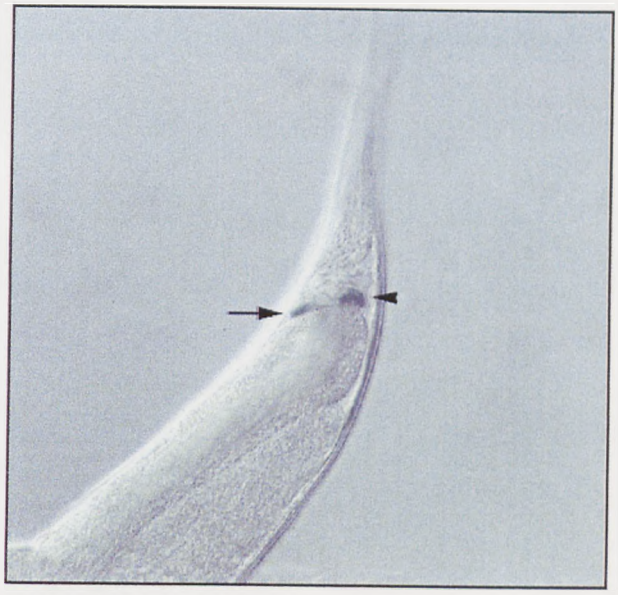


Plate 3.12. *LacZ* fusion expression in UL36 (R107.4).

Expression is evident in the anal sphincter cell (arrowhead) and one of the intestinal muscles (arrow) in the tail of an adult hermaphrodite.

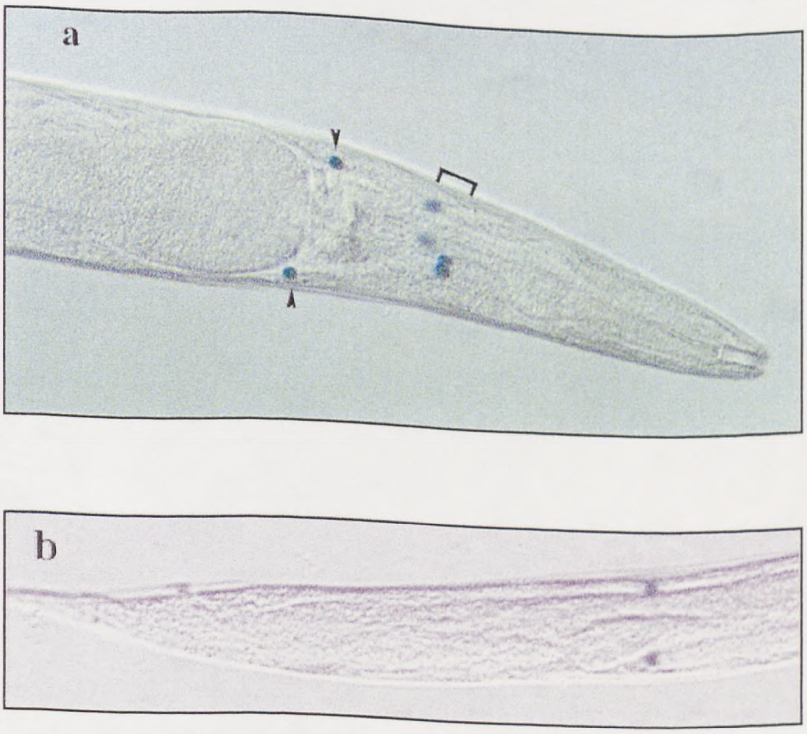


Plate 3.13. *LacZ* fusion expression in UL33 (T23G5.5).

A: Expression in the head of an adult hermaphrodite. The 2 most posterior nuclei (arrowheads) are those of the deirid neurones, ADEL and ADER. The four more anterior nuclei (bracketed) are the cephalic neurones, CEPDL/R and CEPVL/R.
B: Expression in the tail region of an L2 larva comprising the bilaterally symmetrical neurones of the postdeirid sensilla, PDEL and PDER.

The sequence data suggests T23G5.5 encodes a serotonin transporter, but the specificity of transporters is difficult to predict from genomic sequence alone (Reizer *et al.*, 1994). The *lacZ* fusion for T23G5.5 directs expression exclusively in the dopaminergic neurones of the *C.elegans* nervous system, suggesting the gene encodes a dopamine transporter, a testable hypothesis.

ZC21.2

The expression pattern of ZC21.2 has two distinct components. L1 larvae stain in 2 neural cells in the lateral ganglion of the head (Plate 3.14A). The second component is in the vm1 and vm2 vulval muscles of young adults (Plate 3.14B). Expression in all of these cells is cytoplasmic and is noticeably excluded from the nuclei.

ZC21.2 is homologous to the transient-receptor-potential like protein family of plasma membrane cation channels which have been implicated in phototransduction in *Drosophila* (Montell and Rubin, 1989). One form contains sequences for binding calmodulin, so function may include calcium dependent signal transduction as well as simple mediation of calcium ion entry (Phillips *et al.*, 1992) (see pattern description for R08D7.5). *C.elegans* does exhibit a significant response to light (Burr, 1985), but the anatomical or molecular basis for this is unknown (White, 1988). Experiments in-progress with light sensitive mutants have identified the AFDL/R neurones in the lateral ganglion as possible effectors of the light response (Burr, pers. comm.), cells associated with the amphid sensilla (White *et al.*,1986). The two cells seen to express the *lacZ*::ZC21.2 fusion are in the correct area and relative positions to be AFDL and AFDR (White *et al.*, 1986). It is thus possible that ZC21.2 encodes a protein function involved in the photoresponse of *C.elegans*.

ZC21.3

One nucleus in the anterior part of the pharyngeal metacorpus shows expression. The expression is relatively weak and observed at lower frequency than most of the patterns described here, but is reproducible and so is included as a specific gene expression pattern. The position and singularity of expression suggests the staining cell is the pharyngeal interneuron I3 (Albertson and Thomson, 1976) (Plate 3.15).

This gene has no significant homology to other known genes.

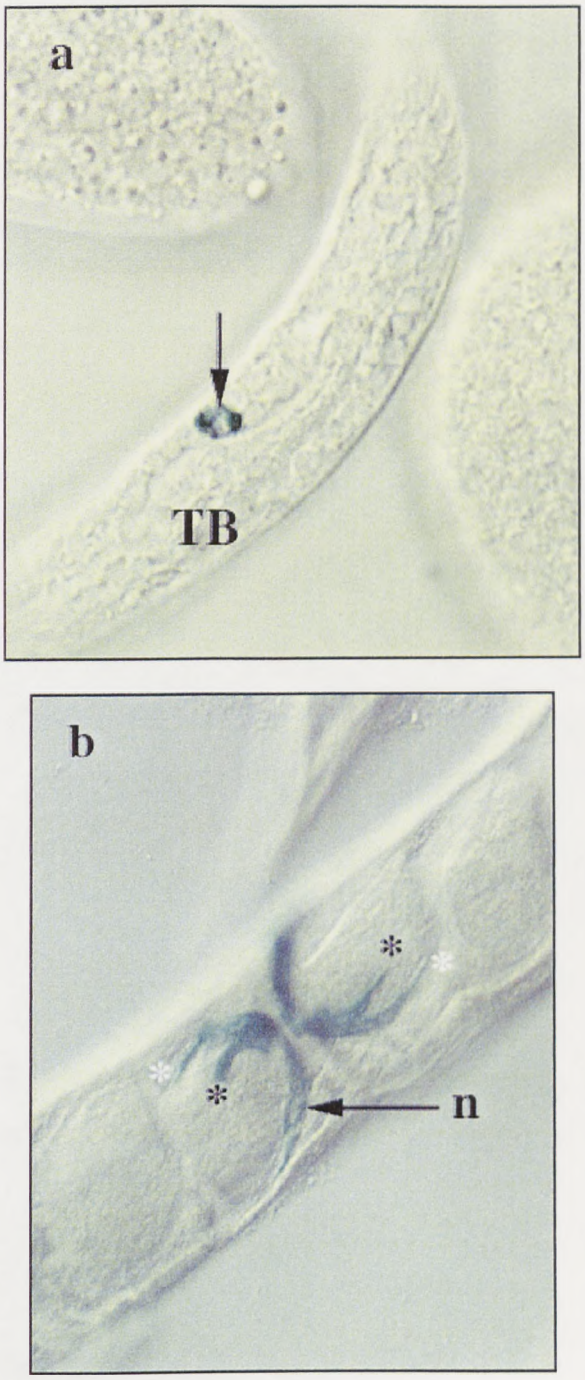


Plate 3.14. *LacZ* fusion expression in UL41 (ZC21.2).

A: Expression in the cell body of a neurone in the head of an L1 larva. Staining is noticeably excluded from the nucleus (arrow). TB - terminal bulb of pharynx.
B: A young adult shows expression in the vulval muscles vm1 (white asterisk) and vm2 (black asterisk). Again, expression cannot be detected in the nuclei (n).

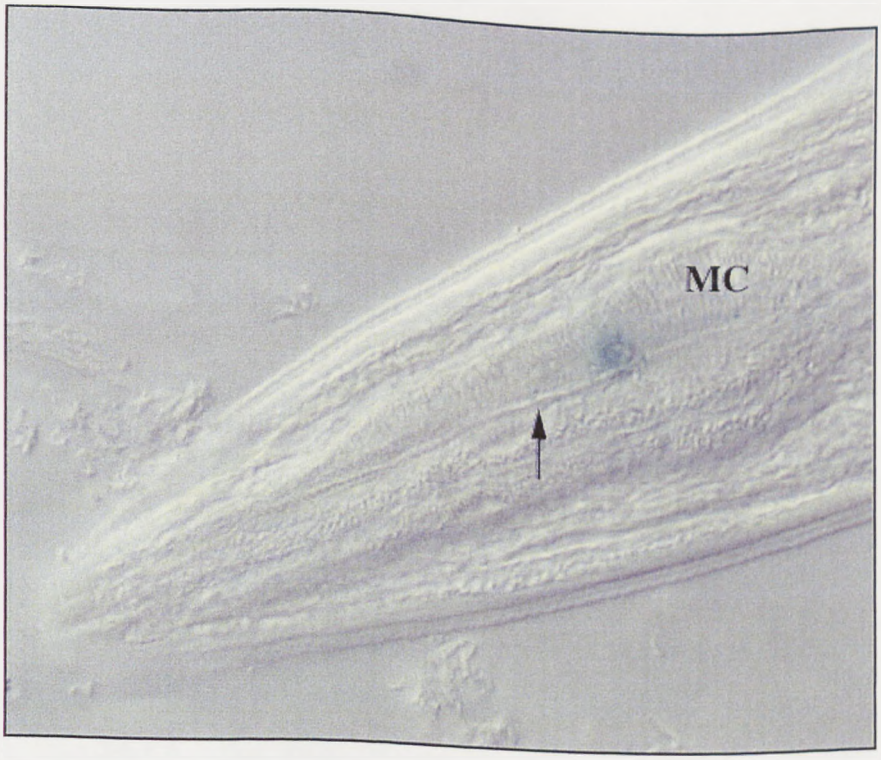


Plate 3.15. *LacZ* fusion expression in UL44 (ZC21.3).

Weak staining can be seen in a single cell body at the anterior margin of the metacorpus (MC) in the pharynx of an adult hermaphrodite. The lumen of the pharynx is clearly visible (arrow) in this image.

ZC21.4

Five nuclei in the terminal bulb of the adult pharynx exhibit expression. The three-dimensional pattern of staining most closely maps to the distribution of the pharyngeal gland cells (two g2 cells and three g1 cells - Albertson and Thomson, 1976) (Plate 3.16A, B and C). The gland cells are thought to secrete molecules aiding the breakdown and digestion of food into the pharyngeal lumen (Albertson and Thomson, 1976).

Sequence homology to the break-point-cluster (bcr) gene of humans (Lifshitz *et al.*, 1988) indicates the probable involvement of ZC21.4 in signal transduction of Ras-related pathways. The bcr gene is a GTPase activating protein (GAP) promoting the exchange of GDP for GTP on p21-Rac, thus activating downstream elements in the signalling pathway (Diekmann *et al.*, 1991). Interestingly, ZC21.4 seems to supply the 3' end of a larger gene, C04D8.1 (see discussion below), which has been shown to encode a GAP molecule with activity in many different Ras-related signalling pathways (Chen *et al.*, 1994). Its promiscuous activity is assumed to result from the lack of several protein domains found in the human bcr gene. ZC21.4's sequence may thus specify its function in a specific set of pathways, presumably only acting in the gland cells of the *C.elegans* pharynx. In the light of the dedicated secretory nature of the gland cells, the implication of one Ras-related subfamily member, Rab, in secretory processes in mammals may prove to be of significance (Chen *et al.*, 1994).

As might be expected from the functional and molecular diversity of the Ras-related proteins (Bourne *et al.*, 1990), there are 9 sequenced genes in *C.elegans* genome sequence similar to ZC21.4 (data in ACeDB). This diversity, both of family members and individual genes, suggests that many more tissue specific expression patterns could be obtained for this gene family.

ZC84.3

This neural pattern covers all non-embryonic stages and accumulates cellular components during development. The expression is localised to the cytoplasm, the resulting smears of staining in cell bodies and processes making it difficult to make definitive cell counts in the tightly populated regions of neuropil. However, by adulthood an estimated 10 neurones in the head, 6 in the tail, and 8/10 in the ventral nerve cord comprise the complete pattern of expression (White *et al.*, 1986) (Plate 3.17A). Stained process bundles serve to sharply define the location of the ventral nerve cord along the animal (Plate 3.17A and B), and indicate the position of the nerve-ring in the head (White *et al.*, 1986) (Plate 3.17B).

This gene has no significant homology to other known genes.

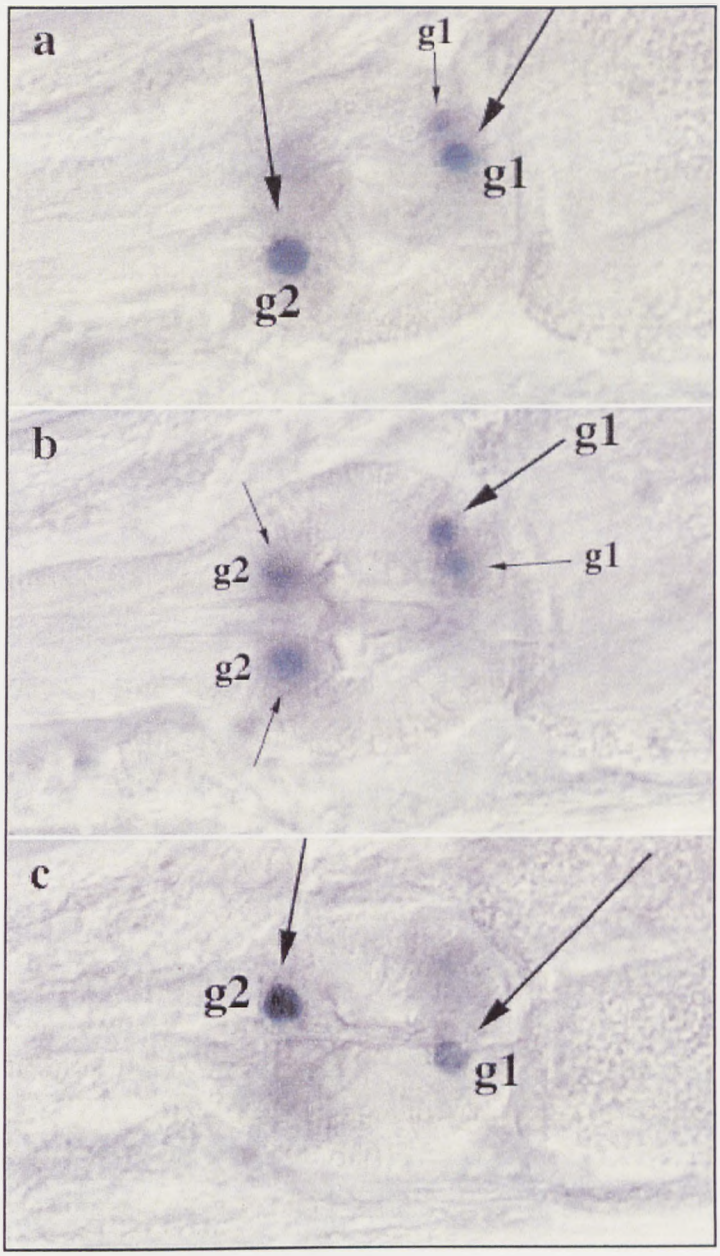


Plate 3.16. *LacZ* fusion expression in UL60 (ZC21.4).

A, B and C: Three focal planes through the terminal bulb of an adult hermaphrodite. Five nuclei can be seen in total, mapping to the positions of the pharyngeal gland cells, g1 and g2.

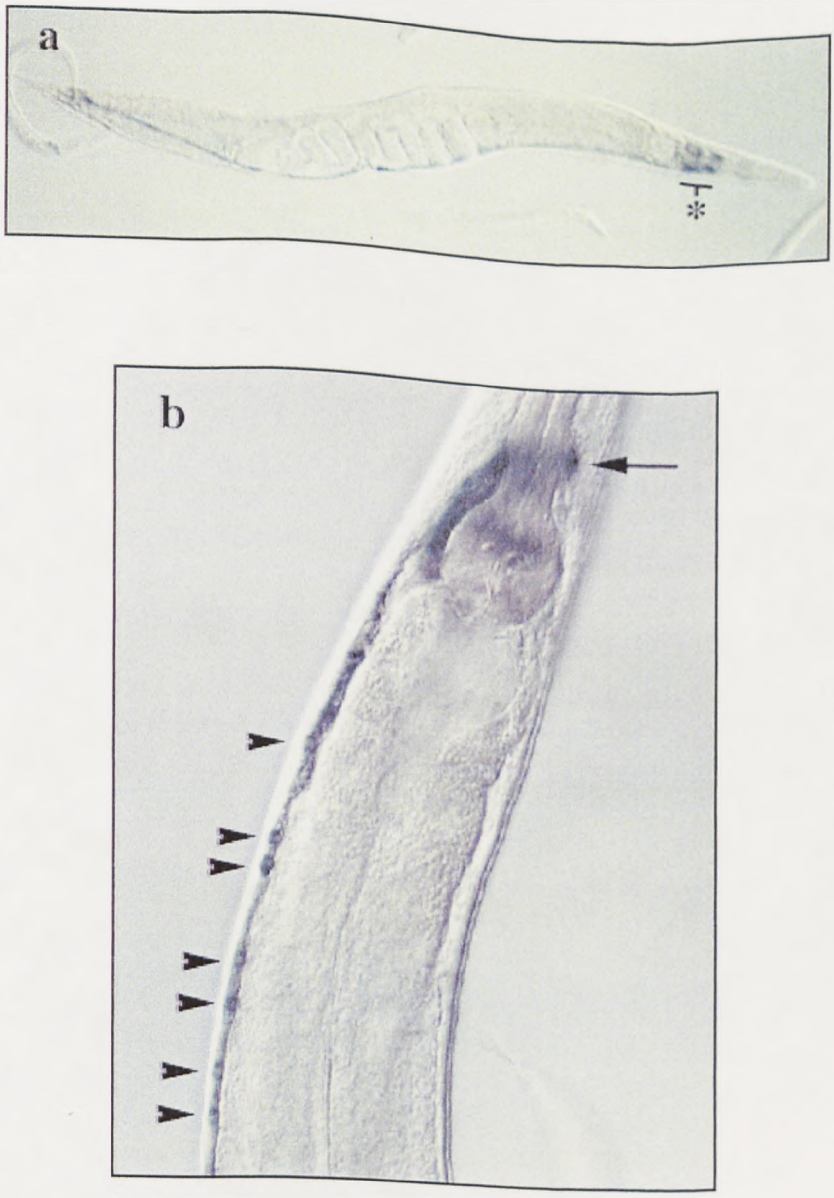
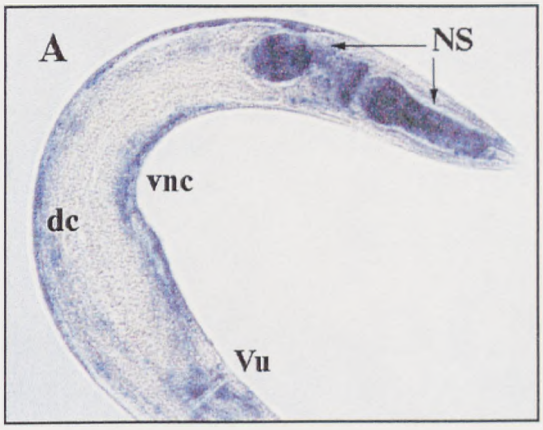


Plate 3.17. *LacZ* fusion expression in UL30 (ZC84.3).

A: The entire expression pattern can be seen in this adult hermaphrodite. Staining is neural in character with neurones in the head (asterisk) and the ventral nerve cord running down the length of the animal clearly visible.
B: A closer view reveals staining in the cell bodies of motoneurones in the ventral nerve cord (arrowheads) and in their projections into the nerve ring in the head (arrow).



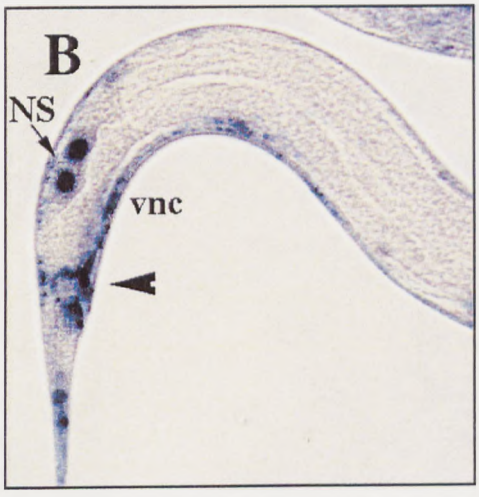


Plate 3.18. *LacZ* fusion expression in UL21 (ZK637.5).

A: This view of the anterior portion of an adult hermaphrodite shows punctate staining down both the ventral (vnc) and dorsal (dc) nerve cords and around the vulva (Vu).
B: The posterior also contains punctate staining in the dorsal and ventral nerve cords and in the neuropil of the tail ganglia (arrowhead).
NS - non-specific staining in the pharynx and posterior intestine..

ZC84.3a

This gene, a possible alternatively spliced product of the ZC84.3 locus, gives an expression pattern indistinguishable from that described for ZC84.3. Thus, although an alternative gene product is expressed, neither spatial nor temporal differences in expression are observed. The phenomenon of alternative splicing for this gene is discussed below.

ZK637.5

Staining in the pharynx, and posterior intestinal cells of adults, indicates that an incomplete promoter may be present in this construct (Mello and Fire, 1995). More specific non-nuclear localised components are present, however, in a subset of neural cells. Punctate staining down processes in the ventral and dorsal nerve cords (Plates 3.18A and B) is seen, and rather disorganised staining was found around cells forming the vulva.

ZK637.8

This gene gives rise to a complicated multicomponent developmental expression pattern. Earliest expression is seen during the gastrulation stage of embryogenesis in the clonal descendants of the E blastomere, the founder cell giving rise to the entire gut of the adult animal (Sulston *et al.*, 1993). Expression begins in Ea and Ep just as gastrulation is starting (Plate3.19A and B), and continues into each of the granddaughters of these two cells. At this stage, the expressing cells clearly outline the emerging form of the gut (Plate 3.19C). This component ends at about the 150/200 cell stage.

The next stage at which expression is evident is during the elongation phase of late embryogenesis when the worm is approximately 2 fold. The nuclei of the M2 motor neurones in the terminal bulb of the pharynx stain strongly (Albertson and Thomson, 1976). More pharyngeal cells show expression as morphogenesis proceeds until at hatching the two I1 interneurones of the metacorpus, either the e2 or m2 cells of the procorpus, and the m8 muscle cell at the pharyngeal-intestinal boundary can all be seen. This pharyngeal pattern persists through the rest of the life cycle, although the m8 expression is lost during early larval stages.

Early larval stages also see the appearance of two more components of expression which last through the rest of the life cycle. The anal sphincter valve stains in the posterior of the worm (Plate 3.19D) (White, 1988). In addition, neural expression is observed in the head ganglia and in a subset of motoneurones in the ventral nerve cord (Plate 3.19E). Older larvae obviously contain more positive motoneurones (Plate 3.19F), a consequence of the

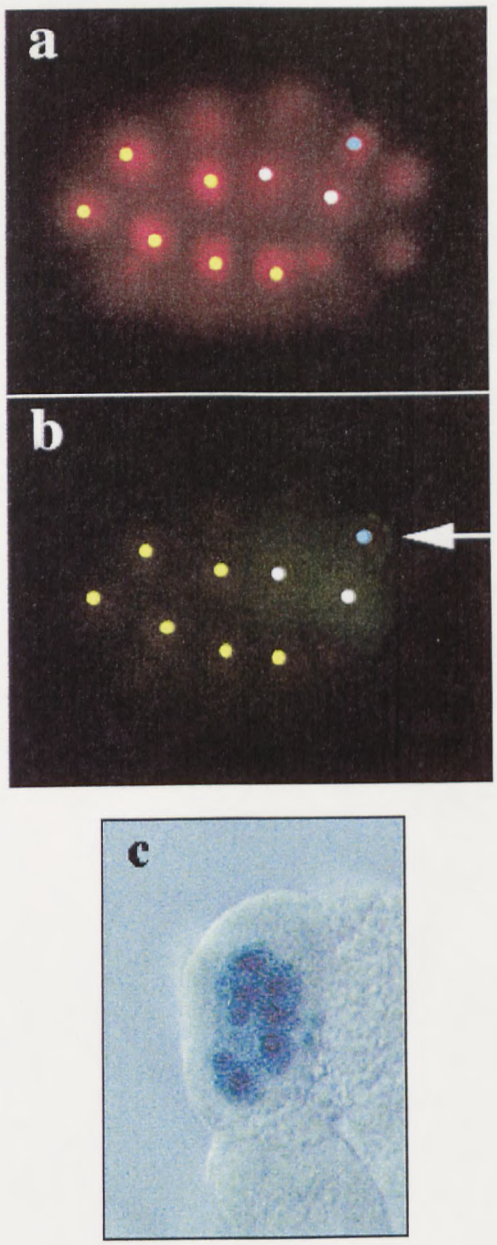


Plate 3.19. *LacZ* fusion expression in UL22 (ZK637.8).

A and B: Images of immunohistochemically stained 28-cell embryo (see Chapter 2) double stained with anti-β-Galacosidase antibody and anti- P-granule antibody. White dots mark the positions of the nuclei of Ea and Ep, blue dots mark P4 and white dots mark MS- and AB-blastomere derived cells used as markers of embryo orientation. Anterior is to the left and the focal lane lies on the ventral side of the embryo. A: The locations of nuclei stained with the DNA specific dye, propidium iodide (Chapter 2). B: The same focal plane showing fluorescence produced by FITC-labelled antibody localisation to Ea and Ep (anti-β-Galacosidase) and P4 (anti-P-granule). P-granule signal is clearly located perinuclearly (arrow).
C: A 150 cell embryo showing expression in the 8 members of the E lineage.

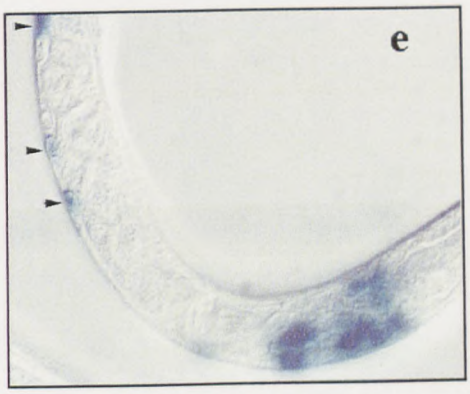


Plate 3.19 contd...

D: Expression in the anal sphincter valve in the tail of an adult hermaphrodite.
E: Neural staining in the ventral nerve cord (arrowheads) and head ganglia of the head of an L1 larva.
F: The neural staining in an L3 larva exhibits considerably more staining cells in the nerve cord. The asterisk marks the position of the gonad primordium in the midbody.

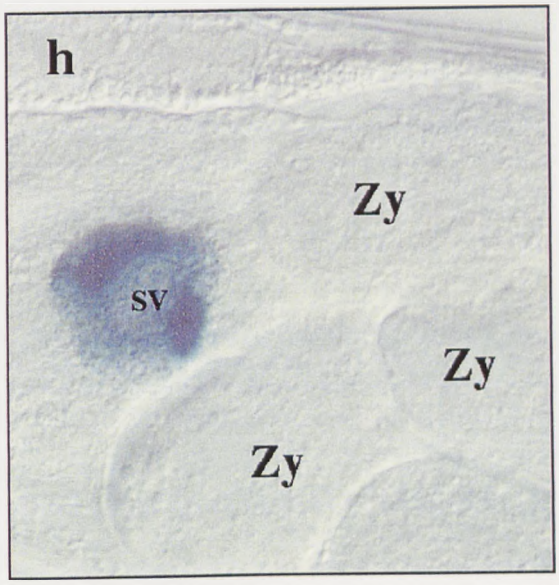


Plate 3.19 contd...

G: Expression in an L4 larva in the D-cells of the developing vulva in the centre of the image, flanked on either side by expression in the developing spermathecal tissues.
H: Spermathecal staining continues into adulthood when the ring of the spermathecal valve (SV) is clearly visible. Zygotes (Zy) which have recently passed through the valve to become fertilised lie to the right of the valve.

continued production of these cells during early larval stages (Sulston and Horvitz, 1977).

The last expression component to appear is in certain cells of the somatic gonad. The D-cells of the vulval labia and unidentified cells of the spermathecal structures begin expression in L4 (Plate 3.19G), whilst gonadal morphogenesis is ongoing (Kimble and Hirsh, 1979). The D-cells do not seem express beyond the first oocyte fertilisations as no zygotes are usually visible when these cells are stained; the spermathecal staining lasting slightly longer into adulthood when a ring structure can occasionally be seen (Plate 3.19H), indicative of the spermathecal valve (White, 1988; Hope, 1991).

ZK637.8 is highly homologous to the ATPase subunit of the vacuolar class of membrane proton pumps (V-ATPases) (Perin *et al.*, 1991). In animal cells they are responsible for the acidification and consequent function of a variety of organelles (reviewed in Forgac, 1989). In clathrin-coated vesicles involved in endocytosis, for example, a low pH is required for dissociation of many receptor-ligand complexes (e.g. Haigler *et al.*, 1980; Posner *et al.*, 1978). The receptors can then be recycled back to the cell membrane for re-use (the low density lipoprotein (LDL) receptor, which binds to extracellular cholesterol, is estimated to be recycled in this way upto 200 times (Brown *et al.*, 1983)). Similar dissociation of receptors bound to proteins in transport to the lysosomal compartment depends on a low pH in vesicles targeted to lysosomes, allowing recycling of the receptors to the Golgi body (Creek and Sly, 1984). Maintanence of a low pH is also known to be required for the optimal activity of hydrolytic enzymes involved in the breakdown of material in lysosomes (Poole and Ohkuma, 1981). A low internal pH in secretory vesicles is required for the correct proteolytic processing of secreted peptides (Rudnick, 1986).

Perhaps most significantly for ZK637.8, concerning the neural component of expression observed, V-ATPases are also involved in the loading of synaptic vesicles with neurotransmitters (Johnson, 1988). Low pH in these vesicles provides the driving force for countertransport of neurotransmitter into the vesicle (Johnson *et al.*, 1979).

Other components of the ZK637.8 expression pattern are more difficult to link to particular V-ATPase functions, as the cellular functions of the expressing cells have no specific correlates with the repertoire of functions described above. It is interesting to speculate, however, on the expression seen in the early embryo, spermathecal valves and vulva. Each of these components concerns cells in a developing structure or tissue. The embryonic and spermathecal expression also occurs, at least in part, at times when the expressing cells are moving: expression in the E lineage begins just as Ea and Ep are about to migrate into the interior of the embryo (Plate 3.19A), the first cell movements to occur in gastrulation (Sulston *et al.*, 1983); spermathecal expression begins as the

developing somatic gonad grows distally from the vulva during the L4 larval stage (Kimble and Hirsh, 1979) (Plate 3.19G). Cell adhesion is known to play an important role in tissue morphogenesis, and molecules supplying this capability at the cell surface have been characterised (Lee and Gumbiner, 1995). Given the extensive involvement of V-ATPases in the transport of proteins to and from the cell membrane described above, the expression of ZK637.8 in developing structures may reflect the transport of morphogenetic information in the same way. Perhaps WP:ZK637.8 acts in a secretory pathway to effect the processing of a signalling molecule on its way out of the cell, or in a receptor-mediated pathway to enable transport of an endocytosed molecule to the nucleus.

ZK643.1

Another gene with a complex expression pattern. Staining can be identified in the cell bodies and processes of neural cells in the head and tail ganglia and in the ventral nerve cord (Plate 3.20A). Closer observation reveals that staining is punctate (Plate 3.20B). These neural components begin at early larval stages, and persist throughout subsequent stages.

A component in the vm1 vulval muscle cells is present throughout adulthood (Plate 3.20C). These muscles are important in opening the vulva during egg-laying (White, 1988). Expression here is targeted to sub-cellular bodies whose location seems to be congruent with the area of overlap with the other set of vulval muscles, the vm2 cells (White, 1988).

ZK643.1 has no homology matches but the sub-cellular localisation of the fusion gene product in the vulval muscles may prove instructive in terms of the gene's function. These muscles are single sarcomere units so this arrangement cannot correspond to attachments for dense bodies or M-line analogues (Waterston, 1988). Nevertheless, expression in discrete foci and restriction of these foci to the area of overlap between the vm1 and vm2 muscle sets suggests that the ZK643.1 gene product does have a structural role, possibly in positioning these cells relative to each other. Localisation to the processes and cell bodies of a large variety of neurones may reflect this supposed function, and identify a molecule serving to attach neural neighbours to each other.

ZK643.3

Expression first becomes evident at the L1 larval stage when the head body-wall muscles (Plate 3.21A) and an unidentified cell(s) in the anal region stain for β-Galactosidase activity. The anal staining lasts only until L2/3, whilst the head muscles continue showing expression right through to adulthood (Plate 3.21B).

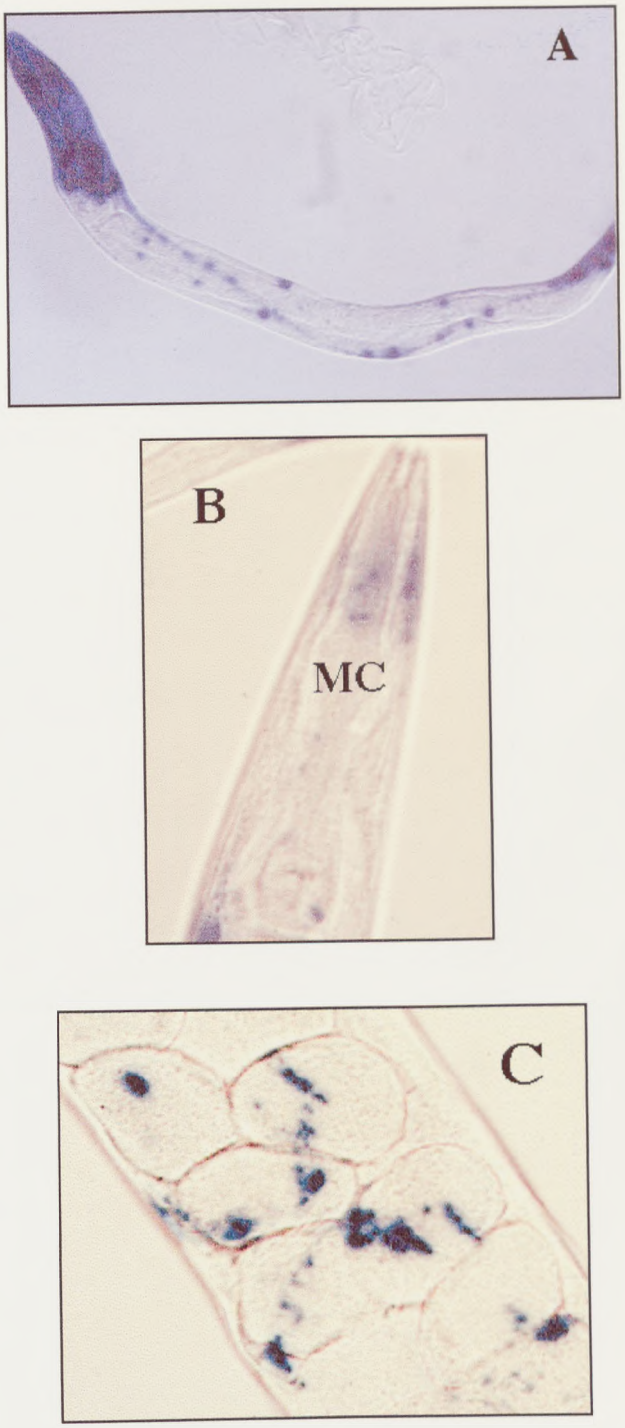


Plate 3.20. *LacZ* fusion expression in UL16 (ZK643.1).

A: The neural components of expression are revealed in this image of an L3 larva; in the head and tail ganglia and down the ventral nerve cord.
B: Staining in the anterior region of the head in an adult hermaphrodite shows a punctate appearance along the processes of sensory neurones.
C: Punctate staining is again seen in the vm1 muscles of an adult hermaphrodite.

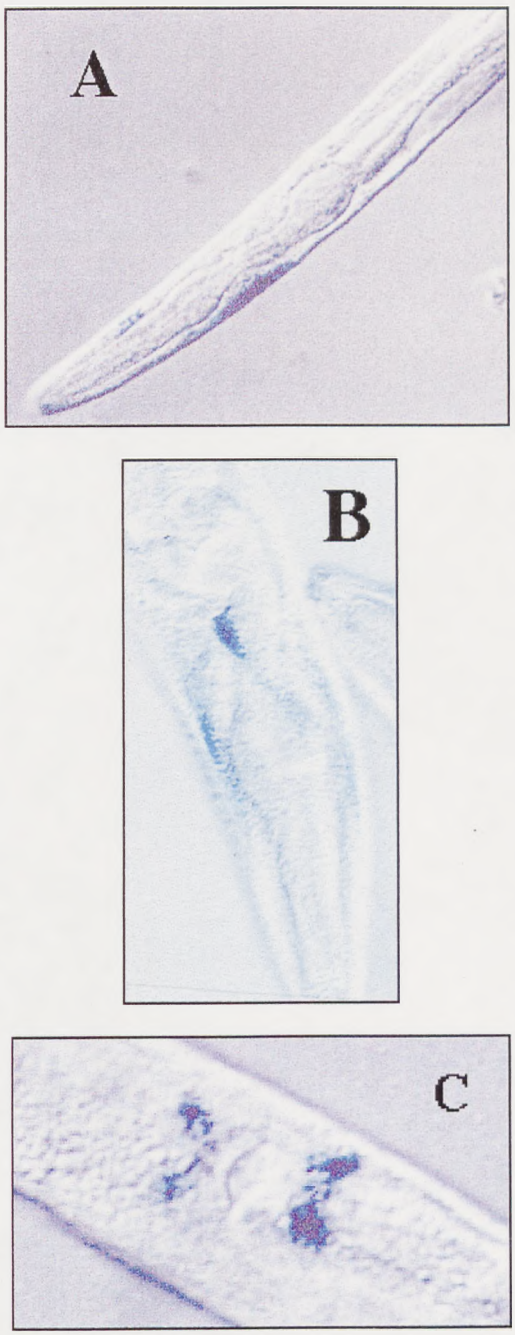
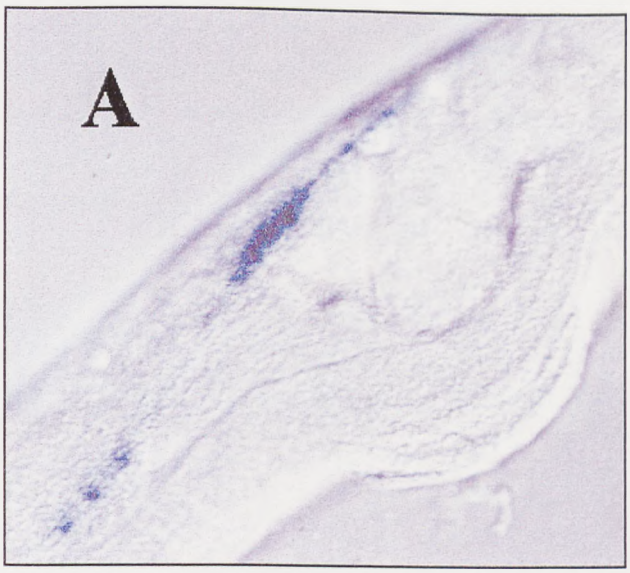


Plate 3.21. *LacZ* fusion expression in UL17 (ZK643.3).

A: An L1 larva showing expression in the head muscles.
B: Muscle staining in the head is still detectable in adult animals.
C: An L4 larva shows the beginning of expression in the vm1 muscles surrounding the vulva.



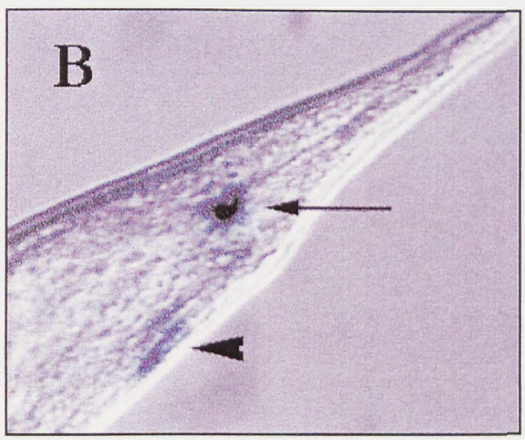


Plate 3.22. *LacZ* fusion expression in UL19 (ZK643.5).

A: Expression is evident in motoneurones of the ventral nerve cord in this adult hermaphrodite.
B: The tail region of another adult hermaphrodite shows expression in the anal depressor muscle (arrow) and the intestinal muscles (arrowhead).

The expression pattern includes a further component in the vm1 vulval muscle cells. These specialised muscle cells attach to the ventral hypodermis and the vulval labia, and are important in opening the vulva during egg-laying. Expression here is transient, seen only for a short time around the L4 to adult molt (Plate 3.21C).

Sequence homology to the calcitonin G-protein coupled receptors (Lin *et al.*, 1991) makes this the third GPCR to be assayed in this screen (see F59B2.13 and C38C10.1). Calcitonin GPCRs are known to regulate calcium homeostasis in mammals, and function via stimulation of adenylyl cyclase to produce the second messenger, cAMP (Lin *et al.*, 1991). They are expressed in many different tissues, including rat skeletal muscle (Albrandt *et al.*, 1993) where they mediate the control of glucose incorporation into glycogen and stimulate sodium pump activity (Andersen and Clausen, 1993). Expression in the muscles of *C.elegans* seen with ZK643.3 may identify an ancestral incidence of these muscle-specific functions.

ZK643.5

Reporter gene expression is seen in neural cells of the ventral nerve cord (Plate 3.22A) and in the anal depressor muscle and intestinal muscles in the tail (Plate 3.22B) Expression for all components stretches from L1 onwards.

This gene has no significant homology to other known genes.

Discussion.

Are *lacZ* reporter expression patterns representative of native gene expression?

The gene density of the *C.elegans* genome is relatively high. Over the entire genome it is estimated that there is one gene per 5kb and in the gene rich centre portions of the autosomes the figure is closer to one gene per 4kb (Wilson *et al.*, 1994). The region covered in this study is in the gene rich cluster of chromosome III and has on average one predicted gene per 4.6kb (Table 3.1). As the average upstream length of DNA sequence in the screened constructs is 4.2kb (Table 3.2), most of a gene's upstream sequence (i.e. that between its own translational start and the upstream neighbour's transcriptional finish) will be contained within each reporter construct. Previous attempts to map control regions in *C.elegans* are few, so general rules about the extent of such regions are difficult to draft. However, the examples so far (e.g. Way *et al.*, 1991; MacMorris *et al.*, 1992; Okkema *et al.*, 1993)) have found that around 3kb of upstream sequence is sufficient to contain all relevant control signals. Thus, the reporter constructs described here are likely to contain enough upstream sequence to include all relevant 5' control elements.

There are circumstances which may lessen the fidelity of reporter expression with respect to a native gene's. Firstly, intragenic (Okkema *et al.*, 1993; Krause *et al.*, 1994) and 3' (e.g. for *fem-3* (Ahringer and Kimble, 1991), *tra-2* (Goodwin *et al.*, 1993), *lin-14* (Wightman *et al.*, 1993) and *glp-1* (Evans *et al.*, 1994)) control elements are known in the worm. Any regulatory sequences in 3' flanking regions would be missing from the 5' translational fusions constructed in this screen. Control elements encoded in intron sequences tend to be located close to the 5' end of the genes under their transcriptional control (Okkema *et al.*, 1993; Krause *et al.*, 1994). The average number of endogenous introns contained in the reporter fusions made here (Table 3.2) is 2.8, however, so any intron-contained transcriptional regulation elements are likely to be included. Secondly, fusion mRNAs and proteins may be subject to different turnover rates than native gene products. Thus, a fusion gene expression pattern is unlikely to be a complete description of native gene expression. An observed expression pattern must reflect, however, at least some aspect of the regulatory elements controlling expression of the predicted native gene (Hope, 1991; Young and Hope, 1993).

Nevertheless, β-galactosidase activity produced from *lacZ* fusion genes has been shown to accurately reflect the expression of many *C.elegans* genes for which the distribution of the native protein product has been directly determined using antibodies (e.g. Krause *et al.*, 1990; Hamelin *et al.*, 1992; Land *et al.*, 1994). In addition, the expression pattern of other *lacZ* fusion genes was completely consistent with mutant phenotype (e.g. Hill and Sternberg, 1992; Cowing and Kenyon, 1992; Mitani *et al.*, 1993). or with expectations

based on expression patterns for homologues in other species (e.g. Stringham *et al.*, 1992; Lincke *et al.*, 1993; Freedman *et al.*, 1993).

Many reporter fusions yield no expression pattern.

22 of the 45 gene specific fusions made gave no observed expression pattern *in vivo*. Such failures may have many causes. Firstly, some predicted genes may not actually be expressed. They may represent pseudogenes, recently duplicated genes which have drifted into non-functionality, a class apparently represented by the gene F58A4.2. It is 200kb away from, and 98% identical to, the gene C38C10.2 on chromosome III, and seems to be a non-transcribed partial copy (Wilson *et al.*, 1994). Genes only expressed under specific conditions (for example during the dauer stage of the life-cycle or in the presence of particular chemicals) would also contribute to this class.

Secondly, certain genes will only be expressed in cells apparently non-permissive for β-Galactosidase gene function, notably the germ-line and all cells of early (i.e less than 12 cell) embryos (e.g. Hope, 1991; Young and Hope, 1993; Seydoux and Fire, 1994).

Polycistronic transcription units have been demonstrated recently in *C.elegans* (Spieth *et al.*, 1993) and will contribute to the third class of non-expressing fusion genes; those that include no 5' upstream sequence. Fusions to downstream members of such polycistronic units may not include the promoter regions upstream of the leading gene. Five genes of the screen are putative downstream genes of predicted polycistronic units as judged by the examples known so far (Spieth *et al.*, 1993; Zorio *et al.*, 1994). The reporter fusions for two of these genes (F59B2.2 and ZK637.10) do not include genomic sequence upstream of the lead gene of the polycistronic unit, and fail to produce an expression pattern (Table 3.3). Of the three fusions which do contain sequence upstream of the lead gene, two (R08D7.5 and ZK637.5) direct reporter expression patterns. Inclusion of sequence 5' to the lead gene of a polycistronic unit thus seems to be necessary for production of an active reporter fusion. It has been estimated that upto 16% of genes in the *C.elegans* genome may be downstream members of polycistronic units (Zorio *et al.*, 1994). Such a representation suggests that the criteria used for gene selection in any continuation screen should include steps to identify polycistronic genes, and avoid those whose upstream regulatory regions cannot be included in a reporter fusion construct.

Alternatively, the prediction for the position of a gene's 5' end may be wrong such that the true transcriptional start lies upstream of the 5' end point of the fusion fragment. For instance, at the time of fragment selection for ZK643.3, the original prediction had the first exon starting at base position 25073 of cosmid ZK643. Later, more thorough computer analysis of this particular gene (Kolakowski, pers. comm.) predicted a more likely start at 20158, which has since been confirmed by cDNA sequencing (Briggs and Coates, pers. comm.). Fortunately, the original 5' restriction enzyme site selected for

gene	5'	pattern
F59B2.2	no	no
F59B2.12	yes	no
R08D7.5	yes	yes
ZK637.5	yes	yes
ZK637.10	no	no

Table 3.3. Genes predicted to be downstream members of polycistronic units.
Columns:-
gene: - predicted gene with characteristics of downstream gene (Zorio *et al.*, 1994)
5': does genomic fragment in lacZ fusion contain at least 1kb of sequence upstream of the lead gene of the polycistronic unit?
pattern: does the *lacZ* fusion drive specific expression (Table 3.2)?

ZK643.3 (Table 3.2) was upstream of this position, the transcriptional start was included, and reporter gene expression was obtained.

Fifthly, β-Galactosidase is known to become inactivated upon passage through a membrane (see above), so would not be expected to show expression for fusion proteins secreted from the cell or having β-Galactosidase fused to an extracellular domain of an integral membrane protein (Fire *et al.*, 1990). Secreted proteins have a signal sequence at their N-terminus which initiates transport of the nascent protein across a biological membrane and is subsequently cleaved off (Strauss and Boime, 1982). Such sequences have a recognisable pattern of amino acids (von Heijne, 1983) which can be searched for in protein sequence and used to identify probable signal sequences (von Heijne, 1986). Integral membrane proteins may have a signal sequence, in which case their N-terminal end will be outside the cell, but also contain hydrophobic transmembrane domains which serve to anchor sections of the protein in the membrane (Strauss and Boime, 1982). Such regions also have specific amino acid character (von Heijne, 1992) which can be searched for to identify probable membrane spanning sequences (Claros and von Heijne, 1994). Identification of the intra/extracellular domains between the predicted transmembrane domains can also be performed by correlation of the number of membrane spanning domains with the predicted location (intra- or extracellular) of the N-terminus (Claros and von Heijne, 1994). Table 3.4 shows the predictions for the genes assayed in the expression screen.

Reporter fusions for each of the three predicted secretory proteins fail to drive expression. Although the dataset for these genes is small, such a finding is consistent with the known inactivation of β-Galactosidase when secreted. As mentioned in the introduction to this chapter, many genes in *C.elegans* genomic sequence may not have their first exon correctly predicted. As signal sequences are located in the 5' end of a coding sequence secreted proteins may be more common than indicated above, increasing the deleterious effect of this phenomenon on the screen.

Of the 20 genes predicted to encode integral membrane proteins, 55% produced active reporter fusions, a figure very similar to the success rate of the screen as a whole. Membrane character does not seem. therefore, to affect the chance of generating an active *lacZ* fusion. A difference in success rate can be discerned, however, between reporter fusions predicted to place β-Galactosidase on the cytoplasmic surface of the membrane (66%) and extracellularly (45%).

These findings suggest several modifications to the selection procedure used in this screen. Prediction of signal sequence and membrane spanning regions should be performed to identify potential secretory and integral membrane proteins. Reporter fusions for genes expected to be secreted should include a synthetic transmembrane

GENE	SIGNAL SEQUENCE	TM	C/E/S	PATTERN
B0303.1	n	n	C	Y
B0303.4	-	**y**	E	N
B0303.11	n	n	C	N
B0303.12	n	**y**	E	Y
B0303.15	n	n	C	N
B0464.1	n	n	C	N
B0464.2	n	n	C	N
B0464.3	n	y	C	N
B0464.4	y	y	C	Y
C38C10.1	y	**y**	E	Y
C38C10.4	y	n	S	N
C40H1.1	n	n	C	N
C40H1.6	n	n	C	Y
F02A9.1	y	**y**	E	N
F02A9.5	n	y	C	Y
F54G8.1	y	n	S	N
F54G8.2	n	n	C	Y
F54G8.3	y	**y**	E	N
F54G8.5	y	**y**	E	N
F59B2.2	n	y	C	N
F59B2.5	y	n	S	N
F59B2.9	n	n	C	N
F59B2.12	y	**y**	E	N
F59B2.13	n	y	C	Y
R08D7.3	n	n	C	Y
R08D7.5	n	n	C	Y
R107.1	y	**y**	E	Y
R107.4	n	**y**	E	Y
T23G5.2	n	**y**	E	N
T23G5.5	n	y	C	Y
T23G5.6	n	n	C	N

Table 3.4 contd....

ZC21.1	n	n	C	N
ZC21.2	n	y	C	Y
ZC21.3	n	n	C	Y
ZC21.4	n	n	C	Y
ZC84.3	n	n	C	Y
ZC84.3a	n	n	C	Y
ZK637.3	y	y	C	N
ZK637.5	n	n	C	Y
ZK637.8	n	y	C	Y
ZK637.10	n	n	C	N
ZK643.1	n	n	C	Y
ZK643.2	n	n	C	Y
ZK643.3	n	y	E	Y
ZK643.5	n	n	C	Y

Table 3.4. Predicted cellular location of translational fusion points.

Columns:-
GENE: gene supplying 5' end of translational fusion to *lacZ*
SIGNAL SEQUENCE: presence of signal sequence suggested by Analysignalase computer program.
TM: predicted transmembrane domains predicted by ToppredII computer program. y denotes that sequence 3' of fusion point will lie extracellularly.
C/E/S: C - cytoplasmic localisation of β-Galactosidase; E - localisation on the extracellular side of the cell membrane; S - secretion of *lacZ* fusion
PATTERN: was the lacZ fusion active?

sequence to hold β-Galactosidase on the cytoplasmic side of the cell membrane (Fire *et al.*, 1990) (Figure 3A). In addition, fusions to genes encoding membrane proteins should still be constructed and assayed, but fusion points placing β-Galactosidase in an intracellular segment should be favoured.

The sixth and final class constitutes those reporter fusions whose expression level is below the threshold sensitivity of the β-Galactosidase assay. Genes with transient expression or very low endogenous expression levels will contribute to this group. Examination of cDNA data, however, suggest that the assay system employed can accomodate variations in gene expression levels. Genes with many ESTs will probably be more highly expressed than those for which only one or no ESTs have been generated (Waterston *et al.*, 1992). ACeDB holds the EST information for each gene. Of the 45 genes in the screen, 36 (Table 3.2) have either 1 or no separate ESTs, and of these 21 produce a reporter expression pattern. Of the five genes with at least 4 ESTs (Table 3.2), indicating relatively high gene expression, only 2 gave expression. Thus high gene expression level, as inferred from cDNA clone frequency, is not indicative of an ability to generate a β-Galactosidase expression pattern in this assay.

The above discussion illustrates that no single explanation can account for the majority of failures to generate an active reporter construct. Several changes to the strategy for gene selection have been proposed, however, and should help to raise the efficiency of a continuation screen. Experiments to investigate the potential for further increases in efficiency are described and discussed in Chapter 4.

The screen as an independent quality control of genome sequence data.

The great dependence of this screen on gene predictions, and the raw genome sequence on which these are based, allows valuable feedback on the quality of genome information. For instance, many restriction digestions of sequenced cosmid DNA are routinely performed as part of the screen. The digestion profiles of the 14 cosmids with a variety of different restriction enzymes reveal that of 596 predicted 6 base pair recognition sequences, only one is absent (the SpeI site at base position 29391 on ZC84). No extra, non-predicted sites were found. The error rate of $3x10^{-4}$ which this signifies compares well to the sequencing project target of 10^{-4} (Waterston and Sulston, 1995).

The exon/gene predictions of the sequencing project have also been addressed. Any error designating an actual intron sequence as an exon sequence would be hard to recognise purely using expression data as there are potentially many reasons why any single reporter fusion will not express (see above). Valuable feedback in cases of uncertainty over particular coding sequences has been possible, however. The gene currently designated ZC84.3 was originally predicted to code for two differentially spliced transcripts. Fusions were designed and made for both, and each produced an identical

expression pattern (Table 3.2). Subsequent sequence annotations in ACeDB, however, do not include one of the alternative transcripts, such that the 3' fusion site for ZC84.3a is now in a predicted intron. Observation of reporter expression with both fusions suggests that the original splicing predictions of alternative splicing were correct.

The gene ZC21.4 (Wilson *et al.*, 1994) was originally annotated in ACeDB as a 5 exon structure occupying a region between coordinates 3490-6742 of cosmid ZC21, and a reporter fusion with a 5' endpoint at position 2406 was found to direct expression (Table 3.2). Subsequently, cDNA sequence data indicated that the original 5 exons of ZC21.4 could provide the 3' end to a much larger gene, C04D8.1; also, the protein encoded by the cDNA of C04D8.1 was synthesised and found to be biochemically active (Chen *et al.*, 1994). The expression data I have generated for the predicted gene ZC21.4 suggests that it may form a genetic unit on its own as well as forming the 3' end of a larger gene (Figure 3.6), and that the originally predicted gene structure does in fact describe a separately definable gene.

The results are consistent with the suitability of this approach for a larger study.

53% of the genes assayed in this study gave a specific expression pattern. Thus, ~35% of the genes predicted in the genomic region covered yielded patterns of expression. Extrapolating to the entire genome, about 4-5000, i.e. 30% of the total gene number of 13-15000 (Waterston and Sulston, 1995), gene expression patterns could be generated, a resource of significant importance for *C.elegans* research.

Patterns including most of the major cell types were obtained: ectodermal lineages of the hypodermal and neural cell types (including neuronal support cells); mesodermal lineages including muscle (bodywall, anal, vulval), the excretory system and the spermathecae; and endodermal lineages of the intestine and pharynx. In addition, all developmental stages are represented apart from the very earliest embryos. Expected modes of expression are also present (Hope, 1991), i.e. transient expression in developing cells/structures, perhaps identifying genes with morphological functions only required at precise times and locations during development, and sustained expression indicative of cells that have terminally differentiated. Many patterns have multiple components, consistent with a view that many genes are employed in different cells and at varying times during development. One of the best examples of this phenomenon for a characterised gene in *C.elegans* is the *glp-1* gene which mediates cellular inductions in early embryos, where it is important for development of pharyngeal tissue (Priess *et al.*, 1987), and also in the adult, where it controls germline cell proliferation in the gonad (Austin and Kimble, 1987).

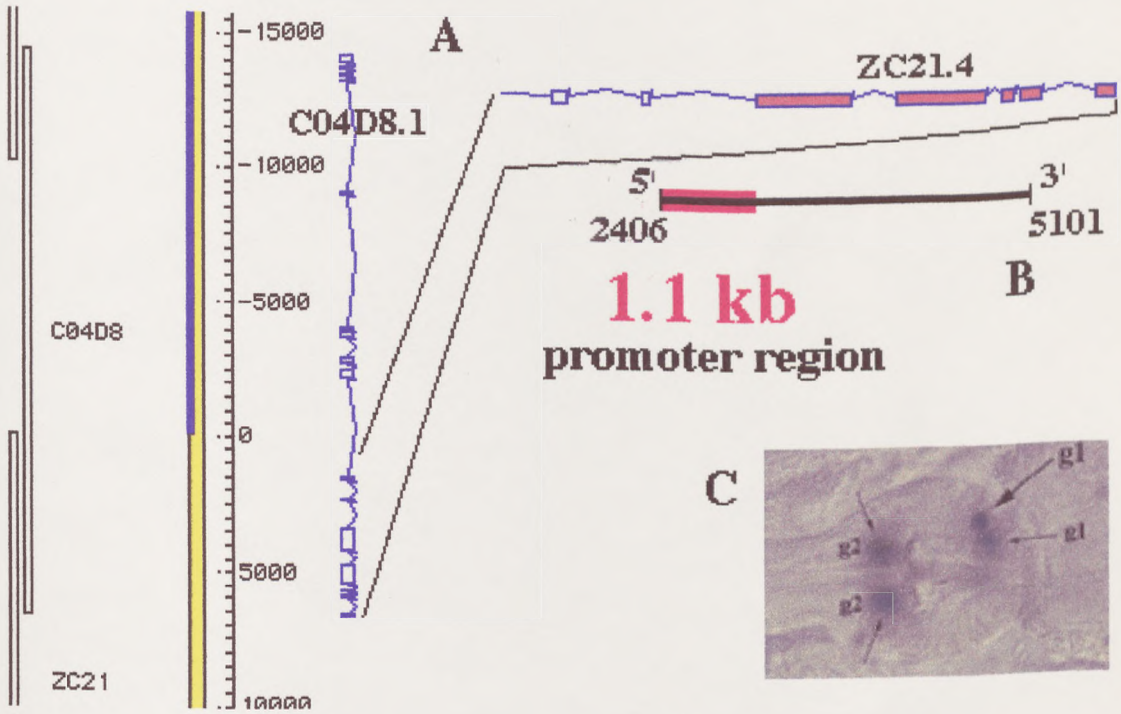


Figure 3.6. Genomic context of the predicted ZC21.4 gene.

A: The five exons of ZC21.4 (pink boxes) supply the 3' end of C04D8.1 (Chen *et al.*, 1994).

B: The genomic insert (stretching from position 2406 to 5101 in the published ZC21 sequence) used to assay expression of ZC21.4 contained only 1.1kb sequence upstream from the predicted start codon (highlighted in red). Promoter elements responsible for expression in the pharynx (C) must reside in this region which is intragenic with respect to C04D8.1.

C: Expression of ZC21.4 can be seen in the pharyngeal gland cells.

Hypotheses of function on the basis of expression patterns observed.

Gene expression pattern information will have an important role in interpretation of genome sequence data. This is particularly valid in *C.elegans* for which expression patterns can be described with single cell (and occasionally subcellular) resolution and protein coding regions are relatively easy to predict. Hypotheses of gene function based on genomic DNA sequence can be refined when the gene expression pattern is known, T23G5.5's homology to a neurotransmitter transporter providing a perfect example. In a similar way, the functional connectivity of the *C.elegans* nervous system, thus far largely inferred from structural connectivity based on electron micrograph reconstructions (White *et al.*, 1986), could be determined from the expression patterns of various nerve-specific proteins including neurotransmitter receptors and transporters. In this study, nervous system restricted expression patterns have been generated for two genes encoding G-protein coupled receptors, C38C10.1 and F59B2.13.

Detailed gene function may be suggested even in the absence of homology when expression patterns include information in the form of a restricted subcellular distribution. As in other screens (Hope, 1991; Young and Hope, 1993), there are many examples of subcellularly localised patterns of expression, presumably descriptive of the actual local context of the gene products. Only that for ZK643.1, however, has enabled functional assertions to be made. The rarity of obvious subcellularly localised expression may be linked to the preponderance of nerve specific patterns; detection of a particular subcellular distribution in neural cells may be precluded by their small size combined with the limited resolution of the β-Galactosidase staining procedure. One possible way to overcome this problem may be to exploit the greater resolution afforded by immunohistochemical techniques.

The potential for functional deductions on the basis of subcellular localisation could be increased by maximising the overall number of expression patterns of that character, perhaps by removal of the NLS from the expression vectors. Such an effect seems unlikely, however, as previous studies indicate that the NLS will be overridden by any other localisation signals present in the reporter fusion (Fire *et al.*, 1990; Roberts *et al.*, 1987). Loss of the NLS is thus likely to result merely in cytoplasmic expression.

The only criterion for selecting genes for analysis was the presence of suitable restriction sites that would facilitate subcloning of upstream regions into the reporter plasmids. Sequence homology to a proton pump ATPase subunit would perhaps not have suggested that the expression pattern for ZK637.8 would be amongst the most complex described here. The results for ZK637.8, and for ZK643.1 which has no database homologues at all, lend credence to the strategy of examining as many genes as possible with no prejudgement of significance based on homology to previously characterised genes. If

applied to the many predicted genes in the *C.elegans* genome, such a strategy would be of great benefit in the push towards a complete functional and developmental description of this important model organism.

Chapter 4

Enhancement of the Primary Screen.

As discussed in Chapter 3, the results of the primary screen suggest that a continuation screen would be capable of generating an excellent resource for study of *C.elegans* biology. There would, however, be obvious benefit in maximising the efficiency of expression pattern generation. Given the experimental approach employed (Figure 3.4) there are two easily assessable routes to this end. First, the number of genes assayed on each cosmid could be increased. Second, the probability of obtaining an expression pattern for any reporter fusion could be maximised. Both of these possibilities were addressed.

Increasing the number of genes assayed on each cosmid.

Introduction.

Approximately half of the genes in the genomic region covered could not be assayed because of the lack of suitable restriction enzyme sites. PCR is a technique which offers the possibility of generating user defined genomic fragments for construction of reporter fusions for genes not accessible by the restriction enzyme based approach, and has been used successfully in *C.elegans* to produce active *lacZ* fusions (Okkema *et al.*, 1993; Laughton, pers. comm.). *LacZ* fusions generated with PCR fragments are potentially more likely to be active than restriction enzyme generated fusions as optimal sites for the 5' and 3' ends of the genomic fragments used can easily be chosen. A 3' site can be chosen in the middle of the largest exon in the predicted gene, increasing the chance of its being in a true coding region (see introduction to Chapter 3). A 5' site can be chosen that is close to the end of the neighbouring upstream gene so as to be almost certain of containing all 5' transcriptional control elements.

There are, however, limitations on the size of genomic fragment that can be reliably amplified using the most common DNA polymerase for PCR, *Taq* polymerase (Saiki *et al.*, 1988), such that it is not consistently possible to generate readily clonable quantities of DNA for fragments greater than 2-3kb (Lawyer *et al.*, 1993). As the typical *C.elegans* gene has 2-3kb of upstream sequence containing its relevant promoter elements (see Chapter 3 discussion), PCR fragments generated using *Taq* would need to have their 3' ends in the most 5' of the gene's coding sequence, restricting the choice of translational fusion sites. *Taq* polymerase also exhibits misincorporation of bases into amplified DNA sequence at the rate of 10^{-4} mutations per base pair per PCR cycle, equivalent to 6 base pair changes in a 2kb sequence over 30 cycles. The danger of producing codon transformations this effect causes would also force translational fusion points to be restricted to gene 5' ends.

It has recently been shown that longer PCR is achievable using a mixture of *Taq* and *Pfu* polymerases, and that the error rate of this mixture is reduced to 10^{-5} mutations per base pair per cycle (Barnes, 1994). Both of these improvements are attributable to the integral

3'-exonuclease activity of *Pfu* polymerase which provides a proofreading function for the DNA polymerisation reaction (Lundberg *et al.*, 1991; Barnes, 1994). All other things being equal, therefore, an approach using both *Taq* and *Pfu* polymerases would be desirable over one employing just *Taq* polymerase. The two polymerase approach does have one significant problem, however, stemming from the unexplained observation that it requires oligonucleotide primers of no less than 35-40nt to produce an amplified DNA fragment (Barnes, 1994), compared to the 20nt primers required for amplification of 2-3kb fragments using 'normal' *Taq*-alone PCR. As the price of primer manufacture is the primary source of cost in PCR, a *Taq*/*Pfu* PCR reaction would therefore be about twice as expensive as a *Taq*-alone reaction (approximately £100 compared to £50, assuming primer costs of £1.25 per nucleotide). As the aim of this experiment was to assess the suitability of a PCR-based strategy for the generation of large numbers of *lacZ* fusions, I chose the *Taq*-alone strategy purely on the basis of expense.

A PCR-based strategy for generation of *lacZ* fusions.

Oligonucleotide primers for PCR were designed using the following criteria: i) primers no more than 2kb apart should be used to stay within the limits for reliable DNA amplification by *Taq* polymerase; ii) the 3' end of the fragment should be close to the translational start of the predicted gene to minimise the amount of coding sequence in the final gene-reporter fusion. This increases the likelihood of the fusion protein being nuclear localised as any gene-encoded trafficking signals are likely to be absent and so unable to overide the vector-encoded NLS (see introduction to Chapter 3) (Fire *et al.*, 1990). The risk of codon transformations due to PCR copying errors is also minimised. As mentioned in the introduction to Chapter 3, extreme 5' ends are the most unpredictable gene region. There was a risk that some 3' PCR fragment ends would thus be in non-coding sequence. However, even if the true 5' end was not predicted for a particular gene, the predicted first exon may still be included in the actual transcript at a more 3' position; iii) restriction enzyme sites were engineered into the primers to enable directional subcloning into the reporter plasmid MCS, as for the restriction enzyme generated fragments described in Chapter 3. To simplify the subcloning step, the same sites were used for all fragments and extra bases were included in the 3' primers as necessary to place all fragments in the same frame reading as the downstream reporter gene (Table 4.1). These steps enabled the same stock of restriction enzyme digested expression vector plasmid to be used for each fragment. Taken together, these three considerations placed the 5' endpoint ~2kb upstream of the gene's predicted translational start (Figure 4.1).

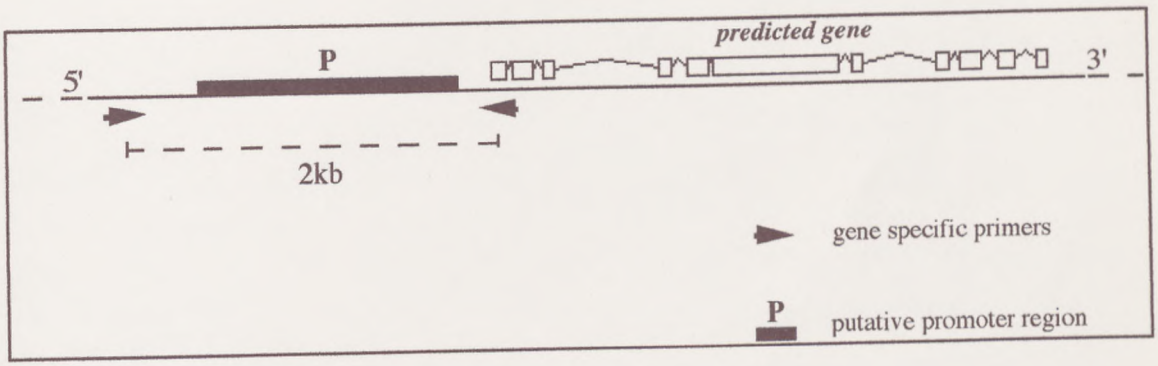


Figure 4.1. Placement of primers for PCR amplification of predicted gene 5' ends.

PCR primers are designed to span the 2kb of sequence just upstream of a predicted gene in *C.elegans* genome sequence.

plasmid	gene	primers	frame	strain
J1	ZK637.11	35684 ATCTGCAGTTACAGACCGTCAAAGTCGC		-
		33677 ATGGATCCATCGCASGAATCCGTTCCAGATG	2	
J2	ZK637.1	3413 ATCTGCAGCGCACTTCTCGATGAGCTTG		-
		1433 AGGGATCCTTGGCCGTTAAATCCACATATGC	0	
J3	ZK643.3	18348 ATCTGCAGAGCTTTTACAGTCGCCTGAC		UL62
		20160 ACGGATCCTCCGTGGATATCGATGGATTC	0	
J5	ZK643.7	3813 TGCTGCAGGTTGAAAGCTACACTTCACCG		-
		1864 ATGGATCCTCATGCATGAGTGTGCAAGCT	0	
J6	ZK643.6	35976 CTGCAGTTCATTGGATGACGCAAC		-
		34219 GGATCCTGTGCCTTTAAACCGAAG	0	
J7	ZK637.2	1825 *GCACCAGCACTATTATTGGTGAG		-
		3820 GGATCCAGTAGCCTCCATCGTTGAG	2	
J9	ZK637.12	34238 CTGCAGTGGCGATGTGCACCTATTTC		-
		36596 GGATCCTCATACGTGGCTTCGACGAG	1	
J10	ZK637.13	40699 CTGCAGATCATACGTGGCTTCGACG		UL48
		39271 GGATCCTTCAAGGGACTTCACACAG	2	

Table 4.1. Primers for generation of PCR fragments from cosmid clones.

Columns:-
plasmid: name of plasmid containing translational fusion of PCR fragment with *lacZ*
gene: name of gene supplying 5' end.
primers: upstream primers are above downstream primers. Coordinates are those of extreme 5' end in Genbank entry for cosmid. Engineered restriction enzyme sites are in red and green. Letters in **bold** signify nucleotidesengineered into the primer sequence to change the frame reading into the reporter gene. * - the 5' primer for ZK637.2 did not include a PstI site as a genomically encoded site just downstream of its position was available for use in subcloning.
frame: the reading frame of the genomic fragment relative to that of *lacZ* - see Table 3.2.
strain: strain designation of transformed lines exhibiting a specific *lacZ* fusion expression pattern.

Results.

Expression patterns of PCR-generated *lacZ* fusions.

Reporter fusions were made for eight genes not covered by the original screen, on two cosmids, ZK637 and ZK643 (Table 4.1). One gene, ZK643.3, which was covered originally but which gave an extremely weak, non-nuclear localised expression pattern was included with the intention of generating a stronger expression pattern. Two of the *lacZ* fusions generated specific patterns of expression.

ZK643.3

Nuclear localised expression is evident in the excretory cell from the threefold stage in late embryogenesis (Plate 4.1A) to the L1 larval stage (Plate 4.1B) and contains no other components. This pattern differs completely to that originally obtained for ZK643.3 (Chapter 3), which exhibited expression apparently localised at the cell membrane in the head muscles, vm1 vulval muscles and a cell in the posterior likely to be the anal sphincter. This unexpected result leads to speculation on the probable structure of the ZK643.3 gene. There seem to be two possible explanations for the differences, based on the two major differences between the reporter fusions responsible for each expression pattern. The first difference concerns the location of the fusion point to *lacZ* (Figure 4.2A). The original fusion, in plasmid pUL#AL3 (Table 3.2), is in exon 4, whilst the PCR fusion, in plasmid pUL#J3 (Table 4.1), is in exon 1. It is formally possible that alternatively spliced transcripts of the gene could exist and the two fusion points could be present in alternatively spliced messages; this, however, seems unlikely. cDNA analysis of the gene has identified only one transcript (Briggs and Coates, pers. comm.). cDNA was prepared from RNA extracted from mixed stage *C.elegans* by RT-PCR. Nested PCR on the cDNA with primers to the transpliced leader SL1 and internal gene specific primers to exons 9 and 6 produced only one cDNA fragment containing SL1 in its expected position just upstream of the predicted start codon. As the ultimate 3' primer used in these experiments was in exon 6, generation of more than one cDNA fragment would have been expected if alternatively spliced messages were extant for the upstream exons 1 and 4. Thus, no evidence of alternatively spliced transcripts was found.

The more likely explanation derives from the differences in potential transcriptional control regions between the two reporter fusions. These differences delineate three such regions (Figure 4.2). Region I (466bp) is in the 5' flanking sequence of ZK643.3 and lies between the 5' end of the pUL#AL3 genomic fragment and the 5' end of the PCR-generated pUL#J3 fragment. Region II (1809bp) is defined by the complete upstream sequence contained within pUL#J3 and also comprises the 3' end of the 5' flanking sequence of pUL#AL3. Region III (4611bp) consists of intron sequences in the 5' end of

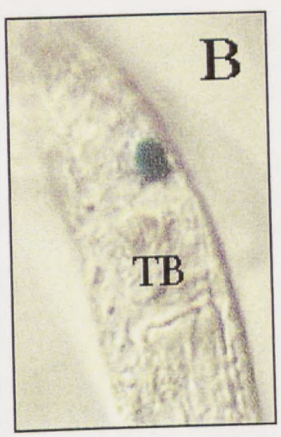


Plate 4.1. *LacZ* fusion expression in UL62 (ZK643.3).

A: A 3-fold embryo just prior to hatching showing expression in the excretory cell nucleus.
B: The same nucleus can be seen in this L1 larva, just after hatching. TB- terminal bulb of pharynx.

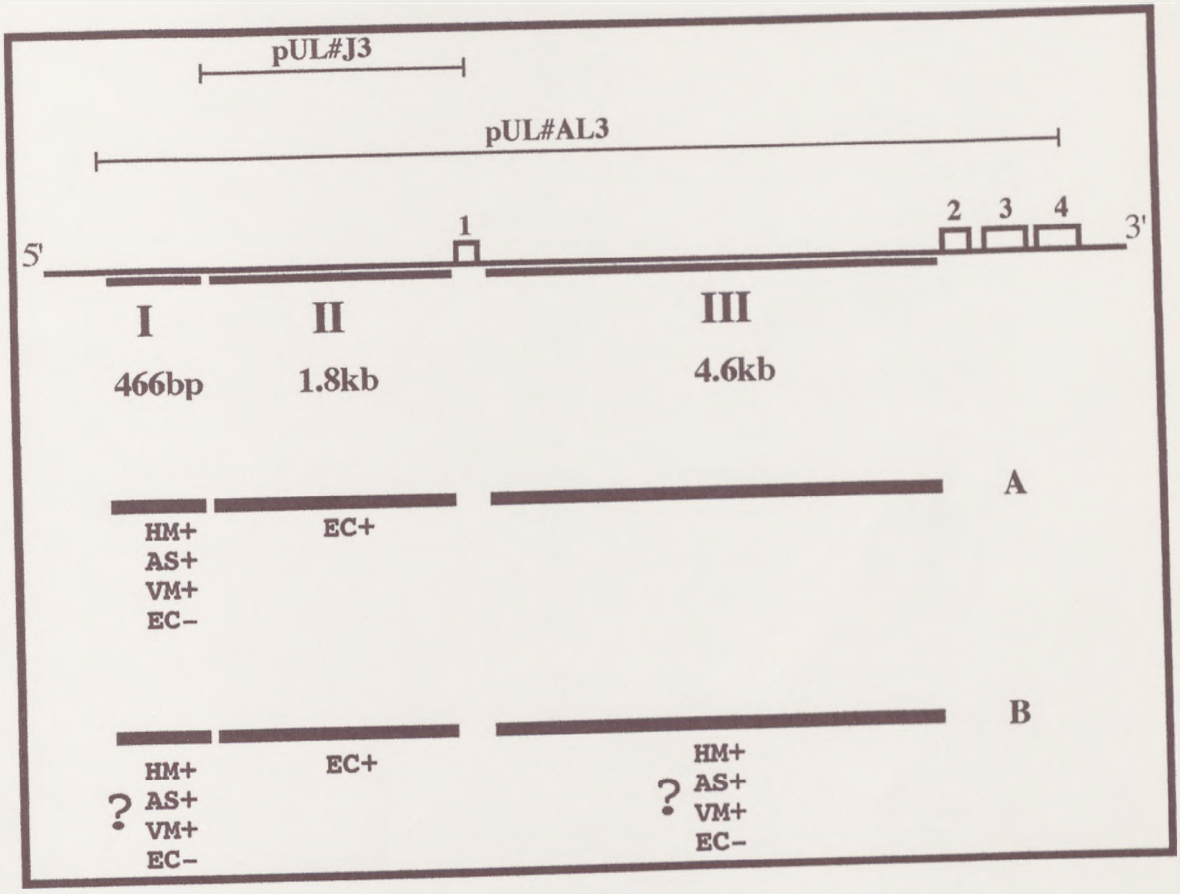


Figure 4.2. Possible organisation of the promoter region of ZK643.3.

The genomic fragments contained within reporter plasmids pUL#AL3 and pUL#J3 define 3 possible discrete regulatory regions, I, II and III. Cell specific regulatory elements are encoded within these regions: HM = head muscle; AS = anal sphincter; VM = vulval muscle; EC = excretory cell. Putative positive elements are marked "+"; negative elements "-". The "?" in section B denotes the possibility of elements being in either Region I or III.

the ZK643.3 gene not included in pUL#J3 but present in pUL#AL3. As related in Chapter 3, regulatory sequences have been found in large 5' introns of *C.elegans* genes.

What models of the promoter structure of ZK643.3 can be mapped onto this framework of DNA sequences given the expression data provided by the two *lacZ* fusions? The simplest arrangement assumes that only 5' flanking regions contain regulatory sequences. In this case, Region II must contain a positive control element for excretory cell expression as this is the only upstream sequence present in pUL#J3. The components of expression driven by pUL#AL3 indicate that positive elements for head muscle, vulval muscle and anal sphincter expression lie in Region I, along with a negative regulatory element for excretory cell expression which overrides the positive element in Region II (Figure 4.2A). This model is the most parsimonious, and so should perhaps be judged the more credible. Another interpretation can be made, however.

pUL#AL3 also contains non-coding sequence in the two large introns at the 5' end of the gene which pUL#J3 does not. The regulatory elements ascribed to Region I above could instead, either in part or in total, be located in the intron, Region III (Figure 4.2B). The available data cannot distinguish between these alternatives; however, a relatively simple set of experiments aimed at specifying the separate regulatory abilities if Regions II and III can be envisaged. These involve construction of two further *lacZ* fusions, one containing only Regions I and II and the other containing Regions I and III. Observation of the resultant expression patterns in *C.elegans* strains containing such reporter fusions should unequivocally identify individual pattern components with a particular sequence region.

ZK637.13

Expression is first observed at the 2-fold stage of embryogenesis in many hypodermal cells, mostly in the anterior portion of the animal (Plate 4.2A). By hatching, expression is restricted to the hyp3 and hyp4 hypodermal cells at the extreme anterior of the animal (Plate 4.2B) (Sulston *et al.*, 1993) though occasional staining is seen in the more posterior hyp 6 hypodermal cells close to the pharyngeal terminal bulb (Plate 4.2C). This pattern of expression continues into adulthood.

ZK637.13 is homologous to the nematode globins (Frenkel *et al.*, 1992; Sherman *et al.*, 1992). These proteins form a family of diverse structure and function (Blaxter, 1993). The most homologous protein is the myoglobin-like isoform expressed in the bodywall of *Nippostrongylus brasiliensis* (Blaxter *et al.*, 1994), a gut parasite of rodents. This protein binds oxygen but deoxygenates when the worm is kept under anaerobic conditions, after which the worm ceases to move (Sharpe and Lee, 1981), indicating that the protein is involved in oxygen storage for the nematode's muscles. The bodywall expression of WP:ZK637.13 is consistent with such a view of globin function.

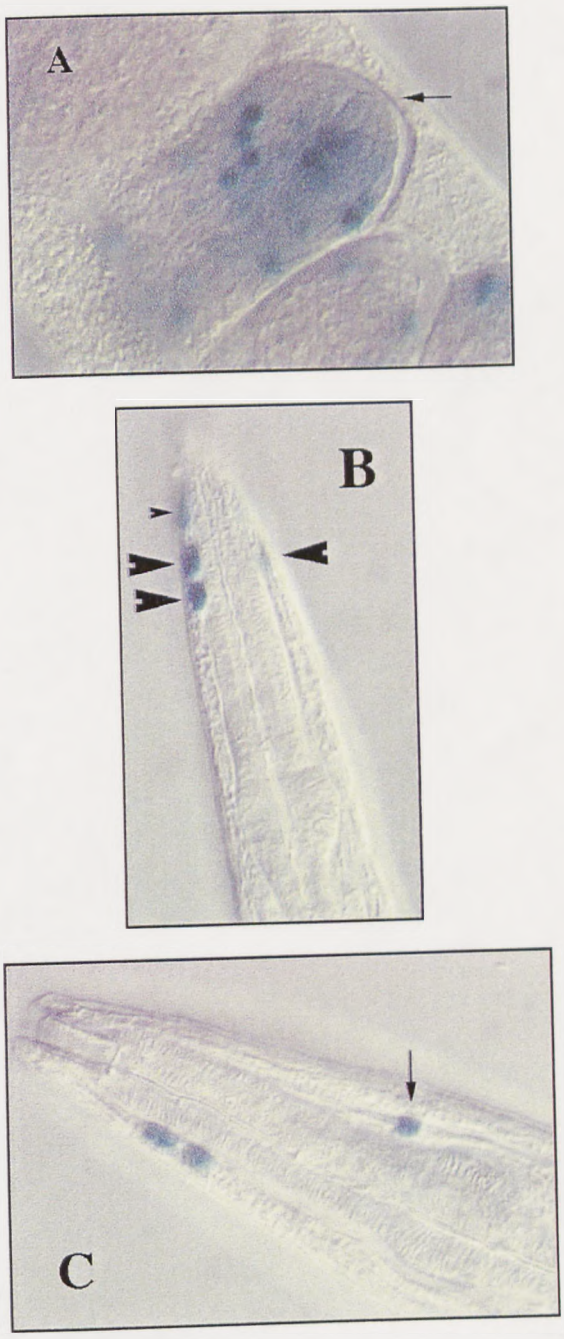


Plate 4.2. *LacZ* fusion expression in UL48 (ZK637.13).

A: Hypodermal nuclei in the anterior of a 3-fold embryo. The arrow marks the tip of the snout of the developing animal.
B: An L3 larva exhibits expression in the hyp3 (large arrowheads) and hyp4 (small arrowhead) hypodermal cells.
C: Expression is also sometimes seen in hyp6 (arrow) hypodermal cells.

One might expect, however, that a protein acting in such a capacity would be expressed in hypodermal cells close to muscle cells in all parts of the nematode. The anterior hypodermal expression of WP:ZK637.13 has an interesting parallel with expression of a red pigment myoglobin-like globin found in another nematode, *Mermis nigrescens* (Burr *et al.*, 1975; Burr and Harosi, 1987). The globin is only present in anterior hypodermal cells and has been proposed to function as a shading chromophore for the light sensitive organs responsible for the phototrophism of the animal. As discussed for ZC21.2 in Chapter 3, the photosensitive structures in *C.elegans* are also suspected of being in the anterior of the worm, in the amphid sensilla which are embedded in the anterior hypoderm (White, 1988). While WP:ZK637.13 is not a coloured protein, its specific expression in these hypodermal cells may represent an ancestral arrangement allowing subsequent acquisition of a function in the phototrophic response.

Discussion.

PCR route is viable.

These experiments were carried out to examine the value of a PCR based assay for generation of gene specific expression patterns. As it stands, however, there are disadvantages arguing against replacing the original strategy with the PCR based strategy. Only 25% of the constructs assayed gave an expression pattern (Table 4.1), a success rate half that of the original approach. However, the small sample size represented here means that more genes would need to be examined to give statistical significance. A lower success rate may be expected, however, given the selection of first exons for providing translational fusion to *lacZ*. The ability to generate PCR fragments containing the fusion point in the centre of larger exons, and thus more likely to be in actual coding sequence (see introduction), may increase the proportion of active PCR generated constructs.

As argued in the introduction, 'long PCR' could supply this ability but seems a prohibitively expensive approach given the requirement of using large (approximately 40nt) oligonucleotide primers. One way of subverting the requirement of primer manufacture is to use restriction enzyme fragments in their stead (Barnes, 1994). If a restriction enzyme fragment was used as the upstream primer in a PCR reaction, a larger downstream primer could be made for the price of the two primers used in the experiments above. Use of these primers in a long PCR reaction should successfully amplify larger genomic fragments enabling sufficient upstream regulatory sequence to be included. As indicated above , the proofreading function provided by *Pfu* polymerase should minimise the risk of introducing novel mutations into coding sequence and so maximise the chance of generating active *lacZ* fusions.

Maximising expression of gene-reporter fusions.

Introduction

It is likely that the expression vectors used in the original screen do not comprise an optimal system for assaying the expression of many *C.elegans* genes, such as those with very low or transient expression levels (discussed in Chapter 3). This concern has recently been addressed by the originators of the plasmids, who are involved in ongoing attempts to improve and diversify the available set of vectors for use in expression studies in *C.elegans*. As related in Chapter 3, it had already been noted that inclusion of a synthetic intron in the 5' end of translational fusions could significantly stimulate the level of expression (Fire *et al.*, 1990). *In situ* experiments had also revealed that transcripts of transgenes integrated into the genome were primarily seen in two spots in the nucleus, presumably identifying the site of integration, whilst the mRNA of endogenous genes was mostly found in the cytoplasm suggesting that reporter fusion messages were inefficiently exported from the nucleus (Okkema and Fire, 1994; Fire, pers. comm.). Acting on the suspicion that this failure could be due to the lack of intron sequences in the *lacZ* gene of the reporter fusions, experiments were performed to assay the effect of intervening intron sequences on *lacZ* fusion expression.

Insertion of multiple artificial intron sequences into the *lacZ* gene sequence of a vector containing a weak *unc-54* promoter resulted in an maximal 2-3 orders of magnitude increase in reporter expression (Fire, pers. comm.), with one intron insertion producing a small increase and twelve intron insertions producing the maximal increase. *In situ* analyses of the transcipts also reveal greater cytoplasmic accumulation of intron-rich reporter genes. Such reporter genes have improved the known expression characteristics of the original vectors in several other ways. For example, mosaicism of expression is decreased with both number of cells in an expression pattern component as well as the staining intensity being increased. For *pes-10* gene fusions expression as early as the 12-cell stage of embryogenesis can be detected, improving significantly (from the 28-cell stage) the coverage of gene expression possible in early embryos, and the pattern of transgene expression seen is close to that observed for the endogenous protein (Fire, pers. comm.; Seydoux and Fire. 1994).

All aspects of expression are enhanced, including the occasional ectopic expression in pharynx and intestine which is thought to result from inappropriate transcriptional activity by cryptic enhancer sequences in the bacteria-derived plasmid backbone of the expression vectors (Krause *et al.*, 1994; Hope, 1991). Large (2-3kb) genomic insertions can usually remove this influence from the *lacZ* gene in the original expression vectors, but may not be enough in every case to prevent ectopic expression of the intron-rich reporter (Fire, pers. comm.). Even if this is the so, however, the pharyngeal and intestinal expression is

easily recognised and can be discounted as specific expression when gene fusion expression is observed.

Test of applicability of intron-rich *lacZ* vectors in an expression pattern screen.

These recent developments have obvious relevance for the type of expression study described in Chapter 3. Assuming that some of the non-expressing genes (~50%) found in the screen to date have revealed no expression due to low expression levels, the percentage of genes giving an expression pattern could be increased by using the new *lacZ* vectors. In addition, the quality of expression data generated could be improved through decreasing the level of mosaicism of reporter expression, and the information content of observed expression patterns expanded by enabling earlier embryonic expression to be detected. I therefore conducted experiments to gauge the scale of improvement that could be expected when using the new expression vectors in the type of approach described in Chapter 3.

Selection of genes for assay.

To estimate the degree of probable impact, several genes (Table 4.2) were selected from those already assayed and tested for increase in expression level with the new *lacZ* reporter gene. Genes which previously gave weak expression patterns were chosen as it would be easier to score the degree of increase in signal for such genes' expression over any novel appearance of expression for a gene showing no expression at all. Two genes were included to assay any improvement in embryonic expression: ZK637.8 has a component starting at the 28-cell stage with the original *lacZ* vectors presenting the possibility of expression prior to this time; ZK637.11 is known to be expressed as a maternal transcript in early embryos by *in situ* hybridisation (Golden, pers.comm.), but no expression was observed for this gene using a PCR generated reporter fusion (Table 4.1).

Genomic 5' ends identical to those present in the original recombinant DNA constructs were used (Table 3.2) by simple substitution of the reporter gene cassette (Figure 4.3) so that any effects seen could be attributed only to the new *lacZ* gene. The reporter gene cassette can be subcloned using either of two sets of restriction enzymes: AgeI and ApaI, or ApaI and KpnI (Figure 4.3). Use of ApaI/AgeI enables the NLS cassette to be retained; use of KpnI/ApaI results in loss of the NLS cassette from the new fusions. As the original genomic fragments for ZK643.3 and B0303.12 contained AgeI restriction enzyme sites, the new recombinant fusions for these genes were constructed using the KpnI/ApaI *lacZ* fragment and so contained no NLS. The original expression patterns of the *lacZ* fusions for both of these genes were not nuclear localised, however, so loss of the NLS was not expected to affect expression.

gene	plasmid	NLS	strain
ZK643.3	pUL#NIL3	no	-
ZK637.8	pUL#NIL10	yes	UL85
B0303.12	pUL#NIL16	no	UL61
C40H1.6	pUL#NIL26	yes	UL86
C38C10.1	pUL#NIL27	yes	-
ZK637.11	pUL#NIL1	yes	UL90

Table 4.2. Genes assayed for expression with an intron-rich *lacZ* gene.
Column:-
gene: name of gene assayed
plasmid: name of plasmid containing genomic fusion to intron-rich *lacZ* gene
NLS: does the plasmid contain a nuclear localisation signal?
strain: if plasmid is active, strain designation.

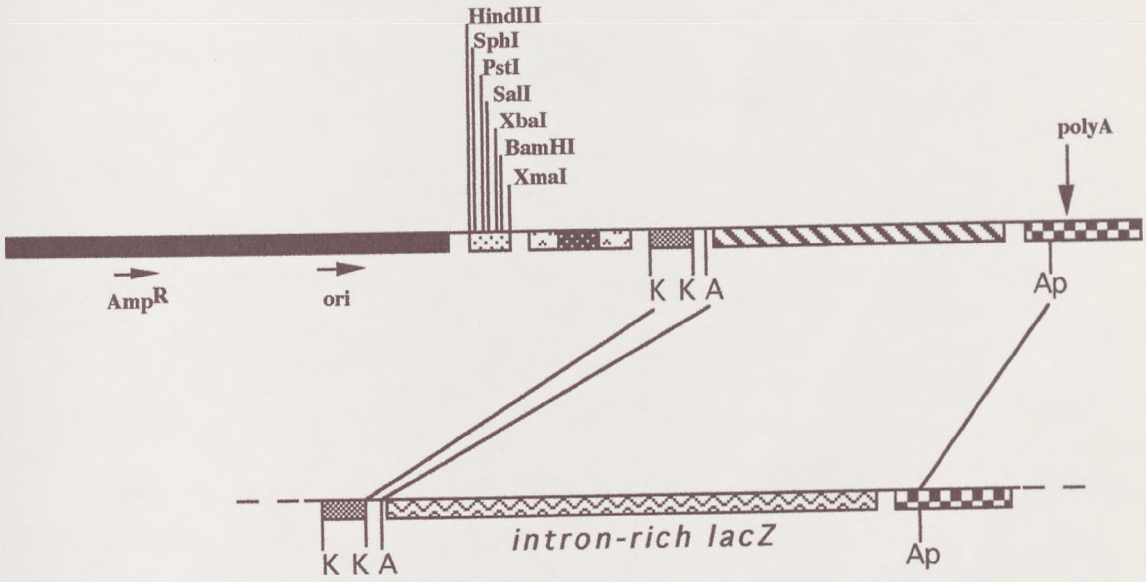


Figure 4.3. Generation of reporter fusions incorporating an intron-rich *lacZ* gene.

The upper plasmid represents the original expression vector plasmids containing an intronless *lacZ* gene (see Figure 3.3A for legend). Substituting an intron-rich *lacZ* gene requires a simple directional ligation with either KpnI (K) and ApaI (Ap), in which case the NLS module is lost, or with AgeI (A) and ApaI, in which case the NLS cassette is retained

Results.

B0303.12

The fusion gene for B0303.12 contained in the intron-enriched plasmid pUL#NIL16 generates expression of the same tissue-specificity as the original reporter construct, pUL#AL16 (Chapter 3); that is, in the excretory system and in the bodywall muscles. The patterns do differ qualitatively: the original expression pattern was of a diffuse, cytoplasmic nature while the new pattern exhibits subcellularly localised expression. This difference can be ascribed to the lack of a NLS in pUL#NIL16 (Table 4.2) allowing transport of B0303.12::*lacZ* to the normal location of the endogenous gene product in the cell. Whilst the subcellular location of gene product allows more detailed description of the expression pattern compared to that generated by pUL#AL16, expression from the intron-rich reporter fusion is also, on a subjective scale, stronger than the original.

The bodywall muscle expression pattern of the newly generated *lacZ* fusion consists of strong staining in discrete foci scattered along their length on the hypodermal interface (Plate 4.3A). Expression is noticeably more pronounced at the anterior and posterior extremities of the muscle rows (Plate 4.3B). Molecular and anatomical aspects of the infrastructural connectivity between muscle, hypodermis and cuticle have been well documented using both electron microscopy and antibody analysis (reviewed by Waterston, 1988). Connections start at the dense bodies and M-lines in bodywall muscles and continue through the hypodermal cell layer to the cuticle. The apparent localisation of B0303.12 gene product to a subset of these structures argues for a role in channelling the force of muscle contraction in nematode movement. That only a subset of such intercellular connections contain WP:B0303.12, and the obvious anterior/posterior bias of the pattern of expression, suggest a specialised role however. One possibility is for WP:B0303.12 to distribute the strain of muscle contraction along the length of each muscle quadrant, the anterior/posterior concentration of protein being analogous to the attachment points at either end of a free-spanning bridge.

The second difference to the original pattern concerns the expression specific to the excretory system. pUL#AL16 gave cytoplasmic expression in the excretory duct cell only; pUL#NIL16 expression covers a wider area of this region, but the subcellularly localised expression and the complex anatomy of these tissues (Nelson *et al.*, 1983) make it difficult to describe in precise cellular terms. As in the muscle component, expression is punctate and seems to include the cells of the excretory system which interface with the hypodermis. Foci in the pore cell can be discerned surrounding the excretory pore (Plate 4.3C). Lines of punctate staining radiate from the pore in four directions (Plate 4.3D), constant with the shape of the excretory gland cell (Nelson *et al.*, 1983). Finally, punctate staining can also be seen along the path of the excretory cell arms running

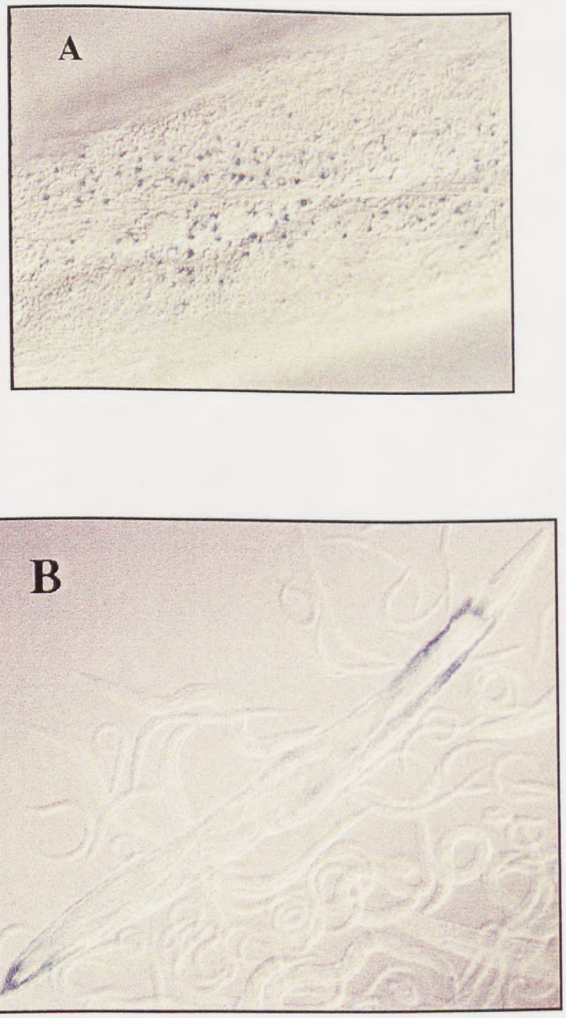


Plate 4.3. *LacZ* fusion expression in UL61 (B0303.12).

A: Punctate staining in bodywall muscles of an adult hermaphrodite.
B: Lower magnification view of the same animal as in the previous image showing a distinct anterior/posterior bias in bodywall muscle expression.

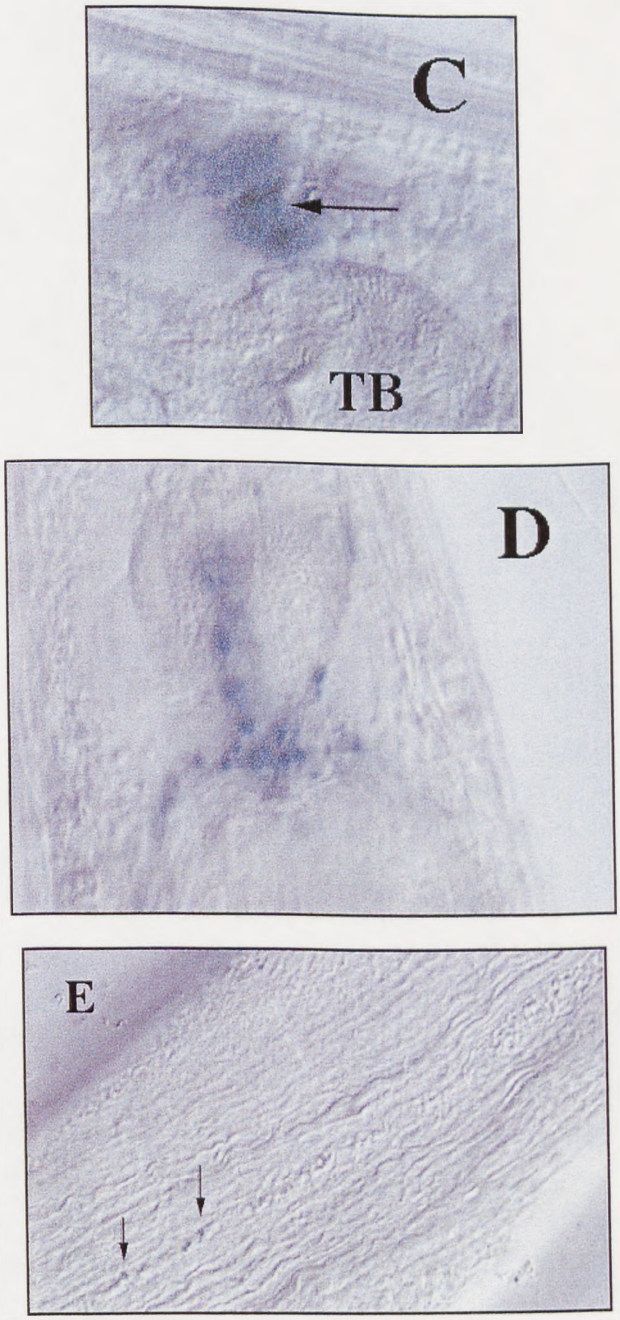


Plate 4.3 contd...

C: Punctate staining around the excretory pore (arrow) in the head of an adult hermaphrodite. TB - terminal bulb of pharynx.
D: A different example reveals lines of punctate staining leading away from the excretory pore.
E: More punctate staining along the line of the lateral processes of the excretory cell in the midbody of an adult hermaphrodite.

laterally along the animal (Plate 4.3E). Thus, as in the bodywall muscles, WP:B0303.12 seems to be involved in structurally connecting the cells of the excretory system directly underlying the hypodermis to the hypodermal cell layer.

ZK637.11

Expression is evident from the 28-cell stage of embryogenesis when all 16 AB-derived cells exhibit expression (Plate 4.4A). The cell Ea also shows expression at this stage, but expression does not continue in the E lineage past the 28-cell stage of embryogenesis. All members of the AB lineage seem to express the fusion until the 64 AB stage (~100 cell embryo) when staining begins to fade (Plate 4.4B), although the large number of cells expressing make it difficult to specifically identify every cellular component of the pattern.

ZK637.11 is a homologue of the yeast cell-cycle gene CDC25. CDC25 is involved in control of the switch from the G2 phase to the M phase in the mitotic cell cycle (Feilotter *et al.*, 1992). The CDC25 protein is a phosphatase, and functions by dephosphorylating threonine and tyrosine residues of the key control gene product, CDC2 (Millar and Russell, 1992). Dephosphorylated CDC2 then induces the progression into the mitotic stage of the cell cycle. The activity of CDC25 is under tight temporal control, allowing CDC2 dephosphorylation to occur only after DNA replication has been completed. Each of the founder cell lineages of *C.elegans*, including that derived from the AB blastomere, has its own unique division period (Sulston *et al.*, 1983). It can thus be inferred that WP:ZK637.11 functions to regulate the rate of mitotic division in the AB lineage during early embryogenesis.

In situ analysis has identified a maternal contribution of ZK637.11 mRNA in the early embryo (Golden, pers. comm.). Transcript was seen to segregate specifically to the anterior AB cell upon division of the fertilised 1 cell egg, and was detected in all AB-derived cells up until the 16 AB cell stage when the *in situ* signal ceased. No zygotic transcripts of ZK637.11 were detected at any stage of embryogenesis. Taken together, the *in situ* and reporter fusion data for ZK637.11 suggest a dynamic view of expression during the first 6 rounds of mitosis in *C.elegans* embryos which I have outlined below.

Translational expression of *lacZ* fusions has never been obtained for germ-line-derived transcripts (Krause, 1995; Seydoux and Fire, 1994), explaining the lack of β-Galactosidase expression for the maternally supplied ZK637.11 transcripts identified by *in situ* hybridisation. This maternal contribution lasts until the 16 AB stage, when zygotic transcripts take over the function of cell-cycle control. A similar switch from maternally-supplied to zygotically-supplied CDC25 has been documented in the *Drosophila* embryo (Edgar and O'Farrell, 1989). The zygotic transcription produces active β-Galactosidase fusions and so ZK637.11::*lacZ* expression is seen from the 16 AB cell stage. Steady state

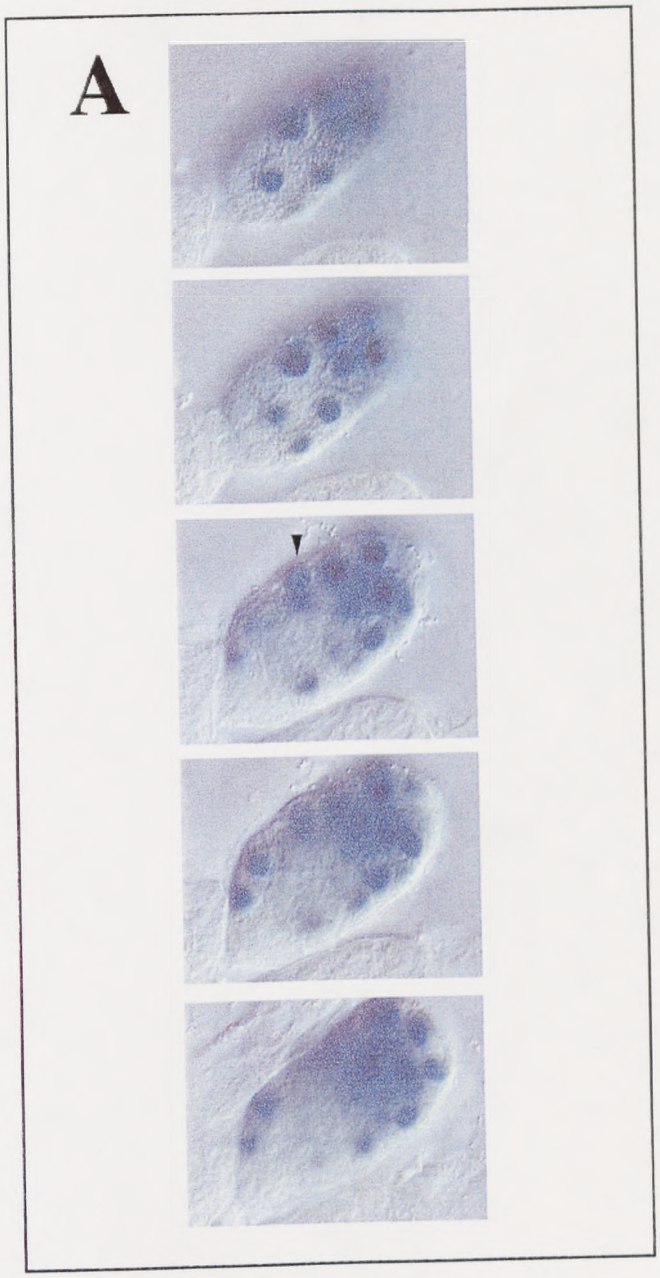


Plate4.4. *LacZ* fusiion expression in UL90 (ZK637.11).

A: A sectional series through a 28-cell embryo showing expression restricted to all cell descendants of the AB blastomere, and also faint expression in the Ea cell (arrowhead). Anterior is to the upper right.

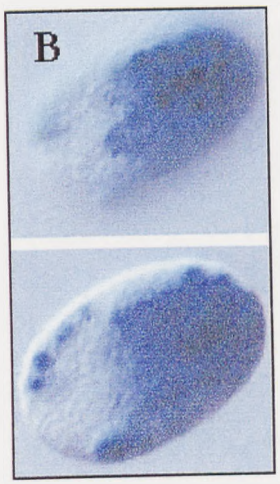


Plate 4.4 contd...

B: 2 sections through an approximately 100-cell embryo showing expression in members of the AB lineage. Anterior is to the upper right.

mRNA levels of zygotic ZK637.11 transcripts must be relatively low as no *in situ* signal is observed at these embryonic stages.

Reporter expression continues until the 64 AB stage when ZK637.11 expression ceases. It is interesting to note that the mitotic cycle producing 128 AB cells is the last to have a short, approximately 5 minute, period covering division of the entire AB-derived cell population (Sulston *et al.*, 1983). Subsequent divisions of the AB-derived cell population are at least 15 minutes in duration. Perhaps the loss of ZK637.11 expression between the 64 and 128 AB cell stages signals the start of cellular differentiation on the molecular scale, as different cells begin to divide at different rates under the control of different cell-cycle control genes.

C40H1.6

The original reporter fusion for C40H1.6 gave weak and mosaic expression in a few hypodermal cells through mid-embryogenesis, and in a nucleus identified as that of the head mesodermal cell through all postembryonic stages. Expression with the new *laZ* gene was stronger and less mosaic in the hypodermal cells, unchanged for the head mesodermal cell, and also included completely novel pattern components compared to the original. Expression was first observed in hypodermal nuclei just as the embryo was entering the elongation phase of embryogenesis (Sulston *et al.*, 1983) and continued in many hypodermal cells throughout the embryo until the 3-fold stage, finally fading just before hatching (Plates 4.5A-C). Expression in the head mesodermal cell was detected in early embryonic stages and continued into adulthood.

The novel component of expression involved all cells of the somatic gonad except for those forming the uterus (Hirsh *et al.*, 1976; Kimble and Hirsh, 1979) (Plate 4.5D), that is: the distal tip cell (Plate 4.5E); the sheath cells of the distal ovary (Plate 4.5F); the endothelial sheath cells of the oviduct (Plate 4.5F); and the cells forming the spermathae (Plate 4.5G). Expression in most cells covered the period of gonad morphogenesis, but the spermathecal cells continued expression into sexually mature adulthood (Plate 4.5H).

C40H1.6 contains no homology to known genes. The expression in hypodermis and epithelial cells of the gonad as those structures are forming may indicate, however, a role in the morphogenesis or final differentiation of entire epithelial tissues.

ZK637.8

The pattern of expression observed with the new *lacZ* fusion for ZK637.8, contained within pUL#NIL10 (Table 4.2), is identical to that observed in strains containing the original reporter construct for this gene, pUL#AL10. In particular, embryonic expression from the 28 cell stage was detected, but non prior to the 28-cell stage. The original pattern was particularly strong and so it is likely that expression of the fusion protein was

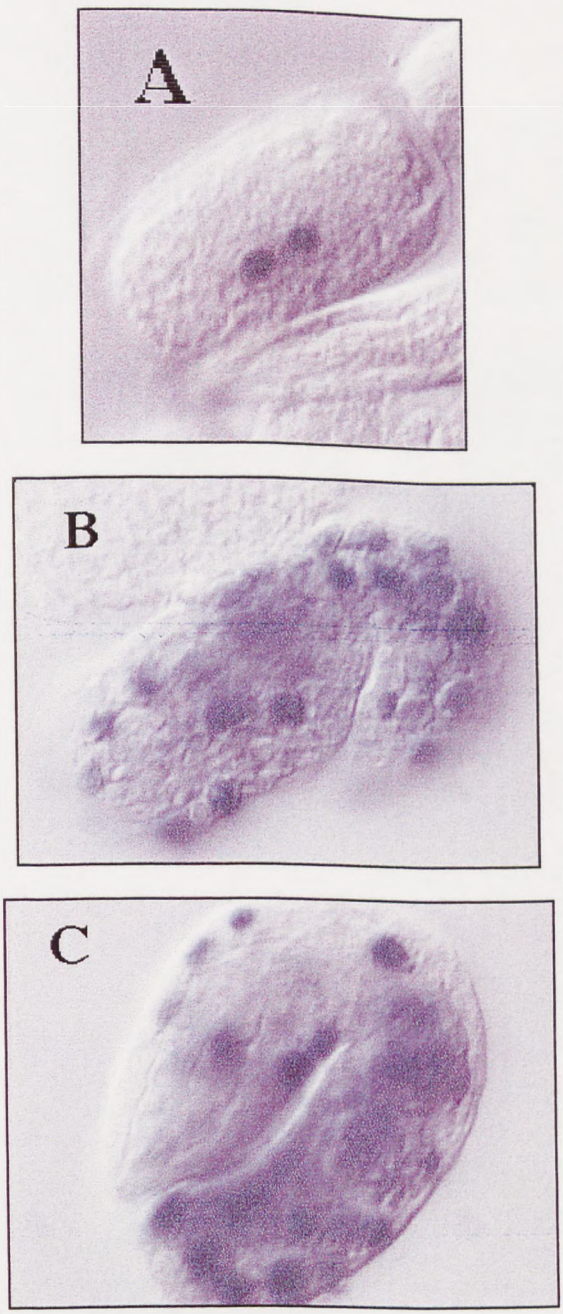


Plate 4.5. *LacZ* fusion expression in UL86 (C40H1.6).

A: A pre-elongation phase embryo showing expression in two cells in the centre of the embryo.
B: A 1.5-fold embryo showing expression in hypodermal cells in all regions.
C: A 3-fold embryo exhibiting continued expression in embryonic hypodermal cells.

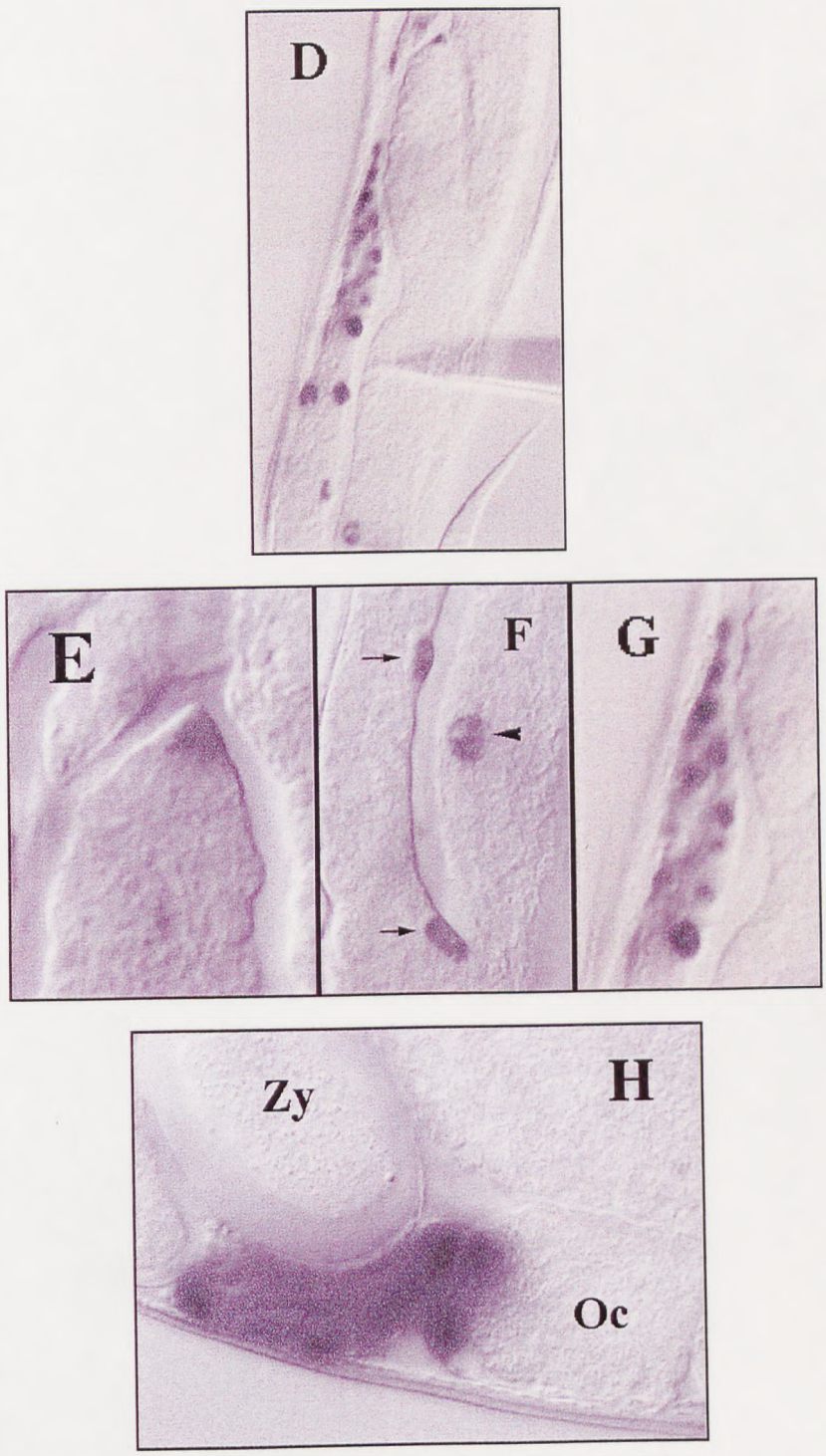


Plate 4.5 contd...

D. Expression in the developing somatic gonad of an L4 larva.
E: A closer view of the expressing distal tip cell.
F: A closer view of the reflex region of the gonad where nuclei of the sheath cells of the distal ovary (arrowhead) and of the more proximal oviduct are seen to stain.
G: A closer view of the nuclei expressing in the developing spermathecal tissue.
H: The cells of the spermathathecae continue expressing into adulthood where they clearly separate the oocytes (Oc) in the oviduct from the fertilised zygotes (Zy) in the uterus.

saturating for β-Galactosidase activity. In this case, increasing fusion protein expression would have no observable impact on β-Galactosidase assay. For ZK673.8 then, the earliest onset of expression is confirmed at the 28-cell stage of embryogenesis.

Finally, the fusions for ZK643.3 and C38C10.1 fail to generate patterns of expression. That for ZK643.3 contains no NLS (Table 4.2). Lack of the NLS in the reporter construct for B0303.12 resulted in a switch of B0303.12::*lacZ* expression from the cytoplasmic compartment to a specific subcellular localisation; a similar effect on ZK643.3::*lacZ* localisation should result in loss of β-Galactosidase activity as *lacZ* is predicted to be fused to an extracellular region of WP:ZK643.3 (Table 3.4), the inactivation of the enzyme on passage across a membrane having been already discussed in Chapter 3. Even though the *lacZ* reporter is also predicted to be in an extracellular region of WP:C38C10.1, this cannot explain the lack of expression with the intron-rich reporter gene in pUL#NIL27 as the NLS is still present in the new reporter construct and directed nuclear localised expression of the original fusion contained in pUL#AL27 (Chapter 3). The failure to observe expression is puzzling, therefore, and remains unexplained.

Discussion.

Intron-rich *lacZ* does promote stronger and less mosaic transgene expression.

Three of the six genes assayed gave better quality expression patterns with the new *lacZ* reporter than with the original version. Signal strength was improved for fusions to B0303.12, C40H1.6 and ZK637.11 but only the improvement observed for ZK637.11 was on a par with the reported 2-3 orders of magnitude increase observed for the weak *unc-54* promoter assayed by Fire and co-workers (see Introduction). The patterns obtained for these three genes also included completely novel cellular components, with expression observed in many cells of the somatic gonad for C40H1.6, in cells of the early embryo for ZK637.11, and in the excretory cell and excretory gland cells for B0303.12. Mosaicism of expression was also greatly improved in the fusion for C40H1.6, which was seen to express in the majority of hypodermal cells during embryogenesis using the intron-rich reporter gene, as opposed to the rather more limited hypodermal expression seen using the original *lacZ*.

Embryonic expression is facilitated by intron-rich *lacZ* fusions.

Of especial importance is the embryonic expression obtained for ZK637.11. The results of the primary screen, described in Chapter 3, were noticeably lacking in early embryonic expression patterns. The ZK637.11 result suggests that a significant contributing reason for this could be the particular *lacZ* reporter gene used, and indicates that use of expression vectors containing an intron-rich *lacZ* gene could increase the number of patterns with embryonic components.

The vector-encoded NLS has significant control over character of expression.

One characteristic of reporter expression seen here, and not in the original screen, concerns the effect on expression of the NLS. It is included in the expression vectors to encourage nuclear localisation of reporter fusions to aid in cell identification and prevent secretion of fusion protein which inactivates the β-Galactosidase enzyme (Chapter 3). The original expression pattern of B0303.12 generated by pUL#AL16 was cytoplasmic in character, indicating that the NLS was overridden by trafficking signals encoded by the reporter fusion protein. When the NLS was removed from the expression construct, as in pUL#NIL16, expression became localised to defined foci at the interface with hypodermis. The original expression seen for ZK643.3 was also cytoplasmic and removal of the NLS in this case resulted in loss of expression, possibly because of the putative extracellular location of β-Galactosidase described above. It thus seems possible that even when the NLS is overridden, its presence in a reporter fusion can still disrupt localisation of gene product.

The data for B0303.12 also emphasises the insight into gene function that is gained when subcellular localisation of gene product is known. The pattern of expression given by pUL#AL16 shed little light on WP:B0303.12 function; it merely identified the expressing cells. The distinct foci of expression in bodywall muscles revealed by pUL#NIL16, however, led to a testable hypothesis of a structural role. One way to test the proposed role of WP:B0303.12 would involve generation of a complete loss-of-function mutant by the transposon-mediated procedure described in Chapter 1 (Zwaal *et al.*, 1993). An uncoordinated (unc) phenotype (Brenner, 1974) would strongly support the proposed role as loss of a close structural relationship between the muscles and the bodywall would be expected to affect the locomotive properties of the animal.

It is interesting to note the large effect on the expression patterns of both of the genes for which the NLS was removed. A positive effect was seen in the case of the subcellular localisation of B0303.12::*lacZ*, whilst the loss of ZK643.3::*lacZ* expression is certainly to be viewed as a negative effect. Any proposals for systematic use of reporter fusions lacking the NLS must take into account these potential consequences, and are discussed in the next section.

Recommended modifications to strategy for expression pattern screen.

Taking all the above findings into account, it is possible to frame a proposal for improvement of the original methodology described in Chapter 3. To improve the quality of expression data generated, and broaden the set of temporal and spatial expression compartments able to be included, it seems reasonable to suggest that all genes covered in a continuation screen should be assayed using the set of reporter plasmids incorporating an intron-rich *lacZ* gene. In addition, those genes which are found to give a specific

pattern of expression which is not subcellularly localised should be further analysed by generation of constructs lacking a NLS, a trivial construction requiring simple excision of the small NLS-containing KpnI restriction enzyme fragment (Figure 4.3). The extra work involved will detract from the total number of genes covered in a screen, but the extra information gained may be useful in terms of proposing functional hypotheses. Indeed, it should in any case be possible to select from amongst those *lacZ* fusions giving non-subcellularly localised expression patterns: first, genes with very high homology which suggests a specific function in those cells comprising the pattern may not benefit from observation of their subcellular localisation (e.g. T23G5.5 in Chapter 3); second, genes whose *lacZ* fusion expression was nuclear localised are not likely to become targeted to a specific cellular compartment as previous studies have shown that any other protein targeting signals present in a protein sequence will prevent nuclear localisation due to the NLS (Roberts, 1987). Therefore, only those fusions exhibiting cytoplasmic expression should be subject to re-assay of expression without a NLS. Given the result for ZK643.3::*lacZ*, it may also seem advantageous to exclude those fusions containing β-Galactosidase in a predicted extracellular protein domain. If this was accepted, however, the subcellularly localised expression for B0303.12 would not have been discovered (Table 3.4).

Chapter 5.

Further analyses of ZK637.8.

Introduction.

One of the basic reasons for performing expression pattern screens is to identify genes of potential relevance in basic biological processes (see Chapter 1). Further analysis of these may then shed light on the mechanisms in which they are involved. For instance, genes expressed in early embryos may be assumed to be involved in the developmental processes of early embryogenesis (Hope, 1991). Such genes can be classed into two groups: regulatory genes which control the genetic program of development, and effector genes which perform the material development of the embryo. Both of these gene types are components of the same genetic pathways, and analysis of both types is crucial for complete pathway elucidation.

ZK637.8 expression

Expression analysis described in Chapter 3 has revealed that ZK637.8 is expressed in early embryos in the gut primordium as the cells of the gut lineage are gathering together prior to full morphogenesis of the intestine. Other expression components in the developing structures of the spermathecae and vulva of young adult worms strengthen the implication that WP:ZK637.8 has a role in early developmental events in some tissues. Very high and extensive sequence homology to the ATPase subunit of vesicular membrane-bound proton pumps (V-ATPases) suggests that ZK637.8 acts at the end of genetic pathways, and is an effector gene in development (see Chapter 3 for discussion of the known biological functions of V-ATPases). In addition to developmental patterns of expression, pattern components are also seen of a constitutive nature in the pharynx, neurones of the head ganglia and ventral nerve cord, and the anal sphincter. The ZK637.8 gene is thus expressed both in a transient, regulatory fashion and a constitutive fashion indicative of terminal differentiation products (Hope, 1991). We would therefore expect ZK637.8 gene expression to be regulated both by regulatory genes in developing tissues, and genes controlling constitutive expression in terminally differentiated cells. The cellular diversity of the expression pattern also presents the possibility of differential regulation in different cell types. The ZK637.8 gene is thus likely to be a regulatory target of several genetic pathways, and was thus chosen as a suitable subject for further molecular analyses. First, confirmation of the predicted exonic structure was sought through sequencing of cDNA clones of the gene. Second, unidirectional deletions of the upstream region of the reporter fusion were performed to identify genomic regions containing promoter elements specific for different expression components. Third, the expression of ZK637.8::*lacZ* in male *C.elegans* was assayed

Extant data for ZK637.8

Before my analyses were begun, a certain amount of information on the structure of ZK637.8 was available (summarised in Figure 5.1). As well as the predicted intron/exon

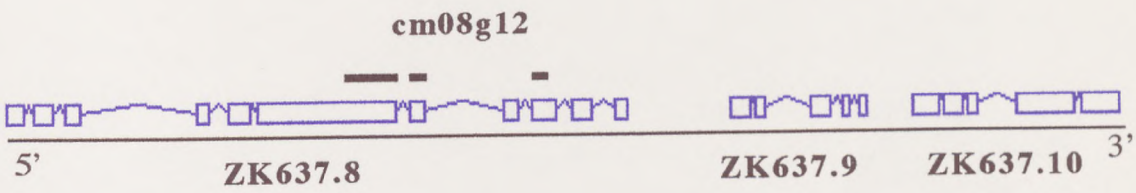


Figure 5.1. Known structure of ZK637.8.

The EST of cDNA clone cm08g12 (thick black line above gene ZK637.8) contains contiguous sequence covering exons 6, 7 and 9 of ZK637.8; exon 8 is not present in this clone. The three genes ZK637.8, ZK637.9 and ZK637.10 are members of a polycistronic transcription unit (see body of text).

structure data in ACeDB, there was also an identified EST for the gene. The cDNA clone cm08g12 had been partially sequenced at its 5' end (Waterston *et al.*, 1992), and had confirmed the splices between predicted exons 6 and 7 but indicated that predicted exon 8 was not present in the transcript, exon 7 being spliced to predicted exon 9 (Figure 5.1). In contrast to the rest of ZK637.8 predicted exon 8 has no significant homology to the family of V-ATPase proton pumps, and its loss does not introduce frameshifts in downstream coding sequence, so it seems unlikely to be a true exon.

It had also been reported that ZK637.8 was the most 5' member of a polycistronic transcription unit, the downstream genes being ZK637.9 and ZK637.10 (Zorio *et al.*, 1994; Figure 5.1). Interestingly, ZK637.10 had been identified as the possible genetic locus of the *unc-32* gene (Brenner, 1974) on the basis phenotypic rescue by transgenic genomic fragments (Pujol and Thierry-Mieg, pers.comm.). The *unc-32* phenotype correlates well with the expression pattern observed for ZK637.8 (and so, by definition, with the likely expression pattern of the co-transcribed ZK637.10 (Zorio *et al.*, 1994)) with the uncoordinated phenotype linking to the motorneurone expression of the ventral nerve cord, egg-laying and fertility defects perhaps being linked to the vulval and spermathecal components of expression, and the embryonic lethality of some *unc-32* alleles perhaps linked to embryonic expression observed in the early E lineage (Thierry-Mieg, pers. comm.).

Results: part one.

Transcript analysis of ZK637.8.

The structure of the ZK637.8 gene was found by sequencing of cDNA clones. As the genomic sequence of the gene was already publicly available, cDNA clones were sequenced on one strand only in order to confirm the intron-exon boundaries and the raw sequence data from the genome sequencing labs.

Sequencing of cDNA 3' end.

The EST clone, cm08g12, covers the 3' end of the ZK637.8 gene and was obtained from the Genome Sequencing Project lab at the Sanger Centre, Cambridge. The vector backbone of cm08g12 is λSHLX2, and so the cDNA fragment insert was obtained in an amp^R plasmid by the pop-out procedure. The cDNA insert was then restriction mapped and suitable fragments identified for subcloning (Figure 5.2) into the M13 phagemids mp18/19 (Yanisch-Perron, 1985). Sequence of the entire cDNA fragment confirmed the predicted splicing pattern for the 3' end of ZK637.8 upto the end of predicted exon 6 (Figure 5.2) and identified a polyadenylation site used *in vivo* as the AGTAAA sequence at cosmid coordinates 27675, 427bp downstream of the predicted stop codon. This is a rare variation of the consensus polyadenylation signal AATAAA (Krause, 1995).

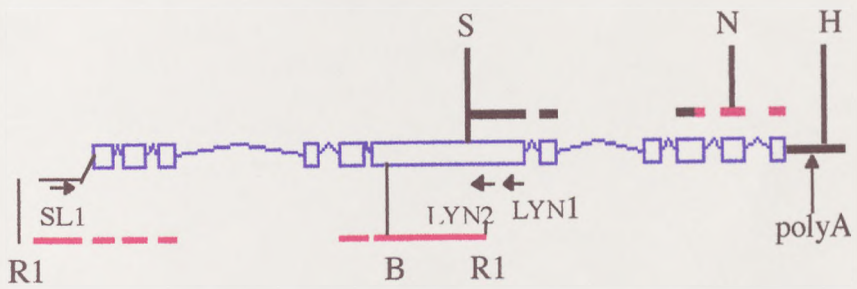


Figure 5.2. cDNA sequencing of ZK637.8.

The predicted gene structure from ACeDB is shown in blue. 3' fragments from the cDNA clone cm08g12 were subcloned into M13 by the sites shown above the predicted gene; S (SacI in the MCS of the cDNA clone), N (NsiI) and H (HindIII in the MCS of the cDNA vector). A polyadenylation site (polyA) was found just downstream of the last exon. EST sequence already available in ACeDB is shown in black, novel sequence is shown in red. 5' cDNA was amplified from a cDNA preparation using the primers LYN1, LYN2 and SL1. The resulting fragment was subcloned into M13 using the restriction sites shown; R1 (vector encoded EcoRI) and B (BglII).

primer	sequence	position
LYN1	TACCTGTCCAGGATACCAAGTG	3483
LYN2	TTCGAACCAGGATACTTGTAGCC	3357
SLI	CGCAGTCGACGGTTTAATTACCCAAGTTTG	-37
PC50	GTCGACAACAGCGTGATGGTCTCGAC	-1364
PC51	GTCGACGAAATGTGAGTCGCCCAATG	-514
PC833	ACTAGTTTCAGCTATGTGAGCTCTG	546
PC843	ACTAGTGCTGAATCGTCGTCCATATC	996

Table 5.1. PCR primers used in experiments with the ZK637.8 gene.
Columns:-
primer: name of primer
sequence: 5'-3' sequence of promer. SalI sites and SpeI sites engineered into the primers to facilitate later subcloning into an expression vector plasmid.
position: base position of the most 5' nucleotide realtive to the first base of ZK637.8 coding sequence(+1).

Sequencing of cDNA 5' end.

The 5' end of ZK637.8 cDNA was obtained by an anchored PCR procedure. Many mature *C.elegans* transcripts are transpliced to a 22mer transpliced leader sequence called SL1 (Spieth *et al.*, 1993). Whilst the biological significance of transplicing is not fully understood, the transpliced sequence itself provides a very useful defined 5' end for many *C.elegans* mRNAs. Using a SL1-specific primer and a nested set of gene-specific primers, cDNA 5' ends can often be easily obtained by PCR of a *C.elegans* cDNA preparation (e.g. Hope, 1994). The success of this procedure depends, of course, on SL1 transplicing of at least some of the transcripts of the gene in question. ZK637.8, as described above, is the most upstream member of a polycistronic unit and as such is transpliced to SL1 (Zorio *et al.*, 1994). This gene is thus a good subject for use of this technique for isolation of 5' cDNA.

The primers used for ZK637.8 are shown in Table 5.1. The 5' end of the gene was clone as a PCR fragment from a whole cDNA preparation donated by Ian Hope (Hope, 1994). The first round of 30 cycles was with the SL1 specific primer and gene specific primer LYN1, designed to be within the 5' endpoint of the cDNA insert of cm08g12, and produced a smear of DNA products with no discernable individual bands. A second round with the SL1 primer and a nested gene specific primer LYN2 produced a single DNA product. The PCR fragment obtained was ligated into vector plasmid pCRII (TA cloning kit), and the insert restriction mapped (Figure 5.2). Suitable restriction fragments were subcloned into the M13 phagemids mp18/19 for dideoxy sequencing. The DNA sequence obtained confirmed the predicted SL1 transplice site and start codon (Figure 5.2). A discrepancy was discovered, however, between the predicted and observed exon splicing pattern downstream of predicted exon 3. The PCR-generated clone was missing predicted exon 4, exon 3 being spliced directly to exon 5. This splicing pattern induces a frame shift in downstream coding sequence leading to loss of the open reading frame, and a translationally non-functional gene product. All other splices in the 5' end of ZK637.8 were as predicted.

Discussion: part one.

Gene structure of ZK637.8.

With one exception all the predicted exons were confirmed by cDNA sequencing. The SL1 transplice site and probable start and stop codons identified also concurred with gene structure predictions in ACeDB. There were no sequence discrepancies over the 2.5kb of confirmed coding sequence. The one splicing discrepancy is therefore likely to be a result generated by error in this analysis.

The only difference between predicted and observed splicing patterns was that immediately downstream of exon 3. Multiple ESTs generated in the Kohara lab, Japan,

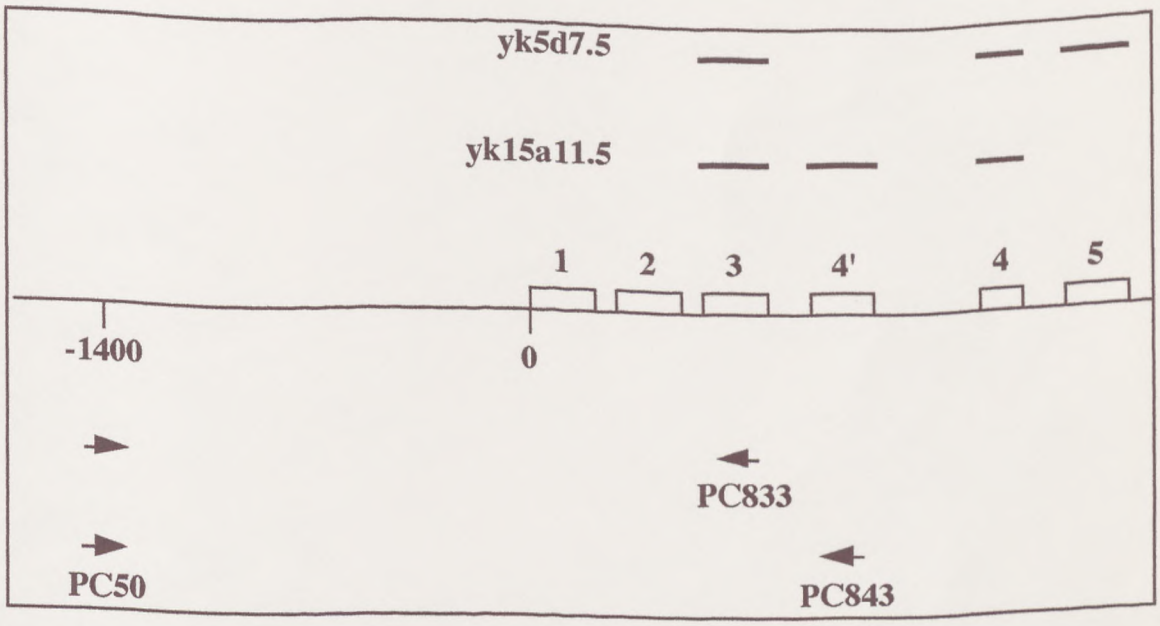


Figure 5.3. The 5' end of ZK637.8.

Two cDNA clones, yk5d7.5 and yk15a11.5, differ in possession of exon 4' suggesting that alternative splicing occurs in the 5' end of ZK637.8 transcripts. Genomic fragments with alternative 3' endpoints were generated by PCR so as to generate translational fusions specific for each of the alternative transcripts represented by yk5d7.5 and yk15a11.5. The two PCR fragments shared a common 5' end 1400bp of the start codon of the gene.

and sequenced subsequent to my analysis (sequences in ACeDB), correspond to the 5' region of ZK637.8. The clones reveal two types of transcript which differ in 3' splicing of exon 3 (Figure 5.3): one species contains exon 3 spliced to the originally predicted exon 4 and is represented by the EST yk5d7.5; the other possesses extra coding sequence between predicted exon 3 and 4, which I have designated exon 4', and is represented by the EST yk15a11.5. The cDNA representation of one of each of these transcript species suggests approximately equal expression levels for the two transcipts and indicates that neither is likely to be a rarely expressed, but biologically tolerated, aberrant transcript. Splicing mechanisms may therefore exist *in vivo* which control alternative splicing reactions for the 3' end of exon 3. The variation in splicing at this point may allow a small amount of aberrant splicing to exon 5 to occur. cDNA library production, involving only 1 cycle of RT-PCR followed by one cycle of reverse strand PCR (Kohara, pers. comm.), would contain the aberrantly spliced transcript at a very low level in the cDNA population. This may explain the apparent lack of ESTs containing direct exon 3/exon 5 splices in the Kohara lab data. The method I used to generate 5' cDNA sequence for ZK637.8 involved 60 cycles of PCR in total. Smaller DNA fragments are more efficiently amplified in a PCR reaction, so it can be easily seen how the shorter aberrant transcript could become hugely enriched in the cDNA population through so many rounds of PCR amplification. Indeed, no evidence of larger DNA fragments was seen in the PCR reactions I conducted.

Expression patterns of the alternative transcripts of ZK637.8.

The fusion point to *lacZ* in pUL#AL10 is in exon 3. The patterns of expression for both of the viable mRNAs from the ZK637.8 locus would thus contribute to the observed pattern as they both include exon 3 (Figure 5.3). Differing patterns of expression for each of the transcripts are conceivable. Exon 4' and exon 4 are remarkable in being the only regions of coding sequence in the whole gene sequence to have no homology to the family of proton pump ATPases. The transcripts represented by yk5d7.5 and yk15a11.5 may thus have evolved to supply separate but related protein functions in *C.elegans*, functions that could be required in different cells and at different times during development. Reporter fusions were used to test for potential differential contributions to the pUL#AL10 expression pattern.

Construction of reporter fusions for the alternative transcripts of ZK637.8.

Only transcript d4' contains a unique exon sequence, so to perform a differential expression assay of the three possible transcripts combinatorial methods were employed. Translational fusions were selected as shown in Figure 5.3: an exon 3 fusion enables both transcripts to contribute to the overall pattern produced (as for pUL#AL10); an exon 4' fusion is specific for the transcript represented by yk15a11.5 enabling any specific

expression components due to that transcript to be deduced from comparison with the pattern produced by the exon 3 fusion.

No convenient restriction sites were present in the relevant exons, so a PCR method was used to generate fragments with the desired 3' exonic end points. To ensure homogeneity of expression control for each DNA fragment the same 5' end point was used, the site chosen for this being at position -1363 to include all control elements necessary for the components of expression observed for ZK637.8 (a position experimentally determined below - Figure 5.4). Primers used for PCR are listed in Table 5.1. Each PCR fragment was generated using cosmid DNA as template, and was ligated into pCRII vector (TA cloning kit) and subsequently subcloned into pPD21.28 (Fire *et al.*, 1990). The same precautions concerning reading frame and site of insertion into the expression vector MCS were taken as in the PCR-based expression screen (described in Chapter 4) to ensure identical molecular contexts for each genomic fragment.

The transcript specific expression patterns are identical.

When strains of *C.elegans* transgenic for each of the described reporter fusions were assayed for β-Galactosidase activity, all showed identical expression patterns to that observed for pUL#AL10. There is thus no evidence to suggest that the alternative transcripts of ZK637.8 have distinct cellular modes of expression. If the alternative transcripts do encode alternative functions in *C.elegans*, these functions are likely to be required in at least some of the same developmental contexts.

Results: part two. Promoter analysis of ZK637.8.

The expression pattern observed for ZK637.8 has several distinct spatial and temporal components, and it would be expected that the promoter region controlling such expression would reflect this complexity. Two main approaches have been used in *C.elegans* previously to identify individual promoter elements: assay of the effect on expression of serial deletion of the upstream regions of reporter fusions (e.g. Aamodt *et al.*, 1991; Okkema *et al.*, 1993); and sequence comparison of gene 5' ends between *C.elegans* genes and those of other, closely related nematode species such as *C.briggsae* and *C.vulgaris* to find regions of conserved sequence potentially identifying conserved control elements required for cell specific regulation of gene expression (e.g. Heschl and Baillie, 1990; Kennedy *et al.*, 1993). The latter approach would involve the cloning and sequencing of genomic DNA corresponding to ZK637.8 from another organism, a procedure requiring no little investment of time and effort.

The plasmid pUL#AL10 was readily available, however, and provided suitable starting material for a promoter deletion approach to probe the structure of the ZK637.8 promoter region, a task performed in two stages. Facile restriction enzyme deletions enabled removal of sequences upstream of position -1646 relative to the translational start of the gene. The lack of suitable restriction enzyme sites in positions more proximal to the start codon necessitated the use of PCR to further define the promoter structure of this gene.

Restriction enzyme deletions.

A set of five restriction enzyme deletions were performed as illustrated in Figure 5.4. DNA constructs Δ1, Δ2, Δ3 and Δ4 represent progressive unidirectional deletions of the upstream promoter region of ZK637.8 contained within reporter fusion plasmid pUL#AL10, Δ5 removing the probable transcriptional start of the gene as well as all of the ZK637.8 coding sequence. Deletions were mostly attained by restriction enzyme digestion of pUL#AL10 followed by intramolecular ligation of the genomically encoded restriction enzyme site to a complementary site in the vector MCS: deletion Δ1 - HindIII site at cosmid coordinate 18968 to MCS HindIII site; deletion Δ2 - BglII site at cosmid coordinate 19823 to BglII site at cosmid coordinate 17745; deletion Δ3 - NsiI site at cosmid coordinate 20013 to MCS PstI site; deletion Δ4 - SphI site at cosmid coordinate 21464 to MCS SphI site; deletion Δ5 - XbaI site at cosmid coordinate 21413 to MCS XbaI site.

Deletions Δ1, Δ2 and Δ3 had no effect on the expression pattern of ZK637.8::*lacZ* in the transgenic strains examined. Δ4 resulted in complete loss of the specific components of the pUL#AL10 expression pattern, and gained non-specific components in the posterior intestine and pharynx. Such non-specific expression is often seen in *C.elegans* when reporter constructs containing incomplete promoters are used, as described in Chapters 3

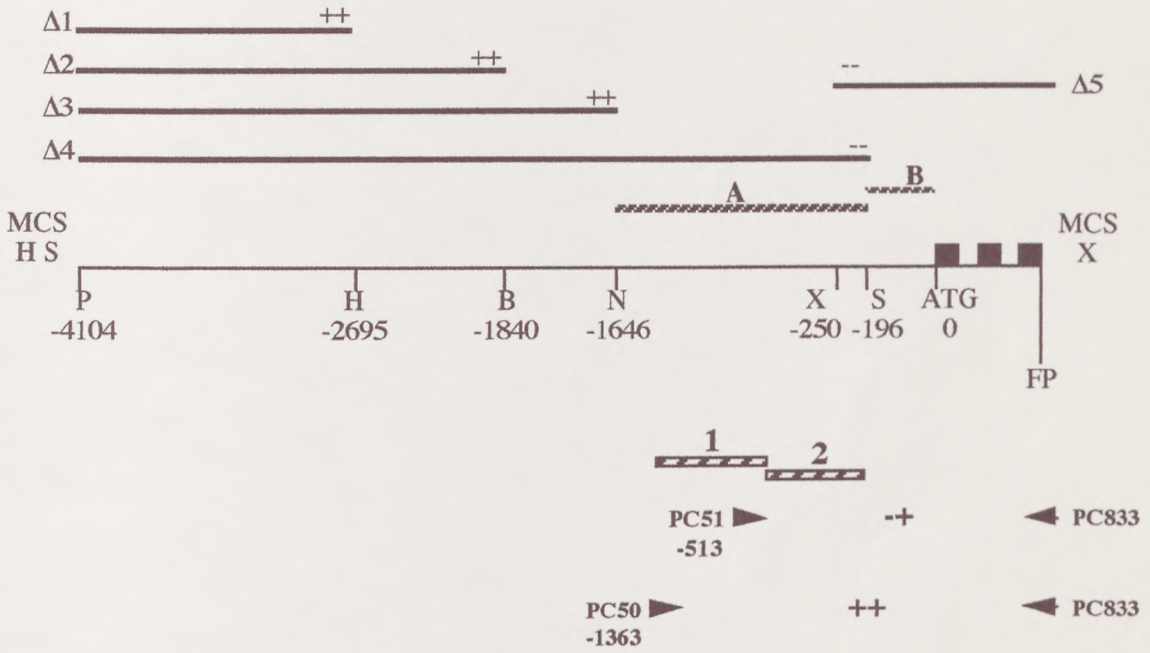


Figure 5.4. Promoter deletions of ZK637.8.

The genomic region containing the 5' end of the ZK637.8 gene in pUL#AL10 is shown. P (PstI), H (HindIII), B (BglII), N (NsiI), X (XbaI), S (SphI), MCS (multiple cloning site) and FP (fusion point to *lacZ*). Restriction enzyme deletions are shown above the sequence, and are scored for ability to drive expression (++:yes, --:no). They define two promoter regions: Region A contains elements necessary for all components of the ZK637.8 expression pattern; Region B contains the transcriptional start of the gene (see text). PCR-generated deletion are shown below the sequence and are scored for ability to drive expression (++:yes, -+:cannot drive expression in the embryo and spermathecae). The define a further two promoter regions: Region 1 contains elements necessary for the embryonic and spermathecal components of the ZK637.8 expression pattern; Region B contains elements necessary for the other components of the ZK637.8 expression pattern.

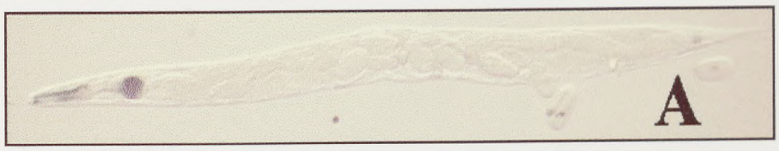


Figure 5.1. *LacZ* fusion expression with the PC51/PC833 PCR-generated genomic 5' end.

A: This adult hermaphrodite has the usual components of ZK637.8 expression in the pharynx and anal sphincter, but is completely missing any embryonic expression.

B:An L4 larva exhibits expression in the vulva, but this is not accompanied by the spermathecal expression seen with non-deleted *lacZ* fusions (e.g. Plate 3.19G).

and 4. Such non-specific expression is blocked when strong genomic promoters are present in a genomic insert, but can occur when a weak or incomplete promoter is present along with a viable transcriptional start. Thus Δ4 represents a genomic deletion removing all of the strong promoter elements of ZK637.8 but leaving in place the transcriptional start of the gene. Δ5 removes the translational fusion to *lacZ* and completely abolishes expression as expected. The lack of non-specific expression for this deletion also suggests that the transcriptional start has been removed.

This analysis indicates a promoter structure for ZK637.8 as shown in Figure 5.4. All of the promoter elements necessary for the gene's specific expression are present in the 1500bp A region between the Δ3 and Δ4 endpoints, and the transcriptional start lies in the 196bp B region between the 3' endpoint of Δ4 and the start codon.

PCR deletions.

PCR primers (Table 5.1) were made for two locations within the 1500bp A region at the positions shown in Figure 5.4. DNA fragments amplified from cosmid genomic DNA clones were subcloned directly into pCRII (TA cloning kit) and subsequently into a reporter plasmid, and the resultant reporter fusions assayed for ability to direct expression *in vivo*. The PCR fragment generated with primers PC50 and PC833 (5' endpoint at cosmid coordinate 20300) directed expression identical to that for pUL#AL10, but the PC51/PC833 fragment (5' endpoint at cosmid coordinate 21150)failed to express the embryonic E-lineage (Plate 5.1A) and spermathecal components (Plate 5.1B). The promoter structure suggested by these results (depicted in Figure 5.4) has regulatory elements necessary for E-lineage and spermathecal expression in the 850bp Region 1 which are discrete from the elements responsible for the anal, vulval, pharyngeal and neural components in the 317bp Region 2.

Discussion: part two.

ZK637.8 promoter structure is consistent with general *C.elegans* promoter structure.

The overall structure of the ZK637.8 promoter is as expected with respect to other characterised *C.elegans* genes. Promoter elements controlling expression in terminally differentiated cells are generally found close to (i.e. within 200-500bp) the translational starts of genes. The components of the ZK637.8 expression pattern of this type are those found in the anal sphincter, the pharynx and neural cells, and I have mapped the elements responsible for these components to Region 2, no more than 513bp from the gene's probable translational start (Figure 5.4). More dynamic types of expression (i.e. transient or recurrent in nature) typical of those genes involved in developmental processes tend to be controlled by elements further upstream of the actual coding sequence. Two of the components of ZK637.8 expression fitting this description, during the first few generations of the embryonic E lineage and in the spermathecae as those structures are

undergoing morphogenesis, are controlled by promoter elements in Region 1, stretching from 1363bp to 513bp upstream of the start ATG. Given this level of broad agreement between the observed and expected promoter structure of ZK637.8 it seems reasonable to predict that the control element(s) specific for vulval expression (another dynamic expression component) will lie in the most 5' section of Region 2, upstream of those responsible for the terminally differentiated modes of expression, and could be tested for by further unidirectional deletion of Region 2.

The mapping of the whole of the promoter region would benefit from such treatment, sequences necessary for each aspect of the observed expression pattern being defined with greater accuracy. Once such sequences are fully delineated, additional analyses can be performed to further define their nature. Defined elements can be identified as being sufficient for expression by their ability to drive correct cell type expression of "naive" promoters, i.e reporter fusions not able to direct expression as of themselves (Aamodt *et al.*, 1991; Krause *et al.*, 1994). Enhancer assays describe the exact regulatory nature of a transcriptional control element through assay of activity of a promoter fragment in both orientations relative to a reporter fusion (Okkema *et al.*, 1993); enhancer sequences are able to drive expression in both orientations, promoter elements can only drive expression when present in a single orientation.

Such defined control elements are suitable tools to begin elucidation of genetic pathways operating in the development and functioning of *C.elegans*. For example, labelled oligonucleotide probes, consisting of DNA sequence found to delineate promoter elements sufficient for cell specific gene expression, have been used to screen a cDNA expression library and positive colonies isolated and identified (Okkema and Fire, 1994). Genes discovered in this way are candidate transcription factors operating to control specific gene expression. Elucidation of many such relationships will enable the molecular basis of genetic control of biological processes to be described.

Results: part three.

ZK637.8 expression in male *C.elegans*.

All the work described thus far concerns gene expression in hermaphrodite *C.elegans*. The species also includes male individuals, present at very low frequency in wildtype populations, which are capable of mating with hermaphrodites enabling sexual reproduction to occur (Wood, 1988). There are large scale anatomical differences between the two sexual forms, principally in the posterior where the male has many extra neural and hypodermal cells contributing to the structure of its copulatory apparatus (Sulston *et al.*, 1980). To investigate the possibility of ZK637.8 expression in male specific cells and tissues, pUL#AL10 was introduced into male worms and these subsequently assayed for reporter fusion expression.

Males transgenic for pUL#AL10 were generated by microinjection of a mutant *C.elegans* strain, *him-5* (high incidence of males). Members of this strain contain a mutation which causes male progeny to be produced at an abnormally high frequency (30% of all progeny as opposed to ~0.2% in wildtype; Hodgkin *et al.*, 1979), enabling routine observation of male specific expression patterns in small populations.

Reporter fusion expression in *him-5* (ZK637.8::*lacZ*).

When assayed for β-Galactosidase activity, males carrying the pUL#AL10 plasmid exhibited expression in the pharynx, the hypodermal seam cells, the ventral nerve cord and the copulatory bursa in the tail (Plate 5.2A). The pharyngeal and ventral nerve cord components are the only pattern components shared with hermaphrodite animals (Chapter 3). The pharyngeal expression is apparently identical to that seen in hermaphrodites, but there seem to be more expressing motoneurones in the ventral nerve cord of male animals (compare Plate 5.2A and Plate 3.19G). The extra staining cells may thus represent the CPn and or CAn male-specific motoneurones found in the ventral nerve cords of male animals (White, 1988), cells which control the reflex arching and turning of the male, mediated by a set of male-specific muscles in the posterior half of the animal, as it attempts to copulate with the hermaphrodite. The seam and tail components are completely novel male specific expression components of the ZK637.8 gene.

Closer examination of the tail expression revealed both neural and hypodermal components. The neural expression is contained within four specific regions (Sulston *et al.*, 1980). Dorsally (Plate 5.2B), expression is observed in the posterior portion of both of the lumbar ganglia and also in the phasmid sensilla. Expression in more ventral locations (Plate 5.2C) is evident in the preanal ganglion and posterior to the cloaca where the postcloacal sensilla are found, the cloaca being an alternative opening of the genital

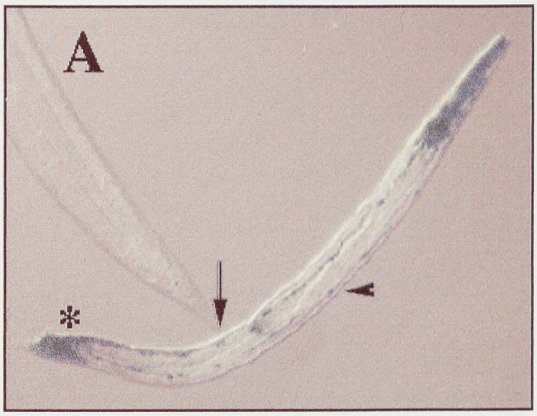


Figure 5.2. *LacZ* fusion expression of ZK637.8 in male *C.elegans*.

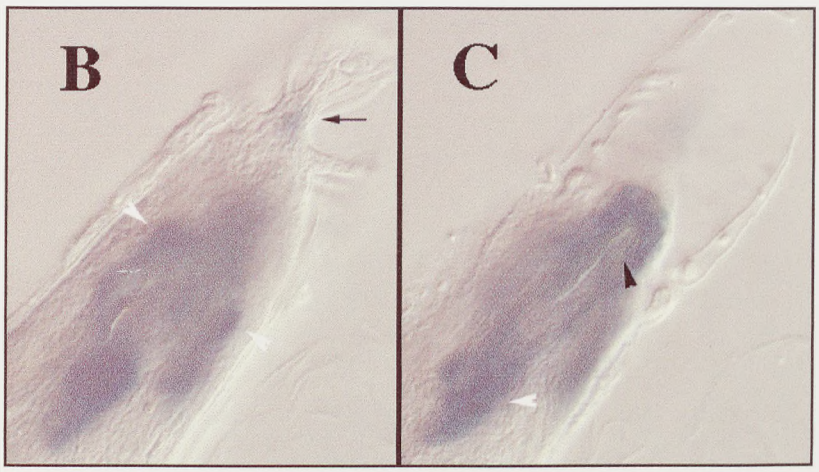
A: A young adult male showing strong expression in the pharynx, the ventral nerve cord (arrow), the hypodermal seam cells (arrowhead) and the specialised tail structures (asterisk).

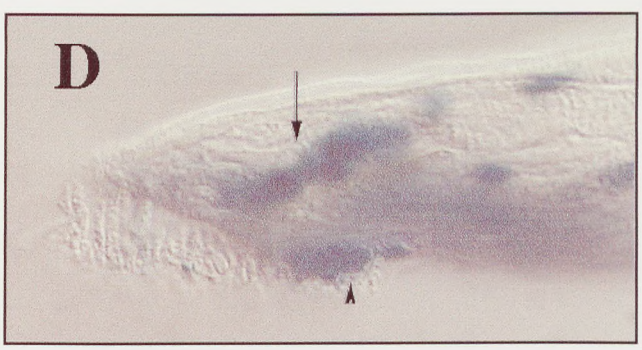
overleaf:

B: A dorsal view of staining in the tail of an adult male animal. Expression is evident in the phasmids at the tip of the tail (arrow), and in the posterior region of the bilaterally symmetrical lumbar ganglia (arrowheads).

C: The same tail in the ventral focal plane. Strong expression in the preanal ganglion (white arrowhead) is seen to lead posteriorly to run around the cloacal opening (black arrowhead).

D: A lateral view of the tail of an L4 male. The fan is clearly not fully developed. Expresssion is seen along the lateral line in a cluster of set cells (arrow), and ventrally in neural tissue (arrowhead).


B
C


D

and alimentary tracts in the male at the position of the anus in hermaphrodites (Sulston *et al.*, 1980). All of the tail neural expression is present from mid-late larval stages throughout adulthood.

Hypodermal expression is restricted to the tail seam, or set, cells in late larval stages as the sex-specific structures of the male tail are forming (Plate 5.2D). These cells are morphologically similar to the hypodermal seam cells of the body and head, which also express ZK637.8::*lacZ* in these strains (Plate 5.2A), and are thought to have a role in morphogenesis of the male tail, specifically in secretion of the cuticular material which forms the tail fan (Sulston *et al.*, 1980). Expression in these cells does not last into adulthood as can be seen by comparison of the expression visible in L4 larvae (Plate 5.2D) and in fully differentiated adult worms (Plates 5.2B and C).

Discussion: part three.

Male expression is consistent with that observed in hermaphrodites.

The overall expression pattern observed in male *C.elegans* fits well with the known hermaphrodite pattern components of ZK637.8 gene expression. First, the pharyngeal and ventral nerve cord components are held in common, though male specific neurones in the nerve cord also seem to express ZK637.8::*lacZ*. The other hermaphrodite pattern components are absent or cannot be confirmed, but each of these observations can be explained by known anatomical differences between the nematode sexes. Obviously, expression in the vulva and spermathecal valves cannot be expected in males as these are hermaphrodite specific cells of the somatic gonad. Similarly, the hermaphrodite anal sphincter is not present in the alimentary tissues of male animals, its role being performed by the proctodeum at the junction of the genital and alimentary tracts (White, 1988). Finally, male specific expression in the E lineage during early embryogenesis cannot be confirmed. It is not possible to anatomically differentiate between male and hermaphrodite individuals at this early stage (i.e before the 150 cell stage), the first detectable cellular difference being observed at the 550 cell stage of embryogenesis at around 470 minutes after fertilisation when cell proliferation has ceased and organismal morphogenesis is underway (Sulston *et al.*, 1983; Sulston, 1988). Embryos which are genetically male probably do have E lineage expression, however, as most embryos in the transgenic *him-5* strains assayed did show embryonic expression. This is only a subjective measurement, but as around 30% of individuals in these strains are male, it seems likely that male embryos are included in those exhibiting E lineage expression.

Novel neural expression.

The overall structure of the nervous system in the male tail has not been reconstructed to the same extent as the entire nervous system of the hermaphrodite, but the nuclear locations of all of its member cells have been mapped and specific regions have been

identified with specific biological functions (Sulston *et al.*, 1980; White, 1988). Such linkages as have been made for regions included in this expression pattern are interesting with respect to the described hermaphrodite neural expression pattern (Chapter 3).

Hermaphrodite expression is nuclear localised in motoneurones of the ventral nerve cord and cytoplasmically localised in neurones of the head ganglia (Plates 3.19F and 3.19G). Male expression, in addition to that seen in hermaphrodites, is nuclear localised in the CAn and CPn motoneurones of the ventral nerve cord and cytoplasmically localised in cells of the preanal and posterior lumbar ganglia. It is difficult to draw conclusions on the relatedness of the two components of neural expression seen in hermaphrodites due to the complexity of nervous tissue in the head ganglia and difficulty in assigning definite cell identities to the expressing nuclei in the ventral nerve cord. The CAn and Cpn motoneurones are easier to identify, however, and it is known that most cells in the posterior of the lumbar ganglia are those of the sensory rays of the tail fan. The rays are known, by laser ablation, to be required for the reflexive arching of the male as it seeks to copulate with the hermaphrodite (White, 1988) and must therefore be part of the same neural circuits as the CAn and Cpn motoneurones. The interneurones connecting the sensory neurones and motoneurones in these circuits are thought to lie in the preanal ganglion where both these sets of cells send processes (Sulston *et al.*, 1980). These expression components may therefore identify complete neural pathways in the male tail from sensory input to behavioural output. The postcloacal expression does not have an obvious place in this scheme as the postcloacal sensilla have no known function, but they are known to project processes into the preanal ganglion and may therefore play an unknown role in male specific locomotion.

The similarity between the hermaphrodite-specific and male-specific neural expression patterns suggests there may be functional homology between the two outlined pathways. Perhaps the head ganglia expression seen in hermaphrodites and males represents the functional correlates of the sensory ray and preanal ganglion cells which express in the male tail, i.e. neural cells conveying information to the muscle cells through motoneurones in the ventral nerve cord.

Novel hypodermal expression.

Set cell expression is also a male-specific pattern component which may have a functional correlation with parts of the hermaphrodite expression pattern. In common with the E lineage, vulval and spermathecal expression in hermaphrodites, set cell expression is transient in a tissue (the lateral hypoderm) involved in morphogenesis. It is also interesting to note that three of these components involve groups of cells in motion; the cells of the early E lineage are the leader cells in the initial movement of gastrulation, the spermathecal cells are part of the developing somatic gonad which grows distally from a

central position in the body, and the set cells withdraw anteriorly as tail morphogenesis proceeds in the L4 male. Although V-ATPase proton pumps are not known to be involved in such processes in other systems (see review of known V-ATPase biological function in Forgac, 1989), these pattern components suggest that ZK637.8 may function in the movement of clusters of developing cells in *C.elegans*.

Implications for promoter structure of ZK637.8

The promoter structure of ZK637.8 deduced earlier in this chapter supports this view. The two hermaphrodite pattern components regulated by promoter Region I (Figure 5.4) are those in the embryonic E lineage and the spermathecae, i.e those developmental components consisting of moving groups of cells. The other developmentally regulated expression component in the vulval D-cells - cells which are not in motion - is not affected by deletion of Region I, indicating its control to be determined by sequences closer to the translational start of the gene in Region II. There is thus a clear separation between the regulatory elements of developmental ZK637.8 expression on the basis of movement of their cellular targets.

Assay of the expression of the promoter deletions described above in male animals would further strengthen this implied link if male set cell expression was also found to rely on Region I sequence. Promoter elements responsible for other male specific expression components could also be mapped in this way. We might expect the neural expression in the posterior of male animals to require Region II sequence, in common with the result found for the seemingly related hermaphrodite neural expression. It is more difficult to predict a promoter region necessary for male expression in the lateral hypodermal seam cells of the body as this component has no obvious correlate in the hermaphrodite. It is observed, however, at the same period as set cell expression and the two cell types are closely related structurally (Sulston *et al.*, 1980). This expression component may therefore be a feature of gene expression in all lateral hypodermal cells of the male, and may be expected to depend upon promoter elements also necessary for set cell expression. I would therefore predict that Region I would contain sequence necessary for seam cell expression of ZK637.8.

As far as I can discover, no work has been done in *C.elegans* on the exact nature of promoter regions with respect to regulation of sex-specific gene expression. The above predictions are based upon the expectation that gene expression in similar cell types - e.g. the differential neural expression pattern components seen in male and hermaphrodite animals with the ZK637.8 gene - would be more efficiently structured in the genome by use of the same regulatory elements in those cell types. It is possible, however, that sex-specific gene expression may be controlled through separate genetic pathways for some genes. As well as identification of promoter structure in the constrained case of

hermaphrodite expression, finer delineation of transcriptional control elements, as described in the previous section, would reveal such structure. It would thus be enlightening to assay the activity of such promoter deletions in both hermaphrodite and male *C.elegans*.

Chapter 6.

Towards an Expression Pattern Database.

Note: some of the work described in this chapter has been published prior to the writing of this thesis (Hope et al., 1996).

Introduction.

As described in Chapter 1, reporter fusion expression patterns have shown their worth in three ways. First, as markers of cell and tissue differentiation, they are invaluable in experiments where the differentiation state of a specified cell or tissue must be known. Second, when ectopic expression of a gene of interest is desired, they can supply promoters capable of driving gene expression in specific cell types. Third, by virtue of the nature of expression reported, they identify candidate genes for involvement in specific biological processes and coincidentally can themselves be used to study their molecular basis.

Expression patterns of known genes have added value.

The directed expression pattern screen I have described has generated expression data for many predicted genes in *C.elegans* genomic sequence, and in continuation will generate many more. This data set has value in addition to those outlined above, determined by the complete molecular and anatomical description that has been, or will shortly be, attained for the nematode. The sequencing projects underway for many different model organisms will provide the complete molecular genetic descriptions of many animals and plants. A future aim of biological science must be to understand the development and function of living organisms in terms of their revealed genetic repertoire. Expression patterns of individual genes provide a direct link between the genetic repetoire of an animal, represented by its genomic sequence, and its developmental anatomy (Hope *et al.*, 1996). *C.elegans* is uniquely suited to this endeavour with reference to metazoan development given the relative simplicity of its completely described anatomy compared to more complex model organisms such as mouse and *Drosophila*, and the precise knowledge of the cellular lineages determining its embryonic and postembryonic development (Kenyon, 1988).

Development of public accessibility of expression data.

As more expression pattern data is generated, the importance of ensuring its public accessibility will increase. To enable the full value of expression data to be realised, databases of expression pattern information are being developed for several organisms. The Flybase database, for example, is being developed for the *Drosophila* genome and contains expression pattern information as a key resource for gene description and definition of specific tissues (Gelbart *et al.*, 1996). A similar database dedicated to mouse development is also available, and has gene expression data in the mouse as a central component (Ringwald *et al.*, 1994). I describe two routes toward a *C.elegans* expression

Author Lynch AS
Date 07/95
Pattern Expression is seen exclusively in the dopaminergic neurones of the worm. The cephalic and deirid sensory nerves in the head stain for beta-galactosidase post L1, while the postdeirids can be seen post L3
Reporter genomic fragment XhoI.20010-XmaI.26412 in pPD22.04 (lacZ)
Reference
Gene
Clone pUL#AL29
Sequence T23G5.5
Cell CEPDL/R, CEPVL/R, ADEL/R, PDEL/R
Cell_group cephalic, deirid, postdeirid, dopaminergic neurones
Life_stage post L1, post L3
Strain UL33
Picture ul33.jpeg
Description The six dopaminergic neurones in the head of an adult hemaphrodite-CEPDL/R CEPVL/R ADEL/R
Picture t23g5_5_plat.jpeg
Description The postdeirid neurones PDEL/R in an L3/4 larva

Figure 6.1. Expression data entry into ACeDB.

Information relevant to the context of the expression data is included in a submission to ACeDB, including the gene tagged, the cells which show expression, a searchable text description of the expression pattern and details of the genomic fragment used to generate the *lacZ* fusion.

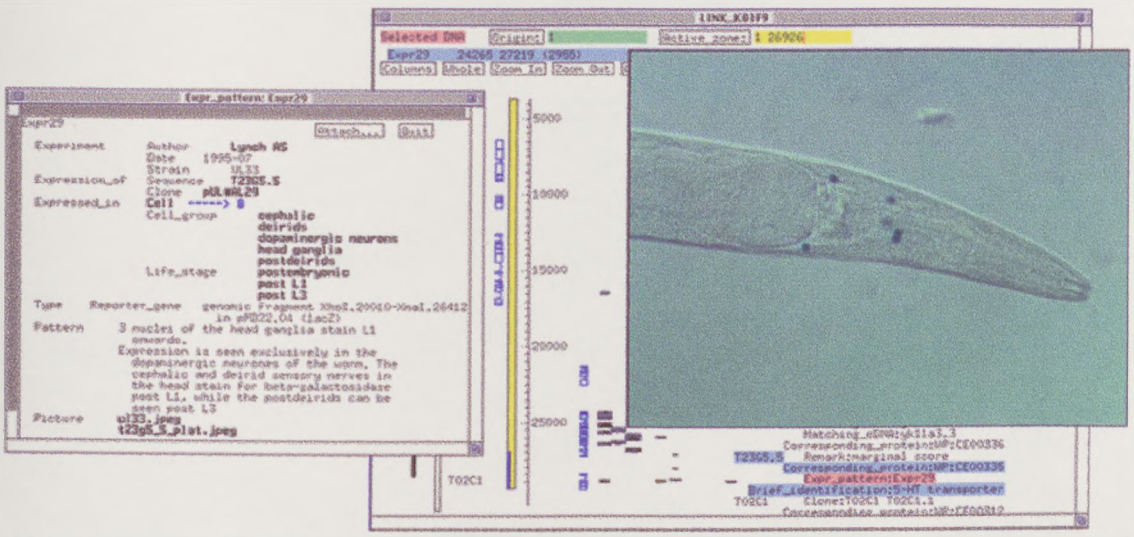


Figure 6.2. Incorporation of expression pattern data in ACeDB.

The figure shows three windows in ACeDB which are linked with respect to expression of the T23G5.5 gene. The rear window is the standard sequence display window, and contains the indication that predicted gene T23G5.5 has expression data associated with it (highlights). Following the link accesses the two front windows: the lefthand one contains textual description of expression and links to related database objects such as cell types and authors; the righthand one contains pictorial information of the pattern of expression.

pattern database which I have developed for the data produced in the directed expression pattern screen.

Results.

Inclusion of expression data in ACeDB.

The ACeDB database is an integrated resource dedicated to the manipulation and presentation of all the molecular and genetic information generated for the *C.elegans* genome, and is the widely used model for many other genome projects (Durbin and Thierry-Mieg, 1991). The built-in database functionality of ACeDB enables cross-referencing of related pieces of information to be easily performed. In concert with several others, I have helped to organise the first inclusion of reporter fusion expression data in ACeDB (Hope *et al.*, 1996). The data defines a new class of information within the database, and is entered into the database in the format shown in Figure 6.1. Automatic linking of expression data with predicted genes is provided by the "Sequence" entry, and with cell and tissue types by the "Cell" and "Cell_group" entries. Any researcher browsing the database for information on a particular gene or cell of interest will thus be directed to relevant expression pattern information. For instance, when perusing information on the sequenced gene T23G5.5, one would find that it has a reporter expression pattern and be able to immediately access details (Figure 6.2). It is important that images of observed expression and details of the strains used are provided to enable independent observation of patterns. Researchers with in depth anatomical knowledge of specific body regions may then be able to correct any wrongly described patterns, or elaborate upon others. Computer mounting of data enables such correction and elaboration to be performed quickly and with immediate effect.

A World Wide Web based archive of expression pattern data.

The World Wide Web (WWW) enables large amounts of computer-held data to be made available to anyone with a computer, and has quickly been adopted by the scientific community for the public presentation of large bodies of biological information. Relevant examples include the Genbank sequence database, against which any DNA or protein sequence can be compared to reveal homologous sequences (http://www2.ncbi.nlm.nih.gov/genbank/query_form.html); the *C.elegans* genome project, which supplies genome sequence data and much related information (http://www.sanger.ac.uk/~sjj/C.elegans_Home.html and http://genome.wustl.edu/gsc/gschmpg.html); and the Caenorhabditis Genetics Centre, which provides details of many available *C.elegans* strains (gopher://elegans.cbs.umn.edu:70/1).

I have developed a WWW site specifically for the expression pattern data I have generated which can be found at http://eatworms.swmed.edu/Worm_labs/Hope. It contains similar data to that found in ACeDB such as text descriptions of the expression pattern of every gene assayed, images of the individual components of each pattern observed, and details of the genomic fragments used to make each reporter fusion. Whilst lacking the relational qualities provided in the ACeDB context the web site does provide access to the data for non-*C.elegans* scientists, who do not have routine access to the *C.elegans* database. As the web site is maintained by those involved in the ongoing effort of generating gene expression patterns, it also tends to be more up to date than the (infrequently updated) ACeDB.

Discussion.

At the present time, in addition to the proven worth of gene specific expression patterns in functional assignation (as described in Chapters 3 and 4), expression pattern information in *C.elegans* is valuable primarily as a source of cell type markers (Young and Hope, 1993; Lynch *et al.*, 1995) and to identify promoter regions important for regulation of specific developmental processes (e.g. Way *et al.*, 1991; Hope, 1994). As has already been realised in the mouse and *Drosophila* research communities, as expression pattern information becomes more and more common it will begin to play a direct role in elucidation of developmental genetic pathways. The amount of information represented by expression pattern data, along with its inherently pictorial nature, discount reliance on publication for public presentation and require the data to be held in computer databases. This has led to the founding of several expression pattern databases (see introduction above), which incorporate the mapping of gene expression patterns onto a representation of the 4D development of an organism. The Flybrain project (Heisenberg and Kaiser, 1995; http://flybrain.uni-freiburg.de), for example, has been developed to contain expression patterns specific to the *Drosophila* brain revealed by enhancer trapping and in situ hybridisation, and mapped onto a series of serial sections and schematic representations of the brain. The Mouse Atlas project (http://glengoyne.hgu.mrc.ac.uk/Genex) offers similar resources for mapping of gene expression patterns onto serial section images of mouse development. When containing a certain threshold of gene information, including gene specific expression profiles, such resources will enable specific questions about gene expression to be asked in the context of anatomical development. One example may be, "which genes are co-expressed in bodywall muscles at this stage of development, and which of these are regulated by the same transcription factors as my favourite gene?". Placement of new genes into known genetic hierarchies, as well as visualisation and appreciation of the same hierarchies, will thus be much facilitated.

C.elegans is uniquely placed to benefit from such an approach as its development is completely described at the cellular level and its anatomy is relatively simple whilst including the major cell types common to all metazoan animals (Kenyon, 1988). There are already available computer technologies for capture of 4D development at cellular resolution (Hird and White, 1993; Fire, 1994). The nascent expression databases I have described in this chapter should therefore be seen as a beginning. Their future development will evolve to link characterised genetic loci with others and specific cells and tissues within a 4D representation of *C.elegans* development. With over a third of *C.elegans* genes conserved between animal phyla (Wilson *et al.*, 1994), understanding of the *C.elegans* developmental program will also make an important contribution to the study of similar mechanisms in other developmental systems.

Chapter 7.

General Discussion

The project which this thesis describes set out with two major aims: to conduct a screen of expression patterns for *C.elegans* genes, and to assess the value of any data generated therein. This chapter discusses how successfully these aims have been met and outlines several potential next steps for genome analysis in *C.elegans*.

Efficiency of the screen.

The efficiency of the pilot screen conducted, that is the percentage of active reporter fusions amongst all those generated, was addressed in Chapters 3 and 4. The initial screen, as the data presented in Chapter 3 shows, generated expression patterns for just over half of the 45 genes assayed. Analysis of the genes themselves found no one reason for the majority of failures, but did identify two major genetic characteristics which could affect the chance of a *lacZ* fusion for a particular gene being active. Accordingly, it was suggested that genes for potential assay by the approach should be screened for membership of polycistronic units and secretory character. Polycistronic genes for which *lacZ* fusions could not be constructed containing genomic sequence upstream of the lead gene of the co-transcribed unit should be excluded from the expression screen; probable secretory genes, i.e. those possessing a predicted signal sequence and no transmembrane domains, should also be excluded as *lacZ* fusions are likely to be secreted from the cell and inactivated.

Such precautions, while likely to decrease the amount of time and effort wasted on fusions doomed to fail, do not address that large fraction of genes apparently excluded from assay. Chapter 4 presented data on one promising route towards the democratisation of the process - use of an intron-rich *lacZ* reporter gene. The results obtained confirmed claims made for this gene, that expression of reporter fusions containing it is much stronger than with the intronless *lacZ* gene used in the initial screen. On this basis, the suggestion that a continuation screen should be performed using expression vectors incorporating the intron-rich *lacZ* gene was made.

A large percentage of genes are likely to remain which will not reveal their expression pattern using a directed assay approach, however. It is possible that the random, promoter trapping approach described in Chapter 1 may now be a more efficient approach in these terms. I discounted it as being less efficient than a directed approach due to the intensive effort required to isolate the active *lacZ* fusions from the vast majority of inactive fusions and to subsequently identify the tagged gene (e.g. Hope, 1991, Hope, 1994). With most of the *C.elegans* genes likely to be sequenced by 1998 (Waterston and Sulston, 1995), however, this last step is becoming increasingly facile. The random approach may also enjoy an advantage in that only active fusions are pursued. Thus, if multiple fusions for a gene exist in the library of randomly generated fusion plasmids, only the active ones will be subject to further analysis. If an active *lacZ* fusion is possible for a particular gene, the

random approach may be the best way of generating it. Even considering these advantages, the random approach still suffers from the need to isolate the active fusions. At present, this is done by microinjection of known combinations of fusion plasmids from a 96 well multiwell plate (Hope, 1991; Young and Hope, 1993), a very laborious and time-consuming procedure. One way to make this step easier is presented by the *C.elegans* genome sequence. Each of the plasmids contained in a multiwell plate consists of *C.elegans* genomic sequenced fused to a *lacZ* gene. The vast majority of fusions are to non-coding sequence, both intra- and intergenic, or are out of frame in coding sequence (Hope, 1991). Active fusions are in-frame with coding sequence. DNA sequencing using *lacZ* specific primers could be conducted for each plasmid, and the partial genomic sequences obtained from the *lacZ* fusion point screened against *C.elegans* genomic sequence. The active plasmids should be easily identified by virtue of their necessary fusion points in-frame with a genetic coding sequence. As these reactions can be done in parallel in a multiwell format (Favello *et al.*, 1995), active fusion identification should be greatly accelerated.

As discussed in Chapters 3 and 4, approaches using b-Galactosidase as a reporter do have inherent drawbacks however. Specifically, *lacZ* fusions do not seem to be active in early embryos and in the germline, and are efficiently inactivated by passage through a membrane. The ability to assay fusion localisation in all contexts would be a great advance for expression pattern screens. One way to do this, briefly mentioned in Chapter 1, is by epitope tagging. Gene fusions are made containing an antibody epitope sequence, and expression thus monitored by immunochemical means (Munro and Pelham, 1987). A proposal for such an approach has been made for *C.elegans* (Hope and Mounsey, pers. comm.) exploiting a technique enabling chromosomal insertion of extrachromosomal DNA sequences developed by Mello and co-workers. High concentrations of single stranded DNA microinjected into the syncitium of the distal arm of the gonad in *C.elegans* was found to cause random chromosomal insertion events (Mello *et al.*, 1991). Hope and Mounsey expect to cause similar random integrations of epitope tag sequences by microinjecting single stranded DNA encoding a *c-myc* tag epitope (Munro and Pelham, 1987). When the DNA integrates into a coding sequence a translational fusion should occur, and the expression pattern visualised using anti-*myc* antibodies. Such fusions should reveal the true expression pattern of the gene as expression will occur from the true chromosomal context of the gene. The gene containing the epitope tag should be easily identified by tag sequencing using *myc* tag specific primers. If successful, such an approach would present a very powerful tool for the description of gene expression patterns in *C.elegans*.

Prediction of gene function.

As reported in Chapters 3 and 4, the data generated by this doctoral project has allowed functional predictions for several genes to be made. These were made on two major bases; in conjunction with information from protein homology, and as a result of the observed subcellular localisation of a *lacZ* fusion. For example, the ZK637.11 gene has homology to a cell cycle control gene in yeast and is expressed solely in the AB cell lineage in early *C.elegans* embryos (Chapter 4), suggesting that ZK637.11 functions to regulate the AB lineage-specific cell cycle which is of a particular rate compared to lineages derived from other founder blastomeres. In contrast, the B0303.12 gene has no known homologies and is expressed in discrete foci at the hypodermal interface of the bodywall muscle and excretory system tissues suggesting a role in cell attachment to the hypodermis. The anterior-posterior bias of expression in the bodywall muscles further suggests a load-bearing function in this tissue (Chapter 4).

As the percentage of genes having a significant homology to others is bound to increase as more sequences are added to the databases (Waterston and Sulston, 1995), assayed genes exhibiting an expression pattern may be increasingly subject to specific functional predictions on the basis of sequence homology. There thus seems no requirement to restrict expression pattern assay to genes having known homology. Subcellular localisation of reporter gene expression can in any case suggest gene function even when no homology to a characterised gene is available. As suggested in Chapter 4, it is thus desirable that information about subcellular localisation should be actively sought by using expression vectors lacking a nuclear localisation signal.

Predictions as to gene function made on these terms suggest further experiments to confirm or deny the predictions. The most useful confirmation will be that conferred by mutational analysis. As described in Chapter 1, targeted gene inactivation is possible in *C.elegans* using a transposon based protocol (Zwaal *et al.*, 1993). The procedure is difficult, however, due to the tendency of inserted transposon sequences to be spliced from mRNAs (Rushforth and Anderson, 1996) and the intensive work required to obtain deletion derivatives (Zwaal *et al.*, 1993). Some genes may have been fortuitously identified as genetic loci. Chapter 5 describes this type of occurrence with the ZK637.8 gene, which is in the polycistronic unit identified as containing the genetically characterised *unc-32* gene. Its components of expression in ventral cord motoneurones and cells of the early E-lineage correlate well with the observed uncoordinated and embryonic lethal phenotypes of *unc-32*.

As related in Chapter 1, however, the large discrepancy between numbers of genetic loci and predicted genes argues against reliance on such correlations. Methods for easier generation of mutant genes will be required before information on phenotype and

expression can become a systematic means of functional analysis in *C.elegans*. One current proposal is particularly interesting as it combines assay of gene expression with production of mutated genes. It is a gene trapping approach, similar to the epitope tagging approach described above; that is, a reporter gene is randomly integrated into the genome such that it is expressed only when under the direct control of endogenous transcriptional control elements. Epitope tag sequences are so small that insertion into a gene sequence does not usually result in disruption of gene function (Munro and Pelham, 1987). Larger insertions are required to produce significant disruption, as in the transposon-based gene trapping approaches used in mouse (Gossler *et al.*, 1989) and *Drosophila* (Spradling *et al.*, 1995). The approach has enabled a project in *Drosophila* to assay expression patterns for those genes mutated by transposon insertion as the transposon used contains a *lacZ* gene which can only be expressed when close to gene specific transcription regulation signals (Spradling *et al.*, 1995). This exact technique is not possible in *C.elegans* due to the frequent excision of transposon sequences from mRNAs described above. An alternative approach utilising retroviral insertions that place a marker gene into the endogenous transcription unit utilizing the normal enhancer and basal promotor elements, resulting in loss-of-function mutations, is now proposed (White and Greenstein, pers. comm.). The technical hurdle for this approach is to engineer a retrovirus capable of crossing *C.elegans* membranes, as no *C.elegans* specific retroviruses are known. Retroviruses usually require a membrane from the host with a suitable ligand for the viral envelope glycoproteins to interact with so that membrane fusion and core particle uncoating can proceed. White and Greenstein expect to enable retroviral entry into *C.elegans* cells by engineering a virus incorporating the G glycoprotein of Vesicular Stomatitis Virus (VSVG) which interacts with the lipid bilayer to allow membrane fusion and enable ingress of the viral genome into the cell (Burns *et al.*, 1993). Microinjection of such an engineered virus containing a *lacZ* reporter gene should allow the concurrent mutation and assay of expression pattern for genes into which the viral genome integrates. Functional predictions will thus be more efficiently and certainly made for genes in *C.elegans* genome sequence.

Provision of promoter sequences.

The ability of the expression pattern screen employed in this study to provide promoter regions of use in elucidation of the regulatory mechanisms of gene expression has been investigated as described in Chapter 5. The ZK637.8 gene was subject to promoter deletion analysis and a region necessary for embryonic expression of the gene defined. Further analysis was suggested to more accurately define the control elements responsible for each component of ZK637.8 expression. These sequences, once defined, will allow experimental progression towards those genes directly regulating expression. There are few examples of such characterisations published for *C.elegans* genes, so such data

represents a pioneering effort in the elucidation of genetic regulatory pathways during *C.elegans* development.

Markers of cellular differentiation.

In common with other screens of gene expression in *C.elegans* (Chapter 1), the results of this screen have provided a selection of differentiation markers which will be of use to the *C.elegans* research community as a whole. Chapter 6 describes the efforts I have made to ensure the data are easily available to the community through the provision of two computer internet resources.

Conclusion.

I have described the generation and interpretation of expression patterns of genes predicted from *C.elegans* genome sequence. The data produced have proved valuable in the suggestion of function for many of the predicted genes assayed, some of which were completely uncharacterised before this analysis. The data also represent a resource of considerable value to the research community in terms of the provision of differentiation markers and provisional identification of promoters involved in specific aspects of gene expression.

As genome analysis in *C.elegans* continues, it will become more important to be able to access related pieces of information easily. For example, researchers working on their favourite gene might find it shares a promoter control element in common with one of those defined for ZK637.8. They should then be able to easily discover the expression pattern of ZK637.8 and perhaps relate it to their gene's. Further connections will be possible as more gene specific data become available, e.g. the promoter control element may have been characterised relative to a specific transcription factor which also has a known pattern of expression. All of this data needs to be available in a suitable format to allow facile cross-referencing of common themes. In *C.elegans*, the obvious repository of such a resource is ACeDB. Such a centralised database is bound to be rather slow in updating, however, so individual internet-based sites such as the expression pattern archive I have placed on the WWW are bound to become more important.

References.

Ahringer, J and Kimble, J. (1991) Control of the sperm-oocyte switch in *Caenorhabditis elegans* hermaphrodites by the *fem-3* 3' untranslated region. *Nature* **349** 346-348.

Albertson, DG and Thomson, JN. (1976) The pharynx of *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **275** 299-325.

Albrandt, K, Mull, E, Brady, EMG, Herich, J, Moore, CX, Beaumont, K. (1993) Molecular cloning of two receptors from rat brain with high affinity for salmon calcitonin. *FEBS Letters* **325** 225-232.

Allen, ND, Cran, DG, Barton, SC, Hettle, S, Reik, W, Surani, MA. (1988) Transgenes as probes for active chromosomal domains in mouse development. *Nature* **333** 852-855

Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215** 403-410.

Andersen, SLV and Clausen, T. (1993) Calcitonin gene-related peptide stimulates active Na+-K+ transport in rat soleus muscle. *Am. J. Physiol.* **264** 419-429.

Anderson, P. (1995) Mutagenesis. Chapter 2 in "*Caenorhabditis elegans*: Modern Biological Analysis of an Organism". *Meth. Cell. Biol.* **48** 31-58.

Ardizzi, JP and Epstein, HF. (1987) Immunochemical localisation of myosin heavy chain isoforms and paramyosin in developmentally and structurally diverse muscle cell types of *C.elegans*. *J. Cell. Biol.* **105** 2763-2770.

Austin, J and Kimble, J. (1987). *glp-1* is required in the germline for regulation of the decision between mitosos and meiosis in *C.elegans*. *Cell* **51** 589-599.

Avery, L and Wasserman, S. (1992) Ordering gene function: The interpretation of epistasis in regulatory hierarchies. *Trends Genet.* **8** 312-316.

Barnes, WM. (1994) PCR amplification of up to 35kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91** 2216-2220.

Bektesh, S, Vandoren, K, Hirsh, D (1988) Presence of the *Caenorhabditis elegans* spliced leader on different mRNAs and in different genera of nematodes. *Genes Dev.* **2** 1277-1283.

Bellen, HJ, O'Kane, CJ, Wilson, C, Grossniklaus, U, Pearson, RK. (1989) P-element mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* **3** 1288-1300.

Berridge, MJ. (1987) Inositol triphosphate and diacylglycerol - 2 interacting 2nd messengers. *Ann. Rev. Biochem.* **56** 159-193

Blaxter, ML. (1993) Nemoglobins: divergent nematode globins. *Parasitol. Today* **9** 353-360.

Blaxter, ML, Ingram, L and Tweedie, S. (1994) Sequence, expression and evolution of the globins of the parasitic nematode *Nippostrongylus brasiliensis. Mol. Biochem. Parasitol.* **68** 1-14.

Bonnerot, C, Rocancourt, D, Briand, P, Grimbler, G, Nicolas, JF. (1987) A beta-Gal hybrid protein targeted to nuclei as a marker for developmental studies. *Proc. Natl. Acad. Sci. USA* **84** 6795-6799.

Bork, P, Ouzounis, C, Sander, C, Scharf, M, Schneider, R, Sonnhammer, E. (1992) Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III.*Protein Sci.* **1** 1677-1690.

Bossy, BLM, Hall, C and Spierer, P. (1983) Genetic activity along 315kb of the *Drosophila* chromosome. *EMBO J.* **3** 2537-2541.

Bourne, HR, Sanders, DA and McKormick, F. (1990) The GTPase superfamily - a conserved switch for diverse cell functions. *Nature* **348** 125-132.

Bowerman, B, Draper, BW, Mello, CC, Priess, JR. (1993) The maternal gene *skn-1* encodes protein that is distributed inequally in early *C.elegans* embryos. *Cell* **74** 443-452.

Bowerman, B, Tax, FE, Thomas, JH and Priess, JR. (1992) Cell interactions involved in development of the bilaterally symmetrical intestinal valve cells during embryogenesis in *Caenorhabditis elegans. Dev.* **116** 1113-1122.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans. Genetics* **77** 71-94.

Brenner, S. (1990) The human genome - the nature of the enterprise. *Ciba Found. Symp.* **149** 6-17.

Brinster, RL Allen, JM, Behringer, RR, Gelinas, RE, Palmiter, RD. (1988) Introns increase transcriptional efficiency in transgenic mice. *Proc. Natl. Acad. Sci. USA* **85** 836-840.

Brown, JR and Doolittle, WF. (1995) Root of the universal tree of life based on ancient amino-acyl-RNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92** 2441-2445.

Brown, MS, Anderson, RG and Goldstein, JL. (1983) Recycling receptors: the round-trip itinerary of migrant membrane proteins. *Cell* **32** 663-667.

Buckman, AR and Berg, P. (1988) Comparison of intron-dependent and intron-independent gene expression. *Mol. Cell. Biol.* **8** 4395-4405.

Bult, CJ, White, O, Olsen, GJ, Zhou, LX, Fleishmann, RD. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273** 1058-1072.

Burns, JC, Friedman, T, Driever, W, Burrascano, M, Yee, JK. (1993) Vesicular stomatitus virus G glycoprotein pseudotyped retroviral vectors - concentration to very high titre and gene transfer into mammalian and non-mammalian cells. *Proc. Natl. Acad. Sci. USA* **90** 8033-8037.

Burr, AH, Schiefke, R and Bollerup, G. (1975) *Biochim. Biophys. Acta* **405** 404-411.

Burr, AH. (1985) The photomovement of *Caenorhabditis elegans*, a nematode that lacks ocelli. Proof that the response is to light not radiant heating. *Photochem. Photobiol.* **41** 577-582.

Burr, AH. (1987) The phototaxis of *Mermis nigrescens. Biophys. J.* **47** 527-536.

Campbell, HD, Schimansky, T, Claudianos, C, Ozsarac, N. (1993) The *Drosophila melanogaster flightless-I* gene in gastrulation and muscle degeneration encodes gelsolin-like and leucine-rich repeat domains and is conserved in *C.elegans* and humans. *Proc. Natl. Acad. Sci. USA* **90** 11386-11390.

Capecchi, MR. (1989) Altering the genome by homologous recombination. *Science* **244** 1288-1292.

Casadaban, MJ and Cohen SN. (1979) Lactose genes fused to exogenous promoters in one step using a Mu-lac bacteriophage: *in vivo* probe for transcriptional control sequences. *Proc. Natl. Acad. Sci. USA* **76** 4530-4533.

Chater, KF. (1989) in *Regulation of Prokaryotic Development* (eds Smith, I, Slepecky, RA and Setlow, P) 277-299 (Am. Soc. Microbiol., Washington).

Chen, W, Blanc, J and Lim, L. (1994) Charaterisation of a promiscuous GTPase-activating protein that has a Bcr-related domain from *Caenorhabditis elegans. J. Biol. Chem.* **269** 820-823.

Claros, MG and von Heijne, G. (1994) Prediction of transmembrane segments in integral membrane proteins, and the putative topologies, using several algorithms. *CABIOS* **10** 685-686.

Claudianos, C and Campbell, HD. (1995) The novel *flightless-I* gene brings together 2 gene families, actin-binding proteins related to gelsolin and leucine-rich-repeat proteins involved in Ras signal-transduction. *Mol. Biol. Evol.* **12** 405-414.

Collins, F and Galas, D. (1993) A new 5 year plan for the United States Human Genome Project. *Science* **262** 43-46.

Collins, FS. (1995) Positional cloning moves from perditional to traditional. *Nat. Genet.* **9** 347-350.

Conrad, R, Thomas, J, Spieth, J, Blumenthal, T. (1991) Insertion of part of an intron into the 5' untranslated region of a *Caenorhabditis elegans* gene converts it into a transpliced gene. *Mol. Cell. Biol.* **11** 1921-1926

Conrad, R, Liou, RF and Blumenthal, T. (1993) Conversion of a transpilced *Caenorhabditis elegans* gene into a conventional gene by introduction of a splice donor site. *EMBO J.* **12** 1249-1255.

Cooley, L, Kelley, R and Spradling, AC. (1988) Insertional mutagenesis of the *Drosophila* genome with insertional P-elements. *Science* **239** 1121-1128.

Coulson, A, Sulston, J, Brenner, S, Karn, J. (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **83** 7821-7825.

Coulson, A, Waterston, R, Kiff, J, Sulston, J, Kohara, Y. (1988) Genome linking with yeast artificial chromosomes. *Nature* **335** 184-186.

Coulson, A, Kozono, Y, Lutterbach, B, Shownkeen, R, Sulston, J, Waterston, R. (1991) Yacs and the *C.elegans* genome. *Bioessays* **13** 413-417.

Coupland, G. (1995) Genes that promote or delay flowering time - *Arabidopsis* as a model sytem to study flowering. *Phil. Trans. Royal. Soc. B* **350** 27-34

Cowing, DW and Kenyon, C. (1992) Expression of the homeotic gene *mab-5* during *Caenorhabditis elegans* embryogenesis. *Dev.* **116** 481-490.

Creek, KE and Sly, WS. (1984) The role of phosphomannosyl receptor in the transport of acid hydrolases to lysosomes. *In Lysosomes in Biology and Pathology* (eds Dingle, J, Dean, R and Sly, WR) New York, Elsevier. pp 63-82.

Daignanfornier, B, Nguyen, CC, Reisdorf, P, Lemeignan, B, Bolotinfukuhara, M. (1994) MBR1 and MBR3, 2 related yeast genes that can suppress the growth defect of HAP2, HAP3 and HAP4 mutants. *Mol. Gen. Genet.* **243** 575-583.

Daniels, DL, Plunkett, G, Burland, V, Blattner, FR. (1992) Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257** 771-778.

Diekmann, D, Brill, S, Garrett, MD, Totty, N, Hsuan, J, Monfries, C, Hall, C, Lim, L, Hall, A. (1991) Bcr encodes a GTPase activating protein for P21Rac. *Nature* **351** 400-402.

Dujon, B. (1996) The yeast genome project: what did we learn? *Trends Genet.* **12** 263-270.

Duncan, I. (1982) *Polycomblike*: a gene required for normal expression of *Bithorax* and *Antennapedia* gene complexes of *D.melanogaster*. *Genetics* **102** 49-70.

Durbin, R and Thierry-Mieg, J. (1991-). A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Elion, EA, Satterberg, B and Kranz, JE. (1993) FUS3 phosphorylates multiple components of the mating signal transduction cascade - evidence for STE12 and FAR1. *Mol. Biol. Cell* **4** 495-510.

Evans, TC, Crittenden, SL, Kodoyianni, V, Kimble, J. (1994) Translational control of maternal *glp-1* mRNA establishes an asymmetry in *Caenorhabditis elegans* embryos. *Cell* **77** 183-194.

Feldmann, H, Aigle, M, Aljinovic, G, Andre, B, Baclet, MC, Barthe, C, Baur, A, Becam, AM, Biteau, N, Boles, E, Brandt, T, Brendel, M, Bruckner, M. (1994) Complete DNA sequence of yeast chromosome II. *EMBO J.* **13** 5795-5809.

Fickett, JW. (1982) Recognition of protein coding regions in DNA sequences. *Nuc. Acids Res.* **10** 5303.

Fields, C. (1990) Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nuc. Acids Res.* **18** 1509-1512.

Fields, S and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340** 245-246.

Fire, A. (1986) Integrative transformation of *Caenorhabditis elegans*. *EMBO J.* **5** 2673-2680.

Fire, A, Harrison, SW and Dixon, D. (1990) A modular set of *lacZ* fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene* **93** 189-198.

Fire, A. (1992) Histochemical techniques for locating *E.coli* beta-galactosidase activity in transgenic organisms. *Genet. Anal. Tech. Appl.* **9** 152-160.

Fire, A. (1994) A 4-dimensional digital archive imaging system for cell lineage tracing and retrospective embryology. *Comp. Appl. Biosci.* **10** 443-447.

Fleischmann, RD, Adams, MD, White, O, Clayton, RA, Kirkness, EF, Kerlavage, AR, Bult, CJ, Tomb, JF, Dougherty, BA, Merrick, JM, McKenney, K. (1995) Whole genome random sequencing and assembly of *Haemophilus influenzae RD*. *Science* **269** 496-512.

Fondrat, C and Kalogeropoulos, A. (1994) Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae* - application to chromosome III. *Curr. Genet.* **25** 396-406.

Forgac, M. (1989) Structure and function of vacuolar class of ATP-driven proton pump. *Physio. Rev.* **69** 765-796.

Fraser, CM, Gocayne, JD, White, O, Adams, MD, Clayton, RA, Venter, JC. (1995) The minimal gene complement of *Mycoplasma genitalium. Science* **270** 397-403.

Freedman, JH, Slice, LW, Dixon, D, Fire, A, Rubin, CS. (1993) The novel metallothionein genes of *Caenorhabditis elegans. J. Biol. Chem.* **268** 2554-2564.

Frenkel, MJ, Dopheide, TAA, Wagland, BM, Ward, CW. (1992) The isolation, characterization and cloning of a globin-like, host-protective antigen from the excretory-secretory products of *Trichostrongylus colubriformis. Mol. Biochem. Parasitol.* 50 27-36.

Froshauer, S, Green, GN, Boyd, D, McGovern, K, Beckwith, J. (1988) Genetic analysis of the membrane insertion and topology of MalF, a cytoplasmic membrane protein of *E.coli. J. Mol. Biol.* **200** 501-511.

Gawienowski,, MC, Szymanski, D, Perera, IY, Zielinski, RE. (1993) Calmodulin isoforms in Arabidopsis encoded by multiple divergent messenger-RNAs. *Plant Mol. Biol.* **22** 215-225.

Gelbart, WM, Rindone, WP, Chillemi, J, Russo, S, Crosby, M. (1996) Flybase - the *Drosophila* database. *Nuc. Acids Res.* **24** 53-56.

Goldstein, LSB. (1993) Functional redundancy in mitotic force generation. *J Cell Biol.* **120** 1-3.

Goodwin, EB, Okkema, PG, Evans, TC, Kimble, J. (1993) Translational regulation of *tra-2* by its 3' untranslated region controls sexual identity in *Caenorhabditis elegans. Cell* **75** 329-339.

Goring, DR, Rossant, J, Clapoff, S, Breitman, ML, Tsui, LC. (1987) In situ detection of beta-galactosidase in the lenses of transgenic mice with a gamma-crystallin/*lacZ* gene. *Science* **235** 456-458.

Gossler, A, Joyner, AL, Rossant, J, Skarnes, WC. (1989) Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. *Science* **244** 463-465.

Greenwald, I and Horvitz, H. (1980) *unc-93(e1500)*: A behaviour mutant of *Caenorhabditis elegans* that defines a gene with a wild-type null phenotype. *Genetics* **96** 147-164.

Greenwald, I, Sternberg, P and Horvitz, HR. (1983) The lin-12 locus specifies cell fates in *Caenorhabditis elegans*. *Cell* **34** 435-444.

Greenwald, I. (1985) *lin-12*, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. *Cell* **43** 583-590.

Guarente, L. (1993) Synthetic enhancement in gene interaction - a genetic tool come of age. *Trends Genet.* **9** 362-366.

Hamelin, M, Scott, IM, Way, JC, Culotti, JG. (1992) The *mec-7* gene of *Caenorhabditis elegans* is expressed in the touch receptor neurons. *EMBO J.* **11** 2885-2893.

Han, M, Golden, A, Han, YM, Sternberg, PW. (1993) *C.elegans lin-45* Raf gene participates in *let-60* ras-stimulated vulvar differentiation. *Nature* **363** 133-140.

Hansen, MF and Cavanee, WK. (1988) Retinoblastoma and the progression of tumour genetics. *Trends Genet.* **4** 125-128.

Harden, N, Yap, SF, Chiam, MA, Lim, L. (1993) A Drosophila gene encoding a protein with similarity to diacylglycerol kinase is expressed in specific neurons. *Biochem. J.* **289** 439-444.

Hasty, P, Ramiro, R-S, Krumlauf, R and Bradley, A. (1991) Introduction of a subtle mutation in the hox-2.6 locus in embryonic stem cells. *Nature* **350** 243-246.

Havukkala, I, Ichimura, H, Nagamura, Y, Sasaki, T. (1995) Rice genome analysis by integration of sequencing and mapping data. *J. Biotech.* **41** 139-148.

Heisenberg, M and Kaiser, K. (1995) The Flybrain project. *Trends Neurosci.* **18** 481-483.

Hill, RJ and Sternberg, PW. (1992) The gene *lin-3* encodes an inductive signal for vulval development in *Caenorhabditis elegans*. *Nature* **358** 470-476.

Hird, SN and White, JG. (1993) Cortical and cytoplasmic flow polarity in early embryonic cells of *Caenorhabditis elegans*. *J. Cell Biol.* **121** 1343-1355.

Hirsh, D, Oppenheim, D and Klass, M. (1976) Development of the reproductive system of *Caenorhabditis elegans*. *Dev. Biol.* **49** 200-219.

Hodgkin, J, Horvitz, HR and Brenner, S. (1979) Nondisjunction mutants of the nematode *C.elegans*. *Genetics* **91** 67-94.

Hodgkin, J. (1988) Sexual dimorphism and sex determination. (In 'The nematode *Caenorhabditis elegans*.' ed Wood, WB.). pp. 243-280. Cold Spring Harbour Laboratory.

Holm, L, Sander, C. (1995) DNA-polymerase beta belongs to an ancient nucloetidyltransferase superfamily. *Trends Biochem. Sci.* **20** 345-347.

Hope, IA. (1991) Promoter trapping in *Caenorhabditis elegans*. *Dev.* **113** 399-408.

Hope, IA. (1994) PES-1 is expressed during early embryogenesis in *Caenorhabditis elegans* and has homology to the fork head family of transcription factors. *Dev.* **120** 505-514.

Hope, IA, Albertson, DG, Martinelli, SD, Lynch, AS, Sonnhammer, E, Durbin, R. (1996) The *C.elegans* expression pattern database: a beginning. *Trends. Genet.* **12** 370-371.

Hoyt, MA, He, L, Loo, KK and Saunders, WS. (1992) 2 *Saccharomyces cerevisiae* kinesin-related gene products required for mitotic spindle assembly. *J. Cell Biol.***118** 109-120.

Huang, LS and Sternberg, PW. Genetic dissection of developmental pathways. Chapter 5 in "*Caenorhabditis elegans*: Modern Biological Analysis of an Organism". *Meth. Cell. Biol.* **48** 97-122.

Inouye, C, Remondelli, P, Karin, M, Elledge, S. (1994) Isolation of a cDNA encoding a metal-response-element binding protein using a novel expression cloning procedure - the one hybrid system. *DNA Cell Biol. 13* 731-742.

Jamieson, AC, Wang, HM and Kim, SH. (1996) Zinc finger directory for high affinity DNA recognition. *Proc. Natl. Acad. Sci. USA* **93** 12834-12839.

Johnson, RG. (1988) Accumulation of biological amines into chromaffin granules - a model for hormone and neurotransmitter transport. *Physiol. Rev.* **68** 232-307.

Johnston, M. (1996) Towards a complete understanding of how a simple eukaryote cell works. *Trends Genet.* **12** 242-243.

Kaback, DB, Angerer, LM and Davidson, N. (1979) Improved methods for the formation and stabilization of R-loops. *Nucleic Acids Res.* **6** 2499-2517.

Kalderon, D, Roberts, BL, Richardson, WD, Smith, AE. (1984a) A short aminoacid sequence able to specify nuclear location. *Cell* **39** 499-509.

Kalderon, D, Richardson, WD, Markham, AF, Smith, AE. (1984b) Sequence requirements for nuclear location of simian virus 40 large-T antigen. *Nature* **311** 33-38.

Karn, J, Brenner, S, Barnett, L. (1983) Protein structural domains in the *Caenorhabditis elegans unc-54* myosin heavy chain gene are not separated by introns. *Proc. Natl. Acad. Sci. USA* **80** 4253-4257.

Kenyon, C. (1988) The nematode *Caenorhabditis elegans*. Science 240 1448-1452.

Kimble, J and Hirsh, D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Dev. Biol.* **70** 396-417.

Klein, P, Kanehisa, M, Delisi, C. (1985) The detection and classification of membrane spanning proteins.*Biochimica et Biophysica Acta* **815** 468-476

Klug, A and Rhodes, D. (1987) "Zinc fingers": a novel protein motif for nucleic acid recognition. *Trens Biochem. Sci.* **12** 464-469.

Tabara, H, Motohashi, T, and Kohara, Y. (1996) A multiwell version of *in situ* hybridisation on whole-mount embryos of *Caenorhabditis elegans*. *Nuc. Acids Res.* **24** 2119-2124.

Koonin, EV, Tatusov, RL and Rudd, KE. (1995) Sequence similarity analysis of *Escherichia coli* proteins - functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA* **92** 11921-11925.

Kramer, J, French, RP, Park, EC, and Johnson, JJ. (1990) The *Caenorhabditis elegans rol-6* gene, which interacts with the *sqt-1* collagen gene to determine organismal morphology, encodes a collagen. *Mol. Cell. Biol.* **10** 2081-2090.

Kraus, JP, Firgaira, F, Novotny, J, Kalousek, F, Williams, KR, Williamson, C, Ohura, T, Rosenberg, LE. (1986) Coding sequence of the precursor of the beta unit of rat propionyl-CoA carboxylase. *Proc. Natl. Acad. Sci. USA* **83** 8049-8053.

Krause, M. (1986) Actin gene expression in the nematode *Caenorhabditis elegans*. PhD thesis, University of Colorado, Boulder, Colorado.

Krause, M and Hirsh, D. (1987) A transpliced leader sequence on actin mRNA in *Caenorhabditis elegans*. *Cell* **49** 753-761.

Krause, M, Fire, A, Harrison, SW, Priess, J, Weintraub, H. (1990) CeMyoD accumulation defines the bodywall muscle cell fate during C.elgans embryogenesis. *Cell* **63** 907-919.

Krause, M, Harrison, SW, Xu, SQ, Chen, LS, Fire, A. (1994) Elements regulation cell and stage-specific expression of the *Caenorhabditis elegans* MyoD homolog *hlh-1*. *Dev. Biol.* **166** 133-148.

Krause, M. (1995) Transcription and Translation. Chapter 20 in "*Caenorhabditis elegans*: Modern Biological Analysis of an Organism". *Meth. Cell. Biol.* **48** 483-512.

Lambie, EJ and Kimble, J. (1991) 2 homologous regulatory genes, lin-12 and glp-1, have overlapping functions. *Dev.* **112** 231-240

Lamphier, MS and Ptashne, M. (1992) Multiple mechanisms mediate glucose repression of the yeast GAL1 gene. *Proc. Natl. Acad. Sci. USA* **89** 5922-5926.

Land, H, Parada, LF and Weinberg, RA. (1983) Tumorigenic conversion of primary embryo fibroblasts requires at least 2 cooperating oncogenes. *Nature* **304** 596-602.

Land, M, Islastrejo, A, Freedman, JH, Rubin, CS. (1994) Structure and expression of a novel, neuronal protein kinase C (PKC1B) fom *Caenorhabditis elegans. J. Biol. Chem.* **269** 9234-9244.

Lee, CH and Gumbiner, BM. (1994) Disruption of gastrulation movements in Xenopus by a dominant negative mutant for C-cadherin. *Dev. Biol.* **171** 363-373.

Lifshitz, B, Fainstein, E, Marcelle, C, Shtivelman, E, Amson, R, Gale, RP, Canaani, E. (1988) Bcr genes and transcripts. *Oncogene* **2** 113-117.

Lin, HY, Harris, TL, Flannery, MS, Aruffo, A, Kaji, EH, Gorn, A, Kolakowski, LF, Lodish, HF, Goldring, SR. (1991) Expression cloning of an adenylate cyclase-coupled calcitonin receptor. *Science* **254** 1022-1024.

Lincke, CR, Broeks, A, The, I, Plasterk, RHA, Borst, P. (1993) The expression of two P-glycoprotein (pgp) genes in transgenic *Caenorhabditis elegans* is confined to intestinal cells. *EMBO J.* **12** 1615-1620.

Lind, SE and Janmey, PA. (1984) Human plasma gelsolin binds to fibronectin. *J. Biol. Chem.* **259** 3262-3266.

Lis, JT, Simon, JA and Sutton, CA. (1983) New heat shock puffs and beta-galactosidase activity resulting from transformation of Drosophila with an *hsp70-lacZ* hybrid gene. *Cell* **35** 403-410.

Lord, EM, Crone, W and Hill, JP. (1994) Timing of events during plant organogenesis - *Arabidopsis* as a model system. *Curr. Top. Dev. Biol.* **29** 325-356.

Luban, J, Bossolt, KL, Franke, EK, Kalpana, GV, Goff, SP. (1993) Human immunodeficiency virus type I gag protein binds to cyclophilinA and cyclophilinB. *Cell* **73** 1067-1078.

Lundberg, KS, Short, JM, Sorge, JA, Mathur, EJ. (1991) A new thermostable polymerase with high fidelity. *Gene* **108** 1-6.

Lynch, AS, Briggs, D and Hope, IA. (1995) Developmental expression pattern screen for genes predicted in the C.elegans genome sequencing project. *Nat. Genet.* **11** 309-313.

Ma, J and Ptashne, M. (1987) Deletion analysis of GAL4 defines two transcriptional activating segments. *Cell* **48** 847-853.

MacMorris, M, Broverman, S, Greenspoon, S, Lea, K, Madej, C, Blumenthal, T, Spieth, J. (1992) Regulation of vitellogenin gene expression in transgenic C.elehgans: short sequences required for the activation of the *vit-2* promoter. *Mol. Cell. Biol.* **12** 1652-1662.

Masai, I, Hosoya, T, Kojima, SI, Hotta, Y.l (1992) Molecular cloning of a Drosophila diacylglycerol kinase gene that is expressed in the nervous sytem and muscle. *Proc. Natl. Acad. Sci. USA* **89** 6030-6034.

Matsubara, K and Okubo, K. (1993) cDNA analyses in the human genome sequence project. *Gene* **135** 265-274.

McCombie, WR, Adams, MD, Kelley, JM, Fitzgerald, MG, Uutterback, TR, Khan, M, Dubnick, M, Kerlavage, AR, Venter, JC, Fields, C. (1992) *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nat. Genet.* **1** 124-131.

Mello, CC, Kramer, JM, Stinchcomb, D, Ambros, V. (1991) Efficient gene transfer in *Caenorhabditis elegans*: Extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10** 3959-3570.

Mello, C and Fire, A. (1995) DNA transformation. Chapter 19 in "*Caenorhabditis elegans*: Modern Biological Analysis of an Organism". *Meth. Cell. Biol.* **48** 451-482.

Miklos, GLG and de Couet, HG. (1990) The mutations previously designated as *flightless-I3, flightless-O2 and standby* are memebers of the W-2 lethal complementation group at the base of the X-chromosome of *D.melanogaster. J. Neurogenet.* **3** 133-151.

Mitani, S, Du, HP, Hall, DH, Driscoll, M, Chalfie, M. (1993) Combinatorial control of touch receptor neuron expression in *Caenorhabditis elegans. Dev.* **119** 773-883.

Montell, C and Rubin, GM. (1989) Molecular characterisation of the *Drosophila* Trp locus -a putative integral membrane protein required for phototransduction. *Neuron* **2** 1313-1323.

Munro, S and Pelham, RB. (1987) A C-terminal signal prevents secretion of luminal ER proteins. *Cell* **48** 899-907.

Nagy, I, Schoofs, G, Vanderleyden, J, Demot, R. (1992) Sequence of a Rhodococcus gene encoding a protein with extensive homology to the mammalian propionyl-CoA carboxylase beta-chain. *Gene* **122** 199-202.

Nakanishi, S. (1986) Structure and regulation of the preprotachykinin gene. *Trends. Neurosci.* **9** 41-44.

Nelson, FK, Albert, PS and Riddle, DL. (1983) Fine structure of the *Caenorhabditis elegans* secretory-excretory system. *J. Ultrastruct. Res.* **82** 156-171.

Nurse, P. (1983) Cell cycle control in yeast. *Trends in Genetics* **1** 51-55.

O'Kane, C and Gehring, WJ. (1987) Detection in situ of genomic regulatory elements in Drosophila. *Proc. Natl. Acad. Sci. USA* **84** 9123-9127.

O'Neil, KT and DeGrado, WF. (1990) How calmodulin binds its targets - sequence independent recognition of amphiphilic alpha-helices. *Trends Biochem. Sci.* **15** 59-64.

Okkema, P, Harrison, SW, Plunger, V, Aryana, A, Fire, A. (1993) Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135** 385-404.

Okkema, PG and Fire, A. (1994) The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Dev.* **120** 2175-2186.

Olesen, J, Hahn, S and Guarente, L. (1987) Yeast HAP2 and HAP3 activators both bind to the CYC1 upstream activation site, UAS2, in an independent manner. *Cell* **51** 953-961.

Oliver, SG. (1992) The complete DNA sequence of yeast Chromosome III. *Nature* **357** 38-46.

Oliver, SG. (1996) From DNA sequence to biological function. *Nature* **379** 597-600.

Oliver, S. (1996) A network approach to the systematic analysis of gene function. *Trends in Genetics* **12** 241-242.

Olson, MV, Dutchik, JE, Graham, MY, Brodeur, GM, Helms, C. (1986) Random clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83** 7826-7830.

Olson, MV. (1991) in "The molecular and cellular biology of the yeast *Saccharomyces cerevisiae*" (eds Broach, JR, Pringle, JR and Jones, EW). pp. 1-39 (Cold Spring Harbour Laboratory, New York).

Orr, SL, Hughes, TP, Sawyers, CL, Kato, RM, Quan, SG. (1994) Isolation of unknown genes from human bone-marrow by differential screening and single-pass cDNA sequence determination. *Proc. Natl. Acad. Sci. USA* **91** 11869-11873.

Park, E and Horvitz, HR. (1986) Mutations with dominant effects on the behaviour and morphology of the nematode *Caenorhabditis elegans*. *Genetics* **113** 821-852.

Perin, MS, Fried, VA, Stone, DK, Xie, XS, Sudhof, TC. (1991) Structure of the 116-kDa polypeptide of the clathrin-coated vesicle/synaptic vesicle proton pump. *J. Biol. Chem.* **266** 3877-3881.

Pernow, B. (1983) Substance P. *Pharmacol. Rev.* **35** 85-141.

Perrimon, N, Smouse, D and Miklos, GLG. Developmental genetics of loci at the base of the X-chromosome of *D.melanogaster*. *Genetics* **121** 313-331.

Perry, M, Li, WQ, Trent, C, Robertson, B, Fire, A, Hageman, JM, Wood, WB. (1993) Molecular characterisation of the her-1 gene suggests a direct role in cell signalling during *Caenorhabditis elegans* sex determination. *Genes Dev.* **7** 216-228.

Phillips, AM, Bull, A and Kelly, LE. (1992) Identification of a Drosophila gene encoding a calmodulin-binding protein with homology to the trp phototransduction gene. *Neuron* **8** 631-642.

Plasterk, RHA. (1995) Reverse genetics. Chapter 3 in "*Caenorhabditis elegans*: Modern Biological Analysis of an Organism". *Meth. Cell. Biol.* **48** 59-80

Poole, B and Ohkuma, S. (1981) Effect of weak bases on the intralysosomal pH in peritoneal macrophages. *J. Cell. Biol.* **90** 665-669.

Posner, BI, Josefsberg, Z and Bergeron, JJ. (1978) Intracellular polypeptide hormone receptors. *J. Biol. Chem.* **253** 4067-4093.

Priess, JR and Thomson, JN. (1987) Cellular interactions in early *C.elegans* embryos. *Cell* **48** 241-250.

Priess, JR, Schnabel, H, and Schnabel, R. (1987) The *glp-1* locus and cellular interactions in early *C.elegans* embryos.*Cell* **51** 601-611.

Ringwald, M, Baldock, R, Bard, J, Kaufman, M, Eppig, JT, Richardson, JE, Nadeau, JH, Davidson, D. (1994) A database for mouse development. *Science* **265** 2033-2034.

Roberts, BL, Richardson, WD, Smith, AE. (1987) The effect of protein context on nuclear location signal function.*Cell* **50** 465-475.

Roof, DM, Meluh, PB and Rose, MD. (1992) Kinesin related proteins required for assembly of the mitotic spindle. *J. Cell Biol.***118** 95-108.

Rothstein, RJ. (1983) One step gene disruption in yeast. *Meth. Enzym.* **101** 202-211.

Rudnick, G. (1986) ATP-driven proton pumping into intracellular organelles. *Annu. Rev. Physiol.* **48** 403-413.

Rushforth, AM, Saari, B and Anderson, P. (1993) Site-selected insertion of the transposon Tc1 into a *Caenorhabditis elegans* myosin light chain gene. *Mol. Cell. Biol.* **13** 902-910.

Rushforth, AM and Anderson, P. (1996) Splicing removes the *Caenorhabditis elegans* transposon Tc1 from most mutant premessenger RNAs. *Mol. Cell. Biol.* **16** 422-429.

Saiki, RK, Gelfand, DH, Stoffel, S, Scharf, SJ, Higuchi, R, Horn, GT, Mullis, KB, Erlich, HA. (1988) Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239** 487-491.

Schmidt, R and Dean, C. (1993) Towards construction of an overlapping YAC library of the *Arabidopsis thaliana* genome. *Bioessays* **15** 63-69.

Seydoux, G and Fire, A. (1994) Soma-germline asymmetry in the *in situ* hybridisation to distributions of embryonic RNAs in *Caenorhabditis elegans*. *Dev.* **120** 2823-2834.

Sharpe, MJ and Lee, DL. (1981) The effect of anaerobiosis on adenosine nucleotide levels in *Nematospiroides dubius* and *Trichostrongylus colubriformis in vitro*. *Parasitol.* **83** 425-433.

Shastry, BS. (1994) More to learn from gene knockouts. *Mol. Cell. Biochem.* **136** 171-182.

Sherman, DR, Kloek, AP, Krishnan, BR, Guinn, B, Goldberg, DE. (1992) Ascaris hemoglobin gene: plant-like structure reflects the ancestral globin gene. *Proc. Natl. Acad. Sci. USA* **89** 11696-11700.

Silhavy, TJ and Beckwith, JR. (1985) Uses of *lacZ* fusions for the study of biological problems. *Microbiol. Rev.* **49** 398-418.

Simchen, G, Kassir, Y. (1989) Genetic regulation of differentiation towards yeast in the yeast *Saccharomyces cerevisiae*. *Genome* **31** 95-99.

Simon, MA, Bowtell, DDL, Dodson, GS, Laverty, TR, Rubin, GM. (1991) Ras1 and Putative Guanine Nucleotide Exchange Factor perform crucial steps in signalling by the *Sevenless* protein tyrosine kinase. *Cell* **67** 701-716.

Soldati, T and Perriard, JC. (1991) Intracompartmental sorting of essential myosin light chains: Molecular dissection and *in vivo* monitoring by epitope tagging. *Cell* **66** 277-289.

Spieth, J, Brooke, G, Kuersten, S, Lea, K, Blumenthal, T. (1993) Polycistronic mRNA precursors are processed by transplicing of SL2 to downstream coding regions. *Cell* **73** 521-532.

Spradling, AC, Stern, DM, Kiss, I, Roote, J, Laverty, T, Rubin, GM. (1995) Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl. Acad. Sci. USA* **92** 10824-10830.

Stinchcomb, DT, Shaw, JE, Carr, SH, Hirsh, D. (1985) Extrachromosomal DNA transformation of *Caenorhabditis elegans*. *Mol. Cell. Biol.* **5** 3484-3496.

Straub, KL, Stella, MC and Leptin, M. (1996) The gelsolin-related *flightless-I* protein is required for actin distribution during cellularisation in Drosophila. *J. Cell Sci.* **109** 263-270.

Strauss, AW and Boime, I. (1982) Compartmentation of newly synthesized proteins. *CRC Crit. Rev. Biochem.* **12** 205-235.

Stringham, EG, Dixon, DK, Jones, D, Candido, EPM. (1992) Temporal and spatial expression patterns of the small heat shock (hsp16) genes in transgenic *Caenorhabditis elegans*. *Mol. Biol. Cell.* **3** 221- 233.

Sulston, J, Dew, M and Brenner, S. (1975) Dopaminergic neurones in the nematode *Caenorhabditis elegans*. *J. Comp. Neurol.***163** 215-226.

Sulston, JE and Horvitz, HR. (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56** 110-156.

Sulston, JE and White, JG. (1980) Regulation and cell autonomy during postembryonic development of *Caenorhabditis elegans*. *Dev. Biol.* **78** 577-597.

Sulston, JE, Albertson, DG and Thomson, JN. (1980) The *Caenorhabditis elegans* male: postembryonic development of non-gonadal structures. *Dev. Biol.* **78** 542-576.

Sulston, JE, Schierenberg, E, White, JG, Thomson, JN. (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100** 64-119.

Sulston, J, Du, Z, Thomas, K, Wilson, R, Hillier, L, Staden, R, Halloran, N, Green, P, Thierry-Mieg, J, Qiu, L, Dear, S, Coulson, A, Craxton, M, Durbin, R, Berks, M, Metzstein, M, Hawkins, T, Ainscough, R, Waterston, R. (1992) The *Caenorhabditis elegans* genome sequencing project: a beginning. *Nature* **356** 37-41.

Tautz, D and Pfiefle, C. (1989) A nonradioactive *in situ* hybridisation method for the localisation of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene *hunchback*. *Chromosoma* **98** 81-85.

Thomas, JH, Birnby, DA and Vowels, JJ. (1993) Evidence of parallel processing of sensory information controlling dauer formation in *Caenorhabditis elegans*. *Genetics* **134** 1105-1117.

Thomas, JH. (1993) Thinking about genetic redundancy. *Trends Genet.* **9** 395-399.

Tugendreich, S, Bassett, DE, McKusick, VA, Boguski, MS, Hieter, P. (1994) Genes conserved in yeast and humans. *Hum. Mol. Genet.* **3** 1509-1517.

von Heijne, G. (1983) Patterns of amino acids near signal sequence cleavage sites. *Eur. J. Biochem.* **755** 17-21.

von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nuc. Acids Res.* **14** 4683.

von Heijne, G. (1992) Membrane-protein structure prediction - hydrophobicity analysis and the positive inside rule. *J. Mol. Biol.* **225** 487-494.

Wagner, A. (1996) Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. *Biol. Cybernet.* **74** 557-567.

Ward, S, Thomson, N, White, JG and Brenner, S. (1975) Electron microscopical reconstruction of the anterior sensory anatomy of the nematode *Caenorhabditis elegans. J. Comp. Neurol.* **160** 313-338.

Ware, RW, Clark, D, Crossland, K and Russell, RL. (1975) The nerve ring of the nematode *Caenorhabditis elegans.J. Comp. Neurol.* **162** 71-110.

Waterston, R. (1988) Muscle. (In 'The nematode *Caenorhabditis elegans.*' ed Wood, WB.) pp 281-335. Cold Spring Harbour Laboratory.

Way, JC, Wang, L, Run, JQ, Wang, A. (1991) The *mec-3* gene contains *cis*-acting elements mediating positive and negative regulation in cells produced by asymmetric *Cell* division in *Caenorhabditis elegans. Genes Dev.* **5** 2199-2211.

Waterston, R, Martin, C, Craxton, M, Huynh, C, Coulson, A, Hillier, L, Durbin, R, Green, P, Shownkeen, R, Halloran, N, Metzstein, M, Hawkins, T, Wilson, R, Berks, M, Du, Z, Thomas, K, Thierry-Mieg, J, Sulston, J. (1992) A survey of expressed genes in *Caenorhabditis elegans. Nat. Genet.* **1** 114-123.

Waterston, R and Sulston, J. (1995) The genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* **92** 10836-10840.

White, J, Southgate, E, Thomson, JN and Brenner, S. (1986) The structure of the nervous system of the nematode *Caenorhabditis elegans. Phil. Trans. R. Soc. B* **314** 1-340.

White, J. (1988) The anatomy. (In 'The nematode *Caenorhabditis elegans.*' ed Wood, WB.) pp 81-122. Cold Spring Harbour Laboratory.

White, JG, Southgate, E and Thomson, JN. (1991) On the nature of undead cells in the nematode *Caenorhabditis elegans. Phil. Trans. R. Soc. Lond. B* **331** 263-271.

Wightman, B, Ha, I and Ruvkun, GB. (1993) Postranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *Caenorhabditis elegans. Cell* **75** 855-862.

Wilson, R, Ainscough, R, Anderson, K, Baynes, C, Berks, M, Burton, J, Connell, M, Bonfield, J, Copsey, T, Cooper, ,J, Coulson, A, Craxton, M, Dear, S, Du, Z, Durbin, R, Favello, A, Fraser, A, Fulton, L, Gardner, A, Green, P, Hawkins, T, Hillier, L, Jier, M, Johnson, L, Jones, M, Kershaw, J, Kirsten, J, Laisster, N, Latreille, P, Lloyd, C, Mortimore, B, O'Callaghan, M, Parsons, J, Percy, C, Rifken, L, Roopra, A, Saunders, D, Shownkeen, R, Sims, M, Smaldon, N, Smith, A, Smith, M, Sonnhammer, E, Staden, R, Sulston, J, Thierry-Mieg, J, Thomas, K, Vaudin, M, Vaughan, K, Waterston, R. (1994) 2.2Mb of contiguous nucleotide sequence from chromosome III of *Caenorhabditis elegans. Nature* **368** 32-38.

Woese, CR, Kandler, O and Wheelis, ML. (1990) Towards a natural system of organisms - proposal for the domains Archaea, Bacteria and Eucarya. *Proc. Natl. Acad. Sci. USA* **87** 4576-4579.

Wolpert, L. (1992) Gastrulation and the evolution of development. *Dev. (Special Suppl).* 7-13.

Wolpert, L. (1994) The evolutionary origin of development - cycles, patterning, privelidge and continuity. *Dev. (Special Suppl).* 79-84.

Wood, WB. (1988) Introduction to *C.elegans* Biology. (In 'The nematode *Caenorhabditis elegans.*' ed Wood, WB.). pp. 1-16. Cold Spring Harbour Laboratory.

Yanagida, M. (1990) Higher order chromosome structure in yeast. *J. Cell. Sci.* **96** 1-3.

Yanisch-Perron, C, Vieira, J and Messing, J. (1985) Improved cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33** 103-119.

Yochem, J and Greenwald, I. (1989) *glp-1* and *lin-12*, genes implicated in distinct cell-cell interactions in *Caenorhabditis elegans*, encode similar transmembrane proteins. *Cell* **58** 553-563.

Young, JM and Hope, IA. (1993) Molecular markers of differentiation in *Caenorhabditis elegans* obtained by promoter trapping. *Devl. Dynam.* **196** 124-132.

Zarkower, D and Hodgkin, J. (1992) Molecular analysis of the *C.elegans* sex-determining gene *tra-1*: a gene encoding two zinc finger proteins. *Cell* **70** 237-249.

Zorio, DAR, Cheng, NSN, Blumenthal, T, Spieth, J. (1994) Operons as a common form of chromosomal organisation in *Caenorhabditis elegans. Nature* **372** 270-272.

Zwaal, RR, Broeks, A, Vanmeurs, J, Groenen, JTM, Plasterk, RHA. (1993) Target-selected gene inactivation in *C.elegans* by using a frozen transposon insertion mutant bank. *Proc. Natl. Acad. Sci. USA* **90** 7431-7435.