The Isolation And Characterisation Of Mouse Genes Containing CAG/CTG Trinucleotide Repeats.

Thomas William Dunlop

University Of Glasgow

Division Of Molecular Genetics

February 1997

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Summary.

Triplet repeats are simple tandem repeats with a basic unit of three nucleotides. The CTG/CAG class, which is over-represented in higher eukaryote genomes, occurs in a number of regulatory genes, often encoding poly-glutamine. More recently, expansion of CAG/CTG triplet repeats in certain human genes has become associated with certain diseases, leading either to loss of function (as the case may be with CTG in Myotonic Dystrophy) or gain of function (CAG encoding glutamine stretches in SCA1, Huntington's disease, Kennedy's Disease). Since the mouse is the best mammalian organism in which to study the biology of triplet repeats and the genes in which they occur, a set of mouse cDNAs containing triplet repeats was isolated and partially characterised.

To this end, DNA sequencing was performed to identify the trinucleotide repeats. Sequences flanking the trinucleotide repeat were used to identify genes which show similarity or identity to previously characterised genes. This was used to compare the lengths of the trinucleotide repeats found in the mouse clones described in this work with those found in (previously characterised) genes which have trinucleotide repeats at identical positions.

To identify clones which showed changes to mRNA abundance during development and adult tissue, reverse dot-blot analysis was performed on a subset of cDNA clones derived from the 8.5 and 12.5 dpc whole mouse embryo cDNA libraries.

Finally, four clones (two high and two low RNA abundance) were selected for further molecular analysis. Of these four clones, the two which exhibited high levels of RNA in

the reverse dot blot experiments, showed a more complex expression pattern in northern analysis and all four clones appear to derive from single copy (per haploid genome) genes.

In conclusion, this work has indicated that mouse genes contained similarly sized trinucleotide repeats compared to those found in human genes; that these repeats are possibly conserved between distantly related species and this may be related to the proper function of the genes. A third of those genes (which were analysed for the existence of RNA derived from the parental gene) displayed detectable amounts of expression and most showed variations in the RNA abundance either at different stages during development or in different adult tissues. Most of those were derived from unknown genes. This indicates that there may be many novel genes containing CAG/CTG trinucleotide in the mouse which may be identified in the future for further characterisation during developmental and organ specific analysis.

One clone (mCTG 63), although not novel, contains the largest perfect CAG/CTG triplet repeat found in this work. The repeat may be transcribed as CTG in the coding strand of this gene. This situation is analogous to the CTG triplet repeat which has a role in myotonic dystrophy (found in the 3'UTR of the DMPK gene). This clone is worth further analysis for this reason.

Acknowledgements

I would like to thank Professor R. Wayne Davies for allowing me to conduct this research in his laboratory.

I would also like to acknowledge the ancillary staff of the department, particularly the prep room ladies for putting up with last minute orders for media and consumables; Lyn and Irene for their technical assistance in the laboratory; Ed Gallagher for his numerous technical, intellectual and social abilities. I would also like to thank Hugh and Fiona, Mo, Maria, Mark and Donna, Mary and Gerry, the three Davids (Hamilton, Walsh and Livingston) and Rab McNab for their friendship in various labs, pubs and flats round about Glasgow. A big thank you also for Tracey Naismith and Roz Reid, without whom, there would be no snow in winter.

I would also like to thank Dr Genevieve Bart for the critical reading of the thesis and Dr Peggy Shelbourne for helpful advice with the introduction of this thesis; Richard Wilson and Mark Bailey for their wisdom as secondary supervisors; Douglas for lots of help with computers and slide making; Kim for letting me use the colour scanner and Dr Marshall Stark for the use of his Phosphoimager and printers.

Thanks also to my parents for their financial support and encouragement throughout my time at University. Finally I would like to thank my wife, Genevieve, for enormous amounts of love, patience help and thesis reading she has given me during the work in the laboratory and writing of the thesis.

"So what would happen if everybody thought the same way you did?"

"I would be a damn fool to think any other way."

from Catch-22

by Joseph Heller

To my Wife and Family

Table Of Contents

Chapter 1
General Introduction

1.1 Introduction.

One of the aims of today's research is to understand how the various parts of an organism function and how they interact together. This has been true from early modern biology where scientists first described phenomena visually and later with the aid of the microscope. As more elaborate and powerful equipment and techniques have been invented, the rate of accumulation of information and the resolution of the mechanisms has increased rapidly.

Molecular biology techniques have been used with great effect in understanding biological systems. For instance, the polymerase chain reaction (PCR) has been applied to the identification of the whole gene content of certain organisms including man and mouse. For example, it has been used in the mapping of genes on to physical and genetical maps (Love *et al*, 1990) and the identification of gene sequence using a modified version (i.e. cycle sequencing using fluorescent dye-terminators as described by Perkin-Elmer; Trower *et al*, 1995) of the DNA dideoxy sequencing protocol (Sanger *et al*, 1978). The second example here highlights another point, that these methods can always be improved.

Mammalian systems are relatively under-characterised as concerning developmental processes. Most work has been done in other organisms, particularly, the fruit fly *Drosophila,* in which it has been easier to dissect development. As a result of this, many mammalian developmental genes have been identified by their homology to *Drosophila* genes. However, this assumes that genes that are used in development, are common to all organisms. This

Figure 1.1

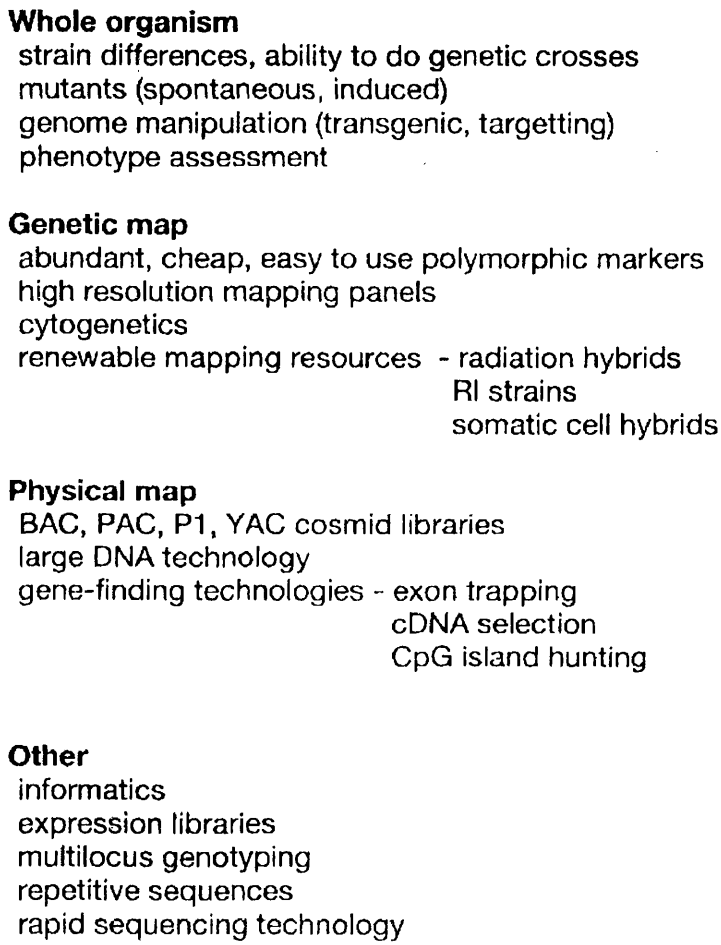


Figure 1.1 Approaches for analysis of mammalian genomes. This Figure details the different ways for the analysis of mammalian genomes, to identify genes and to locate mutations that cause disease. Adapted from Frankel (1995).

appears to be true for the primary developmental genes (such as homeobox containing genes; see Section 1.3). However this does not preclude the action of novel genes in the development of any structures found to be unique to an individual species.

Alternative approaches have been developed that make it easier to dissect development in mammals. A combination of older and more recently developed techniques (see Figure 1.1), make it now possible to dissect mammalian development more effectively. The mouse has an important role in this process as it is the best understood and characterised model system for the investigation of developmental biology of mammals. The early embryological stages post fertilisation are common to both human (a primate) and the mouse. However, in rodents there is a process of "turning" where the embryo is observed to invert with respect to the rostral-caudal axis, with the primordial cell types rearranging themselves. After this process major organogenesis occurs in a way which is similar to human development. A more detailed description of murine development can be found in Kauffman and Kauffman (1992).

1.2 Molecular techniques for finding novel genes.

1.2.1 Mass sequencing experiments.

The most basic of strategies rely on the random sequencing of genetic material. Examples of this would be the human genome project (Dausset *et al*, 1990), its mouse counterpart and mass sequencing of cDNA clones from a particular library.

These projects initially generate a lot of information which is subjected to other criteria to determine interest from a particular

view point. For example, an anonymous gene sequence would require to be analysed for its spatio-temporal expression to assess its possible involvement in developmental processes. This is feasible for a few genes but to apply this to a large number of cloned genes would be very time consuming.

Related to the mass sequencing projects outlined above, there are a group of techniques which take advantage of the differences in transcribed material between different cell types and different stages during development. Among the differentially expressed genes there are genes which are unique for a particular cell type or developmental stage. These genes provide a way into the molecular characterisation of that cell type and of specific developmental processes.

The basic requirement for these techniques is the existence of two distinct pools of expressed genetic information which can be used to describe and identify differences between them. Subtractive hybridisation (Hedrick *et al*, 1984) uses solution hybridisation to remove sequences which are shared between the two pools of information. The remainder is specific to one library. A reciprocal experiment is required to determine the unique portion of the other library. This approach has also disadvantages. For example, genes which have high steady state expression levels may be over-represented in the final subtracted fraction because of an imbalance in input material. This is counteracted by having an excess of the driver library (which constitutes the probing sequences). Unique sequences can also become excluded because of sequences corresponding to specific protein domains which are shared between different proteins like DNA binding domains and motifs for structural proteins. These common sequences between genes can

lead to the removal of cell-type specific gene sequences containing them.

mRNA differential display (Liang and Pardee, 1992) although requiring two different reference points, does not rely on hybridisation. It is PCR based. Essentially PCR is performed with random primer sets on the two cDNA sources, the products are size separated electrophoretically and compared to each other. Differences between the electrophoretic patterns are taken to be specific to one or the other mRNA sources.

All of the above techniques generated much sequence information, but it was necessary to apply other test criteria to enhance the quality of the data for the specific question that was originally asked *i.e.* which gene sequences are specifically expressed in a particular cell type.

1.2.2 Comparative sequence analysis

Another way to identify genes with interesting patterns (*e.g.* sequences which correlate with functional domains in proteins) is the use of sequence comparison. This is based on the matching of residues (amino acids or nucleic acid bases) between different sequences and building up of a score of relatedness based on the matches in sequence. These scoring methods are conducted by the use of algorithm based programmes which add up the scores for a comparison based on a system of marking matches positively, and penalties for mismatches, insertions and deletions. These programmes have been developed to handle sequences in two ways. Global alignments (*e.g.* dot matrix) make comparisons over the whole sequences. This is sensitive but it is long and slow while

making all the possible combinations of comparisons. More widely used in explicit search programmes is the local alignment method which searches for short perfect matches (called words) of a defined minimum size (k-tuple value). Examples of such programmes are the FASTA programme (Lipman and Pearson, 1985) and BLAST (Altschul *et al*, 1990). FASTA is more sensitive than BLAST because after an initial scoring round it then re-scores the data using a PAM (Point Accepted Mutation) table. This assumes that not more than one change has occurred at any one position and allows for more distant relationships between sequences to be ascertained. A more detailed explanation and comparison of available search programmes can be found in Chapter 7 (page 215-248) of the Guide to Human Genome Computing (1994).

The identification of sequence patterns by informatics is a useful technique, but one should be cautious of the results and it is better to independently verify similarities by direct experimentation. For example, repetitive sequence can alter significantly the score of alignments and lead to mis-interpretation of data. However, without computer based sequence comparisons, it would be impossible to handle the large volume of data being generated by the genome sequencing projects. This makes them an invaluable tool for modern biology.

1.3 Semi-specific methods for the identification of genes.

Semi-specific methods are also available for the identification of novel genes. Selection depends on ascertained qualities of individual data. This can be a pattern within the information. For example a motif within DNA which represents a conserved functional domain of a protein, such as a DNA binding motif. Alternatively it can take the form of other types of information, for example expression analysis.

1.3.1 Heterologous screening.

As mentioned above, patterns within the sequence of both RNA and DNA can be used to identify interesting anonymous genes. There are many examples of the use of conserved sequence, which relates to conserved functional domains in proteins. These can be used to identify new members of gene families by way of heterologous screening where similar sequences are picked up in library screens by using relaxed screening conditions.

1.3.1.1 Using heterologous sequences as probes.

An example of this is the homeobox containing genes (reviewed in Gehring, 1987) which are found in eukaryotic organisms (Holland and Hogan, 1986). A homeobox is a 60 amino acid domain which forms a helix-loop-helix structure which mediates sequence specific binding to DNA (Gehring, 1987). The homeobox was first identified in *Drosophila* where mutations in genes which contained homeoboxes were found to be the cause of

developmental (homeotic) mutants (Lewis, 1978). Heterologous screening found more genes in *Drosophila* which contained this homeobox. Homeotic genes are organised in two clusters and the physical order of these genes corresponds to the order in which they are expressed along the anterior-posterior axis of the embryo during development (Harding *et al*, 1985). The most 3' gene having the most anterior expression limit. By further heterologous screening homeobox containing genes have also been found in other organisms which possess a segmented body pattern, (*e.g.* mouse, Duboule *et al*, 1986; Human, Boncinelli *et al*, 1988) where they are similarly organised into clusters and exhibit the cluster position effect described above both in body pattern formation (Graham *et al*, 1989) and further elaborations in structures which develop after the body pattern, *e.g.* hind brain (Wilkinson and Krumlauf, 1990). Heterologous screening does have its limitations. For example it can only be used for the identification of new members of a gene family. It cannot be used to identify new classes of genes which may contain un-described protein domains.

1.3.1.2 Degenerate PCR.

Another method of heterologous screening has been used successfully; it is PCR based and relies on the design of degenerate oligonucleotides which are used as primers to drive amplification of new sequences between the primer pairs. Cloning and sequence analysis is used to establish if the sequences amplified are those from the original sequences or those which are derived from novel members of that particular class of genes. For instance, additional members of specific subgroups of the G protein linked seven trans-

membrane domain receptor super-family have been identified by this method (Libert *et al*, 1989).

1.3.2 Insertional mutagenesis.

1.3.2.1 Enhancer traps.

Another technique which has been used is the introduction of reporter genes into the genetic material of an organism. The reporter gene product is expressed under the control of endogenous regulatory sequences and therefore acts as an assay of activity of regulatory sequences in the organism; the reporter gene expression pattern partially or wholly reflecting the expression of the endogenous genes which are under the control of the regulatory sequences. This approach was first applied to the fruit fly, *Drosophila melanogaster* (O'Kane and Gerhing, 1987). A modified transposon (P-element) was used as the backbone for an integration vector. It also contained a β-galactosidase gene under the influence of a weak *Drosophila* promoter. In the presence of an active transposase, this construct moved to other sites in the *Drosophila* genome, where the β-galactosidase gene was expressed under the influence of regulatory sequences. This construct was termed an enhancer trap. A large number of independent transposition events have been isolated in an attempt to saturate the *Drosophila* genome and characterise the expression of genes.

Initial success in this organism lead other groups to extend the use of the enhancer trap vector approach to other organisms (*C. elegans*; Hope, 1991; Zebrafish, Westerfield *et al*, 1990) including mouse (Gossler *et al*, 1989; Korn *et al*, 1991). Two alternative

approaches to the construction of enhancer trap vectors were used in the mouse. The first construct was based on mammalian retrotransposons (Freidrich and Soriano, 1991). However like the P-element vector used in *Drosophila*, these retrotransposons exhibit a non-random integration pattern and are biased towards sequences which are transcriptionally active at the time of integration. This is restrictive because not all genes are actively expressed at any one time in the cells which are transformed in these experiments (*e.g.* oocytes or murine ES cells).

The second approach was to introduce enhancer traps directly into the genome of embryonic stem (ES) cells (Gossler *et al*, 1989: Allen *et al*, 1990). Stable integration of the constructs was made possible by the addition of a positive selectable marker under the influence of a constitutive promoter. For example, the construct used by Gossler *et al* (1989), contains a β-galactosidase reporter gene linked to a minimal promoter from the heat shock protein gene 68 (hsp68) and the Tn5 neomycin phosphotransferase gene as a selectable marker, whose expression is driven by the mouse thymidine kinase promoter.

Using this construct, Korn *et al* (1992) targeted 59 ES cell lines and detected reporter gene activity in 13 of them. Six of these showed a spatio-temporal restriction in expression pattern in chimaeric embryos. For non-staining targeted cell lines, only 9 out of 46 showed changes in pattern of expression when introduced into donor blastocysts and allowed to contribute to chimaeric embryos. From some of these cell lines which displayed a spatio-temporal pattern of reporter gene expression in chimaeric embryos, putative regulatory genes have been identified. For example ETL (enhancer trap locus)-1 (Soininen *et al*, 1992) a gene which was found in the

vicinity of the enhancer trap insertion site of ES cell line D6-028 (Korn *et al*, 1991), shows homology to the *Drosophila* gene, Brahma (Tamkun *et al*, 1992), which controls homeobox expression. It also has similarity to the RAD54 gene of yeast (Emery *et al*, 1991).

1.3.2.2 Other types of insertional reporter constructs.

Other types of vectors have been developed because the gene(s) which are the targets of the endogenous regulatory elements detected by an enhancer trap vector may be at some distance from the integration site. These constructs are known as gene traps (*e.g.* Gossler *at al*, 1989) or promoter traps (*e.g.* Vonmelchner and Ruley, 1989) as activation of the reporter gene relies on the insertion (of the construct) to be within the gene or the promoter sequences immediately upstream of the endogenous gene. This makes them more efficient for the identification of gene sequences. Gene traps are distinguished from promoter traps in that they have a 3' splice acceptor at the 5' end of the reporter gene. These constructs work by the splicing of the transgene transcript into frame with the endogenous gene sequences to create a hybrid construct which is translated into a chimaeric protein which has reporter protein activity and is expressed in the tissues where the endogenous protein is expressed, thereby reporting the expression profile of the endogenous gene. Promoter traps do not have a splice acceptor and rely solely on integration near endogenous promoters to drive expression of the marker gene.

1.4 A new direction.

There are other sequences which are found in the coding sequence of genes which do not correspond to specific protein domains with defined function *e.g.* amino acid homopeptides. These occur in a variety of protein types, but have been associated with several general functions. There is a bias in the types of proteins in which they occur but this is independent of other specific protein domains. This quality could allow them to be used in a general way to isolate new proteins with unknown structural motifs. For example it is been suggested that certain homopeptides of amino acids have generalised function. In particular polyglutamine has been proposed to be involved in protein-protein interactions by adopting specific structures (Perutz *et al*, 1994). Polyglutamine has also been noted to occur in transcription factors (Gerber *et al*, 1994; Karlin and Burge, 1996). These transcription factors have a variety of binding domains (*e.g.* POU binding domain, Brain 2; zinc finger DNA binding domain; TATA binding proteins; steroid receptors; androgen receptor, glucocorticoid receptor; homeobox domain proteins).

Trinucleotide repeats are the simplest nucleic acid sequence that can be translated into homopeptides in proteins. Therefore probes based on trinucleotide repeats may be used to identify genes which contain homopeptides. There are other reasons which make trinucleotide repeats attractive structures for further investigation. These will be discussed in the rest of the introduction. A screen for genes containing trinucleotide repeats of the mouse may lead to the identification of novel regulatory genes with functional roles for these repeats both as nucleic acid and/or as homopeptides.

1.5 What are trinucleotide repeats?

Trinucleotide Repeats are a subclass of simple sequences and are small tandem repeats of three nucleic bases in constant order. They belong to a class of repetitive DNA, microsatellites (Weber and May, 1989), that are widely dispersed throughout the genomes of eukaryotes. Some classes of repeats have been observed to be over-represented in the genomes of certain organisms *e.g.* both CAG and CGG trinucleotide repeats occur at a greater than expected number in the genome of humans (Han *et al*, 1994).

Microsatellites are also characterised by their ability to change copy number, usually by increments of the basic unit (three nucleotides in the case of trinucleotides). This has made them invaluable as markers for mapping purposes (*e.g.* Sheffield *et al*, 1995). they also have been used to create enriched libraries of clones from eukaryotes bearing repeats of a particular class (*e.g.* Ostrander *et al*, 1992).

More recently trinucleotides have been implicated in a novel group of diseases, those associated with dynamic mutations. Trinucleotide repeats that occur in coding regions give rise to the simplest repeat possible in proteins, homopeptides, which may have functions associated with them.

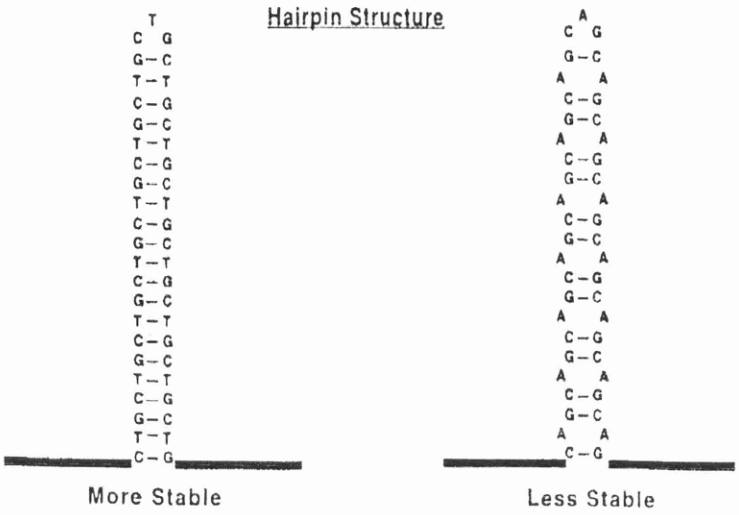
Recent research shows that organisms have used repetitive motifs of nucleic acid in their genes as useful functional patterns. This occurs both at the nucleic acid (both in DNA and RNA) and at the protein level. The following section will discuss that, in higher eukaryotes and mammals in particular, triplet repeats have been used for a variety of functions.

Legend for Figure 1.2 Models for secondary structures formed by trinucleotide repeats.

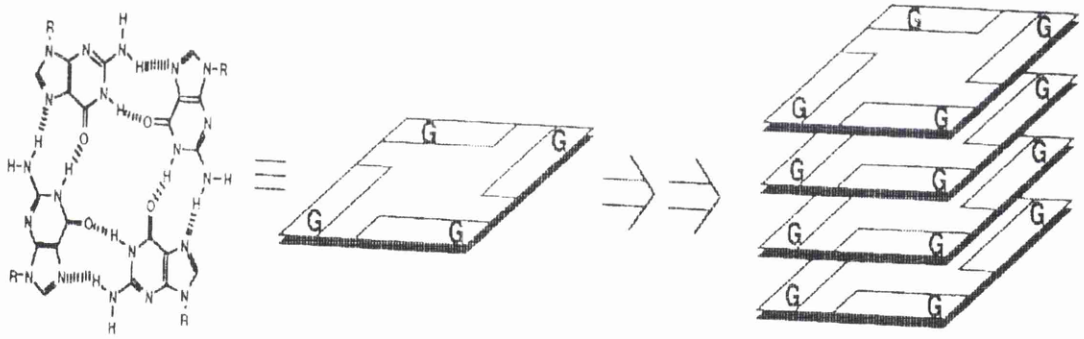
A) Simplified version of hairpin structures that are possible with the CAG/CTG trinucleotide repeat (adapted from Kang *et al*, 1995). B) An example of a quadruplex structure involving bonding between four guanidine residues and how they may stack up on themselves to form the quadruplex (adapted from Williamson, 1993). C) Representation of hydrogen bonding base pairs for AT and CG as proposed by Watson and Crick.

Figure 1.2

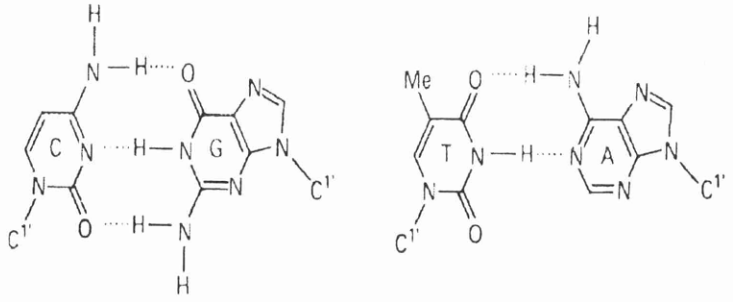
A



B



C



1.6 Trinucleotide repeats in nucleic acids.

1.6.1 Structures that trinucleotide repeats can adopt.

Nucleic acid trinucleotide repeats can adopt structures, other than the double-stranded anti-parallel helix described by Watson and Crick (Watson and Crick, 1953; Crick and Watson, 1954; see Figure 1.2.C). They can form pseudo-helical structures like hairpins (Figure 1.2.A). This hypothesis has been supported by *in vitro* evidence reported by various groups. For example, Gacy *et al* (1995) described the formation of stable hairpins by CG rich trinucleotide repeats (CTG, CAG and CCG). According to them, the stability of these hairpins is dependent on both the length and sequence of the repeat. Imperfections in these repeats de-stabilise these structures. Pearson and Sinden (1996), using CTG and CGG trinucleotide repeats, observed complexes with reduced mobility in polyacrylamide gel electrophoresis experiments. The complexes appeared to be stable at physiological strengths and have a melting temperature of 55°C. They also appeared to contain single-stranded DNA as they are in part sensitive to mung bean nuclease.

Indirect evidence that repeats form hairpins *in vivo* comes from the work of Darlow and Leach (1995), who inserted even and odd number copies of a trinucleotide oligonucleotide into a position in the lambda phage where the formation of a hairpin is necessary for the inhibition of plaque formation. Plaque inhibition was found with even numbers of CAG.CTG and CGG.CCG but not with GAC.GTC inserts. This suggests that not all types of trinucleotide have the potential to form hairpins structures.

What is happening at the molecular level to cause the formation of hairpins? One suggestion is that the nucleotide chains may bend back on themselves to form a stable hydrogen bonding pattern between C and G bases (Smith *et al*, 1995; see also Figure 1.2). Smith *et al*, (1995) propose that CTG repeats are more stable than CAG repeats, because of the difference in size of the middle base (see Figure 1.2.A). On both sides of the central base, two CG base pairs can form. Since thymidine is smaller than adenine, it does not interfere as much with the neighbouring GC hydrogen bonding. This may cause infidelities in replication by a slippage mechanism (Levison and Gutman, 1987; see Figure 1.3). This model has also been proposed to explain the variability observed in microsatellite length, hairpin formation would cause the template or newly synthesised DNA to detach from the partner strand with reassociation at a non-cognate position in the repeat sequence, leading to either shortening or lengthening of the repetitive DNA stretch. However there are problems with this model: not all microsatellites form hairpins (*e.g.* see Darlow and Leach, 1995) and it has been recently shown that other trinucleotide repeats exhibit dynamic instability (GAA trinucleotide repeats in Friedreich's ataxia, Campuzano *et al*, 1996)

Other unusual base pairings have also been implicated in the formation of trinucleotide repeats hairpins, like G:G base pairs. Mitas *et al* (1995) suggested that G(syn).G(anti) base pairs are formed by CGG repeats which are associated with the fragile X sites.

Another structure has been suspected to be involved in the reduced electrophoretic mobility of trinucleotide repeats; there is a suggestion that CGG trinucleotide repeats participate in quadruplex (four-stranded) DNA structures (Fry and Loeb, 1994; Smith *et al*,

1994; see Figure 1.2.B), as the properties of these complexes are consistent with quadruplex structures. Fry and Loeb (1994) observe slower than expected migrating species when $(CGG)_n$:$(CCG)_n$ double-stranded duplexes are electrophoresed. Interestingly the appearance of these slowly migrating species is dependent on the methylation state of the cytosine residue, but only in smaller oligonucleotides. For instance, methylation dependence is observed where the repeat number is five (*i.e.* $(CGG)_5$:$(CCG)_5$), but there is no effect when the repeat number is seven. Methylation involvement is further implicated by the observation that treatment with dimethyl sulphoxide (which inhibits methylation by modification of the N7 position on the pyrimidine ring of cytosine) inhibited complex formation. Smith *et al* (1994) also observed these complexes when using 30mer nucleotides (where the repeat number is 10), they also suggest that there is a secondary structure requirement, as substitution of guanine for ionosine abolishes complex formation.

However, further experiments need to be done to characterise the structures trinucleotide repeats can and cannot adopt. This must include further *in vivo* work, specifically in a mammalian system. This is necessary to understand their importance to human triplet expansion diseases. This is important because recent research has discovered that another repeat, which has hitherto been unsuspected of involvement in secondary structure formation, has been found to be the cause of dynamic disease, which are

Figure 1.3

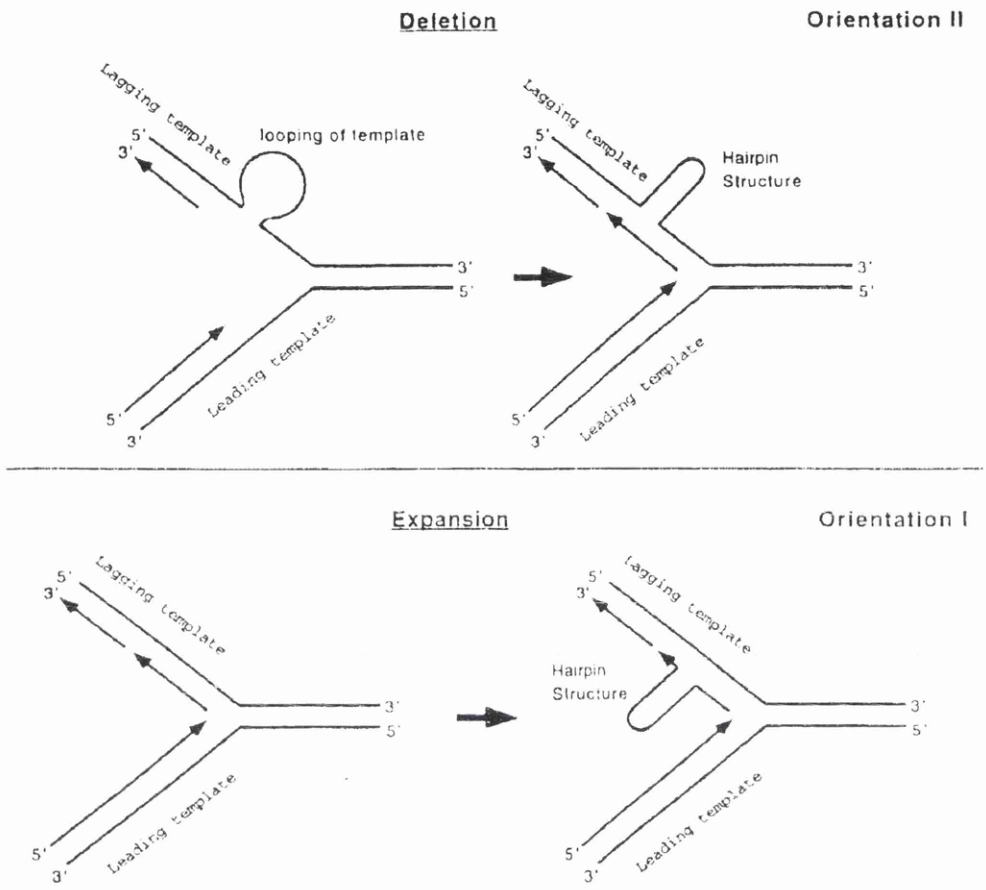


Figure 1.3 A model for deletions and insertions at the DNA synthesis replication fork, involving hairpin structures. Hairpin structures are proposed to form at the replication fork on the lagging strand at the DNA replication fork. This leads to either deletion, where hairpin formation occurs on the template DNA strand (see orientation II) or insertion (or expansion) where they occur on the newly synthesised DNA (see orientation I). This figure was adapted from Kang *et al*, (1995).

characterised by an unstable trinucleotide repeat. A GAA repeat has been found to be implicated in Freidriech's Ataxia (Campuzano *et al*, 1996).

1.6.2 Repeat dynamics.

Several ideas have been put forward to explain the precise mechanism of expansion of GC rich trinucleotide repeats. Ohta and Kimura (1973) describe how trinucleotide repeats (and other repeats) could increase or decrease in length by single trinucleotide units by mis-aligment of the template with the DNA polymerase complex. Streisinger *et al*, (1966) proposed a mechanism of expansion that involves replication slippage in which Okazaki fragments disassociate from the template DNA strand and reassociate at a position different from the original (see Figure 1.3). Crucial to this hypothesis is the formation of hairpin secondary structure in the Okazaki fragments. Experimental evidence (Smith *et al*, 1995) using NMR spectroscopy supports this idea. Kang *et al* (1995) have proposed that this type of hairpin structure could occur in both RNA and DNA to give scenarios for both expansion and deletion events.

All above hypotheses propose mechanisms which can act both ways to create extra triplet repeats (expansion) or to remove them (deletion). However all the human dynamic diseases appear to be due to an overall increase in numbers of trinucleotide repeat units, which suggests that other factors are influencing the fluctuations in repeat number. Kang *et al* (1995) observed expansion and deletions of trinucleotide repeats (those involved in the human dynamic diseases) in *E. coli* that was dependent on the orientation of CTG

repeats relative to the direction of replication. Large units of a trinucleotide repeat were deleted in an orientation where CTG triplets were on the lagging strand relative to the origin of replication of the plasmid. In an analysis of individual bacteria from one culture containing a plasmid with 180 CTG trinucleotide repeat units, it was observed that the deletions had a periodicity of 40 repeats (examples of 180 (un-deleted) 140, 100 and 60 CTG trinucleotide repeat units). When the CTG triplets were in the leading strand, increases in repeat size were observed.

1.7 Proteins that bind repeats in DNA.

Proteins have been discovered which specifically bind repetitive DNA in chromosomes. An example of this is telomeres, which are the physical ends of a chromosome. They are fashioned from tandem hexanucleotide repeats, $(TTAGGG)_n$ (Van Der Ploeg *et al*, 1984), which are conserved throughout higher eukaryotes (Moyzis *et al*, 1988). The sequences were first identified in trypanosomes (Van Der Ploeg *et al*, 1984) and have been found to bind specific proteins. These proteins are distinct from other proteins which exhibit non-specific nucleic acid sequence binding. This indicates a specific role for both the tandem repeats and the proteins which bind to them. The proteins act as a bridge for interactions between chromosomes and the cyto-skeletal structures (*e.g.* Intermediate Filaments; see Traub, 1995 for a review.) which are formed after the chromosomes have been replicated, in preparation for cell division. Telomeres also adopt an unusual secondary structure. They form a quadruplex in a similar fashion as described above for the CGG trinucleotide repeats associated with

human diseases (for review see Brahmachari *et al*, 1995; Figure 1.2 and Section 1.6.1).

1.7.1 Proteins that bind DNA trinucleotide repeats.

Recent work has identified several proteins that interact in a specific fashion with both single and double-stranded DNA trinucleotide repeats. Timchenko *et al* (1996) and Yano-Yanagisawa *et al* (1995) have biochemically characterised individual protein activities that bind to single-stranded CAG and CTG repeats. Timchenko *et al* (1996) used a radio-labelled DNA oligonucleotide ss$(CTG)_8$, in conjunction with HeLa cell protein extract, in a band shift assay experiment, to show the formation of two complexes, ssI (major) and ssII (minor). These were shown to be specific by using unlabelled oligonucleotide competitor; in their presence these complexes were abolished. Challenging the reaction with unlabelled ss$(CAG)_8$ and ss$(CGG)_8$ oligonucleotides had no effect on complex formation. They also found specific protein-DNA interactions with ss$(CAG)_8$ and ss$(CGG)_8$ DNA oligonucleotides. Four complexes and two complexes appeared when using labelled ss$(CAG)_8$ and ss$(CGG)_8$ oligonucleotides, respectively. Comparison of mobilities of the complexes formed by the three oligonucleotides, ss$(CTG)_8$, ss$(CAG)_8$ and ss$(CCG)_8$, lead Timchenko *et al* (1996) to conclude that the major CTG binding complex (ssI) was specific to CTG, and they termed it single-stranded CTG Repeat Recognising Protein (ssCRRP). This binding activity contains at least two polypeptides of differing size because fractionation by size of the HeLa whole cell protein extract abolishes the formation of the complex. Binding activity was

re-established by combining two fractions (4 and 5) of the cell protein together.

In similar experiments, using HeLa nuclear extract as a protein source for oligonucleotide binding assays, Richards *et al* (1993) found a protein activity that interacted specifically with a double-stranded 30mer DNA oligonucleotide, $(CCG)_{10}$. This was competitively inhibited in the presence of cold $(CCG)_{10}$ oligonucleotide but not when cold competitor oligonucleotides representing the other 9 families of trinucleotide repeats were used. This complex was unaffected when two dinucleotide oligonucleotides were used in competitive assays. Interestingly, this complex is destabilised by methylation of the repeat at cytosine residues in CpG dinucleotides. The oligonucleotide $p(CCG)_{10}.p(GGC)_{10}$ was methylated enzymatically by CpG methylase. In band shift assays, the CCG.BP1 complex (which consists of the double-stranded un-methylated oligonucleotide, $p(CCG)_{10}.p(GGC)_{10}$ and sequestered cellular proteins) is not formed, it is replaced by a new higher molecular weight complex that is abolished not only by unlabelled methylated $p(CGG)_{10}.p(GCC)_{10}$, but also by other oligonucleotides containing CpG dinucleotides which were methylase treated *e.g.* $p(ACG)_{10}.p(TGC)_{10}$). Richards *et al* (1993) suggest that this may be the previously described protein, MeCP1 (Meehan *et al*, 1990), which is a sequence non-specific methyl-CpG binding protein. Richards *et al* (1993) also took the other oligonucleotides representing the tri and di-nucleotide families and used them to identify complexes that were specific to other double-stranded DNA trinucleotide and dinucleotide repeats. Some of these complexes have overlapping binding activity with respect to other

oligonucleotides which represent different families of trinucleotide repeat.

As mentioned by Richards *et al* (1993), there are proteins which bind to trinucleotide repeats, as part of a broader spectrum of binding to nucleic acid. The example described by Richards *et al* (1993), is MeCP1, which has a high affinity for methylated CpG pairs. Other proteins have binding preferences CpG dinucleotides. Another protein which has an affinity for (hemi-methylated) CpG dinucleotides is the protein DNA methyltransferase which is required for the maintenance of methylation of cytosines in CpG dinucleotides (Taylor and Jones, 1982). It maintains methylation by the recognition of hemi-methylated double-stranded CpG dinucleotides, which are the product of semi-conservative DNA replication, in which one strand is newly synthesised and does not contain methylated cytosines but the parental strand retains its methylated cytosine status.

CpG dinucleotides are not randomly distributed throughout the genomes of vertebrates. Many CpG dinucleotides are concentrated into regions which are known as CpG islands (for a review see Meehan *et al*, 1992). It has been shown that these islands are associated with genes and are known to influence the activity of the gene, dependent on the methylation status of cytosines. The importance of this covalent modification of cytosine is that it thought to either inhibit the binding of proteins required for transcription, recruit factors which inhibit transcription (Meehan *et al*, 1992), or even induce a regional conformational shift in DNA structure from B form to Z form (Meehan *et al*, 1992). It has been shown that DNA helices which adopt a Z conformation, repress transcriptional activity of genes. It is proposed that this occurs by

the inability of the transcriptional machinery to access the appropriate sites to initiate transcription. This as well as the other theories have been invoked to explain the phenomenon of imprinting (for review see Monk, 1995), where there is only one active copy of a gene, which is consistently inherited from one parental sex.

More recently, this mechanism of gene silencing has been used to explain the Fragile X syndrome of humans. These are characterised by the expansion of CGG and GCC trinucleotide repeats (CGG for FRAXA and FRAXE; Fu *et al*, 1991, Knight *et al*, 1993). The common denominator of these trinucleotide repeats is that they contain CpG dinucleotides. The upward expansion of these repeats, which have been associated with these diseases, is proposed to create a new CpG island which is identified by the methylation machinery of the cell, which acts to methylate the cytosines and leads to a regional inaccessibility and down regulation of gene expression. This leads to an apparent loss of function of a particular gene or genes in the region. In case of FRAXA, the gene affected is *FMR1*. The product of this gene is involved in RNA binding (Siomi *et al*, 1993). In support of the loss of function model, certain individuals have been found to carry point mutations in the *FMR1* gene which leads to a loss of activity of the FMR1 protein (*e.g.* Deboulle *et al*, 1993) and they have similar phenotypes as those with expansions of repeats. Mouse *fmr1* gene knockouts confirmed this hypothesis (Bakker *et al*, 1994). Candidates for the recognition of expanded repeats could be the proteins that participate in the DNA-protein complexes, described by Richards *et al* (1993).

A functional role for repetitive DNA has recently been demonstrated by Wang and Griffith (1995), who showed that long

tandem repeats of the trinucleotide CTG, have a strong effect on the positioning of nucleosomes in DNA. This suggests that the CTG repeat has a structural advantage in promoting interactions with the constitutive proteins of the nucleosome, thereby altering local chromatin structure. This could have an effect on the accessibility of promoter sequences to the cellular transcriptional machinery. Alternatively, other proteins as suggested by Richards *et al* (1993), Timchenko *et al* (1996) and Yano-Yanagisawa *et al*, (1995) may be allowed access to the repeat.

1.8 Proteins that bind repeats in RNA.

Proteins that bind RNA trinucleotide repeats have also been described recently. Timchenko *et al* (1996) found two activities in HeLa cell whole cell extracts that bound to an RNA oligonucleotide, ss(CUG)$_8$. The two complexes were inhibited by excess, unlabelled ss(CUG)$_8$ oligonucleotide. The smaller complex had a similar mobility to that of the ssCRRP complex that exhibited binding to the DNA oligonucleotide ss(CTG)$_8$. This RNA binding was abolished by the addition of ss(CTG)$_8$ DNA oligonucleotide. In contrast, the larger RNA-protein complex was unaffected by the addition of the DNA oligonucleotide or cold ss(CGG) RNA oligonucleotide. Timchenko *et al* (1996) concluded that it exhibited a sequence specific RNA binding specificity and termed it CUG-BP (CUG Binding Protein). CUG-BP protein, appears to be found predominantly in the cytoplasmic fraction of HeLa cells. ssCRRP DNA binding activity was also found to be predominantly cytoplasmic, although some binding activity could be detected in the nuclear extract. CUG-BP was also found in other

cells types; for example, in extracts of fibroblasts and myotubes (Timchenko *et al*, 1996).

The conclusions from the studies described in Sections 1.7 and 1.8 are as follows. 1) There are proteins which have a natural specific affinity for both single-stranded (RNA and DNA) and double-stranded (DNA) triplet repeats; CUG, CTG and CAG specific repeat binding proteins have been observed. 2) Although these are specific, other proteins can also bind to trinucleotide repeats non-specifically. 3) Since Yano-Yanagisawa *et al* (1995) found proteins in the mouse brain that bound single-stranded DNA repeats and Timchenko *et al* (1996), focusing on human cell lines, found these type of activities, it is probable that the human proteins have functional homologues in mouse and vice versa.

1.9 Homopeptides in proteins.

The basic unit of information in the genetic code is the codon. This links the information held in nucleic acids to the protein gene products. The codon is three nucleotides in length and determines the amino acid content of peptides. Homopeptides represent the simplest repetitive pattern amino acids can adopt in proteins. They can also be accommodated in existing genes without disturbing the frame in which the protein is read from the messenger RNA molecule. Increments of trinucleotides are gained or lost in a single step, thus conserving the frame on both sides of the change. However, not all homopeptides are represented in proteins. Some homopeptides have functions within proteins whereas others have not or have a negative effect on the function of the protein.

Several amino acids are encoded by more than one codon that does not belong to the same family. For example CAG encodes glutamine and is a member of the repeat family which will be the focus of the experimental work. However, CAA also encodes glutamine but these two codons do not belong to the same group of trinucleotide on the basis of frame and complementarity. Homopeptides of glutamine can be therefore either be encoded by repeats of CAG or CAA, or even a mixture of the two codons. However, Green and Wang (1994) note from a database survey, that glutamines in repeat stretches are more likely to be encoded by CAG.

Homopeptides observed in proteins can be classified according to which type of side chain the amino acid contains. The largest represented group is the one comprising uncharged polar side chains. According to Karlin and Ghandour (1985) and Karlin and Burge (1996), this group consists of glutamine (Q), asparagine (N), serine (S), threonine (T), proline (P) and histidine (H). The second group consists of small aliphatic amino acids: alanine (A) and glycine (G). The third group consisting of aspartate (D) and glutamate (E) which have an acidic side chain. Karlin and Burge (1996) also observed rare examples of leucine (L) repeats, the only amino acid homopeptide to have a large aliphatic side chain.

1.9.1 Putative functions of homopeptides.

Various ideas about what function these homopeptides may perform in proteins have been proposed. Some are general in respect to amino acid homopeptide whereas others rely on the nature of certain amino acid side chains.

The first hypothesis put forward to explain the occurrence of these homopeptides was that they were spacer domains between functional domains of proteins. Their role would be to separate the critical regions of proteins by an appropriate distance, enabling the other domains to interact in meaningful way. This may account for the occurrence of homopeptides consisting of amino acids with side chains that are believed to be non-reactive and non-polar, *i.e.* alanine and glycine.

Another hypothesis is that they act as interactive domains between proteins. This could explain the runs of acidic residues in general (Mitchell and Tjian, 1989) *i.e.* imperfect or perfect stretches of either glutamate (D) or aspartate (E). The cumulative charge of these regions is believed to facilitate interaction with basic or polar regions in other polypeptides (Sigler, 1988).

Polar amino acid homopeptides can also participate in protein-protein interactions. Serine and threonine can exist both as mixtures or pure homopeptides that form strong polar regions (from accumulative polarity), which could interact with other sections of the polypeptide or other proteins.

1.9.2 Polyglutamine and polyproline.

These represent the two most studied homopeptides with respect to their possible influence on the function of proteins. They also appear to be relatively more abundant than other classes of homopeptides. They are discussed below, relative to these observations.

1.9.2.1 Polyglutamine

Polyglutamine is the most commonly found homopeptide in eukaryotic proteins in reference databases (Green and Wang, 1994) and has been implicated in the facilitation of protein-protein interactions. Green and Wang (1994) also noted that the most common trinucleotide encoding glutamine was CAG. Polyglutamine tracts are the best studied so far, not only because of their function but also because they have been associated with a novel type of human disease, the dynamic mutation diseases (Ross, 1995). Aspects of these diseases will be discussed later.

There is experimental evidence of the functional importance of polyglutamine tracts (Gerber *et al*, 1994). As a prelude to direct experimentation, Gerber *et al*, (1994) conducted searches through the protein database, Swiss-Prot using a polyglutamine sequence (twenty residues, Q20) and the FASTA Programme (Lipman and Pearson, 1985). They observed that a significant proportion of the proteins that contained glutamine homopeptides were transcription factors. From the 40 entries with the highest scores, 82% were transcription factors and 17 were *Drosophila melanogaster* proteins.

Extensive developmental studies in this organism have identified many transcription factors and other developmental control genes (Karlin and Burge, 1996). The repeats are also known in *Drosophila* as *OPA* repeats (Maginnis *et al*, 1984). They were first observed in *Notch*, a neurogenic gene. In a study conducted by Wharton *et al* (1985), the authors also observed a complex hybridisation pattern on Southern analysis with a fragment containing the repeat. They concluded that in the haploid genome of *Drosophila*, there were about 500 individual *OPA* repeats. A probe

containing the repeat also hybridised to multiple poly-adenylated messenger RNA species.

The most abundant type of amino acid encoded by *OPA* repeats in *Drosophila* is glutamine. In this case the third base is either G or A, resulting in the two codons CAG and CAA. These glutamine repeats can be interspersed with the amino acid histidine, which is encoded by one of two codons which are CAC or CAT. Progress has been slower in mammals and proportionally fewer developmental control genes have been identified. It is likely that many more mammalian proteins containing polyglutamine homopeptides remain to be discovered. Li *et al*, (1993), screened a human cerebral cortex cDNA library with a $(CTG)_{10}$ 30mer oligonucleotide and found that of 50000 clones screened with a $(CTG)_{10}$ radioactively labelled probe, 0.28% contained CAG/CTG trinucleotide repeats. Of those which were sequenced and an open reading frame identified (six), four contained putative polyglutamine tracts.

Gerber *et al* (1994) showed that the presence of polyglutamine tracts has a positive effect on activity using *in vitro* transcription complementation experiments (Gerber *et al*, 1992). A series of constructs, containing glutamine stretches of various lengths inserted between the GAL4 DNA binding domain (Giniger *et al*, 1985) and the protein-protein transactivation domain of herpes virus protein, VP16 (Sadowski *et al*, 1988) were generated. The cognate proteins were produced in HeLa cells and were co-precipitated with a reporter construct containing a GAL4 binding site and a reporter gene. Activity was measured by the ability to produce *de novo* transcription from a promoter on the reporter construct. The HeLa cell nuclear extract also provided other proteins

Figure 1.4

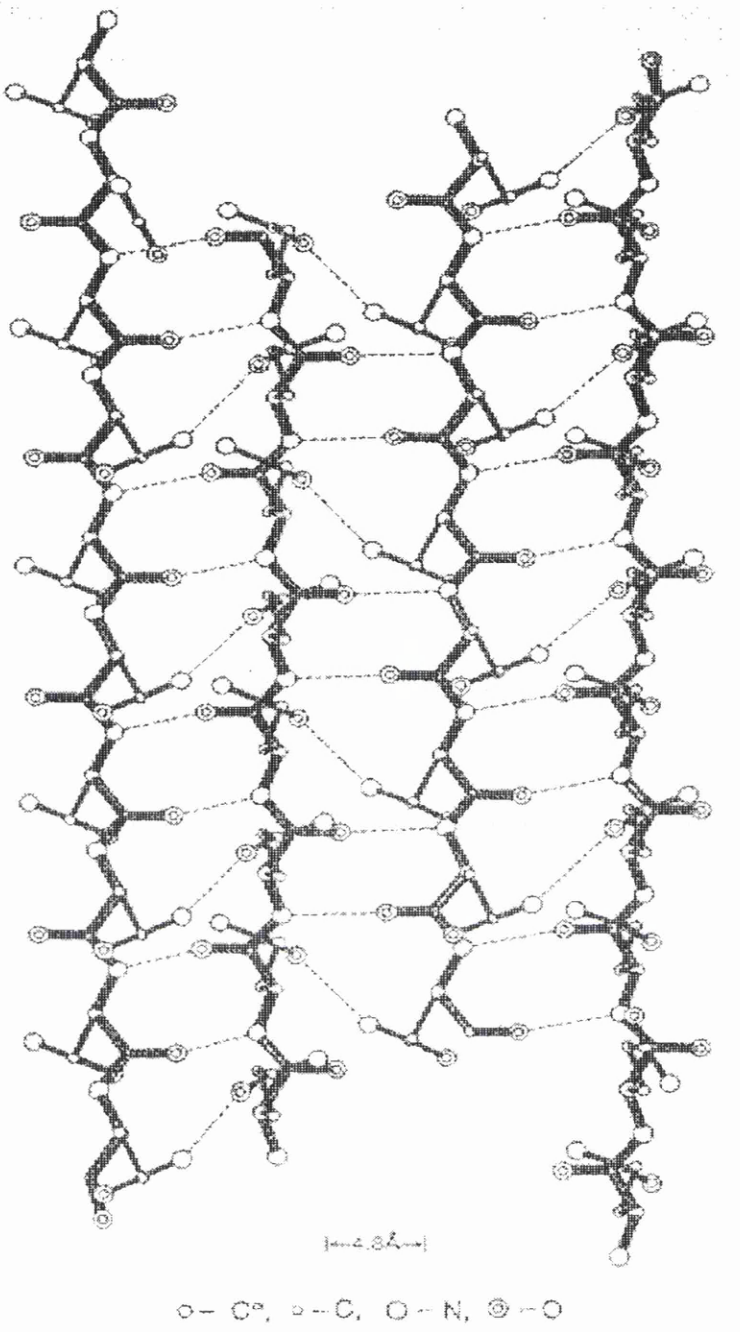


Figure 1.4 Proposed structure of anti-parallel β-sheet adopted by polyglutamine.

Two paired anti-parallel β-strands of poly (L-glutamine) linked together by hydrogen bonds between the main chain and side chain amides are shown. C, carbon, N, nitrogen and O, oxygen. The diagram was adapted from Perutz *et al*, (1994).

involved in transcription. These experiments show that there is a stimulation of basal reporter gene activity with the addition of between 10 and 40 glutamines. Beyond 40 glutamines, there is no observable increase in transcriptional activity. Glutamines alone with a GAL4 DNA binding domain induced a decrease in transcriptional activity if the copy number exceeds 40 residues. Control constructs with glutamate-leucine-glutamine (ELQ)n, or glutamine-glutamine-serine (QQS) repeats showed a weaker response, and activity was independent of the number of repeat units. In a third set of experiments various glutamine rich regions of known human proteins, Oct2 (Clerc *et al*, 1988), Sp1 (Courey and Tjian, 1988 and Courey *et al*, 1989), were fused to the DNA binding domain of GAL4 and they showed stimulation of transcriptional activity of the reporter gene.

However, an increase in length of (and in numbers of residues contained within) a glutamine homopeptide does not always induce an increase in transcriptional activity. For instance, Chamberlain *et al* (1994) found that the deletion of a glutamine stretch in the androgen receptor of humans lead to an increase in transcriptional activity and that the expansion of the glutamine repeat decreases the transcriptional-stimulating activity of the protein. Their suggestion is that the positioning of the glutamine tract relative to other functional regions of proteins may also be important. For example, the glutamine homopeptide in rat and human androgen receptors is different. In the human protein, the polyglutamine tract is upstream of the activation domain. In the rat protein it is juxtaposed to the activation domain and deletion has little functional effect. It therefore appears that polyglutamine function may be context dependent, *i.e.* the influence of polyglutamine is

dependent either on its position in that protein, relative to other critical domains or perhaps on the target molecules with which it interacts.

Recently, polyglutamine stretches have been proposed to take up a defined secondary structure (Perutz *et al*, 1994) which is an anti-parallel beta sheet structure called a polar zipper (Perutz *et al*, 1993: see also Figure 1.4). This occurs by the formation of hydrogen bonds between the main chain and side chain amide groups. Polar zippers are believed to be one way in which proteins can interact with each other. Further to this, Stott *et al* (1995) have experimentally shown that the insertion (or substitution of existing residues) of a glutamine homopeptide, ten residues in length, into a protein (CI2, Chymotrypsin Inhibitor 2; McPhalen *et al*, 1985) promoted the formation of stable dimers and trimers.

1.9.2.2 Polyproline.

Proline homopeptides were first observed in proteins of the fruit fly, *Drosophila melanogaster*. They are encoded by trinucleotide repetitive elements termed PEN repeats (Digan *et al*, 1986). This repeat also encodes other homopeptides, including polyglycine (Haynes *et al*, 1987). These amino acid tracts are encoded by GGN codons. Both polyproline and polyglycine have been noted by Karlin and Burge (1996) to occur in mammalian species, such as rat, mouse and man as well.

Gerber *et al* (1994) also identified a number of proteins from the Swiss-Prot protein database which contain polyproline stretches and observed that a high proportion of them (78%) were transcription factors. The authors showed that medium sized proline

stretches could stimulate transcriptional activity when they were used in an *in vitro* complementation transcription assay; but beyond 10 residues a reduction in activation was observed.

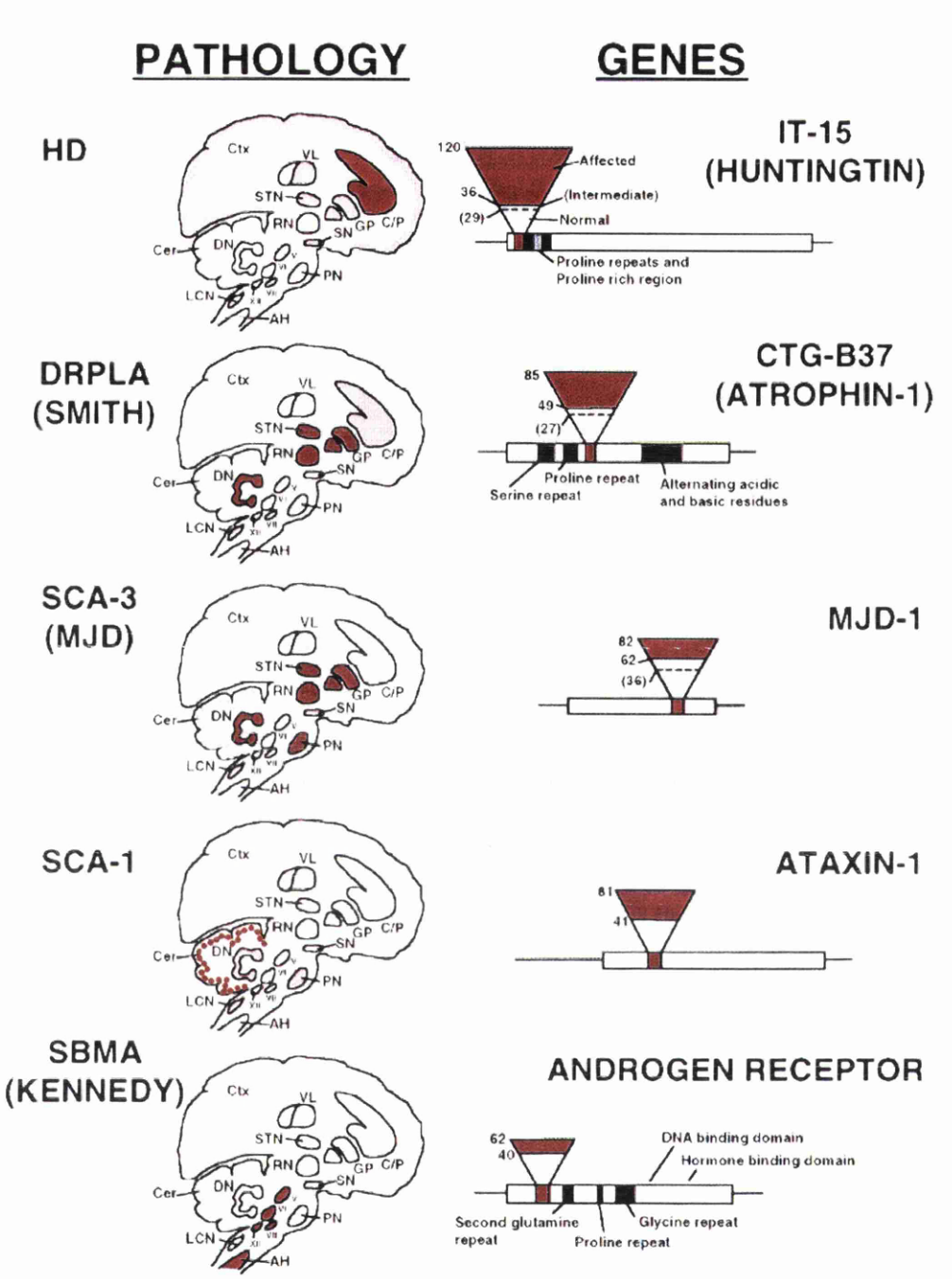
Karlin and Burge (1996) also noted that multiple homopeptides (encoding different amino acids) occur selectively in some genes. They found 68 *Drosophila* proteins, 36 human proteins and 22 mouse proteins with multiple long homopeptides. They suggested that some proteins were more susceptible to the accumulation of repetitive sequences or that, alternatively, these homopeptides were acting in an accumulative manner to add functionality to the proteins.

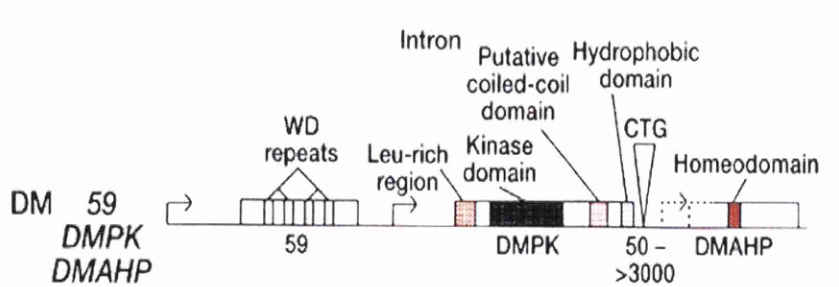
1.10 Diseases associated with CAG/CTG trinucleotide repeats.

A whole new class of genetic mutation leading to disease has been found which is associated with expansion of trinucleotide repeats in humans (for a review see Ashley and Warren, 1995). To date, a number of different trinucleotide repeat expansions have been described; 1) CAG (encoding a polyglutamine in affected proteins), 2) CTG (manifested as CUG RNA triplets), 3) CCG and CGG (which create *de novo* methylation islands and are independent of position with respect to the affected gene) and 4) GAA (present in non-coding intronic DNA). These diseases have been characterised by an upward expansion in the number of triplet repeats with successive generations. This correlates with an increasing severity of the particular disease and a decrease in the age of onset of clinical symptoms specific to that disease. This phenomenon is known as anticipation.

Legend to Figure 1.5 Human diseases which are caused by CAG/CTG trinucleotide repeat expansions.
A) The left hand side indicates the major sites of neuronal degeneration in each diseases. Dark red indicates the site of most severe neuronal loss. Pink indicates areas where the loss of neurones is variable or less pronounced. The circles in the cerebral cortex represent Purkinje cell. The open reading frames of the affected genes in each disease are shown schematically on the right hand side with the glutamine repeats which cause a disease phenotype are indicated in red, with the numbers of glutamines shown at the side. AH, anterior horn; Cer, cerebellar cortex; C/P, caudate/putamen; Ctx, cerebral cortex; DN, dentate nucleus; GP, globus pallidus; LCN, lateral cuneate nucleus; PN, pontine nucleus; RN, red nucleus; SN, substantia nigra; STN, subthalamic nucleus; VL, ventolateral thalamic nucleus; V, VI, VII and XII cranial motor nuclei. B) A representation of the genes DMPK, 59 and DMAHP which lie near the expanded CTG (CUG) triplet repeat which is associated with myotonic dystrophy. Adapted from Ross (1995) and Bailey *et al* (1997).

Figure 1.5


A
PATHOLOGY
GENES
HD
IT-15
(HUNTINGTIN)
120
Affected
36
(Intermediate)
(29)
Normal
Proline repeats and
Proline rich region
DRPLA
(SMITH)
CTG-B37
(ATROPHIN-1)
85
49
(27)
Proline repeat
Serine repeat
Alternating acidic
and basic residues
SCA-3
(MJD)
MJD-1
82
62
(36)
SCA-1
ATAXIN-1
81
41
SBMA
(KENNEDY)
ANDROGEN RECEPTOR
62
40
DNA binding domain
Hormone binding domain
Second glutamine
repeat
Proline repeat
Glycine repeat
Ctx
VL
STN
RN
GP
C/P
SN
Cer
DN
PN
LCN
AH


B
Intron
Putative
coiled-coil
domain
Hydrophobic
domain
WD
repeats
Leu-rich
region
Kinase
domain
CTG
Homeodomain
DM 59
DMPK
DMAHP
59
DMPK
50 –
>3000
DMAHP

The types of repeats which have been associated with CCG (*e.g.* Fragile X, FRAXA) and GAA (Freidreich's Ataxia, Campuzano *et al*, 1996) trinucleotide repeat expansion will not be discussed further in detail, since the focus of this study is the CAG/CTG family of repeats and the pathological mechanism of the disease causation is likely to differ from that of CAG and CTG (CUG) repeats. Furthermore CAG/CTG repeats are more likely to be in the coding regions of genes (Stallings, 1994). The explanation of this over representation of CAG repeats in genes may lie with the positive reasons for the existence of trinucleotide repeats discussed in Sections 1.7 to 1.9. Alternatively it may be because CAG repeats are similar to the splice acceptor tetranucleotide CAGG (Shapiro and Senapathy, 1987), therefore CAG repeats may become excluded from intronic sequence, rather than positive selection. This has been shown in another case, where a minisatellite containing a CAGG tetranucleotide repeat which became incorporated into a human interferon-induced gene by alternative splicing (Turri *et al*, 1995).

1.10.1 CAG repeats, polyglutamine and human diseases.

There are five diseases which are associated with the expansion of CAG trinucleotides that encode polyglutamine: Huntington's Disease (Huntington's Disease Collaborative Research Group, 1993), Spinal Cerebellar Ataxia type 1 (SCA1; Banfi *et al*, 1994), Spinal Cerebellar Ataxia type 3 (SCA3, also known as Machodo-Joseph Disease; Kawaguchi *et al*, 1994), Dentato-Rubral and Pallido-Luysian Atrophy (DRPLA; Koide *et al*, 1994 and Nagafuchi *et al*, 1994) and Spinal and Bulbar Muscular Atrophy (SBMA or Kennedy's Disease; LaSpada *et al*, 1991).

A selective loss of neuronal populations is observed in all these diseases. The set of neurons which are lost is specific to the disease, although some overlap between some of the diseases is observed. For example, in DRPLA (Dentato-Rubral and Pallido-Luysian Atrophy) and MJD (Machodo-Joseph Disease or SCA3), lesions occur in the subthalamic nucleus, globus pallidus, red nucleus and dentatenucleus. The pontine nucleus is the only region to show specific lesion in MJD (SCA3). The regions of lesion for each disease can be found in Figure 1.5 (Ross, 1995).

1.10.2 Polyglutamine pathology.

All these diseases are dominant disorders. They require only one copy of the expanded allele to express the disease phenotype. Gene knockouts in mouse have not replicated the disease pathology (*e.g.* Huntington Disease; Nasir *et al*, 1995), which suggests that these diseases are caused by the gain of a new function by the affected gene product rather than by haplo-insufficiency (*i.e.* a lack of protein with normal function).

The new function could be the result of a modification of the structure of a protein. As discussed earlier in this introduction, polyglutamine stretches are thought to promote protein-protein interactions (Perutz *et al*, 1994). A specific monoclonal antibody made by J.L. Mandel and co-workers (Trottier *et al*, 1995a), which was raised against the glutamine repeat of TBP (TATA binding protein; Trottier *et al*, 1995a), detects specifically mutant proteins with an expanded polyglutamine stretch, but not the normal protein in extracts from patients with HD, SCA1 and SCA3 (MJD). The sensitivity of detection is dependent on the length of the

polyglutamine tract, which suggests that the polyglutamine stretch constitutes a specific epitope and adopts a stable structure in proteins (as suggested by Perutz *et al*, 1994). This antibody has been now used successfully to detect proteins with expanded polyglutamine runs in two other spino-cerebellar ataxias, SCA2 and SCA7 (Trottier *et al*, 1995a), which were expected to be polyglutamine triplet expansion diseases. Recently the existence of a polyglutamine expansion in the gene for SCA2 has been confirmed (Imbert *et al*, 1996).

Another suggestion (Green, 1993) is that the glutamine tract acts as a target for the action of transglutaminases. These enzymes covalently cross-link proteins by the formation of γ-ε glutamyl-lysine dipeptide (Folk and Finlayson, 1977). The polar nature of the homopeptide would position it to the outside of the protein making it accessible to enzymatic attack. It is suggested that the protein product would be degraded but the dipeptide residue would be resistant to proteolysis, which may be lethal to the cell, in a cumulative manner (Green, 1993). The neuronal specific cell death observed in the polyglutamine diseases may be explained this way since there is transglutaminase activity in neurones, which is involved in the regulation of catecholamine release (Pastuszko *et al*, 1986).

This raises another question which is: why this does not lead to the catastrophic loss of all neuronal and other tissues, since the genes involved are all widely expressed and do not match the lesion distribution specific to the disease (see Figure 1.5.A, adapted from Ross, 1995). For example the huntingtin protein is also expressed beyond the areas of lesion observed for the disease (Sharp *et al*, 1995; Trottier *et al.*, 1995b).

There must be some specific component to the particular subpopulations that are lost in each disease. Recent work has started to unravel this conundrum. Various groups have used the yeast two hybrid systems to look for proteins which interact specifically with proteins that contain expanded polyglutamine tracts. For example, by using the the yeast two hybrid methodology, Li *et al* (1995) have identified a rat protein, termed huntingtin associated protein-1 (HAP-1) which interacted specifically with huntingtin proteins which contained expanded glutamine repeats. This protein did not interact either with huntington's proteins that contain glutamine homopeptides that correspond to normal products in humans or with an atrophin-1 protein containing an expanded glutamine homopeptide (atrophin-1 is the protein which contains an expanded glutamine in the disease, DRPLA). This interaction was also dependent on the number of residues contained in the polyglutamine tract. A huntingtin protein with 82 glutamines showed stronger binding compared to Huntingtin with 44 glutamines. In the same study, co-immunoprecipitation experiments were used to test if the interaction occured *in vivo*. Transfected rHAP-1 was coprecipitated with an antibody specific to Huntingtin. Coprecipitation was also observed by repeating the experiment with a HAP-1 specific antibody. This interaction was abolished by the introduction of a Huntingtin peptide antigen. The authors used the rat sequence to isolate the human homologue of HAP-1 by heterologous screening.

1.10.3 Myotonic Dystrophy.

Myotonic dystrophy (DM) is an autosomal dominant disease with variable penetrance (Harper, 1989) which is associated with a CAG/CTG trinucleotide repeat expansion. It is the most common muscular dystrophy in human adults, affecting 1 in 8000 (Harper, 1989). The disease is characterised by clinically variable symptoms and age of onset. Symptoms include the progressive wasting of muscle tissue, myotonia, cardio-respiratory abnormalities, cataracts, mental retardation and gonadal atrophy. The mutation responsible for DM occurs in the 3' UTR of a gene that shows homology to serine/threonine kinases (Brook *et al*, 1992). This was confirmed by Timchenko *et al*, (1995) who expressed the proposed protein and demonstrated serine kinase activity. DM is different from the CAG/polyglutamine diseases described in the previous section in one main respect; the trinucleotide repeat is transcribed but not translated (Brook *et al* 1992).

The effect of the expansion of the repeat is not clear. Differences in DNA (*e.g.* Wang and Griffiths, 1995), RNA (*e.g.* Otten and Tapscott, 1995) or protein levels (*e.g.* Dunne *et al*, 1996) of DMPK have been observed by individual groups and these have been put forward to explain the disease. There is evidence of both up and down regulation of mRNA and protein levels in DM.

1.10.3.1 Potential effects on DNA.

The argument for a DNA level effect centres round the observation that CTG repeats exert a dominant nucleosome positioning effect (Wang and Griffiths, 1995). It is proposed that

due to an altered chromatin structure in the region, transcription is repressed. To support this, Otten and Tapscott (1995) have observed a loss of a DNAse I sensitive site in DM patients with an expanded CTG trinucleotide repeat. However, the lack of sensitivity may be the result of other proteins binding to the repeat. Candidates for these may be the CTG binding protein observed by Timchenko *et al* (1996; see Section 1.8). It is therefore possible that the *de novo* binding of proteins could play a role in disease aetiology.

1.10.3.2 Potential effects on RNA function.

Conflicting evidence of messenger RNA levels has been described by various groups. Fu *et al* (1993) report that the level of DMPK mRNA was inversely proportional to the size of the repeat in adult tissues. Carango and co-workers (1993) fail to detect any alternatively spliced RNA forms, or reduced DMPK mRNA transcript levels in somatic cell hybrids containing affected chromosome 19 from a MD patient family. Carango *et al* (1993) believe the difference lies in the measurement of all transcripts in the case of Fu *et al* (1993), who used a primer pair for RT-PCR analysis which was co-linear with respect to primary and processed messenger RNA, and themselves who used another primer pair which straddled an intron. Contradictory evidence came to light; when using tissues from the congenital form of DM (CMD), Sabourin *et al* (1993) observed an increase in the steady state levels of DMPK mRNA. This probably reflects differences in expression levels of the DMPK gene between adult and developing organisms. More recently Taneja *et al* (1995) observe foci of expanded repeat transcripts in

the nuclei of cells from DM patients. This was not observed in nuclei from unaffected individuals and was specific to the expanded allele transcripts. The authors suggest that the export kinetics of the DMPK mRNA from the nucleus to the cytoplasm may be affected by the expansion.

1.10.3.3 Potential effects on DMPK protein.

DM may also be due to a protein level effect. Since the DMPK has been shown to have kinase activity (Timchenko *et al*, 1995) it has been proposed that it is involved in signal transduction mechanisms important for correct muscle function. Mis-regulation of DMPK, either by a loss of functional isoforms or an overall reduction in protein levels, could possibly lead to some of the clinical phenotypes associated with DM; for example, muscular atrophy, myotonia and heart failure.

Immuno-histochemical analysis has identified several proteins which cross react with anti-DMPK antibodies in muscles and other tissues. Van Der Ven *et al* (1993) localised a 53 kDa protein to the intercalated discs of cardiac muscle and the neuro-muscular junction of skeletal muscles, that might indicate some regulatory function. However, the predicted molecular weight of DMPK is between 70 and 80 kDa, and the 53 kDa protein is likely to be either an isoform of the DMPK protein generated by alternative splicing (as described by Fu *et al*, 1993) or a closely related protein which cross reacts with the antibody. Dunne *et al* (1996) have reported a 64kDa protein from skeletal muscle and a 79kDa protein from brain which is detected by a DMPK specific antibody. The proteolytic profile of the 64kDa protein matches closely that of the

recombinant DMPK protein. Furthermore it has been shown that there is a redistribution of this isoform in DM patients. There is a shift from the triadic region (Dunne *et al*, 1996) to the peripheral sarcoplasmic masses (Sabourin *et al*, 1993).

The complex nature of DM is underlined in recent transgenic knockout studies, in mouse, by Reddy *et al*, (1996). In this study, mice homozygous for the loss of mouse *dmpk* gene were viable and were found to have progressive muscle weakness and myopathy. However no myotonia is observed and heterozygous mice did not display any overt phenotype. This is contrary to the human situation where heterozygous individuals are affected and the effect is dominant. This difference may be due to differences in muscles between species. Alternatively some other factor may be involved, *e.g.* other proteins being involved in the aetiology of DM.

1.10.3.4 Are other genes involved in myotonic dystrophy?

It is known that the region around DMPK has a high density of genes and it has been proposed that DM is an oligo-genic disease (Johnson *et al*, 1996; Harris *et al*, 1996), in which several genes are implicated in the phenotype. In support of this is the contradictory data so far obtained for the DMPK locus (see above). It is proposed that the CTG repeat expansion may not only affect DMPK, but also other genes in the vicinity.

A novel homeodomain protein, DMAHP (DM-associated homeodomain protein; Boucher *et al*, 1996) lies centromeric to DMPK. It has become of interest to DM researchers because of its close proximity to the DMPK locus and the expanded trinucleotide repeat. The last exon of *dmpk* gene overlaps the promoter region of

DMAHP. In addition this gene has been shown to be transcribed in muscle, heart and brain, both in controls and patients with DM. These are tissues that are affected in DM.

The mutant CTG repeat lies within a CpG island that occurs at the 3' end of the DMPK gene and at the 5' end of DMAHP. CpG islands are known to affect gene expression and Harris *et al* (1996) suggest that an expansion of the repeat would disrupt the action of the CpG island and affect DMAHP gene activity. In support of this, the previously mentioned DNAse I hypersensitive site that is lost in patients with an expanded repeat (Otten and Tapscott, 1995), is thought to lie in the DMAHP promoter (Harris *et al*, 1996).

There is another gene on the telomeric side of the DMPK locus, gene 59 (Jansen *et al*, 1995) in mouse or DMR-N9 in man (Shaw *et al*, 1993). It has no known function but is expressed strongly in the testes and brain in mouse. These are also sites of pathology in DM (Harper, 1989).

1.11 Other diseases caused by CAG/CTG trinucleotide repeats.

The discovery of the cause of this kind of disease is recent and it is likely that more diseases will be found to be caused by the expansion of CAG/CTG type trinucleotide repeats.

Attention has also centred on other diseases which display anticipation. This includes some major psychiatric diseases. The discovery of dynamic mutations has lead people to assess the genetic linkage data accumulated on familial forms of schizophrenia and bipolar affective disorder (Bassett and Honer, 1994; Petronis and Kennedy, 1995) which challenges the idea that these are oligo-

genic and poly-genic diseases. Family, twin and adoption studies have indicated that there is a strong genetic component in these diseases (Gottesman and Shields, 1982; Kendler *et al*, 1985). However, the search for individual loci for these diseases has been hampered by the inability to replicate initial reports of linkage by other groups (Bassett and Honer, 1994). Data reassessment has concluded that the dynamic mutation/ anticipation model does also fit the data available for these diseases (Bassett and Honer, 1994; Petronis and Kennedy, 1995).

Initial work identified that repeats were expanded in some patients suffering from psychiatric disorders. For example, Lindblad *et al*, 1995 observed an overall increase in CAG repeats at individual sites in genomic DNA from patients with BPAD (bipolar affective disorder), compared to normal controls. However, some initial data has ruled out certain classes of trinucleotide repeat as having involvement in these diseases. Kauffman *et al*, (1996) working on some schizophrenic pedigrees found no evidence of expansion of CGG trinucleotides, using the Repeat Expansion Detection (RED) method (Schalling *et al*, 1993). Other repeats may be the cause of disease and not only those already associated with dynamic mutations.

The screening of human brain cDNA libraries with trinucleotides which have previously been shown to be associated with dynamic diseases have also identified candidate genes. For example, a cDNA for the affected gene in the disease DRPLA, was identified by Li *et al*, (1993) from a human cerebral cortex cDNA library which was screened with a $(CTG)_{10}$ oligonucleotide. It is likely that other genes that are susceptible to dynamic mutation

will be found within the subset of genes which contain trinucleotide repeats.

1.12 CAG/CTG trinucleotide repeats and the mouse.

The mouse has an unrivalled position in modern molecular biology. It is the best studied mammalian organism and acts as a model in which to study normal function and disease. This is no less true for trinucleotide repeats.

Trinucleotide repeats have been utilised extensively in mouse as microsatellites to map genes (Love *et al*, 1990). Triplet repeats have also been found to occur in mouse genes. In a nucleic acid database search (Genbank and EMBL), Abbott and Chambers, 1994 found a number of mouse genes that contained CAG/CTG type repeats within cDNA sequences. Their position within the genes varied. They were found in the 5' untranslated region, and in both intronic and coding (exonic) sequences.

1.12.1 The mouse genome is rich in un-characterised expressed CAG/CTG repeats.

Furthermore there appear to be many more triplets of this type expressed than has been previously characterised. In support of this, Duboule (1987) took an *OPA* repeat probe derived from *Drosophila melanogaster* gene *Notch*, and used it to probe a northern blot containing Poly-A enriched mouse RNA and found a multiple banding pattern. This suggests that there are many examples of *OPA* type repeats in expressed mouse sequences. Although *OPA* repeats are described as repeats of CAN, a subset of

those detected by Duboule *et al* (1987) will contain perfect repeats of CAG trinucleotides.

More recently, Chambers and Abbott (1996) have screened an adult mouse brain cDNA library with oligonucleotides containing 5 copies of each of the 10 trinucleotide families. They found 3 CAG repeat positive clones per 1000 recombinant phage screened. This is similar to the results obtained by Li *et al* (1993) and Riggins *et al* (1992). Of the five clones presented in the paper (Chambers and Abbott, 1996), only one has similarity to a known gene, a human mitochondrial malate dehydrogenase. Of the remaining four, 3 have similarity to human ESTs, which are the product of random mass sequencing of cDNA libraries. These sequences have no assigned function. The remaining clone shows no similarity to any sequence in the Genbank or EMBL databases. Three of the clones had a repeat sequence which contained 10 tandem copies of CAG/CTG. The remaining trinucleotide repeats were greater than or equal to the size of the probe, $(CTG)_5$. Chambers and Abbott (1996) also looked at the variability of four of the CAG/CTG repeats. Three exhibited variability between different mouse strains derived from *M. musculus*. The fourth, containing five copies of CTG/CAG, did not show any variability even between different species of mouse (*M. musculus, M. spretus and M. caroli*). Overall, Chambers and Abbott (1996) found that 24% of all repeat loci tested (a variety of loci from all ten families of repeat) were found to vary between strains of *M. musculus* and that 64% varied between *M. musculus* and another species (*M. spretus*). For CAG repeat containing clones, 60% were found to vary between inbred strains. These data were compared with the author's previous results (Abbot and Chambers, 1994), where previously characterised mouse cDNAs from nucleic

Table 1.1

			residues/trinucleotides		
Disease	**gene**	**repeat**	**human**	**mouse**	**rat**
DRPLA	atrophin	CAG/Q	7-35 [1]	n/a	5 [2]
SCA1	ataxin	CAG/Q	6-39 [3]	2 [4]	n/a
HD	Huntingtin	CAG/Q	25 [5]	7 [6]	n/a
SBMA	androgen receptor	CAG/Q	12-33 [7]	5 [8]	n/a
MJD/SCA3		CAG/Q	14-34 [9]	n/a	n/a
myotonic dystrophy	DMPK	CTG/ 3' UTR	5-40 [10]	$(CTG)_2(CAG)_2CTG$ [11]	n/a

Table 1.1 Comparison of CAG/CTG repeats in human and rodent genes, which are associated with human disease. The numbers in human, mouse and rat columns represent the number of glutamines (in the case of SCA1, spino-cerebellar atrophy, type 1; DRPLA, dentato-rubral and pallido-luysian atrophy; HD, Huntington's disease; MJD, Machado-Joseph disease (SCA3) and SBMA, spinal and bulbar muscular atrophy) or trinucleotides (in the case of DM, myotonic dystrophy) at identical positions in the same gene, in different species. References, which are indicated by numbers in bold type, are: [1] Li *et al*, 1993; [2] Loev *et al*, 1995; [3] Banfi *et al*, 1994; [4] Banfi *et al*, 1996; [5] Huntington's Disease Collaborative Research Group, 1993; [6] Barnes *et al*, 1994; [7] LaSpada *et al*, 1991; [8] Faber *et al*, 1991; [9] Kawaguchi *et al*, 1994; [10] Brook *et al*, 1992; [11] Jansen *et al*, 1992. n/a indicates that the homologoue gene sequence is not available for that particular species.

acid databases (Genbank and EMBL) were selected on the criteria of having at least 7 CAG trinucleotide repeats. Of the 12 cDNAs which fitted these criteria, 55% varied between inbred strains and 78% varied between *M. musculus* and *M. spretus*. The figures for both studies, 55% (Chambers and Abbott, 1996) and 60% (Abbott and Chambers, 1994) are similar.

Polyglutamine stretches appear to have a positive effect on protein-protein interactions, probably due to the polar nature of the amino acid residues. A high proportion of proteins from a variety of species that contain long stretches of glutamines are transcription factors (Gerber *et al*, 1994). This is may be the case in mouse as well, since a number of mouse transcription factors contain polyglutamine. These include Brain 2 (an octamer repeat binding POU domain protein; Hara *et al*, 1992) and the glucocorticoid receptor (a steroid receptor; Nohno *et al*, 1989). Expressed CAG/CTG repeats in mouse are not exclusively translated and of those that are translated, not all encode polyglutamine. An illustrative example is that described by Theodosiou *et al* (1996) who describe a mouse MAP kinase phosphatase (M3/6) which contains an AGC repeat which is translated as polyserine (AGC is a member of the CAG/CTG codon family). This repeat is part of a greater polyserine tract, which is made up with the addition of an AGT repeat. In addition to this AGY repeat, where Y is a pyrimidine base and can be either cytosine (C) or thymidine, (T), there is a juxtaposed GGN repeat which is presumed (by Theodosiou *et al*, 1996) to encode a polyglycine repeat. The full length of this repeat is $(AGC)_4(GGT)_7(GGC)_{11}(AGC)_4AAC(AGC)_{13}(AGT)_6$ at the nucleotide level. This repetitive region is translated as $S_4G_{18}S_4NS_{19}$. This is another

example of the existence of a CAG/CTG type repeat and a GGN repeat in close proximity.

Using the stringent conditions of Li *et al* (1993, hybridisation solution: 50% formamide, 5xSSPE, 42°C and washes at 60°C in 1xSSC, 0.5% SDS, thrice 20 minutes) will facilitate the isolation of translated CAG type repeats. Compared with the results of Riggins *et al* (1992), Li *et al* (1993) isolated larger CAG (and CCG) trinucleotide repeats, among which there were a large proportion of clones in which the trinucleotide repeat is presumptively translated. Furthermore, 66% of those which were translated encoded glutamine stretches. Indeed this experiment isolated the protein which is associated with the human disease, DRPLA (CTG-B37, Li *et al*, 1993; Kawaguchi *et al*, 1994). There is no information about possible translation available for the mouse repeats isolated by Chambers and Abbott (1996) but they used a smaller probe, which is probably more sensitive to hybridisation conditions. It is known that for oligonucleotides, annealing is dependent on the length of the oligonucleotide, as well as temperature. This may limit the size of trinucleotide repeat isolated.

1.12.2 Human-mouse comparisons in genes associated with dynamic mutation diseases.

In a major effort in trying to understand the nature of genetic diseases involving dynamic mutations in humans genes, the mouse homologues have been cloned to develop animal models. However CAG repeats found in the human genes (Table 1.1) affected are not generally conserved in the mouse homologues. For example the mouse HD gene homologue contains only seven repeats (Barnes *et*

al, 1994), whereas the human HD gene isolated from unaffected control patient contained 25 repeats (Huntington's Disease Collaborative Research Group, 1993). Furthermore the CAG repeat in the mouse gene is corrupted with a CAA trinucleotide, which also encodes glutamine. These corruptions have been shown to have a stabilising effect on trinucleotide repeats (Weber, 1990). It has been noted that in the trinucleotide repeats which are associated with human diseases, stability is conferred by such interruptions. For example Eichler *et al* (1994) noted that the GGC repeat of the *fmr*1 gene (which is affected in FRAXA) is interspersed by two AGG trinucleotides. Instability of this repeat was only observed when one or both of the AGG repeats were replaced with CGG trinucleotides.

Gene knockouts have confirmed that the polyglutamine class of disease is likely to be due to gain of function mutations, since the lack of the proteins in the organism does not replicate the disease phenotype. For example, separate groups which have developed targeted disruptions have found that mice lacking the Huntingtin protein do not display the HD phenotype (Duyao *et al*, 1995; Nasir *et al*, 1995 and Zeitlin *et al*, 1995).

1.12.3 Human transgene studies.

In a further attempt to understand the nature of the polyglutamine associated diseases, human transgenes with the appropriate expanded glutamine tracts have been introduced into the mouse to replicate the disease pathology.

Attempts by different groups, working on different diseases have had varying results. For example, the introduction of human

androgen receptor transgenes with expanded polyglutamine tracts into the mouse failed to replicate the SBMA disease phenotype (Bingham *et al*, 1995). However this transgene was not expressed in the same tissues as the endogenous AR protein. Furthermore, lower levels of the transgenic protein were detected (relative to normal AR protein expression levels). These differences may account for the lack of observable phenotype. For instance, a critical amount of mutant AR protein may be required for the disease phenotype and the lack of mutant protein in the affected tissue types may preclude the disease from occurring.

In contrast to the AR transgene experiment, a disease phenotype that is similar to that found in man has been observed for ataxin (the effected gene in SCA1) transgenes (Burright *et al*, 1995). The authors used the transgene in conjunction with a Purkinje cell specific promoter of the gene pcp2 (Purkinje cell protein 2), which had been shown previously to direct Purkinje cell specific expression (Vandaele *et al*, 1991). It is known that these cells are one of the cell populations that degenerate in SCA1. The age of onset of ataxia in these mice correlated with the level of expression of a transgene containing an expanded glutamine repeat (82 residues). Two transgenic lines which did not display an overt ataxic phenotype in heterozygotes (but in homozygotes) were found to have considerable Purkinje cell degeneration. This observation correlates with previous Purkinje cell disruption in experiments using an SV40 T antigen transgene, where no phenotype was observed until up to 50-75% of the Purkinje cell population was lost (Feddersen *et al*, 1992). This effect is dependent on the level of the expanded glutamine transgene because these transgenic mice, when bred to homozygosity, display an ataxic phenotype.

No ataxic phenotype was observed with a control transgene which contained 30 glutamines, which correlates to a repeat length which gives a normal phenotype in man. This is interesting because it is 28 residues larger than that found in the mouse homologue of Ataxin. This indicates that the pathological mechanism that is activated in the presence of expanded glutamine to cause cell death in humans is also present in the mouse. However no variation in repeat length through the generations of transgenic mice was observed; therefore there may be differences between man and mouse in the level of expansion rates and mice may be protected in some way from run away style expansion observed in humans. This phenomenon was also observed in study using a human transgene carrying an expanded glutamine in the AR, the protein involved in SBMA (Bingham *et al*, 1995). Candidates for these differences would be proteins involved in recombination mechanisms.

It has been observed that in some cases of colorectal cancer, microsatellite loci become hypervariable in size in tumour cell populations (Ionov *et al*, 1993). The basis of this is likely to be a defect in the fidelity of recombination or repair. This is similar to a class of bacterial mutations, known as mutator mutations.

It has been found that the repeat in these human transgenes does not vary from generation to generation as is the case in the human genes. It has from this been suggested that the mouse in some way differs from humans in respect to repeat variability. This is questionable because of the existence of large repeats in mouse and that they are variable in a frequency not too dissimilar to humans (Chambers and Abbott, 1994). Furthermore, the genes which have been introduced into the mouse, are human and lack

introns. It could be possible that important co-sequences are missing.

The effect of random integration of the transgene containing the affected human gene may also be an important factor in determination of the variability of trinucleotide repeats. It is known that the position of integration can have an effect on the expression of transgenes. This position affect has been observed in many eukaryotic species where transgenes are introduced and is thought to be an effect of local chromatin structure, regulatory sequences and other co-sequences which affect regional transcriptional activity.

For Myotonic Dystrophy in which there is thought to be a loss of functional protein, two knockouts of the mouse gene (Jansen *et al*, 1996; Reddy *et al*, 1996) gave phenotypes which are partially what is observed in the human disease. Furthermore, no phenotype was observed in heterozygotes carrying the *dmpk* deficiency. This is contrary to DM in humans in which dominance is observed. Over-expression of DMPK protein in mice did not show any observable phenotype; there were no histological differences or electrophysiological changes (Jansen *et al*, 1996). This data supports the proposition that other genes may be affected in DM (Johnson *et al*, 1996). There is a high degree of conservation of genes in the DM region between mouse and man and it is likely that the mouse will be an important model to examine the co-effectors in DM.

1.12.4 Models for CAG/CTG binding proteins.

There are now three independent groups which have isolated trinucleotide specific binding proteins. Richards *et al* (1993) and

Timchenko *et al* (1996), both working in humans, found proteins which were specific to double-stranded and single-stranded DNA and RNA triplets. Since Yano-Yanagisawa *et al* (1995) found CAG specific proteins in mouse, it is likely that the human proteins have their homologues in mouse. This will provide an animal model in which the function of these proteins can be analysed to see which roles they play in cells as well as repeat expansion mechanisms and disease pathology.

1.13 Aims of this project.

The objective of this research is to isolate, identify and characterise mouse genes which contain CAG/CTG trinucleotide repeats. This is an interesting approach to take because of their potential functions, as described in Sections 1.5 to 1.9 of this Chapter. A general study of CAG triplet repeats in mouse would lay the ground work for an understanding of the biology of this class of repeat in the mouse and in other organisms. Furthermore this class of trinucleotide repeat has become associated with a new form of human disease. An examination of mouse triplet repeats, could help in the understanding of these diseases if any cross species correlation could be made. Finally these triplet repeats can be used to identify new genes which may be involved in regulatory processes.

Chapter 2

Materials And Methods

2.1 Strains and vectors.

2.1.1 Bacterial strains.

Strain	Genotype	Reference
NovaBLUE	*endA1 hsdR17 (rk12$^-$ mk12$^+$) supE44 supF58 thi-1 recA1 gyrA96 relA1 lac (F' proA$^+$B$^+$ lacI^{q}ZΔM15::Tn10 (tet^r))*	Novagen
TG1	*supE hsdD5 thiD(lac-proAB) F(traD36 proAB$^+$ lacI^q lacZΔM15)*	Gibson, 1984
XL1-Blue	*supE44 hsdR17 recA1 endA1 gyrA46 thi relA1 lac$^-$ F' (proAB$^+$ lacI^q lacZΔM15 Tn10(tet^r))*	Bullock *et al*, 1987
Q358	*supE hsdR f80^r*	Karn *et al*, 1980
c600hflA	*supE44 hsdR thi-1 leuB6$^-$ hflA150 (chr::Tn10(tet^r))lacY1 tonA21 F*	Young and Davis, 1983
DH5α	*supE44 ΔlacU169 (ψ80 lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1*	Hanahan, 1983

Table 2.1 List of bacterial strains.

Refer to the individual references for the details of the origin of each strain.

2.1.2 cDNA libraries.

Tissue source	strain	primer	vector	reference
8.5dpc whole mouse embryo	C57BL	oligo-dT	λgt10	Hogan B., unpublished
12.5dpc whole mouse embryo	C57BL	random priming (hexamers)	λgt10	Logan *et al*, 1992
13.0dpc whole mouse embryo	C57BL	oligo-dT	λunizap	Allen N., unpublished
adult mouse brain	C57BL	oligo-dT	pSPORT	Meier-Ewart S.B., unpublished

Table 2.2 cDNA libraries screened with the oligonucleotide $(CTG)_{10}$.

20-40000 clones from each library were plated and screened with a γ-^{32}P-ATP end-labelled oligonucleotide, $(CTG)_{10}$. Column 1) represents the tissue source for the mRNA from which cDNA libraries were constructed; Column 2 (strain) represents the mouse inbred strain from which the tissue was taken; Column 3 (priming) describes the method of first round cDNA synthesis priming; Column 4) (vector) DNA vector in which library is maintained; Column 5) References.

2.1.3 Plasmid Vectors.

pBluescript KS vectors.

Phagemid vectors which contain an ampicillin resistance gene (β-lactamase) for selection, a β-galactosidase polypeptide engineered with an extensive polylinker *in situ* containing unique restriction sites for cloning, and an origin of replication for single stranded DNA production. The full sequence for these phagemids can be accessed through Genbank (accession numbers: X52327, KS+ and X52329, KS-).

pSPORT1.

Available from Gibco BRL-Life Technologies. This plasmid was used as the parent plasmid vector backbone of the adult mouse brain cDNA library described in Table 2.2.

2.1.4 Lambda vectors.

λgt10.

Parent lambda vector used in the construction of the 8.5 dpc and 12.5 dpc whole mouse embryo cDNA libraries used in this work. It was created by Huynh *et al*, (1985) for the purposes of cloning cDNA inserts up to 6 kb in size into a unique *Eco*RI site in the *imm* 434 gene which renders recombinant phage cI^-. Non-recombinant cI^+ lysogenise very efficiently in *hfl*A *E. coli* strains (*e.g.* C600*hfl*A, see above), cI^- phage cannot, but go on to re-infect other bacteria and will form plaques on a bacterial lawn.

λZAP II.

This lambda phage based vector was used in the construction of the 13.0 dpc whole mouse embryo library (N. Allen, unpublished). It contains a number of modifications to the basic phage structure. Discrimination of recombinant and non-recombinant phage is based on the disruption of a *lac*Z α polypeptide when material is cloned into the multiple cloning site contained within the *lac*Z sequences. This vector is available from Stratagene and was constructed by Short *et al*, 1988. It also contains sequences which allow for the *in vivo* excision of phagemids containing the cDNA inserts, using a helper phage.

2.1.5 DNA oligonucleotides used in this study.

The sequence of each of the oligonucleotides used in this work is indicated in Table 2.3.

2.2 Growth and maintenance of bacteria and vectors.

2.2.1 Bacterial growth media.

2YT

For 1 litre, dissolve 16 g of bacto-tryptone, 10 g of yeast extract and 5 g of NaCl into 800 ml dH_2O. Adjust to pH7 with 5 M NaOH. Adjust to 1 litre with distilled H_2O and autoclave to sterilise.

Table 2.3

Primer	sequence	reference
neo10	CGGAAAACGATTCCGAAGCC	Gossler *et al*, unpublished
neo11	AGCCGATTGTCTGTTGTGCC	Gossler *et al*, unpublished
neo1	GGAGAACCTGCGTGCAATCC	Gossler *et al*, unpublished
neo2	GAGTACGACCTCAAGAAGCG	Gossler *et al*, unpublished
T3	ATTAACCCTCACTAAAGGGA	Promega, 1995
T7	TAATACGACTCACTATAGGG	Promega, 1995
bT7	GTAATACGACTCACTATAGGGC	Stratagene, 1995
T3/T7α	AGCGGATAACAATTTCACACAGG	BRL-Life technologies
$(CAG)_{10}$	CAGCAGCAGCAGCAGCAGCAGCAGCAGCAG	Chapter 2, Section 2.20
$(CTG)_{10}$	CTGCTGCTGCTGCTGCTGCTGCTGCTGCTG	Chapter 2, Section 2.20;

Table 2.3 List of DNA oligonucleotides used in this work. Sequences represent the 5' to 3' sequences of each oligonucleotide named. The origin of the primers are indicated by the explicit reference. For further details refer to these references.

LB

For one litre, dissolve 10 g tryptone, 5 g yeast extract, 10 g NaCl, 1 g glucose and 20 mg of thymine in to 800 ml of dH_2O. Make up to 1 litre with distilled H_2O and autoclave. LB was used where it was necessary to propagate with the antibiotic tetracycline, which has been shown to be inhibited by the presence of magnesium ions.

NZCYM

This medium was used in the preparation of phage competent bacteria, except for strain XL1-Blue (Bullock *et al*, 1987). This strain requires tetracycline selection and therefore requires to be grown in Mg^{2+}-free medium for the reason detailed above. To 950 ml of dH_2O, add 10 g NZ amine, 5 g NaCl, 5 g bacto-yeast extract, casamino acids and 2 g of $MgSO_4.7H_2O$. Mix to dissolve solids and adjust to pH 7.0. Make up to one litre with dH_2O and autoclave.

Solid and semi-solid media for bacteria.

As required, the different media were prepared as above. Just before autoclaving bacto-agar was added to the concentration of either 15 g/ litre or 7 g/ litre for plates and top agar respectively.

2.2.2 Growth conditions for Bacteria and Phage.

Liquid cultures.

Bacterial samples were inoculated into the appropriate sterile liquid media (in a flask) and put at 37°C, in an orbital shaking incubator at 200 rpm, for a period of 16 to 20 hours.

Solid Media.

Solid agar was melted in a microwave oven until totally dissolved. The molten agar was allowed to cool until 55°C. Antibiotics were then added (if required) to their working concentration (see Section 2.6). The agar was then poured into petri dishes and allowed to solidify. Before use the agar was dried to remove excess moisture. The plates were stored at 4°C if not used directly. Bacteria were streaked out on to plates containing the appropriate selective antibiotics and colorimetric indicators if required.

2.3 Sterilisation of media.

All solutions were sterilised for 20 minutes at 15 lb/sq. in. on liquid cycle. Heat labile materials were sterilised by filtering through a 0.2 µm filter (Sartorious) and stored in previously sterilised containers.

2.4 Transformation of Bacteria.

2.4.1 Preparation of electro-competent bacteria.

Electro-competent bacteria were prepared using the method described in Current Protocols (Ausubel *et al*, 1990). The cells can be re-suspended directly into H_2O and used directly for electroporation, but it is wise to test each batch of cells with a series of standard dilutions of plasmid before proceeding to use them in any experiments. With the testing procedure lasting 2 days after the preparation of the cells it is better to preserve them at -70°C in the way described below as the cells deteriorate rapidly if left at 4°C or on ice.

Procedure.

1) A single colony from a fresh plate was inoculated into 5 ml of appropriate growth medium with selective antibiotic and was grown for 16- 20 hours (*i.e.* overnight).

2) Dilute 2.5 ml of the overnight culture in 500 ml of pre-warmed 2YT medium in a 2 litre flask (for maximum aeration), without any antibiotics. Measure the O.D. (optical density) at wavelength of 600 nm of the culture regularly until it reaches 0.5-0.6.

3) Chill the culture for 15 minutes in an ice-water combination and decant into 2x 250 ml polypropylene centrifuge tubes. Balance the tubes to within 0.01 g with a fine balance.

4) Pellet the bacteria by spinning the tubes at 4200 rpm for 20 minutes in a JA-14 rotor (Beckman). The rotor must be pre-chilled to 2°C.

5) Decant the supernatant and keep the bacterial pellets at 2°C for the rest of the subsequent procedures.

6) Resuspend the cells gently but fully in an equal volume of pre-chilled distilled H_2O (2°C). This is best done by first resuspending the pellets in 5 ml of water by extensive swirling action and on ice. Once bacterial suspension is achieved add the rest of the water. Do not pipette as this mechanically sheers the cells and will result in a reduced number of cells at the end of the procedure and hence a lower transformation efficiency.

7) Spin the bacterial suspension at 4200 rpm at 2°C for 20 minutes.

8) Decant the supernatant and resuspend in an equal volume of dH_2O in the manner described above.

9) Centrifuge the resuspended bacteria at 4200 rpm for 20 minutes (2°C).

10) Decant the supernatant and resuspend each pellet in 5 ml of 10% glycerol solution. Place in a fresh 45 ml polypropylene tube. Make the volume up to 45 ml with water ice cold 10% (v/v) glycerol solution and spin against a balanced tube (containing water) for 10 minutes at 4200 rpm (2°C) in a JA-20 rotor (Beckman).

11) Pour off the supernatant and replace the tube on ice.

12) Add 500 µl of 10% glycerol and resuspend the bacterial pellet gently.

This procedure should yield about 1 ml of competent bacteria which can be used directly for electroporation or can be frozen for later use. Freezing the electro-competent bacteria can be achieved by placing 40 µl aliquots in pre-chilled 1.5 ml microcentrifuge tubes followed by snap-freezing on dry ice for 15 minutes before placing them at -70°C for long term storage.

2.4.2 Electro-transformation of bacteria.

The bacteria were prepared as described above and transformed in the following way. The equipment was the custom electroporation unit manufactured by Biorad and electroporation was carried out using the manufacturers cuvettes and conditions.

Preparation.

Fill the appropriate amount of 1 ml syringes with 2YT medium, flaming to retain sterility. Dilute the DNA solution if required to reduce the concentration of salt.

Procedure.

1) Take the appropriate numbers of fresh or frozen aliquots of electro-competent bacteria (usually corresponding to the number of individual transformations to be done). Thaw an extra aliquot in case one electroporation attempt fails. Place on ice to thaw slowly.

2) Once thawed, pipette 1 µl of DNA containing solution into the cells. Mix by pipetting up and down twice with a micropipette set at 20 µl and leave on ice for 5 minutes.

3) Transfer the DNA/bacteria solution into an electroporation cuvette (which has been pre-chilled on ice). Avoid the introduction of bubbles into the cuvette chamber as this will produce pockets of conductivity which will cause the reaction to fail and kill the cells. Tap the DNA/bacterial mixture to the bottom of the chamber.

4) Wipe the outside of the cuvette to remove excess moisture which builds up due to condensation when the cuvette is transferred to room temperature after pre-chilling.

5) Place the cuvette in the electroporation chamber, slide into position between the electrode contacts and apply a pulse. For a 0.1 cm diameter cuvette, use the following settings: 1.6 kv, 400 ohms resistance and 25 μF capacitance. 0.2 cm diameter gap cuvettes require 2.5 kv, 400 ohms and 25 μF.

6) Immediately resuspend the cells in 1 ml of 2YT and transfer to a sterile 20 ml universal.

7) Incubate the transformed mixture for 1 hour at 37°C shaking at 200 rpm.

8) Plate out the appropriate amount of transformed bacteria on to pre-prepared plates containing the required antibiotic and colorimetric indicators as required. Incubate the plates overnight at 37°C. The rest of the transformed mixture can be stored at 4°C and plated out the next day, if necessary.

2.5 Lambda phage manipulation.

2.5.1 Preparation of phage competent bacteria.

A fresh colony of bacteria was used to inoculate 100 ml of LB with tetracycline to a final concentration of 50 µg/ml, 10 ml of 20% (w/v) maltose (filter sterilised). Maltose is included to stimulate the accumulation of the maltose receptor (*mal*B), which is the binding site for bacteriophage lambda on the bacterial cell surface, prior to infection.

2.5.2 Infection of bacteria with lambda phage.

Phage-competent bacteria were initially mixed with an appropriate amount of phage then incubated for 20 minutes at 37°C to allow the phage to absorb to the bacteria. This mixture was then mixed with top agar at 48°C in a sterile pre-warmed universal, swirled to mix, and then plated onto a square 10x10 cm plate containing bottom agar (pre-dried to remove excess moisture). After leaving the top agar mixture to solidify for 10 minutes on the bench, the plates were then transferred to the 37°C incubator to allow growth of the bacteria and phage. The plates were left at 37°C until the phage plaques were visible but not touching each other.

2.6 Antibiotics and colorimetric indicators.

2.6.1 Antibiotics

Antibiotic	working concentration	Storage conditions
Ampicillin	100 µg/ml	100 mg/ml in H_2O and stored at -20°C.
Tetracycline	50 µg/ml	50 mg/ml in ethanol. Stored at -20°C.

2.6.2 Colorimetric indicators.

X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside), a chromogenic substrate for the β-galactosidase enzyme, was used with IPTG (isopropyl-β-D-thiogalactopyranoside), an inducer of the *E. coli lac* operon. In the presence of active β-galactosidase, X-gal is converted into a product which gives a strong blue colour.

Xgal Stock. X-gal is made up to a stock solution of 20 mg/ml in H_2O, filter sterilised by the method described above (Section 2.3) and stored at -20°C.

IPTG. IPTG is made up to a 20 mg/ml stock in DMF (di-methyl formamide) and stored at -20°C. IPTG is light sensitive and it is best stored in a sterile universal bottle, wrapped in silver foil.

2.7 Long term maintenance of bacterial, plasmid and phage stocks.

2.7.1 Bacterial stocks.

Parent strains of bacteria and bacteria containing individual plasmids were kept in long term storage as 50% (v/v) glycerol solutions of overnight cultures derived from single colonies from fresh plates. These stocks were stored at -70°C in 2 ml polypropylene screw cap tubes (Nunc).

2.7.2 Maintenance of lambda bacteriophage.

Individual plugs (and individual plaques) were stored at 4°C in 1 ml of SM with 50 µl chloroform.

2.7.3 Bacteriophage reagents and solutions.

SM Add 5.8 g NaCl, 2 g $MgSO_4.7H_2O$, 50 ml 1M TrisHCl, pH 7.5 to 800 ml of dH_2O and allow to dissolve. Add sterile gelatine to a final

concentration of 0.01% (w/v) and adjust to one litre. Aliquot the SM into glass bottles and sterilise by autoclaving (Section 2.3).

2.8 Nucleic acid isolation and manipulation.

2.8.1 Isolation of plasmid DNA.

Miniprep of plasmid DNA was performed using a modified method of the BD miniprep using Wizard miniprep columns (Gareth Griffith, unpublished data; Promega).

1) Inoculate 5 ml of 2YT culture containing the appropriate antibiotics with a single bacterial colony from a fresh plate. Incubate for 16 to 24 hours in a shaker (200 rpm) at 37°C.

2) Pellet 3 ml of above culture into a 1.5 ml Eppendorf tube. To do this, add 1.5 ml of culture and then spin for 15 seconds at full speed in a bench top microfuge. Decant the supernatant and then add another 1.5 ml of culture into the same tube. Decant the supernatant again and remove any remaining medium by aspiration.

3) Add 250 µl of P1 solution to each tube. Vortex to re-suspend the pellet fully.

4) Add 250 µl of P2 solution and mix by gentle inversion of the tube six times. Do not shake the tube or contamination of the recovered plasmid DNA by bacterial genomic DNA will occur.

5) Add 250 µl of P3 solution and mix as in step 4. A viscous white precipitate should appear.

6) Spin at 14000 rpm in a microcentrifuge for 20 minutes. Remove pelleted white scum with a sterile toothpick and spin for a further 10 minutes. For DNA to be sequenced spin times should be adjusted to 30 and 20 minutes respectively.

7) Attach 2.5 ml syringes (Becton Dickinson) to Wizard miniprep columns without plungers. Set up the syringe barrel on the vacuum manifold (Promega), if you have one. Otherwise you will have to retain the plunger for use later.

8) Pour the cleared supernatant into a fresh tube containing 750 µl of Wizard miniprep resin. Mix by inversion and stand for one minute. Pour the mixture into the syringe and apply a vacuum through the manifold or insert the plunger into the syringe and push the liquid through.

9) After the buffer in which the resin was suspended has cleared the column, apply 3 ml of wash buffer to the syringe. Re-apply the vacuum to draw the liquid through. Continue to draw air through the syringe barrel for a further 10 minutes after the wash buffer has passed through. Detach the syringe, remove the column and place it in a fresh 1.5 ml Eppendorf tube. Spin for 30 seconds in a microfuge. Remove the column and place in a fresh tube. Let stand for 15 minutes to dry off excess ethanol. If doing this with a plunger, spin immediately after pushing the wash solution through and stand for 15 minutes in a fresh tube.

10) To each tube add 20 µl of $T_{10mM}E_{1mM}$, pH 8.0 (pre-heated to 85°C) and immediately spin for 15 seconds at 14000 rpm. Repeat elution with a further 20 µl of hot $T_{10mM}E_{1mM}$ and spin again. For sequencing grade DNA substitute distilled H_2O for $T_{10mM}E_{1mM}$, pH8.0.

11) Analyse 1 µl of the miniprep DNA on a 1.0% agarose gel with 7 µl of H_2O and 2 µl of agarose gel loading buffer (x 5).

Regeneration of Wizard miniprep columns.

It is possible to re-use the Wizard miniprep columns by washing them in sterile water. This is done by placing them in a heat resistant flask, shake, and heating the water in a microwave oven. Shake again, cool and replace distilled water with clean batch. Repeat the procedure above, until all the old resin has been removed. Dry the columns in a incubator and mark the columns so as to distinguish them from new ones. Syringes and plungers can be cleaned and dried in the way described above except that it is better to do it in a Pyrex dish rather than in a beaker.

Solutions for Promega Wizard miniprep.

P 1 50 mM TrisHCl (pH 8.0), 10 mM EDTA and 100 µg/ml RNAse (low grade, Sigma). Store at 4°C.

P 2 0.2 M NaOH, 1% SDS. Make this up fresh by combining 8 ml H_2O, 1 ml 2M NaOH solution and 1 ml of 10% SDS. If the SDS precipitates, gently warm this solution.

P 3 3 M CH_3COOH. pH 5.2. Store at room temperature.

Column Wash Solution.

20 mM TrisHCl (pH 7.5), 200 mM NaCl, 5 mM EDTA, 50% (v/v) ethanol. To make 1 litre of stock solution (without ethanol), add 10 ml of 2 M TrisHCl (pH 7.5), 50 ml of 4 M NaCl and 10 ml of 0.5 M EDTA (pH7.5). Make the volume up to one litre with distilled water. Sterilise by autoclaving. Add an equal volume of ethanol before use.

2.8.2 Isolation of λ phage DNA.

λ DNA was prepared in the following way (page 1.13.6, Ausubel *et al*, 1990).

2.8.2.1 Lysate production.

A sample of freshly grown phage strip was put into 400 µl of SM and left for a minimum of 2 hours at 4°C. 50 to 100 µl of this suspension was mixed with 100 µl of phage competent bacteria and incubated for 15 minutes at 37°C to allow infection of bacteria by the phage particles. This phage-bacteria mixture was added to 50 ml of NZCYM broth. This was incubated at 37°C in a shaking incubator (200 rpm) until lysis was observed. This was generally between 6 to 9 hours post infection. After lysis, 500 µl of chloroform was added to kill off the remaining bacterial cells and aid lysis.

2.8.2.2 λ DNA extraction.

The prepared phage lysate was transferred to Nalgene polypropylene tubes and DNase and RNAse were added to the final concentrations, 5 mg/ml and 10 mg/ml, respectively. The tubes were then incubated for one hour at 37°C.

Afterwards the tubes were centrifuged at 4°C in a Beckman J20 rotor at 20000 rpm for 2 hours 15 minutes, to pellet the phage. The supernatant was disposed of and the pellet of phage particles was resuspended in 500 µl 50 mM TrisHCl pH8.0. Protein and lipids were removed by a series of phenol extractions (aqueous phenol, Tris buffered to pH 8.0) until no more white waste material was observed on spinning this mixture at 14000 rpm. A final chloroform extraction was performed to remove any residual phenol.

2.8.3 Isolation of Mammalian genomic DNA.

Genomic DNA was extracted from a variety of sources (*e.g.* cultured cells, adult mouse tissue, cell lysates) with the aid of the Nucleon II kit from Scotlab Bioscience. The manufacturer's conditions were used throughout.

2.8.4 Isolation of total RNA.

Total RNA was extracted by the acid phenol-guanidinium method of Chomczynski and Sacchi (1987). This procedure was used to extract total RNA from whole embryos, adult tissues and cultured cells. The protocol is described below.

You will need a mortar and pestle, 2 Nalgene 45 ml polypropylene tubes, blue and yellow tips and 1.5 ml microfuge tubes. All solutions and materials must treated in the appropriate way to render them Rnase-free.

Glass-ware and other heat resistant materials must be heat-baked at 180 °C for two hours. Heat labile materials must be treated in a 0.01% DEPC solution overnight and autoclaved the following day.

Procedure.

1) Measure the weight of starting material on silver foil on a fine balance.

2) Transfer to mortar, add liquid nitrogen and grind with pestle until tissue is reduced to a fine powder. Add more nitrogen and transfer to an RNAse-free polypropylene tube.

3) Evaporate the remaining liquid nitrogen and add solution D in a proportion of 600 µl of solution D to 100 mg of tissue. Mix by inversion.

4) Homogenise the solution by three 30 second pulses at the yellow setting (8000 rpm, Ultra-Turrax T25 homogeniser (Janke and Kunkel). Keep the tube on ice while this is being done.

5) Add sequentially: 0.2 M NaOAc, pH 4.0 (60 µl per 100 mg of original material), mix by inversion and add water-saturated phenol (600 µl per 100 mg) and chloroform (120 µl per 100 mg). Mix thoroughly and then stand in ice for 5 minutes.

6) Spin for 15 minutes in a Beckman JA-20 rotor at 12000 rpm at 4°C.

7) Transfer upper, aqueous layer to a fresh polypropylene tube, taking note of the total transfer volume.

8) Add equal volume of propan-2-ol, mix and stand on ice for more than 2 hours.

9) Centrifuge at 12000 rpm at 4°C for 15 minutes.

10) Discard supernatant and dry the pellet with tube inverted for 10 minutes.

11) Resuspend RNA pellet in RNase-free 70% ethanol (500 µl per 100 mg of original tissue) and aliquot into 1.5 ml microfuge tubes. Store at -70°C until required.

2.9 Synthesis of oligonucleotides.

All oligonucleotides described in this work were synthesised on a DNA synthesiser by the standard phosphoramidate chemical synthesis. The bases were added in a 3' to 5' manner. Other oligonucleotides were purchased from commercial sources where indicated.

2.10 Quantification of nucleic acids.

Measurements of DNA quantities in solution were performed using a spectrophotometer set at the wavelength 260 nm. At this wavelength it has been calculated that the absorption unit coefficients of various nucleic acids have a relationship to their concentration. For one absorption unit, double-stranded deoxyribonucleic acid is at a concentration of 50 µg/ml, single-stranded DNA at 40 µg / ml and RNA 40 µg/ ml. Readings at the wavelength of 280 nm was also measured. The ratio of 260 nm and 280 nm absorbtions for a particular sample gives a measurement of the purity of the nucleic acid in solution. Both pure DNA and RNA have a ratio of between 1.8 to 2.0 (Maniatis *et al*, 1989).

2.11 Restriction endonuclease digestion of DNA.

All restriction endonucleases described in this work were supplied by Life Technologies and used with the REact buffer system supplied with these enzymes.

2.11.1 Plasmid and lambda phage DNA.

These were digested with a 4-5 fold excess in units of enzyme with the appropriate restriction buffer and in a final volume which diluted the enzyme 10 fold. Enzymatic activity can be inhibited by glycerol concentrations of >5%. Glycerol is a major component of the enzyme storage buffer (50% v/v).

2.11.2 Mammalian genomic DNA.

Genomic DNA was allowed to diffuse in an appropriate volume of sterile H_2O and restriction buffer left on ice or at 4°C. To aid diffusion of the DNA it was mixed regularly with a sterile pipette tip and replaced on ice. When the DNA had diffused, half of the intended amount of enzyme to be used for the digestion was mixed in slowly to the DNA/buffer solution. The tube was then left at 4°C for 5 minutes before being transferred to 37°C for 30 minutes. The rest of the enzyme was then added in the manner described above. When the digestion had been allowed to proceed for 5 hours, a 1 μg sample was taken and used for gel electrophoresis, to see if digestion was complete.

2.12 Generation of a deletion series using exonuclease III.

This procedure was carried out according to the manufacturers protocol with the omission of the step to remove nicked and linear plasmid prior to restriction endonuclease digestion (Erase-a-base System, Promega).

2.13 Agarose gel electrophoresis of nucleic acids.

2.13.1 Agarose gels for DNA electrophoresis.

Solutions and materials.

10x T.B.E.

Dissolve 108 g Tris base, 55 g boric acid and 40 ml 0.5M EDTA (pH8.0) in 800 ml of dH_2O. Adjust to 1 litre final volume. Use at 1x strength as the working concentration.

50x T.A.E.

Dissolve 242 g Tris base and 37.2g $Na_2EDTA.2H_2O$ in 800 ml of dH_2O. Add 57.1 ml glacial acetic acid to adjust the pH to 8.5. Adjust the volume to 1 litre, decant into glass bottles and sterilise by autoclaving (Section 2.3).

Agarose.

Electrophoresis grade agarose was supplied by BRL-Life technologies. Low melting point agarose (used for DNA band recovery from agarose gel matrices) was obtained from Nusieve.

10x gel loading buffer.

Make up to 20% Ficoll 400; 0.1 M EDTA, pH 8.0; 1% SDS; 0.25% bromophenol blue and 0.25% xylene cyanol.

Procedure.

The appropriate amount of agarose (BRL-Life technologies) was mixed in a 1x T.B.E. buffer solution to give the final concentration relative to the size of DNA molecules being used. Generally 1.0% (w/v) agarose was used for plasmid DNA separation; 0.7% agarose for the separation of digested genomic DNA and 0.3% was used for undigested genomic DNA.

Sample preparation.

The appropriate amount of DNA was mixed with dH_2O and 10x loading buffer to a final concentration of one times. Genomic and λ DNA was heated to 65°C for 10 minutes then allowed to cool on ice briefly before being loaded into the wells of a preformed submerged agarose gel.

Running the gels.

A constant voltage was applied to the submerged agarose gel for the appropriate length of time, or until the DNA had migrated sufficiently.

2.13.2 RNA gels.

Preparation of denaturing agarose gel for RNA electrophoresis.

For 200 ml of gel mix, add 2.96 g of agarose to 148 ml of DEPC-treated dH_2O. Microwave to dissolve the agarose and cool to 80°C, add 20 ml of 10x M.O.P.S., mix and leave to cool to 65°C. In a fume hood add 32 ml of formaldehyde, mix thoroughly and pour into the prepared gel cast.

Preparation of RNA gel sample buffer.

For the stock solution, add sequentially 2 ml of RNAse-free glycerol, 3 ml of DEPC-treated H_2O with 20 mg each of methylene blue and bromophenol blue and 2 ml of 10x M.O.P.S. buffer. Mix and store at room temperature. Just before use mix 340 µl of the stock solution with 160 µl of formaldehyde.

Preparation of samples for loading.

The appropriate amount of RNA was mixed with a 1:1 ratio with the gel loading buffer (containing formaldehyde) and incubated for 10 minutes at 65°C. The samples were chilled on ice before being loaded onto a formaldehyde gel.

Running formaldehyde gels.

10 µg samples of total RNA was loaded into each well and electrophoresed in a 1.5% agarose, 1x M.O.P.S. buffered gel in a fume hood for 4 hours at 200 volts. A peristaltic pump was used to circulate the buffer during this time.

2.14 Staining, visualisation and photography of nucleic acids.

2.14.1 DNA visualisation.

Agarose gels containing DNA samples were soaked in a dilute solution of ethidium bromide (50 ng/µl) for 30 minutes, after which the gel was transferred to fresh H_2O for rinsing. The gel was visualised under UV light. If over-staining occurred, the gel was allowed to soak in dH_2O for 30 minute intervals until background staining was lost. Permanent records of the gels were either photographs taken by a Kodak polaroid instamatic camera or by a video-imaging facility.

2.14.2 RNA visualisation.

After transfer of the RNA from the formaldehyde-agarose gel and fixation to nylon membrane (Dupont Genescreen$^+$), the RNA was visualised in the following way. The membrane was washed in 2%

SDS to remove migrating dyes, then in DEPC-treated dH_2O and immersed in an excess volume of 0.5 M sodium acetate (pH 5.2) with 0.04% methylene blue, for 10 minutes. The membrane was then destained with a series of 25% ethanol solutions until the non-specific background staining was removed. The stained membrane was wrapped in cling film and photocopied to provide a permanent record of the distribution of the RNA.

2.15 Recovery of DNA fragments from agarose gels.

Where it was necessary to isolate restricted DNA fragments, the DNA was digested and the fragments were separated by electrophoresis in low melting point (LMP) agarose using 1x T.A.E. buffer at 4°C. The appropriate band(s) were dissected with the minimal amount of agarose and DNA was extracted by spinning through COSTAR spin-X columns in the manner described by the manufacturer. As a final step, the DNA in the column eluant was ethanol precipitated in the manner described in Maniatis *et al* (1989). The recovery of DNA was checked by taking a sample of the re-dissolved precipitate and subjecting it to agarose gel electrophoresis with the relevant DNA molecular weight markers.

2.16 Ligation of lambda cDNA inserts into pBluescript.

5 µg of lambda phage DNA was digested by the enzyme *Eco* RI (as described in Section 2.11.1). After the digestion of the λ DNA had been completed, restriction enzyme activity was destroyed by

heating the reaction to 65°C for 10 minutes. 1 µg of the restricted λ DNA was mixed with 0.2 µg of linearised pBluescript plasmid DNA. This DNA mixture was added to a cocktail solution which contained the appropriate amount of T4 phage DNA ligase (1-3 U/µl, Promega), buffer (x10: 300 mM Tris-HCl, pH 7.8, 100 mM $MgCl_2$ and 100 mM DTT; Promega), 10 mM rATP and dH_2O in a final reaction volume of 10 µl. This reaction was incubated at 4°C for a minimum of 16 hours.

2.17 Transfer of nucleic acids to nylon membranes.

2.17.1 Electrophoretically separated DNA fragments.

Denaturation of DNA.

Prior to the transfer of the DNA to nylon membrane, it is necessary to denature the double-stranded DNA. To achieve this, the gel containing the DNA was placed in a 0.4 M NaOH solution for 30 minutes (in the case of plasmids) or 45 minutes (for genomic DNA). The gels were thoroughly rinsed in dH_2O and immersed in the electro-blotting buffer (1x T.B.E.) for 20 minutes.

Electroblotting DNA from agarose gels.

The DNA was transferred by constant electric current (1.5 Amp) using an electroblot apparatus in 1x T.B.E. buffer solution. This was allowed to proceed for 2 hours in the case of lambda and

plasmid DNA digests. For genomic DNA a 1.0 amp current was applied for 6 hours (or longer) to allow for the larger DNA fragments to transfer to the membrane.

2.17.2 Denatured plasmid DNA for slot blots.

One microgram (spectrophotometrically determined) plasmid DNA was denatured with a 0.1 M NaOH solution, in the presence of 1x agarose gel loading buffer (see Section 2.13.1.1) for 10 minutes. The samples were drawn by a vacuum onto dry nylon membrane (Dupont Genescreen^{+}). The vacuum was continuously drawn for a further 30 minutes to allow the DNA to dry on to the membrane. The DNA was cross-linked to membrane using the UV cross-linking apparatus as described below (Section 2.18).

2.17.3 RNA

This was done by the standard capillary method as described in Maniatis *et al*, (1989), Chapter 7, page 49.

2.18 Covalent cross-linking of transferred nucleic acids to nylon membrane.

Nucleic acids (DNA and RNA from agarose gels and slot-blotted denatured plasmid) were covalently cross-linked to Dupont Genescreen^{+} nylon membrane by UV exposure for 3 minutes in a custom cross-linking apparatus (Stratagene Stratalinker) at a power setting of 1200 W.

2.19 Inverse polymerase chain reaction.

PCR is the selective amplification of DNA between two points defined by sequence specific primers. IPCR (a variation of standard PCR) uses a circularised template from which outwardly directed primers can be used to amplify sequences in the opposite or inverse direction relative to normal PCR (which generally amplifies sequences between two convergent primers). This technique has been used successfully by Korn *et al* (1992) to identify flanking sequences from enhancer trap integration site positive cell line DNA.

Digestion of enhancer trap site (ETS) genomic DNA.

Template DNA was digested to completion with an appropriate restriction enzyme. Digestion was monitored by applying a sample of the DNA to agarose gel electrophoresis (see Sections 2.11.1 and 2.13.1.1). The restriction enzyme was removed by a phenol (pH 8.0) extraction and once with chloroform. The DNA was precipitated (as described by Maniatis *et al*, 1989) and resuspended in dH_2O.

Circularisation of digested DNA.

The DNA was resuspended to a concentration of 1 µg/ml. These conditions favour the formation of circularised monomers of restricted DNA rather than intermolecular ligations reactions. Collins and Weissman (1984) first described these conditions and Maniatis *et al* (1989) describe the relationship between the concentration of DNA (c, measured in µg/ml), the size of the molecule (in base pairs) and the likelihood of intramolecular ligation. The relationship is described by the expression, $1900/c(bp)^{1/2}$. Generally if the calculated value of this ratio is less than 1, circularisation will be favoured.

Therefore the previously digested DNA was allowed to ligate at a concentration of 10 μg/ml in the presence of T4 ligase at a concentration of 1 U/μl. The ligation was allowed to proceed overnight at 16°C. The ligase enzyme was heat killed at 68°C for 15 minutes. The DNA was precipitated and resuspended to a concentration of 25 μg/ml, after a phenol (Tris buffered, pH 8.0) and chloroform extractions.

Amplification conditions.

100 ng (4 μl) of ligated DNA was combined with 30ng each of primers neo10 and neo11 (1 μl each), 4 μl of 1.25 mM dNTPs, 2.5 μl *Taq* amplification buffer (TAB) x10 (Promega), 0.5 μl (5U) *Taq* DNA polymerase (Promega) and dH_2O was added to adjust the final reaction volume to 25 μl.

First round PCR.

DNA was denatured with the appropriate amount of H_2O for 10 minutes at 98°C. The DNA/H_2O solution was placed on ice to cool and any condensate was spun down in a microfuge. 10xTAB and dNTPs were added along with the appropriate primers to the DNA solution. This mixture was overlaid with mineral oil (Sigma) and placed in a Perkin Elmer DNA thermocycler and heated to 80°C for 5 minutes. *Taq* DNA polymerase was then added and the whole mixture was heated to 94°C for 3 minutes. This was followed by 30 rounds of: Annealing 60°C, 30 seconds; Extension 72°C, 2 minutes; denaturation: 94°C 30 seconds. PCR was completed with a 7 minute extension at 72°C, after which reactions were stored at -20°C until required.

Second round PCR.

This was carried as first round PCR except that the template was 1 µl of a 1/1000 dilution of the PCR reaction from round one (see above) and a nested set of primers, neo1 and neo2 were used.

2.20 Radioactive labelling of nucleic acids.

2.20.1 Random priming of DNA fragments.

This procedure was carried out according to the manufacturers recommendation using the Prime-It random priming kit (Stratagene). Generally 50 ng of double stranded, linearised DNA was labelled with 50 µCi of dCTP (specific activity = 3000 Ci/mmol) in a final volume of 50 µl. The reaction was incubated for 30 minutes at 37°C.

2.20.2 End labelling of oligonucleotides with γ-^{32}P-ATP.

1 pmol of oligonucleotide to 10 µCi of γATP (3000 Ci/mmol, NEN) in a 20 µl reaction of 1x T4 polynucleotide kinase buffer (Promega) and 8 units of T4 Polynucleotide kinase (Promega). The reaction was incubated at 37°C for 30 minutes.

2.20.3 Labelling of first strand cDNA.

Annealing and labelling.

This was performed on 10 µg of total RNA, to which 500 ng of oligo-dT_{12-18} primers (Pharmacia) had been annealed. To this mixture, 1 µl reverse transcriptase, Superscript II (Life Technologies) and 4 µl of the 5x first strand buffer were added for the synthesis of complementary first strand molecules. 1 µl cold dNTPs (10 mM each dATP, dGTP and dTTP; final concentration, 2 mM) and 2 µl 0.1M DTT were also added. 70 µCi of α-labelled dCTP (deoxycytosine triphoshphate, 3000 Ci/mmol, NEN Dupont) was used per reaction to label the newly synthesised first strand products. This reaction was allowed to incubate at 42°C for 2 hours.

Cold chase reaction.

After the labelling, a chase step was performed with cold dCTP, to maximise the length of products from the reaction. 2 µl 5x first strand buffer, 2.5 µl 10 mM dCTP, 4.5 µl dH_2O and 1 µl of Superscript II were added sequentially. The reaction was incubated at 42°C for 90 mins.

Removal of RNA by NaOH hydrolysis.

RNA was removed by alkaline hydrolysis by adding 1M NaOH (1.5 µl 1M NaOH and 1 µl 0.5 M EDTA pH 8.0 were added). The treated reaction was then incubated at 50°C for 75 minutes. The

alkali was afterwards neutralised with the addition of the appropriate molar amount of 1 M HCl (1.5 µl 1M HCl).

2.20.4 Separation of labelled nucleic acids from unincorporated radioactively labelled deoxynucleoside triphosphates.

Labelled probes (from random priming, first strand cDNA synthesis and oligonucleotide end-labelling reactions; see Sections 2.20.1 to 2.20.3 above) were separated from unincorporated radioactively-labelled deoxynucleoside triphosphates by passing the reaction over a column containing G50 (or G25 in the case of oligonucleotide seperation) Sephadex (Sigma) which was either purchased from the manufacturer (Pharmacia) or on a home made column. The home made G50 column was constructed from a sterile Pasteur pipette blocked with siliconised glass wool and filled to the top half centimetre with G50 Sephadex slurry which was re-hydrated and autoclaved in STE (100 mM NaCl, 10 mM TrisHCl and 1 mM EDTA, pH 8.0). The G-50 Sephadex slurry was prepared by adding dehydrated G-50 Sephadex to the STE solution. This mixture was autoclaved (Section 2.3) to re-hydrate the Sephadex.

The probe was loaded onto the column and this was followed by 3 ml of S.T.E. Fractions, each containing five drops, were collected as the probe made its way through the column. 1 µl of each of the fractions was then taken to be measured in a scintillation counter (Beckman) using Cherenkov counting. Dpms (disintegrations per minute) were recorded. From this a profile of the elution of

radioactivity from the column was determined. The first peak of radioactivity corresponds to the labelled fragments.

2.21 Hybridisation analysis of nucleic acids.

Materials and solutions.

20xSET

3 M NaCl, 0.4 M TrisHCl (pH 8.0), 20 mM EDTA. To make 2 litres, add 350.6 g NaCl to 1 litre of distilled H_2O, dissolve, then add 800 ml 1 M TrisHCl, pH 8.0 and 80 ml of 0.5 M EDTA, pH 8.0. Adjust the volume to 2 litres and sterilise by autoclaving.

10% SDS.

Add 10 g of SDS to 80 ml H_2O. Dissolve the SDS and adjust to 100 ml.

50x Denhardt's solution.

For 500 ml of stock solution: Add 5 g Ficoll 400, 5 g polyvinylpyrrolidone and 5 g of BSA to 450 ml H_2O. Adjust to 500 ml, filter sterilise and store at -20°C.

De-ionised Formamide.

This was prepared freshly for each experiment by the following procedure. 20 g of de-ionising resin was added per litre of formamide and was left to mix with the aid of a magnetic stir bar for at least one hour. The solution was then filtered through Whatman paper and used immediately for the hybridisation solution.

Sheared salmon sperm genomic blocking DNA.

Salmon sperm DNA was prepared according to the procedure described in Maniatis *et al* (1989, Appendix B, p15) and adjusted to a final concentration of 10 mg/ml. The sheared DNA was aliquoted into 1 ml samples and stored at -20°C until required.

Hybridisation solution.

All probing experiments were conducted in the following solution: 4x SET, 50% formamide, 0.5% SDS, 5x Denhardt's solution and 1 M phosphate buffer, pH 6.8.

Method.

Prehybridisation.

Nylon membranes containing the fixed nucleic acid were treated with an excess amount of hybridisation solution with heat-denatured sheared Salmon sperm DNA at a concentration of 100 µg/ml for a minimum of 2 hours in a cylindrical flask in a Techné hybridisation oven.

Hybridisation.

Prehybridisation solution was replaced with fresh hybridisation buffer along with freshly denatured blocking DNA. The probe was first denatured then added to a concentration of $1X10^6$ cpm/ml relative to the volume of hybridisation solution used.

Calculation of the length of incubation.

The length of hybridisation was allowed to proceed for the length of time corresponding to 3x $Cot_{1/2}$ (Anderson and Young, 1990). $Cot_{1/2}$ was calculated by the equation where $Cot_{1/2} = N= (1/X \times Y/5 \times Z/10) \times 2$ where N is time (in hours), X= the weight of probe added (in mg), Y is equal to the complexity of the probe (for unique sequences, in kilobases) and Z is the total volume of the hybridisation reaction in ml.

Washing regime.

Filters were washed in solutions of 0.1% SDS and first 1x SET and then 0.1x SET at the appropriate temperature indicated in the specific experiment.

Specific modifications for reverse northern hybridisaton.

Slot-blotted plasmid DNA panels were prehybridised and hybridised with the appropriate radio-actively labelled first strand cDNA probes under identical conditions as described above, except for the addition of 100 μg/ml unlabelled trinucleotide repeat oligonucleotides, $(CAG)_{10}$ and $(CTG)_{10}$. This was done to block CAG/CTG repeat loci which may encourage misleading cross-hybridisation reactions.

2.22 Autoradiography.

Autoradiography was performed either using cassettes and film where the membranes were exposed for an appropriate length of time at -70°C (in the case of ^{32}P labelled probes) with intensifier screens with Fuji RX film (Fuji) or the filters were exposed to a light-sensitive screen. The screen was analysed using the MacBAS phosphoimaging system. The files that were generated were stored as Bas type files and manipulated therein. In the case of the use of α-^{35}S-dCTP (manual sequencing), no intensifying screens were used and cassettes were left to expose at room temperature. Exposed autoradiographic film was developed with the aid of a X-OGRAH Compact X2 automated film processor.

2.23 DNA sequencing.

2.23.1 Manual sequencing.

This was performed throughout by the di-deoxynucleotide sequencing method (Sanger *et al* , 1977) using T7 polymerase in a kit as described by Pharmacia (Pharmacia). Double-stranded DNA templates were used and the primers which prime DNA synthesis were obtained either from commercial sources or made on the automated oligonucleotide synthesiser. Samples were run on 7% polyacrylamide gels using the Base Runner gel kit (IBI) at 40 watts power setting in a 1.0x T.B.E. buffer. After the samples were run for the appropriate time, the gel apparatus was dismantled and the gel was recovered by transfering it to a sheet of 3MM Whatman paper. A vacuum was drawn through (and heat applied) the gel to dry it. The gel was exposed to autoradiographic film in the manner described in Section 2.22.

2.23.2 Fluorescent labelled automated sequencing.

This was performed using the Dye-deoxy terminator cycle sequencing kit (Applied Biosystems) using the standard procedures as described by the manufacturer. DNA templates were prepared as above (Section 2.8.1). The primers which were used were those which met the criteria for automated sequencing by this method (Applied Biosystems). The reactions were performed as described

below. Plasmid DNA was adjusted to give the approximate amount of DNA (0.1 to 0.2 μg) as required.

Reaction example:

6.0 μl DNA (= 0.1-0.2 μg DNA)

9.5 μl Reaction mix, ddNTP, dNTP and *Taq* polymerase

4.5 μl H_2O

Total vol. 20 μl

These ingredients were mixed in a thin wall tube and placed in a Perkin Elmer GeneAmp 9600 thermocycler, pre-heated to 96°C. The thermocycle was performed 25 times as: a rapid thermal ramp to 96°C and held at 96°C for 15 seconds; a rapid thermal ramp to 50°C and held at 50°C for 1 second; a rapid thermal ramp to 60°C and held at 60°C for 4 minutes. The labelled DNA was recovered by ethanol precipitation, followed by spinning, washing (in 70% ethanol), spinning for 15 minutes at drying. The pellet was then re-suspended in 4 μl of gel loading solution. The samples were loaded on to a 7% polyacrylamide gel after heating to 85°C for 5 minutes, followed by quenching on ice. The gel was run according to the manufacturer's recommendations. The data was interpreted by using the software provided by the manufacturer.

2.24 Computer-Aided Sequence Analysis.

2.24.1 UWGCG sequence analysis software package.

This was available for use in the laboratory via a Telnet link to UNIX machines (Lenzie and Newton) at the University of Glasgow and the VAX server Sc2a at the University of Geneva. Access was via NCSA Telnet programme on Macintosh computers using a VT100 terminal emulation configuration. See "Programme manual for GCG" for references to specific programmes in the UWGCG (University of Wisconsin Genetics Computing Group; Devereux *et al.*, 1984) package version 8.0 and UNIX users guide (University of Glasgow computing services publication) for a list of commands for the UNIX. Sc2a VAX server in Geneva had UWGCG Version 8.1 installed.

2.24.2 Nucleic acid and protein databases.

Access to the Genbank, DDBJ and EMBL nucleic acid databases and Protein databases (Swissprot and PIR) was gained through the UWGCG package.

2.24.3 Database search algorithms.

Two main search algorithms were used in this work. These were the FASTA hash-coding local alignment algorithm (Lipman and Pearson, 1985) and the BLAST (Basic Local Alignment Search Tool, Altschul *et al*, 1990). Both use local alignment protocols to build up regions of similarity. FASTA was used generally for nucleic acid

database searches. The BLAST derivative programme BLASTX was used to search non-redundant databases with a six-frame translation of the sequence derived from a particular clone. The output was in an amino acid format.

2.24.4 MacVector sequence maintenance programme.

Manual DNA sequence data was transferred into the sequence analysis package via a sequence gel reader from IBI.

2.24.5 MacAlign Sequence alignment programme.

This programme was used in conjunction with the MacVector programme to align sequence into contiguous sequence to generate full sequence of clones to enable work on the nature of the open reading frame of these clones and to assign the putative homopolymer the repeat encodes if it proved to be in an ORF.

2.24.6 GeneJockey (Version 2).

Sequence derived from automated sequencing was handled by this software as it could interpret the ABI derived files.

2.24.7 ABI automated sequence interpretation software package.

This was used to interpret electrophoretogram data of single primer sequence reactions from the automated sequence protocol.

Chapter 3

Screening Of Mouse cDNA Libraries For Inserts Containing CAG/CTG Family Of Repeats; Initial Expression Analysis.

3.1 Introduction.

As described in chapter one, human cDNA libraries (Li *et al*, 1993; Riggins *et al*, 1992) that were screened with CTG_{10} oligonucleotides yielded many clones containing triplet repeats. These studies showed that cDNA libraries derived from brain tissue (foetal brain, Riggins *et al*, 1992 and adult cerebral cortex, Li *et al*, 1993) are rich in CAG/CTG repeats, compared to libraries from other tissues (Riggins *et al*, 1992). From this it can be concluded that complex cDNA libraries are a good source of CAG/CTG triplet repeat containing genes. Duboule *et al* (1987) detected a large number of RNA species in mouse poly-A enriched RNA that cross hybridised with a *Drosophila OPA* repeat-rich probe. This indicates that there is a large number of CAG/CTG-type repeats in the mouse gene repertoire. *OPA* repeats are CAN (where N= A, C, G or T) nucleotide repetitions, which are found in numerous *Drosophila* proteins where they encode amino acid homopeptides. The two most frequent repeated trinucleotides are CAG and CAA, which are the two codons for the amino acid glutamine. The majority of these repeats encode polyglutamine stretches, others encode alanine and serine homopeptides.

To determine experimentally if expressed gene sequences of the mouse contain CTG/CAG trinucleotide repeats, three whole embryo mouse cDNA libraries, derived from 8.5 dpc (Hogan *et al* unpublished); 12.5 dpc (Logan *et al*, 1992) and 13.0 dpc (N. Allen, 1993) embryos were screened. In addition, an adult mouse brain cDNA library (Meier-Ewert *et al*, 1992) was screened. Clones were purified from the 8.5 dpc (Hogan *et al*, 1984) and 12.5 dpc (Logan *et al*, 1990) embryonic cDNA libraries. Reverse dot-blotting was

performed to identify clones with different expression levels at different times in development. This provided a basis for further study of selected clones. The results of these initial experiments are detailed in this Chapter.

3.2 Screening of mouse cDNA libraries.

Various libraries were screened with an oligonucleotide comprising ten copies of the repeat CTG, as described by Li *et al* (1993). The particular conditions used by Li *et al* (1993) were chosen because they yielded a high proportion of longer repeats, many of which were in coding regions, whereas conditions used by Riggins *et al* (1992) gave smaller repeats, which were frequently found in 5' untranslated regions (UTRs) of mRNA. A summary of the results can be found in Table 3.1.

The mouse embryonic cDNA libraries, 8.5 dpc (Hogan *et al*, unpublished) and 12.5 dpc (Logan *et al*, 1992) were screened in tandem, because they share the same vector, λgt10. All clone numbers are given as follows: plate numbers 1-5 were plaques derived from the 8.5 dpc mouse whole embryo cDNA library; plates 6, 7, 8 and 9 contained clones from the 12.5 dpc mouse whole embryo cDNA library. The remaining numbers in the clone names are arbitrary and relate to the order in which the plaques were picked from the individual plate. For instance, Clone 210 is derived from plate 2 and is the tenth clone picked from that plate.

Fifteen thousand plaques from the 8.5 dpc mouse embryo cDNA library (B. Hogan *et al*, unpublished) were screened and 44 replicating signals were identified. This translates to a percentage occurrence of 0.29%, relative to the total plaques screened.

Table 3.1

Library	*Vector*	*p.f.u. screened*	*primary positives*	*percentage occurrence*
8.5 dpc embryonic mouse	**λgt10**	**15000**	**44**	**0.29**
12.5 dpc embryonic mouse	**λgt10**	**25000**	**19**	**0.076**
13.0 dpc embryonic mouse	**λunizap**	**20000**	**33**	**0.165**
adult brain mouse	**pSPORT**	**20000**	**53**	**0.265**

Table 3.1 Summary of results of primary library screens. The results are discussed in this Chapter. p.f.u. refers to the number of plaque forming units; primary positives to the signals found in the primary screen of the libraries and percentage occurrence is the number of primary positives, as a proportion of total plaque forming units screened. Further details of the individual libraries, the vectors and hybridisation conditions of the library screen can be found in Chapter 2, Section 2.20.

Twenty-five thousand plaques from the 12.5 dpc mouse whole embryo cDNA library (Logan *et al* , 1992) were screened and 19 replicating signals were identified in the primary screen. This equates to a percentage abundance of 0.076%.

A 13.0 dpc whole mouse embryo library was obtained from Nick Allen in Cambridge (unpublished) and was screened to check that the reduction in numbers obtained from the 12.5 dpc mouse embryonic library compared to the 8.5 dpc mouse embryonic library was real (Logan *et al*, 1992). Twenty thousand plaques were screened and 33 replicating signals were identified in the primary screen. Thus, 0.165% of the recombinant phage screened contained putative CAG/CTG triplet repeats.

An adult mouse brain cDNA library (Meier-Ewert *et al*, unpublished) was also screened. Fifty-three replicating signals were identified in the primary screen. Twenty thousand clones were screened and this equates to a 0.26% incidence of CAG/CTG triplet repeat containing inserts.

3.3 Subcloning of phage inserts from the 8.5 and 12.5 dpc whole mouse embryo libraries into the phagemid pBluescript.

To enable further manipulation and analysis of the putative trinucleotide repeat-containing cDNA clones described in Section 3.2, the inserts putatively containing CAG/CTG repeats were sub-cloned into the phagemid vector, pBluescript KS- (Stratagene). A shot-gun sub-cloning strategy was adopted. This was because of the large numbers involved. This strategy makes use of general observations about cloning experiments. For example, it is known

that small DNA molecules are preferentially cloned compared with larger DNA molecules and secondly, lambda arms will not be sub-cloned as they have a single restriction enzyme cut at one end. The other end of these fragments contains the lambda phage specific sticky end (cos) and is not compatible with restriction enzyme cleavage sites in a ligation reaction. Because of the procedure's simplicity, this strategy also allowed for the rapid cloning of many inserts in a set of parallel experiments which are easy to set up. Details of the procedure can be found in Section 2.16 of Chapter 2.

Fourteen sub-cloned inserts were used in the preliminary analysis of gene expression described in the next section of this chapter.

3.4 Analysis of gene expression detected by selected clones from the 8.5 and 12.5 dpc mouse whole embryo cDNA libraries during development and in adult brain and liver.

This section describes the initial work to identify developmentally expressed clones from fourteen of the initial sub-cloned inserts from the 8.5 and 12.5 dpc whole mouse embryo libraries. This group represents the first clones purified and sub-cloned in the phagemid vector, pBluescript KS- (Stratagene).

3.4.1 Results of reverse northern experiments.

Reverse dot blot experiments were performed to identify potentially developmentally regulated genes. The clones used here were the first 14 positive lambda clones purified and shotgun sub-

cloned (Section 3.3) into the phagemid vector, pBluescript KS- (Stratagene), from the 8.5 dpc (Hogan *et al*, unpublished) and 12.5 dpc (Logan *et al*, 1992) whole mouse embryo libraries. Those used from the 8.5 dpc library were Clones mCTG 23, 24, 26, 28, 210, 43, 45, 411, 56 and 59. Clones mCTG 61, 63, 82 and 86 derive from the 12.5 dpc cDNA library screen. One microgram (spectrophotometrically determined) of denatured plasmid DNA containing cloned inserts was fixed to nylon membrane strips (Dupont Genescreen^{+}). The membranes were probed with radioactively labelled heterogeneous first strand cDNA, synthesised from total RNA extracted from 8.5 dpc, 12.5 dpc and 17.5 dpc mouse embryos (Chapter 2, Section 2.8.4). A probe was also prepared from embryonic stem (ES) cells (Doetschman *et al*, 1985). This probe represents a time in development of the embryo where little or no organogenesis has taken place. ES cells represent totipotent cells from the inner cell mass of the pre-implantation embryo and are equivalent to 3.5 dpc (Doetschman *et al*, 1985). Probes were also prepared from two adult mouse tissues, brain and liver. The results are shown in Figure 3.1 and quantitatively (densitometric readings) in Figure 3.2.

A total of 5 clones (mCTG 23, 26, 411, 63 and 82) have a high level of steady state mRNA expression and show some variation in expression during development. mRNA from three clones (mCTG 26, 63 and 82) was highly abundant at all developmental time points tested (including ES cells). These genes are also expressed in adult tissues. Of these three clones, mCTG 26 and 63 have reduced expression levels compared to Clone mCTG 82 in the adult tissues. In addition, messenger RNAs from mCTG 26 and 63 have lower abundance in adult liver than in brain. mRNA from Clone 82 is

Legend to Figure 3.1: Reverse dot-blot analysis of expression of selected clones from the 8.5 dpc and 12.5 dpc whole mouse embryo libraries (opposite page).
Denatured plasmid DNA was slot-blotted on to nylon membrane and probed with radioactively-labelled heterogeneous first strand cDNA derived from total RNA extracted from ES cells (3.5dpc); 8.5, 12.5 and 17.5dpc whole embryos; adult mouse brain and liver. horizontal series of numbers represent blotted DNAs from 1 pBluescript KS-; 2 GluT, glutamine transporter; 3 GABA receptor subunit g2; 4) GABA receptor subunit a6. Clones mCTG 23, 26, 411, 63 and 82 are indicated.

Figure 3.1

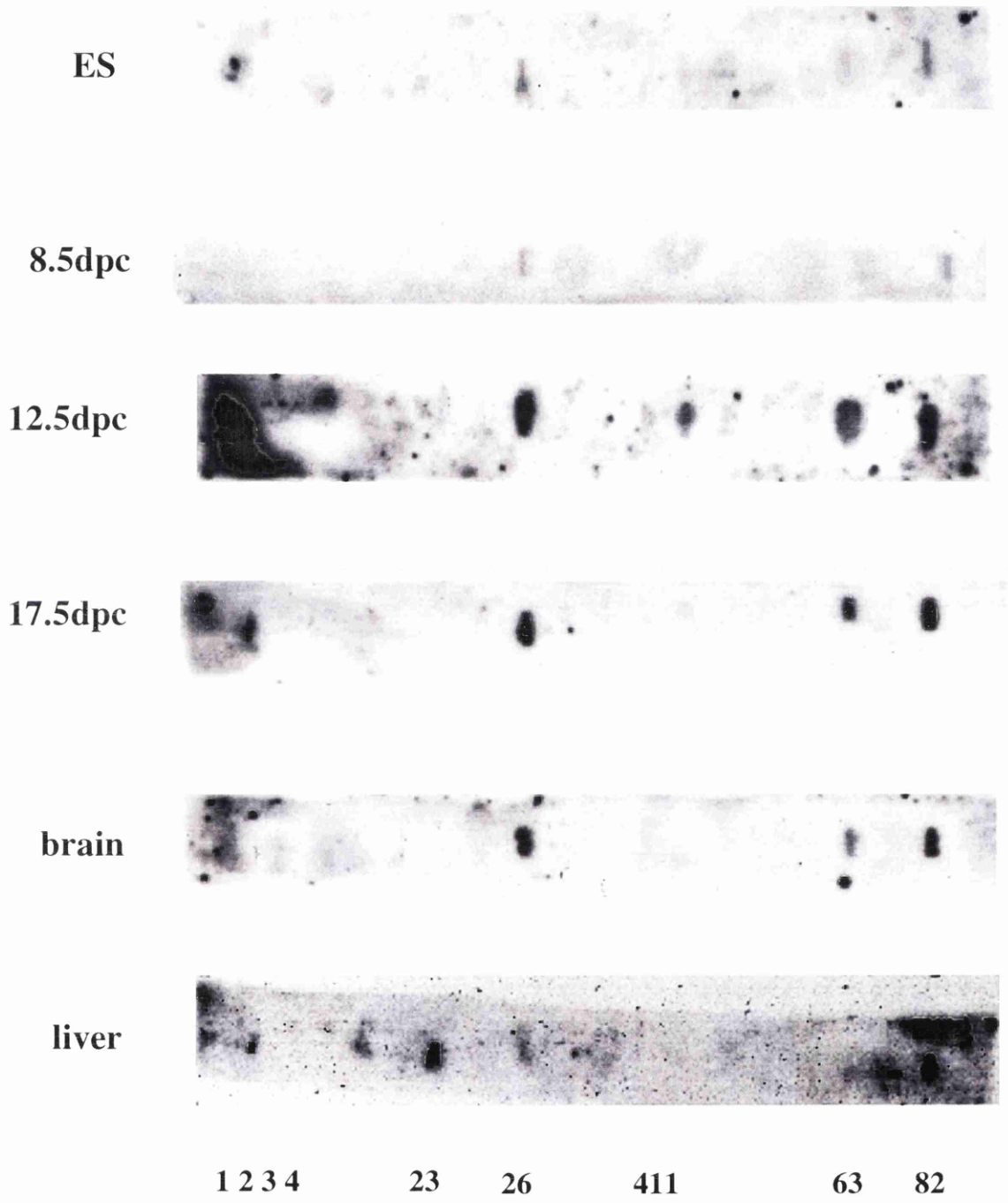


Figure 3.2

0.02	0.01	0.05	0.02	0.03	0.63
0.92	1.05	1.26	0.97	1.03	0.18
0.09	0.04	0.33	0.11	0.04	0.01
0.24	0.11	0.44	0.59	0.28	0.15

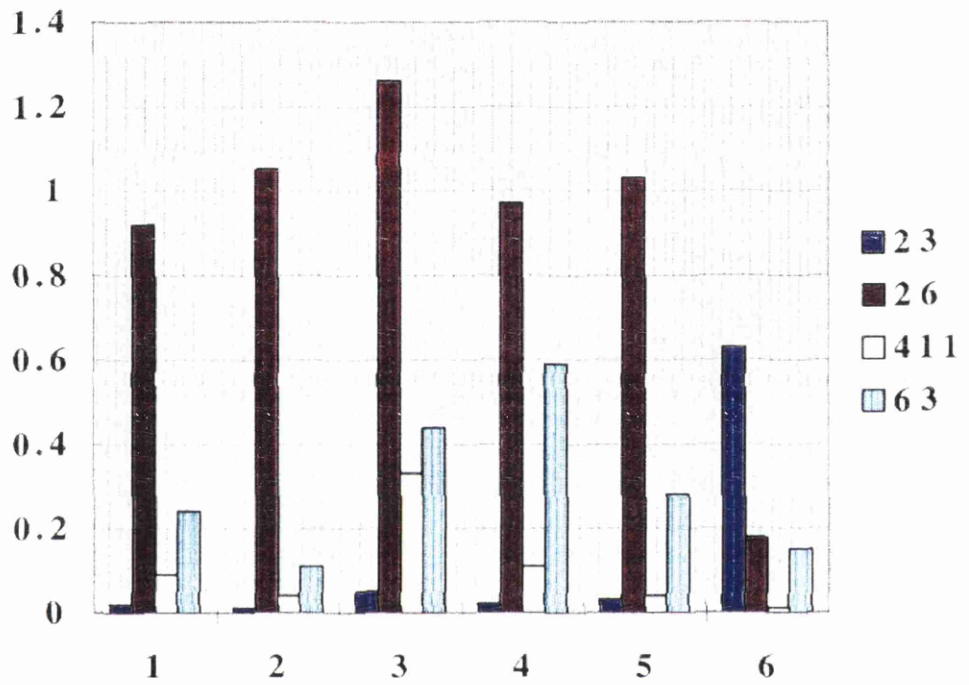


Figure 3.2 Expression of Clones mCTG 23, 26, 411 and 63. The expression of the Clones mCTG 23, 26, 411 and 63 were measured by MacBas imaging programme (Fuji) and their level of expression relative to the expression of Clone mCTG 82 was calculated (vertical axis, a value of one equals the level of expression of Clone mCTG 82; see also top section of this Figure). The horizontal axis refers to the source of the total RNA from which the first strand cDNA probes originated. These were; 1) embryonic stem cells, 2) 8.5 dpc whole mouse embryos, 3) 12.5 dpc whole mouse embryos, 4) 17.5 dpc whole mouse embryos, 5) adult mouse brain and 6) liver.

expressed consistently in all developmental stages and adult tissues tested. This was used as a base line to calculate the abundances of the four clones (mCTG 23, 26, 411 and 63), relative to the expression of Clone mCTG 82. mRNA from Clone mCTG 23 can only be detected in adult liver. Clone mCTG 26 appears to equal or exceed that of Clone mCTG 82 throughout development and in adult brain. However there is a reduction in the liver. mRNA from Clone mCTG 411 is present during development and is more abundant at 12.5 dpc. It was not detectable in adult tissues. mRNA from all other clones tested were undetectable by probes from any developmental time point or adult tissue probes used (brain and liver). The cDNA clones used as post 8.5 dpc positive controls (GABA receptor subunit, $\gamma 2$ (David Livingston, personal gift) and the glutamate transporter, GluT, (M. Sutherland, personal gift) failed to give signals. The negative controls, the parent phagemid, pBluescript KS- acted as a control for background. The GABA receptor, $\alpha 6$ was included as another negative control, except for post natal cerebellum, where its expression is restricted to cerebellar granule neurones. Both of these yielded no signal in this set of experiments.

3.5 Discussion and analysis.

3.5.1 Library screens.

The primary screens of the mouse embryonic cDNA libraries with an oligonucleotide, $(CTG)_{10}$, yielded large numbers of positive signals. This is true for the adult mouse brain cDNA library as well. However the numbers of clones isolated from each library are not the same. A three fold reduction in positive clones is observed from the 12.5 dpc whole mouse embryo cDNA library, in terms of

percentage occurrence, compared to the percentage occurrence in the 8.5 dpc whole mouse embryo cDNA library. In the 13 dpc whole mouse embryo library (Nick Allen, Cambridge) a reduction compared to the 8.5 dpc mouse embryo cDNA library (Hogan *et al*, unpublished) is also observed, but only 1.5 fold. Between the 12.5 and 13 dpc whole embryonic cDNA libraries, there is a two fold increase in observed percentage occurrence of trinucleotide positive replicating signals. The incidence for the mouse adult brain cDNA library (0.26%) is similar to the incidence of replicating signals in the 8.5 dpc mouse embryo library (0.29%). What is the explanation of this changes? There are various possibilities to be considered. These explanations are not mutually exclusive and may overlap in contribution to the phenomenon.

First, these data might just reflect qualitative differences between the individual libraries used for this study. Influencing factors to be considered here are the type of vector used (*i.e.* lambda vectors or plasmids, or even different lambda vectors), and the bacterial strains; *e.g.* whether these tolerate simple repetitive sequences in plasmid vectors. It is known that *E. coli* strains which are deficient in the Mut proteins MutS, MutL and MutH are more tolerant of triplet repeats (Kang *et al*, 1995; Bowater *et al*, 1996 and Rosche *et al*, 1996). In addition it has been reported that other proteins which are involved in mutation repair, replication and recombination show varying effects on repeat stability. These genes include recA, ssb and gyrase (Bowater *et al*, 1996 and Rosche *et al*, 1996).

The method of priming the first strand cDNA product may also have an effect on numbers of clones containing trinucleotide repeats. For example, if CAG/CTG trinucleotides were

disproportionately found in the 5' end of genes, oligo-dT primed libraries would be under-representative in terms of trinucleotide repeats. This, however, is unlikely as there is a large difference between the whole embryo 8.5 dpc, whole embryo 12.5 dpc and mouse adult brain cDNA libraries which are all oligo-dT primed. In addition, the mouse adult brain cDNA library (which is also oligo-dT primed) appears to contain as many CAG/CTG triplet repeats as described for the human libraries (Li *et al*, 1993; Riggins *et al*, 1992). Interestingly, a recent study by Chambers and Abbott (1996) who screened an independent mouse adult brain cDNA library, which was oligo-dT primed, yielded the same proportion of positives as the mouse adult brain cDNA library described here. In the Chambers and Abbott (1996) library, 60 clones per 20000 p.f.u. screened was observed, whereas 53 clones (of 20000 p.f.u. screened) were identified in the mouse adult brain cDNA library screened in this work. This can be taken as a guide to the effectiveness of the experiments detailed here because the numbers of CAG/CTG repeat containing clones isolated from two adult mouse brain cDNA libraries, relative to the total number of clones initially screened in two independent experiments, is similar.

Another interesting point to consider is that this reduction is not the result of changes in expression levels of genes, but a dilution of transcripts of genes from one specific region or organ. The brain, which is known to be a rich source of CTG containing genes (Li *et al*, 1993; Riggins *et al*, 1992), becomes proportionally less in terms of overall volume of the developing embryo. If the levels of transcriptional activity are assumed to be constant, a reduction in positive clones as a percentage of total clones would be observed progressively throughout development.

Taken at face value, the level of primary replicating positives for the mouse adult brain cDNA library screening (0.265%) is similar to that observed for the human foetal and adult (cerebral cortex) brain cDNA library screens, conducted by Li *et al* (1993) and Riggins *et al*, (1992). Respectively, the results in terms of percentage replicating positives for these human experiments were 0.26% and 0.28% for the foetal brain (Li *et al*, 1993) and the cerebral cortex (Riggins *et al*, 1992) libraries. A direct comparison cannot be made because of the differing regions and time points that separate the mouse and human data. However, it seems that brain cDNA libraries from both species are in general a rich source for CAG/CTG trinucleotide repeats. What is the basis of this? Is it that trinucleotide repeat containing RNA is selectively isolated more readily from this source than from other tissues; this seems unlikely as it is known that RNA is more difficult to obtain from brain than from other tissues. An alternative answer to the higher numbers of CAG/CTG repeat-containing clones found in brain cDNA libraries may lie in that the brain is very complex in terms of number of cell types. The larger number of different cells and the larger the brain specific repertoire of genes, leads therefore in a random model of trinucleotide occurrence in genes, to an increased number of trinucleotide repeat containing positives in a library screen.

In addition to brains being a rich source of CAG/CTG repeat containing inserts, from the results presented here, mouse embryo cDNA libraries are an additional rich source of CAG/CTG triplet repeat positive clones. This has provided a useful pool of cloned genes to analyse in terms of development.

It has been observed in the human cDNA library screening studies that brain cDNA libraries yield more CAG/CTG containing

clones (Riggins *et al*, 1992). The results presented here are consistent with this observation. The mouse adult brain cDNA library yielded as many positives as the human brain libraries, as well as the 8.5 dpc whole mouse embryonic library screened in this work. The whole embryo cDNA libraries also contain a significant proportion of cDNAs derived from the developing brain. It is possible to suggest that a large proportion of the embryo derived transcripts are expressed in the brain. To test this a set of experiments was required to determine the transcriptional activity of positive clones derived from embryo library screens was constant and if they were expressed in adult tissues, specifically the brain. These experiments are discussed below.

3.5.2 Analysis of reverse dot blot experiments.

The aim of the reverse dot-blot series of experiments was to identify genes which had variable levels of mRNA transcription levels during development, implying developmental regulation of their expression. In addition, two other tissues were looked at to test for differential expression of the cloned genes in the adult mouse. Initial expression analysis of selected clones by reverse dot-blotting showed that mRNA from five of the Clones, mCTG 23, 26, 411, 63 and 82, was highly abundant. Furthermore some variation in abundance of mRNA was observed between the different developmental stages and also between brain and liver from adult mouse. Four of the Clones (mCTG 26, 411, 63 and 82) do not appear to be associated with organogenesis because their expression level is the same when tested with pre-organogenesis first strand cDNA probe, derived from ES cell RNA. Some have significant differences

in messenger levels between adult tissue types. Clones mCTG 26 and 63 are more abundant in the adult brain compared to adult mouse liver. There is no detectable expression of Clone mCTG 23 messenger RNA at the developmental time points tested or in adult brain of the mouse. However, Clone mCTG 23 mRNA was detected in the liver. This suggests that the gene product of Clone mCTG 23 may have a role in the adult liver. The expression of the other clones tested was below the level of sensitivity of these experiments. Other experimental procedures will have to be used to analyse the level of expression of these clones. For example PCR based technology with increased sensitivity, could be used. RT-PCR using first strand cDNA products as templates to drive PCR could detect very rare products. This technique however requires that sequence information derived from the clones is available to design suitable primers for PCR. This is quickly done for one or a few clones, but laborious for many clones. Run-on experiments could be used to detect transcriptional activity, but requires the isolation of nuclei from different cell types. This is easier for certain cell types than for others. For example, nuclei of white blood cells can be easily retrieved but nuclear isolation from neurones is difficult because of the large proportion of extra-cellular material and lipid membranes from glial cells.

Because the reverse dot-blot experiment did not give information about the numbers of distinct messenger RNA species transcribed from a specific gene, it cannot be ruled out that this apparent difference in expression levels may be due to alternative splicing. The result of this would be the reduced representation of individual or sub-groups of messenger species containing sequences with the up-regulation of others.

The controls described here, (*i.e.* GluT and γ2) are expressed beyond 8.5 dpc embryos, specifically in the brain (GluT, M. Sutherland, Pers. Comm.; γ2, Luntzleybman *et al*, 1993). This should allow for selection of clones displaying pre and post 8.5 dpc expression, because it is observed that there is an increase in differentiation of the nervous system at this stage (Kauffman and Kauffman, 1992). However these genes are not widely or abundantly expressed and in retrospect it would have been better to include moderately and highly expressed well characterised genes as positive controls so as to have a benchmark for measuring the results (*e.g.* GAPDH or β-actin). The GABA receptor subunit α6 is negative apart from the post natal cerebellum (R.W. Davies Pers. Comm.), where it is expressed in low amounts. However, Clone mCTG 82 is consistently expressed through all developmental stages and tissues tested, relative to the activity of the probe in each case. This clone was used as a standard to test differing levels of expression amongst the other clones which were detected in this series of experiments. The background control, the phagemid pBluescript KS- (Stratagene) did not return any signal, therefore it must be concluded that the signals that are detected from hybridisation of probe to the fixed plasmid containing the cDNA inserts are genuine.

Interestingly, the genes corresponding to Clones mCTG 26, 63 and 82 were shown to have detectable amounts of mRNA in development, were also expressed in the brain. This correspondence of developmental and brain expression has been noted also for *Drosophila* proteins with large homopeptides (Karlin and Burge, 1996). They postulate that fewer mammalian genes contain these continuous runs of the same amino acid, because of a negative bias

due to the fact that relatively few developmental or regulatory genes have been sequenced in mammals, compared to *Drosophila.* which has been studied extensively by molecular genetics in development. A screen of mammalian cDNA libraries would therefore identify novel genes involved in development. To assess this, the putative CAG/CTG triplet repeat positives described above were subjected to sequencing analysis to identify novel sequences and characterise the repeat loci of these inserts. That is the subject of the next Chapter.

Chapter 4

Sequence Analysis Of CAG/CTG Repeat Containing cDNAs.

4.1 Introduction

In Chapter 3, a number of recombinant clones, from the 8.5 dpc (Hogan *et al*, unpublished) and 12.5 dpc (Logan *et al*, 1992) mouse embryonic cDNA libraries were sub-cloned. These were purified by consistent hybridisation to a trinucleotide repeat probe, $(CTG)_{10}$. Some of the first sub-cloned inserts were shown to have high expression during development and in adult tissues. To establish if these clones indeed contain CTG trinucleotides, DNA sequencing was performed to identify the repeat(s) in each clone. Some clones may have multiple independent repeat units. This approach was used by both previous human studies (Riggins *et al*, 1992; Li *et al*, 1993) to establish the existence of and identify triplet repeats. The unique sequence which flanks the repeat can be used to identify related sequences found in the large database repositories. The last point is important because we can identify repeats which are conserved between species. This is important when trying to establish if the mouse could be used as a model, in respect to possible diseases. The most relevant cross-species analysis would be to human sequences. This is because man is the only species which has been shown to date to develop naturally diseases associated with trinucleotide repeat expansion. Any cross-species conservation would help in assessing the possibility of using the mouse genes as models for triplet repeat expansion diseases and for studying the role of the triplet repeat in normal function of the gene products.

The two human studies also found a number of novel genes containing CAG/CTG trinucleotide repeats. It is likely that a proportion of the clones sequenced will belong to previously un-

characterised genes. More recently, a new study by Abbott and Chambers (1996), identified a number of trinucleotide repeat containing clones from a mouse adult brain cDNA library. This provides a source of material which can be used as a comparison for the mouse sequence data presented in this Chapter.

4.2 Partial nucleotide sequence of sub-cloned inserts from the 8.5dpc and 12.5dpc whole mouse embryo cDNA libraries.

As described in Chapter 3, clones were isolated from two mouse embryo cDNA libraries. These libraries were derived from RNA from 8.5 and the 12.5 dpc embryos. Some purified lambda positives were sub-cloned into pBluescript KS- phagemid (Stratagene). Due to the large number of potential clones, it was decided to first sequence from the ends of the clones, using standard primers which flank the multiple cloning site of pBluescript. These clones were sequenced first using a manual T7 polymerase kit (Pharmacia), following Sanger's di-deoxynucleotide method (Sanger *et al*, 1978). Double or single-stranded DNA templates were used. The automated sequencing system from Applied Biosystems (Perkin Elmer) was used subsequently to extend the sequence data. For manual sequencing, T3 and T7 primers (Promega) were used; in the case of automated sequence generated data, the extended primers T3/T7α (BRL-Life Technologies) and bT7 (Stratagene) were utilised. The sequences of the four primers described above are described in Chapter 2, Table 2.3.

4.3 Initial manipulation of sequence data.

DNA sequences were analysed using MacVector (IBI) and Gene Jockey version 2 (Cambridge Bioscience) sequence analysis packages. Searches using the nucleic acid databases, Genbank and EMBL were done using UWGCG (Devereux *et al.*, 1984) Version 8.0. To avoid interference with trinucleotide-containing sequence from the database, comparison with sequences from outside the repeat region were used to conduct searches against the databases.

4.3.1 Analysis of Repeats Identified.

Repeats were identified in a large proportion of clones sequenced. Table 4.1, which lists the repeats discovered by this study, shows that CTG/CAG trinucleotide repeats were successfully identified in 21 of the first 23 clones. Further sequencing should identify repeats in the remaining two clones. There was a wide range of sizes of repeats found, ranging from 6 up to 26 (see Figure 4.1) trinucleotides per repeat tract. The average trinucleotide number per independent repeat is calculated to be 13 (39 nucleotides), which is greater than the size of the oligonucleotide probe used in the library screening, which was 30 nucleotides in length ($(CTG)_{10}$).

Three clones have multiple CTG/CAG type repeats. These are clones mCTG 28, 210 and 92. The repeats all tend to be in the same strand and exist as the same subset of repeats. For example a CAG repeat will be followed by CAG/AGC/GCA. This is consistent with the multiple repeats in some human genes found by Li *et al* (1993).

Figure 4.1 CTG repeats identified from clones derived from 8.5 and 12.5 mouse embryo cDNA libraries.

Repeat sizes are derived from sequence data for individual mCTG clones, as described in Section 4.2. Commas indicate discontinuous repeats. Brackets enclose cryptic repeats. + indicates an indeterminate repeat number due to template anomalies.

CLONE	(CTG)n
Single repeats	
2 4	4
5 6	6
1 2	7
4 1 0	8
4 5	8
8 2	9
5 9	1 1
2 3	1 2
4 1 1	1 2
8 1	1 5
4 6	2 5
6 1	2 5
6 3	2 6
Multiple repeats	
8 6	7,3+ and 3+
5 7	6 7
2 8	(1,7,2,4) and (1,2,4)
9 2	(1,3,1,11) (2,5) (1,1,4)
2 1 0	(1,7,2,4)
Cryptic repeats	
4 1 4	(7,1)
2 6	(7,2,2) (1,2,4)
4 3	(2,9)

Figure 4.2 clones with similarity to other genes with proteins of defined function.
Homology was identified using FASTA or BLAST programmes to search non-redundant nucleic acid and protein databases. Similarity indicates the gene to which the mouse sequences are related.

Clone	Similarity	Function	Reference
2 3	70kDa subunit	**ssDNA** binding	Erdile *et al*, 1991
2 4	nuclear receptor co-receptor	**dsDNA** binding	Horlein *et al*, 1995
2 9	ASF-2 like sequence	**RNA** binding	Wang and Manley, 1996
4 3	a) Rad21 b) PW29 c) β-1,4,-galactosyl-transferase	unknown $Ca2^{+}$ binding enzymatic	McKay *et al*, 1996 Yu *et al*, 1995 Nakazawa et al, 1988
4 1 0	rat nucleoporin, P54	**RNA** binding	Hu *et al*, 1996
4 1 1	rat nucleocytoplasmic protein	**RNA** binding	Meier and Blobel, 1992
5 6	a) CW17 mouse protein b) Zfm1 transcription factor c) SF1 splicing factor	unknown **dsDNA** binding **RNA** binding	Wrehkle *et al*, 1993 Toda *et al*, 1994 Arning *et al*, 1996
5 7	a) P300 transcriptional adaptor b) CREB binding protein	**dsDNA** binding **dsDNA** binding	Eckner *et al*, 1994 Chrivia *et al*, 1993
6 3	JC310, suppressor of RAS	RAS binding	Colicelli *et al*, 1993

Legend to Figure 4.3. Summary of alignment analysis of clones which show similarity to other sequences.

The degree of similarity (% homology) to other sequences is expressed as a percentage of the total region used in the analysis. NA indicates a comparison made between nucleic acid sequences and AA indicates a comparison between amino acid sequences. The columns headed comparison region is the size of the region over which the % similarity was calculated. The Species column indicates the origin of the comparison sequence (Mm being *Mus musculus; Hs*, *Homo sapiens; Rn, Rattus Norvegicus*; *Xl, Xenopus laevuus*). Note also that the sequence analysis of clone mCTG 24 is split into different segments to indicate the likelihood that this comparison is made between two related genes from the same species, rather than the same gene because of the higher degree of similarity between these sequences within the coding, compared to the putative 5' UTR sequences. The appendix to clone mCTG 43 shows the overall degrees of similarity of *Mus musculus* Rad21 to the three other sequences which show similarity to clone mCTG 43. Note also that Zfm1 isoform III is omitted from the analysis of clone mCTG 56 as these sequences do not overlap with each other.

Clone	Homology	Species		% Homology		Comparison Region		Insert Size	% Sequenced
				NA	AA	NA	AA		
mCTG 23	Replication Protein A 70 kDa subunit	*Hs*		-	86.5		80	2200	10
		Xl		-	71.2	-	80		
mCTG 24	Steroid Receptor Co-repressor	*Mm*							
			Pre-ATG	61.5	-	117	-		
			Post-ATG	85.7	-	182	-		
			Post Insert	82.5	-	269	-		
			Total	79.4	-	587	-	800	38
mCTG29	H92640	*Hs*		91.4	-	328		2-3000	undetermined
	AA053667	*Hs*		73.6	-	228			
	T55469	*Hs*		78.1	-	228			
	N39431	*Hs*		76.1	-	247			
	T63434	*Hs*		75.0	-	232			
	W39162	*Hs*		76.6	-	248			
	H50597	*Hs*		76.6	-	248			
	H94634	*Hs*		76.6	-	248			

Clone	Homology	Species		% Homology		Comparison Region		Insert Size	% Sequenced
				NA	AA	NA	AA		
mCTG 43	Rad 21	*Mm*		96.5		565		565	100
	Rad21	*Hs*		79.5		565			
	PW29	*Mm*		78.1		406			
	β-1,4-G	*Mm*		82.5		472			
Appendix mCTG43 Data									
Mm Rad21	HsRad21	*Hs*		82		1067			not applicable
	PW29	*Mm*		100		1905			
	β-1,4-G	*Mm*		98		1905			
mCTG 410	p54 nucleoporin	*Rn*		95.1	97	532	178	2-3000	undetermined
mCTG411	Nopp140	*Rn*		86.6	85.5	269	76		22.4%
	Nopp140	*Hs*		74.7	65.5	269	76		
	Nopp130	*Hs*		76.5	64.5	269	76		

Clone	Homology	Species		% Homology		Comparison Region		Insert Size	% Sequenced
				NA	AA	NA	AA		
mCTG 56	CW17R	*Mm*		-	100	-	86	300	100
	Zfm-1-II	*Hs*		-	100	-	86		
	Zfm-1-I	*Hs*		-	100	-	65		
	SF-1 hl1	*Hs*		-	100	-	65		
	SF-1 bo	*Hs*		-	100	-	65		
mCTG57	P300	*Hs*		-	83.1	-	120	1100	32.7
	CBP	*Hs*		-	65	-	120		
	CBP	*Mm*		-	65	-	120		

Clones showing similarity to ESTs

Clone	Homology	Species		% Homology		Comparison Region		Insert Size	% Sequenced
				NA	AA	NA	AA		
mCTG12T3	Z78337	*Hs*		83.5	-	400	-	2000	20
	Z42809	*Hs*		84.7	-	216	-		
	R17799	*Hs*		83.2	-	215	-		
	R91592	*Hs*		83.7	-	215	-		
mCTG12T7	W71508	*Hs*		76.1	-	297	-		
mCTG46	W91417	*Mm*		96.5	-	344	-	344	100
mCTG59	W86172	*Hs*		78.6	-	131	-	1500	8.73
mCTG 63									
3'	AA106210	*Mm*		86.2	-	405	-	2500	40
	AA145456	*Mm*		86.7	-	347	-		
	N79067	*Hs*		74.8	-	250	-		
	R40976	*Hs*		75.6	-	250	-		
	T16564	*Hs*		75.6	-	250	-		
	Z41184	*Hs*		75.6	-	250	-		
	AA199815	*Hs*		75.6	-	250	-		
	W93359	*Hs*		75.6	-	250	-		
	N98931	*Hs*		76.0	-	250	-		
	W81207	*Hs*		76.0	-	250	-		

Clone	Homology	Species		% Homology		Comparison Region		Insert Size	% Sequenced
				NA	AA	NA	AA		
mCTG 63									
3' (continued)	H25927	*Hs*		76.4	-	250	-		
	H27668	*Hs*		76.0	-	250	-		
	N65975	*Hs*		75.6	-	250	-		
	H28318	*Hs*		75.6	-	250	-		
	H05633	*Hs*		74.8	-	250	-		
	R44135	*Hs*		75.2	-	250	-		
	H24291	*Hs*		74.4	-	250	-		
	N62312	*Hs*		73.6	-	250	-		
5'	JC310	*Hs*		85.6	87.6	550	162	2500	40

Figure 4.4 Comparison of repeat numbers in different species. Numbers indicate the size of the repeat in cognates of the same or similar protein in different organisms. Numbers that are interrupted by commas indicate intervening amino acids which represent deviations from the trinucleotide repeat. Asterices indicate that the information on the cognate protein in that organism is not available. Column N represents the CAG/CTG repeat size identified in sequence identified in this study. The mouse column represents previously characterised mouse sequence found in the GenEMBL database.

		Repeat Number				
clone	Protein or NA	N	rat	mouse	human	Xenopus
2 3	RepA	1 2	*	*	1	1
4 3	Rad21	2,9	*	2,4,1	2,1,3,1	*
	PW29		*	2,4,1	*	*
4 1 0	Nopp140	8	8	*	*	*
5 6	CW17	6	*	6	*	*
	Zfm1	*	*	*	6	*
	SF1	*	*	*	6	*
5 7	P300	6	*	*	2	*
	CREB-BP (human)	*	*	*	3	*
	CREB-BP (mouse)	*	*	3	*	*
1 2	EST W78337	7	*	*	6	*

It is not possible to gauge the overall proportion of clones that contain multiple repeats because sequencing has not been pursued to fill in the gaps of sequence. Clone mCTG 43 contains, in addition to a CTG trinucleotide, a GAG type repeat which is adjacent to a hexanucleotide repeat, GAGAAG (Figure 4.8). Clone 92 has a high probability of encoding a polyproline stretch (CCN) preceding the first CAG trinucleotide.

A proportion of these clones correspond to novel genes. Of the initial 23 clones analysed, which were chosen arbitrarily from the 8.5 and 12.5 dpc whole mouse embryo cDNA libraries, there were found to be 11 sequences which had no similarity/identity to any known sequence. A further three sequences had similarity to Expressed Sequence Tags (ESTs). These were: clone mCTG 12, with similarity to a group of related human ESTs (see Figure 4.14); mCTG 46, with similarity to a single mouse EST (see Figure 4.15) and mCTG 59, with similarity to a single human EST (see Figure 4.16). Within the clones with similarity to known gene sequences, 9 had similarity to proteins of known function. These are clones mCTG 23 (replication protein A, 70 kDa subunit), mCTG 24 (mouse nuclear receptor co-repressor); mCTG 29 (alternative splicing factor), mCTG 43 (Rad21, beta-1,4-galactosyltransferase and PW29 calcium binding protein); mCTG 410 (nucleoporin p54); mCTG 411 (Nopp140 nucleophosphoprotein); mCTG 56, (Zfm1 transcription factor, splicing factor 1 and CW17); mCTG 57 (P300 cellular transcriptional activator) and mCTG 63 (JC310 RAS protein inhibitor). A detailed account of each clone and the genes which they are related to is given in Sections 4.4 and 4.5 of this Chapter. Alignments for each clone are shown in Figures 4.5 to 4.16.

Amongst these 9 clones with similarity to genes encoding proteins of known function, 7 are nucleic acid binding proteins. Two of these are known to bind double-stranded DNA (mCTG 24, nuclear receptor co-repressor; mCTG 57, P300 transcriptional adaptor protein), three are RNA-binding proteins (mCTG 29, ASF-2 like gene; 410, nucleoporin p56 and mCTG 411, nucleocytoplasmic protein Nopp140) and one is a single-stranded DNA binding-protein (mCTG 23, 70kDa subunit of replication protein A). The remaining sequence, mCTG 56, has both similarity to a double-stranded DNA binding protein (Zfm1, a transcription factor) and RNA binding protein (splicing factor 1). Clone mCTG 63 was chosen for further study and will be discussed further in Chapter 5.

A comparison of the occurrence and location of trinucleotide repeats in these genes between species, particularly with man, could be useful in assessing the likelihood of mouse trinucleotide repeats acting as models for human disease. In five of the six examples in which a cross-species comparison could be made, human sequences were available to make such a comparison. In all these examples the mouse repeat is larger (clones mCTG 12, 23, 43 and 57) or equal (clone mCTG 56) in size to the human repeat. A summary of these comparisons are described in Figure 4.4. Within the examples where the mouse trinucleotide is larger there is a spectrum of differences. For example, in the case of clone 12 there is only 1 triplet difference compared to the human sequence whereas for clone mCTG 23 (which shows similarity to the 70kDa subunit of replication protein A), the mouse CAG trinucleotide is 11 repeats larger than both the human and *Xenopus* sequences, which only have one CAG trinucleotide, encoding a single glutamine, at the equivalent position in their proteins (see Figure 4.4 and 4.5). In the

case of clone mCTG 410, which displays similarity to a rat gene nucleoporin p54, there is no change in the number of repeats (8).

Novel sequences are classed into 2 subgroups: those which are genuinely unknown (*i.e.* those which show no similarity to any sequence in the nucleic acid databases) and those which are similar to unidentified, incomplete sequences in the database deriving from large total library sequencing projects. For this study, most are assessed to be novel, since no known function can be assigned to the gene product. One exception is mCTG 29 which has similarity to a group of ESTs which have similarity to the splicing factor ASF-2. This is likely to correspond to a new splicing factor.

4.4 Clones with similarity or identity to other sequences.

This section describes the clones which have similarity to or are identical to sequences already deposited in the nucleic acid (Genbank, dBEST and EMBL) or protein (PIR and Swiss-Prot) databases. Similarities were elucidated by using the sequence generated from the clones, in conjunction with the sequence analysis programmes, FASTA (Lipman and Pearson, 1985) and BLAST (Basic Local Alignment Search Tool; Altschul *et al*, 1990) to search nucleic acid and protein databases. The individual details for each case are outlined below and summarised in Figures 4.2 and 4.3.

4.4.1 Clone mCTG 23.

This gene was isolated from the 8.5 day cDNA library. It was found to have significant similarity to a gene encoding the 70 kDa

polypeptide subunit of a human single-stranded DNA-binding protein (Kim *et al*, 1992), replication protein A (Erdile *et al*, 1991). This is the largest of three associated polypeptides and has been shown to contain the DNA binding property (Kim *et al*, 1992). Furthermore in a model system, the activation of the SV40 DNA replication reaction, this subunit cannot itself drive the reaction but has been shown to interact with other proteins, including the DNA polymerase α-primase complex (Dornreiter *et al*, 1992). The heterotrimeric complex has been implicated in processes within cells which require the DNA to be transiently single-stranded, namely replication (Erdile *et al*, 1991), and repair (Liu and Weaver, 1993). It may also have a role in transcription and recombination. This protein is probably an essential gene, as homologous proteins have been found in other eukaryotes (*e.g.* yeast, *Xenopus*). Evidence to support this idea is that Parker *et al* (1997) have reported that this gene is essential for the viability of cells of the fission yeast, *Schizosaccharomyces pombe*. The cloned region is part of the 5' end of the open reading frame (Figure 4.5). The mouse trinucleotide encodes a polyglutamine tract (n=12; see Figure 4.5) as the reading frame is continuous on both sides of the repeat. In the frame generating the glutamines the greatest similarity to the human and *Xenopus* proteins is found (Figure 4.5), but the repeat does not appear in the homologue in these species, just a single glutamine codon.

4.4.2 Clone mCTG 24.

This clone was derived from the 8.5dpc whole mouse embryo library. It was found to contain significant similarity to a previously

characterised mouse gene, nuclear receptor co-repressor (Horlein *et al*, 1995). This protein is involved in the ligand independent repressor activity associated with the thyroid hormone and retinoic acid receptors. The level of identity was found to be 79.4% at the DNA (see Figure 4.6) and 89% protein levels (not shown). This indicates that this sequence may derive from a gene, related to the nuclear receptor co-repressor, which has been previously un-characterised, becuase the sequence of Clone mCTG 24 would be expected to be identical to that of the *M. musculus* published sequence (Horlein *et al*, 1995) if they were derived from the same gene. Analysis of the alignment shows that a deletion has occurred in the sequence of Clone mCTG 24 (nucleotides 431-581 of Figure 4.6). This may represent an alternatively spliced exon. The sequence presented here also extends upstream relative to that published for the nuclear receptor co-repressor.

4.4.3 Clone mCTG 29.

This clone was initially isolated from the 8.5 dpc whole embryo. No repeat has been identified yet in this clone from the available sequence. BLASTN and FASTA database searches with sequence from this clone identified eight human ESTs which all exhibit homology to another human gene, encoding alternative splicing factor, or splicing factor 2 (ASF/SF2), subunit p33 (Krainer *et al*, 1991; for review see Lamond, 1991). The protein product of this gene, in conjunction with another polypeptide (ASF-1, subunit p32), has a role in the prevention of exon-skipping to ensure the accuracy of splicing and regulation of alternative splicing of pre-

Figure 4.5

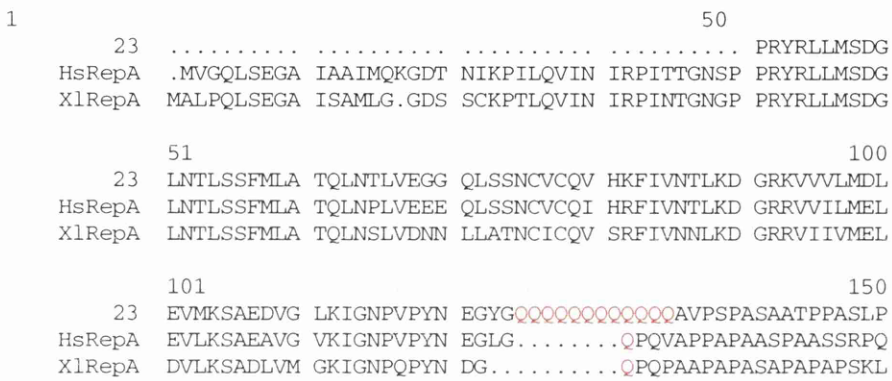


Figure 4.5 Alignment of protein sequences that show similarity to translated sequences derived from Clone mCTG23.

Replication protein A, 70kDa subunits from *Xenopus laevus* (XlRepA) and human (HsRepA) aligned to a translation of sequences derived from clone mCTG 23 (23). The putative glutamine repeat is indicated in red text.

Legend to Figure 4.6. Alignment of sequences derived from Clone mCTG 24 and the mouse nuclear receptor co-repressor protein.

The red coloured sequence represents the triplet repeat as CAG in the DNA sequence. Note that a deletion has occurred in the mCTG 24 sequence presented here, relative to that of the nuclear receptor co-repressor gene. The methionine encoding ATG is indicated in green text and represents the N-terminal of the protein. Clone mCTG 24 sequences are distinguished in blue coloured text.

Figure 4.6

```
     1                                                            50
 24  GAATTCATGT CTTACGGTCA AGGGGCGGCT AAGGTGGCGA AGCTGCGGCG
CoR  .......... .......... .......... .......... ..........

     51                                                           100
 24  GCGGCGGCTG GAATGAGCGC CCAGCTCGCC GCGCTGCCGA ACGAGGGCCG
CoR  .......... .......... .......... .......... ..........

     101                                                          150
 24  GGCGGAGGGC CGCGCGCCGG GGCCGCGCCC CGGTCGTCCG GCCGTGGGGC
CoR  .......... .......... .......... .CCGTGGACT TCAGCGGGGA

     151                                                          200
 24  GCGCCGGCTC TCACTCCGCC AGCTCCCGGT TGCTTCCTAA AAACATGGTT
CoR  GCGTCTGCAA AGGCCCCCGA CGTCCTGAGT GATTTCCTAG AAGCATGGTT

     201                                                          250
 24  GTTTGTTGGA CTTGATCTCA CAGCCCAATG AGGCGTCTT TACTGATAAT
CoR  TTTTGTTGCC CTTGACCTCC CAGCCTGGTG AGGCCTCTG TCCTGAGAAT

     251                                                          300
 24  GTCAAGTTCA GGTTACCCTC CCAACCAGAG GACTTTCAAC ACAAAGCAAA
CoR  GTCAAGTTCA GGTTATCCTC CCAACCAGGG GGCGTTCAGC ACAGAGCAGA

     301                                                          350
 24  GTCGTTATCC TTCACATTCT GTCCAGTATA CCTTTCCCAG TACCCGACAC
CoR  GTCGCTACCC TTCGCACTCG GTTCAGTACA CCTTTCCCAG CGCCCGTCAC

     351                                                          400
 24  CAGCAGGAAT TTGCAGTTCC TGACTACCGC TCTTCTCATA TTGAAATTAG
CoR  CAGCAGGAAT TTGCAGTCCC TGACTACCGT TCTTCTCATC TTGAAGTCAG

     401                                                          450
 24  CCAGGCATCA CAGCTTTTAC AGCAGCAGCA ..........
CoR  TCAGGCGTCA CAGCTCTTGC AGCAGCAGCA GCAGCAGCAG CTTCGCAGAC

     451                                                          500
 24  .......... .......... .......... .......... ..........
CoR  GGCCTTCCCT GCTTTCCGAG TTTCACCCGG GTTCCGACAG GCCCCAGGAA

     501                                                          550
 24  .......... .......... .......... .......... ..........
CoR  AGGAGAAGTG GATACGAGCA GTTCCACCCG GGCCCTTCCC CGGTGGACCA

     551                                                          600
 24  .......... .......... ..........GAGCCAGTT TCTGATGCCC
CoR  TGACTCGCTG GAGTCCAAGC GGCCTCGCCT GGAGCAGGTT TCCGACTCCC

     601                                                          650
 24  ATTTCCAGCG TGTTAGTGCT GCGGTTTTAC CTTTAGTTCA CTCGCTGCCA
CoR  ACTTCCAGCG CATCAGTGCT GCCGTCCTCC CTTTGGTGCA CACGCTGCCA

     651                                                          700
 24  GAAGGCTTGA GGTCGTCTGC AGATGCTAAG AAGGATTCAG CATTTGGAAG
CoR  GAAGGACTGA GGTCTTCTGC CAATGCTAAG AAGGATCCGG CATTTGGAGT

     701                                                          750
 24  CAAACATGAA GCTCCATCCT CTCCTTTGGC TGGGCAACCA TGTGGAGATG
CoR  CAAACATGAA GCTCCTTCCT CTCCCCTCTC TGGGCAGCCA TGCGGAGATG

     751                                                          800
 24  ACCAAAATGC TTCACCTTCA AAGCTTTCAA AGGAGGAGTT AATACAGAGT
CoR  ATCAGAATGC CTCACCTTCA AAACTGTCAA AGGAAGAGCT GATACAGAGC
```

Figure 4.7

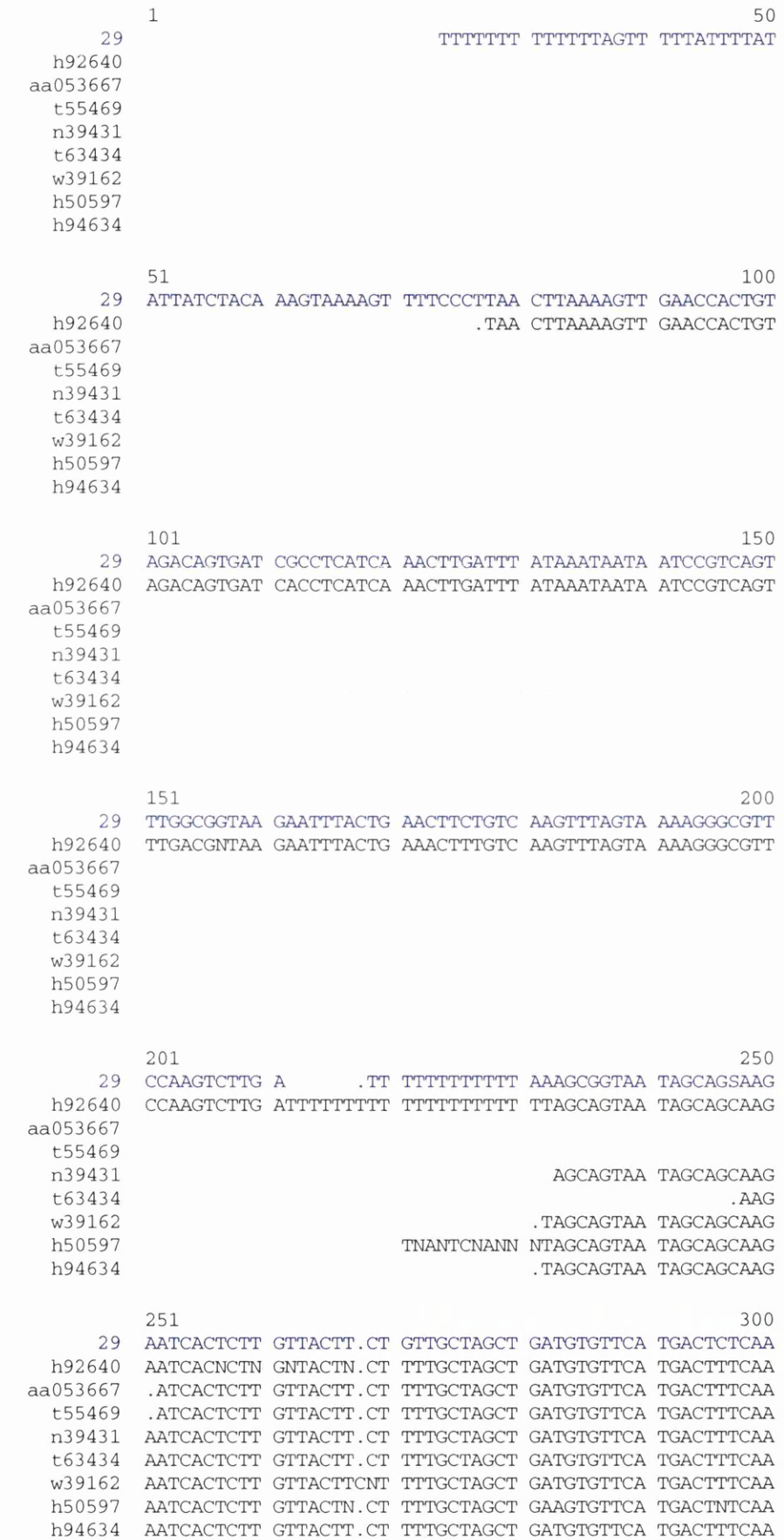

1 50
29 TTTTTTT TTTTTTAGTT TTTATTTTAT
h92640
aa053667
t55469
n39431
t63434
w39162
h50597
h94634
51 100
29 ATTATCTACA AAGTAAAAGT TTTCCCTTAA CTTAAAAGTT GAACCACTGT
h92640 .TAA CTTAAAAGTT GAACCACTGT
aa053667
t55469
n39431
t63434
w39162
h50597
h94634
101 150
29 AGACAGTGAT CGCCTCATCA AACTTGATTT ATAAATAATA ATCCGTCAGT
h92640 AGACAGTGAT CACCTCATCA AACTTGATTT ATAAATAATA ATCCGTCAGT
aa053667
t55469
n39431
t63434
w39162
h50597
h94634
151 200
29 TTGGCGGTAA GAATTTACTG AACTTCTGTC AAGTTTAGTA AAAGGGCGTT
h92640 TTGACGNTAA GAATTTACTG AAACTTTGTC AAGTTTAGTA AAAGGGCGTT
aa053667
t55469
n39431
t63434
w39162
h50597
h94634
201 250
29 CCAAGTCTTG A .TT TTTTTTTTTT AAAGCGGTAA TAGCAGSAAG
h92640 CCAAGTCTTG ATTTTTTTTT TTTTTTTTTT TTAGCAGTAA TAGCAGCAAG
aa053667
t55469
n39431 AGCAGTAA TAGCAGCAAG
t63434 .AAG
w39162 .TAGCAGTAA TAGCAGCAAG
h50597 TNANTCNANN NTAGCAGTAA TAGCAGCAAG
h94634 .TAGCAGTAA TAGCAGCAAG
251 300
29 AATCACTCTT GTTACTT.CT GTTGCTAGCT GATGTGTTCA TGACTCTCAA
h92640 AATCACNCTN GNTACTN.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
aa053667 .ATCACTCTT GTTACTT.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
t55469 .ATCACTCTT GTTACTT.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
n39431 AATCACTCTT GTTACTT.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
t63434 AATCACTCTT GTTACTT.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
w39162 AATCACTCTT GTTACTTCNT TTTGCTAGCT GATGTGTTCA TGACTTTCAA
h50597 AATCACTCTT GTTACTN.CT TTTGCTAGCT GAAGTGTTCA TGACTNTCAA
h94634 AATCACTCTT GTTACTT.CT TTTGCTAGCT GATGTGTTCA TGACTTTCAA

Figure 4.7 (continued)

```
          301                                                       350
      29  GGGCCATTAA AAAATAAATA ACTTCCAGTT TCAGCAAGCA GAGCTGGGGG
   h92640 GGGNTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAGCNGGGGG
 aa053667 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAG.CTGGGG
   t55469 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAN.TGGNGG
   n39431 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAGCTGGGGG
   t63434 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCCAAGC AGACTGGNGG
   w39162 GGGTTATTAA AAAATAAATA ACTTCCAGGT TTCGGCAAGC AGAGCTGGGG
   h50597 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAGCTGGNGG
   h94634 GGGTTATTAA AAAATAAATA ACTTCCAGTT TCGGCAAGCA GAGCT.GGGG

          351                                                       400
      29  TACTTGSAGG GACCTGCAAC AGCACAT.GG AATTATGGAA CAGACACAAG
   h92640 NACCTGCGGG ACTCNGAAAC AGCATATGGG AATTATGGAA TANCCCCCAA
 aa053667 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   t55469 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   n39431 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   t63434 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   w39162 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   h50597 TACCCNCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA
   h94634 TACCTGCGGG ACTCTGAAAC AGCATAT.GG AATTATGGAA TAGCCCCCAA

          401                                                       450
      29  TCCCTACAAT TGCTCTACTC GGTCCAAAAC ATCCTCCACT GAAGTGAGCT
   h92640 GTTCC    .
 aa053667 GTTCCTAAAT GCCTCTACTC .GGTCCAAGT ATCTTCCACT GCAAGTGAAC
   t55469 GTTCCTAAAT GCCTCTACTC .GGTCCAAGT ATCTTCCACT GCAAGTGAAC
   n39431 GTTCCTAAAT GCCTCTACTC .GGTCCAAGT ATCTTCCACT GCAAGTGAAC
   t63434 GTTCCTAAAT GCCTCTACTC .GGTCCAAGT ATCTTCCACT GCAAGTGAAC
   w39162 GTTCCTAAAT GCCTCTACTC GGGTCCAAGT ATCTTCCACT GCAAGTGAAC
   h50597 GTNCCNAAAT GCCTGCTACT NGGNCCAAGT ATCTTCCACT GCAAGTGAAC
   h94634 GTTCCTAAAT GCCT.CTACT CGGTCCAAGT ATCTTCCACT GCAAGTGAAC

          451                                                       500
      29  GTTAGCATCC TATCAGATTG TTAATCAGG.
   h92640
 aa053667 TGTT.AGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA
   t55469 TGTT.AGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA
   n39431 TGTT.AGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA
   t63434 TGTT.AGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA
   w39162 TGTT.AGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA
   h50597 TGTT.AGCAT TCCTATTGGA NGTTACAAGA AATCAACATA TATATTTTAA
   h94634 TGTTAAGCAT TCCTATTGGA TGTTACAAGA AATCAACATA TATATTTTAA

          501                                                       550
      29
   h92640
 aa053667 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGG.TAGAAG
   t55469 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGG.GTAGGA
   n39431 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGG.TAGAAG
   t63434 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGG.TAGAAG
   w39162 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGG.TAGAAG
   h50597 AANCAAAATA AATGAAATTC TAGTNTTCTG NGCTTCCAGT TGGGTAGAAG
   h94634 AAACAAAATA AATGAAATTC TAGTTTTCTG TGCTTCCAGT TGGTAAGAAG

          551                                                       600
      29
   h92640
 aa053667 CAGTAGAAGG GAGGAGGGAA          TT TGGCCTTTCG GTTCCTCCAG
   t55469 AGGCAGTAGG AAGGGGAGGA GGGGAATTTG GGCCTTTCGG GTTCCTCCCG
   n39431 CAGTAGAAGG GAGGAGGG          AATT TGGCCTTTCG GTTCCTCCA.
   t63434 CAGTAGAAGG GAGGAGGG          AATT TGGCCTTTCG GTTCCTCCA.
   w39162 CAGTAGAAGG GAGGAGGG          AATT TGG.CTTTCG GTTCCTCCA.
   h50597 CAGTAGAAGG GAGGGNGG      .GGAATTT G         .
   h94634 CAGTAAGAAG GGAGGAGG      .GGAATTT GGGCCTTTCG GTTCCTCCAG
```

Figure 4.7 (continued)

```
          601                                                   650
      29
   h92640
 aa053667  GGCAGCCTTT ACAACTGCTG TTGG  TGGT GGGCTTGTAC TGAAGGGTGC
   t55469  GGGCAGCCTT ACAACTGCTG TTGGGTGGTG GGGCTTGTAC TGAAGGGTGC
   n39431  GGGCAGCCTT ACAACTGCTG TTGG  TGGT GGGCTTGTAC TGTAAAAAGA
   t63434  GGGCAGCCTT ACAACTGCTG TTGG  TGGT GGGCTTGTAC TGTAAAAAGA
   w39162  GGGCAGCCTT ACAACTGCTG TTGG  TGGT GGGCTTGTAC TGTAAAAAGA
   h50597
   h94634  GGGCAGCCTT ACAACTGCTG TTGG  TGGT GGGCTTGTAC TGAAG    .

          651                                                   700
      29
   h92640
 aa053667  GTCATATA.G GGGAAATGCG CTGTTGTCGA GACTGGAAT. GGGCTGTAGT
   t55469  GTCATATAGG GGGAAATGCG CTGTTGTCGA GACTGGGATG GGGCTGTAGT
   n39431  AAAGCGAACA TTGGAACAGA AAACTGGC
   t63434  AAAGCGNCCA TTGACCC  .
   w39162  AAAAATTGGT GCG      .
   h50597
   h94634

          701                     726
      29
   h92640
 aa053667  GCACTTTGAG CG
   t55469  GCACTTTGAG CGAGGGCGCC TGGGGG
   n39431
   t63434
   w39162
   h50597
   h94634
```

Figure 4.7 Sequence alignment of sequences derived from the mouse Clone mCTG 29 and alternative splicing factor 2-like ESTs.

These sequences were derived from the nucleic acid databases (dBEST). Alignment was performed using the PILEUP programme. Sequences represented are: H92640 (327bp 3' EST), AA053667 (443bp 3' EST from a colon cDNA library clone), T55469 (469bp 3' EST), N39431 (430bp 3' EST), T63434 (404bp 3' EST), W39162 (417bp 5' EST from a cDNA clone derived from a library constructed from parathyroid adenomas), H50597 (353bp 3' EST) and H94634 (403bp 3' EST). Clone mCTG 29 sequences are indicated in blue coloured text.

mRNA molecules in the nucleus (Lamond, 1991). It is known that this gene belongs to the SR (serine-arginine) family of splicing factors. The serine-arginine rich domain, which gives this gene family its name, is required for the selection of splicing events (Wang and Manley, 1995). Recently a transgenic knockout study has shown that ASF/SF2 is essential for cell viability (Wang *et al*, 1996).

4.4.4 Clone mCTG 43.

This clone was originally isolated from the 8.5 dpc library. Overlapping sequence was produced from both ends of the insert from this clone. By using the MacAlign programme a continuous sequence was built. This was re-exported to MacVector and the UWGCG sequence analysis packages for further analysis. The insert was found to be 531 nucleotides in length and contained a cryptic CAG trinucleotide repeat, 11 triplets in length. It was interrupted by a CCA trinucleotide. In addition this insert was also found to contain a polypurine tract, 54 nucleotides in length. This has an internal structure consisting of a GAG repeat, 7 repeats long. This in turn is adjacent to a cryptic hexanucleotide repeat GAGAAG, which is 5 repeats in length with the middle repeat corrupted to GAAAAA. A continuous open reading frame was found throughout this sequence by application of Fickett's testcode algorithm (Fickett, 1982). Upon translation of this ORF from nucleotide to amino acid sequence, the CAG repeat was found to encode a polyglutamine tract. Because of the presence of a flanking CAA trinucleotide and the interrupting CCA repeat the structure of the polypeptide was Q_2PQ_9. In the case of the polypurine tract, it was found that the hexanucleotide repeat would be translated as five copies of alternating amino acids lysine

and glutamic acid and the GAG repeat is translated as glutamic acid. The additional trinucleotide separating these repeats was also found to encode glutamic acid. Therefore the polypurine tract translation in whole was $(EK)_5E_7$ in this reading frame. The full nucleotide sequence is described in Figure 4.8.

The sequence from this clone was then used to search both nucleic acid (Genbank and EMBL) and amino acid databases (Swiss-Prot and PIR). A search using the BLASTX algorithm identified similarities between the sequence from clone 43 and four sequences in the non-redundant sequence database which was searched. Three of these sequences were related, one of which was described as the mouse Rad21 protein, which is involved in the repair of double-strand DNA break lesions. This sequence was described recently by McKay *et al* (1996). This paper also describes the identification of the human homologue of this gene. This represents another sequence identified from this search. The second sequence identified, PW29, has been described by Yu *et al* (1995) as having calcium binding ability (as tested by a calcium blot experiment).

The search of nucleic acid databases with the sequence of clone mCTG 43 using the FASTA algorithm (Lipman and Pearson, 1985), in addition to identifying the sequences reported above, found another independently identified protein coding gene sequence, a mouse beta-1,4-galactosyltransferase. This enzyme synthesises beta-4-N-acetyl-lactosamine structure in glycoconjugates. The protein is required for male mammalian fertility in that it binds to specific O-linked oligosaccharide ligands on the egg coat glycoprotein ZP3. In mammals this protein has an additional metabolic function in the production of lactose in stimulated mammary glands.

From Figure 4.8 which represents a PILEUP multiple sequence alignment analysis, it can be seen that there is identity at the nucleic acid level over the whole sequence of clone mCTG 43 insert to the three gene groups described with known function.

4.4.5 Clone mCTG 410.

Clone mCTG 410 was isolated from the 8.5 dpc cDNA library. When sequence derived from this clone was used to search the database using the BLASTX algorithm (which searches the non-redundant sequences of Genbank and EMBL nucleotide databases, in addition to the protein databases, PIR and Swiss-Prot, with a six frame translation of the input nucleotide sequence) it was found that the sequence showed significant similarity (97% at the amino acid level) to the N terminal portion of a rat nucleoporin, p54 (Hu *et al*, 1996; AC: U63840). There is a similar degree of identity at the nucleic acid level with 95.1% of the bases being identical between the rat and mouse sequences presented here (see Figure 4.9.A). This protein is involved in the transport of RNA from the nucleus to the cytoplasm and has been shown to bind RNA.

The CAG repeat is conserved between the rat and the presumed mouse sequences presented here, both at the nucleotide level and at the amino acid level as a polyglutamine tract, 8 residues long, starting at residue 93 in the rat nucleoporin sequence. This protein also has an unusual dipeptide repeat juxtaposed N-terminal to this polyglutamine tract which consists of alternate glycine residues with larger hydrophilic side chain residues, threonine (T), leucine (L) or phenylalanine (F). There is no conserved repetitive element at the nucleic acid level and this

Legend to Figure 4.8. Sequence analysis of clone mCTG 43.

Nucleotide sequences of *M. musculus* Rad21 cDNA (MmRad21, AC: X98293, McKay *et al*, 1996); Human Rad21 cDNA (HsRad21, AC: X98294, McKay *et al*, 1996); Human calcium binding protein (PW29, Yu *et al*, 1995) and *M. musculus* β-1,4-galactosyltransferase (Mmnb) aligned to sequences derived from clone mCTG 43. The red coloured sequence indicates the position of the CAG trinucleotide repeat in Clone mCTG 43 and the purple coloured triplets indicate corruptions of the CAG/CTG trinucleotide repeat. The green coloured sequence indicates the polypurine stretch. Clone 43 sequences are indicated in blue text.

Figure 4.8

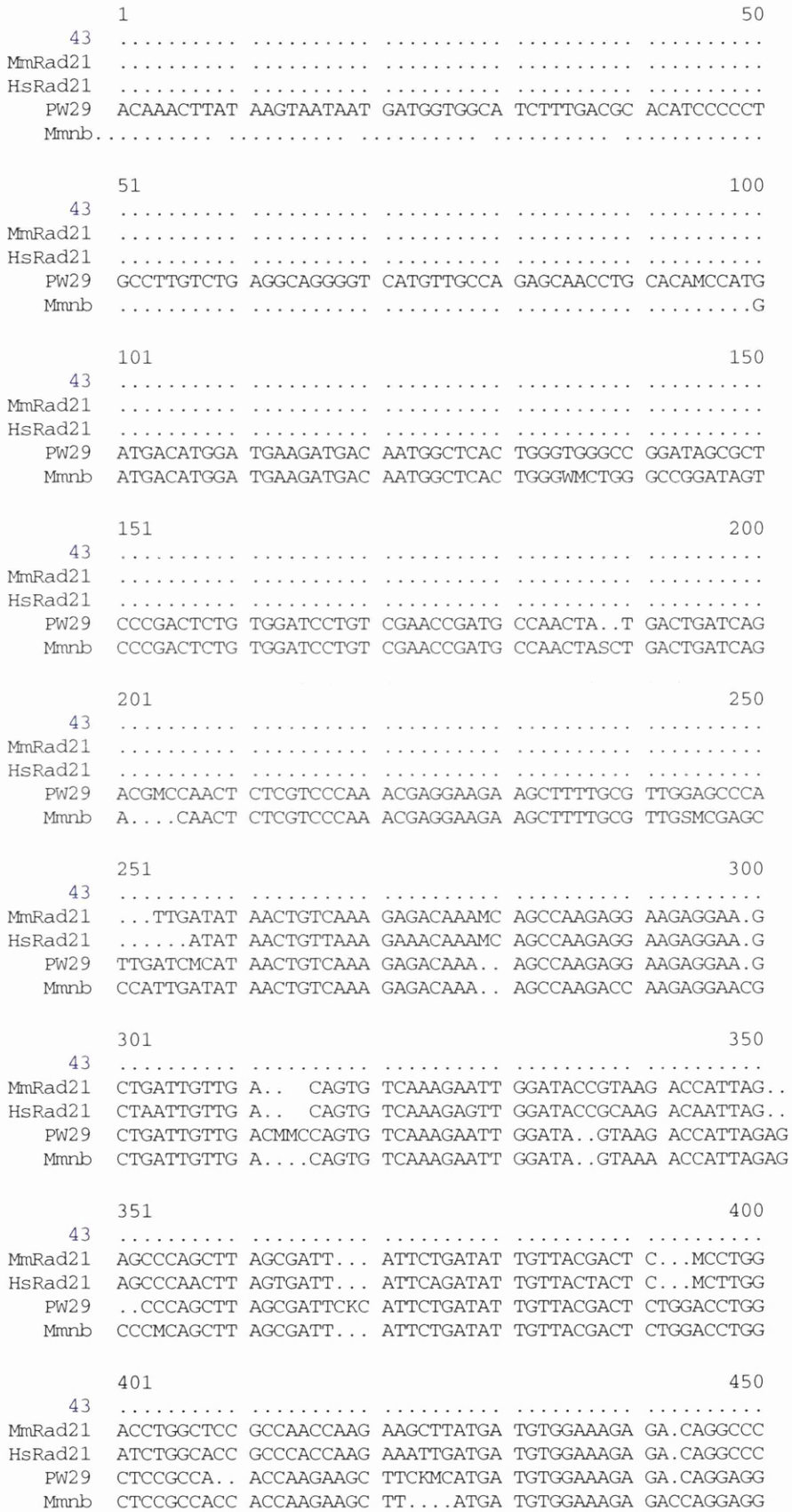

1 50
43 ....
MmRad21 ....
HsRad21 ....
PW29 ACAAACTTAT AAGTAATAAT GATGGTGGCA TCTTTGACGC ACATCCCCCT
Mmnb..... ....
51 100
43 ....
MmRad21 ....
HsRad21 ....
PW29 GCCTTGTCTG AGGCAGGGGT CATGTTGCCA GAGCAACCTG CACAMCCATG
Mmnb ....G
101 150
43 ....
MmRad21 ....
HsRad21 ....
PW29 ATGACATGGA TGAAGATGAC AATGGCTCAC TGGGTGGGCC GGATAGCGCT
Mmnb ATGACATGGA TGAAGATGAC AATGGCTCAC TGGGWMCTGG GCCGGATAGT
151 200
43 ....
MmRad21 ....
HsRad21 ....
PW29 CCCGACTCTG TGGATCCTGT CGAACCGATG CCAACTA..T GACTGATCAG
Mmnb CCCGACTCTG TGGATCCTGT CGAACCGATG CCAACTASCT GACTGATCAG
201 250
43 ....
MmRad21 ....
HsRad21 ....
PW29 ACGMCCAACT CTCGTCCCAA ACGAGGAAGA AGCTTTTGCG TTGGAGCCCA
Mmnb A....CAACT CTCGTCCCAA ACGAGGAAGA AGCTTTTGCG TTGSMCGAGC
251 300
43 ....
MmRad21 ...TTGATAT AACTGTCAAA GAGACAAAMC AGCCAAGAGG AAGAGGAA.G
HsRad21 ......ATAT AACTGTTAAA GAAACAAAMC AGCCAAGAGG AAGAGGAA.G
PW29 TTGATCMCAT AACTGTCAAA GAGACAAA.. AGCCAAGAGG AAGAGGAA.G
Mmnb CCATTGATAT AACTGTCAAA GAGACAAA.. AGCCAAGACC AAGAGGAACG
301 350
43 ....
MmRad21 CTGATTGTTG A.. CAGTG TCAAAGAATT GGATACCGTAAG ACCATTAG..
HsRad21 CTAATTGTTG A.. CAGTG TCAAAGAGTT GGATACCGCAAG ACAATTAG..
PW29 CTGATTGTTG ACMMCCAGTG TCAAAGAATT GGATA..GTAAG ACCATTAGAG
Mmnb CTGATTGTTG A....CAGTG TCAAAGAATT GGATA..GTAAA ACCATTAGAG
351 400
43 ....
MmRad21 AGCCCAGCTT AGCGATT... ATTCTGATAT TGTTACGACT C...MCCTGG
HsRad21 AGCCCAACTT AGTGATT... ATTCAGATAT TGTTACTACT C...MCTTGG
PW29 ..CCCAGCTT AGCGATTCKC ATTCTGATAT TGTTACGACT CTGGACCTGG
Mmnb CCCMCAGCTT AGCGATT... ATTCTGATAT TGTTACGACT CTGGACCTGG
401 450
43 ....
MmRad21 ACCTGGCTCC GCCAACCAAG AAGCTTATGA TGTGGAAAGA GA.CAGGCCC
HsRad21 ATCTGGCACC GCCCACCAAG AAATTGATGA TGTGGAAAGA GA.CAGGCCC
PW29 CTCCGCCA.. ACCAAGAAGC TTCKMCATGA TGTGGAAAGA GA.CAGGAGG
Mmnb CTCCGCCACC ACCAAGAAGC TT....ATGA TGTGGAAAGA GACCAGGAGG

Figure 4.8 (continued)

```
          451                                                   500
     43   .......... .......... ..........    .......... ..........
MmRad21   AGGAGTGGAA AAGCTCTTCT CCTTACCAGC... ACAGCCCCTG TGGAATAACC
HsRad21   AGGAGTAGAA AAACTGTTTT CTTTACCTGC... TCAGCCTTTG TGGAATAACA
   PW29   ...AGTGGAA AAGCTCTTCT CCTTACCAGCCWC ACAGCCCCTG TGGAATAACC
   Mmnb   ...AGTGGAA AAGCTCTCMC TCTCCTTACCAGC ACAGCCCCTG TGGAATAACC

          501                                                   550
     43   .......... .......... .......... .......... ..........
MmRad21   CCMCGGCTAC TGAAGCTCTT CACACGCTGC CTTACCCCAC TTGTACCAGA
HsRad21   CCMCGACTAC TGAAGCTCTT TACACGCTGT CTTACACCGC TTGTACCAGA
   PW29   GGCTACTGAA GCTCTTCACA CGCTGCCTTA CWMCCCCCAC TTGTACCAGA
   Mmnb   GGCTACTGAA G.ACTTTCCC CACACGCTGC CTTACCCCAC TTGTACCAGA

          551                                                   600
     43   .......... .......... .......... .......... ..........
MmRad21   AGACCTCCTT AGGAAGAGAA GGAAAGGGGG AGAGGCAGAT AATCTGGATG
HsRad21   AGACCTCCTT AGAAAAAGGA GGAAAGGAGG AGAGGCAGAT AATTTGGATG
   PW29   AGACCTTAGG AAGAGAAGGA A....AGGGG GAGAGGCACS CGATAATCTG
   Mmnb   AGACCTTAGG AAGAGAAGGA ACCMCAGGGG GAGAGGCA.. .GATAATCTG

          601                                                   650
     43   .......... .......... .......... .......... ..........
MmRad21   AGTTCCTCTM CCAAAGAGTT TGAGAA...T CCAGAGGTTC CCAG....AG
HsRad21   AATTCCTCTM CCAAAGAATT TGAAAA...T CCAGAGGTTC CTAG....AG
   PW29   GATGAGTTCC TCAAAGAGTT TGAGAA...T CCAGAGGTTC CCAGCSMCAG
   Mmnb   GATGAGTTCC TCAAAGAGTT TGAGAATCCT CCAGAGGTTC CCAG....AG

          651                                                   700
     43   .....CAGCA GCCACAGCAG CAGCAGCAGC AGCAGCAGC. AGCAAGATGT
MmRad21   AGGAGCAGCA GCCACAGCCA CAGCAGCAGC CACAGCCGC. AGCGAGATGT
HsRad21   AGGACCAG.. .......... CAGCAGC CACATCAGC. AGCGTGATGT
   PW29   AGGAGCAGCA GCCACAGCAG CAGCAGCCAC AGCCGCAGC. .GAGATGTCA
   Mmnb   AGGAGCAGCA GCCACAGCAG CAGCAGCCAC AGCCGCTMCC AGCGAGATGT

          701                                                   750
     43   CATCGATGAG CCCATTATAG AA....GAGC CMAGCCGCCT CCAGGACTCA
MmRad21   CATCGATGAG CCCATTATAG AACAMCGAGC CAAGCCGCCT CCAGGACTCA
HsRad21   TATCGATGAG CCCATTATTG AAGAGCCAAG CCGCCAMCCT CCAGGAGTCA
   PW29   CCTCGATGAG CCCATTATAG AAGAGCCAAG CC.GCCTCCA GGACTCAGTG
   Mmnb   CATCGATGAG CCCATTATAG AAGAGCCAAG CC.GCCTCCA CACGGACTCA

          751                                                   800
     43   CTGATGGAGG CCAGCAGAAC AACCAT...A GAAGAAT.CA GCCATGCCCC
MmRad21   GTGATGGAGG CCAGCAGAAC AACCATCGCA GAAGAAT.CA GCCATGCCCC
HsRad21   GTGATGGAGG CCAGCAGAAC AAACATAGAT GAGTCAG.CC GCTATGCCTC
   PW29   ATGCMCGAGG CCAGCAGAAC AACCATAGAA GAATCAGCCA TGC..CCCCA
   Mmnb   GTGATGGAGG CCAGCAGAAC AACCATAGAA GAATCAGCCA TGCCAMCCCC

          801                                                   850
     43   CACCACCCCC TCAAGGAGTT AAGCGGAAAG ....CCGGGC AAATAGACCC
MmRad21   CACCACCCCC TCAAGGAGTT AAGCGGAAAG CGMCCCGGGC AAATAGACCC
HsRad21   CACCACCACC TCAGGGAGTT AAGCGAAAAG CTGGACAAAT TGCGMCACCC
   PW29   CCACCCCCTC TCAAGGAGTT AAGCGGAAAG CCGGGCAAAT AGACCCAGAG
   Mmnb   CACCACCCCC TCAAGGAGTT AAGCGGAAAG CCGGGCAAAT AGACCCACGC

          851                                                   900
     43   AGAGCCTTGG ATACCTCCTC AGCAGGTAGA GCAA...ATG GAAATACCAC
MmRad21   AGAGCCTTCG ATACCTCCTC AGCAGGTAGA GCAACMCATG GAAATACCAC
HsRad21   AGAGCCTGTG ATGCCTCCTC AGCAGGTAGA GCAGATGGAA ATACCACMCC
   PW29   .CCTTCGATA CTMCCTCCTC AGCAGGTAGA GCAAATGGAA ATACCACCAG
   Mmnb   .GAGCCTTCG ATACCTCCTC AGCAGGTAGA GCAAATGGAA ATACCACCAG
```

Figure 4.8 (continued)

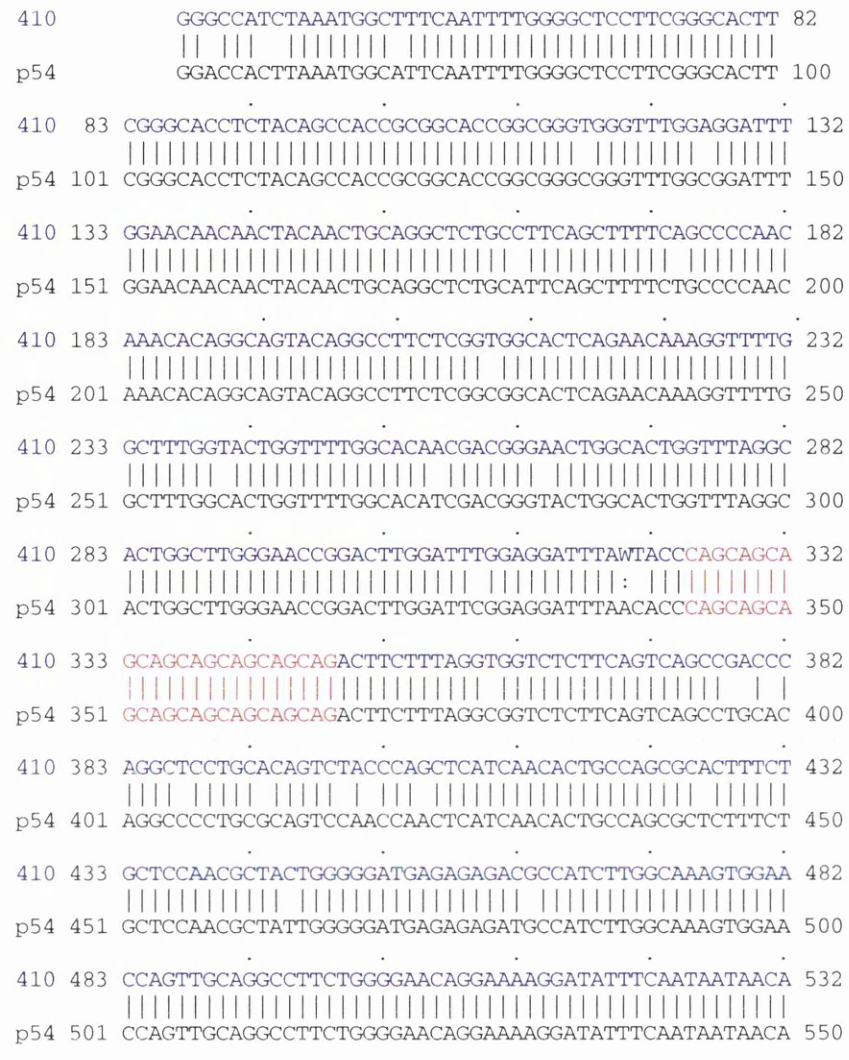
```
          901                                                        950
      43  CAG.TAGAAT TCCCCCCAGA GGAGCCTCCA AATATCTG.. ..TCAGCTTA
MmRad21   CAG.TAGAAC TTCCCCCAGA GGAGCCTCCA AATATCTGCM MCTCAGCTGA
HsRad21   CTG.TAGAGC TTCCCCCAGA AGAACCTCCA AATATCTGTC AGCTAATACC
   PW29   TAGA.ACTTC CCCCATCCGA GGAGCCTCCA AATATCTGTC AGCTGATCCC
   Mmnb   TCGMCAGAAC TTCCCCCAGA GGAGCCTCCA AATATCTGTC AGCTGATCCC

          951                                                       1000
      43  TCCCGGAGTT AGAGCTCCTA CCGGAGAAGG AGAAGGAAAA AG...AGAAG
MmRad21   TCCCGGAGTT AGAGCTCCTA CCGGAGAAGG AGAAGGAAAA AGCKCAGAAG
HsRad21   CMMCAGAGTT AGAACTTCTG CCAGAAAAAG AGAAGGAGAA AGAGAAGGAA
   PW29   GGAGTTAGAG CTCCTACCTC MCGGAGAAGG AGAAGGAAAA AGAGAAGGAG
   Mmnb   GGAGTCMCTA GAGCT.CCTA CCGGAGAAGG AGAAGGAAAA AGAGAAGGAG

          1001                                                      1050
      43  GAGAAGGA.A GAGGAGGAGG AGGAGGAGGA TGAAGATGCT TCAGGG....
MmRad21   GAGAAGGA.A GAGGAGGAGG AGGAGGAGGA TGAAGATGCT TCAGGGCKMC
HsRad21   AAAGCKCA.A GATGATGAAG AGGAAGAGGA TGAAGATGCA TCAGGGGGCG
   PW29   AAGGAAGAG. ...GAGGAGG AGGAGGTTCA TGAAGATGCT TCAGGGGG.T
   Mmnb   AAGGAAGAGC MMCGAGGAGG AGGAGGAGGA TGAAGATGCT TCAGGGGG.T

          1051                                                      1100
      43  GGTGATAAGG ATCAAGAGGA AAGGAGATGG AACAAACGCA CTCAGCAGAT
MmRad21   GGTGATCAGG ATCAAGAGGA AAGGAGATGG AACAAACGCA CTCAGCAGAT
HsRad21   ATCAAGATCK MCCAGGAAGA AAGAAGATGG AACAAAAGGA CTCAGCAGAT
   PW29   GATCAGGATC A......... .......... .......... ..........
   Mmnb   GATCAGGATC AAGACKCGGA AAGAAGATGG AACAAACGCA CTCAGCAGAT

          1101                                                      1150
      43  ...GCTTCAT GGTCTTCAGC GAGCTCTTGC TAAAACTGGA GCAGAGTCTA
MmRad21   CWCGCTTCAT GGTCTTCAGC GAGCTCTTGC TAAAACTGGA GCAGAGTCTA
HsRad21   GCTTCATGGT CTCWCTCAGC GTGCTCTTGC TAAAACTGGA GCTGAATCTA
   PW29   .......... .......... .......... .......... ..........
   Mmnb   GCTTCATGGT CTTCAGCCKM CGAGCTC... .......... ..........

          1151                                                      1200
      43  TCA....GTT TGCTTGAGCT GTGTCGAAAC ACAAACCGAA AGCAGGCAGC
MmRad21   TCACWMCGTT TGCTTGAGCT GTGTCGAAAC ACAAACCGAA AGCAGGCAGC
HsRad21   TCAGTTTGCT TGAGTCWMCT ATGTCGAAAT ACGAACAGAA AACAAGCTGC
   PW29   .......... .......... .......... .......... ..........
   Mmnb   .......... .......... .......... .......... ..........

          1201                                                      1250
      43  AGCAAAG... TTCTACAGCT TTTTG..... .......... ..........
MmRad21   AGCAAAGCSC TTCTACAGCT TTTTGGTTCT TAAGAAGCAG CAAGCCATCG
HsRad21   CGCAAAGTTC TACAGCTTCC SCTTGGTTCT TAAAAAGCAG CAAGCTATTG
   PW29   .......... .......... .......... .......... ..........
   Mmnb   .......... .......... .......... .......... ..........

          1251                                                      1300
      43  .......... .......... .......... .......... ..........
MmRad21   AGCTCACACA CSMCGGAAGA GCCGTACAGT GACATCATTG CAACCCCTGG
HsRad21   AGCTGACACA GGAAGAACCG TACSMCCAGT GACATCATCG CAACACCTGG
   PW29   .......... .......... .......... .......... ..........
   Mmnb   .......... .......... .......... .......... ..........

          1301              1321
      43  .......... .......... .
MmRad21   ACCACGGTTC CATACCTTAT C
HsRad21   ACCAAGGTTC CATATTATA. .
   PW29   .......... .......... .
   Mmnb   .......... .......... .
```

Legend to Figure 4.9. A) Nucleotide sequence and B) amino acid comparison of mCTG 410 sequence and Rat nucleoporin p54.

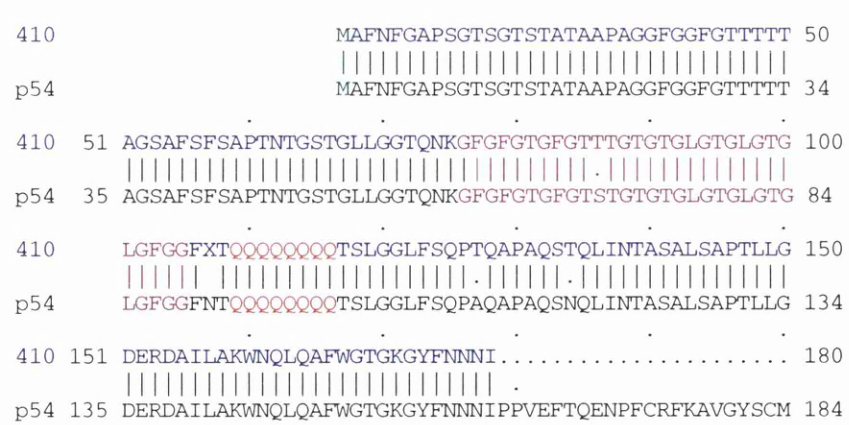
The red sequence indicates the CAG repeat identified in clone 410 (Figure 4.8.A) and translated as a polyglutamine tract (indicated in red text in Figure 4.8.B). The purple coloured sequence represents the alternating glycine repeat. The initial ATG and methionine amino acid is indicated, in green type, on Figures 4.8.B.

Figure 4.9

A)


410 GGGCCATCTAAATGGCTTTCAATTTTGGGGCTCCTTCGGGCACTT 82
p54 GGACCACTTAAATGGCATTCAATTTTGGGGCTCCTTCGGGCACTT 100
410 83 CGGGCACCTCTACAGCCACCGCGGCACCGGCGGGTGGGTTTGGAGGATTT 132
p54 101 CGGGCACCTCTACAGCCACCGCGGCACCGGCGGGCGGGTTTGGCGGATTT 150
410 133 GGAACAACAACTACAACTGCAGGCTCTGCCTTCAGCTTTTCAGCCCCAAC 182
p54 151 GGAACAACAACTACAACTGCAGGCTCTGCATTCAGCTTTTCTGCCCCAAC 200
410 183 AAACACAGGCAGTACAGGCCTTCTCGGTGGCACTCAGAACAAAGGTTTTG 232
p54 201 AAACACAGGCAGTACAGGCCTTCTCGGCGGCACTCAGAACAAAGGTTTTG 250
410 233 GCTTTGGTACTGGTTTTGGCACAACGACGGGAACTGGCACTGGTTTAGGC 282
p54 251 GCTTTGGCACTGGTTTTGGCACATCGACGGGTACTGGCACTGGTTTAGGC 300
410 283 ACTGGCTTGGGAACCGGACTTGGATTTGGAGGATTTAWTACCCAGCAGCA 332
p54 301 ACTGGCTTGGGAACCGGACTTGGATTCGGAGGATTTAACACCCAGCAGCA 350
410 333 GCAGCAGCAGCAGCAGACTTCTTTAGGTGGTCTCTTCAGTCAGCCGACCC 382
p54 351 GCAGCAGCAGCAGCAGACTTCTTTAGGCGGTCTCTTCAGTCAGCCTGCAC 400
410 383 AGGCTCCTGCACAGTCTACCCAGCTCATCAACACTGCCAGCGCACTTTCT 432
p54 401 AGGCCCCTGCGCAGTCCAACCAACTCATCAACACTGCCAGCGCTCTTTCT 450
410 433 GCTCCAACGCTACTGGGGGATGAGAGAGACGCCATCTTGGCAAAGTGGAA 482
p54 451 GCTCCAACGCTATTGGGGGATGAGAGAGATGCCATCTTGGCAAAGTGGAA 500
410 483 CCAGTTGCAGGCCTTCTGGGGAACAGGAAAAGGATATTTCAATAATAACA 532
p54 501 CCAGTTGCAGGCCTTCTGGGGAACAGGAAAAGGATATTTCAATAATAACA 550

B)


410 MAFNFGAPSGTSGTSTATAAPAGGFGGFGTTTTT 50
p54 MAFNFGAPSGTSGTSTATAAPAGGFGGFGTTTTT 34
410 51 AGSAFSFSAPTNTGSTGLLGGTQNKGFGFGTGFGTTTGTGTGLGTGLGTG 100
p54 35 AGSAFSFSAPTNTGSTGLLGGTQNKGFGFGTGFGTSTGTGTGLGTGLGTG 84
410 LGFGGFXTQQQQQQQQTSLGGLFSQPTQAPAQSTQLINTASALSAPTLLG 150
p54 LGFGGFNTQQQQQQQQTSLGGLFSQPAQAPAQSNQLINTASALSAPTLLG 134
410 151 DERDAILAKWNQLQAFWGTGKGYFNNNI..... 180
p54 135 DERDAILAKWNQLQAFWGTGKGYFNNNIPPVEFTQENPFCRFKAVGYSCM 184

primary structure feature probably has some function in the protein. In addition to the above features, this protein also has a threonine repeat 5 residues long, starting at residue 30 in the rat nucleoporin sequence. This is encoded by $(ACA)_2ACTACAACT$ whichis conserved in both the rat and mouse sequence (clone mCTG 410) presented here (see Figure 4.9.A).

4.4.6 Clone mCTG 411.

This sequence (see Figure 4.10) has a strong similarity to a nucleocytoplasmic protein from the rat, Nopp140 (Meier and Blobel, 1992). This protein interacts with nuclear localisation signals (Meier and Blobel, 1990) and is hypothesised to be a chaperone for export and import to and from the nucleus. Interestingly, this protein has a repeat which contains serine tracts. At the nucleic acid level, it is revealed that these serines are encoded by the codon AGC, a member of the family (determined by frame and complementarity) of codons which includes the CTG codon which constituted the probe.

4.4.7 Clone mCTG 56.

Sequence analysis revealed that this clone had one of the smallest CAG repeats identified in this project. The repeat consists of six CAG trinucleotides. This is conserved between sequences which show similarity to this sequence. This clone has identity to a group of related genes; CW17 (Wrehlke *et al*, unpublished), a human transcription factor, Zfm1 (Toda *et al*, 1994), which has been proposed as a candidate for the gene in which mutations cause

multiple endocrine neoplasia type 1 (however, this gene has now been excluded as a candidate for this disease; see Lloyd *et al*, 1997) and a human RNA splicing factor, SF-1 (Arning *et al*, 1996). These proteins all share common N-terminal sequences but have alternatively spliced C-terminal segments as indicated in the multiple sequence alignment of the putative protein sequences derived from the cDNA sequences (Figure 4.11). The region of identity presented in Figure 4.11 overlaps a glutamine and proline rich region common to all these proteins. The CAG repeat found in clone mCTG 56 is conserved between mouse and man and is common to all of the sequences described here except the human Zfm1 isoform L49344. This mRNA is alternatively spliced at a position in the nucleic acid sequence corresponding to amino acid 448 (Figure 4.12, No 1). At a position corresponding to amino acid 528 (Figure 4.12, No 2) there is likely to be another mRNA alternative splicing event when the remaining sequences which were hitherto homologous, are split into two sub-groups which show identity to each other. The first group consists of Zfm1, (isoform Zfm1-I, Figure 4.12) and the two SF-1 isoforms, B0 and h11 (Arning *et al*, 1996). The second group consists of CW17, mCTG 56 and a Zfm1 isoform. This second group of sequences terminate at position 548. There is another mRNA alternative splicing event which leads to a further divergence in amino acid sequences at position 597 in the SF-1 isoform B0.

4.4.8 Clone mCTG 57.

Clone 57, derived from the 8.5 dpc whole mouse embryo library, (Hogan *et al*, 1984) shows identity to a ubiquitously

expressed human transcriptional adaptor, P300. This protein is believed to have a role in the repression of the cell cycle and promotion of differentiation. It has been reported to be a part of a TATA binding complex (Abraham *et al*, 1993.) The human cDNA contains a bromodomain (Eckner *et al*, 1994) which is a putative protein-protein interaction domain (Haynes *et al*, 1992). It also contains 3 cysteine-histidine rich domains. The most carboxyl terminal of these is found to overlap the interaction domain for the adenovirus oncogene E1A (Eckner *et al*, 1994). The sequence of clone mCTG 57 shows similarity to a region corresponding to a region encompassing amino acid residues 2232 to 2351 inclusive of human P300. This corresponds to positions 250 to 369 in the multiple sequence alignment in Figure 4.13.

More recently, P300 has been suggested to be a member of a growing family of proteins which appear to interact with the phosphorylated form of CREB (Lundblad *et al*, 1995), a transcription factor (Chrivia, *et al*, 1993), as well as interacting with the E1A product (Arany *et al*, 1995). The sequence of clone mCTG 57 shows most similarity to human P300, including the gaps between P300 and the sequences which represent the CREB binding proteins of mouse and human (Figure 4.13, Mm-CBP and Hs-CBP respectively, Chrivia *et al*, 1993). It is therefore likely that clone 57 represents a mouse cognate of human P300. The CAG/CTG repeat in the sequence is likely to represent a polyglutamine repeat, six residues long, as indicated in Figure 4.13. This is in contrast to 2 glutamines at the same position found in the human P300 and three glutamines in the sequences of mouse and human CREB binding proteins.

Figure 4.10.A

```
           251                                                  300
      411  .......... .......... .......... .......... ..........
RnNopp140  .......... .......... .......... .......... .........C
HsNopp140  .......... .......... .......... .......... ..........
HsNopp130  .......... .......... .......... .......... ..........

           301                                                  350
      411  .......... .....GCAGG AATTCCATGT CTTACGGTCA AGGGCCCAGG
RnNopp140  ATCCAGTGAC AGCAGTGAGG ACAGCAGTGA GGAAGAGGAC AAAGCCCA..
HsNopp140  ........GT GACAGTGAGG ACAGCAGCGA GGAGGAGGAG GAAGTTCAAG
HsNopp130  CTCATCCAGT GACAGTGAGG ACAGCAGCGA GGAGGAGGAG GAAGTTCAAG

           351                                                  400
      411  GACTTCCCAC ACAGAAGGCT GCCGCACAGG CCAAGCGAGC CAGTGTGCCT
RnNopp140  .AGTTCCCAC ACAGAAGGCT GCCGCCCCTG CCAAGCGAGC CAGTTTGCCT
HsNopp140  GGCCTCCAGC AAAGAAGGCT GCTGTACCTG CCAAGCGAGT CGGTCTGCCT
HsNopp130  GGCCTCCAGC AAAGAAGGCT GCTGTACCTG CCAAGCGAGT CGGTCTGCCT

           401                                                  450
      411  CAGCATGCTG GGAAAGGCAG CAGCAAAGCT TCAGAGAGCA GCAGTAGTGA
RnNopp140  CAGCATGCTG GGAAAGCAGC AGCCAAAGCT TCAGAGAGCA GCAGTAGTGA
HsNopp140  C......CTG GGAAGGCTGC AGCCAAAGCA TCAGAGAGTA GCAGCAGTGA
HsNopp130  C......CTG GGAAGGCTGC AGCCAAAGCA TCAGAGAGTA GCAGCAGTGA

           451                                                  500
      411  AGAATCCAGT GATGAGGAAG AGGAAGAGGA CAAAAAGAAA AAGCCTGTC.
RnNopp140  AGAGTCCAGT GAGGAAGAGG AGGAGAAGGA CAAAAAGAAA AAGCCTGTC.
HsNopp140  AGAGTCCAGT GATGATGATG ATGAGGAGGA CCAAAAGAAA CAGCCTGTC.
HsNopp130  AGAGTCCAGA GATGATGATG ATGAGGAGGA CCAAAAGAAA CAGCCTGTC.

           501                                                  550
      411  ..CAGAAGGC AGCTAAGCCC CAAGCCAAGG CAGTCAGACC TCCTGCGAAG
RnNopp140  ..CAGAAAGC AGTTAAGCCC CAAGCCAAGG CAGTCAGACC TCCTCCGAAG
HsNopp140  ..CAGAAGGG AGTTAAGCCC CAAGCCAAGG CAGCCAAAGC TCCTCCTAAG
HsNopp130  ..CAGAAGGG AGTTAAGCCC CAAGCCAAGG CAGCCAAAGC TCCTCCTAAG

           551                                                  600
      411  AAAGCAGAGA GCTCTGAGTC GGACTCGACA TCGG...... ..........
RnNopp140  AAGGCAGAGA GCTCTGAGTC CGAGTCTGAC TCAAGCTCAG AGGATGAAGC
HsNopp140  AAGGCCAAGA GCTCTGATTC TGATTCTGAC TCAAGCTCCG AGGATGAGCC
HsNopp130  AAGGCCAAGA GCTCTGATTC TGATTCTGAC TCAAGCTCCG AGGATGAGCC

           601            618
      411  .......... ........
RnNopp140  ACCAC..... ........
HsNopp140  ACCAAAGAAC CAGAAGCC
HsNopp130  ACCAAAGAAC ........
```

Figure 4.10.B

```
           101                                                  150
      411  PTQKAAAQAK RASVPQHAGK GSSKASESSS SEESSDEEEE EDKKKKPV.Q
RnNopp140  PTQKAAAPAK RASLPQHAGK AAAKASESSS SEESSEEEEE KDKKKKPVQQ
HsNopp140  PAKKAAVPAK RVGLP..PGK AAAKASESSS SEESSDDDDE EDQKKQPV.Q
HsNopp130  PAKKAAVPAK RVGLP..PGK AAAKASESSS SEESRDDDDE EDQKKQPV.Q

           151                                                  200
      411  KAAKPQAKAV RPPAKKAESS ESDSTS.... .......... ..........
RnNopp140  KAVKPQAKAV RPPPKKAESS ESESDSSSED EAPQTQKPKA AATAAKAPTK
HsNopp140  KGVKPQAKAA KAPPKKAKSS DSDSDSSSED EPPKNQKPKI TPVTVKAQTK
HsNopp130  KGVKPQAKAA KAPPKKAKSS DSDSDSSSED EPPKNQKPKI TPVTVKAQTK

           201                                                  250
      411  .......... .......... .......... .......... ..........
RnNopp140  AQTKAPAKPG PPAKAQPKAA NGKAGSSSSS SSSSSSDDSE EEKKAAAPLK
HsNopp140  APPK...... .PARAAPKIA NGKAASSSSS SSSSSSSDDS EEEKAAATPK
HsNopp130  APPK...... .PARAAPKIA NGKAASSSSS SSSSSSSDDS EEEKAAATPK
```

Figure 4.10 Alignment of A) nucleotide and B) protein sequences which show similarity to clone mCTG 411.
The sequences represented are: those derived from clone 411 (411); *Rattus norvergicus* 140 kDa nucleophosphoprotein (RnNopp140, Meier and Blobel, 1992), *Homo sapiens* 140 kDa nucleophosphoprotein (HsNopp140, Nomura *et al*, 1994) and *Homo sapiens* 130 kDa nucleophosphoprotein (HsNopp130, Pai *et al*, 1995). mCTG 411 sequences are indicated in blue coloured text and the serine repeat which is encoded by a CAG/CTG repeat in the comparison sequences (but is not shown for clone 411 because of unconfirmed sequence data) is indicated in green coloured text.

Figure 4.11

```
            51                                                  100
      56    .......... .......... .......... .......... ..........
   CW17R      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR
 Zfm1-II      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR
 Zfm1- I      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR
SF-1 hl1      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR
 SF-1 b0      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR
Zfm1-III      AYIVQLQIED LTRKLRTGDL GIPPNPEDRS PSPEPIYNSE GKRLNTREFR

            101                                                 150
      56    .......... .......... .......... .......... ..........
   CW17R      TRKKLEEERH TLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI
 Zfm1-II      TRKKLEEERH NLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI
 Zfm1- I      TRKKLEEERH NLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI
SF-1 hl1      TRKKLEEERH NLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI
 SF-1 b0      TRKKLEEERH NLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI
Zfm1-III      TRKKLEEERH NLITEMVALN PDFKPPADYK PPATRVSDKV MIPQDEYPEI

            151                                                 200
      56    .......... .......... .......... .......... ..........
   CW17R      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVEEGKVGR KDGQMLPGED
 Zfm1-II      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVKEGKVGR KDGQMLPGED
 Zfm1- I      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVKEGKVGR KDGQMLPGED
SF-1 hl1      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVKEGKVGR KDGQMLPGED
 SF-1 b0      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVKEGKVGR KDGQMLPGED
Zfm1-III      NFVGLLIGPR GNTLKNIEKE CNAKIMIRGK GSVKEGKVGR KDGQMLPGED

            201                                                 250
      56    .......... .......... .......... .......... ..........
   CW17R      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL
 Zfm1-II      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL
 Zfm1- I      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL
SF-1 hl1      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL
 SF-1 b0      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL
Zfm1-III      EPLHALVTAN TMENVKKAVE QIRNILKQGI ETPEDQNDLR KMQLRELARL

            251                                                 300
      56    .......... .......... .......... .......... ..........
   CW17R      NGTLREDDNR ILRPWQSSET RSITNTTVCT KCGGAGHIAS DCKFQRPGDP
 Zfm1-II      NGTLREDDNR ILRPWQSSGT RSITNTTVCT KCGGAGHIAS DCKFQRPGDP
 Zfm1- I      NGTLREDDNR ILRPWQSSET RSITNTTVCT KCGGAGHIAS DCKFQRPGDP
SF-1 hl1      NGTLREDDNR ILRPWQSSET RSITNTTVCT KCGGAGHIAS DCKFQRPGDP
 SF-1 b0      NGTLREDDNR ILRPWQSSET RSITNTTVCT KCGGAGHIAS DCKFQRPGDP
Zfm1-III      NGTLREDDNR ILRPWQSSET RSITNTTVCT KCGGAGHIAS DCKFQRPGDP

            301                                                 350
      56    .......... .......... .......... .......... ..........
   CW17R      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPAAPA
 Zfm1-II      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPAAPA
 Zfm1- I      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPARPA
SF-1 hl1      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPAAPA
 SF-1 b0      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPAAPA
Zfm1-III      QSAQDKARMD KEYLSLMAEL GEAPVPASVG STSGPATTPL ASAPRPARPA

            351                                                 400
      56    .......... .......... .......... .......... ..........
   CW17R      SNPPPPSLMS TTQSRPPWMN SGPSENRPYH GMHGGGPGGP GGGPHSFPHP
 Zfm1-II      NNPPPPSLMS TTQSRPPWMN SGPSESWPYH GMHGGGPGGP GGGPHSFPHP
 Zfm1- I      NNPPPPSLMS TTQSRPPWMN SGPSESRPYH GMHGGGPGGP GGGPHSFPHP
SF-1 hl1      NNPPPPSLMS TTQSRPPWMN SGPSESRPYH GMHGGGPGGP GGGPHSFPHP
 SF-1 b0      NNPPPPSLMS TTQSRPPWMN SGPSESRPYH GMHGGGPGGP GGGPHSFPHP
Zfm1-III      NNPPPPSLMS TTQSRPPWMN SGPSESRPYH GMHGGGPGGP GGGPHSFPHP
```

Figure 4.11 (continued)

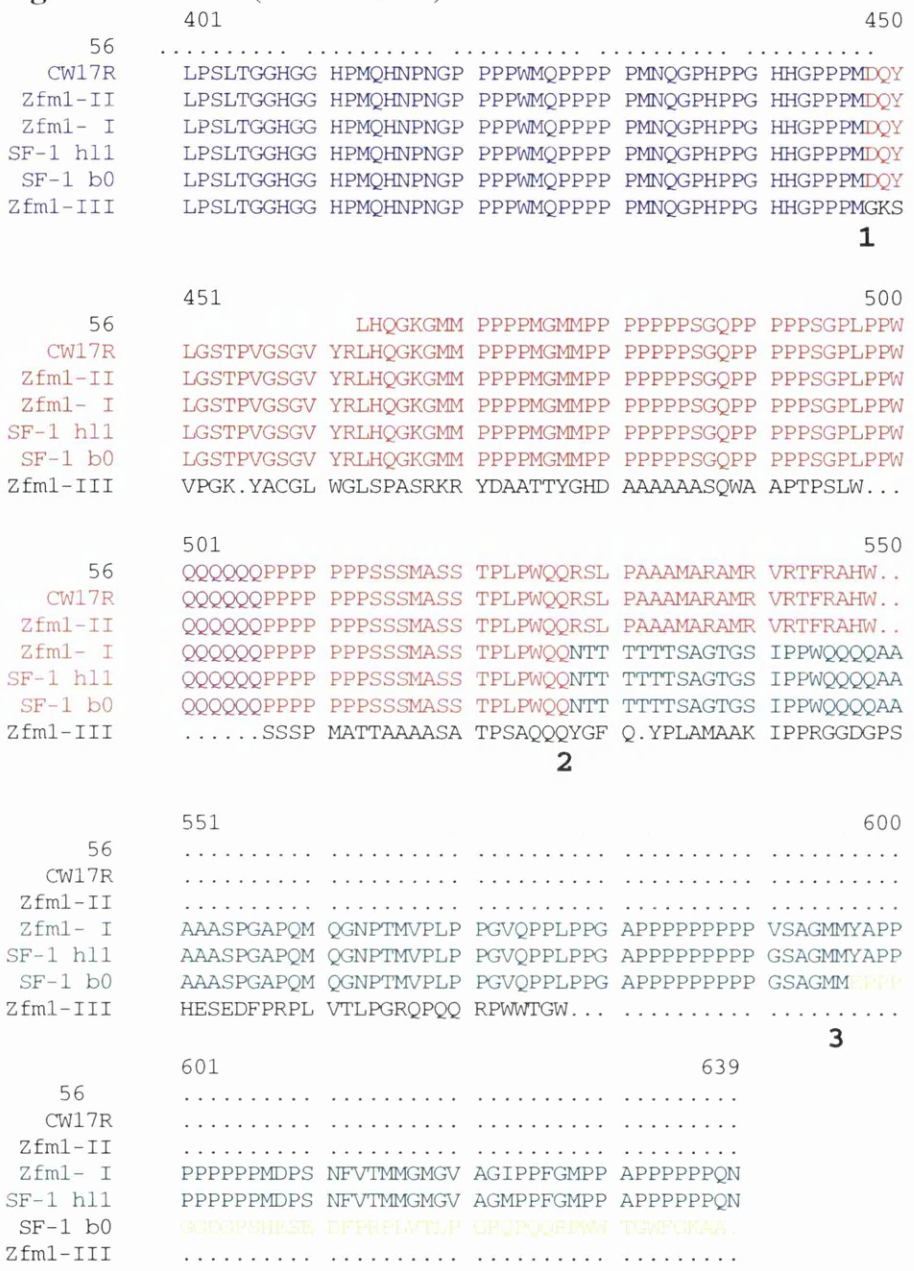


Figure 4.11 Alignment of putative translation of sequences from clone mCTG 56 and proteins which show similarity. The numbers in bold represent where protein sequences diverge. The CAG repeat is represented by a purple Q run. Represented sequences are *Mus musculus* CW17R (Whrekle *et al*, unpublished AC: S52735); *Homo sapiens* Zfm1-I isoform (Toda *et al*, 1994), *Homo sapiens* Zfm1-II isoform (Breviario *et al*, unpublished, AC: L49380); *Homo sapiens* Zfm1-III (Breviario *et al*, unpublished, AC: L49345) and *Homo sapiens* splicing factor 1 isoforms SF-1 hl1 and SF-1 B0 (Arning *et al*, 1996). Colourise text corresponded to the colours given to diverged and related sequences in Figure 4.12.

Figure 4.12

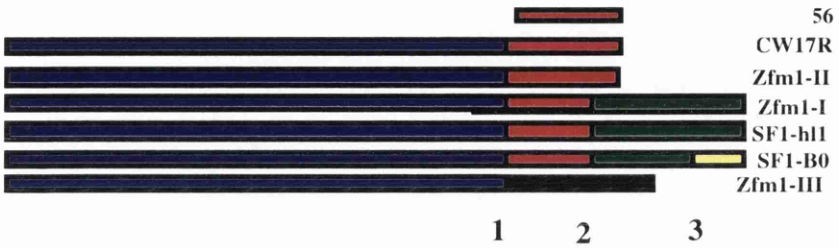


Figure 4.12 Diagrammatic representation of relatedness and putative splicing events (1, 2 and 3) between sequences related to Clone mCTG 56.

Blue filling colour represents homologous sequences shared by Clone mCTG 56, mouse CW17R, splicing factor 1 and human Zfm1 isoforms. Red represents sequences which are common to a group of sequences which includes a putative translation of Clone mCTG 56 sequences. Green filling colour represents sequences common to sequences Zfm1-II (Breviario *et al*, unpublished), SF-1 hl1 and SF-1 B0 isoforms (Arning *et al*, 1996) which starts at splice position 2. Yellow filling colour represents sequences unique to the SF-1 isoform B0 which start at position 3. Black filling colour indicates the unique sequence present in the Zfm1-III isoform (Breviario *et al*, unpublished), after the divergence of this sequence at splice position 1.

Figure 4.13

Figure 4.13 Sequence alignment of translation of clone 57 and protein sequences which show similarity.

A translation of sequences derived from clone mCTG 57 (57) aligned with *Homo sapiens* P300 transcriptional adaptor (HsP300; Eckner *et al*, 1994), *Mus musculus* CREB binding protein (Mm-CBP; Chrivia, *et al*, 1993) and *Homo sapiens* CREB binding protein (Hs-CBP; Chrivia *et al*, 1993).

4.5 Clones with similarity to expressed sequence tags (ESTs).

As described in Section 2 of this Chapter, three clones (mCTG 12, 46 and 59) show similarity to EST sequences derived from mass sequencing projects. In most cases, little is known about the sequence, except the details of the library from which the clone was isolated. It is likely that they represent novel genes.

4.5.1 Clone mCTG 12.

Clone mCTG 12 was isolated from the 8.5 dpc library (Hogan *et al*, unpublished). A search using the BLASTN algorithm (Altschul *et al*, 1990) with sequence obtained by priming with the extended T3 primer (T3/T7a, see Table 2.3, Chapter 2) revealed that this clone had similarity to a group of human and mouse ESTs. In addition to this, a search of databases with the BLASTN search algorithm using sequence obtained by priming with the extended primer bT7 (see Table 2.3, Chapter 2) which generates sequence from the other end of the insert found similarity to a mouse EST, W715089, (Figure 4.14.B).

The EST identified as having the highest similarity to clone mCTG 12 (Z78337, see Figure 4.14.A) was derived from a specific human oriented project which isolated CAG repeat containing clones from cDNA libraries (Neri *et al*, 1996). The clone described by Neri *et al* (1996), ICRFp507H02194, has 6 CAG trinucleotides. This is in agreement with the CAG repeat found in Clone mCTG 12, although the repeat is interrupted by 2 TAG triplets. This repeat may also be larger as it is adjacent to vector sequences in the clone obtained.

4.5.2 Clone mCTG 46.

A BLASTN search found that this clone had similarity to a mouse clone which corresponds to an EST, W91417. This sequence was derived from a 13.5/14.5 dpc whole mouse embryo (inbred strain, C57BL/6J) cDNA library. At the nucleic acid level the similarity was found to be 95.6%. Both sequences were found to contain a large perfect CAG/CTG repeat, 25 trinucleotides in length. There are also two CTG trinucleotides immediately upstream of the large repeat. This repeat may be translated as a polyalanine or polyserine homopeptide. Translation as polyglutamine is unlikely as there is a consistent, in frame stop codon upstream of the repeat in both sequences, adjacent to the CAG repeat. This is represented in Figure 4.15 as the underlined trinucleotide CTA (the opposite strand sequence is TAG). Furthermore, the EST, W91417, represents the 5' end sequences of a clone from a library which had been random primed with oligo-dT. The opposite end contains a poly A tail which restricts the possible reading frames that this trinucleotide repeat can be translated (*i.e.* GCT, CTG or TGC). Alternatively the repeat may lie in an untranslated region.

4.5.3 Clone mCTG 59.

This sequence showed similarity to a single human EST, W91417, derived from a human foetal spleen cDNA library. The homology only extended to a 122 nucleotide overlap (Figure 4.16).

Figure 4.14.A

```
        201                                                         250
   12   CAGTAGCAGC AGCAGCAGTA GCAAATCAAA CGATCAGCTC GGATGTGTGG
z78337  TATCACCAGC AGCAGNAGNA GCAGATCAAA CGGTCAGCCC GCATGTGTGG
z42809  .......... .......... .......... .......... ..........
r17799  .......... .......... .......... .......... ..........
r61592  .......... .......... .......... .......... ..........

        251                                                         300
   12   TGAGTGCGAG GCCTGCCGAT GCACTGAGGA CTGTGGCCAC TGTGACTTCT
z78337  TGAGTGTGAG GCATGTCGGC GCACTGAGGA CTGTGGTCAC TGTGATTTCT
z42809  .......... .......... .......... .......... ..........
r17799  .......... .......... .......... .......... ..........
r61592  .......... .......... .......... .......... ..........

        301                                                         350
   12   GCCGTGACAT GAAGAAGTTT GGGGGCCCCA ACATGATCCG GTAGAAGTGY
z78337  GTCGGGACAT GAAGAAGTTC GGGGGCCCCA ACAAGATCCG GCAGAAGTGC
z42809  .......... .......... .......... .......... ..........
r17799  .......... .......... .......... .......... ..........
r61592  .......... .......... .......... .......... ..........

        351                                                         400
   12   CGGCTTCGTC AGTGTCAGCT GCGGGCACGG GAATCGTACA AGTACTTCCC
z78337  CGGCTGCGCC AGTGCCAGCT GCGGGCCCGG GAATCGTACA AGTACTTCCC
z42809  .......... .......... .......... ....CGTACA AGTACTTCCC
r17799  .......... .......... .......... .....GTACA AGTACTTCCC
r61592  .......... .......... .......... .....GTACA AGTACTTCCC

        401                                                         450
   12   TTCCTCGTTC TCGCCGGTGA CACCCTCAGA GGCCCTGTCA AGGCCCCGTC
z78337  TTCCTCGCTC TCACCAGTGA CGCCCTCAGA GTCCCTGCCA AGGNCCCGCC
z42809  TTCCTCGTTC TCACCAGTGA CGCCCTCAGA GTCCCTGCCA AGGCCCCGCC
r17799  TTCCTCGCTC TCACCAGTGA CGCCCTCAGA GTCCCTGCCA AGGCCCCGCC
r61592  TTCCTCGCTC TCACCAGTGA CGCCCTCAGA GTCCCTGCCA AGGCCCCG.C

        451                                                         500
   12   GGCCACCACC CACTCAACAG TAGCCACAGC AGTCCCAGAA GCTGGGGCGT
z78337  GGGCACTGTC CACCCAACAG GAGCCACAGC CATCACAGAA GTTAGGGCGC
z42809  GGCCACTGCC CACCCAACAG CAGCCACAGC CATCACAGAA GTTAGGGCGC
r17799  GTCCACTGCC CACCCAACAG CAGCCACAGC CATCACAGAA GTTAGGGCGC
r61592  GACCACTGCC CACCCAACAG CAGCCACAGC CATCACAGAA GTTAGGGCGC

        501                                                         550
   12   TTTCGTGAAG ATGAGGGGAC AGCGTTGTCA TCAGTGGTTA AGGAGGCACC
z78337  ATCCGTGAAG ATGAGGGGGC AGTGGCGTCA TCAACAGTCA AGGAGCCTCC
z42809  ATCCNTGAAG ATGAGGGGGC AGTGGCGTCA TCAACAGTCA AGGAGCCTCC
r17799  ATCCGTGAAG ATGAGGGGGC AGTGGCGTCA TCAACAGTCA AGGAGCCTCC
r61592  ATCCGTGAAG ATGAGGGGGC AGTGGCGTCA TCAACAGTCA AGGAGCCTCC

        551                                                         600
   12   AGAGGCTACA GCAACAACTG AGCCACTTTC AGATGAGG.. ACCTAGCACT
z78337  TGAGGCTACA GACACACCTG AGCCACTCTC AGATGAGGGA CCCACCCTCT
z42809  TGAGGCTACA GCCACACCTG AGCCACTTTC AGATGAGG.. ACCTACCTNT
r17799  TGAGGCTACA GCCACACCTG AGCCACTCTC AGATGAGG.. ACCTACCTCT
r61592  TGAGG.TACA GCCACACCTG AGCCACTCTC AGATGAGG.. ACCTACCTCT
```

Figure 4.14.A Alignment of nucleotide sequences derived from clone 12 and ESTs which show similarity.

Alignment of clone mCTG 12 sequences (blue text) and: Z78337, clone ICRFp507H02194 derived from a human foetal brain cDNA library (Neri *et al*, 1996), Z42809 and R17799 (human 5' ESTs).

Figure 4.14.B

Figure 4.14.B. Comparison of sequences derived from clone 12 and mouse EST W71508.

The clone mCTG 12 sequence described here was derived from a T7 primed sequencing reaction and representing sequence from the opposite end of clone mCTG 12 to that described in Figure 4.14.A. The mouse EST described here, W71508, was derived from a 13.5 dpc whole mouse embryo cDNA library. Clone mCTG 12 sequences are indicated in blue text. Bars indicate positions of identity, : indicates similar nucleotides and spaces non-identical nucleic acids at those particular positions.

Figure 4.15

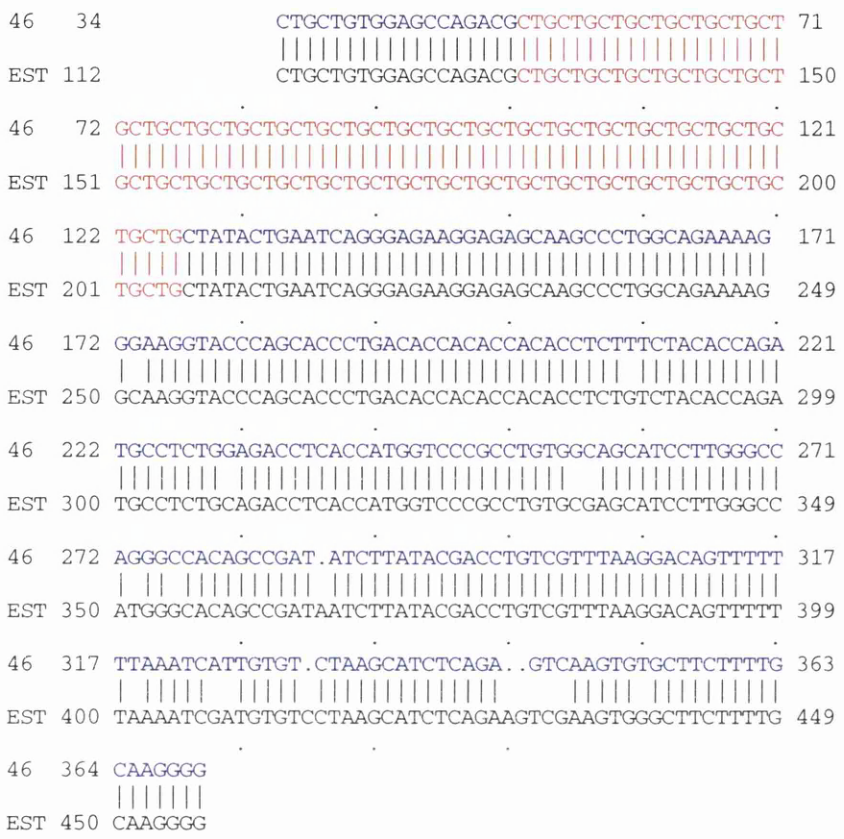


Figure 4.15 Sequence alignment of sequences derived from clone mCTG 46 and mouse 5' EST W91417.
Red sequence represents the CAG/CTG trinucleotide identified in clone mCTG 46 (sequences indicated in blue coloured text) and the mouse EST W91417. The underlined CTA trinucleotide represents a stop codon (in the opposite strand) which is immediately upstream and in frame to prevent translation of the trinucleotide as polyglutamine. For further details refer to the main text. Bars indicate positions of identity, : indicates similar nucleotides and spaces non-identical nucleic acids at those particular positions.

Figure 4.16

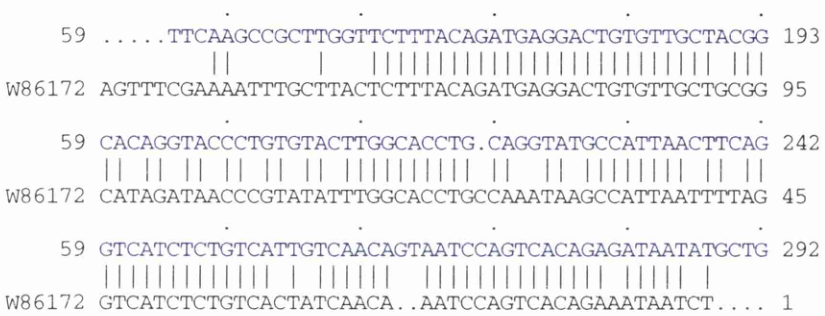


Figure 4.16 Alignment of sequences derived from clone 5 9 and human EST W86172.

This human EST was derived from a foetal liver spleen cDNA library. The sequence represented here represents the 3' end of clone 416399. Clone mCTG 59 sequences are indicated in blue coloured text. Bars indicate positions of identity, : indicates similar nucleotides and spaces non-identical nucleic acids at those particular positions.

4.6 Clones with no similarity to any sequence.

These clones, as of the 6th of November 1996, having been used to search databases with the algorithms, BLASTX, BLASTN (Altschul *et al*, 1990) and FASTA (Lipman and Pearson, 1985) do not show any identity, similarity or homology at either in the nucleic acid or amino acid level to any sequence found in the databases. These are clones mCTG 26, 27, 28, 210, 45, 414, 61, 81, 82, 86 and 92. The sequence of these clones can be found in appendix A of this thesis.

4.7 Discussion.

The results presented in this Chapter prove conclusively that the library screening method adopted in this study was a success. An initial guide was the output numbers from the primary cDNA library screens, described in Chapter 3, and that they hybridised to the repeat oligonucleotide probe consistently through further rounds of purification.

4.7.1 Size, structure and numbers of trinucleotide repeats.

Partial sequencing of the first 23 lambda probe positive inserts sub-cloned into the phagemid vector, pBluescript KS-, identified CAG/CTG triplet repeats in 21 of the inserts sequenced. These CAG/CTG repeats exist in a wide range of sizes ranging from 6 to 26 trinucleotide repeats (see Table 4.1). This spectrum of sizes is also evident in the human study (Li *et al*, 1993) which used the identical screening conditions. The smallest repeat in their study is

5 CAG/CTG triplets in length. The largest uninterrupted repeat was 15 triplets in length.

In this respect the mouse data presented here exceeds this. Eight clones have perfect repeats with length greater than or equal to 15 triplets (clones mCTG 28, 210, 46, 61, 63, 81, 82 and 92). Other large repeat sequences contain imperfections in the repeating unit. These are relatively common. Cryptic or imperfect repeats are described in both the Riggins *et al* (1992) and Li *et al* (1993) human studies, as well as in the data presented in this work. In addition they are observed in mouse in the work of Abbott and Chambers (Abbott and Chambers, 1994; Chambers and Abbott, 1996). These cryptic repeats may have accumulated substitutive nucleotide base mutations (transitions and transversions) in an ancestral repeat during evolution. Alternatively the repeat structure is a product of convergent evolution, where there was an advantage for the accumulation of a cyclical repetition. Insertions and deletions of bases tend to be absent as they would disrupt the open reading frame, if the repeat existed in a translated region. Furthermore, a large number of these mutations tend to be in the 3rd position of codons in a presumptive open reading frame. Due to the nature of the genetic code, this leads to no change in the amino acid encoded by that codon, in many cases. The 3rd position of codons are known to be the most variable position, where different bases can be tolerated in accepting the same amino acid. This effect is due to the anticodon wobble. Accumulation of this type of mutation in translated trinucleotide repeats without disturbing the amino acid homopeptide argues for some functional significance of that homopeptide and thus of the repeat.

Recent work done by Karlin and Burge (1996) noted a preponderance of multiple repeats in certain proteins of higher eukaryotes. These can consist of repeats of the same or different trinucleotide, encoding different amino acids. Their work concentrated on *Drosophila* and human proteins. They mention mouse data, but do not discuss it. The partial sequence data described in this Chapter show some evidence for this in mouse genes. Most of these apparently encode CCN type repeats and are likely to encode polyproline stretches (where N equals A, C, G or T). They were first identified in *Drosophila* proteins, were named PEN repeats (Digan et al, 1986) and are believed to be functionally significant. Gerber *et al* (1994) have shown that polyproline homopeptides have a stimulatory effect on basal transcription. Polyproline and polyglutamine have also been found in close proximity to each other in certain proteins. One example is the human huntingtin protein, which possesses two proline stretches, one of which is adjacent to the glutamine stretch, alterations in the length of which have been associated with changes in the severity of the disease (Pecheux *et al*, 1995).

There are examples of other repeats. For example, there is a GAG repeat in clone mCTG 43 that encodes a glutamic acid repeat. This is in addition to the CAG trinucleotide which encodes a polyglutamine tract (see Figure 4.8). This is part of a larger polypurine stretch of which the remainder forms a hexanucleotide repeat, which encodes an alternate di-amino acid repeat, glutamate-lysine (EK).

4.7.2 Comparison with other CAG/CTG containing cDNA screening studies.

In a recent study, Abbott and Chambers (1996) screened a mouse adult brain cDNA library for CAG/CTG trinucleotide repeats and found a similar number of CAG/CTG trinucleotide repeat positive clones (60 positives per 20000 pfu screened) compared to that observed by myself in the adult mouse brain cDNA library (53 clones per 20000 pfu screened). However, when sequence analysis was performed the largest repeat size was 10 copies of CAG (C. Abbott, personal comm.). This maximum is smaller than the repeat lengths which I observed in clones derived from the 8.5 dpc and 12.5 dpc mouse whole embryo cDNA libraries. It may be possible that CAG/CTG trinucleotide repeats that are expressed are smaller in the adult brain compared to those in development. However, from the reverse dot-blot experiments with some of the clones derived from the 8.5 dpc and 12.5 dpc embryonic cDNA libraries described in Chapter 3 of this work, clones which contain large trinucleotide repeats are expressed not only throughout development but also in adult tissues including mouse brain. Therefore the smaller repeat size derived by Abbott and Chambers (1996), must be caused by some other influencing factor. Abbott and Chambers themselves suggest that this effect may be caused by the fact that the library which they used was derived from oligo-dT primed first strand cDNA synthesis and the resulting library is biased towards 3' transcript clones. An alternative explanation is that the hybridisation conditions under which the libraries were screened were different. Abbott and Chambers (1996) used a wholly aqueous based hybridisation solution combined with a high

temperature of hybridisation (59°C). The hybridisation conditions which I used to screen the cDNA libraries screened here differ in that the hybridisation solution was 50% formamide with a hybridisation temperature of 42°C. Formamide is known to be able to de-stabilise secondary structure. It also known to increase the effective temperature of hybridisation. Therefore the effective temperature of my conditions is 72°C. The secondary structure de-stabilising effect may have contributed to a larger size of repeat being available for hybridisation to the probe $(CTG)_{10}$, used in the experiments described in this work. An additional factor in detecting larger repeat sequences in this work would be the larger size of the oligonucleotide probe used. It is known that nucleic acid hybrid stability is dependent on the length over which complementarity can be achieved. This phenomenon is more pronounced for smaller probes than for larger ones. Therefore a probe which has 10 copies of the trinucleotide CTG, as used in the experiments described in this thesis will form more stable hybrids than the one with only five copies of CTG which was used by Abbott and Chambers (1996).

4.7.3 Cross species relationships.

As an assessment of human-mouse trinucleotide repeat conservation, a comparison of trinucleotide repeat sequences was performed. It was found that in all cases that the size of the repeat sequence of the mouse was greater than or equal to the repeat sequence found in the human homologue (see Figure 4.4).

In a larger repeat (mCTG 23, a homologue of RepA 70 kDa subunit) where a mouse-human comparison could be made there

seems to be more of a divergence between species. In this case, the mouse gene appears to have an expanded CAG repeat compared to both the human and *Xenopus* homologues which do not (See Figure 4.5). This observation also supports the idea of ascertainment bias where the selection of large microsatellite repeats from one organism are smaller than in others rather than the other theory proposed by Rubinzstein *et al* (1995) which proposes that different organisms have different rates of microsatellite dynamics. However from the work presented here, it appears that the mouse repeats are larger than the human counterpart. Only in 1 out of 5 cases where a mouse-human comparison could be made (*i.e.* clone mCTG 56, Figure 4.4, Figure 4.7), is there no change in the number of repeats. Furthermore, there are no examples in the work presented here, where the CAG/CTG trinucleotide repeat identified in mouse sequence is smaller than in the comparable human sequence. This is a small sample number and any general conclusion will have to be tested with more comparisons. The isolation of human homologues for more of the mouse genes containing CAG/CTG trinucleotide repeats presented here in this work would help in extending this observation.

This data is also consistent with the work of Beckmann and Weber (1992) whose survey of rat and human microsatellites indicated that in general, rodent microsatellites were longer than human microsatellites, However, it was not indicated whether rodent microsatellite sequences that were conserved at orthologous positions between distantly related mammalian species were larger. A separate trinucleotide microsatellite study (Stallings, 1994) found that there was some evidence of conservation of trinucleotide

repeats at the same position, but in evolutionarily distant mammal species and those examples studied were a small fraction of the overall number trinucleotide repeat microsatellites analysed and these tended to be in protein coding sequences. Of those examples indicated, one had larger repeats in the rodent gene, one larger in the human gene (these, were both in the same gene, the androgen receptor) and two did not have the repeat conserved in the human gene sequence. The final example had an identical number of amino acids in the human and rodent gene sequence. This non conservation is not a general observation about microsatellites. In separate studies, looking at (GT)n/(CA)n and (CT)n/(GA)n dinucleotides, Stallings (1991 and 1995, respectively) showed that there is some degree of conservation at orthologous positions for these microsatellites and in the work dealing with the (GT)n/(CA)n dinucleotides (Stallings, 1991), it is indicated that there is more conservation between more closely related mammalian species than more distantly related ones.

4.7.4 Clones that are similar to other sequences.

At first sight there appears to be no similarity in general function between any of the clones that have similarity to known proteins. However seven of the nine genes which have known functions, are nucleic acid binding proteins. Three bind DNA (two double-stranded DNA, 24, 57; one single-stranded DNA, 23) and three bind RNA (clones 29, 410, 411). A final clone, mCTG 56, shows identity to both a DNA binding transcription factor (Zfm1; Toda *et al*, 1994) and an RNA binding protein (SF1, a splicing factor; Arning *et al*, 1996)

In addition, all the nine proteins are known to form complexes with other proteins. This should be no surprise, since as mentioned in the Introduction (Section 1.9.1), glutamine homopeptides have been shown to adopt a specific tertiary structure which will enable them to behave as protein-protein interaction domains. This however would only apply to repeats that were proved to be translated into glutamine tracts. Most of the repeats that occur in genes whose proteins have a defined function, appear to encode glutamine homopeptides. Five encode glutamines (clones 23, 24, 410, 56 and 57), three are undefined (clones 29, 46 and 59) and one encodes a serine repeat (clone 411). The final clone 63 occurs in presumed 3' UTR sequences. This clone will be discussed further in Chapter 6.

4.7.5 Multiple hits with genes of known function.

In database searches, it is often the case that there is no similarity with a known sequence and if there is, the sequence found may not have any assigned function. Just as perplexing may be the case when multiple genes of known function are identified by a single sequence. This is the case with clones 43 and 56 in this study. They appear to have remarkable similarity to either two (in the case of clone 56) or three (in the case of clone 43) genes.

For clone mCTG 43, these genes are a mouse beta-1,4-galactosyltransferase, a mouse calcium binding protein PW29 and the Rad21 protein which is involved in double-strand break repair in eukaryotes. All have been characterised by different groups. Clone 56 shows identity to the splicing factor, SF1 (Arning *et al*, 1996): the transcription factor Zfm1 (Toda *et al*, 1994) and a mouse

cDNA, CW17. The sequences all appear to share identical N-terminal domains (Figure 4.12).

This can have occurred because of a few reasons. Firstly that the sequence presented here represents a common domain shared by separate genes. Alternatively this sequence may represent shared exons between these genes. Another alternative explanation is that they are the same gene and separate groups working in separate specific fields of interest have only partially identified the functions of a single gene product. The identification of such cross references is likely to be a by-product of random screens such as this work. There are reports of EST and STS sequences generated from separate mass DNA sequencing experiments being clustered into groups of related sequence (grouped together by programmes such as Beauty; Worley *et al*, 1995). This is exemplified by the multiple ESTs that have been found to have similarity to Clone mCTG 29 (Section 4.4.3 and Figure 4.7) and Clone mCTG 12 (Section 4.5.1 and Figure 4.14.A). However, as far as it is known, the examples presented here are the first grouping of sequences which encode proteins of known functions.

Chapter 5

Further Characterisation Of Selected Clones.

5.1 Introduction

Because this research is focused on developmentally regulated genes which contain CAG/CTG trinucleotide repeats, the criteria for selecting clones for further study were: the size of the repeat (as described in Chapter 3), initial expression data (as described by the reverse northern data in Chapter 3) and presumed novelty. In this last respect, after initial database searches using sequences from these clones, four clones from the 8.5 and 12.5 dpc whole mouse embryo cDNA libraries were selected for further characterisation. These were mCTG 26 and 210 (from the 8.5 dpc whole mouse embryo cDNA library) and clones mCTG 61 and 63, both from the 12.5 dpc whole embryo cDNA library (Logan *et al*, 1992). Subsequent database searches identified sequences in the database which show similarity to some of these cloned genes.

For the characterisation of a gene it is important to know whether it is a single copy gene or belongs to a family of related genes, where and when it is expressed in the embryo and the adult organism, if there are single or multiple transcripts and the size of messenger RNA(s). Reverse dot blotting experiments (Chapter 3, this work), gives valuable information about the expression levels of these genes, but does not give either the size or numbers of mRNA species. Electrophoretic separation and subsequent hybridisation of RNA species (northern blot) with probes derived from specific clones will allow the determination of the size of the messenger RNA species and whether multiple mRNA species exist.

The previous Chapters detailed two routes to characterise clones identified from the 8.5 and 12.5 dpc whole mouse embryo libraries. Firstly, reverse dot-blot experiments were used to identify clones

which showed expression in embryonic or adult mouse tissues. Secondly, DNA sequencing was employed to characterise the repeat. This Chapter presents the further molecular characterisation of four selected clones, mCTG 26, 210, 61 and 63. This will cover the full sequencing of the clones isolated, initial characterisation of the genomic organisation of the sequences and genes from which the cDNA clones were derived from and finally characterise the mRNA species expressed at time points through development and in selected adult tissues (brain and liver).

5.2 Further analysis of Clone mCTG 26.

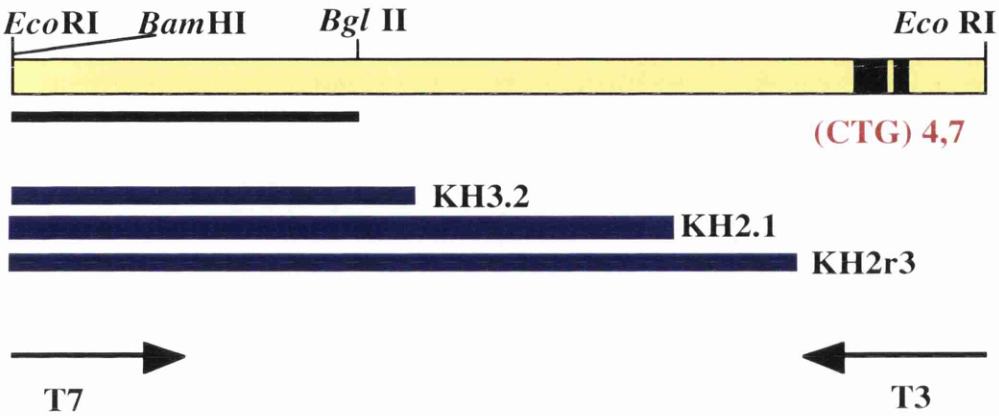
This clone was originally derived from the 8.5 dpc whole mouse embryo library (Hogan *et al*, unpublished). It was found to be present in lower amounts in adult tissues, than in embryos. Within adult tissues it appeared to be more abundant in mouse brain compared to liver. Initial sequence analysis indicated that this clone had a perfect trinucleotide repeat of 12 repetitive units. With this preliminary information, further characterisation was undertaken.

5.2.1 Sequence analysis of Clone mCTG 26.

In order to obtain a complete sequence of this clone, a series of end deletions was generated using the Erase-a-base kit (Promega) and sequenced. A summary of the deletions obtained and clone information is given in Figure 5.1. The ABI automated sequencing apparatus with adapted, extended sequence primers, T3/T7α and bT7 was used. The sequence was assembled using the

Figure 5.1

A)



B)

```
          10                  30                  50
GAATTCATGTCTTACGGTCAAGGGACAACGGCCCCTGGCAGCCGCTCGGCCCCCTCAGAG
---------+---------+---------+---------+---------+---------+    60
CTTAAGTACAGAATGCCAGTTCCCTGTTGCCGGGGACCGTCGGCGAGCCGGGGGAGTCTC
          70                  90                  110
CCCTTTCTGTTTAACAGAGACCGCTCGTTGCCGATGGGTCCTGGTCTGTGCTCCTTGCTC
---------+---------+---------+---------+---------+---------+   120
GGGAAAGACAAATTGTCTCTGGCGAGCAACGGCTACCCAGGACCAGACACGAGGAACGAG
          130                 150                 170
TGTGCAGAGCCTTGTTCTGAGTCCTTGTGCTCCGCCTGCTTCTGAGCTTGGTCCTTGTGG
---------+---------+---------+---------+---------+---------+   180
ACACGTCTCGGAACAAGACTCAGGAACACGAGGCGGACGAAGACTCGAACCAGGAACACC
          190                 210                 230
CTGTCGGCCTTAGTATCCCGCCAGGCCAGGCCCGCTAATTTCCTCAGGGTTGAGAGTGCC
---------+---------+---------+---------+---------+---------+   240
GACAGCCGGAATCATAGGGCGGTCCGGTCCGGGCGATTAAAGGAGTCCCAACTCTCACGG
          250                 270                 290
ATATGGGGAACCTGCGCTGAGGAGGGCAGACTACTCCCGAGACTCAGCACTCAATCAACG
---------+---------+---------+---------+---------+---------+   300
TATACCCCTTGGACGCGACTCCTCCCGTCTGATGAGGGCTCTGAGTCGTGAGTTAGTTGC
          310                 330                 350
CCACTATCTCCCTGGCTGCTCTTAAACCCCTCTGAGGTGCCAGGCTCAGTACTGCTGAAG
---------+---------+---------+---------+---------+---------+   360
GGTGATAGAGGGACCGACGAGAATTTGGGGAGACTCCACGGTCCGAGTCATGACGACTTC
          370                 390                 410
GCAGTGACAGCTGTCCTCCATGAGTCACTGGCGGCGCCCTGCACAGGGAGGGCTTGGCTG
---------+---------+---------+---------+---------+---------+   420
CGTCACTGTCGACAGGAGGTACTCAGTGACCGCCGCGGGACGTGTCCCTCCCGAACCGAC
          430                 450                 470
CTGTTCTCCCAAGATCTCAGTGCCCAGCTGCTGCCAGGCACAATCACATCTTCTCAGACA
---------+---------+---------+---------+---------+---------+   480
GACAAGAGGGTTCTAGAGTCACGGGTCGACGACGGTCCGTGTTAGTGTAGAAGAGTCTGT
          490                 510                 530
CAGGCCGTTCTGCTTTTTAGCAAACCGAGGAGGAATCCGGGATGAAAGGCCCCGACCTTT
---------+---------+---------+---------+---------+---------+   540
GTCCGGCAAGACGAAAAATCGTTTGGCTCCTCCTTAGGCCCTACTTTCCGGGGCTGGAAA
          550                 570                 590
GACAGGCACCTGAGCAGCCTGTTCTTTTTTTCTTCTCTCCTCCTCCAGCAGGCGACGCTG
---------+---------+---------+---------+---------+---------+   600
CTGTCCGTGGACTCGTCGGACAAGAAAAAAAGAAGAGAGGAGGAGGTCGTCCGCTGCGAC
          610                 630                 650
CTTCTTGGTCAGGACTTCAATGAAGCCTTCACCCACCACAGAGCTTGCTTCATTCTCTTC
---------+---------+---------+---------+---------+---------+   660
GAAGAACCAGTCCTGAAGTTACTTCGGAAGTGGGTGGTGTCTCGAACGAAGTAAGAGAAG
```

Figure 5.1.B (continued)

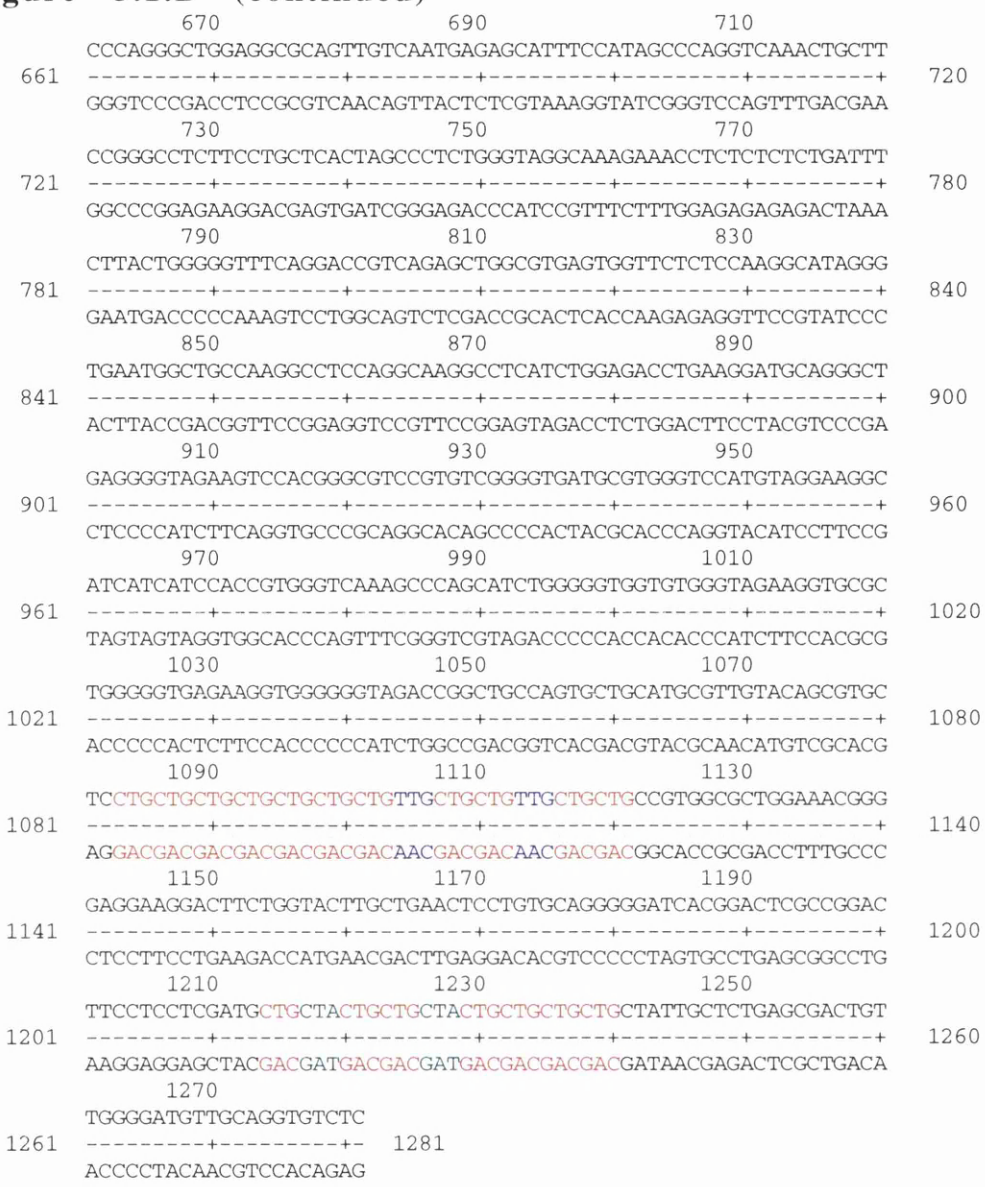


Figure 5.1 Full nucleotide sequence of Clone mCTG 26.

A) Diagrammatic representation of Clone mCTG 26. Yellow bar represents the whole clone, the CAG/CTG trinucleotide repeat represented in black and the numbers of repeats in red text. Blue bars represent the exo III deletion clones used for sequencing the remainder of Clone mCTG 26. The thin black bar represents the 500bp *Bgl*II- *Bam*HI fragment of Clone Δ26KH32 which was used as a probe for subsequent Southern and northern analysis. B) DNA sequence of this clone, derived by the end sequencing of clones derived from an exonuclease III deletion library of Clone mCTG 26 (indicated on Figure 5.1.A) The CAG/CTG trinucleotide repeats are indicated by red coloured sequence. Corruptions of the repeats are in blue (TTG) and green (CTA).

MacAlign package (IBI). The full nucleotide sequence is reproduced in Figure 5.1.B. Sequence derived from clone 26 was initially analysed, firstly for open reading frames using Fickett's method (Fickett, 1982) as part of the MacVector sequence package (IBI). Additional analysis was performed on the UWGCG Gene Analysis Package (Devereux *et al*, 1984) version 8.1. No open reading frame was detected by the application of Fickett's algorithm (Fickett, 1982).

5.2.2 Southern blot analysis with sequence in Clone mCTG 26.

A 500 bp *Bam*HI- *Bgl*II central fragment of a deletion library derived clone, Δ26KH32 (Figure 5.1) was used to probe mouse genomic DNA digested with the restriction enzymes *Eco*RI, *Hin*dIII and *Bam*HI. This yielded single fragments of sizes 14.5, 7.3 and 4.1 kbp respectively. This probe does not contain the trinucleotide which precludes cross reaction of repetitive sequences abundant in the genome. The results are displayed in Figure 5.2.

5.2.3 Further expression analysis of Clone mCTG 26.

In order to characterise the messenger RNA species that are derived from the mCTG 26 cDNA clone, northern blot analysis was performed on electrophoretically separated total RNA. Total RNA was derived from the previous time points used as templates for the first strand probes for the reverse dot-blot experiments

Figure 5.2

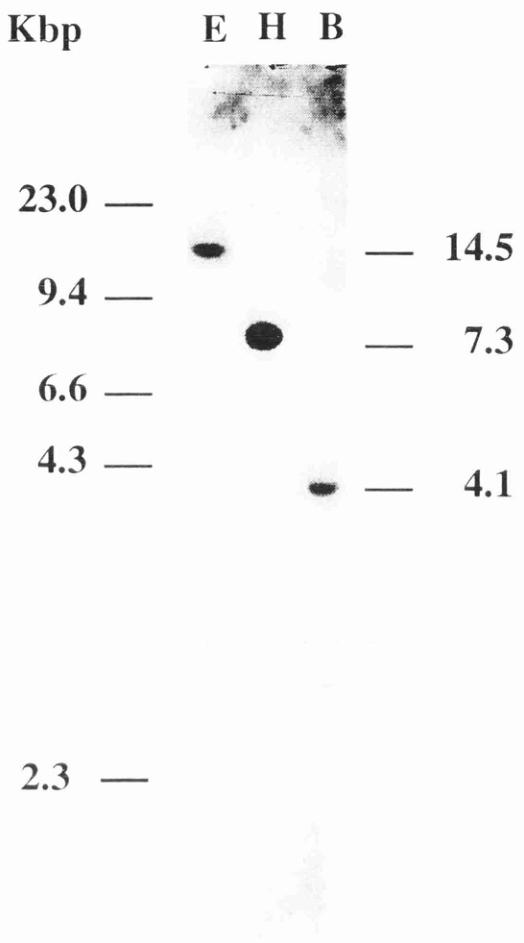


Figure 5.2 Southern analysis of Clone mCTG 26.
10µg of genomic DNA was digested with the restriction enzymes E (*Eco*RI), H (*Hin*dIII) and B (*Bam*HI). They were probed with a 500 bp *Bgl*II-*Bam*HI fragment of clone Δ26KH3.2 which does not contain a trinucleotide repeat.

(Chapter 3). The 500 bp *Bam*HI-*Bgl*II internal fragment was used as a template to derive a random primed, radio-actively labelled probe. This yielded results which were, in the most part, consistent with those of the reverse dot-blot experiment. However there appears to be an absence of expression in the adult liver. The length of the mRNA(s) corresponding to the broad band detected by the 500 bp *Bam*HI-*Bgl*II derived probe was measured to be between 9 to 12 kb (Figure 5.3).

5.3 Further analysis of Clone mCTG 210.

This was a clone derived from the 8.5 dpc whole mouse embryo cDNA library (Hogan *et al*, unpublished). Sequence analysis had revealed that it contained a cryptic CAG/CTG trinucleotide. It had no detectable expression in the reverse dot-blot experiments.

5.3.1 Sequence analysis of Clone mCTG 210.

Fragments of sequence data from both automated sequence (Perkin Elmer) and manual sequence (derived from using T7 polymerase kit, Pharmacia) were assembled together using the MacAlign programme (IBI). A contiguous sequence was derived and exported to the sequence analysis package MacVector (IBI) for further analysis. The sequence of clone mCTG 210 is described in Figure 5.4. The clone was found to be 906 nt in length by sequencing. No open reading frame was identified from application of Fickett's Testcode Algorithm (Fickett, 1982). No sequences were identified when the sequence of clone mCTG 210 was used to search nucleic acid databases with the FASTA algorithm.

Figure 5.3

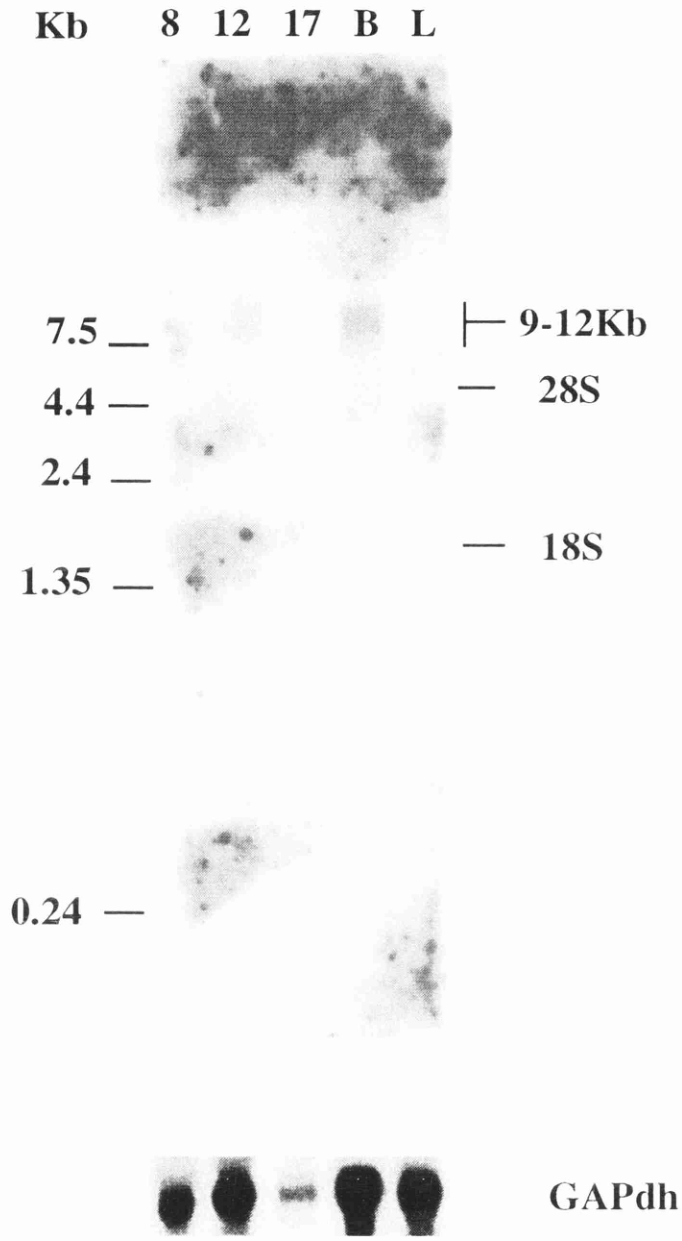


Figure 5.3 Northern blot of Clone mCTG 26.
Approximately 10 µg of total RNA from **8**) 8.5 dpc whole embryos, **12**) 12.5 dpc whole embryos **17**) 17.5 whole embryos, **B**) mouse brain, adult and **L**) adult mouse liver were probed with a 500 bp *Bgl*II-*Bam*HI radioactively labelled fragment of clone Δ26KH3.2 which does not contain a repeat. Measurements are in kilobases (kb). A fragment of a mouse GAPdh cDNA was used as a loading control. 18S and 28S indicate the major ribosomal RNA bands.

5.3.2 Southern analysis of Clone mCTG 210.

A 400 bp *Sma*I-*Eco*RI fragment of the original clone 210 was used to probe mouse embryonic stem cell genomic DNA under stringent hybridisation and wash conditions (detailed in Chapter 2, Section 2.22). This yielded a single band for *Eco*RI (7.3Kb) and *Bam*HI (3.4Kb) respectively (See Figure 5.5.B).

This experiment was repeated, but the sequence which represent the remainder of clone 210 was used as a probe (a *Sma*I-*Eco*RI fragment 550 bp in size). This fragment gave a multiple banding pattern (Figure 5.5.A). This probe contains the region known to contain the trinucleotide repeat (as described in sequence analysis, Chapter 4, this work).

5.4 Further analysis of Clone mCTG 61.

This is a clone derived from the 12.5 dpc whole mouse embryo cDNA library. From initial sequencing data, it was found to contain a large uninterrupted trinucleotide repeat, 25 repeat units in size (Chapter 4, this work). It had no detectable expression levels in the reverse dot-blot experiments. A northern blot experiment was conducted but showed no detectable expression either (data not shown). Southern analysis was also performed to determine the genomic organisation of sequences represented in clone 61. Further sequencing was conducted to fully sequence the clone, in order to discover any patterns that may help in further characterisation of this gene.

Figure 5.4

A)

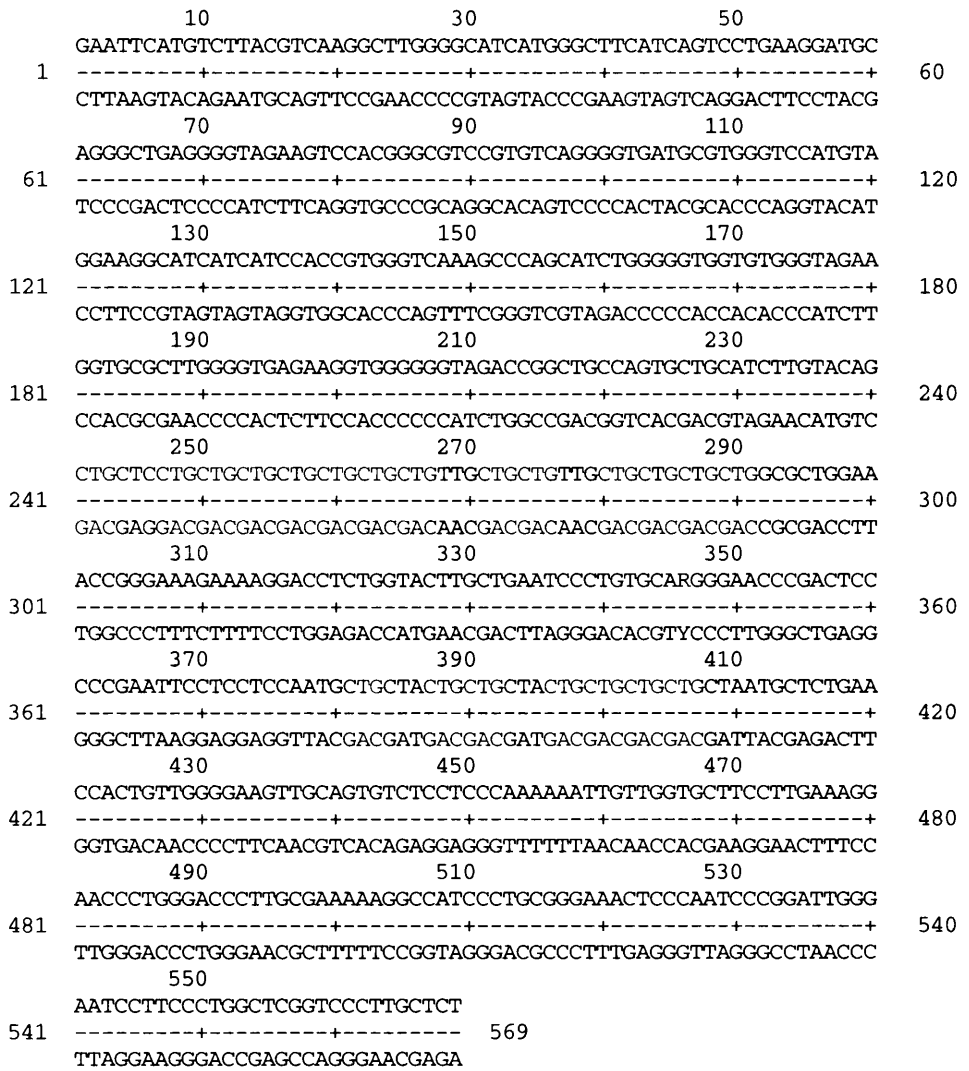

10 30 50
GAATTCATGTCTTACGTCAAGGCTTGGGGCATCATGGGCTTCATCAGTCCTGAAGGATGC
1 60
CTTAAGTACAGAATGCAGTTCCGAACCCCGTAGTACCCGAAGTAGTCAGGACTTCCTACG
70 90 110
AGGGCTGAGGGGTAGAAGTCCACGGGCGTCCGTGTCAGGGGTGATGCGTGGGTCCATGTA
61 120
TCCCGACTCCCCATCTTCAGGTGCCCGCAGGCACAGTCCCCACTACGCACCCAGGTACAT
130 150 170
GGAAGGCATCATCATCCACCGTGGGTCAAAGCCCAGCATCTGGGGGTGGTGTGGGTAGAA
121 180
CCTTCCGTAGTAGTAGGTGGCACCCAGTTTCGGGTCGTAGACCCCCACCACACCCATCTT
190 210 230
GGTGCGCTTGGGGTGAGAAGGTGGGGGGTAGACCGGCTGCCAGTGCTGCATCTTGTACAG
181 240
CCACGCGAACCCCACTCTTCCACCCCCCATCTGGCCGACGGTCACGACGTAGAACATGTC
250 270 290
CTGCTCCTGCTGCTGCTGCTGCTGCTGTTGCTGCTGTTGCTGCTGCTGCTGGCGCTGGAA
241 300
GACGAGGACGACGACGACGACGACGACAACGACGACAACGACGACGACGACCGCGACCTT
310 330 350
ACCGGGAAAGAAAAGGACCTCTGGTACTTGCTGAATCCCTGTGCARGGGAACCCGACTCC
301 360
TGGCCCTTTCTTTTCCTGGAGACCATGAACGACTTAGGGACACGTYCCCTTGGGCTGAGG
370 390 410
CCCGAATTCCTCCTCCAATGCTGCTACTGCTGCTACTGCTGCTGCTGCTAATGCTCTGAA
361 420
GGGCTTAAGGAGGAGGTTACGACGATGACGACGATGACGACGACGACGATTACGAGACTT
430 450 470
CCACTGTTGGGGAAGTTGCAGTGTCTCCTCCCAAAAAATTGTTGGTGCTTCCTTGAAAGG
421 480
GGTGACAACCCCTTCAACGTCACAGAGGAGGGTTTTTTAACAACCACGAAGGAACTTTCC
490 510 530
AACCCTGGGACCCTTGCGAAAAAGGCCATCCCTGCGGGAAACTCCCAATCCCGGATTGGG
481 540
TTGGGACCCTGGGAACGCTTTTTCCGGTAGGGACGCCCTTTGAGGGTTAGGGCCTAACCC
550
AATCCTTCCCTGGCTCGGTCCCTTGCTCT
541 569
TTAGGAAGGGACCGAGCCAGGGAACGAGA

Figure 5.4.B

```
          10                  30                  50
      GAATTCATGTCTTACGGTCAAGGGTGGATCCCGTGTCATCCGAGAGGTGCCAGAGCCTCA
    1 ---------+---------+---------+---------+---------+---------+   60
      CTTAAGTACAGAATGCCAGTTCCCACCTAGGGCACAGTAGGCTCTCCACGGTCTCGGAGT
          70                  90                  110
      GCCTCCTTCCAGGAAGCTTCACAGTTGGGCCTCTGGGCCTGACTACCAGAAACCCACAAT
   61 ---------+---------+---------+---------+---------+---------+  120
      CGGAGGAAGGTCCTTCGAAGTGTCAACCCGGAGACCCGGACTGATGGTCTTTGGGTGTTA
          130                 150                 170
      GGGCAGCATGTTCCGCCAGCACTCCGCTGAGGACAAGGAGGACAAGCCACCCCCACGGCA
  121 ---------+---------+---------+---------+---------+---------+  180
      CCCGTCGTACAAGGCGGTCGTGAGGCGACTCCTGTTCCTCCTGTTCGGTGGGGGTGCCGT
          190                 210                 230
      GAAGTTCATCCAGTCAGAGATGTCCGAGGCCGTGGAACGAGCCCGAAAGCGCAGGGAAGA
  181 ---------+---------+---------+---------+---------+---------+  240
      CTTCAAGTAGGTCAGTCTCTACAGGCTCCGGCACCTTGCTCGGGCTTTCGCGTCCCTTCT
          250                 270                 290
      AGAACAGCGCCGGGCCCAARAGGAGAGCTGGCTGCCCSTGCTGCCAAACTCAAGCAGCTA
  241 ---------+---------+---------+---------+---------+---------+  300
      TCTTGTCGCGGCCCGGGTTYTCCTCTCGACCGACGGGSACGACGGTTTGAGTTCGTCGAT
          310                 330                 350
      AACCATAAGTGCCGGCAGGCTCAGAATGCCAATGAAACCCCGAAGCCAGTGGAAAAGGAG
  301 ---------+---------+---------+---------+---------+---------+  360
      TTGGTATTCACGGCCGTCCGAGTCTTACGGTTACTTTGGGGCTTCGGTCACCTTTTCCTC
          370                 390                 410
      TCCCCGATCCCGGGCATTGAGAAGTCTCCCGCCCAGAGAATGGTGCTGTTGTCCGCAGAG
  361 ---------+---------+---------+---------+---------+---------+  420
      AGGGGCTAGGGCCCGTAACTCTTCAGAGGGCGGGTCTCTTACCACGACAACAGGCGTCTC
          430                 450                 470
      GTCCCCTAATTCCCTGTTCAAGAAACACCTCTTGTTCTTGGAAAAAAACTGCAACATCCC
  421 ---------+---------+---------+---------+---------+---------+  480
      CAGGGGATTAAGGGACAAGTTCTTTGTGGAGAACAAGAACCTTTTTTTGACGTTGTAGGG

      C
  481 -  481
      G
```

C)

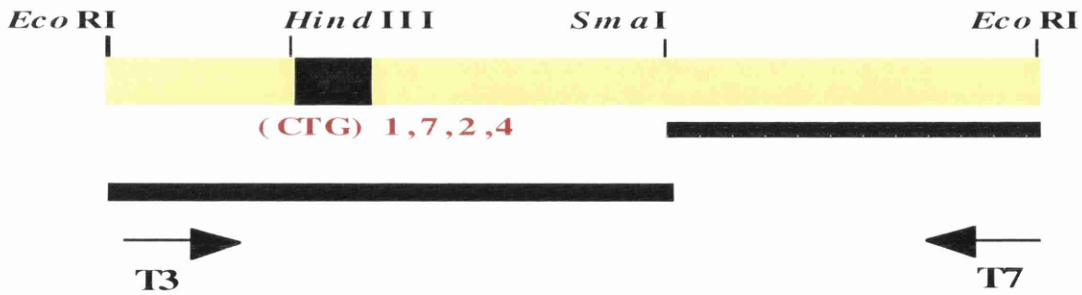


Figure 5.4 Sequence of Clone mCTG 210: A) T3 primed sequence, B) T7 primed sequence and C) diagrammatic representation.

Sequence was derived from priming DNA sequencing reactions using derivatives of the T3 and T7 primers which prime sequence from the multiple cloning site of pBluescript KS-. The CAG/CTG trinucleotide repeats found in this clone are indicated in red sequence. No overlapping sequence was detected. Black bars indicate the 400 and 500 bp *Sma*I-*Eco*RI fragments of Clone mCTG 210 which were used as probes in the Southern experiments.

Figure 5.5

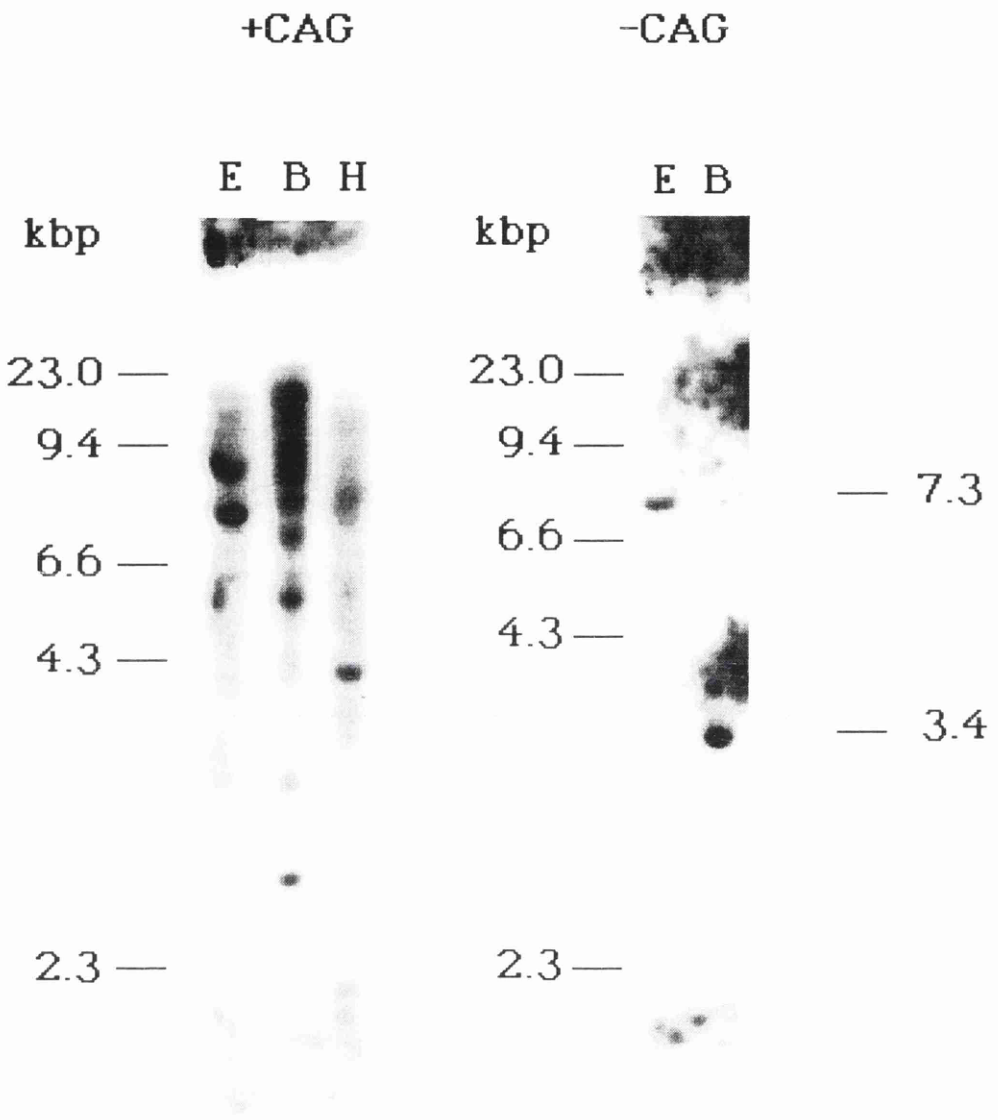


Figure 5.5 Southern analysis of Clone mCTG 210
Ten micrograms of mouse genomic DNA was digested with the restriction enzymes: *Eco*RI (E), *Bam*HI (B) and *Hin*dIII (H) and probed with a 550 bp *Sma*I-*Eco*RI fragment (+CAG) and a 400 bp *Sma*I-*Eco*RI fragment (-CAG).

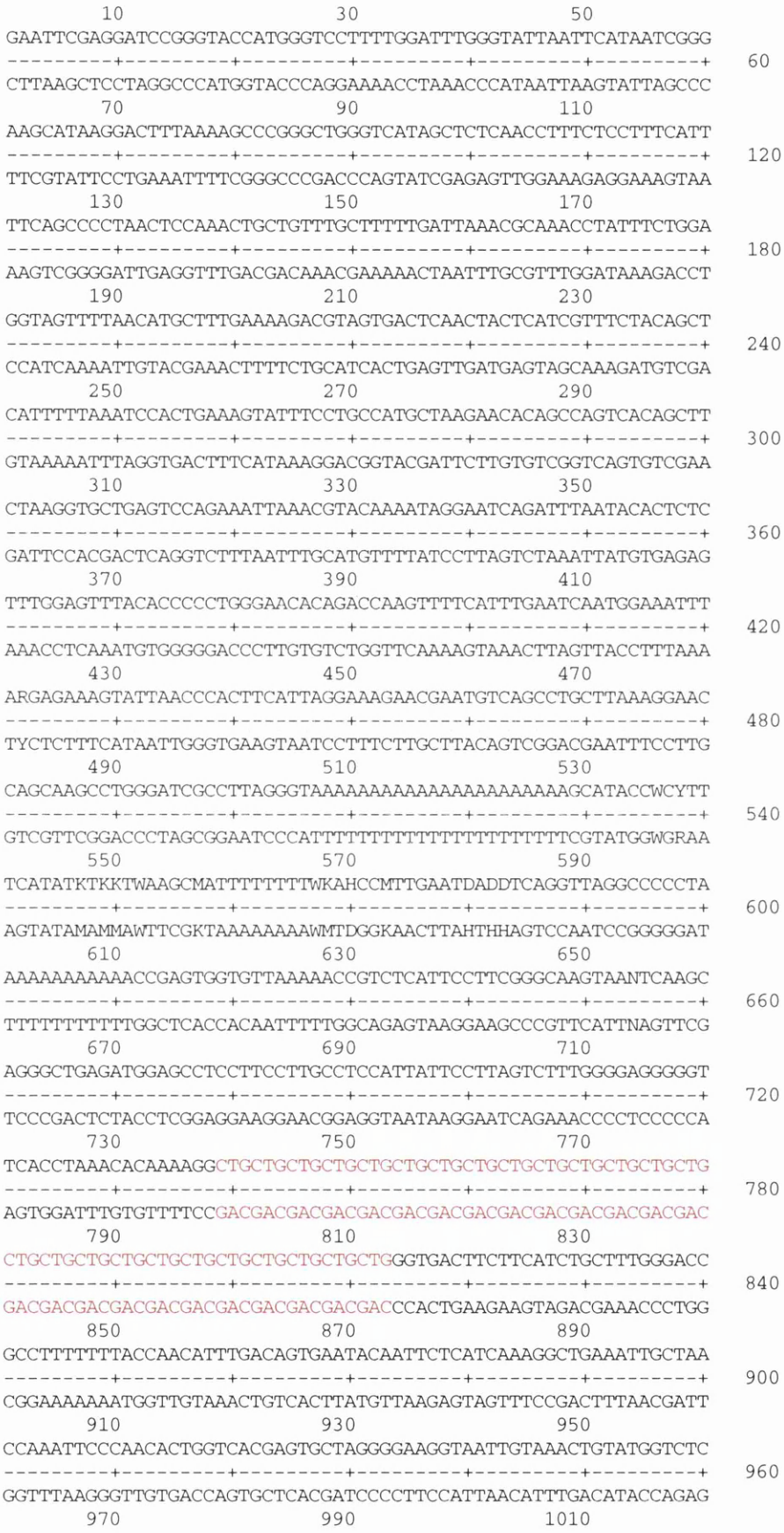
5.4.1 Sequence analysis of Clone mCTG 61.

Manual sequencing did not reveal the whole sequence of this clone. Automated sequencing was performed using the ABI system (Perkin Elmer) to extend and complete the sequence from this clone. The cumulative data for clone 61 was first converted into MacVector (IBI) format sequence files for import and assembly into continuous sequence by the MacAlign sequence assembly software package (IBI). The contiguous sequence was re-exported to MacVector for further analysis. No open reading frame was found using Fickett's algorithm (1982). The sequence is represented in Figure 5.6.

5.4.2 Genomic organisation of Clone mCTG 61.

Southern blot analysis was conducted using a random primed probe, generated from a 650 bp *Hin*dII-*Hin*dII fragment of the original sub-clone of mCTG 61 (Figure 5.6.A). Sequence analysis had determined that the large trinucleotide repeat was not contained within this fragment. Radio-labelled probe from this *Hin*dII fragment identified single bands under stringent hybridisation conditions to restriction enzyme digested mouse genomic DNA, which was size-separated by agarose gel electrophoresis. The individual fragment lengths identified were; 4.2 kb for *Eco*RI, 9.7 kb for *Bam*HI and 3.8 kb for the *Hin*dIII digested genomic DNA (Figure 5.7).

Figure 5.6.A


10 30 50
GAATTCGAGGATCCGGGTACCATGGGTCCTTTTGGATTTGGGTATTAATTCATAATCGGG
1 60
CTTAAGCTCCTAGGCCCATGGTACCCAGGAAAACCTAAACCCATAATTAAGTATTAGCCC
70 90 110
AAGCATAAGGACTTTAAAAGCCCGGGCTGGGTCATAGCTCTCAACCTTTCTCCTTTCATT
61 120
TTCGTATTCCTGAAATTTTCGGGCCCGACCCAGTATCGAGAGTTGGAAAGAGGAAAGTAA
130 150 170
TTCAGCCCCTAACTCCAAACTGCTGTTTGCTTTTTGATTAAACGCAAACCTATTTCTGGA
121 180
AAGTCGGGGATTGAGGTTTGACGACAAACGAAAAACTAATTTGCGTTTGGATAAAGACCT
190 210 230
GGTAGTTTTAACATGCTTTGAAAAGACGTAGTGACTCAACTACTCATCGTTTCTACAGCT
181 240
CCATCAAAATTGTACGAAACTTTTCTGCATCACTGAGTTGATGAGTAGCAAAGATGTCGA
250 270 290
CATTTTTAAATCCACTGAAAGTATTTCCTGCCATGCTAAGAACACAGCCAGTCACAGCTT
241 300
GTAAAAATTTAGGTGACTTTCATAAAGGACGGTACGATTCTTGTGTCGGTCAGTGTCGAA
310 330 350
CTAAGGTGCTGAGTCCAGAAATTAAACGTACAAAATAGGAATCAGATTTAATACACTCTC
301 360
GATTCCACGACTCAGGTCTTTAATTTGCATGTTTTATCCTTAGTCTAAATTATGTGAGAG
370 390 410
TTTGGAGTTTACACCCCCTGGGAACACAGACCAAGTTTTCATTTGAATCAATGGAAATTT
361 420
AAACCTCAAATGTGGGGGACCCTTGTGTCTGGTTCAAAAGTAAACTTAGTTACCTTTAAA
430 450 470
ARGAGAAAGTATTAACCCACTTCATTAGGAAAGAACGAATGTCAGCCTGCTTAAAGGAAC
421 480
TYCTCTTTCATAATTGGGTGAAGTAATCCTTTCTTGCTTACAGTCGGACGAATTTCCTTG
490 510 530
CAGCAAGCCTGGGATCGCCTTAGGGTAAAAAAAAAAAAAAAAAAAAAAGCATACCWCYTT
481 540
GTCGTTCGGACCCTAGCGGAATCCCATTTTTTTTTTTTTTTTTTTTCGTATGGWGRAA
550 570 590
TCATATKTKKTWAAGCMATTTTTTTTWKAHCCMTTGAATDADDTCAGGTTAGGCCCCCTA
541 600
AGTATAMAMMAWTTCGKTAAAAAAAAWMTDGGKAACTTAHTHHAGTCCAATCCGGGGGAT
610 630 650
AAAAAAAAAAACCGAGTGGTGTTAAAAACCGTCTCATTCCTTCGGGCAAGTAANTCAAGC
601 660
TTTTTTTTTTTGGCTCACCACAATTTTTGGCAGAGTAAGGAAGCCCGTTCATTNAGTTCG
670 690 710
AGGGCTGAGATGGAGCCTCCTTCCTTGCCTCCATTATTCCTTAGTCTTTGGGGAGGGGGT
661 720
TCCCGACTCTACCTCGGAGGAAGGAACGGAGGTAATAAGGAATCAGAAACCCCTCCCCCA
730 750 770
TCACCTAAACACAAAAGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
721 780
AGTGGATTTGTGTTTTCCGACGACGACGACGACGACGACGACGACGACGACGACGACGAC
790 810 830
CTGCTGCTGCTGCTGCTGCTGCTGCTGCTGGGTGACTTCTTCATCTGCTTTGGGACC
781 840
GACGACGACGACGACGACGACGACGACGACCCACTGAAGAAGTAGACGAAACCCTGG
850 870 890
GCCTTTTTTTTACCAACATTTGACAGTGAATACAATTCTCATCAAAGGCTGAAATTGCTAA
841 900
CGGAAAAAAAATGGTTGTAAACTGTCACTTATGTTAAGAGTAGTTTCCGACTTTAACGATT
910 930 950
CCAAATTCCCAACACTGGTCACGAGTGCTAGGGGAAGGTAATTGTAAACTGTATGGTCTC
901 960
GGTTTAAGGGTTGTGACCAGTGCTCACGATCCCCTTCCATTAACATTTGACATACCAGAG
970 990 1010

B)

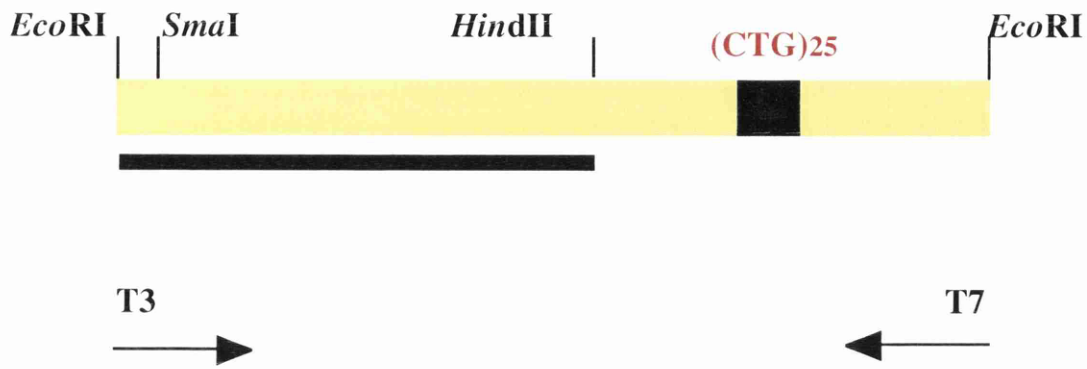


Figure 5.6 Full nucleotide sequence of Clone mCTG 61.
A) The full DNA sequence of Clone mCTG 61. This was derived from an assembly (using the MacAlign programme, IBI) of Clone mCTG 61 derived sequences generated by both manual T7 polymerase sequencing (Pharmacia) and fluorescent-labelling DNA cycle sequencing (Applied Biosystems). The trinucleotide repeat is indicated as red coloured sequence. B) Diagrammatic representation of Clone mCTG 61; the black bar represents sequences used as a probe for the northern and Southern blot experiments (a 650 bp *Hin*dII-*Hin*dII). T3 and T7 indicate the orientation of the sequences represented relative to the multiple cloning site of pBluescript KS-. The trinucleotide in this clone is indicated as a black bar within the yellow section, with the repeat number indicated in red.

5.5 Further analysis of Clone mCTG 63.

Clone mCTG 63 also derives from the 12.5 dpc whole mouse embryo cDNA library. This clone is of primary interest because it contains the longest uninterrupted CAG/CTG repeat identified in this study, extending to 26 trinucleotides. This sequence was also found to have high expression during development and in the adult tissues tested. Reverse slot blot experiments (see Chapter 3) showed that, although expression levels were high they were also variable, compared to other highly expressed sequences.

5.5.1 Sequence analysis of Clone mCTG 63.

Manual sequencing yielded reliable sequence from only one end of this clone. This sequence was found to have similarity to a group of human 3' ESTs derived from independent mass sequencing projects based on cDNA libraries from different tissues (see Figure 5.8.A and B). This is in agreement with the wide-spread expression of the mouse clone during development and in adult tissues as observed from the reverse dot-blot experiments described in Chapter 3. The regions of overlap were found to be similar in all the matches and therefore these probably represent sequence from the same gene. The result are shown as a pileup programme output in Figure 5.8.B. The clone itself had a large poly-adenosine stretch (46nt) with a strong poly-adenylation signal upstream of it (see Figure 5.8.A). Further sequencing showed that it contained the largest repeat found so far in this work (26 repeats) and that this

Figure 5.7

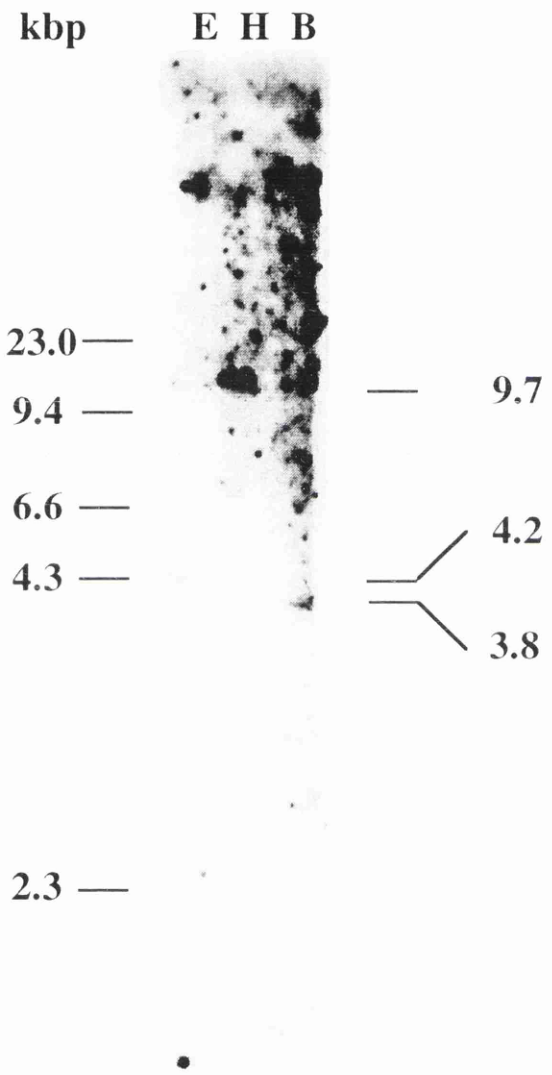


Figure 5.7 Southern analysis of Clone mCTG 61.
Mouse genomic DNA was restricted with three enzymes: *Eco*RI (E), *Bam*HI (B) and *Hin*dIII (H). The digested DNAs were probed with sequence derived from a 650 bp *Hin*dII-*Hin*dII of clone mCTG 61. This fragment does not contain CAG/CTG repeat sequences.

was CTG in the coding strand in the putative 3' end of the transcript. This perhaps could be taken as an analogy for the repeat in the myotonic dystrophy protein kinase (Brook *et al*, 1992), where a large CTG repeat exists in the 3' untranslated region.

Automated sequencing provided sequence data from the opposite end of the clone. Sequence from this end was found to have identity to a human clone JC310, isolated by J. Colicelli and M. Wigler (Colicelli *et al*, 1991) who identified JC310 as a functional inhibitor of the RAS protein. Alignments show that the sequence of clone mCTG 63 shows a high degree of conservation with JC310 at the nucleotide level (Figure 5.8.C). This is also true at the protein level (see Figure 5.8.D). Respectively, the percentages of similarity are 87.5% and 92%.

5.5.2 Genomic organisation of Clone mCTG 63.

A 1.5kb *Hin*dIII-*Eco*RI fragment of the original mCTG 63 clone was used to probe mouse genomic DNA derived from an ES cell line. This sequence represents the probable 3' end of the gene, as it contains a poly-A tail as well as a high quality poly-A signal consensus sequence upstream of the poly-A tail. This yielded three single bands, one for each restricted DNA (with single enzymes *Eco*RI, *Bam*HI and *Hin*dIII). The detected restriction fragments were of sizes 5.2 kb, 3 kb and 2.4 kb respectively (as shown in Figure 5.9). This 1.5 kb *Hin*dIII-*Eco*RI fragment contains the trinucleotide repeat and a poly-A tail. A combination of the trinucleotide repeat and a mononucleotide stretch results in a repetitive banding pattern seen in exposures longer than one day.

Figure 5.8.A

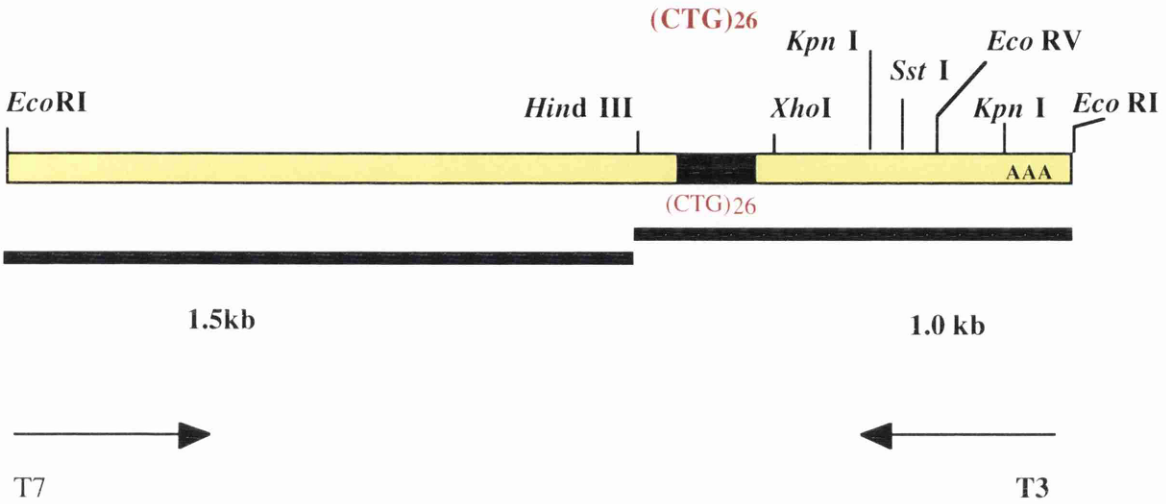


Figure 5.8.A Diagrammatic representation of Clone mCTG 63.

Clone mCTG 63 sequences are represented by a yellow bar. The regions of homology to the human RAS protein inhibitor is indicated by the blue bar and the sequence that shows similarity to various human and mouse ESTs is represented by a thick green bar. The black bars show the fragments which were used to derive radioactively labelled probes for Southern and northern analysis (*i.e.* 1.0 and 1.5 kbp *Eco*RI-*Hin*dIII fragments). The position of the CAG/CTG trinucleotide is indicated by the black bar inset into the yellow bar and the length of the repeat is indicated in red text. T3 and T7 indicate the orientation of the insert relative to the multiple cloning site of pBluescript KS-.

Figure 5.8.B


1 50
63 ~~~~~~~~~~ ~~~~~~~~~~ ~~~~CTGAAA ATAAAATAAT AAAATTATAA
n79067 ~~~~~~~~CA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
r40976 AAAAAAAAAA AATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
t16564 ~~~~~~~~CA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
z41184 ~~~~~~~~CA CATAAATATA CT.GAAANTA AAATTATAGT AAAATTATAG
aa199815 ~~~~~~~~~~ CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
w93359 ~~~~~~~~~~ CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
n98931 ~~~~AAAAAA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
w81207 ~~~~~~~ACA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
h25927 ~~~~~~~~~~ ~~~~~TATAN NT.NCNAATA AAATTATAGT AAAATTATAG
h27668 ~~~~~~~~~~ ~~~~~~TATN AN.ANAAATA AAATTATAGT AAAATTATAG
n65975 ~~~~~~~~~~ ~~~~~~TATA CT.GAAAATA AAATTATAGT AAAATTATAG
h28318 ~~~~~~AAAA CATAAATATA ATGNNNAATA AATTTATAGT AAAATTATAG
h05633 AAAAAAAAAA CATAAATATA CT.GAAAATA AANTTATAGT AAAATTATAG
r44135 ~AAAAAAAAA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
h24291 AAAAAAAAAA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
n62312 ~~~~~~~~CA CATAAATATA CT.GAAAATA AAATTATAGT AAAATTATAG
51 100
63 CTTAWACACT GTTTTAAATT ACTGGTTTAA TAAGTAAATG TTAAAGTTTT
n79067 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
r40976 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
t16564 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
z41184 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
aa199815 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
w93359 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
n98931 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
w81207 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
h25927 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
h27668 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
n65975 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
h28318 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
h05633 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
r44135 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
h24291 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
n62312 CTTATACAGT GTTCTAAATT ACTGGTTTAA TCAGTAAATG TAAAAAGTTT
101 150
63 TATAGATTTA TATTTTAACG ATTTTTAAGT TTATG..TGT CGTTCGATGT
n79067 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
r40976 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
t16564 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
z41184 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
aa199815 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
w93359 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
n98931 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
w81207 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
h25927 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
h27668 CATAGATTTA TATTTTNNCG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
n65975 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
h28318 CATAGATTTA TATTTTNACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
h05633 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
r44135 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
h24291 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT
n62312 CATAGATTTA TATTTTAACG ATTTTTAAGT TTATGTCTAT CGTTCGATGT

Figure 5.8.B (continued)

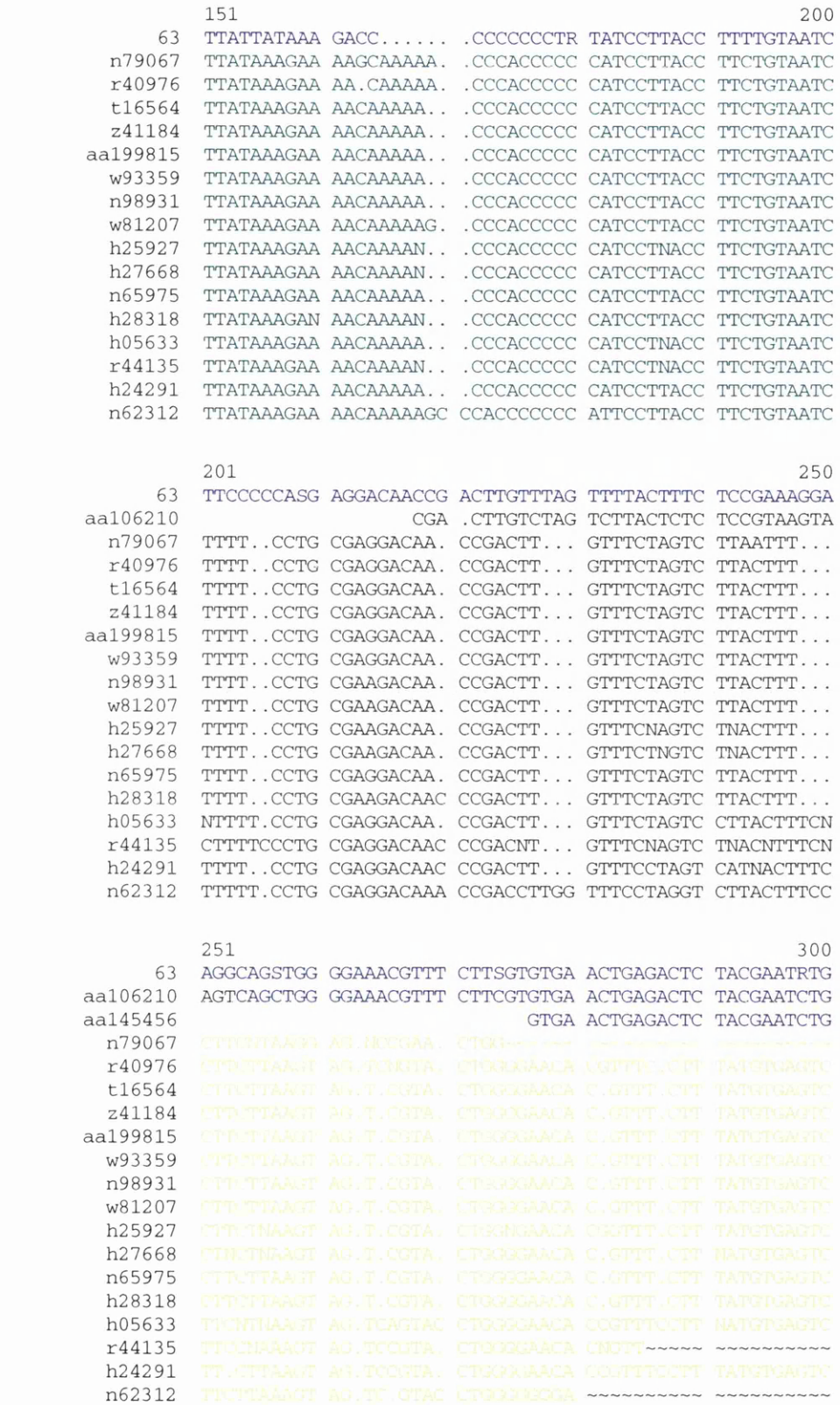


Figure 5.8.B (continued)

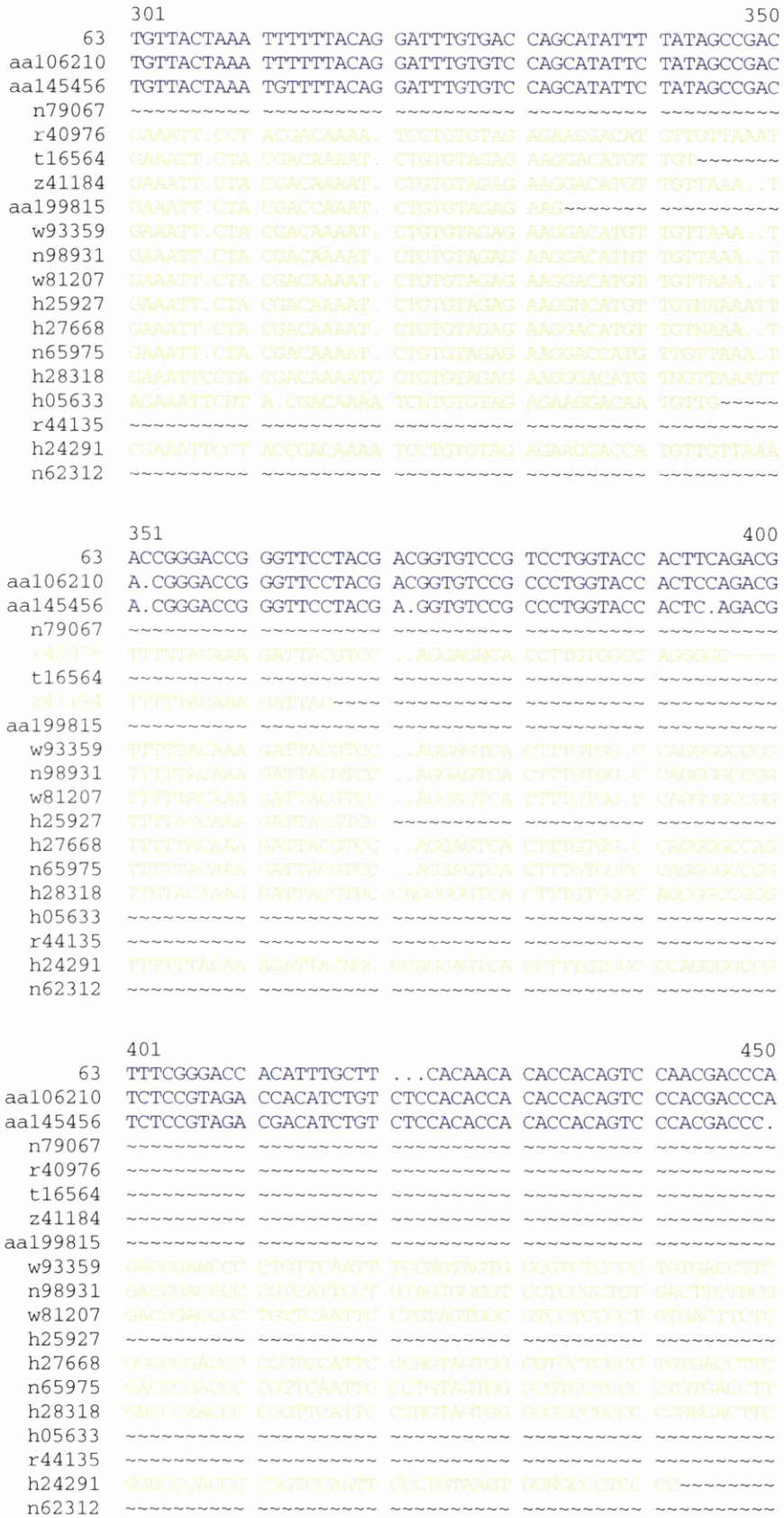


Figure 5.8.B (continued)

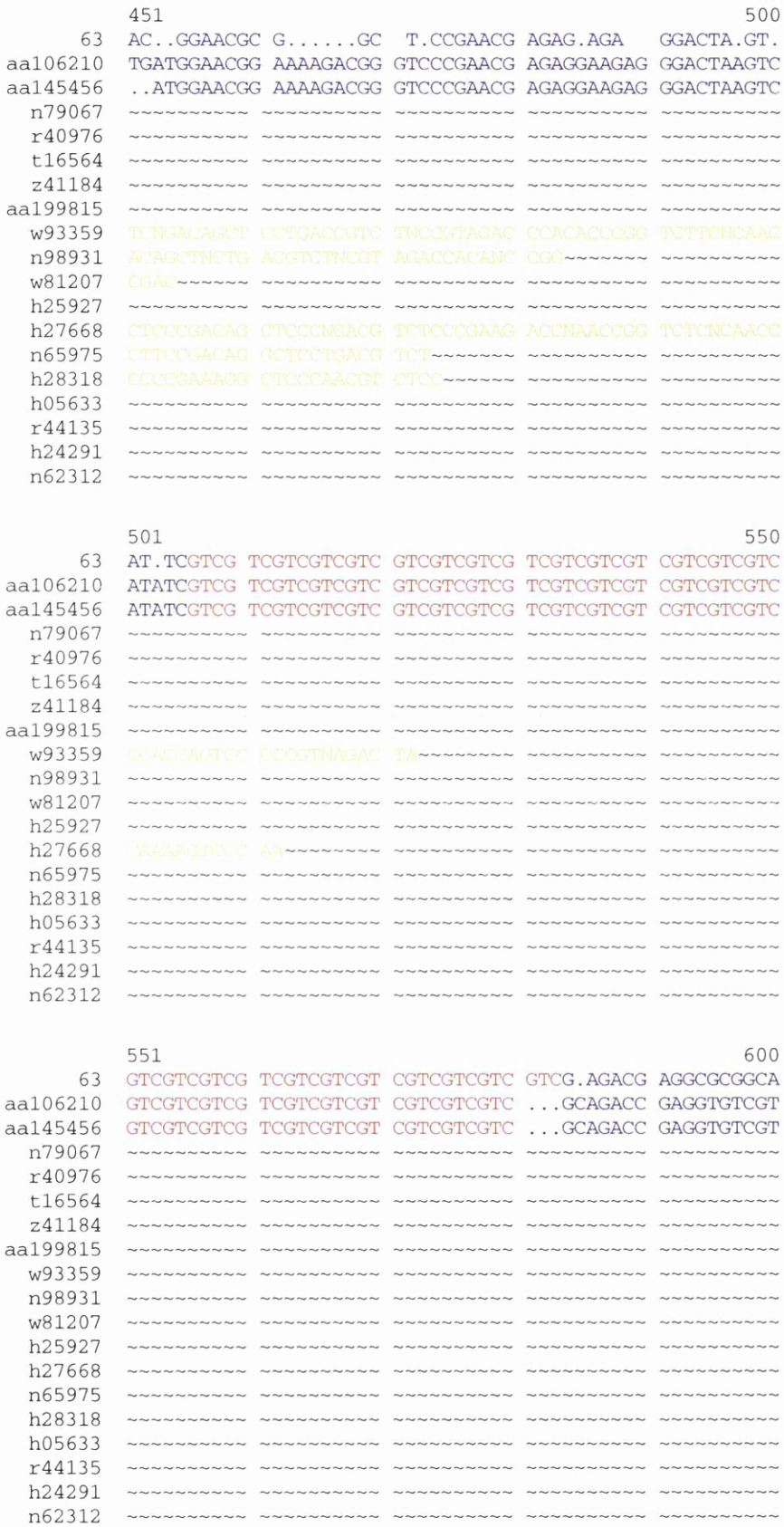


Figure 5.8.B (continued)

```
          601
       63 TTCTACTTAC AGTTCGCGTG AAG
aa106210 CCGTGGCCAT TCTCCTAACC TCTATCCAGT CTCTCCGTC
aa145456 CCGTGGCCAT TCTCCTAACC TCTATCCAGT CTCTCG
  n79067 ~~~~~~~~~~
  r40976 ~~~~~~~~~~
  t16564 ~~~~~~~~~~
  z41184 ~~~~~~~~~~
aa199815 ~~~~~~~~~~
  w93359 ~~~~~~~~~~
  n98931 ~~~~~~~~~~
  w81207 ~~~~~~~~~~
  h25927 ~~~~~~~~~~
  h27668 ~~~~~~~~~~
  n65975 ~~~~~~~~~~
  h28318 ~~~~~~~~~~
  h05633 ~~~~~~~~~~
  r44135 ~~~~~~~~~~
  h24291 ~~~~~~~~~~
  n62312 ~~~~~~~~~~
```

Figure 5.8.B Alignment of EST sequences which show similarity to Clone mCTG 63.

Two sequences (aa106210 and aa145456) are derived from mouse cDNA clones. The rest of the sequences represented here are human in origin and are derived from cDNAs. The trinucleotide repeat is indicated in red text sequence. Note that both mouse ESTs contain a CAG/CTG trinucleotide repeat. The human EST sequences are indicated in two colours to show the sequence which shows similarity to clone mCTG 63 sequences (*i.e.* green coloured text) and those which do not (yellow sequence). Mouse EST sequence which shows similarity to mCTG 63 sequence is indicated in blue text.

Figure 5.8.C

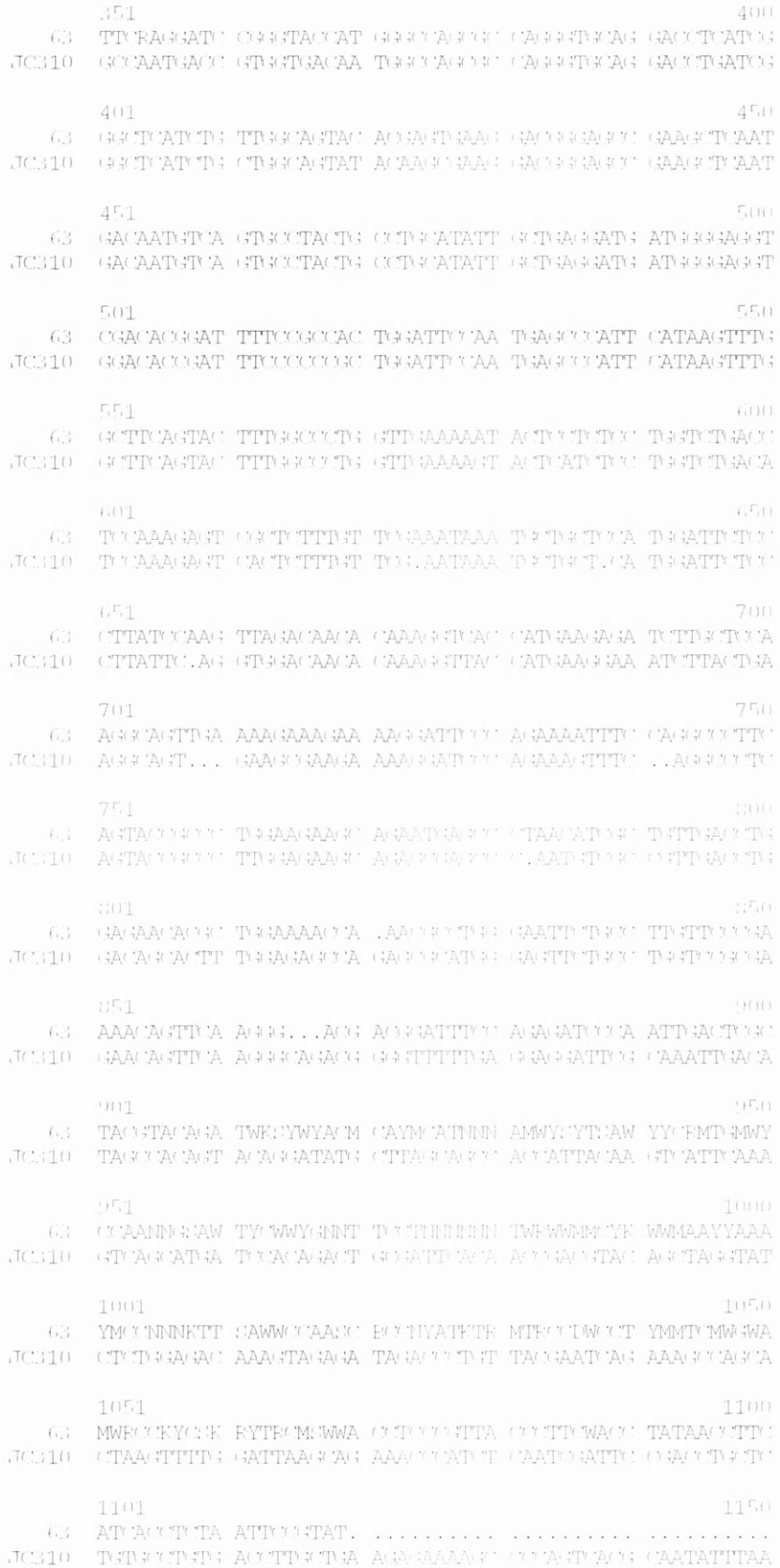


Figure 5.8.D

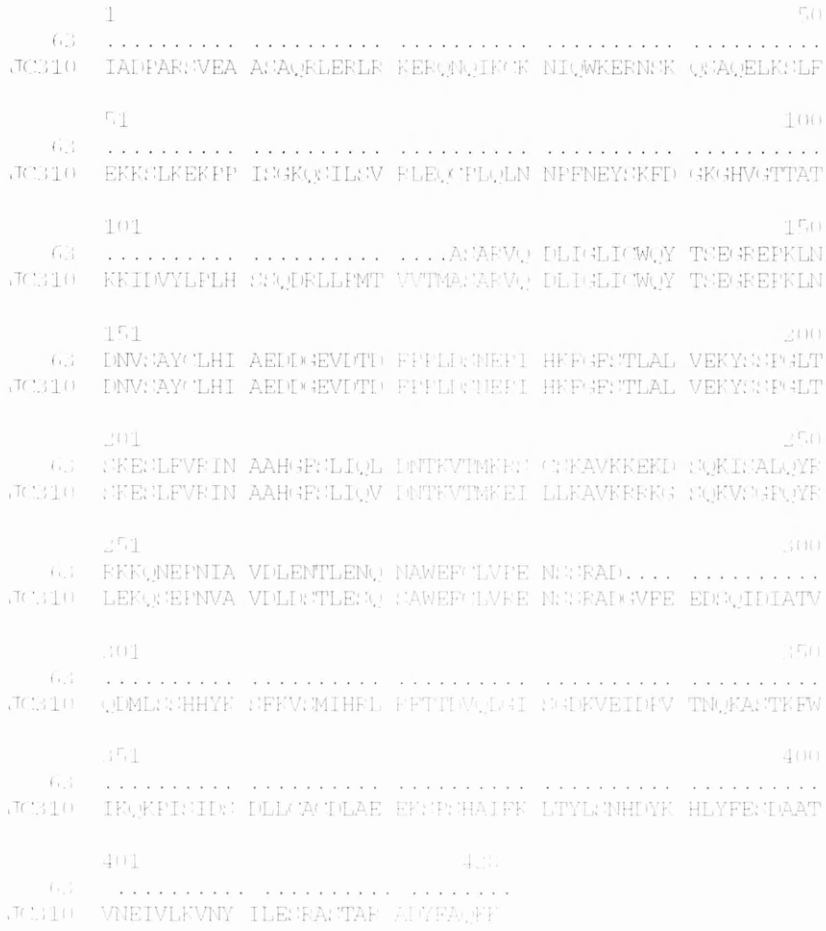


Figure 5.8.C and D Alignment of the RAS inhibiting protein, JC310 and sequences derived from Clone mCTG 63.

The sequences represented here are derived from T3-primed DNA sequences reactions performed on Clone mCTG63 (63) and sequences derived from the RAS protein inhibitor, JC310 (Colicelli *et al*, 1991). C) represents the nucleotide sequence alignment and D) represents translations of these sequences.

Figure 5.9

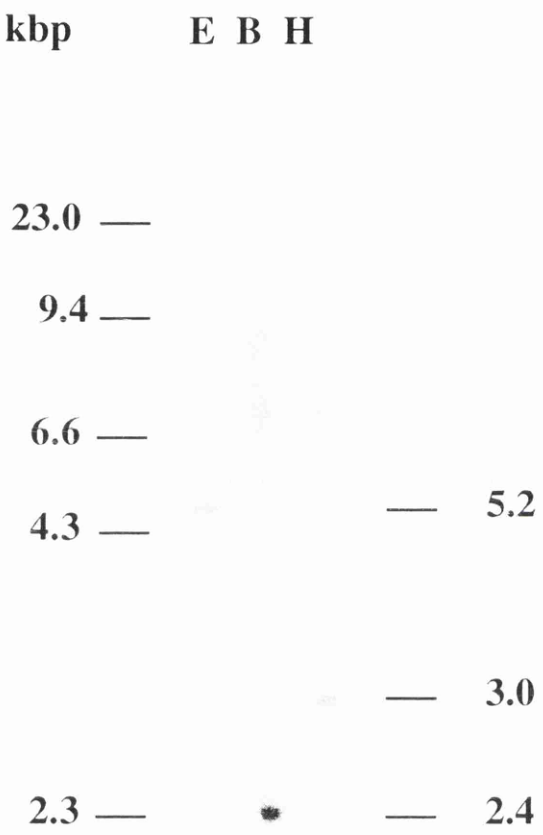


Figure 5.9 Southern analysis of clone mCTG 63.
The 1.5 kbp *Hin*dIII-*Eco*RI fragment was used as a probe to detect mCTG 63 specific bands in *Eco*RI (E), *Bam*HI (B) and *Hin*dIII (H) digested mouse genomic DNA.

Figure 5.10

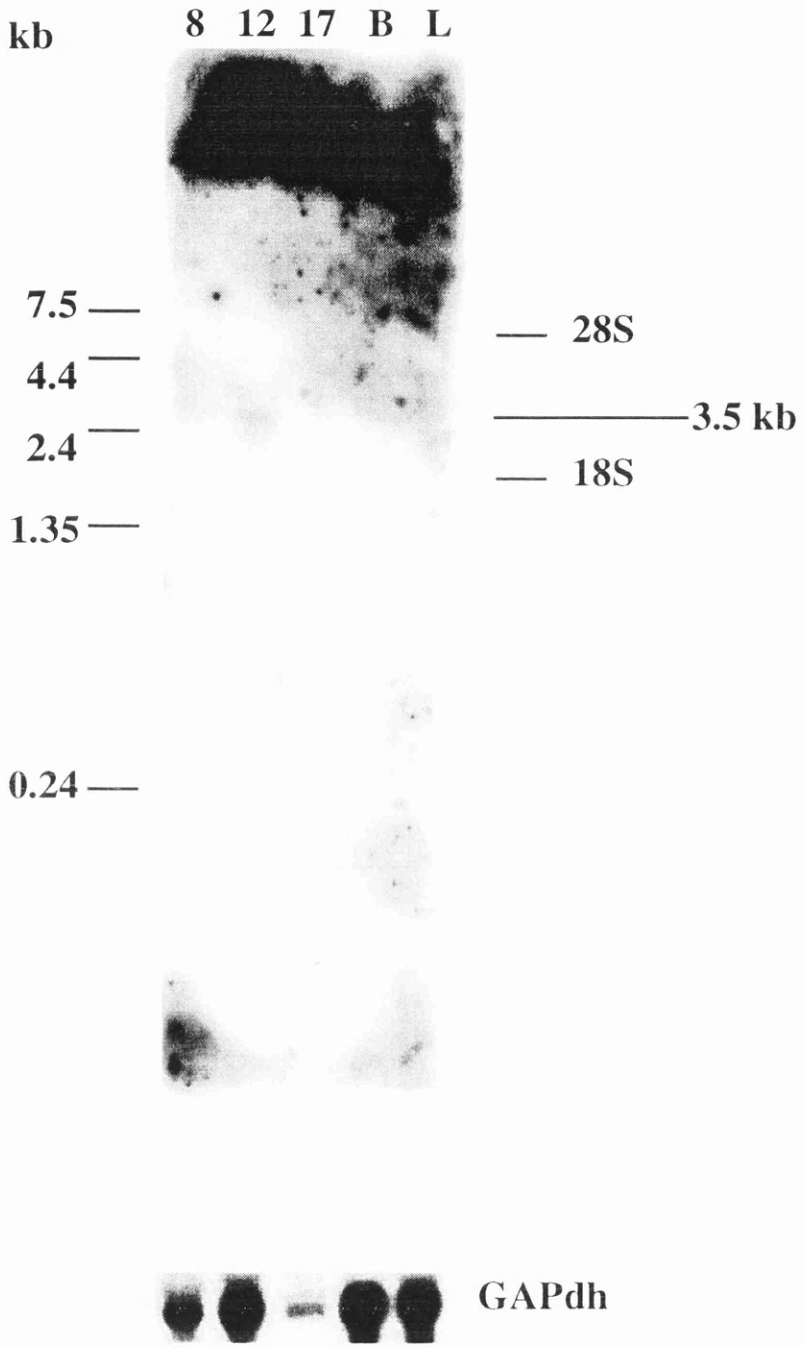


Figure 5.10 Analysis of the expression of Clone mCTG 6 3 with 1.5 kb fragment.

10 µg of total RNA from 8.5 (8), 12.5 (12) and 17.5 (17) dpc whole embryos and from adult brain (B) and liver (L) was probed with radioactively-labelled sequences derived from the 1.5 kb *Hin*dIII-*Eco*RI fragment of Clone mCTG 63. This blot was re-probed with mouse GAPdh sequences as a loading control.

Figure 5.11

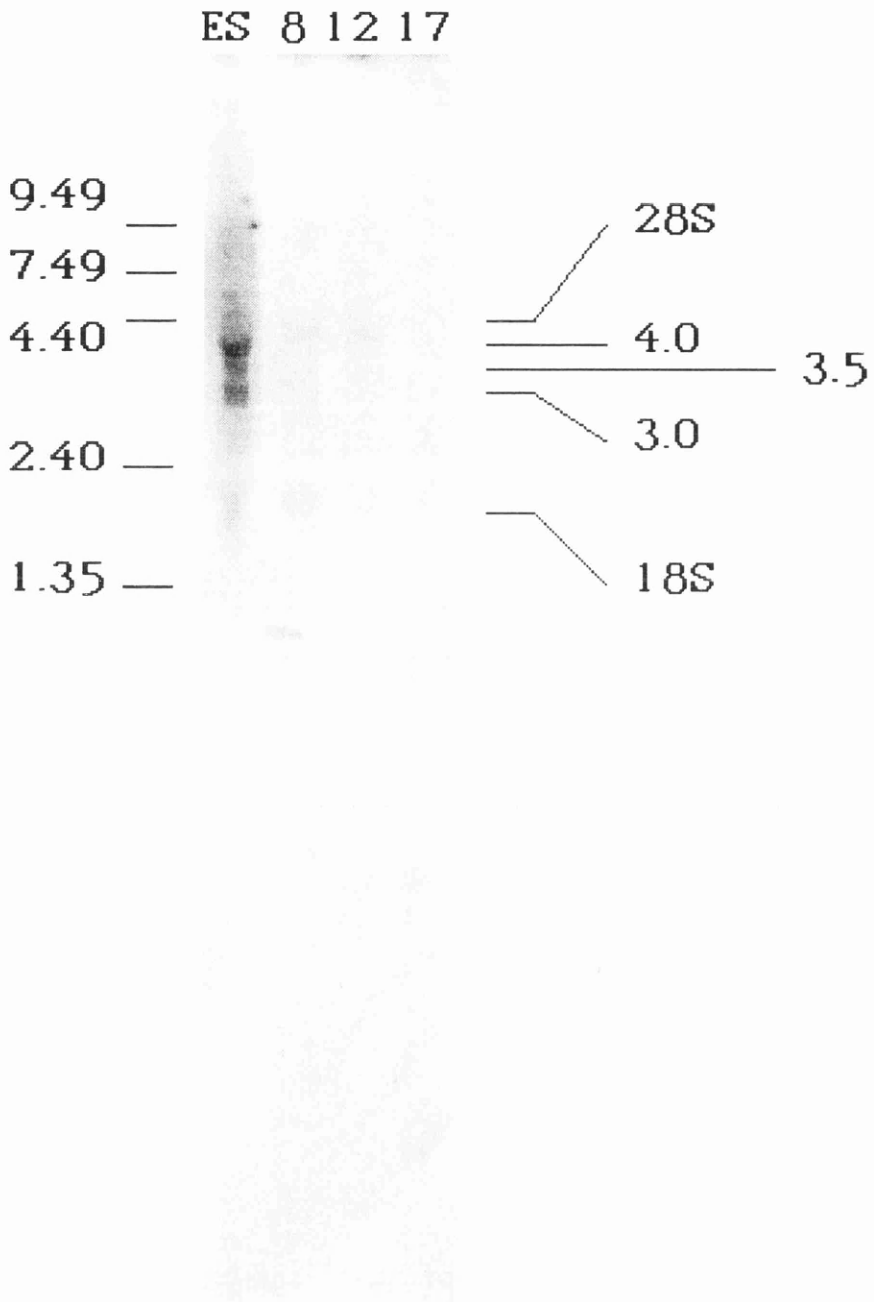


Figure 5.11 Expression analysis of sequences contained within the 1000 bp *Hin*dIII-*Eco*RI fragment of Clone mCTG 63.

10 µg of total RNA from: embryonic stem cell (ES), 8.5 (8), 12.5 (12) and 17.5 (17) dpc whole mouse embryos were probed with sequences from the 1 kb fragment of Clone mCTG 63.

5.5.3 Further expression analysis of Clone mCTG 63.

Radioactively-labelled DNA fragments derived by random priming from a 1.5 kbp *Eco*RI-*Hin*dIII fragment of clone mCTG 63 were used as a probe for hybridisation against total RNA samples. The RNA samples were derived from different embryonic stages (ES cells, equivalent to 3.5 dpc pre-implantation embryo; 8.5 dpc; 12.5 dpc and 17.5 dpc embryos) and adult mouse tissues (brain and liver, derived from inbred strain C57BL/6J). An mRNA calculated to be 3 to 4kb in size was detected in total RNA derived from 12.5 dpc embryos and adult brain and liver. In a separate experiment, using the remainder of clone 63 as a template for a probe, two mRNAs in ES cell total RNA were detected but no expression was observed in RNA samples derived from other sources (8.5, 12.5 and 17.5 dpc whole mouse embryos).

5.6 Discussion.

The further characterisation of a few of the trinucleotide repeat-containing clones described in this Chapter has lead to some general observations. These are: 1) the genes selected are single copy per haploid genome; 2) expression is more complex than first observed for the reverse dot blot experiments described in Chapter 3. Characterisation is incomplete and further work will have to be conducted to resolve the questions raised by the work presented.

All the genes which have been characterised appear from the available genomic Southern blot hybridisation analysis, to be single copy genes.

Messenger RNA expression is more complex. Clone mCTG 210 appears to produce a single transcript. Clone 26 also yielded a single mRNA. However the lack of detectable expression in mouse liver total RNA in the northern blot experiment using a partial fragment of Clone mCTG 26 is contradictory to the result of the reverse dot-blot experiment (Chapter 3) in which a high level of expression was detected with a heterogeneous first strand cDNA probe derived from adult liver. This may be explained by alternative splicing events, where sequences from a single gene are differentially expressed in different tissues and at different stages during development. This is possible because not all the sequences that were used as a target for the reverse dot-blot were used as a probe in the northern experiment described in this Chapter. Therefore the sequences omitted from the northern blot experiment of this Chapter may be expressed in the liver. The complementary experiment, using the other section of the clone, was not done. This was because of the presence of the trinucleotide repeat in this fragment. Repeats can affect the result of an experiment. In this Chapter (Figure 5.5) I have already shown the effect of the existence of a triplet repeat in a probe sequence; a multiple banding pattern is observed in genomic sequence with a probe containing the trinucleotide repeat of Clone mCTG 210. Furthermore Duboule *et al* (1987) observed a multiple banding pattern on poly A enriched mouse RNA with a CAG/CTG rich *Drosophila melanogaster OPA* repeat probe. Attempts to probe with sequences that contained CAG/CTG repeats would yield similar results.

Clone 63 was a single copy gene as shown in genomic Southern results (Figure 5.9) but expression of the sequences in this clone is likely to be complex. This is evident from the northern results

presented in this Chapter. A fragment containing the trinucleotide yielded a single broad band expressed through RNA extracted from development, and adult liver and brain. This replicates the results of the reverse dot blot data for this clone described in Chapter 3 (Section 3.3). However in a separate northern experiment, a probe derived from the whole sequence of Clone mCTG 63 gave two bands in embryonic stem cell total RNA with no detectable expression in total RNA derived from embryo of several development time points tested (Figure 5.11). These two transcripts are similar to those described for the human RAS protein inhibitor, JC310 (Colicelli *et al*, 1991). J. Colicelli observed two transcripts of 3 and 4 kb respectively (J. Colicelli, pers. comm.). This is contradictory to the northern result.

This contradictory evidence from the two northern experiments will require further analysis. For example probing with the remainder of the sequences of Clone mCTG 63 would allow the identification of sequences which contribute to a single band. It may also be possible that this clone is a chimaera. Although this is a rare occurrence, caution must be applied in any interpretation of the results. This can be resolved with the isolation of the remaining sequence from this clone. This would allow for an analysis of open reading frames contained within this clone, any discrepancies in the reading frames may indicate putative alternative splicing events giving rise to different mRNA species and hence different isoforms of the protein.

The effect of repeat sequences (as part of probing sequences) on Northern and Southern analysis is clear from the experiments presented in this Chapter (see Figure 5.5). This is not unexpected because trinucleotides are known to be over-represented in

mammalian genomes and it is not surprising that a probe containing repetitive sequences would tend to pick up many of the sequences sharing the sequence. Indeed this is the basis of the initial isolation of trinucleotide repeat-containing sequences in this work, where an oligonucleotide containing 10 copies of the trinucleotide CTG was used to screen mouse cDNA libraries. In addition triplet repeats were found to be repetitive in eukaryotes by cross hybridisation. The original work on the *Drosophila Notch* gene, described by Wharton *et al* (1985), where a single fragment from a *Notch* cDNA clone gave a multiple banding pattern on genomic DNA. Duboule *et al* (1987) has showed a similar multiple banding pattern by using a *OPA* repeat rich sequence to probe mouse mRNA.

In summary, from the work described in this Chapter, it can be concluded that these genes selected here are single copy with respect to the haploid genome. Their expression however, may prove to be more complex than that observed from the reverse dot-blot data. For example a sub-fragment of clone 26 used as a probe for total RNA in the northern experiment is not expressed in mouse adult liver, whereas the reverse dot-blot experiments of Chapter 3, using the sequences of whole clone indicated that there was expression in mouse adult liver.

More work is required on these clones to take them to a stage where a clearer molecular characterisation is achieved. This will provide a strong base for the further study of these genes and their products at the protein and cellular level. In addition we can use the information to drive towards picking trinucleotide loci which are located in single copy sequence in the genome. This will help in mapping the sequences and the genes from which they are derived from.

Chapter 6

Concluding Remarks.

6.1 Introduction.

The overall aim of this work was to identify previously un-characterised genes that are expressed during development. Two different approaches were used. The first approach involved the analysis of enhancer trap integration sites (see Appendix B), which had been pre-selected for developmental expression in head or neuronal structure by analysis of reporter gene expression patterns in chimaeric embryos.

The second approach was based on particular features of developmentally regulated genes. It involved the identification of expressed sequences from mouse embryonic cDNA libraries which contain CAG/CTG trinucleotide repeats. As discussed in the general introduction there are several reasons to think that CAG/CTG repeats may be found in developmentally regulated genes. For example, it is known that many developmentally regulated genes from *Drosophila melanogaster* contain *OPA* repeats which contain CAG triplet repeats (Wharton *et al*, 1985). Many of these genes encode transcription factors or other types of regulatory proteins. More recently this type of trinucleotide repeat has been shown to be associated with a novel class of human diseases, the dynamic mutations (Ross, 1995). These diseases are characterised by the instability of the trinucleotide repeat, in which an upward expansion of the trinucleotide copy number is associated with an increasing severity of disease. Although these diseases have as yet only been observed to occur naturally in *H. sapiens*, the best animal model in which to study the diseases is the mouse. The characterisation of repeats in mouse will provide a resource for which to study repeat biology.

The major merit of the enhancer trap approach is the early identification of expression patterns that can be used as the basis of an informed choice of which enhancer trap integration site containing ES cell lines to study further (see Appendix B). With the trinucleotide repeat cDNA library screening approach this information comes at the end of the expected process. Screening of embryonic cDNA libraries can be quicker compared to the identification of flanking sequences from the enhancer trap site. In some cases the gene can be distant from the actual ETS and additional cloning steps would be required for complete characterisation of the gene.

Another advantage that the trinucleotide cDNA library screen has over ETS analysis is the high capacity for screening large numbers of clones. This approach allows the identification of many genes which can be analysed for sequence to find anonymous sequences. The ETS analysis requires a lot of work initially to isolate the appropriate gene sequences and then to compare with other sequences to establish whether it is a previously characterised gene or not.

6.2 What have we learnt?

6.2.1 Library screening.

The screening of mouse embryonic cDNA libraries with a $(CTG)_{10}$ oligonucleotide has yielded many positive clones. These are similar in numbers to those isolated from human brain cDNA libraries (Riggins *et al*, 1992 and Li *et al*, 1993) and mouse adult

brain cDNA library (Abbott and Chambers, 1996) which had been previously screened for CAG/CTG trinucleotide repeats.

The size of CAG/CTG trinucleotide repeats found in the mouse embryonic cDNA libraries screened here are most similar to those identified by the Li *et al*, (1993) study which used identical conditions. The range of repeats identified in both these studies was large, although many were larger than 10 trinucleotides in length, the CAG/CTG repeats identified by the other human study (Riggins *et al*, 1992) and the mouse adult brain cDNA library (Abbott and Chambers, 1996) are smaller. The repeats in these studies are all less than or equal to 10 trinucleotides in length. This reflects either less stringent hybridisation conditions, the absence of formamide, a known nucleic acid secondary structure de-stabiliser, in the hybridisation solution, or a smaller oligonucleotide probe used for screening cDNA libraries. In the case of the adult mouse brain cDNA library screened by Abbott and Chambers (1996), a 15mer oligonucleotide was used to probe the library compared to a 30mer oligonucleotide used in this work and in the Li *et al* (1993) study.

However all studies described above identified similar numbers of trinucleotide repeat containing clones irrespective of the screening conditions used. This may be because of the quality of cDNA library screened (C. Abbott, pers. comm.) or the way in which the library was constructed. Alternatively it could be that the differing screening conditions have a predisposition towards the isolation of trinucleotide repeats of a certain size with an upper limiting value, dependent on the screening conditions used (*e.g.* the use of formamide and oligonucleotide probe length).

Partial sequence analysis of clones from the 8.5 dpc (Hogan *et al*, unpublished) and 12.5 dpc (Logan *et al*, 1992) mouse whole

embryo cDNA libraries identified CAG trinucleotide repeats in 21 of the 23 clones analysed in this work. This indicates that the library screening conditions were a success. Many of the repeats were corrupted with base pair substitutions.

6.2.2 Comparative analysis.

The sequence data for the clones were used to search both nucleic acid and protein databases using FASTA (Lipman and Pearson, 1985) and BLAST (Altschul *et al*, 1990). This analysis identified that 10 were entirely anonymous in that there was no similarity to any nucleic acid or protein sequence in the databases that were searched. A further 3 clones had identity to anonymous sequences, ESTs. The remaining 10 clones had some degree of similarity or identity to genes, most of which (9), have assigned functions. Three have identity to previously described mouse genes (mCTG 43, Rad21; mCTG 56, CW17; mCTG 24, nuclear receptor co-repressor).

Comparison of the repeats between species, where possible, shows that in most cases, the mouse repeat is longer than or equal in length to the repeat found in the homologous human gene (see Figure 4.4). This can be taken as evidence for an ascertainment bias where the screened species trinucleotides are larger because of a positive selection for larger repeats over smaller repeats. This is contrary to the idea proposed by Rubinzstein *et al*, (1995), propose that there is a species specific difference in rates of microsatellite expansion in primate species, with humans possessing a marginally higher microsatellite mutation rate. Forty two independent microsatellite loci were examined in a variety of primate species

and it was found that the human loci were significantly larger than those cognate repeats from other primate species.

An ascertainment bias between the two sets of data is a more likely explanation. This hypothesis states that microsatellites from different species selected on a size basis will show larger repeats in the screened species compared to the comparison species. The larger trinucleotide repeats from this work in mouse tend to have smaller counterparts in human genes. For example Clone mCTG 23 which has a trinucleotide repeat 12 units in length; this repeat unit is absent from the human gene (the 70kDa subunit of replication protein A). This repeat encodes a polyglutamine tract in this protein.

This leads to another observation. Most of the translated CAG/CTG trinucleotides encode polyglutamine tracts. In 5 out 6 sequences for which a definite position can be determined for the trinucleotide within the coding region of a gene, by virtue of identity to other genes, the CAG repeat is translated into a polyglutamine homopeptide. This is not entirely unexpected because there is a tendency for this class of repeat to encode glutamines in *Drosophila* (Karlin and Burge, 1996). These are known as *OPA* repeats (Wharton *et al*, 1985). Most of the genes in which these *OPA* repeats occur are expressed in development and in the nervous system (Karlin and Burge, 1996). It has also been observed that glutamine homopeptides are the most abundant in eukaryotic proteins (Green and Wang, 1994). There has been no data published for mouse, but there are examples of proteins which contain polyglutamine tracts, *e.g.* Brain 2 (a POU domain transcription factor, Hara *et al*, 1992).

Another feature that is striking is that a high proportion of the sequences which have identity to previously characterised genes are nucleic acid binding proteins. Seven of the nine known genes identified encode proteins that bind nucleic acids (either DNA or RNA).

As mentioned above, recent research has shown that a large number of DNA binding transcription factors contain CAG trinucleotide repeats, and that these repeats often encode glutamine (Gerber *et al*, 1994; Karlin and Burge, 1996). Within these genes that bind nucleic acids, there are examples of transcription factors (clone mCTG 24 which shows similarity to the nuclear receptor co-repressor and clone mCTG 57 which shows homology to the transcriptional adaptor protein P300). This is however a low overall proportion of the CAG/CTG repeat-containing cDNAs and therefore the suggestion that there is a tendency for CAG repeats that are translated as polyglutamine to occur in transcription factors, may be false or at least have to be modified. It is presumed that the polyglutamine tract has an interactive property with other proteins. With all the proteins identified, it is known that they do interact with other proteins. This may indicate that proteins that interact with nucleic acids do so in conjunction with other proteins as part of complexes in many cases.

6.2.3 Larger CAG/CTG trinucleotide repeats are more likely to exist in novel genes.

The trinucleotide repeats in clones with similarity to genes with known function are generally in the small to medium size spectrum (six to twelve repeats). Larger repeats isolated by this

screen tend to be in clones which are likely to be in previously un-characterised genes. This is an interesting observation. This justifies the screen as an attempt to identify novel mouse genes. However, this may be an artifact, as no comprehensive screen of mouse cDNA libraries for CAG/CTG trinucleotide repeats has been previously conducted and therefore the potential set of CAG/CTG trinucleotide repeat-containing clones is large.

This also indicates that many of the larger CAG/CTG repeats reside in novel mouse genes. One aspect of their usefulness lies in the fact that they could be used more successfully as microsatellites. It has been observed that larger repeats are more likely to be variable than smaller repeats (Weber, 1990). This could also point to particularly unstable repeats in the mouse, which may act as models for unstable human trinucleotide repeats which have been associated with diseases (Ross, 1995).

6.3 Analysis of expression.

A major goal of the work presented here was to identify developmentally regulated genes. This has been partially successful. This section will consider the results of this work.

From the work presented here it is interesting to point out that although I have no information on the majority of the genes examined in this work, of those which I do have information on their expression, some are highly expressed but do not correspond to known genes. This indicates that there are many unknown genes in the mouse which are still waiting to be discovered which are as yet not characterised in organisms which are far more advanced in the analysis of developmental processes.

6.3.1 Reverse dot blots.

In an attempt to identify genes which are expressed during development, a technique was used which had a high capacity. This was necessary to analyse multiple clones at one time. The technique which was used reversed the normal northern procedure by fixing the defined quantities (*i.e.* the clones) to a nylon membrane and probing them with a probe derived from heterologous oligo-dT primed first strand cDNA products. This is known as reverse northern procedure.

Unfortunately the detection threshold of this experiment was high. Although a few clones were detected as having high expression, and some show differences in expression (in the adult tissues), most do not show any signal. Furthermore the positive controls selected did not show any detectable expression. It was concluded that the sensitivity level of this experiment was low.

Nevertheless a high sensitivity test of expression will need to be used in the future if the remaining clones from this work are to be characterised by their expression patterns during development and in adult tissues. One way to achieve this is by enrichment of the RNA sample (from which the first strand cDNA probe is prepared) for poly A^+ mRNA. This will effectively concentrate mRNA molecules which are derived from the genes corresponding to the clones identified here.

RT-PCR would be another route to pursue. This requires sequence from the clones to be known so that primers can be designed which would amplify specific regions of a particular clone

from first strand cDNA products by PCR. This technique has sufficient sensitivity.

In addition if the primers were to include the region containing the CAG/CTG repeat, the primers could be re-used for mapping the gene by using congenic strains of mouse which are hybrids of *M. musculus* and *M. spretus* species of mouse to assign the genes to chromosomes and EUCIB backcross panel DNAs to locate them to 2-5 cM regions (*e.g.* McCallion *et al*, 1996)

6.3.2 Northern blot analysis.

In an attempt to broaden the expression information derived from the reverse northern experiments, northern blots were used to characterise the expression of four selected clones. This uncovered a level of complexity in the expression of the genes from which these sequences are derived, that was not evident from the reverse northern experiments. For example sequences from different fragments of clone mCTG 26 appear to be expressed in differing tissues. It was found from the northern blot using total RNA from various embryos and adult tissues (brain and liver) that the 500bp *Bam*HI-*Bgl*II fragment (which was used to prepare the radio-actively labelled probe) did not detect expression in liver total RNA. This is contrary to the reverse dot-blot experiment where there was a signal detected from the adult mouse liver (Figure 3.1).

6.4 Future work.

6.4.1 Extended molecular characterisation.

A major component of any future work on these clones will be the continued characterisation of these genes at the molecular level. This has been started for a few of the clones. Not all of these clones will fit the criteria set for further analysis, namely the existence of a large repeat which is translated and occurs in a developmentally regulated protein. The clones which do not meet these requirements need not be discarded as there are alternative criteria which could make use of these clones. For example, the repeats could be useful in the mapping of genes onto the mouse genome, or genes could be of importance in adult physiology.

In the latter stages of this research I concentrated on four clones. Three of these have turned out be novel, with the fourth showing similarity to a human RAS inhibitor (Clone mCTG 63, JC310; Colicelli *et al*, 1991). Two of these were widely expressed (clones 26 and 63). This is an interesting parallel to the genes which are affected in human unstable trinucleotide repeat diseases which also show widespread expression, beyond the groups of tissues which are affected by that particular disease. It would therefore be interesting to pursue the isolation of the human counterparts of these mouse genes to ascertain if they also contain trinucleotide repeats. This may allow the identification of the genes in particular human diseases. The trinucleotides may also allow the mapping of these genes to regions of the genome which may contain previously identified genetic disease loci. Furthermore molecular

characterisation of these mouse genes may lead to the identification of novel regulatory genes.

Cloning and analysis of the remaining clones from the 8.5 dpc (Hogan *et al*, unpublished) and 12.5 dpc (Logan *et al*, 1992) mouse embryo cDNA libraries as well as those from the other two libraries, the 13.0 dpc mouse embryo cDNA library and the mouse adult brain cDNA library was not carried out in part due to the large number of clones to be purified. This should be pursued to identify more repeats to improve the statistical significance of the observations concerning 1) the size and structure of repeats in mouse; 2) the observation of an ascertainment bias; 3) that most translated repeats encode glutamine tracts.

An ongoing DNA sequencing project will generate a resource of proteins that are likely to interact with other proteins, and a higher incidence of regulatory genes. This could be tested by taking these anonymous genes and use them to test for stimulation of transcription in whole cell protein fractions as described by Gerber *et al* (1992). This involves the fusion of the sequences containing the CAG/CTG repeat with a GAL4 DNA binding domain which interacts with binding domains in a reporter construct to stimulate transcriptional activity of a reporter gene.

6.4.2 Substrates for trinucleotide binding proteins.

These trinucleotide repeats may have a role in the characterisation of the CAG/CTG nucleic acid trinucleotide repeat binding proteins that have been previously described in both *Homo sapiens* (Richards *et al*, 1993; Timchenko *et al*, 1996) and *Mus musculus* (Yano-Yanagisawa *et al*, 1995). The identification and use

of these isolated CAG/CTG trinucleotide repeats as substrates for the action of these proteins in the mouse could help in the development of *in vivo* models for the role of these proteins in trinucleotide repeat biology and pathology.

6.4.3 Human unstable trinucleotide repeat disease models?

Recently a new class of mutation that causes disease in human has been identified. These are known as dynamic mutations and are caused by the hyper-expansion of trinucleotide repeats. The most common triplet repeat which causes this type of disease are members of the CAG/CTG repeat class. Of these most of the diseases are caused by the expansion of CAG triplets which encode polyglutamine repeats. Some of the repeats identified in mouse during this work are large and may represent models in which an expansion may cause a disease. It has been shown that human transgenes of ataxin, the gene mutated by a CAG/polyglutamine expansion in the disease SCA1, in mouse confer a mutant phenotype in the mouse. This suggests that the necessary cellular mechanisms which are recruited for the cell death that occurs in the human disease are present in the mouse. On this basis it is a valid experiment to expand an endogenous mouse CAG trinucleotide repeat that encodes a polyglutamine tract to look for a disease phenotype as a result of neuronal cell death. All of the diseases that have been found to be caused by a polyglutamine expansion display a disease specific death of certain populations of neurons (see Figure 1.5 of the Introduction).

6.4.4 Large mouse trinucleotide repeats: models for repeat variability and disease.

The existence of hypervariable mouse trinucleotide repeats would help in the study of the human dynamic diseases in that it would help to establish a credible *in vivo* animal model for these diseases. This however, is a controversial proposition as there has been a lack of observed variability in human transgenes which have been introduced into mice, although it can be counter-argued that the reason for this observation is that this approach is problematic because of the hemizygous nature of any introduced transgene. This may preclude the participation of repeats (which exist as part of a transgene which is single copy per diploid genome) in cell division (*i.e.* unable to pair in mitosis and/or meiosis) or even expansion events (*e.g.* if sister chromatids are required for the mechanism of repeat expansion).

Another argument against using the mouse is that the mouse homologues of those human genes which have unstable trinucleotide repeats show little or no repeat structures (see Figure 1.6 in the Introduction). From this people have suggested that mice do not have the particular cellular environment under which trinucleotide repeats can expand in an uncontrollable fashion. Candidate genes for this explanation would be those which encode proteins that are involved in recombination. This argument is weakened by the observation that mouse repeats (and rodent microsatellites in general) are longer and as variable as human repeats. For example, a CAG/CTG trinucleotide repeat (which encodes a polyglutamine homopeptide) in *Sry* (the *Mus musculus domesticus* sex determining gene) is variable in the wild type

mouse population. Repeats of 17 and 14 CAG triplets were found in the *Sry* genes of trapped wild mice (from Denmark and France respectively; Nagamine *et al*, 1992) compared to 12 CAG triplets in the inbred strain C57BL/6J(B6). This suggests that CAG/CTG trinucleotide repeats are variable in the general population. The problem with inbred laboratory mouse strains is that they are not a natural population and are quite unlike the human population. These mice strains are in effect a snap-shot of a single natural mouse at the time of capture. They therefore may contain important co-factors that preclude repeat variability. Alternatively they may represent the fit individuals in the population. This reasoning suggests that mice which are born in to the wild population with CAG/CTG trinucleotide repeats are inherently at a disadvantage and are eliminated within a few generations due to anticipation, where successive generations succumb to a particular disease earlier in their life and more severely. This does not occur in the human species, where CAG/polyglutamine repeat diseases are concerned, because of the unique care we provide for other individuals of the same species (*e.g.* hospitals, medical care, families). It is worth noting that this is not true for all dynamic diseases. In the case of FRAXA males and congenital DM infants, the ability to reproduce is severely diminished and therefore the genotypes of these individuals are eliminated from the gene pool.

Interestingly, the *Sry* CAG repeat has been associated with a sex reversal phenotype in crosses of the C57BL/6J(B6) inbred strain and other *Mus musculus domesticus* strains (Coward *et al*, 1994). The critical feature of this phenomenon appears to be the number of glutamines (and hence the CAG triplets) at this position in the *Sry* protein. Strains which cause a sex reversal phenotype when crossed

with C57BL/6J(B6) mice contain either 11 or 13 glutamines (CAG triplets) at this position, compared with C57BL/6J(B6) which contains 12 glutamines. Strains which do not lead to a sex reversal phenotype contain 12 glutamines (CAG triplets). From this it can be concluded that in certain cases in the mouse, the number of glutamines in a homopeptide can be critical for the proper functioning of proteins. This therefore argues for a critical role for mouse CAG/CTG trinucleotides in certain cases. Therefore a valid route of further experimentation would be to test the variability of these repeats and examine them for variability and potential associations with disease.

6.4.5 A model for Myotonic Dystrophy?

CAG/CTG repeats are not exclusively associated with pathologies that involve the translation of the repeat into polyglutamine. In the case of myotonic dystrophy, the trinucleotide repeat resides in the 3'UTR of the DMPK gene. The trinucleotide identified in clone 63 is the largest repeat identified so far in this study (where there are 26 perfect trinucleotide repeats). It is likely that this repeats lies in the 3'UTR of a gene which is a homologue of a human RAS inhibitor (Colicelli *et al*, 1991). The coding strand sequence is CTG, which would be transcribed as CUG in RNA. This is analogous to the situation in the DM region, where the trinucleotide is similarly found in the 3' UTR of the DMPK gene. This mouse repeat could act as a model system for the effect of the DMPK 3' UTR CTG trinucleotide. However as indicated in the introduction this may be complicated by the existence of other genes which are in the vicinity of the DMPK gene (Johnson *et al*, 1996). The isolation of

a CUG repeat in a gene which may lie in a less gene-rich region may allow the study of a RNA CUG repeat independently of the field effect postulated by Johnson *et al* (1996).

6.5 Summary.

In conclusion this study has been successful in that it has identified mouse CAG/CTG trinucleotide repeats in novel mouse genes which are expressed in development. Within the group of characterised repeats which were identified, larger repeats tend to exist in novel gene sequences. This indicates that there is an enrichment for novel genes within this screening procedure. Furthermore those clones which were selected for further investigation have lead to the identification of genes for which a search for the human homologues would be worthwhile to investigate the existence of the trinucleotide repeat and possible associations with human diseases.

Appendix A

Sequence From Novel Clones.

A.1 Introductory remarks.

As previously mentioned in Chapter 4 (Section 4.5), analysis of the DNA sequence from clones derived from the 8.5 and 12.5 dpc whole mouse embryo cDNA libraries revealed that eleven (out of 23) showed no similarity or identity to sequences in the nucleic acid or protein databases. Three of these (clones mCTG 26, 210 and 61) were selected for further molecular analysis and are described in Chapter 5.

The remaining clones are represented here in this appendix and the CAG/CTG trinucleotide repeats identified within these sequences are indicated (if present). Other sequence motifs are indicated where necessary. For the details of each clone refer to the legend for each Figure.

Figure A.1

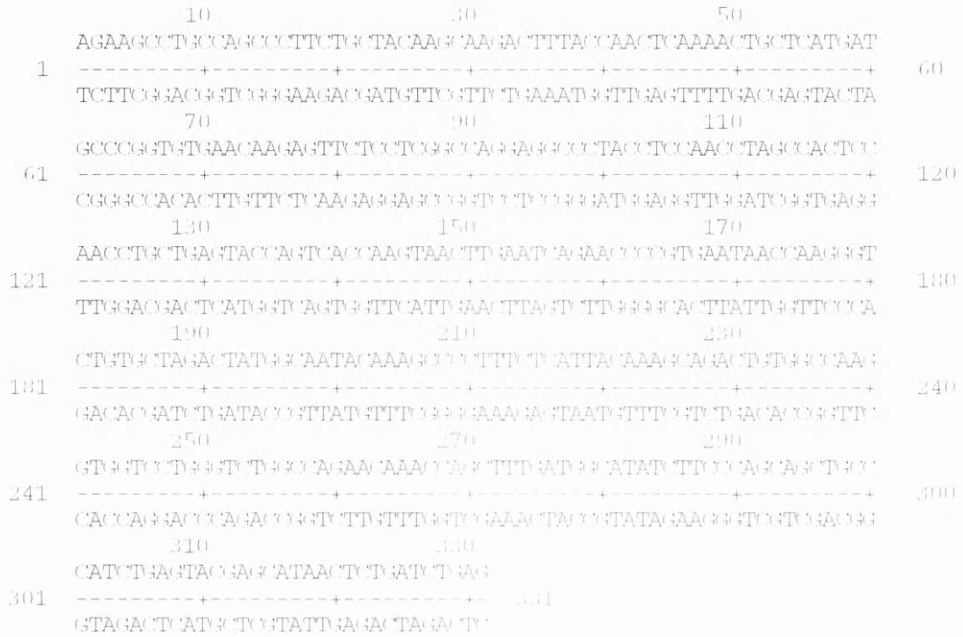


Figure A.1 Partial nucleotide sequence of clone mCTG 27. The sequence presented here was derived from a T7-primed manual sequencing reaction (Pharmacia T7 polymerase kit). No trinucleotide was identified in this clone.

Figure A.2.A

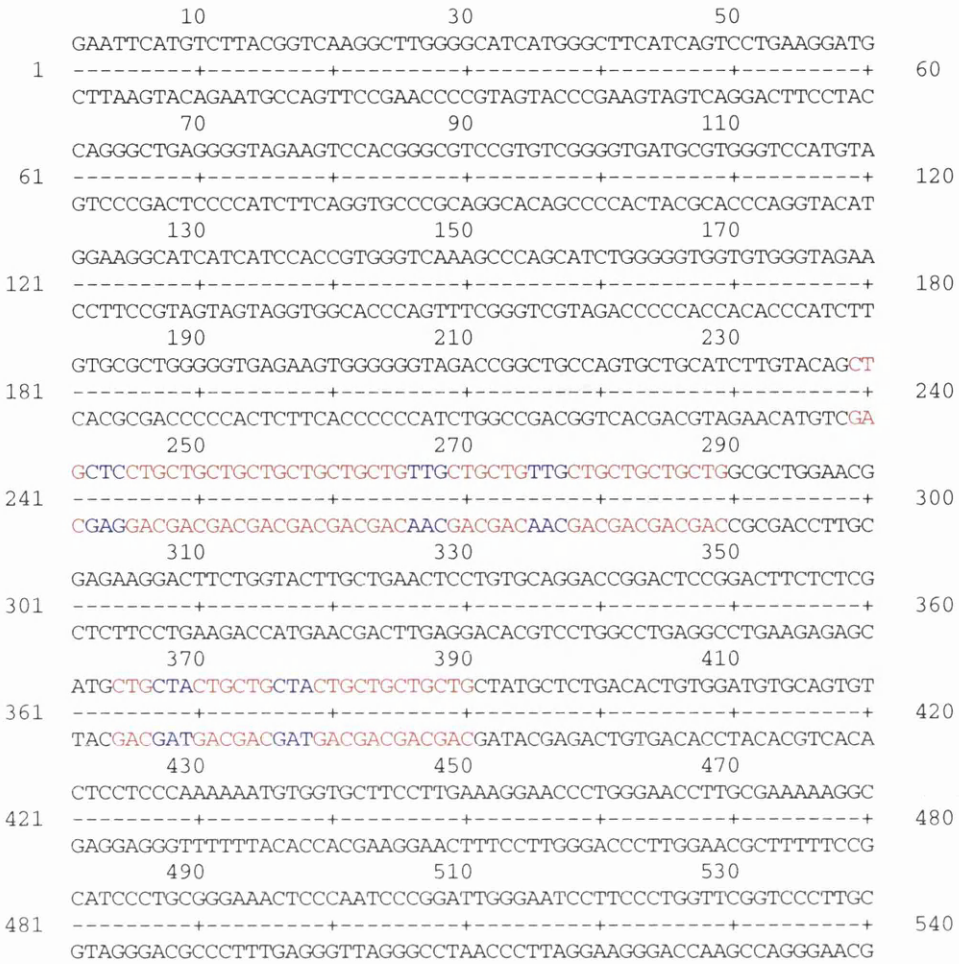


Figure A.2.A Sequence from Clone mCTG 28 using T3/T7α primer.

This sequence was derived from using the T3/T7α oligonucleotide to prime DNA sequencing reactions on the plasmid containing the mCTG 28 insert sub-cloned from lambda. The CAG/CTG trinucleotides are indicated by red coloured text. Note that there are corruptions in the trinucleotide repeat (indicated by blue coloured text). The trinucleotide repeats are CTGctc$(CTG)_7$**t**t**g**$(CTG)_2$**t**t**g**$(CTG)_4$ and atgCTGc**t**a$(CTG)_2$c**t**a$(CTG)_4$cta. The bold small case nucleotides represent the third position codon changes.

Figure A.2.B

Figure A.2.B Nucleotide sequence of Clone mCTG 2 8 derived from T7 primed DNA sequencing.

The sub-cloned mCTG 28 insert (derived from the 8.5 dpc whole mouse embryo cDNA library) was sequenced both with the standard T7 primer (Promega) and the bT7 extended DNA sequencing primers (Stratagene). No trinucleotide repeats were observed in this sequence but trinucleotides were observed in T3 primed DNA sequencing (see Figure A.2.A). The existence of more CAG/CTG trinucleotides cannot be discounted as the clone is not fully sequenced.

Figure A.3

Figure A.3 Nucleotide sequence of Clone mCTG 45 derived from T7 primed DNA sequencing.

Clone mCTG 45 insert (derived from the 8.5 dpc whole mouse embryo cDNA library) was sequenced using T7 based primers (T7, Promega; bT7, Stratagene). The sequences were aligned in the MacAlign programme (IBI). A single CAG/CTG trinucleotide repeat was observed. The sequence was cgc$(CAG)_8$cga and is indicated as red coloured sequence. A CT/AG dinucleotide repeat adjacent to the trinucleotide is coloured green.

Figure A.4

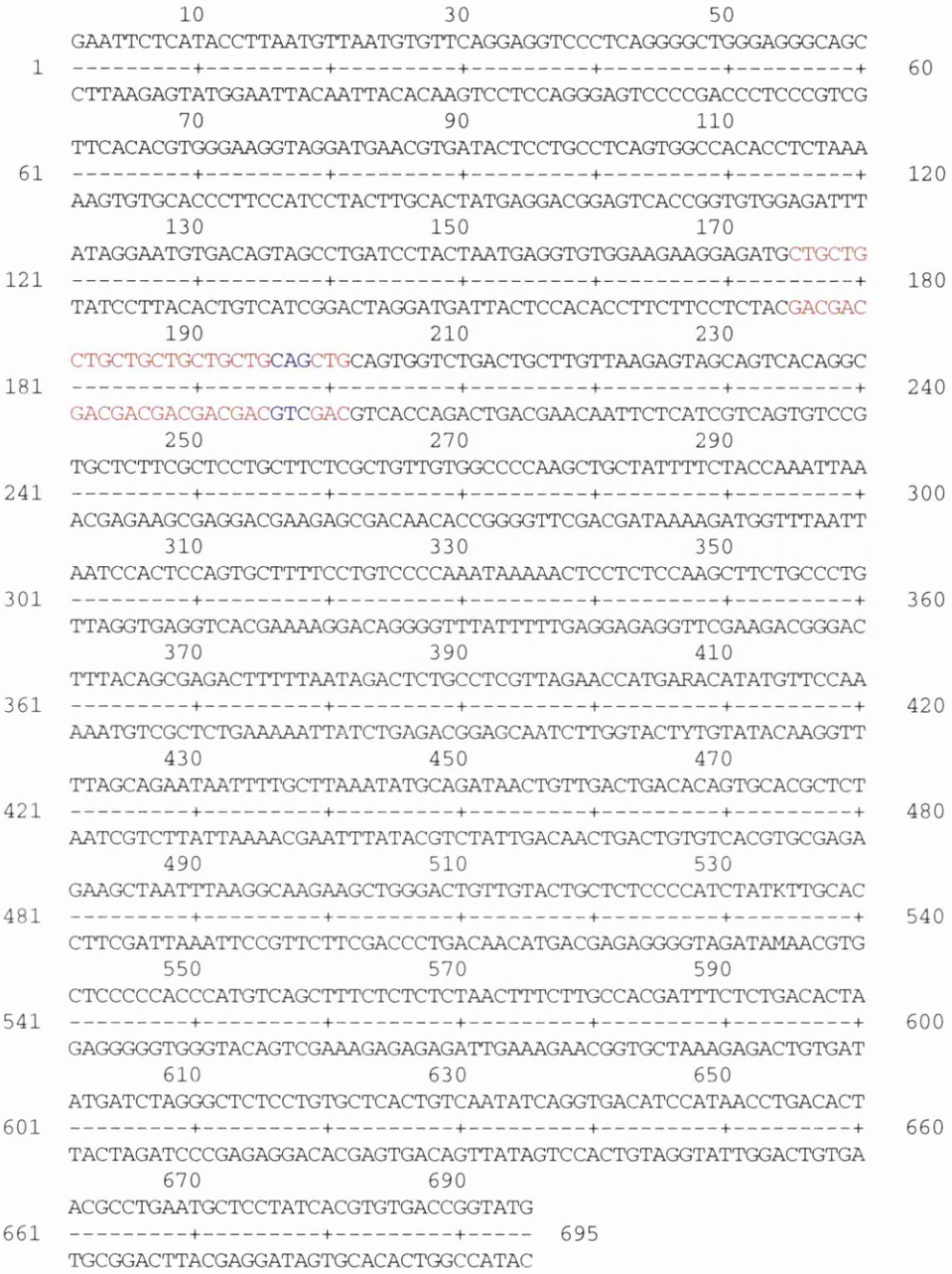


Figure A.4 partial nucleotide sequence of Clone mCTG 414
This clone was isolated from the 8.5 dpc whole mouse embryo cDNA library. the DNA sequence represented here was derived from a sequencing reaction using a PCR cycle sequencing optimised, T7-based primer. This sequence contains the trinucleotide repeat (red coloured text) found in this clone. It reads: atg$(CTG)_7$cagCTG. The reading frame remains to be determined with further sequencing information.

Figure A.5

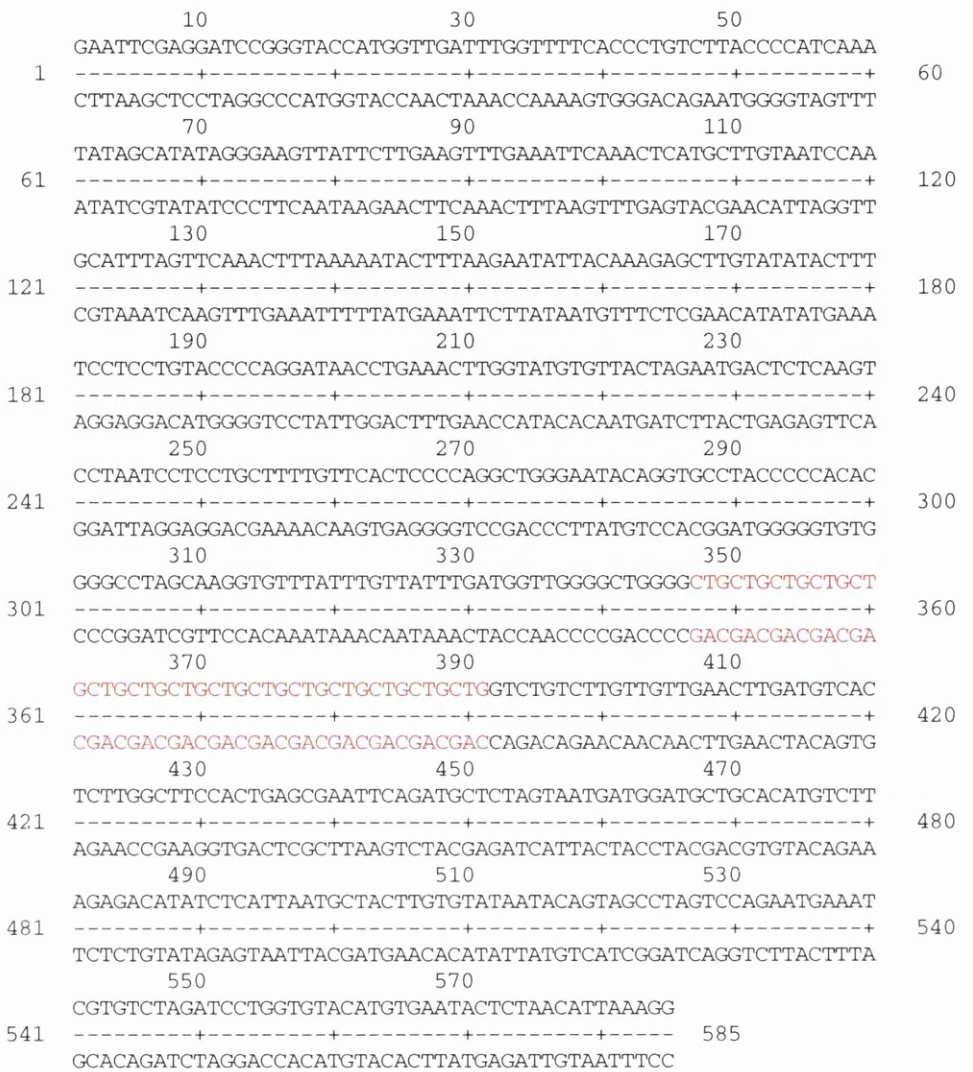


Figure A.5 Nucleotide sequence of Clone mCTG 81.
Clone mCTG 81 was initially isolated from the 12.5 dpc whole mouse embryo cDNA library. The sequence presented here was that obtained from using a PCR cycle-sequencing T3 based primer. The CAG/CTG trinucleotide repeat is represented in red coloured text. The repeat is ggg$(CTG)_{15}$tct.

Figure A.6

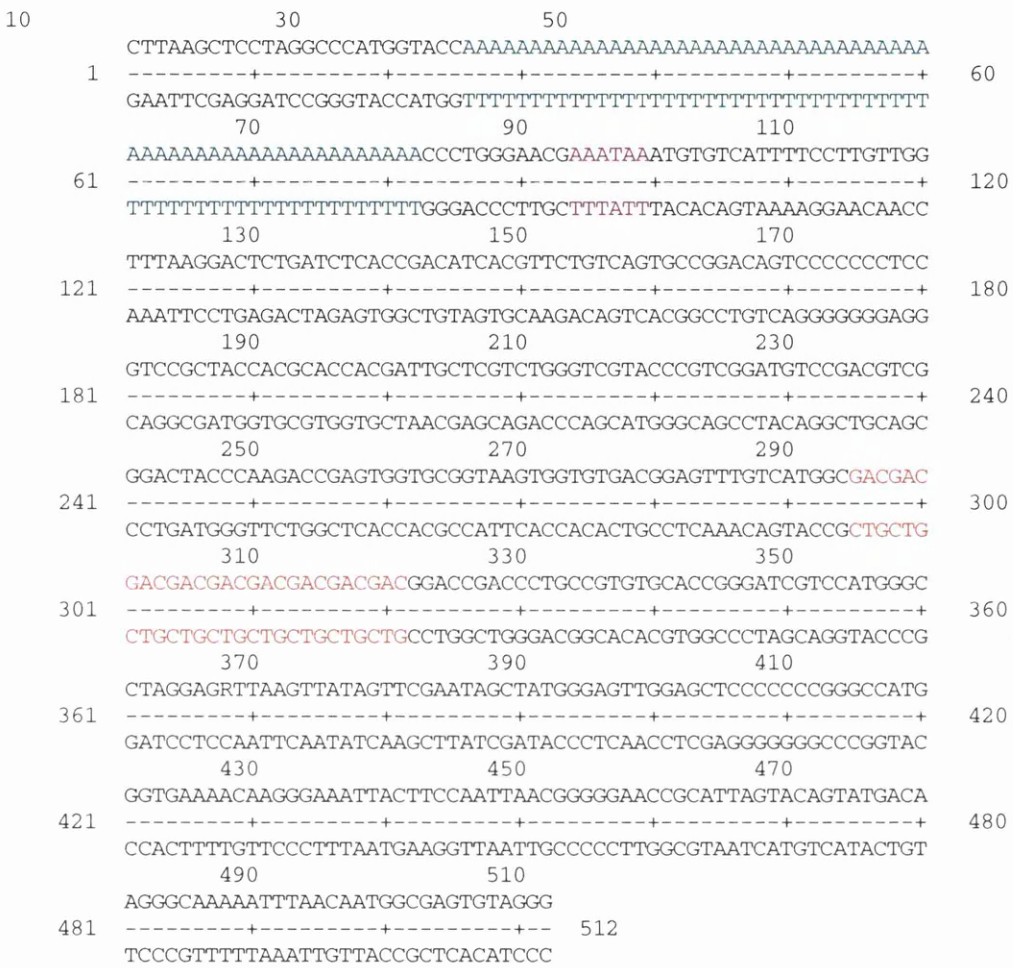
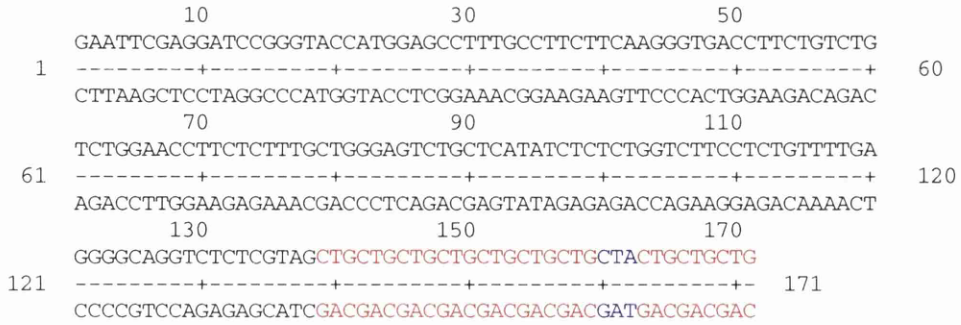


Figure A.6 Partial nucleotide sequence of Clone mCTG 82. Clone 82 was derived from the 12.5 dpc whole mouse embryo cDNA library. DNA represented here describes the 3' region of an mRNA, due to the presence of a large poly-adenosine stretch (n=52; green coloured sequence. Note also the existance of a Polyadenylation signal immediately 5' to the poly A/T stretch). The trinucleotide repeat found in the clone is represented in red text. The sequence of the repeat is agg(CAG)$_9$cgg in the presumed coding strand (*i.e.* that strand in which the polyA/T stretch is read as poly A).

Figure A.7

A)



B)

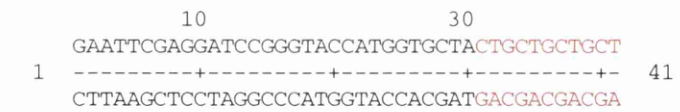


Figure A.7 Partial DNA sequence of Clone mCTG 86.
Clone mCTG 86 was initially isolated from the 12.5 dpc whole mouse embryo cDNA library. DNA sequence from this clone was derived by DNA sequencing reactions primed by T3 and T7 primers, optimised for PCR cycle-sequencing. Note that the available DNA sequence is very short due to secondary structure within the cloned insert. This may involve the CAG/CTG trinucleotide repeats (indicated in red) as the sequence terminates in both reactions in the middle of these repeats. The repeats are also contra-stranded and therefore it is possible that these trinucleotides can dissociate and form paired intra-molecular pseudo-helical hairpins. The trinucleotide repeats are: A) $(CTG)_7cta(CTG)_3$ and B) $(CTG)_3$.

Figure A.8

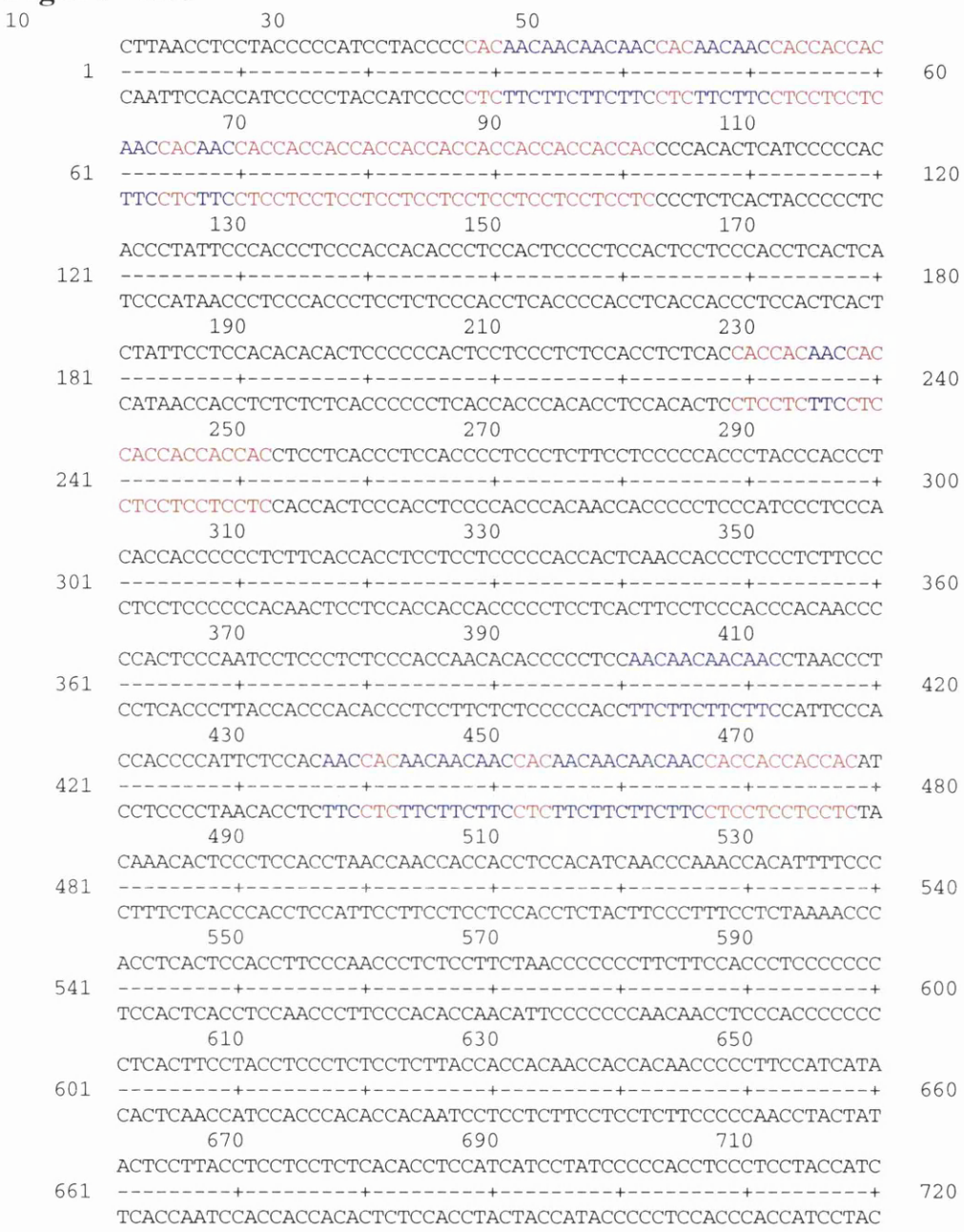


Figure A.8 Partial nucleotide sequence of Clone mCTG 92. DNA sequence was derived from using the cycle sequence optimised primers. The CAG/CTG trinucleotide repeats are indicated in italicised text. The four repeats are highly degenerate and contain equal numbers of CTG (red text) and TTG (blue) trinucleotides. The reverse of TTG is CAA. This is the second triplet which encodes a glutamine codon. Therefore it is highly likely that these repeats encode polyglutamine homopeptides, if they are translated. There is a continuous open reading frame found throughout this sequence.

Appendix B

Isolation And Characterisation Of Genomic DNA From Enhancer Trap Cell lines.

B.1 Introduction.

As described in the general introduction, there are many ways to identify new genes within parameters of interest and each of them have their advantages and disadvantages. Some of them are more suited to developmental gene identification than others. Enhancer trap constructs involve the introduction of transgenes into the genome of various organisms and have been developed to assay for spatial and temporal gene expression in embryos (e.g. Gossler *et al*, 1989 and Korn *et al*, 1992). The aim of the experiments is to detect the expression of a reporter gene with a minimal promoter randomly inserted in the genome. Expression will occur if the reporter gene is in the vicinity of endogenous regulatory sequences. These sequences are responsible for normal transcription from their endogenous target gene (or genes) and therefore this technique can indirectly identify sets of genes which are activated under specific conditions, for example during the development of the nervous system.

The creation of a series of embryonic stem cell lines with independent transgene integrations takes time. In a collaboration with A. Gossler from Germany, our laboratory gained access to a group of embryonic cell lines which contain integrated PLSN enhancer trap transgenes (Figure B.1; Gossler *et al*, 1989). These positive embryonic cell lines had a head and/or a spinal cord *lacz* transgene expression during development, when the cell lines were injected into donor blastocysts and allowed to contribute to chimaeric embryos. The analysis of the unique flanking sequences from the transgene integration sites in the transgene positive cell

lines will help identify genes which are expressed during development in the mouse.

This Chapter will focus on the initial characterisation of the enhancer trap positive cell lines D3-240 and D3-052. A Southern blot analysis was carried out to detect the transgene integration site and also to locate unique restriction enzyme sites in the sequences adjacent to the enhancer trap. It is also necessary to check the cell lines for tandem insertions of the transgene and that the enhancer trap site had not undergone post-integrational rearrangement(s). Flanking sequence was amplified from circularised genomic DNA using inverse PCR (IPCR) with pairs of nested primers; neo1 and neo2; neo10 and neo11 (Chapter 2 and Figure B.1). To characterise these amplified sequences, the products of inverse PCR were sub-cloned into a deoxythymidine-overhang vector and DNA sequenced. Further Southern analysis was carried out using cell lysate genomic DNA, sent from Germany. This showed that DNA derived from different cell lysates gave the same restriction fragments upon digestion with different endonucleases, when hybridised with radioactively labelled probes derived from the reporter genes, *lacz* and *neo*. This indicates that they actually contained the same enhancer trap integration site.

B.2 *Lacz* and neomycin phosphotransferase gene specific probes.

Figure B.1 shows a map of the enhancer trap vector pLSN. It contains the genes *lacz* and neomycin phosphotransferase (*neo*), which are derived from the *E. coli* strain K12 genome *lac* operon

(Maniatis *et al*, 1989) and from the *E. coli* transposon Tn5 respectively.

Restriction sites at both ends of the vector can be used to characterise the integration event and the co-linearity of *lacz* and *neo* genes and to distinguish between unique insertional events in different enhancer trap positive cell lines. Probes derived from both these genes were used.

The *neo* gene specific sequences (1 kbp) were isolated from the plasmid pGT1 (Gossler *et al*, 1989), using *Xba*I restriction endonuclease. A *lacz* gene fragment, (1.8 Kb *Hinc*II-*Hinc*II, 2349 nt to 4177 nt of the whole lac operon of the *E. coli K12* genome, Maniatis *et al*, 1989) was isolated from the *lacz* fusion cassette vector pMC1871 (Shapira *et al*, 1983). This vector does not contain the *neo* gene, which reduces the probability of probe cross contamination, which could confuse the end flanking sequence results. Both fragments were recovered according to the methodology described in Materials and Methods (Chapter 2.15).

B.3 Genomic organisation of enhancer trap cell-line D3-240.

Genomic DNA from this particular enhancer trap positive ES cell-line was extracted by the method described in Chapter 2.8.3 of this thesis, using the Nucleon genomic DNA extraction kit (Scotlab) and restricted with the endonucleases which were known to cut the enhancer trap vector.

10 µg aliquots of D3-240 genomic DNA were restricted with each of the enzymes *Eco*RI, *Bam*HI, *Kpn*I, *Nco*I and *Sal*I. The cut DNA was transferred and fixed to nylon membrane after gel

electrophoresis and probed with *lacz* and *neo* specific probes, to identify restriction sites in the sequences flanking the integration site.

The *neo* specific probe detected single unique fragments containing *neo* sequences for the restriction enzymes, *Eco*RI, *Kpn*I and *Bam*HI. The calculated sizes were 2 kbp, 3.2 kbp and 7.5 kbp respectively (Figure B.2.A). The restriction enzyme *Nco*I, in addition to the enhancer trap *Nco*I internal fragment (5.5 kbp), detected a flanking restriction fragment of 3.5 kbp (Figure B.2.A). There was no *Sal*I fragment detected with the *neo* probe.

Using the *lacz* specific probe, unique flanking restriction sites in the endogenous mouse genome were detected for the enzymes, *Eco*RI, *Kpn*I and *Sal*I. The detected fragment sizes were 10 kbp, 9.5 kbp and 9.0 kbp respectively (Figure B.2.B). *Bam*HI and *Nco*I detected internal restriction fragments, 3.5 and 5.5 kbp respectively (Figure B.2.B). A summary of these results is shown in diagrammatic form in Figure B.3.

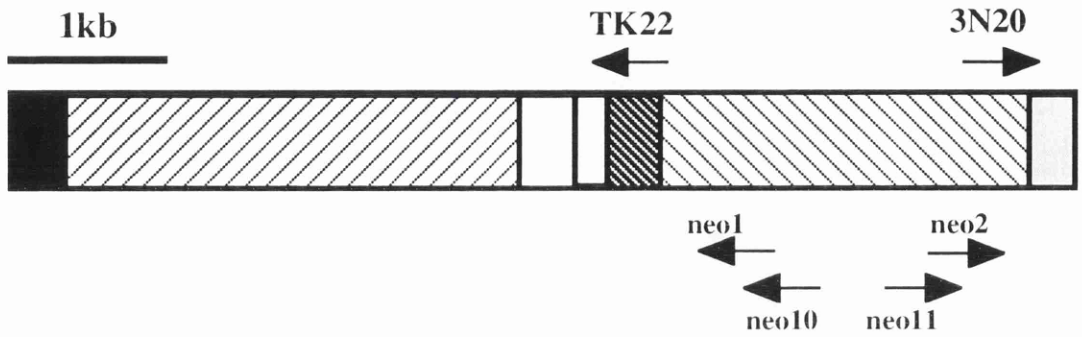
B.4 Isolation of *Eco*RI flanking sequences adjacent to the *neo* gene of the vector pLSN insertion in ES cell-line D3-240.

On the basis of the initial genomic characterisation of the insertion site of enhancer trap ES cell-line D3-240, the nearest flanking restriction site in the mouse genomic DNA at the *neo* gene end of the enhancer trap construct was determined to be *Eco*RI. It was calculated to be 0.5 kbp from the 3' end of the enhancer trap insertion site. This was selected for cloning flanking sequences by the IPCR method (Ochman *et al*, 1990; Chapter 2, Section 2.19).

Figure B.1

pLSN

(enhancer trap vector)



hsp promoter

SupF

thymidine kinase promoter

lacz

neomycin phoshotransferase

SV40 polyadenylation signals

PCR primers used to amplify flanking sequences (*neo*) and prime DNA sequencing (TK22 and 3N20).

Figure B.1 Diagrammatic representation of the enhancer trap vector pLSN.

This enhancer trap vector was constructed by Gossler *et al*, (1989). It contains the *E. coli lacz* gene (as the reporter gene), with a mouse minimal hsp68 promoter; an *E. coli SupF* amber suppressor tRNA gene, for use in plasmid rescue experiments; the Tn5 neomycin phosphotransferase gene (*neo*) as positive selectable marker immediately downstream of a thymidine kinase constitutive promoter. Both genes (*lacz* and *neo*) have SV40 derived poly-adenylation mRNA termination sequences.

Figure B.2

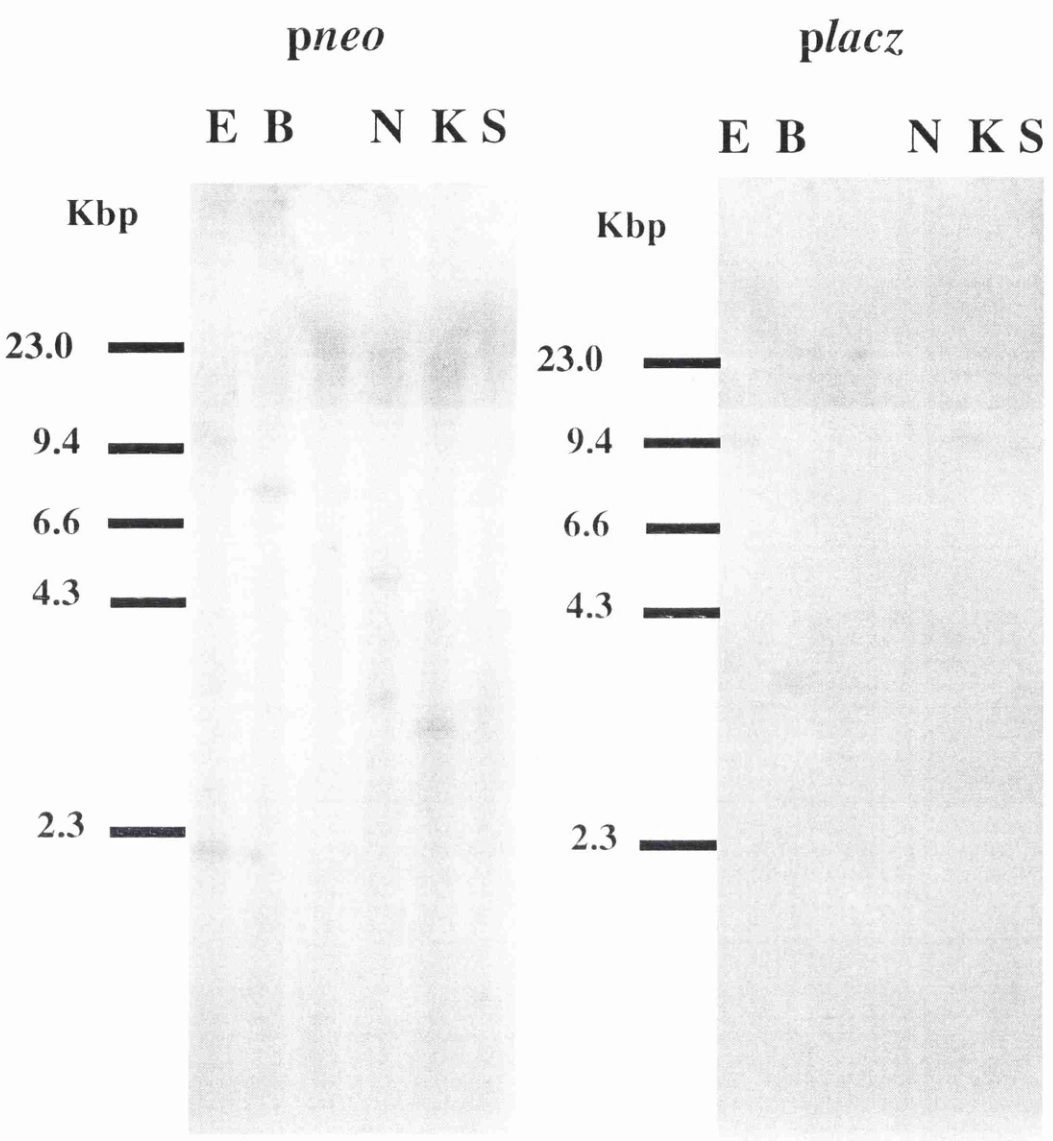
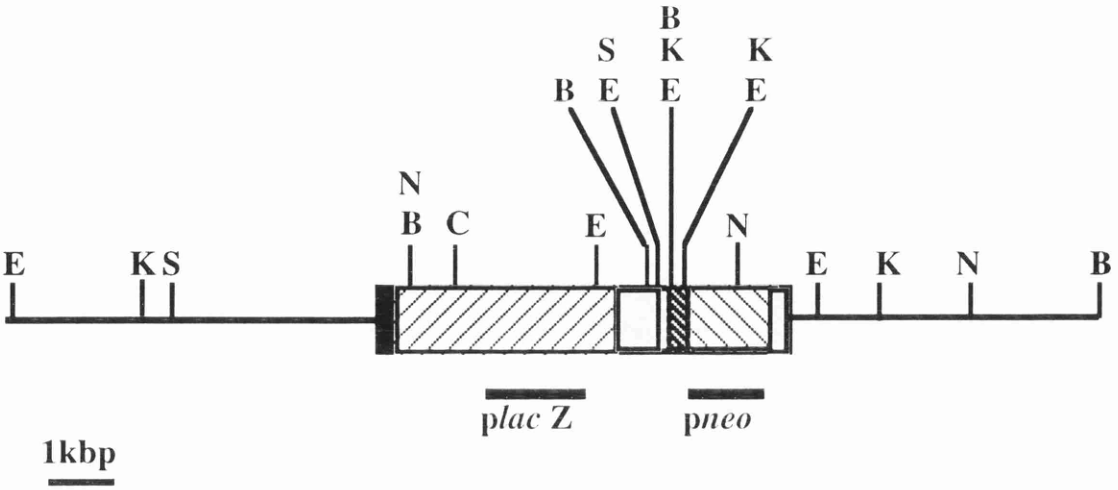


Figure B.2 Southern blot analysis of the enhancer trap integration site of ES cell-line D3-240.
10 µg of genomic DNA extracted from cell-line D3-240 was digested with restriction endonucleases, *Eco*RI (E); *Bam*HI (B); *Nco*I (N); *Kpn*I (K) and *Sal*I (S). The blots were probed with A) *lacz* specific probe and B) *neo* specific probes.

Figure B.3



hsp promoter

supF

thymidine kinase promoter

lac Z

neomycin phosphotransferase

SV40 polyadenylation signals

Figure B.3 Enhancer trap pLSN insertion site of cell-line D3-240 deduced by Southern blot analysis.

Flanking restriction sites were obtained from the combined data of the Southern blot analysis shown in Figure B.2. p*lacz* and p*neo* indicate the positions of the *lacz* and *neo* sequence specific probes, relative to the enhancer trap. The enzymes indicated are: *Eco*RI (E); *Bam*HI (B); *Cla*I (C): *Nco*I (N); *Kpn*I (K) and *Sal*I (S).

B.4.1 Amplification of flanking sequences of D3-240 by inverse PCR.

Inverse PCR was used to amplify endogenous mouse sequences which are adjacent to the *neo* gene at the D3-240 insertion site. Outwardly directed nested primer pairs (shown in Figure B.1), neo1 and 2 and neo10 and 11 were used in two sequential rounds of PCR reactions. The primers neo10 and neo11 were used in the first PCR reaction and the second PCR reaction used neo1 and neo2 primers on a 1/1000th dilution of the first round PCR reaction.

A single 1050 bp fragment was detected after the second round of PCR (Figure B.4). A combination of size of the PCR fragment and the non-amplified sequences (0.85 kbp) gives an estimated size of the template as 1.9 kbp. The circularised PCR template should correspond to the size of the *Eco*RI fragment detected by Southern blot; this was calculated as 2.0 kbp. This is consistant with the idea that the amplified sequences were derived from this *Eco*RI fragment.

B.4.2 Cloning of the D3-240 specific IPCR product.

For cloning purposes, three separate second round PCR reactions were pooled and treated with Promega Wizard DNA clean-up system and the recovered DNA was re-suspended to approximately 50 ng per 3 µl. Recovered amplified DNA was mixed in separate reactions with two types of T-overhang vectors. pBluescript KS-T and pGem-T, were used in the ligations at molar end ratio of 3:1 insert to vector.

Figure B.4

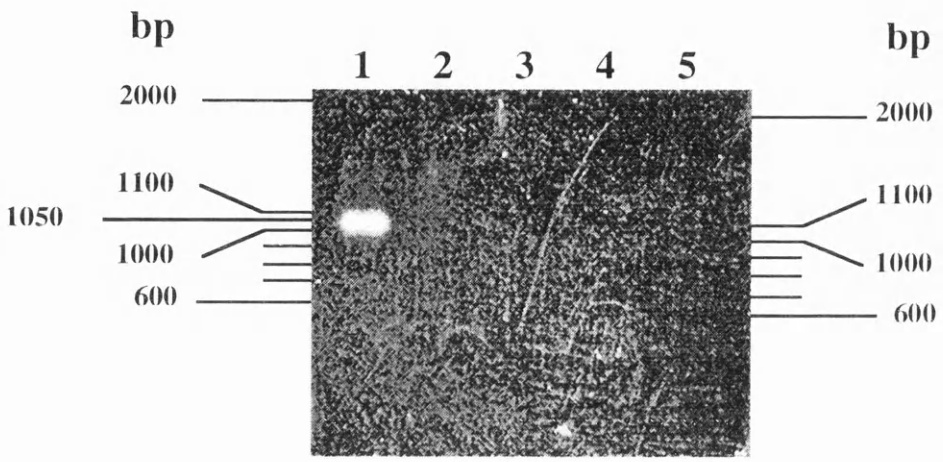


Figure B.4 Amplification of sequences flanking the *neo* gene of the enhancer trap positive cell-line D3-240.
9 µl of 25 µl of second round PCR reactions using combinations of primers were loaded on a 1% agarose gel. Template DNA for PCR was 1 µl of a 1/1000 fold dilution of the first round PCR reaction 1) primers neo1 and neo2 (using primers neo10 and neo11). 2) no primers control; 3) neo1 primer only; 4) neo2 primer only; 5) control without DNA.

DH5α *E. coli* cells were transformed by electroporation with the ligated plasmid and plated on the LB plate containing ampicillin, X-gal and IPTG.

Five positive white colonies were observed from the pBluescript KS- based T-vector and five from the pGem based vector, GEM-T. These were picked and grown in liquid LB in the presence of the appropriate antibiotic markers and grown for 16 hours. DNA was recovered from these bacterial cultures by using the Wizard plasmid DNA miniprep system (Promega). Test digestion of these plasmids with restriction enzymes that would release the captured insert, revealed that all 5 inserts of the pBluescript KS- T-vector had identical 1050 bp inserts. This is the same size as the second round PCR product (Figure B.5).

B.4.3 Sequencing of KS-T and GEM-T 240 IPCR inserts.

To establish if sequences contained within the insert derived from the IPCR reaction were similar to any other sequence in the databases, partial DNA sequencing was undertaken to retrieve insert specific sequences to use for nucleic acid database searches. The sequencing was conducted using the materials as described in Chapter 2 (Section 2.23.1), according to the manufacturers recommendations (Pharmacia). The primers TK22 and 3N20 (Chapter 2, Table 2.3 and Figure B.1 this Chapter) were used to prime the reactions. These primers were designed to minimise the vector specific flanking sequence which were required to be read through before coming across unique flanking sequence. Both primers were designed from sequences which were expected to occur in the cloned IPCR product.

Sequences from both 3N20 and TK22 primed DNA sequencing reactions were recovered and were used to search the Genbank and EMBL database using the FASTA search algorithm (Lipman and Pearson, 1985). From this it was found that sequence derived from 3N20 primed reactions on the KS-T240 plasmid containing the PCR insert (clone 3; see Figure B.5, lane 4) was identical to the 3' UTR of the rabbit β-globin gene (with respect to rabbit β-globin sequences (AC J00660) nucleotides 551-604 inclusive; see Figure B.5.A). Similar data was recovered from the other clones sequences (data not shown).

DNA sequencing, using the TK22 primer which is complementary to the thymidine kinase promoter sequences (TK22, Figure B.1 and Table 2.3), which drives *neo* gene expression in the enhancer trap vector, pLSN (Gossler *et al*, 1989), identified a small fragment of sequence identical to 5' of the thymidine kinase promoter sequences. These are present in the pLSN transgene (Gossler *et al*, 1989) and should also be present in the IPCR product. The remaining sequence identified showed similarity to the rabbit β-globin sequences. Data upstream of the thymidine kinase homologous fragment was found to be identical to the rabbit β-globin gene (Figure B.6.B). This sequence corresponded to sequences 3' in relation to the rabbit β-globin homologous sequences identified by the 3N20 primed DNA sequencing reactions (TK22 primed reaction; nucleotides 970-1150 inclusive of the rabbit β-globin sequence, AC: J00660 and 3N20 primed DNA sequencing reaction, nucleotides inclusive of the rabbit sequence AC: J00660). The relationship between identified sequences is represented in Figure B.6.C.

Figure B.5

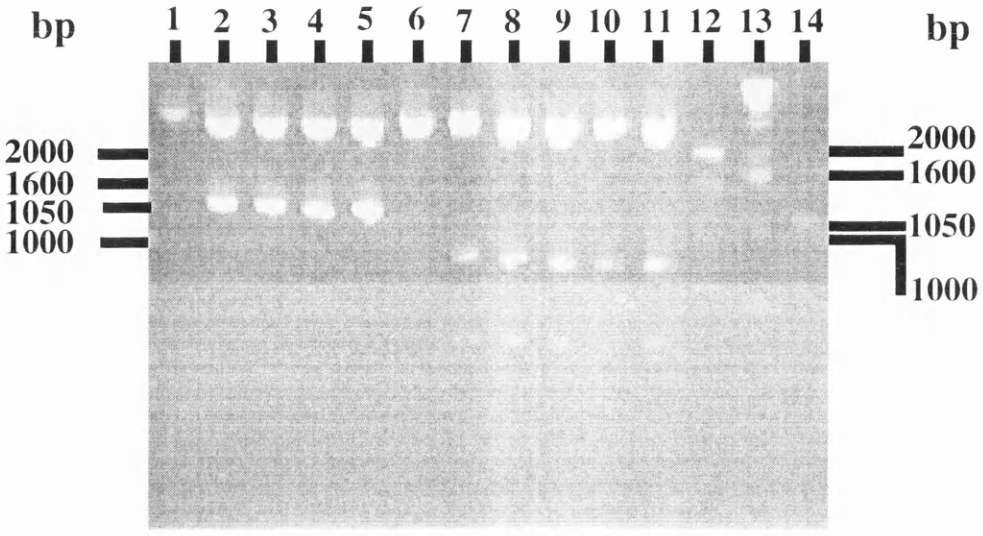


Figure B.5 Analysis of sub-clones derived from ligation of second round PCR product ligated into T-vectors pBluescript KS-T and pGem-T.

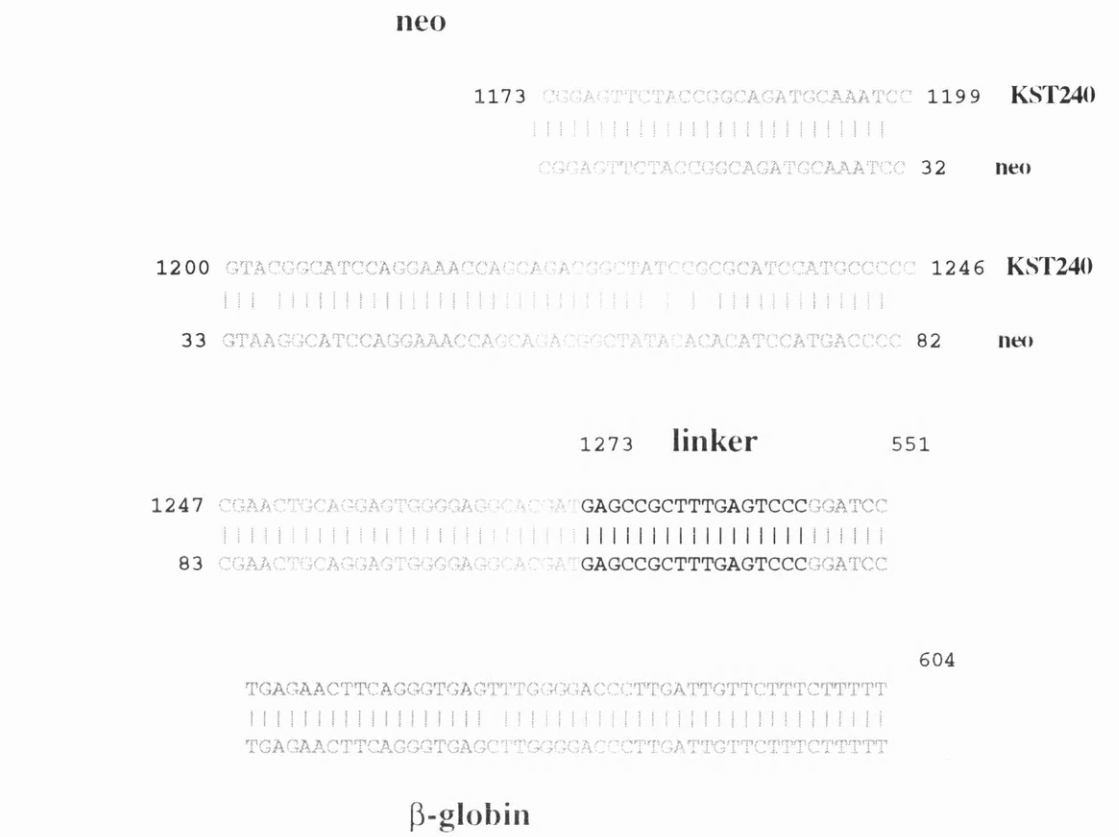
4 µl of plasmid DNA recovered by Promega Wizard miniprep protocol from 3 ml of overnight bacterial cultures was digested with the restriction enzymes, *Kpn*I and *Sma*I (KS-T DNA only) and *Nco*I and *Sal*I (pGem-T DNA only). 10 µl of the digests were then electrophoresed. Lane 1) pBluescript KS- cut with *Kpn*I and *Sma*I; lanes 2) to 5), KS-T clones 2, 3, 5 and 6 digested with enzymes *Kpn*I and *Sma*I; lane 6) pGem-T digested with the restriction enzymes *Nco*I and *Sal*I; lanes 7) to 11), pGem-T 240 clones, 8,9,10,11 and 13 digested with the restriction enzymes *Nco*I and *Sal*I; Lanes 12) 500 ng of 100 bp double stranded DNA marker ladder (BRL); 13). 500 ng of 1 kbp double stranded DNA marker ladder (BRL); lane 14). 3 µl of the second round PCR reaction using primers neo1 and neo 2.

Legend to Figure B.6 Sequence alignment with sequences derived from KS-T-240 clone 3, using primers A) 3N20 and B) TK22.

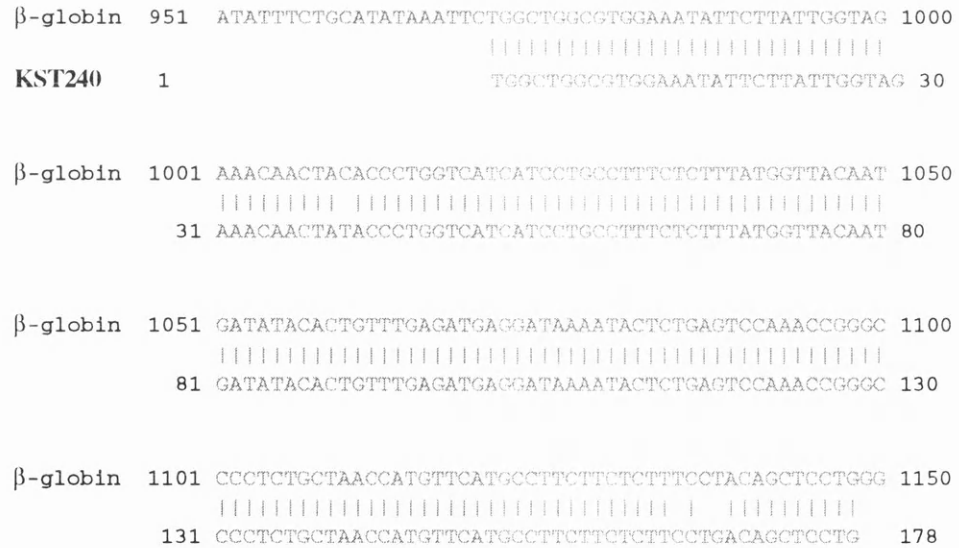
A) Alignment of sequences from KS-T-240, clone 3, using primer 3N20 and rabbit β-globin sequences and Tn5 *neo* sequences (AC: J01834; Beck *et al*, 1982). B) Alignment of sequences from KS-T-240 clone 3, using primer TK22, and rabbit β-globin sequences (AC: J00660). C) Diagrammatic representation of the relationship of sequences derived from TK22 and 3N20 primed DNA sequence reactions performed on clone 3, relative to the structure of KS-T240 clone 3 insert. Arrows indicate the primers; β-globin represents sequences identical to rabbit β-globin sequences (indicated in purple coloured text); tk represent thymidine kinase promoter sequences (blue coloured text) and *neo* (green text) indicates neomycin phosphotransferase specific sequences. The dashed lines indicates unidentified sequence information from the insert of clone 3. Colours are coordinated in the sequence (A and B) and in the diagrammatic representation (C).

Figure B.6

A)


neo
1173 CGGAGTTCTACCGGCAGATGCAAATCC 1199 KST240
CGGAGTTCTACCGGCAGATGCAAATCC 32 neo
1200 GTACGGCATCCAGGAAACCAGCAGACGGCTATCCGCGCATCCATGCCCCC 1246 KST240
33 GTAAGGCATCCAGGAAACCAGCAGACGGCTATACACACATCCATGACCCC 82 neo
1273 linker 551
1247 CGAACTGCAGGAGTGGGGAGGCACGATGAGCCGCTTTGAGTCCCGGATCC
83 CGAACTGCAGGAGTGGGGAGGCACGATGAGCCGCTTTGAGTCCCGGATCC
604
TGAGAACTTCAGGGTGAGTTTGGGGACCCTTGATTGTTCTTTCTTTTTT
TGAGAACTTCAGGGTGAGCTTGGGGACCCTTGATTGTTCTTTCTTTTT
β-globin

B)


β-globin 951 ATATTTCTGCATATAAATTCTGGCTGGCGTGGAAATATTCTTATTGGTAG 1000
KST240 1 TGGCTGGCGTGGAAATATTCTTATTGGTAG 30
β-globin 1001 AAACAACTACACCCTGGTCATCATCCTGCCTTTCTCTTTATGGTTACAAT 1050
31 AAACAACTATACCCTGGTCATCATCCTGCCTTTCTCTTTATGGTTACAAT 80
β-globin 1051 GATATACACTGTTTGAGATGAGGATAAAATACTCTGAGTCCAAACCGGGC 1100
81 GATATACACTGTTTGAGATGAGGATAAAATACTCTGAGTCCAAACCGGGC 130
β-globin 1101 CCCTCTGCTAACCATGTTCATGCCTTCTTCTCTTTCCTACAGCTCCTGGG 1150
131 CCCTCTGCTAACCATGTTCATGCCTTCTTCTCTTCCTGACAGCTCCTG 178

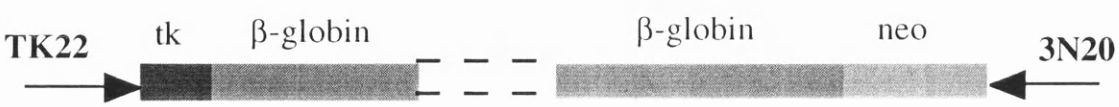

TK22
tk
β-globin
β-globin
neo
3N20

Figure B.7

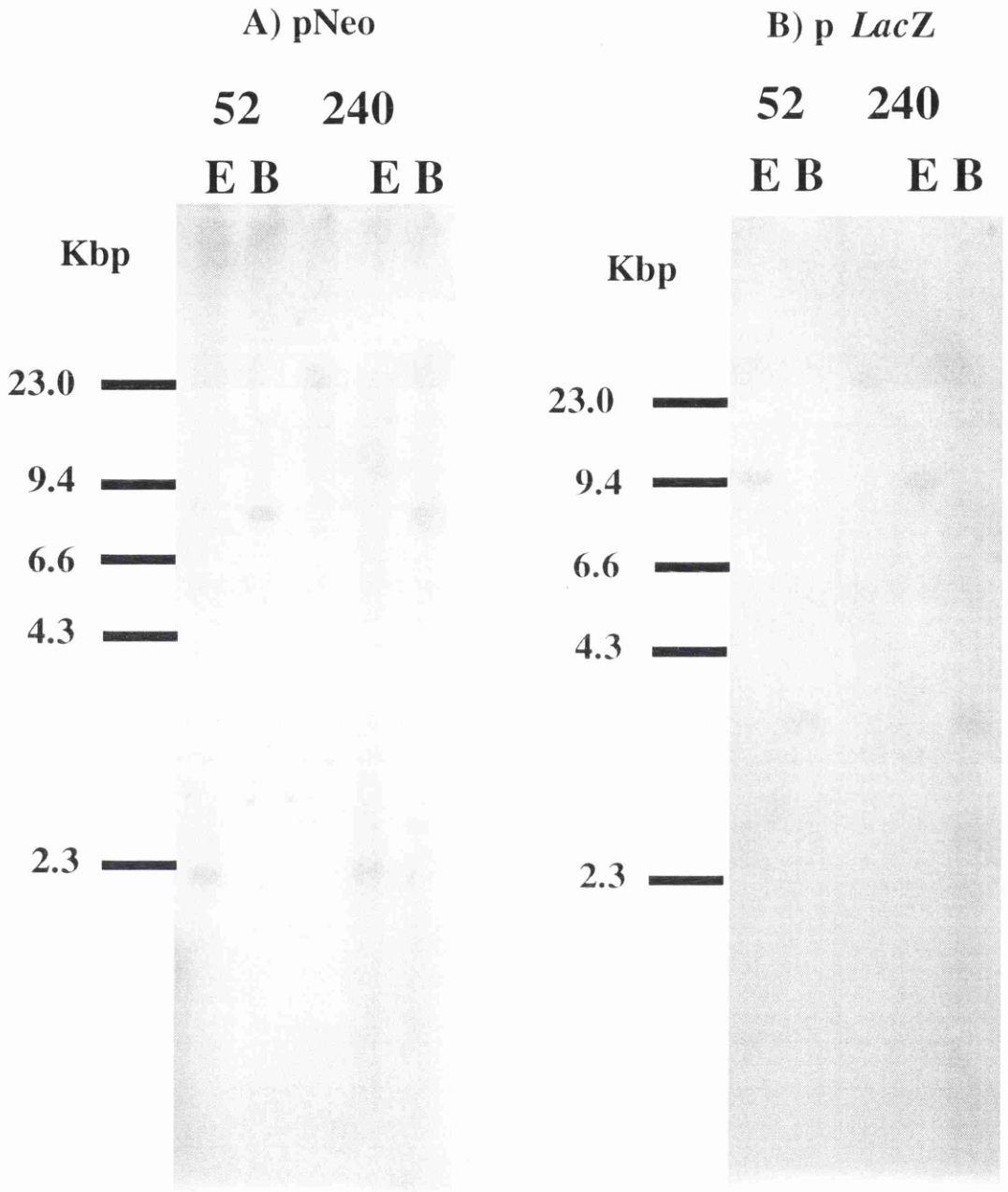


Figure B.7 Southern analysis on two separate genomic DNAs derived from two separate cell lysates.
10 µg of genomic DNA was extracted from the cell lysates D3-052 and D3-240 in the manner described in Chapter 2 and digested with the restriction endonucleases, E (*Eco*RI) and B (*Bam*HI) and probed with A) *lacz* and B) *neo* gene specific probes.

B.5 Southern analysis of genomic DNA derived from cell lysates of enhancer trap lines, D3-052 and D3-240.

Genomic DNA derived from line specific ES cell lysates obtained from Achim Gossler was extracted by the method described in Chapter 2, section 2.8.3. From Figure B.7 it can be observed that in genomic DNA from independent ES cell lines D3-052 and D3-240, fragments of the same size for the restriction enzymes *Eco*RI and *Bam*HI are observed. This is true for when probed with a *neo* specific random primed probe (1kbp *Xba*I-*Xba*I fragment of plasmid incorporating the gene-trap vector PGT1, Gossler *et al*, unpublished data) or *lacz* specific random primed probe (1.8 kbp *Hinc*II-*Hinc*II fragment of plasmid pMC1871, Shapira *et al*, 1983). An *Eco*RI fragment of 2.0 kbp and 7.5 kbp *Bam*HI were detected in the genomic DNA digests using the *neo* specific probe (Figure B.7.A). Likewise 9.5 kbp *Eco*RI and 3.5 kbp *Bam*HI fragments were detected in genomic DNA derived from both the cell lysates, when probed with *lacz* specific sequences (Figure B.7.B).

B.6 Discussion.

The identification of novel genes by the analysis of enhancer trap integrations, has been used successfully by Korn *et al* (1992). The main feature of this approach is the identification of specific expression patterns of a reporter gene driven by endogenous regulatory sequences. The native gene(s) which is (are) the target of these regulatory elements is presumed to be in the vicinity of the insertion site. As a preliminary step to identification of the gene,

cloning of the unique sequences from around the integration site is required to start a search of the surrounding genomic region. This can be achieved by a number of ways such as lambda library construction from the genomic DNA of that particular enhancer trap cell-line and identification of clones which contain insertion site sequences (which are known and characterised). Library construction is time consuming. There are PCR based methods which can and have been used to isolate flanking regions.

B.6.1 Inverse PCR.

Korn *et al* (1992) have already used the inverse PCR (IPCR) procedure to clone and identify mouse endogenous flanking sequences, using outwardly directed PCR primers. IPCR is dependent on the availability of suitable restriction enzyme sites in the endogenous mouse DNA which do not cut between the PCR primer pairs and are not so distant from the insertion site as to reduce the efficiency of the PCR reaction directed from the circularised product of ligation.

IPCR from line D3-240 was carried out and a specific PCR product was detected. However, sequence analysis revealed that the insert of this cloned product had 97.75% identity to the 3' UTR region of the rabbit β-globin gene, in addition to the expected residual *neo* and thymidine kinase promoter sequences derived from the vector. This can have came about in several ways. An obvious point to consider is that a DNA contamination had occurred. This event may have been possible at multiple stages. First to consider is that there has been a contamination of the cell-line with the vector containing the pLSN enhancer trap itself. This is unlikely

because: 1) The vector construct in which pLSN is maintained does not contain a eukaryotic origin of replication (but does contain the plasmid origin of replication; Gossler *et al*, 1989). 2) Personal communications with A. Gossler revealed that the pLSN vector does not contain β-globin poly-adenylation signal sequences, but instead contains SV40 derived poly-adenylation signal sequences.

Alternatively we may be dealing with an enhancer trap integration which has retained the flanking *Eco*RI restriction site which is present in the parent plasmid polylinker. However this is also unlikely because β-globin derived poly-adenylation signal sequences are not present in the pLSN enhancer trap vector (A. Gossler, pers. comm.).

Another explanation to consider is that an air-borne DNA containing both *neo* sequences and the rabbit β-globin sequences may have contaminated the PCR reaction (or genomic DNA derived from the cell-lysates). This is a possibility because of the existence of vectors which have utilised both *neo* gene sequences as a reporter gene and the 3' UTR sequences of the rabbit β-globin gene as poly-adenylation signal sequences. These have been in use, all be it many years ago, in the department in which this work was conducted. The β-globin sequences are used in connection with the poly-adenylation signal sequences which are used to terminate transcription from expression vectors. An example of this type of vector would be pSV2tkneoβg (Okayama and Berg, 1983). This vector contains a *neo* gene which has at its 3' extremity, sequences from intervening sequence 2 (IVS or intron 2) of the rabbit β-globin gene. These sequences are represented in the sequence derived from the sequencing experiments described in this Chapter (see Figure B.6). Furthermore, a hypothetical restriction of this vector

and its parent plasmid vector backbone with *Eco*RI reveals that the restriction product containing the *neo* gene and rabbit β-globin gene sequences is 1.9kb in size. This is similar in size to the *Eco*RI genomic fragment detected in cell-line D3-240 by a *neo* sequence specific probe, (as described in this Chapter; see Section B.2.3 and Figure B.2). This vector (or a closely related derivative) may represent the contamination in this experiment. However this vector does not contain the *lacZ* gene and the remaining genomic DNA Southern blot restriction fragment data is consistent with a pLSN enhancer trap integration. For example the expected fragments internal to the pLSN vector, are detected (see Section B.3).

A contamination may have been introduced at other points in the experimental process. For example the plasmid mentioned above, pSV2tkneoβg (Nicolas and Berg, 1983), contains an SV40 replication origin and this may have been introduced into the cell-line (at sometime in the past) and would have been continually propagated as a replicon by the constant presence of neomycin or an analogue (*e.g.* G418) as a selection agent. Therefore in a genomic DNA isolation from a cell-line, a plasmid contamination would already be present. There is no way to derive the origin of this contamination and it is an unfortunate aberration. This contributed to the abandonment of this particular line of research.

B.6.2 One cell-line?

Another point to consider was the discovery from genomic Southern blot experiments that the same restriction fragments were observed for genomic DNA isolated from lysates from two different

enhancer trap cell lines using the *neo* specific probe. The genomic DNA from lysates from cell lines D3-240 and 52 were observed to have fragments sized at 2 kbp and 7.5 kbp for *Eco*RI and *Bam*HI respectively (Figure B.7). A similar result was observed using the *lacz* specific probe. A 9.5 kbp *Eco*RI fragment and a 3.5 kbp *Bam*HI fragment were detected both for D3-052 and D3-240 (Figure B.7). From this it was concluded that these cell-lysates contained genomic DNA with the same insertion site. This may have occurred due to contamination by a cell-line which over-grew the original cell lines. Alternatively an operator error may have occurred and certain lysates were mis-labelled.

The final piece of evidence which also persuaded me to discontinue this particular line of research was that a student in A. Gossler's laboratory who was working on an enhancer trap cell-line that was supposed to give an expression pattern distinct from that of D3-240, cloned the insertion site flanking sequences of the ETS from that cell-line by IPCR and checked that the sequences were novel. A new gene was identified in the proximity of the insertion site. Using these novel gene sequences, a *in situ* hybridisation experiment identified the expression pattern of this gene. It was found that this gene had a similar expression profile in developing mouse embryo to the original transgene expression pattern described for the enhancer trap positive cell-line D3-240. This supports the idea that the same cell-line, in some way contaminated other ETS cell-line isolates.

It was possible that IPCR could have been attempted, using other restriction fragments. For example the flanking *Kpn*I and *Bam*HI restriction sites (Figure B.3) at the *neo* "end" of the pLSN vector integration could have been used to provide a template for

IPCR, using the *neo* primers. Alternatively, *Eco*RI, *Kpn*I and *Sal*I restriction sites (Figure B.3) were suitably positioned in the unique flanking sequences, beyond the *lacz* "end" of the insertion site (Figure B.3), using *lacz* sequence specific primers used by Korn *et al* (1992). This was deemed as unproductive as it was evident that I was dealing with multiple uncertainties in this experiment, some of which could not be overcome by repeating the approach already described in this Chapter *i.e.* the apparent identity in the restriction fragments detected by Southern blot analysis of genomic DNA from supposedly different cell lines.

Another route which could be used was that using the *E. coli, SupF* amber repressor tRNA gene, to rescue DNA fragments that contained the integration site. This requires the preparation of a lambda (or plasmid) library from the source DNA. Once again, the uncertainty surrounding the source of the DNA which exhibited the same integration pattern, made this line of work unattractive.

B.6.3 Final remarks.

Being unable to find source material to continue this project left me with an important decision. This was whether to continue with this research or identify an alternative approach to the identification of new developmentally expressed genes. The routes available to do this are outlined in Chapter 1 (Section 1.1 to 1.4; see also Figure 1.1). I chose a homologous screening approach in deciding to attempt the isolation of anonymous genes from mouse embryonic cDNA libraries which contained CAG trinucleotide repeats. This was the focus of the rest of the work in this thesis.

Abbreviations.

Measurements

A	ampere
bp	base pair
Ci	curie
cm	centimetre
cM	centimorgan
cpm	counts per minute
°C	degrees centigrade
dpm	disintegrations per minute
g	gram
kb	kilobase
kbp	kilobase pair
kDa	kiloDalton
l	litre
LB	pound
M	molar
mg	milligram
min	minute
ml	millilitre
mm	millimetre
mM	millimolar
ng	nanogram
rpm	revolutions per minute
s	seconds
sq. in.	square inch
µCi	microcurie
µF	microFarad
µg	microgram
µl	microlitre
V	volt

Abbreviations and symbols.

&	and
A	amp or Alanine
AC	Accession number (database identification)
AR	androgen receptor
amp	ampicillin
ATP	adenosine 5'-triphosphate
BLAST	Basic Local Alignment Search Tool
cDNA	complementary DNA
CNS	central nervous system

Abbreviations and symbols (continued).

dATP	deoxyadenosine 5'-triphosphate
dCTP	deoxycytosine 5'-triphosphate
DDBJ	DNA data bank of Japan
ddNTP	dideoxynucleoside 5'-triphosphate
dGTP	deoxyguanosine 5'-triphosphate
dNTP	deoxynucleoside 5'-triphosphate
dTTP	deoxythymidine 5'-triphosphate
DM	myotonic dystrophy
DMSO	dimethyl sulphoxide
DMF	dimethyl formaldehyde
DNA	deoxribonucleic acid
DNAse	deoxyribonuclease
BSA	bovine serum albumin
dpc	days post coitum
DTT	dithiothreitol
DyedNTP	dye labelled dideoxynucleoside 5'-triphosphate.
EDTA	ethylenediaminetetraacetic acid (disodium salt)
e.g.	for example
EMBL	european molecular biology laboratory
ES	embryonic stem
EST	expressed sequence tag
EtBr	ethidium bromide
ExoIII	exonuclease III
ftp	file transfer protocol
UWGCG	university of Winsconsin, genetics computer group
γ-ATP	gamma phosphate labelled phosphate deoxyadenosine 5'-triphosphate
GCG	genetics computer group
GAPdh	glyceraldehyde-3-phosphate dehydrogenase
HSV	herpes simplex virus
HD	Huntingtons disease
IPTG	isopropyl-β–D-thiogalacto-pyranoside
λ	lambda
LMP	low melting point
lacZ	lactose operon (of the bacteria *E. coli*), gene Z
kb	kilobase
kbp	kilobase pair
mRNA	messenger ribonucleic acid
neo	neomycin
nt	nucleotide
O.D.	optical density
ORF	open reading frame
IPCR	inverse PCR.

Abbreviations and symbols (continued).

pA	polyadenylation site
^{32}P	phosphorous, isotope 32
PCR	polymerase chain reaction
p.f.u.	plaque forming unit
PIR	protein identification resource
RNase	ribonuclease
^{35}S	sulphur, isotope 35
SDS	sodium dodecyl sulfate
SM	suspension medium
SV40	simian virus 40
Swiss-Prot	protein sequence databank of Switzerland
TEMED	N,N,N'N',-tetramethylethylenediamine
TK	thymidine kinase
U	units
UTR	untranslated region
U.V.	ultraviolet light
vol	volume
v / v	volume per volume
w / v	weight per volume
W	watt (joule/second)
X-gal	5-bromo-4-chloro-3-indolyl β-D-galactoside

Bibliography

Abbot C., Chambers D. (1994). Analysis of CAG trinucleotide repeats from mouse cDNA sequences. Annals of Human Genetics **58**, 87-94.

Abraham S.E., Lobo S., Yaciuk P., Wang H.G.H., Moran E. (1993). P300, and P300-associated proteins, are components of TATA-binding protein (TBP) complexes. Oncogene **8**, 1639-1647.

Allen N.D., Keverne E.B., Surani M.A. (1990). A position-dependent transgene reveals patterns of gene expression in the developing brain. Developmental Brain Research **55**, 181-190.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology **215**, 403-410.

Anderson M.L.M., Young B.D. (1990). Quantitative filter hybridisation. Nucleic acid hybridisation: a pratical approach (eds. Hames B.D., Higgins S.J.) **1**, 73-111. Press Limited, Oxford, London.

Arany Z., Newsome D., Oldread E., Livingstone D.M., Eckner R. (1995). A family of transcriptional adaptor proteins targets by E1A oncoprotein. Nature **374**, 81-84.

Arning S., Gruter P., Bilbe G., Kramer A. (1996). Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. RNA **2**, 794-810.

Ashley C.T., Warren S.T. (1995). Trinucleotide repeat expansion and human-disease. Annual Review of Genetics **29**, 703-728.

Ausubel F.M., Brent R., Kingston R.E., Moore D.D., Seidman J.G., Smith J.A., Struhl K. (Eds). (1990). Current Protocols in Molecular Biology.

Aylward E.H., Brandt J., Codori A.M., Mangus R.S., Barta P.E., Harris G.J. (1994). Reduced basal ganglia volume associated with the gene for Huntington's disease in aymptomatic at-risk persons. Neurology **44**, 823-828.

Bakker C.E., Verheij C., Willemsen R., Vanderhelm R., Oerlemans F., Vermey M., Bygrave A., Hoogeveen A.T., Oostra B.A., Reyniers E., Deboulle K., Dhooge R., Cras P., Vanvelzen D., Nagels G., Martin J.J., Dedeyn P.P. (1994). Fmr1 knockout mice - a model to study Fragile-X mental retardation. Cell **78**, 23-33.

Banfi S., Servadio A., Chung M.Y., Capozzoli F., Duvick L.A., Elde R., Zoghbi H.Y., Orr H.T. (1996). Cloning and developmental expression analysis of the murine homolog of the spinocerebellar ataxia type 1 gene (*Sca1*). Human Molecular Genetics **5**, 33-40.

Barnes G.T., Duyao M.P., Ambrose C.M., Mcneil S., Persichetti F., Srinidhi J., Gusella J.F., Macdonald M.E. (1994). Mouse Huntingtons disease gene homolog (*hdh*). Somatic Cell and Molecular Genetics **20**, 87-97.

Bassett A., Honer W.G. (1994). Evidence of anticipation in schizophrenia. American Journal of Human Genetics **54**, 864-870.

Beck E., Ludwig G., Auerswald E.A., Reiss B. and Schaller H. (1982). Nucleotide sequence and exact localization of the neomycin phosphotransferase gene from transposon Tn5. Gene **19**, 327-336.

Beckmann J.S., Weber J.L. (1992). Survey of human and rat microsatellites. Genomics **12**, 627-631.

Bingham P.M., Scott M.O., Wang S.P., Mcphaul M.J, Wilson E.M., Garbern J.Y., Merry D.E., Fischbeck K.H. (1995). Stability of an expanded trinucleotide repeat in the androgen receptor gene in transgenic mice. Nature Genetics **9**, 191-196.

Blackwell L.J., Borowiec J.A. (1994). Human replication protein A binds single-stranded DNA in two distinct complexes. Molecular and Cellular Biology **14**, 3993-4001.

Boncinelli E., Somma R., Acampora D., Pannese M., Desposito M., Faiella A., Simeone A. (1988). Organization of human homeobox genes. Human Reproduction **3**, 880-886.

Boucher C.A., King S.K., Carey N., Krahe R., Winchester C.L., Rahman S., Creavin T., Meghji P., Bailey M.E.S., Chartier F.L., Brown S.D., Siciliano M.J. Johnson K.J. (1995). A novel homeodomain-encoding gene is associated with a large CpG island interupted by the Myotonic Dystrophy unstable repeat. Human Molecular Genetics **4**, 1919-1925.

Bowater R.P., Rosche W.A., Jaworski A., Sinden R.R., Wells R.D. (1996). Relationship between *Escherichia coli* growth and center dot CAG triplet repeats in plasmids. Journal of Molecular Biology **264**, 82-96.

Brahmachari S.K., Meera G., Sarkar P.S., Balaguramoarthy P., Tripathi J., Raghavan S., Shaligram U., Pataskar S. (1995). Simple repeat sequences in the genome: structure and functional significance. Electrophoresis **16**, 1705-1714.

Brook J.D., McCurrach M.E., Harley H.G., Buckler A.J., Church D., Aburatani H., Hunter K., Stanton V.P., Thirion J.P., Hudson T., Sohn R., Zemelman B., Snell R.G., Rundle S.A., Crow S., Davies J., Shelbourne P., Buxton J., Jones C., Juvonen V., Johnson K., Harper P.S., Shaw D., Housman D.E. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell **68**, 799-808.

Bullock W.O., Fernandez J.M., Short J.M. (1987). XI1-blue: a high-efficiency plasmid transforming RecA *Escherichia coli* strain with β-galactosidase selection. Biotechniques **5**, 376.

Burright E.N., Clark H.B., Servadio A., Matilla T., Feddersen R.M., Yunis W.S., Duvick L.A., Zoghbi H.Y., Orr H.T. (1995). *Sca1* transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. Cell **82**, 937-948.

Campuzano V., Montermini L., Molto M.D., Pianese L., Cossee M., Cavalcanti F., Monros E., Rodius F., Duclos F., Monticelli A., Zara F., Canizares J., Koutnikova H., Bidichandani S.I., Gellera C., Brice A., Trouillas P., De Michele G., Filla A., De Frutos R., Palau F., Patel P.I., Di Donato S., Mandel J.L., Cocozza S., Koenig M., Pandolfo M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science **271**, 1423-1426.

Carango P., Noble J.E., Marks H.G., Funanage V.L. (1993). Absence of myotonic dystrophy protein kinase (DMPK) mRNA as a result of a triplet repeat expansion in myotonic dystrophy. Genomics **18**, 340-348.

Chamberlain N.L., Driver E.D., Miesfeld L. (1994). The length and location of CAG trinucleotide repeats in androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Research **22**, 3181-3186.

Chambers D.M., Abbott C.M. (1996). Isolation and mapping of novel mouse brain cDNA clones containing trinucleotide repeats, and demonstration of novel alleles in recombinant inbred strains. Genome Research **6**, 715-723.

Chomczynski P., Sacchi N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry **162**, 156-159.

Choy B., Green M. (1993). Eukaryotic activators function during multiple steps of pre-initiation complex assembly. Nature **366**, 531-536.

Chrivia J.C., Kwok R.P.S., Lamb N., Hagiwara M., Montminy M.R., Goodman R.H. (1993). Phosphorylated CREB binds specifically to the nuclear protein CBP. Nature **365**, 855-859.

Clerc R.G., Corcoran L.M., Lebowitz J.H, Baltimore D., Sharp P.A. (1988). The B-cell-specific OCT-2 protein contains POU box-type and homeobox-type domains. Genes and Development **2**, 1570-1581.

Colicelli J., Nicolette C., Birchmeier C., Rodgers L., Riggs M., Wigler M. (1991). Expression of three mammalian cDNAs that interfere with RAS function in *Saccharomyces cerevisiae*. PNAS (USA) **88**, 2913-2917.

Courey A.J., Tjian R. (1988). Analysis of SP1 *in vivo* reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. Cell **55**, 887-898.

Courey A.J., Holtzman D.A., Jackson S.P., Tjian R. (1989). Synergistic activation by the glutamine-rich domains of human transcription factor SP1. Cell **59**, 827-836.

Coward P., Nagai K., Chen D., Thomas H.D., Nagamine C.M., Lau Y-F.C. (1994). Polymorphism of a CAG trinucleotide repeat within *Sry* correlates with B6.Y^{Dom} sex reversal. Nature Genetics **6**, 245-250.

Crick F.H.C., Watson J.D. (1954). The complementary structure of deoxyribonucleic acid. Proceedings of the Royal Society (London). **223**, 80-96.

Darlow J.M., Leach D.R.F. (1995). The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *E. coli* suggest hairpin folding preference *in vivo*. Genetics **141**, 825-832.

Dausset J., Cann H., Cohen D., Lathrop M., Lalouel J.M., White R. (1990). Program description - Center-Detude-du-Polymorphisme-Humain (CEPH) - collaborative genetic mapping of the human genome. Genomics, **6**, 575-577.

Deboulle K., Verkerk A.J.M.H., Reyniers E., Vits L., Hendrick J., Vanroy B., Vandenbos F., Degraaff E., Oostra B.A., Willems P.J. (1993). A point mutation in the FMR-1 gene associated with Fragile-X mental-retardation. Nature Genetics **3**, 31-35.

Doetschman T.C., Eistetter H., Katz M., Schmidt W., Kemler R. (1985). The *in vitro* development of blastocyst-derived embryonic stem cell lines- formation of visceral yolk-sac, blood islands and myocardium. Journal of Embryology and Experimental Morphology **87**, 27.

Devereux J., Haeberli P., Smithies O. (1984). A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Research **12**, 387-395.

Digan, M.E., Haynes, S.R., Mozer, B.A., Dawid, I.B., Forquignon, F., Gans, M., (1986). Genetic and molecular analysis of FS(1)H, a maternal effect homeotic gene in *Drosophila*. Developmental Biology **114**, 161-169.

Dornreiter I., Erdile L.F., Gilbert I.U., Vonwinkler D., Kelly T.J., Fanning E. (1992). Interaction of DNA polymerase-alpha primase with cellular replication protein A and SV40-T antigen. EMBO Journal **11**, 769-776.

Duboule D., Baron A., Mahl P., Galliot B. (1986). A new homeobox is present in overlapping cosmid clones which define the mouse hox-1 locus. EMBO Journal **5**, 1973-1980.

Duboule D., Haenlin M., Galliot B., Mohier E. (1987). DNA sequences homologous to the *Drosophila OPA* repeat are present in murine messenger RNAs that are differentially expressed in fetuses and adult tissues. Molecular and Cellular Biology **7**, 2003-2006.

Dunne P.W., Ma L., Casey D.L., Harati Y., Epstein H.F. (1996). Localization of myotonic dystrophy protein kinase in skeletal muscle and its alteration with disease. Cell Motility and The Cytoskeleton **33**, 52-63.

Duyao M.P., Auerbach A.B., Ryan A., Persichetti F., Barnes G.T., McNeil S.M., Ge P., Vonsattel J.P., Gusella J.F., Joyner A.L., Macdonald M.E. (1995). Inactivation of the mouse Huntingtons disease gene homolog *hdh*. Science **269**, 407-410.

Eckner R., Ewen M.E., Newsome D., Gerdes M., Decaprio J.A., Lawrence J.B., Livingston D.M. (1994). Molecular cloning and functional analysis of the adenovirus E1A-associated 300 kD protein (P300) reveals a protein with properties of a transcriptional adapter. Genes and Development **8**, 869-884.

Emery H.S., Schild D., Kellogg D.E., Mortimer R.K. (1991). Sequence of RAD54, a *Saccharomyces cerevisiae* gene involved in recombination and repair. Gene **104**, 103-106.

Erdile L.F., Heyer W.D., Kolodner R., Kelly T.J. (1991). Characterization of a cDNA encoding the 70kDa single-stranded DNA binding subunit of human replication protein A and the role of the protein in DNA replication. Journal of Biological Chemistry **266**, 12090-12098.

Faber P.W., King A., Vanrooij H.C.J., Brinkmann A.O., Deboth N.J., Trapman J. (1991). The mouse androgen receptor - functional analysis of the protein and characterization of the gene. Biochemical Journal **278**, 269-278.

Feddersen R.M., Ehlenfeldt R., Yunis, W.S., Clark H.B., Orr H.T. (1992). Disrupted cerebellar cortical dvevelopment and progressive degeneration of Purkinje cells in SV40 T antigen transgenic mice. Neuron **9**, 955-966.

Fesus L., Thomazy V., Falus A. (1987). Induction and activation of tissue transglutaminase during programmed cell death. FEBS Letter **224**, 104-108.

Fickett J.W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Research **10**, 5303-5318.

Folk J.E., Finlayson J.S. (1977). The γ-(ε-glutamyl)lysine crosslinks and the catalytic role of transglutaminases. Advances in Protein Chemistry. **31**, 1-33.

Frankel W.N. (1995). Of rats, mice and men. Nature Genetics **9**, 3-4.

Friedrich G., Soriano P (1991). Promoter traps in embryonic stem-cells- a genetic screen to identify and mutate developmental genes in mice. Genes and Development **5**, 1513-1523.

Fry M., Loeb L.A. (1994). The Fragile-X syndrome d(CGG)(n) nucleotide repeats form a stable tetrahelical structure. PNAS (USA) **91**, 4950-4954.

Fu Y.H., Friedman D.L., Richards S., Pearlman J.A., Gibbs R.A., Pizzuti A., Ashizawa T., Perryman M.B., Scarlato G., Fenwick R.G., Caskey C.T. (1993). Decreased expression of myotonin protein kinase messenger RNA and protein in adult form of myotonic dystrophy. Science **260**, 235-238.

Fu Y.H., Kuhl D.P.A., Pizzuti A., Pieretti M., Sutcliffe J.S., Richards S., Verkerk A.J.M.H., Holden J. J. A., Fenwick R.G., Warren S.T., Oostra B. A., Nelson D.L., Caskey C.T. (1991). Variation of the CGG repeat at the Fragile-X site results in genetic instability - resolution of the Sherman paradox. Cell **67**, 1047-1058.

Gacy A.M., Goellner G., Juranic N., Macura S., McMurray C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. Cell **81**, 533-540.

Gehring W.J. (1985). Homeotic genes, the homeo box, and the genetic-control of development. Cold Spring Harbor Symposia On Quantitative Biology, **L**, 243-251.

Gerber H.P., Georgiev O., Harshman K., Schaffner W. (1992). *In vitro* transcription complementation assay with mini-extracts of transiently transfected COS-1 cells. Nucleic Acids Research **20**, 5855-5856.

Gerber H.P., Seipel K., Georgiev O., Höfferer M., Hug M., Rusconi S., Schaffner W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science **263**, 808-811.

Gibson T.J. (1984). Studies on the Epstein-Barr Virus genome. PhD Thesis, University of Cambridge, England, UK.

Giniger E., Varnum S.M., Ptashne M. (1985). Specific DNA-binding of GAL4, a positive regulatory protein of yeast. Cell **40**, 767-774.

Gomes X.V., Wold M.S. (1995). Structural analysis of human replication protein. American Journal of Biological Chemistry **270**, 4534-4543.

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

Graham A., Papalopulu N., Krumlauf R. (1989). The murine and *Drosophila* homeobox gene complexes have common features oorganization and expression. Cell **57**, 367-378.

Green H. (1993). Human genetic-diseases due to codon reiteration - relationship to an evolutionary mechanism. Cell **74**, 955-956.

Green H., Wang N. (1994). Codon reiteration and the evolution of proteins. PNAS (USA) **91**, 4298-4302.

Han J., Hsu C., Zhu Z., Longshore J.W., Finley W.H. (1994). Over-representation of the disease associated (CAG) and (CGG) repeats in the human genome. Nucleic Acids Research **22**, 1735-1740.

Hanahan D. (1983). Studies on transformation of *Escherichia coli* with plasmids. Journal of Molecular Biology **166**, 557-580.

Hara Y., Rovescalli A.C., Kim Y., Nirenberg M. (1992). Structure and evolution of 4 POU domain genes expressed in mouse brain. PNAS (USA) **89**, 3280-3284.

Harding K., Wedeen C., McGinnis W., Levine M. (1985). Spatially regulated expression of homeotic genes in *Drosophila*. Science **229**, 1236-1242.

Harper P.S. (1989). Myotonic Dystrophy. Second edition. W.B. Saunders, London.

Harris S., Moncrieff C., Johnson K. (1996). Myotonic distrophy: will the real gene please step forward! Human Molecular Genetics **5**, 1417-1423.

Haynes S.R., Dollard C., Winston F., Beck S., Trowsdale J., Dawid I.B. (1992). The bromodomain - a conserved sequence found in human, *Drosophila* and yeast proteins. Nucleic Acids Research **20**, 2603.

Haynes S.R., Rebbert M.L., Mozer B.A., Forquignon F., Dawid I.B. (1987). PEN repeat sequences are GGN clusters and encode a glycine-rich domain in a *Drosophila* cDNA homologous to the rat helix destabilizing protein. PNAS (USA) **84**, 1819-1823.

Hedrick S.M., Cohen D.I., Nielsen E.A., Davis M.M. (1984). Isolation of cDNA clones encoding cell specific membrane associated proteins. Nature **308**, 149-153.

Hendricksen L.A., Umbricht C.B., Wold M.S. (1994). recombinant replication protein A - expression, complex formation, and functional characterisation. Journal of Biological Chemistry **269**, 11121-11132.

Holland P.W.H., Hogan B.L.M. (1986). Phylogenetic distribution of antennapedia-like homoeo boxes. Nature **321**, Pp.251-253.

Hope I.A. (1991). Promoter trapping in *Caenorhabditis elegans*. Development **113**, 399-408.

Horlein A.J., Naar A.M., Heinzel T., Torchia J., Gloss B., Kurokawa R., Ryan A., Kamel Y., Soderstrom M., Glass C.K., Rosenfeld M.G. (1995). Ligand independent repression by the thyroid hormone receptor mediated by a nuclear receptor co-repressor. Nature **377**, 397-404.

Hu T.H., Guan T.L., Gerace L. (1996). Molecular and functional characterization of the P62 complex, an assembly of nuclear pore complex glycoproteins. Journal of Cell Biology **134**, 589-601.

Huynh T.V., Young R.A., Davies R.W. (1985). Construction and screening cDNA libraries in λgt10 and λgt11. In DNA Cloning: apratical approach (ed. Glover D.M.) **1**, 49-110. IRL Press Limited, Oxford, London.

Imbert G., Saudou F., Yvert G., Devys D., Trottier Y., Garnier J.M., Weber C., Mandel J.L., Cancel G., Abbas N., Durr A., Didierjean O., Stevanin G., Agid Y., Brice A. (1996). Cloning of the gene for Spinocerebellar Ataxia-2 reveals a locus with high-sensitivity to expanded CAG/glutamine repeats. Nature Genetics **14**, 285-291.

Ionov Y., Peinado M.A., Malkhosyan S., Shibata D., Perucho M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature **363**, 558-561.

Jansen G., Madadevan M., Amemiya C., Wormskamp N., Segers B., Hendriks W., O'Hoy K., Baird S., Sabourin L., Lennon G., Jap P.L., Iles D., Coerwinkel M., Hofker M., Carrano A.V., de Jong P.J., korneluk R.G., Wieringa B. (1992). Characterization of the myotonic dystrophy region predicts multiple protein isoform-encoding mRNAs. Nature Genetics **1**, 261-266.

Jansen G., Bachner D., Coerwinkel M., Wormskamp N., Hameister H., Wieringa B. (1995). Structural organization and developmental expression pattern of the mouse WD repeat gene DMR-N9 immediately upstream of the myotonic dystrophy locus. Human Molecular Genetics **4**, 843-852.

Jansen G., Groenen P.J.T.A., Bächner D., Jap P.H.K., Coerwinkel M., Oerlemans F., van den Broek W., Gohlsch B., Pette D., Plomp J.J., Molenaar P.C., Nederhoff M.G.J., van Echteld C.J.A., Dekker M., Berns A., Hameister H., Wieringa B. (1996). Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nature Genetics **13**, 316-324.

Johnson K.J., Boucher C.A., King S.K., Winchester C.L., Bailey M.E.S., Hamilton G.M., Carey N. (1996). Myotonic dystrophy a single gene disorder? Biochemical Society Transactions **24**, 510-513.

Kang S., Jaworski A., Ohshima K., Wells R.D. (1995). Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E.coli*. Nature Genetics **10**, 213-218.

Kang, S., Ohshima, K., Jaworski A., Wells R.D. (1996). CTG triplet repeats from the myotonic dystrophy gene are expanded in *Escherichia coli* distal to the replication origin as a single large event. Journal of Molecular Biology **258**, 543-547.

Karlin S., Burge C. (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. PNAS (USA) **93**, 1560-1565.

Karlin S., Ghandour G. (1985). Multiple-alphabet amino acid sequence comparison of the immunoglobulin κ-chain constant domain. PNAS (USA) **82**, 8597-8601.

Karn J., Brenner S., Barnett L. and Casereni G. (1980). Novel bacteriophage λ cloning vector. PNAS (USA) **77**, 5172.

Kauffman C.A., Johnson J.E., Ahsan J., Harkavy-Friedman J., Malaspina D., Cleary J., Cloninger C.R., Faraone S.V., Tsuang M.T., Zander C., Lindblad K., Schalling M. (1996). Anticipation and schizophrenia: biology or bias? Cold Spring Harbor Symposium on Quantative Biology **LXI,** 68.

Kawaguchi Y., Okamoto T., Taniwaki M., Aizawa M., Inoue M., Katayama S., Kawakami H., Nakamura S., Nishimura M., Akiguchi I., Kimura J., Narumiya S., Kakizuka A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nature Genetics **8**, 221-227.

Kendler K.S., Masterson C.C., Davis K.L. (1985). Psychiatric illness in 1st degree relatives of patients with paranoid psychosis, schizophrenia and medical illness. British Journal of Psychiatry **147**, 524-531.

Kim C.S., Snyder R.O., Wold M.S. (1992). Binding properties of replication protein A from human and yeast-cells. Molecular and Cellular Biology **12**, 3050-3059.

Knight C.R.L., Rees R.C., Griffin M. (1991). Apotosis: a specific role for cytosolic transglutaminase and its importance in tumour progression. Biochemica et Biophysica Acta **1096**, 312-318.

Knight S.J.L., Flannery A.V., Hirst M.C., Campbell L. Christodoulou Z., Phelps S.R., Pointon J., Middletonprice H.R., Barnicoat A., Pembrey M.E., Holland J., Oostra B.A., Bobrow M., Davies K.E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental-retardation. Cell, **74**, 127-134.

Koide R., Ikeuchi T., Onodera O., Tanaka H., Igarashi S., Endo K., Takahashi H., Kondo R., Ishikawa A., Hayashi T., Saito M., Tomoda A., Miike T., Naito H., Ikuta F., Tsuji S. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nature Genetics **6**, 9-13.

Korn R., Schoor M., Neuhaus H., Henseling U., Soininen R, Zachgo J., Gossler A. (1992). Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimeric embryos with a high frequency and detect endogenous genes. Mechanisms of Development **39**, 95-109.

Krainer A.R., Mayeda A., Kozak D., Binns G. (1991). Functional expression of cloned human splicing factor SF2 - homology to RNA-binding proteins, U1-70K, and *Drosophila* splicing regulators. Cell **66**, 383-394.

Lamond A.I. (1991). ASF/SF2 - A splice site selector. Trends in Biochemical Sciences **16**, 452-453.

LaSpada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., Fishbeck, H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature **352**. 77-79.

Lewis, E.B. (1978). A gene complex controlling segemntation in Drosophila. Nature **276**, 565.

Li S.H., Mcinnis M.G., Margolis R.L., Antonarakis S.E., Ross C.A. (1993). Novel triplet repeat containing genes in human brain - cloning, expression, and length polymorphisms. Genomics **16**, 572-579.

Li X.J., Li S.H., Sharp A.H., Nucifora F.C., Schilling G., Lanahan A., Worley P., Snyder S.H., Ross C.A. (1995). A huntingtin associated protein enriched in brain with implications for pathology. Nature **378**, 398-402.

Liang P., Pardee A.B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science **257**, 967-971.

Libert F., Parmentier M., Lefort A., Dinsart C., Vansande J., Maenhaut C., Simons M.J., Dumont J.E., Vassart G. (1989). Selective amplification and cloning of 4 new members of the G-protein coupled receptor family. Science **244**, 569-572.

Lindblad K., Nylander P.O., Debruyn A., Sourey D., Zander C., Engstrom C., Holmgren C., Hudson T., Chotai J., Mendlewicz J., Vanbroeckhoven C., Schalling M., Adolfsson R. (1995). Detection of expanded CAG repeats in bipolar affective disorder using the repeat expansion detection (RED) method. Neurobiology of Disease, **2**, 55-62.

Lipman D.J., Pearson W.R. (1985). Rapid and sensitive protein similarity searches. Science **227**, 1435-1441.

Liu V.F., Weaver D.T. (1993). The ionizing radiation-induced replication protein A phosphorylation response differs between Ataxia-Telangiectasia and normal human-cells. Molecular and Cellular Biology **13**, 7222-7231.

Lloyd S.E., Pang J.T., Pearce S.H.S., Leigh S.E.A., Thakker R.V. (1997). Exclusion of ZFM1 as a candidate gene for multiple endocrine neoplasia type 1 (MEN1). Human Genetics **99**, 585-589.

Loev S.J., Margolis R.L., Young W.S., Li S.H., Schilling G., Ashworth R.G., Ross C.A. (1995). Cloning and expression of the rat atrophin-I (DRPLA disease gene) homolog. Neurobiology of Disease **2**, 129-138.

Logan C., Hanks M.C., Nobletopham S., Nallainathan D., Provart N.J., Joyner A.L. (1992). Cloning and sequence comparison of the mouse, human, and chicken engrailed genes reveal potential functional domains and regulatory regions. Developmental Genetics **13**, 345-358.

Love, J.M., Knight, A.M., McAleer, M.A., Todd, J.A. (1990). Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. Nucleic Acids Research **18**, 4123-4130.

Lundblad J.R., Kwok R.P.S., Laurence M.E., Harter M.L., Goodman R.H. (1995). Adenoviral E1A-associated protein p300 as a functional homologue of the transcriptional co-activator CBP. Nature **374**, 85-88.

Luntzleybman V., Frostholm A., Fernando L., Deblas A., Rotter A. (1993). GABA(A) benzodiazepine receptor-gamma(2) subunit gene expression in developing normal and mutant mouse cerebellum. Molecular Brain Research **19**, 9-21.

McDonald M.E., Ambrose C.M., Duyao M.P., Myers R.H., Lin C., Srinidhi L., Barnes G., Taylor S.A., James M., Groot N., Macfarlane H., Jenkins B., Anderson M.A., Wexler N.S., Gusella J.F., Bates G.P., Baxendale S., Hummerich H., Kirby S., North M., Youngman S., Mott R., Zehetner G., Sedlacek Z., Poustka A., Frischauf A.M., Lehrach H., Buckler A.J., Church D., Doucettestamm L., Odonovan M.C., Ribaramirez L., Shah M., Stanton V.P., Strobel S.A., Draths K.M., Wales J.L., Dervan P., Housman D.E., Altherr M., Shiang R., Thompson L., Fielder T., Wasmuth J.J., Tagle D., Valdes J., Elmer L., Allard M., Castilla L., Swaroop M., Blanchard K., Collins F.S., Snell R., Holloway T., Gillespie K., Datson N., Shaw D., Harper P.S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntingtons disease chromosomes. Cell **72**, 971-983.

Sambrook J., Fritsch E.F., Maniatis T. (1989). Molecular cloning: a laboratory manual. Second edition, Cold Spring Harbor Laboratory Press.

McGinnis, W., Levine, M.S., Hafen, E., Kuriowa, A., Gehring, W.J. (1984). A conserved DNA sequence in homoeotic genes of the *Drosophila* antennapedia and bithorax complexes. Nature **308**, 428-433.

McKay M.J., Troelstra C., Vanderspek P., Kanaar R., Smit B., Hagemeijer A., Bootsma D., Hoeijmakers J.H.J. (1996). Sequence conservation of the Rad21 *Schizosaccharomyces pombe* DNA double-strand break repair gene in human and mouse. Genomics **36**, 305-315.

McCallion A.S., Guenet J.L., Montague P., Griffiths I.R., Savioz A., Davies R.W. (1996). The mouse gene (mobp) encoding myelin-associated basic-protein maps to distal chromosome 9. Mammalian Genome **7**, 847-849.

McPhalen C.A., Svendsen I., Jonassen I., James M.N.G. (1985). Crystal and molecular-structure of chymotrypsin inhibitor-2 from barley seeds in complex with subtilisin novo. PNAS (USA) **82**, 7242-7246.

Meehan R., Antequera F., Lewis J., Macleod D., Mckay S., Kleiner E., Bird A.P. (1990). A nuclear protein that binds preferentially to methylated DNA *in vitro* may play a role in the inaccessibility of methylated CpGs in mammalian nuclei. Philosophical Transactions of The Royal Society of London Series B **326**, 199-205.

Meehan R., Lewis J., Cross S., Nan X.S., Jeppesen P., Bird A. (1992). Transcriptional repression by methylation of CpG. Journal of Cell Science, **S16**, 9-14.

Meier U.T., Blobel G. (1992). Nopp140 shuttles on tracks between nucleolus and cytoplasm. Cell **70**, 127-138.

Mitas M., Yu A., Dill J., Haworth I.S. (1995). The trinucleotide repeat sequence $d(CGG)_{15}$ forms a heat-stable hairpin containing G(syn).G(anti) base pairs. Biochemistry, **39**, 12803-12811.

Mitchell P.J., Tjian R. (1989). Transcriptional regulation in mammalian-cells by sequence specific DNA binding proteins. Science **245**, 371-378.

Monk M. (1995). Epigenetic programming of differential gene expression in development and evolution. Developmental Genetics **17**, 188-197.

Moyzis R.K., Buckingham J.M., Cram L.S., Dani M., Deaven L.L., Jones M.D., Meyne J., Ratliff R.L., Wu J.R. (1988). A highly conserved repetitive DNA-sequence, (TTAGGG)n, present at the telomeres of human chromosomes. PNAS (USA) **85**, 6622-6626.

Nagafuchi S., Yanagisawa H., Sato K., Shirayama T., Ohsaki E., Bundo M., Taked T., Tadokoro K., Kondo I., Murayama N., Tanaka Y., Kikushima H., Umino K., Kurosawa H., Furukawa T., Nihei K., Inoue T., Sano A., Komure O., Takahashi M., Yoshizawa T., Kanazawa I., Yamada M. (1994). Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nature Genetics **6**, 14-18.

Nagamine, C.M., Nishioka Y., Moriwaki K., Boursot P., Bonhomme F., Lau Y.F.C. (1992). The *musculus*-type Y chromosome of the laboratory mouse is of Asian origin. Mammalian Genome **3**, 84-91.

Nakazawa K., Ando T., Kimura T., Narimatsu H. (1988). Cloning and sequencing of a full length cDNA of mouse N-acetylglucosamine (beta-1-4) galactosyltransferase. Journal of Biochemistry **104**, 165-168.

Nasir J., Floresco S.B., Okusky J.R., Diewert V.M., Richman J.M., Zeisler J., Borowski A., Marth J.D., Phillips A.G., Hayden M.R. (1995). Targeted disruption of the Huntingtons disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell **81**, 811-823.

Neri C., Albanese V., Lebre A.S., Holbert S., Saada C., Bougueleret L., Meier-Ewert S., Legall I., Millasseau P., Bui H., Giudicelli C., Massart C., Guillou S., Gervy P., Poullier E., Rigault P., Weissenbach J., Lennon G., Chumakov I., Dausset J., Lehrach H., Cohen D., Cann H.M. (1996). Survey of CAG/CTG repeats in human cDNAs representing new genes - candidates for inherited neurological disorders. Human Molecular Genetics **5**, 1001-1009.

Nohno T., Kasai Y., Saito T. (1989). Novel cDNA sequence possibly generated by alternative splicing of a mouse glucocorticoid receptor gene transcript from shionogi carcinoma-115. Nucleic Acids Research **17**, 445.

Nomura N., Miyajima N., Sazuka T., Tanaka A., Kawarabayasi Y., Sato S., Nagase T., Seki N., Ishikawa K., Tabata S. (1994). Prediction of the coding sequences of unidentified human genes. DNA Research **1**, 47-56.

Ochman H., Ajioka J.W., Garza D., Hartl D.L. (1990). Inverse Polymerase Chain Reaction. Biotechnology **8**, 759-760.

O'Kane C.J., Gerhing W.J., (1987). Detection *in situ* of genomic regulatory elements in *Drosophila*. PNAS (USA) **84**, 9123-9127.

Okayama H., Berg P. (1983). A cDNA cloning vector that permits expression of cDNA inserts in mammalian cells. Molecular and Cellular Biology **3**, 280-289.

Orr H.T., Chung M.Y., Banfi S., Kwiatkowski Jr. T.J., Servadio A., Beaudet A.L., McCall A.E., Duvick L.A., Ranum L.P.W., Zoghbi H.Y. (1993). Expansion of unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature Genetics **4**, 221-226.

Ostrander E.A., Jong P.M., Rine J., Duyk G. (1992). Construction of small insert genomic DNA libraries highly enriched for microsatellite repeat sequences. PNAS (USA) **89**, 3419-3423.

Otten A.D., Tapscott S.J. (1995). Triplet repeat expansion in Myotonic Dystrophy alters the adjacent chromatin structure. PNAS (USA) **92**, 5465-5469.

Pai C., Chen H., Sheu H., Yeh N. (1995). Cell cycle dependent alterations of a highly phosphorylated nucleolar protein p130 are associated with nucleologenesis. Journal of Cell Science **108**, 1911-1920.

Parker A.E., Clyne R.K., Carr A.M., Kelly T.J. (1997). The Schizosaccharomyces pombe rad11(+) gene encodes the large subunit of replication protein A. Molecular and Cellular Biology, **17**, 2381-2390.

Parrish J.E., Oostra B.A., Verkerk A.J.M.H., Richards C.S., Reynolds J., Spikes A.S., Shaffer L.G., Nelson D.L. (1994). Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. Nature Genetics **8**, 229-234.

Pastuszko A., Wilson D.F., Ericinska M. (1986). Role for transglutaminase in neurotransmitter release by Rat brain synaptosomes. Journal of Neurochemistry **46**, 499-508.

Pecheux C., Mouret J.F., Durr A., Agid Y., Feingold J., Brice A., Dode C., Kaplan J.C. (1995). Sequence analysis of the CCG polymorphic region adjacent to the CAG triplet repeat of the hd gene in normal and HD chromosomes. Journal of Medical Genetics **32**, 399-400.

Perutz M.F., Johnson T., Suzuki M., Finch J.T. (1994). Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. PNAS (USA) **91**, 5355-5358.

Perutz M.F., Staden R., Moens L., Debaere I. (1993). Polar zippers. Current Biology **3**, 249-253.

Petronis A., Kennedy J.L. (1995). Unstable genes - unstable mind? American Journal of Psychiatry **152**, 164-172.

Pillai R., Kytle K., Reyes A., Colicelli J. (1993). Use of a yeast expression system for the isolation and analysis of drug resistant mutants of a mammalian phosphodiesterase. PNAS (USA) **90**, 11970-11974.

Prince V, Lumsden A., (1994). Hoxa-2 expression in normal and transposed rhombomeres: Independent regulation in the neural tube and neural crest. Development **120**, 911-923.

Reddy S., Smith D.B.J., Rich M.M., Leferovich J.M., Reilly P., Davis B.M., Tran K., Rayburn H., Bronson R., Cros D., Balice-Gordon R.J., Housman D. (1996). Mice lacking the myotonic dystrophy protein kinase develop a late onset progressive myopathy. Nature Genetics **13**, 325-335.

Richards R.I., Holman K., Yu S., Sutherland G.R. (1993). Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Human Molecular Genetics **2**, 1429-1435.

Riggins G.J., Lokey L.K., Chastain J.L., Leiner H.A., Sherman S.L., Wilkinson K.D., Warren S.T. (1992). Human genes containing polymorphic trinucleotide repeats. Nature Genetics **2**, 186-191.

Ritchie R.J., Knight S.J.L., Hirst M.C., Grewal P.K., Bobrow M., Cross G.E., Davies K.E. (1994). The Cloning of FRAXF - trinucleotide repeat expansion and methylation at a 3rd fragile site in distal Xqter. Human Molecular Genetics **3**, 2115-2121.

Rosche W.A., Jaworski A., Kang S., Kramer S.F., Larson J.E., Geidroc D.P., Wells R.D., Sinden R.R. (1996). Single-stranded DNA binding protein enhances the stability of CTG triplet repeats in *Escherichia coli*. Journal of Bacteriology **178**, 5042-5044.

Ross C.A. (1995). When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. Neuron **15**, 493-496.

Rubinsztein D.C., Amos W., Leggo J., Goodburn S., Jain S., Li S.H., Margolis R.L., Ross C.A., Ferguson-Smith M.A. (1995). Microsatellite evolution - evidence for directionality and variation in rate between species. Nature Genetics **10**, 337-343.

Rubinsztein D.C., Amos W., Leggo J., Goodburn S., Ramesar R.S., Old J., Bontrop R., McMahon R., Barton D.E., Ferguson-Smith M.A. (1994). Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nature Genetics **7**, 525-530.

Sabourin L.A., Tsilfidis C.T., Mayer P., Puymyrat J.M., Karpati G., Korneluk R.G. (1993). Investigation of the role of the myotonic dystrophy kinase (DMK) in the differentiation of cultured myoblasts. American Journal of Human Genetics **53**, 653.

Sadowski I., Ma J., Triezenberg S., Ptashne M. (1988). GAL4-VP16 is an unusually potent transcriptional activator. Nature **335**, 563-564.

Sanger F., Nicklen S., Coulson A.R. (1978). DNA sequencing with chain-terminating inhibitors. PNAS (USA). **74**, 5463-5464.

Schalling M., Hudson T.J., Buetow K.H., Housman D.E. (1993). Direct detection of novel expanded trinucleotide repeats in the human genome. Nature Genetics **4**, 35-139.

Shapira S.K., Chou J., Richaud F.V., Casadaban M.J. (1983). New versatile plasmid vectors for expression of hybrid proteins coded by a cloned gene fused to *lacz* gene sequences encoding an enzymatically active carboxy terminal portion of beta-galactosidase. Gene **25**, 71-82.

Shapiro M.B., Senapathy P. (1987). RNA splice junctions of different classes of eukaryotes - sequence statistics and functional implications in gene expression. Nucleic Acids Research **15**, 7155-7174.

Sharp A.H., Loev S.J., Schilling G., Li S.H., Li X.J., Bao J., Wagster M.V., Kotzuk J.A., Steiner J.P., Lo A., Hedreen J., Sisodia S., Snyder S.H., Dawson T.M., Ryugo D.K., Ross C.A. (1995). Widespread expression of Huntingtons Disease gene (IT15) Protein Product. Neuron **14**, 1065-1074.

Shaw D.J., Chaudhary S., Rundle S.A., Crow S., Brook J.D., Harper P.S., Harley H.G. (1993). A study DNA methylation in myotonic dystrophy. Journal of Medical Genetics **30**,

Short J.M., Fernandez J.M., Sorge J.A., Huse W.D. (1988). Lambda-zap - a bacteriophage lambda expression vector with in vivo excision properties. Nucleic Acids Research **16**, 7583-7600.

Sheffield V.C., Weber J.L., Buetow K.H., Murray J.C., Even D.A., Wiles K., Gastier J.M., Pulido J.C., Yandava C., Sunden S.L., Mattes G., Businga T., Mcclain A., Beck J., Scherpier T., Gilliam J., Zhong J., Duyk G.M. (1995). A collection of trinucleotide and tetranucleotide repeat markers used to generate high-quality, high-resolution human genome-wide linkage maps. Human Molecular Genetics **4**, 1837-1844.

Sigler P.B. (1988). Acid blobs and negative noodles. Nature **333**, 210-212.

Siomi H., Siomi M.C., Nussbaum R.L., Dreyfuss G. (1993). The protein product of the fragile-X gene, fmr1, has characteristics of an RNA-binding protein. Cell **74**, 291-298.

Smith G.K., Jie J., Fox G.E., Gao X.L. (1995). DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes. Nucleic Acids Research **23**, 4303-4311.

Smith S.S., Laayoun A., Lingeman R.G, Baker D.J., Riley J. (1994). Hypermethylation of telomere-like foldbacks at codon-12 of the human C-Ha-Ras gene and the trinucleotide repeat of the fmr-1 gene of Fragile-X. Journal of Molecular Biology **243**, 143-151.

Soininen R., Schoor M., Henseling U., Tepe C., Kisterswoike B., Rossant J., Gossler A. (1992). The mouse enhancer-trap-locus-1 (ETL-1) - a novel mammalian gene related to *Drosophila* and yeast transcriptional regulator genes. Mechanisms of Development **39**, 111-123.

Stallings L.R., Ford A.F., Nelson D., Torney D.C., Hildebrand C.E., Moyzis R.K. (1991). Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. Genomics **10**, 807-815.

Stallings L.R. (1994). Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: Implications for human genetic disease. Genomics **21**, 116-121.

Stallings L.R. (1995). Conservation and evolution of (CT)n/(GA)n microsatellite sequences at orthologous positions in diverse mammalian genomes. Genomics **25**, 107-113.

Stott K., Blackburn J.M., Butler P.J.G., Perutz M. (1995). Incorporation of glutamine repeats makes protein oligomerize: implication for neurodegenerative diseases. PNAS (USA) **92**, 6509-6513.

Stuart G.W., Vielkind J.R., Mcmurray J.V., Westerfield M. (1990). Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression. Development **109**, 577-584.

Streisinger G., Okada Y., Emrich J., Newton J., Tsugita A., Terzaghi E. and Inouye M. (1996). Frame shift mutations and the genetic code.

In Symposia on Quatative Biology Vol. XXXI, Cold Spring Harbor Laboratory of Quatittative Biology. New York, USA. 77-84.

Tamkun J.W., Deuring R., Scott M.P., Kissinger M., Pattatucci A.M., Kaufman T.C., Kennison J.A. (1992). *Brahma* - a regulator of *Drosophila* homeotic genes structurally related to the yeast transcriptional activator SNF2/SW12. Cell **68**, 561-572.

Taneja, K.L., McCurrach, M., Schalling, M., Housman, D., Singer, R. H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic-dystrophy cells and tissues. Journal of Cell Biology **128**, 995-1002.

Taylor S.M., Jones P.A. (1982). Mechanism of action of eukaryotic DNA methyltransferase - use of 5-azacytosine containing DNA. Journal of Molecular Biology **162**, 679-692.

Theodosiou A. M., Rodrigues N.R., Nesbit M.A., Ambrose H.J., Paterson H., McLellan-Arnold E., Boyd Y., Leversha M.A., Owen N., Blake D.J., Ashworth A., Davies K.E. (1996). A member of the MAP kinase phosphatase gene family in mouse containing a complex trinucleotide repeat in the coding region. Human Molecular Genetics **5**, 675-684.

Timchenko L.T., Monckton D.G., Caskey C.T. (1995). Myotonic Dystrophy- an unstable CTG repeat in a protein kinase gene. Seminars in Cell Biology, **6**, 13-19.

Timchenko L.T., Timchenko N.A., Caskey C.T., Roberts R. (1996). Novel proteins with binding specificity for DNA CTG repeats and RNA CUG repeats: implications for myotonic dystrophy. Human Molecular Genetics **5**, 115-121.

Toda T., Iida A., Miwa T., Nakamura Y., Imai T. (1994). Isolation and characterisation of a novel gene encoding nuclear protein at a locus (D11S636) tightly linked to multiple endocrine neoplasia type-1 (MEN1). Human Molecular Genetics **3**, 465-470.

Traub P. (1995). Intermediate filaments and gene regulation. Physiological Chemistry and Physics and Medical NMR **27**, 377-400.

Trottier Y., Lutz Y., Stevanin G., Imbert G., Devys D., Cancel G., Saudou F., Weber C., David G., Tora L., Agid Y., Brice A., Mandel J.L. (1995a). Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. Nature **378**, 403-406.

Trottier Y., Devys D., Imbert G., Saudou F., An I., Lutz Y., Weber C., Agid Y., Hirsch E.C., Mandel J.L. (1995b). Cellular localization of the Huntingtons Disease protein and discrimination of the normal and mutated form. Nature Genetics **10**, 104-110.

Trower M.K., Burt D., Purvis I.J., Dykes C.W., Christodoulou C. (1995). Fluorescent dye-primer cycle sequencing using unpurified PCR products as templates; development of a protocol amenable to high-throughput DNA sequencing. Nucleic Acids Research **23**, 2348-2349.

Turri M.G., Cuin K.A., Porter A.C.G. (1995). Characterization of a novel minisatellite that provides multiple splice donor sites in an interferon induced transcript. Nucleic Acids Research **23**, 1854-1861.

Van Der Ven P.F.M., Jansen G., Vankuppevelt T.H.M.S.M., Perryman M.B., Lupa M., Dunne P.W., Terlaak H.J., Jap P.H.K., Veerkamp J.H., Epstein H.F., Wieringa B. (1993). Myotonic dystrophy kinase is a component of neuromuscular junctions. Human Molecular Genetics **2**, 1889-1894.

Van der ploeg L.H.T., Liu A.Y.C, Borst P. (1984). Structure of the growing telomeres of Trypanosomes. Cell **36**, 459-468.

Vandaele S., Nordquist D.T., Feddersen R.M., Tretjakoff I., Perterson A.C., Orr H.T. (1991). Purkinje cell protein 2 regulatory regions and transgene expression in cerebellar compartments. Genes and Development **5**, 1136-1148.

Vonmelchner H., Ruley H.E. (1989). Identification of cellular promoters by using a retrovirus promoter trap. Journal of Virology **63**, 3227-3233.

Wang J., Manley J.L. (1995). Over-expression of the SR proteins ASF/SF2 and SC35 influences alternative splicing *in vivo* in diverse ways. RNA-A Publication of The RNA Society **1**, 335-346.

Wang J., Takagaki Y., Manley J.L. (1996). Targeted disruption of an essential vertebrate gene- ASF/SF2 is required for cell viability. Genes and Development **10**, 2588-2599.

Wang Y.H., Griffith J. (1995). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements Genomics **25**, 570-573.

Warren S.T. (1996). The expanding world of trinucleotide repeats. Science **271**, 1374-1375.

Watson J.D., Crick F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribonucleic acids. Nature **171**, 737-738.

Weber J.L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics **7**, 524-530.

Weber J.L., May P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain-reaction. American Journal of Human Genetics **44**, 388-396.

Wharton K.A., Yedvobnick B., Finnerty V.G., Artavanis-Tsakonas S. (1985). *OPA*: a novel family of transcribed repeats shared by the notch locus and developmentally regulated loci in *D. melanogaster*. Cell **40**, 55-62.

Wilkinson D.G., Krumlauf R. (1990). Molecular approaches to the segmentation of the hindbrain. Trends In Neurosciences **13**, 335-339.

Worley K.C., Wiese B.A., Smith R.F. (1995). Beauty - an enhanced blast-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Research **5**, 173-184.

Young R.A., Davis R.W. (1983). Efficient isolation of genes by using antibody probes. PNAS (USA) **80**, 1194-1198.

Yu S., Ozawa M., Naved A.F., Miyauchi T., Muramatsu H., Muramatsu T. (1995). cDNA cloning and sequence analysis of a novel calcium binding protein with oligoproline motif. Cell Structure and Function **20**, 263-268.

Zeitlin S., Lui, J.P., Chapman D.L., Papaioannou V.E., Efstratiadis A. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntingtons disease homolog. Nature Genetics **11**, 155-163.

Zoghbi H.Y., Orr H.T. (1995). Spino-cerebellar ataxia type-1. Seminars In Cell Biology **6**, 29-35.