Human and murine *FMR-1*: alternative splicing and translational initiation downstream of the CGG-repeat

Claude T. Ashley[1], James S. Sutcliffe[1], Catherine B. Kunst[1], Harold A. Leiner[1], Evan E. Eichler[2], David L. Nelson[2] & Stephen T. Warren[1]

Fragile X syndrome is associated with massive expansion of a CGG trinucleotide repeat within the FMR-1 gene and transcriptional silencing of the gene due to abnormal methylation. Partial cDNA sequence of the human *FMR-1* has been reported. We report here the isolation and characterization of cDNA clones encoding the murine homologue, *fmr-1*, which exhibit marked sequence identity with the human gene, including the conservation of the CGG repeat. A conserved ATG downstream of the CGG repeat in human and mouse and an in-frame stop codon in other human 5' cDNA sequences demarcate the *FMR-1* coding region and confine the CGG repeat to the 5' untranslated region. We also present evidence for alternative splicing of the FMR-1 gene in mouse and human brain and show that one of these splicing events alters the *FMR-1* reading frame, predicting isoforms with novel carboxy termini.

[1]*Howard Hughes Medical Institute and Departments of Biochemistry and Pediatrics, Emory University School of Medicine, Atlanta, Georgia 30322, USA*
[2]*Institute for Molecular Genetics and the Human Genome Center, Baylor College of Medicine, Houston, Texas 77030, USA*

J.S.S.present address: Howard Hughes Medical Institute, Institute for Molecular Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

Correspondence should be addressed to S.T.W.

Fragile X syndrome is the most frequent inherited cause of mental deficiency in humans, with a prevalence in males of approximately 1 per 1,000 (ref. 1). The syndrome segregates as an X-linked dominant disorder with reduced penetrance and its map position is coincident with a folate-sensitive fragile site at Xq27.3 (refs 2–4). The molecular basis of the fragile X syndrome[5–8] is governed by a gene termed *FMR-1*, which contains an unusual CGG trinucleotide repeat in the 5' portion of the transcript. Partial cDNAs have predicted a protein sequence dissimilar to any other known sequences and devoid of any obvious domains or motifs[5]. All partial cDNA sequences formed an open reading frame from the 5'-most sequence to a stop codon at nucleotide 1972 and incorporating the CGG-repeat, predicting a polyarginine stretch of 30 residues. As a downstream methionine has been observed, it was unclear if the CGG-repeat encoded FMR-1 protein or resided in the 5' untranslated portion of the message.

What is clear, however, is that the CGG repeat is the site of the mutation responsible for the vast majority of fragile X syndrome cases[5–9]. Amongst normal individuals, this triplet repeat is polymorphic, exhibiting repeat lengths varying from 7 to 52 triplets with a mean of roughly 30 (ref. 10). Fully penetrant individuals with fragile X syndrome (males and females) exhibit a massive expansion of this repeat beyond 230 triplets, usually exceeding 500 repeats. Nonpenetrant males and many carrier females exhibit repeat lengths intermediate between normal and affected. Some carrier females do have lengthy repeats similar to penetrant males but apparently escape the syndrome, presumably due to the lyonization patterns of the normal and fragile X chromosomes. When the CGG repeat is larger than approximately 230 triplets, DNA sequences of and surrounding the repeat are concomitantly methylated, including a CpG-island approximately 250 nucleotides proximal (5') to the repeat[6,11–13]. This abnormal methylation is correlated with the transcriptional silencing of *FMR-1* (refs 12, 13), and the absence of *FMR-1* protein is the presumptive basis for the disorder since FMR-1 expression is high in tissues relevant to the clinical phenotype[14].

The trinucleotide repeat in fragile X families exhibits marked instability, changing in size when transmitted[10]. The probability of expansion beyond 230 repeats into the penetrant range has been shown to be correlated with the maternal repeat length such that a carrier female with 70 triplets, for example, has less chance of having a penetrant offspring than does a carrier female with perhaps 150 repeats. Since the unstable repeat appears to increase rather than decrease in size upon transmission, carrier descendants in a fragile X pedigree have generally larger alleles and therefore more penetrant offspring. This direct correlation between carrier repeat length and penetrance has been referred to as the Sherman paradox[2–3,10] and is fundamentally similar to genetic anticipation where severity or age of onset of a disease appears to increase in subsequent generations. A classic example of genetic anticipation, myotonic dystrophy[15], is also caused by an unstable triplet repeat which undergoes massive expansion in affected individuals[16–18]. In this case, a CTG repeat is involved in the 3' untranslated portion of a gene predicted to encode a protein kinase. A third example of a triplet repeat mutation is spinal and bulbar muscular atrophy (SBMA) which involves a CAG repeat in the coding region of the androgen receptor gene that approximately doubles in length among affected individuals[19]. Very

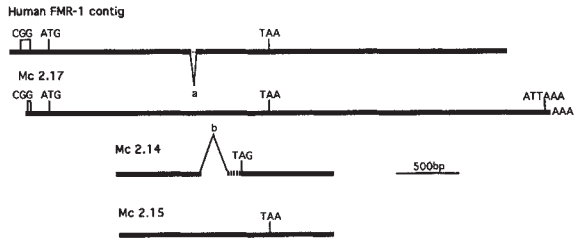

Fig. 1 Position of the mouse cDNA clones relative to the human *FMR-1* contig and each other. CGG denotes the position and relative size of the trinucleotide repeat in the human contig and Mc 2.17; ATG, putative translational start in the human contig and Mc 2.17; TAA, stop codon halting the major open reading frame of the human contig as well as mouse clones Mc 2.17 and Mc 2.15; TAG, stop codon terminating the major open reading frame of Mc 2.14; ATTAAA, consensus poly(A) addition signal of Mc 2.17; AAA, poly(A) tail of Mc 2.17; a refers to a 63 nucleotide deletion in the *FMR-1* contig relative to the three mouse cDNA clones; b denotes a 196 nucleotide deletion observed only in Mc 2.14. The broken line in Mc 2.14 represents the change in reading frame following the deletion. In the human contig the 63 nucleotide insert from human lymphocytes was subsequently cloned and sequenced.

recently, expansion of another CAG repeat sequence has been implicated as the molecular basis of Huntington's disease[20]. These four disorders represent a new mechanism of gene mutation leading to human genetic disease.

Unlike the scenario of myotonic dystrophy and SBMA, there are few clues as to the function of FMR-1 or how its absence leads to mental deficiency in fragile X syndrome. In order to begin elucidation of FMR-1 function, we report the characterization of the murine homologue of *FMR-1* and show that the CGG-repeat, while conserved in evolution, is likely noncoding. Additionally, we demonstrate alternative splicing of *FMR-1* in brain of both human and mouse predicting 12 potential FMR-1 isoforms, a subset of which, having encountered a shift in the FMR-1 reading frame, display unique carboxy termini.

Identification of mouse and human *FMR-1* cDNAs

Mouse *fmr-1* cDNA clones were isolated from a cDNA library constructed from adult BALB/c brain mRNA. Initial screening using the human *FMR-1* cDNA bc22 (ref. 5) resulted in the isolation of clone Mc 2.14 (Fig. 1). The 1,576 bp insert of this clone was subsequently used to rescreen the library, producing three additional clones, Mc 2.15, Mc 2.16 and Mc 2.17. With the exception of Mc 2.16, which displayed recurrent instability, these clones were converted to phagemids by *in vivo* excision[21] and the validity of each insert confirmed by Southern hybridization. The complete nucleotide sequence of the murine clones 2.14 (1576), 2.15 (1761) and 2.17 (4257) was determined by primer walking using automated sequencing techniques, and the sequences obtained were aligned into a contig and compared to the human sequence (Fig. 1). The largest of the mouse cDNAs, Mc 2.17, contained a 3' poly(A) tract preceded 27 nucleotides 5' by a consensus poly(A) addition sequence ATTAAA. Interestingly, this cDNA clone contained a triplet repeat of nine copies near its 5' terminus consisting of eight CGG and one CGA triplet. A putative open reading frame of 655 amino acids was present including the CGG repeat which, if translated, encodes polyarginine. Since the murine open reading frame, like that of the available human *FMR-1* contig, was open to the 5' end of the clone, the need to obtain cDNA clones with additional 5' sequence was apparent.

Human clones containing additional 5' sequence were identified by further analysis of preexisting clones as well as screening of random primed human cDNA libraries. The human cDNA, bc72 reported previously[5], contains additional 5' sequence that had initially failed to sequence reproducibly. Analysis of this clone revealed an additional 47 bases 5' to the former *FMR-1* contig and colinear with a genomic clone, pE 5.1 (Fig. 2). Also, screening of a random-primed, human testis cDNA library with the bc72 insert produced three more cDNA clones, pT1, pT2 and pT4. The nucleotide sequence at the 5' end of the pT2 and pT4 inserts each terminated within the CGG repeat (data not shown), likely due to premature termination by reverse transcriptase within this region. However, the nucleotide sequence of pT1 extended 123 bases upstream of the published *FMR-1* sequence (or 76 nts above the bc72 sequence), still with no divergence from the pE 5.1 genomic sequence (Fig. 2). The beginning of pT1 is 4 nucleotides from a major transcriptional start of FMR-1 in human brain as determined by primer extension and RNase protection assays (C.T.A. *et al.*, manuscript in preparation).

Nucleotide sequence analysis

Alignments of the mouse and human *FMR-1* sequence revealed both marked conservation and potentially important differences (Fig. 2). All of the mouse clones contain a 63 nucleotide (nt) insertion (Fig. 1; nts 1246–1308 of Mc 2.17) which was not reported in the human sequence[5]. Using reverse transcription polymerase chain reaction (RT-PCR) and *FMR-1* specific primers flanking this region, this additional sequence was recovered from total RNA of human lymphocytes and sequenced (Fig. 2 bold type). This 63 basepair (bp) region has since been found to represent a complete exon in humans (exon 12; D.L.N. *et al.*, manuscript in preparation) that is alternatively spliced in human *FMR-1* (ref. 22).

A six nucleotide insertion (nts 1003–1008) and a three nucleotide deletion (after nt 1206 of *fmr-1*) in the mouse were also apparent upon comparison to the human sequence (Fig. 2). The insertion in the mouse cDNAs has since been identified in lymphocyte RNA from 6 individuals, suggesting that the absence of these six nucleotides in the original published *FMR-1* sequence may have been caused by aberration of the bc22 cDNA. In contrast, the three nucleotide deletion appears specific for the mouse and predicts a single amino acid deletion at residue 362 of fmr-1. After correction for these differences, the mouse and human cDNA sequences display notable homology, with a nucleotide identity of 95% within the coding region.

Predicted amino acid sequence

Analysis of the 5' nucleotide sequence of pT1 and bc72 revealed a TGA stop codon sequence located 78 nucleotides upstream of, and in frame with, the CGG repeat of *FMR-1* (Fig. 2). A candidate ATG for translational initiation was identified in both the mouse and human sequence, located 66 and 69 nucleotides

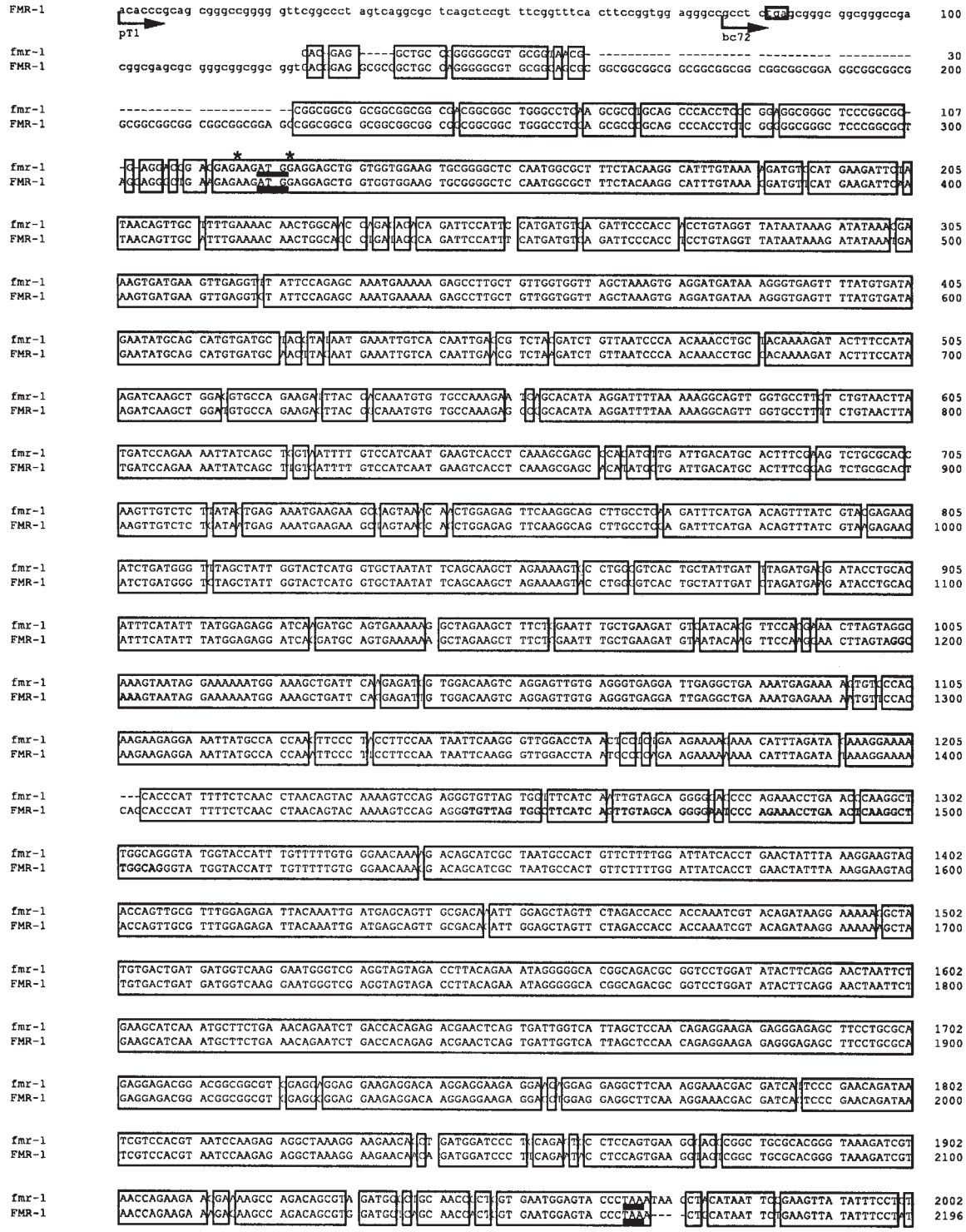

Fig. 2 Nucleotide sequence comparison of the human *FMR-1* contig and the mouse Mc 2.17 cDNA (*fmr-1*). Identities between the two species are enclosed in boxes. The beginning of human cDNA clones pT1 (from testis) and c72 (from brain) are denoted by arrows. Lower case letters represent the additional 5' sequence added to the *FMR-1* contig by these two clones. The bold box enclosing the nucleotides tga (top line) denotes an in frame stop codon found in both pT1 and c72, and the bold box surrounding the sequence ATTAAA (bottom line) denotes the consensus poly(A) addition signal of Mc 2.17. The underlined ATG and TAA demarcate the coding regions of both species. Asterisks over nucleotides denote the key residues in the Kozak consensus for translational starts. Nucleotides in bold type in *FMR-1* at 1198–1203 represent the 6 nucleotide deletion in the published *FMR-1* sequence[5] that has since been shown to be present in the human message. The 63 nucleotides in bold at position 1444–1506 were absent in the published *FMR-1* sequence[5], but have subsequently been found by sequencing RT-PCR products from human lymphocytes.

downstream of their CGG repeats, respectively. In both species, this is the first in-frame methionine codon of the open reading frame. The adjacent nucleotides in both human and mouse are also in close agreement with the Kozak consensus for translational starts (23–25). The conserved adenine (-3) and guanine (+4) have the strongest effects on translational initiation, and the conserved G residues at positions -6 and -9 are the preferred nucleotides at these positions as well (Fig. 2). Furthermore, gaps and non-identities in the nucleotide sequence are relatively more frequent upstream of this ATG, consistent with the assignment of this region as noncoding. We conclude that both the mouse and human *FMR-1* contigs are full length with respect to their coding regions, and that the CGG repeats are within the 5' untranslated regions (5' UTR) of each clone. Alignment of the predicted protein sequence of human with mouse reveals 97.0% amino acid identity and greater than 98% similarity (Fig. 3). A predicted length of 614 amino acids or a protein mass of 68,912 Da is suggested, although, as indicated below, further alternative splicing may produce isoforms of different molecular weights.

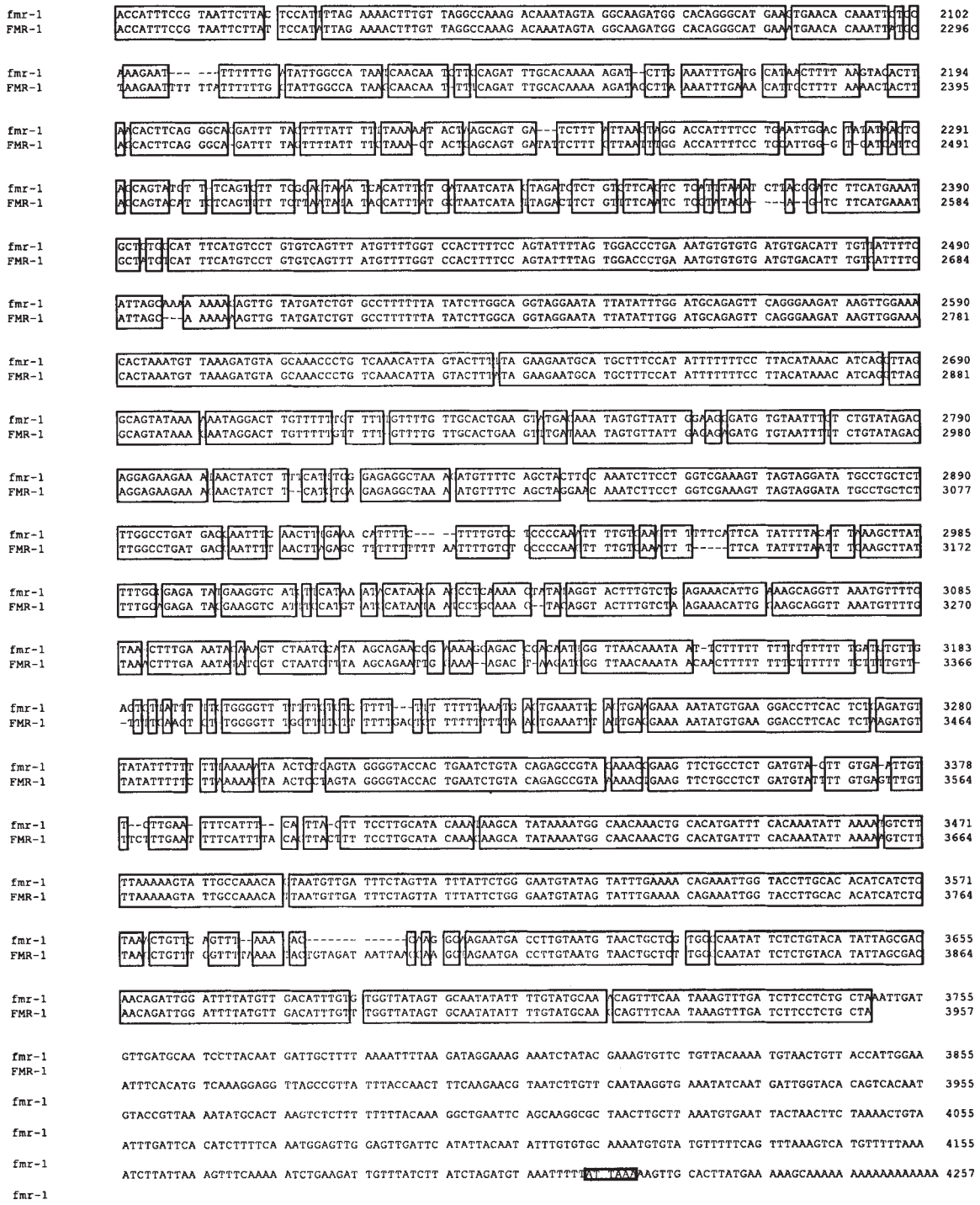

Fig. 2 Continued

```
FMR-1  MEELVVEVRG SNGAFYKAFV KDVHEDSITV AFENNWQPDR QIPFHDVRFP PPVGYNKDIN ESDEVEVYSR ANEKEPCCWW LAKVRMIKGE FYVIEYAACD  100
fmr-1  MEELVVEVRG SNGAFYKAFV KDVHEDSITV AFENNWQPER QIPFHDVRFP PPVGYNKDIN ESDEVEVYSR ANEKEPCCWW LAKVRMIKGE FYVIEYAACD  100

FMR-1  ATYNEIVTIE RLRSVNPNKP ATKDTFHKIK LDVPEDLRQM CAKEAAHKDF KKAVGAFSVT YDPENYQLVI LSINEVTSKR AHMLIDMHFR SLRTKLSLIM  200
fmr-1  ATYNEIVTIE RLRSVNPNKP ATKDTFHKIK LEVPEDLRQM CAKESAHKDF KKAVGAFSVT YDPENYQLVI LSINEVTSKR AHMLIDMHFR SLRTKLSLIL  200

FMR-1  RNEEASKQLE SSRQLASRFH EQFIVREDLM GLAIGTHGAN IQQARKVPGV TAIDLDEDTC TFHIYGEDQD AVKKARSFLE FAEDVIQVPR NLVGKVIGKN  300
fmr-1  RNEEASKQLE SSRQLASRFH EQFIVREDLM GLAIGTHGAN IQQARKVPGV TAIDLDEDTC TFHIYGEDQD AVKKARSFLE FAEDVIQVPR NLVGKVIGKN  300

FMR-1  GKLIQEIVDK SGVVRVRIEA ENEKNVPQEE EIMPPNSLPS NNSRVGPNAP EEKKHLDIKE NSTHFSQPNS TKVQRVLVAS SVVAGESQKP ELKAWQGMVP  400
fmr-1  GKLIQEIVDK SGVVRVRIEA ENEKSVPQEE EIMPPSSLPS NNSRVGPNSS EEKKHLDTKE N-THFSQPNS TKVQRVLVVS SIVAGGPQKP EPKAWQGMVP  399

FMR-1  FVFVGTKDSI ANATVLLDYH LNYLKEVDQL RLERLQIDEQ LRQIGASSRP PPNRIDKEKS YVTDDGQGMG RGSRPYRNRG HGRRGPGYTS GTNSEASNAS  500
fmr-1  FVFVGTKDSI ANATVLLDYH LNYLKEVDQL RLERLQIDEQ LRQIGASSRP PPNRIDKEKG YVTDDGQGMG RGSRPYRNRG HGRRGPGYTS GTNSEASNAS  499

FMR-1  ETESDHRDEL SDWSLAPTEE ERESFLRRGD GRRRGGGGRG QGGRGRGGGF KGNDDHSRTD NRPPNPREAK GRTTDGSLQN TSSEGSRLRT GKDRNQKKEK  600
fmr-1  ETESDHRDEL SDWSLAPTEE ERESFLRRGD GRRRGGGGRG QGGRGRGGGF KGNDDHSRTD NRPPNPREAK GRTADGSLQS ASSEGSRLRT GKDRNQKKEK  599

FMR-1  PDSVDGQQPL VNGVP
fmr-1  PDSVDGIQPL VNGVP
```

Fig. 3 Amino acid alignments of the predicted proteins of the human and mouse contigs, FMR-1 and fmr-1, respectively. Identities between the two polypeptides are shaded. The box represents the amino acids encoded by the 63 nucleotide fragment originally identified in mouse and subsequently isolated in human. Residues in lighter print (also boxed) were obtained by sequencing RT-PCR product from human lymphocytes.

Alternative splicing of *fmr-1*

In addition to the 63 nucleotide insertion observed in all three mouse cDNAs, Mc 2.14 displayed a 196 nucleotide deletion not observed in any of the other mouse or human cDNA clones. Since both of these fragments represent complete exons in the human gene (exon 12 and 14; D.L.N. *et al.*, manuscript in preparation), alternative splicing of these exons seemed likely. To test this notion, RT-PCR of both total mouse and human brain RNA was performed using the primer pairs shown (Fig. 4*a*). In each case, the downstream primer in the amplification step (13r, 16r) was end-labelled, and the RT-PCR products obtained were resolved on denaturing polyacrylamide gels followed by autoradiography.

As depicted (Fig. 4*a*, *b*), amplification of the region including exon 12 in both mouse and human (primers 11f and 13r) produced two major products of the appropriate mobilities to represent messages either including (201 bp, mouse; 205 bp, human) or excluding (140 bp, mouse; 145 bp, human) the 63 nucleotides of exon 12. The slight difference in migration of the bands of mouse and human was expected due to the 3 nucleotide (1 amino acid) deletion in *fmr-1* relative to human. For mouse we chose to designate proteins encoded by messages colinear with Mc 2.17, fmr-1 isoform 1 (iso1), and we named the isoform lacking exon 12, fmr-1 iso7.

To test for alternative splicing of exon 14 and to confirm the validity of the cDNA Mc 2.14, RT-PCR was also performed on both mouse and human using primers 12f and 16r. The bands were identical between mouse and human, and the sizes were in close agreement with predicted values (Fig. 4*a*, *c*). The largest band in each lane (580 bp) corresponded to the size predicted if the template was colinear with that of Mc 2.17 (iso1). The two major bands of 545 and 505 bp were of the appropriate mobilities to represent two alternative splice acceptor sites used in human exon 15 (ref. 22). The predicted proteins encoded by these messages are referred to as fmr-1 iso2 (alternate acceptor 1, exon 15) and fmr-1 iso3 (alternative acceptor 2, exon 15), and the corresponding isoforms lacking exon 12 would be referred to as fmr-1 iso-8 and iso-9. Thus, similar to human *FMR-1* (ref. 22), at least three isoforms are predicted to be generated based upon the three exon 15 acceptor sequences. However, evidence for alternative splicing of exon 14 in conjunction with variable splicing of exon 15

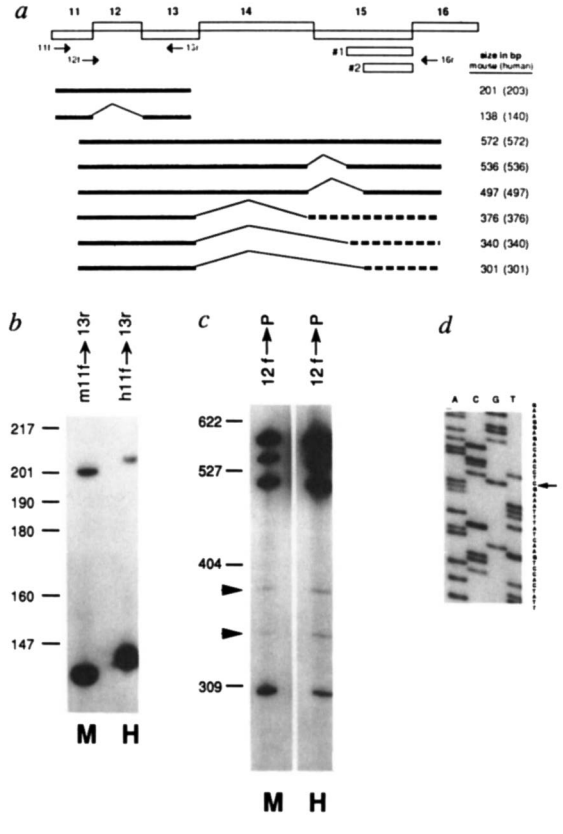

Fig. 4 RT-PCR analysis of alternative splicing of *fmr-1*. *a*, schematic representation of potential products obtained upon RT-PCR analysis of mouse brain total RNA within the regions of alternative splicing. The open boxes at the top denote the *fmr-1* transcript as it would appear if all exons identified were present. The numbers above the boxes represent the particular exons as defined for *FMR-1*. Open boxes preceded by 1 or 2 depict the two alternate splice acceptor sites identified in human exon 15 (ref. 22). The location of primers 11f, 12f, 13r, and 16r are shown. All potential splice products and the predicted size of each are depicted. Broken lines (right side, bottom three predicted products) represent a change in the fmr-1 reading frame. *b* and *c*, Autoradiographs of mouse (M) and human (H) RT-PCR products. Primers used are denoted above figures. Numbers on the left represent the size and location of the labelled pBR322/*Hpa*II molecular weight markers. Arrowheads (*b*) denote minor bands representing 2 of the six expected products. *d*, portion of a sequencing gel of one of the RT-PCR products that has been subcloned into TA vector (Invitrogen) and manually sequenced. The arrow points to the splice junction. This particular product is the bottom predicted product in *a* as well as the lowest band in *c*.

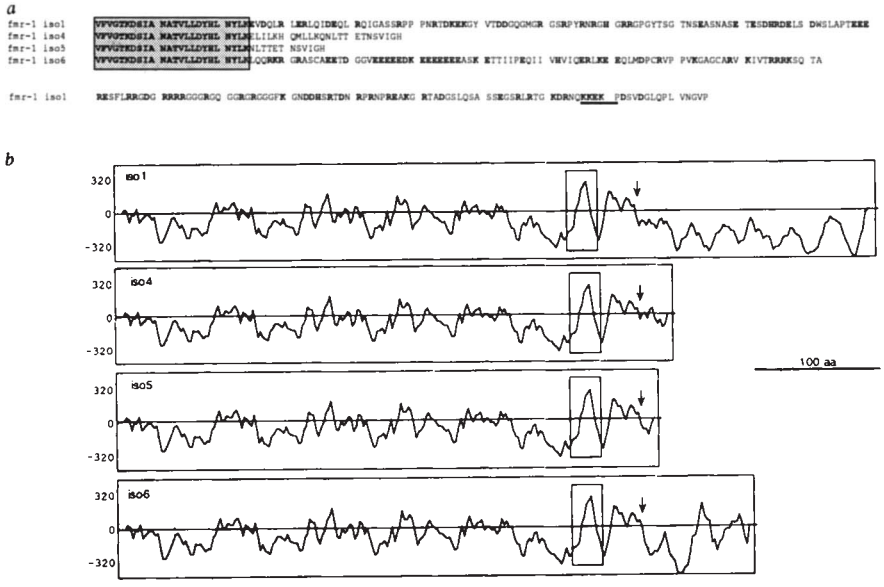

Fig. 5 New C termini of the fmr-1 isoforms which exclude exon 14 compared to the carboxy terminus of iso1. *a*, Amino acid sequence of the carboxy termini of fmr-1 iso1, iso4, iso5 and iso6. The latter three isoforms have exon 14 spliced out resulting in a +1 frameshift in the *fmr-1* reading frame and production of novel C termini. Charged residues in the C termini of iso1 and iso6 are shown in bold. A consensus nuclear translocation signal is underlined. *b*, Hydrophobicity plots (Kyte-Doolittle)[31] of the four fmr-1 isotypes discussed in *a*. Arrows denote the beginning of unique C-terminal sequence for each isoform. Boxed portions denote amino acids belonging to exon 12 which are excluded in fmr-1 isoforms 7-12.

was also demonstrated. The lower major product of 305 bp corresponds in size with that predicted for a message lacking exon 14 and using alternative splice acceptor site 2 of exon 15 (Fig. 4*c*). Two minor products of 385 and 345 bp (Fig. 4*c*, see arrows) correspond to messages lacking exon 14 and using the two other splice acceptors of exon 15, the larger of the two being analogous to the cDNA sequence of clone Mc 2.14. The predicted protein isoforms for these three messages have been termed fmr-1 iso4 (- exon 14), iso5 (- exon 14; alternative acceptor 1, exon 15), and iso6 (- exon 14; alternative acceptor 2, exon 15), and the analogous isoforms in absence of exon 12 fmr-1 iso10 - iso12.

Conclusive evidence for alternative splicing was obtained by repeating the above reactions in mouse using unlabelled primers followed by subcloning and DNA sequencing of selected products. The splice junction of the product corresponding to fmr-1 iso6 is shown in Fig. 4*d*, clearly demonstrating exon 13 joining exon 15 at the 2nd alternative splice acceptor. In the recent work of Verkerk *et al.*[22], alternative splicing of exon 14 was not observed, however the RT-PCR primers used in that study were derived from human exon 14 and would not have detected the absence of this exon.

Novel C termini of *fmr-1* splice products

Analysis of select *fmr-1* isoforms encoded by the alternatively spliced messages described above reveal that exclusion of exon 14 from the *fmr-1* message causes a +1 frameshift in the fmr-1 reading frame downstream of exon 13. For example, cDNA Mc 2.14 would be predicted to encode a truncated fmr-1 polypeptide, iso4, which consists of 448 amino acids and includes 24 novel amino acids at its C terminus (Fig. 5*a*). A similar alternative splice product uses acceptor 1 in exon 15, producing iso5, which would consist of 436 amino acids and exhibits 12 novel amino acids at its C-terminal end (Fig. 5*a*). The final isoform of this type, iso6, excludes exon 14 and has the initial 75 bases of exon 15 up to splice acceptor 2 excluded. In this case, the stop codon terminating iso4 and iso5 has been spliced out and the predicted protein of this isoform displays 88 new amino acids at its C terminus and is 512 amino acids in length (Fig. 5*a*). Of these three isoforms with novel carboxy termini, iso6 appears much more abundant in brain (see Fig. 4*c*). Isoforms 10–12 would presumably display the same carboxy termini as those described above. The predicted amino acid sequences of these newly identified carboxy termini failed to detect significant homology during database searches.

Hydrophobicity profiles of these fmr-1 isoforms just described are shown in Fig. 5*b*. The hydrophobicity profile of iso1 appears trimodal with a highly hydrophilic C-terminal one-third that is separated from a rather unremarkable N-terminal two-thirds by a single hydrophobic peak (Fig. 5*b*). This trimodal characteristic is lost in the profiles of iso4 and iso5, since there is no C-terminal hydrophilic region present (Fig. 5*b*). Interestingly, iso6 regains its hydrophilic tail region (Fig. 5*b*), although the sequence of the C-terminus of this isoform is notably different from that of iso1 (Fig. 5*a*), with frequent glutamic acids rather than the common arginine residues. Also denoted (Fig. 5*b*) is the region of the hydrophobicity profiles which corresponds to exon 12. It is noteworthy that predicted fmr-1 isoforms iso7 – iso12 will not contain this largest hydrophobic peak and will thus appear more bimodal in character. Although none of these hydrophobicity profiles clearly depicts a region of membrane transversion, the significant differences of the profiles of these novel C termini and isoforms lacking exon 12 suggest functional diversity and/or a means of partitioning these isoforms among different tissues or into different structures within a given tissue. It is also noted that a putative nuclear translocation signal, KKXKP (amino acids 597–601 of FMR-1; Fig. 5*a*), is present in predicted isoforms 1–3 and 7–9 and absent in predicted isoforms 4–6 and 10–12, potentially altering the cellular localization of fmr-1 isoforms.

Discussion

Our findings have determined the likely coding regions of the *FMR-1* message of both human and mouse. Significantly, the CGG-repeat is located downstream of an in-frame stop codon and 5' to a translational start motif, and is confined to the 5' UTR of the message. We have also shown that the CGG repeat has been conserved between human and mouse. Evolutionary conservation of this motif within the 5'UTR suggests that it may play an important regulatory role, perhaps as a DNA-protein binding site or, analogous to the translational control

over ferritin and the transferrin receptor[23], a site of interaction with an mRNA binding protein. In favour of the latter is the fact that the mouse and human FMR-1 messages display 5' UTR's that are both relatively long (123 and 318 bp, respectively) and very GC-rich (80% and 83%, respectively), an aspect of mammalian messages that has been linked to posttranscriptional control[24–29]. The presence of regions of strong sequence identity in portions of the 3'UTR between the mouse and human *FMR-1* transcripts is also intriguing, although the functional significance of this homology remains to be determined.

Comparison of the amino acid sequences of the mouse and human *FMR-1* revealed strong homology, with similarity values approaching 99%. This high degree of amino acid identity was not surprising given the high degree of conservation of *FMR-1* across many species[5]. Nevertheless, the strong homology between human and mouse substantiates the functional significance of FMR-1 in normal individuals. The acquisition of the murine sequence, particularly the nucleotide sequence, should also be quite useful for future studies such as genetic disruption of *fmr-1* in transgenic mice that hopefully will provide new information as to the normal function of fmr-1 *in vivo*.

The observation of alternative splicing of the *fmr-1* message is also significant. Alternative exclusion of exon 12 in both human and mouse produces FMR-1 isoforms which lack a major hydrophobic segment of the protein. More interesting are the alternative splicing events which involve exon 14 and result in novel C termini of the predicted isoforms. Individual isoforms could have functional differences and/or may be partitioned to structurally and functionally distinct regions within tissues, particularly in the brain. We have recently shown expression of the *fmr-1* mRNA throughout the murine brain in such areas as the granular layers of the cerebellum and hippocampus, the cerebral cortex, and the habenula[14]. Antibody generated against fmr-1 isoforms with novel C termini could greatly facilitate localization of distinct fmr-1 isoforms to particular regions thereby further advancing attempts to understand normal FMR-1 function and how its absence leads to the mental retardation associated with fragile X syndrome.

Methodology

cDNA library screening and DNA sequencing. cDNA libraries (adult mouse BALB/c, brain, λZAP, Stratagene; human testes, λgt11, Clontech) were plated with appropriate host bacteria (XL1-blue, λZAP; Y1090, λgt11) on 15 cm plates at a density of 40,000–50,000 per plate, and plaque lifts carried out using 15 cm diameter Biodyne nylon filters (Pall). Crosslinked filters were prehybridized for a minimum of 2 h at 65 °C in 12 ml containing 250 mM sodium phosphate, 1 mM EDTA, 250 mM sodium chloride, 7% (w/v) SDS, 10% (w/v) PEG (8000 m.w.), 1% BSA, and 200 mg ml^{-1} denatured salmon sperm DNA. Hybridizations were carried out overnight under the same conditions as above with the exception that fresh hybridization solution containing 25 ng radioactively labelled probe was used. DNA probes used were radiactively labelled by random primed labelling using the Megaprime kit (Amersham) and 10 μCi α-^{32}P-ATP and 25 ng double-stranded probe per reaction. Filters were washed four times for 30 min in 2 × SSC; 0.1% SDS at 65 °C, once for 15 min in .5 × SSC, 0.1% SDS at 65 °C and once for 15 min in 0.2 × SSC, 0.1% SDS at 65°C. Filters were exposed overnight at –80 °C using Kodak XAR film and lightening plus intensifying screens. Secondary screenings were carried out as above except that phage were plated at 500–1000 plaques/15 cm plate. Positive mouse plaques were converted to pBluescript phagemid DNA in pBluescript SK+ via *in vivo* excision[21] as described by the manufacturer (Stratagene). λ phage DNA of positive testis plaques was isolated according to standard techniques[30] and cleaved using *Eco*R1 to liberate the insert, and the inserts were subcloned directly into *Eco*R1 cleaved and CIP treated pBluescript SK+ (Stratagene) without prior purification.

Double-strand sequencing was performed on an ABI 373A automated DNA sequencer using the *Taq* dye deoxy cycle sequencing kit (ABI) as described by the manufacturer. Primers used initially to obtain sequence from the ends of the inserts were the commercially available pBluescript vector primers SK and KS. The remainder of the inserts were sequenced via primer walking using 21-24mer oligonucleotide primers. DNA sequencing of bc72 was carried out manually using the Bst premixed 7-deaza-dGTP sequencing kit (Biorad) and the protocol of the supplier.

DNA and protein sequence analysis. DNA sequences obtained were analysed and alignments performed using the Geneworks version 2.1 software (Intelligenetics). Databank searches (Genbank, Swissprot) were carried out using the 'Fasta' program of the GCG package (Genetics Computer Group, Madison). Hydrophobicity plots were compiled[31] using the Geneworks 2.1 software (Intelligenetics).

RT-PCR analysis of alternative splice products. RT-PCR experiments were carried out using the Genamp RNA PCR kit (Perkin-Elmer) as described by the manufacturers. In each RT-PCR reaction, 1 μg total mouse brain (adult BALB/c) RNA was used as template. First strand synthesis was performed using random hexamer primers. PCR amplification was carried out as described by the manufacturer with the exception that 2.2 mM $MgCl_2$ was used. 13 pmol of each of the primers m11f (5'-TCAAGGGTTGGACCTAACTCCTC-3'; nts 1150–1172 of *fmr-1*), h11f (5'-CAAGGGTTGGACCTAATGCCCC-3'; nts 1346–1367 of *FMR-1*), and 13r (5'-GATGCTGTCCTTTGTTCCCAC-3'; nts 1330–1350 of *fmr-1*) were used per reaction in analysis of splicing of exon 12. 10 pmol of each of the primers 12f (5'-CCAGAGGGTGTTAGTGGTTTC-3'; nts 1239–1259 of *fmr-1*) and 16r (5'-GTGGACGATTATCTGTTCGGGA-3'; nts 1789–1810 of *fmr-1*) were used per reaction in analysis of splicing of exon 14. Primers 13r and 16r had been end-labelled using T4 kinase and 100 μCi γ-^{32}P-ATP[30] prior to amplification. A Perkin-Elmer Model 480 thermal cycler was used for 35 cycles of amplification under conditions of denaturation: 94 °C for 1 min; annealing: 64 °C for 1 min (primers m11f, h11f and 13r) or 62 °C for 1 min (primers 12f and 16r); extension: 72 °C for 3 min, followed by a final extension for 7 min at 72 °C. A volume of 80 μl formamide dye solution (80% formamide, 0.1% xylene cyanol, 0.1% bromophenol blue) was added to each 100 μl reaction containing radiactively labelled products, and 3–5 μl of each were electrophoresed in a 5% Long Ranger polyacrylamide gel (AT Biochem) at 80 W for 1.25 h. Exposure time was 3–48 h at –80 °C. For subcloning and DNA sequence analysis, RT-PCR reactions were run exactly as described above with the exception that primer 16r was not end-labelled. PCR reactions (100 μl) were reduced by vacuum to approximately 40 μl each, then purified from a 0.7% LMP agarose gel (BRL) using the Qiaex DNA extraction kit (Qiagen). The purified products were then cloned into TA vector (Invitrogen) as described by the manufacturer. Sequencing of recombinants was done manually using the Sequenase kit (USB) and the manufacturer's protocol.

RT-PCR of exon 12 of human lymphocyte total RNA was carried out as described above. Total RNA was isolated from human lymphocytes as described[32], and 1 μg/reaction amplified using 40 pmol each of *FMR-1*-specific primers J (5'-CACTTTCGGAGTCTGCGCAC-3'; nts 880–899 of *FMR-1*) and K (5'-TAGCTCCAATCTGTCGCAACTGC-3'; nts 1635–1657 of *FMR-1*). PCR products were phenol-chloroform extracted, purified from primers and truncated products using Ultrafree-MC 30,000 spin columns (Millipore), and subcloned into TA vector (Invitrogen) as described. Sequencing was done manually using the Sequenase kit (USB).

Received 25 January; accepted 10 March 1993.

Acknowledgements

We are grateful to Greg Riggins, Doug Price and Fuping Zhang for their support and technical assistance, and Keith Wilkinson and Jeremy Boss for helpful discussion. C.T.A. is a predoctoral fellow of the March of Dimes Birth Defects Foundation (18-F492-0951). S.T.W. is an investigator of the Howard Hughes Medical Institute. This work was supported in part from grants from the NICHD (HD29256 to D.L.N. and HD20521 to S.T.W.).

1. Brown, W.T. The fragile X: progress towards solving the puzzle. *Am. J. hum. Genet.* **47**, 175–180 (1990).
2. Sherman, S.L. *et al.* The marker (X) syndrome: a cytogenetic and genetic analysis. *Ann. Hum. Genet.* **48**, 21–37 (1984).
3. Sherman, S.L. *et al.* Further segregation analysis of the fragile X syndrome with special reference to transmitting males. *Hum. Genet.* **69**, 289–299 (1985).
4. Lubs, J.A., Jr. A marker X chromosome. *Am. J. hum. Genet.* **21**, 231–244 (1969).
5. Verkerk, A.J.M.H. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
6. Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fargile X syndrome. *Science* **252**, 1097–1102 (1991).
7. Kremer, E.J. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CGG)n. *Science* **252**, 1711–1714 (1991).
8. Vincent, A. *et al.* Abnormal pattern detected in fragile-X patients by pulsed-field gel electrophoresis. *Nature* **349**, 624–626 (1991).
9. Yu,S. *et al.* Fragile X genotype characterized by an unstable region of DNA. *Science* **24**, 1179–1181 (1991).
10. Fu, Y.H. *et al.* Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**, 1047–1058 (1991).
11. Bell, M.V. *et al.* Physical mapping across the fragile X: hypermethylation and clinical expression of fragile X syndrome. *Cell* **64**, 861–866 (1991).
12. Pieretti, M. *et al.* Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* **66**, 817–822 (1991).
13. Sutcliffe, J.S. *et al.* DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum. molec. Genet.* **1**, 397–400 (1992).
14. Hinds, H. L. *et al.* Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. *Nature Genet.* **3**: 36–43 (1993).
15. Harper, P.S., Harley, H.G., Reardon, W. & Shaw, D.J. Anticipation in myotonic dystrophy: new light on an old problem. *Am. J. hum Genet.* **51**, 10–16 (1992).
16. Brook, J.D. *et al.* 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 (1992).
17. Fu, Y.H. *et al.* An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**, 1256–1258 (1992).
18. Mahadevan, M. *et al.* Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**, 1253–1255 (1992).
19. La spada, A.R. *et al.* Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
20. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
21. Short, J.M., Fernandez, J.M., Sorge, J.A. & Huse, W.D. lZAP: a bacteriophage l expression vector with *in vivo* excision properties. *Nucl. Acids Res.* **16**, 7583–7600 (1988).
22. Verkerk, A. J. M. H. *et al.* Alternative splicing in the fragile X (FMR-1) gene. *Hum. molec. Genet.* **2**, 399–404 (1993).
23. Theil, E.C. Regulation of ferritin and transferrin receptor mRNAs. *J. Biol. Chem.* **265**, 4771–4774 (1990).
24. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl. Acids Res.* **15**, 8125–8148 (1987).
25. Kozak, M. Leader length and secondary structure modulate mRNA function under conditions of stress. *Molec. cell. Biol.* **8**, 2737–2744 (1988).
26. Kozak, M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. biol. Chem.* **266**, 19867–19870 (1991).
27. Muller, A.J. & Witte, O.N. The 5' noncoding region of the human leukemia-associated oncogene BCR/ABL is a potent inhibitor of in vitro translation. *Molec. cell. Biol.* **9**, 5234–5238 (1989).
28. Godeau, F., Persson, H., Gray, H.E. & Pardee, A.B. *c-myc* expression is dissociated from DNA synthesis and cell division in *Xenopus* oocyte and early embryonic development. *EMBO J.* **5**, 3571–3577 (1986).
29. Taylor, M.V.M., Gusse, M., Evan, G.I., Dathan, N. & Mechali, M. *Xenopus myc* proto-oncogene during development: expression as a stable maternal mRNA uncoupled from cell division. *EMBO J.* **5**, 3563–3570 (1986).
30. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular cloning: a laboratory manual* 2nd edn (Cold Spring Harbor Press, New York, 1989).
31. Kyte, J. & Doolittle, R.F. A simple method for displaying hydropathic character of a protein. *J. molec. Biol.* **157**, 105–132 (1982).
32. Chomczynski, P. & Sacchi, N. Single step method of RNA isolation by acid quanidinium thiocyanate -phenol-chloroform extraction. *Analyt. Biochem.* **162**,156–159 (1987).