An Unstable Triplet Repeat in a Gene Related to Myotonic Muscular Dystrophy

Y.-H. Fu, A. Pizzuti, R. G. Fenwick, Jr., J. King, S. Rajnarayan, P. W. Dunne, J. Dubel, G. A. Nasser, T. Ashizawa, P. de Jong, B. Wieringa, R. Korneluk, M. B. Perryman, H. F. Epstein, C. Thomas Caskey*†

Synthetic oligonucleotides containing GC-rich triplet sequences were used in a scanning strategy to identify unstable genetic sequences at the myotonic dystrophy (DM) locus. A highly polymorphic GCT repeat was identified and found to be unstable, with an increased number of repeats occurring in DM patients. In the case of severe congenital DM, the paternal triplet allele was inherited unaltered while the maternal, DM-associated allele was unstable. These studies suggest that the mutational mechanism leading to DM is triplet amplification, similar to that occurring in the fragile X syndrome. The triplet repeat sequence is within a gene (to be referred to as myotonin–protein kinase), which has a sequence similar to protein kinases.

THE MOLECULAR BASIS OF GENETIC "anticipation," defined as the appearance of increasing disease severity in subsequent generations of a family with an inherited disorder was recently elucidated for the fragile X syndrome, where the phenomenon is referred to as the Sherman paradox (*1, 2*). A triplet (CGG) in the 5′ region of the FMR-1 gene is tandemly repeated 6 to 54 times in normal individuals, but occurs more than 200 times in patients with fragile X syndrome. The triplet repeat becomes unstable, particularly during female meiosis, once the repeat number exceeds about 52. The FMR-1 transcript that contains the repeat is not expressed in patients with the fragile X syndrome (*3*). A second disorder, spinal and bulbar muscular atrophy (Kennedy disease), has disease features associated with a CAG repeat expansion within the coding sequence of the androgen receptor gene (*4, 5*). Thus, fragile X and Kennedy syndromes are the consequence of a novel mechanism of mutation, expansion of GC-rich short tandem repeats within gene transcripts.

Myotonic dystrophy (DM) is a common autosomal dominant myopathy with pleiotropic effects including prolonged muscle contractions, cataracts, and cardiac arrhythmias (*6*). The severity of DM also increases over multiple generations (anticipation), suggesting a mutational mechanism similar to that found in fragile X and Kennedy syndromes. The symptoms of DM range from asymptomatic adult heterozygotes to newborns who have severe disease associated with hypotonia and retardation. These different clinical manifestations can occur within single families.

Myotonic dystrophy has been mapped to 19q13.2–13.3 by genetic linkage (*7, 8*) and a consensus genetic and physical map of the region is being prepared (*9–11*). We subcloned YAC clones 231G8 and 483E7 (*11*) into cosmids and human clones were identified by the presence of common repeat sequences (*12*). A mixture of four oligonucleotides consisting of tandemly repeated GC-rich trinucleotides (CAC, GCT, TCC, and TCG) hybridized to two overlapping cosmids out of 300 (MDY1 and MDY2). The set of four repeat oligonucleotide (each 21 nucleotides in length) included 24 of 60 possible triplet repeats with emphasis on GC-rich units. The CGG repeat was examined separately. A 1.4-kb Bam HI fragment that specifically hybridized to the GCT repeat was identified and subcloned into pBluescript (pMDY1). The sequence of pMDY1 was determined by dideoxynucleotide termination and an ABI 373 automated fluorescent DNA sequencer (Fig. 1) (*13, 14*). As predicted by oligonucleotide hybridization, a region containing 11 repeats of the GCT triplet was identified. This triplet is known to be highly polymorphic (*4*) and unstable (*15*) in the androgen receptor gene.

We chose to test genetic instability at the DM locus by studying families with congenital DM children born to affected DM mothers (Fig. 2). Family 1860 (Fig. 2B) shows a three-generation transmission of DM with progressive enlargement (8.8 to 12.7 kb) of an Nco I fragment. Sequence enlargements were also observed in the affected mothers from families 1127 and 1800. Other restriction endonucleases, including Ban I and Taq I, also identified fragment enlargements (Fig. 2A). With the resolution afforded by Nco I, we found 9 of 9 congenital DM patients and 14 of 16 adult DM patients who had fragment enlargements. An apparent exception is shown in Fig. 2B (family

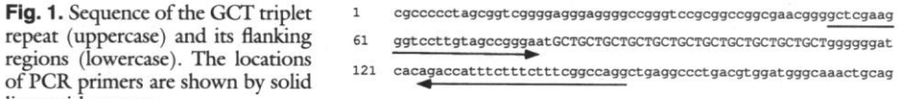
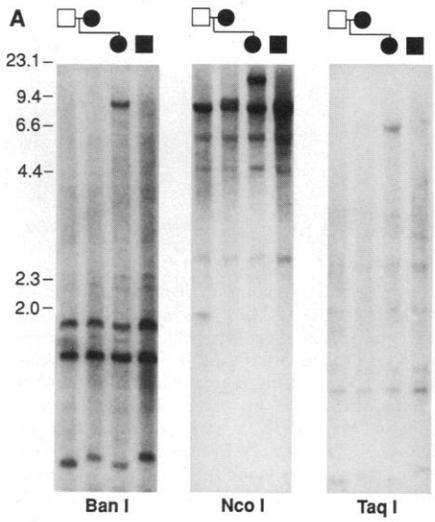


Fig. 1. Sequence of the GCT triplet repeat (uppercase) and its flanking regions (lowercase). The locations of PCR primers are shown by solid lines with arrows.



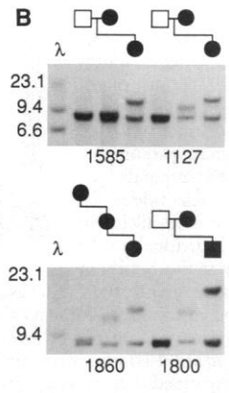


Fig. 2. Southern analysis of leukocyte DNA with probes containing the GCT repeat from MDY1. (**A**) After digestion with the three restriction enzymes indicated, DNA was hybridized with a probe containing the 1.4-kb Bam HI fragment from cosmid MDY 1. The enlarged sequence was detected in neither parent and, by examination of the Ban I data, was at least 6 kbp larger than sequences detected in the parents. (**B**) Samples from families in which a congenitally affected child had been born. The subcloned 1.4-kb Bam HI fragment from cosmid MDY1 was hybridized to Nco I–digested DNA. Sizes of markers are shown at the left in kilobase pairs.

Y.-H. Fu, A. Pizzuti, R. G. Fenwick, Jr., S. Rajnarayan, Institute for Molecular Genetics, Baylor College of Medicine, Houston, TX 77030.
J. King and C. T. Caskey, Institute for Molecular Genetics and Human Genome Center, Baylor College of Medicine, Houston, TX 77030.
P. W. Dunne, J. Dubel, G. A. Nasser, T. Ashizawa, H. F. Epstein, Department of Neurology, Baylor College of Medicine, Houston, TX 77030.
P. de Jong, Lawrence Livermore Laboratory, University of California, Livermore, CA 94550.
B. Wieringa, Department of Cell Biology and Histology, University of Nijmegen, Nijmegen, The Netherlands.
R. Korneluk, Department of Genetics, Children's Hospital, Eastern Ontario, Canada.
M. B. Perryman, Department of Cardiology, Baylor College of Medicine, Houston, TX 77030.

*To whom correspondence should be addressed.
†Also at Howard Hughes Medical Institute, Baylor College of Medicine, Houston, TX 77030.

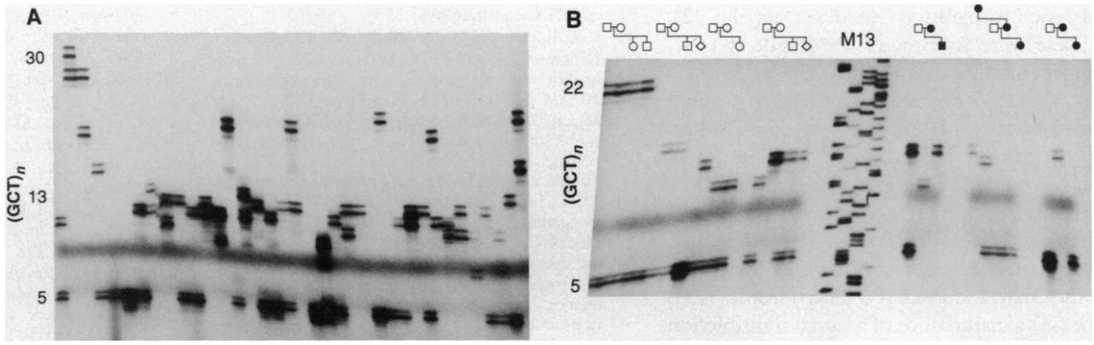


Fig. 3. (**A**) Polymorphic nature of the GCT locus in normal human genomic DNAs. Amplification of genomic DNA was carried out as described (*16*) and analyzed on a denaturing DNA sequencing gel. The electrophoretic pattern of each allele consists of two or three bands spaced at 3-base intervals. (**B**) GCT alleles determined by PCR in control and myotonic dystrophy families.

1585). However, subsequent studies with Bam HI revealed a small wild-type fragment (1.4 kb) that resolved sequence enlargements of 200 to 300 bp in patients whose alterations could not be detected in Nco I digests.

There were no fragment enlargements or reductions among 31 controls examined. Since each congenital DM patient had a unique enlarged restriction fragment that cannot be attributed to the parents, we conclude that this DNA sequence expansion is the mutational basis of DM. In each of these families, non-parentage was excluded by the linkage study.

To delineate the sequence involved in the DNA expansion, we examined the variation in GCT repeat size by amplification with the polymerase chain reaction (PCR) (*16*). Synthetic oligonucleotides that directly flank the GCT repeat (Fig. 1) were used to obtain the radioactive amplification products. The region is highly polymorphic (Fig. 3A). The most common allele consisted of five repeats with a range of 5 to 30 from 40 normal individuals analyzed; the heterozygote frequency was 85%. This length polymorphism was also observed by agarose gel analysis but with less detailed resolution. Examination of this sequence polymorphism in three DM and four control families is shown in Fig. 3B. Unaffected individuals had the expected frequency of pairs of alleles, while DM patients had only one allele, from the unaffected parent. Mendelian inheritance of alleles was observed in the control families. Thus, in these family studies, the DM GCT allele (as measured by PCR) was not detectable; a repeat sequence of >1 kb is beyond our current ability to amplify by PCR. DNA hybridization analysis indicated that each affected individual has a large expanded fragment. The simplest interpretation of these data is that the GCT repeat at the DM locus has meiotic instability and is responsible for the mutation in DM. We have examined by PCR the regions immediately flanking the GCT repeats indicated in Fig. 1, and have found them to be nonpolymorphic and unaltered in DM families.

We looked for open reading frames in the pMDY1 sequence by the computer program GRAIL (*17*). This program revealed an "excellent" exon identification score, possibly biased by the inclusion of triplet repeat sequences. PCR amplification of brain and skeletal muscle mRNA that had been reverse-transcribed into cDNA was used to identify products of the expected size supporting the computer prediction. The pMDY1 probe hybridized to several cDNA clones, including one of 3186 bp. Complete sequence of this cDNA clone revealed a unique predicted protein sequence, highly similar to numerous protein kinases (Fig. 4) (*18, 19*). The sequence has been submitted to GenBank (accession number M87312). As the putative protein from the DM cDNA contained a G-rich consensus signature that identifies an ATP binding domain (*20*) and a second consensus sequence for a serine-threonine protein kinase (*21*), it is likely that the protein is a protein kinase. We therefore designate it myotonin–protein kinase. Full molecular and cellular confirmation of this predicted function is requisite.

We propose that the sequence expansion associated with DM is an amplification of the GCT repeat which, from sequence data, resides in the 3′ untranslated portion of the putative protein kinase mRNA (Fig. 5). This would be consistent with the 3-base spacing of the polymorphic alleles (Fig. 3A) and direct sequence (Fig. 1). Present sequencing technology is incapable of confirming this speculation, because the DM expansions exceed 1 kb and direct sequence of GC-rich repeat regions is technically limited to 250 to 450 bp. A deletion or insertion mutation is considered to be unlikely since the sequence is unstable, expands in families, and is resistant to restriction endonuclease digestion (two enzymes that recognize 6-base elements and one that has a

```
                                    *********
MT-PK     80  DFEILKVIGRGAFSEVAVVKMKQTGQVYAMKIMNK
Bov-PKC   43  QFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDK
TPK2      69  DFQIMRTLGTGSFGRVHLVRSVHNGRYYAIKVLKK
                                  **********
MT-PK    199  LGYVHRDIKPDNILLDRCGHIRLADFG
Bov-PKC  161  LDLIYRDLKPENLLIDQQGYIQVTDFG
TPK2     187  HNIIYRDLKPENILLDRNGHIKITDFG
```

Fig. 4. Amino acid sequence comparison between myotonin-PK and two protein kinases. The letters in bold represent amino acid identities. TPK2 is the catalytic subunit of *Saccharomyces cerevisiae* cAMP-dependent protein kinase (*18*). Bov-PKC is the c_α form of the catalytic subunit of bovine cAMP-dependent protein kinase (*19*). The two domains contain the protein kinase signatures sequences (asterisks). The G-rich consensus signature is (LIV)-G-*x*-G-*x*-(FY)-(SG)-*x*-(LIV) and identifies an ATP-binding position (*20*). The second consensus sequence (function unknown) is (LIVMFYC)-*x*-(HY)-*x*-D-(LIVMFY)-K-*x*-*x*-N-(LIVMFC). This consensus identifies serine/threonine protein kinases (*21*). In the consensus for tyrosine kinases, R or A/S/T/ in position 7 substitutes for K.

```
                2467
.............. AAC CCT AGA ACT GTC TTC GAC TCC GGG GCC CCG TTG GAA GAC TGA GTG CCC GGG
.............. asn pro arg thr val phe asp ser gly ala pro leu glu asp STOP
2521  GCA CGG CAC AGA AGC CGC GCC CAC CGC CTG CCA GTT CAC AAC CGC TCC GAG CGT GGG TCT
2581  CCG CCC AGC TCC AGT CCT GTG ACC GGG CCC GCC CCC TAG CGG CCG GGG AGG GAG GGG CCG
2641  GGT CCG CGG CCG GCG AAC GGG GCT CGA AGG GTC CTT GTA GCC GGG AAT GCT GCT GCT GCT
2701  GCT GGG GGG ATC ACA GAC CAT TTC TTT CTT TCG GCC AGG CTG AGG CCC TGA CGT GGA TGG
2761  GCA AAC TGC AGG CCT GGG AAG GCA GCA AGC CGG GCC GTC CGT GTT CCA TCC TCC ACG CAC
2821  CCC CAC CTA TCG TTG GTT CGC AAA GTG CAA AGC TTT CTT GTG CAT GAC GCC CTG CTC TGG
2881  GGA GCG TCT GGC GCG ATC TCT GCC TGC TTA CTC GGG AAA TTT GCT TTT GCC AAA CCC GCT
2941  TTT TCG GGG ATC CCG CGC CCC CCT CCT ACT TGC GCT GCT CTC GGA GCC CCA GCC GCT CCG
3001  CCC GCT TCG GCG GTT TGG ATA TTT ATT GAC CTC GTC CTC CGA CTC GCT GAC AGG CTA CAG
3061  GAC CCC CAA CAA CCC CAA TCC ACG TTT TGG ATG CAC TGA GAC CCC GAC ATT CCT CGG TAT
3121  TTA TTG TCT GTC CCC ACC TAG GAC CCC CAC CCC CGA CCC TCG CGA ATA AAA GGC CCT CCA
3181  TCT GCC AAA AAA AAA AAA
```

Fig. 5. Sequence of the last 15 codons (including the stop) and the 3′ untranslated region of the MT-PK cDNA. The GCT repeat is underlined.

4-base recognition sequence) (Fig. 2). These were features of the fragile X CGG repeat amplification.

We propose several hypotheses that might explain the variability of symptoms in heterozygotes: (i) an effect of the amount of gene product, such as occurs with the low density lipoprotein receptor defect in type II hypercholesterolemia (*22*); (ii) differential parental inheritance of mutations, such as in Angelman and Prader-Willi syndromes (*23*); or (iii) a disturbance of a signal transduction pathway in which myotonin–protein kinase is only one of the disease-producing factors. Each of these hypotheses can be directly examined given our new molecular knowledge of myotonin–protein kinase.

These studies provide a simple method for identification of unstable genetic elements in the human genome. Although we used oligonucleotides as probes and nuclear DNA clones as targets, it is logical to search for other unstable genes by screening cDNA libraries for GC-rich triplet repeats. The lessons of fragile X syndrome, Kennedy disease, and now DM are consistent. Heritable disorders that exhibit the feature of anticipation or molecular imprinting (*24, 25*) would appear worthy of investigation as reported here for DM. Furthermore, since somatic genetic instability is demonstrated for the CGG repeat in the fragile X syndrome, genes containing unstable repeats may be involved in neoplasia and possibly aging, in which somatic mutations are implicated in disease.

REFERENCES AND NOTES

1. A. Verkerk *et al.*, *Cell* **65**, 905 (1991).
2. Y.-H. Fu *et al.*, *ibid.* **67**, 1047 (1991).
3. M. Pieretti *et al.*, *ibid.* **66**, 817 (1991).
4. A. Edwards, A. Civitello, H. A. Hammond, C. T. Caskey, *Am. J. Hum. Genet.* **49**, 746 (1991).
5. A. R. La Spada *et al.*, *Nature* **352**, 77 (1991).
6. P. S. Harper, *Myotonic Dystrophy* (Saunders, London, ed. 2, 1979).
7. K. Johnson *et al.*, *Am. J. Hum. Genet.* **46**, 1073 (1990).
8. H. J. Smeets *et al.*, *Genomics* **9**, 257 (1991).
9. G. Shutler *et al.*, *ibid.*, in press.
10. G. Jansen *et al.*, *ibid.*, in press.
11. C. Aslanidis *et al.*, *Nature* **355**, 548 (1992).
12. YACs 231G8 and 483E7 DNA were partially digested by Sau 3A and cloned into cosmid vector "Super Cos" (Stratagene) as described by the manufacturer. Human clones, identified by hybridization with radiolabeled total human DNA, were selected and arrayed on a gridded plate. Duplicate filter lifts were screened for specific triplet repeats by their hybridization to a mixture of four radiolabeled oligonucleotides. Two positive cosmids (MDY1 and MDY2) were identified on the grid and were found to contain sequences in common, including Bam HI fragments of 1.4 and 1.35 kb. The 1.4-kb Bam HI fragment contains the triplet repeat sequence.
13. R. A. Gibbs, P.-N. Nguyen, A. Edwards, A. Civitello, C. T. Caskey, *Genomics* 7, 235 (1990).
14. Sequence of pMDY1 was determined with a combination of dideoxynucleotide termination sequencing and the Taq DyeDeoxy terminator cycle sequencing reaction (Applied Biosystems). The sequencing reactions were analyzed on an automated DNA sequencer (ABI 373).
15. J. R. Lupski and C. T. Caskey, unpublished data.
16. Genomic DNAs (100 ng) were mixed with 3 pmol of each primer in a total volume of 15 μl containing 10 mM tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 200 μM of each of the 4 dNTPs, 4 μCi of α-[^{32}P]dCTP, and 0.75 units of AmpliTaq DNA polymerase. The reactions were heated to 95°C for 10 min and followed by 25 cycles of denaturation (95°C, 1 min), DNA reannealing (54°C, 1 min), and elongation (72°C, 2 min). The radioactive PCR products were combined with 95% formamide loading dye and then heated to 95°C for 2 min before electrophoresis through a 6% denaturing DNA sequencing gel. Allele sizes were determined by migration relative to an M13 sequencing ladder. For analysis by 3% agarose gel electrophoresis, 200 ng of genomic DNA were amplified in a final volume of 100 μl using the same buffer, 250 μM of the 4 dNTPs and 0.5 unit of AmpliTaq DNA polymerase. The reactions were heated to 95°C for 5 min and then subjected to 32 cycles of 94°C for 1 min 57°C for 1 min, and 72°C for 3 min.
17. GRAIL (Gene Recognition and Analysis Internet Link) computer searches are available to general users via the Oak Ridge National Laboratory File server at GRAIL@ornl.gov.
18. T. Toda *et al.*, *Cell* **50**, 277 (1987).
19. S. Shoji *et al.*, *Biochemistry* **22**, 3702 (1983).
20. M. P. Kamps, S. S. Taylor, B. M. Sefton, *Nature* **310**, 589 (1984).
21. S. K. Hanks, A. M. Quinn, T. Hunter, *Science* **241**, 42 (1988).
22. J. L. Goldstein and M. S. Brown, in *Metabolic Basis of Inherited Disease*, C. Scriver, A. L. Beaudet, D. Valle, W. Sly, Eds. (McGraw-Hill, New York, 1989), chap. 48.
23. J. G. Hall, *Am. J. Hum. Genet.* **46**, 357 (1990).
24. R. M. Ridley, C. D. Frith, L. A. Farrer, P. M. Conneally, *J. Med. Genet.* **28**, 224 (1991).
25. H. Y. Zoghbi *et al.*, *Ann. Neurol.* **23**, 580 (1988).
26. Abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

27. A.P. is a Muscular Dystrophy Association Fellow; C.T.C. is a Howard Hughes Medical Institute Investigator, S.R. is supported by NIH grant 5-M01-RR00350 to the Baylor College of Medicine General Clinical Research Center at the Methodist Hospital. This work was supported by U.S. Department of Energy grant DE-FG05-88ER60692 (C.T.C.), a New Neuromuscular Disease Research Development Grant (R.G.F.), a Veterans' Administration Merit review award (T.A.), a Muscular Dystrophy Association Task Force Grant (H.F.E.), and National Heart, Blood, and Lung Institute grant P50HL42267-01 (H.F.E. and M.B.P.). Automated sequencing was supported by NIH grant P30-HG00210 and the W. M. Keck Center for Computational Biology at Baylor College of Medicine and Rice University. A.P. is on leave from Instituto di Clinica Neurologica Università di Milano, Centro "Dino Ferrari," Milano, Italy. We thank J. F. Hejtmancik for contributions to the early phases of this research.

21 January 1992; accepted 14 February 1992

The Linguistic Basis of Left Hemisphere Specialization

DAVID P. CORINA,* JYOTSNA VAID, URSULA BELLUGI

In humans the two cerebral hemispheres of the brain are functionally specialized with the left hemisphere predominantly mediating language skills. The basis of this lateralization has been proposed to be differential localization of the linguistic, the motoric, or the symbolic properties of language. To distinguish among these possibilities, lateralization of spoken language, signed language, and nonlinguistic gesture have been compared in deaf and hearing individuals. This analysis, plus additional clinical findings, support a linguistic basis of left hemisphere specialization.

THE LEFT HEMISPHERE OF THE HUman brain is specialized for language. The underlying basis of this specialization has been controversial, and it has not been clear if this brain system is uniquely designed for language processing or if it derives from a more general specialization based on motor control (*1*) or symbolization (*2*). Until recently most of our knowledge regarding hemispheric specialization for language has come from the study of spoken languages. In contrast, we have now addressed these competing hypotheses by studying native users of American Sign Language (ASL) (*3, 4*).

ASL is a natural language with structural properties akin to those of spoken languages (*5–10*). After left hemisphere injury deaf signers exhibit sign language aphasia, and right hemisphere damage can result in severe visuospatial disruption but leaves signing intact (*3*). Thus, despite auditory deprivation, deaf users of a signed language show a complementary hemispheric specialization like that of spoken language users. Some researchers have used this evidence to suggest that the left hemisphere is uniquely predisposed for mediation of language, both spoken and signed (*11*). Others argue that left hemisphere specialization for signed and spoken language derives from the left hemisphere's more general role in controlling changes in the position of oral and manual articulators (*12*). Under this interpretation, any skilled motoric movement, such as the execution

D. P. Corina and U. Bellugi, the Salk Institute for Biological Studies, La Jolla, CA 92037.
J. Vaid, Texas A&M University, College Station, TX 77843.

*Present address: Program in Neural, Informational, and Behavioral Sciences, University of Southern California, HNB 18C, University Park, Los Angeles, CA 90089–2520.