Additional Sequence Complexity in the Muscle Gene, *unc-22*, and Its Encoded Protein, Twitchin, of *Caenorhabditis elegans*

Guy M. Benian,* Steven W. L'Hernault† and Mary E. Morris*,[1]

**Department of Pathology and †Department of Biology, Emory University, Atlanta, Georgia 30322*

Manuscript received July 14, 1992
Accepted for publication April 2, 1993

ABSTRACT

Null mutations of the *Caenorhabditis elegans unc-22* gene cause a pronounced body surface twitch associated with impaired movement and disruption of muscle structure. Partial sequence analysis of *unc-22* has previously revealed that its encoded polypeptide, named twitchin, consists of a single protein kinase domain and multiple copies of both an immunoglobulin-like domain and a fibronectin type III-like domain. This paper reports additional DNA sequence information that has revealed the transcription start of *unc-22,* the N terminus of twitchin, and an explanation for the weak phenotype of a transposon insertion allele. These new data indicate that the *unc-22* gene is 18 kb larger than previously reported and has a transcription unit of 38,308 bp. These data add 791 amino acids to the twitchin N terminus for a complete polypeptide size of 6,839 amino acids and a predicted molecular weight of 753,494. This new polypeptide sequence includes four additional copies of the above-mentioned immunoglobulin-like domains and also includes a glycine-rich sequence that might form a flexible hinge. The additional coding sequence reveals that the insertion of the Tc*1* transposon, in the *unc-22* allele, *st139,* should disrupt twitchin structure because it is located in an exon. However, cDNA sequencing has revealed that several cryptic splice donors and acceptors adjacent to the Tc*1* insertion site are used to splice the transposon out of *unc-22(st139)* mRNA. One of these splicing events produces a near wild-type mRNA that deletes only six amino acids from twitchin, and this might explain the unusually mild phenotype associated with this mutation.

unc-22 was originally identified mutationally as one of about 40 genes that are important for the assembly and function of muscle in *Caenorhabditis elegans* (Waterston 1988). Worms carrying mutant alleles of *unc-22* show varying degrees of impaired movement and muscle structure disorganization, but all alleles show an intriguing constant twitch of the body surface that originates in the underlying muscle (Waterston, Thomson and Brenner 1980; Moerman 1980). "Twitching" suggested that the *unc-22* product was somehow involved in regulating contraction, and genetic reversion analysis suggested that the *unc-22* product might interact with myosin (Moerman *et al.* 1982). The *unc-22* gene was cloned by transposon tagging (Moerman, Benian and Waterston 1986) and shown to specify a very large polypeptide (Moerman *et al.* 1988). This protein, called "twitchin," was localized to muscle A-bands by immunofluorescence (Moerman *et al.* 1988). DNA sequence analysis showed that twitchin consists of a single protein kinase domain and multiple copies of both a fibronectin type III-like domain (motif I) and an immunoglobulin-like domain (motif II) (Benian *et al.* 1989).

Twitchin was the first intracellular protein recognized as a member of the immunoglobulin superfamily. Subsequently, a number of intracellular proteins from several animal species have been shown to be composed of multiple copies of motif I and motif II in their polypeptide sequences. Presently, this family includes the vertebrate proteins smooth muscle (Olson *et al.* 1990) and non-muscle (Shoemaker *et al.* 1990) myosin light chain kinases (MLCK), telokin or kinase-related protein (Gallagher and Herring 1991; Collinge *et al.* 1992) titin (Labeit *et al.* 1990), C-protein (Einheber and Fischman 1990), 86-kD protein (Fischman *et al.* 1991), skelemin (Price 1987; Price, Brooks and Gomer 1990), and M-protein (Noguchi *et al.* 1992), and the insect protein called projectin (Ayme-Southgate *et al.* 1991; Fyrberg *et al.* 1992). Telokin is an abundant protein in gizzard smooth muscle, encoded by a portion of the MLCK gene, and is identical to the C-terminal 155 residues of smooth muscle MLCK. The recent report of the X-ray crystal structure of telokin (Holden *et al.* 1992) indicates that the Ig-like sequences of members of this muscle protein family form Ig folds. Extracellular and cell surface proteins engaged in recognition or adhesion contain domains that are very similar to motifs I and II of these muscle proteins, suggesting that these motifs are involved in protein binding inside muscle cells. There is evidence that bacterially expressed titin motifs interact *in vitro* with myosin and C-protein (Labeit *et al.* 1992).

[1] Present address: Program in Neuroscience/Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115.

Initial sequence analysis suggested that twitchin is comprised of 6,048 amino acids with a molecular weight of 668,520 (BENIAN *et al.* 1989). Here we report the 5′ end of the *unc-22* message and the N terminus of twitchin. This adds 791 amino acids to twitchin, resulting in a polypeptide composed of 6839 amino acids with an estimated molecular weight of 753,494. This makes twitchin the largest polypeptide for which both the complete amino acid sequence and the complete exon/intron organization are known.

MATERIALS AND METHODS

Nematode strains, RNA and protein isolation: The wild-type strain of *C. elegans* var. Bristol was obtained from the *C. elegans* Stock Center. The wild-type revertant of the Tc1 transposon allele *st139* was kindly provided by R. H. WATERSTON. Worms were maintained and grown in large quantity by standard procedures (BRENNER 1974). Total RNA was isolated by slight modification of published procedures (CHIRGWIN *et al.* 1979). A high salt extract that was enriched for myofilament proteins was prepared by minor modifications of the procedure described by EPSTEIN, WATERSTON and BRENNER (1974). Worms that had been stored at −20° in 50% glycerol were washed twice in 7–10 volumes (1 volume = 1 volume of packed worms) of wash buffer (EPSTEIN, WATERSTON and BRENNER 1974), which involved thorough mixing and then pelleting at 3000 × *g* for 3 min. The pellet was suspended with a glass rod in 5 volumes of low salt buffer (LSB: 100 mM KCl, 2 mM $MgCl_2$, 5 mM EGTA, 1 mM 2-mercaptoethanol, 10 mM Tris-HCl, 2 mM $Na_2P_2O_7$, pH 6.8), and passed three times through a French pressure cell at 10,000–12,000 psi. After adding 1 volume of LSB and mixing, the homogenate was centrifuged at 3000 × *g* for 10 min and the supernatant discarded. The pellet was resuspended in 3 volumes of LSB, again centrifuged at 3000 × *g* for 10 min and the supernatant discarded. This washing of the pellet was repeated, and the final pellet was resuspended in 2 volumes of ice-cold "extracting solution" (0.6 M KCl, 2 mM $MgCl_2$, 1 mM EDTA, 10 mM imidazole-HCl, pH 7.0, 0.5 mM dithiothreitol) and incubated on ice for 20 min. This mixture was centrifuged at 15,000 × *g* for 1 hr, and the resulting supernatant was called "high salt extract."

DNA sequencing, cDNA library screen and polymerase chain reactions (PCR): The genomic sequence of 7825 bp just upstream of the previously published *unc-22* region sequence was determined by dideoxy sequencing (SANGER, NICKLEN and COULSON 1977) of both strands of pSL8 and pSL13, which are Bluescript subclones of cosmid C13G4 (COULSON *et al.* 1986) The cDNA clone pSL12 was obtained from a λZAP cDNA library (kindly provided by R. BARSTEAD) screened with the pSL13 insert. A Bluescript plasmid version of pSL12 was obtained by utilizing standard procedures for *in vivo* excision (Stratagene) and this plasmid was sequenced on one strand. PCR on first-strand cDNA was used to determine coding sequence between the end of pSL12 cDNA at 9439 and coding sequence deduced from the genomic sequence at 21,952. First-strand cDNA was generated with avian myeloblastosis virus (AMV) reverse transcriptase at 42° from 10 μg of total RNA with the antisense oligonucleotide GB3L (GGACGTGAGTAGGTACACTTC; position 22448–22428). This oligo and all other oligos were synthesized at the Emory University Microchemical Facility. The PCR included 10% of the above first-strand cDNA product as template DNA and 10–30 pmol each of GB3L and the sense oligonucleotide PCR3 (GTGGAAGAAGAACTCATGAAGC; position 9391–9412) and was performed for 35 cycles of denaturation at 94° for 1 min, annealing at 63° for 2 min and extension at 72° for 3 min. Conditions for PCR were essentially those described by Perkin-Elmer Cetus. Two prominent bands of 1.1 and 0.45 kb were synthesized. After separation on an agarose gel, each band was isolated by Geneclean (BIO 101 Inc., LaJolla, California), end-repaired with Klenow fragment, phosphorylated with T4 polynucleotide kinase, ligated into *Sma*I-cut and phosphatased mp10M13 (Amersham) and transformed into competent JM101 cells. The sequence of the 0.45-kb product did not correspond to any sequence in the 55-kb *unc-22* region and was not further analyzed. The sequence of the 1.1-kb product contained both primers and corresponded to the genomic sequence as indicated in RESULTS.

Reverse transcription PCR was also used to sequence the *unc-22* mRNAs from *st139*, the *st139* revertant (RW2418) and wild type. cDNA was generated with AMV reverse transcriptase at 42° from 10 μg of total RNA with the antisense oligonucleotide ST139R (TAGTCACACTTTGCATGACATTT; position 9160–9138). One-tenth of this cDNA was used as template DNA for PCR with ST139R and the sense oligonucleotide ST139F (GTCGAATCGGTGGATCTGTC; position 8281–8300) and was performed for 40 cycles of denaturation at 94° for 1 min, annealing at 57° for 2 min and extension at 72° for 30 sec. ST139F and ST139R lie on either side of the 678-bp intron no. 7 and were expected to produce a PCR product of 202 bp from mRNA. Indeed, single bands of about 200 bp were synthesized from both wild type and the *st139* revertant. From *st139*, two prominent bands close to 200 bp and at least three larger bands of lesser abundance were produced. Approximately 10% of the PCR products from each strain were ligated to the pT7Blue(R) T-vector using the single 3′ A-nucleotide (nt) overhangs created by Taq polymerase in the PCR products (Novagen) and transformed into the manufacturer's NovaBlue competent *Escherichia coli*. Plasmid inserts from the following number of colonies were sequenced from each strain: nine from wild type, three from revertant and nine from *st139*.

Primer extension: The following two antisense oligonucleotides were synthesized from sequence near the 5′ end of cDNA pSL12: primer PEX1 (TGGGTGAAGCGCGGTGCGCCAACC; position 4112–4089) and primer PEX2 (CGATCAAACTGTCTCTTCTATGACG; position 4049–4025). They were end-labeled with [γ-^{32}P]ATP via polynucleotide kinase. Ten micrograms of total RNA were mixed with 20 ng of labeled oligonucleotide and annealed for 10 min at 65° and then for 1 hr at 37°, as described by DRISCOLL *et al.* (1989). cDNA was synthesized by using avian myeloblastosis virus reverse transcriptase (Bethesda Research Laboratories) for 40 min at 45° essentially as described by KRUG and BERGER (1987). After heat denaturation of the enzyme, the DNA was ethanol precipitated with carrier tRNA and resolved on a urea-5% acrylamide sequencing gel. The size of the primed cDNA products was determined by comparison with the sequence of M13mp18 (−40 universal primer) run on the same gel.

Expression of the first unique domain of twitchin, generation of antiserum and Western blotting: The 167 amino acid-long first unique domain of twitchin (residues 207–373) was expressed as a glutathione *S*-transferase (GST) fusion protein by use of the pGEX-2T vector (SMITH and JOHNSON 1988). DNA encoding this segment was synthesized by two rounds of PCR templated from pSL12. The first round used a 26-mer including 20 nt of *unc-22* sequence plus added *Bam*HI site at the 5′ end, and an antisense 28-mer including 21 nt of *unc-22* sequence plus in-frame stop codon and added *Eco*RI site at the 3′ end. The second round of PCR utilized similar primers missing 8 nt from their 3′ ends but having 4 nt of unrelated sequence added 5′ of the *Bam*HI or *Eco*RI sites. The amplified 500-bp fragment was purified from an agarose gel, ligated into *Bam*HI/*Eco*RI cut pGEX-2T, and transformed into BL21 (DE3) *E. coli*. Candidate recombinant clones were sequenced to assure no errors were introduced by PCR. The *GST* fusion protein of one clone, "500GEX1," was expressed and purified with a glutathione-agarose batch method essentially as described by SMITH and JOHNSON (1988). Isopropyl-β-D-thiogalactopyranoside induction was conducted at room temperature because pilot experiments showed that 37° induction resulted in insoluble protein. One rabbit antiserum was generated by dividing approximately 300 μg of fusion protein emulsified with Hunter's TiterMax #R-1 adjuvant (CytRx Corp., Norcross, Georgia) among three to four intramuscular injections, and repeating this inoculation after 1 month. Immune serum was collected 8 days after the last injection and subjected to affinity purification. The affinity column was prepared by coupling 15 mg of fusion protein from 500GEX1 to 1 ml of a 50/50 mixture of Affi-Gel-10 and Affi-Gel-15 (Bio-Rad) using the manufacturer's directions. This column was prewashed with 10 ml volumes of each of the following, in succession: 10 mM Tris, pH 7.5, 100 mM glycine, pH 2.5, 10 mM Tris, pH 8.8, 100 mM triethylamine acetate, pH 11.5, and 10 mM Tris, pH 7.5, until pH returned to 7.5. One milliliter of the antiserum was cleared of anti-GST antibodies by performing three consecutive immunoprecipitations with 400 μg each of pure GST. Cleared serum was diluted to 5 ml with 10 mM Tris, pH 7.5, and recirculated through the affinity column 10 times. The column was then washed with 20 ml of 10 mM Tris, pH 7.5, followed by 20 ml of 10 mM Tris, pH 7.5, and 500 mM NaCl to remove nonspecifically bound proteins. The bound antibodies were eluted into a tube containing 2 ml of 1 M Tris, pH 7.5, by passage of 10 ml of 100 mM glycine, pH 2.5, washing with 10 ml of 10 mM Tris, pH 8.8, and passage of 10 ml of 100 mM triethylamine acetate, pH 11.5. These affinity-purified antibodies were then concentrated to a volume of 1 ml in a Centriprep-30 concentrator (Amicon), and washed twice with 10 mM Tris, pH 7.5.

Western blotting was conducted as follows. Approximately 20 μl of "high salt extract" (see above), mixed 1:1 with 2 × Laemmli loading buffer, were loaded per lane, separated on a 5% polyacrylamide sodium dodecyl sulfate (SDS) Laemmli gel (LAEMMLI 1970), transferred to nitrocellulose membrane (TOWBIN, STAEHELIN and GORDON 1979) for 2 hr according to manufacturer's instructions (Bio-Rad), stained for 5 min with Ponceau S Solution (Sigma), destained in 5% acetic acid, and rinsed with water. Blocking was performed at 37° for 4–5 hr in 5% non-fat dry milk, 1% bovine serum albumin; this solution was also used for dilution of primary and secondary antibodies. The affinity-purified anti-500 GEX antibodies were used at a 1:200 dilution, and the "positive control" R11–3 antibodies, previously shown to react to twitchin (MOERMAN *et al.* 1988) were used at a 1:400 dilution. The secondary antiserum was a peroxidase-conjugated donkey anti-rabbit immunoglobulin, and detection was by enhanced chemiluminescence (Amersham).

RESULTS

The 5′ end of the *unc-22* mRNA and the N terminus of twitchin: Previously, analysis of the 5′ end of the *unc-22* gene was based solely on genomic sequence, and it suggested that the initiator methionine codon was located at position 21,952 (BENIAN *et al.* 1989; Figure 1). The result placed the translation start about 13 kb from the known position of transposon insertion of the *unc-22* allele *st139* (MORI *et al.* 1988). Subsequently, Northern hybridizations suggested that the transcription start was near position 4,000 (Figure 1) and an additional 7,825 bp of genomic sequence was determined in order to search for additional exons (GenBank accession no. L10351). During a screen for cDNA clones for the nearby *spe-17* gene, which affects spermatogenesis (L'HERNAULT, BENIAN and EMMONS 1993), one cDNA clone, pSL12 (1598-bp insert), that hybridized to the *unc-22* giant message was obtained (data not shown). This cDNA was sequenced and comparison to genomic sequence revealed that it contained 9 exons distributed between positions 3960 and 9439 (see Figure 1). The remaining gap between the 3′ end of cDNA pSL12 and coding sequence beginning at 21,952 was determined by PCR on first-strand cDNA, and this resulted in a single product of 1.1 kb. The sequence of this PCR-generated cDNA defines 10 new exons, the smallest being 54 bp (Figure 1). The 17 new introns conform to typical *C. elegans* introns (BLUMENTHAL and THOMAS 1988) in their donor and acceptor sequences, high AT content, and except for the remarkable 7402-bp intron no. 10, their small sizes. A methionine codon lies near the beginning of the pSL12 sequence (genomic position 4087) and its upstream sequence matches at three of four positions a consensus sequence for translation initiation for *C. elegans* (M. D. PERRY, G. Z. HERTZ and W. B. WOOD, personal communication). The pSL12 sequence begins with the last 9 nt of the 22-nt SL1 *trans*-spliced RNA leader (KRAUSE and HIRSH 1987), indicating that we have defined the 5′ end of the *unc-22* transcript. The mRNA size was also determined by primer extension analyses. Two different primers (5′ primer ends: 4112 and 4049 on sequence in Figure 1) yielded, respectively, major products of 175 and 112 nt (Figure 2). The size of these primer extension products corresponds precisely with the mRNA size predicted by cDNA pSL12, if the full length of the SL1 *trans*-spliced leader is included. This predicts that SL1 is *trans*-spliced onto *unc-22* genomic sequence at position 3960, and this position is preceded by the sequence TTCCAG, which is presumed to function as an intron acceptor. These results indicate that the *unc-22* 5′-untranslated sequence is 149 nt, including the 22-nt SL1 *trans*-spliced RNA leader. The pSL12 and PCR-derived cDNA sequences add 2522 nt to the *unc-22* message and 791 amino acids to the deduced twitchin polypeptide sequence. This results in an mRNA of 21,614 nt and a polypeptide of 6839 amino acids with a predicted molecular weight of 753,494 (Figure 3). The pI is predicted to be 5.50 (IBI Pustell sequence analysis programs, 1989).

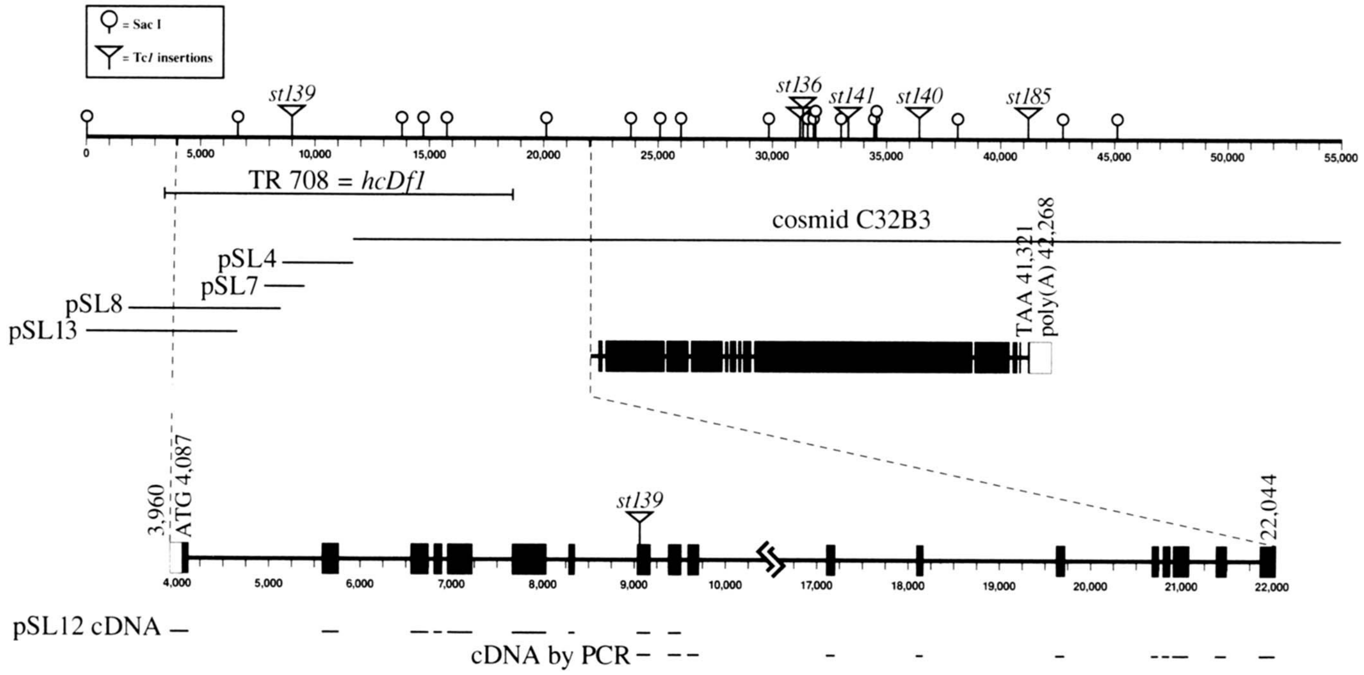


FIGURE 1.—Restriction map, sites of transposon Tc1 insertion, extent of TR708 deletion, DNA clones and exon map of the *unc-22* gene. The insertion sites for the Tc*1* alleles shown have been sequenced by MORI *et al.* (1988). The TR708 = *hcDf1* deletion (courtesy of JOHN COLLINS and PHIL ANDERSON) causes both *unc-22* and *spe-17* phenotypes. Thin lines, introns; open boxes and black boxes, exons of noncoding and coding regions, respectively. The genomic sequence between positions 3960 and 22044 has been expanded 4-fold to more clearly show the numerous small exons in this region. The cosmid C32B3 and plasmids pSL4 and pSL7 had been sequenced previously and were thought to probably contain most of *unc-22* (BENIAN *et al.* 1989). However, that analysis suggested a likely initiator methionine codon at position 21,952, which is 13 kb downstream of the transposon insertion site in the *unc-22* allele *st139*. Moreover, Northern blots indicated that part of *unc-22* coding sequence was well beyond the left end of cosmid C32B3. During the hunt for a *spe-17* cDNA (which is in the region deleted by TR708), an *unc-22* cDNA, called pSL12, was fortuitously recovered with the genomic clone pSL8. Consequently, additional genomic clones (pSL8 and pSL13) were sequenced (giving a new total of 54,963 bp of continuous genomic sequence, including the correction of several previous errors) and the distribution of cDNA sequence derived from PCR (total of 1100 bp) and from cDNA clone pSL12 (total of 1598 bp) were determined.

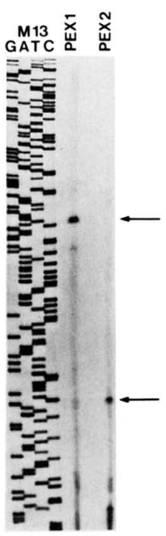


FIGURE 2.—Primer extension of the *unc-22* mRNA. Lane 5 shows the 175-nt fragment obtained with primer PEX1 (begins at 4112) and lane 6 shows the 112-nt fragment obtained with primer PEX2 (begins at 4049), both indicated by arrows. Lanes 1–4 are the sequence of M13 mp18 with the −40 universal primer. These are the sizes of primer extension products expected from the pSL12 cDNA sequence beginning its correspondence to genomic sequence at 3960 and *trans*-splicing to the 22-nt SL1 RNA.

Additional sequence domains near the N terminus of twitchin: Figure 3 is an up-dated schematic of the deduced amino acid sequence of twitchin. Dark boxes correspond to motif I copies and light boxes correspond to motif II copies. The dashed horizontal line indicates the new sequence that has been added to the N terminus. This includes four more copies of motif II, and an alignment with the previously derived consensus sequence for motif II is presented in Figure 4; there are a total of 30 copies of motif II. The new sequence also has two segments of 167 residues and 189 residues (clear boxes between motif II copies 2′ and 3′, and 3′ and 4′ in Figure 3) that have no significant homologies to previously determined sequences in twitchin (BENIAN *et al.* 1989) or to sequences in the computer databases (using FASTA (PEARSON and LIPMAN 1988) on SWISS-PROT release no. 20 and GenPept release no. 70). These unique sequences appear to be enriched for basic amino acids (22% and 20%, respectively), proline (11.4% and 7.4%) and serine (23% and 17%). The twitchin sequence outside these unique domains has only 13.9% basic residues, 6.9% proline and 5.5% serine. Just preceding the five tandem copies of motif

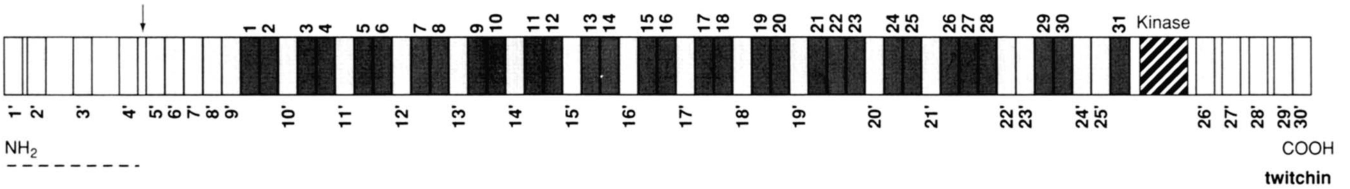
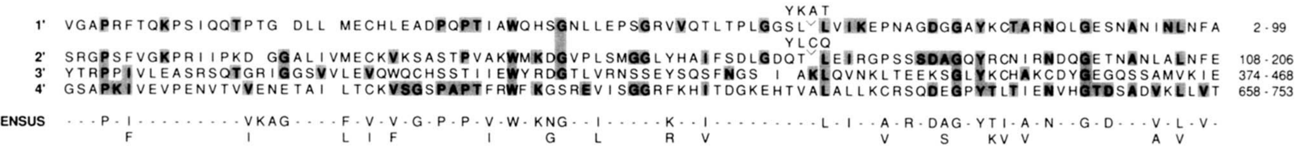


FIGURE 3.—Up-dated schematic of domains in the deduced amino acid sequence of twitchin. Dark boxes correspond to motif I and light boxes correspond to motif II. The dashed horizontal line indicates the 791 amino acids that we have added to the N terminus. This includes four more copies of motif II and two segments of 167 amino acids and 189 amino acids which have no homologies to sequences in twitchin or to sequences in the databases. An arrow indicates a glycine-rich segment (9/11 amino acids beginning at position 777) which could possibly form a flexible hinge.



```
                                                                       YKAT
1'         VGAPRFTQKPSIQQTPTG DLL MECHLEADPQPTIAWQHSGNLLEPSGRVVQTLTPLGGSL LVIKEPNAGDGGAYKCTARNQLGESNANINLNFA   2-99
                                                                       YLCQ
2'         SRGPSFVGKPRIIPKD GGALIVMECKVKSASTPVAKWMKDGVPLSMGGLYHAIFSDLGDQT LEIRGPSSSDAGQYRCNIRNDQGETNANLALNFE   108-206
3'         YTRPPIVLEASRSQTGRIGGSVVLEVQWQCHSSTIIEWYRDGTLVRNSSEYSQSFNGS I AKLQVNKLTEEKSGLYKCHAKCDYGEGQSSAMVKIE   374-468
4'         GSAPKIVEVPENVTVVENETAI LTCKVSGSPAPTFRWFKGSREVISGGRFKHITDGKEHTVALALLKCRSQDEGPYTLTIENVHGTDSADVKLLVT   658-753

CONSENSUS  ---P-I---------VKAG----F-V-V-G-P-P-V-W-KNG--I-----K--I---------L-I--A-R-DAG-YTI-A-N--G-D---V-L-V-
                F         I       L I F       I     G  L     R  V              V    S   KV V          A V
```

FIGURE 4.—The four new copies of motif II aligned with the previously derived consensus sequence (BENIAN *et al.* 1989).

II (N-terminal to motif II copy 9′) is a glycine-rich segment, indicated by an arrow in Figure 3; 9/11 amino acids are glycine beginning at residue 777.

Of the 10 sites which match the optimal substrate sequence for cAMP-dependent protein kinases (R-R/K-*X*-S/T; KENNELLY and KREBS 1991), 9 fall at the ends of twitchin (positions 216, 281, 337, 599, 628, 629, 636, 797, 6516), and within unique segments (not motif I or II or kinase sequences). The one exception occurs in motif II copy 13. Searches for recognition motifs for smooth muscle and skeletal muscle myosin light chain kinases and myosin-I heavy chain kinase (KENNELLY and KREBS 1991) were negative. A PROSITE (IntelliGenetics, Inc.) search yielded many possible sites with no interesting patterns for protein kinase C, casein kinase II and tyrosine kinases.

Because the new coding sequence is distributed over an 18-kb region, we wanted to be certain that it is indeed part of the same *unc-22* gene. The first piece of evidence is that, as mentioned above, the pSL12 cDNA hybridizes to the large *unc-22* mRNA on Northern blots. The second piece of evidence involves expression of the first unique portion of twitchin (residues 207–373) in *E. coli* as a glutathione *S*-transferase fusion protein. A rabbit polyclonal antiserum raised against and purified to this fusion protein (*a*500GEX) reacts with twitchin on a Western blot (Figure 5).

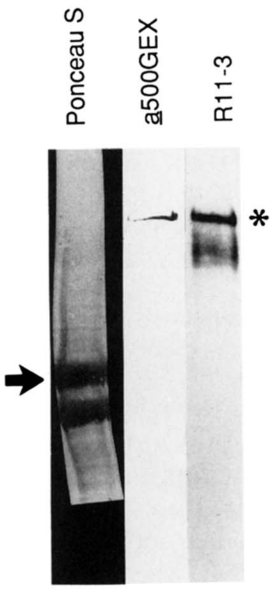


FIGURE 5.—Antibodies to polypeptide sequence encoded by cDNA pSL12 react with intact twitchin. Samples from a whole worm "high salt extract" were separated on a 5% polyacrylamide SDS gel, transferred to nitrocellulose, and either stained with Ponceau S for total protein, or reacted with the indicated antibodies and visualized with enhanced chemilumenescence. *a*500GEX is a rabbit antiserum raised against the first unique portion of the deduced twitchin polypeptide sequence (residues 207–373) expressed in *E. coli* as a glutathione-*S*-transferase fusion protein. R 11-3 is a rabbit antiserum made to a central portion of *unc-22* coding sequence and shown previously, by analysis of *unc-22* mutants, to react specifically with twitchin (MOERMAN *et al.* 1988). An asterisk indicates the position of twitchin (750 kD). The arrow indicates the position of nematode myosin heavy chain (210 kD). The second prominent band on Ponceau S staining likely represents a degradation product of myosin heavy chain.

Molecular basis for the phenotype of the transposon allele *st139*: Our data indicate that the insertion site of the Tc*1* transposon in the *unc-22* allele *st139* (MORI *et al.* 1988) is within the *unc-22* coding region, residing in the TAC codon for threonine (amino acid 416, in the middle of the third copy of motif II, Figure 3), between genomic positions 9038 and 9039. Of all the *unc-22* alleles examined, *st139* has the weakest phenotype. In contrast to 5 other Tc1 transposon alleles that result in larger, Tc*1-unc-22* composite messages, *st139* results in a wild-type-sized message, not containing Tc*1* sequences (MOERMAN *et al.* 1988). The weak phenotype of *st139* might result from a large fraction of the composite messages undergoing a splicing-out of Tc*1*, since the Tc*1* resides just 7 bp downstream of an intron (no. 7). This suspicion was confirmed by reverse transcriptase PCR

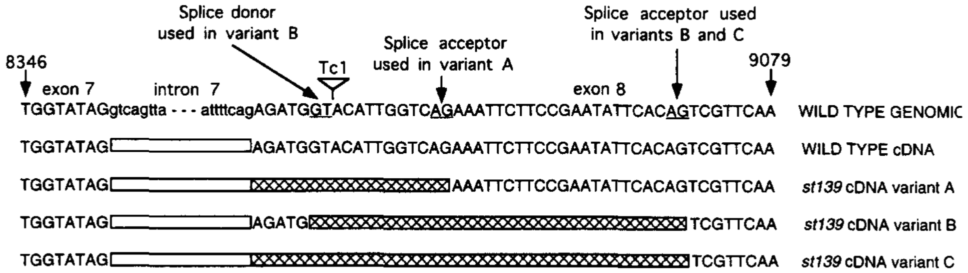

Splice donor used in variant B
Splice acceptor used in variant A
Splice acceptor used in variants B and C
8346
9079
Tc1
exon 7
intron 7
exon 8
TGGTATAGgtcagtta - - - attttcagAGATGGTACATTGGTCAGAAATTCTTCCGAATATTCACAGTCGTTCAA
WILD TYPE GENOMIC
TGGTATAG AGATGGTACATTGGTCAGAAATTCTTCCGAATATTCACAGTCGTTCAA
WILD TYPE cDNA
TGGTATAG AAATTCTTCCGAATATTCACAGTCGTTCAA
st139 cDNA variant A
TGGTATAG AGATG TCGTTCAA
st139 cDNA variant B
TGGTATAG TCGTTCAA
st139 cDNA variant C

FIGURE 6.—Splicing events near intron no. 7 in wild type and *st139*. Genomic sequence between positions 8346 and 9079, including a portion of intron no. 7, is depicted. Exonic sequence is shown in uppercase and intronic sequence is shown in lowercase letters. The insertion site for Tc*1* in *st139* is taken from MORI *et al.* (1988). Nine independently derived cDNAs from wild type gave the same, expected sequence. Of the nine cDNAs sequenced from *st139*, three gave one event (variant A, an 18-bp deletion which maintains the reading frame) and two show different events (variants B and C, that are out-of-frame deletions). Alternative splice donors and acceptors used by *st139* mRNAs are underlined. Intron 7 is spliced out of all pictured cDNAs. Additional sequences lost during *st139* splicing are indicated by hatched boxes.

on RNA isolated from *st139*. Primers that bracketed the Tc1 insertion site yielded the expected, single 202-bp product from wild-type RNA. When the same primers were used to amplify cDNA from *st139*, two prominent bands, one slightly larger and one slightly larger than 202 bp were produced, together with several somewhat larger bands of lesser abundance. The PCR products from *st139* RNA were ligated into a vector, and nine clones were sequenced. Nearly all of Tc*1* had been spliced-out, in all cases. In one case, the first 37 bp of Tc1 were left behind. In five of nine clones, intron no. 7 was extended 18 or 40 bp to use a cryptic splice acceptor site, or an additional cryptic intron of 35 bp was spliced-out (see Figure 6). The mild phenotype of *st139* is probably due to the high percentage of these splicing events (3 of 9) being 18 bp 3′ extensions of intron no. 7 which maintain the reading frame. By sequencing a wild-type isogenic revertant for *st139* called RW2418 (MOERMAN *et al.* 1988), we find an insertion of 3 bp (TTG), just preceding a TAC codon, which presumably adds an in-frame leucine. Given the observed variability in length of motif IIs in twitchin and Ig domains in general, the wild-type phenotype is not surprising. In addition, all *unc-22* alleles that had previously been positioned on the physical map are now known to alter the 21,614-nt twitchin transcript.

DISCUSSION

We report completion of the full sequence of *unc-22* which transcribes an mRNA of 21,614 nt that encodes a twitchin polypeptide of 6839 amino acids. A cDNA that includes the SL1 *trans*-spliced leader was recovered, and this localized the 5′ end of the *unc-22* message. SL1 is a 22 nt sequence that is added to the 5′ end of the messages of about 10% of *C. elegans* genes, but its presence has no known functional significance (BLUMENTHAL and THOMAS 1988). We are also reasonably certain about the initiator methionine at genomic position 4087; its upstream sequence is similar to a consensus sequence for translation initiation for *C. elegans*, and in the cDNA sequence, there are no other in-frame ATGs further upstream.

These results add 791 amino acids to the N terminus of the previously reported twitchin sequence (BENIAN *et al.* 1989). Most interesting are the three segments (167, 189 and 52 amino acids) that are unique and have no significant homology to each other or to other sequences in the databases. The third unique segment has a glycine-rich sequence (9 of 11 residues are glycine). It has been suggested that such glycine-rich regions are lacking in secondary structure and perhaps function as flexible "hinges" within a protein (BEACHY, HELFAND and HOGNESS 1985). We are presently looking for such a hinge by electron microscopic observations of isolated twitchin. The first two N-terminal unique segments have a high percentage of proline, basic amino acids and serine. It is intriguing that 9 of 10 sites within the twitchin sequence that match an optimal substrate consensus sequence for cAMP dependent protein kinases are located near the amino and carboxyl termini of the predicted twitchin protein sequence; eight occur in these three new unique sequences and one occurs in the third unique segment of 62 amino acids that follows the protein kinase domain of twitchin (between motif II copies 27 and 28). Whether phosphorylation occurs at these sites *in vivo* is presently unknown.

Twitchin is the largest polypeptide for which a complete exon/intron map is known. A striking feature is that the coding sequence for the first 896 amino acids (or 13% of the total coding sequence) is interrupted by 19 of the 30 total introns, which results in this coding sequence being distributed across 18,544 bp of chromosome *IV* (or 50% of the total gene). The largest intron in *unc-22*, no. 10, consists of 7402 bp, and very likely contains exons for another gene–a male germline-specific 1.2-kb message (L'HERNAULT, BENIAN and EMMONS 1993). Interestingly, a large intron of 6.5 kb in the *cha-1* gene contains the complete coding sequence for a separate gene, *unc-17* (RAND *et al.* 1991). Of the 30 introns in *unc-22*, six fall at the boundaries of motif I or motif II copies (introns 8, 19, 20, 23, 24 and 26), and 2 more introns fall within 42 bp of a motif in unique segments (introns 14 and 17).

st139 has the weakest phenotype of all the transposon Tc*1* alleles of *unc-22* examined; it is necessary to use nicotine to reliably score the homozygote (MOERMAN *et al.* 1988). The *st139* insertion site was believed to reside in 5′ upstream control sequence until all of the *unc-22* exons were defined. It is now clear that it lies in coding sequence, and that its mild phenotype is very likely due to an interesting mechanism. Examination of cDNAs generated by PCR revealed that the wild type-sized *unc-22* mRNA detected on Northern blots of *unc-22(st139)* (MOERMAN *et al.* 1988) is due to imprecise splicing-out of Tc1 from the primary transcript. In only one of nine cDNAs was a Tc*1* sequence left behind (the first 37 bp). In all other cases, Tc*1* was completely excised. The mild phenotype of *st139* is likely the result of a high fraction of the splicing events (three of nine) being 18-bp 3′ extensions of intron no. 7, which maintain the reading frame, and delete 6 amino acids from the twitchin polypeptide. This slightly smaller twitchin probably has nearly wild-type activity because twitchin, being a long polypeptide composed of repeating units, tolerates even larger in-frame deletions. By examining revertants of *unc-22* resulting from imprecise germline Tc*1* excision from one site in the middle of *unc-22*, KIFF *et al.* (1988) have shown that a nearly wild-type phenotype can be seen when up to six motifs have been deleted from the polypeptide.

The splicing of a transposon from the pre-mRNA of a gene into which it has inserted may be rather common, and may help the evolutionary survival of the transposon. It has been found in maize with the *Ds* and *dSpm* transposons (WESSLER 1989, 1991), in *Drosophila melanogaster* with retroposons (FRIDELL, PRET and SEARLES 1990; PRET and SEARLES 1991) and found more recently in *C. elegans* with Tc*1* (RUSHFORTH, SAARI and ANDERSON, 1993), Tc*3* and Tc*5* (J. COLLINS, M. MILLS, P. OLSEN and K. NORMAN, personal communication), and Tc*4* (LI, HERMAN and SHAW 1992). The donor-acceptor pair, either site being normal or cryptic, can reside within the termini of the transposon, within the gene in which it has inserted, or combinations of sites within the transposon and the target gene. The previous examples with Tc*1* and Tc*3* were small (6–16 amino acids) in-frame deletions or insertions in the polypeptide encoded by the target gene. The bias toward in-frame alterations is likely due to the fact that either revertants of mutations were being examined, or a mutant phenotype had not been selected.

We wish to thank ROBERT WATERSTON and DON MOERMAN for st139 and strain RW2418, JUNYI LEI for RNA and a Northern blot, ROBERT CHURCH for help in the calculation of pI and BARRY YEDVOBNICK and MARIO VALENZUELA for comments on the manuscript. This work was supported by grants from the National Institutes of Health to G.M.B. and S.W.L'H. G.M.B. is an Established Investigator of the American Heart Association.

LITERATURE CITED

AYME-SOUTHGATE, A., J. VIGOREAUX, G. BENIAN and M. L. PARDUE, 1991 *Drosophila* has a twitchin/titin-related gene that appears to encode projectin. Proc. Natl. Acad. Sci. USA **88:** 7973–7977.

BEACHY, P. A., S. L. HELFAND and D. S. HOGNESS, 1985 Segmental distribution of bithorax complex proteins during *Drosophila* development. Nature **313:** 545–551.

BENIAN, G. M., J. E. KIFF, N. NECKELMANN, D. G. MOERMAN and R. H. WATERSTON, 1989 Sequence of an unusually large protein implicated in regulation of myosin activity in *C. elegans*. Nature **342:** 45–50.

BLUMENTHAL, T., and J. THOMAS, 1988 Cis and trans mRNA splicing in *C. elegans*. Trends Genet. **4:** 305–308.

BRENNER, S., 1974 The genetics of *Caenorhabditis elegans*. Genetics **77:** 71–94.

CHIRGWIN, J. M., A. E. PRZYBYLA, R. J. MACDONALD and W. J. RUTTER, 1979 Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. Biochemistry **18:** 5294–5299.

COLLINGE, M., P. E. MATRISIAN, W. E. ZIMMER, R. L. SHATTUCK, T. J. LUKAS, L. J. VAN ELDIK and D. M. WATTERSON, 1992 Structure and expression of a calcium-binding protein gene contained within a calmodulin-regulated protein kinase gene. Mol. Cell. Biol. **12:** 2359–2371.

COULSON, A., J. SULSTON, S. BRENNER and J. KARN, 1986 Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. USA **83:** 7821–7825.

DRISCOLL, D. M., J. K. WYNNE, S. C. WALLIS and J. SCOTT, 1989 An *in vitro* system for the editing of apolipoprotein B mRNA. Cell **58:** 519–525.

EINHEBER, S., and D. A. FISCHMAN, 1990 Isolation and characterization of a cDNA clone encoding avian skeletal muscle C-protein: an intracellular member of the immunoglobulin superfamily. Proc. Natl. Acad. Sci. USA **87:** 2157–2161.

EPSTEIN, H. F., R. H. WATERSTON and S. BRENNER, 1974 A mutant affecting the heavy chain of myosin in *Caenorhabditis elegans*. J. Mol. Biol. **90:** 291–300.

FISCHMAN, D. A., K. VAUGHAN, F. WEBER and S. EINHEBER, 1991 Myosin binding proteins: intracellular members of the immunoglobulin superfamily, pp. 211–222 in *Frontiers of Muscle Research; Myogenesis, Muscle Contraction and Muscle Dystrophy*, edited by E. OZAWA, T. MASAKI and Y. NABESHIMA. Elsevier, Amsterdam.

FRIDELL, R. A., A.-M. PRET and L. L. SEARLES, 1990 A retrotransposon 412 insertion within an exon of the *Drosophila melanogaster* vermilion gene is spliced from the precursor RNA. Genes Dev. **4:** 559–566.

FYRBERG, C., S. LABEIT, B. BULLARD, K. LEONARD and E. FYRBERG, 1992 *Drosophila* projectin: relatedness to titin and twitchin and correlation with *lethal(4) 102cda* and *bent-Dominant* mutants. Proc. R. Soc. Lond. B **249:** 33–40.

GALLAGHER, P. J., and B. P. HERRING, 1991 The carboxyl terminus of the smooth muscle myosin light chain kinase is expressed as an independent protein, telokin. J. Biol. Chem. **266:** 23945–23952.

HOLDEN, H. M., M. ITO, D. J. HARTSHORNE and I. RAYMENT, 1992 X-ray structure determination of telokin, the C-terminal domain of myosin light chain kinase, at 2.8 Å resolution. J. Mol. Biol. **227:** 840–851.

KENNELLY, P. J., and E. G. KREBS, 1991 Consensus sequences as substrate specificity determinants for protein kinases and protein phosphatases. J. Biol. Chem. **266:** 15555–15558.

KIFF, J. E., D. G. MOERMAN, L. A. SCHRIEFER and R. H. WATERSTON, 1988 Transposon-induced deletions in *unc-22* of *C. elegans* associated with almost normal gene activity. Nature **331:** 631–633.

KRAUSE, M., and D. HIRSH, 1987 A trans-spliced leader sequence on actin mRNA in *Caenorhabditis elegans*. Cell **49:** 753–761.

KRUG, M. S., and S. L. BERGER, 1987 First-strand cDNA synthesis primed with oligo(dT), pp. 316–325 in *Guide to Molecular Cloning Techniques*, edited by S. L. BERGER and A. R. KIMMEL. Academic Press, San Diego, Calif.

LABEIT, S., D. P. BARLOW, M. GAUTEL, T. GIBSON, J. HOLT, C.-L. HSIEH, U. FRANCKE, K. LEONARD, J. WARDALE, A. WHITING and J. TRINICK, 1990 A regular pattern of two types of 100-residue motif in the sequence of titin. Nature **345:** 273–276.

LABEIT, S., M. GAUTEL, A. LAKEY and J. TRINICK, 1992 Towards a molecular understanding of titin. EMBO J. **11:** 1711–1716.

LAEMMLI, U. K., 1970 Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature **227:** 680–685.

L'HERNAULT, S. W., G. M. BENIAN and R. B. EMMONS 1993 Genetic and molecular characterization of the *Caenorhabditis elegans* spermatogenesis defective gene *spe-17*. Genetics **234:** 796–780.

LI, W., R. K. HERMAN and J. E. SHAW, 1992 Analysis of the *Caenorhabditis elegans* axonal guidance and outgrowth gene *unc-33*. Genetics **132:** 675–689.

MOERMAN, D. G., 1980 A genetic analysis of the *unc-22* region in *C. elegans*. Ph.D Thesis, Simon Fraser University, Burnaby, British Columbia.

MOERMAN, D. G., G. M. BENIAN and R. H. WATERSTON, 1986 Molecular cloning of the muscle gene *unc-22* in *Caenorhabditis elegans* by Tc1 transposon tagging. Proc. Natl. Acad. Sci. USA **83:** 2579–2583.

MOERMAN, D. G., S. PLURAD, R. H. WATERSTON and D. L. BAILLIE, 1982 Mutations in the *unc-54* myosin heavy chain gene of *Caenorhabditis elegans* that alter contractility but not muscle structure. Cell **29:** 773–781.

MOERMAN, D. G., G. M. BENIAN, R. J. BARSTEAD, L. SCHREIFER and R. H. WATERSTON, 1988 Identification and intracellular localization of the *unc-22* gene product of *Caenorhabditis elegans*. Genes Dev. **2:** 93–105.

MORI, I., G. M. BENIAN, D. G. MOERMAN and R. H. WATERSTON, 1988 The transposon Tc1 of *C. elegans* recognizes specific target sequences for integration. Proc. Natl. Acad. Sci. USA **85:** 861–864.

NOGUCHI, J., M. YANAGISAWA, M. IMAMURA, Y. KASUYA, T. SAKURAI, T. TANAKA and T. MASAKI, 1992 Complete primary structure and tissue expression of chicken pectoralis M-protein. J. Biol. Chem. **267:** 20302–20310.

OLSON, N. J., R. B. PEARSON, D. S. NEEDLEMAN, M. Y. HURWITZ, B. E. KEMP and A. R. MEANS, 1990 Regulatory and structural motifs of chicken gizzard myosin light chain kinase. Proc. Natl. Acad. Sci. USA **87:** 2284–2288.

PEARSON, W. R., and D. J. LIPMAN, 1988 Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:** 2444–2448.

PRET, A.-M., and L. L. SEARLES, 1991 Splicing of retrotransposon insertions from transcripts of the Drosophila melanogaster vermilion gene in a revertant. Genetics **129:** 1137–1145.

PRICE, M. G., 1987 Skelemins: cytoskeletal proteins located at the periphery of M-discs in mammalian striated muscle. J. Cell Biol. **104:** 1325–1336.

PRICE, M. G., C. A. BROOKS and R. H. GOMER, 1990 Skelemins are members of a family of myosin-associated proteins with immunoglobulin superfamily C2 and fibronectin type III domains. J. Cell Biol. **111:** 170a.

RAND, J. B., A. ALFONSO, K. GRUNDAHL, J. R. MCMANUS and J. M. ASBURY, 1991 Alternative splicing of choline acetyltransferase transcripts in the nematode *C. elegans*. Soc. Neurosci. Abst. **17:** 1525.

RUSHFORTH, A. M., B. SAARI and P. ANDERSON, 1993 Site-selected insertion of the transposon Tc*1* into a Caenorhabditis elegans myosin light chain gene. Mol. Cell. Biol. **13:** 902–910.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SHOEMAKER, M. O., W. LAU, R. L. SHATTUCK, A. P. KWIATKOWSKI, P. E. MATRISIAN, L. GUERRA-SANTOS, E. WILSON, T. J. LUKAS, L. J. VAN ELDIK and D. M. WATTERSON, 1990 Use of DNA sequence and mutant analyses and antisense oligodeoxynucleotides to examine the molecular basis of nonmuscle myosin light chain kinase autoinhibition, calmodulin recognition, and activity. J. Cell Biol. **111:** 1107–1125.

SMITH, D. B., and K. S. JOHNSON, 1988 Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. Gene **67:** 31–40.

TOWBIN, H., T. STAEHELIN and J. GORDON, 1979 Electrophoretic transfer of proteins form polyacrylamide gels to nitrocellulose sheets: Procedure and some application. Proc. Natl. Acad. Sci. USA **76:** 4350–4354.

WATERSTON, R. H., 1988 Muscle, pp. 281–336 in *The Nematode Caenorhabditis elegans*, edited by W. B. WOOD. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

WATERSTON, R. H., J. N. THOMSON and S. BRENNER, 1980 Mutants with altered muscle structure in *Caenorhabditis elegans*. Dev. Biol. **77:** 271–302.

WESSLER, S. R., 1989 The splicing of maize transposable elements from pre-mRNA–a minireview. Gene **82:** 127–133.

WESSLER, S. R., 1991 The maize transposable *Ds1* element is alternatively spliced from exon sequences. Mol. Cell. Biol. **11:** 6192–6196.

Communicating editor: R. K. HERMAN