Am. J. Hum. Genet. 48:621–627, 1991

Rapid Detection of CA Polymorphisms in Cloned DNA: Application to the 5′ Region of the Dystrophin Gene

C. A. Feener, F. M. Boyce, and L. M. Kunkel

Division of Genetics, Howard Hughes Medical Institute, Children's Hospital, Boston

Summary

To identify CA repeats in genomic sequences which had been previously subcloned into plasmids, we performed PCR using a $(CA)_n$ primer and a flanking vector primer on the genomic inserts. By incorporation of a restriction enzyme site into the $(CA)_n$ primer, we have been able to subclone the genomic DNA so that the sequence flanking the CA repeat is readily determined. Primers can then be designed to amplify across the CA repeat in patient DNA samples. Application of this technique to genomic DNAs surrounding the upstream "brain" promoter of the dystrophin gene has led to the discovery of four new CA repeats. Three of these repeats are highly polymorphic, with PICs ranging from .586 to .768. The location of these markers at the extreme 5′ terminus of the dystrophin gene, together with their high degree of polymorphism and ease of assay, makes them ideal for linkage analysis in families with Duchenne muscular dystrophy.

Introduction

Duchenne and Becker muscular dystrophies (DMD and BMD) are allelic, X-linked disorders resulting in progressive muscle wasting. DMD occurs with an incidence of one in 3,500 male births, making it the most common lethal X-linked disorder in man. A collaborative effort involving many laboratories has led to the identification and cloning of the DMD/BMD gene (Monaco et al. 1986; Burghes et al. 1987). This gene contains approximately 65 exons spanning a distance of 2.4 million bp on the human X chromosome (Koenig et al. 1987; Den Dunnen et al. 1989) and encodes a 427-kD protein named dystrophin (Hoffman et al. 1987). Isolation of the complete cDNA for dystrophin (Koenig et al. 1987) has led to the ability to detect about 60% of DMD mutations by Southern blot analysis (Forrest et al. 1988; Koenig et al. 1989) or, more recently, by PCR methods (Chamberlain et al. 1988; Beggs et al. 1990). The remaining 40% of DMD mutations do not have gross deletions and must be traced indirectly through linkage analysis. Linkage analysis in Duchenne families without detectable deletions has been conventionally performed by Southern blot analysis using probes flanking and within the DMD gene (Bakker et al. 1985; Darras and Francke 1988).

Simple dinucleotide sequence repeats, such as CA repeats, have been shown to be highly polymorphic and useful in genetic linkage analysis, since they are readily assayed by PCR (Litt and Luty 1989; Smeets et al. 1989; Tautz 1989; Weber and May 1989). These repeats may occur adjacent to the coding regions of genes, in introns within genes, or within untranslated regions (Hamada et al. 1984; Weber and May 1989). CA repeats occur in 6–30 units, sometimes containing minor interruptions, with longer repeats tending to be more polymorphic (Weber 1990). Tandem repeat polymorphisms have been detected in the 3′ untranslated region of the dystrophin gene and found to be highly polymorphic (Roberts et al. 1989; Beggs and Kunkel 1990; Oudet et al. 1990).

We have developed a simple PCR method for detecting CA repeats within genomic sequences which are subcloned into plasmids. We have used this technique to screen genomic clones from the most 5′ region of the dystrophin gene. This region encodes the promoter of the dystrophin gene utilized in the brain, and is located far upstream of the "muscle" promoter (Boyce et al., in press). Application of the PCR strategy

Received August 30, 1990; revision received November 7, 1990.

Address for correspondence and reprints: L. M. Kunkel, Division of Genetics, Howard Hughes Medical Institute, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115.

© 1991 by The American Society of Human Genetics. All rights reserved.

0002-9297/91/4803-0023$02.00

to this region has revealed three new polymorphic CA repeats. These CA repeats will be useful in linkage analysis of Duchenne families because of their high degree of polymorphism, ease of assay by PCR, and location at the extreme 5′ terminus of the dystrophin gene.

Material and Methods

Identification and Cloning of CA Repeats

Subclones of the dystrophin upstream promoter region (Boyce et al., in press) in the plasmid vector Bluescript SKII + (Stratagene) were analyzed by PCR using the primers BAM-CA (5′-CCCGGATCC$(AC)_9$-3′), BAM-TG (5′-CCCGGATCC$(TG)_9$-3′), T3 (5′-CACGCTCGAGATTAACCCTCACTAAAG-3′), and T7 (5′-ACCGCGGATCCAATACGACTCACTATAG-3′) in various combinations (BAM-CA/T3, BAM-CA/T7, BAM-TG/T3, and BAM-TG/T7). Amplification conditions were as follows: denaturation at 94°C for 1 min, annealing at 55°C for 1 min, and extension at 72°C for 3 min, for 20 cycles. Products of the PCR reactions were analyzed by agarose gel electrophoresis and stained with ethidium bromide. Reactions containing visible products were extracted with phenol and chloroform, digested with *Bam*HI, extracted with phenol and chloroform, and the products subcloned into Bluescript SKII + (Stratagene). Cesium chloride–purified plasmids were sequenced using the Sequenase version 2.0 kit (U.S. Biochemical) in conjunction with T3 or T7 primers (Stratagene), depending upon the orientation of the subclone. Copies of the sequence flanking each repeat are available upon request. Primers were then designed from the sequence flanking the CA repeat. Phage DNA from this region was screened by hybridization of Southern blots using a $(CA)_{12}$ primer labeled with gamma ^{32}P-ATP (New England Nuclear) and T4 polynucleotide kinase (U.S. Biochemical). Prehybridization was at 45°C in 6 × SSC, 5 × Denhardt's solution, 0.05% sodium pyrophosphate, 0.025 mg salmon sperm DNA/ml, 0.5% SDS, with hybridization at 42°C in 6 × SSC, 1 × Denhardt's solution, 0.05% sodium pyrophosphate, 0.1 mg tRNA/ml, and 0.025 mg salmon sperm DNA/ml, and with washing at 37°C in 6 × SSC, 0.05% sodium pyrophosphate.

PCR of Genomic DNA Samples

Each PCR reaction contained 400 ng of genomic DNA, 200 ng of each unlabeled primer of the set (table 1), 8 μl of 1.25 mM dNTP, 5 μl of 10 × reaction buffer (Cetus), 0.5 μl of native *Taq* polymerase (Cetus) in a volume of 50 μl and overlaid with 50 μl of mineral oil. In addition, each reaction contained 2,000–20,000 cpm of one of the primers labeled with gamma ^{32}P-ATP (New England Nuclear) and T_4 polynucleotide kinase (U.S. Biochemical). The thermocycle programs consisted of a single denaturation at 94°C for 7 min, 24 cycles of amplification, and a final cycle with an extension time of 10 min. Two different amplification conditions were used: for the primer sets I and IV denaturation was carried out at 94°C for 30 seconds; and annealing and extension at 65°C for 4 min (conditions described by Beggs et al. [1990]). Amplification conditions for primer sets II and III were denaturation at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 63°C for 4 min (conditions described by Chamberlain et al. 1988). The reaction products were diluted with 3 vol sequence loading buffer, boiled for 2 min, and quenched on ice, and 4 μl/lane were loaded onto an 8% sequencing gel (gel mix 8; BRL). Electrophoresis was at 35 mA constant current on a model S2 sequencing apparatus (0.4 × 31.0 × 38.5 cm; BRL). When the xylene cyanol dye in the loading buffer left the gel, electrophoresis was stopped. The gels were dried and then autoradiographed at − 80°C using an intensifying screen. Typical exposure times were 1–3 d.

Results

We have previously isolated the genomic sequences surrounding the dystrophin promoter used in the brain (Boyce et al., in press). Since this region maps upstream of the dystrophin muscle promoter, it represents the extreme 5′ terminus of the gene, and we reasoned that it might be valuable to obtain polymorphisms of this region for use in linkage analysis. Rather than extensive restriction mapping, subcloning, and sequencing, we designed a PCR method to rapidly identify the locations of CA repeats and to determine the genomic sequence immediately surrounding them (fig. 1). The method involves using a $(CA)_n$ or $(TG)_n$ oligonucleotide, together with an oligonucleotide primer from the flanking vector sequence, to amplify a portion of the genomic region. The length of this PCR product corresponds to the distance of the CA repeat from the flanking vector primer. Furthermore, the PCR product divides the genomic region at the CA repeat so that subcloning of this PCR product allows one to directly sequence across the CA repeat from the flanking vector sequences (fig. 2*A*). By performing PCR using $(TG)_n$ as well as $(CA)_n$ primers the sequence on either side

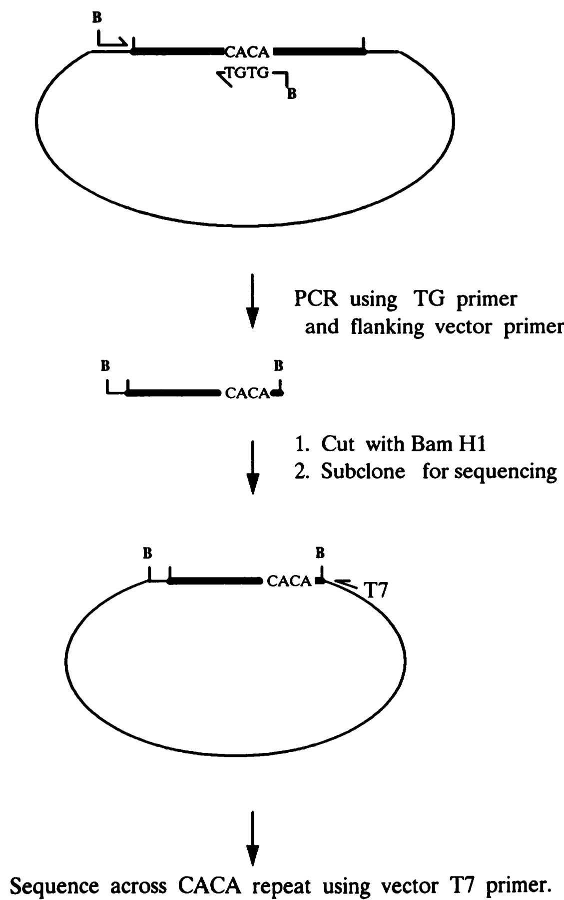

Figure 1 PCR technique used to identify and map CA repeats in cloned genomic samples. A portion of genomic DNA (thick line) cloned into a plasmid vector (thin line) is subjected to PCR using a $(TG)_9$ primer containing a *Bam*HI site (B) and a primer in the flanking vector sequence. This allows determination of the location and orientation of the $(CA)_n$ repeat. For simplicity, the $(CA)_n$ repeat is illustrated as a $(CA)_2$ repeat. Samples which contain CA repeats are visualized by gel electrophoresis of the PCR products. The PCR products are then digested with *Bam*HI and subcloned into a plasmid vector. This allows the genomic sequence flanking the CA repeat to be determined using a primer (T7) which binds to the polylinker adjacent to the CA repeat. The entire process may also be performed with a $(CA)_9$ primer instead of a $(TG)_9$ primer to determine the genomic sequence on the other side of the CA repeat. From these flanking sequences, primers which specifically amplify the CA repeat are obtained.

of the CA repeat is obtained. Alternatively, once the sequence on one side of the CA repeat is obtained, a primer can be designed to sequence back across the other side of the CA repeat (fig. 2*B*). Primers are then designed from the sequences flanking the repeat and used to amplify the repeat from genomic DNA samples by PCR, with polymorphisms of the repeat apparent in the sizes of the products obtained.

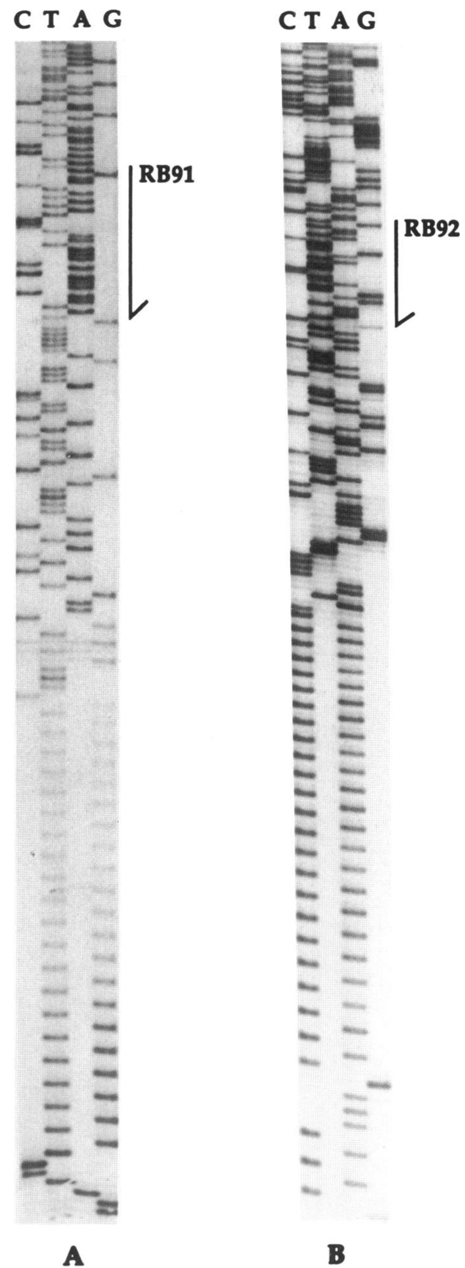

Figure 2 Example of DNA sequence used for design of flanking primers A $(TG)_9$-*Bam*HI primer and used in conjunction with a vector primer to amplify portion of genomic sequence surrounding 5′DYS-II. The resulting product was then subcloned into a plasmid vector. *A*, DNA sequence of this subclone, using primer in vector, allowing determination of sequence flanking dinucleotide repeat. *B*, DNA sequence in opposite direction, using primer RB91 (designed from sequence in *A*) with original genomic clone as template. This allows determination of the opposite flanking sequence, and design of primer RB92 for amplification of the entire CA repeat.

Table 1

DNA Sequence of CA Repeats and Flanking Primers

Repeat Name	5′ Primer	Repeat Sequence	3′ Primer
5′DYS-I	ACTGTAAATGAAATTGTTTTCTAAGTGCC	$(TA)_{11}(CA)_6$	GTTAACAAAATGTCCTTCAGTTCTATCC
5′DYS-II	TCTTGATATATAGGGATTATTTGTGTTTGTTATAC	$(CA)_{23}$	ATTATGAAACTATAAGGAATAACTCATTTAGC
5′DYS-III	TTTTTTAGGTATAACTTACATACAATAAACC	$(TA)_{11}(CA)_{10}$TACA	GTGACAATAAGCATATCAGTGGCTGCC
5′DYS-IV	AAACAAAGTAATATTGTTAGACATGTATCC	$(TG)_5(TA)_4$TG	GTTAAGGTTCCTCTTAGATCATGGTGTCC

NOTE.—Primer sequences are written in the 5′-to-3′ direction.

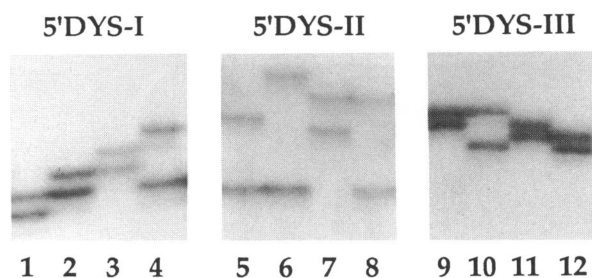

Figure 3 Examples of dinucleotide repeat length polymorphisms in normal females. PCR amplification was performed on genomic DNA from normal human females using primers flanking each new repeat (see table 1), and the products were resolved on a denaturing polyacrylamide gel. The dinucleotide repeat is as indicated above each panel; each sample is numbered at the bottom.

We detected four CA repeats (5′DYS-1, 5′DYS-II, 5′DYS-III, and 5′DYS-IV) in approximately 19 kb of the subcloned dystrophin 5′ terminus using our PCR analysis. We determined the sequence of each repeat, as well as the sequence of the flanking region, in our genomic subclones. The smallest of these repeats (5′DYS-IV) contained only five CA dinucleotides (table 1), indicating that the PCR technique is very sensitive, since only 10 bp of CA sequence are apparently required for primer annealing. The remaining repeats detected by PCR analysis include examples of compound repeats (5′DYS-I, 5′DYS-III) as well as a very long perfect repeat (5′DYS-II). To determine whether this region contained other CA repeats not detectable by PCR, we probed the subcloned plasmids by using conventional hybridization with a CA repeat oligonucleotide and failed to detect additional CA repeats (data not shown).

Primers were then designed from the flanking sequences (table 1) and used to amplify the segment from unrelated females by PCR to determine if these segments were polymorphic in length. Examples of this analysis for three of the repeats are shown in figure 3. These three new markers were found to be polymorphic with PIC values ranging from .586 to .768 (table 2). 5′DYS-IV, which contained the shortest repeat, was not found to be polymorphic in the 54 chromosomes assayed. 5′DYS-II, which contained a very long perfect repeat, was the most polymorphic in our population with eight alleles observed in the 78 chromosomes assayed. When taken together, these three new markers are informative for linkage in over 90% of the females studied (table 2).

The two most polymorphic markers, 5′DYS-I and

Table 2

Informativeness of New 5′ Dystrophin Markers

Marker and Allele	Length (bp)	Allele Frequency[a]	% Heterozygous[b]	PIC[c]
5′DYS-I:				
A1	185	.048		
A2	183	.071		
A3	181	.310	78.6	.608
A4	179	.536		
A5	177	.036		
5′DYS-II:				
A1	228	.090		
A2	226	.077		
A3	224	.103		
A4	222	.026	82.0	.768
A5	220	.103		
A6	218	.192		
A7	216	.013		
A8	214	.397		
5′DYS-III:				
A1	225	.057		
A2	223	.529	51.0	.586
A3	221	.057		
A4	219	.357		

[a] Population assayed consisted predominantly of white females.
[b] Thirty-eight of 42 females were informative for at least one marker.
[c] PIC for these X-linked markers was calculated by the formula PIC $= 1 - \sum_{i=1}^{n} P_i^2$, where P_i is the frequency of allele i.

5′DYS-II, are located just upstream of the transcriptional start site for the brain promoter, making them true 5′ flanking markers. 5′DYS-I is located approximately 3.5 kb 5′ to the transcriptional start site, while 5′DYS-II is located approximately 1.2 kb 5′ to the transcriptional start site (data not shown). 5′DYS-III and 5′DYS-IV map within the first intron, located 3.5 and 4.2 kb 3′ to the transcriptional start site, respectively.

Analysis of these three new simple sequence polymorphisms in Duchenne families reveals that in each case the polymorphisms as detected by PCR are inherited in a Mendelian manner. An example of this analysis for repeat 5′DYS-I is shown in figure 4. In this family, the 185-bp allele is linked to DMD, as evidenced by its presence in the affected male (lane 7). Three of his sisters are carriers of this allele (lanes 3, 4, 6), and were previously known to be carriers of DMD by prior linkage testing and by their increased serum creatine kinase levels. Thus, these polymorphisms are linked to DMD in the families studied and therefore should prove useful in linkage analysis of Duchenne families.

Discussion

We have demonstrated a new PCR-based technique for mapping of simple sequence repeats in subcloned

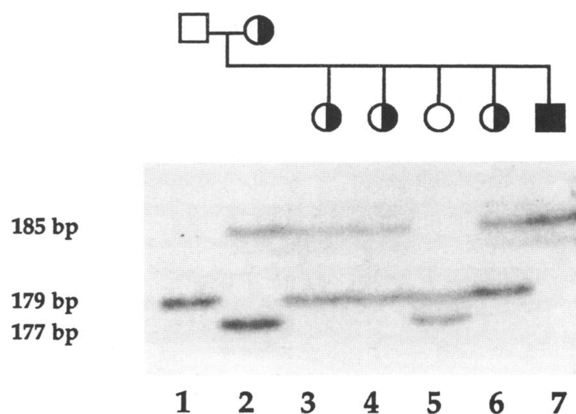

Figure 4 Linkage analysis of CA repeat in family with DMD. Genomic DNA from a family with DMD was used in PCR reactions using primers for CA repeat 5′DYS-I. The 185-bp allele segregates with DMD in the affected male (lane 7). Three of the female siblings (lanes 3, 4, 6) are carriers of the 185-bp allele which segregates with DMD in this family.

DNAs and rapid determination of the sequences flanking these repeats. This allows design of flanking primers to amplify the repeat from genomic DNA samples in linkage analysis, a technique which is much easier than Southern blot–based RFLP analysis. Using this technique, we have found three new polymorphic markers for the newly described 5′ terminus of the DMD gene.

There are several limitations to our method. We have used genomic sequences which are already subcloned into plasmids, thus allowing the CA repeats to be within the length limit which is easily obtainable by PCR. In our study, we were able to successfully detect repeats as far as 4 kb away from the flanking vector sequence (not shown). It might be possible to use a similar technique in genomic lambda phages, but only a portion of the phage insert would be within range of PCR from the flanking vector primer. Alternatively, primers for Alu repeats could be used in conjunction with CA repeats to obtain amplification of internal sequences. However, for instances in which large genomic pieces are not yet subcloned from cosmid or yeast artificial chromosome vectors, it is probably easier to use the method of shearing the DNA into small fragments, shotgun subcloning of these fragments, and hybridization screening of these subclones with a $(CA)_{12}$ oligonucleotide probe (Weber 1990). Thus, our technique is best suited to cases where genomic DNAs for a particular locus are already subcloned into pieces. One application of this technique would be to find CA polymorphisms in probes previously used to detect RFLPs. Since such probes are already well defined on linkage maps, detection of CA repeats in these probes would allow PCR techniques to replace the more laborious and less informative RFLP technique. Although we have chosen to subclone our PCR products for sequencing, it is also possible to sequence PCR products directly (see, e.g., Casanova et al. 1990).

Application of our PCR technique to the 5′ region of the Duchenne muscular dystrophy gene resulted in the identification of three new polymorphic markers for linkage analysis. We feel it is unlikely that our technique did not detect additional CA repeats present in this region, since the PCR screening gave an identical map as did conventional hybridization screening. Furthermore, we were able to detect CA repeats as short as 10 bp (repeat IV). However, hybridization screening did suggest the presence of at least one additional CA repeat outside the area analyzed by PCR. Since CA repeats are believed to occur on average every 25–100 kb in the genome, this region appears unusually heavily represented with these simple sequence repeats. The reason for this is not clear, although these repeats have the potential for forming Z-DNA, which could affect regulation of this promoter region (Naylor and Clark 1990).

The three new polymorphic markers which we describe should prove very useful in linkage studies of DMD. Linkage analysis is still required in the 40% of families that do not have deletions of the dystrophin gene detectable by Southern blotting or PCR. The location of these markers is unique, since they are in the newly described brain promoter of dystrophin, which is at the extreme 5′ terminus of the gene and far upstream of the dystrophin muscle promoter (Boyce et al., in press). Thus, these markers are ideal 5′ flanking markers and are the first polymorphic markers described for this region of the dystrophin gene. It is important to note that additional linkage markers within the gene and at the 3′ terminus are also necessary for linkage analysis of DMD, since there is significant recombination across the dystrophin gene (Abbs et al. 1990). Toward this goal, PCR-based linkage markers for the dystrophin 3′ terminus have been described elsewhere (Roberts et al. 1989; Beggs and Kunkel 1990; Oudet et al. 1990).

Ideally, methods for the direct detection of Duchenne mutations will obviate the need for linkage analysis. The large size of the DMD gene and its cDNA currently make this an exceedingly difficult task, particularly for point mutations. Nevertheless, the use of PCR and other new technologies may one day achieve this goal.

Acknowledgments

We thank Alan Beggs and Judy Snyder for DNA samples, and Paul Neumann for his help with PIC calculations. We thank Mark Fleming of the Howard Hughes Medical Institute biopolymers laboratory for synthesis of oligonucleotides. Special thanks to Alan Beggs and the members of our laboratory for helpful discussions and critical readings of this manuscript. We thank Nicole Picard for assistance with the manuscript and Dixon Yun for preparation of the figures. F.M.B. is the Betty Q. Banker Fellow of the Muscular Dystrophy Association. This work was supported by grants from the National Institutes of Health, the Muscular Dystrophy Association, and Howard Hughes Medical Institute to L.M.K.

References

Abbs S, Roberts RG, Mathew CG, Bentley DR, Bobrow M (1990) Accurate assessment of intragenic recombination

frequency within the Duchenne muscular dystrophy gene. Genomics 7:602–606

Bakker E, Hofker MH, Goor N, Mandel JL, Wrogemann K, Davies KE, Kunkel LM, et al (1985) Prenatal diagnosis and carrier detection of Duchenne muscular dystrophy with closely linked RFLPs. Lancet 1:655–658

Beggs AH, Koenig M, Boyce FM, Kunkel LM (1990) Detection of 98% of DMD/BMD deletions by PCR. Hum Genet 86:45–48

Beggs AH, Kunkel LM (1990) A polymorphic CACA repeat in the 3′ untranslated region of dystrophin. Nucleic Acids Res 18:1931

Boyce FM, Beggs AH, Feener C, Kunkel LM. Dystrophin is transcribed in brain from a distant upstream promoter. Proc Natl Acad Sci USA (in press)

Burghes AHM, Logan C, Hu X, Belfall B, Worton RG, Ray PN (1987) A cDNA clone from the Duchenne/Becker muscular dystrophy gene. Nature 328:434–437

Casanova JL, Pannetier C, Jaulin C, Kourilisky P (1990) Optimal conditions for directly sequencing double-stranded PCR products with Sequenase. Nucleic Acids Res 18:4028

Chamberlain JS, Gibbs RA, Ranier JE, Nguyen PN, Caskey CT (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res 16:11141–11156

Darras BT, Francke U (1988) Normal human genomic restriction-fragment patterns and polymorphisms revealed by hybridization with the entire dystrophin cDNA. Am J Hum Genet 43:612–619

Den Dunnen JT, Grootscholten PM, Bakker E, Blonden LAJ, Ginjaar HB, Wapenaar MC, van Paassen HMB, et al (1989) Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. Am J Hum Genet 45:835–847

Forrest SM, Cross GS, Flint T, Speer A, Robson KJH, Davies KE (1988) Further studies of gene deletions that cause Duchenne and Becker muscular dystrophies. Genomics 2: 109–114

Hamada H, Petrino MG, Kakunaga T, Seidman M, Stollar BD (1984) Characterization of genomic poly(dT-dG)·poly(dC-dA) sequences: structure, organization, and conformation. Mol Cell Biol 4:2610–2621

Hoffman EP, Brown RH, Kunkel LM (1987) Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell 51:919–928

Koenig M, Beggs AH, Moyer M, Scherpf S, Heindrich K, Bettecken T, Meng G, et al (1989) The molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion. Am J Hum Genet 45:498–506

Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM (1987) Complete cloning of the Duchenne muscular dystrophy gene. Cell 50:509–517

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM (1986) Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 323:646–650

Naylor LH, Clark EM (1990) d(TG)·d$(CA)_n$ sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. Nucleic Acids Res 18:1595–1601

Oudet C, Heilig R, Mandel JL (1990) An informative polymorphism detectable by polymerase chain reaction at the 3′ end of the dystrophin gene. Hum Genet 84:283–285

Roberts RG, Montandon AJ, Bobrow M, Bentley DR (1989) Detection of novel genetic markers by mismatch analysis. Nucleic Acids Res 17:5961–5971

Smeets HJM, Brunner HG, Ropers HH, Wieringa B (1989) Use of variable sequence motifs as genetic markers: application to study of myotonic dystrophy. Hum Genet 83: 245–251

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic markers. Nucleic Acids Res 17:6463–6471

Weber JL (1990) Informativeness of human (dC-dA_n·(dT-$dG)_n$ polymorphisms. Genomics 7:524–530

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396