New 5′-(CGG)ₙ-3′ Repeats in the Human Genome*

Laurence Mangel, Thomas Ternes, Birgit Schmitz, and Walter Doerfler‡

From the Institute of Genetics, University of Cologne, D-50931 Koeln, Germany

(Received for publication, June 17, 1998, and in revised form, August 26, 1998)

We identified new, potentially unstable loci in the human genome containing 5′-(CGG)ₙ-3′ trinucleotide repeats by screening a human subgenomic library as well as a chromosome 16 library with a 5′-(CGG)₇-3′ oligodeoxynucleotide probe. Five different clones were isolated, two from the chromosome 16 library and three from the subgenomic library. Determinations of the nucleotide sequences have revealed that the E7 clone displayed, in addition to the 5′-(CGG)₉-3′ trinucleotide repeat, a 5′-(CGA)₉-3′ and a 5′-(CCT)₉-3′ trinucleotide repeat. Two clones, CL16-1 and P5-5, had homologies to known genes, the human casein kinase II α subunit (chromosome 16) and the human calcium-activated potassium channel (chromosome 10), respectively. Clones E7 and P4 were assigned to chromosome 6, whereas CL16-8 mapped to chromosome 16. Their potential coding capacities were assessed by RNA transfer (Northern blotting) experiments. Four different transcripts were identified by using the E7 clones as hybridization probes, three of them being brain-specific. The P4 clone was expressed in placenta and skeletal muscle. Minor polymorphisms within the repeats were observed in normal and in fragile X individuals. Lung and colon carcinoma cell lines in which some microsatellites were shown to be unstable were also investigated. Expansions of the 5′-(CGG)ₙ-3′ repeats were not found.

Fragile sites can be classified into two major groups, rare and common ones. Among the rare fragile sites, at least 19 are folate-sensitive. So far, five folate-sensitive fragile sites have been characterized at the molecular level: FRAXA, FRAXE, FRAXF, FRA16A, and FRA11B. They are associated with an expansion of 5′-(CGG)₉-3′ trinucleotide repeats, with three of them being linked to clinical phenotypes. Whereas expansion at the FRAXA locus results in the fragile X syndrome, expansion at the FRAXE locus leads to mild mental retardation, and the FRA11B locus can be involved in the Jacobsen syndrome. The amplified FRA16A and FRAXF loci are not linked to known phenotypes, and genes have not yet been found near these fragile sites.

Several severe human diseases have been associated with the expansion of naturally occurring triplet repeats in the human genome (for recent reviews, see Refs. 1–5). The cause of these instabilities in human DNA has remained unknown. In eight different loci, a 5′-(CAG)ₙ-3′ repeat as part of the coding region of genes can be amplified and cause a lengthening of a normal polyglutamine sequence. Severe neurological disorders, spinal and bulbar muscular atrophy, spinocerebellar ataxia types 1, 2, 6, and 7, Huntington disease, dentatorubral-pallidoluysian atrophy (Haw River syndrome), and Machado–Joseph disease (spinocerebellar ataxia type 3) are associated with the expansion of polyglutamine tracts. In Friedreich’s ataxia, the only disease in this group with an autosomal recessive inheritance pattern, a 5′-(GAA)₉-3′ repeat is expanded in the first intron of the X25 gene (6). Surprisingly, in myotonic dystrophy, the expanded 5′-(CTG)₉-3′ repeat lies in the 3′-untranslated region, allegedly affecting the structure of the 3′-region of the messenger RNA (7, 8). But the same region also constitutes the 5′-segment of the myotonic dystrophy-associated homeobox gene (9, 10).

The fragile X syndrome has been the first disease to be linked to a fragile site in the human genome. The FRAXA fragile site on the X chromosome in location A can be elicited by culturing the cells in a medium devoid of folic acid and thymidine (for reviews, see Refs. 11 and 12). By now, in excess of 100 fragile sites have been identified in the human genome, their recognition being dependent on the addition of specific chemicals to the culture medium, such as 5-bromo-deoxyuridine, thymidine, distamycin A, aphidicolin, 5-azacytidine, and others. Most of these fragile sites exhibit no genetic phenotype, e.g. those at chromosomes XF (FRAXF) or 16A (FRA16A). The fragile sites FRAXA at Xq27.3 or FRAXE at Xq28 are linked to varying degrees of mental retardation. At several of the fragile sites in the human genome, expansions of naturally occurring 5′-(CGG)₉-3′ repeats have been documented. The pathologically expanded repeats and adjacent nucleotide sequences can become de novo 5′-(CG)₉-3′-methylated (13).

In the FRAXA syndrome, typically presenting with severe mental retardation, facial dysmorphologies, macroorchidism, and joint hyperextensibility, the 5′-(CGG)₉-3′ repeat amplification is located in the 5′-untranslated region and promoter regions of the FMR1 gene. The 5′-(CG)₉-3′ methylation is associated with the inactivation of the FMR1 promoter. In a few cases, this methylation has been absent despite a marked 5′-(CGG)₉-3′ expansion, the FMR1 gene continued to be expressed, and the probands with the full expansion have failed to exhibit the FRAXA phenotype (14). Recent results of genomic sequencing experiments in the FMR1 repeat region indicate that methylation mosaics exist in healthy and some fragile X individuals.¹

We have initiated a search for regions in the human genome with 5′-(CGG)₉-3′ repeats. Because most of the previously isolated 5′-(CGG)₉-3′ repeat-containing fragments are located in noncoding regions, genomic libraries were screened with a ⁸⁻P-

* This research was supported by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie, Bonn, Germany; and the Center for Molecular Medicine Köln, TP13. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) AJ001215, AJ001216, AJ001217, AJ001218, and AJ001219.

† To whom correspondence should be addressed: Institute of Genetics, University of Cologne, Weyertal 121, D-50931 Köln, Germany. Tel.: 49-221-470-2386; Fax: 49-221-470-5163; E-mail: doerfler@scan.genetik.uni-koeln.de.

¹ B. Genç, H. Müller-Hartmann, M. Zeschning, H. Deissler, B. Schmitz, F. Majewski, A. von Gentard, and W. Doerfler, submitted for publication.

This paper is available on line at http://www.jbc.org
labeled 5′-(CGG)_{n-3}-repeats oligodeoxynucleotide. We report here the isolation of new genomic sequences containing 5′-(CGG)_{3}-repeats, mostly from the 5′-upstream segments of gene-rich regions.

**Experimental Procedures**

**Cell Lines**—The cell lines KB (human oral carcinoma), HeLa (human cervix carcinoma), 293 (adenovirus type 5-transformed human embryonic kidney cells), HEp-2 (human epidermoid larynx carcinoma), HeLa G2 (human hepatocellular carcinoma), A549 (human lung carcinoma), LoVo and SW480 (human colon carcinomas), and A172 and U138 (human glioblastoma cell lines) were cultivated in Dulbecco’s medium containing 10% fetal calf serum.

**DNA and RNA Extraction, Southern (DNA), and Northern (RNA) Transfer Hybridizations**—DNA was prepared from peripheral blood as described (16). RNA was isolated from cultured cells according to published protocols (17, 18). Restriction enzyme cleavage, gel electrophoresis, 

\[2^{32}\text{P}-\text{Labeling of hybridization probes, hybridization and autoradiography were carried out as described (19, 20). For Southern transfer (21) and Northern blotting (22)}\]

DNA filters were incubated for 5 min in 2 mM sodium citrate) and 5% SDS and dried for 2 min in a microwave oven for 30 s. The probe hybridization mix, which was incubated for 10 min at 37 °C before adding the mix to the chromosome preparation. The chromosome DNA was counterstained with propidium iodide or 4′, 6-diamidino-2-phenylindole when two different probes had to be detected.

**RESULTS**

**Genomic DNA Clones Containing 5′-(CGG)_{3}-Repeats**—A subgenomic library was constructed by ligating EcoRI-cleaved genomic DNA from human white blood cells into the Zap express and Zap II λ vectors as described under “Experimental Procedures.” In addition, chromosome-specific libraries (22) were screened, and we began to isolate positive clones from the chromosome 16 library, since the DNA on this chromosome had already shown expansions in the 5′-(CGG)_{3}-containing region of the locus FRA16A at 16p13.11 in some individuals (24, 25). All the clones isolated so far are schematically presented in Fig. 1. The total lengths of the nucleotide sequences and the nature and locations of the 5′-(CGG)_{n}-repeats actually detected in each of the cloned DNA segments are also indicated. In addition to the 5′-(CGG)_{3}-repeats, repeats with different sequences were also present in two of these DNA segments. The actual nucleotide sequences were not reproduced here. They are available via accession numbers AJ001215, AJ001216, AJ001217, AJ001218, and AJ001219 from the EMBL data base for clones P4, E7, P5-5, CL16-1, and CL16-8, respectively. Table I summarizes the homologies found among the individual DNA segments in the GenBank using the Blast program.

**Chromosomal Localizations of Human DNA Segments Carrying 5′-(CGG)_{3}-Repeats**—The isolated human DNA segments with 5′-(CGG)_{3}-repeats have been mapped to specific locations on the human chromosomes (i) by hybridization to a somatic cell hybrid panel (Oncor), (ii) by fluorescent in situ hybridization of the biotinylated BAC clones to human chromosome spreads (Fig. 2), and (iii) by hybridization to radiation hybrid panels (31). The CE16-1 and the CL16-8 clones were confirmed to be derived from human chromosome 16, and the P5-5 clone was located on chromosome 10 (data not shown). The mapping results of the CL16-1 and P5-5 clones were in line with the locations of their homologous sequences (Table I). The E7 and P4 clones mapped to chromosome 6q. The E7 genomic fragment mapped to 6q25-26 (Fig. 2a) and the P4 segment to 6q22-23 (Fig. 2b). The clone P4 has been more precisely assigned to chromosome 6q using EcoRI/Nor1 digestion and radiation hybrid mapping that has been performed at the Sanger Center, Cambridge, UK. The P4 clone is closely linked to the marker stSG11083, which lies between markers D6S1004 and D6S1620 on the long arm of chromosome 6 at position 510.7 cm from the Sanger Center’s radiation hybrid map (data not shown).

**Transcription of the Newly Identified DNA Segments Containing 5′-(CGG)_{3}-Repeats**—The 2^{32}P-labeled subfragments of the clones were hybridized to a multiple tissue Northern blot
(CLONTECH), to a Master blot (CLONTECH), or to a Northern blot on which poly(A)-selected RNAs from different tumor cell lines were immobilized. The human DNA segment in clone P4 was transcribed in skeletal muscle into RNAs of 5.1, 3.3, 2.8, and 1.8 kb and in placenta into RNA of 2.8-kb lengths (Fig. 3a).

This apparently tissue-specific 2.8-kb RNA was also present in the human tumor cell lines A172, 293, A549, KB, HeLa, and Hep2 (Fig. 3b).

In all the tissues tested, the DNA segment in the E7 clone was transcribed into a 5-kb RNA molecule (Fig. 4, a and b). None of the tested tumor cell lines expressed this segment. The brain-specific transcripts, detected by the EcoRI-EcoRI fragment as hybridization probe were 9.6, 4.8, and 3.8 kb in lengths (Fig. 4c). The 5-kb transcript present in all tested tissues (Fig. 4b) was not detected with the EcoRI-EcoRI fragment of E7 in Fig. 4c. To assess the extent of this brain-specific expression, we probed a master blot carrying dot-blotted poly(A) RNAs from 50 different human tissues (Fig. 4d). Transcripts were mainly found in different parts of the brain, with a stronger signal in cerebellum. Significant expression was also seen in some glandular tissues like the pituitary gland (D4), the thymus (E5), and to a lesser extent in the thyroid (D6) and the adrenal gland (D5), as well as in fetal brain, kidney, and thymus (Fig. 4c). Transcription analyses were not conducted...
for the CL16-1 and P5-5 clones. However, the CL16-8 clone did not allow the detection of RNA molecules by hybridization to the multiple tissue Northern or the master blot.

**Lack of Expansions of 5’-(CGG)n-3’ Repeats in Fragile X Patients—**We next searched for polymorphisms within the repeats in the DNA loci E7, P4, and CL16-8 from healthy individuals. The possibility existed that in fragile X patients 5’-(CGG)n-3’ repeats outside the bona fide location on Xq27.3 would also become subject to expansions. Therefore, DNAs from several full expansion fragile X patients were analyzed. Southern blotting experiments did not reveal abnormal mobility of the corresponding P4 (Fig. 5, a–c), E7 (Fig. 5, d–f), or CL16-8 bands (data not shown) in normal individuals (Fig. 5, a and d) or in fragile X individuals (Fig. 5, b and e). Minor variations in the repeat-containing regions were assessed by PCR using primers indicated in the scheme of each clone (see maps in Fig. 1, a and b). Changes in sizes could not be detected in the tested DNAs. Because the trinucleotide repeats observed were cryptic and complex, we cloned some of the PCR products from P4 and regions R2 plus R3 of the E7 DNA segment in a suitable vector to sequence individual clones, as described under “Experimental Procedures.” Only minor polymorphisms were observed in region R2 of the E7 locus in healthy and fragile X individuals (Table II). The control PCR products from the E7 plasmid remained constant throughout the cloning procedure. It is concluded that the 5’-(CGG)n-3’ repeats in various segments of the human genome outside the Xq27.3 locus are not expanded in fragile X patients. The P4 and E7 loci are not markedly polymorphic.

**No Rearrangements in the P4 and E7 Loci in Human Tumor Cell Lines—**Because of the differences in expression of the E7 and P4 DNA segments between different human tissues and human tumor-derived cell lines, we decided to analyze repeats in the DNAs from these cell lines as well as from two colon carcinoma cell lines, A549 and HeLa. The latter is known to exhibit (CA)n microsatellite instabilities. Analyses of the DNAs from these tumor cell lines by Southern blotting (Fig. 5, c and f) or by PCR experiments (data not shown) did not reveal any major rearrangements in either

---

**Table I**

| Clones     | Homology to references |
|------------|------------------------|
| CL16–1     | Human casein kinase II α’ subunit mapped on 16p13.2-p13.3, HSA1CKII 26 |
| P5–5       | Human calcium-activated potassium channel mapped on 10q22.3 28 |
| E7         | DAN 15 EMBL: Y08266 30 |
| P4 EST g1364621 |
| CL16–8     | None |

---

**Fig. 2.** Chromosomal localizations of (a) the E7 (chromosome 6q25-26) and (b) the P4 (chromosome 6q22-23) clones. In a, the E7-matching BAC clone was biotinylated (cy3-red) and hybridized to spread metaphase chromosomes together with the digoxigenin-labeled chromosome 6-specific probe (fluorescein isothiocyanate green), the 6q27 TBP (Oncor). In b, the corresponding P4-biotinylated BAC clone was hybridized to spread metaphase (fluorescein isothiocyanate signal) chromosomes. Chromosomal DNA was stained with (4’, 6-diamidino-2-phenylindole) (a) or propidium iodide (b).

**Fig. 3.** Transcription pattern of the P4 DNA segment. a. P4 DNA transcription was assessed by hybridizing the 4-0 and 4-1 subfragments (see map in Fig. 1) to RNA preparations from different human organs on an RNA transfer blot. b, the same probes as used in a were hybridized to 8 μg of poly(A)-selected RNAs from various human tumor cell lines as indicated. These RNAs were fixed to a Qiagen nylon membrane. A172 and U138 are glioblastoma cell lines, 293 is an Ad5-transformed primary embryonal kidney cell line, A549 is a lung carcinoma cell line, KB is an oral epidermoid carcinoma cell line, HeLa is a cervix epitheloid carcinoma cell line, and HEp2 is a larynx epidermoid carcinoma cell line. The filter-bound RNAs were subsequently hybridized to a β-actin probe as a control after the previous probes had been removed by boiling for 10 min in 0.1× SSC, 0.1% SDS (autoradiogram in the bottom part of b).
locus. Therefore, the observed differential expression (loss of tissue specificity) could be attributed to alterations of DNA methylation in the P4 region. However, some deletions occurring upstream and/or downstream the 6170-bp E7 fragment and/or mutations in the E7 sequence could still be responsible for the complete absence of E7 expression in tumor cell lines.

**DISCUSSION**

In several human genomic DNA fragments we have identified 5′-(CGG)$_n$-3′ repeats located 5′ to the previously identified cDNA sequences of the genes for the Ca$^{2+}$-activated potassium channel and the casein kinase II α′ subunit (CSNK2A2). This enzyme is known to have three possible isoforms α$_2$β$_2$, ααβ$_2$, and α′β$_2$. The α and α′ polypeptides are the catalytic subunits. The α′ cDNA can detect several mRNAs in Northern blot experiments (26). Alterations of the repeat-rich region located 5′ to the CSNK2A2 gene might play a role in the regulation of different levels of mRNAs.

In the case of spinocerebellar ataxia 6, the gene had been known for a long time before it could be assigned to a disease (32–34). Interestingly, the human DNA segments investigated here exhibit a combination of several types of repeats, CGG, CAG, CCT, GGA, and GAG. Within a 1.8-kb fragment, the DNA segment E7 clone carries most of the identified repeat sequences.

So far, none of the new 5′-(CGG)$_n$-3′ repeat-containing segments have been found expanded in fragile X patients, nor are these repeats highly polymorphic in the tested healthy individuals. The instability observed at the expanded FRAXA locus in fragile X patients is most likely not the result of a widespread instability in the genome but rather confined to this single locus. A similar conclusion has been reached by Goellner et al.

**FIG. 5.** Assessment of the stability of the E7 and P4 loci. Variations in size at the P4 (a–c) and E7 loci (d–f) were investigated by cleaving 15 μg of genomic DNA with EcoRI plus PspI in (a–e), whereas in f, DNA was cut with EcoRI plus PstI. In the E7 locus, the 2162 and 2180 bp fragments contain both the repeats R1 and R2; repeats R3 and R4 are located in the 2976- and 2734-bp fragments. The 1394-bp fragment in the P4 locus contains the repeat. In a and d, cleavage patterns of DNA from normal individuals, in b and e from fragile X individuals, and in c and f from human tumor cell lines are shown. The lanes E7 and P4 are plasmid controls. In panels a to c, the 32P-labeled DNA fragment 4-0 was used to probe the DNA on the membranes, and the 7-3 and 7-42 fragments were used in panels d to f. Letter combinations refer to DNAs from control (a, d) or fragile X individuals (b, e).

**TABLE II**

The relevant part of R2 is reproduced here; the sequences were analyzed after sequencing the single clones. The 10 isolated clones from the plasmid control E7 did not show any variations.

| R2 Consensus | (CTT)$_4$(CGG)$_n$ | (CTT)$_4$(CGG)$_n$ | (CTT)$_4$(CGG)$_n$ | (CTT)$_4$(CGG)$_n$ | (CTT)$_4$(CGG)$_n$ |
|-------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| N2          |                   |                   |                   |                   |                   |
| RF          |                   |                   |                   |                   |                   |

and only a part of the complete 5′-(GCC)$_n$-3′-rich region is included in the different transcripts reported.

Clone CL16-1 is a genomic DNA fragment comprising the 5′ region of the casein kinase II α′ subunit (CSNK2A2). This clone contains the 5′- (GCC)$_n$-3′ repeats located 5′ to the previously identified cDNA sequences of the genes for the Ca$^{2+}$-activated potassium channel (K$_{Ca}$) proteins are a large and diverse family of ion channels. All the different isoforms of the K$_{Ca}$ are generated by alternative splicing of a single gene mapped on 10q22.3 (28, 29). Although the expression of the K$_{Ca}$ is widely distributed, the highest levels of expression are found in brain, aorta, and skeletal muscle (28). So far, only the different cDNA sequences of the gene were reported. The genomic clone P5-5 isolated here contains the 5′ region of the different spliced mRNA molecules.
(35) on the instabilities within simple tandem repeats (STR) at 10 loci in 29 Huntington disease families that have been compared with instabilities at the same loci in 29 colon cancer patients with defects in mismatch repair enzymes. The hypothesis that secondary structures adopted by CAG, CGG, or GAA repeats in the trinucleotide repeat diseases might mediate instability is supported by the work of Gacy et al. (36). This group has suggested that the improper DNA structure at the repeat region is responsible for GAA instability in Friedreich's ataxia.

The tissue-specific transcription displayed by the P4 DNA segment is apparently erased in the tumor cell lines. In only one of the two glioblastoma cell lines, the 2.8-kb transcript has been identified. The location of the P4 clone, 6q22-23, and its strong muscle expression renders it a possible candidate gene for muscular dystrophy. Indeed it maps close to some other genes also expressed in muscle, e.g., lamA2 and lamA4. The lamA2 gene is involved in merosin-deficient congenital muscular dystrophy (37).

Surprisingly, the 5-kb transcripts ubiquitous in many tissues were absent in all tumor cell lines tested. Could this region be involved in the process of tumorigenesis because of its location at 6q25-26? Several authors reported that the telomeric part of the long arm of chromosome 6 might contain tumor suppressor genes (15, 38–39).

The region R3 with its 5’-(CAG)3’-3’ triplet repeat and region R4 of the E7 locus were transcribed. The EST1 clone, which is homologous to the E7 region detecting the brain-specific transcripts, was only 1,283 nucleotide pairs long. The sequence available in the data base from the DAN15 clone is 575 bp long. The region R3 with its 5’-(CAG)3’-3’ triplet repeat region is responsible for GAA instability in Friedreich's ataxia.

Acknowledgments—We are very grateful to Andy Mungall from the Sanger Center, Cambridge, UK for help with the chromosome 6 detailed mapping. We thank Helmut Deissler, Institute of Cell Biology, University of Essen, Essen, Germany, for cell lines SW480 and LoVo. We also thank Petra Böhm for expert editorial work.

REFERENCES

1. Ashley, C. T., Jr., and Warren, S. T. (1995) Annu. Rev. Genet. 29, 703–728.
2. Warren, S. T. (1996) Science 271, 1374–1375.
3. Koshy, B. T., and Zoghbi, H. Y. (1997) Trends Genet. 13, 364–372.
4. La Spada, A. R. (1997) Brain Pathol. 7, 313–321.
5. Thornton, C. A., Wymer, J. P., Simmons, Z., McClan, C., and Moxly III, R. T. (1997) Nat. Genet. 16, 407–409.
6. Kleset, T. R., Otten, A. D., Bird, T. D., and Tassop, J. (1997) Nat. Genet. 16, 402–406.
7. Sutherland, G. R. (1991) Genet. Anal. Tech. Appl. 8, 161–166.
8. Sutherland, G. R., and Richards, R. I. (1995) Curr. Opin. Genet. Dev. 5, 323–327.
9. Oostra, B. A. (1996) Acta Genet. Med. Gemellol. (Roma) 45, 93–108.
10. Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, D., de Jong, P., Jansen, G., Anvret, M., Riley, B., Williamson, R., and Doel, T. (1995) Science 269, 1259–1267.
11. Canis, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., P. W., Dubel, J., and Kuroda, Y., Nishimura, M., and Nakamura, Y. (1997) Hum. Mol. Genet. 39, 252, 1097–1102.
12. Saito, S., Sirahama, S., Matsushima, M., Suzuki, M., Sagae, S., Kudo, R., Saito, J., Noda, K., and Nakamura, Y. (1996) Cancer Res. 56, 5585–5589.