SHORT COMMUNICATION

Characterization of Four Novel CAG Repeat-Containing cDNAs

JIN-XING JIANG,[1] RONALD H. LEKANNE DEPREZ,[2] ELLEN C. ZWARTHOFF,[3] AND PETER H. J. RIEGMAN

Department of Pathology, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Received April 4, 1995; accepted July 14, 1995

Stretches of CAG nucleotides coding for the amino acid glutamine are an important feature of many transcription factors and genes that are involved in neurodegenerative disorders. In an attempt to isolate CAG repeat-containing cDNAs expressed in nervous tissue, we screened a human fetal brain cDNA library with a probe containing a CAG repeat. Five different clones were characterized and found to contain CAG repeats. Sequence data revealed that four of these cDNAs were derived from novel genes. These cDNAs were designated CAG6, CAG12, CAG24, and CAG40 and were found to correspond to transcripts of 5.0, 7.5, 4.4, and 15 kb, respectively. The genes encoding CAG6, CAG12, CAG24, and CAG40 were assigned to chromosomes 12, 16, X, and 12, respectively. For the 5th gene, CAG26/pRHpA, a localization on two different chromosomes was established: 16 and X. None of the repeats showed any length polymorphisms in human DNA.

Over the last few years CAG repeats have received a lot of attention, mainly because of their surprising role in neurodegenerative disorders. So far five different neurodegenerative disorders have been reported in which expansion of CAG repeats within the open reading frame is the cause of the disease: spinal and bulbar muscular atrophy (SBMA) (5), Huntington disease (HD) (3), spino-cerebellar ataxia type 1 (SCA1) (8), dentatorubral and pallidoluysian atrophy (DRPLA) (7), and Machado–Joseph disease (MJD) (4). In all cases the repeats show a considerable degree of length polymorphism in the population in general. They are often interrupted by alternative triplets, and it has been suggested that the alternative triplets play a role in the stabilization of the repeat (1). Only for SBMA, in which the CAG repeat in the gene encoding the androgen receptor is affected, is a clearcut function for the gene causing the disease known. However, since over 80% of the proteins containing glutamine stretches are transcription factors (2), it is most likely that this will also be the case for at least some of the other genes.

In an attempt to isolate additional genes containing CAG repeats, we screened a human fetal brain cDNA library (Stratagene 936206) with a 0.5-kb, CAG repeat-containing, *Pst*I fragment from the human *MN1* gene (6). From these five unique cDNAs were selected for further characterization. These cDNAs were designated CAG6, CAG12, CAG24, CAG26, and CAG40. The clones were sequenced over the CAG repeat (see Fig. 1). Four of the five cDNAs showed no significant homology apart from the repeat with other genes in the GenBank (release 86) and EMBL (release 41) databases and are therefore derived from genes that have not been described so far. Sequence analysis showed that the insert from CAG26 is a composite of two different cDNAs, of which one is identical to an earlier described CAG repeat-containing cDNA, pRHpA (9). The sequence data of the four novel cDNAs were submitted to the EMBL data library, and the accession numbers for the sequences are CAG6, X85326; CAG12, X85324; CAG24, X85323; and CAG40, X85325. All four novel CAG repeats were tested for length polymorphisms on DNA isolated from 44 nonrelated normal individuals. None of them showed length variation (results not shown). This had already been established for pRHpA (9).

The transcript length and expression levels in several human tissues were determined by hybridizing Northern blots with probes that were adjacent to the CAG repeats in the cDNAs; the results are shown in Fig. 2. Transcript lengths of 5.0, 7.5, 4.4, and 15 kb were found for clones CAG6, CAG12, CAG24, and CAG40, respectively. Expression of CAG24 and CAG40 was very low. CAG40 was detectable only in skeletal muscle, whereas CAG6, CAG12, and CAG24 were found in all five tissues tested. The expression of the gene for pRHpA could not be tested, because

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession Nos. X85323–26.

[1] Present address: Pediatric Research Institute, Cardinal Glennon Children's Hospital, 3662 Park Avenue, Saint Louis, Missouri 63110.

[2] Present address: Department of Anatomy and Embryology, Academic Medical Center, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands.

[3] To whom correspondence should be addressed. Telephone: (3110) 4087929. Fax: (3110) 4366660. E-mail: Zwarthoff@pa1.fgg.eur.nl.

GENOMICS **30,** 91–93 (1995)
0888-7543/95 $12.00

CAG6:

ctaggtgccccagaagtgtgcaaccaggtcacctcatcccagtctgtccgggggcttctc

ccttgtactgcc**CAGCAGCAGCAGCAGCAGCAGCAGCAG**cttcctgctctcccaccc

acgcct**CAGCAGCAG**ccacccttgaataatcacatgatctcacagccagtcccggctctg

CAG12:

tgccatgcagagtaactctccatcccaggaa**CAGCAGCAGCAGCAG**caa**CAGCAGCAG**ca

a**CAGCAGCAGCAG**caacaa**CAG**agcattttattcagtaatcagaataccatggctagaat

CAG24:

tgcgtgcagctgaggagtccacagcccgtaaccttaac**CAGCAGCAGCAG**caacaacaa**C**

AGCAGCAGCAGcaacaacaa**CAGCAGCAG**caacaacaa**CAGCAGCAGCAGCAGCAGCAG**g

tggcagcagctgcaggggcacgcacagagcatccaccaggcccccccaacttcggcgactg

ttatgagcacatgcacccactcccttaaaaaccccttattaagtagggcatatgcacacct

gtccccccttcccc**aaccatggccccaggtttga**gctttgtggtcttgcatcttgtggtga

CAG26/pRHpA:

taacttggagaagcaaaaccatactccaagaaagcatcatcaacatcaccac**CAGCAGCA**

Gcaccac**CAGCAG**caa**CAGCAGCAG**ccgccaccacc//caatacctgcaaatggcaacag

CAG40:

agagcgctcagcccatgggctctttaa**CAGCAG**ctt**CAGCAGCAGCAGCAG**ctgcaa**CAG**c

aa**CAG**caacttt**CAGCAGCAGCAGCAGCAGCAG**ctacaa**CAG**caa**CAG**caacttt**CAGCAG**c

aa**CAG**cttcaa**CAGCAG**caa**CAGCAGCAGCAG**cttcaacaa**CAGCAGCAG**caa**CAG**cttc

aa**CAGCAG**caa**CAGCAG**ctacaa**CAG**caa**CAG**caacaacaa**CAGCAGCAG**tttcaa**CAGC**

AGCAGcaa**CAGCAGCAG**atgggccttttaaaccagagtcgaactttactgtctcctcagc

FIG. 1. Sequence of the five clones over the CAG repeat. The primers used for PCR are underlined. CAGs belonging to a CAG repeat are shown in uppercase and boldface letters. The accession numbers for the novel sequences are CAG6, X85326; CAG12, X85324; CAG24, X85323; and CAG40, X85325.

it was not possible to derive an appropriate probe from CAG26 due to its composite nature.

To map the genes of the different cDNAs in the human genome, DNA isolated from a panel of human/mouse and human/hamster hybrid cell lines (BIOS Laboratories) was used for PCR over the CAG repeat. This approach allows an unequivocal assignment of a given PCR product to a human chromosome. However, with primers for the CAG repeat genes described in this paper, it appeared that PCR products were generated not only from human but also from mouse and hamster DNA (see Table 1). This is probably due to a high degree of conservation of the primer sequences in rodent DNA. The genes for cDNAs CAG24 and CAG40 could be assigned to human chromosomes X and 12, respectively, because the length of the human PCR product differed from that produced on rodent DNA. We were able to remove the rodent signal by employing a higher temperature during the annealing phase of the PCR reaction in the case of CAG12, which resulted in its assignment to chromosome 16. For CAG6 and CAG26/pRHpA, raising the annealing temperature did not resolve the problem. Discrimination between human and rodent PCR products could be achieved only after the PCR products had been separated under conditions used for single-strand conformation analysis. The results thus obtained clearly assigned

CAG6 to chromosome 12. For CAG26/pRHpA we obtained signals from two different human chromosomes, 16 and X. This means either that there presumably are two closely related CAG26/pRHpA genes in the human genome or that one of the two is a pseudogene.

The CAG repeats in the genes described here are not polymorphic in the human population, which renders them unlikely candidates for genes involved in neurodegenerative disorders. Variation in the length of the CAG repeats in these genes may not be compatible with life, and therefore a more stringent selection on their length may occur. The finding that the flanking sequences were conserved enough to allow amplification in rodent DNA resulting in PCR products of comparable or even equal length suggests that the repeat might be conserved between humans and rodents.

Interruptions of the CAG repeat of the SCA1 gene by CAA triplets were shown to have a stabilizing effect on the repeat length (1). The same interruptions were found in CAG12 and CAG24, and it is possible that in these cases the CAA triplets stabilize the integrity of the CAG repeat. In the three other clones the CAG stretches are interrupted by other triplets. There are two reasons why we think it is likely that at least some of the repeats in all four novel genes are part of the open reading frame and encode glutamine stretches. First, although the sequence of the open reading frame of these genes is not complete, the data obtained so far show that in all cases the frame containing the CAGs is the only open reading frame of the six possibilities, and second, the interruptions of the CAG repeats are in all cases multiple-of-three sequences. Since 82% of the proteins containing glutamine stretches are involved in activation of transcription (2), some of these genes could, if our hypothesis is correct, play a role in regulation of transcription.

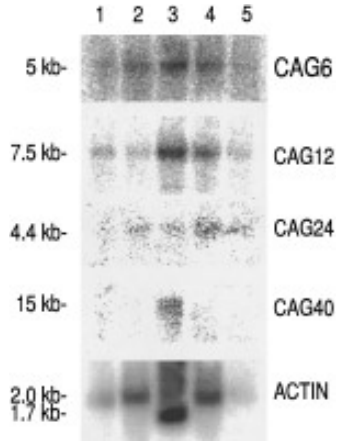

FIG. 2. Expression of the CAG genes in several human tissues. The poly(A)$^+$ RNA samples in lanes **1** to **5** are from lung, liver, skeletal muscle, kidney, and pancreas, respectively.

TABLE 1

Chromosomal Location and Conservation of CAG Repeats

cDNA	Location on human chromosome	Length of PCR product (bp)		
		Human	Mouse	Hamster
CAG6	12	104	104	104
CAG12	16	106	106	106
CAG24	X	265	—	220
CAG26/pRHpA	16 and X	105	105	105
CAG40	12	290	450	450

REFERENCES

1. Chung, M. Y., Ranum, L. P., Duvick, L. A., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1993). Evidence for a mechanism predisposing to integrational CAG-repeat instability in spinocerebellar ataxia type I. *Nature Genet.* **5:** 254–258.
2. Gerber, H. P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263:** 808–811.
3. Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971–983.
4. Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizewa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S., and Kakizuka, A. (1994). CAG expansions in a novel gene for Machado–Joseph disease at chromosome 14q32.1. *Nature Genet.* **8:** 221–228.
5. La Spada, A. R., Roling, D. B., Harding, A. E., Warner, C. L., Spiegel, R., Hausmanowa-Petrusewicz, I., Yee, W. C., and Fischbeck, K. H. (1992). Meiotic stability and genotype–phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. *Nature Genet.* **2:** 301–304.
6. Lekanne Deprez, R. H., Riegman, P. H. J., Groen, N. A., Warringa, U. L., van Biezen, N. A., Molijn, A. C., Bootsma, D., de Jong, P. J., Menon, A. G., Kley, N. A., Seizinger, B. R., and Zwarthoff, E. C. (1995). Cloning and characterization of MN1, a gene from chromosome 22q11, which is disrupted by balanced translocation in a meningioma. *Oncogene* **10:** 1521–1528.
7. Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., Tanaka, Y., Kikushima, H., Umino, K., Kurosawa, H., Furukawa, T., Nihei, K., Inoue, T., Sano, A., Komure, O., Takashi, M., Yoshizawa, T., Kanazawa, I., and Yamada, M. (1994). Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6:** 14–18.
8. Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P., and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG-repeat in spinocerebellar ataxia type 1. *Nature Genet.* **4:** 221–226.
9. Riggins, G. J., Lokey, L. K., Chastain, J. L., Leiner, H. A., Sherman, S. L., Wilkinson, K. D., and Warren, S. T. (1992). Human genes containing polymorphic trinucleotide repeats. *Nature Genet.* **2:** 186–191.