Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 417–421, January 1996
Neurobiology

Codon repeats in genes associated with human diseases: Fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration

PHILIPPE DJIAN*, JOHN M. HANCOCK†, AND HARJINDER S. CHANA†

*Centre National de la Recherche Scientifique, Centre de Recherche sur l'Endocrinologie, Moléculaire et le Développement, 9, rue Jules Hetzel, 92190 Meudon, France; and †Gene and Genome Evolution Group, Medical Research Council Clinical Sciences Center, Royal Postgraduate Medical School, Hammersmith Hospital, London W12 0NN, United Kingdom

Communicated by Howard Green, Harvard Medical School, Boston, MA, September 22, 1995

ABSTRACT Five human diseases are due to an excessive number of CAG repeats in the coding regions of five different genes. We have analyzed the repeat regions in four of these genes from nonhuman primates, which are not known to suffer from the diseases. These primates have CAG repeats at the same sites as in human alleles, and there is similar polymorphism of repeat number, but this number is smaller than in the human genes. In some of the genes, the segment of poly(CAG) has expanded in nonhuman primates, but the process has advanced further in the human lineage than in other primate lineages, thereby predisposing to diseases of CAG reiteration. Adjacent to stretches of homogeneous present-day codon repeats, previously existing codons of the same kind have undergone nucleotide substitutions with high frequency. Where these lead to amino acid substitutions, the effect will be to reduce the length of the original homopolymeric stretch in the protein.

Five human diseases are due to an excessive number of codon reiterations. These diseases are X-linked spinal and bulbar muscular atrophy (1), Huntington disease (2), type 1 spinocerebellar ataxia (3), dentatorubral–pallidoluysian atrophy (4), and Machado–Joseph disease (5). All five genes contain a region of CAG reiteration. In the normal human population, the average number of CAGs is around 20, and a number in excess of 35–40 leads to the disease. Beyond 40, the number tends to increase with successive generations and can reach values of up to 100; in such cases, clinical symptoms tend to appear very early in life. We report here the nucleotide sequences of PCR fragments containing the sites of codon reiteration in three of the homologous genes of hominoid primates: those associated with Huntington disease, spinal and bulbar muscular atrophy, and Machado–Joseph disease.‡ We have also determined the number of reiterated codons in spinocerebellar ataxia type 1 gene by sizing the PCR fragments on denaturing polyacrylamide gels. We find that the size of reiterants has tended to increase during the evolution of anthropoid primates, particularly in the human lineage and that nucleotide substitutions have occurred frequently at the sites of reiteration.

MATERIALS AND METHODS

DNA was prepared (6) from three chimpanzees (La, Li, and Ga), three gorillas (F, K, and M), and a gibbon. The sources of the DNAs have been described earlier for the chimpanzees (7), the gorillas (8), the gibbon (9), and the cercopithecine monkeys (10).

A 200- to 300-bp fragment encompassing the sites of poly(CAG) expansion was amplified by PCR. For amplification of the fragments of the Huntington disease, spinal and bulbar muscular atrophy, and Machado–Joseph disease genes, PCR conditions were similar to those described earlier (11), except that initial denaturation temperature was 97°C and dAza-dGTP concentration was 187.5 μM. For Huntington disease and the androgen receptor genes, the temperature profile was that described earlier (11): starting annealing temperature of 70°C decreasing by 1°C for each of the first five cycles to reach 65°C, a temperature which is retained for the remaining cycles. For Machado–Joseph disease, initial annealing temperature was 65°C and was reduced by 1°C per cycle for each of the first five cycles to reach 60°C. For the Huntington disease gene, the two primers correspond to codons 8–15 and 85–92, respectively (2) (see Fig. 1). For spinal and bulbar muscular atrophy, the two primers correspond to codons 23–31 and 108–116 (12) (see Fig. 3). For Machado–Joseph disease we used primers MJD52 and MJD25 (5). After amplification, PCR fragments were purified by agarose gel electrophoresis, then cloned directly by A/T cloning (13) into pGEM-T. The sequence of the cloned PCR fragments was subsequently determined with T7 DNA polymerase. dITP was used during the labeling part of the reaction but not during the termination. Two clones were sequenced for each PCR product.

For spinocerebellar ataxia type 1, the number of reiterated codons was estimated from the size of the PCR fragments (166–199 bp depending on the number of CAGs). PCR reactions were carried out with fluorescent dUTP. Thermocycling conditions were 35 cycles at 95°C for 30 sec, 55°C for 30 sec, and 72°C for 45 sec, followed by a final extension at 72°C for 5 min. Primers Rep-2 and Rep-1 were used (3). For sizing, samples were loaded on a denaturing 6% polyacrylamide gel. Electrophoresis was performed on an ABI 373 automatic sequencer. Fragment sizes were determined by using GENESCAN software and ABI markers.

RESULTS

The Size of the Reiterants. *Huntington disease gene.* A fragment of the Huntington disease gene corresponding to codons 8–84 of the published human sequence (2) was amplified by PCR from DNA samples from a human, three chimpanzees, three gorillas, and a gibbon. The nucleotide sequences of the PCR fragments are illustrated in Fig. 1. This region contains four sites of codon reiteration. Site 1 contains a CAG reiterant that is highly polymorphic in normal humans and, when excessively long, causes the disease. Sites 2, 3, and 4 contain CCG reiterants unrelated to disease. Site 2, adjacent to the CAG reiteration site, is also polymorphic in the normal human population (14, 15). All sites contain a homogeneous reiterant (underlined in Fig. 1) flanked by variant codons differing from the homogeneous reiterant by a single nucleotide.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡The sequences reported in this paper have been deposited in the GenBank data base (accession nos. L49348–L49375).

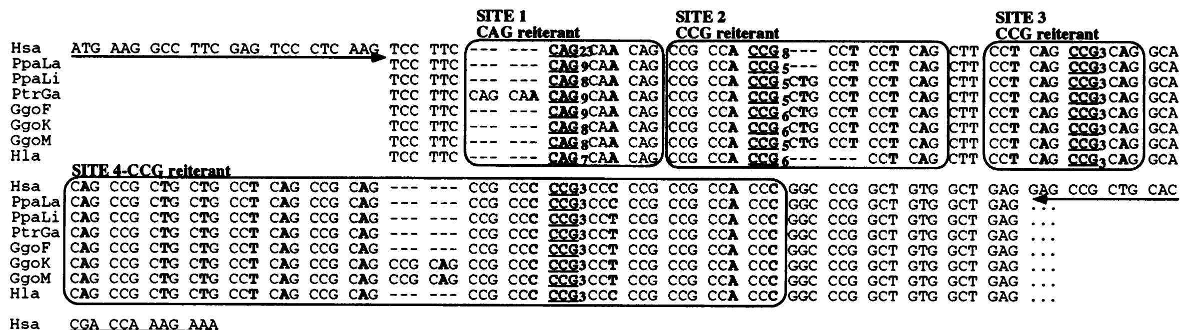

FIG. 1. The gene associated with Huntington disease: nucleotide sequences of the reiteration sites of hominoid primates. PCR fragments containing the three reiteration sites were generated from the first exon of the Huntington disease gene. The fragments correspond to codons 8–84 of the human gene (2). The sequences of the PCR fragments were determined in one human (Hsa), two *Pan paniscus* individuals (La and Li, abbreviated PpaLa and PpaLi, respectively), one *Pan troglodytes* individual, three *Gorilla gorilla* individuals (GgoF, -K, and -M), and one *Hylobates lar* individual (Hla). Nucleotide sequences of the PCR primers are above the arrow in the human sequence. Sites of codon reiteration are framed. Each reiteration site contains a homogeneous stretch of at least three codons (boldface and underlined), flanked by numerous codons differing from the homogeneous reiterant by a single nucleotide (boldface). Site 1 is the CAG reiteration site associated with the disease. Sites 2, 3, and 4 are CCG reiteration sites not associated with the disease. Nucleotide substitutions in previously homogeneous reiterants are indicated in boldface.

The number of CAG codons at site 1 in the normal human that we have examined is 23. In the initial report of the range for normal human alleles, the number varied between 11 and 34 (2). From examination of large numbers of alleles, a normal modal number of 19–20 was later reported (16). Chimpanzees and gorillas, like the human, are polymorphic for number of CAG repeats. The three chimpanzees, the three gorillas, and the gibbon have between seven and nine CAG codons, a number smaller than the smallest number reported in humans. These numbers are in agreement with those reported by Rubinsztein *et al.* (17).

The number of CCG reiterants at site 2 tends to be smaller in the hominoid primates than in the human, but the difference is much less than for the CAG codons. In the human, although the range is 7–12, most alleles possess either 7 or 10 CCG codons (14), whereas in the African apes we find 5 or 6. It is interesting that in nonhuman primates the number of CAGs at site 1 is only slightly larger than that of CCGs at the adjacent site 2: 7–9 versus 5 or 6. Since in the human there has been a large increase at site 1 and a very small increase at site 2, expansion in the human lineage has been targeted preferentially to the CAG reiterant. It has been reported that Huntington disease alleles with large numbers of CAGs tend to possess small numbers of CCGs (17). In these genes, the reiteration mechanism has been entirely focused on the CAG codons.

Both site 3 and site 4 contain a homogeneous stretch of three CCGs that is conserved by all the species examined.

Machado–Joseph disease gene. The nucleotide sequence containing the unique poly(CAG) site is given in Fig. 2. The single human allele that we sequenced contains 16 CAGs, consistent with the normal range of between 13 and 36 (5). In the five alleles of *Pan* (one *P. troglodytes* and four *P. paniscus*) the range of CAGs was between 2 and 13. The gorilla and gibbon had 2–7. The 5′-most CAG trinucleotide belongs to the 3′ end of the intron and is part of the splice site.

Spinal and bulbar muscular atrophy: the androgen receptor gene. The nucleotide sequence of the PCR fragments is given in Fig. 3. Two sites of CAG reiteration are present. A third site, containing five CAGs in the human, was not examined because it is located downstream of the amplified sequence. The upstream site (site 1) is associated with the disease. The single normal human allele that we sequenced has 25 CAGs, consistent with the range in normal humans of 11–31. In the hominoid primates, we found a range of 4–17 CAG codons. In the human, less than 4% of alleles have fewer than 17 CAGs (18). The human is not polymorphic at the downstream site (site 2), and the number of CAGs at this site in other hominoids is similar, though perhaps smaller.

Spinocerebellar ataxia type 1. The number of CAG codons in this gene of the nonhuman primates was determined by sizing short PCR fragments on a polyacrylamide gel. The values obtained for the three chimpanzees, the three gorillas, the gibbon, and four cercopithecines are shown in Table 1. In the human population, normal alleles can vary widely in the size of the CAG reiterant: numbers of 6–39 (19) and 19–36 (20) have been reported. However, the modal number of CAGs is 30; 65% of alleles contain 28–30 repeats and 95% contain 25–33 repeats (20). Although the average values found in *Pan, Gorilla, Hylobates*, and cercopithecines would be within the normal range for the human, they are smaller than the modal

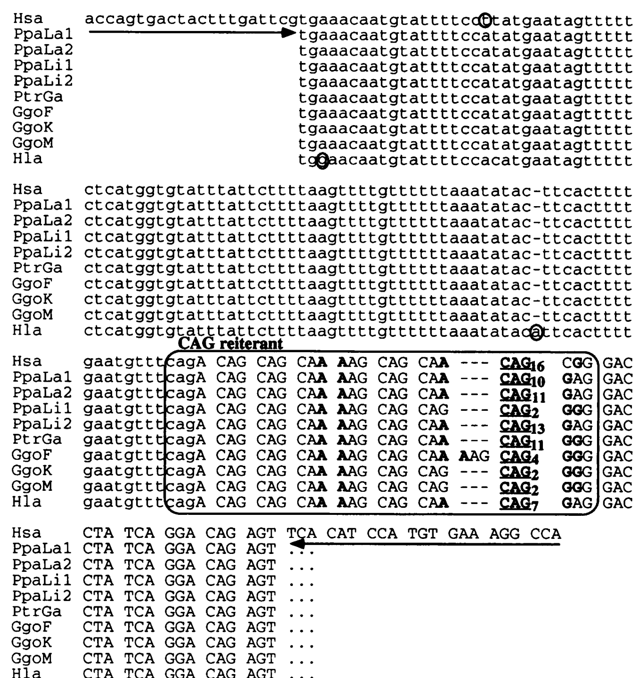

FIG. 2. The gene associated with Machado–Joseph disease: nucleotide sequences of the reiteration sites of hominoid primates. The PCR fragments correspond to part of an intron (in lowercase) and codons 292–330 (5). Symbols are as in the legend to Fig. 1. Single-base mismatches in the intron are circled. When two different alleles were found in one sample, they are designated as 1 or 2—i.e., PpaLi1 and PpaLi2.

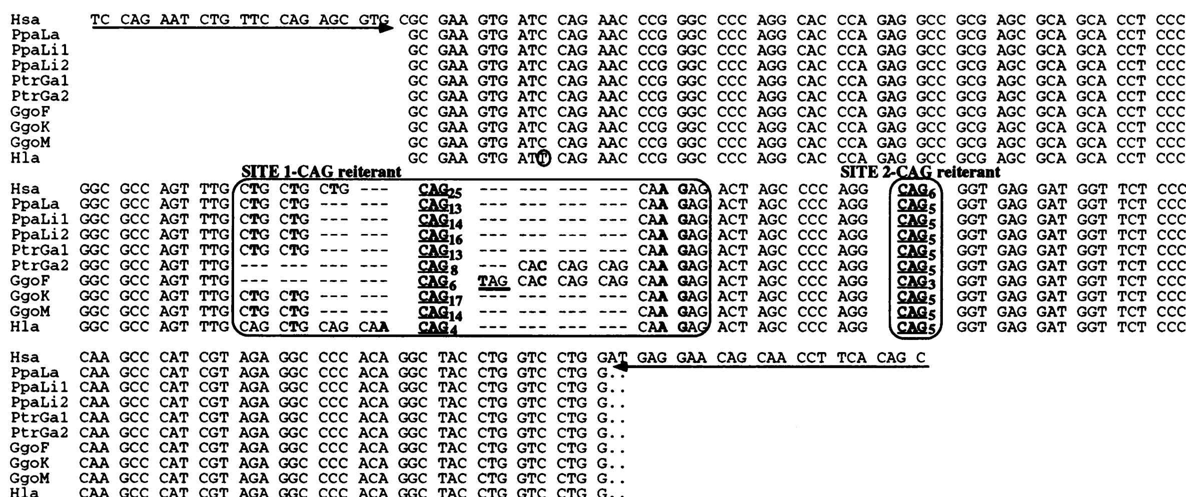

FIG. 3. The gene associated with spinal and bulbar muscular atrophy: nucleotide sequences of the reiteration sites of hominoid primates. The sequences correspond to codons 23–115 of the human gene (12). Symbols are as in the legend to Fig. 1. When two different alleles were found in the same animal, they are designated 1 or 2. A TAG termination codon in GgoF has a thick underline. A single-base mismatch in the coding region upstream of the reiteration site is circled.

number of the human (Table 2). This is particularly true for the gibbon and the cercopithecines.

Other Types of Mutations at the Reiteration Sites. Nucleotide substitutions are clustered at the reiteration sites. As clusters do not occur elsewhere in the sequences of the PCR fragments, they must be related to the reiteration process. For the most part, these mutations could be regarded as single-nucleotide substitutions in CAG or CCG codons. Some of these substitutions are found in the sequences of all the primates, and others are present only in some species or individuals. Some of the mutations would lead to amino acid replacement, but a significant number would be silent.

Huntington disease gene. The four reiteration sites contain 16 substitutions common to all alleles examined: 8 are silent and 8 are missense (Fig. 1). Other mutations are found only in some alleles. At site 1, PtrGa possesses nine CAG codons and two additional glutamine codons, one encoded by CAG, the other by CAA. It is likely that the CAA codon originated by a G to A transition in a CAG codon that was part of an uninterrupted series of 11 CAGs. At site 2, PpaLi, PtrGa, and the three gorillas have an extra CTG codon following the CCG reiterant. This CTG may have been generated by a C to T transition in a preexisting CCG. *Hylobates* is missing a CCT, which must have been either deleted in the lineage leading to the gibbon or added by duplication of the neighboring CCT in the lineage ancestral to the African apes and man. At site 4, there is a silent G to T transversion in the codon located just downstream of the homogeneous CCG reiterant. There is also a duplication of a CCG CAG hexanucleotide in GgoK and -M.

Table 1. The gene associated with spinocerebellar ataxia type 1: number of reiterated CAG codons in hominoid primates and Old World monkeys

Sample	No. of CAGs
PpaLa	24 and 25
PpaLi	24 and 29
PpaGa	23 and 24
GgoF	21
GgoK	20 and 21
GgoM	20 and 21
H. lar	14
Macaca fascicularis	14 and 15
Macaca mulatta	15
Cercopithecus aethiops	14 and 18
Cercopithecus hamlyni	14

Numbers were estimated by sizing short PCR fragments on denaturing polyacrylamide gels.

Machado–Joseph disease gene. The CAG reiterant contains one silent and one missense substitution, both common to all alleles (Fig. 2). An isolated A nucleotide begins the exonic part of the reiteration site in all genes examined; it splices to an AA dinucleotide from the previous exon thus forming the codon AAA (5). The A nucleotide could have resulted either from an insertion of an A or from a deletion of C and G nucleotides in a preexisting CAG codon.

Other mutations are found only in some alleles. Two variant codons (CAA and AAG) are present immediately upstream of the homogeneous CAG reiterant. In the CAG position, seven alleles have converted to CAA. The AAG codon is an extra codon found in GgoF alone. It was probably generated by a single substitution in one of the CAGs of the adjacent homogeneous reiterant. Immediately downstream of the reiterant, there have been numerous substitutions in the first two nucleotides; three variant codons are found: CGG, GGG, and GAG. There is a total of five variant codons in the three positions immediately adjacent to the homogeneous reiterant. These must have been generated by at least five independent nucleotide substitutions, a frequency of 5/60 or 0.083 substitutions per nucleotide. Two nucleotide substitutions and one single base deletion or insertion are present in the intronic part of the reported sequence, a frequency of 3/850 or 0.003 mutations per nucleotide. The mutation rate at the two codons adjacent to the poly(CAG) is therefore 27 times higher than in the intron, or 40 times higher if one counts only substitutions. The difference is particularly striking as the intron is itself repetitive and might be expected to undergo replication slippage with high frequency. Substitutions in and adjacent to the site of expansion have been observed in some human alleles by Kawaguchi *et al.* (5); the segment of CAG repeats was broken by two interspersed CAAs and one AAG, and at the 3′ boundary of the reiteration site CGG and GGG variant codons were present.

The androgen receptor gene. The CAG reiterants contain only two substitutions shared by all alleles: a silent transition and a missense transversion, both present at site 1 (Fig. 3). Numerous other mutations are found within reiteration-site 1, but they are not shared by all the genes examined. Two variant codons are present upstream of the homogeneous reiterant (CAA or CTG), and three are present downstream (CAA, CAC, and TAG). A total of 14 codons have either been deleted or inserted at the reiteration site. In the rest of the coding sequence reported, there is only one nucleotide substitution

Table 2. Number of codon repeats in mammalian genes homologous to those associated with four human diseases

	Huntington disease*		Spinal and bulbar muscular atrophy (androgen receptor)*		Machado–Joseph disease*	Spinocerebellar ataxia type 1†
	Site 1 CAG	Site 2 CCG	Site 1 CAG	Site 2 CAG	CAG	CAG
Normal human						
Range	11–34	7–12	11–31	6	13–36	6–39
Modal number	19–20		21		19	30
Pan	8 or 9	5 or 6	8–14	5	2–13	23–29
Gorilla	8	6	6–17	3–5	2–4	20–21
Hylobates	7	6	4	5	7	14
Cercopithecines	—	—	—	—	—	14–18
Mouse	7	3 or 4	1	2	—	2
Rat	—	—	1	2	—	—

Sites 1 and 2 are the two polymorphic reiteration sites found in the genes associated with Huntington disease and spinal and bulbar muscular atrophy, respectively (see Figs. 1 and 3). For the human, normal modal numbers and normal ranges are given; for the other species, only ranges are given because of the small sample sizes. For the normal human, the sources for the number of codon repeats are as follows: Huntington disease gene (16, 21), gene for spinal and bulbar muscular atrophy (1, 18), Machado–Joseph gene (5), spinocerebellar ataxia gene (19, 20), the values for the mouse and rat androgen receptor gene (22–24), the values for the mouse Huntington disease gene (25), and the values for the mouse spinocerebellar ataxia type 1 gene (44).
*Value obtained by sequencing of PCR fragments.
†Value obtained by sizing of PCR fragments.

and no insertions or deletions. PtrGa2 and GgoF have obvious similarities: (*i*) they do not possess CTG codons at the 5′ boundary, (*ii*) they have a small number of CAG reiterants, and (*iii*) they contain additional trinucleotides at the 3′ boundary (CAC CAG CAG for PtrGa2 and TAG CAC CAG CAG for GgoF). It is probable that the CAC codon was generated by a G to C transversion and the TAG termination codon by a C to T transition. However, since gorilla F was a female, this mutation may have been of no consequence. Although a number of mutations have been identified in the androgen receptor gene of patients suffering from testicular feminization (26, 27), it is only recently that a mutation similar to that described here has been found. A CAG to TAG substitution within the poly(CAG) (at glutamine 60) has been identified in two 46 XY siblings with complete testicular feminization syndrome. In these patients, a truncated receptor containing the C terminus was made because of initiation at a downstream methionine (28, 29). We do not know whether downstream initiation occurs in the gorilla mutant. A stop codon occurring within an expanded poly(CAG) would probably prevent spinal and bulbar muscular atrophy, but in a male the neurological disease would be replaced by testicular feminization syndrome.

DISCUSSION

In a previous study of sizes of reiterants in the Huntington disease genes of various anthropoid primate species, Rubinsztein *et al.* (17) observed that nonhuman primates have smaller numbers of repeats than humans. They derived the length of the reiterant by sizing the PCR fragments but did not determine the nucleotide sequence of the PCR fragments.

Two main conclusions can be derived from our more extensive study: (*i*) There has been a trend of expansion of CAG repeats in the four genes of higher primates; and (*ii*) in the three genes for which we have sequence information, nucleotide substitutions are particularly frequent within or at the boundaries of the expanded sites.

Evidence supporting the first point is summarized in Table 2. In all four genes, CAG reiterants are smaller in the monkeys and apes than in the human. Differences in number could be due either to a higher rate of expansion in the human than in the nonhuman primates or to contractions in the nonhuman primates, but, in view of the tendency toward expansion observed in the diseases, the first alternative is more likely. Expansion during evolution has proceeded differently in the different genes and at different sites within the same genes. The mouse homolog of the Huntington disease gene possesses seven CAGs (25), a number very similar to that in nonhuman primates (Table 2). This indicates that the large number of CAG repeats found in normal human alleles (up to 34) is due primarily to expansion in the human lineage alone, after it diverged from all other hominoid lineages. In the androgen receptor, expansion of the number of CAG repeats began earlier in the hominoid lineage. Although the mouse and rat genes contain a CAG reiterant elsewhere in the gene they contain no CAG reiterant at this site (22–24), whereas the great apes contain up to 17 CAGs and the normal human up to 26 (Table 2). Here, expansion likely began prior to the ape–human divergence, but must have continued to occur in the human after its divergence from the apes (22–24). Another example of expansion of a CAG repeat during evolution is provided by the TATA-binding protein. The size of the reiterant increased from *Xenopus* to chicken and to the human (30).

Poly(CAG) could expand rapidly because there are signals in its vicinity that target it for expansion. However, as four sites in four unrelated genes at widely different locations have expanded in the higher primates (Table 2), one may postulate that expansion is affected by the products of other genes. The fact that the endogenous androgen receptor gene of the mouse does not contain reiterated CAG at the disease-related site (23) and that the segment of codon repeats of the human gene is stable in transgenic mice (31) seems to indicate that any genes promoting the reiteration process at this site in the human have little activity in the mouse. As microsatellites tend to be longer in humans than in other primates (32) and the size of simple sequence repeats also increases during evolution (33), the rates of expansion of microsatellites, of simple sequences, and of the reiteration sites associated with genetic diseases might be controlled by similar genes.

What is the significance of the high mutation rate associated with the process of trinucleotide expansion? In the past several years, we have investigated the evolution of involucrin, a protein encoded by a gene containing an unusually large number of CAG codons and codons similar to CAG. Because of this unusual codon composition and because the entire coding region of mammalian involucrin genes is encoded within a single exon, it was postulated that the first step in the generation of the coding region was the formation of poly-(CAG) (34, 35). The poly(CAG) was thereafter modified by nucleotide substitutions. Addition of CAGs to the involucrin

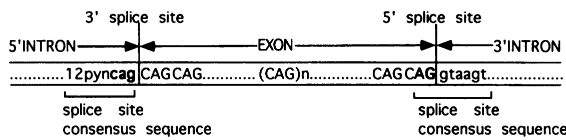

FIG. 4. Poly(CAG) contains part of the consensus sequence for 3′ and 5′ intronic splice sites.

gene has not stopped in all lineages since there is a stretch of recently added CAGs in the mouse gene (36). CAG codon expansion/nucleotide substitution must be the result of a mechanism used by many genes for the generation of new coding region, but this mechanism has been detrimental in the genes associated with the neurological diseases.

It has been noted that poly(CAG) is excluded from introns, possibly because of its similarity to splice signals (37). The 5′-most CAG provides the most conserved part of an intronic 3′ splice site, while the 3′-most CAG provides the AG dinucleotide of the 5′ splice site (Fig. 4). In the Machado–Joseph disease gene, the 3′ splice site is indeed part of the poly(CAG) (Fig. 2). No other reiterated codon could provide splice site sequences as close to the consensus. The possibility is suggested that a stretch of poly(CAG) in a noncoding region could lead to the generation of an exon. Because the most common reiterant found in coding regions of all genes is poly(CAG), even out of frame (37–39), it has been postulated that DNA strand-specific signals predispose to CAG reiteration (38).

In the gene associated with spinocerebellar ataxia type 1, the presence of interspersed CAT codons appears to stabilize long poly(CAG) stretches (40). The same is likely to be true for the clustered mutations that flank the homogeneous poly(CAG) stretches in the genes for the other three neurological diseases (Figs. 1–3). The high frequency of point mutations at and around the region of reiteration in these genes raises the possibility that this region is subject to a specific mechanism that induces or retains mutations, and, as a result, the regions are stabilized with respect to further expansion. In the poly-(CAG) regions of the Huntington disease, Machado–Joseph disease, and spinal and bulbar muscular atrophy, the mutations have occurred only at the ends of the poly(CAG). It may be that these ends are the oldest part of the reiteration or that the ends have been specifically targeted for mutation.

Unsubstituted stretches of polyglutamine are almost certainly toxic to neurons as a dominant gain of function; the toxicity has been attributed either to the postulated ability of proteins containing polyglutamine to act as substrates of transglutaminase, thereby forming insoluble cross-linked products with other proteins (41), or to the insolubility of long glutamine reiterants resulting from their ability to form multimers (42, 43). Silent nucleotide substitutions at reiteration sites might prevent further expansion of the poly(CAG) by mechanisms such as unequal recombination but would not interrupt the stretch of polyglutamine in the protein. Missense mutations, such as are present in the reiterants of the androgen receptor and of Machado–Joseph disease, would be expected to counteract the toxic effect of reiteration by reducing the size of the expanded polyglutamine in the protein.

We thank Mrs. Régine Frangne for meticulous technical assistance and Mauro Santibáñez Koref for assistance with sizing of SCA1 PCR products. This investigation was aided by grants from the Centre National de la Recherche Scientifique, the Association Française contre les Myopathies, and the Medical Research Council (United Kingdom).

1. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. (1991) *Nature (London)* **352,** 77–79.
2. Huntington Disease Collaborative Research Group (1993) *Cell* **72,** 971–983.
3. Orr, H. T., Chung, M.-y., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P. W. & Zoghbi, H. Y. (1993) *Nat. Genet.* **4,** 221–226.
4. Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. & Tsuji, S. (1994) *Nat. Genet.* **6,** 9–13.
5. Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S. & Kakizuka, A. (1994) *Nat. Genet.* **8,** 221–227.
6. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
7. Djian, P. & Green, H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 8447–8451.
8. Teumer, J. & Green, H. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 1283–1286.
9. Djian, P. & Green, H. (1990) *Mol. Biol. Evol.* **7,** 220–227.
10. Djian, P. & Green, H. (1992) *Mol. Biol. Evol.* **9,** 417–432.
11. Goldberg, Y. P., Andrew, S. E., Clarke, L. A. & Hayden, M. R. (1993) *Hum. Mol. Genet.* **2,** 635–636.
12. Tilley, W. D., Marcelli, M., Wilson, J. D. & McPhaul, M. J. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 327–331.
13. Kovalic, D., Kwak, J.-H. & Weisblum, B. (1991) *Nucleic Acids Res.* **19,** 4560.
14. Andrew, S. E., Goldberg, Y. P., Thellmann, J., Zeisler, J. & Hayden, M. R. (1994) *Hum. Mol. Genet.* **3,** 65–67.
15. Barron, L. H., Rae, A., Holloway, S., Brock, D. J. H. & Warner, J. P. (1994) *Hum. Mol. Genet.* **3,** 173–175.
16. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, L., *et al.* (1993) *Nat. Genet.* **4,** 387–392.
17. Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Ramesar, R. S., Old, J., Bontrop, R., McMahon, R., Barton, D. E. & Ferguson-Smith, M. A. (1994) *Nat. Genet.* **7,** 525–530.
18. Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. (1992) *Genomics* **12,** 241–253.
19. Matilla, T., Volpini, V., Genis, D., Rosell, J., Corral, J., Davalos, A., Mollins, A. & Estivill, X. (1993) *Hum. Mol. Genet.* **2,** 2123–2128.
20. Ranum, L. P., Chung, M.-y., Banfi, S., Bryer, A., Schut, L. J., Ramesar, R., Duvick, L. A., McCall, A., Subramony, S. H., Goldfarb, L., Gomez, C., Sandkuijl, L. A., Orr, H. T. & Zoghbi, H. Y. (1994) *Am. J. Hum. Genet.* **55,** 244–252.
21. Zühlke, C., Riess, O., Bockel, B., Lange, H. & Thies, U. (1993) *Hum. Mol. Genet.* **2,** 2063–2067.
22. Chang, C., Kokontis, J. & Liao, S. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 7211–7215.
23. He, W. W., Fischer, L. M., Sun, S., Bilhartz, D. L., Zhu, X., Young, C. Y.-F., Kelley, D. B. & Tindall, D. J. (1990) *Biochem. Biophys. Res. Commun.* **171,** 697–704.
24. Faber, P. W., King, A., van Rooij, H. C. J., Brinkmann, A. O., de Both, N. J. & Trapman, J. (1991) *Biochem. J.* **278,** 269–278.
25. Lin, B., Nasir, J., MacDonald, H., Hutchinson, G., Graham, R. K., Rommens, J. M. & Hayden, M. R. (1994) *Hum. Mol. Genet.* **3,** 85–92.
26. Pinsky, L., Trifiro, M., Kaufman, M., Beitel, L. K., Mhatre, A., Kazemi-Esfarjani, P., Sabbaghian, N., Lumbrusco, R., Alvarado, C., Vasiliou, M. & Gottlieb, B. (1992) *Clin. Invest. Med.* **15,** 456–472.
27. Wilson, J. D. (1992) *Biol. Reprod.* **46,** 168–173.
28. Zoppi, S., Wilson, C. M., Harbison, M. D., Griffin, J. E., Wilson, J. D., McPhaul, M. J. & Marcelli, M. (1993) *J. Clin. Invest.* **91,** 1105–1112.
29. Wilson, C. M. & McPhaul, M. J. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1234–1238.
30. Hancock, J. M. (1993) *Nucleic Acids Res.* **21,** 2823–2830.
31. Bingham, P. M., Scott, M. O., Wang, S., McPhaul, M. J., Wilson, E. M., Garbern, J. Y., Merry, D. E. & Fischbeck, K. H. (1995) *Nat. Genet.* **9,** 191–196.
32. Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Jain, L., Li, S.-H., Margolis, R. L., Ross, C. & Ferguson-Smith, M. A. (1995) *Nat. Genet.* **10,** 337–343.
33. Hancock, J. (1995) *J. Mol. Evol.*, in press.
34. Eckert, R. L. & Green, H. (1986) *Cell* **46,** 583–589.
35. Green, H. & Djian, P. (1992) *Mol. Biol. Evol.* **9,** 977–1017.
36. Djian, P., Phillips, M., Easley, K., Huang, E., Simon, M., Rice, R. H. & Green, H. (1993) *Mol. Biol. Evol.* **10,** 1136–1149.
37. Stallings, R. L. (1994) *Genomics* **21,** 116–121.
38. Green, H. & Wang, N. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 4298–4302.
39. Hancock, J. M. (1995) in *Human Genome Evolution*, eds. Jackson, M. S., Dover, G. & Strachan, T. (Bios, Oxford), in press.
40. Chong, S. S., McCall, A. E., Cota, J., Subramony, S. H., Orr, H. T., Hughes, M. R. & Zoghbi, H. Y. (1995) *Nat. Genet.* **10,** 344–350.
41. Green, H. (1993) *Cell* **74,** 955–956.
42. Perutz, M., Johnson, T., Suzuki, M. & Finch, J. T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5355–5358.
43. Stott, K., Blackburn, J. M., Butler, P. J. G. & Perutz, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6509–6513.
44. Burright, E. N., Clark, H. B., Servadio, A., Matilla, T., Feddersen, R. M., Yunis, W. S., Duvick, L. A., Zoghbi, H. Y. & Orr, H. T. (1995) *Cell* **82,** 937–948.