Hum Genet (1997) 100 : 114–122

ORIGINAL INVESTIGATION

Russell L. Margolis · Meena R. Abraham
Shawn B. Gatchell · Shi-Hua Li · Arif S. Kidwai
Theresa S. Breschel · O. Colin Stine · Colleen Callahan
Melvin G. McInnis · Christopher A. Ross

cDNAs with long CAG trinucleotide repeats from human brain

Received: 12 January 1997 / Accepted: 5 March 1997

Abstract Twelve diseases, most with neuropsychiatric features, arise from trinucleotide repeat expansion mutations. Expansion mutations may also cause a number of other disorders, including several additional forms of spinocerebellar ataxia, bipolar affective disorder, schizophrenia, and autism. To obtain candiate genes for these disorders, cDNA libraries from adult and fetal human brain were screened at high stringency for clones containing CAG repeats. Nineteen cDNAs were isolated and mapped to chromosomes 1, 2, 4, 6, 7, 8, 9, 12, 16, 19, 20, and X. The clones contain between 4 and 17 consecutive CAG, CTG, TCG, or GCA triplets. Clone H44 encodes 40 consecutive glutamines, more than any other entry in the nonredundant GenBank protein database and well within the range that causes neuronal degeneration in several of the glutamine expansion diseases. Eight cDNAs encode 15 or more consecutive glutamine residues, suggesting that the gene products may function as transcription factors, with a potential role in the regulation of neurodevelopment or neuroplasticity. In particular, the conceptual translation of clone CTG3a contains 18 consecutive glutamines and is 45% identical to the C-terminal 306 residues of the mouse numb gene product. These genes are therefore candidates for diseases featuring anticipation, neurodegeneration, or abnormalities of neurodevelopment.

R. L. Margolis · M. R. Abraham · S. B. Gatchell · S.-H. Li
A. S. Kidwai · T. S. Breschel · O. C. Stine · C. Callahan
M. G. McInnis · C. A. Ross
Laboratory of Molecular Neurobiology,
Johns Hopkins University School of Medicine,
600 N. Wolfe Street, Baltimore, MD 21287, USA

R. L. Margolis (✉) · C. A. Ross
Department of Psychiatry,
Johns Hopkins University School of Medicine, Meyer 2-181,
600 N. Wolfe Street, Baltimore, MD 21287, USA
Tel.: +1-410-614-4262; Fax: +1-410-955-8233;
e-mail: rmargoli@welchlink.welch.jhu.edu

Introduction

Trinucleotide repeat expansion mutation is now know to cause 12 diseases, most with neuropsychiatric features (Linblad and Schalling 1996; Paulson and Fischbeck 1996; Ross 1995; Zoghbi 1996). Seven of these are known as the type 1 disorders – spinocerebellar ataxia type 1 (SCA1, Orr et al. 1993), SCA2 (Imbert et al. 1996; Pulst et al. 1996; Sanpei et al. 1996), Machado-Joseph disease (MJD or SCA3, Kawaguchi et al. 1994), SCA6 (Zhuchenko et al. 1997), dentatorubral pallidoluysian atrophy (DRPLA, Koide et al. 1994; Nagafuchi et al. 1994), Huntington's disease (HD, Huntington's Disease Collaborative Research Group 1993), and spinal and bulbar muscular atrophy (SBMA, La Spada et al. 1991). Each is caused by a $(CAG)_n$ expansion in an open reading frame, resulting in an expanded glutamine repeat. The properties of the repeats in the other (type 2) expansion mutation diseases vary widely. Myotonic dystrophy is caused by a 3′ untranslated $(CTG)_n$ expansion (Brook et al. 1992; Fu et al. 1992; Mahadevan et al. 1992), the A and E forms of fragile X syndrome (Fu et al. 1991; Knight et al. 1993; Kremer et al. 1991; Verkerk et al. 1991) and some cases of Jacobsen's syndrome (Jones et al. 1995) result from 5′ untranslated region $(CCG)_n$ expansions, and Friedreich's ataxia is caused by an intronic $(GAA)_n$ expansion (Campuzano et al. 1996). Expandable trinucleotide repeats therefore are found in translated, transcribed but untranslated, and intronic regions; they may be G-C or A-T rich and range from minimal to highly variable in length in the normal population.

At least four lines of evidence indicate that additional disorders may arise from trinucleotide repeat expansion mutations. First, an antibody (IC2) that specifically recognizes expanded glutamine repeats detects an expansion segregating with SCA7 (Trottier et al. 1995). Second, indirect evidence of CAG expansion has been detected using rapid expansion detection (RED, Schalling et al. 1993) in a pedigree with SCA7, and less clearly in heterogeneous populations of patients with bipolar affective

disorder and schizophrenia (Linblad et al. 1996; Linblad and Schalling 1996; O'Donovan et al. 1995). Third, several neurodegenerative disorders, including SCA4, SCA5, SCA7, and familial Parkinson disease, are phenotypically similar to the type I expansion mutation disorders. Fourth, anticipation, the phenomenon of increasing phenotypic severity or decreasing age of onset in successive generations affected by a disease (McInnis 1996; Ross et al. 1993), is found in most of the expansion mutation diseases. Anticipation has been detected in a disparate group of other diseases, including affective disorder (Engstrom et al. 1995; McInnis et al. 1993; Nylander et al. 1994), schizophrenia (Chotai et al. 1995; Gorwood et al. 1996; Stober et al. 1995; Thibaut et al. 1995), autism (Stine 1993), familial Parkinsonism (Bonifati et al. 1995; Markopoulou et al. 1995; Payami et al. 1995; Plante-Bordeneuve et al. 1995), familial leukemias (Horwitz et al. 1996), Crohn's disease (Polito et al. 1996), Meniere's disease (Morrison 1995), torsion dystonia (LaBuda et al. 1993), rheumatoid arthritis (McDermott et al. 1996), facioscapulohumeral muscular dystrophy (Tawil et al. 1996), Holt-Oram syndrome (Newbury-Ecob et al. 1996), and familial spastic paraplegia (Raskind et al. 1997).

We have sought to identify candidate genes for these disorders by screening cDNA libraries for the presence of DNA fragments containing CAG, CCG, CCA, and AAT trinucleotide repeats (Li et al. 1993; Margolis et al. 1995 a, b). Our description of CTG-B37, a cDNA fragment with a highly polymorphic CAG repeat located within an open reading frame on chromosome 12, directly led to the finding that an expansion mutation within the CTG-B37 repeat causes DRPLA (Koide et al. 1994; Nagafuchi et al. 1994). This same strategy of screening cDNA libraries for trinucleotide repeats was later employed to identify the MJD gene (Kawaguchi et al. 1994) and the SCA6 gene (Zhuchenko et al. 1997). Screening genomic contigs for trinucleotide repeats was used to clone the gene for SCA2 (Pulst et al. 1996). Based on the repeats that expand to cause disease, repeats with the highest likelihood of undergoing expansion mutation consist of at least six consecutive CAG or CTG triplets in the transcribed portions of genes expressed in brain. To identify genes with these features, we have screened human adult frontal cortex and fetal brain cDNA libraries at high stringency for the presence of CAG or CTG repeats. We now report the identification and mapping of 19 of these cDNA fragments.

Materials and methods

cDNA cloning

Adult human frontal cortex (clones with CTG prefixes) and human fetal brain (all others) cDNA libraries (Strategene) was screened with either a $(CTG)_{20}$ or a $(CTG)_{15}$ radiolabeled oligonucleotide using standard techniques as previously described (Margolis et al. 1995 a). In brief, approximately 200 000–500 000 plaques were screened in each of four sets of primary screens. Hybridization was at 52° (CTG and F clones) or 60° (H and L clones) with 50% formamide with a 70° (CTG and F clones) or 72° (H and L clones) wash at 0.2 × SSC/0.5% SDS for 15 min or until activity was below about 5 counts/s. After plaque purification, plasmids were excised using an in vivo excision procedure and purified.

Inserts were released by *Eco*RI digest, and Southern blots were prepared after electrophoresis through 1% agarose. Southern blots were hybridized as described for library screening, and positive inserts were sequenced using ABI automated technology until the repeat was detected or until a blastn search of the GenBank nonredundant, dbest, or dbsts libraries indicated that the clone had been previously described. Clones H32, 44, and 45 represent consensus double-stranded sequences of multiple clones. The contigs were formed in part by other clones identified in the library screen described above and in part by rescreening the fetal brain library with 45mer oligonucleotides antisense to sequence immediately adjacent to the repeat of each clone. H45: 5′ CCTGGGCACTCATTGGAGTCATGGTACTTATCATGGGTGTCTGCT 3′; H44: 5′ AGGTGGAATGGAGATGAGTCCCTGACGCTGAAGGCTGAGCAGATG 3′ and GGCCGCCCGTTCACCTTGATGAGACCATTGCCGCGTTTGCAGTGC; H32: 5′ AGGTGTGATGGTTTTTCCAGGGAGCTGAACTCCTGTGGCTTGGGA 3′.

Mapping

Clones were assigned to a chromosome or a specific locus using a variety of methods as indicated in Table 1. The NIGMS panel 2 monochromosomal human-rodent hybrid cell lines were screened using PCR primers that spanned the repeat as indicated in Table 2 (Dubois and Naylor 1993). Radiation hybrid mapping was performed by Research Genetics on either the GeneBridge 4 or Stanford G3 Radiation Hybrid panels using primers listed in Table 2. For L69 and F28, primers amplifying a region flanking the repeat were used to avoid crosshybridization with hamster genomic DNA (L69: 5′ CTCCGAGTATCGTGACAGGTG 3′, 5′ AAGCAGAGGTGAGGGAGAGTC 3′, 59° annealing temperature, 157-bp product; CAGF28: 5′ GCTGAAGGTGTAAAACCG 3′, 5′ GAAAGTGACTCCAGAGAC 3′, 52° annealing temperature, 86-bp product). CTG3a was mapped by analysis of repeat polymorphism in CEPH pedigrees 1332, 1347, 1362, and 884 (Weissenbach et al. 1992) with the LINKAGE program (Ott 1991). As indicated, other clones were assigned to a locus either through a GenBank match with a previously mapped ESTs (expressed sequence tags) or through use of Gene Map (Schuler et al. 1996).

Polymorphism analysis

Variability of repeat length was determined by size separation of radiolabeled PCR products on a 6% denaturing polyacrylamide gel, as previously described (Margolis et al. 1995 a). Template consisted of 40–100 ng human genomic DNA from either a set of unrelated CEPH parents or a panel of unrelated individuals with movement disorders of unknown etioloy, and reaction conditions were as described in Table 2. Allele lengths of additional individuals with movement disorders were tested for size variation in L69 using automated detection (ABI) of fluorescence-labeled PCR product.

Results

Table 1 summarizes the initial characterization of the cDNA clones identified by screening adult human frontal cortex or human fetal brain cDNA libraries with a $(CAG)_{15}$ or $(CAG)_{20}$ probe. Those clones for which data on both polymorphism and chromosomal location are available are included, and clones corresponding to known genes were excluded except where noted, leaving

Table 1 Repeat sequence, mapping, and homologies

Clone	Genbank accession	Longest repeat	Conceptual translation of repeat region	Mapping method[a]	Chromosomal location	Homology
CTG3a	U80758	6	Q_{18}	E	19q13.13–19q13.2 D19S224–D19S228	mouse numb (U70764)
CTG4a	U80744	8	L_4PL_9	A	1	
CTG7a	U80745	8	$Q_5LQ_8HLIKLHHQNQ_6LQRIAQLQLQ_{16}$	A	20	
CTG20a	U80749	13	$PQ_{23}P_3$ (or $A_{13}TA_{11}$)	A	6	
F28	U80735	7	Q_8, 54% Q in 68 residue region	B	7q36.3 qter of D7S2781	mcag32 (U23862)
H1	U80738	13	$(AQ)_{11}ASQASQ_{15}$ or L_{15}	D	12p12 D12S328–D12S89	
H3	U80747	5′= 8 3′= 8	5′: Q_{14} 3′: Q_{18}	C	4	2.119 (5′ repeat only)
H16	U80737	12	Q_{28}	D	20q13.13 D20S891–D20S109	
H26	U80750	4	$Q_2PQ_8PQ_4PQ_3PQPQPQ_4PQ_3PQ$	A	16	mcag19 (U23860)
H32	U80743	15	Q_{30}	B	12q24.33 qter of D12S367	
H38	L13744	10	S_{42}	C	9q22	AF-9 (L13744)
H39	U80740	17	none	B	Xp11.4 AFMA212ZG1–DXS6745	
H44	U80741	7	$Q_{40}HPGKQAKEQ_{10}$	B	6q14–15 D6S1291	
H45	U80742	7	$Q_{25}YHIQ_6ILRQ_{25}HQ_6GP(XQ)_5$	C	1	KIAA0192 (D83783)
L69	U80752	13	none	B	2q37.3 D2S2704	
L85	U80755	10	S_{10} or A_{15} or Q_{11}	A	8	
L114	U80765	6	Q_{11}	B	12q13.12 WI-7107-DS12S1363	
L234	U80751	9	none	B	2q21.2 AFMA153WBI	
L237	U80757	15	none	B	16q24.3 D16S520-WI-12410	

[a] Mapping methods: A, NIGMS panel 2; B, radiation hybrid panel; C, identity to mapped cDNA; D, Gene Map; E, linkage

a total of 20 repeat regions from 19 indepdendent cDNA clones. The number of consecutive triplets ranges from 4 to 17 (mean = 9.8). Of the 20 repeats, 16 appear to be located in extended open reading frames. Of these 16, conceptual translation indicates that 12 (and possibly 14) encode polyglutamine. In each case the CAG repeat is interrupted frequently by CAA triplets also encoding glutamine. Eight of these polyglutamine regions consist of 15 or more consecutive glutamines, including Q_{28} (H16), Q_{30} (H32), and two adjacent stretches of Q_{25} (H45). The longest consecutive glutamine repeat, Q_{40}, is present in H44 (Fig. 1).

The cDNAs correspond to loci on 12 different chromosomes, as determined by the methods indicated in Table 1: one each from chromosomes 4, 7, 8, 9, 19 and X; two each from chromosomes 1, 2, 6, 16, and 20; and three from chromosomes 12. Blastn and blastp homology searches (Altschul et al. 1990) of GenBank revealed that the repeats in F28 and H26 are present in short cDNA fragments termed mcag32 and mcag19 (Yamagata et al. 1996). H45 is identical to the 3′ portion of a longer cDNA fragment (D83783) isolated from the KG-1 human cell line. D83783 is ubiquitously expressed, and contains both a DNA topoisomerase II motif and an N terminal transmembrane domain. A short murine amino acid sequence termed the mopa box protein (A26892) is also similar to H45. The first repeat in H3 is contained in clone 2.119 (Neri et al. 1996). H38 is a portion of the human gene AF-9, identified in relationship to a t(9:11) translocation in which AF-9 fuses with ALL-1 to induce acute lymphocytic leukemia (Prasad et al. 1993). AF-9 may function in the transcriptional complex (Cairns et al. 1996; Welch and Drubin 1994). CTG3a is similar to the mouse numb gene (Zhong et al. 1996), with 44.6% amino acid identity and 85.3% amino acid conservation (ALIGN, Pearson and Lipman 1988) over the length of CTG3a, which extends from m-numb amino acid 267 to 585 (Fig. 2). A comparison of CTG3a with both *Drosophila* and mouse numb by Block Maker (Henikoff et al. 1995) suggested five possible blocks, corresponding to CTG3a amino acid residues 26–52, 57–68, 85–104, 123–133, and 256–264. The Gibbs alignment (Lawrence et al. 1993) suggested three blocks, corresponding to CTG3a amino acid residues 29–60, 188–226, and 256–278. A search of these alignments against the Blocks database using LAMA (Pietrokovsi 1996) revealed a match only to one block

Table 2 Polymorphism analysis

Clone	Longest nucleotide repeat	PCR primers	Anneal temperature	Product length (in bp)	No. of chromosomes tested	Allele bp: frequency	Heterozygosity
CTG3a	6	GGGCACTGGGGCCACTGAGG CCTGGGCACAAGCGGACACC	64°	147	32	141:0.03 147:0.66 150:0.31	62%
CTG4a	8	CGTCCCGCTGTCTTCTGCTTC AGGCGAACCCAGTCGTTCTCC	62°	112	28	100:0.25 109:0.68 112:0.07	21%
CTG7a	8	TGCAGCTCCAACAACAGCAAC CTGCTGCATCGGTGGCTGCTG	62°	101	40	86:0.02 101:0.80 104:0.18	20%
CTG20a	13	CACCATGTCGCTGAAGCCCC CGCCGGGCTTGCGGACATTG	61° (7-deaza-dGTP)	125	38	125:0.97 128:0.03	15%
F28	7	GAAAGTGACTCCAGAGAC GTAAGGTTGCTGAGAGAT	49°	145	32	145:1.00	0%
H1	13	GGAGACTGGAAGTGTGGT CTCCCAGGCATCACAGCA	47° (DMSO)	82	30	82:0.83 85:0.17	20%
H3	5′= 8	GCACAGCAGCAACAAAGG GTCCTAAGGGAGACCAAT	47° (DMSO)	118	32	118:0.84 124:0.03 127:0.06 130:0.03 133:0.03	31%
	3′= 8	ACAAAAACCCCTTGATGC CCTGGAGCTGTGGAGGTG	53°	102	32	102:1.00	0%
H16	12	GGGTGGCTATGATGATGC TGAAGACCTGGGGTTGCT	54°	120	30	120:0.53 123:0.43 129:0.03	60%
H26	4	ATCCTAATAACGGCACTTCC TTCCTGGTGGCGAAATCGTG	54°	224	32	215:0.19 224:0.81	25%
H32	15	ACATTTCCAGCTTCTCAGGCA TCTGTTGCTGCTGCTGCTGTT	57°	113	32	110:0.53 113:0.41 119:0.03 128:0.03	38%
H38	10	TTAATTTGTGAGGCTTTGAAAAAC CTAATAGGAGTATTCATACCAGCA	54°	174	32	174:0.88 177:0.06 180:0.06	25%
H39	17	CTCTGATTGCTTAGTGGACAA TTCAGTACATTGCTGCTGCTG	62°	101	26	89:0.23 98:0.69 101:0.08	31%
H44	7	CAGCAAGAGCAGTTACAT GCTTGCTTTCCAGGATGC	52°	167 (bp 424–590)	32	167:1.00	0%
		TGCTGCTGGAAGACAAGC GAAAGCAAGCGAAAGAGC	52° (DMSO)	79 (bp 581–659)	32	79:1.00	0%
H45	7	TGCTGCTGCCGGATGTGGTA GCAGGCGTCCTGTCAACA	59°	131 (bp 2388–2518)	28	131:1.00	0%
L69	13	AGGTGAGGGAGAGTCCAT TATGGGCTCCTGTTCCTG	55°	104	88	83:0.02 92:0.09 104:0.20 107:0.02 110:0.45 113:0.01 116:0.03 122:0.15 126:0.01	75%
L85	10	GCTCTCGGGCATTGATGTTA AGCAGCAGCGGTAGCAGTGG	58°	136	26	133:0.78 136:0.15 139:0.08	46%
L114	6	GGGACCCAGAACTTGGTTTG CGAAGCCTCGGACCTGATTC	57°	195	28	195:1.00	0%

Table 2 (continued)

Clone	Longest nucleotide repeat	PCR primers	Anneal temperature	Product length (in bp)	No. of chromosomes tested	Allele bp: frequency	Heterozygosity
L234	9	ACAAGAGGAAAAGGGACATAGC GCAAAACCAAATACCAGGTCTC	51°	94	32	124:0.88 127:0.09 130:0.03	14%
L237	15	GGTTCCCTGCACAGAAACCATC AGATGCCACCGCATTCGG	59°	120	26	96:0.04 99:0.04 108:0.04 117:0.58 120:0.08 123:0.23	33%

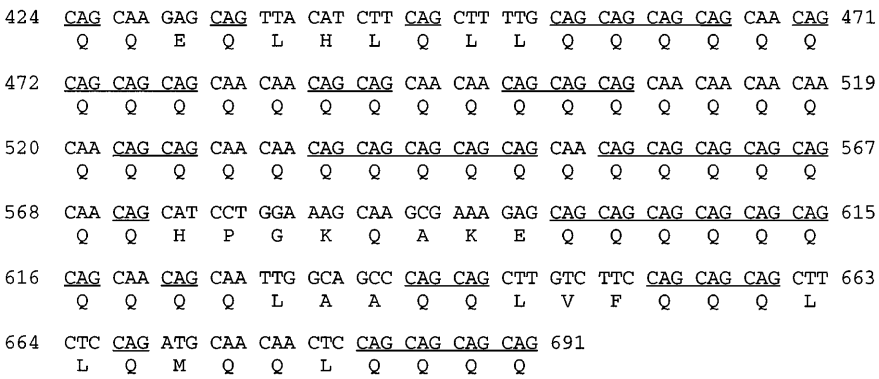

Fig. 1 Sequence and conceptual translation of the glutamine-rich region of H44. Forty consecutive glutamines are encoded by a CAG repeat interrupted in seven places by $(CAA)_{1–4}$

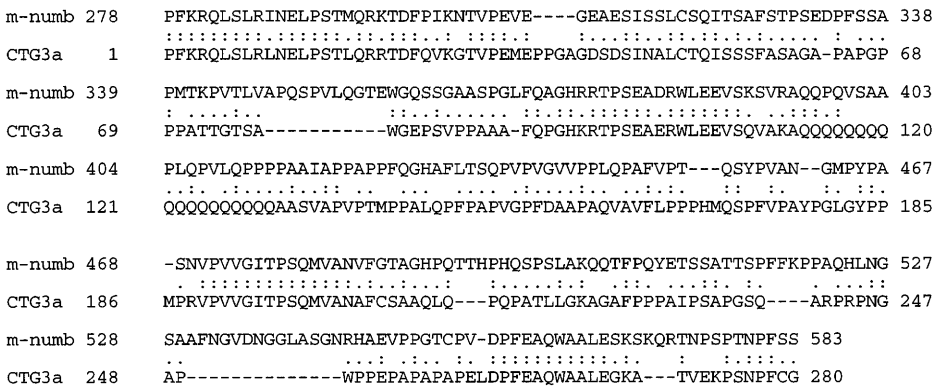

Fig. 2 Homology between the mouse numb gene product (*m-numb*) and the conceptual translation of *CTG3a*. There is 44.6% identity and 85.3% conservation between the sequences

formed from the synapsin family. This alignment is based on the abundance of glutamine and proline residues in the query block, and is of questionable biological significance. CTG3a and m-numb share a common PKC phosphorylation site (SLR at CTG3a amino acid residues 7–9) and three potential CamKinase II phosphorylation sites (CTG3a residues 28–31, 95–98, and 97–100). No other motif was detected and no other cDNA had significant homologies to other known genes at the DNA or protein level.

Table 2 summarizes the polymorphism analysis of this set of loci, as determined by radiolabeled PCR of CEPH individuals or individuals with movement disorders of unknown etiology. Heterozygosity ranged from 0% to 75%, with a mean of 25%. The 75% heterozygosity of L69 corresponded to the presence of 9 alleles, with the repeat presumably ranging in length from 6 to 20 triplets. Three other loci, CTG3a, H16, and L85 had fewer alleles but heterozygosity over 45%. Testing more individuals would probably have resulted in the detection of additional alleles. Several of the clones encoding long stretches of polyglutamine are not polymorphic; in each case the glutamine repeat is encoded by a combination of CAG and CAA codons.

To test the correlation of repeat length with the extent of heterozygosity, the modal repeat length was determined for each clone, and also for clones containing CAG repeats previously identified in our laboratory and for genes with CAG or CTG repeats in which expansion mutation is known to occur (Fig. 3). The repeat in ataxin-2, which expands to cause SCA2, is not highly polymorphic, and its ratio of heterozygosity to repeat length is similar to a number of the clones listed in Tables 1 and 2. All other CAG/CTG repeats associated with a disease are highly polymorphic with fairly long repeat lengths; the heterozygosity of L69, MAB21L1 (a gene previously identified in

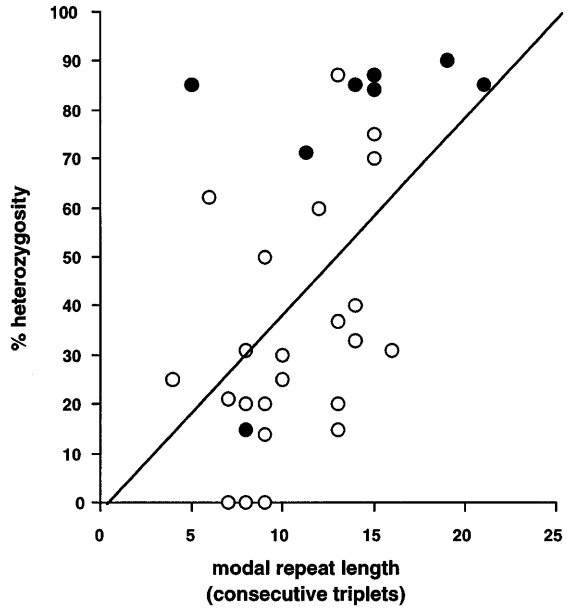

Fig. 3 The relationship between heterozygosity and modal repeat length. The modal repeat length and heterozygosity at the loci for CAG-containing clones identified in our laboratory (unfilled circles) were combined with similar information for the loci associated with CAG or CTG expansion mutations (filled circles). There is a marginal positive correlation between the two variables ($r^2 = 0.28$). Note that the repeat causing SCA2 is similar in length (8 triplets) and heterozygosity (15%) to many of the repeats listed in Tables 1 and 2

our laboratory and originally termed CAGR1, Margolis et al. 1996), and CTG-B33 (Li et al. 1993) are in a similar range.

Discussion

We now report the identification of 19 cDNA clones containing 20 CAG repeats. The repeats contain a mean of 10 consecutive triplets, with a maximum of 17. Most of the repeats are in open reading frames, contain the triplet CAA interspersed with longer runs of consecutive CAG triplets, and encode stretches of polyglutamine. The clones correspond to loci mapped to 12 different chromosomes and have a heterozygosity of up to 75%.

Four of the clones encode tracts of 25 to 40 consecutive glutamine residues. These are among the most glutamine-rich regions yet identified in the proteins of any species; in comparison, the ten most significant matches (blastp) in GenBank to a 40 glutamine query sequence consisted of proteins or protein fragments with 18 to 38 consecutive glutamines, only two of which were human. The longest stretches of consecutive glutamines were found in the human transcription factor TFIID (P20226, Q_{37} or Q_{38}, Imbert et al. 1994) and the yeast transcription factor SNF5 (P18480, Q_{37}), neither of which is homologous to any of the cDNA clones described here. The next longest glutamine repeats were present in the mouse mopa protein fragment (A26892, $Q_{33}YHIRQ_7MLRQ_{27}PHQ_5$) homologous to H45 and the uncharacterized yeast protein YM8520.13c (S54522, $Q_{30}HQ_5$). The remaining six proteins, with substantially fewer consecutive glutamines, were the *Drosophila* developmental regulatory protein notch (S09358, $Q_{13}HQ_{21}$), the fungal transcription factor white collar-1 (1480115, $HQHQ_{28}HQHQ_4$), a murine AT-motif binding factor (1585921, Q_{19}), a long human protein fragment nearly identical to H45 (D83783), the *Drosophila* ecdysone-induced protein E74A (B53225, Q_{18}), and a putative African malarial mosquito nucleic acid binding protein (S27770, Q_{20}). This list indicates that few glutamine repeats longer than 25 residues have yet been identified, and that proteins with long glutamine repeats, such as the gene products of H16, H32, H44, and H45, may function as transcription factors.

Clone H44, located on 6q14–15, is of particular interest since it encodes 40 consecutive glutamine residues and is present in the same length in all tested individuals. Repeat expansions resulting in as few as 21 consecutive glutamines in the α_{1A}-voltage-dependent calcium channel causes SCA6, and 35 consectuvie glutamines in ataxin-2 causes SCA2 (Imbert et al. 1996; Pulst et al. 1996; Sanpei et al. 1996). The minimal repeat length for disease onset is 35 or 36 glutamines in HD (Barron et al. 1993) and 40 glutamines in SBMA (La Spada et al. 1991) and SCA1 (Ranum et al. 1995). Normal alleles with more than 35 or 36 residues have not been detected in DRPLA or MJD, though the minimum expansion associated with disease is 49 glutamines for DRPLA and 68 glutamines for MJD (Kawaguchi et al. 1994; Koide et al. 1994; Nagafuchi et al. 1994). These findings, coupled with our data that a putative protein exists with 40 consecutive glutamines and the presence of at least 38 consecutive glutamines in TFIID, demonstrate that the development of an expansion mutation disease is not solely determined by the length of the glutamine repeat within a given gene. Other features of protein structure or expression must be relevant. The available data suggests, however, the presence of a threshold effect: it may be that a glutamine repeat of 60 or more residues in any protein leads to cell damage, whereas shorter stretches of glutamines are only toxic in a protein with certain as yet unidentified structural features or peculiarities of spatial or temporal expression. Conversely, shorter polyglutamine repeats may be inherently neurotoxic, with unknown aspects of structure, function, or expression serving a protective role in many, but not all, proteins. Comparative anylsis of TFIID and H44 with the α_{1A}-voltage-dependent calcium channel, ataxin-2, huntingtin, and the androgen receptor in cell toxicity assays (Ikeda et al. 1996) may help elucidate these features.

The CAG repeat-containing clones listed in Tables 1 and 2 can be considered candidate genes for neuropsychiatric disorders on the basis of both structure and function. From a structural standpoint, several of the repeats have either a high heterozygosity index (e.g., L69) or a fairly large number of alleles (H3 and L237), features of most of the repeats known to undergo expansion mutation. As depicted in Fig. 3, the ratio of modal repeat length to heterozygosity in many of the other cDNA clones is quite similar to ataxin-2; like ataxin-2, many of these cDNAs

have moderate stretches of CAG repeats interspersed with occasional CAA triplets. The presence of uninterrupted CAG repeats in expanded ataxin-1 and ataxin-2 demonstrates that CAA can mutate into CAG, leading to repeat length instability (Imbert et al. 1996; Orr et al. 1993; Pulst et al. 1996; Sanpei et al. 1996). A review of OMIM indicates that a number of diseases, though none with established anticipation, map near the loci listed in Table 1. Of particular interest, two developmental disorders, triphalangeal thumb-polysyndactyly syndrome (OMIM 190605) and sacral agenesis (OMIM 176450) map to 7q36 (clone F28), noninsulin-dependent diabetes mellitus 1 (OMIM 601283) and brachydactyly type E map to 2q37 (clone L69) progressive foveal dystrophy (OMIM 136550) maps to 6q14–16.2 (clone H44), and type 1 congenital nephrosis, Finnish type (OMIM 256300), maps to the same marker on 19q13 as CTG3a.

From the standpoint of function, we and others have postulated that developmental abnormalities may underlie at least some forms of schizophrenia, autism, and bipolar affective disorder (Keshavan et al. 1994; Margolis et al. 1994; Peterson 1995; Ross and Pearlson 1996; Weinberger 1995). Both *Drosophila* and murine numb appear to specify cell fate during early stages of neurodevelopment (Spana and Roe 1996; Zhong et al. 1996). The CTG3a gene product may share this capacity, though the extent of sequence identity (45%) between CTG3a and mouse numb indicates that the CTG3a gene product is not the human numb homolog but is instead a numb family member. Many of the other putative genes identified here may also have a specific role in the process of neuronal cell-fate determination, migration, or differentiation through regulation of transcription. A mutation, expansion or otherwise, would serve to alter this function. In particular, glutamine repeat length has been shown to modulate the efficieny of transcription (Chamberlain et al. 1994; Gerber et al. 1994). It is tempting to speculate that, though disease is induced by marked glutamine expansions, normal variations in glutamine repeat length may underlie normal variations in phenotype, modify the consequences of mutations in other genes, or influence the response to a nongenetic insult.

Acknowledgements The authors would like to thank Duane Bartley, Roxann Ashworth, and Dr. Alan Scott from the Johns Hopkins Genetics Core facility for assistance with sequencing, and John Kleiderlein, Kirsten Bottoms, and Hae Yoon for technical assistance. This work was supported in part by the National Alliance for Research on Schizophrenia and Depression, NIH MH02175-10A1, NIH NS34172, and NIH MH50763.

References

Altschul AF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barron LH, Warner JP, Porteous M, Holloway S, Simpson S, Davidson R, Brock DJ (1993) A study of the Huntington's disease associated trinucleotide repeat in the Scottish population. J Med Genet 30:1003–1007

Bonifati V, Fabrizio E, Vanacore N, De Mari M, Meco G (1995) Familial Parkinson's disease: a clinical genetic analysis. Can J Neurol Sci 22:272–279

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Davies J, Shelbourne P, Buxton J, Jones C, Juvonen V, Johnson K, Harper PS, Shaw DJ, Housman DE (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell 68:799–808

Cairns BR, Henry NL, Kornberg RD (1996) TFG/TAF30/ANC1, a component of the yeast SWI/SNF complex that is similar to the leukemogenic proteins ENL and AF-9. Mol Cell Biol 16:3308–3316

Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Canizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, Demichele G, Filla A, Defrutos R, Palau F, Patel PI, Didonato S, Mandel JL, Cocozza S, Koenig M, Pandolfo M (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271:1423–1427

Chamberlain NL, Driver ED, Miesfelf RL (1994) The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Res 22:3181–3186

Chotai J, Engstrom C, Ekholm B, son Berg ML, Adolfsson R, Nulander PO (1995) Anticipation in Swedish families with schizophrenia. Psychiatr Genet 5:181–186

Dubois BL, Naylor SL (1993) Characterization of NIGMS human/rodent somatic cell hybrid mapping panel 2 by PCR. Genomics 16:315–319

Engstrom C, Thornlund AS, Johansson EL, Langstrom M, Chotai J, Adolfsson R, Nylander PO (1995) Anticipation in unipolar affective disorder. J Affect Disord 35:31–40

Fu Y-H, Kuhl DPA, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkerk AJMH, Holden JJA, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047–1058

Fu Y-H, Pizzuti A, Fenwick RG, King J, Rajnarayan S, Dunne PW, Dubel J, Nasser GA, Ashizawa T, De Jong P, Wieringa B, Korneluk R, Perryman MB, Epstein HF, Caskey CT (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science 255:1256–1258

Gerber H-P, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W (1994) Transciptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811

Gorwood P, Leboyer M, Falissard B, Jay M, Rouillon F, Feingold J (1996) Anticipation in schizophrenia: new light on a controversial problem. Am J Psychiatry 153:1173–1177

Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S (1995) Automated contruction and graphical presentation of protein blocks from unaligned sequences. Gene 163:17–26

Horwitz M, Goode EL, Jarvik GP (1996) Anticipation in familial leukemia. Am J Hum Genet 59:990–998

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

Ikeda H, Yamaguchi M, Sugai S, Aze Y, Narumiya S, Kakizuka A (1996) Expanded polyglutamine in the Machado-Joseph disease protein induces cell death in vitro and in vivo. Nat Genet 13:196–202

Imbert G, Trottier Y, Beckmann J, Mandel JL (1994) The gene for the TATA binding protein (TBP) that contains a highly polymorphic protein coding CAG repeat maps to 6q27. Genomics 21:667–668

Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier J-M, Weber C, Mandel J-L, Cancel G, Abbas N, Durr A, Didierjean O, Stevanin G, Agid Y, Brice A (1996) Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nat Genet 14:285–291

Jones C, Penny L, Mattina T, Yu S, Baker E, Voullaire L, Langdon WY, Sutherland GR, Richards RI, Tunnacliffe A (1995) Association of a chromosome deletion syndrome with a fragile site with the proto-oncogene CBL2. Nature 376:145–149

Kawaguchi Y, Okamoto T, Taniwaki T, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I, Kimura J, Narumiya S, Kakizuka A (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet 8:221–228

Keshavan MS, Anderson S, Pettegrew JW (1994) Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The Feinberg hypothesis revisited. J Psychiatr Res 28:239–265

Knight SJL, Flannery AV, Hirst MC, Campbell L, Christodoulou Z, Phelps SR, Pointon J, Middletonprice HR, Barnicoat A, Pembrey ME, Holland J, Oostra BA, Bobrow M, Davies KE (1993) Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell 74:127–134

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) Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nat Genet 6:9–12

Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, Richards RI (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science 252:1711–1714

LaBuda MC, Fletcher NA, Korczyn AD, Inzelberg R, Harding AE, Pauls DL (1993) Genomic imprinting and anticipation in idiopathic torsion dystonia. Neurology 43:2040–2043

La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352:77–79

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wooton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262:208–214

Li S-H, McInnis MG, Margolis RL, Antonarakis SE, Ross CA (1993) Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. Genomics 16:572–579

Linblad K, Schalling M (1996) Clinical implications of unstalbe DNA repeat sequences. Acta Paediatrica 85:265–271

Linblad K, Savontaus M-L, Stevanin G, Holmberg M, Digre K, Zander C, David G, Benomar A, Nikoskelainen E, Trottier Y, Holmgren G, Ptacek LJ, Anttinen A, Brice A, Schalling M (1996) CAG repeat expansions in SCA7. Am J Hum Genet 59:A269

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Nelville C, Narang M, Barcelo J, O'Hoy K, Leblond S, Earle-MacDonald J, DeJong PJ, Wieringa B, Korneluk RG (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. Science 255:1253–1255

Margolis RL, Chuang D-M, Post RM (1994) Programmed cell death: implications for neuropsychiatric disorders. Biol Psychiatry 35:946–956

Margolis RL, Breschel TS, Li S-H, Kidwai AS, Antonarakis SE, McInnis MG, Ross CA (1995a) Identification and characterization of cDNA clones containing CCA trinucleotide repeats derived from human brain. Somat Cell Mol Genet 21:279–284

Margolis RL, Breschel TS, Li S-H, Kidwai AS, McInnis MG, Ross CA (1995b) Polymorphic (AAT)n trinucleotide repeats derived from a human brain cDNA library. Hum Genet 495–496

Margolis RL, Stine OC, McInnis MG, Ranen NG, Rubinsztein DC, Leggo J, Jones Brando LV, Kidwai AS, Loev SJ, Breschel TS, Callahan C, Simpson SG, Depaulo JR, McMahon FJ, Jain S, Paykel ES, Walsh C, DeLisi LE, Crow TJ, Torrey EF, Ashworth RG, Macke JP, Nathans J, Ross CA (1996) cDNA cloning of a human homologue of the *Caenorhabiditis elegans* cell fate-determining gene mab-21: expression, chromosomal localization and analysis of a highly polymorphic (CAG)n trinucleotide repeat. Hum Mol Genet 5:607–616

Markopoulou K, Wszolek ZK, Pfeiffe RF (1995) A Greek-American kindred with autosomal dominant, levodopa-responsive parkinsonism and anticipation. Ann Neurol 38:373–378

McDermott E, Khan MA, Deighton C (1996) Further evidence for genetic anticipation in familial rheumatoid arthritis. Ann Rheum Dis 55:475–477

McInnis MG (1996) Anticipation: an old idea in new genes. Am J Hum Genet 59:973–979

McInnis MG, McMahon FJ, Chase GA, Simpson SG, Ross CA, DePaulo JR (1993) Anticipation in bipolar affective disorder. Am J Hum Genet 53:385–390

Morrison AW (1995) Anticipation in Meniere's disease. J Laryngol Otol 109:499–502

Nagufuchi 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, Osamu K, Takahashi M, Yoshizawa T, Kanazawa I, Yamada M (1994) Expansion of an unstable CAG trinucleotide on chromosome 12p in dentatorubral and pallidoluysian atrophy. Nat Genet 6:14–17

Neri C, Albanese V, Lebre A-S, Holbert S, Saada C, Bougueleret L, Meier-Ewert S, Le Gall I, Millasseau P, Bui H, Giudicelli C, Massart C, Guillou S, Gervy P, Poullier E, Rigault P, Weissenbach J, Lennon G, Chumakov I, Dauseet J, Lehrach H, Cohen D, Cann HM (1996) Survey of CAG/CTG repeats in human cDNAs representing new genes: candidates for inherited neurological disorders. Hum Mol Genet 5:1001–1009

Newbury-Ecob RA, Leanage R, Raeburn JA, Young ID (1996) Holt-Oram syndrome: a clinical genetic study. J Med Genet 33:300–307

Nylander PO, Engstrom C, Chotai J, Wahlstrom J, Adolfsson R (1994) Anticipation in Swedish families with bipolar affective disorder. J Med Genet 31:686–689

O'Donovan MC, Guy C, Craddock N, Murphy KC, Cardno AG, Jones LA, Owen MJ, McGuffin P (1995) Expanded CAG repeats in schizoprenia and bipolar disorder. Nat Genet 10:380–381

Orr HT, Chung M-Y, Banfi S, Kwiatkowski TJ, Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LPW, Zoghbi HY (1993) Expansion of an unstable trinucleotide repeat in spinocerebellar ataxia type 1. Nat Genet 4:221–226

Ott J (1991) Analysis of human genetic linkage, 2nd edn. The Johns Hopkins University Press, Baltimore

Paulson HL, Fischbeck KH (1996) Trinucleotide repeats in neurogenetic disorders. Annu Rev Neurosci 19:79–107

Payami H, Bernard S, Larsen K, Kaye J, Nutt J (1995) Genetic anticipation in Parkinson's disease. Neurology 45:135–138

Pearson WR, Lipman DJ (1988) Improved tools for biological squence analysis. Proc Natl Acad Sci USA 85:2444–2448

Peterson BS (1995) Neuroimaging in child and adolescent neuropsychiatric disorders. J Acad Child Adolesc Psychiatry 34:1560–1576

Pietrokovski (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res24:3836–3845

Plante-Bordeneuve V, Taussig D, Thomas F, Ziegler M, Said G (1995) A clinical and genetic study of familial cases of Parkinson's disease. J Neurol Sci 133:164–172

Polito JM, Rees RC, Childs B, Mendeloff AI, Harris ML, Bayless TM (1996) Preliminary evidence for genetic anticipation in Crohn's disease. Lancet 347:798–800

Prasad R, Canaani O, Kamada N, Gale RP, Lange B, Crist WM, Nowell PC et al. (1993) Genes on chromosomes 4, 9, and 19 involved in 11q23 abnormalities in acute leukemia share sequence homology and/or common motifs. Proc Natl Acad Sci USA 90:4631–4635

Pulst S-M, Nechiporuk A, Nechiporuk T, Gispert S, Chen X-N, Lpes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A, DeJong P, Rouleau GA, Auburger G, Korenberg JR, Figueroa C, Sahba S (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet 14:269–276

Ranum LP, Lundgren JK, Schut LJ, Ahrens MJ, Perlman S, Aita J, Bird TD, Gomez C, Orr HT (1995) Spinocerebellar ataxia type 1 and Machado-Joseph disease: incidence of CAG expansions among adult-onset ataxia patients from 311 families with dominant, recessive, or sporadic ataxia. Am J Hum Genet 57:603–608

Raskind WR, Pericak-Vance MA, Lennon F, Wolff J, Lipe HP, Bird TD (1997) Familial spastic paraparesis: evaluation of locus heterogeneity, anticipation, and haplotype mapping of the SPG4 locus on the short arm of chromosome 2. Am J Med Genet 74:26–36

Ross CA (1995) When more is less: pathogenesis of glutamine repeat neurodegenerative diseases. Neuron 15:493–496

Ross CA, Pearlson GD (1996) Schizophrenia, the heteromodal association neocortex and development: potential for a neurogenetic approach. Trends Neurosci 19:171–176

Ross CA, McInnis MG, Margolis RL, Li S-H (1993) Genes with triplet repeats: candidate mediators of neuropsychiatric disorders. Trends Neurosci 16:254–260

Sanpai K, Takano H, Igarashi S, Sato T, Oyake M, Sasaki H, Wakisaka A, Tashiro K, Ishida Y, Ikeuchi T, Koide R, Saito M, Sato A, Tanaka T, Hanyu S, Takiyama Y, Nishizawa M, Shimizu N, Nomura Y, Segawa M, Iwabuchi K, Eguchi I, Tanaka H, Takahashi H, Tsuji S (1996) Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. Nat Genet 14:277–284

Schalling M, Hudson TJ, Buetow KH, Housman DE (1993) Direct detection of novel expanded trinucleotide repeats in the human genome. Nat Genet 4:135–139

Schuler GD, et al (1996) A gene map of the human genome. Science 274:540–546

Spana EP, Roe CQ (1996) Numb antagonizes notch signaling to specify sibling neuron fates. Neuron 17:21–26

Stine OC (1993) Is autism associated with an X-linked trinucleotide repeat? Psychiatr Genet 3:129

Stober G, Franzek E, Lesch KP, Beckmann H (1995) Periodic catatonia: a schizophrenic subtype with major gene effect and anticipation. Eur Arch Psychiatry Clin Neurosci 245:135–141

Tawil R, Forrester J, Griggs RC, Mendell J, Kissel J, McDermott M, King W, Weiffenbach B, Figlewicz D (1996) Evidence for anticipation and association of deletion size with severity in facioscapulohumeral muscular dystrophy. Ann Neurol 39:744–748

Thibaut F, Martinez M, Petit M, Jay M, Campion D (1995) Further evidence for anticipation in schizophrenia. Psychiatry Res 59:25–33

Trottier Y, Lutz Y, Stevanin G, Imbert G, Devys D, Cancel G, Saudou F, Weber C, David G, Tora L, et al. (1995) Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. Nature 378:403–406

Verkerk AJMH, Pieretti M, Sutcliffe JS, Fu Y-H, Kuhl DPA, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang F, Eussen BE, Ommen van G-JB, Blonden LAJ, Riggins GJ, Chastain JL, Kunst CB, Galjaard H, Caskey CT, Nelson DL, Oostra BA, Warren ST (1991) Identification of a gene (FMR-1) containing CGG repeat coincident with a breakpoint cluster region exhibiting length variation in Fragile X syndrome. Cell 65:905–914

Weinberger DR (1995) From neuropathology to neurodevelopment. Lancet 346:552–557

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second-generation linkage map of the human genome. Nature 359:794–801

Welch MD, Drubin DG (1994) A nuclear protein with sequence similarity to proteins implicated in human acute leukemias is important for cellular morphogenesis and actin cytoskeletal function in *Saccharomyces cerevisiae*. Mol Biol Cell 5:617–632

Yamagata K, Takeda J, Menzel S, Chen X, Eng S, Lim LR, Concannon P, Hanis CL, Spielman RS, Cox NJ, Bell GI (1996) Searching for NIDDM suscpetibility genes: studies of genes with triplet repeats expressed in skeletal muscle. Diabetologia 39:725–730

Zhong W, Feder JN, Jiang M-M, Jan LY, Jan YN (1996) Asymmetric localization of a mammalian numb homolog during mouse cortical neurogenesis. Neuron 17:43–53

Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC (1997) Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the a1A-voltage dependent calcium channel. Nat Genet 15:62–69

Zoghbi HY (1996) The expanding world of ataxins. Nat Genet 14:237–238