J Hum Genet (2001) 46:413–422

ORIGINAL ARTICLE

Yaeko Ichikawa · Jun Goto · Masahira Hattori
Atsushi Toyoda · Kazuo Ishii · Seon-Yong Jeong
Hideji Hashida · Naoki Masuda · Katsuhisa Ogata
Fumio Kasai · Momoki Hirai · Patrícia Maciel
Guy A. Rouleau · Yoshiyuki Sakaki · Ichiro Kanazawa

The genomic structure and expression of *MJD*, the Machado-Joseph disease gene

Received: March 7, 2001 / Accepted: April 17, 2001

Abstract Machado-Joseph disease (MJD) is an autosomal dominant neurodegenerative disorder that is clinically characterized by cerebellar ataxia and various associated symptoms. The disease is caused by an unstable expansion of the CAG repeat in the *MJD* gene. This gene is mapped to chromosome 14q32.1. To determine its genomic structure, we constructed a contig composed of six cosmid clones and eight bacterial artificial chromosome (BAC) clones. It spans approximately 300 kb and includes *MJD*. We also determined the complete sequence (175,330 bp) of B445M7, a human BAC clone that contains *MJD*. The *MJD* gene was found to span 48,240 bp and to contain 11 exons. Northern blot analysis showed that *MJD* mRNA is ubiquitously expressed in human tissues, and in at least four different sizes; namely, 1.4, 1.8, 4.5, and 7.5 kb. These different mRNA species probably result from differential splicing and polyadenylation, as shown by sequences of the 21 independent cDNA clones isolated after the screening of four human cDNA libraries prepared from whole brain, caudate, retina, and testis. The sequences of these latter clones relative to the *MJD* gene in B445M7 indicate that there are three alternative splicing sites and eight polyadenylation signals in *MJD* that are used to generate the differently sized transcripts.

Key words Machado-Joseph disease (MJD) · 14q32.1 · CAG repeat · Genome structure · Alternative splicing · mRNA expression

Y. Ichikawa · J. Goto (✉) · H. Hashida · N. Masuda · K. Ogata · I. Kanazawa
Department of Neurology, Graduate School of Medicine, The University of Tokyo, 7-3-3 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan
Tel. +81-3-5800-8672; Fax +81-3-5800-6548
e-mail: gotoj-tky@umin.ac.jp

Y. Ichikawa · J. Goto · S.-Y. Jeong · H. Hashida · N. Masuda · K. Ogata · I. Kanazawa
CREST, Japan Science and Technology Corporation, Saitama, Japan

M. Hattori · A. Toyoda · K. Ishii · Y. Sakaki
RIKEN Genomic Science Center, Yokohama, Japan

F. Kasai · M. Hirai
Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

P. Maciel
UnIGENe, IBMC University of Porto, Porto, Portugal

G.A. Rouleau
Centre for Research in Neuroscience, McGill University and the Montreal General Hospital Research Institute, Montreal, Québec, Canada

Introduction

Machado-Joseph disease (MJD) is an autosomal dominant neurodegenerative disorder clinically characterized by cerebellar dysfunction and various associated symptoms. The responsible gene, denoted as *MJD*, maps to chromosome 14q32.1. Disease is associated with an unstable expansion of the CAG trinucleotide repeat contained in *MJD* (Kawaguchi et al. 1994). Mutations leading to the unstable expansion of CAG repeats in other genes have also been reported in nine other neurodegenerative disorders, including Huntington's disease (HD), (The Huntington's Disease Collaborative Research Group 1993), spinocerebellar ataxia type 1 (SCA1) (Orr et al. 1993), SCA2 (Pulst et al. 1996; Imbert et al. 1996; Sanpei et al. 1996), SCA6 (Zhuchenko et al. 1997), SCA7 (David et al. 1997), SCA12 (Holmes et al. 1999), dentatorubral-pallidoluysian atrophy (DRPLA) (Nagafuchi et al. 1994; Koide et al. 1994), and spinal and bulbar muscular atrophy (SBMA) (La Spada et al. 1991). The CAG repeats lead to the expression of a protein containing multiple glutamines, and pathology is thought to arise as a consequence of a gain of function by such proteins. As the truncated protein containing the expanded polyglutamine sequence induces apoptotic cell death in vitro and the gene products are cleaved by caspase-3, a proapoptotic cysteine protease (Ikeda et al. 1996; Goldberg et al. 1996), it may be that the expanded polyglutamine promotes apoptosis in the brain. This could result in the clinical phenotype associated with these so-called polyglutamine diseases. Indeed, neuronal nuclear

aggregates are commonly found in the brains of those people suffering from polyglutamine diseases. Transgenic mice generated to understand the pathogenesis of the polyglutamine diseases include two strains of *MJD* transgenic mice. One of these strains, which selectively expresses the expanded polyglutamine stretch in Purkinje cells, develops a severely atrophic cerebellum and has an ataxic phenotype (Ikeda et al. 1996). The other transgenic strain, the YAC transgenic mouse that carries the full-length human *MJD* gene, shows a mild and slowly progressive cerebellar deficit (Cemal et al. 1999a; Cemal et al. 1999b).

Despite these studies, the molecular mechanisms by which CAG repeat expansion causes disease remain unclear. Consequently, to better understand of the molecular basis of *MJD*-associated pathology, we have studied the genomic structure of *MJD* and its transcription. Previous work has also attempted to do this. Kawaguchi et al. (1994) isolated and characterized a cDNA clone denoted as MJD1a and reported that *MJD* consists of 1776 bp and may be composed of four exons (Kawaguchi et al. 1994). In addition, in previous work, we isolated two *MJD* cDNA clones that have alternative carboxyl terminus exon sequences (Goto et al. 1997). In this study, we have aimed to more thoroughly understand the structure of *MJD* and its expression. We report here the complete structure of *MJD* and show that *MJD* is expressed as several differently sized transcripts resulting from differential splicing and polyadenylation.

Materials and methods

Cosmid library screening

The cosmids were obtained from the Los Alamos gridded chromosome 14-specific cosmid library (Xie Y.-G. et al. 1998). The library was screened by direct hybridization, using 14 gridded filters, each of which contains 1152 clones. These filters were constructed as described in previous reports (Xie Y.-G. et al. 1998; Lafrenière et al. 1995). The condition used for isolation of cosmids was stringent, hybridization and washes were performed at 65°C, and final washes in 0.1 × standard saline citrate buffer and 0.1% sodium lauryl sulfate.

Bacterial artificial chromosome (BAC) library screening

The RPCI11 human BAC library (Rosewell Park Cancer Institute, New York, NY, USA) was used for screening. Successive pools of BAC DNA in a series of microtiter wells were screened by polymerase chain reaction (PCR), using the primers BAC-F (5′-TGA CTT GAA GTG CTA ATA GCA CAG-3′) and BAC-R (5′-AAT CAG GTA TAA ACC TAA GGT GCT CT-3′). These primers were designed on the basis of the sequence of intron 1. Two replicate microtiter plates, Z1-4 and Z5-8, were screened. The first amplification was performed with a hot start at 70°C, followed by 40 thermal cycles, consisting of 1 min at 94°C, 30 s at 55°C, and 30 s at 72°C. The second amplification was carried out under the same conditions, except that 25 thermal cycles were used. Replica plates bearing positive clones were screened by an identical method.

Cosmid and BAC DNA preparation

Cosmid and BAC clones were grown overnight at 37°C in Luria-Bertani (LB) or in 2 × YT media containing appropriate antibiotics (10 μg/ml ampicillin for cosmid, 50 μg/ml chloramphenicol for BAC). Cosmid DNA was prepared by the alkaline lysis method (Sambrook et al. 1989) and BAC DNA was prepared using the Qiagen plasmid maxi kit (Qiagen, Chatsworth, CA, USA) with one minor alternation; namely, that the elution buffer was used after being preheated at 65°C.

Sequencing

Sequencing reactions were performed using a DYEnamic Direct sequencing kit with T7 and T3 primers (Amersham Pharmacia Biotech, Buckinghamshire, UK) or a BigDye Terminator Cycle Sequencing reaction kit (Perkin-Elmer, Foster City, CA, USA), both used according to the manufacturer's instructions. The sequencing products were analyzed by a 377 automated fluorescence sequencer (Applied Biosystems, Foster City, CA, USA). The nested deletion method (Hattori et al. 1997) was used to sequence the cosmid clones. The sequence of the BAC clone B445M7 was determined by the shot-gun strategy for large-scale sequencing, at Hitachi (Tokyo, Japan) (Fleischmann et al. 1995).

Data analysis

Sequence data were analyzed with Sequencing Analysis software (version 3.0; Perkin-Elmer), Sequencher software (version 3.0; Genecodes, Ann Arbor, MI, USA) and Genetyx-Mac software (version 10; Software Development, Tokyo, Japan). Database searches were carried out using the BLAST program through the National Center for Biotechnology Information (NCBI). Searches for motif and transcription factor sequences in *MJD* were performed using the MOTIF and the TFSEARCH programs in the Genome Net WWW Server Japan.

Fluorescence in situ hybridization (FISH) and fiber FISH

Chromosome mapping of the cosmid and BAC clones was carried out by FISH, as previously described (Hirai et al. 1996). A contig of the *MJD* gene region was constructed by the fiber FISH method (Mann et al. 1997). The stretched DNA fibers were hybridized with two probes, one labeled with biotin (fluorescein isothiocyanate [FITC]; green) and

the other with digoxigenin (rhodamine; red). The cosmid clones 24C7 (cos2) and 76F4 (cos5) were used as references to determine the chromosomal locations of the BAC clones.

cDNA libraries and screening

Four human cDNA libraries were screened for the presence of *MJD* cDNA clones. The mRNA for the libraries originated from caudate, retina, whole brain, and testis. The caudate library and its screening have been described previously (Goto et al. 1997). The retina library was purchased from Clontech (Palo Alto, CA, USA) and was constructed on a λ gt10. The whole brain and testis cDNA libraries were constructed using a ZAP Express cDNA vector (Stratagene, La Jolla, CA, USA). Two probes were generated. The first (used to probe the retina library) consists of a 185-bp cDNA fragment of *MJD* that was amplified with the MJD13F and MJD197R primers, and the second (used to probe the brain and testis libraries) is composed of a 434 bp cDNA fragment of *MJD* that was amplified with the MJD74F(52) and MJD507R(485) primers (see below for primer sequences and PCR conditions). The probes were labeled with [α-^{32}P] dCTP (6000 Ci/mmol) or digoxigenin (DIG)-dUTP. Hybridization and washing were performed according to the standard method (Sambrook et al. 1989). The cDNA inserts of the positive λ gt10 clones were subcloned into pBluescript SKII(+) (Stratagene), while the positive λ ZAP XR and λ ZAP Express clones were self-excised, and the corresponding plasmids were obtained according to the manufacturer's instructions (Stratagene).

PCR

The primer pairs used to amplify the exons were as follows: exon 4: 277F (5′-GCA ATG CCT TGA AAG TTT GG-3′) and 355R (5′-ATA GGA TCG ATC CTG AGC CT-3′); exons 7 and 8: 585F (5′-CAA CAG ATG CAT CGA CCA AA-3′) and 694RK (5′-CCT GAG CCA TCA TTT GCT TCT AAT A-3′); exon 10: MJD52 (5′-CCA GTG ACT ACT TTG ATT CG-3′) and MJD70 (5′-CTT ACC TAG ATC ACT CCC AA-3′); exon 11: 5R2 (5′-GAT TAC AGC ATA GGG GTC CAC-3′) and 5R1 (5′-CAC TGG AGC ACA CGG TAT AC-3′).

The primer pairs used to prepare probes for cDNA library screening were as follows: retina cDNA library: MJD13F (5′-CCG TTG GCT CCA GAC AAA TA-3′) and MJD197R (5′-AGT AAC TCC TCC TTC TGC CA-3′); whole brain and testis cDNA libraries: MJD74F(52) (5′-ACG AGA AAC AAG AAG GCT CAC T-3′) and MJD507R(485) (5′-AAG TGC AAG ATA TGT ATC TGA TAT TAA TTC T-3′).

Probe and exon amplifications were performed by 30 cycles of 1-min denaturation at 94°C, 1-min annealing at 56°C, and 1 min extension at 72°C.

The primers used for reverse transcription (RT)-PCR (see below) were 1818F (5′-GCG TTC CTA AAC TCT GAA ATC AGC CTT GCA CAA GTA CT-3′) and 164664R (5′-TGA TAC TAT GGT TAC TTG CAC TGG CAT CTT TTC ATA CTG GC-3′).

Northern blot analysis

Three kinds of multiple tissue northern blots (heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas); human brain II (cerebellum, cerebral cortex, medulla, spinal cord, occipital pole, frontal lobe, temporal lobe, and putamen); and human brain IV (amygdala, caudate nucleus, corpus callosum, hippocampus, whole brain, substantia nigra, and thalamus), were purchased (Clontech). Northern blot hybridization was performed according to the manufacturer's instructions. The probe used was the same as that used to screen the human whole brain and testis cDNA libraries. After removal of the probe by 0.5% sodium dodecylsulfate (SDS), the filters were rehybridized with labeled human β actin cDNA as a control.

RT-PCR

Total RNA was extracted from peripheral blood leukocytes, using Trizol reagent (GIBCO-BRL, Rockville, MD, USA). One microgram of total RNA was treated with RNase-free DNaseI (GIBCO-BRL). Reverse transcription was performed using SuperScript II reverse transcriptase (GIBCO-BRL) with oligo(dT) primer. PCR amplification was carried out using Advantage cDNA polymerase mix (Clontech) with the primer pairs 1818F and 164664R (see above). PCR conditions consisted of one cycle for 30 s at 94°C, followed by five cycles of 5 s at 94°C and 4 min at 72°C, followed by another five cycles of 5 s at 94°C and 4 min at 70°C, and ending with 25 cycles of 5 s at 94°C and 4 min at 68°C.

Results

Cosmid and BAC contig construction

Thirteen cosmid clones were obtained when the human chromosome 14-specific genomic library was screened. These clones are denoted as 4H9 (cos1), 24C7 (cos2), 62B11 (cos3), 74C8 (cos4), 76F4 (cos5), 79H4 (cos6), 87A8 (cos8), 111F12 (cos9), 113A3 (cos10), 137H5 (cos11), 161H1 (cos12), 168A1 (cos13), and 10B9 (cos14). FISH analysis revealed that the cos1, cos2, cos3, cos5, cos10, cos11, cos13, and cos14 sequences are located on chromosome 14q32.1 and the other 5 clones on 14q21–22. Among the cosmid clones that map to 14q32.1, exon sequences of exon 4, exons 7 and 8, and exon 10 and exon 11 were amplified by PCR. The sequence of exon 4 was contained in cos2, cos13, cos11, and cos3, that of exons 7 and 8 in cos11, cos1, cos5, and cos3, and that of exons 10 and 11 in cos1 and cos5 (data not shown). The fiber FISH method revealed that 6 cosmid clones mapped to 14q32.1 aligned from centromere to telomere in the following order: cos2, cos11, cos13, cos5, cos14, and cos1 (Fig. 1). Cos2, cos11, and cos13 were un-

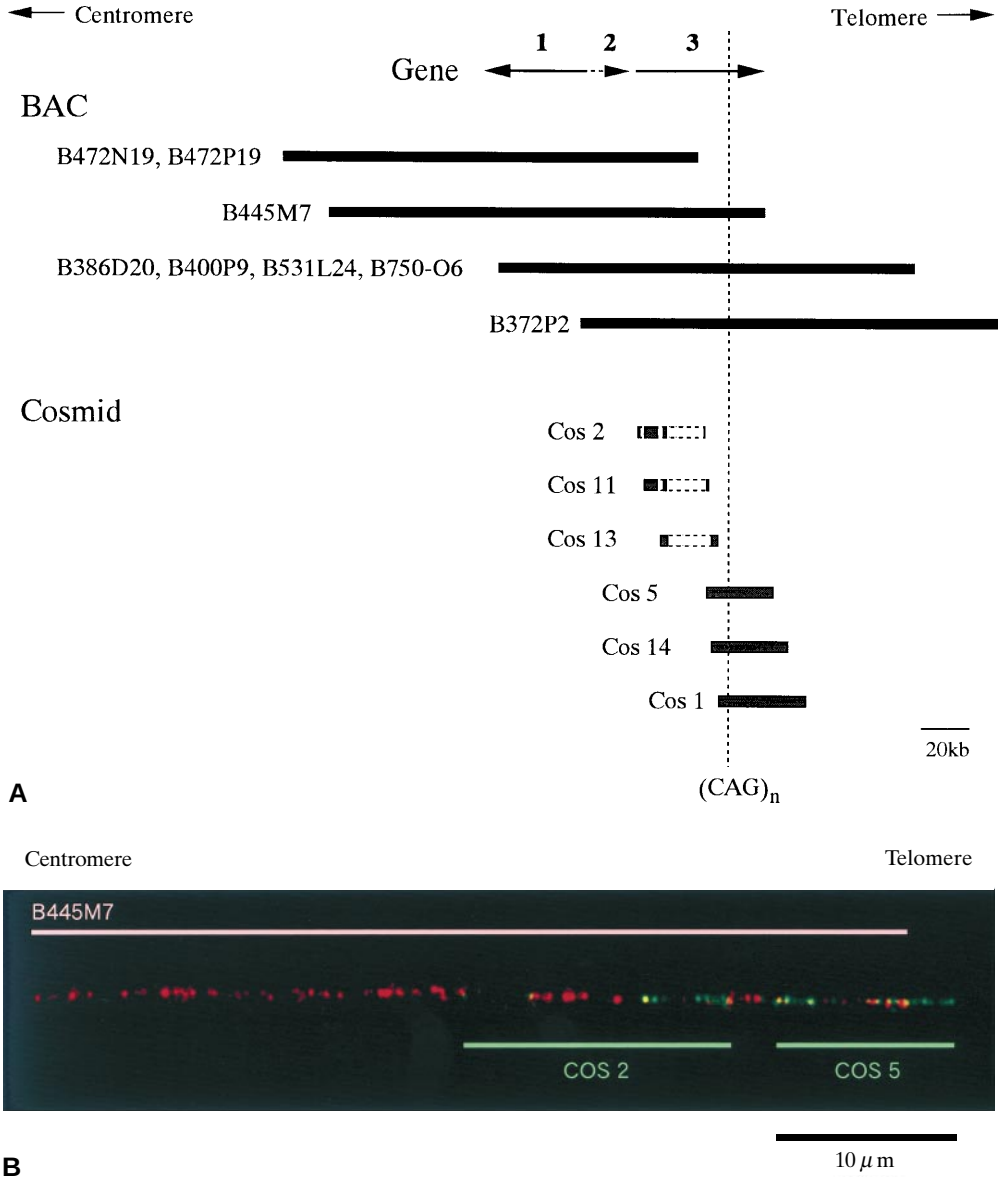


Fig. 1A,B. Structure and analysis of the *MJD* gene contig. **A** Schematic representation of the contig. Gene 1: human homologue of the *Bos taurus* cleavage and polyadenylation specificity factor (*BTCPSFSU*). Gene 2: *NDUFB1*. Gene 3: *MJD*. The *arrows* denote the directions of transcription. Cos2, cos11, and cos13 were partially deleted. The deleted region of the cosmid is denoted by *dotted lines*. **B** Two-color fiber fluorescence in situ hybridization (FISH) analysis of BAC B445M7 and the cosmid clones cos2 and cos5. B445M7 was visualized by the digoxigenin/rhodamine system (*red*) and cos2 and cos5 by the biotin/fluorescein isothiocyanate (FITC) system (*green*). Overlapping regions of the clones are shown by *yellow* points

stable and partially deleted. Cos3 and cos10 clones did not overlap with any of the above 6 cosmid clones.

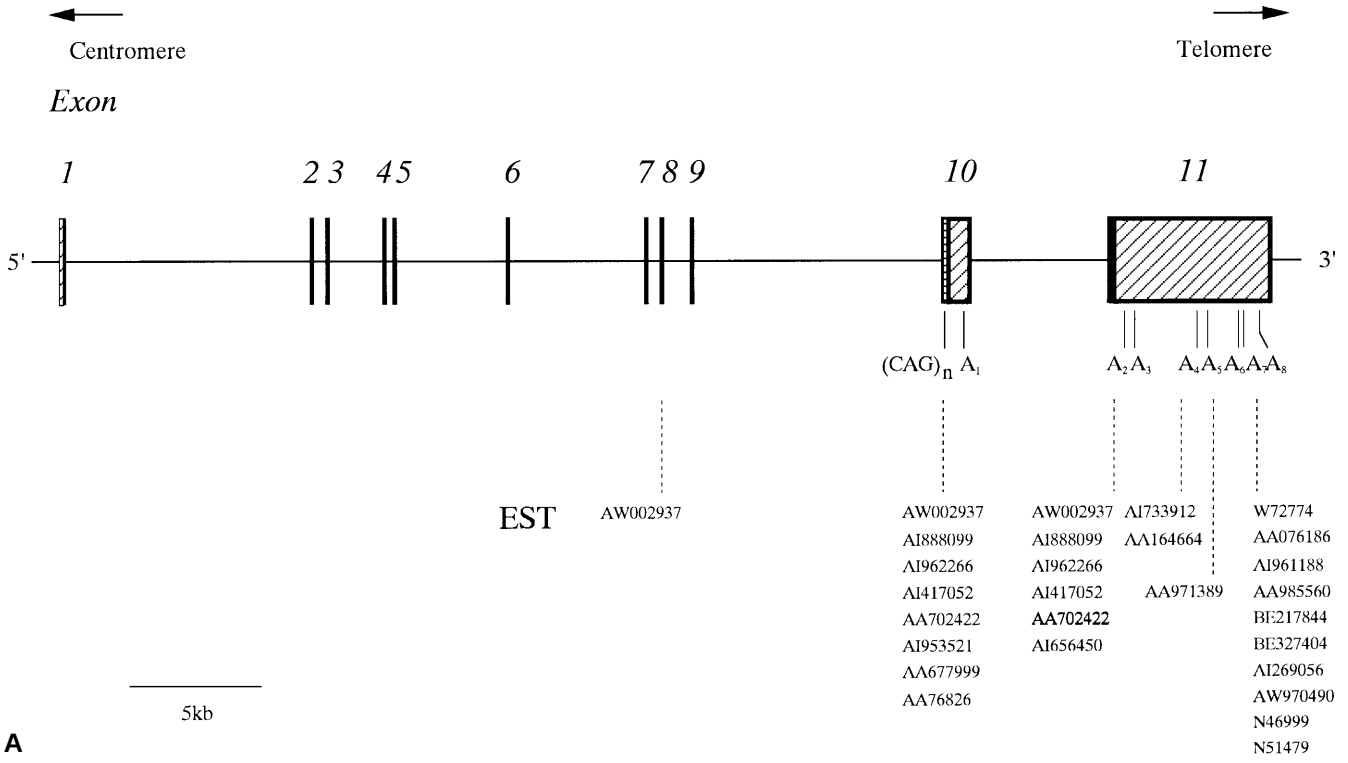
Screening of the RPCI11 BAC library was subsequently carried out, and eight BAC clones were obtained: B472N19, B472P19, B445M7, B386D20, B400P9, B531L24, B750-O6, and B372P2. All BAC clones mapped to 14q32.1. By the fiber FISH method, we constructed a contig, on the basis of the six cosmid and eight BAC clones, that comprises an approximately 300-kb region containing *MJD* (Fig. 1).

Sequence analysis

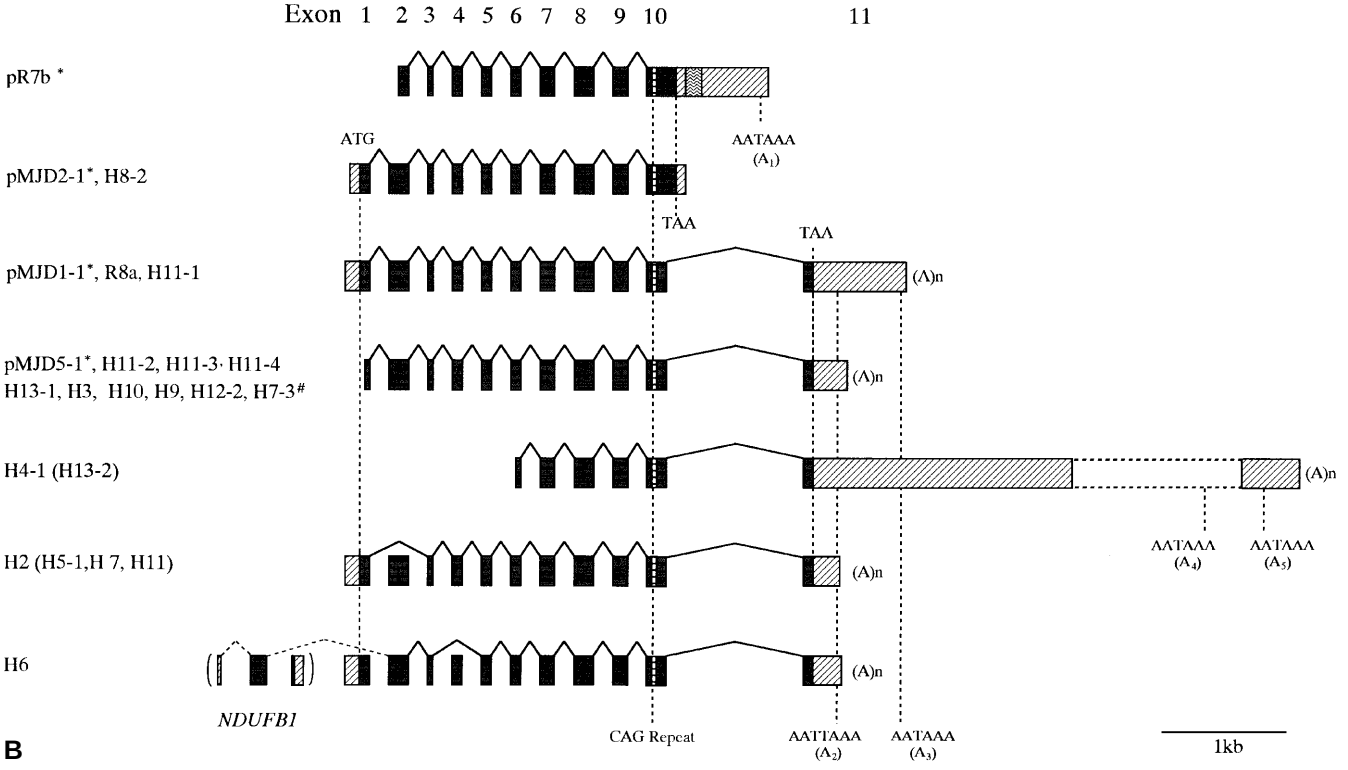
The eight cosmid clones mapping to 14q32.1 and the B445M7 BAC clone were sequenced. The complete sequence of B445M7 was found to be composed of 175,330 bp, and has been submitted to the DDBJ/GenBank/EMBL database (accession number, AB038653). Within B445M7 is the *MJD* gene, which spans the nucleotide (nt) region 126,768–175,007. The gene is approximately 48.2 kb in size and is composed of 11 exons (Fig. 2A). The exon/intron borders are presented in Table 1.

The 5′-end sequences of five independent cDNA clones isolated from human cDNA libraries (see below) indicate that the transcription initiation site is probably at nucleotide position 126,768 in B445M7. The *MJD* translation initiation site in B445M7 is at nt 126,826. An Alu element is found 437 bp upstream of the presumed transcription start site, and between these two sites, four GC box consensus sequences were identified. Neither TATA boxes nor CCAAT boxes are found in this region. With regard to transcription factors, Sp1, Ap2, USF and SRY binding sequences were identified. The translation initiation site is located at nt 126,826 in B445M7, and is in good agreement with Kozak's consensus sequence (Kozak 1987).

As will be described further below, either exon 10 or exon 11 is used by alternative splicing to form the 3′-end of *MJD* mRNA. Exon 10 is composed of 1099 bp and includes



A



B

Fig. 2A–C. The genomic structure and transcripts of *MJD*. **A** The genomic structure of *MJD*. Exons are numbered 1 to 11 and are presented as *boxes*. *Filled boxes* indicate the coding regions and *hatched boxes* represent the 5′- and 3′-untranslated regions (UTRs). A_1 *to* A_8 are polyadenylation consensus sequences. Approximate positions of expressed sequence tags (EST) sequences are also depicted. **B** cDNA clones and their usages of the exons. Representative cDNA clones are shown with a *superscript asterisk*. Others are cDNA clones that are shorter than the representative clones in length. cDNA clone H7-3 (*superscript hash sign*) is a chimera clone in its 5′ region. cDNA clones *shown in parentheses* are identical to the representative clones. *Boxes* denote exons; *filled boxes* represent coding regions; *hatched boxes* represent untranslated regions; the *wavy area* represents Alu; and the *dotted line* in exon 10 represents CAG repeat sequences. The translation initiation site is presented as ATG and termination sites are shown as TAA. Polyadenylation signals (AATAAA and AATTAAA) are presented as A_1 *to* A_5, as in **A**. **C** Representative amino-acid sequences of the *MJD* products. *Dashes* represent the same amino acid residue as that in the MJD1a sequence (Kawaguchi et al. 1994). *Blank* is a deleted residue. *Dots* indicate polymorphic sites associated with amino-acid substitution. The sequences of pMJD2-1, pMJD5-1, and pMJD1-1 were previously reported (Goto et al. 1997). H2 has been submitted to the DDBJ/GenBank/EMBL database, under accession number AB050194

```
 MJD1a     MESIFHEKQEGSLCAQHCLNNLLQGEYFSPVELSSIAHQLDEEERMRMAEGGVTSEDYRTFLQ     63
pMJD2-1    ------------------------------------------------------------
pMJD5-1     -----------------------------------------------------------
pMJD1-1    ------------------------------------------------------------
H2         --------                                                       8

 MJD1a     QPSGNMDDSGFFSIQVISNALKVWGLELILFNSPEYQRLRIDPINERSFICNYKEHWFTVRKL    126
pMJD2-1    ------------------------------------------------------------
pMJD5-1    ------------------------------------------------------------
pMJD1-1    ------------------------------------------------------------
H2         ------------------------------------------------------------     71

 MJD1a     GKQWFNLNSLLTGPELISDTYLALFLAQLQQEGYSIFVVKGDLPDCEADQLLQMIRVQQMHRP    189
pMJD2-1    ------------------------------------------------------------
pMJD5-1    ------------------------------------------------------------
pMJD1-1    ------------------------------------------------------------
H2         ------------------------------------------------------------    134
                                 •
 MJD1a     KLIGEELAQLKEQRVHKTDLERMLEANDGSGMLDEDEEDLQRALALSRQEIDMEDEEADLRRT    252
pMJD2-1    ----------------------V-------------------------------------
pMJD5-1    ------------------------------------------------------------
pMJD1-1    ----------------------V-------------------------------------
H2         ----------------------V-------------------------------------    197

 MJD1a     IQLSMQGSSRNISQDMTQTSGTNLTSEELRKRREAYFEKQQQKQQQQQQQQQQQQQQQQQQQ    315
pMJD2-1    ----------------------------------------------------          305
pMJD5-1    ------------------------------------------------------------    315
pMJD1-1    ----------------------------------------------------            305
H2         ----------------------------------------------------            250
              •                                          •
 MJD1a     QQ RDLSGQSSHPCERPATSSGALGSDLGKACSPFIMFATFTLYLT*                  360
pMJD2-1       G-----------------------------------------YELHVIFALHYSSFPL*    364
pMJD5-1    --Q--------------------------                                   344
pMJD1-1       G-------------------------                                    331
H2            G-------------------------                                    276

pMJD5-1    DAMSEEDMLQAAVTMSLETVRNDLKTEGKK*     374
pMJD1-1    DAMSEEDMLQAAVTMSLETVRNDLKTEGKK*     361
H2         DAMSEEDMLQAAVTMSLETVRNDLKTEGKK*     306
```

C

Fig. 2A–C. Continued

one polyadenylation signal consensus sequence (A1 in Fig. 2A,B). Exon 11 is composed of 6046 bp and contains seven polyadenylation signal consensus sequences (A_2 to A_8 in Fig. 2A,B). As shown by the cDNA clones, the A_2, A_3, and A_5 polyadenylation signals at least are used (see below and discussion). A8 is found 2037 bp downstream of A_5 in several ESTs: AA985560, AI961188, N51479, N46999, BE217844, BE327404, AI269056, and AW970490.

The BLAST search indicated that additional gene sequences apart from *MJD* are present within B445M7. The nt region 102,427–70,540 contains the human homologue of the *Bos taurus* cleavage and polyadenylation specificity factor (*BTCPSFSU*). Several partial sequences of *NDUFB1* were also found (regions 111,694–111,736, 114,409–114,498, 115,736–115,880, and 117,149–117,255) (Fig. 1A). The 114,781–115,255 region contains the EST: AA456166, while EST: AA203662 can be found in the region comprised by nt 116,067–116,662.

cDNA cloning

In total, 5×10^6 pfu of four human cDNA libraries derived from caudate, retina, whole brain, and testis mRNA were screened. This resulted in the isolation of 27 *MJD* cDNA clones, 3 clones from the caudate library, 4 from the retina library, 9 from the whole brain library, and 11 from the testis library. Sequence analysis revealed that 21 of the 27 clones were independent of each other and could be classified into four types according to the splicing patterns used to generate them (Fig. 2B). Type 1 clones consist of pMJD2-1, pR7b, and H8-2, and are similar in structure to the cDNA clone MJD1a that was originally described by Kawaguchi et al. (1994). These clones use exon 10 to provide the 3′-terminal sequence. Relative to MJD1a, our three type 1 clones carry a single nucleotide substitution in the stop codon as described previously (Goto et al. 1997).

Table 1. Exon-intron boundaries of the *MJD* gene

Exon Intron	Exon position in B445M7	Length (bp) Exon	 Intron	Acceptor site	Donor site
Exon 1	126,768–126,849	82			CCACGAGAAA/gtgagtgtcc
Intron 1			9,682		
Exon 2	136,540–136,704	165		gaaatattag/CAAGAAGGCT	GTTTTTACAG/gtactgattt
Intron 2			536		
Exon 3	137,241–137,285	45		aaattaacag/CAGCCTTCTG	CTCTATTCAG/gtaagtagtc
Intron 3			2,260		
Exon 4	139,546–139,631	86		ttcttaacag/GTTATAAGCA	TCGATCCTAT/gtaagattct
Intron 4			427		
Exon 5	140,059–140,125	67		tctttttag/AAATGAAAGA	AGGAAAACAG/gtaacatttc
Intron 5			4,434		
Exon 6	144,560–144,647	88		tttttccag/TGGTTTAACT	CAACAGGAAG/gtaagtaacg
Intron 6			5,471		
Exon 7	150,119–150,251	133		ttttttttag/GTTATTCTAT	AAGAGCAAAG/gtaaaaatga
Intron 7			659		
Exon 8	150,911–151,077	167		tgttctttag/AGTCCATAAA	AGTATGCAAG/gtaaagacat
Intron 8			1,238		
Exon 9	152,316–152,412	97		ttgttttcag/GTAGTTCCAG	ACTTTGAAAA/gtaaagtagt
Intron 9			9,910		
Exon 10	162,323–163,421	1,099[a]		aatgtttcag/ACAGCAGCAA	
	162,323–162,441	119[b]		aatgtttcag/ACAGCAGCAA	AGTGATCTAG/gtaaggcctg
Intron 10			6,520		
Exon 11	168,962–175,007	6,046		ttcccaacag/GTGATGCTAT	

[a] Length of exon 10 in type 1 cDNA clones
[b] Length of exon 10 in type 2 cDNA clones

There are 14 type 2 clones (pMJD1-1, pMJD5-1, R8a, H11-1, H11-2, H11-3, H11-4, H13-1, H9, H12-2, H3, H10, H4-1, and H13-2). All contain exon 11 and skip their 3′-regions of exon 10 due to the use of the splicing donor site in exon 10, as described in a previous report (Goto et al. 1997). Of the type 2 clones, the H4-1 and H13-2 clones are identical and carry the longest 3′-untranslated region (UTR). They do not contain a 1510-bp fragment (nts 171,111–172,620 in B445M7) in their 3′-UTR (Fig. 2A) but this may be an artifact, because the RT-PCR products amplified with the primers 1818F and 164664R include this fragment (data not shown), and two expressed sequence tags (ESTs) (AI733912 and AA164664) that include this fragment have been registered in GenBank. There are not splice consensus sequences or repeats at the boundaries of this region. Four polyadenylation signal consensus sequences, denoted as A_2 to A_5 (Fig. 2A,B), could be identified in the long 3′-UTR of the type 2 clones.

Type 3 clones consist of four identical clones, H2, H5-1, H7, and H11. These clones originate from the testis library. In these clones, exon 2 is skipped, presumably by alternative splicing, but this does not cause a shift in the reading frame. The sequence of H2 has been submitted to the DDBJ/GenBank/EMBL database (accession number, AB050194).

The type 4 clone category is a miscellaneous collection of clones that cannot be classified into any of the prior three types. R3d and R6a are short and are composed of only exons 2 to 9. H6 and H7-3 are chimeras, as their 5′-regions are derived from genes other than *MJD*. With regard to H6, the 173 bp at its 5′-end is derived from nicotinamide adenine dinucleotide (NADH), reduced dehydrogenase (ubiquinone) 1 beta subcomplex, 1 (7 kD; MNLL); *NDUFB1* (Fig. 2B). *NDUFB1* is thought to be divided into four parts (as shown by comparing *NDUFB1* with the genome sequence of B445M7) and the 173 bp at the 5′-end of H6 consists of a partial sequence from the second part of *NDUFB1* linked to the 3′-end of the third part. The 5′-end of the *MJD*-derived sequence of H6 consists of the 5′-end of the second *MJD* exon. Exon 4 of *MJD* is skipped in H6. In the case of the chimeric clone H7-3, the 349 bp at its 5′-end is highly homologous to human glutaminase liver isoform mRNA (AF223944.1; 95% similarity), while the *MJD*-derived region is identical to the 453-1372 sequence of pMJD1-1, except in the number of CAG repeats. The remaining type 4 clone, H5-2, is not a cDNA for *MJD* but is highly homologous. Its homology to pMJD1-1 is 79.6% at the nucleotide level and 70% at the amino-acid level. A homology search revealed that the sequence of H5-2 is highly homologous (99.7%) to a genomic sequence (nts 52,269–53,529) within BAC GSHB-600G8, an Xp22 BAC clone. The sequence of H5-2 has been submitted to the DDBJ/GenBank/EMBL database (accession number, AB050195).

All four of the type 3 clones (H2, H5-1, H7, and H11) and one of the type 2 clones, pMJD1-1, have equivalent 5′-ends and the longest 5′-UTRs. The sequences of these clones start at the nt 126,768 position of the BAC clone B445M7. This is 23 nt upstream of the first 5′ nt in MJD1a (Kawaguchi et al. 1994) and is probably the transcription initiation site of *MJD*.

From the results of cDNA cloning, we identified two alternative splicings. We previously reported three polymorphisms in the *MJD* gene (Goto et al. 1997). From these findings, we estimated that there were at least five *MJD* gene products (Fig. 2C). Four of five type amino acid

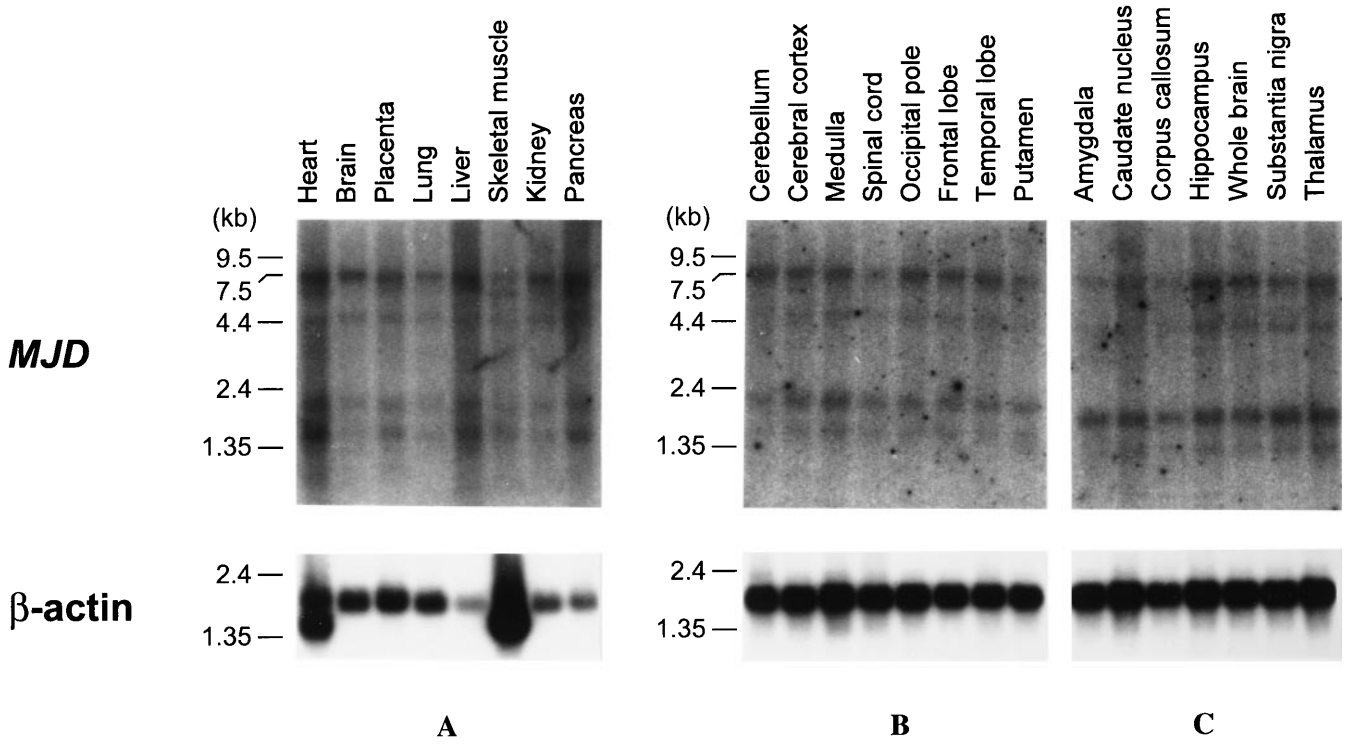


Fig. 3A–C. Northern blot analysis of human *MJD*. **A** Various adult human organs. **B** and **C** Subregions of adult human brain. Each lane contains 2 μg of poly(A)$^+$ RNA

sequences of gene products were previously reported (Goto et al. 1997). The amino acid sequence of H2 (H5-1, H7 and H11) clone is the same as of sequence of pMJD1-1 except for 55 amino acids missing due to exon2 skipping by alternative splicing.

Northern blot analysis

Figure 3 shows the Northern blots of various human tissues to determine where in the body *MJD* is expressed. *MJD* was found to be ubiquitously expressed both in the brain and in non-nervous tissues. Several differently sized transcripts, of approximately 1.4, 1.8, 4.5, and 7.5 kb were detected.

Discussion

In this study we have characterized the genomic structure of the *MJD* gene and assessed the location and nature of its transcription. We first constructed a contig composed of six cosmid clones and eight BAC clones. Sequencing of B445M7, a BAC clone obtained from the RPCI11 human BAC library, allowed us to further characterize the *MJD*. As a result, we know that the *MJD* gene, which spans the nt region 126,768–175,007 in B445M7, is 48,240 bp in size and is composed of 11 exons. The CAG repeats are found in exon 10.

To study the transcription of the *MJD* gene, four human cDNA libraries, established from caudate, retina, whole brain, and testis cDNA, were screened for *MJD* sequences. This resulted in the isolation of 21 independent *MJD* cDNA clones. Five of these clones have longer 5′-UTRs than MJD1a, and are the longest of all the 21 clones. As the 5′-regions of these five clones all start at nt 126,768 in B445M7, we believe that the transcription initiation site of *MJD* is probably at this position.

As described further below, exons 10 and 11 are used variably in the transcripts of *MJD*, probably by alternative splicing mechanisms. Exon 10 is composed of 1099 bp and includes one polyadenylation signal consensus sequence (A_1 in Fig. 2A,B). Exon 11 is 6046 bp in size and contains a long 3′-UTR, of 5954 bp. This is not uncommon, as the transcripts of *HD* and *SCA1* have also been shown to carry long 3′-UTRs, of 3921 bp and 7277 bp, respectively (Banfi et al. 1994; Lin et al. 1993).

We found, by Northern blot analysis, that the *MJD* gene is ubiquitously transcribed, as all the adult human tissues we examined contained detectable levels of *MJD* mRNA. This supports the observations made by Nishiyama et al. (1996), who showed by in situ hybridization analysis that *MJD* mRNA was expressed in the brains of both patients with MJD and control individuals (Nishiyama et al. 1996). Such ubiquitous expression of a gene whose defects primarily lead to neurodegeneration is not unusual, as Paulson et al. (1997) have demonstrated by immunohistochemical studies that ataxin-3, the protein of *MJD*, is expressed not only in the brain but also throughout the body (Paulson et al. 1997). Other genes — *HD*, *SCA1*, *SCA2*, *SCA7*, and *DRPLA* — also appear to be ubiquitously expressed (The Huntington's

Disease Collaborative Research Group 1993; Strong et al. 1993; Orr et al. 1993; Pulst et al. 1996; Sanpei et al. 1996; Imbert et al. 1996; David et al. 1997; Onodera et al. 1995).

Our Northern blot analysis also revealed that there were at least four different species of *MJD* transcripts. Their sizes were estimated to be approximately 1.4, 1.8, 4.5, and 7.5 kb. This is probably partly as a result of differential splicing. This is suggested by the cDNA clone sequences, as the 3′-ends of the clones were composed of either exon 10 or exon 11. Furthermore, there is one polyadenylation signal consensus sequence (A_1 in Fig. 2A,B) in exon 10, and there are seven consensus sequences in exon 11 (A_2–A_8 in Fig. 2A,B). The four species of mRNA seem to correspond to transcripts that are generated by both alternative splicing of the last two exons of *MJD* and by alternative polyadenylation in exon 11, as described below.

The transcript that uses the A_1 polyadenylation signal and terminates at exon 10 is calculated to be 1930 nt long, apart from a poly-A tail. Three cDNA clones (pMJD2-1, pR7b, and H8-2), denoted as type 1 clones in the "Results" section, represent this transcript. Transcripts that terminate at exon 11 can vary in size from 1350 nt to 6887 nt because of alternative polyadenylation. When A_2 is used, the transcript will be 1350 nt long, while A_3, A_4, A_5, A_6, A_7, and A_8 usage, respectively, yields 1860-, 4349-, 4845-, 6065-, 6159-, and 6887-nt-long mRNA species. We isolated a number of cDNA clones that use exon 11 (denoted as type 2 clones). Of these, pMJD5-1 contains the transcript using A_2, while pMJD1-1 consists of the transcript using A_3, and H4-1 and H13-2 consist of the transcript using A_5. Altogether, it seems as though the 1.4-kb band appearing in the Northern blotting work corresponds to the transcript using the A_2 polyadenylation site, while the 1.8-kb band is the transcript using the A_1 and/or A_3 sites. The 4.5-kb band may be generated by the use of the A_5 and, possibly, A_4 sites, while the 7.5-kb band may arise from the use of the A_8 and, possibly, A_6 and A_7 sites.

Other *MJD* cDNA clones suggest that exons other than 10 or 11 may be skipped in transcription. Four identical cDNA clones (H2, H5-1, H7, and H11), denoted as type 3 clones, are characterized by their deletion of exon 2. Despite this, the open reading frame is preserved, and the result is the deletion of 33 amino acid residues. Clones skipping either exon 3 or exon 4 have also been reported previously (Kawaguchi et al. 1994; Paulson et al. 1997). One of the so-called type 4 clones isolated in this study, namely, H6, was also found to have skipped exon 4. Although the skipping of exons 2 or 3 does not shift the reading frame, exon 4 skipping causes a frameshift and results in truncation.

Another type 4 clone, H5-2, was found not to be an *MJD* cDNA, but, rather, to have high homology to *MJD*. This suggests that there may be *MJD*-related gene(s) in the human genome. Supporting this notion is that Kawaguchi et al. (1994) have reported that there are three other homologous genomic loci (8q23, 14q21, and Xp22.1) apart from *MJD* (*MJD1*) on 14q32.1, and indeed, in this study, we obtained several cosmid clones mapping to 14q21. Homology searching also showed that the BAC clones, R-10A2 (AL359212) and R-403E (AL359332), which are mapped to the centromere of 14q, contain partial homology to *MJD* exon 4–9 sequences. With regard to the H5-2 cDNA clone, it bears high homology to a genomic sequence in Xp22. However, this region of Xp22 has no intron, and it seems to be a processed type of pseudogene sequence. Finally, the draft sequence of the BAC clone, RP11-238I10, which is mapped to chromosome 8, has 67.9% homology to the cDNA clone pMJD1-1, except for exon 3. This homologous region of RP11-238I10 also has no intron. These findings suggested that there is a processed pseudogene of *MJD* on chromosome 8.

When the whole sequence of the B445M7 BAC clone was sequenced and subjected to homology searches, several other gene sequences were identified. One was the human homologue of *BTCPSFSU*. The cleavage and polyadenylation specificity factor (CPSF) plays a central role in the 3′ cleavage of transcripts and subsequent polyadenylation (Jenny et al. 1994). The partial sequence of *NDUFB1* was also found in B445M7. *NDUFB1* is one member of the enzyme complex in the electron transport chain of mitochondria and its gene is mapped to chromosome 14 (locus ID, 5707; online mendelian inheritance in man [OMIM], 603837). It is identical to the human CI-MNLL homologue gene expressed in human $CD34^+$ hematopoietic stem/progenitor cells (Mao et al. 1998). Relative to the sequence contained in B445M7, *NDUFB1* can be divided into at least four parts. The genome sequences between its third and fourth parts start with GT and end with AG. These characteristics are consistent with the GT-AG consensus sequences of splice sites.

Our studies revealed the genome structure of *MJD*, and showed that *MJD* coded several transcripts by alternative splicing and polyadenalation. Through the sequencing of the B445M7 BAC clone, it was shown that the homologue of *BTCPSFSU* and *NDUFB1* was located on 14q32.1.

Acknowledgments We thank Dr Y. Misumi for providing the cDNA library; the technical staff of RIKEN Genomic Science Center for assistance and useful suggestions for genome sequencing; and Ms M. Koizumi, Ms K. Matsuba, and Ms N. Tsuji for technical assistance. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, and Culture of Japan, and by a grant from CREST, the Japan Science and Technology Corporation. Y.I. was also supported by research fellowships of the Japan Society for the Promotion of Science for Young Scientists.

References

Banfi S, Servadio A, Chung M-Y, Kwiatkowski TJ Jr, McCall AE, Duvick LA, Shen Y, Roth EJ, Orr HT, Zoghbi HY (1994) Identification and characterization of the gene causing type 1 spinocerebellar ataxia. Nat Genet 7:513–520

Cemal CK, Huxley C, Chamberlain S (1999a) Insertion of expanded CAG trinucleotide repeat motifs into a yeast artificial chromosome containing the human Machado-Joseph disease gene. Gene 236:53–61

Cemal C, Huxley C, McGuigan A, Chamberlain S (1999b) Transgenic animals carrying pathological alleles at the MJD1 locus exhibit a mild and slowly progressive cerebellar deficit (abstract). Am J Hum Genet 65(Suppl):104

David G, Abbas N, Stevanin G, Dürr A, Yvert G, Cancel G, Weber C, Imbert G, Saudou F, Antoniou E, Drabkin H, Gemmill R, Giunti P, Benomar A, Wood N, Ruberg M, Agid Y, Mandel J-L, Brice A (1997) Cloning of the *SCA7* gene reveals a highly unstable CAG repeat expansion. Nat Genet 17:65–70

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–498

Goldberg YP, Nicholson DW, Rasper DM, Kalchman MA, Koide HB, Graham RK, Bromm M, Kazemi-Esfarjani P, Thornberry NA, Vaillancourt JP, Hayden MR (1996) Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. Nat Genet 13:442–449

Goto J, Watanabe M, Ichikawa Y, Yee S-B, Ihara N, Endo K, Igarashi S, Takiyama Y, Gaspar C, Maciel P, Tsuji S, Rouleau GA, Kanazawa I (1997) Machado-Joseph disease gene products carrying different carboxyl termini. Neurosci Res 28:373–377

Hattori M, Tsukahara F, Furuhata Y, Tanahashi H, Hirose M, Saito M, Tsukuni S, Sakaki Y (1997) A novel method for making nested deletions and its application for sequencing of a 300 kb region of human APP locus. Nucleic Acids Res 25:1802–1808

Hirai M, Kusuda J, Hashimoto K (1996) Assignment of ADP ribosylation factor (ARF) genes *ARF1* and *ARF3* to chromosome 1q42 and 12q13, respectively. Genomics 34:263–265

Holmes SE, O'Hearn EE, Mclnnis MG, Gorelick-Feldman DA, Kleiderlein JJ, Callahan C, Kwak NG, Ingersoll-Ashworth RG, Sherr M, Sumner A., Sharp AH, Ananth U, Seltzer WK, Boss MA, Vieria-Saecker AM, Epplen JT, Riess O, Ross CA, Margolis RL (1999) Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. Nat Genet 23:391–392

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

Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier J-M, Weber C, Mandel J-L, Cancel G, Abbas N, Dürr 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

Jenny A, Hauri HP, Keller W (1994) Charcterization of cleavage and polyadenylation specificity factor and cloning of its 100-kilodalton subunit. Mol Cell Biol 14:8183–8190

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) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet 8:221–228

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 6:9–13

Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15:8125–8148

Lafrenière RG, DeJong PJ, Rouleau GA (1995) A 405-kb cosmid contig and *Hin*dIII restriction map of the progressive myoclonus epilepsy type 1 (EPM1) candidate region in 21q22.3. Genomics 29:288–290

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

Lin B, Rommens JM, Graham RK, Kalchman M, MacDonald H, Nasir J, Delaney A, Goldberg YP, Hayden MR (1993) Differential 3′polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. Hum Mol Genet 2:1541–1545

Mann SM, Burkin DJ, Grin DK, Ferguson-Smith MA (1997) A fast, novel approach for DNA fibre-fluorescence in situ hybridization analysis. Chromosome Res 5:145–147

Mao M, Fu G, Wu J-S, Zhang Q-H, Zhou J, Kan L-X, Huang Q-H, He K-L, Gu B-W, Han Z-G, Shen Y, Gu J, Yu Y-P, Xu S-H, Wang Y-X, Chen S-J, Chen Z (1998) Identification of genes expressed in human CD34+ hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. Proc Natl Acad Sci USA 95:8175–8180

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, Takahashi M, Yoshizawa T, Kanazawa I, Yamada M (1994) Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosone 12p. Nat Genet 6:14–18

Nishiyama K, Murayama S, Goto J, Watanabe M, Hashida H, Katayama S, Nomura Y, Nakamura S, Kanazawa I (1996) Regional and cellular expression of the Machado-Joseph disease gene in brains of normal and affected individuals. Ann Neurol 40:776–781

Onodera O, Oyake M, Takano H, Ikeuchi T, Igarashi S, Tsuji S (1995) Molecular cloning of a full-length cDNA for dentatorubral-pallidoluysian atrophy and regional expressions of the expanded alleles in the CNS. Am J Hum Genet 57:1050–1060

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

Paulson HL, Das SS, Crino PB, Perez MK, Patel SC, Gotsdiner D, Fischbeck KH, Pittman RN (1997) Machado-Joseph disease gene product is a cytoplasmic protein widely expressed in brain. Ann Neurol 41:453–462

Pulst S-M, Nechiporuk A, Nechiporuk T, Gispert S, Chen X-N, Lopes-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

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Sanpei 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

Strong TV, Tagle DA, Valdes JM, Elmer LW, Boehm K, Swaroop M, Kaatz KW, Collins FS, Albin RL (1993) Widespread expression of the human and rat Huntington's disease gene in brain and nonneural tissues. Nat Genet 5:259–265

The 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

Xie Y-G, Rochefort D, Brais B, Howard H, Han F-Y, Gou L-P, Maciel P, The BT, Larsson C, Rouleau GA (1998) Restriction map of a YAC and cosmid contig encompassing the oculopharyngeal muscular dystrophy candidate region on chromosome 14q11.2–q13. Genomics 52:201–204

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 α_{1A}-voltage-dependent calcium channel. Nat Genet 15:62–69