Somatic Cell and Molecular Genetics, Vol. 20, No. 2, 1994, pp. 87–97

Mouse Huntington's Disease Gene Homolog (*Hdh*)

Glenn T. Barnes, Mabel P. Duyao, Christine M. Ambrose, Sandra McNeil, Francesca Persichetti, Jayalakshmi Srinidhi, James F. Gusella and Marcy E. MacDonald

Molecular Neurogenetics Unit, Massachusetts General Hospital, Charlestown, Massachusetts 02129; and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114

Received 28 February 1994

Abstract—*The incurable neurodegenerative disorder, Huntington's disease (HD), is caused by an expanded, unstable CAG repeat encoding a stretch of polyglutamine in a 4p16.3 gene (*HD*) of unknown function. Near the CAG repeat is a polyproline-encoding CCG repeat that shows more limited allelic variation. The mouse homologue,* Hdh, *has been mapped to chromosome 5, in a region devoid of mutations causing any comparable phenotype. We have isolated overlapping cDNAs from the* Hdh *gene and compared their sequences with the human transcript. The consensus mouse coding sequence is 86% identical to the human at the DNA level and 91% identical at the protein level. Despite the overall high level of conservation,* Hdh *possesses an imperfect CAG repeat encoding only seven consecutive glutamines, compared to the 13–36 residues that are normal in man. Although no evidence for polymorphic variation of the CAG repeat was seen, a nearby CCG repeat differed in length by one unit between several strains of laboratory mouse and* Mus spretus. *The absence of a long CAG repeat in the mouse is consistent with the lack of a spontaneous mouse model of HD. The information presented concerning the sequence of the mouse gene should facilitate attempts to create such a model.*

INTRODUCTION

Huntington's disease is a dominant neurodegenerative disorder involving a premature loss of specific neurons that is most extensive in the striatum (1). The hallmark of the disorder is a characteristic chorea that begins insidiously and progresses until it consumes all parts of the body. Accompanying or even sometimes preceding the motor disturbance are psychiatric changes and gradual cognitive decline. The onset of HD typically occurs in middle age (mean ~40 years), but a minor proportion of cases begin in childhood, or late in life. Family studies using DNA polymorphisms mapped the *HD* gene near the tip of the chromosome 4 short arm, creating the possibility of identifying the genetic defect gene by a location cloning approach (2–4). We have recently demonstrated that HD is caused by an expanded, unstable CAG repeat (5).

The CAG repeat is polymorphic on normal chromosomes, varying from approximately 11 to 34 units (6–8). It is found near the 5′ end of a novel gene, next to a less polymorphic CCG repeat that varies from 6 to 12 units (9). Both repeats are located within the predicted coding sequence, where they encode stretches of polyglutamine and polyproline, respectively, near the *N*-terminus of the gene product, named huntingtin (5). On HD chromosomes, the CAG repeat is expanded to a range of 37–100 or more

0740-7750/94/0300-0087$07.00/0

units. In this expanded state, the repeat segment is unstable, with more than 80% of meiotic transmissions involving a change in the number of CAGs (6). The greatest size changes are seen in paternal transmissions. The length instability appears to occur during gametogenesis, without leading to appreciable somatic mosaicism (10). Age of onset of the disease is inversely correlated with the length of the CAG repeat, with the juvenile onset cases, which most often result from paternal transmission, displaying the longest repeats (6–8).

We have mapped *Hdh,* the mouse homologue of the *HD* gene, to the proximal portion of mouse chromosome 5, in a region containing no known mutants with phenotypes similar to the human disease (11). To set the stage for creation of a mouse model of HD, we have isolated and sequenced cDNAs for the murine gene. Our results indicate that the gene is highly conserved, with mouse and human huntingtin being 91% identical. However, the two differ in the crucial *N*-terminus, where the mouse protein contains a much shorter stretch of seven glutamine residues, encoded by an interrupted CAG repeat. A nearby CCG repeat is also present in the mouse, where it provides a polymorphic marker for analysis of interspecies crosses. The relatively short CAG segment in the mouse gene is consistent with the absence of any naturally occurring HD model in this species. However, the extent of the conservation of the remainder of mouse huntingtin suggests that engineering of an expanded CAG repeat at the mouse locus might mimic the effects of the HD mutation.

MATERIALS AND METHODS

Hybridization. Northern blots (Clontech polyA+) were hybridized with ^{32}P-labeled (12) human cDNA clone IT15B (5). 1 spanning nt 5345–10366 of the composite IT15 cDNA sequence (GenBank L12392). Hybridization conditions were: 50% formamide, 10% dextran sulfate, 0.8 M NaCl, 5× Denhardt's, 50 mM Tris pH 7.5, 0.5% SDS, 100 μg/ml sheared single-stranded fish DNA and 0.1% sodium pyrophosphate. Filters were hybridized for 48 h at 42°C, then washed in 0.5× SSC, 0.1% SDS at 65°C.

The PCC4 embryonal carcinoma phage cDNA library (Stratagene) was hybridized with a pool of ^{32}P-labeled (12) human PCR and cDNA probes representing nt 933–1899 and 3028–10366. The 129 genomic phage library was screened similarly using a pair of probes flanking the CAG and CCG repeats prepared by PCR amplification from PCC4-8. The following primer pairs were used to amplify segments 5′ and 3′ to the repeats, respectively: primer set 1, 5′-GAAAAGCTGATGAAGGCT-3′ and 5′-CTGCTGAAACGACTTGAG-3′; primer set 2, 5′-CACCGCCGCTGCCAGGTC-3′ and 5′-GGTCGGTGCAGCGGTTCC-3′. Hybridization and washing were performed as above except 40% formamide, 1 M NaCl, and 1× Denhardt's were used and washing was at room temperature.

DNA Sequencing. Double-stranded cDNA clones (1 μg), the 129-1 genomic phage clone (40 μg), and six pBSKII subclones (1 μg) of PCR product from *M. spretus* were sequenced by dideoxy chain termination (13) using custom primers (Biosearch Cyclone) and 7-deazaguanosine/sequenase 2.0 (USB). Sequence comparisons with the human gene were performed using the GCG package (14).

Polymorphism Analysis. Genomic mouse DNA (200 ng) was amplified using the following primers flanking the CAG-CCG repeat region: 5′-CTGATGAAGGCTTTCGAGTCGCTCAAGTCG-3′ and 5′-CCTTCTTTGGTCGGTGCAGCGGTTCCTCTG-3′. Reaction conditions were 200 μM dNTPs, 10% DMSO, 1 μCi [^{32}P]dATP, 2 units Taq polymerase (Boehringer/Mannheim), and the buffer supplied by the manufacturer. The cycling program

was: 1× 2′ min at 94°C; 30× (1 min at 94°C; 1 min at 60°C; 1 min at 72°C). Labeled PCR products were displayed on 6% denaturing polyacrylamide gels.

For subcloning of the *M. spretus* PCR product, the above primers were resynthesized with (CUA)4 and (CAU)4, respectively, on the 5′ ends. After amplification, the product was cut from low-melt agarose and subcloned using uracil DNA glycosylase (UDG) (Gibco/BRL) into pBSKII modified by digestion with EcoRV and PCR amplification using primers 5′-AGUAGUAGUAGAUCAAGCTTATCGATACC-3′ and 5′-AUGAUGAUGAUGAUCGAATTCCTGCAGCC-3′.

RESULTS

Northern Blot Analysis of Hdh *Expression.* The *HD* gene is expressed in all human tissues tested to date as two different mRNAs of 10.5 and 13.5 kb that encode the same huntingtin protein, but differ in their 3′ untranslated regions (UTRs) due to alternative polyadenylation (15, 16). In DNA blot analyses used to map the *Hdh* locus on chromosome 5, we determined that the mouse gene is sufficiently conserved to be easily detected using a human probe. Figure 1 displays the results of hybridizing a human *HD* probe to Northern blots containing poly(A)$^+$ RNA from a variety of mouse tissues. The pattern of expression is remarkably similar to the expression of HD in man, with two different RNAs also of 10.5 and 13.5 kb. These RNAs are expressed in all tissues tested, but at varying ratios. As in man, mouse brain RNA displays the highest proportion of the larger transcript. A novel band of variable intensity is also seen on the mouse northern blots at ~7 kb. This signal is removed by stringent washing suggesting the possibility of a related locus.

Isolation of Overlapping cDNA Clones for Hdh. To permit direct comparison of the human and mouse homologes, we screened a mouse PCC4 embryonal carcinoma cDNA

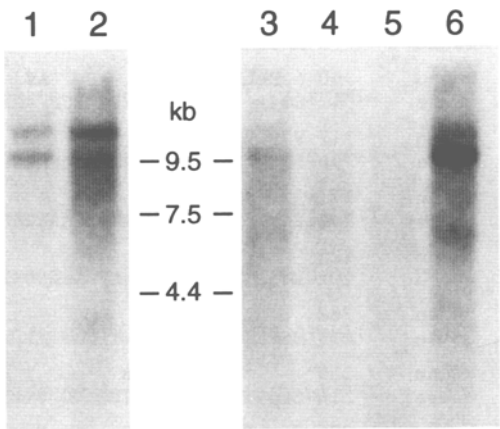

Fig. 1. Northern blots analysis of mouse *Hdh* mRNAs. Northern blots containing 2 μg of poly(A)$^+$ mRNA from various adult mouse tissues were hybridized with human IT15B.1 Transcript sizes were estimated from RNA size markers as shown. Lanes: 1, heart; 2, brain; 3, liver; 4, skeletal muscle; 5, kidney; 6, testis.

library with a pool of cDNA and PCR probes spanning almost the entire published composite IT15 sequence of 10,366 bp (GenBank #L12392) to isolate overlapping clones representing the *Hdh* mRNAs. A summary of the cDNAs obtained is displayed in Fig. 2. Three cDNA clones provided overlapping sequence coverage from a few bases 3′ to the initiator ATG codon through the entire coding sequence. Clone PCC4-3 also possessed a

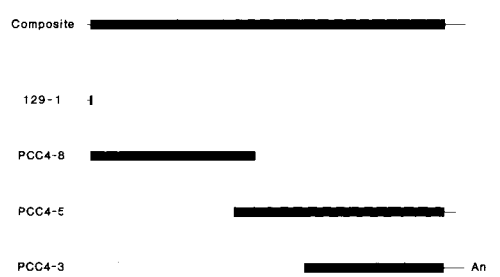

Fig. 2. Schematic diagram of mouse *Hdh* clones. The composite mouse Hdh cDNA sequence deposited in GenBank as accession #L28827 is shown schematically over the clones from which it was derived. The 5′ UTR and 3′ UTR sequences are shown as thin lines, while the predicted coding sequence is depicted as a filled box. The sequences provided by each clone are: 129-1 genomic phage, nt 1–133; cDNA PCC4-8, nt 102–4469; PCC4-5, nt 3906–9765; and PCC4-3, nt 5781–9998. Only the latter clone displayed a poly(A) tail.

```
human     1                    QQQQQQQQQQQQQ        QQ P         -  L          59
mouse     1 MATLEKLMKAFESLKSF--------------QQQQQQQPPPQPPPPPPPPPPQPPQPPPQ      46

human    60 A PLLPQPQP        P   AV                                           116
mouse    47 GQ--------PPPPPPPLPGP--AEEPLHRP 1 KKELSATKKDRVNHCLTICENIVAQS      93

human   117 V                                                                 170
mouse    94 LR 2 NSPEFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIK 3 ALMDSNLPRLQL     147

human   171                                                                   224
mouse   148 ELYKEIKK 4 NGAPRSLRAALWRFAELAHLVRPQKCR 5 PYLVNLLPCLTRTSKRPEE     201

human   225                                                  I                281
mouse   202 SVQETLAAAVPKIMASFGNFANDNEIKV 6 LLKAFIANLKSSSPTVRRTAAGSAVSICQ     258

human   282          S               V D                Y                     338
mouse   259 HSRRTQYFYNWLLNVLLG 7 LLVPMEEEHSTLLILGVLLTLRCLVPLLQQQVKDTSLKG     315

human   339               A                                                   395
mouse   316 SFGVTRKEMEVSPSTEQLVQ 8 VYELTLHHTQHQDHNVVTGALELLQQLFRTPPPELLQ     372

human   396 T  AV  I    AAK  SG  S         I                                 448
mouse   373 ALTTPGGLGQLTLVQEEARGRGRSGSIVELLA 9 GGGSSCSPVLSRKQKG 10 KVLLG     425

human   449                      LT        D  S                A             504
mouse   426 EEEALEDDSESRSDVSSSAFAA 11 SVKSEIGGELAASSGVSTPGSVGHDIITEQPRSQ     481

human   505           AS     S                  V                             564
mouse   482 HTLQADSVDLSGCDLTSAATDGDEEDILSHSSSQFSAVPSDPAMDLNDGTQASSPISDSS     541

human   565                              T N    L      DE   - T I PD A EA     619
mouse   542 QTTTEGPDSAVTPSDSSEI 12 VLDGADSQYLGMQIGQPQEDDEEGAAGVLSGEVSDVF     597

human   620     M            KN S C        V  F L   AT PG Q N                671
mouse   598 RNSSLA 13 LQQAHLLERMGHSRQPSDSSIDKYVTRDEVAEASDPESK 14 PCRIKGD     649

human   672    ST                        G NV                        V       727
mouse   650 IGQPNDDDSAPLVHCVRLLSASFLLTGEKKA 15 LVPDRDVRVSVKALALSCIGAAVAL     705

human   728         K       D   YP                                      IC    783
mouse   706 HPESFFSRLYKVPLNTTESTE 16 EQYVSDILNYIDHGDPQVRGATAILCGTLVYSILS     761

human   784     FH  D M T                 A       R                        N  835
mouse   762 RSRLRVGEWLGNIRTLTG 17 NTFSLVDCIPLLQKTLKDESSVTCKLACTAVR 18 HC     813

human   836  M         E      I  V T R              E                       N  891
mouse   814 VLSLCSSSYSDLGLQLLIDMLPLKNSSYWLVRTELLDTLAEIDFR 19 LVSFLEAKAES     869

human   892                L              H                A  I              943
mouse   870 LHRGAHHYTG 20 FLKLQERVLNNVVIYLLGDEDPRVRHVAATSLTR 21 LVPKLCDQ     921

human   944                                                  .     N           999
mouse   922 GQAFYKDPVVAVARDQSSVYLKLLMHETQPPSHFSVSTITR 22 IYRGYSLLPSITDVT     977

human  1000          I                                      T     I           1051
mouse   978 MENNLSRVVAAVSHELITSTTRALT 23 FGCCEALCLLSAAFPVCTWSLGWHCG 24 V     1029

human  1052                    T                                              1107
mouse  1030 PPLSASDESRKSCTVGMASMILTLLSSAWFPLDLSAHQDALILAGNLLAA 25 SAPKSL     1085

human  1108      A       P   K   V        A   M                        A      1167
mouse  1086 RSSWTSEEEANSAATRQEEIWPALGDRTLVPLVEQLFSHLLKVINICAHVLDDVTPGPAI     1145

human  1168                                      V L    GS                    1219
mouse  1146 K 26 AALPSLTNPPSLSPIRRKGKEKEPGEQASTPMSPKKVGEASAA 27 SRQSDTSG     1197
```

Fig. 3. Comparison of the predicted sequences of human and mouse huntingtin. The mouse amino acid sequence for huntingtin is shown in full. For the human huntingtin sequence, only residues different from the mouse are shown. Dashes represent gaps in the matching of the two sequences. The sequences are numbered with the predicted initiation Met codon of each protein as #1. The human sequence is that containing 21 CAG repeat units (5) deposited in GenBnk #L12392. Numbers 1–66 inserted within the protein sequence indicate the positions of introns in the human *HD* gene (15).

```
human  1220    T                                                          1275
mouse  1198 PVTASKSSSLGSFYHLPSYLKLHDVLKATHANYK 28 VTLDLQNSTEKFGGFLRSALDV 1253

human  1276                          I                                   1327
mouse  1254 LSQILELATLQDIGK 29 CVEEVLGYLKSCFSREPMMATVCVQQ 30 LLKTLFGTNLA 1305

human  1328               G                                          N  1387
mouse  1306 SQFDGLSSNPSKSQCRAQRLGSSSVRPGLYHYCFMAPYTHFTQALADASLRNMVQAEQER 1365

human  1388  T              T                                           1439
mouse  1366 DASG 31 WFDVLQKVSAQLKTNLTSVTKNRAD 32 KNAIHNHIRLFEPLVIKALKQYT 1417

human  1440    C                                                         1491
mouse  1318 TTTSVQLQKQVLDLLAQLVQLRVNYCLLDSDQ 33 VFIGFVLKQFEYIEVGQFR 34 E 1469

human  1492                                                              1547
mouse  1470 SEAIIPNIFFFLVLLSYERYHSKQIIGIPKIIQLCDGIMASGRKAVTHA 35 IPALQPI 1525

human  1548                                                              1603
mouse  1526 VHDLFVLRGTNKADAGKELETQKEVVVSMLLRLIQYHQ 36 VLEMFILVLQQCHKENED 1581

human  1604         I                                                    1659
mouse  1582 KWKRLSRQVADIILPMLAKQQ 37 MHIDSHEALGVLNTLFEILAPSSLRPVDMLLRSMF 1637

human  1660 V  N                                        S         Y I  T  1715
mouse  1638 ITPSTM 38 ASVSTVQLWISGILAILRVLISQSTEDIVLCRIQELSFSPHLLSCPVINR 1693

human  1716   D DSTS  E H     I N   E                                 E  1771
mouse  1694 LRGGGGNVTLGECSEGKQ-KSLPEDTFSR 39 FLLQLVGILLEDIVTKQLKVDMSEQQH 1748

human  1772                                          R    G      D   L A  1827
mouse  1749 TFYCQELGTLLMCLIHIFKSG 40 MFRRITAAATRLFTSDGCEGSFYTLESLNARVRSM 1804

human  1828 IT                V    Y                   S  L S  M       D  1883
mouse  1805 VPTHPALVLLWCQILLLINHTDHRWWAEVQQTPK 41 RHSLSCTKSLNPQKSGEEEDSG 1860

human  1884 L  K                                                        1939
mouse  1860 SAAQLGMCNREIVRRGALILFCDYV 42 CQNLHDSEHLTWLIVNHIQDLISLSHEPPVQ 1916

human  1940      V                              M                       1995
mouse  1917 DFISAIHRNSAASGLFIQAIQSRCENLST 43 PTTLKKTLQCLEGIHLSQSGAVLTLYV 1972

human  1996     C    V      I                                M        Y  S  2051
mouse  1973 DRLLGTPFRALARMVDTLACRRVEMLLAANLQ 44 SSMAQLPEEELNRIQEHLQNSGLA 2028

human  2052                        M      S   S         V                V  2103
mouse  2029 QR 45 HQRLYSLLDRFRLSTVQDSLSPLPPVTSHPLDGDGHTSLETVSPDK 46 DWYL 2080

human  2104 H  K                                A   N                    S  2159
mouse  2081 QLVRSQCWTRSDSALLEGAELVNRIPAEDMNDFMMSS 47 EFNLSLLAPCLSLGMSEIA 2136

human  2160 G    A       E T A  SGT        H    E  A  A   S     F       AA  2215
mouse  2137 NGQKSPLFEAARGVILNRVTSVVQQLPAVHQVFQPFLPIEPTAYWNKLNDLLG 48 DTT 2192

human  2216 L    PT           V  L S         K I     A L                  2271
mouse  2193 SYQSLTILARALAQYLVVLSKVPAHLHLPPEKEGDTVKFVVMTVE 49 ALSWHLIHEQI 2248

human  2272                       L    S V  T F        Y  H     V         E  2327
mouse  2249 PLSLDLQAGLDCCCLALQVPGLWGVLSSPEYVTHACSLIHCVRFILEAI 50 AVQPGDQ 2304

human  2328   S  R TN  K ISEEE     PNT       PKYI A       E                 2383
mouse  2305 LLGPESRSHTPRAVR--KEEVDSDIQN 51 LSHVTSACEMVADMVESLQSVLALGHKRN 2358

human  2384  GV A   PL R  I                                              A   2439
mouse  2359 STLPSFLTAVLKNIVISLARLPLVNSYTRVPPL 52 VWKLGWSPKPGGDFGTVFPEIPV 2414
```

Fig. 3. (Continued)

poly(A) tail attached to a 3′ UTR similar in length to that of the shorter of the two human transcripts. Although the larger mouse *Hdh* transcript, like that in man, is probably generated by alternative polyadenylation, we did not obtain any cDNA clones to confirm this supposition.

The 5′ UTR and the first few bases of

```
human  2440        VF                    S                                    2493
mouse  2415 EFLQEKEILKEFIYRINTLG 53 WTNRTQFEETWATLLGVLVTQPLVMEQEESPPEE 54   2468

human  2494          N                                                       2549
mouse  2469 DTERTQIHVLAVQAITSLVLSAMTVPVAGNPAVSCLEQQPRNKPLKALDTR 55 FGRKL   2524

human  2550  I           A   K   I    LY          S              E         2605
mouse  2525 SMIRGIVEQEIQEMVSQRENTATHHSHQAWDPVPSLLPATTG 56 ALISHDKLLLQINP   2580

human  2606    L S                        S                A A   S   T    2661
mouse  2581 EREPGNMSYKLGQ 57 VSIHSVWLGNNITPLREEEWDEEEEEESDVPAPTSPPVSPVNS   2636

human  2662                                    S     A                       2712
mouse  2637 R 58 KHRAGVDIHSCSQFLLELYSRWILPSSAARRTPVILISEVVRS 59  LLVVSDL   2687

human  2713      N   L  V                 A                        A        2768
mouse  2688 FTERTQFEMMYLTLTELRRVHPSEDEILIQYLVPATCKAAAVLGM 60 DKTVAEPVSRL   2743

human  2769              RV     V                      I                    2824
mouse  2744 LESTLRSSHLPSQIGALHGILYVLECDLLDDTAKQLIPVVSDYLLSNLKGIAH 61 CVN   2799

human  2825                    I                 I                         2880
mouse  2800 IHSQQHVLVMCATAFYLMENYPLDVGPEFSASVIQ 62 MCGVMLSGSEESTPSIIYHCA   2855

human  2881                    A                 H                      V   2936
mouse  2856 LRGLERLLLSEQLSRLDTESLVKLSVDRVNVQSPHRAMAALGLMLTCMYTG 63 KEKAS   2911

human  2937    T   N  A                                                     2992
mouse  2912 PGRASDPSPATPDSESVIVAMERVSVLFDR 64 IRKGFPCEARVVARILPQFLDDFFPP   2967

human  2993   I                                          T                  3048
mouse  2968 QDVMNKVIGEFLSNQQPYPQFMATVVYK 65 VFQTLHSAGQSSMVRDWVMLSLSNFTQR   3023

human  3049 A      T       F          A                                    3104
mouse  2924 TPVAMAMWSLSCFLVSASTSPWVSAI 66 LPHVISRMGKLEQVDVNLFCLVATDFYRHQ   3079

human  3105      L         L               T  R                            3144
mouse  3080 IEEEFDRRAFQSVFEVVAAPGSPYHRLLACLQNVHKVTTC*                       3119
```

Fig. 3. (Continued)

coding sequence were not recovered in any cDNA clones. To obtain these sequences, we isolated a genomic clone by screening a 129-phage library with probes from the 5′ end of PCC4-8. Direct sequencing of the phage insert provided *Hdh* sequence from 89 bp upstream of the initiator ATG through the first 44 bases of the coding sequence.

Composite Hdh *cDNA and Huntingtin Sequences.* The composite DNA sequence generated from the clones shown in Figure 2 spans 9998 nucleotides and has been deposited in GenBank, with accession #L28827. The putative initiator ATG codon at nt 90 and the TGA stop codon at nt 9447 bracket an open reading frame that predicts a mouse huntingtin protein of 3119 amino acids (Fig. 3). Like human huntingtin, mouse huntingtin has a region with stretches of polyglutamine and polyproline near its *N*-terminus (see below). Across the coding sequence 5′ (nt 90–143) and 3′ (nt 267–9446) to the glutamine/proline-rich region, respectively, the DNA sequence is 90% and 86% identical to the human cDNA (not shown as a figure). In the 89 nt of 5′ UTR, identity to the human sequence declines to 67%, with the mouse sequence having an insert of seven bases 48 nt upstream from the ATG (Fig. 4A). The 552 nt of 3′ UTR is less conserved overall (64%) with many gaps required to match the sequences from the two species (Fig. 4B). Interestingly, however, the 75-nt segment immediately upstream from the site of poly(A) addition shows 90% identity in man and mouse, perhaps indicative of a common structural basis for alternative polyadenylation at this site.

At the protein level, human and mouse huntingtin are 100% and 91% identical *N*-terminal and *C*-terminal to the glutamine/proline-rich region, respectively. The mouse

protein is shorter than the human protein, owing largely to a smaller glutamine/proline-rich region (see below). The remainder of the mouse protein displays 238 conservative amino acid substitutions, 29 nonconservative substitutions, 5 residue deletions, and 1 residue addition relative to its human counterpart (Fig. 3).

Critical features of DNA encoding the glutamine/proline-rich region of human huntingtin are the polymorphic CAG repeat that is expanded on disease chromosomes and the adjacent polymorphic CCG repeat. In Fig. 4A, we have compared the human and mouse sequences across the glutamine/proline-rich region. In man, the polyglutamine stretch varies from 13 to 36 residues and is encoded almost entirely by CAG except for a penultimate CAA codon. The mouse gene encodes seven consecutive glutamines in an imperfect repeat with a CAA codon flanked on the 5′ and 3′ sides by two and four CAG codons, respectively. In both species, the glutamine stretch is followed by a segment with runs of proline with the occasional glutamine or other amino acid residue interspersed. In man, the CCG repeat located just downstream from the polymorphic CAG repeat is also polymorphic (9).

Polymorphisms of a CCG Repeat in Hdh. To determine whether the repeat sequences displayed polymorphism in the mouse comparable to that on human chromosomes, we amplified the corresponding region from various strains of laboratory mouse and from *M. spretus.* A typical result is shown in Fig. 5, in which 129 (represented by clone PCC4-8), C57BL/6J, and CBA/J all yield an identical product. The shorter product generated from *M. spretus* was sequenced for comparison with PCC4-8. The difference in length is not due to any change in CAG number, but rather to a decrease of one CCG in the *M. spretus* (see Fig. 4A). Thus, the CAG repeat is not only shorter in mouse than in man, it also does not display any evidence of significant length variation.

DISCUSSION

The mouse *Hdh* gene is located on chromosome 5, in a region of synteny conservation with human chromosome 4 (11, 17). Although the genomic structure of *Hdh* is unknown, the human *HD* gene contains 67 exons, spread across 180 kb of 4p16.3 (15). The human and mouse genes are extremely similar, showing an overall amino acid identity of more than 90% over most of the predicted protein. A comparison of the differences (Fig. 3) indicates that they are not confined to alterations affecting a few exons, but are found throughout the gene. However, neither are they equally distributed. The regions encompassing amino acids 373–403, 567–641, 1684–1717, and 2136–2374 seem particularly rich in amino acid substitutions in contrast to segments such as residues 60–372 and 1190–1637. The latter might indicate the locations of critical functional domains of huntingtin, but without any reasonable clue as to the function of the protein, this is impossible to evaluate. More extensive differences are found in the DNA of the 5′ and particularly the 3′ UTRs, suggesting less stringent selective pressures overall on these sequences. However, the relatively high level of conservation of the DNA sequence immediately upstream from one site of poly(A) addition in man indicates that this segment may be involved in regulating alternative polyadenylation (Fig. 4B). Although the existence of alternative polyadenylation has not been demonstrated unequivocally in the mouse, the ubiquitous expression of two *Hdh* RNAs comparable in size to the human *HD* RNAs supports this supposition.

While the work reported here was in its final stages, Lin et al. (18) reported a cDNA sequence for mouse *Hdh* spanning 9992 bp, also encoding a huntingtin protein of 3119 residues. However, the composite cDNA sequence reported here has notable differences. These are best compared at the

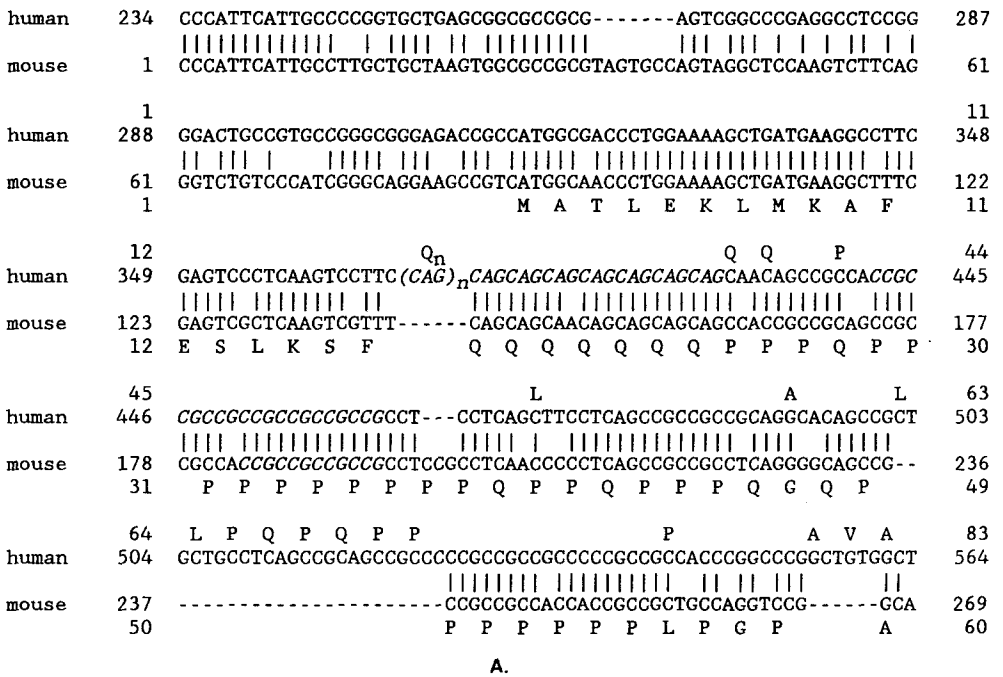

Fig. 4. Comparison of human and mouse 5′ and 3′ DNA sequences. (A) The 5′ UTR and CAG-CCG region. The human and mouse DNA sequences across the 5′ UTR and CAG-CCG-rich region were aligned, and gapped (dashes) to maximize similarity. Vertical lines join identical bases. Bases shown in italic indicate the regions subject to polymorphism in each species. $(CAG)_n$ (encoding Q_n) indicates the polymorphic repeat segment. Numbering is based on 21 CAG repeats as in Fig. 2. The mouse protein sequence, beginning with the initiator Met as codon 1 is shown below the DNA sequence. Mismatches in the human protein sequence are shown above. (B) The 3′ UTR. The human and mouse 3′ UTR DNA sequences were aligned and displayed as above. Both sequences were halted at the site of poly(A) addition.

protein level, where the two sequences differ at 28 residues spread across the entire protein, from position 2 to position 3096. In 24 of these cases, our mouse huntingtin sequence matches the amino acid sequence found in man. In the remaining four cases, Lin et al. (18) matches the human sequence. Our 3′ UTR sequence has seven mismatches, additions, or deletions of single bases compared to Lin et al. (18). In addition, both of our clones PCC4-3 and PCC4-5 contain a stretch of 35 bp not present in their sequence. These investigators also found that in their cDNA clones the CCG repeat beginning at codon 32 varies between C57BL6 and random outbred laboratory mice, displaying three and four repeat units, respectively. While our amplification of genomic DNA agrees with the site of this polymorphism, it does not yield the same strain-specific pattern. In our experiments, all three strains of laboratory mice, including 129, C57BL/6J, and CBA/J, possessed four CCG units, while only *M. spretus* revealed three CCGs. Finally, Lin et al. (18) reported the identification of a putative alternative splicing event that removed nt 4652–6091, and therefore amino acids 1522–2001 from the protein. An examination of Figure 2 reveals that this segment begins in exon 35 and ends in exon 44 of the human gene. Thus, unless the exon structure of the mouse gene differs radically from that of the human gene, the clone isolated by Lin et al. (18) cannot be explained by simple alternative splicing. Indeed, these authors suggest that the same variant sequence is also

```
human   9748  ----TGAGCGCCATGGTGGGAGAGACTG---TGAGGCGGCAGCTGGGGCCGGAGCCTTTG  9800
mouse   9457  TGAGTAGTGCCTGTGGGACAAAAGGCTGAAAGAAGGCAGCTGCTGGGGCCTGAGCCTCCA  9507

human   9801  GAAGTCTGTGCCCTTGTGCCCTGCCTCCACCGAGCCAGCTTGGTCCCTATGGGCTTC-CG  9859
mouse   9508  GGAGCCTG--------CTCCAAGCTTCTGCTGGGGCTGCCTTGGCCGTGCAGGCTTCACT  9558

human   9860  CACATGCCGCGGGCGGCCAGGCAACGTGCGTGTCTCTGCCATGTGGCAGAAGTGCTCTTT  9919
mouse   9559  TGTGTCAAGTGGACAGCCAGGCAA----------------TGGCAGGAGTG--CTTT  9597

human   9920  GTGGCAGTGGCCAGGCAGGGAGTGTCTGCAGTCCTGGTGGGGCTGAGCCTGAGGCCTTCC  9979
mouse   9598  GCAATGAGGGCTATGCAGGGAACATGCAC---TATGTTGGGGTTGAGCCTGAGTCCTGGG  9654

human   9980  AGAAAGCAGGAGCAGCTGTGCTGCACCCCATGTGGGTGACCAGGT-----CCTTTCTCCT  10034
mouse   9655  TCCTGGC-CTCGCTGCAGCTGGTGACAGTGCTAGGTTGACCAGGTGTTTGTCTTTTTCCT  9713

human   10035 GATAGTCACCTGCTGGTTGTTGCCAGGTTGCAGCTGCTCTTGCATCTGGGCCAGAAGTCC  10094
mouse   9714  AGTGTTCCCCTGGCCATAGTCGCCAGGTTGCAGCTGCCCTGGTATGTGGATCAGAAGTCC  9773

human   10095 TCCCTCCTGCAGGCTGGCTGTTGGCCCCTCTGCTGTCCTGCAGTAGAAGGTGCCGTGAGC  10154
mouse   9774  TAGCTCCTGCCAGATGGTTCTGAGCCNGCCTGCTCCACTGGGCTGGAGAGCTCCCT----  9829

human   10155 AGGCTTTGGGAACACTGGCCTGGGTCTCCCTGGTGGGGTGTGCATGCCACGCCCCGTGTC  10214
mouse   9830  ----------CCCACATTTACCCAGTAGGCATACCTGCCAC-ACCAGTGTC  9869

human   10215 TGGATGCACAGATGCCATGGCCTGTGCTGGGCCAGTGGCTGGGGGTGCTAGACACCCGGC  10274
mouse   9870  TGGACACAAA-----TGAATGGTGTGTGGGGCTGGGAACTGGGGCTGCCAGGTGTCCAGC  9924

human   10275 ACCATTCTCCCTTCTCTCTTTTCTTCTCAGGATTTAAAATTTAATTATATCAGTAAAGAG  10334
mouse   9925  ACCATTTTCCTTTCTGTGTTTTCTTCTCAGGAGTTAAAATTTAATTATATCAGTAAAGAG  9984

human   10335 ATTAATTTTAACGT  10348
mouse   9985  ATTAATTTTAATGT  9998
```

B.

Fig. 4. (Continued)

expressed in man. For this to occur would require a complex change, including the recognition of a segment in exon 35 as a splice donor, the use of a different sequence in exon 44 as a splice acceptor, and the bypassing of conventional splicing signals in exons 36–44. Alternatively, the sequence reported by Lin et al. could have resulted from a cloning artifact.

Human huntingtin is predicted to be a large protein of greater than ~3130 amino acids that does not display significant homology to any known protein. The high level of conservation of mouse huntingtin (91% identity) suggests that there are tight evolutionary constraints on its sequence. However, the absence of recognizable conserved motifs, except for a questionably significant leucine zipper motif at position 1424, makes it hard to predict a function for huntingtin. The protein has yet to be isolated, and its *N*-terminus determined with certainty, but

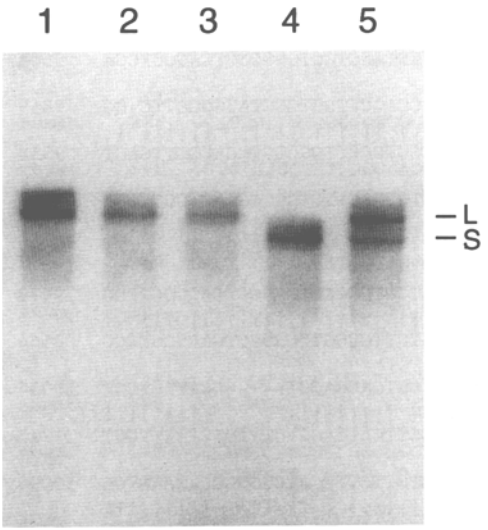

Fig. 5. CCG polymorphism in *Mus spretus*. Mouse genomic or cloned DNA was amplified using PCR primers flanking the CAG-CCG-rich region near the 5′ end of the *Hdh* gene. Products were displayed on a 6% denaturing polyacrylamide gel. Lanes: 1: PCC4-8 cDNA; 2, C57BL/6J; 3, CBA/J; 4, *M. spretus;* 5, C57BL/6J + *M. spretus*. The laboratory mouse (L = 190 bp) and *M. spretus* (S = 187 bp) products differ by one CCG repeat unit (see Fig. 4A) as confirmed by DNA sequence analysis.

the decline in DNA sequence conservation upstream from the putative initiator ATG suggests that the coding sequence indeed begins as predicted and includes the polyglutamine segment encoded by the CAG repeat.

The fact that mouse huntingtin also contains a short stretch of polyglutamine argues for a role of this segment in the normal function of the protein. However, there must be considerable leeway in the fulfillment of this role and of the role of the adjacent polyproline stretch, given the extensive CAG repeat variation on normal human chromosomes. The failure to observe a similar variation in the mouse gene, with the exception of a one codon change in a CCG repeat, may indicate a stronger selective pressure for maintaining the length of these repeats in the model organism. Alternatively, the variation in the human repeats may indicate that a greater mean length, the particular chromosomal context in which they are found, or species differences in characteristics of the replication process produce a higher mutation rate in man.

The mechanism by which the expanded CAG repeat leads to neuronal loss in HD is uncertain, but there are a number of arguments for its operating at the protein level. Because RNA is produced at normal levels from the *HD* allele and heterozygous disruption of the gene by translocation does not produce any phenotype (15), the expanded CAG mutation does not entail simple *cis*-inactivation of the *HD* gene (although its effects on adjacent genes remain to be determined). The dominant nature of the HD phenotype (19, 20) indicates that the effect of the expanded repeat must include either *trans*-inactivation of the normal product or conferring of a new property on the abnormal product. Both of these possibilities seem more likely to operate at the protein than at the RNA level, particularly since initial antibody studies of huntingtin have not revealed grossly altered expression in HD (21). The small size and apparent stability of the CAG repeat in mouse is consistent with the absence of an HD-like disorder in this model organism. However, the overall conservation of the *Hdh* gene suggests that genetic manipulation in the mouse, either to produce homozygous "knock-outs" or to introduce an expanded CAG repeat, provides a reasonable hope of resolving the mechanistic issues and of generating an accurate animal model of HD.

ACKNOWLEDGMENTS

We thank Dr. A. Joyner for the 129 genomic phage library and V. Ramesh for mouse DNAs. This work was supported by NIH grants NS16367 (Huntington's Disease Center Without Walls), and by grants from Bristol-Myers Squibb, Inc., the Hereditary Disease Foundation Collaborative Research Agreement, and the Huntington's Disease

Society of America. C.M.A. and M.P.D. received fellowship support from the Andrew B. Cogan Fellowship of the Hereditary Disease Foundation and from the Huntington's Disease Society of America.

LITERATURE CITED

1. Martin, J.B., and Gusella, J.F. (1986). *N. Engl. J. Med.* **315:**1267–1276.
2. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonilla, E., and Martin, J.B. (1983). *Nature* **306:**234–238.
3. Gusella, J.F., Tanzi, R.E., Anderson, M.A., Hobbs, W., Gibbons, K., Raschtchian, R., Gilliam, T.C., Wallace, M.R., Wexler, N.S., and Conneally, P.M. (1984). *Science* **225:**1320–1326.
4. Gusella, J.F., and MacDonald, M.E. (1993). In *Molecular Genetic Medicine, Vol. II,* (ed.) Friedmann, T. (Academic Press, San Diego), pp. 139–158.
5. The Huntington's Disease Collaborative Research Group, Group 1: MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.H., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M.A., Wexler, N.S., and Gusella, J.F.; Group 2: Bates, G.P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.M., and Lehrach, H.; Group 3: Buckler, A.J., Church, D., Doucette-Stamm, L., O'Donovan, M.C., Riba-Ramirez, L., Shah, M., Stanton, V.P., Strobel, S.A., Draths, K.M., Wales, J.L., Dervan, P., and Housman, D.E.; Group 4: Altherr, M., Shiang, R., Thompson, L., Fielder, T., and Wasmuth, J.J.; Group 5: Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., and Collins, F.S.; Group 6: Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D., and Harper, P.S. (1993). *Cell* **72:**971–983.
6. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franza, M., Abbott, M., Gray, J., Conneally, P., Young, A., Penney, J., Hollingsworth, Z., Shoulson, I., Lazzarini, A., Falek, A., Koroshetz, W., Sax, D., Bird, E., Vonsattel, J., Bonilla, E., Alvir, J., Bickham Conde, J., Cha, J.H., Dure, L., Gomez, F., Ramos, M., Sanchez-Ramos, J., Snodgrass, S., de Young, M., Wexler, N., Moscowitz, C., Penchaszadeh, G., MacFarlane, H., Anderson, M., Jenkins, B., Srinidhi, J., Barnes, G., Gusella, J.F., and MacDonald, M.E. (1993). *Nature Genet.* **4:**387–392.
7. Snell, R.G., MacMillan, J.C., Cheadle, J.P., Fenton, I., Lazarou, L.P., Davies, P., MacDonald, M.E., Gusella, J.F., Harper, P.S., and Shaw, D.J. (1993). *Nature Genet.* **4:**393–397.
8. Andrew, S.E., Goldberg, Y.P., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., Graham, R.K., and Hayden, M.R. (1993). *Nature Genet.* **4:**398–403.
9. Rubinsztein, D.C., Leggo, J., Barton, D.E. and Ferguson-Smith, M.A. (1993). *Nature Genet.* **5:**214–215.
10. MacDonald, M.E., Barnes, G., Srinidhi, J., Duyao, M.P., Ambrose, C.M., Myers, R.H., Gray, J., Conneally, P.M., Young, A., Penny, J., Shoulson, I., Hollingsworth, Z., Koroshetz, W., Bird, E., Vonsattel, J.P., Bonilla, E., Moscowitz, C., Penchaszadeh, G., Brzustowics, L., Alvir, J., Bickham-Conde, J., Cha, J.-H., Dure, L., Gomez, F., Ramos-Arroyo, M., Sanchez-Ramos, J., Snodgrass, S.R., de Young, M., Wexler, N.S., MacFarlane, H., Anderson, M.A., Jenkins, B., and Gusella, J.F. (1993). *J. Med. Genet.* **30:**982–986.
11. Grosson, C.L.S., MacDonald, M.E., Duyao, M.P., Ambrose, C.M., Roffler-Tarlov, S., and Gusella, J.F. (1994). *Mamm. Genome* (in press).
12. Feinberg, A.P., and Vogelstein, B. (1984). *Anal. Biochem.* **137:**266–267.
13. Sanger, T., Nicklen, S., and Coulson, A.R. (1977). *Proc. Natl. Acad. Sci. U.S.A.* **74:**5463–5467.
14. Genetics Computer Group (1991). Program manual for the GCG package, version 7, April 1991, 575 Science Drive, Madison, Wisconsin 53711.
15. Ambrose, C.M., Duyao, M.P., Barnes, G., Bates, G.P., Lin, C.S., Srinidhi, J., Baxendale, S., Hummerich, H., Lehrach, H., Altherr, M., Wasmuth, J., Buckler, A., Church, D., Housman, D., Berks, M., Micklem, G., Durbin, R., Dodge, A., Read, A., Gusella, J., and MacDonald, M.E. (1994). *Somat. Cell. Mol. Genet.* **20:**27–38.
16. Lin, B., Rommens, J.M., Graham, R.K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y.P., and Hayden, M. (1993). *Hum. Mol. Genet.* **2:**1541–1545.
17. Cheng, S.V., Martin, G.R., Nadeau, J.H., Haines, J.L., Bucan, M., Kozak, C.A., MacDonald, M.E., Lockyer, J.L., Ledley, F.D., Woo, S.L.C., Lehrach, H., Gilliam, T.C., and Gusella, J.F. (1989). *Genomics* **4:**419–426.
18. Lin, B., Nasir, J., MacDonald, H., Hutchinson, G., Graham, R.K., Rommens, J.M., and Hayden, M. (1994). *Hum. Mol. Genet.* **3:**85–92.
19. Wexler, N.S., Young, A.B., Tanzi, R.E., Travers, H., Starosta-Rubenstein, S., Penney, J.B., Snodgrass, S.R., Shoulson, I., Gomez, F., Ramos-Arroyo, M.A., Penchaszadeh, G., Moreno, R., Gibbons, K., Faryniarz, A., Hobbs, W., Anderson, M.A., Bonilla, E., Conneally, P.M., and Gusella, J.F. (1987). *Nature* **326:**194–197.
20. Myers, R.H., Leavitt, J., Farrer, L.A., Jagadeesh, J., McFarlane, H., Mark, R.J., and Gusella, J.F. (1989). *Am. J. Hum. Genet.* **45:**615–618.
21. Hoogeveen, A.T., Willemsen, R., Meyer, N., de Rooij, K.E., Roos, R.A.C., van Ommen, G., and Galjard, H. (1993). *Hum. Mol. Genet.* **2:**2069–2073.