Somatic Cell and Molecular Genetics, Vol. 20, No. 1, 1994, pp. 27–38

Structure and Expression of the Huntington's Disease Gene: Evidence against Simple Inactivation Due to an Expanded CAG Repeat

Christine M. Ambrose,[1] Mabel P. Duyao,[1] Glenn Barnes,[1] Gillian P. Bates,[2] Carol S. Lin,[1] Jayalakshmi Srinidhi,[1] Sarah Baxendale,[2] Holger Hummerich,[2] Hans Lehrach,[2] Michael Altherr,[3] John Wasmuth,[3] Alan Buckler,[1] Deanna Church,[1] David Housman,[4] Mary Berks,[5] Gos Micklem,[5] Richard Durbin,[5] Alan Dodge,[6] Andrew Read,[6] James Gusella,[1,7] and Marcy E. MacDonald[1]

[1]*Molecular Neurogenetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114;* [2]*Imperial Cancer Research Fund, Lincoln's Inn Fields, London, U.K.;* [3]*Department of Biological Chemistry, University of California, Irvine, California 92717;* [4]*Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139;* [5]*Sanger Centre, Hinxton Hall, Hinxton, Cambridge, CB10 1RQ, U.K.;* [6]*Department of Medical Genetics, St. Mary's Hospital, Manchester, U.K.; and* [7]*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114*

Received 6 December 1993

Abstract—*Huntington's disease, a neurodegenerative disorder characterized by loss of striatal neurons, is caused by an expanded, unstable trinucleotide repeat in a novel 4p16.3 gene. To lay the foundation for exploring the pathogenic mechanism in HD, we have determined the structure of the disease gene and examined its expression. The* HD *locus spans 180 kb and consists of 67 exons ranging in size from 48 bp to 341 bp with an average of 138 bp. Scanning of the* HD *transcript failed to reveal any additional sequence alterations characteristic of HD chromosomes. A codon loss polymorphism in linkage disequilibrium with the disorder revealed that both normal and HD alleles are represented in the mRNA population in HD heterozygotes, indicating that the defect does not eliminate transcription. The gene is ubiquitously expressed as two alternatively polyadenylated forms displaying different relative abundance in various fetal and adult tissues, suggesting the operation of interacting factors in determining specificity of cell loss. The* HD *gene was disrupted in a female carrying a balanced translocation with a breakpoint between exons 40 and 41. The absence of any abnormal phenotype in this individual argues against simple inactivation of the gene as the mechanism by which the expanded trinucleotide repeat causes HD. Taken together, these observations suggest that the dominant HD mutation either confers a new property on the mRNA or, more likely, alters an interaction at the protein level.*

INTRODUCTION

Huntington's disease is a dominant neurodegenerative disorder of mid-life onset in which progressive chorea and dementia result from the premature, specific loss of neurons in the basal ganglia (1). The genetic defect in HD was mapped to chromosome 4 by linkage analysis in 1983 (2). A decade of progressively detailed physical and genetic mapping ensued, producing overlapping cosmid and yeast artificial chromosome clones sets spanning the *HD* region in 4p16.3 (3). Using haplotype analysis to focus the search

0740-7750/94/0100-0027$07.00/0

(4) and exon amplification to identify genes (5, 6), we recently discovered an expanded, unstable trinucleotide repeat at the *HD* locus in affected individuals (7). This polymorphic CAG triplet repeat is located within the predicted coding sequence of IT15, a novel transcript from 4p16.3.

The predicted product of the IT15 transcript, huntingtin, is an ~348-kDa protein with a variable polyglutamine stretch near its *N*-terminus. It does not appear to be related to any previously described protein, giving no clue to its normal function. To provide a foundation for investigating the specific neuronal cell death in HD due to expansion of the trinucleotide repeat, we have defined the genomic organization of the *HD* gene and examined its expression. The *HD* gene contains 67 exons spread across 180 kb. The final exon encoding the 3′UTR gives rise to two different transcripts due to alternative polyadenylation, suggesting that a detailed assessment of the consequences of CAG repeat expansion in HD may require an independent appraisal of the mutation's effects on each transcript. Neither major transcript is confined to the neurons targeted for degeneration. Rather, each is expressed widely in a variety of fetal and adult tissues, suggesting that interacting factors must determine the specificity of cell loss. Both normal and HD alleles are represented in the mRNA population, indicating that the expanded triplet repeat does not eliminate transcription. Finally, in one phenotypically normal individual, the *HD* locus is bisected by a balanced translocation, indicating that heterozygous disruption of this gene does not produce a severe developmental effect and making it unlikely that the neuronal loss in HD is due to direct inactivation of the gene by the triplet repeat expansion. Thus, the HD mutation probably acts to alter or promote an interaction of the IT15 mRNA, or more likely the huntingtin protein, that is specific to striatal neurons and leads to their eventual death.

MATERIALS AND METHODS

Cell Lines. Cell lines from normal individuals, from HD heterozygotes and homozygotes, and from the balanced t(4;12) carrier were established by EBV transformation of blood lymphocytes (8). Somatic cell hybrids have been described previously (9, 10).

Exon Amplification. The exon amplification procedure was used to isolate coding sequences from a contig cosmid spanning the location of the *HD* gene (11). Exon products were obtained from either BamHI-BglII digests cloned into the first generation pSPL1 vector or by BamHI-BglII or PstI digests cloned into the second generation pSPL3 as described (5, 6, Church et al., submitted). PCR-amplified exon products were cloned into pBSKII and sequenced using the double-stranded template by dideoxy chain termination (12).

Exon–Intron Structure. Exon–intron boundaries were sequenced directly from cosmid DNA using specific primers designed from the IT15 cDNA (13). Placement of the boundaries was achieved by comparison of the genomic and cDNA sequences. Some exon–intron boundaries were confirmed by sequencing of random cosmid subclones. To place exons on the physical map, 60 ng of two oligomers (21–23 base pairs each) corresponding to sequences at the edges of each exon were end-labeled with [γ-^{32}P]ATP using T4 polynucleotide kinase and were hybridized independently at 42°C to nylon filter membrane Southern blots of cosmid DNAs digested with EcoRI, followed by washing in 6× SSC at room temperature for 15 min. Oligonucleotides for PCR, DNA sequencing, and hybridization were synthesized using an automated DNA synthesizer (Applied Biosystems). Exons and flanking intronic sequences are being deposited in GenBank.

Blot Analyses. DNA was prepared from cultured cells and Southern blots were prepared and hybridized as described (2, 14). Northern blots were purchased from Clontech Laboratories, Inc., and were hybridized

using the conditions provided by the manufacturer. Probes for Southern and northern analyses were labeled with [α-^{32}P]dATP by the random priming method (15).

Scanning for Polymorphism. To scan for polymorphism, first-strand cDNA was prepared by oligo(dT) priming of 1 μg of lymphoblast mRNA using cloned MuLV reverse transcriptase (BRL) as described (5, 16). The composite IT15 sequence 3′ to the CAG was then amplified by PCR in overlapping segments of ~1 kb using specific primer sets based on the cDNA sequence. Each PCR product was directly sequenced (12) and was also used as template for production of ~200–300 bp ^{32}P-labeled PCR products for SSCP analysis. PCR reactions, direct sequencing, and SSCP analysis were all carried out as described previously (16). PCR products that were refractory to direct sequencing were subcloned into pT7Blue(R) (Novagen). Several independent subclones were then sequenced for each product.

RESULTS

Exon–Intron Structure of HD *Gene.* In the initial search for the *HD* gene, we cloned exons from cosmids spanning a region of 4p16.3 that displayed a common haplotype on approximately ⅓ of HD chromosomes (4). Initially, we employed the first-generation exon amplification system developed by Buckler et al. (5) to produce cloned exons from individual cosmids isolated by sequential walking steps from *D4S180* and *D4S156* (11). These exons were used to identify the IT15 cDNA clones (7).

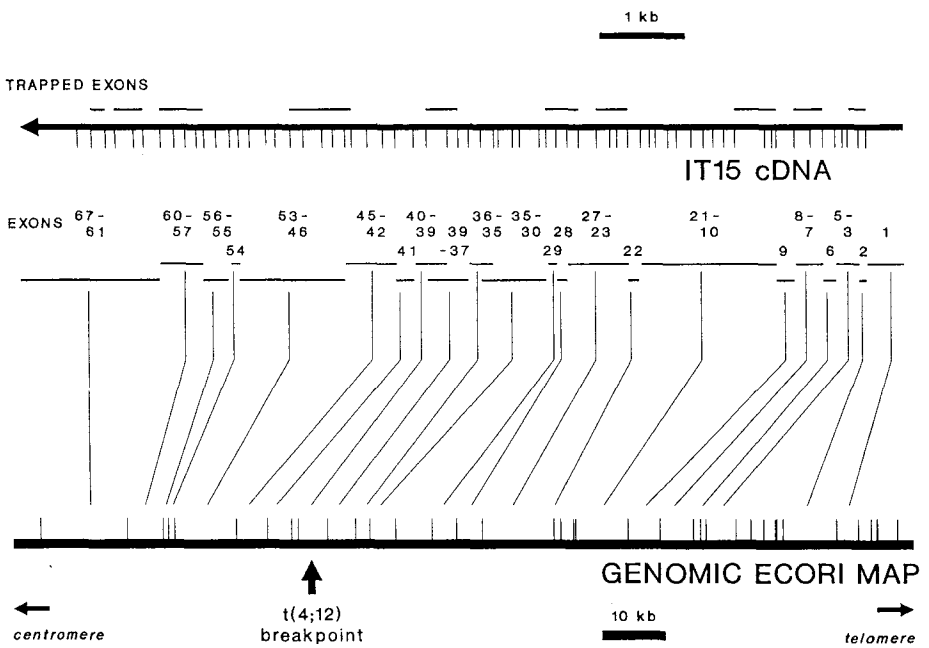

Fig. 1. Exon structure of the *HD* gene. A schematic of the IT15 cDNA is shown (top) with an arrow to denote the direction of transcription. Vertical lines below the cDNA correspond to exon junctions. Horizontal lines above the cDNA span those exons that were recovered from genomic DNA as cloned products in the exon amplification procedure. The genomic EcoRI map of the *HD* region of 4p16.3 (bottom) is shown, with vertical lines denoting EcoRI sites and centromere–telomere orientation provided below (11). Between the cDNA and the genomic map, exons 1–67 are shown as horizontal lines under the corresponding exon number. Contiguous horizontal lines denote exons that map to the same genomic EcoRI fragment as demonstrated by the vertical lines connecting to the physical map. The position of the t(4;12) breakpoint between exons 40 and 41 is shown by the vertical arrow below the genomic map.

Table 1. Exon–Intron Junctions of *HD* Gene

	5′ INTRON SEQUENCE	EXON FIRST BASE	EXON LENGTH (bp)	EXON LAST BASE	3′INTRON SEQUENCE
EXON 1			**	584 TGCACCGACC	GTGAGTTTGGGCCCGCTGCAGCTCCCTGTC
EXON 2	TATTAATTTCCTTCTTTTTTTTATTTTTAG	585 AAAGAAAGAA	84	668 AGTCTGTCAG	GTAATTGCACTTTGAACTGTCTAGAGAAAA
EXON 3	CTTGACAGTTTCTCTTCTTTTTTTGCTTAG	669 AAATTCTCCA	121	789 AGTTATCAAA	GTAAGAACCGTGTGGATGATGTTCTCCTCA
EXON 4	CTTCCATAAATCTCTTGTGATTTGTTGTAG	790 GCTTTGATGG	60	849 AATTAAAAAG	GTGGGCCTTGCTTTTCTTTTTTAAAAATGT
EXON 5	CTTAATGCAACCCTCATTGCACCCCCTCAG	850 AATGGTGCCC	80	929 AGAAATGCAG	GTAAGTTGTACACTCTGGATGTTGGTTTTT
EXON 6	AGAATGACTTGCGTTCTTTTGCATACACAG	930 GCCTTACCTG	139	1068 TGAAATTAAG	GTATGATTGTTGCCTCAGGTCACAAACATG
EXON 7	TTTTATCTACTTGGACTTTTGCTTCCGTAG	1069 GTTTTGTTAA	142	1210 GTGCTCTTAG	GTAAGGTGGAGGCATATGAGTGGAAGAGTC
EXON 8	TGTTAAGATGTCTTGCTTCCACCCCCACAG	1211 GCTTACTCGT	179	1389 GCTTGTCCAG	GTAGGAGCACAGGGTTTACTCTAGGAACTG
EXON 9	ACCAGAACACCTGTGTTTCTCTGTTTCTAG	1390 GTTTATGAAC	205	1594 GAACTTATAG	GCAAGTTATTAGCAAGGTCTACACTTACAA
EXON 10	ACTTTATCTGTCACTTTCTGTGATTTGCAG	1595 CTGGAGGGGG	48	1642 AAACAAAAAG	GTGATTATTTCAGAAATCAGAGTCTTGTGT
EXON 11	TAAAAGGAATGTTGGTACATTATTTACTAG	1643 GCAAAGTGCT	81	1723 GCCTTAACAG	GTAGTTCTCACTAGTTAGCCGCTGGTGTGG
EXON 12	TTTGACAAATGAGTGTTTCTCTGTCTTCAG	1724 CCTCAGTGAA	341	2064 TTCTGAAATT	GTAAGTGGGCAGAGGGGCCTGACATCTTTT
EXON 13	AATTCTCACAGCCCCCCTTGAACCGTTTAG	2065 GTGTTAGACG	124	2188 TCTTCCATGG	GTATGTGGACTACAGGTGATGCGCTACAAA
EXON 14	CACTTAATCTTGATTTCTCTGTTTTTAAAG	2189 CCCTTCAACA	119	2307 AGAAAACAAG	GTGAGGGACATAGGCTTGAGACGACTTGGT
EXON 15	GACAAACAAGTGTCATTGTCTCCTTTCTAG	2308 CCTTGCCGCA	112	2419 GGAAAAAATG	GTGAGTACAAAAGGGGATGTGCACAGTTGA
EXON 16	CTGAAGGTGGCTTGGGTGATTTCTTGGCAG	2420 TGCTGGTTCC	138	2557 GAATACCCTG	GTATGTTAAAAGTTCACATCTGATGTGCTC
EXON 17	GTTCCATGGCTGAGCAATTTATCTCCACAG	2558 AGGAACAGTA	159	2716 ACCCTCACAG	GTAACGGCCAGTTTTTCAGCTGTGTTTTTT
EXON 18	ATGATGTTTGTTGCTTGTTCTTCTGGTTAG	2717 GAAATACATT	98	2814 AGCTGTGAGG	GTGAGCATAATCTTCTGTGGAACCATTTCT
EXON 19	TGTCCTCTTGCCTTGGACCTTGTGTTCCAG	2815 AACTGTGTCA	140	2954 TTGACTTCAG	GTAAGTGAGTCACATCCATTAGATTTCATG
EXON 20	ATTTCATTGTTAAATGTGCTCTTTTGTTAG	2955 GCTGGTGAGC	64	3018 TTATACAGGG	GTAAGCAGTTTATTTTTGTGAGATGCTGTT
EXON 21	TGTTTATTTTTATTATCCTTCTCTCTAAAG	3019 CTTTTAAAAC	101	3119 CACTAATTAG	GTATTTACCAATATTTTATCTCTTTTCCTT
EXON 22	TTAAGCAAATTAACCTTACTTTTGTGTTAG	3120 GCTTGTCCCA	147	3266 CAATAACCAG	GTATGCTGACCCAGTGGCATCTTCACATTG
EXON 23	TATTTTAAGTCTCTATATTTTTGTTATTAG	3267 AATATATAGA	121	3387 AGCACTCACA	GTAAGTCTCTTTCTTGATCGGTCTTACTGA
EXON 24	GGTGTGATTTTATTGTTTCTTTCTTCTCAG	3388 TTTGGATGCT	77	3464 GGCACTGTGG	GTATGTATTTTCCTCAGTATATATTAATAG
EXON 25	TAATTTGACTTTGCAAATGTCTGCTTCCAG	3465 AGTGCCTCCA	152	3616 TTGCTTGCAG	GTACTGGTACTGAGTTGAAACAGGGACTCC
EXON 26	GGAGAGGTNNTGTCTGTGCCCATATCACAG	3617 CCAGTGCTCC	203	3819 CGCAATAAAG	GTAATGTCCCACTTGGGTGCTGGATTCATA
EXON 27	TTGTTTTTTGTTTTTGTTTTTCTATTTTAG	3820 GCAGCCTTGC	127	3946 GCCAGTGCAG	GTAGGAAACAGCGTGGGGAAGGGAGGGACA
EXON 28	AGTTTATCTTTTGTGTGCATATTTTTAAAG	3947 CTTCTAGACA	128	4074 TAACTACAAG	GTATGGGCCTCTGCATCTTTTAAAAATATA
EXON 29	ACCGTGTGTTCTCTCCTTCACCTTCCCAAG	4075 GTCACGCTGG	111	4185 CATTGGGAAG	GTTTGTGTCTTGTTTTTTCTCCTTGGGTTG
EXON 30	TCGCTTAATGTCTCACTTGTCTTTCTACAG	4186 TGTGTTGAAG	78	4263 TGTTCAACAA	GTAAGAGCTTCATTCTTTTCCTCTTCTGTT
EXON 31	ATTGTTGATGCCTCATTTTTTTCACTGTAG	4264 TTGTTGAAGA	224	4487 ACACCTCGGG	GTAACAGTTGTGGCAAGAATGCTGTCGTTG
EXON 32	CTCTGCTTCCCTTTTATTCCCATTTGGCAG	4488 ATGGTTTGAT	79	4566 TGCAGATAAG	GTAAATGGTGTTGTTTGTGGATGTGAACTC
EXON 33	ATTCTTTCTTTCTTTTTTTCTTTTTTATAG	4567 AATGCTATTC	162	4728 TTCAGATCAG	GTTTGTCACTTTTATCTTTCATCCATCATA

		4729			4784		
EXON 34	TTGATGTAAATTTTATTTTCCTTCCTGTAG		GTGTTTATTG	56	GCCAGTTCAG		GTAATAGCATTTTATTATTTTAGATTTTTT
		4785			4933		
EXON 35	AAGGATCTAAATGGATGTTTTTGTTTCTAG		GGAATCAGAG	149	GTGACACATG		GTAACNGGACACACCTTTCACTGTCGTCTT
		4934			5070		
EXON 36	CCTGATAAGGGTACCCTTTTGTCCCCACAG		CCATACCGGC	137	GTACCATCAG		GTAAGAGGAATGTATGTTGGAACTGTCGTG
		5071			5187		
EXON 37	CAGACTTTCTAATTGTGCACGCTCTTATAG		GTGTTGGAGA	117	CAAACAGCAG		GTTTGTCCCCGCAGCCTTGGCTTGTTGTTG
		5188			5310		
EXON 38	TAGAAATGTTTGTGGTGTCTAATTCCACAG		ATGCACATTG	123	AAACACAATG		GTGAGTCTCTCGCCTGGCTCAGCAGATGAA
		5311			5546		
EXON 39	GCTGTGACTTATGTATTATGTTTATTTTAG		GCGTCCGTGA	236	CATTTTCAAG		GTATGCTTTCTATCTGAGCCTATAACTAAC
		5547			5689		
EXON 40	TTCACTGTCATCTTTTTTGTTTCTTGGAAG		GTTTCTATTA	143	TTCAAGTCTG		GTAGGTGAATCACATTAGTCTTCCTGGAGT
		5690			5897		
EXON 41	AAAGACATTTCTCCTTAACTTTGTTTCTAG		GAATGTTCCG	208	AGACCCCGAA		GTAGGTTCATAATGCCCCACAGCCCAGGGC
		5898			6039		
EXON 42	CATTGTCAATGCATCTGTTGCTCCTTCTAG		AAGACACAGT	142	TGATTATGTC		GTAAGTTTGAAATGCCTGTAAACGGGGTTG
		6040			6219		
EXON 43	AAATGAATCTCTCATCATATTTTTCCTTAG		TGTCAGAACC	180	CCTTTCAACT		GTACGTCTTCATCCTGCCGACTATTGCCAG
		6220			6396		
EXON 44	ATCTTTTGTTCTTTTCCTTCTTGCTGTTAG		CCAACCATGC	177	AAATTTACAG		GTATTGGGAAGAGAAACCCTGATATTGATT
		6397			6473		
EXON 45	CAAACACACTAATGTGTTTTTGTCTATTAG		AGCAGCATGC	77	TCGCTCAGAG		GTAATGCTGGAAACACAGGTCGTCCTTGTG
		6474			6612		
EXON 46	ACTGTAATTTCATTTTTATTTGTATTTTAG		ACACCAAAGG	139	TCCGGACAAA		GTAAGTGTCCAGCGTGTCTGCATGGGAGGC
		6613			6735		
EXON 47	TGTTCCCCTTATCCATTTTTTTCTTCCCAG		GACTGGTACG	123	GATGAACTCG		GTACGGGGGGAGCAGTGGAGGCAAGGAATC
		6736			6949		
EXON 48	GTTTGTTAACCTTTAATGCTCTGATTTCAG		GAGTTCAACC	214	GATCTGTTTG		GTAATTAAAATTAAAATTTATCTTATTTTA
		6950			7095		
EXON 49	GCACCCACCCACGAGGTCCTTCTGTTTCAG		GGGATGCTGC	146	AACCCTTGAG		GTAAGAGGCAGCTCGGGAGCTCAGTGTTGC
		7096			7273		
EXON 50	GGCATTCTGTGACTCGGTACTTCCCTTTAG		GCCCTGTCCT	178	CTGGAGGCCG		GTGAGTCCCCGTCCATGAACGGTGGGTTCC
		7274			7375		
EXON 51	ATTCTTCTCTTTGTTCTGTTGTAATTTTAG		TTGCAGTGCA	102	AACACACAGA		GTAAGTCTCAGGACCCATTTTTTTCTTACA
		7376			7563		
EXON 52	AAAGTCCTCTCTTAACCGTTGCTTGTTTAG		ATCCTAAGTA	188	GCCCCCACTG		GTGAGTCTGCTCGTTCCTTGCAGAAGACCA
		7564			7690		
EXON 53	GATGATGTCACTTCCTTTTCATCTTCTCAG		GTGTGGAAGC	127	AACACACTAG		GTACTCTTGGGGCCTCTCCTTCAGGTCACC
		7691			7791		
EXON 54	CACTCTCTCATGTAACATTTATATTTCTAG		GCTGGACCAG	101	CCCACCAGAA		GTAAGGCCACACCCTGTGCTGGTTGGCACA
		7792			7946		
EXON 55	GCTCTTGTTACATGTGGGCTCTCCTTCCAG		GAAGACACAG	155	TCGACACCAG		GTTTGCTTGAGTTCCCACGTGTCTCTGGGA
		7947			8086		
EXON 56	AACACTCTTTACCTTTTTTCTAAAATGTAG		GTTTGGGAGG	140	GCTACTACAG		GTACCTGAGGGAAAGGGAGCGGGGGAGCGG
		8087			8169		
EXON 57	GATCAAGACTCAGGGTGCTGGTGTTCACAG		GTGCCCTCAT	83	ACTCGGCCAG		GTCAGTCTCGCGNNCCCGCCGCCTGGCCTC
		8170			8300		
EXON 58	ACACTGAGCAGTGCCCCGTTTCTGTGGCAG		GTGTCCATAC	131	TCAACTCCAG		GTTTGCAGATGGCCTTTTTATTTTTAACAG
		8301			8430		
EXON 59	TGGAAAATACCCATCTCGCATATTCCACAG		GAAACACCGG	130	GGTCAGATCC		GTAAGTGAGCCTTCCCATTCCCCTCACACC
		8431			8586		
EXON 60	CCTTGCCCTCCTGGTTTTCCACATCTCCAG		CTTCTAGTGG	156	CCTTGGGATG		GTAAGTGACAGGTGGCACAGAGGTTTCTGT
		8587			8777		
EXON 61	ATGCAGCAGCTTTTGTCTGTGTGTGCCTAG		GACAAGGCCG	191	GGATCGCCCA		GTGAGTGGGAGCCTGGCTGGGGCTGGGGCG
		8778			8892		
EXON 62	CTGAGCCTGGATGCTGTCTCCCGTTTTCAG		CTGCGTGAAC	115	AATAATACAG		GTGAGTGGGCCCTGGCTGTCTTCCTCTGCA
		8893			9106		
EXON 63	TTTGACACAGAGGCCTTTCTCCCTGTGCAG		ATGTGTGGGG	214	ATGTACACAG		GTGAGCATGTACACGGTGCCCATAAGGCCA
		9107			9212		
EXON 64	TAACCTTCGTACTGAACACTTTTGTTACAG		GAAAGGAGAA	106	TTTTTGATAG		GTAAGAAGCGAANCCCATCCCTCAGCCCGT
		9213			9375		
EXON 65	TCAGTCTCTGACCTGCGTCCCTCCTCCCAG		GATCAGGAAA	163	GGTGTATAAG		GTGAGGTTGCATGTGGGATGGGGATGGAGT
		9376			9536		
EXON 66	TGACACTCAGGCGCCTGCTTGCTCTTGCAG		GTGTTTCAGA	161	TCGCGGCGAT		GTATCCTCTCTGGNTCCCTGGTNCTGGCCC
		9537					
EXON 67	GCCGGCCTTTTTCCTTAACTCCTGCACCAG		CCTCCCACAT	**			

To determine intron–exon junctions, we used DNA primers located every 200–300 bp in the cDNAs to directly sequence the corresponding cosmid DNAs and designed new primers as needed based on the evolving knowledge of the exon structure. As this work progressed, a second-generation vector system that eliminated false-positive products and allowed cloning of genomic DNA with multiple restriction enzymes was ap-

plied in multiple experiments to saturate the region with cloned exons (Church et al., submitted). The products obtained in this system have the additional advantage that 5′–3′ orientation is immediately discernible. To position all exons on the physical map, two primers from each exon were hybridized to EcoRI digests of all overlapping cosmids from the region, representing an average threefold redundancy.

The composite IT15 cDNA sequence corresponds to a genomic segment of 180 kb and is encoded in 67 exons as shown in Fig. 1. The sequences of exon–intron junctions are listed in Table 1. The internal exons ranged in size from 48 bp to 341 bp with an average of 138 bp. All cloned, sequenced exons are aligned with the composite cDNA sequence in Fig. 1, and together constitute 36% of the transcript. Of the 65 internal exons, 27 were trapped by exon amplification using PstI or BamHI-BglII digests, 15 as single exon products and 12 as multiple adjacent exons spliced together in the amplification procedure. The minimum and maximum sized exons were both represented in this collection, which averaged 139 bp/exon, indicating no apparent size bias in the procedure.

Codon Loss Polymorphism in IT15. To search for DNA changes other than the trinucleotide repeat expansion that might also be associated with HD, we compared the normal and HD transcripts by sequence analysis of partial cDNA clones and by single-strand conformational polymorphism analysis (SSCP) of PCR products from first-strand cDNA (17). Sequencing of individual normal cDNAs revealed four single base pair differences from the consensus sequence, at positions 1849 (C to G, Leu to Val), 2372 (C to G, Ser to Cys), 4034 (G to A, Arg to Lys), and 8677 (A to G, Ile to Val). Each of these changes could represent polymorphisms or reverse transcriptase errors either in the published sequence or in the variant cDNA. They were not pursued, since they only occurred on normal chromosomes. No sequence differences other than the CAG repeat length were found exclusively in the HD cDNAs.

For SSCP analysis, we scanned two HD homozygotes of different haplotypes, both alleles from a single normal individual, and the normal sequence represented in a corresponding cDNA clone. Variant SSCP bands were detected in exons 58, 60, and 67. The exon 67 difference involved a choice of either C or T at position 9809 of the composite cDNA sequence (GENBANK #L12392). Because this change occurred in the 3′ untranslated region and both forms were represented on at least one normal and one HD allele, we did not analyze it further. The exon 60 difference was found only on one of the normal alleles and was therefore abandoned without sequencing.

The exon 58 difference (Fig. 2A) was present in the HD homozygote representing the most common disease haplotype, but absent from an HD homozygote of another haplotype (4, 18). Sequence analysis of multiple cloned PCR products revealed the loss of a single codon from a run of four consecutive GAG (Glu) codons at positions 2642–2645 of the predicted amino acid sequence (7). For convenience, we refer to this change as Δ2642. We developed a genomic PCR assay for Δ2642 in order to scan additional HD and normal chromosomes to test its disease specificity (Fig. 2B). This analysis revealed that the codon loss represents a normal, infrequent polymorphism with allele frequencies of 0.93 and 0.07 for presence or absence or codon 2642, respectively ($N = 175$ normal chromosomes). The Δ2642 change showed linkage disequilibrium with HD ($\chi^2 = 37.47$, 1 *df*, $P < 0.0001$), where the codon loss was represented on 38% of disease chromosomes ($N = 80$ independent HD chromosomes).

Both HD *Alleles are Expressed.* The Δ2642 polymorphism provided a ready means to assay whether both alleles of the *HD* gene are expressed in the cells of affected individu-

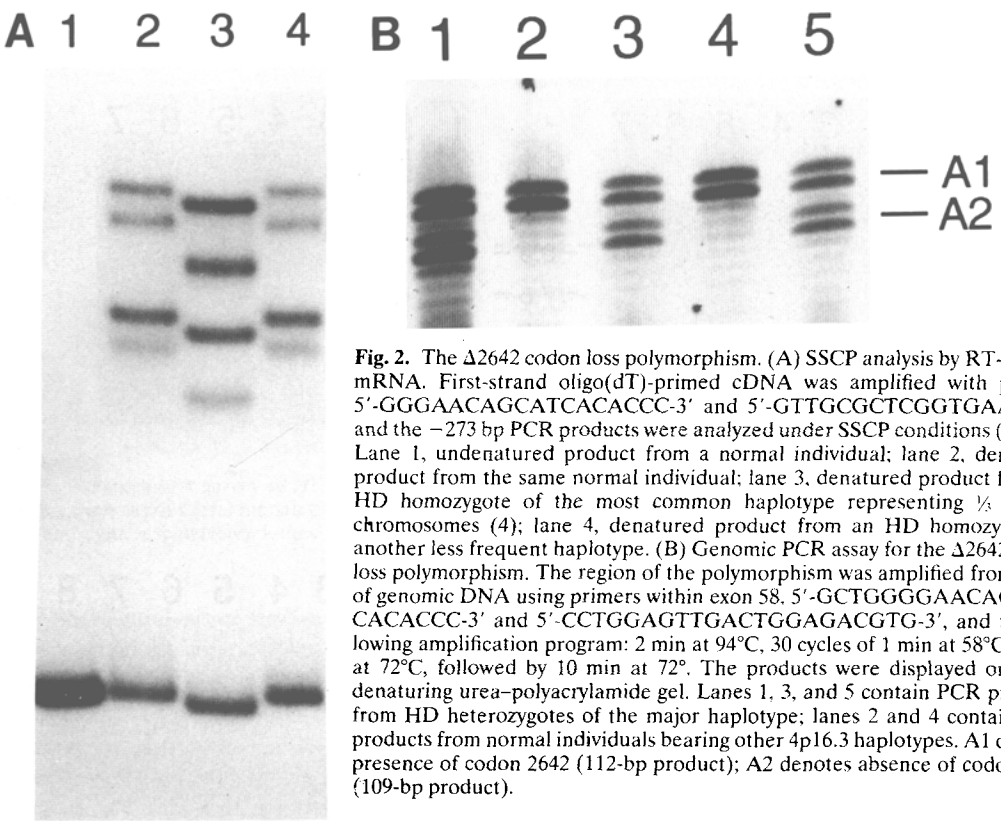

Fig. 2. The Δ2642 codon loss polymorphism. (A) SSCP analysis by RT-PCR of mRNA. First-strand oligo(dT)-primed cDNA was amplified with primers 5′-GGGAACAGCATCACACCC-3′ and 5′-GTTGCGCTCGGTGAACA-3′ and the −273 bp PCR products were analyzed under SSCP conditions (16, 17). Lane 1, undenatured product from a normal individual; lane 2, denatured product from the same normal individual; lane 3, denatured product from an HD homozygote of the most common haplotype representing ½ of HD chromosomes (4); lane 4, denatured product from an HD homozygote of another less frequent haplotype. (B) Genomic PCR assay for the Δ2642 codon loss polymorphism. The region of the polymorphism was amplified from 10 ng of genomic DNA using primers within exon 58. 5′-GCTGGGGAACAGCAT-CACACCC-3′ and 5′-CCTGGAGTTGACTGGAGACGTG-3′, and the following amplification program: 2 min at 94°C, 30 cycles of 1 min at 58°C, 1 min at 72°C, followed by 10 min at 72°. The products were displayed on a 6% denaturing urea–polyacrylamide gel. Lanes 1, 3, and 5 contain PCR products from HD heterozygotes of the major haplotype; lanes 2 and 4 contain PCR products from normal individuals bearing other 4p16.3 haplotypes. A1 denotes presence of codon 2642 (112-bp product); A2 denotes absence of codon 2642 (109-bp product).

als. Figure 3 shows the analysis of two independent preparations of first strand cDNA from lymphoblast lines of four unrelated HD patients, two of whom were heterozygous for the polymorphism, with the codon loss segregating with the disease chromosome. Both of these individuals clearly expressed both the normal and disease alleles. Similar results have been observed in RNA from normal individuals, HD heterozy-

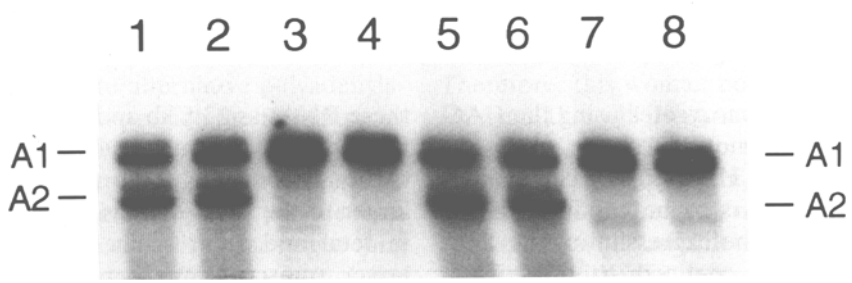

Fig. 3. Expression of both normal and HD alleles in mRNA. First-strand cDNA primed with oligo(dT) (lanes 1–4) or an IT15-specific primer (5′-CAGGTACTGAGCGAGGAT-3′) (lanes 5–8) was amplified using the same primers described in Fig. 2B. The PCR products spanning the Δ2642 codon polymorphism were resolved on a 6% denaturing urea–polyacrylamide gel. Lanes 1, 5 and 2, 6 represent two different HD heterozygotes with the major HD haplotype; lanes 3, 7 and 4, 8 represent two different HD heterozygotes with two other HD haplotypes. A1 denotes presence of codon 2642 (112 bp product); A2 denotes the absence of codon 2642 (109 bp product).

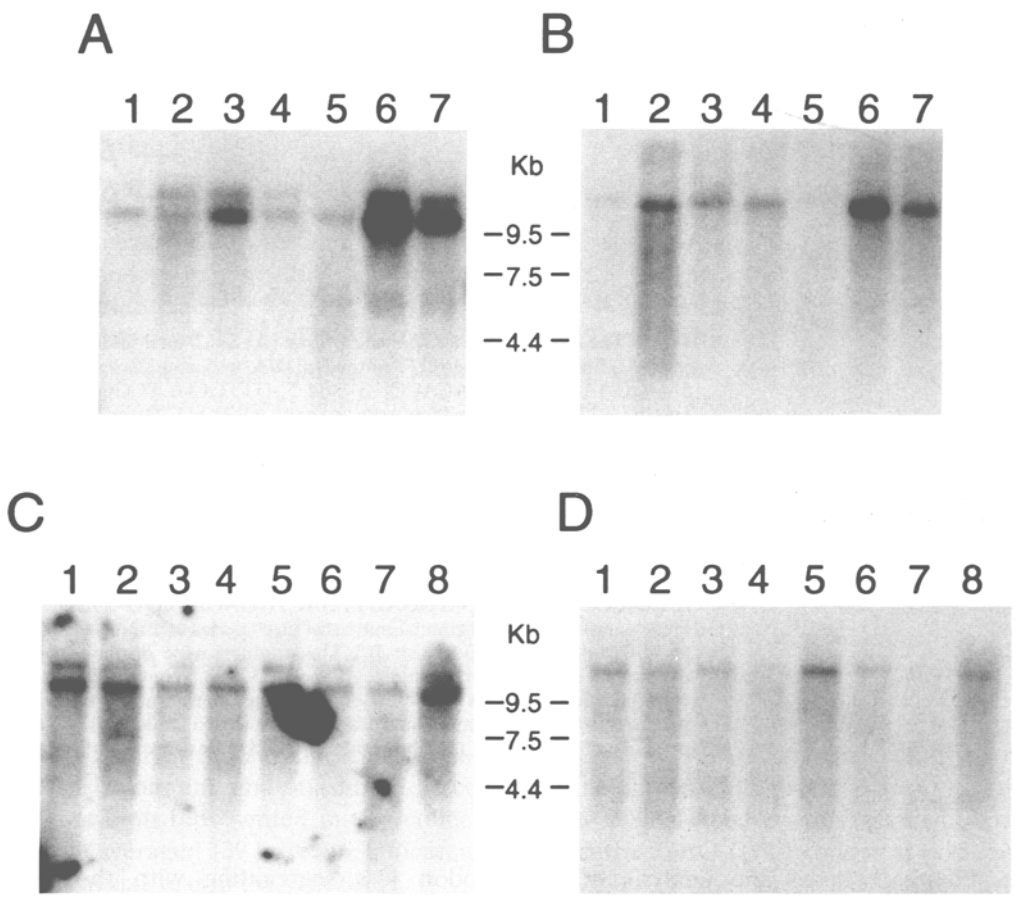

Fig. 4. Northern blot survey of HD gene transcripts in adult tissues. Northern blots containing 2 μg of poly(A)$^+$ mRNA from various adult human tissues were hybridized with two probes. Panels A and C were hybridized with coding region 2841-bp probe made by EcoRI digestion of cDNA clone IT15B. This probe spans nucleotides 2028–4868 of the published IT15 sequence (7). Panels B and D represent the same blots hybridized with a 292-bp genomic probe produced by PCR of cosmid L120D5 DNA using primers 5′-GGAGAACACAGTCGTCTGTG-3′ and 5′-CGTGTAAAGTATGTGAATCGC-3′. This probe derives from the sequence immediately 3′ to the end of the 3′UTR reported in the published IT15 sequence. Panels A and B, lanes 1, heart; 2, brain; 3, placenta; 4, lung; 5, liver; 6, skeletal muscle; 7, kidney. Panels C and D, lanes 1, spleen; 2, thymus; 3, prostate; 4, testis; 5, ovary; 6, small intestine; 7, colon; 8, peripheral blood leukocyte. Transcript sizes were estimated from RNA size markers as shown.

gotes, and HD homozygotes using the CAG repeat assay (data not shown).

Expression of HD *mRNA.* The pathology of HD appears to be confined to the brain. However, the expression of the IT15 transcript is not confined to this tissue. Figures 4 and 5 show a northern blot survey of 15 adult and five fetal tissue RNAs, respectively. Hybridization with an IT15 probe revealed two RNA species that were present in all tissues tested but varied in relative abundance. We estimate the size of these RNAs as 13.5 kb and 10.5 kb, with the latter being the more abundant in most tissues. Interestingly, the apparent ratio of larger to the smaller transcript was greatest in fetal and adult brain. By contrast, the larger transcript was barely detectable in adult liver and colon.

We presumed that the smaller RNA species corresponds to the composite cDNA sequence previously reported (7) and that the larger could result either from alternative splicing or from alternative polyadenylation.

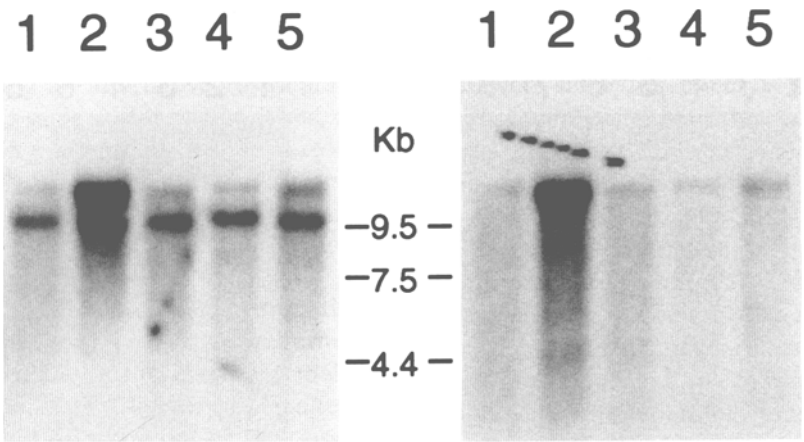

Fig. 5. Northern blot survey of *HD* gene transcripts in fetal tissues. A northern blot containing 2 μg of poly(A)$^+$ mRNA from various fetal human tissues were hybridized sequentially with the same two probes described in Fig. 4 (left and right, respectively). Lanes 1, heart; 2, brain; 3, lung; 4, liver; and 5, kidney.

Our SSCP analysis of first-strand cDNA had failed to yield any evidence of extensive alternative splicing, and an exon 2 probe detected both RNA species (data not shown). Therefore, we prepared a genomic probe from the region of cosmid L120D5 located immediately downstream from the sequence at the site of the poly(A) tail in cDNA clone IT15B (7, 8). Hybridization of this second probe to the northern blots is also shown in Figures 4 and 5. The extended 3′ probe detected only the larger of the two IT15 RNA species, suggesting that this transcript arises by use of a downstream poly(A) addition site. Indeed, cDNA library screening has recently identified two types of clones differing in their 3′UTRs, which were presumed to be due to alternative polyadenylation (19). Our data provide the formal proof for this presumption.

Balanced Translocation Disrupting HD *Gene.* The *HD* gene search produced a panel of somatic cell hybrid lines dissecting 4p into several regions (9, 10). One of the chromosomes from this panel has a t(4p16.3; 12p13.3) with a breakpoint between *D4S180* and *D4S127* (20). To establish whether this chromosome bisects the *HD* gene, we hybridized exon probes to genomic blots of DNA from a lymphoblast cell line (CV066) with the balanced translocation and from a hybrid line (HHW1071) containing only the region of 4p16.3 between the translocation breakpoint and the 4p telomere as part of the der(12) chromosome. Exons 41–67 are absent from the hybrid, indicating that the breakpoint maps between exons 40 and 41. Indeed, the EcoRI and HindIII fragments containing exon 40 are altered in size in CV066 and in HHW1071 (Fig. 6) positioning the t(4;12) breakpoint within the *HD* gene as depicted in Fig. 1.

The CV066 lymphoblast line was derived from a balanced carrier of the t(4;12) who was first identified as the mother of a Wolf-Hirschhorn child produced by transmission of only the der(4) chromosome (20). Therefore, this woman possesses one intact *HD* gene that will produce a normal product, and a bisected *HD* gene that at best could produce a partial protein, or partial fusion protein. This balanced translocation is not associated with any detectable abnormal phenotype either in the woman or in one of her offspring. Thus, heterozygous disruption of the *HD* gene does not have catastrophic consequences for development or cause juvenile HD. Moreover, this translocation makes it unlikely that the expanded CAG repeat in HD acts by simply inactivating the

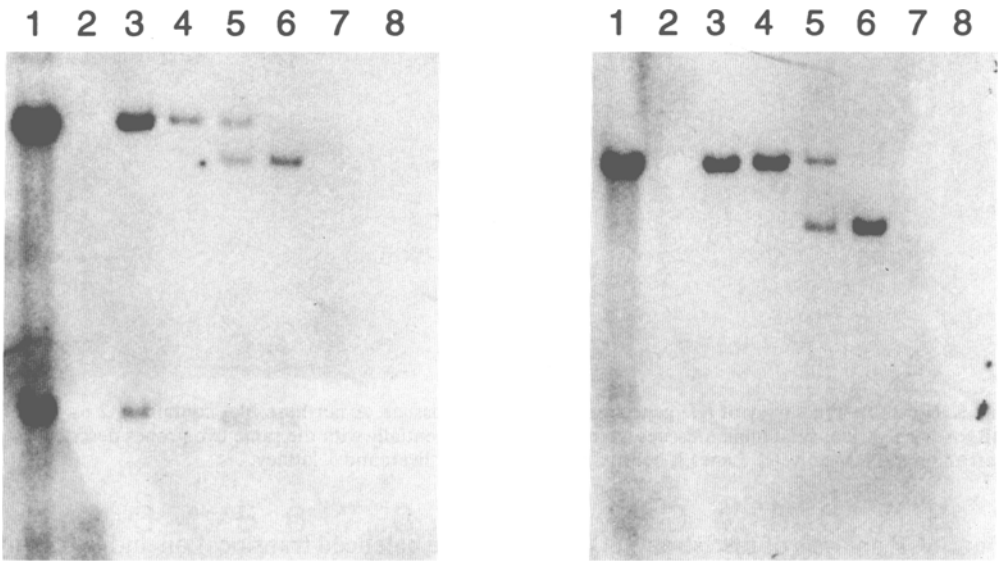

Fig. 6. The t(4;12) translocation disrupts the *HD* gene. Southern blots of HindIII- (panel A) and EcoRI- (panel B) digested DNAs were hybridized with a 210-bp probe consisting of all of exon 40 and that portion of exon 39 3′ to the EcoRI site contained within this exon. The probe was made by PCR from the cDNA using primers 5′-CTTCAACG-CTAGAAGAAC-3′ and 5′-CAGACTTGAAGATGTGGATC-3′. Lane 1, normal human genomic lymphoblastoid cell DNA; lane 2, hamster DNA; lane 3, DNA from human–hamster hybrid HHW416 containing only human chromosome 4; lane 4, DNA from human–hamster hybrid HHW661 containing only a human t(4p15.1;5p15.1) chromosome; lane 5, DNA from lymphoblast line CV066 from the balanced t(4p16.3;12p13.3) carrier (20); lane 6, DNA from human–hamster hybrid HHW1071 containing the der(12) from CV066; lane 7, DNA from human-hamster hybrid HHW842 containing a chromosome 4 with an interstitial deletion the removes the entire HD gene; lane 8, DNA from human–hamster hybrid HHW847 containing a t(4;21) chromosome from which all of 4p16.3 is missing (9, 10). Both EcoRI and HindIII fragments are altered in CV066 and HHW1071. Since exons 39 and 40 reside on the same EcoRI fragment but different HindIII fragments (the small unaltered HindIII fragment is detected by exon 39), the t(4;12) breakpoint must map within the EcoRI fragment but proximal to both exons.

allele containing it. At age 46, the woman, who possesses only one intact copy of this locus, is already beyond the age of onset of the majority of HD cases and does not display any signs of the disorder. She has also passed the balanced translocation to one offspring who is similarly phenotypically normal.

DISCUSSION

The number of exons comprising the *HD* gene is one of the highest reported to date for any human locus. However, the exons are arrayed across a relatively compact genomic region of 180 kb. The initial identification and detailed analysis of this locus was aided tremendously by the development of the exon amplification procedure (5, 6). Cloned trapped exons provided probes for the isolation of cDNA clones and multiple sequenced, oriented entry points for aligning the cDNAs. The knowledge of the cosmid of origin of each trapped exon included in a cDNA clone gave an immediate assessment of genomic coverage, and provided the basis for complete sequence analysis and rapid determination of exon–intron junctions. The fact that 42% of the internal exons susceptible to exon amplification were recovered as cloned segments demonstrates that it is remarkably easy to isolate a significant portion of a gene using this procedure. In fact, we only employed two of several possible enzyme combinations for

cloning the genomic DNA. It is likely that many of the exons that were missed could be isolated using an alternative restriction digest with the same vector system. Thus, exon amplification appears to be an excellent means of saturating a particular genomic region with expressed sequences and quickly relating the corresponding transcripts to the physical map.

The *HD* gene is expressed in every tissue tested to date, with at least two alternative forms that differ in the extent of their 3′ untranslated region. There might be alternative splicing of the transcript in some tissues, but RNA-PCR SSCP analysis of lymphoblastoid cell RNA failed to reveal any evidence of alternative forms within the coding sequence. Moreover, the exon trapping did not yield any other putative exons from this region that could be a part of a transcript from this gene. Finally, all of the overlapping cDNAs so far isolated from brain and other tissues have been colinear, except when they contain unspliced intronic sequence. Thus, if alternative splicing occurs, it is unlikely to be extensive unless it is restricted to a specific cell type not yet explored.

The ubiquitous expression of the *HD* gene requires that special considerations must be invoked to explain the specific and selective nature of loss of striatal neurons. The expanded CAG repeat does not appear to interfere with transcription since HD homozygotes, HD heterozygotes, and normal individuals all express RNA from both alleles. However, these studies were performed in a peripheral tissue. It remains conceivable that a subset of cells in the striatum might be affected distinctly at the transcriptional level. Testing of this possibility awaits the examination of HD brain tissue prior to any cell loss, which is extremely difficult to obtain.

The HD mutation could conceivably alter translation, stability, or localization of the mRNA, an effect that could act on either or both of the two alternatively polyadenylated mRNA forms. The need to examine both of these transcripts independently thus complicates assessment of such mechanisms. Moreover, since the effect is likely to be tissue-specific, in view of the pattern of cell death, the lack of appropriate HD tissue is again a stumbling block.

If, as seems probable, the HD mutation acts at the protein level, it does not create a gene product that is directly toxic, since cell death occurs selectively in the striatum. In this gain-of-function scenario, some special aspect of the metabolism of striatal neurons must be particularly sensitive to the effects of the altered product, perhaps via a specific protein–DNA, protein–RNA, or protein–protein interaction.

Of course, it is also possible that the mutation causes a loss of function of the protein. To explain the complete phenotypic dominance observed in HD, this loss of function must either cross below a critical threshold of activity that cannot be adequately compensated by the normal allele or must *trans*-inactivate the normal product as well. In either loss-of-function scenario, the striatal cells must be peculiarly sensitive to such a loss. The observed correlation between CAG repeat length and severity of disease would dictate that a range of functional loss must occur in HD (21–23). The identification of an individual with an interrupted *HD* gene indicates that 50% of the normal gene product can apparently be lost without immediate dire consequence. If this disruption halved the activity of the *HD* gene, then the threshold hypothesis would predict that juvenile onset HD would ensue. However, if the partial protein retained some function, or if *trans*-inactivating effects are required, then an HD phenotype might occur later or not at all. Because of the typically late onset of the disorder, careful long-term clinical evaluation will be required to establish whether this structural disruption can mimic the effects of an expanded trinucleotide repeat, causing the characteristic symptoms and neuropathology of HD.

ACKNOWLEDGMENTS

This work was supported by NIH grant NS16367 (Huntington's Disease Center Without Walls) and by grants from Bristol-Myers Squibb, Inc., the Hereditary Disease Foundation Collaborative Research Agreement, the Huntington's Disease Society of America, and the Deutsche Forschungsgemeinschaft. Sequencing at the Sanger Centre was supported by the Wellcome Trust. 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., and MacDonald, M.E. (1993). In: Molecular Genetic Medicine, *Vol.* II, (ed.) Friedmann, T. (Academic Press, San Diego), pp. 139–158.
4. MacDonald, M.E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., Lehrach, H., Collins, F.S., Wasmuth, J.J., Frontali, M., and Gusella, J.F. (1992). *Nature Genet.* **1**:99–103.
5. Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A., and Housman, D.E. (1991). *Proc. Natl. Acad. Sci. U.S.A.* **88**:4005–4009.
6. Church, D.M., Banks, L.T., Rogers, A.C., Graw, S.L., Housman, D.E., Gusella, J.F., and Buckler, A.J. (1993). *Hum. Mol. Genet.* (in press).
7. 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.
8. Anderson, M.A., and Gusella, J.F. (1984). *In Vitro* **20**:856–858.
9. Smith, B., Skarecky, D., Bengtsson, U., Magenis, R.E., Carpenter, N., and Wasmuth, J.J. (1988). *Am. J. Hum. Genet.* **42**:335–344.
10. Lin, C.S., Altherr, M., Bates, G., Whaley, W.L., Read, A.P., Harris, R., Lehrach, H., Wasmuth, J.J., Gusella, J.F., and MacDonald, M.E. (1991). *Somat. Cell Mol. Genet.* **17**:481–488.
11. Baxendale, S., MacDonald, M.E., Mott, R., Francis, F., Lin, C., Kirby, S.F., James, M., Zehetner, G., Hummerich, H., Valdes, J., Collins, F.S., Deaven, L.J., Gusella, J.F., Lehrach, H., and Bates, G.P. (1993). *Nature Genet.* **4**:181–186.
12. Sanger, T., Nicklen, S., and Coulson, A.R. (1977). *Proc. Natl. Acad. Sci. U.S.A.* **74**:5463–5467.
13. McClatchey, A.I., Lin, C.S., Wang, J., Hoffman, E.P., Rojas, C., and Gusella, J.F. (1992). *Hum. Mol. Genet.* **1**:521–527.
14. Gusella, J.F., Varsanyi-Breiner, A., Kao, F.T., Jones, C., Puck, T.T., Keys, C., Orkin, S., and Housman, D.E. (1979). *Proc. Natl. Acad. Sci. U.S.A.* **76**:5239–5243.
15. Feinberg, A.P., and Vogelstein, B. (1984). *Anal. Biochem.* **137**:266–267.
16. Ambrose, C., James, M., Barnes, G., Lin, C., Bates, G., Altherr, M., Duyao, M., Groot, N., Church, D., Wasmuth, J.J., Lehrach, H., Housman, D., Buckler, A., Gusella, J.F., and MacDonald, M.E. (1992). *Hum. Mol. Genet.* **1**:697–703.
17. Orita, M., Suzuki, Y., Sekiya, T., and Hayashi, K. (1989). *Genomics* **5**:874–879.
18. 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.
19. 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.
20. McKeown, C., Read, A.P., Dodge, A., Stecko, O., Mercer, A., and Harris, R. (1987). *J. Med. Genet.* **24**:410–412.
21. 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.
22. 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.
23. 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.