© 1999 Oxford University Press
Human Molecular Genetics, 1999, Vol. 8, No. 9 **1807–1815**

A putative *Drosophila* homolog of the Huntington's disease gene

Zhen Li[+], Chris A. Karlovich[+], Matthew P. Fish[1], Matthew P. Scott[1] and Richard M. Myers[§]

Department of Genetics and [1]Departments of Developmental Biology and Genetics and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305-5120, USA

Received May 18, 1999; Revised and Accepted June 7, 1999

The Huntington's disease (HD) gene encodes a protein, huntingtin, with no known function and no detectable sequence similarity to other proteins in current databases. To gain insight into the normal biological role of huntingtin, we isolated and sequenced a cDNA encoding a protein that is a likely homolog of the *HD* gene product in *Drosophila melanogaster*. We also determined the complete sequence of 43 125 contiguous base pairs of genomic DNA that encompass the *Drosophila HD* gene, allowing the intron–exon structure and 5'- and 3'-flanking regions to be delineated. The predicted *Drosophila* huntingtin protein has 3583 amino acids, which is several hundred amino acids larger than any other previously characterized member of the HD family. Analysis of the genomic and cDNA sequences indicates that *Drosophila HD* has 29 exons, compared with the 67 exons present in vertebrate *HD* genes, and that *Drosophila* huntingtin lacks the polyglutamine and polyproline stretches present in its mammalian counterparts. The *Drosophila HD* mRNA is expressed in a broad range of developmental stages and in the adult, a temporal pattern of expression similar to that observed for mammalian *HD* transcripts. We can discern five regions of high similarity from multiple sequence alignments between *Drosophila* and vertebrate huntingtins. These regions may define functionally important domains within the protein.

INTRODUCTION

Huntington's disease (HD) is an autosomal dominant neurodegenerative disease that affects 4–10 in every 100 000 individuals in North America and Europe (1–4). The clinical features include an involuntary movement disorder, dementia and cognitive decline, and an onset in adult life. The pathological hallmark of the disease is atrophy of the caudate and putamen within the basal ganglia. The *HD* gene product, huntingtin, has a cluster of CAG-encoded glutamine repeats that is increased in size in patients with the disease (5). Chromosomes from HD patients have a CAG repeat length of 36–121 in their *HD* genes, whereas almost all unaffected individuals have <30 repeats (6). The association of an increased number of glutamines with disease is a feature shared by the gene products of a large group of neurodegenerative disorders, including spinocerebellar ataxias 1, 2, 3, 6 and 7, dentatorubral-pallidoluysian atrophy (DRPLA), and spinal and bulbar muscular atrophy (SBMA) (reviewed in refs 7,8).

Considerable progress has been made toward understanding the cellular changes that take place in neurons affected by HD. An important advance in our knowledge came from experiments in which an N-terminal *HD* transgene containing 115–150 CAG repeats was overexpressed in mice. The mice developed a progressive neurological phenotype similar to many of the features of HD, including involuntary stereotypic movements, tremor, epileptic seizures and weight loss (9). Neuropathological analysis of the mice revealed a single inclusion body in the nuclei of neurons but not glia in several brain regions (10). These neuronal intranuclear inclusions (NIIs) are recognized by antibodies against ubiquitin and the N-terminus of huntingtin. Similarly, ubiquitin-positive inclusions have been found in the affected regions of the brain in patients with HD (11) or other CAG expansion disorders (12–14). Although the function of these aggregates has yet to be determined, two recent studies argue that their role may be protective rather than pathogenic (15,16).

In spite of these advances, little is known about the normal biological function of huntingtin. While human huntingtin contains 3144 amino acid residues, no significant similarity to proteins outside the HD family has been found in current databases, nor have any proven functional domains or motifs been identified over any stretch of this very large protein. Some insight into huntingtin function has come from immunohistochemical studies, in which the protein has been localized around secretory vesicles (17,18), suggesting a possible role in vesicular trafficking. Although these localization studies as well as yeast two-hybrid (19) and other experiments investigating the biochemistry of HD (20) have been informative, the precise function of huntingtin has remained elusive.

One means of identifying potentially important functional domains of a protein is to isolate orthologs from other organisms. In the case of the HD protein, vertebrate homologs from mouse (21,22), rat (23), pufferfish (24) and zebrafish (25) have all been identified. However, the high degree of amino acid sequence similarity between these tested vertebrate proteins has made it difficult to determine potential functional domains. Sequence comparison with a more distantly related species, in particular that of an invertebrate, could be more informative.

[+]These authors contributed equally to this work
[§]To whom correspondence should be addressed. Tel: +1 650 725 9687; Fax: +1 650 725 9689; Email: myers@shgc.stanford.edu

One of the most powerful tools for understanding the function of a protein is to use genetic techniques for generating specific mutations in the gene encoding the protein. Shortly after the mouse *HD* gene was cloned, several groups generated mice nullizygous for the gene and studied the loss-of-function phenotype (26–28). In all these studies, the mice died early in embryogenesis during gastrulation, suggesting that the *HD* gene product has a necessary role in development of the embryo. Because it is expressed in a wide range of adult tissues (29–31), the huntingtin protein almost surely has functions in the adult. The lethal phenotype at an early developmental stage of mice containing a knockout mutation has made it difficult to define a specific function of huntingtin in cellular processes. While conditional, tissue-specific mutations can be constructed in the mouse genome, it is clear that our understanding of the functions of the huntingtin protein would be greatly aided by identifying a homolog of the gene in an easily manipulable organism, such as baker's yeast, the nematode worm or the fruit fly.

In this study, we report the isolation of a putative homolog of the *HD* gene from *Drosophila* and describe a detailed analysis of its genomic structure. The low overall similarity we find between the predicted *Drosophila* HD protein and the vertebrate family of huntingtins allows us to pinpoint five areas of high evolutionary conservation that are likely to contain functionally important parts of the protein. The availability of a *Drosophila* homolog will make it possible to use classical genetic approaches toward an understanding of the normal role of huntingtin in the cell.

RESULTS

Isolation of the *Drosophila HD* cDNA

We carried out a BLAST search of the NCBI dbEST database with the zebrafish *HD* gene product as a query to identify a *Drosophila* homolog of the HD gene. One of the searches produced an alignment with the predicted translated peptide product from a *Drosophila* expressed sequence tag (EST) sequence known as LD23533, with an E value of 10^{-6}. We subsequently sequenced the entire 1.95 kb LD23533 cDNA, and found that the predicted amino acid sequence was highly similar to the N-terminus of vertebrate huntingtins, with the pairwise alignment to human HD returning a probability score of 2×10^{-21} using BLOSUM62 as the scoring matrix.

We hybridized *Drosophila* polytene chromosomes with a biotinylated probe from LD23533 and mapped the cDNA to the 98E1–E2 region of the third chromosome. Genomic P1 clones known to map to this region were obtained and screened by PCR with primers specific to LD23533. From the screen, we identified one P1 clone, DS04556, that contained the LD23533 cDNA. To obtain more sequence information from DS04556, we constructed a library of fragments ~3 kb in length from the sheared insert of the P1 clone. By randomly sequencing subclones from the library and performing BLAST searches, we identified six clones that showed sequence similarity to the *HD* gene family. More specifically, predicted open reading frames (ORFs) from these clones contained sequences having similarity scattered across the zebrafish HD protein at amino acid positions 635–670, 1290–1365, 1461–1564, 2423–2483, 2713–2765 and 2958–3021. We isolated sequence for most of the remainder of the predicted coding region of this gene by designing primers to predicted exons from the six library clones and performing RT–PCR on *Drosophila* poly(A)$^+$ RNA. Six overlapping RT–PCR fragments were obtained (Fig. 1). From the sequence of these fragments and of LD23533, we were able to build a contig that corresponds to the cDNA of a *Drosophila HD*-like gene. Because of the strong similarity to the *HD* gene family, we have designated this gene *Drosophila HD*.

We identified additional cDNA clones by screening the 0–22 h stage embryonic cDNA library from which LD23533 was identified, and by screening a *Drosophila* head cDNA library. Two overlapping cDNA clones, 3.3 and 4.8 kb in length, were recovered and sequenced (Fig. 1). These contain stretches of sequence identical to those obtained by RT–PCR, except for mismatches at some positions (data not shown). To explore the nature of these sequence differences, we sequenced both strands from multiple independent RT–PCR clones. In each case, the sequence differences remained, demonstrating that the observed discrepancy between RT–PCR and cDNA clones was not the result of base substitutions introduced by PCR. Instead, it is likely that these sequence alterations are the result of polymorphisms between the fly strain used to construct the embryonic cDNA library and the strain used for the RT–PCR experiments. Because the genomic library and the cDNA library were constructed from the same strain of *Drosophila*, only the RT–PCR fragments contain the variant sequences. The cDNA sequence predicted from the genomic sequence has been deposited in GenBank (accession no. AF146362).

We performed rapid amplification of cDNA ends (RACE) to verify the 5' and 3' ends of the *Drosophila HD* transcript obtained from the sequencing of cDNA clones. We obtained two fragments from the 5' RACE experiments, one having the same 5' end as LD23533 and the other having an additional 25 nucleotides at its 5' end. Therefore, there may be two transcriptional start sites for the *Drosophila HD* gene. Only one product was generated in the 3' RACE experiments. The 3' end of this fragment is identical to that of the two cDNA clones identified from the cDNA library screen. A polyadenylation signal [AATAAA (32)] is located 18 nucleotides upstream of the 3' end, making the entire predicted 3'-untranslated region (UTR) of *Drosophila HD* 737 bp in length. In contrast, human and mouse *HD* have two transcripts produced by alternative polyadenylation (33). In human *HD*, the 3'-UTRs of the two transcripts are 600 and 3921 bp in length.

Genomic structure of the *Drosophila HD* gene

We used four-color fluorescent, dideoxy DNA sequencing to determine the complete contiguous sequence of a 43 125 bp region of genomic DNA (GenBank accession no. AF147779) encompassing the *Drosophila HD* gene and its flanking segments. This sequence contig includes 3.2 kb of sequence upstream of the putative transcription start sites and 1.8 kb of sequence downstream of the predicted transcription stop site. The *Drosophila HD* gene is spread over a much smaller genomic region than is human *HD* (5,24), which covers ~170 kb of the genome. However, it is considerably larger than the *Fugu HD* gene, which spans only 23 kb (24).

The *Drosophila HD* cDNA is 11 579 bp in length. As deduced from the genomic sequence, there are only 29 exons compared with the 67 exons seen in all vertebrate sequences examined so far (24). The *Drosophila HD* exons range in length from 82 to 1151 bp compared with a range of 48 to 341 bp for exons from human *HD*. Of the 38 kb of transcribed sequence in *Drosophila HD*, 26 534 bp encodes intron sequences. In *Drosophila* genes, most introns are relatively small, with the median length being 79 bp (34). However,

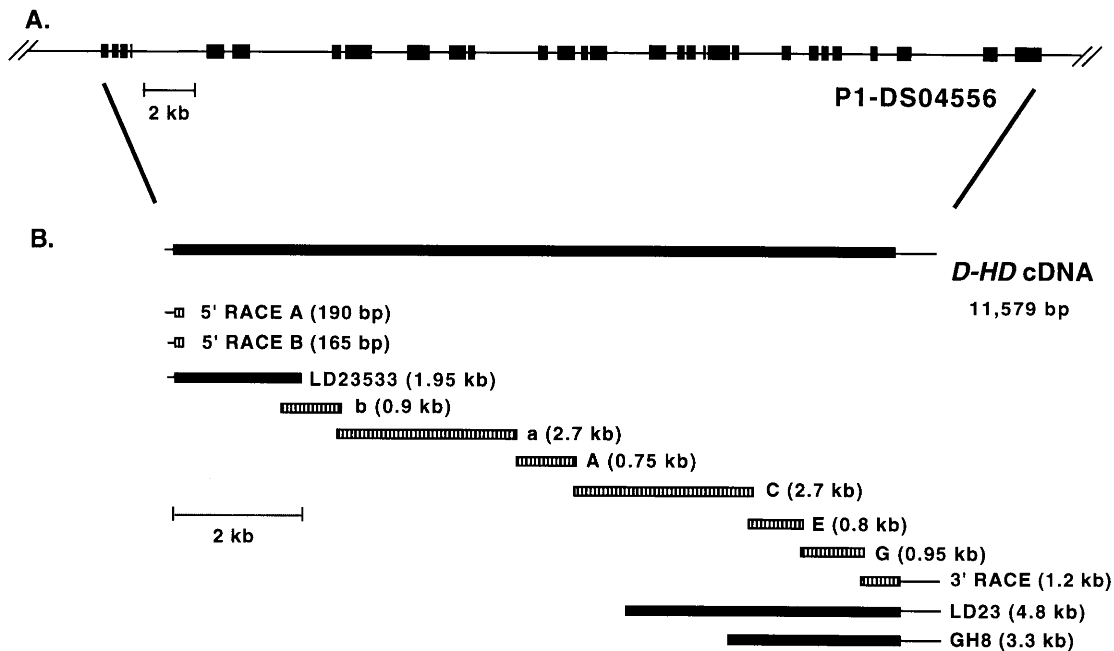

Figure 1. *Drosophila HD* genomic region and cDNA contig. (**A**) The 43 125 kb genomic region of P1 clone DS04556 showing the intron–exon structure of *Drosophila HD*. The solid segments indicate exons. (**B**) The *Drosophila HD* cDNA. The 5'- and 3'-untranslated regions of this 11 579 bp cDNA are indicated by lines, whereas coding regions are shaded. Coding regions of cDNA fragments are shown as solid boxes; RT–PCR products are hatched boxes. Clone LD23533, the clone initially identified in a BLAST search of dbEST, is likely to be the result of a mispriming event (see Materials and Methods).

the introns of *Drosophila HD* are much larger. Only three of the introns from the *Drosophila HD* sequence are <80 bp in length and there are 10 introns of >1000 bp in length (Table 1).

Predicted amino acid sequence of *Drosophila* HD

The *Drosophila HD* cDNA is predicted to encode a protein of 3583 amino acids and to have a mol. wt of 394 kDa. An in-frame stop codon is located 36 nucleotides upstream of a potential initiator methionine codon at nucleotide 91. The *Drosophila* HD protein is thus predicted to be several hundred amino acids larger than the other members of the HD family. When the *Drosophila* and vertebrate huntingtins are aligned, most of the additional amino acids from the *Drosophila* sequence appear within the N-terminal one-third of the protein. In human huntingtin, the glutamine repeats are located in the extreme N-terminus, beginning at amino acid 18. While the low similarity of *Drosophila* and human HD proteins in this region makes alignment difficult, *Drosophila* HD appears to have at most one glutamine, suggesting that the glutamine stretch is a feature unique to vertebrates and may not be necessary for the normal function of the protein. Indeed, even in vertebrates, it appears that a long stretch of continuous glutamine residues near the N-terminus of huntingtin is not important for function, as mice have only seven and zebrafish only four glutamines in this position. The proline-rich stretch immediately following the glutamines in mammalian HD proteins is completely absent in *Drosophila* HD.

Expression of *Drosophila HD* at different developmental stages

We looked for the expression of the *Drosophila HD* transcript at different developmental stages by using northern blot analysis. A transcript of the predicted size, ~12 kb in length (Fig. 2), was detected in all embryonic stages examined, in third-instar larvae and in adults. A second, fainter band of ~5 kb was also observed when a 5' riboprobe was used to probe the blot, but not when riboprobes from other regions of *Drosophila HD* were used. The transcript is expressed widely, which is similar to the case in mammalian *HD* genes (35). We detect a hybridization signal only in poly(A)$^+$ fractions and not in total RNA from *Drosophila* tissues. The *Drosophila HD* transcript is present at very low levels, consistent with the results of our embryonic cDNA library screens where we found only three partial length positive clones among 3×10^6 plaques screened.

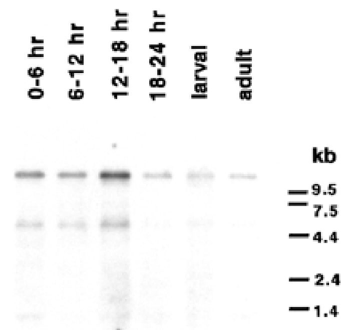

Figure 2. Northern blot analysis of poly(A)$^+$ RNA fractions at various developmental stages. The lane assignments are as follows: lane 1, 0–6 h embryos; lane 2, 6–12 h embryos; lane 3, 12–18 h embryos; lane 4, 18–24 h embryos; lane 5, larvae; lane 6, adult. The blot was probed with an antisense riboprobe corresponding to nucleotides 114–1045 from *Drosophila HD*. A 3 µg aliquot of poly(A)$^+$ RNA was loaded in each lane.

Table 1. Intron–exon organization of the *Drosophila HD* gene

Exon no.	Exon size (bp)	5' Splice donor consensus[a]		Intron size (bp)	3' Splice acceptor consensus[b]		Codon phase[c]	Amino acid[d]
		(C/A)AG	gt(a/g)agt		(c/t)ag	G		
1	164	TGC**AG**	**gt**ggga	266	cac**ag**	CCAAA	II	25 Ser
2	258	CAG**AG**	**gtaag**g	63	cac**ag**	ATCCC	II	111 Arg
3	217	TCG**AG**	**gt**agaa	61	tgc**ag**	ACTTT	0	183/184 Glu/Thr
4	130	GATG**G**	**gtaag**a	2934	tgc**ag**	AGCTC	I	227 Glu
5	669	TGA**AG**	**gt**atgg	434	tcc**ag**	GGGGA	I	450 Gly
6	603	TGC**AG**	**gtgag**c	3494	tcc**ag**	CTTCG	I	650 Ala
7	324	CAAT**G**	**gtaagt**	58	ctt**ag**	CTAAT	I	759 Ala
8	923	TGAAG	gtaagt	1640	tgc**ag**	GGCTT	0	1066/1067 Lys/Gly
9	771	ATG**AG**	**gtaag**c	1009	tgc**ag**	GCGCA	0	1323/1324 Glu/Ala
10	591	TGC**AG**	**gtaaat**	88	cac**ag**	TTTGG	0	1520/1521 Gln/Phe
11	184	GGCAG	gtgagt	2822	tgc**ag**	GTGCC	I	1582 Gly
12	278	ACAAG	gtaagt	447	ccc**ag**	ATAAC	0	1674/1675 Lys/Ile
13	552	TGAC**G**	**gtaagt**	377	ttc**ag**	CTCTT	0	1858/1859 Thr/Leu
14	158	GAT**AG**	**gtaagt**	104	tat**ag**	AAATG	II	1911 Arg
15	622	AGC**AG**	**gtgggt**	1950	tgc**ag**	GGCGT	0	2118/2119 Gln/Gly
16	678	GTGC**G**	**gtaagt**	601	tgc**ag**	GGCTG	0	2344/2345 Ala/Gly
17	168	GCAG**G**	**gtaagt**	89	ccc**ag**	CAATC	0	2400/2401 Arg/Gln
18	290	CTTCT	**gtaagt**	366	cgc**ag**	ACTGA	II	2497 Leu
19	82	CGC**AG**	**gtaaat**	135	gac**ag**	CTAAT	0	2524/2525 Gln/Leu
20	885	CATT**G**	**gtaagt**	85	tcc**ag**	TCGCT	0	2819/2820 Leu/Ser
21	188	AAG**AG**	**gtgagt**	1823	tcc**ag**	GCTCT	II	2882 Arg
22	249	ATT**AG**	**gtgagt**	930	ttc**ag**	CCTGA	II	2965 Ser
23	364	TCG**AG**	**gtaagt**	240	gac**ag**	CTGAA	0	3086/3087 Glu/Leu
24	138	ATC**AG**	**gtaagt**	209	tac**ag**	CACAT	0	3132/3133 Gln/His
25	263	GAG**AG**	**gtgagt**	1312	tgc**ag**	CCCCC	II	3220 Ser
26	175	TGC**A**T	**gtgagt**	1115	gac**ag**	GGCCT	0	3278/3279 His/Gly
27	235	TGTG**G**	**gtgagt**	3164	ctc**ag**	GCTCG	I	3357 Gly
28	269	ATC**AG**	**gtgagt**	718	cgc**ag**	GTCTT	0	3446/3447 Gln/Val
29	1151							

[a]The DNA sequence around the 5' intron–exon boundary, with exon sequences in upper case and intron sequences in lower case. Nucleotides which match the consensus sequence are in bold.
[b]The DNA sequence around the 3' intron–exon boundary, designated in the same manner as described above.
[c]Phase 0 introns are introns that do not split codon triplets, phase I introns are inserted after the first nucleotide of the triplet, and phase II introns are inserted after the second nucleotide.
[d]The amino acid residues encoded at the splice sites and their number in the translated cDNA sequence.

Alignment of *Drosophila* HD with other huntingtins

A hallmark of the vertebrate HD proteins has been a very high level of conservation at the amino acid level, with human and zebrafish huntingtins, for example, sharing 70% identity (25). We used the CLUSTAL W algorithm (36) with BLOSUM62 as the scoring matrix to align the *Drosophila* HD protein with zebrafish and human huntingtins. The resulting multiple sequence alignment identified three relatively large areas and two smaller segments as regions of highest similarity between *Drosophila* HD and the vertebrate family of huntingtin proteins (Fig. 3). The first of these, region a, encompasses amino acids 1–314 of *Drosophila* huntingtin in an area that includes what would be the glutamine-rich region from human huntingtin. In this region, there is 27% identity and 53% similarity with human huntingtin. The two smaller regions, regions b and d, span amino acids 794–839 and 2862–2944, respectively, and may define small motifs of functional importance. Region c includes amino acids 1650–2072 from the center of the protein. Region e (amino acids 3066–3504) lies at the extreme C-terminus and shares 22% identity and 49% similarity with the corresponding human region. A pairwise alignment performed between human and fly sequences in this region gives a BLASTP probability score of 4×10^{-21}.

To optimize our multiple sequence alignment, we adjusted the gap penalty parameters of CLUSTAL W and experimented with different scoring matrices (see Materials and Methods). In all of the alignments we generated, *Drosophila* HD sequences which lie outside of the five highly similar regions showed only poor overall similarity to vertebrate huntingtins. For example, the N-terminal one-third of the *Drosophila* protein,

excluding regions a and b, aligns especially poorly with the vertebrate proteins with many gaps introduced, most likely because this area contains most of the additional residues which make *Drosophila* huntingtin larger than vertebrate huntingtins. Nevertheless, the fact that regions a, c and e define large contiguous stretches of high conservation and are distributed over the entire length of the *Drosophila* HD protein argues that the gene reported here is very likely to encode a *Drosophila* homolog of the huntingtin family.

DISCUSSION

In this study, we have characterized a putative *Drosophila* homolog of the *HD* gene. The multiple sequence alignments that were performed between *Drosophila* HD and vertebrate sequences show a high degree of similarity in three large, contiguous regions at the N-terminus, middle and C-terminus of the protein. In the ~1200 amino acids that comprise these regions, there is an overall identity of 24% and similarity of 49% between fly and human sequences. The *Drosophila HD* transcript is expressed at all developmental stages tested from early embryogenesis to adulthood, a pattern that mirrors expression of the mammalian *HD* transcripts. The segments of amino acids in the *Drosophila* HD protein that are most highly similar to the vertebrate huntingtins are located in the same order as those found in the vertebrate proteins and appear throughout the entire length of the protein. These properties, and the fact that the *Drosophila* HD protein and the vertebrate huntingtins are unusually large proteins, suggest that that they not only share homologous domains, but are likely to be true orthologs.

We did not isolate other *HD*-like genes in our cDNA and genomic screens. Similarly, we did not observe any signals outside of the 98E1–E2 region where *Drosophila HD* maps in our chromosomal *in situ* hybridization experiments with *Drosophila HD* probes. We therefore consider it unlikely that there are other *Drosophila* huntingtins, suggesting that *Drosophila HD* is the *Drosophila* ortholog of *HD*.

By adding *Drosophila* huntingtin to the multiple alignment analysis, we were able to identify five relatively discrete regions that are highly conserved areas of the huntingtin family of proteins. To investigate the possibility that these regions encode functional domains or motifs present in existing databases, we used the *Drosophila* HD sequence as a query and employed search algorithms such as PSI-BLAST to search for domains that might not be uncovered with conventional BLAST searches. We also queried the BLOCKS and PROSITE databases of protein domains and motifs with the *Drosophila* HD sequence, and with motifs that were identified from the multiple sequence alignment we generated. None of these database searches identified any new domains or motifs in the huntingtin protein family. Nonetheless, the high degree of conservation in the five highly similar regions as compared with the remainder of the protein suggests that these areas are of functional importance. This issue can be explored by using biochemical approaches in conjunction with transformation-based experiments in the fly.

The complete sequences of the genomes of *Saccharomyces cerevisiae* and *Caenorhabditis elegans* have become available recently. If *HD*-like genes exist in these organisms, they might be identified by querying the yeast and worm databases with an *HD* sequence from a lower eukaryotic species such as *Drosophila*. We therefore used the *Drosophila HD* sequence to search for *HD*-related genes in both *S.cerevisiae* and *C.elegans*. Although we identified several large predicted proteins from *C.elegans*, they all had probability scores >0.5 when aligned with *Drosophila* HD, and in every case a large number of gaps were introduced into the alignment. Our searches thus found no compelling matches to the huntingtin family of proteins in either yeast or nematode databases.

In the multiple sequence alignment of huntingtin proteins, the N-terminus emerges as one of the most highly conserved regions, similar in both human and *Drosophila*. This is intriguing given that N-terminal huntingtin fragments have been detected in brains from HD patients (11) and that cells transfected with N-terminal huntingtin constructs containing expanded polyglutamine stretches undergo apoptosis and aggregate formation more readily than transfected full-length proteins with a similar polyglutamine expansion (37). It has been suggested that the presence of several consensus caspase cleavage sites at the N-terminus of mammalian huntingtin could facilitate cleavage of the N-terminus *in vivo* (38,39). It is unclear whether a fly huntingtin would be expected to be cleaved similarly. Although the *Drosophila HD* gene product does not contain a polyglutamine tract, this may not be a prerequisite for cleavage as cleavage occurs in the normal human huntingtin *in vitro* (39). It is noteworthy that we did not find any consensus caspase cleavage sites in the N-terminus of *Drosophila* HD. Perhaps the *Drosophila* protein is not cleaved by caspases or, if cleavage occurs, it is achieved by some other mechanism.

In the normal human *HD* gene product, the polyglutamine stretch at the N-terminus has an average of 18 glutamines (5). Mice have only seven glutamines and fish have only four glutamines at the N-terminus. While there are glutamines present in the N-terminus of *Drosophila* HD, there is no identifiable stretch of even as few as two consecutive glutamine residues in the glutamine repeat region (Fig. 3). Therefore, we speculate that the glutamine tract was acquired by vertebrate huntingtins over evolutionary time, with a pronounced expansion occurring in the human protein. While the glutamine stretch may impart a function to vertebrate huntingtins that is not observed in the fly, it is also possible that the glutamines are not required for the normal biological function of huntingtin. There are many examples of large CAG repeats in *Drosophila* transcripts (40), so the lack of a predicted polyglutamine tract in the N-terminus of *Drosophila* HD is not because of a general absence of long glutamine and non-coding CAG stretches in flies.

The human *HD* gene product also has a proline-rich stretch of sequence just distal to the polyglutamine tract (5). This stretch is of unknown function, but may play a role in aggregation of huntingtin protein via interactions with the SH3 domain-containing protein SH3GL3. SH3GL3 was found to promote the formation of polyglutamine-containing aggregates in transfected COS cells, and the proline-rich region was found to be essential for an interaction between SH3GL3 and HD exon 1 in yeast two-hybrid studies (41). A proline-rich stretch of sequence was not found in the N-terminus of the zebrafish protein (25), and our results here indicate that it is also not present in the *Drosophila* protein.

Recently, *Drosophila* was shown to be a useful system for studying glutamine repeat-induced neurodegeneration. In these experiments, gene segments encoding long stretches of polyglutamine were injected into flies, producing a neuro-

degenerative phenotype (42,43). In addition to clarifying the normal biological function of huntingtin, the availability of a *Drosophila* homolog of *HD* could further contribute to our understanding of the molecular aspects of the pathology of HD. The future availability of mutations in *Drosophila HD* will make it possible to analyze the corresponding mutant phenotype and conduct suppressor and enhancer screens for genetic interactors. Such genetic approaches could identify components that function together with huntingtin in a common biochemical pathway. As it has been suggested that the selective neuropathology of HD is related to the restricted expression of a gene product with which huntingtin normally associates (27,44), these future studies could shed light on the pathogenesis of this devastating disorder.

MATERIALS AND METHODS

Library construction

We isolated the P1 plasmid clone DS04556 (45) by using a maxiprep protocol (Qiagen, Valencia, CA) following the manufacturer's instructions. Purified plasmid DNA was sheared into 3 kb fragments with a HydroShearer device (GeneMachines, Palo Alto, CA). We then cloned the resulting fragments into the plasmid pJS101 to make a library, according to a protocol available from the Stanford Human Genome Center (www-shgc.stanford.edu).

RT–PCR and RACE

We isolated poly(A)$^+$ RNA from 6–12 or 12–18 h *Drosophila* embryos and performed reverse transcription reactions with Superscript II reverse transcriptase (Gibco BRL, Gaithersburg, MD) and internal primers, following the manufacturer's instructions. The resulting cDNA–RNA product was then used in PCR assays. We cloned the RT–PCR fragments b, a, A, C, E and G, shown in Figure 1, into the TA cloning vector (Invitrogen, Carlsbad, CA) for sequencing.

We mapped the 5' and 3' ends of the *Drosophila HD* cDNA by RACE with poly(A)$^+$ RNA as template by using the 5' and 3' RACE system (Gibco BRL) following the manufacturer's protocol. The amplification products were cloned into the TA cloning vector and sequenced.

We sequenced the insert of clone LD23533 by first generating a set of $\gamma\delta$ transposon insertions as previously described (Stanford Human Genome Center website: www-shgc.stanford.edu/seq/Protocols/BacMat.html).
Our sequence analysis of LD23533 revealed that the stretch of poly(A) in this cDNA clone begins immediately after a stretch of amino acids that are in the predicted ORF. The cDNA clone appears to have been the result of an oligo(dT) mispriming event during cDNA synthesis. Therefore, we performed a number of other experiments, including cDNA library screening, RT–PCR walking and RACE, to obtain clones covering the entire coding and untranslated regions.

cDNA fragments obtained by RT–PCR and RACE were sequenced on both strands from multiple clones. Sequencing was done with a dye-labeled primer sequencing kit (Amersham, Arlington Heights, IL) on an ABI Prism 377 Sequencer. BLAST searches were conducted by using the on-line service from the National Center for Biotechnology Information (NCBI). The sequences were assembled into contigs by using GeneWorks (Intelligenetics, Campbell, CA) sequence analysis software.

Northern analysis

We isolated total RNA from *Drosophila* tissues by using the Trizol (Gibco BRL) reagent following the manufacturer's instructions. Approximately 20 mg of total RNA from each developmental stage examined was loaded onto oligo(dT) cellulose columns (Gibco BRL) and poly(A)$^+$ RNA was isolated according to the manufacturer's protocol. A 3 µg aliquot of poly(A)$^+$ RNA from each stage was loaded onto a 1% formaldehyde gel and electrophoresis was carried out at 80 V for 3 h. The gel was transferred overnight by capillary transfer onto a Hybond-N+ (Amersham) nylon membrane.

Radiolabeled antisense riboprobes were synthesized from linearized DNA templates by *in vitro* transcription in the presence of [α-^{32}P]UTP. The riboprobes were then hybridized to northern blots by using Ultrahyb (Ambion, Austin, TX) hybridization buffer at 68°C and washed according to the

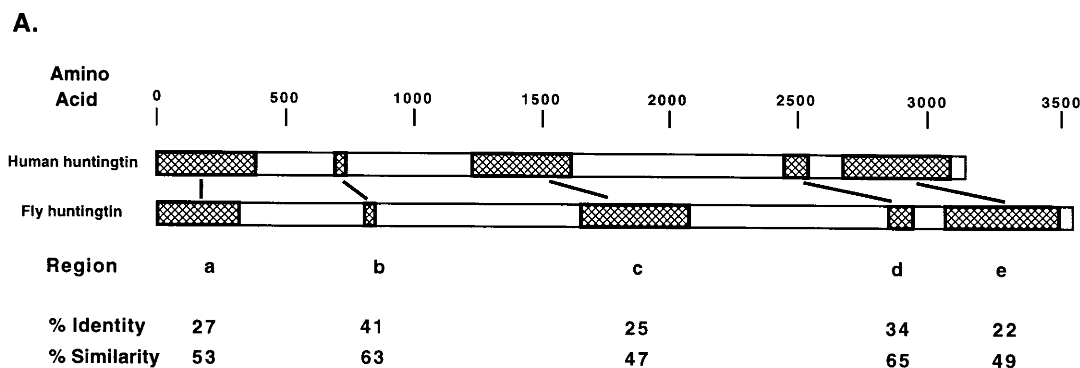

Figure 3. Regions of high similarity between the *Drosophila* HD predicted protein product and vertebrate huntingtins. (**A**) Diagram outlining the relative positions of the regions of highest similarity between *Drosophila* and human huntingtins. (**B**) Alignment of *Drosophila*, zebrafish and human huntingtins in regions a, b, c, d and e. Region a contains the polyglutamine tract from the vertebrate proteins. The alignments were generated using the CLUSTAL W multiple sequence alignment program. Residues that are identical in the *Drosophila* protein and any other proteins are shaded black; conservative differences between the *Drosophila* protein and other proteins are shaded gray.

B.

Region a

Drosophila 16
Zebrafish 23
Human 80

Drosophila 88
Zebrafish 103
Human 160

Drosophila 168
Zebrafish 183
Human 240

Drosophila 245
Zebrafish 263
Human 320

Drosophila 314
Zebrafish 325
Human 382

Region b

Drosophila 839
Zebrafish 680
Human 730

Region c

Drosophila 1728
Zebrafish 1278
Human 1306

Drosophila 1805
Zebrafish 1358
Human 1386

Drosophila 1882
Zebrafish 1433
Human 1461

Drosophila 1959
Zebrafish 1513
Human 1541

Drosophila 2039
Zebrafish 1559
Human 1587

Drosophila 2072
Zebrafish 1594
Human 1622

Region d

Drosophila 2934
Zebrafish 2482
Human 2512

Drosophila 2944
Zebrafish 2492
Human 2522

Region e

Drosophila 3134
Zebrafish 2715
Human 2744

Drosophila 3214
Zebrafish 2792
Human 2821

Drosophila 3294
Zebrafish 2862
Human 2891

Drosophila 3374
Zebrafish 2919
Human 2942

Drosophila 3448
Zebrafish 2999
Human 3022

Drosophila 3504
Zebrafish 3056
Human 3079

Figure 3. *Continued*

manufacturer's suggestions. Filters were exposed to X-ray film for 24 h at –80°C.

cDNA library screening

We isolated the cDNA clone LD23 by screening 3×10^6 phages from an oligo(dT)-primed cDNA library made from 0–22 h stage *Drosophila* embryos. The library filters were probed with a radiolabeled DNA probe synthesized from RT–PCR product G, a fragment from the 3' end of *Drosophila HD* (Fig. 1). A total of 1×10^6 phages from a second oligo(dT)-primed library derived from *Drosophila* heads were also screened with radiolabeled fragment G as probe. The cDNA clone GH8 was the longest clone among 20 positives obtained as determined by PCR. We carried out all library hybridizations at 42°C in the presence of 50% formamide, 5× SSC, 5× Denhardt's solution, 0.2% SDS and 100 µg/ml denatured salmon sperm DNA.

In situ hybridization to polytene chromosomes

Salivary glands of wandering third-instar Canton S larvae were dissected in 0.7% NaCl and fixed in 45% glacial acetic acid for 1 min. Glands were then squashed to spread the chromosomes in 3:2:1 glacial acetic acid:water:lactic acid. Approximately 100 ng of DNA from LD23533 was randomly labeled with biotinylated-16-dUTP (Enzo, Farmingdale, NY) to be used as a probe. The probe was boiled for 3 min just prior to use and allowed to hybridize to denatured chromosome squashes overnight at 58°C. The probe was detected on the chromosomes by incubating with avidin–horseradish peroxidase secondary antibody (Sigma, St Louis, MO) and developing with 0.5 mg/ml DAB/H_2O_2. Mapping was done with a Nikon microscope using a 100× objective.

Sequence analysis

For the CLUSTAL W multiple sequence alignment, we tested BLOSUM30, 45, 62, 80 and 100 as well as PAM250 and 500 as scoring matrices. We also used the Smith–Waterman algorithm to generate pairwise alignments between *Drosophila* HD and individual vertebrate family members. It was determined by visual inspection that the CLUSTAL W algorithm with BLOSUM62 as the scoring matrix yielded the best alignment. The portions of huntingtin outside regions a, b, c, d and e were of low similarity and often generated very different alignment outputs depending on the scoring matrix.

ACKNOWLEDGEMENTS

We thank Drs Jonathan Eisen, Len Pennacchio, Shirin Khambata-Ford, Joe Weiss, Keith Wharton, Gavin Sherlock, Doug Brutlag and Mike Cherry for helpful advice and discussions, and Dr Gerald Rubin for making available cDNA clone LD23533 and for providing the LD embryonic and GH adult head cDNA libraries. This work was supported by an award from the Wills Foundation to R.M.M., by a fellowship from the Huntington's Disease Society of America to C.A.K. and by a traineeship from the Stanford Genome Training Program (NIH grant HG00044) to Z.L.

REFERENCES

1. Harper, P. (1996) *Huntington's Disease*. W.B. Saunders, London, UK.
2. Albin, R.L. and Tagle, D.A. (1995) Genetics and molecular biology of Huntington's disease. *Trends Neurosci.*, **18**, 11–14.
3. Ross, C.A., Margolis, R.L., Rosenblatt, A., Ranen, N.G., Becher, M.W. and Aylward, E. (1997) Huntington disease and the related disorder, dentatorubral-pallidoluysian atrophy (DRPLA). *Medicine (Baltimore)*, **76**, 305–338.
4. Beal, M.F. (1994) Neurochemistry and toxin models in Huntington's disease. *Curr. Opin. Neurol.*, **7**, 542–547.
5. 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.
6. Kremer, B., Goldberg, P., Andrew, S.E., Theilmann, J., Telenius, H., Zeisler, J., Squitieri, F., Lin, B., Bassett, A., Almqvist, E., Bird, T.D. and Hayden, M.R. (1994) A worldwide study of the Huntington's disease mutation: the sensitivity and specificity of measuring CAG repeats. *N. Engl. J. Med.*, **330**, 1401–1406.
7. Paulson, H.L. and Fischbeck, K.H. (1996) Trinucleotide repeats in neurogenetic disorders. *Annu. Rev. Neurosci.*, **19**, 79–107.
8. Ross, C.A. (1997) Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative disease? *Neuron*, **19**, 1147–1150.
9. Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S.W. and Bates, G.P. (1996) Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, **87**, 493–506.
10. Davies, S.W., Turmaine, M., Cozens, B.A., DiFiglia, M., Sharp, A.H., Ross, C.A., Scherzinger, E., Wanker, E.E., Mangiarini, L. and Bates, G.P. (1997) Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. *Cell*, **90**, 537–548.
11. DiFiglia, M., Sapp, E., Chase, K.O., Davies, S.W., Bates, G.P., Vonsattel, J.P. and Aronin, N. (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, **277**, 1990–1993.
12. Becher, M.W., Kotzuk, J.A., Sharp, A.H., Davies, S.W., Bates, G.P., Price, D.L. and Ross, C.A. (1998) Intranuclear neuronal inclusions in Huntington's disease and dentatorubral and pallidolluysian atrophy: correlation between the density of inclusions and IT15 CAG triplet repeat length. *Neurobiol. Dis.*, **4**, 387–397.
13. Paulson, H.L., Perez, M.K., Trottier, Y., Trojanowski, J.Q., Subramony, S.H., Das, S.S., Vig, P., Mandel, J.L., Fischbeck, K.H. and Pittman, R.N. (1997) Intranuclear inclusions of expanded polyglutamine protein in spinocerebellar ataxia type 3. *Neuron*, **19**, 333–344.
14. Skinner, P.J., Koshy, B.T., Cummings, C.J., Klement, I.A., Helin, K., Servadio, A., Zoghbi, H.Y. and Orr, H.T. (1998) Ataxin-1 with an expanded glutamine tract alters nuclear matrix-associated structures. *Nature*, **389**, 971–974.
15. Klement, I.A., Skinner, P.J., Kayton, M.D., Yi, H., Hersch, S.M., Clark, H.B., Zoghbi, H.Y. and Orr, H.T. (1998) Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in *SCA1* transgenic mice. *Cell*, **95**, 41–53.
16. Saudou, F., Finkbeiner, S., Devys, D. and Greenberg, M.E. (1998) Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell*, **95**, 55–66.
17. Trottier, Y., Devys, D., Imbert, G., Saudou, F., An, I., Lutz, Y., Weber, C., Agid, Y., Hirsch, E.C. and Mandel, J.L. (1995) Cellular localization of the Huntington's disease protein and discrimination of the normal and mutated form. *Nature Genet.*, **10**, 104–110.
18. DiFiglia, M., Sapp, E., Chase, K., Schwarz, C., Meloni, A., Young, C., Martin, E., Vonsattel, J.-P., Carraway, R., Reeves, S.A., Boyce, F.M. and Aronin, N. (1995) Huntingtin is a cytoplasmic protein associated with vesicles in human and rat brain neurons. *Neuron*, **14**, 1075–1081.
19. Wellington, C.L. and Hayden, M.R. (1997) Of molecular interactions, mice and mechanisms: new insights into Huntington's disease. *Curr. Opin. Neurol.*, **10**, 291–298.
20. Burke, J.R., Enghild, J.J., Martin, M.E., Jou, Y.-S., Myers, R.M., Roses, A.D., Vance, J.M. and Strittmatter, W.J. (1996) Huntingtin and DRPLA proteins selectively interact with the enzyme GAPDH. *Nature Med.*, **2**, 347–350.

21. Barnes, G.T., Duyao, M.P., Ambrose, C.M., McNeil, S., Persichetti, F., Srinidhi, J., Gusella, J.F. and MacDonald, M.E. (1994) Mouse Huntington's disease gene homolog (Hdh). *Somat. Cell Mol. Genet.*, **20**, 87–97.
22. Lin, B., Nasir, J., MacDonald, H., Hutchinson, G., Graham, R.K., Rommens, J.M. and Hayden, M.R. (1994) Sequence of the murine Huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat. *Hum. Mol. Genet.*, **3**, 85–92.
23. Schmitt, I., Bachner, D., Megow, D., Henklein, P., Hameister, H., Epplen, J.T. and Riess, O. (1995) Expression of the Huntington disease gene in rodents: cloning the rat homologue and evidence for downregulation in non-neuronal tissues during development. *Hum. Mol. Genet.*, **4**, 1173–1182.
24. Baxendale, S., Abdulla, S., Elgar, G., Buck, D., Berks, M., Micklem, G., Durbin, R., Bates, G., Brenner, S., Beck, S. and Lehrach, H. (1995) Comparative sequence analysis of the human and pufferfish Huntington's disease genes. *Nature Genet.*, **10**, 67–76.
25. Karlovich, C.A., John, R.M., Ramirez, L., Stainier, D.Y.R. and Myers, R.M. (1998) Characterization of the Huntington's disease (HD) gene homolog in the zebrafish *Danio rerio*. *Gene*, **217**, 117–125.
26. Duyao, M.P., Auerbach, A.B., Ryan, A., Persichetti, F., Barnes, G.T., McNeil, S.M., Ge, P., Vonsattel, J.-P., Gusella, J.F., Joyner, A.L. and MacDonald, M.E. (1995) Inactivation of the mouse Huntington's disease gene homolog *Hdh*. *Science*, **269**, 407–410.
27. Nasir, J., Floresco, S.B., O'Kusky, J.R., Diewert, V.M., Richman, J.M., Zeisler, J., Borowski, A., Marth, J.D., Phillips, A.G. and Hayden, M.R. (1995) Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell*, **81**, 811–823.
28. Zeitlin, S., Liu, J.-P., Chapman, D.L., Papaioannou, V.E. and Efstratiadis, A. (1995) Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nature Genet.*, **11**, 155–163.
29. Sharp, A.H., Loev, S.J., Schilling, G., Li, S.-H., Li, X.-J., Bao, J., Wagster, M.V., Kotzuk, J.A., Steiner, J.P., Lo, A., Hedreen, J., Sisodia, S., Snyder, S.H., Dawson, T.M., Ryugo, D.K. and Ross, C.A. (1995) Widespread expression of Huntington's disease gene (IT15) protein product. *Neuron*, **14**, 1065–1074.
30. Li, S.-H., Schilling, G., Young, W.S., Li, X.-J., Margolis, R.L., Stine, O.C., Wagster, M.V., Abbott, M.H., Franz, M.L., Ranen, N.G., Folstein, S.E., Hedreen, J.C. and Ross, C.A. (1993) Huntington's disease gene (IT15) is widely expressed in human and rat tissues. *Neuron*, **11**, 985–993.
31. Hoogeveen, A.T., Willemsen, R., Meyer, N., de Rooij, K.E., Roos, R.A.C., van Ommen, G.B. and Galjaard, H. (1993) Characterization and localization of the Huntington's disease gene product. *Hum. Mol. Genet.*, **2**, 2069–2073.
32. Sheets, M.D., Ogg, S.C. and Wickens, M.P. (1990) Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res.*, **18**, 5799–5805.
33. Lin, B., Rommens, J.M., Graham, R.K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y.P. and Hayden, M.R. (1994) Differential 3' polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.*, **2**, 1541–1545.
34. Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O. and Fields, C. (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.*, **20**, 4255–4262.
35. Strong, T.V., Tagle, D.A., Valdes, J.M., Elmer, L.W., Boehm, K., Swaroop, M., Kaatz, K.W., Collins, F.S. and Albin, R.L. (1993) Widespread expression of the human and rat Huntington's disease gene in brain and nonneural tissues. *Nature Genet.*, **5**, 259–265.
36. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
37. Martindale, D., Hackam, A., Wieczorek, A., Ellerby, L., Wellington, C., McCutcheon, K., Singaraja, R., Kazemi-Esfarjani, P., Devon, R., Kim, S.U., Bredesen, D.E., Tufaro, F. and Hayden, M.R. (1998) Length of huntingtin and its polyglutamine tract influences localization and frequency of intracellular aggregates. *Nature Genet.*, **18**, 150–154.
38. Wellington, C.L., Ellerby, L.M., Hackam, A.S., Margolis, R.L., Trifiro, M.A., Singaraja, R., McCutcheon, K., Salvesen, G.S., Propp, S.S., Bromm, M., Rowland, K.J., Zhang, T., Rasper, D., Roy, S., Thornberry, N., Pinsky, L., Kakizuka, A., Ross, C.A., Nicholson, D.W., Bredesen, D.E. and Hayden, M.R. (1998) Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J. Biol. Chem.*, **273**, 9158–9167.
39. Goldberg, Y.P., Nicholson, D.W., Rasper, D.M., Kalchman, M.A., Koide, H.B., Graham, R.K., Bromm, M., Kazemi-Esfarjani, P., Thornberry, N.A., Vaillancourt, J.P. and Hayden, M.R. (1996) Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nature Genet.*, **13**, 442–449.
40. Wharton, K.A., Yedvobnick, B., Finnerty, V.G. and Artavanis-Tsakonas, S. (1985) opa: a novel family of transcribed repeats shared by the Notch locus and other developmentally regulated loci in *D.melanogaster*. *Cell*, **40**, 55–62.
41. Sittler, A., Walter, S., Wedemeyer, N., Hasenbank, R., Scherzinger, E., Eickhoff, H., Bates, G.P., Lehrach, H. and Wanker, E.E. (1998) SH3GL3 associates with the huntingtin exon 1 protein and promotes the formation of polygln-containing protein aggregates. *Mol. Cell*, **2**, 427–436.
42. Warrick, J.M., Paulson, H.L., Gray-Board, G.L., Bui, Q.T., Fischbeck, K.H., Pittman, R.N. and Bonini, N.M. (1998) Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in *Drosophila*. *Cell*, **93**, 939–949.
43. Jackson, G.R., Salecker, I., Dong, X., Yao, X., Arnheim, N., Faber, P.W., MacDonald, M.E. and Zipursky, S.L. (1998) Polyglutamine-expanded human huntingtin transgenes induce degeneration of *Drosophila* photoreceptor neurons. *Neuron*, **21**, 633–642.
44. Gusella, J.F. and MacDonald, M.E. (1994) Huntington's disease and repeating trinucleotides. *N. Engl. J. Med.*, **330**, 1400–1451.
45. Smoller, D.A., Petrov, D. and Hartl, D.L. (1991) Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75–100 kilobase pairs. *Chromosoma*, **100**, 487–494.