Phylogenetic Comparison of Huntingtin Homologues Reveals the Appearance of a Primitive polyQ in Sea Urchin

Marzia Tartari, Carmela Gissi,† Valentina Lo Sardo,* Chiara Zuccato,* Ernesto Picardi,‡ Graziano Pesole,‡ and Elena Cattaneo**

*Department of Pharmacological Sciences, University of Milan, Milano, Italy; †Department of Biomolecular Sciences and Biotechnology, University of Milan, Milano, Italy; and ‡Department of Biochemistry and Molecular Biology, University of Bari and CNR Biomedical Technology Institute, Bari, Italy

Huntingtin is a completely soluble 3,144 amino acid (aa) protein characterized by the presence of an amino-terminal polymorphic polyglutamine (polyQ) tract, whose aberrant expansion causes the progressively neurodegenerative Huntington's disease (HD). Biological evidence indicates that huntingtin (htt) is beneficial to cells (particularly to brain neurons) and that loss of its neuronal function may contribute to HD. The exact protein domains involved in its neuroprotective function are unknown. Evolutionary analyses of htt primary aa have so far been limited to a few species, but its thorough assessment may help to clarify the functions emerging during evolution. We made an extensive comparative analysis of the available htt protein homologues from different organisms along the metazoan phylogenetic tree and defined the presence of 3 different conservative blocks corresponding to human htt aa 1–386 (htt1), 683–1,586 (htt2), and 2,437–3,078 (htt3), in which HEAT (Huntingtin, Elongator factor3, the regulatory A subunit of protein phosphatase 2A, and TOR1) repeats are well conserved. We also describe the cloning and sequencing of sea urchin htt mRNA, the oldest deuterostome homologue so far available. Multiple alignment shows the first appearance of a primitive polyQ in sea urchin, which predates an ancestral polyQ sequence in a nonchordate environment and defines the polyQ characteristic as being typical of the deuterostome branch. The fact that glutamines have conserved positions in deuterostomes and the polyQ size increases during evolution suggests that the protein has a possibly Q-dependent role. Finally, we report an evident relaxing constraint of the N-terminal block in *Ciona* and drosophilids that correlates with the absence of polyQ and which may indicate that the N-terminal portion of htt has evolved different functions in *Ciona* and protostomes.

Introduction

In humans, huntingtin (htt) is a completely soluble protein of 3,144 amino acids (aa) that carries a polyglutamine (polyQ) tract in its N-terminus. When this expands over 36 units, the protein becomes toxic and Huntington's disease (HD) develops, with the subsequent preferential death of striatal GABAergic and cortical neurons and dysfunctions in other cell types and tissues. At least 8 other disease-causing proteins (Everett and Wood 2004) share the presence of an expanded polyQ with htt, and the fact that each of these diseases is characterized by the loss of a different subset of neurons is a clear indication that the normal sequences surrounding the polyQ tract play a critical pathogenic role (Cattaneo et al. 2001). The contribution of normal htt protein to HD has also aroused considerable attention because specific molecular abnormalities have been described in HD that are similar to those observed in cells and mice lacking htt (Cattaneo et al. 2005, review; Zuccato et al. 2007). A more thorough understanding of the protein's normal functions and potential domains should therefore improve our knowledge as to whether and how an expanded polyQ in the protein pathologically represses or enhances its normal function. However, one major difficulty is the fact that htt has no sequence similarity with any other known proteins and is ubiquitously expressed throughout the lifetime of humans and rodents. It is most expressed in central nervous system (CNS) neurons and the testes, and its particular enrichment in neurons suggests that it plays an important role in the nervous system. The constitutive knockout of htt in mice is embryonically lethal at gastrulation, thus indicating its necessary role early in development and specifically in nonneuronal cells. In addition, its precise removal from or overexpression in neuronal cells and mouse brain, respectively, decreases or increases neuronal cell survival (O'Kusky et al. 1999; Dragatsis et al. 2000; Rigamonti et al. 2000). The idea that htt plays a role in the brain has recently been reinforced by findings of a link between htt level and brain derived neurotropic factor, an important neurotrophin for the striatal neurons that die during the disease (Zuccato and Cattaneo 2007, review). These discoveries have led to suggest that htt acts in different cell types in order to coordinate multiple intracellular pathways (Sipione and Cattaneo 2001) and that it has evolved neuronal functions that gradually accumulate along deuterostomes and become specific to this evolutionary branch (Cattaneo et al. 2005, review).

Coincidentally, the function of htt during mammalian development seems to reflect its evolutionary steps. The early nonneuronal activity of htt in htt knockout mice can be likened to its ancestral function in species with no or a poorly organized nervous system, whereas it might have acquired activities in higher vertebrates that were important for the newly formed nervous system and essential for postmitotic neurons. In line with this view, htt from *Drosophila melanogaster* is very divergent, has no polyQ, and is characterized by 5 regions distributed along the entire length of the protein that are 20–50% conserved (Li et al. 1999). Furthermore, it has approximately 300 aa insertion in its N-terminal portion that may indicate a differently evolved function in drosophilids.

The reconstruction of htt evolution along the deuterostomes has long suffered from the fact that only partial sequences are available from a few deuterostome invertebrates, such as the tunicata *Halocynthia roretzi* and 2

Key words: huntingtin, polyQ, sea urchin, evolution, homologues, deuterostomes.

E-mails: elena.cattaneo@unimi.it; graziano.pesole@biologia.uniba.it.

Mol. Biol. Evol. 25(2):330–338. 2008
doi:10.1093/molbev/msm258
Advance Access publication November 28, 2007

echinodermata *Strongylocentrotus purpuratus* (SP) and *Heliocidaris erythrogramma* (Kauffman et al. 2003). The only exceptions are htts from *Ciona intestinalis* and *Ciona savignyi*, whose entire sequences have recently been reported by us (Gissi et al. 2006). We found that the 2 *Ciona* species have an aromatic aa group instead of a polyQ region and fewer HEAT repeats (Huntingtin, Elongator factor3, the regulatory A subunit of protein phosphatase 2A, and TOR1), these being the only consensus sequences found in htt and known to be important for protein–protein interactions (Andrade and Bork 1995). Moreover, there is an accumulation of substantial differences in the first part of the gene compared with other chordates, which suggests that its 5′ end is fast evolving.

We describe here a more distant homologue along the deuterostome branch from the sea urchin SP. Sea urchin htt has a hydrophilic NHQQ group in the same position as that of the vertebrate polyQ, and its characteristics and gene structure indicate that it is more similar to vertebrate than *Ciona* htt. We also present the first thorough bioinformatic recovery, analysis, and comparison of the available primary sequences of htt homologues in deuterostomes and protostomes, these latter represented by 4 insect species. These analyses revealed 3 main constraints of conservation along the protein in the aa portions corresponding to human htt positions 1–386 (htt1), aa 683–1586 (htt2), and aa 2437–3078 (htt3). In addition, the greater divergence of the N-terminal portion of *Ciona* and *Drosophila* htt and the lack of a polyQ in their sequences, together with the progressive increase of the polyQ size along deuterostomes and the conservation of some functionally important residues at the extreme N-terminus, suggest a specific function associated with the N-terminal portion that may have evolved along deuterostomes. Finally, our HEAT repeat analysis showed that, in addition to the ancestral repeats, there are others specific to the deuterostome or vertebrate groups, which may indicate selective pressure in guaranteeing specific protein–protein interactions.

We conclude that the glutamine insertion in the htt sequence was born at the base of the deuterostome branch and was conserved and subsequently expanded at the same time as there was a progressive refinement in the N-terminal aa environment, whereas organisms that lack the polyQ also have considerably different N-terminal portions.

Materials and Methods

Cloning and Sequencing SP Transcripts

To clone and sequence the sea urchin htt messenger, a reverse transcriptase–polymerase chain reaction (RT-PCR) cloning plan was established taking advantage of the messenger prediction produced by the sea urchin genomic data analysis. Using human htt as the query sequence, we first used TBlastN to screen the last assembly of the sea urchin genome project through the human genome sequencing center sea urchin web site. Then, a transcript prediction was produced (see Materials and Methods) to plan the RT-PCR cloning. Using total RNA from 70-h fertilized SP eggs (corresponding to about 10-dpc of mouse embryo), we produced, cloned, and sequenced 18 partially overlapping fragments of about 1,000–2,000 bp, reconstructing the entire sea urchin htt coding sequence (CDS). As each fragment was represented by a mean of 3 clones and the starting RNA represents multiple individuals, we also annotated all point variations by evaluating their frequency of appearance (see Materials and Methods for accession numbers).

On the basis of the genomic sequence, we performed a PCR at the 5′ of the messenger and a 3′ rapid amplification of cDNA ends–polymerase chain reaction (RACE-PCR) to extend the sequence to the untranslated regions (UTRs). Total RNA was retrotranscribed into cDNA using random examers (100 ng) and SuperscriptIII according to the customer protocol. For 3′RACE-PCR, a GENERACER kit (Invitrogen Carlsbad, CA) was used according to the customer protocol with 4 different gene-specific primers for the primary amplification and 3 different primers for nested PCR. All the combinations of primers produced almost the same 3 bands, which were cloned and sequenced to reveal the 3 different 3′UTRs. The primers to obtain the entire SP htt messenger are available upon request, and a total of 38 clones and 81 sequences were produced (accession numbers: AM398482–AM398562). The entire mRNA was reconstructed and the point variations of the CDS were annotated as possible allelic variants; the majority (99.5%) were polymorphisms present in the normal population (no change in aa). Base calling was defined as the most frequent base in the sequenced clones, excluding the variations present in a single clone.

Determination of htt Gene Structure in SP

The htt gene structure in SP was determined by aligning the experimentally inferred cDNA sequence with the available genome data at the National Center for Biotechnology Information (NCBI) repository. The exon–intron boundaries were refined using Spidey's alignment. Supplementary table 1 (Supplementary Material online) shows the accession numbers of the different genome sequences used to infer the exon–intron structure of the htt gene.

Sequence Retrieval and Alignment

In order to recover all the available protein htt homologues, using NCBI and University of California, Santa Cruz resources, we carried out BLAT and TBlastN searches against genomic databases of the 19 organisms listed in supplementary table 3 (Supplementary Material online), using as probes the htt proteins of human (P42858) and *D. melanogaster* (AAF03255). The sequences in the multiple alignment were computationally predicted from genomic data of 6 species, that is, *Apis mellifera*, *Tribolium castaneum*, *Drosophila pseudoobscura*, *Canis familiaris*, *Bos taurus*, and *Monodelphis domestica*. These species were selected as insect representatives or because of their crucial position in the mammalian phylogenetic tree. The matching genomic contigs of the above species were analyzed by several bioinformatics tools, such as Genscan (http://genes.mit.edu/GENSCAN.html), Genomescan (http://genes.mit.edu/genomescan.html), and Genewise (http://www.ebi.ac.uk/Wise2/) to obtain reliable predictions of the mRNA-coding

portion and of the exon–intron gene structure. The htt predictions/mRNA sequences of remaining species are from Gissi et al. (2006).

The multiple alignment of all complete htt protein available, listed in supplementary table 3 (Supplementary Material online), has been constructed starting from the alignment published in Gissi et al. (2006). The program PROMALS (Pei and Grishin 2007) helped the alignment update and was followed by manual adjustments using Seal and SeaView programs.

Identification of Conserved Blocks

Given a multiple sequence alignment (MSA), the conserved blocks were identified using our own BlockP software, which detects clusters of conserved sites in a window whose size is defined by the user.

Conserved MSA sites are classified into 3 categories: 1) semiconservative, 2) conservative, and 3) identical (labeled in the alignment in supplementary fig. 1 [Supplementary Material online] as "." or ":" or "*" as detailed in the ClustalW documentation) (http://www.ebi.ac.uk/clustalw). Assigning arbitrary scores of 1, 2, and 3 to semiconservative, conservative, and identical sites (nonconserved sites = 0), an overall quality score for the alignment can be computed as follows:

$$Q(\mathrm{MSA}) = \frac{1}{L}\sum_{i=1}^{L} \mathrm{score}_i,$$

where L is the length of the MSA.

The blocks conserved in the MSA were detected using a threshold conservation score of 1.0 and a minimum window size of 10 sites. All the overlapping windows fulfilling above criteria define the final block size.

Identification of HEAT Repeats

HEAT repeats were identified using the REP program (Andrade et al. 2001) available at http://www.embl-heidelberg.de/~andrade/papers/rep/search.html and the htt protein multiple alignment shown in supplementary figure 3 (Supplementary Material online). The HEAT score and *E* value were calculated by running the program in the single-sequence mode.

If very significant HEAT repeats ($P < 10^{-5}$) were detected in a specific aligned region of one or more species, we assumed the presence of the HEAT repeat in all of the other species provided that they showed unambiguous and significant alignment in the same region.

Phylogenetic Analysis

The phylogenetic analysis was carried out by means of MrBayes program using the JTT model (Jones et al. 1992, Ronquist and Huelsenbeck 2003) with the invariant plus gamma option. One cold and 3 incrementally heated chains were run for 1,000,000 generations. The trees were sampled every 100 generations from the last 500,000 generated (well after chain stationarity), and 5,000 trees were used for inferring posterior probabilities.

Results

Sea Urchin htt mRNA Cloning and Sequencing

Full-length sea urchin htt mRNA, determined as described in the Material and Methods and expressed in 70-h fertilized eggs of SP, is 10,532 nt long and consists of a 5′UTR of 78 nt, a CDS of 9,159 nt, and a 3′UTR of 1,371 nt (deposited in GenBank under accessions AM398482–AM398562).

Interestingly, the 5′UTR contains 2 consecutive out-of-frame upstream AUGs that could potentially affect the translation efficiency of htt mRNA (Iacono et al. 2005).

A 55-nt repeat is located in the 3′UTR just downstream of the stop codon. The 3′RACE experiments (see Materials and Methods) showed 3 alternative polyadenylation sites generating 3′UTRs of 332 nt, 876 nt, and 1,417 nt (see supplementary fig. 1, Supplementary Material online). The longest 3′UTR was further confirmed by an expressed sequence tag (EST) from a sea urchin radial nerve cDNA library (GenBank accession number: EC438761). Alignment of the alternative UTRs with the corresponding genomic clone (GenBank accession number: AAGJ02123117) revealed several indels, the largest of which overlapped the second copy of the 55-nt repeat (see supplementary fig. 2, Supplementary Material online), thus suggesting a high rate of sequence variability in this region. In fact, no significant conserved region was found by database search of the 3′UTR against currently available cDNA/ESTs from other species.

Gene Structure

Comparison of the inferred htt transcript with the available genomic clones indicated that SP htt mRNA consists of at least 58 exons, against the 67 exons in humans and other vertebrates, and the 61 coding exons in the ascidian *C. intestinalis* (Gissi et al. 2006) (see supplementary table 1 [Supplementary Material online] for the details of the genomic clones used in this analysis). There are 3 gaps in the available SP genomic sequence corresponding to putative exons 30, 35, and 53. All of the splicing sites obey the canonical GT/AG rule. It is interesting to note that the gene structure of SP htt is much more similar to that of vertebrate htt than it is to that of the tunicate *C. intestinalis*. Despite its higher genetic divergence, 16/58 positionally conserved exons have exactly the same length and phase and 51/58 share the same phase (see supplementary table 2, Supplementary Material online), whereas only 5 exons positionally conserved between vertebrate and *C. intestinalis htt* (Gissi et al. 2006) have identical lengths. At least 3 intron gains can be predicted in human *htt* as human exons 11, 12, and 13 (totaling 546 bp) correspond to the 546-bp exon 10 in sea urchin and human exons 22 and 23 (totaling 268 bp) correspond to the 268-bp exon 17 in SP. As such intron gains are also observed in *htt* of other analyzed vertebrates, it can be argued that the intron gain events predate the radiation of vertebrates. These intron gain/loss events may

Table 1
Overall Sequence Conservation of the htt Multialignment Including Only the 11 Vertebrate, 13 Chordate, 14 Deuterostome, or all 18 Metazoan Sequences

	Alignment Size	Identical	Conservative	Semiconservative	% Conserved Residues	Conserved Blocks	Alignment Quality
Vertebrata	3262	1805	553	201	0.87	2829	2.06
Chordata	3405	721	674	272	0.56	1916	1.11
Deuterostomia	3452	545	619	231	0.46	1585	0.90
Metazoa	4346	151	363	149	0.15	646	0.31

NOTE.—For each multialignment, the table shows alignment size; the number of identical, conservative, and semiconservative sites; the percentage of conserved residues; the total length of the conserved blocks (detected as described in Materials and Methods); and the alignment quality (see Materials and Methods).

provide critical information concerning the evolutionary relationships between species at large evolutionary distances and help to reconstruct the evolutionary history of htt gene structure.

The htt genes of insects have much fewer exons and a much more heterogeneous gene structure. We estimated 24 exons in *T. castaneum*, 13 in *A. mellifera* (honeybee), and 29 in the 2 species of *Drosophila* (*melanogaster* and *pseudoobscura*).

Comparative Analysis of htt Protein in Metazoa

The SP htt protein is 3,052 aa long, slightly shorter than the human homologue (3,144 aa), but longer than the tunicate counterparts. The length of htt in insects is remarkably heterogeneous, ranging from 2,679 aa in *T. castaneum* to 3,758 in *D. pseudoobscura*. This suggests that structural (and possibly functional) constraints are stronger in vertebrate and sea urchin htts than in insect htts, which also show greater sequence variability despite overall conservation along the entire protein (see below).

In order to study the evolution of htt in metazoa, we constructed a high-quality multialignment of all available htt sequences, including some protein data that is still not available in public databases but could be computationally inferred from available genomic sequences. The htt multialignment, shown in supplementary figure 3 (Supplementary Material online), includes 17 sequences (listed in supplementary table 3, Supplementary Material online) from vertebrates (11), tunicates (2), and insects (4).

Table 1 shows the overall sequence conservation of htt protein calculated from multialignments including only sequences from vertebrates, chordates, deuterostomes, or all available metazoans.

InterProScan analysis of human and sea urchin htt proteins detected 3 pairs of PRINTS blocks (ID PR00375) diagnostic of htt and totaling 124 aa (blue boxes in fig. 1, panel B) in both species. Furthermore, SMART analysis detected several stretches of intrinsic disorder regions (Linding et al. 2003) totaling 567 aa in human htt and 532 aa in SP htt (green boxes in fig. 1, panel B). However, htt conservation across lineages from vertebrate to invertebrates is much more extended than the annotated PRINTS domains. Sequence conservation spans the entire protein length, with several conserved blocks. The conserved blocks (identified as described in Materials and Methods) showed a high degree of overall conservation throughout the protein length in vertebrates (fig. 1, panel C) and 3 main conserved regions in both deuterostomes (fig. 1, panel D) and metazoans (fig. 1, panel E) that largely overlapped HEAT clusters: an N-terminal domain overlapping HEAT

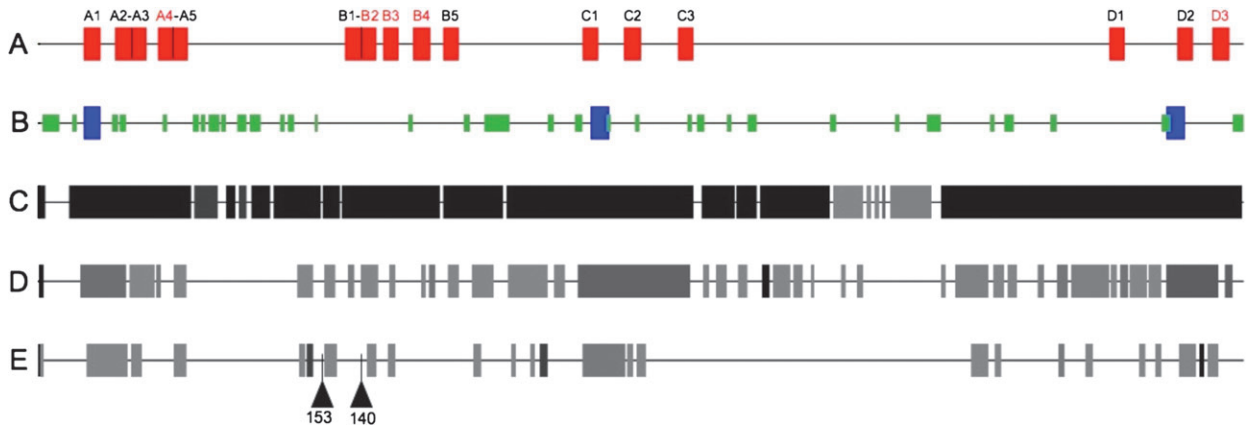

FIG. 1.—The 16 HEAT repeats identified in the htt proteins organized into 4 clusters (A1–A5, B1–B5, C1–C3, and D1–D3). The red labels denote not fully conserved HEAT elements (see table 3) (panel A). Prints signatures (blue boxes) and intrinsic disorder regions (green boxes) (panel B). Conserved blocks detected as described in the Material and Methods section in vertebrate (panel C), deuterostome (panel D), and all metazoa (panel E) htts. Black triangles denote 2 large expansion regions of 153 and 143 aa in drosophilids.

Table 2
Pairwise Protein Distances (lower) and Standard Deviations (upper) Calculated According to the JTT Matrix on the Ungapped Multiple Alignment Positions by the MEGA Software (Tamura et al. 2007)

	Hsa	Rno	Mmu	Ssc	Bta	Cfa	Mdo	Gga	Dre	Tni	Fru	Csa	Cin	Spu	Tca	Ame	Dps	Dme
Hsa	0	0,005	0,005	0,006	0,006	0,005	0,007	0,007	0,01	0,011	0,01	0,031	0,033	0,027	0,039	0,035	0,073	0,074
Rno	0,072	0	0,003	0,007	0,007	0,006	0,007	0,007	0,011	0,011	0,011	0,033	0,034	0,028	0,041	0,035	0,073	0,074
Mmu	0,064	0,021	0	0,007	0,006	0,006	0,007	0,007	0,011	0,011	0,011	0,032	0,033	0,027	0,039	0,033	0,073	0,074
Ssc	0,088	0,106	0,101	0	0,006	0,006	0,009	0,008	0,011	0,011	0,011	0,033	0,034	0,028	0,039	0,037	0,073	0,074
Bta	0,083	0,101	0,094	0,084	0	0,006	0,008	0,008	0,011	0,011	0,011	0,032	0,033	0,028	0,038	0,036	0,071	0,073
Cfa	0,063	0,085	0,079	0,086	0,086	0	0,008	0,007	0,011	0,011	0,011	0,032	0,033	0,027	0,039	0,036	0,071	0,073
Mdo	0,094	0,111	0,102	0,127	0,126	0,104	0	0,006	0,01	0,01	0,01	0,034	0,033	0,028	0,039	0,035	0,07	0,07
Gga	0,108	0,125	0,119	0,14	0,138	0,119	0,085	0	0,01	0,009	0,009	0,033	0,033	0,027	0,039	0,034	0,073	0,072
Dre	0,221	0,24	0,233	0,245	0,248	0,237	0,203	0,198	0	0,009	0,009	0,035	0,035	0,028	0,039	0,036	0,071	0,071
Tni	0,223	0,242	0,235	0,247	0,247	0,235	0,212	0,201	0,133	0	0,003	0,035	0,036	0,028	0,04	0,038	0,069	0,07
Fru	0,221	0,244	0,235	0,246	0,246	0,231	0,207	0,198	0,13	0,027	0	0,035	0,036	0,028	0,041	0,037	0,069	0,07
Csa	1,081	1,096	1,091	1,085	1,084	1,089	1,073	1,067	1,071	1,087	1,088	0	0,011	0,043	0,054	0,051	0,086	0,087
Cin	1,073	1,091	1,082	1,08	1,092	1,085	1,06	1,056	1,068	1,086	1,08	0,271	0	0,042	0,055	0,05	0,086	0,084
Spu	0,832	0,851	0,854	0,852	0,855	0,859	0,842	0,827	0,817	0,819	0,813	1,212	1,195	0	0,043	0,038	0,073	0,072
Tca	1,302	1,321	1,308	1,301	1,299	1,314	1,297	1,3	1,307	1,283	1,301	1,548	1,537	1,32	0	0,03	0,074	0,073
Ame	1,167	1,181	1,165	1,165	1,172	1,169	1,155	1,158	1,161	1,167	1,178	1,483	1,457	1,178	0,93	0	0,074	0,074
Dps	2,197	2,203	2,19	2,172	2,19	2,191	2,158	2,179	2,19	2,168	2,159	2,424	2,426	2,253	2,117	2,125	0	0,01
Dme	2,189	2,202	2,19	2,165	2,186	2,179	2,159	2,167	2,165	2,147	2,147	2,436	2,394	2,246	2,101	2,11	0,157	0

NOTE.—*Homo sapiens* (Hsa); *Rattus norvegicus* (Rno); *Mus musculus* (Mmu); *Sus scrofa* (Ssc); *Bos taurus* (Bta); *Canis familiaris* (Cfa); *Monodelphis domestica* (Mdo); *Gallus gallus* (Gga); *Danio rerio* (Dre); *Tetraodon nigroviridis* (Tni); *Fugu rubripes* (Fru); *Ciona savignyi* (Csa); *Ciona intestinalis* (Cin); *Strongylocentrotus purpuratus* (Spu); *Tribolium castaneum* (Tca); *Apis mellifera* (Ame); *Drosophila pseudoobscura* (Dps); and *Drosophila melanogaster* (Dme).

cluster A, a central domain overlapping HEAT clusters B and C, and a C-terminal domain overlapping HEAT cluster D. These 3 domains correspond to residues 1–386 (htt1), 683–1,586 (htt2), and 2,437–3,078 (htt3) in human htt (see supplementary table 4, Supplementary Material online).

Table 2 shows the corrected (Kimura's [1980] method) and uncorrected pairwise distances calculated on the ungapped sites of the multialignment in supplementary figure 3 (Supplementary Material online). It can be seen that there is a striking increase in evolutionary constraints in vertebrates. The average genetic distance between proteins of mammals and fish, separated by 450 MYA (Blair-Hedges and Kumar 2003), is 0.23 substitutions per site, whereas the genetic distance between echinoderms and vertebrates is 0.83 substitutions per site (almost 4-fold) although the divergence time is less than double (Blair-Hedges and Kumar 2003). It is also worth noting that there is an even greater distance between vertebrate and tunicate htt proteins (an average of 1.10 substitutions per site) despite the fact that tunicates are Chordata and thus more closely related to vertebrates than echinoderms (also see the phylogenetic tree in fig. 2). An accelerated rate of evolution of htt protein can therefore be observed along the tunicate lineages, but there is an even higher acceleration rate along the drosophilid lineages whose genetic distance from vertebrate htt proteins is 2.2 substitutions per site, as against the 1.18 substitutions per site of *A. mellifera* and the 1.31 substitutions per site of *T. castaneum* (also see the branch lengths in the tree in fig. 2). Furthermore, drosophilid htt proteins contain many unique stretches that are not present in any other taxa (see supplementary fig. 3, Supplementary Material online).

The pairwise distances calculated separately on the 3 domains (htt1–htt3) were not very different although the N- and C-terminal domains were slightly more conserved than the central domain (see supplementary table 4, Supplementary Material online). It is interesting to note that congeneric pairwise comparisons of *C. intestinalis* versus *C. savignyi* and *D. melanogaster* versus *D. pseudoobscura* showed a more conserved C-terminal than N-terminal domain. The pairwise distance between the 2 *Ciona* species is 0.22 substitutions per site in the N-terminal domain and 0.18 substitutions per site in the C-terminal domain and that between the 2 drosophilids is 0.06 substitutions per site in the C-terminal domain and 0.12 substitutions per site in the N-terminal domain. On the contrary, in warm-blooded vertebrates (mammals and birds), the N-terminal domain is on average 1.7 times more conserved than the C-terminal domain, whereas a similar rate is observed for fish.

Identification of HEAT Repeats

The deuterostome htts listed in supplementary table 3 (Supplementary Material online) showed a total of 16 HEAT repeats, all of which belonged to the AAA group (Andrade et al. 2001). They are organized into the 4 clusters labeled A–D in figure 1 (panel A): the first 2 clusters both contained 5 repeats (A1–A5, B1–B5) and the last 2 contained 3 repeats each (C1–C3, D1–D3). Table 3 shows all HEAT repeats detected in human and SP htt, with their relative positions and taxonomic distribution. Only HEAT A4 was present in all vertebrates and SP but not in *Ciona*. The large majority of HEAT repeats (14/16) seem to be also conserved in insects as only HEAT A4 and B4 were not detected in the 4 arthropod htts. Some of the HEAT repeats observed in insects were missing in Drosophilidae (B2, D3) or *T. castaneum* (B2), but all were present in *A. mellifera*, which is in line with the finding that the honeybee has the slowest evolving of the insect genomes so far sequenced (Honeybee Genome Sequencing Consortium 2006) and

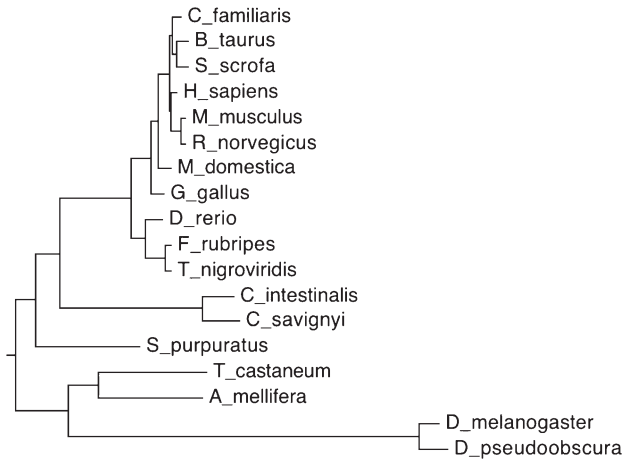

FIG. 2.—Phylogenetic tree of metazoan htts reconstructed using the MrBayes program. The posterior probabilities supporting the tree nodes are only shown when less than 100. Arthropoda were used as the outgroup. The lengths of the branches are proportional to the number of aa substitutions per site.

suggests that HEAT repeats are an ancestral character in htt evolution.

Phylogenetic Analysis

The phylogenetic analysis was made on the basis of the htt multiple alignment shown in supplementary figure 3 (Supplementary Material online) using a Bayesian method (see Materials and Methods). The resulting phylogenetic tree (fig. 2) fully resolves all branches with a 100% posterior probability and is completely congruent with the current view of animal phylogeny within mammals (Springer et al. 2004) and between vertebrates, tunicates, echinoderms, and arthropods (Telford 2006). Long branches are evident in the 2 tunicates and (especially) the 2 drosophilids, which may have experienced a remarkable acceleration in evolution.

Evolution of the polyQ Region in Metazoa

The most interesting characteristic of htt is the polyQ tract, about which our analysis revealed a real evolutionary story (fig. 3). At the base of the protostome–deuterostome divergence, the ancestor possessed one htt with a single Q or no Q in the corresponding position, and only deuterostome homologues show a double Q that is maintained until the vertebrates, in which the first real "polyQ" tract was established (QQQQ). In particular, the NHQQ sequence in sea urchin consists of a group of 4 hydrophilic aa that can be considered biochemically comparable to the 4 glutamines (QQQQ) present in vertebrates. The *Ciona* genus lost this characteristic (*Ciona* htt has no polyQ tract, which is replaced by an aromatic group) and also evolved other specific and typical tracts (Gissi et al. 2005). The 4 glutamines in vertebrates are stably maintained in fish, amphibians, and birds. The polyQ expands gradually from opossum to *Sus*, to join the longest and most polymorphic Q in humans (which spans from 15–21 to 36 in normal htt). Interestingly, rodents show a shorter polyQ (7 and 8 Q in mouse and rat, inverting the evolutionary trend) and a differently organized polyP. The polyP stretch is in fact interrupted by a Q or L aa, and the interruption seems to be conserved in at least 4 positions. Aligning polyPs with these 4 conserved points of interruption reveals a different regional structure in rodents that seem to extend more toward polyQ than in the other mammals, in which polyP expands equally to the left and right of the 4 central Q positions. The fact that the polyP tract is present only in mammalian htt (and not in nonmammalian vertebrate, *Ciona*, or sea urchin htt) again confirms that it is a recent and sudden acquisition in htt evolution.

Finally, the first 17 aa of human htt (with the 3 lysines that participate in determining its intracellular distribution between the cytoplasm and nucleus in vertebrates [Steffan et al. 2004; Rockabrand et al. 2007]) are strongly conserved in vertebrates, but sea urchin and *Ciona* have a shorter and less conserved sequence (14 aa). The 3 lysines are conserved in *Ciona* (K3, K6, and K12), whereas in sea urchin the first and third lysines (K3 and K12) are conserved with a conservative substitution in the second position (R6). This proportion of 2/3 conserved lysines also seems to be maintained in protostomes, with the third residue always conserved.

Table 3
HEAT Repeats Detected in the htt Multiple Alignment Provided in supplementary figure 3 (Supplementary Material online)

HEAT ID	*Homo sapiens*	*Strongylocentrotus purpuratus*	Number of Significant Hits out of 18	Taxonomic Range
A1	124–162	73–111	14	Deu, Pro
A2	205–243	154–192	15	Deu, Pro
A3	247–285	196–234	14	Deu, Pro
A4	317–355	266–305	8	Vrt, Spu
A5	353–391	303–341	1	Deu, Pro
B1	803–841	689–729	12	Deu, Pro
B2	845–883	733–771	1	Deu, Ame
B3	904–943	792–829	15	Deu, Pro
B4	984–1025	871–912	4	Deu
B5	1062–1100	985–1023	3	Deu, Pro
C1	1425–1463	1347–1385	11	Deu, Pro
C2	1534–1575	1456–1497	2	Deu, Pro
C3	1672–1710	1594–1632	4	Deu, Pro
D1	2798–2836	2715–2752	8	Deu, Pro
D2	2975–3013	2895–2933	2	Deu, Pro
D3	3068–3107	2988–3027	2	Deu, Ame, Tca

NOTE.—The repeats are named on the basis of their relative positions along the multialignment, using the same letter for repeats closer than 150 aa. The table also shows the location of the HEATs detected in *H. sapiens* and *S. purpuratus* htt, the number of significant hits (out of the 18 aligned sequences), and their taxonomic distribution (Deu, deuterostomes; Pro, protostomes; Vrt, Vertebrates; Spu, *S. purpuratus*; Ame, *Apis mellifera*; and Tca, *Tribolium castaneum*).

Discussion

The aim of this study was to reconstruct htt evolution by making a wide-ranging comparative analysis of htt homologues in both deuterostome and protostome branches and comparing the primary sequence of the protein homologues through multiple alignment. To add an important

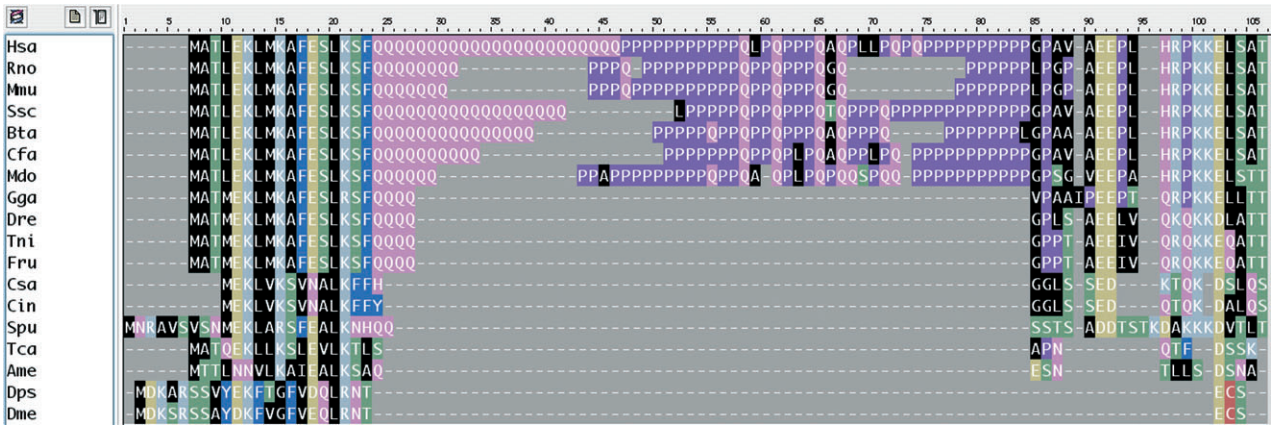

FIG. 3.—The evolution of the polyQ: details of the multiple alignment (supplementary fig. 3, Supplementary Material online) in the polyQ region. The N-terminus aa sequences are listed following the phylogenetic tree, and the residues are color labeled according to their physicochemical properties. Species abbreviations are like in table 2.

point in evolution, we also cloned the most ancient deuterostome homologue (i.e., sea urchin htt), which is present at the base of the deuterostome–protostome divergence and is one of the oldest still living deuterostome organism.

Our findings show that the structures of sea urchin htt messenger and gene are similar to those of the vertebrate homologue. Like the vertebrate homologue, the sea urchin messenger has upstream AUGs in the 5′UTR, possibly endowed of a regulatory role, and alternative 3′UTRs from 332 to 1,417 nt. Furthermore, the sea urchin gene has a large number of exons that are conserved in phase and length with respect to the human gene.

If we consider the protein starting with the second methionine, sea urchin htt N-terminus is also much similar to vertebrate htt at protein level. In particular, looking at the extreme N-terminus, the position of 2/3 lysines (which are critical for htt's subcellular localization [Steffan et al. 2004; Rockabrand et al. 2007]) and 12/17 residues are conserved. Comparison with our previously cloned *C. intestinalis* htt (Gissi et al. 2006) allowed us to show that the *Ciona* homologue diverges more than expected, whereas, although older, sea urchin htt has a higher degree of conservation. This suggests that sea urchin htt may endow specific functions that are closer to those of vertebrate htt than those of the Ciona protein.

Comparison of the gene structure of the entire group of homologues also showed that the gene has evolved along the deuterostome branch by allowing a progressive increase in the number of exons depending on phylogenetic distance, whereas the evolution of the gene (and protein) in the protostome branch is more heterogeneous. This suggests that the protostome branch has less stringent functional constraints and that the function of the protein in protostomes may be dispensable or involved in different biological functions. It is interesting to note that, among the protostomes, honeybee htt is more similar to the deuterostome homologues, thus indicating an older and more conserved htt function. The intron gains in the vertebrate genes, which are particularly concentrated at the 5′ end, may be correlated to the likely functional shift of htt in this lineage and are in accordance with the previous observation of a fast evolution of the 5′end of this gene (Gissi et al. 2006).

The multiple alignment also highlighted a number of other important aspects: 1) htt consists of 3 major conserved regions corresponding to blocks 1–386 (htt1), 683–1,586 (htt2), and 2,437–3,078 (htt3) of human htt; 2) it follows a more progressive and linear evolution along the deuterostome branch and is more heterogeneous in the protostome branch; 3) the polyQ evolution is a characteristic typical of deuterostomes whose appearance dates back to sea urchin divergence and whose position is conserved, whereas its length increases; 4) the *Ciona* genus has lost the polyQ while accumulating more differences in its N-terminal fragment; 5) the drosophilids accumulate differences in the N-terminal portion of the protein due to a large aa insertion without any polyQ; and 6) when polyQ length increases along vertebrates and couples with the polyP tract, the conservation of the N-terminal domain becomes more stringent.

We speculate that the evolution of the primary htt aa sequence parallels the particular evolution of the nervous system. At a biological level: 1) the sea urchin nervous system is poorly organized in comparison with that of vertebrates; 2) although belonging to the chordates, *Ciona* has a totally differently organized nervous system from that of vertebrates; 3) vertebrates all share the same structural organization of the nervous system, whose complexity increases progressively with the development of the most anterior brain structures (telencephalon); and 4) the structuring of the nervous system along the protostome branch has followed a different type of developmental program (metamerism). In line with this, htt has evolved differently (drosophilids) or only slightly (honeybee).

At anatomical level, the evolution of the nervous system along the deuterostome branch has progressively increased its anterodorsal positioning. On these grounds and given the biological evidence that human htt has a major neuronal function (Dragatsis et al. 2000), we suggest that, along the deuterostome branch, htt may have become progressively more important for nervous system development, maturation, and maintenance. We also speculate that its critical domain resides in the N-terminal portion of the protein, in which the polyQ first arose >450 MYA and has been specifically maintained (except in *Ciona*) in the same position although gradually expanding.

In addition, an innovative theory in evolution suggests that copy number variation of repeats in the coding region of genes involved in embryo development can be on the basis of rapid and biased development of embryo morphology (Arthur 2004; Ruden et al. 2005). A limited variation in the number of repeats may influence timing, place, type, or amount of expression of the developmental genes involved (Arthur 2004; Ruden et al. 2005), possibly determining a rapid and directed evolution. We could then speculate that glutamine repeats introduced a bias into the evolution of the protein affecting the embryo development in deuterostomes, leading them to acquire a chordate-like structure. It is also possible that, further in the evolution, htt in vertebrates has developed glutamine-dependent functions, which are particularly important for neurons and that are possibly mediated by the dynamic intracellular distribution of the protein (Arango et al. 2006).

Lastly, our analysis of HEAT repeats suggests a correspondence between HEAT repeats conservation and the 3 identified constrained blocks in the protein. As HEAT repeats may indicate a general propensity of the protein to interact with other proteins, it is possible that domain-related protein function has evolved in parallel with the appearance of specific interactors in the deuterostome branch. The 2 rounds of whole-genome duplication occurring in the ancestral vertebrates (Dehal and Boore 2005) have amplified the number of proteins (and possible interactors) and prompted the variation necessary for the development of new protein specificity and the diversification of an originally common function. Identifying the HEAT repeats that are specifically present or absent in some homologues could help define the interactors that have been typically acquired or lost in some organisms during the >450 MYA of history of this protein.

Supplementary Material

Supplementary figures 1–3 and tables 1–4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). New sequence accession numbers are AM398482–AM398562.

Acknowledgments

We would like to thank Dr Eric Davidson and Dr Victor Vacquier for the sea urchin RNA material. This study was supported by Fondazione Telethon (GGP06250), Fondo Italiano Riccrca di Base RBLA03AF28-006 to E.C. and RBLA039M7M-002 to G.P. (Ministry of University and Research), and Fondo Interno Ricerca Scientifica e Tecnologica (University of Milan) to C.G.

Literature Cited

Andrade MA, Bork P. 1995. HEAT repeats in the Huntington's disease protein. Nat Genet. 11:115–116.

Andrade MA, Petosa C, O'Donoghue SI, Muller CW, Bork P. 2001. Comparison of ARM and HEAT protein repeats. J Mol Biol. 309:1–18.

Arango M, Holbert S, Zala D, et al. 2006. CA150 expression delays striatal cell death in overexpression and knock-in conditions for mutant huntingtin neurotoxicity. J Neurosci. 26:4649–4659.

Arthur W. 2004. Biased embryos and evolution. Cambridge: Cambridge University press

Blair-Hedges S, Kumar S. 2003. Genomic clocks and evolutionary timescales. Trends Genet. 19:200–206.

Cattaneo E, Rigamonti D, Goffredo D, Zuccato C, Squitieri F, Sipione S. 2001. Loss of normal huntingtin function: new developments in Huntington's disease research. Trends Neurosci. 24:82–188.

Cattaneo E, Zuccato C, Tartari M. 2005. Normal huntingtin function: an alternative approach to Huntington's disease. Nat Rev Neurosci. 6:919–930 [Review].

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Dragatsis I, Levine MS, Zeitlin S. 2000. Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. Nat Genet. 26:300–306.

Everett CM, Wood NW. 2004. Trinucleotide repeats and neurodegenerative disease. Brain. 127:2385–2405.

Gissi C, Pesole G, Cattaneo E, Tartari M. 2006. Huntingtin gene evolution in Chordata and its peculiar features in the ascidian Ciona genus. BMC Genomics. 7:288.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 443:931–49.

Iacono M, Mignone F, Pesole G. 2005. uAUG and uORFs in human and rodent 5′untranslated mRNAs. Gene. 349:97–105.

Jones DT, Taylor WR, Thormton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–82.

Kauffman JS, Zinovyeva A, Yagi K, Makabe KW, Raff RA. 2003. Neural expression of the Huntington's disease gene as a chordate evolutionary novelty. J Exp Zool B Mol Dev Evol. 297:57–64.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Li Z, Karlovich CA, Fish MP, Scott MP, Myers RM. 1999. A putative Drosophila homolog of the Huntington's disease gene. Hum Mol Genet. 8:1807–1815.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. Structure. 11:1453–1459.

O'Kusky JR, Nasir J, Cicchetti F, Parent A, Hayden MR. 1999. Neuronal degeneration in the basal ganglia and loss of pallido-subthalamic synapses in mice with targeted disruption of the Huntington's disease gene. Brain Res. 818:468–479.

Pei J, Grishin NV. 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 23:802–808.

Rigamonti D, Bauer JH, De-Fraja C, et al. (15 co-authors). 2000. Wild-type huntingtin protects from apoptosis upstream of caspase-3. J Neurosci. 20:3705–3713.

Rockabrand E, Slepko N, Pantalone A, Nukala VN, Kazantsev A, Marsh JL, Sullivan PG, Steffan JS, Sensi SL, Thompson LM. 2007. The first 17 amino acids of Huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. Hum Mol Genet. 16:61–77.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Ruden DM, Garfinkel MD, Xiao L, Lu X. 2005. Epigenetic regulation of trinucleotide repeat expansion and contractions and the "biased embryos" hypothesis for rapid morphological evolution. Curr Genomics. 6:145–155.

Sipione S, Cattaneo E. 2001. Modeling Huntington's disease in cells, flies, and mice. Mol Neurobiol. 23:21–51.
Springer MS, Stanhope MJ, Madsen O, de Jong WW. 2004. Molecules consolidate the placental mammal tree. Trends Ecol Evol. 19:430–438.
Steffan JS, Agrawal N, Pallos J, et al. (13 co-authors). 2004. SUMO modification of Huntingtin and Huntington's disease pathology. Science. 304:100–104.
Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596–9.
Telford MJ. 2006. Animal phylogeny. Curr Biol. 16(23):R981–R985.
Zuccato C, Belyaev N, Conforti P, et al. (11 co-authors). 2007. Widespread disruption of repressor element-1 silencing transcription factor/neuron-restrictive silencer factor occupancy at its target genes in Huntington's disease. J Neurosci. 27:6972–6983.
Zuccato C, Cattaneo E. 2007. Role of brain-derived neurotrophic factor in Huntington's disease. Prog Neurobiol. 81:294–330 [Review].

Billie Swalla, Associate Editor

Accepted November 18, 2007