Published online 28 January 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 10* ***4273–4287***
doi:10.1093/nar/gks011

Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks

Martin H. Schaefer[1], Erich E. Wanker[2] and Miguel A. Andrade-Navarro[1,*]

[1]Computational Biology and Data Mining and [2]Neuroproteomics, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany

Received October 31, 2011; Revised December 15, 2011; Accepted December 23, 2011

ABSTRACT

Expanded runs of consecutive trinucleotide CAG repeats encoding polyglutamine (polyQ) stretches are observed in the genes of a large number of patients with different genetic diseases such as Huntington's and several Ataxias. Protein aggregation, which is a key feature of most of these diseases, is thought to be triggered by these expanded polyQ sequences in disease-related proteins. However, polyQ tracts are a normal feature of many human proteins, suggesting that they have an important cellular function. To clarify the potential function of polyQ repeats in biological systems, we systematically analyzed available information stored in sequence and protein interaction databases. By integrating genomic, phylogenetic, protein interaction network and functional information, we obtained evidence that polyQ tracts in proteins stabilize protein interactions. This happens most likely through structural changes whereby the polyQ sequence extends a neighboring coiled-coil region to facilitate its interaction with a coiled-coil region in another protein. Alteration of this important biological function due to polyQ expansion results in gain of abnormal interactions, leading to pathological effects like protein aggregation. Our analyses suggest that research on polyQ proteins should shift focus from expanded polyQ proteins into the characterization of the influence of the wild-type polyQ on protein interactions.

INTRODUCTION

Polyglutamine (polyQ) stretches of expanded pathological length in human proteins have been observed to cause neurodegenerative diseases such as Huntington's disease (HD) or several ataxias (1). The formation of insoluble protein aggregates is a key feature of all known polyQ diseases (2–4). Biochemical and cell biological experiments have demonstrated that expanded polyQ tracts drive the spontaneous assembly of insoluble protein aggregates in disease model systems (5), suggesting that polyQ-mediated protein misfolding and aggregation are critical for disease development. However, it remains unclear whether polyQ-mediated aggregation of proteins is the cause or the consequence of progressive neurodegeneration in polyQ diseases (6,7). Further theoretical and experimental studies are required to address the role of polyQ-containing proteins under pathological and non-pathological conditions (8).

Besides their association with disease development, polyQ sequences may have normal functions. Indeed, they are present in more than 60 human proteins (9), and some lower organisms, such as the amoeba *Dictyostelium discoideum* (10), possess several hundred. However, their function has not yet been determined (11). It was hypothesized that they form a flexible spacer between protein domains like other low complexity regions (12–14). Anecdotal experimental evidence suggests a role of polyQ tracts in activation of gene transcription (15). Accordingly, statistical studies revealed that proteins containing a polyQ stretch are biased toward functions related to transcriptional regulation and nuclear localization in several species (12,16,17). Also, a more general role in mediating protein–protein interactions has been suggested (18).

In order to advance our understanding of the functions of polyQ regions in proteins, we investigated their potential properties from a systemic point of view, with a particular focus on phylogenetic conservation, presence in protein interaction networks and functional aspects of protein families containing polyQ tracts. Our analyses suggest that the normal function of polyQ regions in proteins is to stabilize protein–protein interactions (PPIs) and that the pathological effect of polyQ expansion would then be due to a gain of abnormal interactions eventually leading to protein aggregation.

We present our results starting with the analysis of CAG repeats at the nucleotide level, moving on to the analysis of protein sequences with polyQ tracts and phylogenetic studies of protein families, to the investigation of

*To whom correspondence should be addressed. Tel: +49 30 9406 4250; Fax: +49 30 9406 4240; Email: miguel.andrade@mdc-berlin.de

protein interaction networks of polyQ-containing proteins and finally to an examination of features of the sequences adjacent to polyQ tracts.

MATERIALS AND METHODS

Sequence data and annotation

Human genomic sequence data and gene annotations were retrieved from the UCSC (human: GRCh37/hg19, rat: Baylor 3.4/rn4, mouse: NCBI37/mm9 and fly: BDGP R5/dm3) (19). Conflicting assignments for a genomic region due to different splicing forms were resolved by giving priority to assignments in the following order: protein-coding exon, UTR, intron, intergenic region. Protein sequences were downloaded from UniProt (version 15.6) (20), which consists of the manually curated and non-redundant Swiss-Prot database and a largely automatically annotated TrEMBL data set. Protein domain annotations were taken from Pfam (version 23.0) (21) and protein family definitions from the TreeFam database (22). For the analyses of human protein–protein interactions (PPIs), the HIPPIE database was used (23). HIPPIE is a large integrated PPI network that consists of most commonly used human experimental PPI databases. For the PPI analysis of other species, we retrieved the BioGRID database (24) and removed genetic interactions (*Saccharomyces cerevisiae* v3.1.74 and *Drosophila melanogaster* v3.1.69).

Definition of the polyQ set

We examined a selection of proteins with a stretch of consecutive Qs. We found that a threshold of 10 was the minimum length that recognized the set of nine known human polyQ disease proteins (1). Ataxin-7 is, among those disease proteins, the one with the minimum polyQ length of 10 residues in its non-expanded form. We allowed for one mismatch (independent of its position within the polyQ tract) taking into account that polyQ stretches are often interspersed with single amino acids (as in the known polyQ *D. melanogaster* Homeobox proteins Deformed and Antennapedia).

Where possible, we restricted our analyses of species-specific polyQ sets to proteins from the manually curated and non-redundant Swiss-Prot. For the analysis of functional annotations and sequence features enriched in the polyQ set, we defined a canonical set of representative species that had at least 750 protein entries in Swiss-Prot (to guarantee a sufficiently large coverage of their proteomes in the sequence database) and at least eight polyQ proteins (to allow for conclusive enrichment statistics). From the resulting 13 species, we removed two yeast members (*Kluyveromyces lactis* and *Candida albicans*) and kept only *S. cerevisiae*. Thereby, we came up with a compact though heterogeneous set of 11 representative species: *Homo sapiens*, *Bos taurus*, *Rattus norvegicus*, *Mus musculus*, *Danio rerio*, *D. melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *S. cerevisiae*, *D. discoideum* and *Neurospora crassa*.

To assemble the set of species for the comparison of polyQ frequencies (Figure 2), we choose the same selection criterion on minimum number of protein entries in Swiss-Prot as before (>750 protein entries) without any threshold on the polyQ number. For clarity reasons, we removed several Bacteria since they have no polyQ proteins.

Only for the purely quantitative analysis of domains correlated with the presence of polyQ (Table 1; Supplementary Tables S1 and S2) and for the analysis and discussion of frequencies within non-model organisms which are much less represented in Swiss-Prot, we additionally included the automatically curated TrEMBL.

Randomization test for the detection of domains enriched in polyQ proteins

To detect overrepresented domains in the set of proteins interacting with the polyQ proteins in the PPI network, a randomization test was applied. As a test statistic, the number of interactions between the polyQ or the random set, and the domain set was calculated. The background distribution was generated by selecting a set of non-polyQ proteins with the same or more interacting partners than those in the polyQ set and containing the same amount of transcription factors. A similar randomization procedure was applied in the test for enrichment of polyQ proteins in the set of proteins that interact with polyQ proteins.

Gene ontology analysis of proteins with many partners

We assigned the number of interactions in which each protein is participating (their degree) to all human and yeast proteins and retrieved the top 86 proteins (corresponding to the size of the human polyQ set) as well as the 10% with the highest degree for each species. From these lists, we removed all polyQ proteins. We then calculated the enrichment of GO terms among these proteins as compared to proteins that have at least one interaction partner as a background using the web tool DAVID (25). The resulting GO terms were corrected for multiple testing with the Benjamini–Hochberg method.

RESULTS

PolyQ repeats have functional and evolutionary features that have been proposed to be relevant at different inter-related molecular levels: nucleotide sequences, protein sequences, protein structures and protein interaction networks. For this reason, we have studied and combined analyses at those levels to obtain a systematic overview of polyQ function and evolution.

Distribution of CAG repeats in the human genome

Glutamines in proteins can be encoded in human genes by either CAG or CAA codons. However, the polyQ stretches that are enlarged in human disease proteins are encoded almost exclusively by pure CAG runs, while CAA repeats have not been observed (1). This led to the suggestion of a possible mechanism for their generation by DNA slippage and hairpin formation during DNA

replication facilitating length extensions to which CAG but not CAA repeats are prone (26). Are CAG repeats, and the polyQ encoded by them, just artifacts of faulty DNA replication without biological function?

This question can be answered by examining the genomic location of CAG repeats. If their genomic location was solely determined by random processes such as copy errors during DNA replication, and they underwent no evolutionary selection, they should be evenly distributed in the genome. On the contrary, if they had a biological function, their distribution should correspond to the molecular level of action: a function on RNA level would bias their genomic position toward transcribed genomic regions whereas a function in proteins would shift their distribution toward protein-coding exons.

We studied the distribution of CAG repeats of 10 or more consecutive trinucleotides in the human genome. To measure whether the observed distribution is random or biased toward specific elements, we calculated the relative number of repeats falling into different regions such as protein-coding exons, introns, untranslated regions (UTRs) and intergenic regions. These frequencies were normalized by the fraction of the genome covered by the respective region type (Figure 1). Indeed, of 136 CAG repeats considered, 33 are in protein coding exons as previously described (27) (~43-fold enrichment over a random expectation). Although those 33 CAG repeats in coding regions could potentially encode three types of amino acid repeats depending on the reading frame (polyQ, polyS and polyA, for codons CAG, AGC and GCA, respectively), 28 coded for polyQ. This suggests that even if CAG repeats are accidental, they are selected for the encoding of polyQ in proteins, suggesting that polyQ has a biological function. Accordingly, the number of CAG repeats in introns and intergenic regions is close to random expectation (8 and 89, respectively; Figure 1). However, 6 CAG repeats are in UTRs (8-fold over random expectation), including the known disease locus in the 5′-UTR of the gene PPP2R2B causing spinocerebellar ataxia type 12 (28), suggesting that they have a function at the transcript level. We also found CAG repeats enriched in UTRs and protein coding exons in rat, mouse and fly, though to a lower degree as in human (UTR enrichment ranges from 1.7- to 2.5-fold and exon enrichment from 3.1- to 5.3-fold) (Supplementary Figure S1a–c).

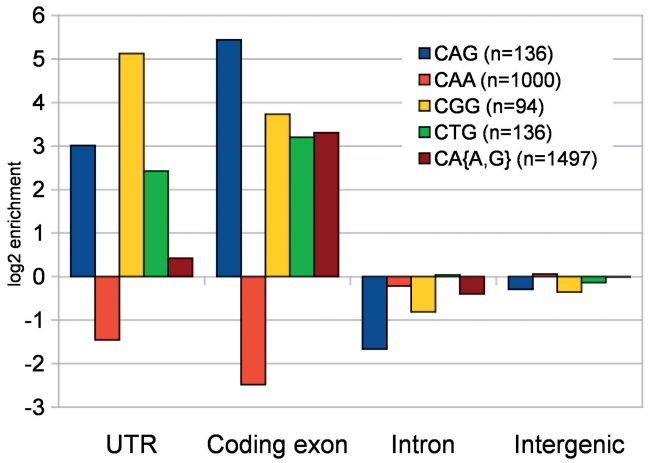

Figure 1. Frequency of trinucleotide repeats in the human genome. The *y*-axis represents the log2 of the ratio between the relative number of repeat runs observed (considering runs of at least 10 consecutive trinucleotides) and the proportion of the genome that is covered by the respective genomic region type.

For comparison, we did an analysis considering consecutive runs composed of both codons encoding glutamine (CAG or CAA). These mixed trinucleotide repeats are 11 times more frequent in the human genome than pure CAG repeats. Like CAG repeats, the mixed repeats were enriched in exons and randomly distributed outside transcripts; their presence in UTRs, unlike CAG repeats, was close to random expectation. Together, these results suggest that CAG repeats have a function both at the protein and the transcript level.

We also analyzed the frequencies of pure CAA repeats in the different genomic region types. We found them to be generally more frequent in the human genome as compared to pure CAG repeats (1000 versus 136) but largely absent from protein coding regions (just one CAA repeat falls into a translated region encoding a polyQ stretch in the human protein ZFHX3). We note that there is a 2-fold enrichment of genes encoding tRNAs with an anticodon for CAG as compared to tRNAs matching CAA (21 versus 11) and that the CAG codon abundance is almost 3-fold higher in human exons (29), but these numbers alone do not explain the 243-fold higher relative amount of CAG repeats in human protein coding regions.

Similarly, we did a calculation for CTG repeats in the human genome, which, like CAG repeats, are CG rich and when expanded are known to cause diseases of altered RNA function (1) such as Myotonic dystrophy type 1 (30). Of a total of 136 CTG repeats, 7 were found in coding regions: 4 encoding for polyL (CTG codon), 3 for polyA (GCT codon) and none for polyC (TGC codon). Comparison to random expectation indicates selection for protein function. As for CAG, the number of CTG repeats found in UTRs was significantly above random expectation, suggesting also that they have a biological function in transcripts. Recent evidence indicates that CTG and CAG repeats form RNA–DNA hybrids (R loops) and it has been hypothesized that these structures may have a biological role (31,32).

As CGG repeats of length 6 have been previously described as being the most strongly overrepresented trinucleotide repeats in human exons (27), we also compared their distribution in the human genome to that of CAG repeats. We observed a similar distribution as the one for CAG and CTG with strong enrichment in UTRs and protein coding exons. In summary, whereas CAG repeats are clearly selected because they code polyQ in human proteins, there is some evidence of their function in non-coding parts of transcripts. We observed this in other mammals and for other CG rich repeats expanded in disease such as CTG.

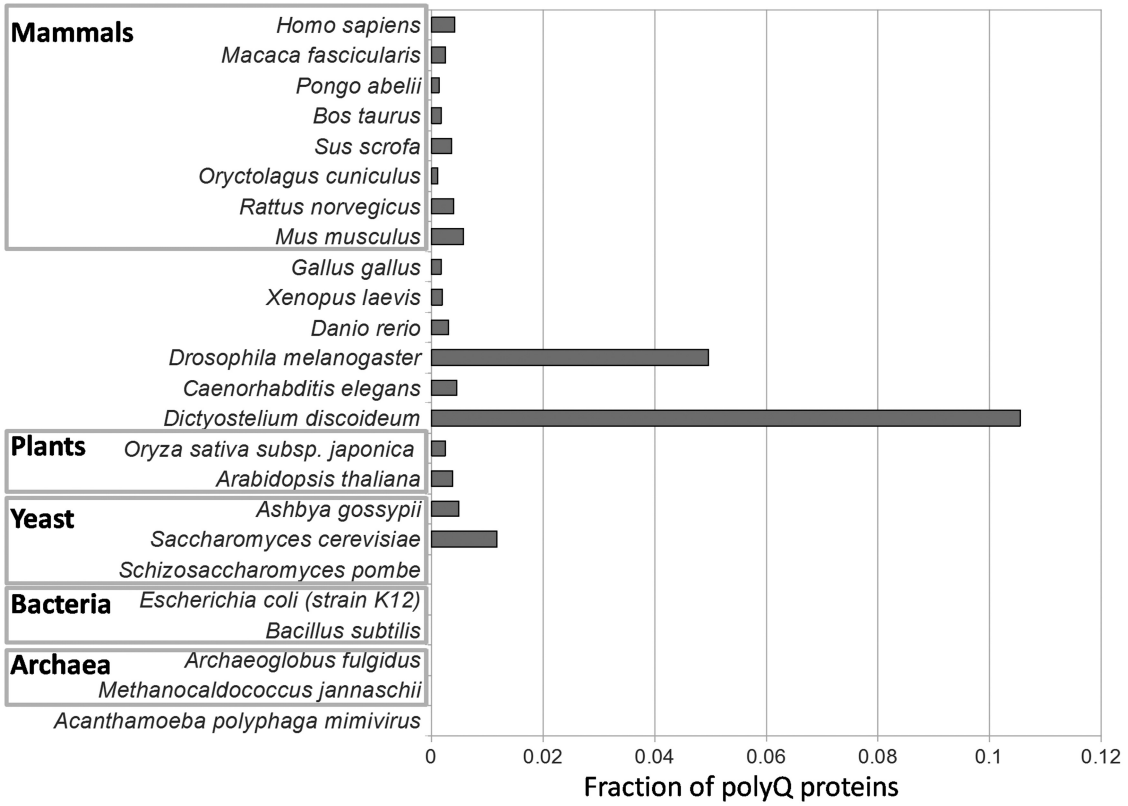

Figure 2. Relative amount of polyQ proteins in a representative set of species. The graph represents the fraction of proteins of each species' available proteome that contains a polyQ tract. Species included had more than 1000 protein sequences in the UniProt/Swiss-Prot database (version 15.6) (20). For simplicity, just two bacterial species were included in the plot since all of those analyzed had very few or no polyQ proteins.

Table 1. Correlation of domains to polyQ presence over species

Pfam identifier	Description	Class	Correlation on eukaryotic subset
PF03810	Importin-beta N-terminal domain		0.530
PF01302	CAP-Gly domain		0.522
PF12171	Zinc-finger double-stranded RNA-binding	ZF	0.494
PF02207	Putative zinc finger in N-recognin (UBR box)	UBX, ZF	0.493
PF01363	FYVE zinc finger	PI, ZF	0.479
PF03731	Ku70/Ku80 N-terminal alpha/beta domain		0.476
PF01151	GNS1/SUR4 family		0.470
PF08389	Exportin 1-like protein		0.464
PF00787	PX domain	PI	0.447
PF00153	Mitochondrial carrier protein		0.435
PF00169	PH domain	PI	0.432
PF00613	Phosphoinositide 3-kinase family, accessory domain (PIK domain)	PI	0.431
PF09336	Vps4 C terminal oligomerization domain		0.428
PF01585	G-patch domain		0.423
PF05047	Mitochondrial ribosomal protein L51 / S25 / CI-B8 domain		0.417
PF00620	RhoGAP domain		0.408
PF00566	TBC domain		0.400

Columns are (1) Pfam identifier, (2) Pfam description, (3) functional and structural classes: zinc finger (ZF), ubiquitin (UBX) or phosphatidylinositol (PI), (4) (Spearman) correlation over 133 eukaryotic species.

Investigating the relative number of polyQ proteins in different organisms

We determined the frequency of polyQ containing proteins in a large number of species belonging to a wide taxonomic range (20). For this analysis (and hereafter, unless otherwise indicated), we identified as polyQ proteins those containing at least one polyQ stretch with a minimum length of 10 glutamines allowing for one

mismatch. This threshold was chosen in order to account for all known human polyQ disease proteins (see 'Materials and Methods' section for details).

We observed that the fraction of proteins having a polyQ stretch deviates largely among different species. Figure 2 displays polyQ frequencies in several representative species of the manually curated part of UniProt (Swiss-Prot) and Supplementary Table S1 lists the amount for a larger selection of species from the entire UniProt (Swiss-Prot and TrEMBL). This suggests that abundance of polyQ proteins is not a random feature but depends on properties variable between species.

While the proteomes of bacteria and archaea typically contain no proteins with polyQ tracts at all, lower and higher eukaryotes on an average have 0.1% proteins with polyQ tracts. The *H. sapiens* fraction of polyQ proteins is above the average (0.34%) but much lower than that of many other organisms such as the yeast *S. cerevisiae* (1.1%), the fly *D. melanogaster* (3.8%) or the slime mold *D. discoideum* (10.5%).

In many cases, taxonomically related species have similar content of polyQ proteins but this is by no means the rule. For example, one can observe extreme differences between yeasts: the fission yeast *Schizosaccharomyces pombe* has only three polyQ proteins (out of 4974, <0.1%) whereas the baker's yeast *S. cerevisiae* has 79 (out of 6552, 1.1%), with other yeasts having even higher frequencies, e.g. *N. crassa* (2.7%) and *Lodderomyces elongisporus* (6.8%). Variation of polyQ protein content can be significant even within species of the same genus. For example, in the 12 *Drosophilae* species that were analyzed the fraction ranges from 2.7% in *D. simulans* to the 8.9% of *D. grimshawi* (median 4.2%). Variation of polyQ protein content between species is indeed important but we observed that it is limited when we compare closely related species, suggesting that it is tied to evolution. For example, while the three strains of the yeast *Paracoccidioides brasiliensis* analyzed had slightly different numbers of proteins, their overall polyQ frequencies were found to be similar (around 1.1%).

To find out whether there are species-specific functions that associated with polyQ protein content, we studied the frequency of polyQ proteins in a species in relationship to the presence of other proteins with particular domains. Protein domains are good indicators of particular protein functions and subcellular locations. Numerous and accurate annotations of domains known or predicted to be present in proteins can be easily obtained from several databases like Pfam (21). We calculated the correlation between the relative number of proteins containing a polyQ stretch and the relative number of proteins containing a given protein domain (Table 1 and Supplementary Table S2). For this investigation, we used the protein annotations stored in the Pfam database [version 23.0; (33)].

In a first analysis, we computed the correlations of 4088 domains found in human proteins over all bacterial and eukaryotic species with at least 5000 protein entries in Pfam (for a total of 428 species, 133 of them eukaryotic and 295 bacterial). Since polyQ proteins are almost absent from prokaryotes, many domains appeared to be correlated to polyQ protein frequency simply because they were exclusive to eukarya. Therefore, we did a second analysis of correlation only on the 133 eukaryotic species. We found 40 domains having a (Spearman) correlation value over all species >0.8 and a correlation value on eukaryotic species >0.3 indicating that these domains are significantly enriched in the proteomes of species with many polyQ proteins (see the most highly correlated domains in Table 1 and a full list of all 40 domains including additional information in Supplementary Table S2).

Among the most highly correlated domains were the FYVE and PX domains. Remarkably, they are the only domains known to bind phosphatidylinositol 3-phosphate (PI3P) (34). The current version of the SMART database of domain annotations (35) indicates that these domains do not co-occur in any of the current set of annotated proteins. This suggests that the identification of these two domains is based on independent sequences. The functional implication is that there is a true association to polyQ proteins, more precisely, that the presence of polyQ proteins in a species is likely to be connected with processes that use PI3P, possibly in relation to signaling and transport mechanisms in which this molecule is involved.

We can point to further striking functional and structural similarities between the other 38 correlated domains supporting associations to polyQ proteins to particular functions. Three domains have a function in the phosphatidylinositol (PI) signaling system (CRAL/TRIO, PH, Phosphoinositide 3-kinase family accessory domain). Two domains are related to ubiquitin (UBR box and UBX). Finally, we also observed seven domains that belong to the zinc finger domain class (FYVE, Zinc-finger of the MIZ type in Nse subunit, UBR box, Zinc-finger double-stranded RNA-binding, Zinc finger ZZ type, PHD-finger, HIT zinc finger). Together, these observations suggest that polyQ proteins seem to be present in high numbers in species rich in proteins with roles related to PI signaling, the protein degradation system and molecular interactions.

PolyQ emergence in protein families

Human non-pathogenic huntingtin contains an N-terminal polyQ tract of variable length [ranging from 11 to 34 glutamines (1): Q11–Q34]. Such N-terminal polyQ appreciably and progressively shortens in orthologs from species increasingly distant from human along the chordate lineage (Q10 in dog, Q7 in mouse, Q6 in opossum, Q4 in *Xenopus* and fish; Figure 3a, left box). We noted that the *Drosophila* huntingtin protein in various Drosophilae does not contain any N-terminal polyQ stretch but has several in two other regions of the protein (e.g. *D. yakuba* GenPept ID:195503512, has a Q10 at positions 625–634 and a Q12 in a stretch of 14 amino acids at positions 1118–1131), which are absent in the human protein (Figure 3a). This indicates that huntingtin proteins in ancestral species along the chordate and *Drosophilae* lineages have experienced independent events of insertion of polyQ tracts. This would suggest that the huntingtin protein is under evolutionary pressure to

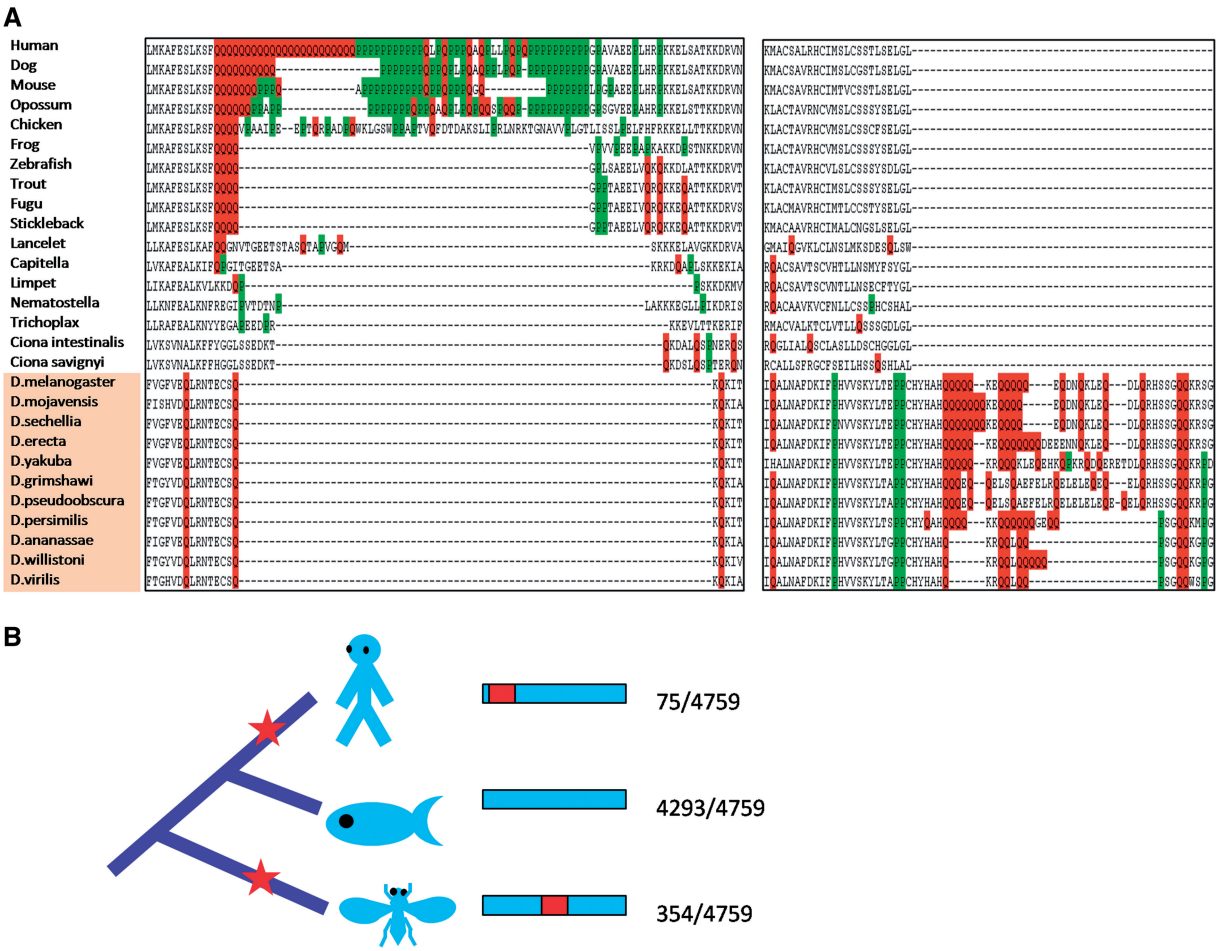

Figure 3. Protein families with multiple events of polyQ insertion. (**A**) Fragments of a multiple sequence alignment of huntingtin orthologs from several species, with glutamines and prolines marked in red and green, respectively. Left box: N-terminal polyQ region progressively enlarged along the chordate lineage and missing in Drosophilae. Note how this region is followed by polyP in some species where the polyQ length is above four. Right box: very variable polyQ rich insertion specific to Drosophilae at another, distant position in huntingtin. (**B**) A total of 4759 protein families with members in human, zebrafish and fly was studied. We found 75 families having at least one human protein with a polyQ stretch, 354 families having at least one fly protein with a polyQ stretch, and 4293 having no Q-rich region in the fish proteins (see main text for details). For a total of 14 families (including huntingtin), both the human and the fly sequences had polyQ tracts (red boxes within the blue boxes) but not the zebrafish one, indicating multiple events of polyQ insertion along separate lineages (stars). By randomizing the identity of the polyQ sets in human and fly, we found the number of selected families to be significantly higher than random expectation ($P < 0.05$).

accept polyQ insertions, but that this pressure would not seem to act on the precise position of those insertions in the sequence.

To test whether this finding in huntingtin is unique or whether there are other protein families that underwent similar events during their evolution, we examined the distribution of polyQ-containing proteins in families of proteins with members in human (*H. sapiens*), zebrafish (*D. rerio*) representing another chordate and the fly (*D. melanogaster*) representing a non-chordate organism. Given a protein family, existence of a polyQ in the human and fly proteins but not in the zebrafish one will suggest that at least two independent events of polyQ insertion occurred: one outside the Chordate lineage and another within the chordate lineage, after the divergence of zebrafish and human.

We obtained 4759 protein families with at least one member from each of human, fly and fish according to the database of phylogenetic trees TreeFam (22). We then selected those families in which the more distantly related species (human and fly) both had at least one homolog with a polyQ stretch (here requiring 8 Q in a range of 10 residues) while the zebrafish homologues were required to have no polyQ stretch at all (less than 5 Q in a window of size 10), considering them as families with evidence of multiple evolutionary events of polyQ insertion (see Figure 3b). A total of 14 protein families fulfilled this conservative criterion (Supplementary Table S3). This number was significant ($P < 0.05$, randomization test). Considering also that in most cases the polyQ stretches appear at different protein positions within the aligned protein family, we conclude that the most likely explanation for the distribution of polyQ regions in the protein families analyzed is that polyQ emerged independently at different time points during evolution rather than that being lost in zebrafish.

The emergence of a significant number of protein families where insertion of polyQ tracts occurs in

multiple ancestral proteins suggests that functional selection for the insertion of polyQ tracts at the protein level is a significant factor affecting the evolution of polyQ tracts. The fact that these insertions may be located at different positions in the protein suggests that polyQ performs a function that is not bound to a particular sequence. Insertion of polyQ tracts, however, does not seem to be absolutely necessary and therefore its function, while advantageous, must depend on some pre–existing, more important functional context. PolyQ could be selected to modulate already existing protein interactions of the protein in which it is inserted.

PolyQ in protein complexes

To find if the functional context of polyQ tracts is related to protein interactions, we investigated whether polyQ-containing proteins are enriched among proteins that form complexes. Among 1825 human protein complexes [defined as described in ref. (36)], we identified 130 having at least one protein containing a polyQ stretch (using the same polyQ definition as above: repeat length of 10 glutamines allowing for one mismatch).

These 1825 human complexes are formed by 8797 components; among them 149 are polyQ proteins, showing a 4-fold enrichment with respect to the frequency of polyQ proteins in the human proteome. In the non-redundant list of 2541 proteins forming part of complexes, the enrichment is still significant (2.1-fold). This suggests that polyQ proteins even have a tendency to form part of multiple complexes. This is the case of human proteins such as CBP and TBP.

To test whether there is a significant tendency to find multiple polyQ proteins within individual protein complexes, we applied a randomization test. We randomized the polyQ annotations and observed whether, we obtained an equal or larger amount of complexes containing two or more polyQ proteins, which happened in 52 of 1000 tests ($P = 0.052$). For less restrictive polyQ threshold selections, the results were even more significant (e.g. eight Qs in a window of 10 residues resulted in a P-value <0.001). This suggested that polyQ containing proteins are not randomly distributed among complexes but that the chance of seeing one polyQ protein increases significantly the chance of finding at least one other polyQ containing protein in the same protein complex. For example, the RSmad complex contains a total of 10 proteins. Among them are three polyQ containing proteins: ARID1B, CBP and NCOA3. In summary, protein complexes are enriched in polyQ proteins suggesting that polyQ function is related to protein interactions.

PolyQ tracts are associated to proteins with many partners

To further investigate the association of polyQ tracts with protein interactions, we compared the distribution of polyQ tract containing proteins and the number of protein interacting partners (according to the HIPPIE database of human PPI data (23)). We observed that proteins containing polyQ tracts have significantly more interactions than proteins that do not ($P = 5e\text{-}09$, Wilcoxon–Mann–Whitney test). However, we observed that polyQ proteins have a longer than average length (1253 residues versus 550 residues) and that longer proteins have more interaction partners as compared to short proteins, probably due to their higher number of potential interaction interfaces (e.g. the longest 25% of all human proteins have a mean value of 9.1 interaction partners while the shortest 25% have only 4.9). Therefore, we repeated the test ensuring that the randomly chosen non-polyQ proteins used for comparison were at least as long as the average polyQ containing protein. The resulting P-value of 0.007 was again significant.

Since it is likely that transcription factors have more interactions than the average protein and polyQ proteins are enriched in transcription factors, we repeated the test comparing the interaction distribution of the polyQ proteins to that of the set of human transcription factors [as defined by UniProt annotations (20)] without a polyQ tract (Figure 4a). The resulting P-value of 0.009 was once more significant.

We carried out the same analysis in the proteins of *S. cerevisiae* and observed that its polyQ proteins have a significantly larger number of interacting partners than those that do not have polyQ (P-value 2e-12), even when filtering for transcription factors (P-value 0.041) or proteins of higher length (P-value 0.0003) (Figure 4b). These results confirm that our findings are not species-specific and that our observations in human proteins are not due to a bias in the PPI network arising from researchers focusing on particular disease related proteins.

To test whether there is an effect of the length of the polyQ stretch on the number of interactions, we binned either all human or all yeast proteins into the three categories: lacking polyQ tract, having a small polyQ (length between 5 and 14 amino acids), and having a long polyQ tract (longer than 14 amino acids). As in the analyses described earlier, we counted the number of interactors for each of these proteins (Figure 4c and d). The differences between degree distributions were significant ($P < 0.01$) and increased with the length of the polyQ tract both for human and for yeast proteins. This observation suggests a correlation between the length of polyQ and the interaction capacity of the hosting protein.

In summary, these analyses demonstrate that polyQ proteins have more protein interactions than proteins lacking a polyQ tract. Although there is a component in that effect related to polyQ proteins having longer than average length and being associated to particular functions, these properties of polyQ proteins alone are not responsible for the whole effect. We interpret these results as indicating that polyQ tracts favor PPIs.

Function of polyQ proteins

It has already been noted that polyQ proteins are biased toward functions related to transcriptional regulation and nuclear localization (12,16,17). To make a comprehensive analysis of the association of polyQ function to particular functions in the proteins containing it, we collected the polyQ proteins from 11 eukaryotic organisms of different taxa including plants, fungi, nematoda and Chordata

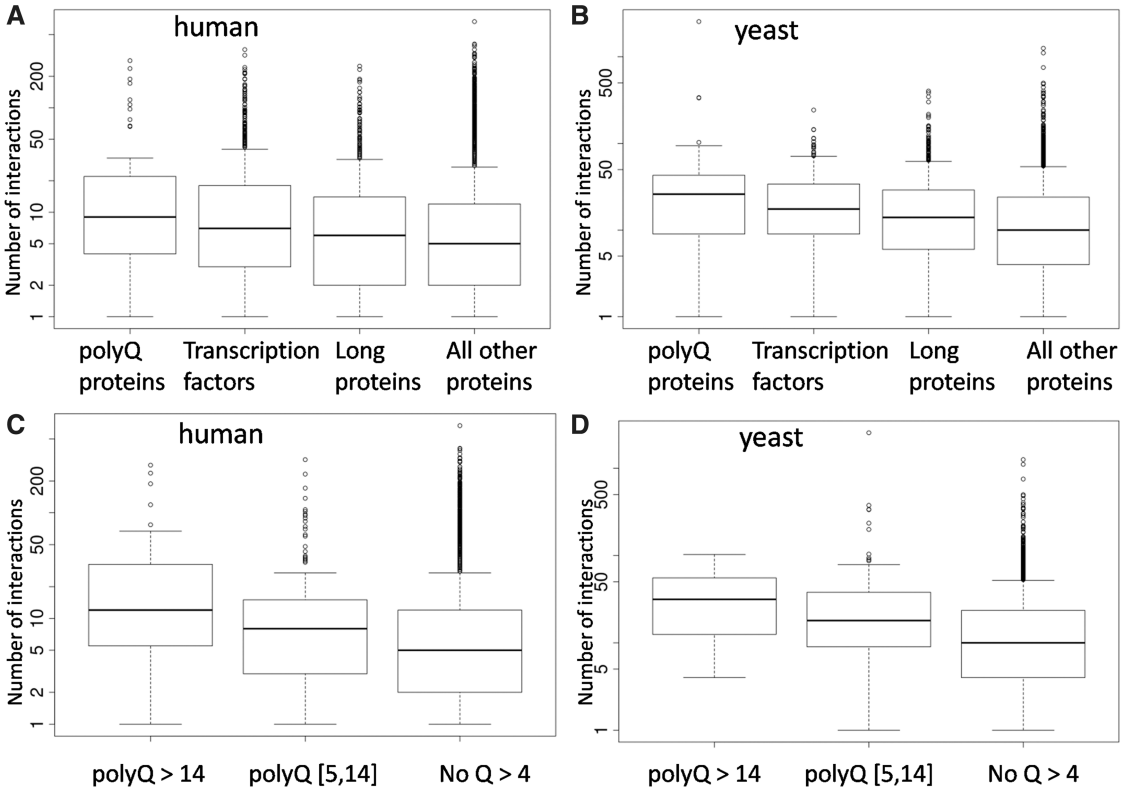

Figure 4. Protein interaction degree distribution for different protein sets. Box plots of the distribution of protein interaction partners for different protein sets. (**A** and **B**) Comparison of polyQ proteins, transcription factors without polyQ, large proteins without polyQ and all non-polyQ proteins, for human and yeast, respectively. (**C** and **D**) Comparison of proteins with long polyQ, short polyQ, or no polyQ, for human and yeast, respectively. All pairwise differences within a species were significant ($P < 0.01$) except for the comparison between medium and long polyQ length in yeast (P-value 0.056). This exception was due to an outlier in the medium set: one of the proteins has a degree of 2549 which is more than twice as high as the second highest degree. Removing it results in significant differences for all comparisons.

(*H. sapiens, B. taurus, R. norvegicus, M. musculus, D. rerio, D. melanogaster, C. elegans, A. thaliana, S. cerevisiae, D. discoideum* and *N. crassa*) and studied their functional annotations using two complementary approaches.

Firstly, we computed the enrichment in Gene Ontology annotations associated to these polyQ sets with respect to the total protein set. We analyzed each of the 11 species independently using the web tool DAVID (25). Annotations significantly enriched in the polyQ sets of several of these species included nuclear related functions (e.g. transcription, splicing), but interestingly also protein dimerization, which correlates with our findings above suggesting that polyQ is involved in protein interactions (Table 2).

Secondly, we evaluated functional enrichment according to protein annotations where annotated domains are known [Pfam version 23.0; (33)]. A total of 31 domains were significantly overrepresented in polyQ proteins in at least 2 out of the 11 species ($P < 0.05$; see 'Materials and Methods' section for details; Supplementary Table S5). PolyQ tracts were not significantly located near them (data not shown).

Overall, most of these domains group into five functional categories: transcription regulation, protein binding, chromatin maintenance, RNA binding and signaling. For example, among the 15 most frequently colocalized domains (those which are overrepresented in the polyQ sets of at least three species), we found 6 domains involved in protein–protein interactions (PAS fold, Bromodomain, PHD finger, PH, PDZ, SAM). Furthermore, many domains frequently present in proteins with polyQ fulfill functions in the nucleus (Supplementary Table S5).

In agreement with the association of polyQ protein content to phosphatidylinositol signaling that we found at genomic level, here we also identified some domains related to this function including, again, the PH domain. Many PH domains certainly bind PI (10–20%), other lipids, as well as peptides and proteins (37). Perhaps more specifically to PI, we also found the enrichment of the ENTH and ANTH domains. These two domains bind lipids (including PI) to recognize and manage vesicle coat components (38). Interestingly, the top domains found in relation to this function in the genomic association analysis (FYVE, PX) were not found here, indicating that PI signaling might be a function related to polyQ proteins but not directly performed by polyQ proteins.

Other functions found to be enriched in polyQ proteins were not found in the genomic analysis, indicating a more direct association than PI signaling. We observed several domains associated with chromatin maintenance. The Bromodomain, e.g. associated with polyQ in seven of

Table 2. Frequently overrepresented[a] functional annotations among polyQ proteins from 11 eukaryotic species

Category[b]	HS	BT	RN	MM	DR	DM	CE	AT	SC	DD	NC
Transcription-related	✓		✓	✓	✓	✓		✓	✓	✓	✓
Nucleus	✓		✓	✓	✓	✓	✓	✓	✓		✓
(RNA and nitrogen) metabolic or biosynthetic process	✓		✓	✓		✓		✓	✓	✓	✓
Compositionally biased region (Ser,Gly,Pro,Ala)	✓	✓	✓	✓		✓			✓		
Protein phoshporylation	✓			✓		✓			✓	✓	
Alternative splicing	✓			✓		✓		✓			
Protein dimerization activity						✓		✓		✓	
Developmental protein						✓	✓	✓			

[a]Using the web tool DAVID (25).
[b]We merged the resulting species-specific lists of functional terms (applying a *P*-value threshold of 0.05 after multiple testing correction with the Benjamini–Hochberg method) and replaced similar terms by representative substitutes.
HS = *Homo sapiens*, BT = *Bos taurus*, RN = *Rattus norvegicus*, MM = *Mus musculus*, DR = *Danio rerio*, DM = *Drosophila melanogaster*, CE = *Caenorhabditis elegans*, AT = *Arabidopsis thaliana*, SC = *Saccharomyces cerevisiae*, DD = *Dictyostelium discoideum*, NC = *Neurospora crassa*.

the analyzed species, is involved in the recognition of acetylated-lysines, e.g. in the N-termini of histones, and is found in chromatin associated proteins (39).

Function of proteins interacting with polyQ proteins

We found that polyQ proteins are associated with interactions between proteins and that polyQ proteins are enriched in some general functions related to the nucleus such as transcriptional regulation and chromatin maintenance among others. We wondered whether there would be particular domains or functions specific to the proteins interacting with polyQ proteins. These would account for other indirect functional associations of polyQ proteins.

To investigate this, we measured the significance of over-representation of predicted domains in non-polyQ proteins that interact with polyQ proteins (see 'Materials and Methods' section for details).

In our analysis of proteins interacting with polyQ proteins, we found 17 domains significantly enriched (occurring in more than 10 polyQ-protein interacting proteins; $P < 0.01$). The list of domains was manually curated to remove redundant entries and obvious false positive predictions (as detailed in Supplementary Methods). The set was different from the set of domains found to be enriched in polyQ proteins with the exception of nuclear hormone receptor domains (NHR and Zinc finger C4 type). In general functional terms, the list of domains contained again an important fraction of domains with nuclear functions (NHR, bZIP, MH1, Zinc finger MIZ type). However, it also included domains with non-nuclear functions (Ubiquitin family, AAA, EGF). The curated set of domains enriched in proteins interacting with the polyQ set is listed in Table 3.

For comparison, we tested the enrichment of protein domains in proteins interacting with polyP tracts (defined as for polyQ tracts: minimum number of 10 consecutive P allowing for one mismatch). The known polyP interacting domains SH3 and WW were among the top 15 most enriched domains (additionally, Actin, RhoGEF, SH2, BAR, Spectrin, Arf, FF, PX, FCH, WH1, CH, RhoGAP and the UBX domain) all being significantly associated to polyP ($P < 0.01$). The strength of these associations is comparable to that of the top polyQ associated domains.

PolyQ as a motif for protein interaction

Multiple observations presented above seem to indicate that polyQ is involved in PPI: polyQ proteins are related to dimerization, proteins with longer polyQ tracts tend to have more interaction partners, and many human protein complexes contain multiple polyQ proteins.

In addition, among the 86 human polyQ containing proteins, we counted 49 interactions where both interacting proteins contained polyQ tracts. This enrichment was as significant as the enrichment determined for domains present in proteins interacting with polyQ proteins (Table 3; *P*-value of 0.0023, see 'Materials and Methods' section).

Moreover, among the list of domains enriched in proteins that interact with human polyQ proteins, we observed the bZIP domain, which can form a coiled-coil. It was recently shown that polyQ regions overlap with coiled-coils regions in a set of polyQ containing proteins and are also found in their interaction partners (40). Coiled-coil domains are involved in oligomerization. This would explain the function of polyQ observed above as a motif of protein interaction ubiquitously used.

To determine whether the association of polyQ tracts and coiled-coil regions is a general phenomenon, we systematically studied the overlap of polyQ regions and predicted coiled-coil regions in polyQ proteins. For the prediction of coiled-coils, we applied the tool Coils (41), which detects hydrophobic heptad repeats in protein sequences. We considered only high-confidence predictions (over a probability threshold of 0.8). We observed a significant enrichment in human and in the other 10 eukaryotic species analyzed (Supplementary Table S4). For example, of the 109 polyQ tracts in 86 human proteins, 54 (50%) overlapped with a coiled-coil region and 5 more were in very close proximity (distance of 10 amino acids or less) (*P*-value 3.8e-15).

We found that the distribution of coiled-coils is extremely biased toward the N-terminus of the polyQ tract. In this respect, one has to note that the amino acid composition of the regions surrounding polyQ

Table 3. Domains overrepresented in proteins that interact with human polyQ proteins

Domain name	*P*-value	Pfam accs.	Reason for merging[a]	Interactions
Nuclear hormone receptor associated	0[b]	PF00104[c], PF00105[c]	Colocalization	95
EGF	0[b,d]	PF00008, PF07645	Colocalization and overlap	29
Zinc finger MIZ type	0.0016	PF02891, PF11789	Overlap	12
ATPase family associated with various cellular activities (AAA)	0.00	PF00004, PF05496	Colocalization and overlap	24
Ubiquitin family	0.002	PF00240, PF11976	Overlap	25
MH1 domain	0.007	PF03165, PF03166, PF10401	Colocalization and overlap	30
Basic Leucine Zipper Domain (bZIP)	0.0088	PF00170, PF07716	Overlap	37

[a]Details in Supplementary Methods.
[b]*P*-value remains under a significant level of 0.05 even after Benjamini and Hochberg correction for multiple testing.
[c]Over-represented in polyQ proteins (Supplementary Table S5).
[d]Result is reproducible in *Drosophila melanogaster*.

tracts is biased for some amino acids. For example, we could detect enrichment in proline and histidine around polyQ tracts in several organisms including human, fly and yeast. The described amino acid bias is not evenly distributed to both sides of the polyQ stretch. In human sequences, the most extreme case is found for prolines, which often appear as polyP tracts almost exclusively C-terminally to the polyQ. For example, in the set of 86 human polyQ proteins, 13 proteins contain a polyP run of at least three residues at a maximum distance of three amino acids from the polyQ. Of those, we found 12 C-terminally and only 1 N-terminally to the polyQ stretch. For the other enriched amino acids (aspartic, methionine, histidine), the number of single amino acid tracts surrounding polyQ was actually too small to assess any distributional bias. This finding corresponds to the findings of Bhattacharyya *et al.* (42) that polyproline (polyP) stretches inhibit polyQ-dependent aggregate formation only when located at the C-terminus of the polyQ tract and support the ability of polyP to stabilize protein conformation situated near their N-terminus (43).

To exclude the possibility that the bias of coiled-coils toward the N-terminus of polyQ tracts is simply due to C-terminal polyP, we analyzed the position of coiled-coils with respect to polyQ tracts excluding cases where polyP was present. The bias was still observed both in human (34 N-terminal versus 6 C-terminal) and yeast (14 N-terminal versus 1 C-terminal), suggesting that the association of coiled-coils to polyQ tracts is asymmetric.

Finally, we could establish that also non-polyQ proteins interacting with polyQ proteins are significantly enriched in coiled-coil regions. This enrichment was stronger than the enrichment found for protein domains ($P < 2.2e-16$). To exclude that the observed colocalization of coiled-coils with polyQ stretches is an artifact of the prediction tool applied (e.g. over-predicting spurious coiled-coil regions on polyQ stretches), we repeated the coiled-coil prediction on the human polyQ set with a different prediction tool [Paircoil2 (44)]. We found, again, a significant enrichment of coiled-coils in the polyQ set (for a tool specific threshold of 0.025, we observed 30 proteins with a coiled-coil among the 86 human polyQ proteins at a background prediction rate of 13%; $P < 4e-9$).

To further substantiate our observations, we deleted the polyQ stretches from the sequences and repeated the coiled-coil prediction in 45 human proteins hosting 54 polyQ stretches that were either overlapping a coiled-coil or in close proximity of one. We excluded from this analysis those proteins where the coiled-coil was predicted to be within a polyQ stretch and those with a C-terminal polyP. We counted how often we could still observe a coiled-coil prediction in the 10 residues flanking each side of the 54 deletion sites. In 11 of the 54 cases, a coiled-coil was predicted, which corresponds to a 6-fold enrichment over the background frequency of predicted coiled-coil regions in all human protein sequences. This enrichment was significant ($P = 2.0e-6$; probability of observing 11 or more coiled-coil regions under the Binomial distribution). This result proves that the association of polyQ to coiled-coil regions is not just due to the presence of polyQ but that also its flanking sequences have significant coiled-coil forming potential.

In summary, we found a significant association of polyQ tracts after coiled-coil regions. This agrees with a function of polyQ tracts related to protein interactions. We wondered if functions previously noted to be associated to polyQ proteins (12,16,17) could be just a secondary effect and explained simply by the fact that those functions (e.g. transcriptional regulation) require more protein interactions than other functions (e.g. metabolism). Therefore, we tested if we observed a similarly high enrichment in certain GO terms when we compared proteins with many interactions to all proteins with at least one known interaction (see 'Materials and Methods' section for details). Indeed, many functions associated to the polyQ set are also enriched in the set of proteins with the 10% highest number of interaction partners. For example, both in yeast and human, we observed in the set of proteins with many partners a significant enrichment of the GO term GO:0031981-nuclear lumen (*P*-values 6.3e-53 and 4.2e-28) and in human proteins of the term GO:0008134-transcription factor binding (*P*-value 1.7e-23). This effect is independent of the precise protein set size and can be reproduced, e.g. with the 86 highest degree proteins in human (a cutoff chosen in accordance with the size of the human polyQ set).

DISCUSSION

PolyQ tracts in protein sequences have been researched mostly because of their pathogenic expansion in multiple human genetic diseases. However, their presence in many wild-type proteins across a variety of species is intriguing and suggests that normal polyQ tracts might have a function.

Following this idea, we provide evidence collected at multiple inter-related biological levels that collectively and consistently indicates that polyQ tracts are involved in protein interactions, e.g. because of their enrichment in protein complexes, and their association with coiled-coil regions. Through our analyses, we noted other features of polyQ tracts, which may not be directly related to their function as an interaction motif, but to the pathogenic effects of their abnormal expansion.

At the nucleotide level, we could observe selection of CAG repeats in exons of human, mouse, rat and fly genes. Intriguingly, we found them also enriched, although at a lower level, in untranslated regions of transcripts (UTRs). It was noted that both CAG and CTG repeats can form RNA • DNA hybrids (R loops) that could have a biological function (31,32). In agreement with this, we found CTG and CGG repeats similarly enriched in UTRs but not so much in exons.

Along these lines, we found that whereas only 13% of prolines in human proteins are encoded by the rare codon CCG, this fraction is higher in prolines forming polyP (of length three or more) (23%), and even higher (43%; $n = 48$ codons) if the polyP is near uninterrupted polyQ sequences of minimum length 10 (at a maximum distance of three amino acids). We observed a related effect in polyQ, which are encoded more frequently by CAG codons when the polyQ sequences are close to polyP tracts ($n = 156$) as compared to other polyQ ($n = 1169$) (90% versus 79%). This inter-dependence between GC rich codons hints at an effect at the transcript level. In summary, we interpret these results as indicating that CAG repeats are under positive selection at the nucleotide level. Abnormally expanded CTG repeats bind muscleblind resulting in Myotonic dystrophy type 1 (30). CAG and CTG repeats might bind to proteins in their wild-type transcripts.

The mechanisms that have been proposed to originate regions encoding poly-amino acid repeats are not well known (45). Many polyQ tracts are composed exclusively of CAG repeats in mammals (46) (the other possible codon encoding Q being CAA); this is interpreted as evidence of their formation due to trinucleotide expansion by gene slippage, resulting from the formation of an abnormal loop of the CAG repeat via CG pairings. According to this, it was shown that polyQ-coding pure CAG repeats are expanded from mouse to human (47) while expansion does not occur if they are formed by a mix of CAG and CAA codons. In contrast, in some non-mammalian organisms polyQ tracts tend to be encoded by pure CAA repeats [e.g. *Drosophila* (46) and *D. discoideum* (10)] actually suggesting that they are selected to resist slippage.

The abundance of proteins with polyQ tracts across different species is highly variable, being, e.g. absent from prokaryotic organisms. In a few species, polyQ tracts are among the most frequent amino acid repeats (14,48). This variability may be related to the inability of some species to deal with these aggregation-prone repeats. Therefore, analysis of the correlation between systemic properties of species and presence of polyQ proteins might hint at the mechanisms by which species deal with polyQ proteins and at the origin of their pathogenic effects. We investigated this systematically and observed a huge variation between species in content of polyQ proteins (e.g. highest in *D. discoideum* and *D. melanogaster* versus, e.g. very low for *Xenopus* or *D. rerio*). We observed that those species with high polyQ protein content have a higher number of proteins bearing domains with functions related to phosphatidylinositol (PI) signaling and ubiquitin-directed protein degradation. Interestingly, both ubiquitin and phosphatidylinositol play a role in the clearance of polyQ aggregates: aggregates containing proteins with an expanded polyQ stretch have been shown to be ubiquitinated (49). Phosphatidylinositol-binding domains are involved in targeting polyQ aggregates to membranes during the process of macroautophagy. For example, the FYVE domain containing human protein Alfy promotes the degradation of huntingtin in mammalian cells (50). Therefore, this association between high content of polyQ proteins at the genomic level to both PI signaling and ubiquitin-directed protein degradation could be explained by the need of the cell to effectively degrade polyQ containing aggregation-prone proteins. We speculate that differential selection explains the high content of polyQ proteins in some organisms (47): organisms that can select polyQ co-evolve the appropriate machinery to clear polyQ protein aggregates, whereas organisms lacking strong clearance mechanisms for protein-aggregates might not tolerate polyQ proteins at all; this could explain the absence of polyQ proteins in the prokaryotic kingdom.

When analyzing the variability of polyQ protein occurrence in a large variety of species in more detail, we observed that polyQ tracts are not a feature characteristic of particular gene families. Variability of polyQ protein content among species is therefore due to orthologs of a protein having a polyQ tract in one species and not in another. Moreover, we demonstrated that particular protein families show multiple events of emergence of polyQ and that they can happen at different positions of the sequence. For example, the human huntingtin has a polyQ tract situated near the N-terminus of the protein, whereas many lower organisms have none, e.g. *Ciona* (51). However, the huntingtin of the *Drosophila* genus has multiple polyQ tracts, none of them in the N-terminus. This suggests that particular protein families, including huntingtin, are under selective pressure to accumulate polyQ tracts. In summary, the fact that some organisms have orthologs lacking the polyQ tract indicates that the protein can fulfill its tasks without this feature: its function cannot be essential. In addition, the fact that the polyQ tracts can occupy different positions in the sequence suggests that polyQ tracts perform a function without

strong positional requirements. On the other hand, they do have some function specific to particular protein families since evolutionary pressure to insert the polyQ tracts leads to this occurring in distantly related clades. In terms of speed, the evolutionary expansion of a polyQ tract is much slower than a pathological expansion. For example, the expansion of the N-terminal polyQ tract in human huntingtin could be estimated to have evolved at an average rate of one Q per 30 million years. This is indicative of how delicate the effect of modification of polyQ tract length can be (Figure 3a, left box). On the other hand, the large variations of the huntingtin mid-of-sequence polyQ tracts in the different *Drosophilae* indicate that fast evolution of polyQ tracts is also possible (Figure 3a, right box).

There is already some experimental evidence suggesting that the function of polyQ could be to modulate protein–protein interactions (PPIs). For example, a polyQ sequence in TBP modulates its interaction with TFIIB (52), and a glutamine-rich activation domain in SP1 directly interacts with TAF4 in *Drosophila* (53). It was observed in an *in vitro* experiment that mouse Sp1 and some components of the core transcription apparatus (e.g. TFIID and TFIIF) are direct targets inhibited by mutant huntingtin in a polyglutamine-dependent manner (54). In addition, mutant proteins with enlarged polyQ tracts aggregate, which also points to a relation between polyQ and protein interactions (2–4). Given this evidence, we wondered whether the non-specific function of wild-type polyQ could be to modulate protein–protein interactions (PPIs): such a role is coherent with our observations described earlier. We were able to provide further evidence to support this hypothesis at multiple levels. For example, we could detect that polyQ proteins have more protein interaction partners than non-polyQ proteins and have a higher tendency to interact with other polyQ proteins than non-polyQ proteins.

In agreement to previous studies (55), we identified an over-representation of protein domains related to nucleus-based functions in polyQ proteins but also in proteins that interact with them. In fact, we could demonstrate that these functions are also over-represented in proteins with many interactions. Therefore, we deduce that the functional biases observed in polyQ and polyQ-interacting proteins are due to the involvement of polyQ in protein interactions.

Until now, the structural basis for the possible modulation of protein interactions by polyQ is not clear. To begin with, the precise structure of polyQ itself is unknown and suggested conformations of both synthetic polyQ peptides and naturally occurring proteins with polyQ tracts include alpha helix, random coil, and extended loop [e.g. huntingtin exon1 (56)]. This might be due to polyQ adopting an unstable context-dependent structure. Part of this context can be flanking sequences, which have been shown to influence both the structure (56) and aggregation properties of polypeptides with polyQ tracts (57). Length expansions of polyQ stretches seem to be accompanied by a transition of a random coil into a beta sheet structure (58) (which would account for its pathogenic effect). In addition, polyQ tracts seem to be able to modify the conformation of structured domains nearby in sequence (59). Such interactions could be dependent on the presence of other interacting proteins, and it has recently been suggested that the mechanisms by which polyQ modulate protein interactions might be the expansion of sequence-adjacent coiled-coil regions upon interaction of the coiled-coil region with another protein (40).

In support of this view, we found a very strong association between polyQ and coiled-coil regions: both are found in the same sequence more often than random expectation, overlapping or at very short distance, as well as in proteins that interact with each other. This association was by far more significant than the association of polyQ to any protein domain. In summary, our results underline that polyQ expansions are selected in evolution to extend coiled-coil regions that take part in protein–protein interactions.

We found a strong bias for coiled-coil regions to be situated N-terminally of polyQ tracts. At the same time, polyP is sometimes found near polyQ and if so, often C-terminally to the polyQ tract. This is in agreement with the finding that polyP stabilizes the structure of adjacent polyQ when located C-terminally but not when located N-terminally of it. This directional property of polyP to influence conformation is known in contexts other than polyQ proteins [see (42) and references therein]. In an X-ray study of huntingtin exon 1, the polyP was found to adopt a classical poly-proline helix structure (left-handed helix) (56). The conformational extension by polyQ of the coiled-coil region is then stabilized and paused at the polyP region. According to this evidence, we propose that polyQ tracts have a tendency to follow a coiled-coil region that they expand upon protein interaction and in turn to be followed by a capping sequence which, like polyP, acts directionally to stabilize and stop the growth of the helical region (Figure 5).

With respect to the properties of the sequences surrounding polyQ tracts, it is interesting to note that in a recent computational study polyQ tracts were associated to the presence of disordered regions (60). Indeed, we could confirm a significant enrichment of disordered regions [as predicted by the tool RONN (61)] nearby polyQ compared to all human proteins, though this was smaller than the one found for coiled-coil regions (3.3-fold versus 6-fold, data not shown).

In summary, our results lead to the following general picture of the function of polyQ: its activity as a motif for protein interaction is tightly related to the length of the polyQ tract itself, the character of the sequences adjacent to it and to the concentration of interacting protein partners. We assume that the normal interplay of all these elements would lead to an enhanced, highly stable and specific interaction. However, the complexity of this system also suggests that small perturbations could lead to pathological interactions either with altered affinities or with different partners. The complex interaction of factors influencing the function of polyQ tracts perhaps explains why so many processes have been found to contribute to the pathomechanism of polyQ diseases including transcriptional dysregulation (62), RNA

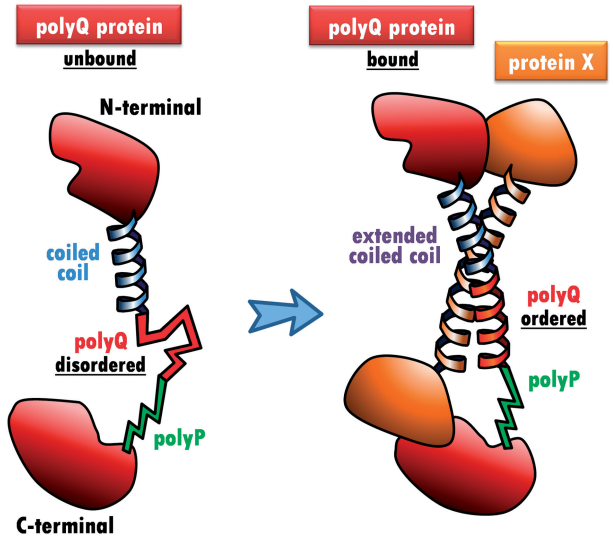

Figure 5. Cartoon of proposed polyQ function in protein interaction. Left: a polyQ protein contains a coiled-coil (blue), followed by a polyQ region (red) and a polyP region (green). In the unbound state, the polyQ region is disordered. Right: upon interaction with a protein partner X, the polyQ region adopts a coiled-coil structure that extends the original coiled-coil. The polyP region remains unstructured capping precisely the extension of the coiled-coil.

toxicity (63), impairment of the ubiquitin-proteasome system (64–66), mitochondrial dysfunction (67) and disturbed calcium signaling (68).

Our results also suggest that a given species may accumulate an abundance of polyQ proteins to modulate many protein interactions. However, this may come at no small expense: protein networks with abundant polyQ proteins may be in a delicate balance in which aggregation can occur depending on the concentration of many molecules. This balance might be lost in specific tissues and circumstances as mechanisms to keep protein aggregates in check get challenged in ageing cells [as it has been observed in *C. elegans* (69)]. This may explain why neurons of the elderly are particular prone to anomalous polyQ expansion and in turn neurodegeneration.

We suggest that the study of the function of wild-type and pathogenic polyQ proteins will require experiments to test the variation in functionality that removing or expanding particular polyQ stretches will produce. Specifically, it needs to be investigated how these modifications influence the interaction abilities of the polyQ protein. Gain or loss of interactions with other proteins with coiled-coil regions and polyQ should be paid special attention. Explaining and predicting the effects of polyQ tracts will require elaborate analysis for each particular situation. The recent analysis of SCA1, where dramatic differences in the effects of the pathogenic protein were observed between brain regions (70), supports this idea.

In conclusion, our work has approached the study of CAG/glutamine repeats integrating analyses at the nucleotide and protein level with phylogenetic data, genomic analyses and studies of protein interaction networks. The wild-type function of polyQ tracts is to modulate protein interactions in dependency of their molecular context. Therefore, their study will require correlating modification of this context to modifications in the protein interaction network.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figure 1 and Supplementary Methods.

ACKNOWLEDGEMENTS

We thank Sigrid Schnögl (MDC, Berlin) for critical reading of the manuscript.

FUNDING

Program of medical genome research NGFNp by the German Ministry of Education and Research (BMBF) (reference numbers 01GS08169-171 and 01GS0844 to E.E.W. and 01GS08170 to M.A. A.-N.), DFG Collaborative Research Centre grants (SFB577, SFB740 to E.E.W; SFB618 to E.E.W. and M.A. A.-N.), Helmholtz Alliance for Systems Biology grant to E.E.W and M.A. A.-N., Huntington's Disease Society of America (HDSA) and Cure Huntington's Disease Initiative (CHDI) grants to E.E.W. Funding for open access charge: Max Delbrück Center for Molecular Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Gatchel,J.R. and Zoghbi,H.Y. (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat. Rev. Genet.*, **6**, 743–755.
2. Tran,P.B. and Miller,R.J. (1999) Aggregates in neurodegenerative disease: crowds and power? *Trends Neurosci.*, **22**, 194–197.
3. Kopito,R.R. (2000) Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol.*, **10**, 524–530.
4. Ross,C.A. (1997) Intranuclear neuronal inclusions: a common pathogenic mechanism for glutamine-repeat neurodegenerative diseases? *Neuron*, **19**, 1147–1150.
5. 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.
6. Chai,Y., Shao,J., Miller,V.M., Williams,A. and Paulson,H.L. (2002) Live-cell imaging reveals divergent intracellular dynamics of polyglutamine disease proteins and supports a sequestration model of pathogenesis. *Proc. Natl Acad. Sci. USA*, **99**, 9310–9315.
7. Kuemmerle,S., Gutekunst,C.A., Klein,A.M., Li,X.J., Li,S.H., Beal,M.F., Hersch,S.M. and Ferrante,R.J. (1999) Huntington aggregates may not predict neuronal death in Huntington's disease. *Ann. Neurol.*, **46**, 842–849.
8. Pennuto,M., Palazzolo,I. and Poletti,A. (2009) Post-translational modifications of expanded polyglutamine proteins: impact on neurotoxicity. *Hum. Mol. Genet.*, **18**, R40–R47.
9. Butland,S.L., Devon,R.S., Huang,Y., Mead,C.L., Meynert,A.M., Neal,S.J., Lee,S.S., Wilkinson,A., Yang,G.S., Yuen,M.M. *et al.* (2007) CAG-encoded polyglutamine length polymorphism in the human genome. *BMC Genomics*, **8**, 126.
10. Eichinger,L., Pachebat,J.A., Glockner,G., Rajandream,M.A., Sucgang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q.

et al. (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.
11. von Mikecz,A. (2009) PolyQ fibrillation in the cell nucleus: who's bad? *Trends Cell Biol*., **19**, 685–691.
12. Karlin,S. and Burge,C. (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl Acad. Sci. USA*, **93**, 1560–1565.
13. Huntley,M. and Golding,G.B. (2000) Evolution of simple sequence in proteins. *J. Mol. Evol*., **51**, 131–140.
14. Faux,N.G., Bottomley,S.P., Lesk,A.M., Irving,J.A., Morrison,J.R., de la Banda,M.G. and Whisstock,J.C. (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*., **15**, 537–551.
15. Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, **245**, 371–378.
16. Alba,M.M. and Guigo,R. (2004) Comparative analysis of amino acid repeats in rodents and humans. *Genome Res*., **14**, 549–554.
17. Harrison,P.M. (2006) Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. *BMC Bioinformatics*, **7**, 441.
18. Hands,S., Sinadinos,C. and Wyttenbach,A. (2008) Polyglutamine gene function and dysfunction in the ageing brain. *Biochim. Biophys. Acta*, **1779**, 507–521.
19. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
20. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.
21. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*., **30**, 276–280.
22. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al*. (2008) TreeFam: 2008 Update. *Nucleic Acids Res*., **36**, D735–D740.
23. Schaefer,M.H., Fontaine,J.F., Vinayagam,A., Porras,P., Wanker,E.E. and Andrade-Navarro,M.A. (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*., **7**, e31826.
24. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*., **39**, D698–D704.
25. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*., **4**, P3.
26. Strand,M., Prolla,T.A., Liskay,R.M. and Petes,T.D. (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, **365**, 274–276.
27. Kozlowski,P., de Mezer,M. and Krzyzosiak,W.J. (2010) Trinucleotide repeats in human genome and exome. *Nucleic Acids Res*., **38**, 4027–4039.
28. Holmes,S.E., Hearn,E.O., Ross,C.A. and Margolis,R.L. (2001) SCA12: an unusual mutation leads to an unusual spinocerebellar ataxia. *Brain Res. Bull*., **56**, 397–403.
29. Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*., **37**, D93–D97.
30. Miller,J.W., Urbinati,C.R., Teng-Umnuay,P., Stenberg,M.G., Byrne,B.J., Thornton,C.A. and Swanson,M.S. (2000) Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *EMBO J*., **19**, 4439–4448.
31. Reddy,K., Tam,M., Bowater,R.P., Barber,M., Tomlinson,M., Nichol Edamura,K., Wang,Y.H. and Pearson,C.E. (2010) Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res*., **39**, 1749–1762.
32. Lin,Y., Dent,S.Y., Wilson,J.H., Wells,R.D. and Napierala,M. (2010) R loops stimulate genetic instability of CTG.CAG repeats. *Proc. Natl Acad. Sci. USA*, **107**, 692–697.
33. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
34. Stenmark,H., Aasland,R. and Driscoll,P.C. (2002) The phosphatidylinositol 3-phosphate-binding FYVE finger. *FEBS Lett*., **513**, 77–84.
35. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res*., **37**, D229–D232.
36. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V. *et al*. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*., **36**, D646–D650.
37. DiNitto,J.P. and Lambright,D.G. (2006) Membrane and juxtamembrane targeting by PH and PTB domains. *Biochim. Biophys. Acta*, **1761**, 850–867.
38. Duncan,M.C. and Payne,G.S. (2003) ENTH/ANTH domains expand to the Golgi. *Trends Cell Biol*., **13**, 211–215.
39. Dhalluin,C., Carlson,J.E., Zeng,L., He,C., Aggarwal,A.K. and Zhou,M.M. (1999) Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, **399**, 491–496.
40. Fiumara,F., Fioriti,L., Kandel,E.R. and Hendrickson,W.A. (2010) Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell*, **143**, 1121–1135.
41. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
42. Bhattacharyya,A., Thakur,A.K., Chellgren,V.M., Thiagarajan,G., Williams,A.D., Chellgren,B.W., Creamer,T.P. and Wetzel,R. (2006) Oligoproline effects on polyglutamine conformation and aggregation. *J. Mol. Biol*., **355**, 524–535.
43. Hinderaker,M.P. and Raines,R.T. (2003) An electronic effect on protein structure. *Protein Sci*., **12**, 1188–1194.
44. McDonnell,A.V., Jiang,T., Keating,A.E. and Berger,B. (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, **22**, 356–358.
45. Kovtun,I.V. and McMurray,C.T. (2008) Features of trinucleotide repeat instability in vivo. *Cell Res*., **18**, 198–213.
46. Alba,M.M., Santibanez-Koref,M.F. and Hancock,J.M. (2001) The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila. *J. Mol. Evol*., **52**, 249–259.
47. Hancock,J.M. (1995) The contribution of slippage-like processes to genome evolution. *J. Mol. Evol*., **41**, 1038–1047.
48. Karlin,S., Brocchieri,L., Bergman,A., Mrazek,J. and Gentles,A.J. (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl Acad. Sci. USA*, **99**, 333–338.
49. Suhr,S.T., Senut,M.C., Whitelegge,J.P., Faull,K.F., Cuizon,D.B. and Gage,F.H. (2001) Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *J. Cell. Biol*., **153**, 283–294.
50. Filimonenko,M., Isakson,P., Finley,K.D., Anderson,M., Jeong,H., Melia,T.J., Bartlett,B.J., Myers,K.M., Birkeland,H.C., Lamark,T. *et al*. (2010) The selective macroautophagic degradation of aggregated proteins requires the PI3P-binding protein Alfy. *Mol. Cell*, **38**, 265–279.
51. Gissi,C., Pesole,G., Cattaneo,E. and Tartari,M. (2006) Huntingtin gene evolution in Chordata and its peculiar features in the ascidian Ciona genus. *BMC Genomics*, **7**, 288.
52. Friedman,M.J., Shah,A.G., Fang,Z.H., Ward,E.G., Warren,S.T., Li,S. and Li,X.J. (2007) Polyglutamine domain modulates the TBP-TFIIB interaction: implications for its normal function and neurodegeneration. *Nat. Neurosci*., **10**, 1519–1528.
53. Hoey,T., Weinzierl,R.O., Gill,G., Chen,J.L., Dynlacht,B.D. and Tjian,R. (1993) Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. *Cell*, **72**, 247–260.
54. Zhai,W., Jeong,H., Cui,L., Krainc,D. and Tjian,R. (2005) In vitro analysis of huntingtin-mediated transcriptional

repression reveals multiple transcription factor targets. *Cell*, **123**, 1241–1253.
55. Whan,V., Hobbs,M., McWilliam,S., Lynn,D.J., Lutzow,Y.S., Khatkar,M., Barendse,W., Raadsma,H. and Tellam,R.L. (2010) Bovine proteins containing poly-glutamine repeats are often polymorphic and enriched for components of transcriptional regulatory complexes. *BMC Genomics*, **11**, 654.
56. Kim,M.W., Chelliah,Y., Kim,S.W., Otwinowski,Z. and Bezprozvanny,I. (2009) Secondary structure of Huntingtin amino-terminal region. *Structure*, **17**, 1205–1212.
57. Dehay,B. and Bertolotti,A. (2006) Critical role of the proline-rich region in Huntingtin for aggregation and cytotoxicity in yeast. *J. Biol. Chem.*, **281**, 35608–35615.
58. Perutz,M.F. (1996) Glutamine repeats and inherited neurodegenerative diseases: molecular aspects. *C Opin. Struct. Biol.*, **6**, 848–858.
59. Ignatova,Z. and Gierasch,L.M. (2006) Extended polyglutamine tracts cause aggregation and structural perturbation of an adjacent beta barrel protein. *J. Biol. Chem.*, **281**, 12959–12967.
60. Simon,M. and Hancock,J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.*, **10**, R59.
61. Yang,Z.R., Thomson,R., McNeil,P. and Esnouf,R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.
62. Truant,R., Atwal,R.S. and Burtnik,A. (2007) Nucleocytoplasmic trafficking and transcription effects of huntingtin in Huntington's disease. *Prog. Neurobiol.*, **83**, 211–227.
63. Li,L.B., Yu,Z., Teng,X. and Bonini,N.M. (2008) RNA toxicity is a component of ataxin-3 degeneration in Drosophila. *Nature*, **453**, 1107–1111.
64. Chai,Y., Koppenhafer,S.L., Shoesmith,S.J., Perez,M.K. and Paulson,H.L. (1999) Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro. *Hum. Mol. Genet.*, **8**, 673–682.
65. Bence,N.F., Sampat,R.M. and Kopito,R.R. (2001) Impairment of the ubiquitin-proteasome system by protein aggregation. *Science*, **292**, 1552–1555.
66. Waelter,S., Boeddrich,A., Lurz,R., Scherzinger,E., Lueder,G., Lehrach,H. and Wanker,E.E. (2001) Accumulation of mutant huntingtin fragments in aggresome-like inclusion bodies as a result of insufficient protein degradation. *Mol. Biol. Cell*, **12**, 1393–1407.
67. Lin,M.T. and Beal,M.F. (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, **443**, 787–795.
68. Tang,T.S., Slow,E., Lupu,V., Stavrovskaya,I.G., Sugimori,M., Llinas,R., Kristal,B.S., Hayden,M.R. and Bezprozvanny,I. (2005) Disturbed Ca2+ signaling and apoptosis of medium spiny neurons in Huntington's disease. *Proc. Natl Acad. Sci. USA*, **102**, 2602–2607.
69. David,D.C., Ollikainen,N., Trinidad,J.C., Cary,M.P., Burlingame,A.L. and Kenyon,C. (2010) Widespread protein aggregation as an inherent part of aging in *C. elegans*. *PLoS Biol.*, **8**, e1000450.
70. Jafar-Nejad,P., Ward,C.S., Richman,R., Orr,H.T. and Zoghbi,H.Y. (2011) Regional rescue of spinocerebellar ataxia type 1 phenotypes by 14-3-3epsilon haploinsufficiency in mice underscores complex pathogenicity in neurodegeneration. *Proc. Natl Acad. Sci. USA*, **108**, 2142–2147.