Computational and Structural Biotechnology Journal 18 (2020) 306–313

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

The importance of definitions in the study of polyQ regions: A tale of thresholds, impurities and sequence context

Pablo Mier [a,*], Carlos Elena-Real [b], Annika Urbanek [b], Pau Bernadó [b], Miguel A. Andrade-Navarro [a]

[a] Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, Hans-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany
[b] Centre de Biochimie Structurale (CBS), INSERM, CNRS, Université de Montpellier, 29, rue de Navacelles, 34090 Montpellier, France

ARTICLE INFO

Article history:
Received 23 September 2019
Received in revised form 13 December 2019
Accepted 30 January 2020
Available online 4 February 2020

Keywords:
Homorepeat
polyQ
Glutamine
Sequence context
Codon usage

ABSTRACT

Polyglutamine (polyQ) regions are one of the most prevalent homorepeats in eukaryotes. It is however difficult to evaluate their prevalence because various studies claim different results. The reason is the lack of a consensus to define what is indeed a polyQ region. We have tackled this issue by studying how the use of different thresholds (i.e., minimum number of glutamines required in a protein region of a given size), to detect polyQ regions in the human proteome influences not only their prevalence but also their general features and sequence context. Threshold definition shapes the length distribution of the polyQ dataset, and changes the observed number and position of impurities (amino acids other than glutamine) within polyQ regions. Irrespective of the chosen threshold, leucine and proline residues are enriched both within and around polyQ. While leucine is enriched at the N-terminus of polyQ and specially at position −1 (amino acid preceding the polyQ), proline is prevalent in the C-terminus (positions +1 to +5, that is, the first five amino acids after the polyQ). We also checked the suitability of these thresholds for other species, and compared their polyQ features with those found in humans. As the sequence context and features of polyQ regions are threshold-dependent, we propose a method to quickly scan the polyQ landscape of a proteome. We complement our results with a summarized overview about which biases are to be expected per threshold when studying polyQ regions.

1. Introduction

Homorepeat regions, homopolymeric stretches or polyX are defined as runs of a given amino acid (X) in a protein sequence. They are present in around 15% of eukaryotic proteins [1–3]. Some amino acids never or rarely form homorepeats (R, C, V, I, M, F, Y, W), while others are distinctly prevalent depending on the species [4]. Cases of extreme abundance are polyN in *Plasmodium falciparum* [5,6], polyQ and polyN in *Dictyostelium discoideum* [7], and polyA in *Chlamydomonas reinhardtii* [4]. Repeats of polyQ and polyA are prone for expansion via strand slippage of the encoding DNA and have been linked to several inherited diseases in human when the number of consecutive residues in the run exceeds a certain threshold [8–13].

It has been established that homorepeats with lengths of five amino acids or more occur non-randomly [14,15], meaning that they might be under evolutionary selection pressure due to functional or structural constraints. In order to detect homorepeats, several thresholds have been used in the literature, providing different results. They all depend on a sliding window with a defined length (L) and a minimum number of a given amino acid (N); if at least one window in a region meets the threshold N/L, the region is considered to be a homorepeat. Therefore, the relative abundance and length of homorepeats depend directly on that threshold. Indeed, definition problems are general to low complexity regions in proteins [16].

Here we take polyQ regions as an example due to their abundance in human proteins, the growing interest in their function and their connection to ten neurodegenerative diseases, including Huntington's disease and several ataxias [9–11,17]. While initial interest in polyQ focused on their association to the pathologies elicited by glutamine-encoding CAG trinucleotide expansions [18,19], polyQ regions have also been associated with the stabilization of protein-protein interactions [20].

Quantitatively, polyQ is the most prevalent homorepeat in eukaryotes using either thresholds 6/6 [21] and 4/7 [22], and the second most prevalent when using threshold 7/7 [23]. Specifically, in metazoans, they are the most prevalent (threshold 6/6 [21]) and the sixth most prevalent (threshold 8/10 [4]) homorepeat, while in

* Corresponding author.
E-mail address: munoz@uni-mainz.de (P. Mier).

https://doi.org/10.1016/j.csbj.2020.01.012

humans polyQs are the seventh (thresholds 5/5 [1] and 8/10 [4]) and the eighth (threshold 5/5 [15]) most prevalent. The different order in the last example, even with the same threshold, may be due to the use of a different human proteome dataset. From these reports, it becomes obvious that there is no consensus in how to define polyQ repeats and that using different thresholds significantly changes their detection and consequently their associated properties.

The purpose of the present work is to evaluate how the threshold used to identify polyQ regions influences the main features of the retrieved polyQ datasets. Actually, we find that the amino acid and nucleotide sequence context of a polyQ region are largely related to its length and purity, both determined by the threshold. We performed our study using human sequences and complemented it with the proteomes of seven additional model organisms to identify these polyQ features that are species-dependent.

2. Methods

2.1. Data retrieval

The complete reference human proteome (73928 proteins) was downloaded from the UniProtKB database release 2019_02 (Uniprot > Proteomes; https://www.uniprot.org), and taken as the default amino acid dataset for most of the calculations. Similarly, we downloaded from the same site the complete reference proteomes of seven model organisms from diverse taxonomic groups: *Mus musculus* (54425 proteins), *Danio rerio* (46926 proteins), *Drosophila melanogaster* (21923 proteins), *Caenorhabditis elegans* (26893 proteins), *Saccharomyces cerevisiae* (6049 proteins), *Arabidopsis thaliana* (39380 proteins) and *Dictyostelium discoideum* (12746 proteins).

All protein coding transcripts of the human genome (GRCh38.p12; 92090 sequences) were retrieved from the Ensembl database (Ensembl > Biomart), by using "Gene type: protein_coding" and "Transcript type: protein_coding" as filters. The same procedure was followed to download the transcripts of *M. musculus* (GRCm38.p6; 59780 sequences), *D. rerio* (GRCz11; 51249 sequences), *D. melanogaster* (BDGP6.22; 30504 sequences), *C. elegans* (WBcel235; 33391 sequences), *S. cerevisiae* (R64-1-1; 6600 sequences), *A. thaliana* (TAIR10; 48321 sequences) and *D. discoideum* (dicty_2.7; 13233 sequences).

2.2. Threshold usage

Several thresholds of a minimum number of glutamines (N) in a window of a given size (L), noted as N/L, were used to consider whether a region is annotated as a polyQ. These were: 4/4, 4/6, 6/6, 6/8 and 8/10. Sequences were scanned with an in-house Perl script using the thresholds independently. This means that a region may be considered to be a polyQ with one or more thresholds (i.e. a pure stretch of six glutamines will be found with thresholds 4/4, 4/6, 6/6 and 6/8). Once the script finds a window matching the threshold, it slides forward until a window does not match; then, the polyQ region is annotated from the first matching window until the last one (Fig. 1). A set of polyQ regions was obtained per dataset and per threshold.

An additional constraint is taken into consideration to avoid nonspecific polyQ starts and ends. As an example, given the threshold 4/6 in the sequence "RRRAQAQQQRRR" the polyQ could be taken as "AQAQQQR" (both windows "AQAQQQ" and "QAQQQR" match the threshold). However, we chop off the beginning and the end of the region until a glutamine is found; the polyQ itself is taken then as "QAQQQ". This implies that the polyQ may be shorter than the window used to look for it, as long as it contains the minimum number of specified glutamines; in this case, the length is five but it has already four glutamines. The chopped amino acids are indeed the positions −1 and +1 with respect to the polyQ.

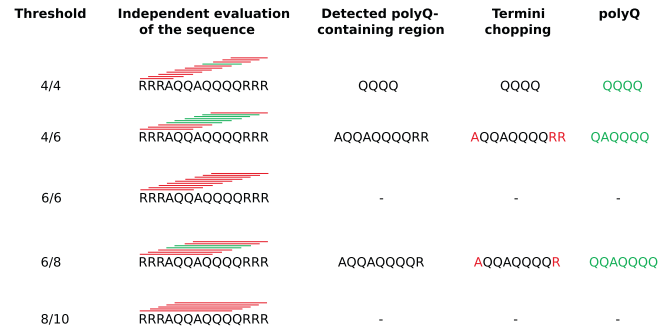

Fig. 1. Example of threshold usage. Scanning of sequence RRRAQQAQQQQRRR with thresholds 4/4, 4/6, 6/6, 6/8 and 8/10 to look for polyQ regions. Green and red lines depict windows matching and not matching the threshold, respectively. Their lengths depend on the window length (L) of the threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In addition, nucleotide sequences were also scanned with the above-described procedure, but using triplets as units instead of amino acids, in order to look for polyCA[G/A] regions coding for polyQ. Here the threshold is defined as the minimum number of codons CAG or CAA in a window of a given number of codons (keeping the protein coding frame).

2.3. Sequence context calculation

Once a polyQ is detected in a region, we extract both the preceding and the following ten amino acids defining its sequence context. Regarding impurities, non-glutamine amino acids within the polyQ, we studied their proportion and relative position. The latter was calculated as a function of the length of the polyQ; in a 10-amino acid polyQ, an impurity in the fifth position would have a relative position of 5/10 = 0.50.

Background composition for each amino acid and codon was calculated using the respective proteome (e.g. human) and the set of protein coding transcripts, respectively, as reference.

3. Results & discussion

3.1. The features of polyQ regions are threshold-dependent

PolyQ regions have been generally defined as a part of a protein sequence with several consecutive glutamines, but there is no consensus neither about which minimum number is to be considered as "several" nor whether non-glutamine amino acids should be considered part of the polyQ regions. Here, we have used five different thresholds to look for short and long, and pure and impure, polyQ in human proteins. The use of these thresholds results in very different numbers of detected polyQ (Table 1) that obviously

Table 1
Number of polyQ tracts and polyQ-containing proteins found with different thresholds in the human proteome.

Threshold	PolyQ tracts	PolyQ-containing proteins
4/4	1458	1062
4/6	5200	3315
6/6	488	378
6/8	958	696
8/10	400	309

affect their length distributions (Fig. 2a) and fraction of impurities (Fig. 2b), calculated as the sum of all impurities divided by the sum of all polyQ lengths. The majority of polyQ regions are impure when the threshold used allows impurities (Fig. 2c). Not surprisingly, the fraction of such impurities decreases for the more restrictive thresholds (4/6 > 6/8 > 8/10) (Fig. 2d).

In order to study the nature of these impurities, the abundance of each amino acid within the polyQ region was analyzed. Consistently in the three thresholds that allow impurities, leucine and proline were the most prevalent amino acids (Fig. 2e) [24]. We found prolines preferentially towards the C-terminus within longer polyQ (Fig. 2f). This is consistent with prolines being usually found as part of polyP or Pro-rich regions positioned at the C-termini of polyQ [4,20]. Conversely, we find leucine residues towards the N-terminus of longer polyQ [24]. Both amino acids have a relative position of ~0.6 in threshold 4/6, and diverge to a ~0.3 and ~0.7 for leucine and proline, respectively, when using the 8/10 threshold. This suggests that in long polyQ tracts the selective pressure acts not only over the nature of the impurities but also over their position. These data reveal that the distributions of length and number of impurities in polyQ are closely linked to the threshold used to detect them. It is difficult to separate both concepts, as one leads inevitably to the other. For example, a restrictive threshold (8/10) allowed us to find biases in the position of the polyQ impurities that were not detected when using more relaxed cut-offs, such as 4/6 and 6/8.

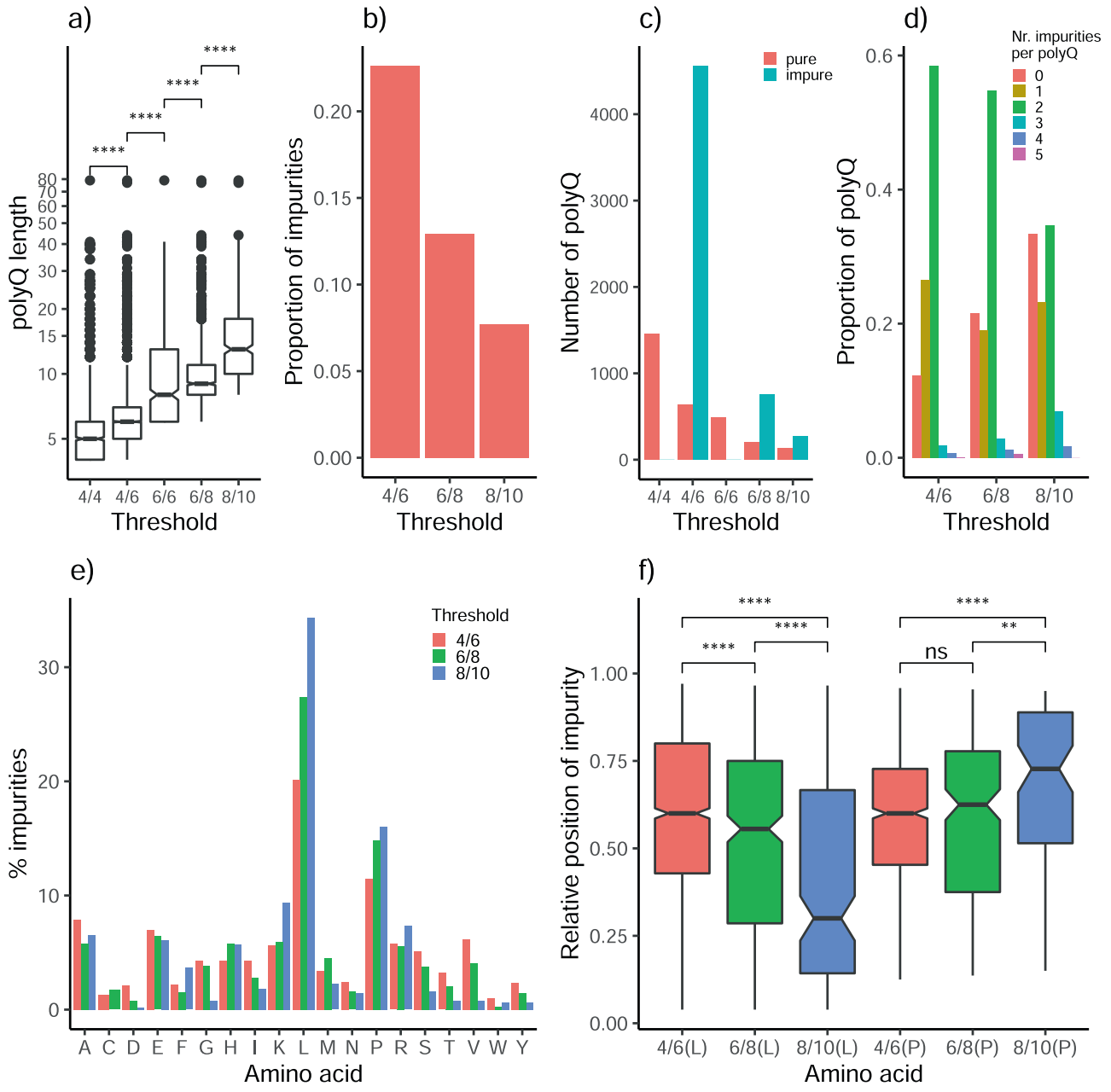

Fig. 2. Threshold-dependent polyQ characterization. a) Distribution of polyQ lengths. b) Proportion of impurities in polyQ regions. c) Number of pure and impure polyQ per threshold. d) Distribution of number of impurities per polyQ region. e) Percentage of impurities in polyQ regions per type of amino acid. f) Relative position of impurities (leucine, L and proline, P) within polyQ regions. Non-parametric Mann-Whitney U statistical tests were performed to compare the distributions (**** P-value $\leq$ 1e−4; ** P-value $\leq$ 0.01; ns = not significant).

3.2. The amino acid context of polyQ regions is length- and purity-dependent

Since polyQ length and the presence of impurities have been proven to be threshold-dependent, we wondered whether the sequence context of the polyQ is also dependent on the threshold used to define the tracts. To study this feature we calculated the amino acid composition per position of polyQ flanking regions for each of the five thresholds. The most prevalent amino acid around the polyQ regions for all thresholds is glutamine (data not shown), suggesting that there is a soft transition between polyQ and Q-rich regions surrounding polyQ.

Two other amino acids are significantly enriched in the proximity of the human polyQ: leucine and proline. This is consistent with the most abundant amino acids found as impurities within the polyQ described in the previous section. Likewise, the presence of leucine and proline around polyQ has different position dependencies. While the percentage of leucine is increased in position −1 for all thresholds and remains above the background abundance up to position −4 (Fig. 3a), the higher abundance of proline residues extends in the opposite direction, from position +1 to +5, and it is especially significant in longer polyQ (Fig. 3b). This suggests different biological functions for leucine and proline. Indeed, experimental data obtained for the human Androgen Receptor shows that leucines upstream of a polyQ tract makes it α-helical and therefore less aggregation prone [25]. A subsequent study showed that this effect is caused by the hydrophobic leucine side chain that protects a hydrogen bond involving the glutamine side chain and the carbonyl of the preceding leucine [26].

PolyP tracts have also been proposed as a mechanism to attenuate the aggregation propensity of long polyQ regions. However, aggregation is a complex phenomenon that depends on large variety of parameters, such as the presence of chaperones [27,28] and the nature of the flanking regions [29,30]. In huntingtin, the causative agent of HD, it has been shown that the proline-rich region following the polyQ exerts a protective role *in vitro* and *in vivo* most probably due to the conformational restrictions that prolines impose to neighbouring glutamines [31]. Notably, this protective effect is directional, as N-terminal polyP does not attenuate the aggregation of polyQ peptides [32,33].

3.3. The nucleotide landscape of polyQ regions

Since the amino acid context, length and purity of the polyQ tract depend on the threshold used, we examined whether the nucleotide sequences coding for polyQ tracts and their flanking regions are also threshold-dependent. This is key to understand the presence of CAG trinucleotide expansions in evolution and disease [20]. Therefore, we focused on the nucleotide context, the usage of polyCA[G/A] within polyQ and the characteristics of the impurities in all protein coding transcripts of the human genome.

The same five thresholds used in previous sections were used to detect polyCA[G/A]. Once a polyCA[G/A] was detected, five codons N- and C-terminal to the polyQ were retrieved. We calculated the codon abundance per position −5 to +5 and normalized it with the codon abundance in the complete dataset of human transcripts.

Regarding the codons composing the polyCA[G/A] itself, they are highly enriched in CAG over CAA (Fig. 4a) regardless of the threshold used, as previously described [34]. Interestingly, the prevalence of impurities (codons that are neither CAG nor CAA, see "Other" in Fig. 4a) in threshold 4/6 is higher than the value for CAA codon. As discussed above, most of the impurities are proline and leucine (Fig. 2e). As to the proline codons found as impurities, CCG is positively enriched in the three thresholds in a length-dependent manner (Fig. 4b). This is not the case for leucine codons, with no length-dependent prevalence for any codon (Fig. 4c).

Since leucine and proline are also enriched around polyQ regions (Fig. 3), we next analyze in detail the positional enrichment of each codon for proline and leucine respect to background in the positions surrounding polyQ (Fig. 4d). In the case of proline, CCG is up to 8-fold the background in positions +1 to +5. Although to a lesser extent, this CCG enrichment occurs at position −1. At first this result seems striking, given that CCG is the least prevalent of the codons for proline in humans (compare background levels in Fig. 4b). The fact that CCG is one point mutation away from CAG may be the reason for this bias, suggesting that these CCG encoded prolines might have evolved from CAG-rich glutamine tracts; we do not expect this to be a general phenomenon. As an example, CTG (leucine) is not enriched in the vicinity of the polyQ [24].

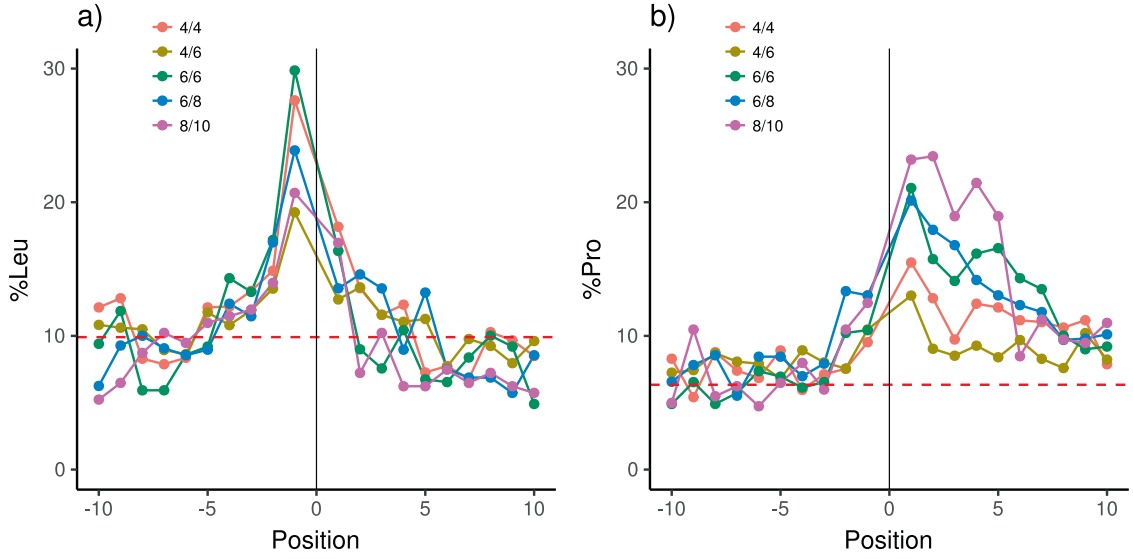

Fig. 3. Amino acid context of polyQ regions. a) Leucine and b) proline abundance around human polyQ regions. Dashed red lines refer to the background composition of the amino acid. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

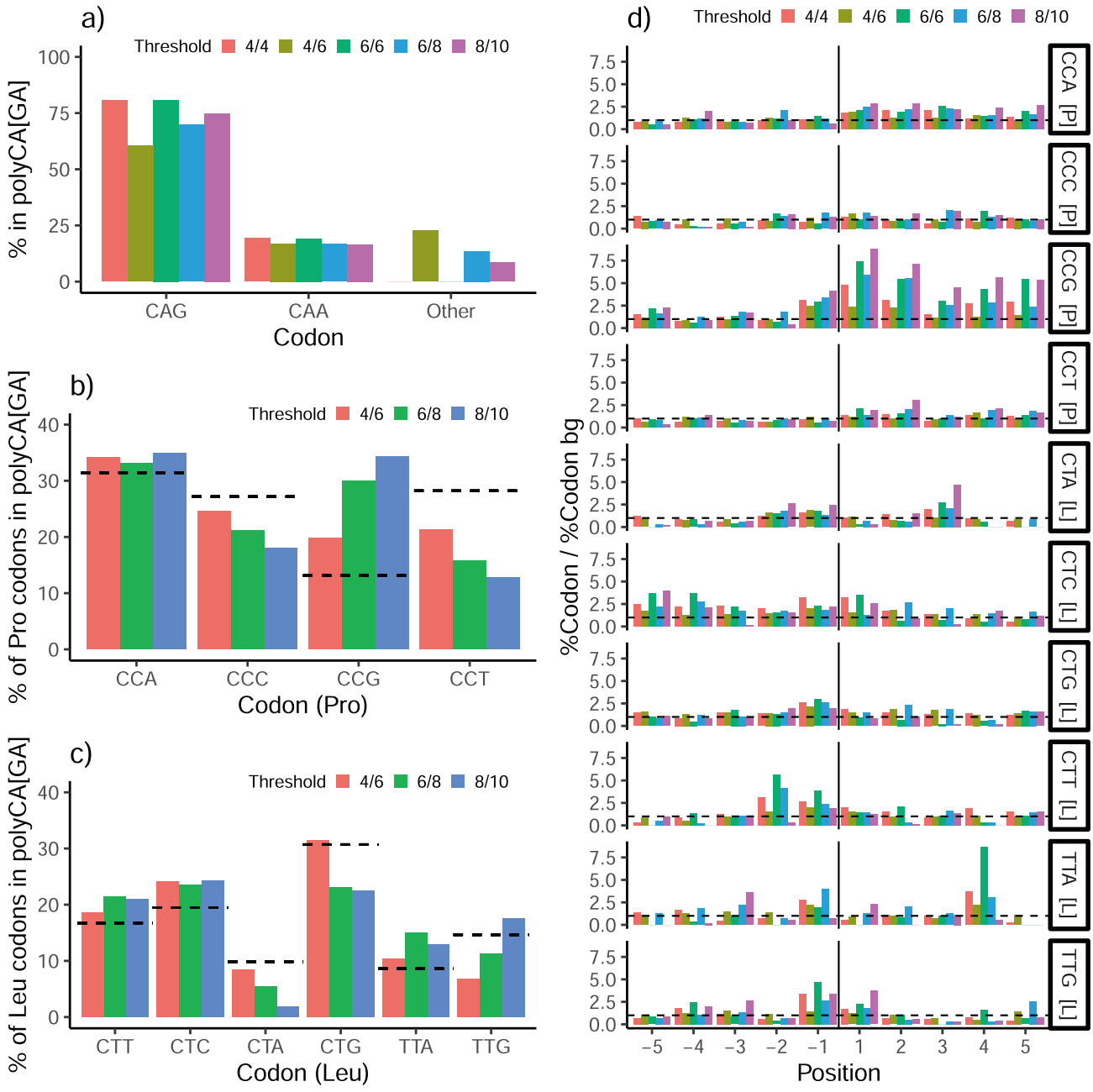

Fig. 4. Nucleotide context and codon usage of polyCA[G/A] in the human genome. a) Codon usage in polyCA[G/A]; "Other" refers to codons that are neither CAG nor CAA found in the polyQ (impurities). Percentage of b) proline and c) leucine codons in polyCA[G/A] regions. Dashed lines refer to the background composition of the respective amino acid. d) Normalized codon usage in positions −5 to + 5 with respect to human polyQ regions, for leucine and proline codons; the horizontal dashed line indicates no enrichment (%Codon/%Codon bg = 1), and the vertical line the position of the polyCA[G/A].

3.4. PolyQ features are species-dependent

Up to this point we have studied the threshold-dependencies of human polyQ features. Factors such as proteome size or genomic biases may modify the features related to polyQ in other species, and some of the thresholds may not be as useful as in human to extract conclusions. To illustrate this caveat, we located polyQ regions in seven additional phylogenetically diverse species, and we studied their features.

The number of polyQ (pure or impure) in *M. musculus*, *D. rerio* and *C. elegans* is similar to the polyQ abundance in human (Fig. 5a). In other species, this distribution is perturbed due to a higher prevalence of short pure polyQ (*D. melanogaster* and *D. discoideum*) or to a lower number of long polyQ (*S. cerevisiae* and *A. thaliana*). These results indicate that there is a large diversity in the number of polyQ tracts depending on the organism studied: the thresholds applied may have a different effect in different organisms. The study of the polyQ length distribution indicates differences between species and reflects the importance of threshold choice (Figs. 2a and 5b). For example, polyQ selected with threshold 4/6 have the highest average length in *D. discoideum* but not for the more restrictive 8/10. *C. elegans* seems to have shorter polyQ than most other species studied at all thresholds.

The amino acid context of polyQ regions is also species-dependent. For example, with the exception of *C. elegans*, leucine enrichment at position −1 of polyQ is present in all of the studied

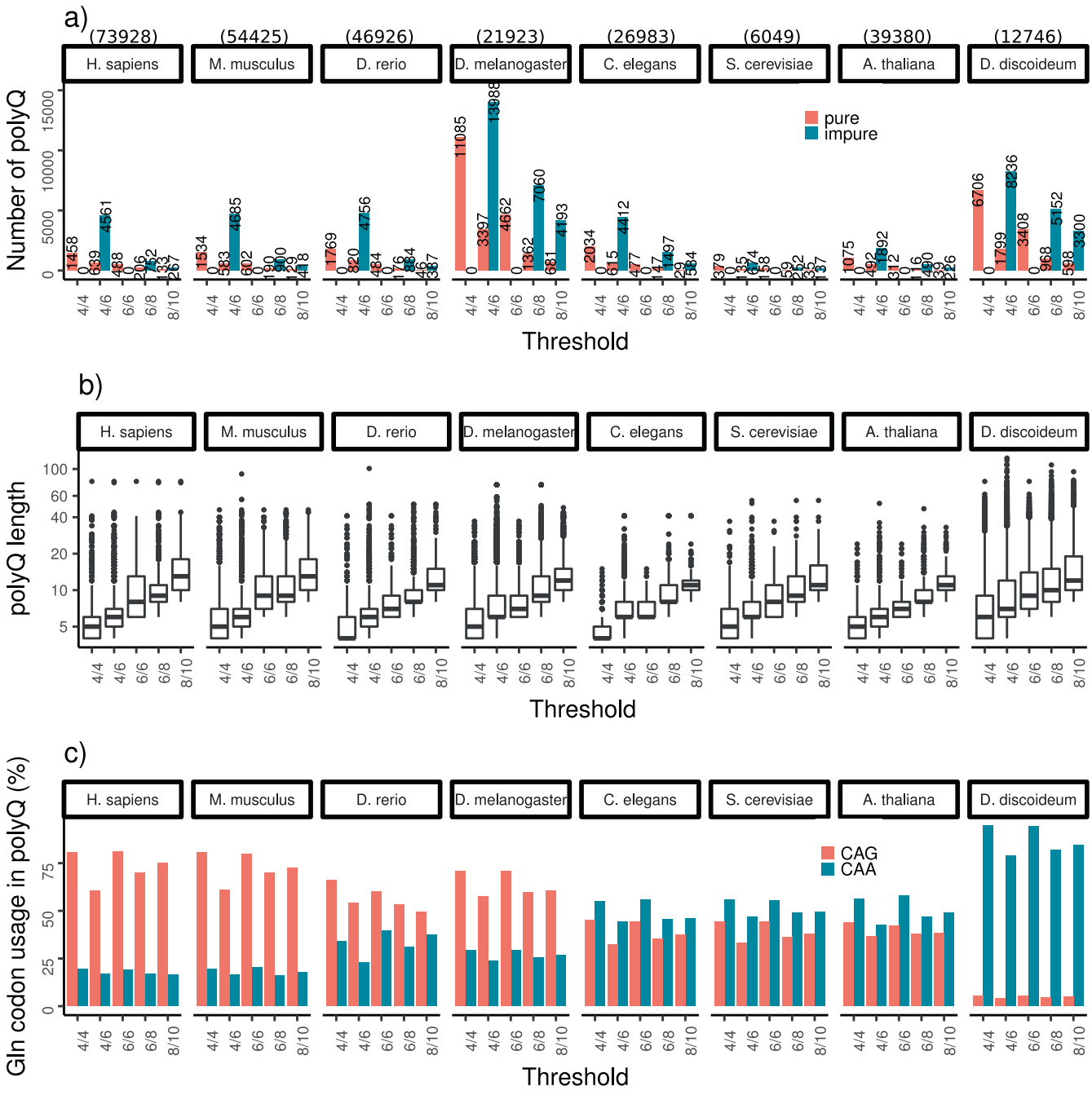

Fig. 5. PolyQ characterization in several eukaryotic species. a) Number of pure and impure polyQ, b) polyQ length distribution per threshold, and c) glutamine codon usage in polyQ, in the complete reference proteome of *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Dictyostelium discoideum*. The numbers in brackets refer to the number of proteins in the respective proteome.

proteomes (Supplementary Fig. 1). The signal is stronger in vertebrates, and in longer polyQ in *A. thaliana*. Interestingly, in *D. melanogaster* the signal for position −1 is similar to that of position +1. Regarding the proline enrichment in positions +1 to +5 of polyQ, none of the proteomes shows a signal as clear as in human (Supplementary Fig. 2). A few of them are enriched only in position +1 (*C. elegans*, *S. cerevisiae* and *D. discoideum*), while others are not significantly enriched (*D. rerio*, *D. melanogaster* and *A. thaliana*). *M. musculus* is the only one with an extended enrichment up to position +5, suggesting that this feature may have appeared late along evolution.

With respect to the glutamine codon usage in the polyQ regions of those species, results show three clearly differentiated groups (Fig. 5c). First, vertebrates and *D. melanogaster* are enriched in CAG, independently of the threshold used. Second, slightly higher values of CAA than CAG are found in *C. elegans*, *S. cerevisiae* and *A. thaliana*. Finally, polyQ regions in *D. discoideum* are up to a 94% composed of CAA codons. Although to a much lower extent, these groups and trends are maintained when studying the glutamine codon usage in the complete proteome (Supplementary Fig. 3). In the first and the latter group the CAG/CAA ratio in polyQ is much extreme than in the whole genome, but is genome-

dependent regarding the direction of said enrichment: CAG/CAA ratio in *H. sapiens*, *M. musculus*, *D. rerio* is around 3:2 in the genome and 3:1 in polyQ, and in *D. discoideum* is 1:4 in the genome and 1:9 in polyQ.

These results point to possible functional and structural properties of polyQ that are species-specific. Therefore, future findings relating to polyQ found in one species should be cautiously translated to other species, carefully considering the threshold used for their detection.

4. Conclusions

PolyQ regions consist of consecutive repetitions of glutamine residues in a protein sequence. At the nucleotide level, these repetitions are composed of a mix of codons, CAG and CAA, which in human are in at least a 3:1 ratio [34]. To detect a polyQ region in a sequence, one must define a threshold regarding the minimum length of the homorepeat and its purity. This threshold should be selected specifically for each study, depending on the properties under investigation. Length-, purity- and species-dependent behaviours may be faded or lost otherwise (see Fig. 2, Fig. 3, Supplementary Fig. 1 and Supplementary Fig. 2).

As a rule of thumb, if there are not many cases of polyQ in a given proteome, one might aim at capturing the majority of them. The threshold 4/6 could be used in that case, since this threshold is less restrictive and captures polyQ in a significantly different structural environment [35]. If the focus is on retrieving functional polyQ with high precision, then we recommend using the thresholds 6/8 or 8/10. However, this rule might not apply to some species as we have seen that the properties of the polyQ are species-dependent even when using the same threshold.

While we acknowledge that it does not make sense to carry out every study on homorepeats with all the thresholds shown here, or even with others, we recommend a fast scanning similar to the one in Fig. 5. Depending on the results one could make an informed decision and thus choose a threshold depending both on the needs of the study and on the dataset; i.e. if the dataset does not contain many long polyQ, choosing 8/10 will in principle be useless. Furthermore, if there were not many polyQ, whichever the threshold, it would be preferable to use 4/6 to capture as many regions as possible. To carry out the proposed fast scanning, we have prepared a simple script called sQannner (Supplementary File 1) that takes a proteome (a protein dataset in FASTA format) and generates a plot with the number of pure and impure polyQ, and the polyQ length distribution per threshold. The whole process takes around 10 s in a standard laptop computer for the proteome of *S. cerevisiae* (6049 proteins), which is provided with the script as an example, and 2 min for the human proteome (74449 proteins). In addition, sQanner is available as a simple web tool in http://cbdm-01.zdv.uni-mainz.de/~munoz/sQanner/.

It is important to know the biases associated with each threshold (Table 2). The dataset of polyQ regions resulting from the selected threshold will have inherent properties, such as containing long polyQ if 8/10 is used or containing only pure polyQ if 4/4 is used. Although Table 2 can be applied to all proteomes, the way the thresholds shape their length and purity distributions, as well as their amino acid and nucleotide sequence contexts are species-dependent. For example, at position −1 of human polyQ regions there is a clear enrichment in leucine abundance for all thresholds, especially in pure polyQ; however, in *A. thaliana* this enrichment is only detected when using threshold 8/10 (Supplementary Fig. 1).

From this work it becomes clear that studying an apparently simple entity such as polyQ regions is more complicated than it seems. The choice of a suitable threshold to locate them may change the results of the study simply because polyQ features are length- and purity-dependent. We have shown here that polyQ have a biased amino acid and nucleotide context. Although we have centered our study in humans, the exploratory analysis of other model species demonstrates that polyQ has experienced a taxa-specific evolutionary pressure. This suggests that these homorepeats have distinct functional roles in different species. However, more work is still necessary to have a comprehensive characterization of polyQ regions in all sequenced species. Our contributions in this respect are practical methodological and computational guidelines to evaluate the polyQ landscape in any species. Importantly, these guidelines may be also applied to other homorepeats with relevant roles in biology.

Table 2
Overview of thresholds used to look for polyQ regions, and associated biases.

Threshold	What is found	Biases
4/4	All pure polyQ	• More short than long polyQ (see Fig. 1a); very short polyQ may not be functional. • No impure polyQ are captured.
4/6	All polyQ regions	• Many possible impurities (up to 2/6 = 33%, mean = 22.6%, see Fig. 1b); some regions matching the threshold may not function as polyQ, and thus they could be false positives.
6/6	Pure functional polyQ	• Short pure polyQ are missed. • PolyQ with extremely low number of impurities are considered separately (i.e. in Ataxin-1 [UniProt:Q14119] positions 197–225, sequence QQQQQQQQQQQQHQHQQQQQQQQQQQQQQ would be taken as two different polyQ, separated by the Histidine residues).
6/8	Most polyQ	• Short polyQ are missed.
8/10	Long polyQ	• Short polyQ are missed. • Low amount of impurities allowed (up to 2/10 = 20%, mean = 7.7%, see Fig. 1b).

Acknowledgments

Not applicable.

Funding

This work was supported by Deutsche Forschungsgemeinschaft [AN735/4-1 to M.A.A.N.] and by the European Research Councilunder the European Union's H2020 Framework Programme (2014–2020)/ERC Grant agreement nr. [648030], and Labex EpiGenMed, an "Investissements d'avenir" program (ANR-10-LABX-12-01) awarded to PB. The CBS is a member of France-BioImaging (FBI) and the French Infrastructure for Integrated Structural Biology (FRISBI), two national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively).

Competing interests

Declarations of interest: none.

Authors' contributions

PM conceived and carried out the project. PB and MAAN supervised the project. PM drafted the manuscript. All authors contributed in writing the final manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.01.012.

References

[1] Albà MM, Guigó R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res 2004;14(4):549–54. https://doi.org/10.1101/gr.1925704.

[2] Chavali S, Chavali PL, Chalancon G, de Groot NS, Gemayel R, et al. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. Nat Struct Mol Biol 2017;24(9):765–77. https://doi.org/10.1038/nsmb.3441.

[3] Galzitskaya OV, Lobanov MY. Proteome-scale understanding of relationship between homo-repeat enrichments and protein aggregation properties. PLoS ONE 2018;13(11):. https://doi.org/10.1371/journal.pone.0206941e0206941.

[4] Mier P, Alanis-Lobato G, Andrade-Navarro MA. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins 2017;85(4):709–19. https://doi.org/10.1002/prot.25250.

[5] Aravind L, Iyer LM, Wellems TE, Miller LH. Plasmodium biology: genomic gleanings. Cell 2003;115(7):771–85. https://doi.org/10.1016/s0092-8674(03)01023-7.

[6] Davies HM, Nofal SD, McLaughlin EJ, Osborne AR. Repetitive sequences in malaria parasite proteins. FEMS Microbiol Rev 2017;41(6):923–40. https://doi.org/10.1093/femsre/fux046.

[7] Kuspa A, Loomis WF. The genome of dictyostelium discoideum. Methods Mol Biol 2006;346:15–30. https://doi.org/10.1385/1-59745-144-4:15.

[8] Darling AL, Uversky VN. Intrinsic disorder in proteins with pathogenic repeat expansions. Molecules 2017;22(12). https://doi.org/10.3390/molecules22122027.

[9] Walker FO. Huntington's disease. Lancet 2007;369(9557):218–28. https://doi.org/10.1016/S0140-6736(07)60111-1.

[10] Shao J, Diamond M. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum Mol Genet 2007;16(2):R115–23. https://doi.org/10.1093/hmg/ddm213.

[11] Williams AJ, Paulson HL. Polyglutamine neurodegeneration: protein misfolding revisited. Trends Neurosci 2008;31(10):521–8. https://doi.org/10.1016/j.tins.2008.07.004.

[12] Amiel J, Trochet D, Clément-Ziza M, Munnich A, Lyonnet S. Polyalanine expansions in human. Hum Mol Genet 2004;13(2):R235–43. https://doi.org/10.1093/hmg/ddh251.

[13] Hughes JN, Thomas PQ. Molecular pathology of polyalanine expansion disorders: new perspectives from mouse models. Methods Mol Biol 2013;1017:135–51. https://doi.org/10.1007/978-1-62703-438-8_10.

[14] Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 2001;18(7):1161–7. https://doi.org/10.1093/oxfordjournals.molbev.a003903.

[15] Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A 2002;99(1):333–8. https://doi.org/10.1073/pnas.012608599.

[16] Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltesz B, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform 2019. https://doi.org/10.1093/bib/bbz007.

[17] Adegbuyiro A, Sedighi F, Pilkington 4th AW, Groover S, Legleiter J. Proteins containing expanded polyglutamine tracts and neurodegenerative disease. Biochemistry 2017;56(9):1199–217. https://doi.org/10.1021/acs.biochem.6b00936.

[18] Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci 2007;30:575–621. https://doi.org/10.1146/annurev.neuro.29.051605.113042.

[19] Blum ES, Schwendeman AR, Shaham S. PolyQ disease: misfiring of a developmental cell death program?. Trends Cell Biol 2013;23(4):168–74. https://doi.org/10.1016/j.tcb.2012.11.003.

[20] Schaefer MH, Wanker EE, Andrade-Navarro MA. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res 2012;40(10):4273–87. https://doi.org/10.1093/nar/gks011.

[21] Lobanov MY, Galzitskaya OV. Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol Biosyst 2012;8(1):327–37. https://doi.org/10.1039/c1mb05318c.

[22] Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 2005;15(4):537–51. https://doi.org/10.1101/gr.3096505.

[23] Jorda J, Kajava AV. Protein homorepeats sequences, structures, evolution, and functions. Adv Protein Chem Struct Biol 2010;79:59–88. https://doi.org/10.1016/S1876-1623(10)79002-7.

[24] Ramazzotti M, Monsellier E, Kamoun C, Degl'Innocenti D, Melki R. Polyglutamine repeats are associated to specific sequence biases that are conserved among eukaryotes. PloS One 2012;7(2):e30824. https://doi.org/10.1371/journal.pone.0030824.

[25] Eftekharzadeh B, Piai A, Chiesa G, Mungianu D, García J. Sequence context influences the structure and aggregation behavior of a PolyQ tract. Biophys J 2016;110(11):2361–6. https://doi.org/10.1016/j.bpj.2016.04.022.

[26] Escobedo A, Topal B, Kunze MBA, Aranda J, Chiesa G, et al. Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. Nat Commun 2019;10(1):2034. https://doi.org/10.1038/s41467-019-09923-2.

[27] Tam S, Spiess C, Auyeung W, Joachimiak L, Chen B, et al. The chaperonin TriC blocks a huntingtin sequence element that promotes the conformational switch to aggregation. Nat Struct Mol Biol 2009;16(12):1279–85. https://doi.org/10.1038/nsmb.1700.

[28] Kakkar V, Mansson C, de Mattos EP, Bergink S, van der Zwaag M, et al. The S/T-rich motif in the DNAJB6 chaperone delays polyglutamine aggregation and the onset of disease in a mouse model. Mol Cell 2016;62(2):272–83. https://doi.org/10.1016/j.molcel.2016.03.017.

[29] Lin HK, Boatz JC, Krabbendam IE, Kodali R, Hou Z, et al. Fibril polymorphism affects immobilized non-amyloid flanking domains of huntingtin exon1 rather than its polyglutamine core. Nat Commun 2017;8:15462. https://doi.org/10.1038/ncomms15462.

[30] Thakur AK, Jayaraman M, Mishra R, Thakur M, Chellgren VM, et al. Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism. Nat Struct Mol Biol 2009;16(4):380–9. https://doi.org/10.1038/nsmb.1570.

[31] Shen K, Calamini B, Fauerbach JA, Ma B, Shahmoradian SH, et al. Control of the structural landscape and neuronal proteotoxicity of mutant huntingtin by domains flanking the polyQ tract. Elife 2016;5. https://doi.org/10.7554/eLife.18065.

[32] Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, et al. Oligoproline effects on polyglutamine conformation and aggregation. J Mol Biol 2006;355(3):524–35. https://doi.org/10.1016/j.jmb.2005.10.053.

[33] Darnell G, Orgel JP, Pahl R, Meredith SC. Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing PPII-like helix structure. J Mol Biol 2007;374(3):688–704. https://doi.org/10.1016/j.jmb.2007.09.023.

[34] Mier P, Andrade-Navarro MA. Glutamine codon usage and polyQ evolution in primates depend on the Q stretch length. Genome Biol Evol 2018;10(3):816–25. https://doi.org/10.1093/gbe/evy046.

[35] Totzeck F, Andrade-Navarro MA, Mier P. The protein structure context of polyQ regions. PLoS One 2017;12(1):. https://doi.org/10.1371/journal.pone.0170801e0170801.