Polyglutamine and Polyalanine Tracts Are Enriched in Transcription Factors of Plants

Nina Kottenhagen[1*], Lydia Gramzow[1*], Fabian Horn[2], Martin Pohl[3], and Günter Theißen[1]

1 **Friedrich-Schiller-Universität Jena, Lehrstuhl für Genetik, Philosophenweg 12, D–07743 Jena, {nina.kottenhagen,lydia.gramzow,guenter.theissen}@uni-jena.de**
2 **Leibniz-Institut für Naturstoffforschung und Infektionsbiologie e.V. - Hans-Knöll-Institut, Systembiologie/ Bioinformatik, Beutenbergstr. 8, D–07745 Jena, fabian.horn@hki-jena.de**
3 **Friedrich-Schiller-Universität Jena, Lehrstuhl für Bioinformatik, Ernst-Abbe-Platz 2, D–07743 Jena, m.pohl@uni-jena.de**

Abstract

Polyglutamine (polyQ) tracts have been studied extensively for their roles in a number of human diseases such as Huntington's or different Ataxias. However, it has also been recognized that polyQ tracts are abundant and may have important functional and evolutionary roles. Especially the association of polyQ and also polyalanine (polyA) tracts with transcription factors and their activation activity has been noted. While a number of examples for this association have been found for proteins from opisthokonts (animals and fungi), only a few studies exist for polyQ and polyA stretches in plants, and systematic investigations of the significance of these repeats in plant transcription factors are scarce. Here, we analyze the abundance and length of polyQ and polyA stretches in the conceptual proteomes of six plant species and examine the connection between polyQ and polyA tracts and transcription factors of the repeat-containing proteins. We show that there is an association of polyQ stretches with transcription factors in plants. In grasses, transcription factors are also significantly enriched in polyA stretches. While there is variation in the abundance, length, and association with certain functions of polyQ and polyA stretches between different species, no general differences in the evolution of these repeats could be observed between plants and opisthokonts.

1998 ACM Subject Classification J.3 Life and Medical Sciences

Keywords and phrases tandem repeats, molecular evolution, GO annotation

Digital Object Identifier 10.4230/OASIcs.GCB.2012.93

1 Introduction

In general, amino acid repeats (AARs) in a protein can have very different consequences, ranging from causing severe diseases over neutral polymorphisms being suitable as genetic markers to "tuning knobs" of evolution [27, 28]. Polyglutamine (polyQ) tracts have found special interest because they are found in various severe human neurodegenerative or neuromuscular hereditary diseases [55]. In the case of Huntington's disease, aberrantly extended polyQ tracts in the HUNTINGTIN protein cause abnormal folding, subsequent protein aggregation and neuronal loss (reviewed in [41]). The repeat length and the severity of the disease are positively correlated: The longer the repeat, the earlier the age of onset and also

* These authors contributed equally to this work.

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 93–107
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the more severe the symptoms. However, apart from their roles in diseases, the functional and evolutionary importance of polyQ stretches has also already been recognized, especially when they occur in transcription factors (TFs) [4, 17, 18, 19, 45].

Concerning the functional significance of polyQ stretches in TFs, their role in breeding and evolution has found special interest. Fondon and Garner [17], for example, studied protein repeats in RUNT-RELATED TRANSCRIPTION FACTOR 2 (RUNX-2), which is a main regulator of osteoblast differentiation, in various dog breeds. The authors found different lengths of adjacent polyalanine (polyA) and polyQ tracts in RUNX-2 in the different dog breeds, where the length ratio of these sequences correlates with the dorsoventral nose bend and midface length across breeds. As the repeat lengths evolve fast, Fondon and Garner concluded that repeats contributed to an acceleration of the morphological evolution of limbs and skulls in different dog breeds [17].

Other striking examples, where polyA or polyQ stretches in TFs may lead to phenotypic variability, were found in a broad range of species. In insects, the acquisition of a polyA transcription repression domain in the HOX protein ULTRABITHORAX may have contributed to the suppression of abdominal limbs during arthropod evolution and hence to the macroevolution of a body plan [45, 19]. A polymorphism at a polyA stretch in the protein Hoxd-13 may contribute to variation in sesamoid bone formation in amniotes [4]. The longer the polyQ tract in White Collar-1 (WC-1), a protein which influences the circadian clock of the fungus *Neurospora crassa*, the shorter are the circadian periods [18].

One important molecular mechanism by which differences in repeat lengths lead to phenotypic variation is likely due to the capacity of polyA and polyQ tracts to modulate transcription factor activity. Specifically, polyA tracts are thought to decrease and polyQ stretches are thought to increase transcriptional activation [20]. Hence, changes in the lengths of polyA and polyQ stretches of TFs may alter the transcription rate and implicate changes in the expression of a set of target genes. This could generate variation upon which selection acts and thus contribute to morphological changes [17].

Based on this exemplary evidence for the importance of polyA and polyQ tracts in TFs, it was hypothesized that, in general, polyA and polyQ stretches predominantly occur in TFs [20]. Subsequently, several systematic studies have been conducted as well. For a number of proteomes, the entirety of polyA and polyQ tracts was analyzed and the association of the repeat-containing proteins with functions in transcriptional activation was studied [29]. In yeast, polyQ stretches belong to the most abundant amino acid repeats found beside repeats of asparagine (N), aspartic acid (D), glutamic acid (E), and serine (S) [2]. Furthermore, repeats of acidic and polar amino acids, to which Q belongs, were found to be significantly associated with TFs and protein kinases. Similar results were obtained for other species. In rodents, humans, fruit flies, and nematodes, a functional bias of proteins with repeats, including A and Q repeats, was observed, where TFs were consistently overrepresented [1, 2, 23, 51].

In plants, however, systematic studies on polyA and polyQ tracts and their association with certain functions of the repeat-containing proteins are scarce [49, 59]. Plants, including Glaucophytes, Red algae, Green algae and Embryophytes (land plants), are one of the major eukaryotic groups whereas animals and fungi, for which association studies between amino-acid repeats and the function of the containing proteins are common, belong to another major group, the opisthokonts [58]. The most recent common ancestor of opisthokonts and plants was recently estimated to have lived at least one billion years ago [12]. Considering this long period of independent evolution between plants and opisthokonts, it is questionable whether polyA and polyQ tracts show the same abundance and variability and the same

functional bias of the repeat-containing proteins in these two major eukaryotic groups.
Amino-acid repeats are thought to expand or contract mainly by replication slippage or recombination of the corresponding protein-coding DNA [32, 43]. These mechanisms result in a rapid evolution of repeats. However, selection may act on repeat length as has been shown for opisthokonts [36]. For polyA and polyQ stretches in TFs, a correlation between repeat length and transcriptional activation has been shown [20] and hence the existence of repeats may be favored in some cases. On the other hand, long polyA and polyQ tracts may have negative effects, such as aggregation of the containing proteins, and thus there may be selection against polyA and polyQ stretches that are too long [10, 35]. All these factors, i.e., the mutational mechanism leading to the expansion and contraction of repeats, the capacity of polyQ and polyA to modulate transcription, and the negative effects of long repeats, may be different in plants. Hence, whether polyA and polyQ tracts are as important for phenotypic variation and morphological evolution in plants as in opisthokonts remains to be clarified.
Here, we analyze A and Q repeats in six proteomes of diverse species which span the phylogeny of land plants. We obtain the total number of polyA and polyQ tracts and their lengths in these proteomes and study the association between the containing proteins and a function in transcriptional regulation. We compare our findings to those found for opisthokonts and hypothesize on similarities and differences of the importance of polyA and polyQ stretches in opisthokonts and plants.

2 Materials and methods

To study the evolution of polyQ and polyA stretches in plants, six species spanning the phylogeny of land plants were selected with respect to their taxonomic placement and the availability of their proteome and corresponding GO annotations. These species are the moss *Physcomitrella patens*, the lycophyte *Selaginella moellendorffii*, the eudikotyledonous angiosperms *Arabidopsis thaliana* and *Populus trichocarpa* and the monocotyledonous angiosperms *Sorghum bicolor* and *Oryza sativa*. As an animal reference set, *Homo sapiens*, *Danio rerio*, and *Anopheles gambiae* were selected. We decided not to take our data from databases like COPASAAR [13], GENPEPT [14], TRIPS [29], ProtRepeatsDB [25], or ProRepeat [34], as they proved not to be customizable for our research. Therefore, conceptual proteome datasets for the selected species were obtained from the Ensembl, EnsemblMetazoa and EnsemblPlants project [16] (versions: May, 22nd 2012, `http://www.ensembl.org/index.html`, `http://metazoa.ensembl.org/index.html`, `http://plants.ensembl.org/index.html`). The datasets chosen contained all *ab initio* predicted protein sequences as well as manually curated sequences. Ensembl distinguishes between „known“, „novel“, and „putative“ proteins. For our analyses, all three classes were used. We selected only the first splice variant to avoid a bias due to overrepresentation of proteins originating from genes with multiple splice variants. The number of protein sequences included in our analyses for each species is given in Table 1.
Here, we define an AAR as a stretch of at least five identical amino acids in a row which is significant according to Karlin *et al.* (2002) [26]. The number and length of polyQ and polyA stretches within each species was determined using a custom Perl-script (all scripts are available upon request). The length of an amino acid repeat is defined as the number of residues forming the repeat. For comparison, the number and length of polyasparagine (polyN) stretches were also determined.
Gene Ontology (GO) annotation files [5] were downloaded with the help of Ensembl BioMart

[30] (version: May, 22nd 2012, v0.7, `http://www.biomart.org/biomart/martview/`). To check whether polyQ, polyA or polyN stretches show a specific enrichment pattern, a GO enrichment analysis was performed using a custom R-script based on the topGO package available from the Bioconductor website [3]. An exact Fisher test [15] was used to investigate whether certain GO categories are enriched in sequences containing a polyA, polyQ, or polyN stretch. The obtained p-values were corrected for multiple testing using the Benjamini-Hochberg method [8]. We focused on the GO category GO:0003700 (sequence-specific DNA binding transcription factor activity) as it has been shown for fungi and animal species that polyQ as well as polyA stretches are associated with this GO category [1, 2, 22, 51].
The significant results were assigned a rank according to their p-value. The lower the p-value was, the lower was the assigned rank. We use the p-value as a sign of the strength of the association. To compare the magnitude of the association between the studied repeats and the GO category GO:0003700, the average ranking of this GO category was compared between plants and animals. If the category was not identified as an enriched GO category, a penalty score was assigned to be able to calculate the average ranking. This penalty score was chosen to be the number of the highest rank of the category GO:0003700 across all studied plant and animal species plus one.

Table 1 Overview on source data. Species are in taxonomic order, animals are highlighted in gray.

Species	Number of protein sequences	Number of first splice variants	Number of annotated first splice variants	Reference
Arabidopsis thaliana	35386	27416	27155	[24]
Populus trichocarpa	45778	41377	26091	[56]
Oryza sativa	68619	57995	16994	[37]
Sorghum bicolor	36338	34496	21744	[38]
Selaginella moellendorffii	34825	34799	22650	[7]
Physcomitrella patens	38354	32273	8931	[42]
Homo sapiens	100354	22400	22400	[33]
Danio rerio	42171	26212	26212	[54]
Anopheles gambiae	14324	12670	8873	[21]

3 Results

3.1 Number and length of polyQ, polyA and polyN stretches

We studied amino acid repeats which had a length of at least five residues. On average, polyQ stretches were shorter in plants than in animals (Table 2). Also the maximum number of residues in a polyQ stretch was lower in plants than in animals. In *P. trichocarpa*, the longest polyQ stretch contained 33 residues, in *A. gambiae*, there were an exceptional 132 residues in the longest repeat. The total number of polyQ stretches was on average lower in plant proteins than in animal proteins, which is due to the exceptionally high number in *A. gambiae*.

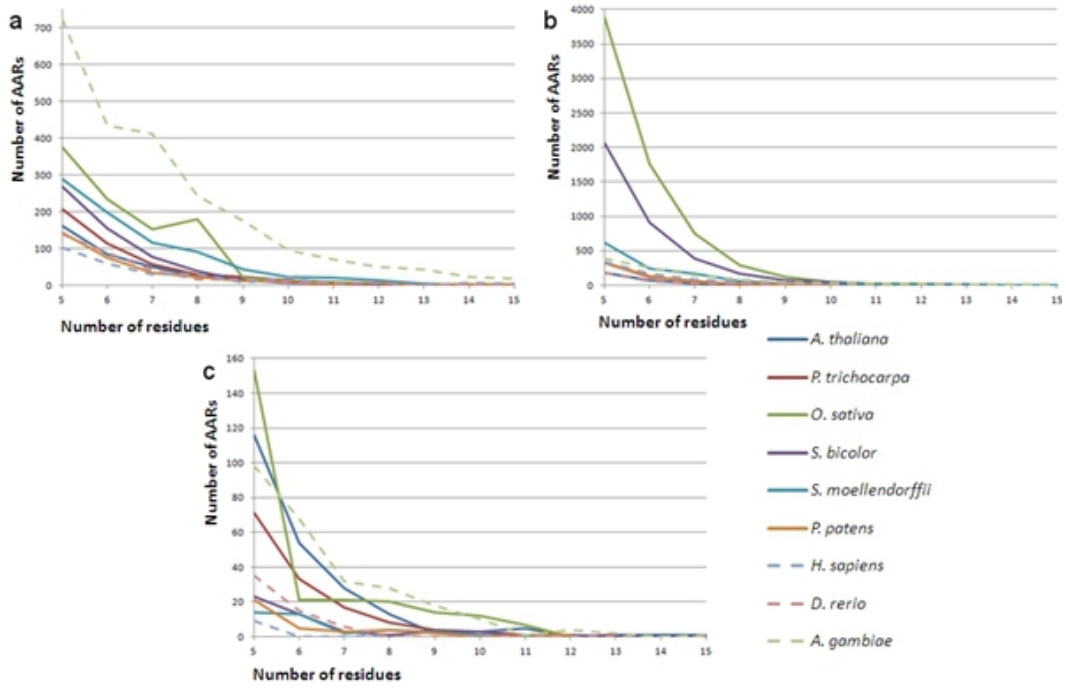

Figure 1 Length distribution of polyQ (a), polyA (b) and polyN (c) stretches in the conceptual proteomes of plant and animal species. Plant species are indicated by solid lines, broken lines represent animal species.

Similar to polyQ stretches, the average polyA stretch in plant proteins was a bit shorter than in animal proteins (Table 2). Also the maximum number of residues in a polyA stretch was lower in plants than in animals. Both numbers were lower as compared to polyQ stretches. The total number of polyA stretches varied widely within both kingdoms with plants harboring a higher number of them. PolyA stretches were also more commonly found than polyQ stretches in all species except *A. gambiae*.

PolyN was used as a control for polyQ because both amino acids are encoded by two codons and are chemically quite similar; their side chains differ from one another only by one methyl group. Unlike polyQ and polyA stretches, a higher average length was found in plant proteins than in animal proteins for polyN stretches (Table 2). The maximum number of residues in a polyN stretch was also higher in proteins of plants than of animals as well as the total number of polyN stretches. Across all species, fewer polyN stretches were found than polyQ and polyA stretches. However, the trends for polyN stretches were less clear than for polyQ and polyA stretches in both, plants and animals.

The length distributions for polyQ, polyA and polyN stretches followed the same trend across all species: While there are many short repeats the number of long repeats is low (Figure 1). *A. gambiae* harbored more polyQ stretches of every length than any other species. Remarkebly, however, also *O. sativa* showed an increased number of polyQ stretches of eight residues. PolyA stretches were most abundant in the grasses *O. sativa* and *S. bicolor*. For polyN stretches, no clear pattern could be found. In *O. sativa*, a plateau for polyN stretches of six to eleven residues was observed.

Table 2 Number, average repeat length and maximum length of polyQ, polyA, and polyN stretches in the conceptual proteomes of plant and animal species. Species are in taxonomic order, animals are highlighted in gray.

Amino acid	Species	Number of stretches	Average length of stretches	Longest stretch
Q	*A. thaliana*	376	6.61	24
	P. trichocarpa	462	6.44	33
	O. sativa	1005	6.49	27
	S. bicolor	571	6.08	24
	S. moellendorffii	811	6.74	25
	P. patens	322	6.63	22
	H. sapiens	307	8.79	40
	D. rerio	334	6.79	41
	A. gambiae	2373	7.68	132
A	*A. thaliana*	296	5.65	17
	P. trichocarpa	545	5.7	17
	O. sativa	6926	5.77	19
	S. bicolor	3705	5.83	18
	S. moellendorffii	1193	6.09	19
	P. patens	540	5.67	10
	H. sapiens	800	6.86	21
	D. rerio	335	6.12	18
	A. gambiae	1068	6.92	23
N	*A. thaliana*	226	6.13	20
	P. trichocarpa	138	5.99	13
	O. sativa	249	6.18	19
	S. bicolor	53	10.02	61
	S. moellendorffii	38	6.39	12
	P. patens	37	6.19	12
	H. sapiens	10	5.3	8
	D. rerio	57	5.58	11
	A. gambiae	263	6.56	16

3.2 Abundance of amino acid repeats within sequences annotated as transcription factors

The overall percentage of proteins related to transcription factor activity (GO:0003700) was quite low in plants as well as in animals, except in *A. thaliana* for which 10% of its proteins were annotated as TFs (Table 3). The percentage of polyQ-containing sequences in plants and animals was also quite low. Only in *A. gambiae* an exceptional 9% of its proteins contained polyQ stretches. The percentage of proteins annotated as TFs and containing polyQ stretches ranged from 4% to 10% in plants and animals. However, an exception was again *A. gambiae* with 26% of its putative TFs containing polyQ stretches.

Very few proteins harbored polyA stretches, not even *A. gambiae* proteins. In contrast, of the proteins annotated as TFs between 2% and 35% contained polyA stretches. They were most commonly found in proteins of *O. sativa*, *S. bicolor*, *H. sapiens* and *A. gambiae*.

Except for in *A. gambiae* with 2%, polyN stretches were hardly found. Proteins annotated as TFs contained to a slightly higher percentage polyN stretches (~1%). *A. thaliana* and *A. gambiae* constituted the exceptions with 4% and 13%, respectively, of their proteins annotated as TFs and containing polyN stretches. Hence, the control amino acid asparagine

Table 3 Percentages of sequences annotated as transcription factors (TFs) and sequences containing polyQ, polyA or polyN stretches. Species are ordered according to taxonomy, animals are highlighted in gray.

Species	% of proteins annotated as TFs	% of proteins containing polyQ stretches	% of TFs containing polyQ stretches	% of proteins containing polyA stretches	% of TFs containing polyA stretches	% of proteins containing polyN stretches	% of TFs containing polyN stretches
A. thaliana	10.06	1.02	3.77	0.21	2.07	0.80	3.59
P. trichocarpa	2.10	0.89	4.26	0.06	2.65	0.32	1.27
O. sativa	0.88	1.49	7.48	0.31	35.24	0.43	0.39
S. bicolor	1.95	1.41	8.2	0.46	23.85	0.15	0.45
S. moellendorffii	0.97	0.75	4.79	0.07	7.35	0.11	<0.01
P. patens	1.15	1.81	9.52	0.10	8.52	0.10	<0.01
H. sapiens	4.14	1.00	4.20	0.59	14.22	0.04	<0.01
D. rerio	2.88	1.01	3.58	0.13	4.64	0.22	0.80
A. gambiae	2.15	8.72	26.47	0.49	22.79	1.74	12.87

was less prevalent in both, proteins annotated as TFs and in stretches, than glutamine and alanine.

Thus, even though only a low percentage of proteins contained polyQ or polyA stretches and only a low percentage of proteins were annotated as TFs, a high percentage of proteins annotated as TFs contained polyQ or polyA stretches in both plants and animals (Table 3).

3.3 PolyQ and polyA stretches significantly associated with transcription factors

To find out whether TFs are significantly overrepresented in plant proteins containing polyQ and polyA stretches, we carried out GO enrichment analyses. The significant results were assigned a rank according to their p-value. Thereby, the rank was lower when the p-value was lower.

In all plant species examined, protein sequences containing polyQ tracts were found to be significantly associated with the GO category GO:0003700 "sequence-specific DNA binding transcription factor activity" (Table 4 in appendix). The mean rank value equaled 4.13.

In animals, proteins annotated as having transcription factor activity were also significantly enriched in polyQ stretches. However, other transcription-related categories were found at even lower p-values except in *A. gambiae* where categories related to different binding activities were found at the lowest ranks. In comparison to plants, the category GO:0003700 was found at lower ranks in animals. However, other more specific GO categories related to transcription regulation like GO:0044212 "transcription regulatory region DNA binding" exhibited lower p-values in animals.

Protein sequences containing polyA stretches also showed a significant association with the GO category GO:0003700 in all species except in *P. trichocarpa*. Instead, the association of proteins annotated with GO categories related to catalytic activity was found on the first ranks in this species, while in the other plants and in animals proteins belonging to various

categories connected to binding activity were found on the first ranks. In humans, protein sequences annotated with more specific GO categories such as "transcription regulatory region DNA transcription activity" appeared at ranks shortly after the more general category GO:0003700. Thus, except for *P. trichocarpa*, plant and animal proteins belonging to GO categories related to transcription factor activity were enriched in polyA stretches.
Unlike the two other types of amino acid repeats, polyN-containing proteins were hardly associated with GO categories related to transcription factor activity. Only in *A. thaliana* and *P. trichocarpa* such an association was found. In the other species, either no significant association with any category was found or with proteins belonging to GO categories related to different kinds of binding and catalytic activities.

4 Discussion

PolyQ and polyA tracts in the conceptual proteomes of different plant and animal species were investigated and their association with TFs was analyzed to determine whether there are differences in the evolution of these repeats between plants and animals. As a control, we also studied the occurrence of polyN stretches.

4.1 No major difference in the abundance and length of polyQ and polyA stretches between plants and animals

On average, the length distributions of polyQ, polyA and polyN stretches varied between the different species but no major difference between plants and animals could be observed (Figure 1). The abundance of repeats generally decreased with increasing repeat lengths for all species. The length distributions were different for the different amino acids. PolyQ and polyA stretches were found far more often, with longer repeat lengths and a higher average repeat length than polyN stretches. These findings correspond well with the findings of Faux *et al.* [14]. who found more polyQ and polyA stretches than polyN stretches in proteins of *H. sapiens*, *D. rerio*, *A. thaliana*, and *O. sativa*. Siwach *et al.* [51] also found length distributions of many shorter and few longer amino acid repeats. Of the proteins they analyzed, 84% contained repeats consisting of 10 residues or less. At least in part, this may simply be due to the fact that the DNA sequences encoding long repeats have a higher likelihood of being split into shorter repeats by non-synonymous point mutations. As we have only analyzed pure repeats, those long impure repeats escaped our statistics. On the other hand, long polyQ and polyA stretches are probably also selected against, possibly because the containing proteins have a tendency of aggregation [10, 35].
Even though Q is encoded by only two codons and found less often than A, an amino acid encoded by four codons [31, 47], it occurs in quite long repeats and a high number of repeats in both, plants and animals. N, also encoded by two codons and approximately as frequent as Q in the proteomes of *A. thaliana* and *O. sativa* [31, 47], forms fewer and shorter repeats than the other two types of amino acids in both kingdoms. Furthermore, polyQ stretches were found to a higher percentage in proteins than polyA and polyN stretches. This indicates that there are differences in the rate at which the different repeats are elongated and contracted and/ or in the selection forces acting on the different repeats. There are two main mechanisms which are thought to contribute to the expansion of repeats, replication slippage and unequal recombination [50]. While replication slippage is supposed to be the major mechanism for the extension of polyQ stretches, polyA stretches are thought to be mainly extended by unequal recombination, at least in animals [9]. Selection has also been shown to contribute to the generation and extension of amino acid repeats in animals [36, 52] and in simulation studies

[46]. Analyses of the codons contributing to the different repeats will provide insights into the relative contributions of mutation and selection to the observed frequencies of polyQ, polyA and polyN stretches. Replication slippage and unequal recombination generally lead to amino acid repeats which are encoded by the same codon [57]. In contrast, if selection has contributed to the conservation of a polyQ or polyA stretch, one would expect that the amino acid repeat may be encoded by different synonymous codons rather than by the same codon [46].

A clear trend was that polyQ and polyA stretches are longer on average in animals than in plants (Table 2). Also the maximum number of residues in a polyQ stretch is markedly higher in animals than in plants. In contrast, a higher abundance of polyQ, polyA and polyN stretches is observed in plants (excluding the outlier *A. gambiae*). These differences may be caused by differences in the underlying mutation rates and/ or repair mechanisms or by differences in the selection patterns between plants and animals. Selection against long polyQ and polyA stretches may be stronger while a higher number of amino acid repeats is tolerated by or selected for in plants. It is possible that, instead of having long repeats, plant proteins may have a number of consecutive shorter repeats to fulfill the same function (if any). However, we did not study the distribution of the repeats within the proteins and additional studies are required to test this hypothesis.

Further studies, taking into account the underlying codons of the polyQ and polyA stretches and the distribution of these repeats within the proteins will permit a more thorough characterization of the relative contributions of the mutational mechanisms and selection regimes to the generation and maintenance of repeats. Moreover, studies including, for example, additional plant, animal, fungal and possibly also bacterial genomes will allow more general conclusions as to which of the observed abundances and length distributions of polyQ and polyA stretches are characteristic for certain species and which patterns are observed in a broader taxonomic range.

In other words, some differences in the number and length distributions of polyQ, polyA and polyN stretches were found but they currently cannot be extrapolated to major differences distinguishing plants and animals.

4.2 PolyQ and polyA stretches are enriched in transcription factors also in plants

The genomes of all analyzed species were published several years ago [7, 21, 24, 33, 38, 40, 42, 54, 56] and have been studied for some time now [37, 48, 53] resulting in a good quality of the sequence assembly and annotation. Many TFs belong to transcription factor families which each are defined by highly conserved DNA-binding domains (DBD). These DBD are used to predict TFs in newly sequenced species. Thus, many common TFs are reasonably predictable and are likely annotated even in newly released proteomes. In animals, more proteins annotated as TFs (GO:0003700) were found than in plants. This may be ascribed to the fact that the animal proteins are completely (*H. sapiens* and *D. rerio*) or to a high percentage annotated [16]. Annotation of proteins of *A. thaliana* is also nearly complete (99%). *A. thaliana* has previously been shown to have a higher amount of TFs than investigated animal and fungal species [44]. When the annotations of the other species become more elaborated, the number of TFs may change. Most annotations have been assigned automatically and may improve with manual curation. However, the quality of automatic annotations has been shown to be quite high for *H. sapiens* and several other model organisms [52]. Hence, despite possible differences in the progress of the assembly, the amount of TFs and the completeness of protein annotation in the different species, we do not assume that our findings are considerably biased.

PolyQ as well as polyA stretches are found more often in TFs than one would expect from their individual occurrence rates in both, plants and animals. An exceptionally high percentage of TFs of *O. sativa* and *S. bicolor* contain polyA stretches. The high GC content of grass genomes [11] and the fact that A is encoded by GC-rich codons (GCN) may contribute to this phenomenon.
Our GO enrichment analyses confirm that polyQ and polyA tracts are associated with proteins annotated as TFs (GO category GO:0003700). Hence, our findings corroborate that the correlation between polyQ and polyA stretches and TFs found in animals and fungi [20] also holds true for plants. The ranks of this category are a bit lower in plants than in animals. However, this probably does not indicate a major difference between plants and animals because the other associations at low ranks in animals are also categories involved in transcription activity. Thus, differences are rather species-specific than kingdom-specific.
Unlike polyQ and polyA, polyN, used as a control here, was not found to be associated with TFs in several plant and animal species. Thus, the overrepresentation of polyQ and polyA hints at a function of these stretches in TFs. Selection may often favor polyQ and polyA stretches, at least up to a certain length, whereas polyN stretches may be neutral or even deleterious except for *A. thaliana* and *P. trichocarpa*. In these two species, polyN-rich regions, just like polyQ-rich regions, may have a role in mediating protein-protein interactions [39]. For vertebrates, selection increasing the retention of amino-acid repeats including polyQ and polyA has been found recently [36]. It has also been shown before that polyQ tracts enhance transcriptional activation in animals in a length dependent manner [20]. This result was recently extended to fungi [6], indicating a role for polyQ tracts in the modulation of transcription in opisthokonts. Our findings now make it appear likely that polyQ stretches also have such a function in plants. PolyA stretches have been hypothesized to repress transcriptional activity in animals [17]. The overrepresentation of polyA stretches also in plant TFs again hints at a conservation of this function between plants and opisthokonts. Further support for the role of these repeats in transactivation could be gained from an analysis of the position of the repeats within the corresponding proteins. The repeats would be expected to occur outside of the DNA-binding domain where transcriptional activity can mainly be modulated. In our analyses, we observed some species-specific differences in the abundance and length distributions of polyQ and polyA stretches. However, there seem to be no major differences in the evolution of polyQ and polyA stretches between plants and opisthokonts. Hence, the mutational mechanisms for the generation, expansion and contraction as well as the selection pressure on polyQ and polyA stretches seem to be similar in respective species, or differences in mutation and selection compensate each other. Furthermore, the association between TFs and polyQ and polyA stretches was also found for the plant species examined. These stretches may have similar roles in plants and opisthokonts. Hence, we provide data suggesting that polyQ and polyA tracts act as "evolutionary tuning knobs" [28, 27] not only for opisthokonts but also for land plants.

Acknowledgements We would like to thank all members of the Theißen lab for their kind support during the preparation of this publication.

References

1 M. M. Albà and R. Guigo. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.*, 14(4):549–54, 2004.

2 M. M. Albà, M. F. Santibáñez-Koref, and J. M. Hancock. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. *J Mol Evol*, 49(6):789–797, 1999.

3 A. Alexa and J. Rahnenfuhrer. topGO: Enrichment analysis for Gene Ontology. *R package version 2.4.0*, 2010.

4 K. Anan, N. Yoshida, Y. Kataoka, M. Sato, H. Ichise, M. Nasu, and S. Ueda. Morphological change caused by loss of the taxon-specific polyalanine tract in Hoxd-13. *Mol Biol Evol*, 24(1):281–287, 2007.

5 M. Ashburner et al. Gene Ontology: tool for the unification of biology. *Nat Genet.*, 25(1):25–29, 2000.

6 L. Atanesyan, V. Gunther, B. Dichtl, O. Georgiev, and W. Schaffner. Polyglutamine tracts as modulators of transcriptional activation from yeast to mammals. *Biol Chem.*, 393(1-2):63–70, 2012.

7 J. A. Banks et al. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332(6032):960–3, 2011.

8 Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

9 L. Y. Brown and S. A. Brown. Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet.*, 20(1):51–8, 2004.

10 S. Caburet, A. Demarez, L. Moumné, M. Fellous, E. De Baere, and R. A. Veitia. A recurrent polyalanine expansion in the transcription factor FOXL2 induces extensive nuclear and cytoplasmic protein aggregation. *J Med Genet*, 41(12):932–936, 2004.

11 N. Carels and G. Bernardi. The compositional organization and the expression of the *Arabidopsis* genome. *FEBS Lett.*, 472(2-3):302–6, 2000.

12 D. Chernikova, S. Motamedi, M. Csürös, E. V. Koonin, and I. B. Rogozin. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol Direct*, 6:26, 2011.

13 D. P. Depledge and A. R. Dalby. COPASAAR–a database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics*, 6:196, 2005.

14 N. G. Faux, S. P. Bottomley, A. M. Lesk, J. A. Irving, J. R. Morrison, M. G. de la Banda, and J. C. Whisstock. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res*, 15(4):537–551, 2005.

15 R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society.*, 85(1):87–94, 1922.

16 P. Flicek. Ensembl 2012. *Nucleic Acids Res.*, 40(D1):D84–D90, 2012.

17 J. W. Fondon and H. R. Garner. Molecular origins of rapid and continuous morphological evolution. *Proc Natl Acad Sci U S A*, 101(52):18058–18063, 2004.

18 A. C. Froehlich, Y. Liu, J. J. Loros, and J. C. Dunlap. White Collar-1, a circadian blue light photoreceptor, binding to the frequency promoter. *Science*, 297(5582):815–819, 2002.

19 R. Galant and S. B. Carroll. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature*, 415(6874):910–913, 2002.

20 H. P. Gerber, K. Seipel, O. Georgiev, M. Höfferer, M. Hug, S. Rusconi, and W. Schaffner. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science*, 263(5148):808–811, 1994.

21 R. A. Holt et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591):129–49, 2002.

22 M. Huntley and G. B. Golding. Evolution of simple sequence in proteins. *J Mol Evol*, 51(2):131–140, 2000.

23 M. A. Huntley and A. G. Clark. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol*, 24(12):2598–2609, 2007.

24 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815, 2000.

25 M. K. Kalita, G. Ramasamy, S. Duraisamy, V. S. Chauhan, and D. Gupta. Protrepeatsdb: a database of amino acid repeats in genomes. *BMC Bioinformatics*, 7:336, 2006.

26 S. Karlin, L. Brocchieri, A. Bergman, J. Mrazek, and A. J. Gentles. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A*, 99(1):333–8, 2002.

27 Y. Kashi, D. King, and M. Soller. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet*, 13(2):74–78, 1997.

28 Yechezkel Kashi and David G King. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*, 22(5):253–259, 2006.

29 M. V. Katti, R. Sami-Subbu, P. K. Ranjekar, and V. S. Gupta. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci*, 9(6):1203–1209, 2000.

30 R.J. Kinsella. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011, 2011.

31 N. Kottenhagen. *Frequency and distribution of amino acid repeats in the proteomes of Arabidopsis thaliana and Oryza sativa.* Diploma thesis, Friedrich-Schiller-Universität Jena, 2009.

32 S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A*, 95(18):10774–10778, 1998.

33 E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

34 H. Luo, K. Lin, A. David, H. Nijveen, and J. A. Leunissen. ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Res.*, 40(Database issue):D394–9, 2012.

35 J. F. Morley, H. R. Brignull, J. J. Weyers, and R. I. Morimoto. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*, 99(16):10417–10422, 2002.

36 L. Mularoni, A. Ledda, M. Toll-Riera, and M. M. Albà. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res.*, 20(6):745–54, 2010.

37 S. Ouyang et al. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, 35(Database issue):D883–7, 2007.

38 A. H. Paterson et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(7229):551–6, 2009.

39 M. F. Perutz, B. J. Pope, D. Owen, E. E. Wanker, and E. Scherzinger. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. *Proc Natl Acad Sci U S A.*, 99(8):5596–600, 2002.

40 International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, 2005.

41 A. Reiner, I. Dragatsis, and P. Dietrich. Genetics and neuropathology of Huntington's disease. *Int Rev Neurobiol*, 98:325–372, 2011.

42 S. A. Rensing et al. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859):64–9, 2008.

43 G. F. Richard and F. Paques. Mini- and microsatellite expansions: the recombination connection. *EMBO Rep.*, 1(2):122–6, 2000.

44 J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman, and G. Yu. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–10, 2000.

45 M. Ronshaugen, N. McGinnis, and W. McGinnis. Hox protein mutation and macroevolution of the insect body plan. *Nature*, 415(6874):914–917, 2002.

46 M. M. Rorick and G. P. Wagner. The origin of conserved protein domains and amino acid repeats via adaptive competition for control over amino acid residues. *J Mol Evol.*, 70(1):29–43, 2010.

47 P. Seeber. *Frequency and distribution of amino acid repeats in transcription factors.* Diploma thesis, Friedrich-Schiller-Universität Jena, 2010.

48 M. V. Sharakhova, M. P. Hammond, N. F. Lobo, J. Krzywinski, M. F. Unger, M. E. Hillenmeyer, R. V. Bruggner, E. Birney, and F. H. Collins. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol.*, 8(1):R5, 2007.

49 N. Sharopova. Plant simple sequence repeats: distribution, variation, and effects on gene expression. *Genome*, 51(2):79–90, 2008.

50 R. R. Sinden, V. N. Potman, E. A. Oussatcheva, C. E . Pearson, Y. L. Lyubchenko, and L. S. Shlyakhtenko. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J Biosci.*, 27(1):53–65, 2002.

51 P. Siwach, S. D. Pophaly, and S. Ganesh. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. *Mol Biol Evol*, 23(7):1357–1369, 2006.

52 N. Skunca, A. Altenhoff, and C. Dessimoz. Quality of computationally inferred gene ontology annotations. *PLoS Comput Biol.*, 8(5), May 2012.

53 D. Swarbreck et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, 36(Database issue):D1009–14, 2008.

54 The *Danio rerio* Sequencing Project. (http://www.sanger.ac.uk/projects/d_rerio/).

55 Y. Trottier, Y. Lutz, G. Stevanin, G. Imbert, D. Devys, G. Cancel, F. Saudou, C. Weber, G. David, and L. Tora. Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. *Nature*, 378(6555):403–406, 1995.

56 G. A. Tuskan et al. The genome of black cottonwood, *Populus trichocarpa* (Tor. & Gray). *Science*, 313(5793):1596–604, 2006.

57 E. Viguera, D. Canceill, and S. D. Ehrlich. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.*, 20(10):2587–95, 2001.

58 Hwan Su Yoon, Jessica Grant, Yonas Tekle, Min Wu, Benjamin Chaon, Jeffrey Cole, John Logsdon, David Patterson, Debashish Bhattacharya, and Laura Katz. Broadly sampled multigene trees of eukaryotes. *BMC Evolutionary Biology*, 8(1):14, 2008.

59 Y. Zhou, J. Liu, L. Han, Z. G. Li, and Z. Zhang. Comprehensive analysis of tandem amino acid repeats from ten angiosperm genomes. *BMC Genomics.*, 12:632, 2011.

A Appendix

Table 4 GO enrichment analysis. Shown are the three significant results with the lowest p-values of the analysis for polyA, polyN, and polyQ tracts for all species. If the GO category related to transcription factor activity (GO:0003700) did not belong to the first three entries, this category with its corresponding rank is shown in addition, if found significant. Species are in taxonomic order, animals are highlighted in gray. T – transcription, A – activity, TFA – transcription factor activity, if no significant results were found, this is indicated by "–".

Repeat	Species	Rank	GO-ID	GO category	P-value
polyQ	*A. thaliana*	1	GO:0001071	nucleic acid binding TFA	1E-28
		2	GO:0003700	sequence-specific DNA binding TFA	1E-28
		3	GO:0003677	DNA binding	1.8E-25
	P. trichocarpa	1	GO:0003676	nucleic acid binding	1E-28
		2	GO:0003677	DNA binding	1E-28
		3	GO:0001071	nucleic acid binding TFA	7.8E-11
		4	GO:0003700	sequence-specific DNA binding TFA	7.8E-11
	O. sativa	1	GO:0003677	DNA binding	2.5E-28
		2	GO:0003676	nucleic acid binding	2.5E-24
		3	GO:0001071	nucleic acid binding TFA	8.4E-12
		4	GO:0003700	sequence-specific DNA binding TFA	8.4E-12
	S. bicolor	1	GO:0003677	DNA binding	1E-28
		2	GO:0003676	nucleic acid binding	1E-28
		3	GO:0001071	nucleic acid binding TFA	2.3E-21
		4	GO:0003700	sequence-specific DNA binding TFA	2.3E-21
	S. moellendorffii	1	GO:0003677	DNA binding	1E-28
		2	GO:0003676	nucleic acid binding	1.3E-26
		3	GO:0005515	protein binding	3.7E-20
		5	GO:0003700	sequence-specific DNA binding TFA	3.4E-16
	P. patens	1	GO:0003676	nucleic acid binding	1.7E-14
		2	GO:0046983	protein dimerization A	2.6E-10
		3	GO:0003677	DNA binding	8.7E-09
		6	GO:0003700	sequence-specific DNA binding TFA	6.1E-06
	H. sapiens	1	GO:0003676	nucleic acid binding	1.2E-14
		2	GO:0003677	DNA binding	1.3E-14
		3	GO:0044212	T regulatory region DNA binding	6.7E-12
		9	GO:0003700	sequence-specific DNA binding TFA	2.1E-11
	D. rerio	1	GO:0003676	nucleic acid binding	3,00E-19
		2	GO:0003677	DNA binding	6.3E-17
		3	GO:0003712	transcription cofactor A	3.4E-07
		6	GO:0003700	sequence-specific DNA binding TFA	1.3E-06
	A. gambiae	1	GO:0008270	zinc ion binding	7.1E-26
		2	GO:0005488	binding	7.6E-24
		3	GO:0005515	protein binding	1.1E-23
		8	GO:0003700	sequence-specific DNA binding TFA	8.1E-16
polyA	*A. thaliana*	1	GO:0004124	cysteine synthase A	3.6E-05
		2	GO:0001071	nucleic acid binding TFA	6.7E-05
		3	GO:0003700	sequence-specific DNA binding TFA	6.7E-05
	P. trichocarpa	1	GO:0004124	cysteine synthase A	9.00E-09
		2	GO:0004970	ionotropic glutamate receptor A	1.3E-06
		3	GO:0005231	excitatory extracellular ligand-gated ion channel A	1.3E-06
	O. sativa	1	GO:0050824	water binding	1E-28
		2	GO:0050825	ice binding	1E-28
		3	GO:0001071	nucleic acid binding TFA	1E-28
		4	GO:0003700	sequence-specific DNA binding TFA	1E-28
	S. bicolor	1	GO:0001071	nucleic acid binding TFA	7.1E-21
		2	GO:0003700	sequence-specific DNA binding TFA	7.1E-21
		3	GO:0003677	DNA binding	2.9E-19
	S. moellendorffii	1	GO:0003676	nucleic acid binding	7.5E-13
		2	GO:0003677	DNA binding	3.2E-09
		3	GO:0005488	binding	1.6E-07
		6	GO:0003700	sequence-specific DNA binding TFA	3.2E-06

continued on next page

	P. patens	1	GO:0001071	nucleic acid binding TFA	5.2 E-06
		2	GO:0003700	sequence-specific DNA binding TFA	5.2 E-06
		3	GO:0003844	1,4-alpha-glucan branching enzyme A	7.8E-04
	H. sapiens	1	GO:0003676	nucleic acid binding	1E-28
		2	GO:0043565	sequence-specific DNA binding	1E-28
		3	GO:0003677	DNA binding	1E-28
		5	GO:0003700	sequence-specific DNA binding TFA	1E-28
	D. rerio	1	GO:0003676	nucleic acid binding	1.5E-23
		2	GO:0003677	DNA binding	6.1E-09
		3	GO:0003700	sequence-specific DNA binding TFA	1.8E-08
	A. gambiae	1	GO:0003676	nucleic acid binding	1E-28
		2	GO:0003677	DNA binding	3.4E-26
		3	GO:0005488	binding	6.5E-26
		7	GO:0003700	sequence-specific DNA binding TFA	4.9E-16
polyN	*A. thaliana*	1	GO:0001071	nucleic acid binding TFA	1E-28
		2	GO:0003700	sequence-specific DNA binding TFA	1E-28
		3	GO:0003677	DNA binding	4.1E-19
	P. trichocarpa	1	GO:0003676	nucleic acid binding	1.7E-07
		2	GO:0043565	sequence-specific DNA binding	1.4E-07
		3	GO:0003677	DNA binding	5.6E-04
		7	GO:0003700	sequence-specific DNA binding TFA	1.0E-02
	O. sativa	—	—	—	—
	S. bicolor	—	—	—	—
	S. moellendorffii	1	GO:0015018	galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase A	5.7E-04
		2	GO:0015020	glucuronosyltransferase A	1.7E-03
	P. patens	—	—	—	—
	H. sapiens	—	—	—	—
	D. rerio	1	GO:0003676	nucleic acid binding	1.6E-02
	A. gambiae	1	GO:0003677	DNA binding	1.6E-18
		2	GO:0001071	nucleic acid binding TFA	5,00E-17