Author/s:
Polling, S;Ormsby, AR;Wood, RJ;Lee, K;Shoubridge, C;Hughes, JN;Thomas, PQ;Griffin, MDW;Hill, AF;Bowden, Q;Böcking, T;Hatters, DM

Title:
Polyalanine expansions drive a shift into α-helical clusters without amyloid-fibril formation

Date:
2015-12-01

Citation:
Polling, S., Ormsby, A. R., Wood, R. J., Lee, K., Shoubridge, C., Hughes, J. N., Thomas, P. Q., Griffin, M. D. W., Hill, A. F., Bowden, Q., Böcking, T. & Hatters, D. M. (2015). Polyalanine expansions drive a shift into α-helical clusters without amyloid-fibril formation. Nature Structural and Molecular Biology, 22 (12), pp.1008-1015. https://doi.org/10.1038/nsmb.3127.

Polyalanine expansions drive a shift into α-helical clusters without amyloid fibril formation.

Saskia Polling[1,2], Angelique R. Ormsby[1,2], Rebecca J. Wood[1,2], Kristie Lee[3], Cheryl Shoubridge[3], James N. Hughes[4], Paul Q. Thomas[4], Michael D.W. Griffin[1,2], Andrew F. Hill[5], Quill Bowden[6], Till Böcking[6,7] and Danny M. Hatters[1,2,7]

[1]Department of Biochemistry and Molecular Biology, The University of Melbourne, VIC Australia

[2]Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, VIC Australia

[3]Robinson Research Institute, School of Paediatrics and Reproductive Health, The University of Adelaide, SA. Australia,

[4]School of Molecular and Biomedical Science, The University of Adelaide, SA. Australia

[5]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, VIC, Australia

[6]Single Molecule Science and Department of Anatomy, School of Medical Sciences, The University of New South Wales, NSW. Australia

[7]Australian Research Council Centre of Excellence in Advanced Molecular Imaging. Australia

Corresponding author: D.M.H. (dhatters@unimelb.edu.au) Ph: +61 38344 2530.

ABSTRACT

Polyglutamine (polyGln) expansions in nine human proteins, which result in neurological diseases, induce their tendency to form β-rich amyloid fibrils and intracellular deposits. Less well known are at least nine other human diseases caused by polyalanine (polyAla) expansion mutations in different proteins. The mechanisms for how polyAla aggregates under physiological conditions remain unclear and controversial. We show here that expanded polyAla aggregation is mechanistically dissimilar to polyGln and hence amyloid kinetics. PolyAla expansion leads to a spontaneous assembly into α-helical clusters of diverse oligomeric state. Such clustering was pervasive in cells irrespective of visible aggregate formation and it disrupted the normal physiological oligomeric state of two human proteins natively containing polyAla, ARX and SOX3. This self-assembly pattern indicates polyAla expansions chronically disrupt protein behaviour by the imposition of a deranged oligomeric status.

The human genome contains repetitive nucleotide sequences that can expand in sequence length and cause disease. Some are in non-coding sequences (*e.g.* C9ORF72 in amyotrophic lateral sclerosis [1], FMR1 in Fragile X syndrome [2] and DMPK in muscular dystrophy type 1 (ref. 3)) leading to abnormal mRNA sequences that can form RNA aggregates or abnormal non-ATG initiated protein translation products [4]. Others are found in coding regions and translate to expanded polyamino acid sequences. The most well-known class arise by CAG trinucleotide expansion mutations (that encode polyglutamine (polyGln)) in nine genes and which cause a range of neurodegenerative diseases including Huntington's disease, Dentatorubropallidoluysian atrophy, Kennedy's disease and several spinocerebellar ataxias [5]. PolyGln expansions beyond a

range of approximately 35 glutamines, depending on host protein, are typically associated with a gain of toxic function manifested by abnormal protein aggregation in cells upon aging [5]. Less well-known are GCN trinucleotide expansion mutations that lead to polyalanine (polyAla) expansions in proteins, which, when expanded beyond approximately 19 alanines, also cause at least nine different diseases spanning a wide range of pathologies from developmental defects to neurological disorders [6,7]. PolyAla-expansions in FOXE1 may also promote the risk for certain cancers [8,9].

The biophysical features of polyGln sequences and the impact of varying length above and below the disease threshold have been extensively studied [10–12]. In essence, polyGln is initially monomeric in a disordered, heterogeneous conformation and in longer lengths can self-associate via rate-limiting nucleation steps (which may initiated by the formation of altered monomer conformations and-or small oligomers) to ultimately form amyloid-like fibrils rich in β-sheet [10,13]. Aggregation rates positively correlate to glutamine length and can be affected by flanking regions [14,15,16]. Oligomeric intermediates to fibrils as well as possibly the monomer in an altered conformation, have been postulated to underlie a gain of toxicity mechanism of pathogenesis [10,11]. Indeed, small soluble "oligomers" that form transiently as intermediates to fibrils have been postulated to have features inherently toxic to cells regardless of protein sequence [17].

Much less is understood about the impact of polyAla expansions on disease-associated proteins that natively contain polyAla sequences. Because of the apparent similarities in mutation patterns to the polyGln expansion-diseases it is tempting to postulate that polyAla, like polyGln, promotes misfolding and aggregation and indeed this mechanism has been proposed in the field [18]. However, this model is not entirely consistent with other studies suggesting polyAla aggregates into α-helical structures of an amorphous nature [19]. We also found that co-expression of 37Ala and 72Gln fusions to EGFP in the same mammalian cell resulted in spatially and morphologically distinct aggregates suggesting that the molecular bases for aggregation are different [20]. In addition, the expanded polyAla proteins seemed to cluster into oligomers independently to whether cells had visible aggregates or not. As a consequence we posited that polyAla may aggregate via a non-amyloid mechanism. We set out to define the mechanism of polyAla aggregation and how it impacted disease-relevant proteins in the cellular context. Our results reveal polyAla does not form amyloid under normal cellular conditions but instead immediately forms α-helical clusters that provides an explanation to account for how polyAla mutations disrupt normal protein function.

RESULTS

Spontaneous clustering of long polyAla peptides *in vitro*

First, we investigated the aggregation kinetics of a purified model polyAla sequence appended to EGFP as a convenient tag for visual tracking and compared the behaviour to polyGln, which forms amyloid fibrils [10]. Based on previous schemes [21–23], the polyAla and polyGln motifs were also flanked, via a TEV protease cleavage site, with maltose binding protein (MBP)) as a tool to sterically suppress aggregation until the MBP cleaved (Fig 1a). After TEV cleavage, the 7Ala and 25Gln constructs remained soluble over an 8-day time course (Fig 1b). In contrast EGFP-72Gln progressively formed insoluble material to completion as anticipated for aggregation into fibrils. EGFP-37Ala also formed pelletable aggregates to 20% of the input protein levels.

To gain insight into the mechanism of aggregation, we investigated the oligomeric heterogeneity using fluorescence adapted sedimentation velocity analysis (SVA) in an analytical

ultracentrifuge. SVA evaluates the real-time rate of sedimentation of (in our case) fluorescently-labelled proteins under centrifugal force and enables molecular size and heterogeneity to be determined [22]. We previously showed EGFP-37Ala formed oligomeric clusters in cell lysates with SVA[20], but it remained unclear whether this was attributable to self-association or co-aggregation with other cellular proteins. Under low gravitational force (3,000 rpm rotor speed), which sediments only higher protein masses, there was no sedimentation for EGFP-7Ala (Supplementary Fig 1), consistent with an unaggregated soluble state. In contrast, 20% of the fluorescence signal EGFP-37Ala sedimented (Supplementary Fig 1), matching the proportion pelleted in Fig 1b. The sedimentation data could not be fitted to standard size-distribution models that assume non-interaction of species (*e.g.* c(*s*) or ls-g*(*s*) [24]), suggesting non-ideality from the oligomers changing in size during the hours timescale of the experiment.

The oligomeric heterogeneity of the soluble pool was examined by SVA at higher centrifugal force (50,000 rpm rotor speed), which sedimented all remaining proteins (Supplementary Fig 2). c(*s*) size distribution analysis revealed EGFP-7Ala to be monodisperse at *s*=2.4 S (Fig 2a). The mass was estimated by fitting to a non-interacting discrete species model, which takes into consideration the effects of diffusion. The best fit mass of 38 kDa was consistent with monomer of 31 kDa predicted mass. The 37Ala counterpart formed predominantly oligomers in the sedimentation coefficient range of 5–30 S (Fig 2a). The 25Gln and the 72Gln counterparts both yielded single components with sedimentation coefficients of 2.4 S for each (Fig 2a). This corresponded to masses of 39 and 33 kDa respectively, which are sufficiently similar to the predicted monomer masses of 34 and 40 kDa (Fig 2a). No oligomers were detected for EGFP-72Gln; instead the monomer peak diminished after 4 days of incubation as a result of the formation of high molecular weight species that immediately pelleted, and hence were not detectable at this centrifugal speed. This result led us to two conclusions. First was that shorter lengths of polyAla and polyGln are intrinsically monomeric and do not aggregate. Second was that upon expansion polyGln aggregated via monomer recruitment to the fibrils, which is consistent with a nucleated mode of aggregation. Expanded polyAla instead spontaneously adopted a diverse array of higher order oligomers.

We next examined the reversibility of polyAla association into oligomers. SVA of material diluted prior to sedimentation revealed partial EGFP-37Ala dissociation into monomers indicative of reversibility in EGFP-37Ala self-association (Fig 2a). We further investigated the mechanism of association by size exclusion chromatography (SEC). After 1 day TEV cleavage of the fusions (at 5 μM), EGFP-37Ala separated into monomers eluting at 17 mL (as calibrated by the EGFP-7Ala elution time), a major oligomeric population (low mass oligomers) eluting at 12 mL and a shoulder of higher mass oligomers (Fig 2b). Dilution of the EGFP-37Ala sample just prior to analysis by SEC led to the reappearance of a small monomer peak, in agreement with the SVA result showing some back-exchange to monomer (Fig 2b). The fractions corresponding to the monomers (~0.05 μM), low mass oligomers (~1 μM) and high mass oligomers (0.1 μM) were collected and rerun on SEC the same day (Fig 2c) and after an additional overnight incubation (Fig 2d). The monomer fraction remained predominantly monomeric on the same-day rerun but approximately half of this pool converted to oligomer after overnight incubation (Fig 2c, 2d). The low mass oligomers contained a minor fraction of monomers in the same-day rerun but monomers were subsequently depleted after overnight incubation. Collectively these data suggested a mechanism by which monomers and small oligomers interconverted early in an association reaction yet over time the oligomers converted into less reversible forms.

PolyAla aggregates are α-helical

While others have suggested that polyAla can form amyloid fibrils [26,27], there is also conflicting evidence that aggregates are α-helical [19]. Hence we applied circular dichroism (CD) spectroscopy to assess the secondary structure of our polyAla and polyGln models. As anticipated [25], four-day incubated 72Gln formed β-sheet structure after subtracting out the EGFP contribution (Fig 3a). Four-day incubated 37Ala on the other hand was α-helical. Electron microscopy confirmed fibrils for 72Gln whereas 37Ala adopted only non-fibrillar structures, which are likely the largest aggregated states (the small soluble oligomers that make up the bulk of the protein may be the features in the background) (Fig 3b). The 7Ala and 25Gln showed negligible net CD signal, and neither formed any detectable aggregates by electron microscopy.

Long polyAla sequences extensively self-cluster in cells

We next examined the extent of polyAla clustering in the cellular environment. First we performed SVA on lysates from human AD293 cells transfected with the EGFP-polyAla constructs, which revealed long polyAla formed larger oligomeric states (Fig 4a); a result that is consistent with our previous study [20]. The sizes of the oligomers were similar to that produced from the purified counterparts in Fig 2a suggesting they are homo oligomers. To examine whether they were homo oligomers we applied single molecule imaging approaches. Diluted crude lysate was applied to a coverslip and allowed to non-specifically absorb the oligomers into discrete fluorescence spots. The bulk of the adsorbed species appeared as diffraction-limited spots of varying fluorescence intensity (Supplementary Fig 3a). Large aggregates were rare (less than one per field of view) and appeared as very bright, non-diffraction-limited objects with saturated pixel values; these objects could not be quantified and were excluded from analysis. Calibration of the imaging system via single-molecule photobleaching (Supplementary Fig 3b) allowed us to determine the number of EGFP molecules in each oligomer. This analysis confirmed that EGFP-7Ala was monomeric and that EGFP-37Ala (soluble) oligomers were heterogeneous of mostly between 1 and 20 molecules (Fig 4b).

To further define the extent of the self-association patterns in intact cells, we devised a fluorescence resonance energy transfer (FRET) system by coexpressing two forms of each polyAla (or polyGln) construct together; one fused to Venus fluorescence protein (as FRET donor) with one fused to mCherry (as FRET acceptor). Visual inspection of plated cells prior to FRET analysis by flow cytometry indicated approximately 5% of cells expressing the 37Ala constructs to have visible aggregates, while about 10% of cells with 72Gln had visible aggregates similar to our previously reported findings [20] (not shown). Despite the low frequency of cells with visible aggregates, a strong FRET signal was detected in almost all poly37Ala-expressing cells whereas there was no FRET in cells with the 7Ala counterpart (Fig 4c). The FRET signal trailed off for the cells with lowest levels of expression, which is consistent with self-association being driven in a concentration-dependent manner. In contrast, most cells expressing the 72Gln proteins were FRET negative, which is consistent with most of these cells containing monomeric 72Gln.

To further investigate whether polyAla self-association was related to the presence of visible aggregates in cells we reanalysed the flow cytometry data using Pulse Shape Analysis (PulSA). PulSA can discriminate cells that have visible punctate aggregates (inclusions; i population) from those lacking visible aggregates (non-inclusion; ni population) by monitoring changes in the fluorescence width and height parameters[28]. For 72Gln, only cells with inclusions (the i

population) correlated tightly with those that were FRET positive, suggesting 72Gln is mostly monomeric in cells until inclusions form (Fig 4d). This pattern is consistent with a nucleated mechanism of aggregation driving a bimodal population of aggregation states in cells; *i.e.* cells contained either all-aggregated polyGln or were monomer-enriched. For 37Ala, the i population was difficult to unequivocally identify, which was likely due to the poor abundance of cells with inclusions (<5% of cells) and the reduced sensitivity of PulSA for defining less punctate aggregate structures (as is the case for polyAla [20]). Nonetheless almost all cells in the ni population had a strong FRET signal which indicated homo-oligomerization occurred pervasively in each cell irrespective to visible aggregate formation. Cells coexpressing 7Ala and 37Ala displayed no FRET, suggesting the longer polyAla lengths do not recruit short lengths into the aggregate (Supplementary Fig 4). In contrast some co-aggregation was observed for polyGln, which is consistent with prior reports showing aggregation-prone polyGln sequences can recruit shorter, normally non-aggregating polyGln sequences into the aggregates [29].

PolyAla expansion perturbs SOX3 and ARX oligomeric states

To determine whether expanded polyAla –mediated clustering can explain dysfunction of endogenous disease-associated polyAla expansion mutations, we investigated two transcription factors, SRY-related HMG-box 3 (SOX3) and Aristaless related homeobox (ARX) that both have four native polyAla tracts. Expansion of the first tract in human SOX3 from 15 to 22 or 26 alanines in causes X-linked hypopituitarism [6,30,31]. Expansion in either of the first two tracts of human ARX (PA1 from 16 alanine repeats to beyond 18 and PA2 from 12 repeats to beyond 20) causes X-linked intellectual disability and seizures [32,33].

First we assessed human SOX3 fused C-terminally to EGFP transfected in AD293 cells by imaging. Wild-type SOX3 (15Ala) was restricted to the nucleus whereas the 26Ala mutant formed cytoplasmic puncta in approximately 50% of cells (Supplementary Fig 5a), consistent with previous findings [34]. We applied SVA on crude cell lysate of these cells to examine how polyAla expansion affected oligomeric state. Three parallel-run SVA regimes were chosen that we previously found enabled oligomers of vastly different sizes (*i.e*. different sedimentation coefficients (S)) to be measured [20,22]. Small oligomers (*e.g.* monomers up to approximately 10–20 mers; ~1–200 S) require assay at high centrifugal force (50,000 rpm) whereas larger complexes (~200–1500 S) require low centrifugal force (3,000 rpm). Very large aggregates (with a sedimentation coefficient > ~1500 S) are resolvable by a further increase in solution viscosity by the supplementation of 2 M sucrose.

First, we examined the largest complexes that SOX3 adopted in cells with the 3,000 rpm and 2 M sucrose regime. 50% of wild-type SOX3 sedimented under these conditions, suggesting a large extent of SOX3 engaged with very large structures in the cell such as cytoskeleton, chromosomes or potentially incompletely lysed nuclei (Supplementary Fig 6a). The mutant had a small, but significant increase in the proportion of sedimenting material under these conditions (Fig 5a; based on proportions calculated as shown in Supplementary Figure 6a). Notably, fits to a size distribution model (fits of raw data shown in Supplementary Fig 6a) revealed a shift in the distribution of the mutant to smaller masses (Fig 5b). This result suggested polyAla expansion displaced the natural ligand-bound forms of SOX3, which are very large, and promoted abnormal self-aggregates that are on average smaller in size.

The SVA regime focusing on intermediate oligomeric sizes (~200–1500 S) revealed the presence of fast sedimenting material for the 26Ala mutant that was lacking with the wild-type SOX3

(Supplementary Fig 6b). This material sedimented too fast (within 3 min of sedimentation) to enable size-distribution analysis, however its abundance was measurable (Fig 5a) and its presence is consistent with our other data showing polyAla promoting abnormally clustered states.

The SVA regime focusing on smallest oligomeric sizes (~1–200 S) produced data that could not be well-fitted to simple size distribution models, which may reflect exchange between different oligomeric states during sedimentation (which takes place over 150 minutes). However, analysis of the early time points of sedimentation (first 15 min) separately to the later timepoints (40 min – 135 min) overcame this issue. This may be because the early timepoints focussed more on the faster moving larger soluble oligomers (<15 S), whereas the later timepoints on the slower moving smallest oligomers and monomer (>15 S) that exchange at different rates (fits overlaying the raw data shown in Supplementary Fig 6c). Wild-type SOX3 formed a diverse and complex array of low order native oligomeric states which may include ligand-bound forms. The mutant reduced the heterogeneity of the oligomeric states, suggesting polyAla expansion hindered normal SOX3 interactions (Fig 5c). In addition, the mutant decreased the collective abundance of low order oligomers as calculated by the areas under the size distribution curves (Fig 5a). Collectively, these results supported the conclusion that polyAla expansion in SOX3 impaired normal ligand and self-associations and promoted abnormally clustered states.

Our final set of experiments was to examine the influence of polyAla expansion on the clustering patterns of SOX3 and ARX in intact cells. This was achieved by measuring the granularity of polyAla-containing protein within the nucleus of cells lacking otherwise visible aggregates. Using fluorescence micrographs of cells stained for our proteins of interest, we defined a region of interest (ROI) that encompassed the nucleus only for each cell. The pixels in each ROI was measured for mean fluorescence (for protein levels) and standard deviation (for variation in localized protein levels) based on the prediction that these parameters will correlate differently with different clustering patterns (strategy and rationale shown in Fig 6a). As a positive control EGFP-polyAla was targeted to the nucleus with a nucleus-localization sequence (NLS) (Supplementary Fig 5b). The 37Ala variant showed a much greater standard deviation than the 7Ala across medium to high expression levels, which validated this approach (Fig 6b). In transfected cell lines both EGFP-SOX3 (detected by GFP fluorescence; analysis shown in Fig 6c from raw data shown in Supplementary Fig 5a) and human ARX mutants (detected by immunofluorescence; analysis shown in Fig 6d from raw data shown in Supplementary Fig 5c) displayed a significant increase in standard deviation relative to wild-type counterparts for equivalent protein levels, consistent with polyAla expansions causing elevated levels of clustering.

To test whether the increase in clustering also occurred under physiological expression conditions, we performed the same imaging experiments on tissue slices of two human ARX knock-in mouse models of ARX expansion (PA1 and PA2) [35]. ARX expression is known to be most pronounced in the ventrolateral mantle zone of the telecephalon of 12.5 dpc embryonic mice [35]. Hence we examined ARX clustering in this tissue of mutant versus age matched wild-type littermates. In agreement with the transfected cells, ARX mutants displayed elevated standard deviation of fluorescence relative to comparable expression levels of wild-type ARX (analysis shown in Fig 6e from raw data shown in Supplementary Fig 7). Collectively these results suggested that the abnormal clustering patterns caused by polyA-expansion underlie a general mechanism for displacement of normal protein functionality.

DISCUSSION

Our data show that the behaviour of polyAla is fundamentally dissimilar to that of polyGln and amyloid kinetics. Hence, polyAla is unlikely – with the one exception of PABPN1, which is discussed in more detail below – to lead to mechanisms of dysfunction proposed for amyloid proteins such as "toxic oligomers" [12,36]. Instead, our findings suggest a simple and straightforward mechanism by which the polyAla expansions promote the inappropriate self-assembly of these sequences together by α-helical clustering (summarized as a model in Fig 7). This process could be driven by hydrophobic and dipole couplings of α-helices, which are present in these oligomers and appears to have two phases. Initially, clusters may form loosely with a degree of reversibility to dissociate back to monomer but over time the oligomers restructure to become more stable. This non-native clustering could influence the function of proteins in many ways but at the most fundamental level it would be expected to chronically interfere with normal host protein behaviour including normal ligand binding interactions and normal modes of self-association (*i.e*. the depletion of "good oligomers"). Because clustering happens extensively as low-order oligomers the presence of visible or pelletable aggregates would only indicate the extreme end of the deranged clustering spectrum. This was evidenced by our ARX mouse data where we observed no visible aggregates (Fig 6e and Supplementary Fig 7). This is a crucial point because, apart from PABPN1 (which is discussed in more detail below), polyAla expansions do not usually result pathologically as visible protein aggregates except when proteins are highly overexpressed in cell culture [20,37–41]. However, even with PABPN1 there is evidence to support this mechanism in that polyAla expansion leads to loss of ligand binding capacity (ARIH2 E3-ligase) in the soluble state [42].

Our findings reconcile a number of observations more appropriately than an amyloid-type mechanism of aggregation. First is that rather than a gain-of-toxicity function seen in amyloid-related aggregation, most polyAla diseases confer a loss of function [5,6]. A loss of function is a logical consequence of an induced deranged oligomeric status leading to the disruption of normal ligand interactions. Such forms may be expected to be targeted for clearance by the quality control machinery and indeed polyAla-expanded proteins have been shown to colocalize with cellular markers of degradation (reviewed in [18]). Disease is also usually fully present during development and early in life, which is consistent with a chronic dysfunction of the host protein, whereas most amyloid diseases have age-dependent onset and progression – which is more consistent with a nucleated mode of aggregation and a gain-of-toxic mechanism [7,36,40].

PABPN1 is one of the most well-studied polyAla-containing proteins but also is the only one that has been shown to form amyloid fibrils *in vitro* [43–45]. It is also the only polyAla-expanded protein that causes late onset disease. We propose that PABPN1 is distinct to other polyAla-proteins because it can cluster by both polyAla means of self-association and an independent amyloid-mediated pathway. PolyAla expansion has been previously found not necessary in itself to enable PABPN1 to form amyloid fibrils [44]. However, fibrillization has been proposed to be dependent in part on two proposed oligomerizing domains, which suggests that oligomerization helps overcome the rate limiting step of nucleation in amyloid kinetics [46]. Such a two-step mechanism of aggregation has precedence with other amyloid forming proteins [47]. Our data here implicates that under *in vivo* conditions, the polyAla-expansion causes an initial derangement of PABPN1 by abnormally clustering PABPN1 molecules together in a manner that favours a second step of amyloid assembly (Fig 7). The second amyloid assembly step would likely arise in protein domains located elsewhere in the PABPN1 sequence to the polyAla sequence [45].

Further support for this mechanism was observed in a report of PABPN1 N-terminal fragment amyloid fibrillization showing polyAla expansion to confer a net increase in α-helical structure and in this context, polyAla expansion enhanced PABPN1 fibrillization rates [44].

Perhaps because of the known fibrillization of PABPN1 into amyloid, other workers have described polyAla forming amyloid fibrils [19,26,27]. However, these studies described experimental conditions far from those *in vivo* where polyAla is α-helical. The formation of polyAla into β-sheet aggregates has only been reported to happen when placed at high temperatures above about 55 °C or under other extreme solution conditions [19,26,27]. At more physiological temperatures and buffer conditions polyAla strongly favours α-helix formation [19]. Indeed rather than preferentially adopting β-sheet, polyAla has been an exemplar for α-helix structure in the study of peptides, especially in non-polar environments, in part because alanine has the highest propensity of all amino acids to form α-helix [48,49]. Furthermore, there is no evidence of visible aggregate formation in mice expressing 26Ala-SOX3 at endogenous levels [50], consistent with our findings in the ARX knock-in mice (Supplementary Fig 7). Indeed, the SOX3 mutant mice exhibit a reduction in SOX3 protein which, given our findings, would appear to result from the formation of abnormal soluble oligomers that are targeted for degradation. This finding is consistent with ubiquitin-proteasome associated clearance mechanisms proposed for other polyAla expanded proteins [18].

In conclusion, our findings of an alternative explanation for how α-helical mediated clustering of proteins relates to protein dysfunction is an important step forward in efforts to understand and treat these diseases.

ACKNOWLEDGMENTS

This work was funded by grants from the Australian Research Council Discovery: Project grant DP120102763 (DMH), Future Fellowships FT120100039 (DMH), FT100100411 (TB), FT100100560 (AFH); and National Health and Medical Research Council Project grants APP1049458 (DMH), 628946 (AFH), 465401 (PQT) and APP1049459 (DMH and TB). The GFP-Tag-NLS vector was kindly provided by Gregory Mosely from the Department of Biochemisty and Molecular Biology, The University of Melbourne, VIC, Australia.

There are no competing financial interests to declare.

AUTHOR CONTRIBUTIONS

SP, ARO, RJW, KL, CS, JNH, PQT, MGWD, QB and TB performed experiments and–or analysed data. SP, ARO, RJW wrote parts of the manuscript. DMH oversaw the project, analyzed data and wrote the manuscript. AFH, TB, PQT, CS and RJW provided critical feedback and guidance in experimental design.

REFERENCES FOR MAIN TEXT

1. Renton, Alan E. et al. A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
2. Verkerk, A.J.M.H. et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
3. Brook, J.D. et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **68**, 799–808 (1992).
4. Zu, T. et al. Non-ATG–initiated translation directed by microsatellite expansions. *Proceedings of the National Academy of Sciences* **108**, 260-265 (2011).
5. Orr, H.T. & Zoghbi, H.Y. Trinucleotide Repeat Disorders. *Annual Review of Neuroscience* **30**, 575–621 (2007).
6. Hughes, J. & Thomas, P. Molecular Pathology of Polyalanine Expansion Disorders: New Perspectives from Mouse Models. in *Tandem Repeats in Genes, Proteins, and Disease*, Vol. 1017 (eds. Hatters, D.M. & Hannan, A.J.) 135–151 (Humana Press, 2013).
7. Albrecht, A. & Mundlos, S. The Other Trinucleotide Repeat: Polyalanine Expansion Disorders. *Current Opinion in Genetics & Development* **15**, 285 (2005).
8. Kallel, R. et al. Genetic investigation of FOXE1 polyalanine tract in thyroid diseases: new insight on the role of FOXE1 in thyroid carcinoma. *Cancer Biomarker* **8**, 43–51 (2010).
9. Watkins, W.J. et al. An investigation into FOXE1 polyalanine tract length in premature ovarian failure. *Molecular Human Reproduction* **12**, 145–149 (2006).
10. Wetzel, R. Physical Chemistry of Polyglutamine: Intriguing Tales of a Monotonous Sequence. *Journal of Molecular Biology*.
11. Hatters, D.M. Putting huntingtin "aggregation" in view with windows into the cellular milieu. *Current Topics in Medicinal Chemistry* **12**, 2611–22 (2012).
12. Lotz, G. & Legleiter, J. The role of amyloidogenic protein oligomerization in neurodegenerative disease. *Journal of Molecular Medicine* **91**, 653–664 (2013).
13. Vitalis, A. & Pappu, R.V. Assessing the contribution of heterogeneous distributions of oligomers to aggregation mechanisms of polyglutamine peptides. *Biophysical Chemistry* **159**, 14–23 (2011).
14. Scherzinger, E. et al. Self-assembly of Polyglutamine-containing Huntingtin Fragments into Amyloid-like Fibrils: Implications for Huntington's Disease Pathology. *Proceedings of the National Academy of Sciences* **96**, 4604–4609 (1999).
15. Chen, S., Berthelier, V., Yang, W. & Wetzel, R. Polyglutamine Aggregation Behavior *in vitro* Supports a Recruitment Mechanism of Cytotoxicity. *Journal of Molecular Biology* **311**, 173–182 (2001).
16. Duennwald, M.L., Jagadish, S., Muchowski, P.J. & Lindquist, S. Flanking Sequences Profoundly Alter Polyglutamine Toxicity in Yeast. *Proceedings of the National Academy of Sciences* **103**, 11045–11050 (2006).
17. Bucciantini, M. et al. Inherent Toxicity of Aggregates Implies a Common Mechanism for Protein Misfolding Diseases. *Nature* **416**, 507–511 (2002).
18. Messaed, C. & Rouleau, G.A. Molecular mechanisms underlying polyalanine diseases. *Neurobiology of Disease* **34**, 397–405 (2009).

19. Bernacki, J.P. & Murphy, R.M. Length-Dependent Aggregation of Uninterrupted Polyalanine Peptides. *Biochemistry* **50**, 9200–9211 (2011).
20. Polling, S. et al. Misfolded polyglutamine, polyalanine, and superoxide dismutase 1 aggregate via distinct pathways in the cell. *Journal of Biological Chemistry* **289**, 6669–6680 (2014).
21. Ramdzan, Y.M. et al. Conformation sensors that distinguish monomeric proteins from oligomers in live cells. *Chemistry & Biology* **17**, 371–9 (2010).
22. Olshina, M.A. et al. Tracking mutant huntingtin aggregation kinetics in cells reveals three major populations that include an invariant oligomer pool. *Journal of Biological Chemistry* **285**, 21807–16 (2010).
23. Robertson, A.L., Bate, M.A., Buckle, A.M. & Bottomley, S.P. The Rate of PolyQ-Mediated Aggregation Is Dramatically Affected by the Number and Location of Surrounding Domains. *Journal of Molecular Biology* **413**, 879–887 (2011).
24. Schuck, P. & Rossmanith, P. Determination of the sedimentation coefficient distribution by least-squares boundary modeling. *Biopolymers* **54**, 328–341 (2000).
25. Perutz, M.F., Johnson, T., Suzuki, M. & Finch, J.T. Glutamine Repeats as Polar Zippers: Their Possible Role in Inherited Neurodegenerative Diseases. *Proceedings of the National Academy of Sciences* **91**, 5355–5358 (1994).
26. Shinchuk, L.M. et al. Poly-(L-alanine) expansions form core β-sheets that nucleate amyloid assembly. *Proteins: Structure, Function, and Bioinformatics* **61**, 579–589 (2005).
27. Blondelle, S.E., Forood, B., Houghten, R.A. & Pérez-Payá, E. Polyalanine-Based Peptides as Models for Self-Associated β-Pleated-Sheet Complexes. *Biochemistry* **36**, 8393–8400 (1997).
28. Ramdzan, Y.M. et al. Tracking protein aggregation and mislocalization in cells with flow cytometry. *Nature Methods* **9**, 467–70 (2012).
29. Ren, P.-H. et al. Cytoplasmic penetration and persistent infection of mammalian cells by polyglutamine aggregates. *Nature Cell Biology* **11**, 219–225 (2009).
30. Laumonnier, F. et al. Transcription Factor SOX3 Is involved in X-linked Mental Retardation with Growth Hormone Deficiency. *American Journal of Human Genetics* **71**, 1450–5 (2002).
31. Woods, K.S. et al. Over- and Underdosage of SOX3 Is Associated with Infundibular Hypoplasia and Hypopituitarism. *American Journal of Human Genetics* **76**, 833–49 (2005).
32. Shoubridge, C., Fullston, T. & Gecz, J. ARX spectrum disorders: making inroads into the molecular pathology. *Human Mutations* **31**, 889–900 (2010).
33. Marques, I. et al. Unraveling the pathogenesis of ARX polyalanine tract variants using a clinical and molecular interfacing approach. *Molecular Genetics & Genomic Medicine* **3**, 203–214 (2015).
34. Wong, J., Farlie, P., Holbert, S., Lockhart, P. & Thomas, P.Q. Polyalanine Expansion Mutations in the X-linked Hypopituitarism Gene SOX3 Result in Aggresome Formation and Impaired Transactivation. *Frontiers in Biosciences* **12**, 2085–95 (2007).
35. Lee, K., Mattiske, T., Kitamura, K., Gecz, J. & Shoubridge, C. Reduced polyalanine-expanded Arx mutant protein in developing mouse subpallium alters Lmo1 transcriptional regulation. *Human Molecular Genetics* **23**, 1084–1094 (2014).

36. Chiti, F. & Dobson, C.M. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry* **75**, 333–366 (2006).
37. Albrecht, A.N. et al. A Molecular Pathogenesis for Transcription Factor Associated Poly-alanine Tract Expansions. *Human Molecular Genetics* **13**, 2351–9 (2004).
38. Bachetti, T. et al. Distinct pathogenetic mechanisms for PHOX2B associated polyalanine expansions and frameshift mutations in congenital central hypoventilation syndrome. *Human Molecular Genetics* **14**, 1815–24 (2005).
39. Caburet, S. et al. A recurrent polyalanine expansion in the transcription factor FOXL2 induces extensive nuclear and cytoplasmic protein aggregation. *Journal of Medical Genetics* **41**, 932–6 (2004).
40. Shoubridge, C. & Gecz, J. Polyalanine tract disorders and neurocognitive phenotypes. *Advances in Experimental Medicine and Biology* **769**, 185–203 (2012).
41. Utsch, B. et al. Molecular characterization of HOXA13 polyalanine expansion proteins in hand-foot-genital syndrome. *American Journal of Medical Genetics. Part A* **143A**, 3161–8 (2007).
42. Raz, V. et al. A Novel Feed-Forward Loop between ARIH2 E3-Ligase and PABPN1 Regulates Aging-Associated Muscle Degeneration. *The American Journal of Pathology* **184**, 1119–1131 (2014).
43. Winter, R., Kühn, U., Hause, G. & Schwarz, E. Polyalanine-independent Conformational Conversion of Nuclear Poly(A)-binding Protein 1 (PABPN1). *Journal of Biological Chemistry* **287**, 22662–22671 (2012).
44. Scheuermann, T. et al. Trinucleotide expansions leading to an extended poly-l-alanine segment in the poly (A) binding protein PABPN1 cause fibril formation. *Protein Science* **12**, 2685–2692 (2003).
45. Winter, R., Liebold, J. & Schwarz, E. The unresolved puzzle why alanine extensions cause disease. *Biological Chemistry* **394**, 951–63 (2013).
46. Fan, X., Dion, P., Laganiere, J., Brais, B. & Rouleau, G.A. Oligomerization of polyalanine expanded PABPN1 facilitates nuclear protein aggregation that is associated with cell death. *Human Molecular Genetics* **10**, 2341–2351 (2001).
47. Ellisdon, A.M., Thomas, B. & Bottomley, S.P. The Two-stage Pathway of Ataxin-3 Fibrillogenesis Involves a Polyglutamine-independent Step. *Journal of Biological Chemistry* **281**, 16888–16896 (2006).
48. Nick Pace, C. & Martin Scholtz, J. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical Journal* **75**, 422–427 (1998).
49. Liu, L.-P. & Deber, C.M. Uncoupling Hydrophobicity and Helicity in Transmembrane Segments: α-HELICAL PROPENSITIES OF THE AMINO ACIDS IN NON-POLAR ENVIRONMENTS. *Journal of Biological Chemistry* **273**, 23645–23648 (1998).
50. Hughes, J. et al. Mechanistic Insight into the Pathology of Polyalanine Expansion Disorders Revealed by a Mouse Model for X Linked Hypopituitarism. *PLoS Genetics* **9**(2013).
51. Shoubridge, C., Cloosterman, D., Parkinson-Lawerence, E., Brooks, D. & Gecz, J. Molecular pathology of expanded polyalanine tract mutations in the Aristaless-related homeobox gene. *Genomics* **90**, 59–71 (2007).
52. Moseley, G.W. et al. Dual modes of rabies P-protein association with microtubules: a novel strategy to suppress the antiviral response. *Journal of Cell Science* **122**, 3652–62 (2009).

53. Schuck, P. Size-Distribution Analysis of Macromolecules by Sedimentation Velocity Ultracentrifugation and Lamm Equation Modeling. *Biophysical Journal* **78**, 1606–1619 (2000).
54. Ormsby, A.R., Ramdzan, Y.M., Mok, Y.-F., Jovanoski, K.D. & Hatters, D.M. A Platform to View Huntingtin Exon 1 Aggregation Flux in the Cell Reveals Divergent Influences from Chaperones hsp40 and hsp70. *Journal of Biological Chemistry* **288**, 37192–37203 (2013).
55. Kitamura, K. et al. Three human ARX mutations cause the lissencephaly-like and mental retardation with epilepsy-like pleiotropic phenotypes in mice. *Human Molecular Genetics* **18**, 3708–24 (2009).
56. Lee, K. et al. Congenital hydrocephalus and abnormal subcommissural organ development in Sox3 transgenic mice. *PLoS One* **7**, e29041 (2012).
57. Miura, H., Yanazawa, M., Kato, K. & Kitamura, K. Expression of a novel aristaless related homeobox gene 'Arx' in the vertebrate telencephalon, diencephalon and floor plate. *Mechanisms of Development* **65**, 99–109 (1997).
58. Kitamura, K. et al. Mutation of ARX causes abnormal development of forebrain and testes in mice and X–linked lissencephaly with abnormal genitalia in humans. *Nature Genetics* **32**, 359–69 (2002).

FIGURE LEGENDS

Figure 1. Distinct aggregation kinetics of recombinant polyAla and polyGln fusions to EGFP. (a) Expression system. Coomassie-blue stained SDS-PAGE of purified proteins showing the TEV cleavage process (representative of more than 3 experiments). The uncropped gel is shown in Supplementary Data Set 1. **(b)** Kinetics of aggregation by pelleting. Samples of protein were incubated at 37 °C at 5 μM for the times shown after TEV protease cleavage and then assessed for aggregation by pelleting (18,000 *g;* 30 min). *n*=3 technical replicates (same stock protein but treated in parallel tubes upon TEV cleavage), mean ± range shown.

Figure 2: Expanded polyalanine self-associates via an equilibrium shift to large oligomers rather than through a nucleated recruitment of monomers. **(a)** Data shows the fits of sedimentation velocity analysis (SVA) data collected at 50,000 rpm to c(*s*) continuous size distributions. Times refer to the incubation time point after TEV protease cleavage at which SVA was performed. Samples were assessed at two protein concentrations (after incubation; and diluted just prior to SVA) to assess for concentration dependence of aggregation state. **(b)** Size-exclusion chromatography (SEC) profile of EGFP-37Ala assessed at two protein concentrations diluted just prior to SEC, after 1 day incubation post-cleavage. Data show EGFP fluorescence intensity normalised to area under the curve (note the 7Ala is plotted at 0.63× normalized intensity for clarity in the comparison). Inset is zoomed in to show monomer peak for the 37Ala proteins. **(c)** 5 μM EGFP-37Ala incubated for 1 day post-cleavage was assessed by SEC (total, 5 μM) and separated into high mass oligomers, low mass oligomers and monomers. Each species was rerun on the same day (immediate) or after an additional day incubation (overnight, (**d**)). Values in **d.** are raw fluorescence; inset is zoomed in to show monomer peak.

Figure 3. Expanded polyAla assembles into α-helical non-fibrillar aggregate clusters. (a) Circular dichroism spectra of purified (after TEV protease cleavage) EGFP-tagged proteins (3.5 μM) incubated for 4 days at 37 °C. Mean (solid lines) and range (dashed) shown of *n*=3 technical replicates (same stock protein incubated in separate tubes over 4 days). Spectra is corrected for EGFP alone under the same conditions. **(b)** Negative stained electron micrographs of EGFP-72Gln aggregates and EGFP-37Ala.

Figure 4. Expanded polyAla is pervasively homo-oligomerized in cells regardless of visible aggregation state. (a) Best-fit sedimentation velocity analysis (SVA) c(*s*) size distributions of high speed (50,000 rpm) data of AD293 cell lysate from cells transfected with EGFP-polyAla. Cell lysates were adjusted to different total cellular protein concentrations (by bicinchoninic acid assay) as shown; each concentration curve was diluted from the same starting pool of cell lysate. **(b)** Histograms of EGFP units within single oligomeric clusters in the lysate. Data were collected by single molecule step-wise bleaching of EGFP on oligomers immobilized on a coverslip diluted to diffraction limited spots. **(c)** Fluorescence resonance energy transfer (FRET) analysis of intact AD293 cells by flow cytometry. Cells were co-transfected with Venus and mCherry-tagged versions of the constructs shown. Plots show acceptor fluorescence (Venus) against FRET fluorescence (excitation of Venus and emission of mCherry) **(d)** Flow cytometry data reanalysed by PulSA to identify cells with visible aggregates (i; inclusions) from those without visible aggregates (ni: non-inclusions).

Figure 5. Chronic disruption of SOX3 oligomeric state by polyAla expansion. (a) Summary of abundances of different oligomeric classes of EGFP-SOX3 in AD293 cell lysates, as assessed

by sedimentation velocity analysis (SVA) performed under three conditions as shown. Individual measurements (n=3 biological replicates of separate transfections) and mean shown. Data shows the P values calculated for t-test comparisons; however the >1500S comparison failed the Shapiro-Wilk Normality test and the 15–200S comparison failed the equal variance test (P < 0.05). The Mann-Whitney Rank Sum Test for these comparisons both gave P = 0.100. **(b)** SVA c(s) size distributions of the low speed SVA experiments on lysates supplemented with 2 M sucrose. Data shows mean (solid lines) ± range (shaded regions) of n=3 biological replicates (separate transfections). **(c)** c(s) size distributions calculated from the high speed experiment and analysed separately for the early scans versus the later scans. Data shows mean (solid lines) ± range (shaded regions) of n=3 biological replicates (separate transfections). Brackets indicate points of major changes in the size distribution.

Figure 6. Increased granularity of nuclear-localized proteins containing a polyAla-expansion enhances clustering of SOX3 and ARX in cells lacking visible nuclear aggregates. (a) Microscopic images of cells were analysed by measuring pixel intensity variation (as standard deviation) in a nucleus region of interest (ROI). Cells with visible nuclear aggregates were excluded from analysis. Each ROI is made up of pixels (shown as squares), each with different fluorescence intensity values. Homogenous protein distribution results in homogenous pixel intensity (*i.e*. low standard deviation, σ, across the ROI), whereas granularity results in some pixels with higher fluorescence and others with lower fluorescence generating higher σ relative to the mean fluorescence intensity (μ). The histograms further illustrate this concept. For panels (**b–d)**, linear regressions are shown (solid lines). (**b)** Analysis of EGFP fluorescence on cells transfected with EGFP-polyAla-nuclear localization sequence. P = 2e–27 for comparison of regression slopes. (**c**) Analysis of EGFP fluorescence on cells transfected with EGFP-SOX3. P = 6e–6 for comparison of regression slopes. (**d)** Analysis of ARX immunofluorescence on cells transfected with ARX. P = 0.04 for comparison of regression slopes (wildtype vs both mutants). **(e)** Analysis of endogenous ARX *in vivo* by comparison of PA1 and PA2 knockin mice to the wild-type mice. P = 0.408 for PA1 vs wildtype and P = 0.206 for PA2 vs wildtype for comparison of slopes, P = 0.017 and 0.027 respectively for offsets of slopes. Data shows values collected from ventrolateral mantle zone of telecephalon of 12.5 dpc embryonic male mice stained by immunofluorescence.

Figure 7: A schematic summarizing how extended polyAla sequences drive an abnormally clustered state with multiple impacts on structure and function. In this model extended polyAla is α-helical and drives an association into a distribution of clusters of diverse sizes. Our data in Fig 2 suggest these clusters can initially partially dissociate into monomers but convert over time to more stable clusters. We propose that the abnormal clustering interferes with normal ligand interactions of the host protein, including normal self-association patterns, and can trigger targeting for degradation. For ARX and SOX3 the final reaction products are non-amyloid aggregates. For PABPN1 we propose that the α-helically-driven clustered states lower the energetic barrier to amyloid fibril nucleation, which are driven by sequences outside the polyAla stretch. (For simplicity we have not shown the non-amyloid sequences in the diagram of the fibrils).

ONLINE METHODS

Expression constructs. The mammalian polyAla and polyGln constructs were made by cloning cDNA sequences into pEGFP-C1 and pEGFP-C2 vectors respectively. The encoded proteins are shown in Supplementary Table 1. For the bacterial constructs, cDNA sequences were cloned into pET21d(+) to express sequences shown in Supplementary Table 1. Fluorescent proteins moieties were swapped by PCR-mediated cloning methods. SOX3 constructs cloned into pcDNA3.1 as described previously [34] were excised by EcoRI and cloned into the EcoR1 site of pEGFP-C2. ARX constructs were prepared by cloning ARX cDNAs into a pCMV vector with an N-terminal myc tag as described previously [51].For the NLS-tagged constructs, the polyAla-EGFP cDNAs were swapped into the GFP-Tag-NLS vector. This construct expresses the NLS (residues 110–135) of the SV40 large-T antigen [52]. In essence the polyAla-cDNA sequences from Supplementary Table 1 were amplified by PCR with primers NheI-EGFP-NLS forward primer (5' GCTAGCGCTACTTGTCGCC) and the AgeI-EGFP-NLS reverse (5' ACCGGTGGGGTCTTCTACCT). The PCR product was subcloned into NheI–AgeI sites of EGFP-NLS. All constructs were sequenced for verification.

Recombinant protein expression and purification. pET21d(+)-EGFP-polyAla or polyGln constructs were transformed into BL21(DE3) cells and cultured in LB-broth with 100 μg/mL ampicillin. Cells were grown to a density of 0.6 AU and then induced overnight at 20 °C with 1 mM isopropyl β-D-1-thiogalactopyranoside. Cells were pelleted (20 min at 5,000 *g* and 4 ºC) and resuspended in ice-cold 100 mM Tris pH 8.0 containing EDTA-free Complete protease inhibitor cocktail (Roche). Phenylmethylsulfonyl fluoride was added to 2 mM followed by lysis with the addition of hen egg-white lysozyme to 1 mg/ml. Cell lysates were stored at –20 ºC for (at least) overnight. Lysate was thawed on ice and Benzonase (Novagen) was added to 8 U/mL. Lysate was centrifuged (30 min at 35,000 *g* and 4 °C) and the supernatant loaded onto a Nickel-affinity column pre-equilibrated with 20 mM Tris, 250 mM NaCl, 25 mM imidazole, pH 8.0. The column was washed in several column volumes of 20 mM Tris, 250 mM NaCl, 50 mM imidazole, pH 8.0 and eluted with 20 mM Tris, 250 mM NaCl, 150 mM imidazole, pH 8.0. The eluted protein was buffer exchanged into 20 mM Tris, 150 mM NaCl, 12.5 mM imidazole, pH 7.4 with a PD10 column (GE Healthcare). Proteins were aliquoted and snap-frozen in liquid nitrogen. Protein concentration was determined using an EGFP extinction coefficient value of 56,000 $M^{-1}cm^{-1}$ at 488 nm.

Pelleting assays. Aliquots of protein were thawed, pelleted (30 min at 22,000 *g* and 4 ºC) to remove pre-existing aggregates, and diluted to 5 μM after verification of concentration. Cleavage of MBP was initiated by addition of home-made TEV protease for 20 min at room temperature as described previously [22]. Samples (300 μL per point) were incubated at 37 °C for various time points of the time course. One portion of the samples (100 μL) was assayed for fluorescence in a plate-reader for total protein yield. A second portion (150 μL) was pelleted (18,000 *g* for 30 min at room temperature) and 100 μL of the supernatant assayed for fluorescence as "unaggregated" yield.

Cell culture. HEK293T and AD293 cell-lines (from lab cultures orginally obtained from ATCC) were used in this study and tested and cleared for mycoplasma. Cells were not tested for authenticity of cell-line. They were maintained in Dulbecco's modified Eagles Medium (DMEM) supplemented with 1 mM Glutamine, 200 U/mL penicillin-streptomycin and 10% v/v fetal bovine serum. Cells were cultured at 37 ºC in a humidified incubator with 5% atmospheric CO_2. For microscopy experiments 5×10^4 AD293 cells were plated on poly-L-lysine coated 8-well μ-

slides (Ibidi) for all constructs except those involving ARX, for which 8×10^5 HEK293T cells were plated on 6-well plates. For the FRET, flow cytometry and single molecule imaging experiments 2×10^5 AD293 cells were plated on 12-well plates. For SVA, 3×10^6 AD293 cells were plated on T75 flasks. Cells were transfected using Lipofectamine 2000 according to manufacturer's instructions with 0.5 μg DNA for the microscopy experiments, 0.25 μg DNA with the ARX experiments, 1.6 μg DNA for the flow cytometry and single molecule imaging experiments, and 24 μg DNA for SVA using a DNA: lipofectamine ratio of 1:2.5 w/v (Life Technologies). Media was refreshed 6 h after transfection.

Sedimentation velocity analysis. Protein samples and cell lysates were assayed in an XLA analytical ultracentrifuge (Beckman) equipped with a fluorescence detection system (Aviv) as described previously [22] using the Macromolecular Interactions Facility at the Bio21 Institute. The mammalian lysates were prepared from transfected cells, which were detached in phosphate buffered saline (PBS) at 24 hours expression using a cell scraper and lysed as described previously [22]. Data were fitted in the Sedfit software (Peter Schuck; http://www.analyticalultracentrifugation.com), including size distribution models of c(*s*) and ls-g*(*s*) and the non-interacting discrete species model [53].

Size exclusion chromatography. Aliquots of protein were thawed, pelleted to remove pre-existing aggregates and cleaved to remove MBP as described for pelleting assays. Samples were incubated at 37°C at 5 μM overnight. After 24-hour incubation samples were assessed by size exclusion chromatography (SEC) at two different protein concentrations (diluted just prior to size exclusion chromatography). SEC was performed on a Superdex 200 10/300 GL (GE Life sciences) using 20 mM Tris, 150 mM NaCl, 12.5 mM imidazole, pH 7.4 as the running buffer using an Akta explorer FPLC (GE Life Sciences). 500 μl of sample were run with one column volume of running buffer at a flow rate of 0.5 mL/min with 250 μl fractions collected. The column was washed with one column volume of running buffer between each run. The fluorescence of each fraction was assessed using a CLARIOstar microplate reader (BMG Labtech, excitation 488 nm, emission 523 nm, 10 nm bandwidth).

Electron microscopy. 5 μM solutions of TEV protease-treated samples were incubated at 37 ºC 4 days prior to application onto freshly glow discharged carbon-coated copper grids. After 1 min application, solutions were wicked with filter paper, stained with 1.5% w/v uranyl acetate, pH 4.5 for 1 min, wicked again to remove excessive liquid, and air-dried. Samples were imaged at the Bio21 Insitute Advanced Microscopy Unit using a FEI Tecnai TF30 Transmission Electron Microscope (FEI-Company, Eindhoven, The Netherlands) operating at 200 kV. Images were recorded digitally using a Gatan US1000 2k × 2k CCD camera.

Circular dichroism spectroscopy. Proteins were thawed, pelleted and cleaved from MBP as per the pelleting assay. TEV protease and MBP were removed by application through a his-trap spin column and the eluant assessed by SDS-PAGE to validate purity of the polyamino acid EGFP-fusions. The eluant proteins were buffer exchanged into 20 mM sodium phosphate, 100 mM NaCl, pH 7.4, adjusted to 3.5 μM and incubated for 4 days at 37 ºC. Circular dichroism spectra were collected on samples in 1 mm pathlength cuvettes on a Chirascan-plus spectrometer (Applied Photophysics). Data were converted to mean residue ellipticity and corrected for the EGFP contributions to the spectra.

Single molecule imaging. Glass coverslips (#1.5 Warner Instruments) were cleaned in 70% (v/v) ethanol, dried at 70 ºC and exposed to plasma cleaning for 10 min (Harrick Plasma). Cells

were lysed as described previously [22] and protein was diluted in imaging buffer (PBS with addition of 0.5 mg/mL BSA, 1mM Trolox and 10mM Cysteamine) and absorbed onto freshly cleaned coverslips for 2–3 min. Coverslips were then rinsed with imaging buffer to remove unbound protein and imaged using an inverted TIRF microscope (TILL Photonics) with Zeiss α Plan-apochromat 100× oil objective (1.46 NA), 488nm laser line and electron-multiplying charge-coupled device camera (Andor). Images were aquired using 50 ms exposure time. Spots in the fluorescence image were detected as local maxima and background-corrected fluorescence intesities were extracted using home-written image analysis software implemented using MATLAB.

Code Availability. The Matlab protocols used for the single molecule imaging will be provided upon request.

FRET and PulSA. Cells were harvested 72 hours after transfection by gentle pipetting in PBS. Cells were analyzed at a high flow rate in an LSRFortessa flow cytometer, equipped with 405 nm, 488 nm and 561 nm lasers (BD Biosciences). 100,000 events were collected, using a forward scatter threshold of 5,000. Data were collected in pulse height, width and area parameters for each channel. For Venus fluorescence, data were collected with the 488 nm laser and FITC filter. For mCherry fluorescence, data were collected with the 561 nm laser and PE-Texas Red filter. FRET data were collected with the 488 nm laser and the PI filter, and the FRET gate was set using the FRET negative controls as a reference. All flow cytometry data were analyzed and plotted with FlowJo (Tree Star Inc). For PulSA, we analyzed the data as described previously [28,54], and used the non-aggregating protein samples to set the boundary for the ni-i gates.

Imaging. For all experiments involving cultured cells, cells were imaged 24 or 48 hours after transfection. For all experiments except those involving ARX, cells were stained with 2.5 µg/ml Hoescht 33342 and fixed with 4% (w/v) paraformaldehyde in PBS for 15 minutes, with media replaced by PBS. For the SOX3, polyAla and polyGln imaging experiments, cells were imaged on the Leica TCS SP5 Confocal microscope for mCherry (excitation 561 nm, emission 600–700 nm) and EGFP fluorescence exciation 488 nm, emission 500–555 nm) using a 63× objective lens. For the ARX experiments ARX was detected by immunostaining. In essence, cells were fixed with 3.7% (v/v) formaldehyde in PBS and permeabilized with 0.2% (v/v) Triton X-100 in PBS for 5 mins. Cells were blocked with 5% (w/v) skim-milk powder in Tris buffered saline and 0.5% (v/v) Tween (TBS-T) for 1 hour at room temperature. ARX was stained with a previously described primary mouse monoclonal antibody to ARX [51] (2 µg/mL final) overnight at 4 ºC. Cells were rinsed twice in high volume and then washed for 5 mins at room temperature with TBS-T. The DonkeyαMouse IgG Alexa 488 (Life technologies: A21202) secondary antibody was added at 1:1000 diluion for 1 hour at room temperature. Nuclei were counterstained and mounted for microscopy with ProLong Gold antifade reagent with DAPI (Molecular Probes, Invitrogen). Cells were imaged on the Zeiss Axio Imager M2 microscope with data collected using the 488 nm laser and FITC filter.

Mouse models of Arx. *ARX*$^{GCG7/+}$ (BRC number: 03654) and *ARX*$^{432455dup/+}$ (BRC number: 03653) from RIKEN Bioresource Centre, Japan [55] were maintained in the C57BL/6 background. All experiments were approved by Animal Ethics Committee of the University of Adelaide, Australia. Pregnant dams were euthanized by cervical dislocation followed by decapitation of embryos at 12.5 dpc. Whole mount frozen tissue from hemizygous male mice was prepared as described previously [56] and sectioned at 10 mm thick using Microm HM505E. Frozen tissue

sections were air-dried for 1 hour at room temperature with all subsequent procedures performed in a humidified chamber to prevent drying of tissue sections. Sections were permeabilized in PBS + 0.5% (v/v) Triton for 5 min, blocked with blocking solution (10% (v/v) horse serum and 10% (w/v) BSA in PBS + 0.1% Triton) for 1 hour at room temperature. Incubation in primary Arx antibody (rabbit anti-ARX at 1:500, [57,58]) in blocking solution was at 48 °C overnight followed by 1:400 donkey anti-rabbit IgG Alexa488 secondary antibody (Life Technologies: ab 150073) at room temperature for 4 hours with washing (PBS + 0.01% (v/v) Tween 20 for 10 min three times) after each incubation and mounted with ProLong Gold Antifade Reagent with DAPI (Life Technologies). All images were analysed using Olympus IX81 inverted microscope equipped with CellSens 1.3 Software. Immunofluorescence images were acquired by Olympus XM10 black and white camera, while bright field images were captured using Olympus DP70 digital colour camera. No statistical method was used to predetermine sample size. The experiments were not randomized and were not performed with blinding to the conditions of the experiments.

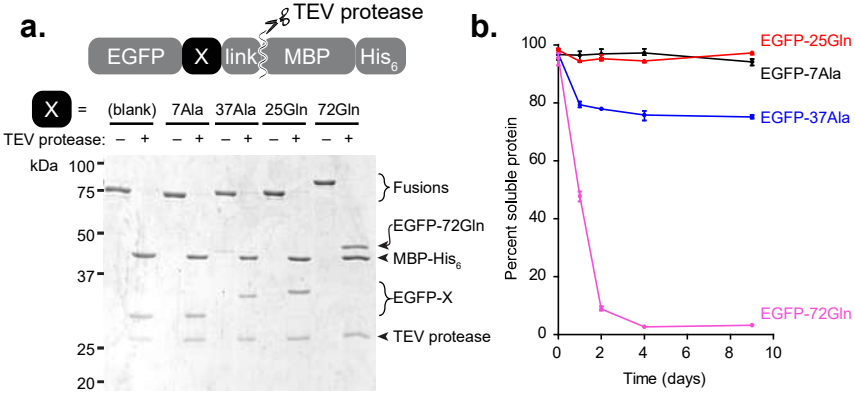
Statistics. The *P* values reported on figures were calculated by the two-tailed Students t-test method using Prism (Graphpad) or Sigmaplot (Systat) software packages. In the case of failed Normality Test (Shapiro-Wilk) or Equal Variance Test (which are indicated in the figures), the Mann-Whitney Rank Sum Test was also performed. For the comparisons of regressions, t-tests on slopes for two independent samples were performed with the Real Statistics Resource Pack software (Release 3.5), Charles Zaiontz. www.real-statistics.com or with Prism software.

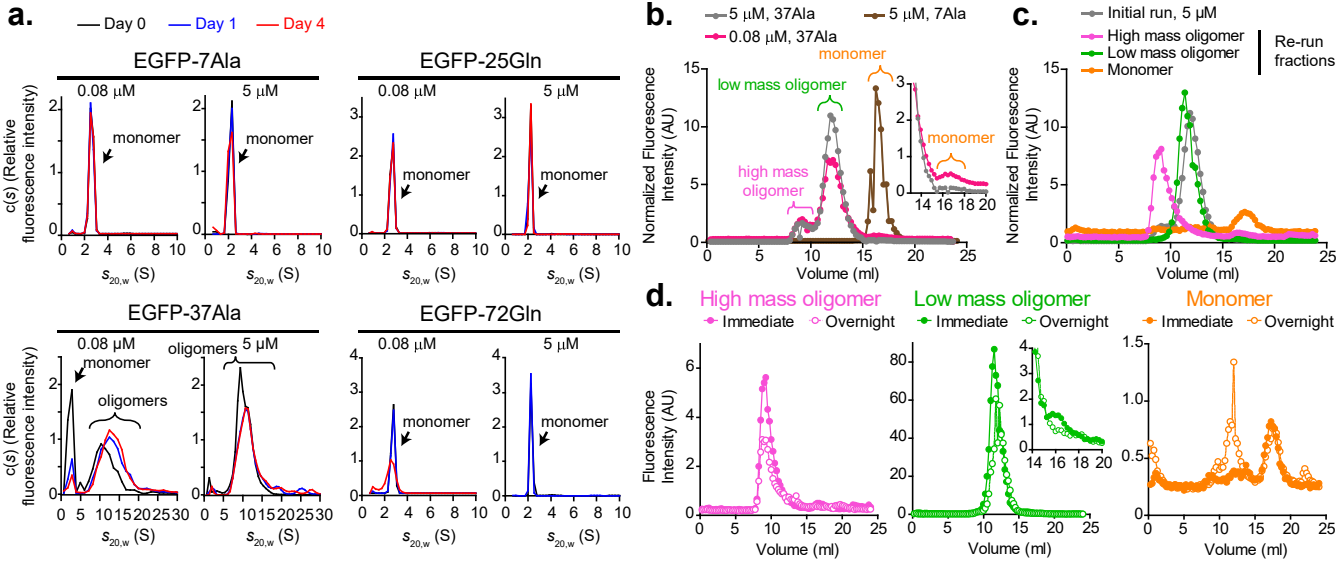
SUPPLEMENTARY INFORMATION

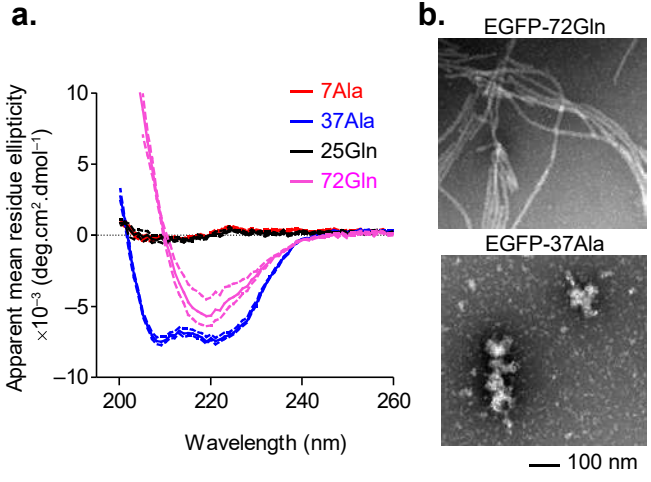
Supplementary Figures 1 to 7 are provided in the template.

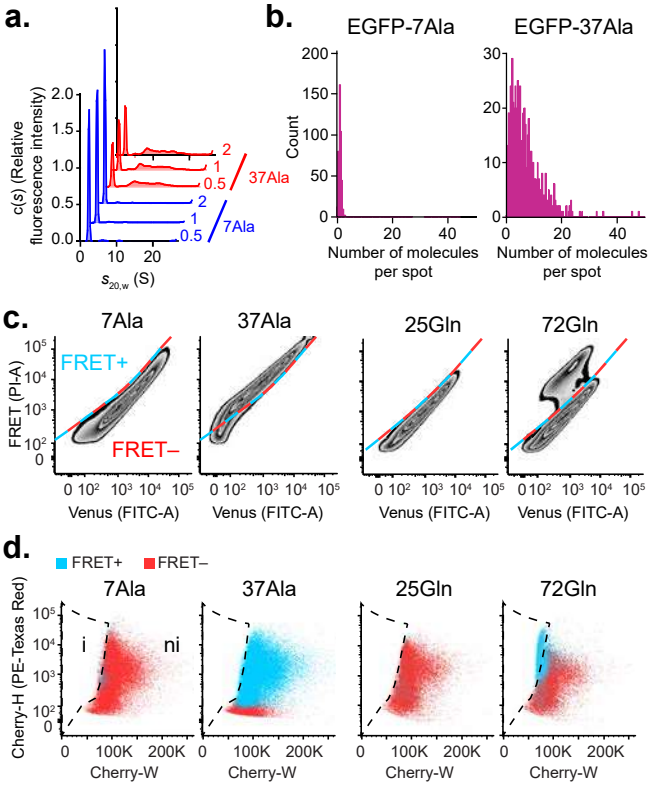
Supplementary Table 1 is included as a PDF.

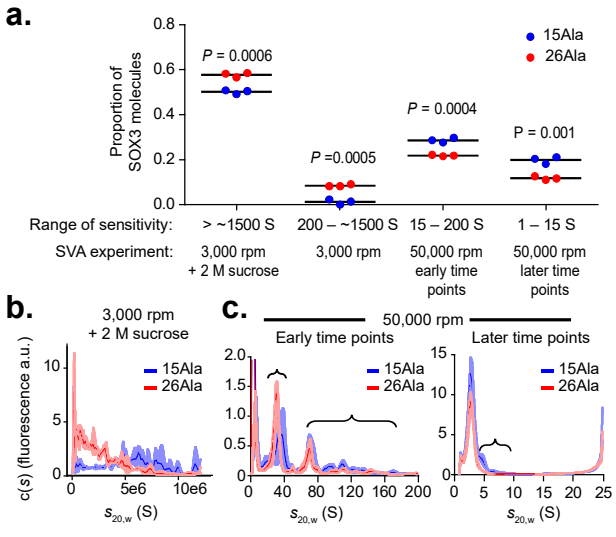
Supplementary Data Set 1 is provided as a PDF.

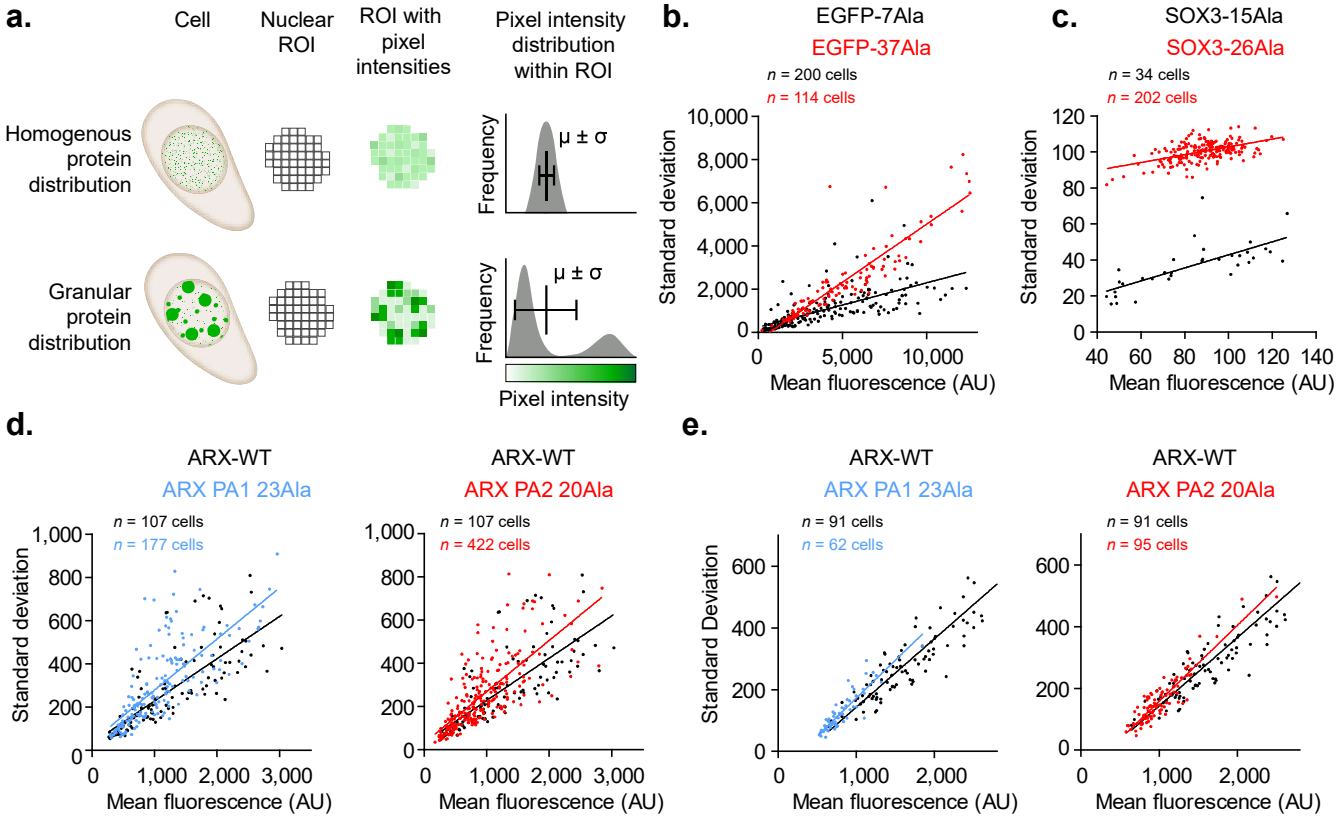

a.
TEV protease
EGFP
X
link
MBP
His6
X = (blank) 7Ala 37Ala 25Gln 72Gln
TEV protease: – + – + – + – + – +
kDa
100
75
50
37
25
20
Fusions
EGFP-72Gln
MBP-His6
EGFP-X
TEV protease
b.
Percent soluble protein
Time (days)
EGFP-25Gln
EGFP-7Ala
EGFP-37Ala
EGFP-72Gln

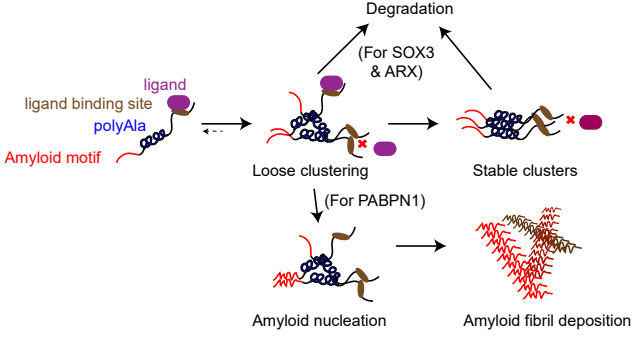

a.
Day 0
Day 1
Day 4
EGFP-7Ala
EGFP-25Gln
EGFP-37Ala
EGFP-72Gln
0.08 μM
5 μM
monomer
oligomers
c(s) (Relative fluorescence intensity)
$s_{20,w}$ (S)
b.
5 μM, 37Ala
5 μM, 7Ala
0.08 μM, 37Ala
low mass oligomer
high mass oligomer
Normalized Fluorescence Intensity (AU)
Volume (ml)
c.
Initial run, 5 μM
High mass oligomer
Low mass oligomer
Monomer
Re-run fractions
d.
High mass oligomer
Low mass oligomer
Monomer
Immediate
Overnight
Fluorescence Intensity (AU)


a.
b.
EGFP-72Gln
EGFP-37Ala
7Ala
37Ala
25Gln
72Gln
Apparent mean residue ellipticity
×10⁻³ (deg.cm².dmol⁻¹)
10
5
0
−5
−10
200
220
240
260
Wavelength (nm)
100 nm


a.
b.
EGFP-7Ala
EGFP-37Ala
c(s) (Relative fluorescence intensity)
$s_{20,w}$ (S)
37Ala
7Ala
Count
Number of molecules per spot
c.
7Ala
37Ala
25Gln
72Gln
FRET+
FRET–
FRET (PI-A)
Venus (FITC-A)
d.
FRET+
FRET–
Cherry-H (PE-Texas Red)
Cherry-W
i
ni


a.
Proportion of SOX3 molecules
0.8
0.6
0.4
0.2
0.0
P = 0.0006
P =0.0005
P = 0.0004
P = 0.001
15Ala
26Ala
Range of sensitivity:
> ~1500 S
200 – ~1500 S
15 – 200 S
1 – 15 S
SVA experiment:
3,000 rpm + 2 M sucrose
3,000 rpm
50,000 rpm early time points
50,000 rpm later time points
b.
3,000 rpm + 2 M sucrose
c(s) (fluorescence a.u.)
15Ala
26Ala
0
5e6
10e6
$s_{20,w}$ (S)
c.
50,000 rpm
Early time points
Later time points
0
40
80
120
160
200
0
5
10
15
20
25


a.
Cell
Nuclear ROI
ROI with pixel intensities
Pixel intensity distribution within ROI
Homogenous protein distribution
Granular protein distribution
Frequency
μ ± σ
Pixel intensity
b.
EGFP-7Ala
EGFP-37Ala
n = 200 cells
n = 114 cells
Standard deviation
Mean fluorescence (AU)
c.
SOX3-15Ala
SOX3-26Ala
n = 34 cells
n = 202 cells
Standard deviation
Mean fluorescence (AU)
d.
ARX-WT
ARX PA1 23Ala
n = 107 cells
n = 177 cells
Standard deviation
Mean fluorescence (AU)
ARX-WT
ARX PA2 20Ala
n = 107 cells
n = 422 cells
Mean fluorescence (AU)
e.
ARX-WT
ARX PA1 23Ala
n = 91 cells
n = 62 cells
Standard Deviation
Mean fluorescence (AU)
ARX-WT
ARX PA2 20Ala
n = 91 cells
n = 95 cells
Mean fluorescence (AU)


Degradation
(For SOX3
& ARX)
ligand
ligand binding site
polyAla
Amyloid motif
Loose clustering
Stable clusters
(For PABPN1)
Amyloid nucleation
Amyloid fibril deposition