Database, 2016, 1–8
doi: 10.1093/database/baw021
Database update

Database update

PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins

Chen Li[1], Jeremy Nagel[1], Steve Androulakis[2], Jiangning Song[1,3,*] and Ashley M. Buckle[1,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, [2]Monash Bioinformatics Platform, Monash University, Melbourne, Vic. 3800, Australia, and [3]National Engineering Laboratory of Industrial Enzymes and Key Laboratory of Systems Microbial Biotechnology, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

*Corresponding author: Tel: +61 3 9902 9313; Fax: +61 3 9902 9500; Email: Ashley.Buckle@monash.edu

Correspondence may also be addressed to Jiangning Song. Tel: +61 3 9902 9304; Fax: +61 3 9902 9500; Email: Jiangning.Song@monash.edu

Citation details: Li,C., Nagel,J., Androulakis,S. *et al.* PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins. *Database* (2016) Vol. 2016: article ID baw021; doi:10.1093/database/baw021

Received 13 August 2015; Revised 8 February 2016; Accepted 8 February 2016

Abstract

Proteins with expanded polyglutamine (polyQ) repeats are involved in human neurodegenerative diseases, via a gain-of-function mechanism of neuronal toxicity involving protein conformational changes that result in the formation and deposition of β-sheet-rich aggregates. Aggregation is dependent on the context and properties of the host protein, such as domain context and location of the repeat tract. In order to explore this relationship in greater detail, here we describe PolyQ 2.0, an updated database that provides a comprehensive knowledgebase for human polyQ proteins. Compared with the previous PolyQ database, our new database provides a variety of substantial updates including detailed biological annotations and search options. Biological annotations in terms of domain context information, protein structural and functional annotation, single point mutations, predicted disordered regions, protein–protein interaction partners, metabolic/signaling pathways, post-translational modification sites and evolutionary information are made available. Several new database functionalities have also been provided, including search using multiple/combinatory keywords, and submission of new data entries. Also, several third-party plug-ins are employed to enhance data visualization in PolyQ 2.0. In PolyQ 2.0 the proteins are reclassified into 3 new categories and contain 9 reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins (reviewed by UniProt curators). We envisage that this updated database will be a useful resource for functional and structural investigation of human polyQ proteins.

Database URL: http://lightning.med.monash.edu/polyq2/

Introduction

The polyglutamine (polyQ) repeat containing proteins harbour a stretch of multiple consecutive glutamines (1). Expansion of the polyQ tract can lead to a toxic gain-of-function via a conformational change within the protein and the deposition of β-sheet-rich amyloid-like fibrils (2–4). As such, polyQ repeats are implicated in several neurodegenerative diseases, including Huntington disease and spinocerebellar ataxia (5–11). The length and domain context (i.e. the domains flanking the polyQ tract) is critical to the pathogenesis of the polyQ repeat (12–15). In addition to aggregation, other pathogenic mechanisms including calcium signaling (16), uncommon protein interaction (17) and proteasome dysfunction are also responsible for polyQ diseases (18–20). PolyQ 2.0 provides an improved tool to understand these diseases, and is a collection of all currently known human polyQ repeat-containing proteins. Nine of these proteins are implicated in pathogenesis, with the precise repeat threshold to pathogenesis varying within the disease subset (21–23).

Given the importance of polyQ repeats and their domain context information, we recently performed a bioinformatics investigation of the protein context of polyQ repeats (24), and constructed a web-accessible database of all human proteins containing a polyQ repeat greater than seven glutamines in length (25). Although the PolyQ database provides basic information for each entry, it lacks in both depth and breadth of annotation as well as functionality. Here, we present PolyQ 2.0, a substantially updated knowledgebase for human polyQ proteins. Compared with PolyQ, PolyQ 2.0 contains a variety of structural and functional annotations (such as polyQ protein disease models in mouse, protein 3D structure, Pfam domain, post-translational modification sites, single point mutations and complementary protein annotations), and domain context of polyQ repeats. Here, domain refers to a consecutive protein sequence motif. In addition, the usability of the web interface has also been improved, including the availability of database search with multiple keywords and user data submission with multiple levels of annotations including gene/protein basic information, protein structural and functional annotations. PolyQ 2.0 updates the MySQL relational database that stores entries, and enhances the web interface through the use of modern Javascript tools for visualization and interaction. Apache Tomcat mediates users' access to the database through Java Servlets and JavaServer Pages.

Update of database entries

Although PolyQ contained two types of polyQ proteins, namely disease and non-disease-associated, in PolyQ 2.0 all entries are categorized into three groups according to the annotation of disease involvement and completeness of review by UniProt curators. Here, disease-associated proteins refer to those proteins causing neurodegenerative diseases due to the abnormal expansion of polyQ repeats (e.g. Huntingtin; UniProt ID: P42858; PolyQ 2.0 ID: PD00043) rather than other proteins with common disease-associated mutations (e.g. CREB-binding protein; UniProt ID: Q92793; PolyQ 2.0 ID: PD00019). These groups are: reviewed disease-associated polyQ proteins, reviewed non-disease polyQ proteins and un-reviewed polyQ proteins. We first validated all the data entries in the previous PolyQ database with their UniProt annotation in order to ensure that only high quality data entries are included in PolyQ 2.0. Proteins were included as reviewed entries according to their annotation in the UniProt database. We incorporated polyQ proteins that have not been manually verified from UniProt as un-reviewed polyQ proteins for potential future reference. As a result, we obtained 9 reviewed disease-associated polyQ proteins, 105 reviewed non-disease polyQ proteins and 146 un-reviewed polyQ proteins (Figure 1A). We envisage that our identification and

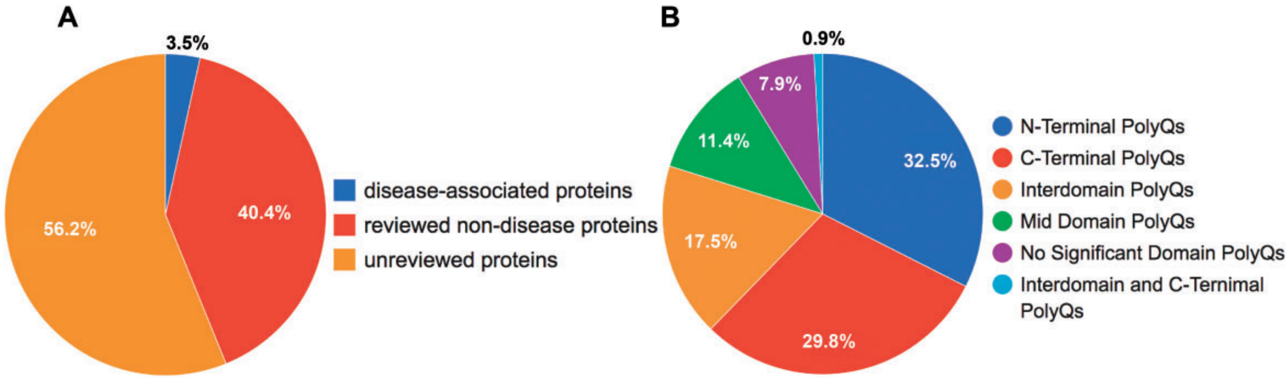


Figure 1. Statistics of data entries in PolyQ 2.0. (**A**) Distribution of disease-associated proteins, reviewed non-disease proteins and un-reviewed proteins. (**B**) Distribution of the sequence context of different types of polyQ domains for reviewed entries only.

annotation of polyQ proteins in un-reviewed sequences may provide opportunities for disease discovery and further investigation. Since un-reviewed sequences are automatically annotated by UniProt but not manually verified, we will endeavor to keep entries updated as their annotation in UniProt evolves.

Following the classification system set out previously in PolyQ, we further classified all reviewed 114 sequences into 6 categories based on the locations and context of polyQ repeats relative to Pfam domains (26): (i) N-Terminal PolyQs—the first polyQ repeat appears before all Pfam domains (e.g. Ataxin-1; ID: PD00011), (ii) C-Terminal PolyQs—the last polyQ repeat appears after all the Pfam domains (e.g. CREB-binding protein; ID: PD00019), (iii) Interdomain PolyQs—the polyQ tracts appear between the first Pfam and last Pfam domain (e.g. BMP-2-inducible protein kinase; ID: PD00016), (iv) Mid Domain PolyQs—the polyQ repeat appears in the middle of a Pfam domain or overlaps with a Pfam domain (e.g. Atrophin-1; ID: PD00010), (v) No Significant Domain PolyQs—sequences that do not contain any significant Pfam domains (e.g. Mastermind-like domain-containing protein 1; ID: PD00053) and (vi) Unclassified PolyQs—sequences that do not fit into any of the above categories. The majority of polyQ domains are either N- or C-Terminal PolyQs, whereas only 7.9% of the reviewed polyQ containing entries do not harbor any significant Pfam domains (Figure 1B).

Update of content and annotation

For PolyQ 2.0, the information content and annotations for entries have been significantly improved and expanded. The updated content includes basic protein information, protein structural information, predicted disordered regions, protein–protein interaction partners, metabolic/signaling pathways, single point disease- and non-disease associated mutations, and protein post-translational modification sites. In addition, we also performed BLAST search and generated multiple sequence alignments (MSAs) in order to provide evolutionary information for each protein entry. A comparison of protein annotations provided in PolyQ and PolyQ 2.0 is shown in Table 1.

Table 1. A comparison of protein annotation in PolyQ and PolyQ 2.0

Content	PolyQ	PolyQ 2.0
Gene information	No	Links to show overall gene information, gene sequences and variations
Protein information	Sequence and unstructured FASTA headers	Structured protein information (function, gene name, protein accession, etc.)
Protein 3D structure	No	Yes
Pfam domain	Yes	Yes
Protein disordered regions	No	Yes
Protein interaction partner	No	Yes
Metabolic/signaling pathway	No	Yes
Single point mutation	No	Yes, incorporating both disease-associated and nonsense mutations
Post-translational modification sites	No	Yes
Disease models on mouse	No	Yes
MSA	No	Yes

Table 2. Summary of the database contents and annotations of PolyQ 2.0 and PolyQ

Annotation	PolyQ 2.0	PolyQ
Number of proteins	260	128
Number of protein structures	356	0
Number of protein interactions	4081	0
Number of single point mutations	704	0
Number of KEGG pathways	25	0
Number of Pfam domains	175	132
Number of post-translational modification sites	569	0

Annotations were extracted and reviewed from a variety of different publicly available resources, including ENA (27), 1000 Genome Project database (28), UniProt (29), Protein Data Bank (30), BioGrid (31), KEGG (32), SUPERFAMILY (33), Pfam (Version 27.0) (26) and Mouse polyQ database (34). We employed VSL2B (35) to annotate predicted disordered regions. Homologous sequence search was conducted using PSI-BLAST (36) (with an *E*-value of 0.001) against the UniProt database (http://www.uniprot.org/downloads). MSAs were generated using Clustal Omega (37). A summary of the database contents and annotations of PolyQ 2.0 and PolyQ is shown in Table 2.

We performed a statistical analysis of the database contents in terms of the distribution of disease-associated mutations, post-translational modification sites and number of protein–protein interaction partners. From a total of 704 single point mutations within the 114 reviewed entries, 458 (65.1%) mutations are disease-associated, whereas 246 (34.9%) mutations are polymorphisms (Figure 2A). We further defined two mutation patterns, i.e. $<A, C, \ldots, V> \rightarrow X$ and $X \rightarrow <A, C, \ldots, V>$ to identify trends in the wild-type amino acid that is mutated (Figure 2B, left) and the amino acid after mutation (Figure 2B, right). By analyzing the

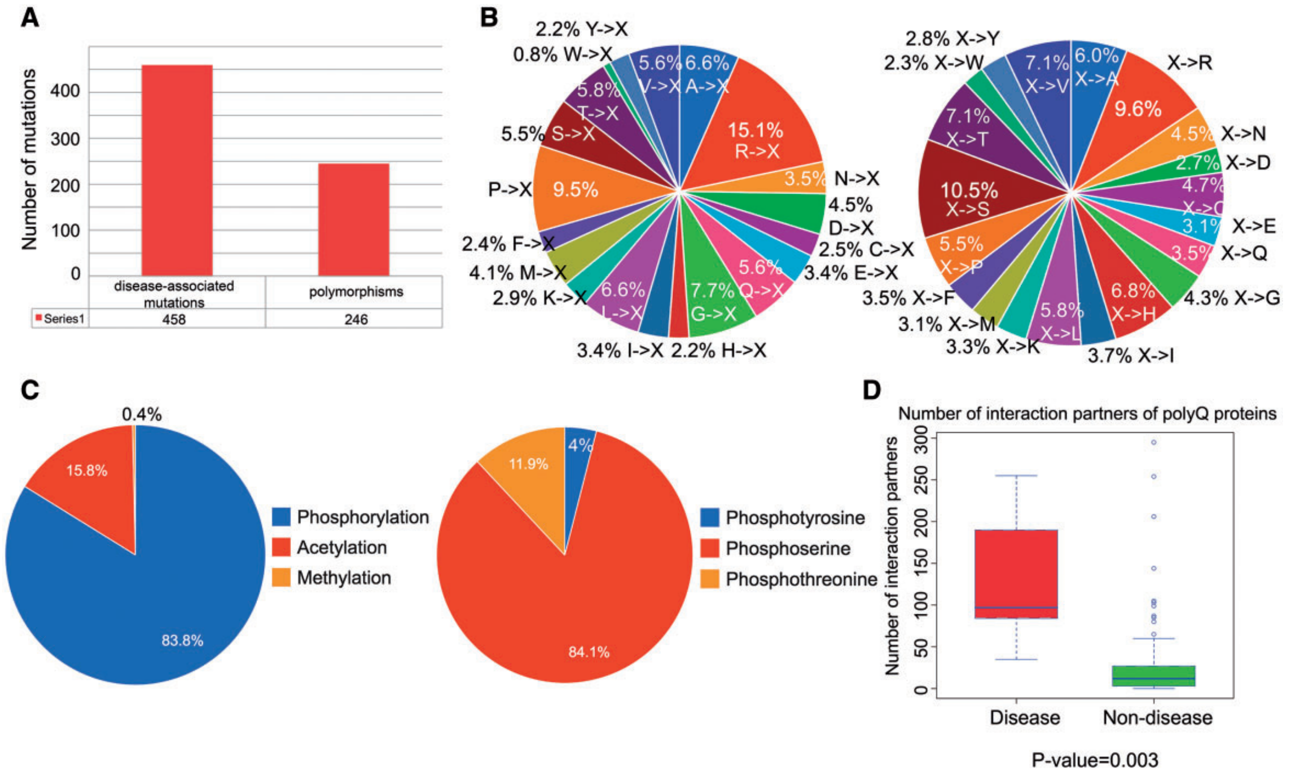


Figure 2. Statistical analysis of database content in terms of distributions of disease-associated mutations, post-translational modification site and number of protein–protein interaction partners. (**A**) Distribution of disease-associated mutation and polymorphism. (**B**) Distribution of the number of mutations with respect to two mutation patterns (where X means any amino acid). (**C**) Distribution of types of protein post-translational modification with detailed distribution of sub-types of phosphorylation. (**D**) Number of protein–protein interaction partners of reviewed polyQ disease-associated proteins and non-disease proteins.

Table 3. Database functionality comparison between PolyQ and PolyQ 2.0

Functionality	PolyQ	PolyQ 2.0
Database search		
Database ID/UniProt ID	No	Yes
Gene name	No	Yes
Protein name	Yes	Yes
Pfam domain	Yes	Yes
Disease	No	Yes
PTM	No	Yes
PTM kinase	No	Yes
Interaction partner	No	Yes
Combinatory search	No	Yes
User submission	No	Yes

distribution of different types of mutations associated with polyQ proteins, we found that arginine is the most frequently mutated amino acid (accounting for approximately 15% of the mutated residues; Figure 2B, left). No trend was apparent for *X*→<*A*, *C*,..., *V*> type mutations (Figure 2B, right). Phosphorylation is the most frequently observed post-translational modification (Figure 2C). Disease-associated polyQ proteins have significantly more protein interaction partners than non-disease polyQ proteins (*P*-value = 0.003; Figure 2D). However, interpretation of the biological relevance of these observations should be interpreted with caution as the dataset may contain bias towards disease-associated polyQ proteins due to their relatively greater interest compared with non-disease polyQ proteins. At the present time, these biological observations are based on the current dataset of PolyQ 2.0 and may vary when the database is updated.

Database functionality and web interface improvements

PolyQ 2.0 features several important improvements of the user interface as well as new functionality, including database search with multiple types of keywords and new entry submission. A comparison of database functionality between PolyQ and PolyQ 2.0 is listed in Table 3.

The search functionality in PolyQ 2.0 has been considerably improved, with search options available using multiple keywords, in addition to the options of protein name and Pfam domain offered by the previous version. The database can be searched by PolyQ/UniProt ID, protein/

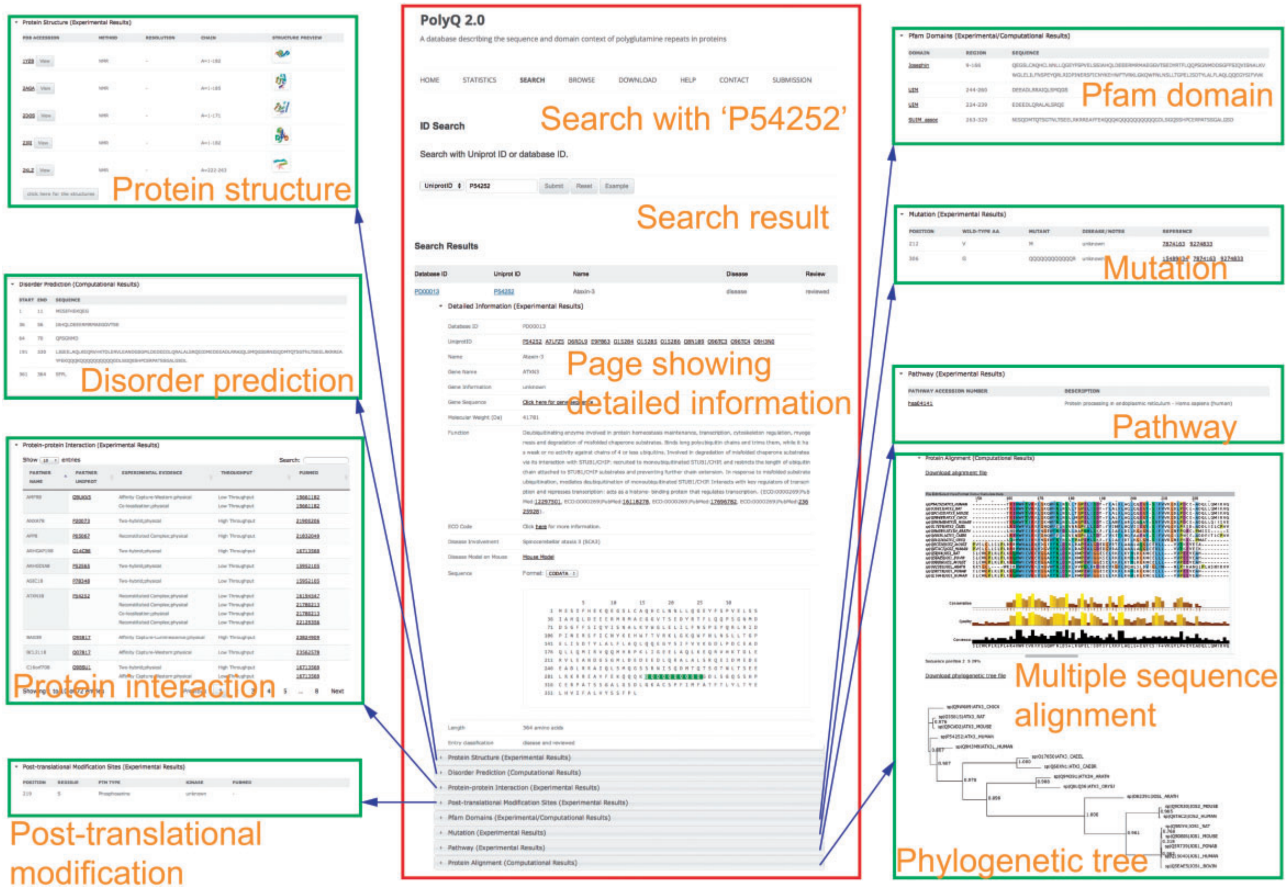
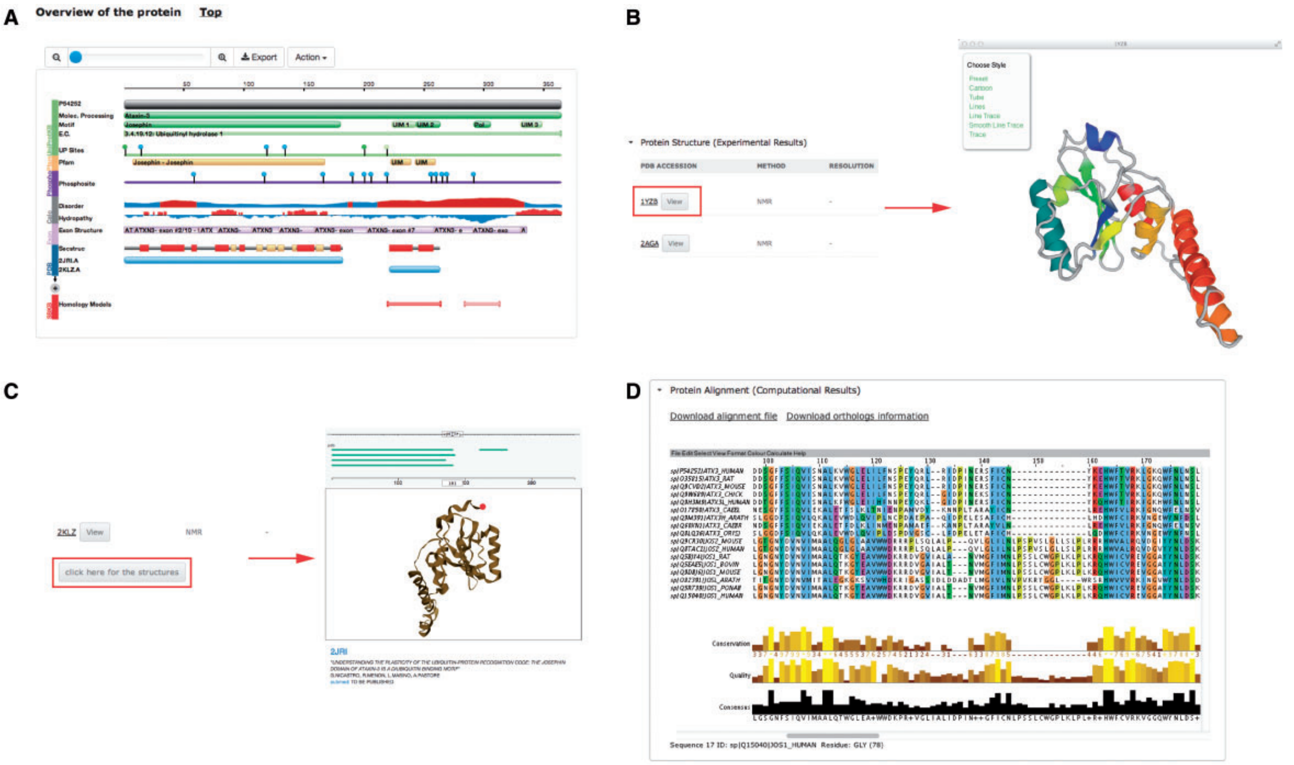

PolyQ 2.0
A database describing the sequence and domain context of polyglutamine repeats in proteins
HOME STATISTICS SEARCH BROWSE DOWNLOAD HELP CONTACT SUBMISSION
ID Search
Search with Uniprot ID or database ID.
UniprotID P54252 Submit Reset Example
Search Results
Search with 'P54252'
Search result
Page showing detailed information
Protein structure
Disorder prediction
Protein interaction
Post-translational modification
Pfam domain
Mutation
Pathway
Multiple sequence alignment
Phylogenetic tree

Figure 3. Typical search results in PolyQ 2.0 using the UniProt ID P54252 as an example. The website consists of nine sections showing detailed information for each entry, including gene/protein information, protein structure, metabolic/signalling pathway, protein interaction, post-translational modification site, Pfam domain, disorder region prediction, protein mutation and MSA.

gene name, Pfam domain, disease, type of protein post-translational modification sites/kinase, protein–protein interaction partner name, and combinatory keywords including disease, PTM type and polyQ domain context. The PolyQ ID is composed of 'PD' followed by five digits. As there are in total 260 entries in PolyQ 2.0, the PolyQ ID ranges from 'PD00001' to 'PD00260'. An example of the result of database search with UniProt ID = P54252 (Ataxin-3) is shown in Figure 3, comprising nine main sections related to different annotations.

Several plug-ins were employed to enhance visualization of database entries. In the protein basic information section, we embedded a protein feature view plug-in in order to show protein functional sites/domains and basic structural information (Figure 4A). PV (http://biasmv.github.io/pv/) and pViz (38) were also used to allow detailed examination of protein structures (Figure 4B and C). MSA is displayed using JalView (39) to visualize sequence conservation (Figure 4D). In addition to MSA, we generated the phylogenetic tree for each protein entry in PolyQ 2.0 using FastTree (40) and visualized the result using the jsPhyloSVG package (41). This provides a good visualization of the evolutionary relationships among the protein and its closest homologs. We also provided links to PhylomeDB (42) to show phylogenetic trees using a more comprehensive tree structure view.

Browsing of data entries has also been improved. The entries can now be categorized in terms of disease involvement and completeness of review and annotation. In addition, detailed context annotations, which show the distribution of polyQ domain, protein superfamily domain, Pfam domain and protein post-translational modification sites are available. The legend providing the explanations of the shapes and colours used can also be found at this page. In the figure provided, overlapped domains have been aligned horizontally based on their ranges. A webpage showing database statistics is available, giving users a one-page snapshot of database contents as well as convenient navigation around the database. Detailed user help and instructions are also provided. Finally, we made available a data submission page, where users can conveniently deposit their new polyQ sequences with detailed sequence/structural/functional annotations if applicable, including protein basic information, protein structural information, protein–protein interaction,



PDB ACCESSION	METHOD	RESOLUTION
1YZB View	NMR	-
2AGA View	NMR	-

2KLZ View	NMR	-

Figure 4. Plug-ins in PolyQ 2.0 to enhance database visualization. (**A**) Protein feature view plug-in. (**B**) PV showing protein structure. (**C**) pViz for visualizing multiple structures. (**D**) Jalview displaying MSAs.

mutation, functional domains/sites and pathway information. In addition, the submission for new entries does not require a specific format or style for the annotations. Users can provide simple sentences and references to describe the structural and functional annotations. All the submitted data will be formatted, structured and made publicly available once it has been carefully checked, curated and approved by a database administrator.

Conclusions

Based on our previous PolyQ database for human polyQ proteins, in the present study we have developed an updated database, PolyQ 2.0, to provide comprehensive protein functional, structural and evolutional annotations together with domain context information for human polyQ proteins. Integrating publicly available annotations and computational resources, PolyQ 2.0 offers a variety of annotations in terms of protein basic information, protein structure, predicted intrinsically disordered domain, protein–protein interaction, protein functional site/domain, single point mutation, metabolic/signaling pathway and MSA. Given that the third-party databases (e.g. Pfam, UniProt, etc.) integrated in PolyQ 2.0 update regularly we will endeavor to update our database on a regular basis. This is particularly important since the categories of polyQ proteins defined in our study are based on the context information of polyQ and Pfam domains. In future revisions, we will cross-reference the NCBI Conserved Domains Database (43) to incorporate more functional domains into PolyQ 2.0 to enrich the annotations. We anticipate that this updated knowledgebase will benefit functional and structural studies of human polyQ proteins and their role in neurodegenerative diseases.

Acknowledgements

We would like to thank Dr Andreas Prlić for his help with the protein feature view plug-in. J.S. is a recipient of the Hundred Talents Program of CAS. A.M.B. is an NHMRC Senior Research Fellow.

Funding

National Natural Science Foundation of China (61202167, 61303169); the Hundred Talents Program of the Chinese Academy of Sciences (CAS); the Knowledge Innovative Program of CAS (KSCX2-EW-G-8). Funding for open access charge: National Health and Medical Research Council (1022688).

Conflict of interest. None declared.

References

1. Fan,H.C., Ho,L.I., Chi,C.S. *et al.* (2014) Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant.*, 23, 441–458.
2. Perutz,M.F., Johnson,T., Suzuki,M. *et al.* (1994) Glutamine repeats as polar zippers: their possible role in inherited

neurodegenerative diseases. *Proc. Natl. Acad. Sci. U S A*, 91, 5355–5358.

3. Chen,S., Berthelier,V., Hamilton,J.B. *et al.* (2002) Amyloid-like features of polyglutamine aggregates and their assembly kinetics. *Biochemistry*, 41, 7391–7399.
4. Robertson,A.L., Horne,J., Ellisdon,A.M. *et al.* (2008) The structural impact of a polyglutamine tract is location-dependent. *Biophys. J.*, 95, 5922–5930.
5. Kawaguchi,Y., Okamoto,T., Taniwaki,M. *et al.* (1994) CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.*, 8, 221–228.
6. Lin,B., Nasir,J., MacDonald,H. *et al.* (1994) Sequence of the murine Huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat [corrected]. *Hum. Mol. Genet.*, 3, 85–92.
7. Zuhlke,C., Hellenbroich,Y., Dalski,A. *et al.* (2001) Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia. *Eur. J. Hum. Genet.*, 9, 160–164.
8. Nakamura,K., Jeong,S.Y., Uchihara,T. *et al.* (2001) SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.*, 10, 1441–1448.
9. Silveira,I., Miranda,C., Guimaraes,L. *et al.* (2002) Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus. *Arch. Neurol.*, 59, 623–629.
10. Banfi,S., Servadio,A., Chung,M.Y. *et al.* (1994) Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat. Genet.*, 7, 513–520.
11. Quan,F., Janas,J. and Popovich,B.W. (1995) A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. *Hum. Mol. Genet.*, 4, 2411–2413.
12. Stefani,M. and Dobson,C.M. (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, 81, 678–699.
13. Saunders,H.M. and Bottomley,S.P. (2009) Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng. Des. Sel.*, 22, 447–451.
14. Ellisdon,A.M., Thomas,B. and Bottomley,S.P. (2006) The two-stage pathway of ataxin-3 fibrillogenesis involves a polyglutamine-independent step. *J. Biol. Chem.*, 281, 16888–16896.
15. DiFiglia,M., Sapp,E., Chase,K.O. *et al.* (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, 277, 1990–1993.
16. Bezprozvanny,I. (2009) Calcium signaling and neurodegenerative diseases. *Trends Mol. Med.*, 15, 89–100.
17. Thompson,L.M. (2008) Neurodegeneration: a question of balance. *Nature*, 452, 707–708.
18. Zheng,C., Geetha,T. and Babu,J.R. (2014) Failure of ubiquitin proteasome system: risk for neurodegenerative diseases. *Neurodegener. Dis.*, 14, 161–175.
19. Jansen,A.H., Reits,E.A. and Hol,E.M. (2014) The ubiquitin proteasome system in glia and its role in neurodegenerative diseases. *Front. Mol. Neurosci.*, 7, 73.
20. Dantuma,N.P. and Bott,L.C. (2014) The ubiquitin-proteasome system in neurodegenerative diseases: precipitating factor, yet part of the solution. *Front. Mol. Neurosci.*, 7, 70.
21. Goto,J., Watanabe,M., Ichikawa,Y. *et al.* (1997) Machado-Joseph disease gene products carrying different carboxyl termini. *Neurosci. Res.*, 28, 373–377.
22. Padiath,Q.S., Srivastava,A.K., Roy,S. *et al.* (2005) Identification of a novel 45 repeat unstable allele associated with a disease phenotype at the MJD1/SCA3 locus. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 133B, 124–126.
23. Li,W., Serpell,L.C., Carter,W.J. *et al.* (2006) Expression and characterization of full-length human huntingtin, an elongated HEAT repeat protein. *J. Biol. Chem.*, 281, 15916–15922.
24. Robertson,A.L., Bate,M.A., Buckle,A.M. *et al.* (2011) The rate of polyQ-mediated aggregation is dramatically affected by the number and location of surrounding domains. *J. Mol. Biol.*, 413, 879–887.
25. Robertson,A.L., Bate,M.A., Androulakis,S.G. *et al.* (2011) PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Res.*, 39, D272–D276.
26. Finn,R.D., Bateman,A., Clements,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, 42, D222–D230.
27. Leinonen,R., Akhtar,R., Birney,E. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, 39, D28–D31.
28. Abecasis,G.R., Auton,A., Brooks,L.D. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.
29. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, 43, D204–D212.
30. Rose,P.W., Prlic,A., Bi,C. *et al.* (2015) The RCSB protein data bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, 43, D345–D356.
31. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, 43, D470–D478.
32. Kanehisa,M., Goto,S., Sato,Y. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42, D199–D205.
33. Wilson,D., Pethica,R., Zhou,Y. *et al.* (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, 37, D380–D386.
34. Szlachcic,W.J., Switonski,P.M., Kurkowiak,M. *et al.* (2015) Mouse polyQ database: a new online resource for research using mouse models of neurodegenerative diseases. *Mol. Brain*, 8, 69
35. Peng,K., Radivojac,P., Vucetic,S. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7, 208
36. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389–3402.
37. Sievers,F., Wilm,A., Dineen,D. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7, 539.
38. Mukhyala,K. and Masselot,A. (2014) Visualization of protein sequence features using JavaScript and SVG with pViz.js. *Bioinformatics*, 30, 3408–3409.
39. Waterhouse,A.M., Procter,J.B., Martin,D.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25, 1189–1191.

40. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.
41. Smits,S.A. and Ouverney,C.C. (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, 5, e12267.
42. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P. *et al.* (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.*, 42, D897–D902.
43. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, 43, D222–D226.