RESEARCH ARTICLE

Proteostasis networks in aging: novel insights from text-mining approaches

Diogo Neves · Sara Duarte-Pereira · Sérgio Matos · Raquel M. Silva

Abstract Aging is a topic of paramount importance in an increasingly elderly society and has been the focus of extensive research. Protein homeostasis (proteostasis) decline is a hallmark in aging and several age-related diseases, but which specific proteins and mechanisms are involved in proteostasis (de) regulation during the aging process remain largely unknown. Here, we used different text-mining tools complemented with protein–protein interaction data to address this complex topic. Analysis of the integrated protein interaction networks identified novel proteins and pathways associated to proteostasis mechanisms and aging or age-related disorders, indicating that this approach is useful to identify previously unknown links and for retrieving information of potential novel biomarkers or therapeutic targets.

Diogo Neves and Sara Duarte-Pereira have contributed equally to this work.

D. Neves · S. Duarte-Pereira · R. M. Silva (✉)
Department of Medical Sciences & iBiMED, University of Aveiro, Aveiro, Portugal
e-mail: rmsilva@ucp.pt

S. Duarte-Pereira · S. Matos
IEETA, University of Aveiro, Aveiro, Portugal

S. Matos
DETI, University of Aveiro, Aveiro, Portugal

R. M. Silva
Universidade Católica Portuguesa, Faculdade de Medicina Dentária, Centro de Investigação Interdisciplinar em Saúde, Estrada da Circunvalação, 3504-505 Viseu, Portugal

Introduction

Aging is the progressive physiological and functional decline of organisms that affects the tissues and organs of the whole body. It is best characterized as a multifactorial process that comprises the interaction of cellular and molecular mechanisms, however, individuals do not age at the same rate (Broz and Dixit 2016). Since it is estimated that by 2050, more than 20% of the world´s population will be over 65 years of age (Beard et al. 2016), finding healthy aging biomarkers and therapeutic targets is of outmost importance to predict the biological age of individuals and significantly impact the socio-economic burden caused by an increasingly elderly society.

Maintaining protein homeostasis (proteostasis) is crucial for protein structure, stability, and functional properties. The proteostasis network includes chaperones, co-chaperones, the ubiquitin-proteasome system (UPS) and the autophagic machinery as protein

quality control mechanisms (Díaz-Villanueva et al. 2015). In addition to these processes, organelle-specific systems exist and encompass the heat-shock response (HSR), as well as the endoplasmic reticulum (ER) and mitochondria unfolded protein response (UPR^{ER} and UPR^{mito}). These processes act in a coordinated fashion to ensure correct protein folding (Díaz-Villanueva et al. 2015; Read and Schröder 2021). When these cellular and molecular protein quality control machineries are overwhelmed, the formation of unfolded proteins increases the tendency of proteins to aggregate.

Earlier in the past decade, *Lopez-Otín* et al. included defective proteostasis, or proteotoxicity, in a panel of hallmarks of aging and age-related disorders (López-otín et al. 2013). Despite numerous publications that revealed that unbalanced proteostasis might be a consequence or a cause of aging (Thibaudeau et al. 2018; Luna et al. 2018; Kelmer Sacramento et al. 2020), there is still a lack of proteins that combine these two phenomena. As a better understanding of the molecular pathways involved in proteostasis (de)regulation during aging might lead to the identification of biomarkers and targets for therapeutic intervention, the objective of this work was to provide an integrative view of the regulatory networks involved in proteostasis and the aging process, using two complementary text-mining approaches.

Material and methods

Automated text-mining dataset

Automated text mining analysis was performed using the Agilent Literature Search plugin for Cytoscape (Cline et al. 2007). This plugin retrieves documents for a user query and extracts protein–protein associations from these texts using a set of lexicons that define protein names and association terms. A set of queries was built up to include "Proteostasis", "protein aggregation" and "proteotoxicity" as the terms selected and "aging", "ageing" and "senescence" as the context query. The species *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans* and *Drosophila melanogaster* were considered, resulting in the retrieval and analysis of 769 full-text biomedical research papers and the creation of a list of proteins for each of the species.

Under extraction controls the option "relaxed" was chosen, which means that a more permissive set of association terms to identify potential protein interactions was considered.

Curated text-mining dataset

A curated dataset of protein interactions was generated using EGAS (Campos et al. 2014; Matos et al. 2016). Pre-processed abstracts containing automatically identified concepts (proteins, drugs, species, bioprocesses) and relations (protein–protein and protein-chemical interactions) were revised by the curators using the interactive tool. The 1920 results obtained from the PubMed search ((protein aggregation OR proteostasis OR proteotoxicity) AND (ageing OR aging OR senescence)) were prioritized using a linear Support Vector Machine (SVM) classifier. The classifier was initially trained with data from the BioCreative III protein–protein interaction (PPI) article classification task (Krallinger et al. 2011), consisting of 6280 PubMed abstracts classified as containing PPI information (n = 1822) or not (n = 4458). This classifier was then refined through two steps of user feedback in which the 300 highest ranked abstracts, according to classification probability, were manually verified and annotated as containing relevant information (positive) or otherwise (negative). After retraining the classifier using these manually verified documents the full list of 1920 PubMed results was re-ranked. Curators then annotated the first documents of this prioritized list, resulting in a curated corpus of 108 abstracts. Specifically, this annotation process was performed using EGAS, a web-based text-mining and assisted curation tool. EGAS uses dictionary matching and machine learning to automatically identify and annotate concepts and concept relations mentioned in text document. These automatic annotations were then inspected, corrected or removed manually by curators, who also added new annotations. After exporting the curated annotations to a tabular text format, a manual revision step was required to assign unique UniProt identifiers (47 IDs) to protein annotations that had not been normalized during the interactive curation step. The final PPI pairs were imported to Cytoscape for network visualization.

Network analysis

Cytoscape version 3.4.0 (Shannon et al. 2003) was used to visualize and analyze the protein networks resulting from the two text-mining approaches. A third network was generated using the Cytoscape "import—network—from public databases" tool. The list of UniProt accession IDs was used as input to retrieve the protein–protein interaction (PPI) network from databases of the IMEX consortium (Orchard et al. 2012) (IntAct, MINT, I2D, InnateDB, MPIDb). The human proteins were selected and the PPIs from the two text-mining approaches that were not present in databases were identified. For each network, self-loops and duplicated edges were removed and a column was added to the edge table, indicating the number of underlying edges to each interaction. The EGAS and Agilent filtered networks (2NET) were merged, and these were merged with the network built with information retrieved from public databases (3NET). For key proteins (nodes) the first neighbors were selected, and subnetworks were created containing only edges connected to the central node, to evaluate which interactions were identified by each approach. In the merged network of the three approaches, a clustering analysis was performed using MCODE App from Cytoscape (Bader and Hogue 2003). The protein sub-network consisting of the novel protein interactions (excluding databases) was visualized with STRING (http://string-db.org/), using evidence data from text-mining and experiments with a medium confidence score (0.4) and no interactors in the first or second shell.

Enrichment and pathway analyses

The list of UniProt IDs of each dataset was analyzed in the Reactome Pathway Database (Croft et al. 2011). All non-human identifiers were converted to their human counterparts. The statistical overrepresentation test was performed in PANTHER (www.pantherdb.org), using PANTHER GO-Slim Biological Process dataset (Mi et al. 2017). Gene enrichment analyses were performed with the Functional annotation clustering tool in DAVID (https://david.ncifcrf.gov/), using the default parameters (Sherman et al. 2022; Huang et al. 2009) These analyses provided statistical measures for the association of our datasets to gene ontology terms, giving an insight of the functional biological meaning of the results.

Results

For the automated text-mining dataset, the Agilent Literature Search plugin for Cytoscape was used to generate a list of proteins for all the species considered. The resulting protein networks consist of 155 nodes (proteins) and 284 edges (interactions) for *D. melanogaster*, 74 nodes and 105 edges for *C. elegans*, 325 nodes and 629 edges for *H. sapiens* and 265 nodes and 466 edges for *M. musculus*. All proteins from each species were analyzed in the Reactome Pathway Database and similar results were obtained for all species. The integrated analysis of the statistically significant processes shows an overrepresentation of pathways related to the immune system, cellular responses to stress and programmed cell death (Supplementary Fig. 1).

The curated text-mining dataset was obtained using EGAS. A total of 110 protein–protein interactions (PPIs), nine protein-chemical interactions (PCIs) and 62 protein-pathway relations were identified, involving 91 unique proteins that were mapped to their UniProt accession IDs. Fifty-six protein names were not mapped and appear as annotated in the text. For network analysis, only PPIs were considered.

The results of the statistical overrepresentation test, performed on the 91 UniProt IDs from the EGAS dataset, showed that the main processes represented were apoptotic process, response to stress, intracellular signal transduction and regulation of biological process. The analysis in the Reactome Pathway Database revealed that the top four pathways represented were related to cellular responses to stress, particularly to heat stress and involving heat shock transcription factor 1 (HSF1). A total of 34 pathways had a significant p-value (< 0.001) and were involved mainly in signal transduction, immune system, programmed cell death and disease. Overall, the results from both automated and curated datasets were similar.

To compare the output obtained from the literature searches and text mining tools with the information described in specific databases of protein–protein interactions, a network using the automatic search of public databases in Cytoscape was created. As an input, the smaller dataset of protein IDs, consisting of

91 UniProt accession numbers obtained from EGAS was used and only the databases from IMEX Consortium, which are curated and standardized, were selected. A network with a total of 5414 interactions involving 3578 proteins was obtained.

After filtering from the text-mined PPIs the interactions already contained in the public databases, the curated network of PPIs contained 88 non-duplicated edges connecting 116 nodes. The same was done to the automated dataset of human proteins and a network of 323 nodes and 596 edges was obtained. Comparing these two, 28 of the identified proteins (nodes) were common to both datasets. Of note, 27 of these appear to directly interact and participate in a proteostasis-related network that has not been previously acknowledged (Fig. 1A). Functional enrichment of the 28 proteins present in both text-mining datasets indicates 13 significant clusters (Supplementary Table 1), mainly associated with protein degradation or folding, apoptosis and neuronal survival. Although with fewer genes, Alzheimer's Disease and NAD-dependent ribosylation are among the clusters with higher fold enrichment (137.8% and 64.6%, respectively).

Considering only the proteins mapped to UniProt IDs, the two text-mining approaches together allowed the identification of 570 PPIs that were not present in the network obtained with database information. Examples are shown in Fig. 1B and 1C, representing the subnetworks centered in HSF1 and in α-synuclein (SNCA) and tau (MAPT). The subnetworks comprise all interactions obtained from the three different methods and show 61 HSF1 interactions, four identified by EGAS and 12 by Agilent (Fig. 1B). HSF1, NEDD4, SNCA, SIRT1 and their interactions can be retrieved from the subnetwork. In the second example (Fig. 1C), key proteins with a role in aging related disorders, such as Alzheimer's disease and Parkinson's disease are shown. Although most of the interactions are already found in the databases, EGAS and Agilent provided new information as well.

Discussion

Proteostasis and aging seem to be closely related processes. Despite its most prevalent association to pathologies such as diabetes or neurodegenerative diseases, the aging process itself is associated with a severe decline in the proteostasis machinery. Here a text-mining strategy was followed, towards a most comprehensive overview of the molecular pathways involved both in proteostasis and in the aging process to identify predictive biomarkers. Two distinct but complimentary approaches were applied, using the Agilent Literature Search plugin for Cytoscape and EGAS that provide, respectively, automated and curated methods based on the analysis of the literature generated from terms of interest. Given the words that provide an association between proteostasis and aging as described in the methods section, networks of proteins were generated and further analyzed using Cytoscape, Reactome and PANTHER. The results demonstrate the valuable contribution of text-mining approaches to identify novel hypothesis and targets that can be further validated, as discussed next.

Considering the two text-mining strategies, the Agilent tool is less restrictive when compared to the manually curated approach and allows the use of full-text. On the other hand, manual literature curation results are more reliable, but since annotating full-text documents would require a much greater curation effort, only abstracts were annotated with EGAS. Accordingly, more information was retrieved with the automated method. Using EGAS, new information could still be extracted, namely, the identification of 82 PPIs that were not present in either the Agilent or the public databases networks.

Merging the networks from the two text-mining approaches revealed 28 common nodes (proteins) and 6 common edges (interactions). These included proteins such as PARK7, MAPK8, SQSTM1, SIRT1, VCP, STAT3, SIRT6, APP, DDIT3 and CLU, showing that these methods provide new, and literature supported information. For example, APP interaction with Superoxide Dismutase 1 (SOD1), a key participant in apoptotic signaling and oxidative stress, was detected. SOD1 had been associated with Amyotrophic lateral sclerosis (ALS) and was recently linked to Alzheimer's disease (Muresan and Ladescu 2016). Clusterin also referred as apolipoprotein J (ApoJ), a small heat shock protein that acts as a molecular chaperone and promotes cell survival, has been largely reported in the literature associated with Alzheimer's disease and other neurodegenerative diseases (Yuste-Checa et al. 2022; Jackson et al. 2019). According to the literature, CLU is a good target for the development of therapeutic approaches (Wilson
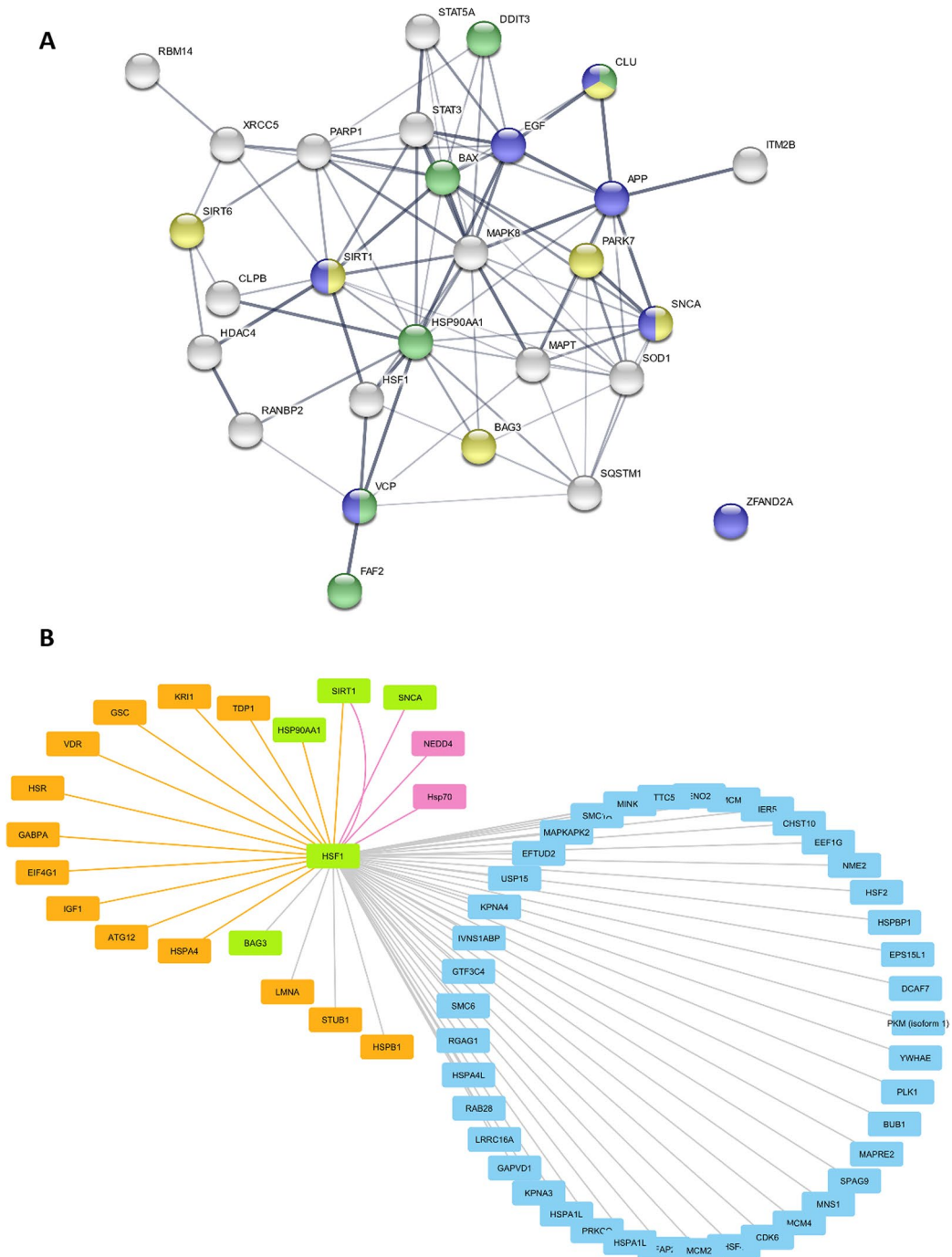
Fig. 1 Subnetwork of proteins and protein–protein interactions identified by text-mining approaches as associated with proteostasis and aging. **A** Nodes (proteins) identified by both datasets were visualized using STRING and shown to participate in proteostasis-related processes, such as response to topologically incorrect protein (GO:0035966, green), regulation of protein stability (GO:0031647, yellow) or positive regulation of proteolysis (GO:0045862, blue). **B**, **C** Subnetworks centered in HSF1, and in MAPT and SNCA. Nodes (proteins) and Edges (interactions) from the automated and curated dataset networks are shown in orange and pink, respectively. Common proteins are represented in green. Nodes and edges retrieved from public databases are shown in blue and gray. The Group Attributes Layout based on the origin of the dataset (Curated, Automated or Databases) and the Prefuse Forced Directed Layout were used to visualize the network
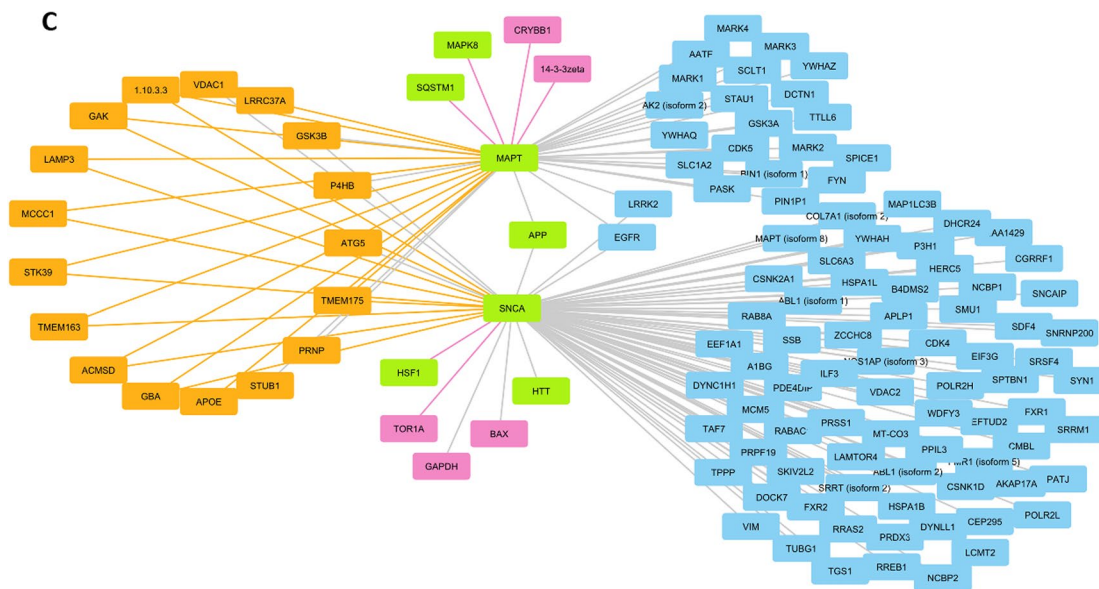
Fig. 1 (continued)

and Zoubeidi 2017), however its interaction with APP was not found in the public databases network but was identified by the EGAS text-mining tool. In addition, both text-mining approaches identified the interaction of amyloid-β (APP) with BCL2 Associated Athanogene 3 (BAG3). BAG3 is involved in chaperone-assisted selective autophagy and was already associated to Alzheimer's disease (Lei et al. 2015). Also, a BAG1 to BAG3 switch was observed during aging, being BAG1 associated with the removal of polyubiquitinated proteins from the proteasome and BAG3 related with the turnover of polyubiquitinated proteins from the autophagic-lysosomal system (Gamerdinger et al. 2009).

The enrichment analysis revealed the pathways that were overrepresented in our set of results. This means that there was a significant number of proteins/genes associated to those pathways, and this number is higher than we would have expected to find by chance, using the entire human genome as reference. The top 4 pathways were related to cellular responses to stress, particularly involving the HSF1 regulation or transactivation. EGAS provided two more HSF1 interactions with SNCA and NEDD4 that were not found in public databases. NEDD4 (Neural Precursor Cell Expressed, Developmentally Down-Regulated 4, E3 Ubiquitin Protein Ligase) is an enzyme that targets proteins for ubiquitination. Its involvement in

neurodegeneration and interaction with HSF1 were described (Kim et al. 2016). Interestingly, the HSF1-SIRT1 interaction was acknowledged by both text-mining approaches. SIRT1, a NAD-dependent deacetylase well studied in aging, had a subnetwork with 136 edges, most of them found in databases (data not shown).

From the statistically significant processes represented, it should also be noted the overrepresentation of processes related to the immune system. The NLRP3 inflammasome is the major inflammatory complex (Broz and Dixit 2016) and its activation triggers the caspase-1 mediated processing of IL-1β and IL-18 into their bioactive cytokine forms and pyroptosis (Broz and Dixit 2016). It is not clear whether NLRP3 activation is a cause or a consequence of mitochondrial dysfunction however it is known that this complex can be activated through several damage-associated molecular patterns (DAMPs) such as ATP, DNA, heat shock proteins and protein aggregates or even mitochondrial DAMPs such as cathepsin B and reactive oxygen species (ROS) (Gurung et al. 2015). Moreover, the activation of mitochondrial biogenesis that is a pathway related to the inflammasome activation is also observed. NLRP3 activation is enhanced in many age-related diseases and is also associated with aging itself (Gritsenko et al. 2020). The acetylation status of NLRP3 is

affected during aging. NLRP3 is normally modified by SIRT2-dependent deacetylation in macrophages, a process that is dysregulated during aging. This switch towards an acetylated status activates the inflammasome leading to chronic inflammation and insulin resistance (He et al. 2020).

Conclusion

In conclusion, the work presented here shows how text-mining tools can be used to extract knowledge from available data. To face the increasing number of published papers in a topic such as aging and the need to progress with directed studies to identify new potential biomarkers or therapeutic targets, text-mining tools provide a useful approach as a first step in the research process.

Declarations

Ethical approval Not applicable.

Competing interests The authors declare no competing interests.

References

Bader GD, Hogue CWV (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 4:2. https://doi.org/10.1186/1471-2105-4-2

Beard JR, Officer A, de Carvalho IA, Sadana R, Pot AM, Michel J-P et al (2016) The World report on ageing and health: a policy framework for healthy ageing. Lancet (London, England) 387:2145–2154. https://doi.org/10.1016/S0140-6736(15)00516-4

Broz P, Dixit VM (2016) Inflammasomes: mechanism of assembly, regulation and signalling. Nat Rev Immunol 16:407–420. https://doi.org/10.1038/nri.2016.58

Campos D, Lourenço J, Matos S, Oliveira JL (2014) Egas: a collaborative and interactive document curation platform. Database 2014:bau048. https://doi.org/10.1093/database/bau048

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C et al (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2:2366–2382. https://doi.org/10.1038/nprot.2007.324

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L et al (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691–D697. https://doi.org/10.1093/nar/gkq1018

Díaz-Villanueva JF, Díaz-Molina R, García-González V (2015) Protein folding and mechanisms of proteostasis. Int J Mol Sci 16:17193–17230. https://doi.org/10.3390/ijms160817193

Gamerdinger M, Hajieva P, Behl C (2009) A switch from BAG1 to BAG3 during ageing triggers the enhanced use of the autophagic-lysosomal system for the degradation of polyubiquitinated proteins. FASEB J 23:668. https://doi.org/10.1096/fasebj.23.1_supplement.668.1

Gritsenko A, Green JP, Brough D, Lopez-Castejon G (2020) Mechanisms of NLRP3 priming in inflammaging and age related diseases. Cytokine Growth Factor Rev 55:15–25. https://doi.org/10.1016/j.cytogfr.2020.08.003

Gurung P, Lukens JR, Kanneganti T-D (2015) Mitochondria: diversity in the regulation of the NLRP3 inflammasome. Trends Mol Med 21:193–201. https://doi.org/10.1016/j.molmed.2014.11.008

He M, Chiang H-H, Luo H, Zheng Z, Qiao Q, Wang L et al (2020) An acetylation switch of the NLRP3 inflammasome regulates aging-associated chronic inflammation and insulin resistance. Cell Metab 31:580–591. https://doi.org/10.1016/j.cmet.2020.01.009

Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. https://doi.org/10.1038/nprot.2008.211

Jackson RJ, Rose J, Tulloch J, Henstridge C, Smith C, Spires-Jones TL (2019) Clusterin accumulates in synapses in Alzheimer's disease and is increased in apolipoprotein E4 carriers. Brain Commun 1:fcz003. https://doi.org/10.1093/braincomms/fcz003

KelmerSacramento E, Kirkpatrick JM, Mazzetto M, Baumgart M, Bartolome A, Di Sanzo S et al (2020) Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. Mol Syst Biol 16:e9596. https://doi.org/10.15252/msb.20209596

Kim E, Wang B, Sastry N, Masliah E, Nelson PT, Cai H et al (2016) NEDD4-mediated HSF1 degradation underlies α-synucleinopathy. Hum Mol Genet 25:211–222. https://doi.org/10.1093/hmg/ddv445

Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-aryamontri A, Winter A et al (2011) The protein-protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinform 12:S3. https://doi.org/10.1186/1471-2105-12-S8-S3

Lei Z, Brizzee C, Johnson GVW (2015) BAG3 facilitates the clearance of endogenous tau in primary neurons. Neurobiol Aging 36:241–248. https://doi.org/10.1016/j.neurobiolaging.2014.08.012

López-otín C, Blasco MA, Partridge L, Serrano M (2013) Europe PMC Funders Group The Hallmarks of. Aging 153:1194–1217. https://doi.org/10.1016/j.cell.2013.05.039.The

Luna E, Decker SC, Riddle DM, Caputo A, Zhang B, Cole T et al (2018) Differential α-synuclein expression contributes to selective vulnerability of hippocampal neuron subpopulations to fibril-induced toxicity. Acta Neuropathol 135:855–875. https://doi.org/10.1007/s00401-018-1829-8

Matos S, Campos D, Pinho R, Silva RM, Mort M, Cooper DN et al (2016) Mining clinical attributes of genomic variants through assisted literature curation in Egas. Database. https://doi.org/10.1093/database/baw096

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D et al (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res 45:D183–D189. https://doi.org/10.1093/nar/gkw1138

Muresan V, Ladescu MZ (2016) Shared molecular mechanisms in Alzheimer's disease and amyotrophic lateral sclerosis: neurofilament-dependent transport of sAPP, FUS, TDP-43 and SOD1, with endoplasmic reticulum-like tubules. Neurodegener Dis 16:55–61

Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S et al (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 9:345–350

Read A, Schröder M (2021) The unfolded protein response: an overview. Biology (Basel). https://doi.org/10.3390/biology10050384

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC et al (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 50:W216–W221. https://doi.org/10.1093/nar/gkac194

Thibaudeau TA, Anderson RT, Smith DM (2018) A common mechanism of proteasome impairment by neurodegenerative disease-associated oligomers. Nat Commun 9:1097. https://doi.org/10.1038/s41467-018-03509-0

Wilson MR, Zoubeidi A (2017) Clusterin as a therapeutic target. Expert Opin Ther Targets 21:201–213. https://doi.org/10.1080/14728222.2017.1267142

Yuste-Checa P, Bracher A, Hartl FU (2022) The chaperone Clusterin in neurodegeneration-friend or foe? Bioessays 44:e2100287. https://doi.org/10.1002/bies.202100287