Journal of Alzheimer's Disease xx (20xx) x–xx
DOI 10.3233/JAD-200206
IOS Press

A Coordinated Approach by Public Domain Bioinformatics Resources to Aid the Fight Against Alzheimer's Disease Through Expert Curation of Key Protein Targets

Lionel Breuza[a], Cecilia N. Arighi[b,c], Ghislaine Argoud-Puy[a], Cristina Casals-Casas[a], Anne Estreicher[a], Maria Livia Famiglietti[a], George Georghiou[d], Arnaud Gos[a], Nadine Gruaz-Gumowski[a], Ursula Hinz[a], Nevila Hyka-Nouspikel[a], Barbara Kramarz[e], Ruth C. Lovering[e], Yvonne Lussi[d], Michele Magrane[d], Patrick Masson[a], Livia Perfetto[d], Sylvain Poux[a], Milagros Rodriguez-Lopez[d], Christian Stoeckert[f], Shyamala Sundaram[a], Li-San Wang[f], Elizabeth Wu[g], Sandra Orchard[d,*] and IMEx Consortium, UniProt Consortium

[a]*Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland*
[b]*Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA*
[c]*Protein Information Resource, University of Delaware, Newark, DE, USA*
[d]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Campus, Hinxton, Cambridge, UK*
[e]*Functional Gene Annotation, Preclinical and Fundamental Science, Institute of Cardiovascular Science, University College London (UCL), London, UK*
[f]*Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[g]*Alzforum, Cambridge, MA, USA*

Accepted 5 June 2020

Abstract.

Background: The analysis and interpretation of data generated from patient-derived clinical samples relies on access to high-quality bioinformatics resources. These are maintained and updated by expert curators extracting knowledge from unstructured biological data described in free-text journal articles and converting this into more structured, computationally-accessible forms. This enables analyses such as functional enrichment of sets of genes/proteins using the Gene Ontology, and makes the searching of data more productive by managing issues such as gene/protein name synonyms, identifier mapping, and data quality.

Objective: To undertake a coordinated annotation update of key public-domain resources to better support Alzheimer's disease research.

Methods: We have systematically identified target proteins critical to disease process, in part by accessing informed input from the clinical research community.

Results: Data from 954 papers has been added to the UniProtKB, Gene Ontology, and the International Molecular Exchange Consortium (IMEx) databases, with 299 human proteins and 279 orthologs updated in UniProtKB. 7,45 binary interactions were added to the IMEx human molecular interaction dataset.

*Correspondence to: Sandra Orchard, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Campus, Hinxton, Cambridge CB10 1SD, UK. E-mail: orchard@ebi.ac.uk. ORCID: 0000-0002-8878-3972

ISSN 1387-2877/20/$35.00

Conclusion: This represents a significant enhancement in the expert curated data pertinent to Alzheimer's disease available in a number of biomedical databases. Relevant protein entries have been updated in UniProtKB and concomitantly in the Gene Ontology. Molecular interaction networks have been significantly extended in the IMEx Consortium dataset and a set of reference protein complexes created. All the resources described are open-source and freely available to the research community and we provide examples of how these data could be exploited by researchers.

Keywords: Alzheimer's disease, Cytoscape network analysis, data curation, database, neurobiology, protein

INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disease characterized by loss of memory, inability to process new information, loss of language function, a disturbed perception of space, inability to do calculations, indifference, depression, delusions, and eventually death. Inheritable AD (familial AD) represents less than 5% of AD cases of which 10–15% have a family history of autosomal dominant inheritance; whereas the more common, sporadic, AD with complex polygenic risk inheritance accounts for more than 90% of cases [1]. Worldwide, at least 50 million people are currently believed to be living with AD or other dementias and this number could exceed 152 million by 2050 (https://www.who.int/news-room/fact-sheets/detail/dementia). The global cost of AD and dementia is estimated to be $605 billion, which is equivalent to 1% of the entire world's gross domestic product. Globally, governments and medical charities spend millions of taxpayer and fundraiser dollars on biomedical research into this condition. It is therefore critical that the data generated by AD research is collated, organized and available in data resources and tools to increase the pace of discovery and innovation.

AD is a complex disease which needs to be studied at many levels, from the molecular mechanisms to the cellular composition and physiology of the brain [2]. Conditions such as vascular damage and neuroinflammation are also believed to play important roles in disease initiation and progression. Our current understanding of the causes, risk factors, and sub-types of these devastating conditions have been reviewed extensively elsewhere (for example, [2–4] and they are not the subject of this manuscript. However, a number of key processes known to play a role in disease etiology and progression are briefly described to showcase the representation of selected proteins in UniProtKB and demonstrate how users can access information about both physiological and pathological aspects of the molecules.

Central to AD disease pathology are two processes: the extracellular formation of senile plaques in the grey matter of the brain which are primarily composed of amyloid-β precursor protein (APP)-derived amyloid-β (Aβ) [5, 6], and intracellular accumulation of hyperphosphorylated tau/Microtubule-associated protein tau (MAPT) protein to form neurofibrillary tangles [7, 8]. Aβ oligomers are believed to contribute to cell death by interfering with neuron-to-neuron communication at synapses [9] and restricting the source of oxygen and nutrients [10], while tau tangles block the transport of nutrients and other essential molecules inside neurons [11]. Whilst the relationship between Aβ and tau in AD is not fully understood, abnormal species of tau protein are believed to spread in a 'prion-like' manner between cells and its uptake may be potentiated by extracellular Aβ [12, 13]. Aβ peptides can be cleared intracellularly by microglia and other cell types [14–16], by transcytosis across the blood-brain barrier [17, 18], or by Aβ degrading enzymes, such as insulin-degrading enzyme (IDE) and neprilysin (MME) [19, 20]. Tau has been shown to be degraded via the ubiquitin-proteasome system as well as the autophagy lysosome system [21]. Disorders in clearance of Aβ and tau play a key role in the development of neurodegenerative disorders such as AD while overloading of the microglial system results in chronic inflammation [22, 23]. However, evidence has been emerging that aggregation of Aβ and tau may not be the underlying causes of disease, but may be the outcome of perturbations in cellular homeostasis in the brain, occurring years to decades prior to disease onset [2, 24]. Normal brain function may be compromised by the decreased ability of the brain to metabolize glucose and aberrant lipid metabolism, such as sluggish cholesterol transport [25]. To date, over 350 human proteins have been associated with the development of AD as researchers move toward an understanding of the underlying cellular mechanisms that drive the formation of the protein aggregates and the downstream effect these have on the brain.

The analysis and interpretation of data generated from increasing large-scale examination of patient-derived clinical samples relies on access to high-quality bioinformatics resources. The scientific content of these resources is maintained and updated by professional biocurators who extract knowledge from unstructured biological data described in free-text journal articles and convert it into both more easily digestible, high-level summaries and a structured, computable form. The latter both enables large-scale data analyses, for example functional enrichment of sets of genes/proteins using the Gene Ontology (GO) [26, 27], and also helps to make the searching of data more productive by managing issues such as the problems caused by gene/protein name synonyms, identifier mapping, and minimizing the effect of poor quality, redundant, or misleading data. The work of these data resources helps researchers overcome known bottlenecks in data analysis, namely the time spent in discovering and collating required information, manually verifying it, and integrating it into analysis pipelines [28]. We here describe a coordinated approach to updating key public domain resources with the aim of supporting AD research, starting with the update of genes/proteins with a known role in AD biology. Accessing informed input from the clinical research community was an essential part of this process and was critical in defining where curation effort was focused. We also illustrate the way this coordinated update can be used by researchers to answer questions pertaining to the complex etiology of AD.

METHODS AND MATERIALS

Identifying disease-related proteins

A recent initiative by the UniProt Knowledgebase of protein sequences and annotations [29] to update the proteins which play a role in the initiation and development of AD, coordinated with the curation of their interactions and the complexes they form, has been funded by the NIH National Institute on Aging (NIA). At the start of this annotation project, curators were faced with two main problems—an accurate description of the various forms of AD and identification and prioritization of the proteins associated with the disease. AD is generally classified into early and late-onset forms, with genetic variants or risk alleles [30] associated with each condition providing a further sub-classification. In order to identify key AD-related proteins appropriate for update and reannotation, UniProt curators reached out to members of the AD clinical and research communities, leveraging contacts made through the NIH NIA programs and a collaboration with the Alzheimer's Research UK (ARUK) funded GO project at University College London (UCL) [31, 32]. Workshops were organized to help database providers understand how their resources are used by the research community, and conversely for the research community to directly input into the curation process. Attendees were asked to identify proteins which played a key role in the disease, or which had been associated with disease even if a clear molecular mechanism explaining this association had yet to be identified. Additional candidates were provided by Alzforum (http://www.alzforum.org), the Agora portal (https://agora.ampadportal.org), collected from targeted research groups, and from literature searching. The main pathway resource consulted was WikiPathways which provided an overview of the disease process (http://www.wikipathways.org/index.php/Pathway:WP2059). Drug target resources included the ChEMBL database [33] and the OpenTargets platform [34], taking only high scoring (0.8 to 1) targets associated with AD from the latter. To build the AD-centric protein-protein interaction network, data was downloaded from the IntAct molecular interaction database [35], limited to interactors with an MIscore of >0.45 (see explanation below). Proteins were prioritized for curation following a ranking system, i.e., 1) proteins known to play a functional role in AD pathways and known drug targets for AD, 2) proteins known to have an association to AD, e.g., through a genome wide association study (GWAS) study but for which a molecular mechanism has yet to be identified, and 3) proteins that physically interact with those defined in (1) or (2). A copy of this list, as of UniProt release 2019_10 is available as Supplementary Table 1.

Protein annotation

Data from selected publications were transferred into the UniProtKB, GO, IntAct molecular interaction, and the Complex Portal databases, as appropriate, as previously described [26, 32, 35–37].

Producing an AD-centric molecular interaction network

Seed proteins were identified by searching the UniProt website (Release 2019_08) for reviewed

entries containing the keyword 'Alzheimer disease'. (keyword: "Alzheimer disease [KW-0026]" AND reviewed: yes). As this keyword is only added to human entries, there was no need to further restrict the search by species. The final list is available in Supplementary Table 2.

Interactors of this list of proteins were obtained from IntAct using the PSICQUIC client app in Cytoscape Version: 3.7.1 [38]. To return an isoform- and post-processed chain- specific network the following query was used: (id:P37840* OR id:P49810* OR id:P49768* OR id:O14672* OR id:P03886* OR id:Q8IZY2* OR id:Q16643* OR id:P02649* OR id:P05067* OR id:Q92673* OR id:P03891* OR id:O95185*) AND annot: "imex curation".

This gave a raw network containing 1461 nodes and 2671 edges.

The network was then filtered to: a) remove non-human interactors; b) remove duplicated interactions; c) select interactions having MIscore > 0.45.

A MIscore of >0.45 can only be achieved by interacting pairs having at least a single interaction evidence showing that the two molecules directly interact or two or more evidences of a physical interaction. The filtered isoform- and post-processed chain-specific Network contains 152 nodes and 277 edges.

To enable users to access a detailed view of this network, a copy has been deposited at the NDEx data repository (http://www.ndexbio.org/#/network/49e43d68-939b-11ea-aaef-0ac135e8bacf) [39]. Users may alternatively download an updated set of the data used to derive an AD-focused interaction network by pasting the query annot: "dataset:Alzheimers" into the IntAct website (http://www.ebi.ac.uk/intact). Using the Advanced Search capabilities will enable further filtering of the results of this query.

To perform the ClueGO functional enrichment analysis, all isoform and post-processed chains were collapsed to the canonical identifiers in UniProtKB, all leaves (proteins not directly connected in the network) were removed and the complexes were demerged into protein subunits. The Cytoscape APP ClueGO version 2.5.0 [40] was then used, implementing the following parameters:

Organism analyzed: Homo Sapiens [9606].
Identifier types used: UniProtKB
#Genes in custom reference set: 3001 human proteins extracted from UniProt having tissue-specificity='brain'
Ontology used: GO_BiologicalProcess-EBI-QuickGO-GOA_20.11.2017_00h00 and REACTOME_Reactions_20.11.2017
Evidence codes used: All
Statistical Test Used = Enrichment (Right-sided hypergeometric test), Correction Method Used = Bonferroni step down
Min GO Level = 8
Max GO Level = 20
GO Fusion = true
GO Group = true
Kappa Score Threshold = 0.4
Over View Term = SmallestPValue
Group By Kappa Statistics = true
Initial Group Size = 1
Sharing Group Percentage = 60.0

RESULTS

All known human protein-coding genes have been curated by experts within the UniProtKB database (http://www.uniprot.org) with, as far as possible, all the protein products encoded by one gene described in a single reviewed entry [29, 36]. Each entry groups all the protein isoforms expressed by that gene, with positional features such as binding domains, post-translational modifications and amino-acid variants mapped to a representative sequence. Isoforms yet to be integrated are maintained in unreviewed entries but are accessible as part of the complete human proteome reference set (UniProt Proteome UP000005640) and can also be viewed in the corresponding reviewed entry on the website as a result of an automatic gene-centric mapping. Expert curators summarize knowledge extracted from biomedical literature in sections describing different aspects of protein biology relevant to those gene products, these can include function, enzymatic activity, subcellular location, and links to disease conditions. For example, over the period of this annotation project PSEN1 (UniProtKB P49768) had data from 43 publications added to its entry in UniProtKB, enhancing the 'Function' section, and including details of the functional roles played by specific domains within the protein. Information on disease linked variants and the effects of point mutations on protein behavior were also added.

Proteins do not operate in isolation and details of their interactions with other molecules are manually curated by the IMEx Consortium of interaction databases (http://www.imexconsortium.org) [41] via

the IntAct database [35], from where a filtered subset of high confidence binary protein-protein interactions is imported back into the 'Interaction' section of the corresponding UniProtKB entries. Proteins also form higher-order, functional assemblies and descriptions of stable protein complexes are curated into the Complex Portal (http://www.ebi.ac.uk/complexportal), giving details of complex content, stoichiometry, and topology in addition to function and 3D structure, when available [37]. Again, these data can be accessed from the appropriate UniProtKB records. In parallel, biocurators link these proteins and protein complexes to appropriate GO terms describing their biological function, the cellular processes in which they play a role, and the cellular compartment in which they are found. The GO is a biomedical ontology which describes these aspects of protein behavior in a consistent and computer-accessible manner [26, 27]. Linking gene products to GO terms means that researchers can use the resulting annotations to interpret high-throughput datasets using GO term enrichment.

The NIA-funded annotation project resulted in data from 954 papers being added to the UniProtKB, GO, and IMEx databases, with 299 human proteins and 279 orthologues updated in UniProtKB. 7,045 binary interactions were added to the IMEx human molecular interaction dataset.

Understanding the function of AD-associated proteins

UniProt curators provide high-quality literature sourced annotations for experimentally characterized proteins across diverse protein families. These data are presented both in free text fields and in structured mappings to the underlying protein sequence to enable users to understand how, for example, a post-translational modification to a specific residue can drive a change in protein behavior. The proteins identified by AD domain experts were subjected to an intense literature review and corresponding update of the relevant annotation fields in order to help researchers understand both the physiological role these entities play in a cell, and how this relates to the pathological disease condition. As described above, this includes a full review of both protein isoforms and protein chains formed by post-translational processing of the full-length gene product. This is particularly important in the case of AD-related proteins as amyloid plaque formation is a consequence of disregulated protein cleavage [42]. APP (UniProtKB P05067) is a ubiquitously expressed type I transmembrane protein which functions as a cell surface receptor with roles in neurite growth, neuronal adhesion, and axonogenesis. The protein consists of a large ectodomain, a single membrane spanning domain and a short cytoplasmic tail. The ectodomain comprises two highly conserved E1 and E2 domains, involved in metal (copper and zinc) and heparin binding. APP undergoes extensive post-translational modification and proteolytic processing to generate peptide fragments. The cleavage products of APP are all described at the residue level in the UniProtKB database, with stable identifiers allowing unambiguous recognition of each proteoform when described (Fig. 1).

As detailed in the appropriate UniProtKB records, APP processing is initiated either by α-secretase/ADAM10 (UniProtKB O14672) cleavage within the Aβ region, or by β-secretase (BACE1/2, UniProtKB P56817/Q9Y5Z0) cleavage at the N-terminus of Aβ, leading to the secretion of large soluble ectodomains, termed soluble APPα (APPsα, UniProtKB PRO_0000000089) and soluble APPβ (APPsβ, UniProtKB PRO_0000000090), respectively. Subsequent processing of the C-terminal fragments by the γ-secretase complex (Complex Portal:CPX-2176/CPX-4231/CPX-4232/CPX-4233), as well as processing along non-canonical pathways, results in numerous fragments, which have different and partially opposite functional properties. During amyloidogenic processing, APP is sequentially cleaved by β- and γ-secretases to mainly generate $A\beta_{40}$ (UniProtKB PRO_0000000093), and $A\beta_{42}$ (UniProtKB PRO_0000000092) fragments.

Many of the AD-associated proteins prioritized for update (Supplementary Table 1) are enzymes, which may be responsible for the proteolytic processing of longer protein chains as described above, catalysis of metabolic reactions, or generation/removal of post-translational modification sites. Enzymatic function is now described in UniProtKB using Rhea (http://www.rhea-db.org), a comprehensive and non-redundant resource of expert-curated biochemical reactions [43], as a vocabulary to annotate and represent enzyme-catalyzed reactions. Rhea uses the ChEBI (Chemical Entities of Biological Interest) ontology to describe reaction participants, their chemical structures, and chemical transformations [44]. Additional small molecule interactions, such as cofactor binding sites are also described within UniProt using ChEBI. Sophisticated searches within UniProtKB now allow the researcher to

PTM / Processing[i]

Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Signal peptide[i]	1 – 17	3 Publications	Add BLAST		17
Chain[i] (PRO_0000000088)	18 – 770	Amyloid-beta precursor protein	Add BLAST		753
Chain[i] (PRO_0000000089)	18 – 687	Soluble APP-alpha	Add BLAST		670
Chain[i] (PRO_0000000090)	18 – 671	Soluble APP-beta	Add BLAST		654
Chain[i] (PRO_0000381966)	18 – 286	N-APP	Add BLAST		269
Chain[i] (PRO_0000000091)	672 – 770	C99	Add BLAST		99
Chain[i] (PRO_0000000092)	672 – 713	Amyloid-beta protein 42	Add BLAST		42
Chain[i] (PRO_0000000093)	672 – 711	Amyloid-beta protein 40	Add BLAST		40
Chain[i] (PRO_0000000094)	688 – 770	C83	Add BLAST		83
Peptide[i] (PRO_0000000095)	688 – 713	P3(42)	Add BLAST		26
Peptide[i] (PRO_0000000096)	688 – 711	P3(40)	Add BLAST		24
Chain[i] (PRO_0000384574)	691 – 770	C80	Add BLAST		80
Chain[i] (PRO_0000000097)	712 – 770	Gamma-secretase C-terminal fragment 59	Add BLAST		59
Chain[i] (PRO_0000000098)	714 – 770	Gamma-secretase C-terminal fragment 57	Add BLAST		57
Chain[i] (PRO_0000000099)	721 – 770	Gamma-secretase C-terminal fragment 50 By similarity	Add BLAST		50
Chain[i] (PRO_0000000100)	740 – 770	C31	Add BLAST		31

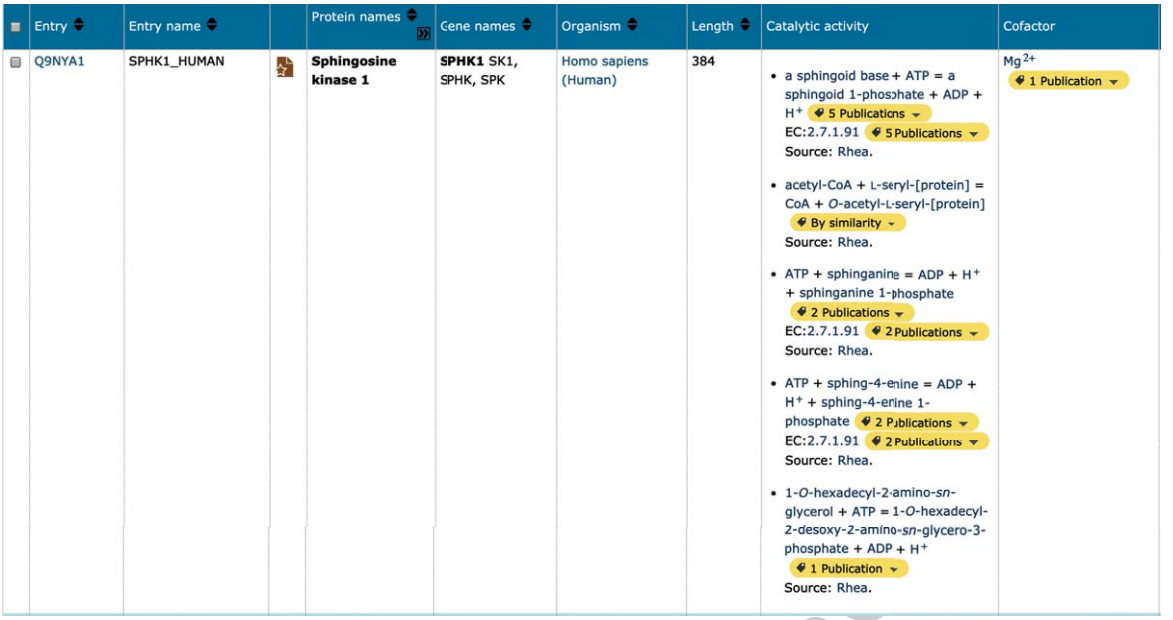
Fig. 1. Screenshot showing the UniProtKB description of the products of amyloid-beta precursor protein post-transcriptional modifications and processing. This information is available in the UniProtKB P05067 entry for amyloid-beta precursor protein (APP).

identify metabolic networks and predict new pathways for drug production. For example, alterations in sphingolipid metabolism have been detected in AD, with levels of SPHK1 (UniProtKB Q9NYA1) downregulated and, conversely, levels of SPHK2 (UniProtKBQ9NRA0) upregulated [45]. Both entries for these proteins have been updated in UniProtKB, where it is now possible to visualize the chemical reaction, balanced for mass and charge (at an arbitrary pH of 7.3) as described by Rhea, and cofactors linked to the corresponding entry in ChEBI (Fig. 2).

Tau/MAPT (UniProtKB P10636) is a microtubule-associated protein predominantly expressed in the axons of neurons [46]. Tau is a naturally unfolded protein with an extended structure; however, in AD brains, tau is accumulated in a hyperphosphorylated state in a unique filamentous structure, paired helical filaments of 10 nm diameter with 80 nm periodicity [47]. The phosphorylation of tau regulates both its functional ability to assemble and stabilize microtubules and also its pathological structure [48], and the 441 amino acid isoform of tau (UniProtKB P10636-8) has 45 serine, 35 threonine, and 5 tyrosine residues, resulting in a total of 85 potential phosphorylation sites [49]. CDK5 (UniProtKB Q00535) is one enzyme known to play a role in the phosphorylation of tau [50], priming tau for further phosphorylation events by the hierarchical kinase GSK3B (UniProtKB P49841) by modifying an upstream +4 (or +3) site, (S/T)xx(x)p(S/T). Again, this chemical reaction has been updated in UniProt (Fig. 2B), where it is also possible to identify the resulting phosphorylated residues in the corresponding entry for tau. CDK5 is activated by p35/CDK5R1 (Q15078), the resulting complex (Complex Portal:CPX-2201) then being recruited to membranes via the N-terminal p35 myristoylation site (Fig. 2C) [51]. p35/CDK5R1 is a protein with a short-life span which is cleaved by calpain (Complex Portal:CPX-2674/CPX-4302) into a p25 C-terminal fragment (UniProtKB PRO_0000004795) when neurons suffer from stress or encounter death signals. p25/CDK5R1 has a longer half-life and this complex (Complex Portal:CPX-3142) dissociates from the plasma membrane into the nucleus, where it can phosphorylate additional proteins [52].

Linking amino acid variation to functional consequence

AD-causing mutations in APP (UniProtKB P05067), PSEN1 (UniProtKB P49768), and PSEN2 (UniProtKB P49810) affect the generation of Aβ peptides, changing the relative ratio of $A\beta_{42}$ to $A\beta_{40}$ peptide [53]. The longer $A\beta_{42}$ peptides seem to be more prone to aggregation, and increased ratios of $A\beta_{42}/A\beta_{40}$ are thought to play a role in AD pathogenesis. It is therefore important to document all APP, PSEN1, and PSEN2 variants that lead to a change in this ratio. About 1% of AD cases develop as a result of mutations within APP or the genes encoding the PSEN1 and PSEN2 proteins



Entry	Entry name	Protein names	Gene names	Organism	Length	Catalytic activity	Cofactor
Q9NYA1	SPHK1_HUMAN	**Sphingosine kinase 1**	**SPHK1** SK1, SPHK, SPK	Homo sapiens (Human)	384	• a sphingoid base + ATP = a sphingoid 1-phosphate + ADP + H^+ 5 Publications EC:2.7.1.91 5 Publications Source: Rhea. • acetyl-CoA + L-seryl-[protein] = CoA + *O*-acetyl-L-seryl-[protein] By similarity Source: Rhea. • ATP + sphinganine = ADP + H^+ + sphinganine 1-phosphate 2 Publications EC:2.7.1.91 2 Publications Source: Rhea. • ATP + sphing-4-enine = ADP + H^+ + sphing-4-enine 1-phosphate 2 Publications EC:2.7.1.91 2 Publications Source: Rhea. • 1-*O*-hexadecyl-2-amino-*sn*-glycerol + ATP = 1-*O*-hexadecyl-2-desoxy-2-amino-*sn*-glycero-3-phosphate + ADP + H^+ 1 Publication Source: Rhea.	Mg^{2+} 1 Publication

Fig. 2. Screenshot showing the results of a UniProtKB search for human Sphingosine kinase 1. The UniProtKB database was queried for the term 'SPHK1'. The top hit (human) in the results table is displayed. It is possible to customize this view to select additional data fields from the UniProt record, in this case the column options 'cofactor' and 'catalytic activity' were added to the results table.

present in the γ-secretase complex; however, those inheriting a known AD-associated APP or PSEN1 variant will develop the disease, whereas a slightly lower risk (95%) is associated with inheriting a known AD variant in PSEN2 [54]. Individuals with AD mutations in any of these three genes tend to develop early-onset disease, with symptoms developing before the age of 65, sometimes as early as age 30. Understanding how a genetic variation changes protein function or expression levels is essential for our understanding of genetic disease and the ability to identify those variants which are causal. UniProtKB curators capture nonsynonymous variants described in the literature with, when available, detail on the phenotypic or pathogenic consequences on the amino acid change. UniProt also receives input (publications and suggested annotations) from expert groups, e.g., Alzforum, who collects detailed variant information about AD proteins from the literature. To date, UniProtKB records contain information on over 30,000 variants linked to Mendelian diseases in more than 13,000 human protein sequence records [55] and work is ongoing to standardize variant interpretations through the incorporation of American College of Medical Genetics and Genomics (ACMG) guidelines and the ClinGen pathogenicity calculator into the curation workflow. Cross references to variant resources such as dbSNP (http://www.ncbi.nlm.nih.gov/snp/) and Ensembl (http://www.ensembl.org), and disease-specific databases such as NIAGADs (https://www.niagads.org/) are added. Additional variant data is imported from large-scale studies such as 1000 Genomes and ExAC, and again mapped to the protein sequence and made available via the Proteins API (http://www.ebi.ac.uk/proteins/api/doc/).

UniProtKB acts as an integrative layer, enabling users to align genomic variants with enzyme active sites, modified residues, the phenotypic consequence of site-directed mutagenesis and binding domains mapped to the residue level. An exact mapping of the Ensembl translation to a UniProtKB sequence enables the calculation of UniProtKB positional annotations to their genomic coordinates and these mappings are continually reviewed and updated by both UniProt and Ensembl curation teams [56]. Thirty-four different positional annotation types are currently aligned with the genome sequence. An additional 17,371 mutations which map to the genome have been supplied by the IMEx Consortium which captures the effects of point mutations on molecular interactions, using controlled vocabulary terms to describe whether these increase, disrupt, or cause an interaction to occur [57]. Again, these site-directed

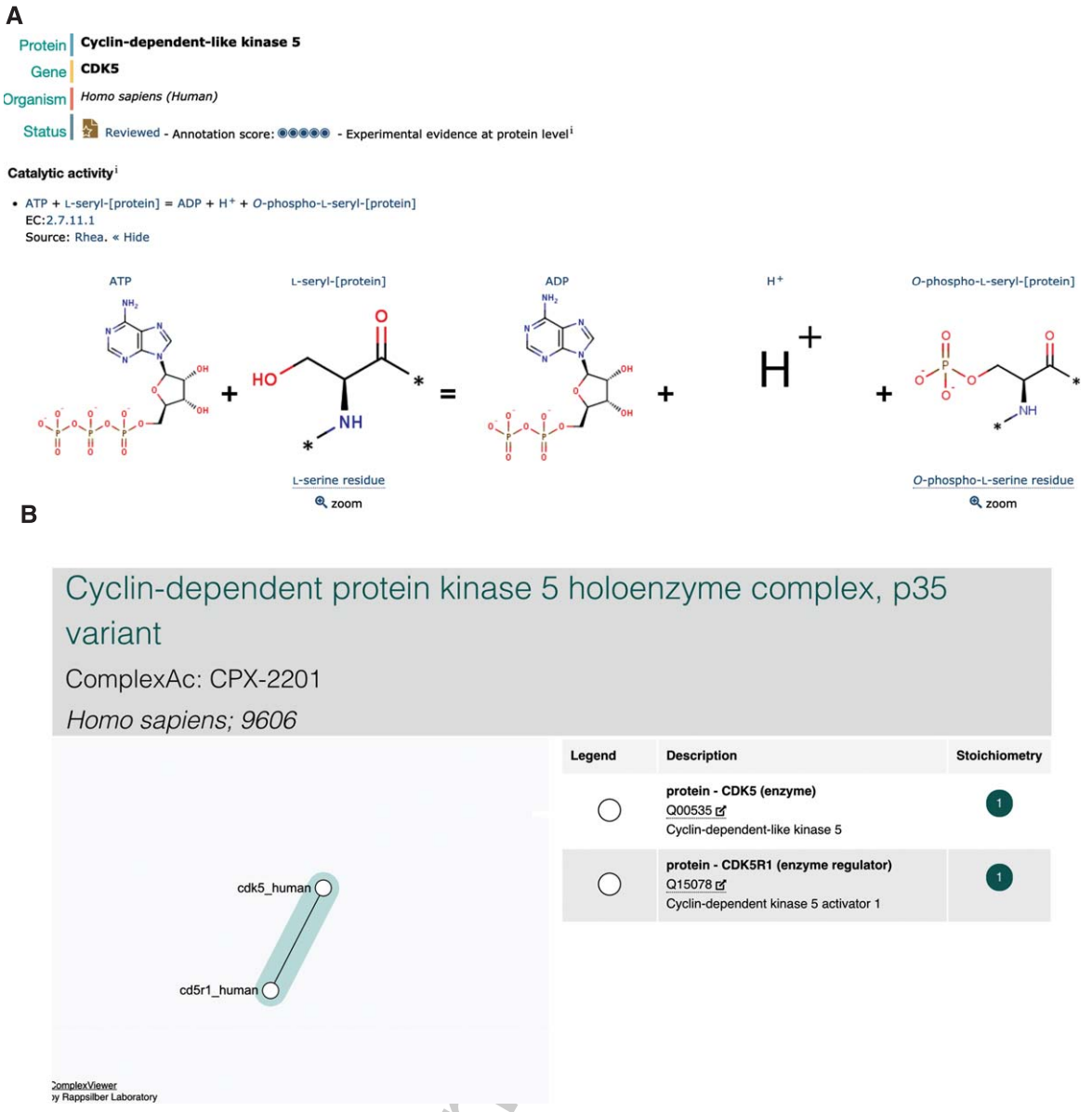


Fig. 3. Representation of CDK5 in UniProtKB and the Complex Portal. A) The representation of CDK5 catalytic activity by Rhea within the UniProtKB entry. B) The Complex Portal display of the CDK5-p35/CDK5R1 complex which can be found by searching for either of the proteins, or by complex name.

mutations have been mapped to the underlying UniProt protein sequence and can be used to understand the effect a genomic variant may have on a local protein interaction network. Further to this, in collaboration with PDBe through the Structure Integration with Function, Taxonomy, and Sequences resource (SIFTS; http://pdbe.org/sifts/), UniProtKB maps between protein structure and protein sequence, so that a knowledge of protein conformation can contribute to an understanding of protein function [58]. These data are all displayed in UniProtKB using the ProtVista visualization tool [59] which allows the graphical alignment of sequence feature data to the linear protein sequence and from there to the 3D structure (Fig. 4).

Late-onset AD is observed in >90% of patients, and the APOE (UniProtKB P02649) allele E4 is strongly associated with these cases. APOE is a plasma

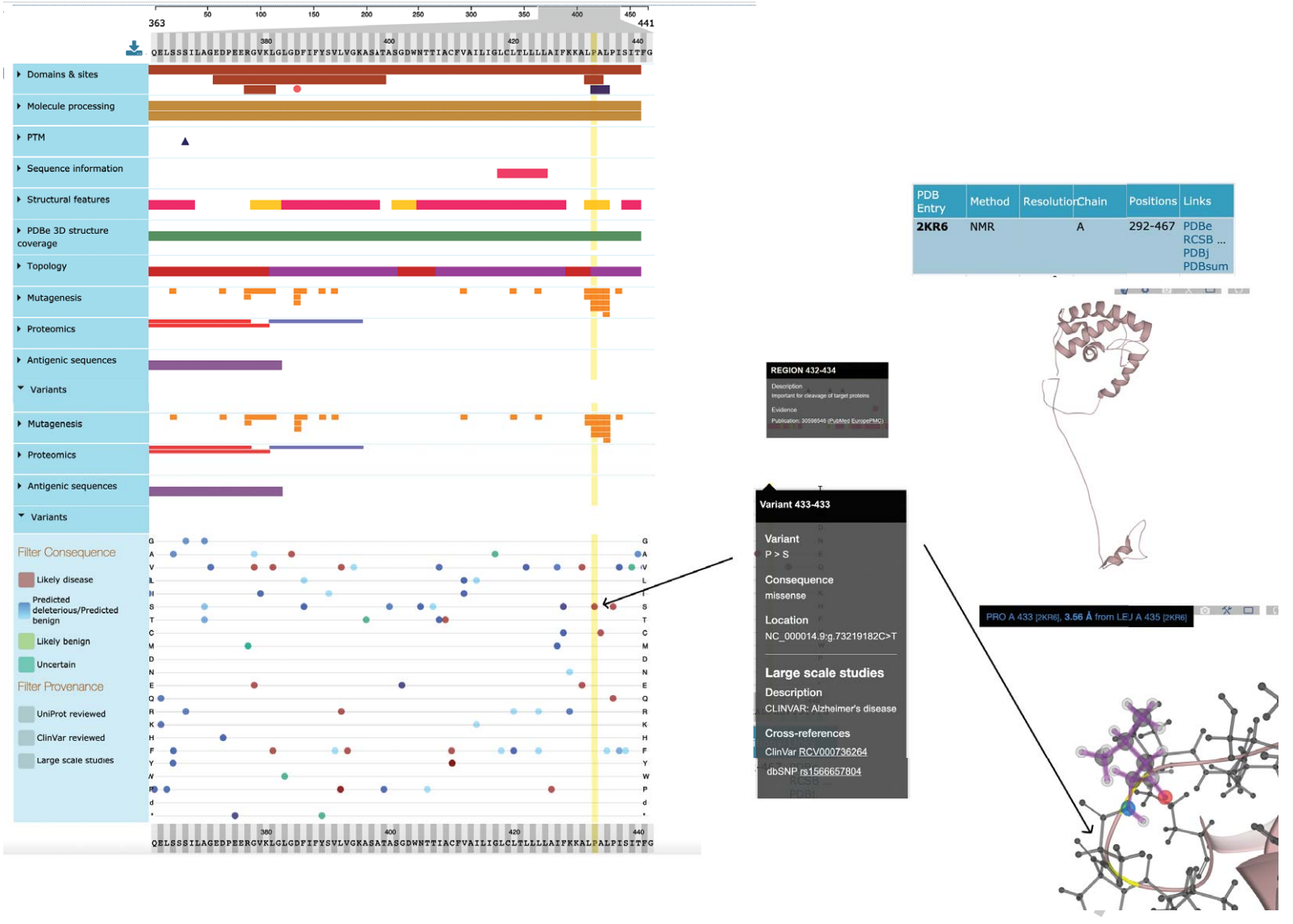


Fig. 4. Simplified view of ProtVista for Human PSEN1 (UniProtKB P49768). To investigate the effect of a specific variant (p.Pro433Ser) of human PSEN1 protein, the user can look at its potential effect on active sites and domain. Clicking on the variant at position 433 shows it to be positioned in the PAL domain, required for normal active site conformation and also in a region important for cleavage of this protein. The position of this variant is also highlighted in the NMR structure of this protein.

lipoprotein which transports lipids between cells and tissues. Abnormal cholesterol metabolism associated with allele E4 is believed to mediate cell type-specific AD pathology, including Aβ upregulation and impaired synaptic function in neurons, reduced synapse elimination activity in astrocytes, impaired remyelination in oligodendrocytes, and Aβ accumulation and inflammatory response in microglia [60]. The most common allele in the human population, and that present on the reference genome GRCh38, APOE*3 is the displayed sequence in the UniProtKB entry, with all three possible alleles fully described in the Polymorphism section of the entry. Sequence variants, single amino acid polymorphisms, and other sequence annotations, have then been described relative to that allele with the alignment of the APOE sequence to the reference genome then allowing the integration of genomic and protein data. In the recent curation project, APOE had information from 40 new references added to the entry.

Enabling functional Insights Into large-scale AD datasets through network analysis

AD is not a single disease but a number of separately-triggered conditions [61] which share the same pathological phenotype, suggesting that these conditions may have many downstream processes in common. Understanding how proteins associated with AD are linked in the interacting network of molecules that drive cellular processes may help to identify proteins which are critical for initiating or driving the disease condition as potential therapeutic targets. Network-based analysis is a powerful technique for extracting biological insights from large datasets, enabling researchers to identify clusters of interacting molecules which participate in the same biological process or are members of the same physical complex. Protein interaction networks can help researchers understand the interconnectivity of both intra- and extracellular signaling, while studying network topology can give information about biological function and properties of the component molecules. Merging external 'omics data, such as transcriptomics, proteomics, and genome-wide association (GWA) studies, with the network can indicate tightly associated nodes of co-regulated proteins. An understanding of the processes associated with these networks can be further investigated by using GO annotations or Complex Portal data.

The IMEx Consortium curates to a detailed curation model, i.e., all aspects of an interaction experiment, including host organism, interaction detection, and participant identification methodologies and full details of the constructs, including binding domains and the effects of site-directed mutations, are captured [41, 57]. All this information is accurately mapped to controlled vocabulary terms, in particular those described by the HUPO PSI-MI CV. Interactions are not limited to protein:protein but increasingly also include protein-small molecule, protein-protein complex, protein-ncRNA, and protein-gene interactions using identifiers from ChEBI, Complex Portal, RNAcentral (http://www.rnacentral.org), and Ensembl, respectively, to identify the respective entities. This enables the IMEx databases to fully capture the differences in interacting molecules observed with different APP isoforms (UniProtKB P05067-4/P05067-8 IntAct:EBI-21132406/EBI-21132308) [62] or by monomeric (UniProtKB PRO_0000000092) versus oligomeric (Complex Portal CPX-1134) $A\beta_{42}$ (IntAct:EBI-20818781/EBI-20821761) [63]. The effects of mutagens, site directed to mimic known variants can also be described, for example the interactome of MAPT/Tau (UniProtKB P10636) p.Pro618Leu variant (dbSNP:rs63751273) with a known link to frontotemporal dementia [64], which reduces the ability of MAPT/Tau to promote microtubule assembly and accelerates aggregation of tau into filaments has been compared to that of the wild-type protein (IntAct:EBI-20800792/EBI-20799058) [65]. Data on the effect of site-directed mutations on molecular interactions is available as a downloadable file from the IntAct website (ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/mutations.tsv) and is also exported to the UniProtKB ProtVista viewer to provide additional understanding of how a particular amino acid variant may affect protein function.

Any protein interaction network built using current data will at best be partial, as we are far from having achieved full coverage of the human interactome. However, a more immediate concern is the quality of the networks being used for analysis, which are currently often created by combining data from many resources with little attention to the source(s) of the binary interactions and the methodology by which they were generated. The detailed curation model of the IMEx curation enables data filtering on many levels and thus enables the building of high-quality networks. The addition of AD-relevant protein interactions as a part of the curation marathon described above has enriched the interactome of AD-related proteins by several thousand binary interactions and

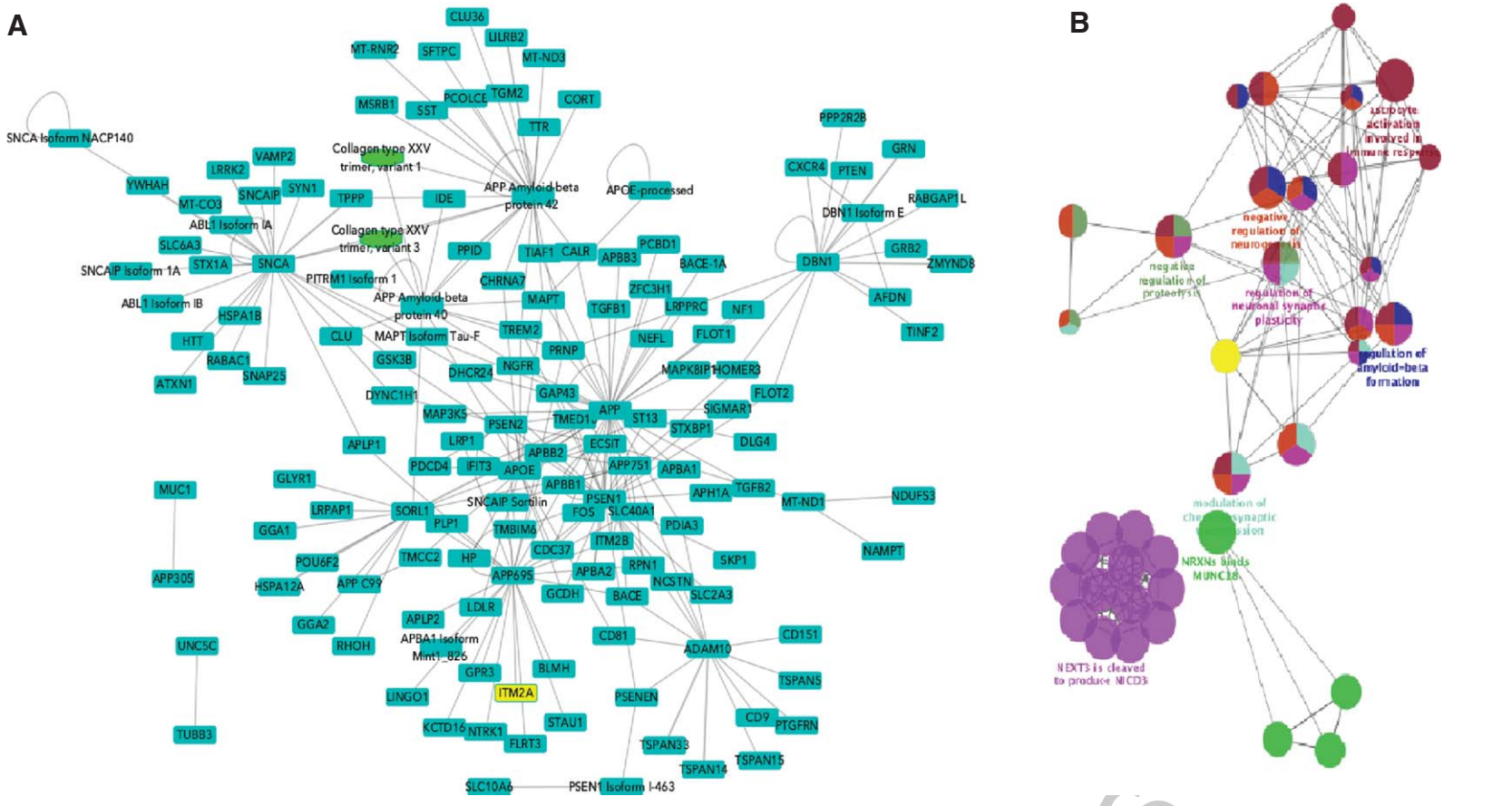


Fig. 5. Networks built from Alzheimer's disease relevant proteins. A) High-confidence network with isoforms and post-processed chains acting as distinct nodes. Seed proteins are those to which the Alzheimer Disease keyword has been added in UniProtKB. Blue squares represent proteins, green ovals represent protein complexes. Nodes have been collapsed to canonical sequence/gene level. B) ClueGO functional enrichment analysis of network shown in B.

is a significant addition to previous work by the IMEx curators in building the APP interactome [66]. To demonstrate the utility of these data for AD researchers, high confidence interaction networks could be built using both protein interactors described at the isoform/post-processed chain level and also following the collapse of this level of detail to the consensus sequence selected (Fig. 5A). In both cases, the seed proteins were those to which the Alzheimer Disease keyword has been added in UniProtKB. The raw network contained 1,461 nodes and 2,671 edges. This was then filtered by MI score > 0.45 [67] to produce a high-confidence sub-network, restricted to human-only interactions and redundant interaction evidences were merged, reducing this to 152 nodes and 277 edges. Remapping isoforms and post-processed chains to the canonical sequence level further reduced this to 136 nodes and 179 edges. This final network was analyzed using ClueGO, a Cytoscape App that visualizes non-redundant biological GO terms for large clusters of genes in a functionally grouped network (Fig. 5B). In this case, the network was filtered for 'Biological Process' term enrichment. Terms such as 'regulation of amyloid-beta formation', 'regulation of synaptic plasticity', 'astrocyte activation' (linked to AD pathology [68]), and child terms of Notch1 signaling (known to be altered in AD [69]) were overexpressed in comparison to a full list of human brain proteins, suggesting that this is a biologically relevant network. This network, and subsequent ongoing expansions to the dataset, is now freely available to the research community to enable network analysis of generated data and can easily be extended to encompass, for example, all the proteins known to be expressed in the human brain by performing the relevant queries on the IntAct website. This resource will facilitate interrogation of large-scale GWAS, transcriptome and proteomics clinical datasets, and allow users to explore novel biology and enhance our understanding of the disease process [70].

The Reactome database of curated biological pathways provides a tool for visualizing user-supplied expression data as an overlay on manually curated pathway diagrams [71]. Pathways are authored by biologists who are recruited for their expertise in the area, in this case biocurators involved with the curation of AD-associated papers in UniProtKB. As a result of this curation marathon, a number of AD-related pathways are in the process of being created and will be available to researchers as another tool enabling large-scale 'omics analysis. Reactome pathways can be further extended by adding IMEx quality filtered protein interactions to extend out the networks and these additional molecules can be included in subsequent representation analysis, a statistical (hypergeometric distribution) test that determines whether certain Reactome pathways are over-represented (enriched) in any submitted dataset.

Further enhancing the Gene Ontology to improve interpretation of AD data

UniProtKB biocurators are the single largest contributing group to GO manual annotations, both as a whole but in particular for the annotation of human proteins. The recent focus on AD proteins has added to work by the UCL Functional Gene Annotation group, funded by ARUK, to associate GO terms to proteins, protein complexes, and microRNAs relevant to processes involving amyloid-beta and tau, concomitantly creating many new GO terms in the process to further enrich those branches of the ontology relevant to neuronal biology. As a proof-of-concept of the benefit of a focused annotation effort, a functional analysis was performed by Kramarz et al. in November 2018 [31] on a hippocampal proteomic dataset, identifying proteins that were differentially expressed in AD versus age-matched controls. Analyzing the data against the GO in 2018 versus an earlier version archived in 2016 showed an almost doubling of enriched GO terms and highlighted new processes with a potential role in AD, for example 23% of dysregulated hippocampal proteins now showed a contribution to a heightened immune response. The work on curating proteins and protein complexes to GO terms is being continued by the UniProt, Complex Portal, and UCL annotation teams, while the UCL team are additionally contributing GO annotation of microRNAs regulating the expression of microglial AD relevant proteins [32].

One advantage of the improved GO representation of processes related to AD, is that it can be used as a tool to search for lists of proteins relevant to a particular aspect of the disease. It is now widely acknowledged that neuroinflammation plays a key role in the pathogenesis of AD, for example through the elevation of amyloidogenesis. The list of proteins involved in any inflammatory response is long, but searching the UniProt or QuickGO (http://www.ebi.ac.uk/QuickGO) websites for proteins annotated to the GO term "neuroinflammatory response" (GO:0150076) and limiting the search to human proteins, retrieves a list of 42 reviewed

protein entries (GOA release 2020-04-22), which may be connected to the disease process. The protein list can be downloaded from the QuickGO website in CSV format, along with all the GO annotations and publications from which the evidence was extracted.

DISCUSSION

AD is a progressive brain disorder that damages and destroys brain cells, leading to loss of memory, disregulated brain function, and eventually death. In addition to the profound human suffering caused by the condition, AD and other dementias are creating an enormous pressure on both health care systems and national budgets. To understand the molecular mechanisms both triggering and subsequently driving the development of AD, researchers have designed numerous high-throughput transcriptomic, proteomic, metabolomic, and GWA studies generating vast amounts of data. The subsequent analyses and interpretation of the results from such experiments is completely dependent on functional annotation data provided by bioinformatic resources. Resources such as UniProt, the GO, and the IMEx molecular interaction networks enable researchers to take lists of genes/proteins identified in large-scale 'Omics experiments and, for example, find clusters of co-regulated genes which may represent processes or protein complex members involved in a particular process or pathway.

The content of these core data resources is dependent on the work of skilled biocurators, reading and evaluating the scientific literature and transferring key facts to the appropriate entries. Expert manual curation is undeniably expensive, but is essential to make this information readily available to the researcher, the clinician, and to the computational biologist. By working collaboratively, contributing data to multiple specialist resources and working together to develop shared curation tools [72], the biocuration community is taking a lead in giving funders the best possible return on their investment [28]. The AD focused biocuration project described here has benefitted from governmental funding, charitable funding, from pharmaceutical company funding through a public-private partnership [32, 34] and also from previously funded work into other neurological conditions [73, 74]. While in this case the shared funding pool was serendipitous, it suggests that actively managed collaborations between funding bodies could be at least equally successful in increasing both the quantity and quality of information freely available in biomedical databases. As a result of these efforts, researchers can now access 299 disease-relevant human protein records updated in UniProtKB (as of release 2019_10), with experimental GO annotation also added, where possible. An additional 7045 binary molecular interactions have been added to the IMEx dataset, significantly increasing the abilities of researchers to perform network analysis on large-scale datasets.

Once the data is in these resources, it is also the responsibility of database managers to ensure that users can find and access it as easily as possible. The UniProt Consortium is already working to release a disease-specific entry point to those proteins of interest which will enable researchers to navigate the network of molecules that play a role in this condition and easily find information on the function of each. An AD portal will be the first of these released. The data is also being made available through other public domain biomedical resources such as the Open Targets platform (http://www.opentargets.org) [34] which integrates evidence from genetics, genomics, transcriptomics, drugs, animal models, and scientific literature to score and rank target-disease associations for drug target identification. The UniProt Consortium is also looking to improve the ability of both scientists and clinicians to navigate from genomic disease variant to amino acid polymorphism to effect of protein structure and/or function with both graphical visualization and computational access readily available. Variant data will become more structured, thus making it more computationally accessible [55]. The value of metabolomics data derived from AD-patients will be significantly enhanced by the work on enhancing the content of Rhea and ChEBI, and ensuring that appropriate data are incorporated into UniProt and improved and updated protein sequences will increase the number of identifications made by mass spectrometry-based proteomics groups.

In conclusion, the work described above represents a significant increase in the content of a number of public domain resources specifically focused on the molecules which play a key role in AD. Many of these proteins also play a role in other neurological disorders and are, of course, of fundamental importance to the normal physiology of the brain. These ongoing and future data updates will help clinical researchers to provide insights into the molecular mechanisms underlying the development of dementia and enable more in-depth analysis of 'Omics'-level datasets, thus

supporting the development of novel treatments and tools for early diagnosis.

DATA AVAILABILITY

UniProtKB records in which disease is caused by mutations affecting the gene represented in that entry can be found by searching www.uniprot.org with the term "keyword:"Alzheimer disease [KW-0026]". An introduction to the QuickGO Gene Ontology browser can be found at www.ebi.ac.uk/training/online/course/goa-and-quickgo-quick-tour. Tutorials on how to search UniProt and use the tools made available by this resource and how to access data pertaining to AD in the GO are available [75, 76]. Data required to create AD-focused molecular interaction network can be obtained by pasting the query annot: "dataset:Alzheimers" into the IntAct website (www.ebi.ac.uk/intact) with further details on how to use this resource available at www.ebi.ac.uk/training/online/course/intact-molecular-interactions-ebi. Extensive tutorial materials on Cytoscape network building and analysis are available at https://github.com/cytoscape/cytoscape-tutorials/wiki, the use of ClueGO is specifically described by Bindea et al. [40, 77]. How to use the Complex Portal is described by Meldal et al. [78].

ACKNOWLEDGMENTS

This work was supported by the National Eye Institute (NEI), National Human Genome Research Institute (NHGRI), National Heart, Lung, and Blood Institute (NHLBI), National Institute on Aging (NIA), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of General Medical Sciences (NIGMS), National Cancer Institute (NCI) and National Institute of Mental Health (NIMH) of the National Institutes of Health under Award Number [U24HG007822]. Research reported in this publication was additionally supported by the National Human Genome Research Institute (NHGRI) and the National Institute on Aging (NIA) of the National Institutes of Health under Award Number [3U24HG007822-05S1] (the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health).

IntAct, the Complex Portal and other EMBL-EBI-based authors also received funding from EMBL core funding, Open Targets (grant agreements OTAR-044 and OTAR02-048) and the Wellcome Trust grant INVAR (grant ref: 212925/Z/18/Z). Authors based in the Swiss-Prot Group, SIB Swiss Institute of Bioinformatics also receive funding from the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI). The University College London functional annotation team is supported by ARUK-NSG2016-13, ARUK-NAS2017A-1 and the National Institute for Health Research University College London Hospitals Biomedical Research Centre

The authors would like to thank Dr Rina Bandopadhyay and Profs John Hardy and (UCL, UK), Profs Nigel Hooper and David Brough (U.Manchester, UK), Profs Casey Brown, Li-San Wang, Christian Stoeckert (U. Penn, US), Prof. Michael MacCoss (U. Washington, US), Prof. Hans-Ulrich Klein (Columbia U., US), Prof. Christopher Martens (U. Delaware, US), Prof. Thomas Wingo (Emory U. US), Dr. Christopher Khalid-Janney (Delaware State U. US) amongst others for their help in identifying gene candidates for annotation.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-0206r1).

SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-200206.

REFERENCES

[1] Armstrong RA (2013) What causes Alzheimer's disease? *Folia Neuropathol* **51**, 169-188.
[2] De Strooper B, Karran E (2016) The cellular phase of Alzheimer's disease. *Cell* **164**, 603-615.
[3] Chávez-Gutiérrez L, Szaruga M (2020) Mechanisms of neurodegeneration - insights from familial Alzheimer's disease. *Semin Cell Dev Biol*, doi: 10.1016/j.semcdb.2020.03.005
[4] Sengoku R (2020) Aging and Alzheimer's disease pathology. *Neuropathology* **40**, 22-29.
[5] Hardy JA, Higgins GA (1992) Alzheimer's disease: The amyloid cascade hypothesis. *Science* **256**, 184-5.
[6] Selkoe DJ, Hardy J (2016) The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* **8**, 595-608.
[7] Delacourte A, Defossez A (1986) Alzheimer's disease: Tau proteins, the promoting factors of microtubule assembly, are major components of paired helical filaments. *J Neurol Sci* **176**, 173-186.

[8] Nolan A, De Paula Franca Resende E, Petersen C, Neylan K, Spina S, Huang E, Seeley W, Miller Z, Grinberg LT (2019) Astrocytic tau deposition is frequent in typical and atypical Alzheimer disease presentations. *J Neuropathol Exp Neurol* **78**, 1112-1123.

[9] Jarosz-Griffiths HH, Noble E, Rushworth JV, Hooper NM (2016) Amyloid-β receptors: The good, the bad, and the prion protein. *J Biol Chem* **291**, 3174-3183.

[10] Nortley R, Korte N, Izquierdo P, Hirunpattarasilp C, Mishra A, Jaunmuktane Z, Kyrargyri V, Pfeiffer T, Khennouf L, Madry C, Gong H, Richard-Loendt A, Huang W, Saito T, Saido TC, Brandner S, Sethi H, Attwell D (2019) Amyloid β oligomers constrict human capillaries in Alzheimer's disease via signaling to pericytes. *Science* **365**, eaav9518.

[11] Brion JP (1998) Neurofibrillary tangles and Alzheimer's disease. *Eur Neurol* **40**, 130-140.

[12] Hasegawa M (2016) Molecular mechanisms in the pathogenesis of Alzheimer's disease and tauopathies-prion-like seeded aggregation and phosphorylation. *Biomolecules* **6**, 24.

[13] Shin WS, Di J, Cao Q, Li B, Seidler PM, Murray KA, Bitan G, Jiang L (2019). Amyloid β-protein oligomers promote the uptake of tau fibril seeds potentiating intracellular tau aggregation. *Alzheimers Res Ther* **11**, 86.

[14] Wyss-Coray T, Lin C, Yan F, Yu GQ, Rohde M, McConlogue L, Masliah E, Mucke L (2001) TGF-beta1 promotes microglial amyloid-beta clearance and reduces plaque burden in transgenic mice. *Nat Med* **7**, 612-618.

[15] Kanekiyo T, Liu CC, Shinohara M, Li J, Bu G (2012) LRP1 in brain vascular smooth muscle cells mediates local clearance of Alzheimer's amyloid-β. *J Neurosci* **32**, 16458-16465.

[16] Kanekiyo T, Cirrito JR, Liu CC, Shinohara M, Li J, Schuler DR, Shinohara M, Holtzman DM, Bu G (2013) Neuronal clearance of amyloid-β by endocytic receptor LRP1. *J Neurosci* **33**, 19276-19283.

[17] Zhao Z, Sagare AP, Ma Q, Halliday MR, Kong P, Kisler K, Winkler EA, Ramanathan A, Kanekiyo T, Bu G, Owens NC, Rege SV, Si G, Ahuja A, Zhu D, Miller CA, Schneider JA, Maeda M, Maeda T, Sugawara T, Ichida JK, Zlokovic BV (2015) Central role for PICALM in amyloid-β blood-brain barrier transcytosis and clearance. *Nat Neurosci* **18**, 978-987.

[18] Bell RD, Sagare AP, Friedman AE, Bedi GS, Holtzman DM, Deane R, Zlokovic BV (2007) Transport pathways for clearance of human Alzheimer's amyloid beta-peptide and apolipoproteins E and J in the mouse central nervous system. *J Cereb Blood Flow Metab* **27**, 909-918.

[19] Leal MC, Magnani N, Villordo S, Buslje CM, Evelson P, Castaño EM, Morelli L (2013) Transcriptional regulation of insulin-degrading enzyme modulates mitochondrial amyloid β (Aβ) peptide catabolism and functionality. *J Biol Chem* **288**, 12920-12931.

[20] Hama E, Shirotani K, Iwata N, Saido TC (2004) Effects of neprilysin chimeric proteins targeted to subcellular compartments on amyloid beta peptide clearance in primary neurons. *J Biol Chem* **279**, 30259-30264.

[21] Lee MJ, Lee JH, Rubinsztein DC (2013) Tau degradation: The ubiquitin-proteasome system versus the autophagy-lysosome system. *Prog Neurobiol* **105**, 49-59.

[22] Subhramanyam CS, Wang C, Hu Q, Dheen ST (2019) Microglia-mediated neuroinflammation in neurodegenerative diseases. *Semin Cell Dev Biol* **94**, 112-120.

[23] Pereira CF, Santos AE, Moreira PI, Pereira AC, Sousa FJ, Cardoso SM, Cruz MT (2019) Is Alzheimer's disease an inflammasomopathy? *Ageing Res Rev* **56**, 100966.

[24] Makin S (2018) The amyloid hypothesis on trial. *Nature* **559**, S4-S7.

[25] Di Paolo G, Kim TW (2011) Linking lipids to Alzheimer's disease: Cholesterol and beyond. *Nat Rev Neurosci* **12**, 284-296.

[26] The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338.

[27] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29.

[28] International Society for Biocuration (2018) Biocuration: Distilling data into knowledge. *PLoS Biol* **16**, e2002846.

[29] UniProt Consortium (2019) UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515.

[30] Van Cauwenberghe C, Van Broeckhoven C, Sleegers K (2016) The genetic landscape of Alzheimer disease: Clinical implications and perspectives. *Genet Med* **18**, 421-430.

[31] Kramarz B, Roncaglia P, Meldal BHM, Huntley RP, Martin MJ, Orchard S, Parkinson H, Brough D, Bandopadhyay R, Hooper NM, Lovering RC (2018) Improving the gene ontology resource to facilitate more informative analysis and interpretation of Alzheimer's disease data. *Genes (Basel)* **9**, 593.

[32] Kramarz B, Huntley RP, Rodríguez-López M, Roncaglia P, Saverimuttu SCC, Parkinson H, Bandopadhyay R, Martin MJ, Orchard S, Hooper NM, Brough D, Lovering RC (2020) Gene ontology curation of neuroinflammation biology improves the interpretation of Alzheimer's disease gene expression data. *J Alzheimers Dis*, doi:10.3233/jad-20020

[33] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR (2019) ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res* **47**, D930-D940.

[34] Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, Carmona M, Faulconbridge A, Hercules A, McAuley E, Miranda A, Peat G, Spitzer M, Barrett J, Hulcoop DG, Papa E, Koscielny G, Dunham I (2019) Open targets platform: New developments and updates two years on. *Nucleic Acids Res* **47**, D1056-D1065.

[35] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-D363.

[36] Breuza L, Poux S, Estreicher A, Famiglietti ML, Magrane M, Tognolli M, Bridge A, Baratin D, Redaschi N; UniProt consortium (2016) The UniProtKB guide to the human proteome. *Database (Oxford)* **2016**, bav120.

[37] Meldal BHM, Bye-A-Jee H, Gajdoš L, Hammerová Z, Horácková A, Melicher F, Perfetto L, Pokorný D, Lopez MR, Türková A, Wong ED, Xie Z, Casanova EB, Del-Toro N, Koch M, Porras P, Hermjakob H, Orchard S (2019) Complex Portal 2018: Extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res* **47**, D550-D558.

[38] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504.

[39] Pillich RT, Chen J, Rynkov V, Welker D, Pratt D (2017) NDEx: A community resource for sharing and publishing of biological networks. *Methods Mol Biol* **1558**, 271-301.

[40] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J (2009) ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093.

[41] Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H (2012) Protein interaction data curation: The International Molecular Exchange (IMEx) consortium. *Nat Methods* **9**, 345-350.

[42] Murphy MP, LeVine H 3rd (2010) Alzheimer's disease and the amyloid-beta peptide. *J Alzheimers Dis* **19**, 311-323.

[43] Morgat A, Lombardot T, Axelsen KB, Aimo L, Niknejad A, Hyka-Nouspikel N, Coudert E, Pozzato M, Pagni M, Moretti S, Rosanoff S, Onwubiko J, Bougueleret L, Xenarios I, Redaschi N, Bridge A (2017) Updates in Rhea - an expert curated resource of biochemical reactions. *Nucleic Acids Res* **45**, D415-D418.

[44] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**, D1214-1219.

[45] Mielke MM, Lyketsos CG (2010) Alterations of the sphingolipid pathway in Alzheimer's disease: New biomarkers and treatment targets? *Neuromolecular Med* **12**, 331-340.

[46] Iwata M, Watanabe S, Yamane A, Miyasaka T, Misonou H (2019) Regulatory mechanisms for the axonal localization of tau protein in neurons. *Mol Biol Cell* **30**, 2441-2457.

[47] Arima K (2006) Ultrastructural characteristics of tau filaments in tauopathies: Immuno-electron microscopic demonstration of tau filaments in tauopathies. *Neuropathology* **26**, 475-483.

[48] Barbier P, Zejneli O, Martinho M, Lasorsa A, Belle V, Smet-Nocca C, Tsvetkov PO, Devred F, Landrieu I (2019) Role of tau as a microtubule-associated protein: Structural and functional aspects. *Front Aging Neurosci* **11**, 204.

[49] Kimura T, Sharma G, Ishiguro K, Hisanaga SI (2018) Phospho-tau bar code: Analysis of phosphoisotypes of tau and its application to tauopathy. *Front Neurosci* **12**, 44.

[50] Kimura T, Ishiguro K, Hisanaga SI (2014) Physiological and pathological phosphorylation of tau by Cdk5. *Front Mol Neurosci* **7**, 65.

[51] Lee MS, Kwon YT, Li M, Peng J, Friedlander RM, Tsai LH (2000) Neurotoxicity induces cleavage of p35 to p25 by calpain. *Nature* **405**, 360-364.

[52] Patrick GN, Zukerberg L, Nikolic M, de la Monte S, Dikkes P, Tsai LH (1999) Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. *Nature* **40**, 615-622.

[53] Tanzi RE, Bertram L (2005) Twenty years of the Alzheimer's disease amyloid hypothesis: A genetic perspective. *Cell* **120**, 545-555.

[54] Goldman JS, Hahn SE, Catania JW, LaRusse-Eckert S, Rumbaugh M, Strecker MN, Roberts JS, Burke W, Mayeux R, Bird T (2011) Genetic counseling and testing for Alzheimer disease: Joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors. *Genet Med* **13**, 597-605.

[55] Famiglietti ML, Estreicher A, Breuza L, Poux S, Redaschi N, Xenarios I, Bridge A; UniProt Consortium (2019) An enhanced workflow for variant interpretation in UniProtKB/Swiss-Prot improves consistency and reuse in ClinVar. *Database (Oxford)* **2019**, baz040.

[56] McGarvey PB, Nightingale A, Luo J, Huang H, Martin MJ, Wu C; UniProt Consortium (2019) UniProt genomic mapping for deciphering functional effects of missense variants. *Hum Mutat* **40**, 694-705.

[57] IMEx Consortium Curators, Del-Toro N, Duesbury M, Koch M, Perfetto L, Shrivastava A, Ochoa D, Wagih O, Piñero J, Kotlyar M, Pastrello C, Beltrao P, Furlong LI, Jurisica I, Hermjakob H, Hermjakob H, Orchard S, Porras P (2019) Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat Commun* **10**, 10.

[58] Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S (2019) SIFTS: Updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* **47**, D482-D489.

[59] Watkins X, Garcia LJ, Pundir S, Martin MJ, UniProt Consortium (2017) ProtVista: Visualization of protein sequence annotations. *Bioinformatics* **33**, 2040-2041.

[60] Jeong W, Lee H, Cho S, Seo J (2019) ApoE4-induced cholesterol dysregulation and its brain cell type-specific implications in the pathogenesis of Alzheimer's disease. *Mol Cells* **42**, 739-746.

[61] Karch CM, Goate AM (2015) Alzheimer's disease risk genes and mechanisms of disease pathogenesis. *Biol Psychiatry* **77**, 43-51.

[62] Andrew RJ, Fisher K, Heesom KJ, Kellett KAB, Hooper NM (2019) Quantitative interaction proteomics reveals differences in the interactomes of amyloid precursor protein isoforms. *J Neurochem* **149**, 399-412.

[63] Wang H, Muiznieks LD, Ghosh P, Williams D, Solarski M, Fang A, Ruiz-Riquelme A, Pomès R, Watts JC, Chakrabartty A, Wille H, Sharpe S, Schmitt-Ulms G (2017) Somatostatin binds to the human amyloid β peptide and favors the formation of distinct oligomers. *Elife* **6**, e28401.

[64] Clark LN, Poorkaj P, Wszolek Z, Geschwind DH, Nasreddine ZS, Miller B, Li D, Payami H, Awert F, Markopoulou K, Andreadis A, D'Souza I, Lee VM, Reed L, Trojanowski JQ, Zhukareva V, Bird T, Schellenberg G, Wilhelmsen KC (1998) Pathogenic implications of mutations in the tau gene in pallido-ponto-nigral degeneration and related neurodegenerative disorders linked to chromosome 17. *Proc Natl Acad Sci U S A* **95**, 13103-13107.

[65] Gunawardana CG, Mehrabian M, Wang X, Mueller I, Lubambo IB, Jonkman JEN, Wang H, Schmitt-Ulms G (2015) The human tau interactome: Binding to the

ribonucleoproteome, and impaired binding of the proline-to-leucine mutant at Position 301 (P301L) to chaperones and the proteasome. *Mol Cell Proteomics* **14**, 3000-3014.

[66] Perreau VM, Orchard S, Adlard PA, Bellingham SA, Cappai R, Ciccotosto GD, Cowie TF, Crouch PJ, Duce JA, Evin G, Faux NG, Hill AF, Hung YH, James SA, Li QX, Mok SS, Tew DJ, White AR, Bush AI, Hermjakob H, Masters CL (2010) A domain level interaction network of amyloid precursor protein and Abeta of Alzheimer's disease. *Proteomics* **10**, 2377-2395.

[67] Villaveces JM, Jiménez RC, Porras P, Del-Toro N, Duesbury M, Dumousseau M, Orchard S, Choi H, Ping P, Zong NC, Askenazi M, Habermann BH, Hermjakob H (2015) Merging and scoring molecular interactions utilising existing community standards: Tools, use-cases and a case study. *Database (Oxford)* **2015**, bau131.

[68] Hussaini SMQ, Jang MH (2018) New roles for old glue: Astrocyte function in synaptic plasticity and neurological disorders. *Int Neurourol J* **22**, S106-S114.

[69] Nagarsheth MH, Viehman A, Lippa SM, Lippa CF (2006) Notch-1 immunoexpression is increased in Alzheimer's and Pick's disease. *J Neurol Sci* **244**, 111-116.

[70] Malhotra A, Younesi E, Sahadevan S, Zimmermann J, Hofmann-Apitius M (2015) Exploring novel mechanistic insights in Alzheimer's disease by assessing reliability of protein interactions. *Sci Rep* **5**, 13634.

[71] Jupe S, Fabregat A, Hermjakob H (2015) Expression data analysis with Reactome. *Curr Protoc Bioinformatics* **49**, 8.20.1-8.20.9.

[72] Orchard S, Hermjakob H (2015) Shared resources, shared costs–leveraging biocuration resources. *Database (Oxford)* **2015**, bav009.

[73] Porras P, Duesbury M, Fabregat A, Ueffing M, Orchard S, Gloeckner CJ, Hermjakob H (2015) A visual review of the interactome of LRRK2: Using deep-curated molecular interaction data to represent biology. *Proteomics* **15**, 1390-1404.

[74] Foulger RE, Denny P, Hardy J, Martin MJ, Sawford T, Lovering RC (2016) Using the gene ontology to annotate key players in Parkinson's disease. *Neuroinformatics* **14**, 297-304.

[75] Pundir S, Magrane M, Martin MJ, O'Donovan C; UniProt Consortium (2015) Searching and Navigating UniProt Databases. *Curr Protoc Bioinformatics* **50**, 1.27.1-10.

[76] Kramarz B, Lovering RC (2019) Gene ontology: A resource for analysis and interpretation of Alzheimer's disease data. In *Alzheimer's Disease*. Codon Publications, Brisbane.

[77] Mlecnik B, Galon J, Bindea G (2019) Automated exploration of gene ontology term and pathway networks with ClueGO-REST. *Bioinformatics* **35**, 3864-3866.

[78] Meldal BHM, Orchard S (2018) Searching and extracting data from the EMBL-EBI complex portal. *Methods Mol Biol* **1764**, 377-390.