

A Molecular View of Amyotrophic Lateral Sclerosis Through the Lens of Interaction Network Modules

Klaus Højgaard Jensen

Intomics A/S

Anna Katharina Stalder

University Hospital Basel: Universitatsspital Basel

Rasmus Wernersson (✉ rwe@intomics.com)

Intomics A/S <https://orcid.org/0000-0003-4417-9842>

Tim-Christoph Roloff-Handschin

University Hospital Basel: Universitatsspital Basel

Daniel Hvidberg Hansen

Intomics A/S

Peter MA Groenen

Idorsia Pharmaceuticals Ltd

Research article

Keywords: Amyotrophic lateral sclerosis, ALS, neurodegenerative disease, protein-interaction networks, PPI, data integration, Network Biology, molecular disease mechanism

Posted Date: July 19th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-656827/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: Despite the discovery of familial cases with mutations in Cu/Zn-superoxide dismutase (SOD1), Guanine nucleotide exchange C9orf72, TAR DNA-binding protein 43 (TARDBP) and RNA-binding protein FUS as well as a number of other genes linked to Amyotrophic Lateral Sclerosis (ALS), the etiology and molecular pathogenesis of this devastating disease is still not understood. As proteins do not act alone but rather in networks, conducting an analysis of ALS at the system level may provide new insights into the molecular biology of ALS around these potential ALS-associated genes and put it into relationship to other neurological diseases.

Methods: A set of ALS-associated genes/proteins were collected from publicly available databases and text mining of scientific literature. We used these as seed proteins to build protein-protein interaction (PPI) networks serving as a scaffold for further analyses. From the collection of networks, a set of core modules enriched in seed proteins were identified. The core modules were investigated for overrepresentation of Gene Ontology categories and association to other diseases. To assess the core modules' ability to describe unknown or less well-studied ALS biology, they were queried for proteins more recently associated to ALS and not involved in the primary analysis.

Results: We describe a set of 26 ALS core modules enriched in ALS-associated proteins. Our workflow is robust against individual highly connected proteins often dominating this type of analysis due to study bias or experimental artefacts. We show that these ALS core modules not only capture most of the current knowledge about ALS, but they also allow us to suggest biological interdependencies. In addition, new associations of ALS networks with other neurodegenerative diseases, e.g. Alzheimer's, Huntington's and Parkinson's disease were found. A follow-up analysis of 140 ALS-associated proteins identified since 2014 reveals a significant overrepresentation of new ALS proteins in these 26 disease modules.

Conclusions: Using protein-protein interaction networks offers a relevant approach for broadening the understanding of the biological context of known ALS-associated genes. Using a bottom-up approach for the analysis of protein-protein interaction networks is a useful method to avoid bias caused by over-connected proteins. Our ALS-enriched modules cover most known biological functions associated with ALS. The presence of recently identified ALS-associated proteins in the top network highlights the potential for using the ALS-associated networks as a scaffold for identification of novel ALS disease mechanisms, investigation of complex networks and hypothesis generation.

Background

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease clinically driven by deterioration of motor functions and affects about 1–2 individuals per 100,000 per year. It is a rapidly progressing disease with a life expectancy after diagnosis of 3–5 years. ALS is classified into two categories, which are clinically very similar. The most common form of ALS, with over 90% of the cases, is the sporadic form (SALS), where no mutations can be identified. The less frequent (10 %) familial forms (FALS) are often autosomal dominant, caused by a number of mutations in a heterogeneous set of genes [1, 2].

The main mutations currently described are in SOD1 (20% of FALS), C9orf72 (30–40% of FALS), TARDBP and FUS (together 5% of FALS)[3]. There are also other genes that have been associated with ALS, but they represent a small proportion of cases [4]. However, there is an increasing rate of new discoveries of ALS-associated genes, for which the supporting biology is not well established. During the last 5 years, more than 7 new mutations have been identified which are associated with one or more pathological mechanisms associated with ALS [5] (MATR3, CHCHD10, TBK1, TUBA4A, NEK1, C21orf2, CCNF). These pathological mechanisms include, among others, dysfunction in global protein homeostasis resulting from abnormal protein aggregation or a defect in the protein clearance pathway, mitochondrial dysfunction, altered RNA metabolism, impaired cytoskeletal integrity, altered axonal transport dynamics, and DNA damage accumulation due to defective DNA repair. However, despite active research, the pathogenesis and etiology of ALS remain largely unknown, and while an increasing number of genetic factors are being identified it remains unclear how these cellular events lead to the wide range of proposed mechanisms that underlie this complex and devastating neurodegenerative disease.

The clinical manifestation of ALS is not the consequence of a single mutated protein, but rather represents a complex imbalance of a dynamic network of proteins which links intra-cellular processes to inter-cellular processes. Network biology offers an approach to explore the molecular complexity of a disease and define disease modules and relevant pathways. With this approach ALS-associated mutations can be investigated in the relevant biological context of the molecular pathways they are involved in, thereby linking them to potentially new genes that may become future drug targets.

The concept of spectrum disorders has been introduced to link phenomenologically distinct

diseases based on a related underlying etiology or genetic link. The most accepted spectrum

disorders are in the autism spectrum, where this concept is well accepted. Also, other

mental diseases, such as Lewy body dementias and Parkinson's disease are thought to be part of a spectrum rather than separate pathologies [6]. Similarly, ALS and frontotemporal dementia [7] or other motor neuron diseases have been discussed to be the extreme ends of a spectrum, based on imaging similarities and genetic and neuropathological findings [8]. Based on the ALS core modules we are analyzing other neurological disorders for molecular overlaps and are relating them to the molecular biology of ALS.

ALS in the Network Biology perspective

Genes and proteins do not work in isolation, and insight about disease biology of complex diseases can be gained from investigating the interaction partners of known disease associated genes/proteins [9]. This could for example be by investigating which pathways appear to be perturbed (either by containing known disease associated proteins or by analysis of experimental data such as GWAS and gene expression studies). Pathway data from sources such as

KEGG [10], WikiPathways [11] and Reactome [12] is generally of high quality, and is indeed a very valuable resource for biological interpretation of data. However, it is important to realize that pathways represent what could be considered “text-book knowledge” or “established knowledge” – a curated and aggregated summary of what is considered to be the consensus in the scientific literature. Consequently, truly novel interactions of ALS-associated genes/proteins are not expected to be found relying solely on pathway information.

Protein-protein interaction (PPI) networks can be used as an alternative avenue for analysis. PPI networks have a higher coverage of human genes/proteins compared to pathways (up to 85% coverage of proteome [13, 14]) and are especially strong in their ability to cover the areas in-between the well-known pathways because they are based on high-throughput experimental screens, rather than curation of literature knowledge. Biological modules found from network analysis of PPI data could be considered “proto-pathways” that may mature into established pathways after experimental follow-up.

When working with PPI data, it is important to be aware of experimental artifacts that may carry over into the protein networks, and which therefore may impact the interpretation of results. We explore this in detail in the supplement, as there are many factors to consider. Here it will be sufficient to emphasize that certain proteins have an inflated number of interactions in PPI networks, either due to experimental artefacts (see [15] for a review) or due to study bias. We here refer to these as over-connected proteins. This situation can adversely affect results from network analysis if not addressed properly. A protein that is highly connected due to these factors in the global interactome, will have a very high likelihood of appearing in the network neighborhood around any set of disease associated proteins, no matter if the study is investigating ALS, Heart disease, Diabetes, Cancer or any other class of disease. It will also be very likely, that such an “over-connected” protein will have a high number of interaction partners in the local network around the disease proteins being studied, without it necessarily being of particular importance for the disease *per se*. This is important, since network analysis studies that are built around the concept of the “centrality-lethality rule” [16] where the node degree (number of interaction partners) is important, are likely to pick up these over-connected proteins, unless a node-degree normalization scheme is used. UBC (Polyubiquitin-C) is a good example of such an artificially over-connected protein (see Supplement for details), and we notice that this very protein is singled out as being important for ALS in recent study by Mao et al [17], due to its high node degree in the local network around ALS associated proteins. In that study UBC’s high node degree in the entire interactome was not taken into account, and we speculate it to be a false positive.

Here, we present an approach to network analysis of known ALS-associated genes/proteins that handles the issues concerning over-connected proteins in the network in several ways. First of all, the network modules are built using a bottom-up approach, where the first-order networks around each individual ALS-associated protein are first extracted and then merged if they overlap. By doing this, rather than having a first step, where we pull out a very large network surrounding all ALS proteins, we avoid having to run a topological network clustering algorithm (such as MCODE [18] or CLUSTERONE[19]) which are prone to emphasize features in the network influenced by the over-connected proteins. As a second measure we have built in a pruning step following the initial network extraction where proteins are removed if their global degree (total number of interactions) is much larger than their local degree (number of interactions in the network being investigated), which again will down-weight the influence of proteins that are highly and unspecifically connected in the global interactome. The result is a set of 26 ALS core modules that capture most of the current knowledge about ALS and provide new insight into the biology and etiology of this complex disease.

Methods

Our analysis strategy is shown in Fig. 1 and is based on the concept of finding network modules of moderate size significantly enriched in known ALS-associated proteins, and where other proteins in the modules therefore are likely also involved in ALS-related processes. This also allows for a detailed analysis of the biology represented in each module, and for generating a module collection that can be used for further study.

Compiling the list of known ALS-associated genes

The dataset of known ALS-associated genes, was generated by combining information from the sources listed below. The lists were selected in order to be as comprehensive and inclusive as possible in order to capture all reported ALS genes. Entries were mapped to reviewed UniProt [20] entries in order to facilitate analysis in the context of PPI networks. Entries that could not be mapped (e.g. intergenic SNPs) were omitted from the analysis.

Amyotrophic Lateral Sclerosis online Database (ALSoD)

ALSoD [21] is a repository combining information genotype, phenotype and geographical information.

Data was downloaded from ALSoD on May 27th 2014. The dataset consisted of 114 genes, out of which 111 could be mapped to 111 reviewed UniProt proteins.

ALSGene

ALSGene is a database of genes associated with ALS through GWAS.

The ALSGene “Top Results” dataset [22] was downloaded on May 27th 2014. The dataset consists of 22 genetic polymorphisms, out of which 17 were mapped to genes. All 17 genes could be mapped to reviewed UniProt proteins.

Genotator

Genotator [23] combines data from 11 external sources to provide information on human diseases.

All data associated with “Amyotrophic Lateral Sclerosis” was downloaded from Genotator on April 30th 2014. The dataset consists of 294 loci, out of which 289 genes could be mapped to 289 reviewed UniProt proteins.

Genetic Association Database (GAD)

GAD contains genetic associations from complex diseases.

The entire dataset from GAD [24, 25] was downloaded on May 2nd 2014. Afterwards all genes associated with “Amyotrophic Lateral Sclerosis” (and subcategories) were extracted yielding a dataset consisting of 275 loci, out of which 226 could be mapped to 228 reviewed UniProt proteins.

HuGE Navigator, GWAS

The full gene level GWAS [26] dataset was downloaded from the HuGE Navigator website on May 1st 2014. Each line containing the text string “Amyotrophic lateral sclerosis” was extracted yielding a dataset of 76 records which after redundancy reduction and removal of intergenic regions gave rise to a list of 46 genes, out of which 43 could be mapped to reviewed UniProt proteins.

Furthermore, a list of genes in the 200kb regions surrounding ALS-associated SNPs was downloaded on May 6th 2014. Following redundancy reduction 101 genes were indicated, out of which 97 could be mapped to reviewed UniProt proteins.

The 200kb region list was combined with the gene level list to yield a final list of 118 proteins.

HuGE Navigator, Gene prospector

A dataset of all genes associated to “Amyotrophic Lateral Sclerosis” was downloaded from the HuGE Navigator website on May 1st 2014 via the “Gene Prospector” search interface [27]. The dataset consisted of 194 genes, out of which 191 could be mapped to reviewed UniProt proteins. In total 224 UniProt proteins were indicated due to mapping of HLA-B to multiple allelic variants.

Text mining

Finally, we used the InBio Know [28] text-mining solution from Intomics A/S (see the text mining section below for details) to build a list of genes significantly associated with ALS based on the May 2014 set of PubMed abstracts. The list consisted of 164 proteins (already based on UniProt IDs thus no further mapping was needed).

Combined ALS-associated list

All lists of ALS gene/protein associations were aggregated to yield a total list of 656 proteins – see Supplementary Table S 1.

ALS-associated genes added since 2014 data survey

An updated dataset of ALS associated genes were downloaded from ALSod and HuGE Navigator Genopedia on April 27, 2020. Genes were mapped to UniProt IDs. A new round of text mining was conducted to identify new proteins associated with ALS in the scientific literature.

A total of 10 and 91 new proteins (i.e. not present in the ALS-databases at the time of the original data survey) were found in ALSod and HuGE Navigator Genopedia, respectively, while 61 new proteins were identified through text mining. Due to overlap between gene sets from the three resources, a total set of 140 unique proteins were found to be associated with ALS since May 2014 (Supplementary Table S 2).

Proteins present in ALS core modules were tested for overrepresentation of new ALS-associated proteins using a Fisher’s exact test.

For the remaining databases no new datasets were obtained. The ALSGene database has not been updated since 2011 and Genotator and GAD databases have been discontinued.

Generating ALS-associated PPI networks

The PPI resource inBio Map™/InWeb_IM [13] was chosen as the source of PPIs for building a comprehensive collection of ALS related networks. Briefly, inBio Map is a large, robust, high confidence database of inferred human physical PPIs gathered from multiple databases of experimental evidence. Human PPIs were obtained from the February 2014 version of inBio Map. Low confidence interactions were filtered out using a confidence score cutoff of 0.1, resulting in 130,746 high confidence interactions between 11,900 proteins.

The ALS network collection was built by considering all first order networks around the ALS-associated proteins (the aggregated inclusive list, across all datasets, we consider these the “seed proteins”), then pruning the networks for over-connected (proteins with a high number of interactions outside the current network), and finally merging all networks with a high degree of overlap. This yielded a collection of 282 networks with the size distribution shown in Supplementary Figure S 1. This collection of networks is also available as a Cytoscape [29] data file as part of the supplementary materials.

Defining ALS core modules

From the full set of 282 networks, we define the sub-set of networks significantly enriched in ALS-associated genes/proteins as the set of **ALS core modules**. Each of the modules in the core set, is more likely to be close to known ALS biology, and has been investigated further with regards to both molecular biology as well as overlap with genes/proteins associated with other diseases, that significantly share these core modules with ALS.

The set of five modules significant after Bonferroni correction are in the following termed ‘Tier 1’, while the remaining modules significant only after the less strict Benjamini-Hochberg correction are termed ‘Tier 2’. Tier 1 and Tier 2 are collectively referred to as the ‘ALS core modules’.

Biological annotation of the ALS network collection

We investigated the biology represented in each of the ALS-associated disease modules using two lines of evidence: 1) overrepresentation of Gene Ontology [30, 31] 'Biological processes' terms and 2) overlap with other diseases.

In both cases of overrepresentation analysis, the gene/protein identifiers were extracted from the individual ALS-associated modules and used as the study set in the analysis and the entire human interactome as the background set. Fisher's Exact test was used to test for overrepresentation of GO terms/diseases among the network proteins. The significance threshold was adjusted for multiple testing using Bonferroni correction.

After GO annotation 3 filtering steps were applied to select the most relevant GO terms representing each disease module:

1. GO term must be 'minimal'. When a p-value for a GO term has been calculated, it is compared to the p-values for all child GO terms. The p-value is then said to be "minimal" if it is less than all the p-values for all the child GO terms (or if the GO term does not have any children).
2. GO term overrepresentation must be significant after Bonferroni correction ($p_{\text{corr}} < 0.05$)
3. GO term must contain < 1000 genes

GO terms were manually curated by biological experts, to remove very generic GO terms (e.g. GO:0022402, "cell cycle process" or GO:000077 "DNA damage checkpoint"), or irrelevant GO terms based on a biological assessment, resulting in a total of 182 GO terms being overrepresented in the top 26 modules. To reduce the redundancy inherently present in the GO hierarchy (some GO terms have a high degree of overlap, or multiple GO terms can describe the same biology), the 182 GO terms were then manually collected into GO biological classes based on their labels. GO classes were furthermore collected into GO super-classes representing broad biological classes or functions (Supplementary Table S 3).

Visualization of networks

All visualizations of disease modules, including the visualization of protein-level metadata (known ALS-associated genes/proteins, biological categories) were performed using Cytoscape [29].

Text mining of diseases co-mentioned with ALS

The inBio Know™ software suite [28] provided by Intomics A/S was used to find associations between 1) diseases vs. genes and 2) diseases vs. diseases in PubMed abstracts. Manually curated synonyms for all human genes/proteins as well as synonyms for all Disease Ontology [32, 33] diseases provided by Intomics A/S was used.

The text mining software was used as follows:

1. For each item of interest (e.g. a human disease) all matches in PubMed for any of its curated synonyms were recorded. Multiple hits in a single PubMed entry would be counted as one. From this the background frequency in the entire pool of PubMed abstract was calculated.
2. For pairs of items of interest (e.g. a disease and a gene/protein), the number of abstracts co-mentioning the terms was counted, and enrichment over background frequencies (observed / expected) was calculated using Fisher's Exact test.
3. Within each sub-study (e.g. human disease vs. all human genes/proteins), the significance threshold was adjusted for multiple testing using Bonferroni correction.

Disease vs. disease comparison

An initial round of text mining to find diseases co-mentioned with ALS in PubMed abstracts was conducted, and from this we extracted all diseases with at least one co-mentioning with ALS. The significance of each disease/ALS-association was calculated using a Fisher's exact test. From the total set of diseases, the top 50 most significantly overrepresented hits with respect to the association with ALS were selected as the pool of diseases to investigate further.

Disease vs. genes/proteins comparison

Each of the top 50 diseases identified above, were text-mined for co-mentioning with human genes/proteins in PubMed abstracts (Supplementary Table S 4). A p -value threshold of 10^{-10} was used to call a significant association between a disease and a gene/protein - notice that this is even more restrictive than the standard Bonferroni correction of testing a theoretical maximum of up to 20,500 genes/proteins per disease ($p < 2.4 \times 10^{-6}$). The 10^{-10} threshold corresponds to the high-confidence subset of disease/gene associations in the inBio Know™ software suite.

Results

Overlap of ALS-associated genes in different data bases

A comprehensive analysis of 7 databases for canonical ALS-associated proteins, yielded 656 proteins linked to ALS (Fig. 2, Supplementary Table S 1). There was a surprisingly low overlap between the ALS-associated proteins obtained from the 7 sources we used to build the dataset (see Fig. 2). Even considering the diversity of the sources, this appears to indicate a level of uncertainty whether these genes are truly associated with ALS. A set of only 29 proteins had a high level of agreement in 5 out of 7 data sources (Table 1), indicating the most comprehensively studied subset of ALS related genes. Among these genes are the known players of ALS pathology such as SOD1, C9ORF72, TARDBP, as well as many less well-established genes, which are thought to constitute

additional risk factors for causation, modification or progression of ALS (for example SQSTM1 and VCP [4]). Other putative ALS-associated genes are found only in one database or in literature and their contribution to ALS pathogenesis needs to be studied further.

Table 1: ALS-associated genes/proteins supported by 5 or more resources

29 proteins found in ≥ 1 data source. 8 out of the most frequent proteins were found in one or more of the 26 ALS core modules.

Gene name	Resource							Protein information	
	ALSGene (n = 17)	ALSoD (n = 111)	GAD (n = 224)	Genotator (n = 289)	HuGE GWAS (n = 118)	HuGE GeneProspector (n = 224)	Text Mining (n = 164)	UniProt entry	Protein name
C9ORF72	•	•	•	•	•	•	•	CI072_HUMAN	Protein C9orf72
DPP6		•	•	•	•	•	•	DPP6_HUMAN	Dipeptidyl aminopeptidase-like protein 6
ITPR2		•	•	•	•	•	•	ITPR2_HUMAN	Inositol 1,4,5-trisphosphate receptor type 2
KIFAP3		•	•	•	•	•	•	KIFA3_HUMAN	Kinesin-associated protein 3
SOD1		•	•	•	•	•	•	Tier 1 SODC_HUMAN	Superoxide dismutase [Cu-Zn]
UNC13A		•	•	•	•	•	•	UN13A_HUMAN	Protein unc-13 homolog
ALS2		•	•	•		•	•	ALS2_HUMAN	Alsin
ALS6, FUS		•	•	•		•	•	Tier 1 FUS_HUMAN	RNA-binding protein FUS
ANG		•	•	•		•	•	ANG1_HUMAN	Angiogenin
CHMP2B		•	•	•		•	•	CHM2B_HUMAN	Charged multivesicular body protein 2b
CST3		•	•	•		•	•	Tier 2 CYTC_HUMAN	Cystatin-C
DCTN1		•	•	•		•	•	Tier 1 DCTN1_HUMAN	Dynactin subunit 1
DYNC1H1		•	•	•		•	•	DYHC1_HUMAN	Cytoplasmic dynein 1 heavy chain 1
ELP3		•	•	•		•	•	ELP3_HUMAN	Elongator complex protein 3
FGGY		•	•	•		•	•	FGGY_HUMAN	FGGY carbohydrate kinase domain-containing protein
HFE		•	•	•		•	•	HFE_HUMAN	Hereditary hemochromatosis protein 1
LIPC		•	•	•	•	•	•	Tier 1 LIPC_HUMAN	Hepatic triacylglycerol lipase
PON1		•	•	•		•	•	PON1_HUMAN	Serum paraoxonase/arylesterase 1
PON2		•	•	•		•	•	PON2_HUMAN	Serum paraoxonase/arylesterase 2
PON3		•	•	•		•	•	PON3_HUMAN	Serum paraoxonase/lactonase 3
PRPH		•	•	•		•	•	PER1_HUMAN	Peripherin
SMN2		•	•	•		•	•	SMN_HUMAN	Survival motor neuron protein
SUSD1		•	•	•	•	•	•	SUSD1_HUMAN	Sushi domain-containing protein 1
TARDBP		•	•	•		•	•	Tier 1 TADBP_HUMAN	TAR DNA-binding protein 43
VAPB		•	•	•		•	•	Tier 2 VAPB_HUMAN	Vesicle-associated membrane protein-associated protein B/C
VCP		•	•	•		•	•	Tier 1 TERA_HUMAN	Transitional endoplasmic reticulum ATPase
VEGF _A		•	•	•		•	•	VEGFA_HUMAN	Vascular endothelial growth factor 1

VEGFA								A
VPS54	•	•	•	•	•		VPS54_HUMAN	Vacuolar protein sorting-associated protein 54
ZFP64	•	•	•	•	•		ZF64A_HUMAN	Zinc finger protein 64 homolog, isoforms 1 and 2
ZFP64	•	•	•	•	•		ZF64B_HUMAN	Zinc finger protein 64 homolog, isoforms 3 and 4

ALS-associated genes with network support

Without biological context, individual genes that are identified to be associated with ALS may be useful for diagnosis but do not contribute to the understanding of the molecular pathophysiology and the subsequent search for prevention or treatment [34]. However, if these genes are part of networks which are significantly enriched in ALS-associated proteins, it can help to reinforce the evidence for more weakly supported proteins. To investigate this further, we evaluated a collection of 282 PPI network modules for overrepresentation of ALS-associated proteins (see Methods for details). 26 ALS modules were significantly enriched for ALS-associated proteins after multiple testing correction using the Benjamini-Hochberg procedure ($q < 0.1$). Five of these were also significant after correcting for multiple testing using Bonferroni correction ($p < 0.05/282$). The set of five modules significant after Bonferroni correction are in the following termed 'Tier 1', while the remaining 21 modules are termed 'Tier 2'. Tier 1 and Tier 2 modules are collectively referred to as 'ALS core modules'.

36 ALS-associated genes/proteins are supported by Tier 1 modules (Table 2). Nine of these proteins were only mentioned in one source, often found only by text mining. The link to Tier 1 ALS core modules strengthens the likelihood that these genes are indeed ALS-relevant genes. Further 108 genes/proteins have Tier 2 support for a total of 144 genes/proteins having Tier 1/2 network support (see Supplementary Table S 1 for full list).

Table 2: ALS-associated proteins in Tier 1 Modules

36 ALS-associated proteins present in at least of the five Tier 1 networks. A full list of ALS-associated proteins found in ≥ 1 of the ALS core modules is found in Supplementary Table S 1..

Gene name	Resource							Overlap	Protein information	
	ALSGene (n = 17)	ALSoD (n = 111)	GAD (n = 224)	Genotator (n = 289)	HuGE GWAS (n = 118)	HuGE GeneProspector (n = 224)	Text Mining (n = 164)		UniProt entry	Protein name
SOD1		•	•	•	•	•	•	6	SODC_HUMAN	Superoxide dismutase [Cu-Zn]
DCTN1		•	•	•			•	5	DCTN1_HUMAN	Dynactin subunit 1
ALS6, FUS		•	•	•			•	5	FUS_HUMAN	RNA-binding protein FUS
LIPC		•	•	•	•	•		5	LIPC_HUMAN	Hepatic triacylglycerol lipase
TARDBP		•	•	•			•	5	TADBP_HUMAN	TAR DNA-binding protein 43
VCP		•	•	•			•	5	TERA_HUMAN	Transitional endoplasmic reticulum ATPase
APOE		•	•	•				4	APOE_HUMAN	Apolipoprotein E
ATXN2		•		•			•	4	ATX2_HUMAN	Ataxin-2
SQSTM1		•		•			•	4	SQSTM_HUMAN	Sequestosome-1
MAPT		•		•			•	4	TAU_HUMAN	Microtubule-associated protein tau
UBQLN2		•		•			•	4	UBQL2_HUMAN	Ubiquilin-2
APOA4			•	•				3	APOA4_HUMAN	Apolipoprotein A-IV
LPA			•	•				3	APOA_HUMAN	Apolipoprotein(a)
APOB			•	•				3	APOB_HUMAN	Apolipoprotein B-100
APOC3			•	•				3	APOC3_HUMAN	Apolipoprotein C-III
CCS		•		•			•	3	CCS_HUMAN	Copper chaperone for superoxide dismutase
CETP			•	•				3	CETP_HUMAN	Cholesteryl ester transfer protein
CNTF		•		•			•	3	CNTF_HUMAN	Ciliary neurotrophic factor
LDLR			•	•				3	LDLR_HUMAN	Low-density lipoprotein receptor
LPL			•	•				3	LIPL_HUMAN	Lipoprotein lipase
RNF19A		•		•			•	3	RN19A_HUMAN	E3 ubiquitin-protein ligase RNF19A
HNRNPA1		•					•	3	ROA1_HUMAN	Heterogeneous nuclear ribonucleoprotein A1
CHGB			•	•				3	SCG1_HUMAN	Secretogranin-1
UBQLN1				•			•	3	UBQL1_HUMAN	Ubiquilin-1
ARHGEF28, RGNEF		•						2	ARG28_HUMAN	Rho guanine nucleotide exchange factor 28
P4HB							•	2	PDIA1_HUMAN	Protein disulfide-isomerase
BCL2							•	1	BCL2_HUMAN	Apoptosis regulator Bcl-2
CASP3							•	1	CASP3_HUMAN	Caspase-3
DERL1							•	1	DERL1_HUMAN	Derlin-1
HDAC6							•	1	HDAC6_HUMAN	Histone deacetylase 6
HECW1							•	1	HECW1_HUMAN	E3 ubiquitin-protein

				ligase HECW1
HS3ST3A1	•	1	HS3SA_HUMAN	Heparan sulfate glucosamine 3-O-sulfotransferase 3A1
MYLIP	•	1	MYLIP_HUMAN	E3 ubiquitin-protein ligase MYLIP
NEFM	•	1	NFM_HUMAN	Neurofilament medium polypeptide
PARK7	•	1	PARK7_HUMAN	Protein deglycase DJ-1
SORT1	•	1	SORT_HUMAN	Sortilin

An effort to identify new ALS-associated proteins through a combination of text mining and database searches revealed a set of 140 proteins not present in the initial data survey (Supplementary Table S 2). 17 (12.1 %) new proteins are found in one or more Tier1 + Tier2 disease modules, which is a significant overlap ($p = 0.03$).

ALS network collection

The collection of all 282 ALS-associated PPI networks, offers the opportunity to investigate the biology of networks closely associated with ALS related genes, as well as a framework for mapping experimental data (e.g. gene expression data) to the networks. The networks are available for download as a Cytoscape session file as part of the supplementary materials. A separate Cytoscape session is available for download containing only the **ALS core modules**, with Tier 1 vs. Tier 2 clearly marked in the overview. Furthermore, it contains all metadata and graphical styles needed to generate the visualization of the disease modules shown in this publication, thus allowing for further exploration of the **ALS core modules**.

ALS core modules in overview

Investigating the spectrum of molecular biology represented in the 26 **ALS core modules** (Fig. 3), by evaluating the Gene Ontology categories overrepresented in them, leads to the following observations: *Apoptosis* is represented in most (19) of the modules, as is *protein degradation* (19). A large proportion (19) of the modules are enriched for genes/proteins involved in *protein-modification* (15) or *-localization* (11). *Axon guidance*, and *immune response* are represented in 12 and 13 core modules, respectively.

Some GO terms were only represented in Tier 2 core modules and not in Tier 1 modules. These include GO terms are centered around muscle, nervous system, synapse and glutamate, which are classically linked to ALS.

With a focus on the five Tier 1 core modules this analysis showed that most core modules are representing apoptosis and most often linking it to protein degradation or core module specific additional GO terms, for example core module 93 exclusively contains *lipoprotein*.

Table 3: Top 50 diseases most commonly co-mentioned with ALS in PubMed abstracts

List of the 50 diseases most commonly co-mentioned with ALS in scientific literature. Diseases have manually been categorized into 'Disease types' guided by the tree structure in Disease Ontology [32, 33]. ID: Disease Ontology ID. Abbreviation: Short name for disease used in Figure 5-9. Name: Disease name in Disease Ontology. Occurrences: Number PubMed abstracts mentioning disease. Co-mentionings: Number of PubMed abstracts where both disease and ALS is mentioned. Overrep. ratio: Overrepresentation ratio of co-mentionings. P-value: Significance of hypergeometric test of the overlap between disease and ALS. Disease type: Manual classification of diseases.

ID	Abbreviation	Name	Occurrences	Co-mentionings	Overrep. ratio	P-value	Disease type
DOID:11949	CJD	Creutzfeldt-Jakob disease	4,508	61	14.6	1.59E-48	Brain disease
DOID:12680	PseudoBulbPalsy	pseudobulbar palsy	392	22	60.5	7.41E-32	
DOID:12859	Chorea	choreatic disease	3,001	38	13.7	7.75E-30	
DOID:5702	PleoLiposac	pleomorphic liposarcoma	4,293	69	17.3	1.54E-59	Cancer
DOID:3939	LipoCanc	lipomatous cancer	4,378	69	17.0	5.59E-59	
DOID:769	NeuroBlast	neuroblastoma	27,962	99	3.8	5.56E-28	
DOID:2476	HS Paraplegia	hereditary spastic paraplegia	1,690	103	65.7	8.73E-147	Central nervous system disease
DOID:607	Paraplegia	paraplegia	9,017	114	13.6	5.10E-86	
DOID:1307	Dementia	dementia	136,817	2,639	20.8	~0	Cognitive disorder
DOID:9255	FT Dementia	frontotemporal dementia	5,737	1,217	228.8	~0	
DOID:12217	LB Dementia	Lewy body dementia	3,916	81	22.3	4.24E-78	
DOID:5408	PBD	Paget's disease of bone	1,532	54	38.0	6.84E-65	Connective tissue disease
DOID:205	BoneHyp	hyperostosis	6,705	54	8.7	5.47E-32	
DOID:4953	Poliomyelitis	poliomyelitis	7,184	95	14.3	1.04E-73	Infectious disease
DOID:4952	PPMSyn	postpoliomyelitis syndrome	535	35	70.6	4.20E-52	
DOID:438	AIDNeuro	autoimmune disease of the nervous system	14,913	189	13.7	1.10E-141	Immune system disease
DOID:12842	GBSyn	Guillain-Barre syndrome	6,675	106	17.1	5.36E-90	
DOID:437	MyaGrav	myasthenia gravis	8,408	105	13.5	7.67E-79	
DOID:2033	ComDis	communication disorder	26,572	129	5.2	9.33E-50	Mental health
DOID:0060046	Aphasia	aphasia	9,060	57	6.8	2.56E-28	
DOID:700	MitoMetaD	mitochondrial metabolism disease	5,245	45	9.3	4.82E-28	Metabolic disease
DOID:683	MotorNeu	motor neuritis	1,915	139	78.3	2.97E-208	Motor neuron disease
DOID:681	ProgBulbPalsy	progressive bulbar palsy	432	97	242.2	1.37E-196	
DOID:0060161	SBMA	Kennedy's disease	787	96	131.6	6.91E-167	Movement disease
DOID:678	ProgSupraPalsy	progressive supranuclear palsy	2,782	119	46.1	7.26E-151	
DOID:767	MuscAtrophy	muscular atrophy	14,482	897	66.8	~0	Muscle tissue disease
DOID:3429	IncBodyMyositis	inclusion body myositis	1,525	88	62.2	1.52E-123	
DOID:11722	MyoDystT1	myotonic dystrophy type 1	3,772	74	21.2	7.35E-70	
DOID:11723	DuchMuscDys	Duchenne muscular dystrophy	7,844	82	11.3	4.30E-56	Neurodegenerative
DOID:12858	HD	Huntington's disease	11,696	642	59.2	~0	

DOID:14330	PD	Parkinson's disease	69,957	1,725	26.6	~0	disease
DOID:10652	AD	Alzheimer's disease	92,611	1,658	19.3	~0	
DOID:2377	MS	multiple sclerosis	49,375	686	15.0	~0	
DOID:2478	SC Degen	spinocerebellar degeneration	8,342	232	30.0	1.35E-249	
DOID:1441	SC Ataxia	spinocerebellar ataxia	7,104	211	32.0	4.70E-233	
DOID:11870	PickD	Pick's disease	976	65	71.8	7.94E-96	
DOID:9277	PC Degen	primary cerebellar degeneration	1,134	48	45.7	1.14E-61	
DOID:12705	FR Ataxia	Friedreich ataxia	2,119	56	28.5	2.50E-60	
DOID:10595	CMTD	Charcot-Marie-Tooth disease	8,334	647	83.7	0	Neuromuscular disease
DOID:3602	ToxEnceph	toxic encephalopathy	26,094	293	12.1	1.44E-204	Nervous system disease
DOID:12697	LockedInSyn	locked-in syndrome	5,900	72	13.2	6.12E-54	
DOID:913	AtroMuscD	atrophic muscular disease	5,802	575	106.9	~0	Peripheral nervous system
DOID:5214	DemyelinPN	demyelinating polyneuropathy	3,391	69	22.0	2.59E-66	
DOID:5213	CIDMPRN	chronic inflammatory demyelinating polyradiculoneuropathy	2,061	51	26.7	1.14E-53	
DOID:4308	PR Neuropathy	polyradiculoneuropathy	9,689	74	8.2	1.09E-41	
DOID:2491	SP Neuropathy	sensory peripheral neuropathy	3,520	41	12.6	1.12E-30	
DOID:11162	Resp. Failure	respiratory failure	57,385	418	7.9	1.53E-220	Respiratory system disease
DOID:2733	Skin Atrophy	skin atrophy	12,885	80	6.7	1.16E-38	Skin disease
DOID:318	ProgMusc atrophy	progressive muscular atrophy	303	141	501.9	~0	Spinal cord disease
DOID:0050881	IBMPFD	inclusion body myopathy with Paget disease of bone and frontotemporal dementia	105	26	267.1	3.84E-55	Syndrome

Based on the 50 diseases most significantly co-mentioned with ALS in PubMed abstracts (Table 3) an overrepresentation analysis was performed. Disease-associated genes were then overlapped with the ALS core modules to identify connections to other diseases. A total of 37 diseases were overlapping with at least 1 ALS core module (Fig. 4). From the matrix of disease overrepresentation in ALS core modules some clear trends are seen. First of all, the well-known ALS comorbidity *Dementia* is strongly evident from the matrix: *Dementia, broad term* (13 modules), the clinically closely associated *Frontotemporal Dementia* (6 modules, 3 of which Tier 1) and *Lewy Body Dementia* (3 modules). Among the group of other nervous system diseases, the following conditions are also associated with the ALS core modules: *Alzheimer's Disease* (12 modules of which 3 are in Tier 1), *Parkinson's Disease* (8 modules), *Huntington's Disease* (5 modules) and *Muscular Atrophy* (4 modules, 3 of which are in the Tier 1 collection). The remaining diseases have 2 or fewer modules associated – including *Multiple Sclerosis* with only 1 module (184), being significantly associated. The only other non-degenerative CNS-diseases being prominently represented by mostly overlapping ALS core modules are *neuroblastoma* (10 modules) and *toxic encephalopathy*, which is likely due to the many modules described by *apoptosis* GO terms and containing a significant enrichment of brain-associated proteins. It is interesting to note, that motor diseases (such as spastic paraplegia, paraplegia, and Friedreich's ataxia) are not represented by any of the ALS core modules, while muscle diseases, such as atrophic muscular disease, muscular atrophy, myotonic dystrophy type 1 and inclusion body myositis are significantly represented by at least one ALS core module.

ALS-Tier 1 core modules

The Tier 1 core modules were then investigated for known ALS disease biology and proteins associated with other neurodegenerative diseases (Fig. 5–9).

Module 83 – very ALS specific, contains SOD1, links oxidative stress and protein folding, (Fig. 5).

The identification of causative mutations in SOD1 gene was the first evidence of genetically inherited forms of ALS [35]. SOD1, with its many mutations is therefore the best studied protein in this disease and has been linked to two main pathogenic mechanisms which are thought to lead to ALS pathology. Both potential mechanisms are reflected in the underlying biology represented in this network. Mutations in SOD1, a ubiquitously expressed peroxide dismutase, have been linked to oxidative stress, either by a gain of function of this catabolic enzyme or also as a direct regulator of the NADPH dependent oxidation of RAC1 [36]. The network contains many other proteins playing part in the oxidative stress response, therefore the main GO term associated with this network is *oxidative stress* (Fig. 5B). Alternatively, mutations in SOD1 have been reported to induce its misfolding and aggregation (GFER, CCS, PDIA2) and thus to lead to loss of function [37]. Protein misfolding elicit a number of cellular mechanisms to protect the cell against the accumulations of aggregates. Representative of these rescue mechanisms are the large number of heat shock chaperones (for example HSPH1, HSPA2-6, DNAJB2) [38], where PARK7 is by itself redox sensitive. Ubiquitin ligases are also present in the proteasomal pathways (HECW1, STUB1, RNF19a). In the **Tier 1 collection**, module 83 is the most specific to ALS and shows minimal overlap with other neurological diseases. Changes in SOD1 associated function leading to a concomitant deficit in proteostasis may therefore be a unique feature of ALS pathology and its close relative frontotemporal dementia.

Module 196 Represented also in muscular atrophy, linked to protein degradation and apoptosis (F)

Represented also in muscular atrophy, linked to protein degradation and apoptosis (Fig. 6)

Module 196 is centered around HSPB1 (HSP27), which has a variety of functions relevant to ALS. This network shows a molecular link to HSPB1 to the crystallin chaperones, which are Zn^{2+} dependently activated and upregulated in neurological diseases [39]. Crystallin chaperones are also associated with myopathies consistent with their abundant expression in muscle where they stabilize Desmins [40]. HSPB1 oligomerization induced by stress, also TNF induced inflammatory stress. The TNF induced apoptotic signaling pathway is activated through MAPKAP, where HSPB1 deactivates DAXX [41]. Apart from its role in apoptosis, HSPB1 is also important in the proper function of proteasomes and can modulate reactive oxygen species. With this focus on responses to oxidative (and inflammatory stress) this network remains specific for ALS and with its interactive link to the crystallin chaperones makes the muscle particularly sensitive to dysregulation. This is reflected in the link of this network to muscular atrophy and Charcot Marie Tooth disease (Fig. 6) another neuropathy which is characterized by progressive muscular loss and genetic link to HSPB1 [42].

Module 93 Represented also in Alzheimer's disease, *linked to lipoproteins and lipid metabolism* (Fig. 7)

Module 93 is the only Tier 1 network significantly linked to lipid metabolism (Fig. 7, panel B) through the presence lipoprotein receptors (LPRx), which are part of the cholesterol pathway genes as well as the APO protein family. Lipids and Lipoproteins are implicated in a whole range of biological process, where they are involved as energy substrates, building blocks, structural machinery and bioactive molecules [43]. In ALS, and AD, lipid metabolism has been thought to underlay denervation, mitochondrial dysfunction, excitotoxicity neural transport, cytoskeletal defect and impaired neurotransmitter release [43]. In the context of ALS, the energy metabolism, in particular, may have increased needs and in muscle a switch from glucose to lipid energy has been described [44], as well as changes in glycosphingolipids [45]. In addition, the brain strongly depends on fatty acid oxidation [46]. High fat and ketogenic diet in animals prolonged survival, while caloric restriction was detrimental in SOD1 transgenic mice [47, 48]. Therapeutically, this hypothesis has been tested with a high fat diet in a small clinical trial, which suggests that nutritional intervention needs to be followed up [49].

The high number of APO and LRP proteins in this network potentially drives the significant association with Alzheimer's Disease, for which genotype of APOE is the main risk factor. While APOE in Alzheimer has been proposed to play a role in many processes [50], we suggest, based on this ALS core module, that its role in CNS lipid homeostasis is similar in ALS and AD, The use of high fat diets in AD has been discussed controversially, however.

Module 128 – Represented in many neurodegenerative diseases linked to protein metabolism and apoptosis (Fig. 8).

Module 128 represents a large network that contains a wide range of proteins. It is overlapping with 83 (SOD1, HDAC6, BCL2, SQSTM1)) as well as with 196 (SNCA, MAPT) as well as with 21 (HDAC6, SQSTM1, TARDBP). The high degree of disease overrepresentation in this network may be due to the fact that it is the only network that has neuronal function associated proteins, such as the GABA receptors (GABAx), Glutamate receptors (GRIA1,3, GRIN2a) and neuronal related growth factors (NGF) and the microtubular system (HTT, MAPs). Interestingly, however, it is not associated with multiple sclerosis, suggesting that the dysfunction seen in the clinical presentation of MS is more strongly driven by a different mechanism such as immune dysfunction.

Similar to the other networks, this network contains proteins involved in ubiquitination (KEAP, TRIMs). As diseases are often caused by disturbance of homeostatic functions, these stress networks are found in many diseases activated, which may make this network so important also in non-degenerative diseases, such as neuroblastoma.

Module 21 – Represented in many neurodegenerative diseases, highest density of ALS genes, but little significant biology (Fig. 9).

This network is almost exclusively made up of ALS-associated genes. It directly links many ALS-risk genes (HDAC6, VCP, HNRNPA1, SQSTM1, ATXN2) into the same network with the major causative genetic mutations (TARDBP and FUS). In fact, mutations in most of the proteins in this network have been proposed to be linked in one or the other way to ALS. This may strengthen the importance of these genes in the overall ALS pathogenesis [51]. This module *functionally* links many ALS-associated genes into one network which may partly explain why such a large variety of mutations and risk factors lead to the same pathological and clinical features. This network links the two major pathogenetic theories about ALS that are currently discussed: defects in RNA processing [52] and proteasomal malfunction.. Dysfunctional mRNA processing, in addition to loss of function of these transcripts, may lead to an overload of the protein degradation system and thus to cellular dysfunction, independent of the aggregate per se. The finding that TDP-43 (the protein product of TARDBP), is also involved in low molecular weight neurofilament processing and aggregation [53], represents a very interesting insight into how general biological principles can become organ-, here neuron, specific pathologies. In recent years, many neurodegenerative diseases have been recognized and grouped as proteinopathies. Apart from the RNA-related malfunction that comes with mutations of FUS and TARDBP, some studies have recently suggested that these

proteins contain prion like structures [54], which makes them prone to seeding and aggregation with other proteins or lead to dysfunction of the protein degradation pathway causing other proteins to aggregate [55]. In particular TDP43 is also found aggregated in other neurodegenerative diseases [56]. This module is therefore strongly associated with other neurodegenerative diseases that have protein deposits (Fig. 9C).

Discussion

While there is a high rate of new genes that become described as potentially relevant to ALS-disease pathogenesis, their role often remains unclear and confirmation is lacking. Network analysis is one way to link individual proteins into functional networks. Being part of a functional network with a biological relevance for ALS may strengthen the association of a described ALS gene to the disease. In this study we identified 656 proteins that have previously been associated with ALS. Of these, 144 genes were connected in 26 ALS core modules, suggesting a functional association with ALS. Some of the previously less well described genes, such as HDAC6 and SQSTM were shown to be part of the majority of the Tier 1 modules, linking them closely to the well-known ALS genes TARDBP, FUS and SOD1.

Several studies have utilized PPI networks in order to understand the biological complexity underlying ALS. In a 2017 study, Mao et al. [17] found proteins connected to known ALS-causative genes in order to identify common downstream proteins. However, despite the apparent overlap with our approach there are some key differences allowing us to reduce the number of false-positive hits and home in on a much more specific biological interpretation.

Using a bottom-up approach, where networks are generated by including first-order neighbors of the ALS associated proteins and merging highly overlapping networks, provides us with a set of distinct disease modules with a well-defined biological annotation.

Another important difference is the consideration of local degree/global degree ratio in order to handle the problem of over-connected proteins. Highly connected proteins (e.g. UBC) will have a higher chance of showing up in any given network than less connected proteins and are known to introduce noise in network analyses. Filtering out noisy proteins gives a much clearer picture of the key proteins involved in any particular biological process (see Introduction for further details).

A wide range of cellular processes have been implicated in ALS pathogenesis, as reviewed recently [57]. These include neuronal-specific processes, including hyper- and hypo-excitability, glutamate excitotoxicity, and neuronal branching defects [58], proteostasis pathways with impairments in ubiquitin-proteasome systems, autophagy and lysosomal function as well as dysfunction in the endoplasmic reticulum (ER) and mitochondria [59]. More recently, altered RNA processing/metabolism, RNA splicing transcriptional defects have been shown to be linked in particular to TARDBP [60]. Furthermore, dysregulation of cytoskeletal dynamics, leading to impairment of vesicular trafficking including nuclear-cytoplasmic transport [61], between ER and Golgi [62], as well as transport along axons [63] have been found to be part of the ALS pathology.

The highly significant networks, represented in the Tier 1 and Tier 2 modules which we presented in more detail in this study, capture many of these pathophysiological aspects of ALS biology, in particular oxidative stress and proteostasis, but also other plausible biological mechanisms with less research focus such as lipid metabolism (module 93) and neuron specific functions such as neurotransmission and synapse (module 128). Lipid metabolism dysfunctions as a driver of ALS pathology is currently much debated and aberrant lipid metabolism is proposed to underlie denervation of neuromuscular junctions mitochondrial dysfunction, excitotoxicity, impaired neuronal transport, cytoskeletal defects, inflammation and reduced neurotransmitter release [43]. The finding that one of our five ALS-core networks, one is linked to lipid metabolism is strengthening this view on the disease and may support the initiation of clinical trials investigating not only high fat diets, but also modulation of the balance between fatty acids and glucose oxidation [44] and modulation of sphingolipids [64].

The emphasis on proteostasis (protein degradation, modification and folding), which is part of all the presented networks and links genetic evidence (C9orf72, VCP, SQSTM, Dynactin, TBK) to altered chaperone functions (HSPs, crystallins, module 196) and autophagy (lysosomal degradation, modules 21 and 128) is also very interesting. The proteostasis network is a complex regulatory network that maintains protein homeostasis. It consists of several pathways that control protein biosynthesis, folding, trafficking, and degradation and responds to specific protein stress pathways such as the unfolded protein response (UPR), oxidative stress, inflammatory stress with regulation of autophagic lysosomal pathways, chaperones and heat shock response genes [59]. Therefore, dysregulation of this pathway can come therefore from many disturbances associated with ALS. Network 196 is associated with the proteasome pathway via HSPB1, which oligomerizes under oxidative stress, and TNF (regulation of an apoptotic response via DAXX) which links it to oxidative stress and inflammation. Network 21 contains FUS and TARDBP, associating this network with RNA modifications and the autophagy process, which is another component of protein homeostasis [65]. Most of the genetic variations of ALS (C9orf72, TARDBP and FUS) have recently been shown to be prone to aggregation, with possible a prion like mechanisms in ALS [66]. Apart from the loss of proper function of these proteins, these self-propagating aggregations are further impacting the proteasomal system [55]. Collapse of proper proteostasis due to failure to refold, degrade or effectively sequester and compartmentalize aggregation-prone, misfolded or potentially toxic proteins is detrimental to any cell type. Neuronal cells appear to be particularly vulnerable to disturbances in proteostasis most probably because they are long-lived, large post-mitotic cells that are not able to dilute out protein aggregates during cell divisions. In addition, the individual networks contain protein that are over expressed in particular cell types, such as neurobiology associated proteins (module 128) microtubular proteins (module 128), or c (module 196), which suggests that not only neurons suffer from disturbances in these systems but also muscular cells, with high crystallin expression functionally linked to microtubular and intermediate filament integrity [67–69] may directly be impaired.

There are a few genes that are part of the majority of core modules. In particular HDAC6 is found in 4 of the 5 Tier 1 networks, suggesting a central role in many biological pathways underlying ALS biology. HDAC6 plays a role in RNA metabolism, cytoskeletal dynamics and proteodynamics and its regulation in ALS is linked to FUS and TARDBP [70]. Our networks suggest an overlap between these molecular pathways making it difficult to identify a single causative pathway. It is however remarkable that HDAC6, together with FUS and TARDBP is part of the novel core module 21, which is almost exclusively made up of ALS-associated genes. While the biology of this network is only weakly classified, it suggests that mutations in all these genes are leading to a highly overlapping

pathomechanism. Recently, pharmacological inhibition of HDAC6 has been shown to restore axonal transport defects *in vitro* and also ER to Golgi transport by increased acetylation of α -tubulin [71]. Similarly, SQSTM and HSPH1 are found in 3 of the 5 core network modules.

This analysis was performed on the network level and therefore it is not limited to the previously identified ALS-associated proteins. The network modules also contain novel proteins not previously associated with ALS and these proteins are candidates for being identified involved in ALS-related processes as well.

An investigation of the proteins identified through an updated text mining and database search revealed, that while none of the new proteins from ALSod were found in the Tier 1 or Tier 2 disease modules, 11 proteins from HuGE Navigator and 11 proteins identified through text mining of the scientific literature were found in at least one of the Tier 1 or Tier 2 modules (Supplementary Table S 2). This finding further strengthens our approach and supports a functional role of these new genes in ALS.

It is remarkable that most of the genes currently associated with ALS are found in the five Tier 1 disease modules. This suggests that there is a functional link between these mutated proteins, which leads to the common clinical phenotype of ALS, independent of the individual mutations. In this context network 21 is highly interesting, as it suggests a direct link between genes involved in RNA modifications and genes that are part of protein homeostasis, which is currently the most discussed mechanism underlying the pathogenesis of ALS.

TARDBP is a good example how deregulation of the proper function of a single protein is linking to many intracellular pathways. In line with its nuclear and cytoplasmic functions TDP-43 is pivotal in multiple cellular functions from RNA processing steps to misfolding and granule formation in the cytoplasm. Its pathological translocation to the cytoplasm on the one hand leads to a loss of nuclear function with dysregulated transcription, splicing, stabilization and RNA transport downstream leading to a dysregulation of a large number of dependent proteins. Its pathological presence in the cytoplasm on the other hand leads to a gain of function with formation of toxic aggregates (stress granules), an overload of the misfolded Protein response leading to a proteinopathy, which is present in 95% of all ALS cases [72]. Along these lines, this recent review puts the TARDBP-linked proteinopathy in the center of a spectrum of neurodegenerative diseases [73].

It is also interesting that C9orf72, which is frequently mutated in ALS, is not part of network 21 or any of the other core networks. This could suggest that the gain of toxic function due to the repeats in the intronic part of the *C9orf72* gene resulting in the formation of Dipeptide Repeats (DPR) might be more prominent than the concomitant loss of physiological function of the C9orf72 protein in autophagy. In addition, this RNA repeats- and DPR-mediated toxic mechanism is likely to be part of another biological process than the one covered by network 21.

When interpreting the disease overlap of the modules (Fig. 4), it is important to consider that the genes associated with each disease were found using text mining of the scientific literature, and that the diseases investigated are known to often be mentioned in ALS related abstracts (Table 3). In the case of this study, extra care must be taken since all 50 diseases we used for comparison, were defined as the top 50 diseases most often being co-mentioned with ALS. However, this will also be true for any study working with well-known associations such as FD, AD, PD, HD and MS.

Following this word of caution, the most striking observation is actually the lack of overlap with most of the top 50 diseases, as mentioned in the results section. For example, only a single Tier 2 module has an overlap with MS biology, while 12, including 3 of the 5 Tier1 have an overlap with AD. This indicates that the overlap matrix in Fig. 4 is not just driven by a generic overlap in literature between ALS and the top 50 diseases, but it represents a clear trend towards the overlap with a specific subset of diseases.

Biologically, it is interesting that these core modules are so selectively represented in specific diseases. Oxidative stress shows a certain specificity to ALS (and associated FTD), while only module 128, with its many neurobiology associated genes shows a broader overlap. The modules often overlap with AD, either exclusively or together with other conditions, or they show a representation with muscular atrophy. This is an important hint on how general cellular processes may lead to specific diseases, and what pathways may be particularly vulnerable in both muscle and neurons such as the crystallin chaperone system (module 196). An interesting finding is the relatively high overlap of our core ALS networks in toxic encephalopathy. Toxic encephalopathy is a heterogeneous clinical disease of brain dysfunction caused by toxic substances of a wide variety [74]. MalaCards [75] reports several biological pathways (perinuclear transport, neuronal projection and membrane raft proteins) as targets of neurotoxic substances leading to clinical problem of encephalopathy. Many of these mechanisms are also central to ALS. This may support the speculation that pesticides and other environmental toxins can lead to ALS, which has been suggested as an explanation for the high incidence of ALS among military people and certain population groups as the Chamorro people of Guam through the exposure to the β -methylamino-L-alanine (BMAA) found in fishes [76].

Conclusions

In the case of complex diseases, discovering and describing the molecular systems responsible for the phenotype is extremely difficult, since a complex disease is not caused by a single gene, but is rather a perturbation of a biological system. Since the disease-causing genetic factors differ between individuals, it is crucial to understand how they are connected. Network analysis is an effective approach for investigating the functional interactions between molecules. Analyzing a comprehensive ALS dataset in context of protein-protein interactions allowed us to get a unique top-level understanding of ALS biology. By consolidating the networks into modules with known players in focus, we were able to extract a comprehensive and rich set of ALS modules.

When focusing on the five most significant modules (Tier 1), the represented biology is covering the main hypothesis around the pathogenesis of ALS, including oxidative stress, energy metabolism, proteasome dysfunctions and mRNA processing changes. Some of the modules are generic and shared with other neurological diseases. These involve the functional response to stress, be it oxidative or linked to protein- or energy metabolism. Many known ALS mutations lead to a dysregulation of the proper production of proteins, potentially starting at mRNA processing and resulting in a disturbed proteostasis, with a central role of TARDBP. As neurons may be particularly sensitive to these failures, several networks are part of other neurodegenerative diseases (modules 128

and 21) and may be the molecular basis of a proteinopathy spectrum of diseases from dementias to muscular atrophies. Other networks, in particular the ones around SOD, are associated only with ALS (modules 83, 93 and 196). A follow-up study including recently identified ALS-associated proteins found that the networks we defined as “Core Modules” (Tier 1 and Tier 2) contained a significantly higher proportion of new ALS genes than expected.

List Of Abbreviations

AD Alzheimer’s Disease

ALS Amyotrophic Lateral Sclerosis

ALSoD ALS online Database

AP-MS Affinity Purified Mass Spectrometry

CNS Central Nervous System

ER Endoplasmatic Reticulum

fALS Familial Amyotrophic lateral sclerosis

FTD Frontotemporal dementia

GWAS Genome Wide Association Studies

HD Huntigton’s Disease

KEGG Kyoto Encyclopedia of Genes and Genomes, <https://www.genome.jp/kegg/>

LBD Lewy Body Dementia

MND Motor Neuron Disease

MAPs Microtubule Associated Proteins

MS Multiple Sclerosis

PD Parkinson’s Disease

PPI Protein-protein interaction

sALS Sporadic Amyotrophic lateral sclerosis

UniProt A hub for protein Information, www.uniprot.org

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data generated or analyzed during this study are included in this published article [and its supplementary information files]

Competing interests

DHH, KHL and RWE are employees of Intomics A/S, and RWE is a co-founder of the company. The inBio Map™ and inBio Know™ data resources used in the study are both produced by Intomics A/S.

AS, TCRH and PG are present and former employees of Actelion Pharmaceuticals and Idorsia Pharmaceuticals.

Funding

The study was partially funded by Actelion Pharmaceuticals Ltd. and Idorsia Pharmaceutical Ltd. with Intomics A/S providing additional computational analyses.

Authors' contributions

DHH, KHJ and RW performed the data analysis with AS and PG driving the biological interpretation of the results. TCRH contributed with ideas for figures and analyses. The study was initiated by and lead by PG and AS with DHH, KHJ and RW driving the algorithmic development. The manuscript was drafted by RW, KHJ, AS and PG. All authors read and approved the manuscript

Acknowledgements

The authors would like to thank Elodie Angot and Michel Steiner for their critical review of and useful comments on the manuscript.

References

1. Robberecht W, Philips T. The changing scene of amyotrophic lateral sclerosis. *Nat Rev Neurosci*. 2013;14:248–64. doi:10.1038/nrn3430.
2. Andersen PM, Al-Chalabi A. Clinical genetics of amyotrophic lateral sclerosis: what do we really know? *Nat Rev Neurol*. 2011;7:603–15. doi:10.1038/nrneurol.2011.150.
3. National Library of Medicine. ghr.nlm.nih.gov. Accessed 1 Jun 2020.
4. Morgan S, Shoai M, Fratta P, Sidle K, Orrell R, Sweeney MG, et al. Investigation of next-generation sequencing technologies as a diagnostic tool for amyotrophic lateral sclerosis. *Neurobiol Aging*. 2015;36:1600.e5-1600.e8. doi:10.1016/j.neurobiolaging.2014.12.017.
5. Chia R, Chiò A, Traynor BJ. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. *Lancet Neurol*. 2018;17:94–102. doi:10.1016/S1474-4422(17)30401-5.
6. Jellinger KA, Korczyn AD. Are dementia with Lewy bodies and Parkinson's disease dementia the same disease? *Bmc Med*. 2018;16:34,34.
7. Strong MJ, Abrahams S, Goldstein LH, Woolley S, McLaughlin P, Snowden J, et al. Amyotrophic lateral sclerosis - frontotemporal spectrum disorder (ALS-FTSD): Revised diagnostic criteria. *Amyotroph Lateral Scler Front Degener*. 2017;18:153–74. doi:10.1080/21678421.2016.1267768.
8. Yedavalli VS, Patil A, Shah P. Amyotrophic Lateral Sclerosis and its Mimics/Variants: A Comprehensive Review. *J Clin Imaging Sci*. 2018;8:53,53.
9. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. *Cell*. 2011;144:986–98. doi:10.1016/j.cell.2011.02.016.
10. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47:D590–5. doi:10.1093/nar/gky962.
11. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*. 2018;46:D661–7. doi:10.1093/nar/gkx1064.
12. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2014;42:D472–7. doi:10.1093/nar/gkt1102.
13. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkovic G, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat Methods*. 2017;14:61–4. doi:10.1038/nmeth.4083.
14. Rahmati S, Abovsky M, Pastrello C, Jurisica I. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. *Nucleic Acids Res*. 2017;45:D419–26. doi:10.1093/nar/gkw1082.
15. Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, et al. The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat Methods*. 2013;10:730–6. doi:10.1038/nmeth.2557.
16. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411:41–2. doi:10.1038/35075138.
17. Mao Y, Kuo S-W, Chen L, Heckman CJ, Jiang MC. The essential and downstream common proteins of amyotrophic lateral sclerosis: A protein-protein interaction network analysis. *PLoS One*. 2017;12:e0172246. doi:10.1371/journal.pone.0172246.
18. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2. doi:10.1186/1471-2105-4-2.
19. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods*. 2012;9:471–2. doi:10.1038/nmeth.1938.
20. Apweiler R. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2004;32:115D – 119. doi:10.1093/nar/gkh131.
21. Wroe R, Wai-Ling Butler A, Andersen PM, Powell JF, Al-Chalabi A. ALSOD: The Amyotrophic Lateral Sclerosis Online Database. *Amyotroph Lateral Scler*. 2008;9:249–50. doi:10.1080/17482960802146106.
22. ALSgene. http://www.alsgene.org/top_results. Accessed 2 May 2014.
23. Wall DP, Pivovarov R, Tong M, Jung J-Y, Fusaro VA, DeLuca TF, et al. Genotator: A disease-agnostic tool for genetic annotation of disease. *BMC Med Genomics*. 2010;3:50. doi:10.1186/1755-8794-3-50.
24. Becker KG, Barnes KC, Bright TJ, Wang SA. The Genetic Association Database. *Nat Genet*. 2004;36:431–2. doi:10.1038/ng0504-431.

25. GAD full data set. <https://geneticassociationdb.nih.gov/data.zip>. Accessed 2 May 2014.
26. Yu W, Yesupriya A, Wulf A, Hindorf LA, Dowling N, Khoury MJ, et al. GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur J Hum Genet.* 2011;19:1095–9. doi:10.1038/ejhg.2011.91.
27. Yu W, Wulf A, Liu T, Khoury MJ, Gwinn M. Gene Prospector: An evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics.* 2008;9:528. doi:10.1186/1471-2105-9-528.
28. inBio Know - text mining site. <https://inbio-discover.com/know.html>. Accessed 20 Jul 2017.
29. Shannon P. Cytoscape. A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 2003;13:2498–504. doi:10.1101/gr.1239303.
30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25:25–9. doi:10.1038/75556.
31. Carbon S. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 2017;45:D331–8. doi:10.1093/nar/gkw1108.
32. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 2012;40:D940–6. doi:10.1093/nar/gkr972.
33. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* 2015;43:D1071–8. doi:10.1093/nar/gku1011.
34. Diaz-Ortiz ME, Chen-Plotkin AS. Omics in Neurodegenerative Disease: Hope or Hype? *Trends Genet.* 2020;36:152–9. doi:10.1016/j.tig.2019.12.002.
35. Rosen DR, Siddique T, Patterson D, Figlewicz DA, Sapp P, Hentati A, et al. Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature.* 1993;362:59–62,59–62.
36. Harraz MM, Marden JJ, Zhou W, Zhang Y, Williams A, Sharov VS, et al. SOD1 mutations disrupt redox-sensitive Rac regulation of NADPH oxidase in a familial ALS model. *J Clin Invest.* 2008;118. doi:10.1172/JCI34060.
37. Seetharaman SV, Prudencio M, Karch C, Holloway SP, Borchelt DR, Hart PJ. Immature copper-zinc superoxide dismutase and familial amyotrophic lateral sclerosis. *Exp Biol Med.* 2009;234:1140–54,1140–1154.
38. Kalmar B, Lu C-H, Greensmith L. The role of heat shock proteins in Amyotrophic Lateral Sclerosis: The therapeutic potential of Arimocloamol. *Pharmacol Ther.* 2014;141:40–54.
39. Capponi S, Geuens T, Geroldi A, Origone P, Verdiani S, Cichero E, et al. Molecular Chaperones in the Pathogenesis of Amyotrophic Lateral Sclerosis: The Role of HSPB1. *Hum Mutat.* 2016;37:1202–8,1202–1208.
40. Wang J, Martin E, Gonzales V, Borchelt DR, Lee MK. Differential regulation of small heat shock proteins in transgenic mouse models of neurodegenerative diseases. *Neurobiol Aging.* 2008;29:586–97,586–597.
41. Arrigo A-P. Mammalian HspB1 (Hsp27) is a molecular sensor linked to the physiology and environment of the cell. *Cell Stress Chaperones.* 2017;22:517–29.
42. Weeks SD, Muranova LK, Heirbaut M, Beelen S, Strelkov SV, Gusev NB. Characterization of human small heat shock protein HSPB1 α -crystallin domain localized mutants associated with hereditary motor neuron diseases. *Sci Rep.* 2018;8:688. doi:10.1038/s41598-017-18874-x.
43. Tracey TJ, Steyn FJ, Wolvetang EJ, Ngo ST. Neuronal Lipid Metabolism: Multiple Pathways Driving Functional Outcomes in Health and Disease. *Front Mol Neurosci.* 2018;11. doi:10.3389/fnmol.2018.00010.
44. Palamiuc L, Schlagowski A, Ngo ST, Vernay A, Dirrig-Grosch S, Henriques A, et al. A metabolic switch toward lipid use in glycolytic muscle is an early pathologic event in a mouse model of amyotrophic lateral sclerosis. *EMBO Mol Med.* 2015;7:526–46. doi:10.15252/emmm.201404433.
45. Henriques A, Huebecker M, Blasco H, Keime C, Andres CR, Corcia P, et al. Inhibition of β -Glucocerebrosidase Activity Preserves Motor Unit Integrity in a Mouse Model of Amyotrophic Lateral Sclerosis. *Sci Rep.* 2017;7:5235. doi:10.1038/s41598-017-05313-0.
46. Ebert D, Haller RG, Walton ME. Energy Contribution of Octanoate to Intact Rat Brain Metabolism Measured by ^{13}C Nuclear Magnetic Resonance Spectroscopy. *J Neurosci.* 2003;23:5928–35. doi:10.1523/JNEUROSCI.23-13-05928.2003.
47. Dupuis L, Oudart H, Rene F, de Aguilar J-LG, Loeffler J-P. Evidence for defective energy homeostasis in amyotrophic lateral sclerosis: Benefit of a high-energy diet in a transgenic mouse model. *Proc Natl Acad Sci.* 2004;101:11159–64. doi:10.1073/pnas.0402026101.
48. Mattson MP, Cutler RG, Camandola S. Energy Intake and Amyotrophic Lateral Sclerosis. *NeuroMolecular Med.* 2007;9:17–20. doi:10.1385/NMM:9:1:17.
49. Wills A-M, Hubbard J, Macklin EA, Glass J, Tandan R, Simpson EP, et al. Hypercaloric enteral nutrition in patients with amyotrophic lateral sclerosis: a randomised, double-blind, placebo-controlled phase 2 trial. *Lancet.* 2014;383:2065–72. doi:10.1016/S0140-6736(14)60222-1.
50. Verghese PB, Castellano JM, Holtzman DM. Apolipoprotein E in Alzheimer's disease and other neurological disorders. *Lancet Neurol.* 2011;10:241–52. doi:10.1016/S1474-4422(10)70325-2.
51. Polymenidou M, Lagier-Tourenne C, Hutt KR, Huelga SC, Moran J, Liang TY, et al. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat Neurosci.* 2011;14:459–68. doi:10.1038/nn.2779.
52. Gascon E, Gao F-B. The Emerging Roles of MicroRNAs in the Pathogenesis of Frontotemporal Dementia–Amyotrophic Lateral Sclerosis (FTD-ALS) Spectrum Disorders. *J Neurogenet.* 2014;28:30–40. doi:10.3109/01677063.2013.876021.
53. Oberstadt M, Claßen J, Arendt T, Holzer M. TDP-43 and Cytoskeletal Proteins in ALS. *Mol Neurobiol.* 2018;55:3143–51. doi:10.1007/s12035-017-0543-1.
54. Iglesias V, Paladin L, Juan-Blanco T, Pallarès I, Aloy P, Tosatto SCE, et al. In silico Characterization of Human Prion-Like Proteins: Beyond Neurological Diseases. *Front Physiol.* 2019;10. doi:10.3389/fphys.2019.00314.

55. McAlary L, Plotkin SS, Yerbury JJ, Cashman NR. Prion-Like Propagation of Protein Misfolding and Aggregation in Amyotrophic Lateral Sclerosis. *Front Mol Neurosci*. 2019;12. doi:10.3389/fnmol.2019.00262.
56. Porta S, Xu Y, Restrepo CR, Kwong LK, Zhang B, Brown HJ, et al. Patient-derived frontotemporal lobar degeneration brain extracts induce formation and spreading of TDP-43 pathology in vivo. *Nat Commun*. 2018;9. doi:10.1038/s41467-018-06548-9.
57. Balendra R, Isaacs AM. C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nat Rev Neurol*. 2018;14:544–58. doi:10.1038/s41582-018-0047-2.
58. Fogarty MJ. Driven to decay: Excitability and synaptic abnormalities in amyotrophic lateral sclerosis. *Brain Res Bull*. 2018;140:318–33. doi:10.1016/j.brainresbull.2018.05.023.
59. Webster CP, Smith EF, Shaw PJ, De Vos KJ. Protein Homeostasis in Amyotrophic Lateral Sclerosis: Therapeutic Opportunities? *Front Mol Neurosci*. 2017;10. doi:10.3389/fnmol.2017.00123.
60. Dorn T, Kornherr J, Parrotta EI, Zawada D, Ayetey H, Santamaria G, et al. Interplay of cell–cell contacts and RhoA/ MRTF -A signaling regulates cardiomyocyte identity. *EMBO J*. 2018;37. doi:10.15252/embj.201798133.
61. Kim HJ, Taylor JP. Lost in Transportation: Nucleocytoplasmic Transport Defects in ALS and Other Neurodegenerative Diseases. *Neuron*. 2017;96:285–97. doi:10.1016/j.neuron.2017.07.029.
62. Soo KY, Halloran M, Sundaramoorthy V, Parakh S, Toth RP, Southam KA, et al. Rab1-dependent ER–Golgi transport dysfunction is a common pathogenic mechanism in SOD1, TDP-43 and FUS-associated ALS. *Acta Neuropathol*. 2015;130:679–97. doi:10.1007/s00401-015-1468-2.
63. De Vos KJ, Hafezparast M. Neurobiology of axonal transport defects in motor neuron diseases: Opportunities for translational research? *Neurobiol Dis*. 2017;105:283–99. doi:10.1016/j.nbd.2017.02.004.
64. Dodge JC, Treleaven CM, Pacheco J, Cooper S, Bao C, Abraham M, et al. Glycosphingolipids are modulators of disease pathogenesis in amyotrophic lateral sclerosis. *Proc Natl Acad Sci*. 2015;112:8100–5. doi:10.1073/pnas.1508767112.
65. Wang D, Peng Z, Ren G, Wang G. The different roles of selective autophagic protein degradation in mammalian cells. *Oncotarget*. 2015;6:37098–116. doi:10.18632/oncotarget.5776.
66. Verma A. Protein aggregates and regional disease spread in ALS is reminiscent of prion-like pathogenesis. *Neurol India*. 2013;61:107–10. doi:10.4103/0028-3886.111109.
67. Dubin RA, Wawrousek EF, Piatigorsky J. Expression of the murine alpha B-crystallin gene is not restricted to the lens. *Mol Cell Biol*. 1989;9:1083 LP – 1091. doi:10.1128/MCB.9.3.1083.
68. Ito H, Kamei K, Iwamoto I, Inaguma Y, Tsuzuki M, Kishikawa M, et al. Hsp27 suppresses the formation of inclusion bodies induced by expression of R120GαB-crystallin, a cause of desmin-related myopathy. *Cell Mol Life Sci C*. 2003;60:1217–23. doi:10.1007/s00018-003-3024-9.
69. Brady JP, Garland DL, Green DE, Tamm ER, Giblin FJ, Wawrousek EF. αB-Crystallin in Lens Development and Muscle Integrity: A Gene Knockout Approach. *Invest Ophthalmol Vis Sci*. 2001;42:2924–34.
70. Iaconelli J, Xuan L, Karmacharya R. HDAC6 Modulates Signaling Pathways Relevant to Synaptic Biology and Neuronal Differentiation in Human Stem-Cell-Derived Neurons. *Int J Mol Sci*. 2019;20:1605. doi:10.3390/ijms20071605.
71. Guo W, Van Den Bosch L. Therapeutic potential of HDAC6 in amyotrophic lateral sclerosis. *Cell Stress*. 2018;2:14–6. doi:10.15698/cst2018.01.120.
72. Aulas A, Vande Velde C. Alterations in stress granule dynamics driven by TDP-43 and FUS: a link to pathological inclusions in ALS? *Front Cell Neurosci*. 2015;9:423. doi:10.3389/fncel.2015.00423.
73. de Boer EMJ, Orié VK, Williams T, Baker MR, De Oliveira HM, Polvikoski T, et al. TDP-43 proteinopathies: a new wave of neurodegenerative diseases. *J Neurol Neurosurg Psychiatry*. 2020;:jnnp-2020-322983. doi:10.1136/jnnp-2020-322983.
74. Kim Y, Kim JW. Toxic Encephalopathy. *Saf Health Work*. 2012;3:243–56. doi:10.5491/SHAW.2012.3.4.243.
75. MalaCards. <https://www.malacards.org>. Accessed 1 Jul 2020.
76. Su F-C, Goutman SA, Chernyak S, Mukherjee B, Callaghan BC, Batterman S, et al. Association of Environmental Toxins With Amyotrophic Lateral Sclerosis. *JAMA Neurol*. 2016;73:803. doi:10.1001/jamaneurol.2016.0594.
77. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*. 2002;417:399–403. doi:10.1038/nature750.

Supplementary

Supplemental Files S1 and S2 are not available with this version.

Figures

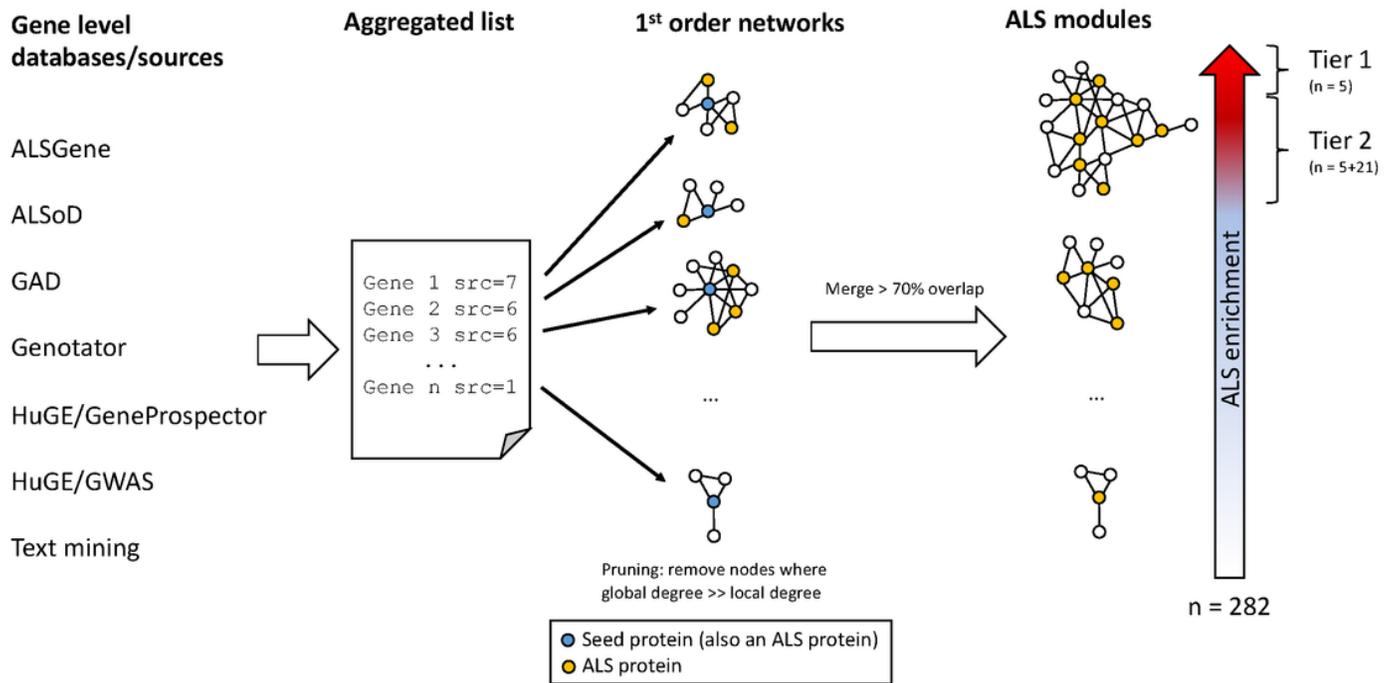


Figure 1

Analysis workflow Outline of the workflow followed in this study. Briefly, lists of genes associated to ALS were collected from 7 data sources, and combined into an aggregated list, keeping note of how many sources supported each gene. Next, genes were mapped to their corresponding protein products, and for each protein, a protein-protein interaction (PPI) network of 1st order interaction partners was found using the inBio Map PPI database. Networks with a large overlap were merged, producing a final set of 282 "ALS modules". Finally, the strength of the association to ALS for each module was investigated via over-representation analysis for enrichment of ALS associated proteins. From the over-representation analysis we defined 26 modules to be "Core modules" (subdivided into two tiers of significance).

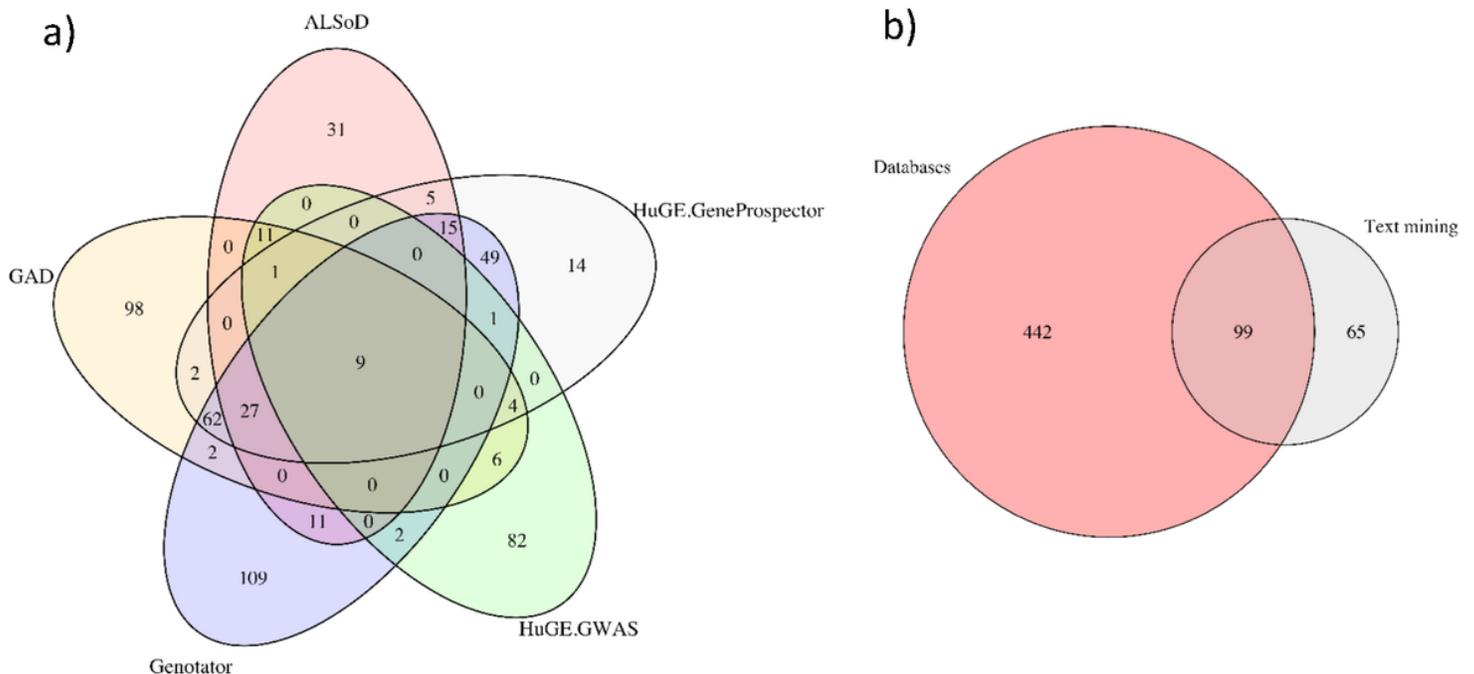


Figure 2

Page 20/26

Overlap between ALS resources Overlap between proteins suggested to be ALS-associated. Panel A: Overlap between the 5 largest ALS databases. Panel B: Overlap between the combined databases and the text mining-based list.

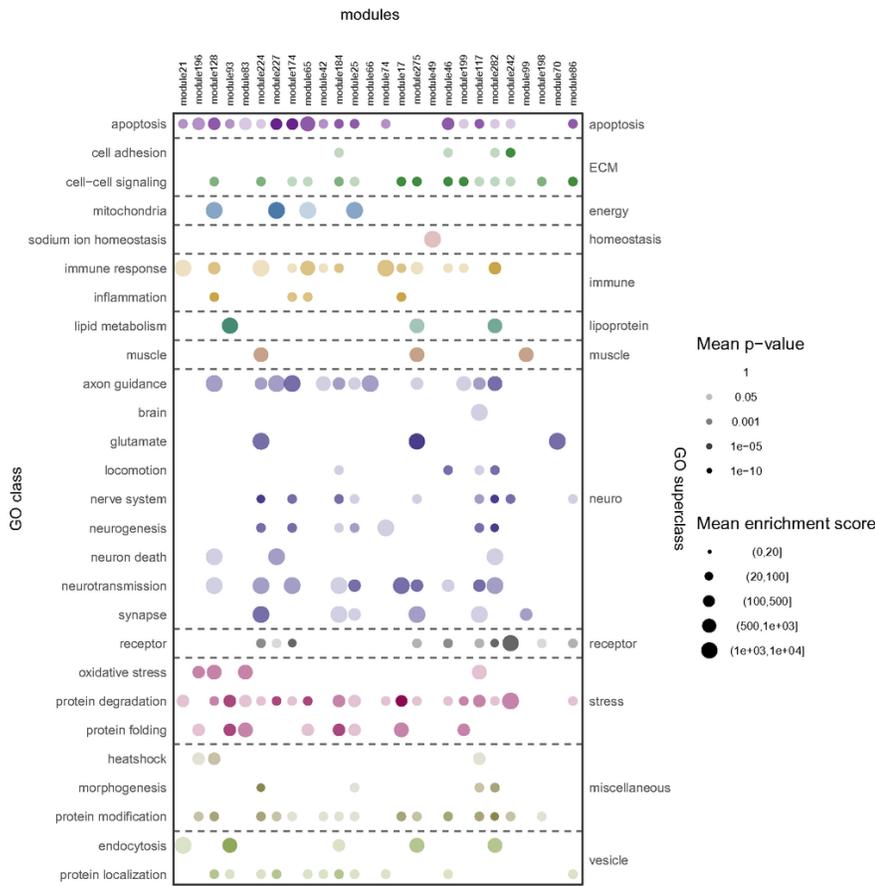


Figure 3

Relevant Gene Ontology categories overrepresented in ALS core modules Top row: Module IDs – sorted on significance of ALS overrepresentation (most significant to the left). Left: Significantly overrepresented Gene Ontology categories were manually collected in GO classes. GO class significance and enrichment score is calculated as the geometric mean of the p-values and mean of the enrichment scores, respectively, for all contained GO categories. Right: GO super-classes is a top-level descriptor for the GO classes and categories.

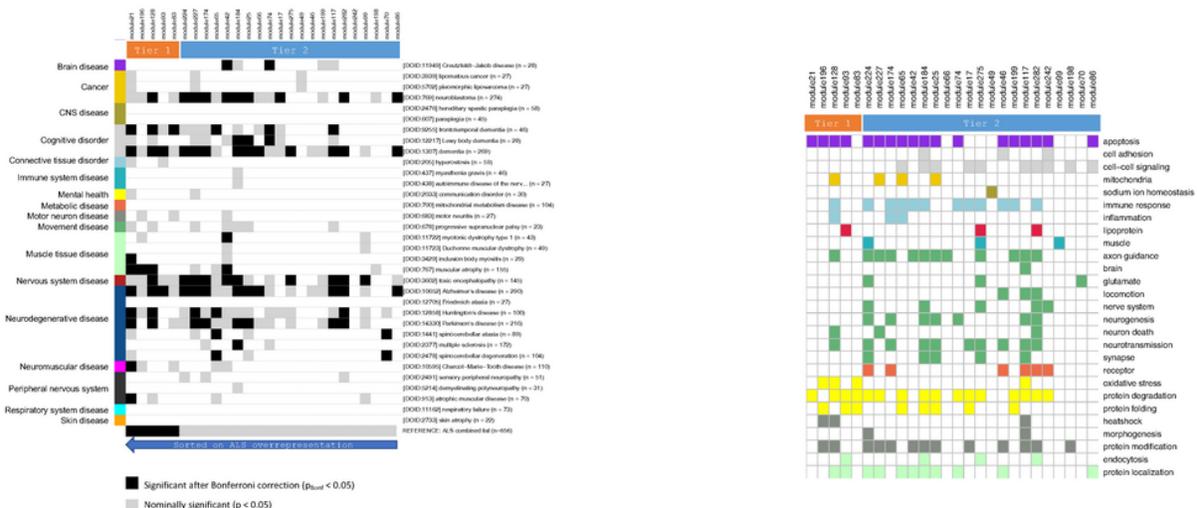
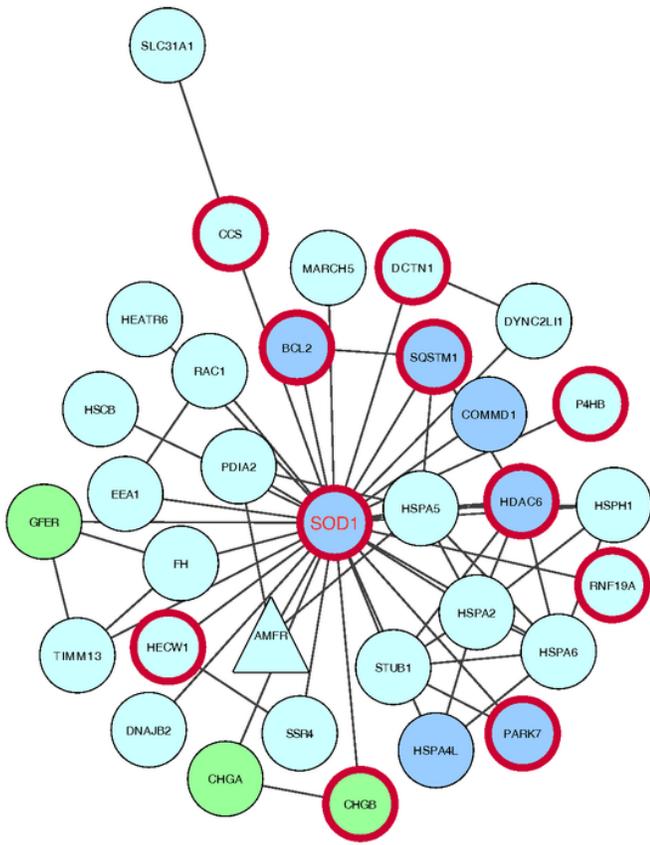


Figure 4

Disease overlap matrix for ALS core modules Top row: Module IDs – sorted on significance of ALS overrepresentation (most significant to the left), and Tier 1 / Tier 2 subsets indicated. Right side: Disease description: Disease Ontology ID, name (truncated to 35 letters), number of genes/proteins associated to the disease via text mining. Only diseases with a least 20 associated genes/proteins included in this figure. Left: Disease categories as defined in Table 3.

A) Module 83: Overview



B) Module 83: Top 25 biological processes

GO accession	Name	GO size	Overlap	Enrichment	p-value
GO:0035966	response to topologically incorrect protein	148	9	38.70	6.12E-11
GO:0006886	response to unfolded protein	137	7	32.52	6.80E-08
GO:0034976	response to endoplasmic reticulum stress	183	6	20.87	9.51E-06
GO:0015036	disulfide oxidoreductase activity	26	4	97.91	9.66E-06
GO:1903201	regulation of oxidative stress-induced cell death	33	4	77.14	1.99E-05
GO:0044087	regulation of cellular component biogenesis	593	8	8.59	3.15E-05
GO:0051130	positive regulation of cellular component organization	673	8	7.57	7.00E-05
GO:0016567	protein ubiquitination	473	7	9.42	8.47E-05
GO:0031396	regulation of protein ubiquitination	150	5	21.21	8.97E-05
GO:1900409	positive regulation of cellular response to oxidative stress	11	3	173.57	1.52E-04
GO:0042026	protein refolding	13	3	146.86	2.10E-04
GO:0033043	regulation of organelle organization	828	8	6.15	2.53E-04
GO:0032434	regulation of proteasomal ubiquitin-dependent protein catabolic process	83	4	30.67	3.10E-04
GO:1902175	regulation of oxidative stress-induced intrinsic apoptotic signaling pathway	17	3	112.31	3.54E-04
GO:0051129	negative regulation of cellular component organization	403	6	9.48	3.93E-04
GO:2000377	regulation of reactive oxygen species metabolic process	96	4	26.52	4.74E-04
GO:0035967	cellular response to topologically incorrect protein	106	4	24.02	6.31E-04
GO:0061630	ubiquitin protein ligase activity	108	4	23.57	6.67E-04
GO:0000209	protein polyubiquitination	259	5	12.29	7.13E-04
GO:0043254	regulation of protein-containing complex assembly	284	5	11.20	1.00E-03
GO:0051129	cellular response to oxidative stress	136	4	18.72	1.29E-03
GO:0055114	oxidation-reduction process	833	7	5.35	1.70E-03
GO:0051603	proteolysis involved in cellular protein catabolic process	341	5	9.33	1.96E-03
GO:0031334	positive regulation of protein-containing complex assembly	160	4	15.91	2.05E-03
GO:0051881	regulation of mitochondrial membrane potential	43	3	44.40	2.19E-03

C) Module 83: Disease overlap



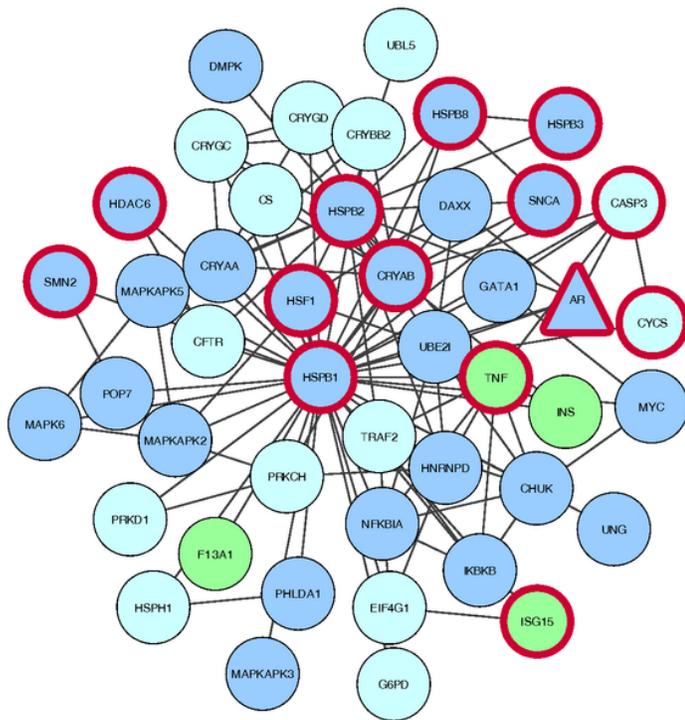
Borderline:
 AD
 PD
 Toxic Enceph.
 Neuroblastoma

Frontotemporal dementia

Figure 5

Module 83 – Protein folding and oxidative stress Panel A: thick red border: gene/protein annotated to be ALS-associated on any of the 7 lists. Large red label: gene/protein with GWAS evidence. Triangular shaped nodes: receptors (UniProt keyword). Node colors indicate subcellular location (UniProt keywords): Blue = Nuclear, Green = Secreted, Purple = conflicting nuclear/secreted, cyan = neither nuclear nor secreted (likely cytosolic). Panel B: Gene Ontology overrepresentation analysis. Panel C: Thick border: ALS-associated (as panel A), Colored node: Disease-associated gene/protein.

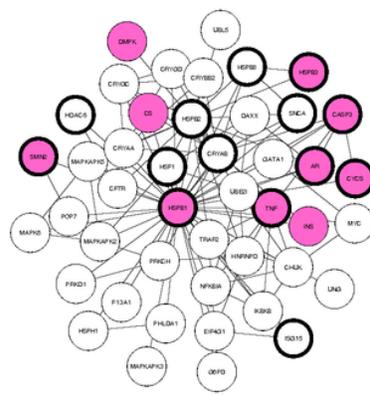
A) Module 196: Overview



B) Module 196: Top 25 biological processes

GO accession	Name	GO size	Overlap	Enrichment	p-value
GO:0060548	negative regulation of cell death	554	14	11.44	1.31E-10
GO:0043066	negative regulation of apoptotic process	495	12	10.97	7.72E-09
GO:0051092	positive regulation of NF-kappaB transcription factor activity	127	8	28.51	1.19E-08
GO:0042327	positive regulation of phosphorylation	679	13	8.66	1.84E-08
GO:0035556	intracellular signal transduction	956	14	6.63	9.12E-08
GO:0001934	positive regulation of protein phosphorylation	647	12	8.39	1.19E-07
GO:0006468	protein phosphorylation	822	13	7.16	1.48E-07
GO:000135	regulation of cellular response to stress	525	11	9.48	1.82E-07
GO:000165	MAPK cascade	315	9	12.93	4.02E-07
GO:0051347	positive regulation of transferase activity	448	10	10.10	4.99E-07
GO:0034599	cellular response to oxidative stress	136	7	23.29	5.37E-07
GO:0031329	regulation of cellular catabolic process	600	11	8.30	5.87E-07
GO:0034097	response to cytokine	760	12	7.15	5.90E-07
GO:0035966	response to topologically incorrect protein	148	7	21.40	8.81E-07
GO:0033674	positive regulation of kinase activity	390	9	10.44	1.99E-06
GO:0004674	protein serine/threonine kinase activity	409	9	9.96	2.83E-06
GO:0007249	I-kappaB kinase/NF-kappaB signaling	47	5	48.14	3.62E-06
GO:0006915	apoptotic process	739	11	6.74	3.82E-06
GO:0071902	positive regulation of protein serine/threonine kinase activity	226	7	14.02	1.01E-05
GO:1902533	positive regulation of intracellular signal transduction	661	10	6.85	1.20E-05
GO:0006986	response to unfolded protein	137	6	19.82	1.29E-05
GO:0032147	activation of protein kinase activity	244	7	12.98	1.56E-05
GO:0051129	negative regulation of cellular component organization	403	8	8.98	2.72E-05
GO:1900034	regulation of cellular response to heat	78	5	29.01	2.72E-05
GO:0035690	cellular response to drug	167	6	16.26	3.35E-05

C) Module 196: Disease overlap



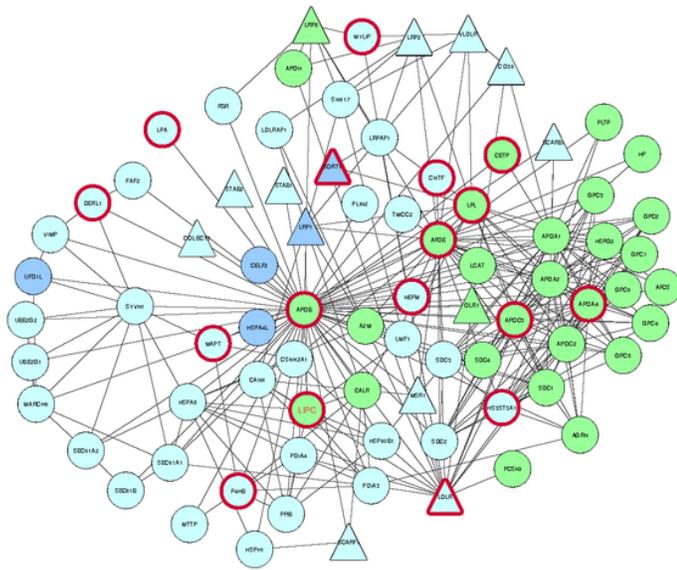
Borderline:
 AD
 HD
 Toxic Enceph.
 Neuroblastoma

Muscular Atrophy

Figure 6

Module 196 – apoptosis and oxidative stress related Panel A: thick red border: gene/protein annotated to be ALS-associated on any of the 7 lists. Large red label: gene/protein with GWAS evidence. Triangular shaped nodes: receptors (UniProt keyword). Node colors indicate subcellular location (UniProt keywords): Blue = Nuclear, Green = Secreted, Purple = conflicting nuclear/secreted, cyan = neither nuclear nor secreted (likely cytosolic). Panel B: Gene Ontology overrepresentation analysis. Panel C: Thick border: ALS-associated (as panel A), Colored node: Disease-associated gene/protein.

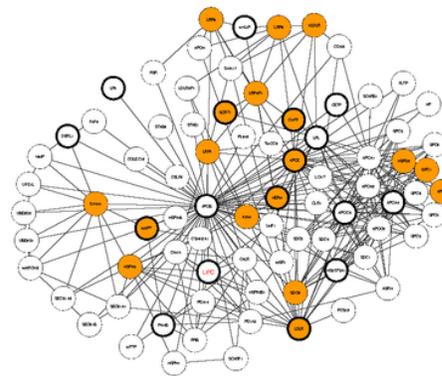
A) Module 93: Overview



B) Module 93: Top 25 biological processes

GO accession	Name	GO size	Overlap	Enrichment	p-value
GO:0034378	chylomicron assembly	9	9	257.78	0.00E+00
GO:0034372	very low-density lipoprotein particle remodeling	7	7	257.78	0.00E+00
GO:0030229	very low-density lipoprotein particle receptor activity	3	3	257.78	0.00E+00
GO:0010903	negative regulation of very-low-density lipoprotein particle remodeling	3	3	257.78	0.00E+00
GO:0001523	retinoid metabolic process	89	23	66.62	1.24E-34
GO:0009706	regulation of plasma lipoprotein particle levels	60	21	90.22	1.69E-34
GO:0006629	lipid metabolic process	962	38	10.18	3.21E-28
GO:00071827	plasma lipoprotein particle organization	36	15	107.41	1.95E-25
GO:0034369	plasma lipoprotein particle remodeling	22	13	152.33	3.42E-24
GO:0034370	triglyceride-rich lipoprotein particle remodeling	12	11	236.30	5.68E-23
GO:0042632	cholesterol homeostasis	38	14	94.97	8.45E-23
GO:0034381	plasma lipoprotein particle clearance	35	13	95.75	3.99E-21
GO:0006898	receptor-mediated endocytosis	176	19	27.83	5.31E-21
GO:0030301	cholesterol transport	41	13	81.74	3.67E-20
GO:1901565	organonitrogen compound catabolic process	665	27	10.47	1.48E-19
GO:0043691	reverse cholesterol transport	15	10	171.86	7.42E-19
GO:0034375	high density lipoprotein particle remodeling	15	10	171.86	7.42E-19
GO:0006027	glycosaminoglycan catabolic process	54	13	62.06	1.52E-18
GO:0038024	cargo receptor activity	36	11	78.77	1.06E-16
GO:1905952	regulation of lipid localization	81	13	41.37	2.85E-16
GO:0034371	chylomicron remodeling	9	8	229.14	7.30E-16
GO:0006024	glycosaminoglycan biosynthetic process	93	13	36.03	1.61E-15
GO:0030228	lipoprotein particle receptor activity	10	8	206.23	1.82E-15
GO:0032374	regulation of cholesterol transport	37	10	69.67	1.95E-14
GO:0008203	cholesterol metabolic process	92	12	33.62	6.22E-14

C) Module 93: Disease overlap



Borderline:
Toxic Enceph.
FT dementia

Alzheimer's Disease

Figure 7

Module 93 – lipid metabolism related Panel A: thick red border: gene/protein annotated to be ALS-associated on any of the 7 lists. Large red label: gene/protein with GWAS evidence. Triangular shaped nodes: receptors (UniProt keyword). Node colors indicate subcellular location (UniProt keywords): Blue = Nuclear, Green = Secreted, Purple = conflicting nuclear/secreted, cyan = neither nuclear nor secreted (likely cytosolic). Panel B: Gene Ontology overrepresentation analysis. Panel C: Thick border: ALS-associated (as panel A), Colored node: Disease-associated gene/protein.

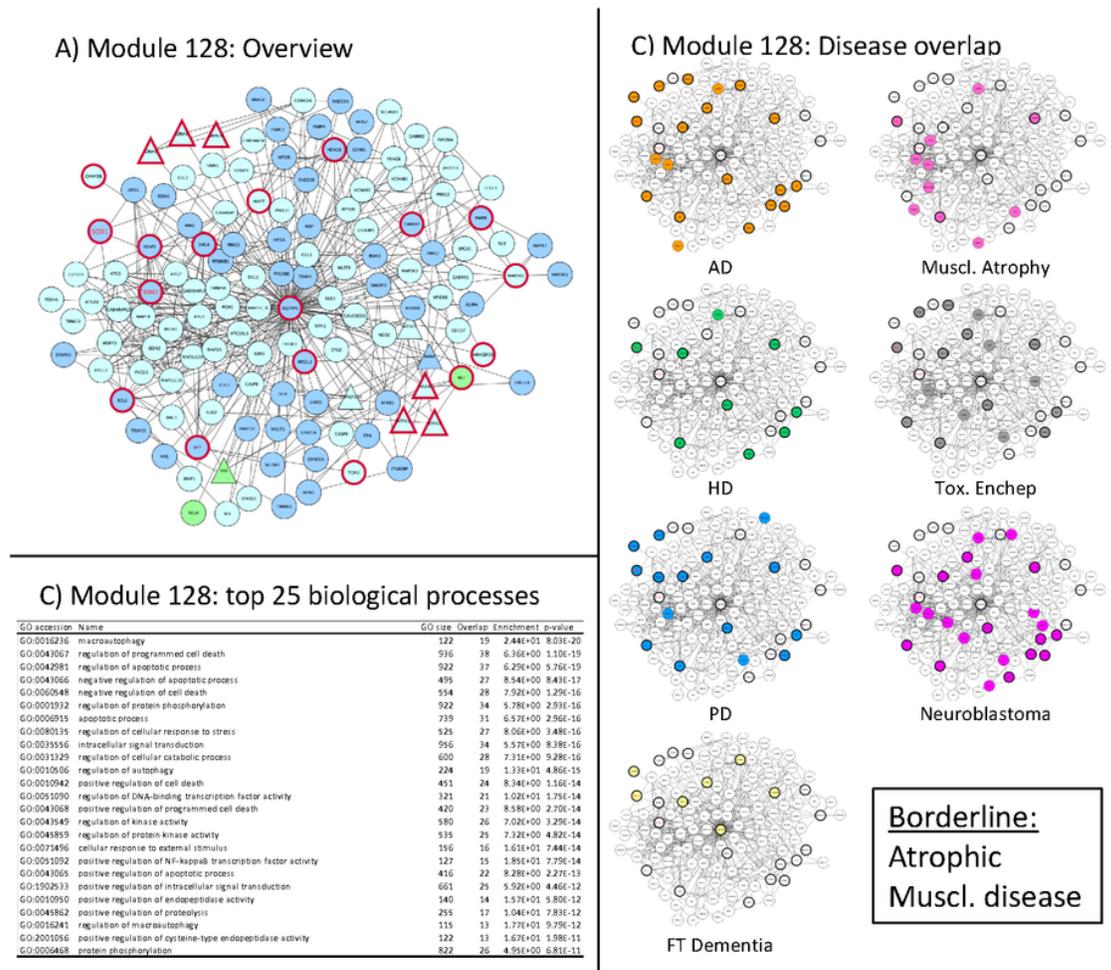
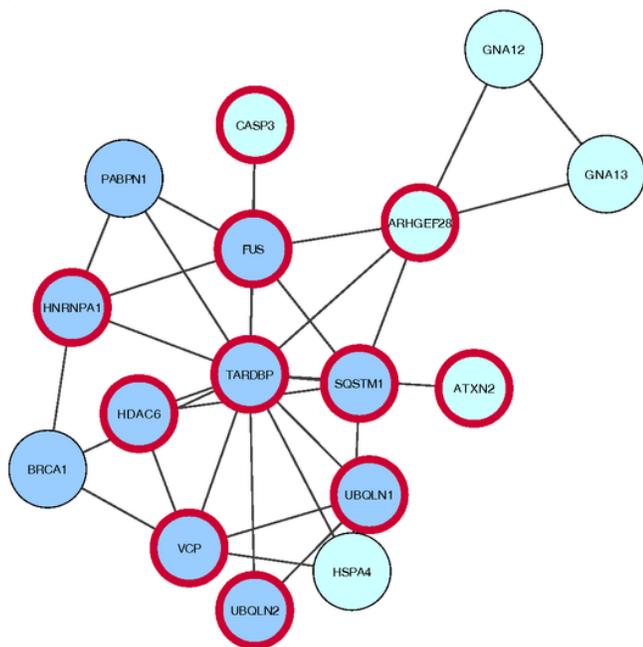


Figure 8

Module 128 – stress and apoptosis related Panel A: thick red border: gene/protein annotated to be ALS-associated on any of the 7 lists. Large red label: gene/protein with GWAS evidence. Triangular shaped nodes: receptors (UniProt keyword). Node colors indicate subcellular location (UniProt keywords): Blue = Nuclear, Green = Secreted, Purple = conflicting nuclear/secreted, cyan = neither nuclear nor secreted (likely cytosolic). Panel B: Gene Ontology overrepresentation analysis. Panel C: Thick border: ALS-associated (as panel A), Colored node: Disease-associated gene/protein.

A) Module 21: Overview



B) Module 21: Top 25 biological processes

GO accession	Name	GO size	Overlap	Enrichment	p-value
GO:0031329	regulation of cellular catabolic process	600	8	16.97	1.89E-07
GO:0043067	regulation of programmed cell death	936	8	10.88	3.75E-06
GO:0006914	autophagy	237	5	26.85	3.17E-05
GO:0032434	regulation of proteasomal ubiquitin-dependent protein catabolic process	83	4	61.34	3.85E-05
GO:0031331	positive regulation of cellular catabolic process	262	5	24.29	4.69E-05
GO:0042981	regulation of apoptotic process	922	7	9.66	5.03E-05
GO:0044265	cellular macromolecule catabolic process	599	6	12.75	7.86E-05
GO:0051603	proteolysis involved in cellular protein catabolic process	341	5	18.66	1.30E-04
GO:0010243	response to organonitrogen compound	398	5	15.99	2.35E-04
GO:0051129	negative regulation of proteasomal ubiquitin-dependent protein catabolic process	403	5	15.79	2.66E-04
GO:1903320	regulation of protein modification by small protein conjugation or removal	173	4	29.43	3.36E-04
GO:0046907	intracellular transport	933	6	8.19	6.17E-04
GO:0080135	regulation of cellular response to stress	525	5	12.12	6.68E-04
GO:0030433	ubiquitin-dependent ERAD pathway	48	3	79.55	7.01E-04
GO:1903311	regulation of mRNA metabolic process	234	4	21.76	8.09E-04
GO:0032434	positive regulation of proteasomal ubiquitin-dependent protein catabolic process	52	3	73.43	8.21E-04
GO:0006511	ubiquitin-dependent protein catabolic process	296	4	17.20	1.59E-03
GO:0022411	cellular component disassembly	300	4	16.97	1.65E-03
GO:0008380	RNA splicing	364	4	13.99	2.85E-03
GO:1903322	positive regulation of protein modification by small protein conjugation or removal	108	3	35.36	3.44E-03
GO:0016241	regulation of macroautophagy	115	3	33.20	3.88E-03
GO:0006397	mRNA processing	426	4	11.95	4.43E-03
GO:0051054	positive regulation of DNA metabolic process	146	3	26.15	6.16E-03
GO:0035973	aggrephagy	6	2	424.27	6.30E-03
GO:0031396	regulation of protein ubiquitination	150	3	25.46	6.48E-03

C) Module 21: Disease overlap

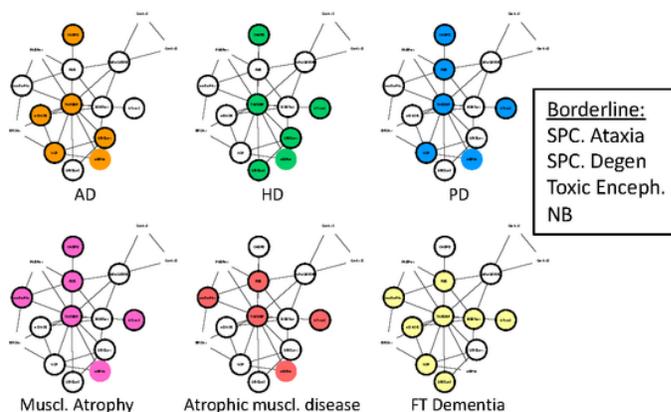


Figure 9

Module 21 – weakly classified biology, strong disease overlap Panel A: thick red border: gene/protein annotated to be ALS-associated on any of the 7 lists. Large red label: gene/protein with GWAS evidence. Triangular shaped nodes: receptors (UniProt keyword). Node colors indicate subcellular location (UniProt keywords): Blue = Nuclear, Green = Secreted, Purple = conflicting nuclear/secreted, cyan = neither nuclear nor secreted (likely cytosolic). Panel B: Gene Ontology overrepresentation analysis. Panel C: Thick border: ALS-associated (as panel A), Colored node: Disease-associated gene/protein.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPLEMENTARYMATERIALS.docx](#)
- [SupplementaryTableS1ALSgenelistnwsupport.xlsx](#)
- [SupplementaryTableS2ALSproteinlistnew2020.xlsx](#)
- [SupplementaryTableS3GOCLASSSUPERCLASS.xlsx](#)
- [SupplementaryTableS4TmTop50table.xlsx](#)
- [SupplementaryTableS5MostConnectedProteins.xlsx](#)