Identification of potentially functional modules and diagnostic genes related to amyotrophic lateral sclerosis based on the WGCNA and LASSO algorithms

Narges Daneshafrooz\textsuperscript{1,2}, Masumeh Bagherzadeh Cham\textsuperscript{3}, Mohammad Majidi\textsuperscript{4} & Bahman Panahi\textsuperscript{5}*

Amyotrophic lateral sclerosis (ALS) is a genetically and phenotypically heterogeneous disease results in the loss of motor neurons. Mounting information points to involvement of other systems including cognitive impairment. However, neither the valid biomarker for diagnosis nor effective therapeutic intervention is available for ALS. The present study is aimed at identifying potentially genetic biomarker that improves the diagnosis and treatment of ALS patients based on the data of the Gene Expression Omnibus. We retrieved datasets and conducted a weighted gene co-expression network analysis (WGCNA) to identify ALS-related co-expression genes. Functional enrichment analysis was performed to determine the features and pathways of the main modules. We then constructed an ALS-related model using the least absolute shrinkage and selection operator (LASSO) regression analysis and verified the model by the receiver operating characteristic (ROC) curve. Besides we screened the non-preserved gene modules in FTD and ALS-mimic disorders to distinct ALS-related genes from disorders with overlapping genes and features. Altogether, 4198 common genes between datasets with the most variation were analyzed and 16 distinct modules were identified through WGCNA. Blue module had the most correlation with ALS and functionally enriched in pathways of neurodegeneration-multiple diseases', 'amyotrophic lateral sclerosis', and 'endocytosis' KEGG terms. Further, some of other modules related to ALS were enriched in 'autophagy' and 'amyotrophic lateral sclerosis'. The 30 top of hub genes were recruited to a LASSO regression model and 5 genes (BCLAF1, GNA13, ARL6IP5, ARGLU1, and YPEL5) were identified as potentially diagnostic ALS biomarkers with validating of the ROC curve and AUC value.

Amyotrophic lateral sclerosis (ALS) is an adult-onset neurodegenerative disorder. This motor neuron disease is characterized by rapidly progressive paralysis, atrophy, and respiratory failure leading to death typically within 1 to 5 years after symptoms onset. The incidence of ALS is approximately 2 to 3 per 100,000 people each year\textsuperscript{1}. When symptoms are beginning, it is difficult to distinguish ALS from other disease especially for those with motor neuron signs. Consecutive years the El Escorial criteria, which essentially rely on motor signs, is using for ALS diagnosis and progression. However, it has estimated a false-positive of 8–10% and a false-negative of approximately 45% for applying this measure to evaluate ALS patients\textsuperscript{2}.

\textsuperscript{1}Department of Neuroscience, Faculty of Advanced Technologies in Medicine, Iran University of Medical Science, Tehran, Iran. \textsuperscript{2}School of Rehabilitation Science, Iran University of Medical Science, Tehran, Iran. \textsuperscript{3}Neuromusculoskeletal Research Center, Physical Medicine and Rehabilitation Department, Firoozgar Hospital, School of Medicine, Iran University of Medical Science, Tehran, Iran. \textsuperscript{4}Department of Biotechnology, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran. \textsuperscript{5}Department of Genomics, Branch for Northwest & West Region, Agricultural Biotechnology Research Institute of Iran (ABRII), Agricultural Research, Education and Extension Organization (AREEO), Tabriz, Iran. *email: b.panahi@abrii.ac.ir
ALS appears in 5–10% of cases with family history (fALS), while almost 90% of ALS patients are affected without family history (sALS). Genetic studies have revealed more than 30 genes contributing in ALS among which, C9orf72, SOD1, TARDBP, and FUS are more frequently involved. Increasing evidence suggests that several ALS-related genes represent a considerable phenotypic pleiotropy. Previous studies indicated mutations in some identical genes including C9orf72, TARDBP, FUS, CHCHD10, UBQLN2, SQSTM1, VCP and TBK1 have associated with the both ALS and FTD. Furthermore, C9orf72 variants have been detected in Parkinson’s disease or Alzheimer’s dementia patients. TBK1 and KIF5A linked to several other neurodegenerative diseases beyond ALS, especially to cerebellar ataxia syndrome and hereditary spastic paraplegia, respectively.

Despite the known genes related to ALS, the precise cause and underlying mechanism have not completely elucidated. Vast majorities of efforts are made in multiple areas to determine an acceptable biomarker as if meanwhile explain the pathogenic mechanism of ALS disease. Transcriptomics, a method to reveal differences in gene activity through mRNA expression, has been used in assessing the differential expression of mRNA in several disorders including neurodegenerative diseases. During recent years, transcriptome analysis provides means to explore and determine underlying molecular and cellular processes using high throughput technologies especially gene expression microarray and RNA-seq transcriptome profiling.

Nowadays, there is no standard approach for transcriptomic analyses, however network analysis is a novel strategy to analyze intricate datasets and to research the complex and heterogeneous diseases such ALS. Weighted gene co-expression network correlation-based analysis (WGCNA) is a precise approach of network analysis that assigns interconnected genes to the groups called ‘modules’, determines the most representative genes of each module as ‘eigengene’, and links the modules of co-expressed genes to a clinical phenotype. In addition, WGCNA provides the investigation of the degree of module preservation in other datasets. In this study, we used WGCNA to Meta-analyze the previously published circulating transcriptome datasets of ALS and control subjects.

To date, circulating mRNA profiling of ALS patients through microarray and RNA-seq have applied to identify disease biomarkers, although the limited sample size of these studies decreases the reliability of their results. More recently, a large cohort dataset containing 1117 participants (GSE112681) has been generated by et al. and re-analyzed by others. They used different strategies to identify different expressed genes that produced poor overlap results, while the identified DEGs did not present significant changes in expression (FDR < 0.1, fold change > 1.1 and < 9.9 for up regulated and down regulated DEGs, respectively).

In recent years, WGCNA has been used to identify circulating mRNA biomarkers in several conditions, including neurodegenerative diseases. In present study we conducted an integrated analysis of the blood transcriptome among ALS patients to identify circulating hub genes using weighted gene co-expression network analysis (WGCNA), a free R package, frequently used to explore the complex relationships between genes and phenotypes via simultaneously constructing gene networks and detecting gene modules and identifying the hub genes as central players within modules. The most important advantage of WGCNA is transforming gene expression data into co-expression modules, providing insights into signaling networks that may be responsible for phenotypic traits of interest. Here, co-expression network on gene expression datasets using WGCNA was employed to improve the level of significance in understanding biological mechanisms, including those related to diseases. Subsequently, we established a diagnostic-related gene signature set associated with ALS using logistic regression and least absolute shrinkage and selection operator (LASSO) analysis. The obtained model was verified using a receiver operating characteristic (ROC) curve.

One of the main challenges in determining biomarkers for ALS is finding a marker that can differentiate between ALS and some disorders with close symptoms. For example, FTD has been represented as one of two extremes of the ALS-FTD spectrum. The discovery of shared genetic disorders and the subsequent common pathological pathways links ALS and FTD in a number of ways. Ongoing evidence from epidemiological studies imply that 30–50% of patients with ALS diagnosis have developed cognitive/behavioral impairments during the course of disease. Further, there are 10 to 30% of patients with FTD who have demonstrated a varying range of a few motor neuron dysfunction symptoms up to the diagnosis of ALS. On the other hand, ALS patients have symptoms of motor dysfunctions at initial presentation that may confuse clinicians in distinguishing ALS from other motor diseases with similar symptoms.

Therefore, a biomarker with capability of differentiating ALS from diseases with similar early symptoms would certainly be more efficient. Here, we evaluated the different patterns of gene co-expression between ALS and ALS-mimic diseases using a dataset containing gene expression profiling of patients with ALS mimicking symptoms. Besides, we sought to find the genes capable of discriminating ALS genotypes from FTD by analyzing a dataset of gene expression GSE140830.

Methods

Dataset retrieval. The datasets GSE112681; including datasets GSE112676 and GSE112680, and GSE140830 were downloaded from the GEO database. These datasets generated from whole blood gene expression profiling via platforms Illumina HumanHT-12 V3.0 (the first) and HumanHT-12 V4.0 expression BeadChip arrays (two latter), respectively. Since the raw data idat files were not deposited at GEO, we got expression data including intensity and detection P-value as supplementary files. Expression information was extracted using the GEOquery R package (Version, 2.58.0). Sample information was extracted from the series matrix file. Subsequently, for each dataset, the respective ‘preQC-nonnormalized.txt’ files were processed via limma R package (version 2.9.8).

Correcting platform effects using algorithms aimed to eliminate batch effects directly modifies data, and inevitably, statistical tests after batch effect removal will be conducted on modified data. Therefore, we decided to select one of the platforms as a training set and the other as a validation set instead of platform effect removal.
Probe filtering and data processing. In both platforms, the probes that met the following criteria were excluded: (1) probes whose detection was statistically significant \((p < 0.05)\) in at least three samples. (2) Control probes and those with no symbol, (3) probes that are not referring to protein-coding genes; those without ‘NP’ or ‘NM’ prefix in RefSeq identifier in annotation files, (4) Probes with more than 10% NA value of samples. A microarray annotation file was used to assign probes with corresponding gene symbols. Probes with more than one gene symbol were eliminated and the average value was calculated out for those genes corresponding to more than one probes. Finally, the overlapped gene symbols between the datasets with CV > 20% remained as expression data matrices.

Weighted gene correlation network analysis. The network co-expression analysis was performed using the WGCNA R package (version 1.71). The flashClust package (version 1.01-2) was used to perform the cluster analysis of samples with the appropriate threshold value, followed by plotting the sample clusters against clinical traits. Soft-thresholding powers tested ranging from 1 to 20 and the appropriate power value was determined. The co-expression similarity matrix calculated just as Pearson test for all pairs of genes and then transformed into an adjacency matrix by soft-thresholding power \((\beta)\) as the smallest value to retain scale free topology. The adjacency matrix was then converted to a topological overlap matrix (TOM) for prospecting the modular structures of the co-expression network. The topological overlap dissimilarity, then was calculated using TOM, followed by constructing the gene clusters by hierarchical clustering in order to identify separated modules of interconnected genes. Subnetworks identified as co-expression modules are highly interconnected gene clusters that would be the true representatives of interactive genes contributing to biological pathways. Then, the module eigengene (ME), as the first principal component of each module was calculated. The ME is an expression representative of all genes in one module. Then, the module membership as a correlation between ME and gene expression pattern of a module was calculated. Finally, the gene significance (GS) of each gene in the module, which represented the correlation between gene and the clinical trait, was further calculated.

Network preservation analysis. To evaluate the module preservation, we applied a Z-summary permutation testing using the modulePreservation() function in the WGCNA package. In addition to the module preservation test between the training and validation datasets, we conducted the testing to find the modules not preserved in diseases mimicking ALS and FTD patients based on data presented in datasets GSE112680 and GSE112680.

As mentioned above, the heterogeneity of ALS can be considered in both genotypic and phenotypic overlapping with other diseases such as several movement disorders and frontotemporal dementia. Obviously, there are similarities in gene expression profiles due to the common underlying pathological pathways involved in neurological and neurodegenerative diseases. Finding the co-expressed genes not correlated to ALS, can distinguish it from disorders that have mechanisms in common with ALS.

Pathway enrichment analysis. To characterize the function of the significant gene modules, we performed KEGG pathway enrichment analysis\(^23\) using clusterProfiler and DOSE packages\(^24,25\). The threshold level of \(q\)-value < 0.05 were employed to indicate a significance for KEGG pathways.

Visualization of hub genes using protein–protein interaction network. The hub genes of main modules with MM > 0.8 and GC > 0.5 were imported into the STRING database (https://www.string-db.org/) and protein–protein interaction (PPI) networks were built. The results then were exported to Cytoscape Version 3.9.1 to visualize the PPI network. The CytOHubba app in Cytoscape was used to calculate the degree of each node and screen out the top hub genes.

Screening and validation of hub genes. The gene signature for the prediction of ALS was performed through least absolute shrinkage and selection operator (LASSO) logistic regression analysis. In order to validate the hub genes acquired from the WGCNA method, the LASSO was conducted using the cvglmnet() function (with tenfold cross-validation and alpha = 0.5) in R. The Lasso model shrinks the less important feature's coefficient to zero with the aim of retaining the good features. By L1 regularization, the LASSO regression adds a penalty equal to the absolute value of the magnitude of coefficients. It can minimize the coefficients and yield a spares model. The strength of the penalty term in Lasso regression is controlled by a tuning parameter \(\lambda\). When \(\lambda = 0\), no coefficients (features) are eliminated. By increasing the value of \(\lambda\), the more coefficients are set to zero and eliminated. In the present study, the hub genes of blue and black modules with gene significance > 0.5 and module membership > 0.8 were validated using this approach.

Subsequent to identifying genes that could potentially be diagnostic markers of ALS, receivers operating characteristic (ROC) curve analyses were conducted with the pROC R package. The area under the ROC curve (AUC) values were utilized to assess the predictive utility of identified key predictive genes. The AUC is a criterion for estimating the probability that a classifier (predictor), i.e. gene-set can predict better than randomly selected classifiers\(^26\).

Results

Demographic information of participants. The cohort included three large datasets, GSE112676, GSE112680, and GSE140830 with two platforms. GSE112676 contains 233 ALS diseases and 508 control subjects, while GSE112680 contains 164 ALS diseases, 137 control subjects, and 75 ALS-mimic diseases. GSE140840 consists of 172 FTD diseases (including 82 bvFTD, 46 nfvPPA, and 44 svPPA) and 281 control subjects. The
phenotypic data (demographic data and clinical assessments) extracted from series matrix files, were contained diagnostic and sex information of all samples, in addition of site of onset, age of onset and duration of disease, as well as genetic state of C9orf72 for ALS patients. The sample information has been summarized in Supplementary Table S1.

**Gene expression data processing.** Gene expression and annotation data were retrieved through the GEOquery R package. Gene expression data included raw expression intensities and detection P-values for all three datasets. Using data.table and limma packages, the raw expression and annotation files converted to matrices are appropriate for further analysis. The nque() function was then used to apply a normal-exponential convolution model for background correction followed by quantile normalization and log2 transformation to the intensities. The control probes, those with no symbol, and those that failed to detect expression were filtered out. Finally, only probes remained that referring protein-coding genes; those with ‘NP’ or ‘NM’ prefix in RefSeq identifier in annotation files. Following the last filtering step, GSE112676, GSE112680, and GSE140840 datasets included 35,000, 33,765, and 31,980 probes, respectively. Values for replicate probes were then replaced with their average using the averagep() function.

Identifying the outliers is a necessary step before sample clustering. So, the expression values upper and lower than 1.5 times the interquartile range (IQR) were equalized to ‘NA’ in final matrices of expression data. The probes of > 10% ‘NA’ were excluded as outliers. After averaging and outlier removal, the datasets GSE112676, GSE112680, and GSE140830 contained 22,691, 24,818, and 23,574 genes, respectively.

When network analyzing, the genes with the most variation in expression provides more information, so we selected a set of genes with coefficient of variation (CV) ≥ 0.2. Then, only the genes overlapped between three datasets remained for following network analysis. Consequently, either datasets contained 4198 common genes.

**Removing unwanted variance.** Batch effects, technical variation and other sources of unwanted variation are ever-present in big data, especially so in high-throughput gene expression profiling. Accordingly, we found an obviously inter-platform batch effect in three datasets using the principle component analysis (PCA) (Supplementary Fig. 1). Various algorithms that try to neutralize batch effects directly modify the data, therefore, preparing the data in a separate step of the analysis may lead to artifacts that could affect the reliability of the analysis. So, instead of correcting batch effects for meta-analysis, we decided to analyze GSE112676 as a training set and the GSE112680 as a validation set to compare the gene co-expression between ALS and control subjects. As well, comparing the gene co-expression between ALS and ALS mimic patients was possible by preservation analysis of these datasets. Further, by preservation analysis of dataset GSE140830, we managed to compare the gene co-expression between ALS and FTD patients.

We also evaluated the intra-platform batch effects in the datasets separately and found out an intra-platform batch effect in the GSE112676 dataset (Supplementary Fig. 2a). Furthermore, there was a direct correlation between the GSM indices and PC1 (cor = 0.69, P = 2.2e−16) (Supplementary Fig. 2b).

There are multiple ways to eliminate such variations including empiricalBayesLM() functions in the WGCNA package. So we applied empirical Bayes moderated linear model adjustment in function empiricalBayesLM() function. By applying the PCA on adjusted data, the batch effect, in addition to correlation between PC1 and GSM indices were removed (Supplementary Fig. 2d, c).

**Weighted-gene co-expression network analysis.** After removing unwanted variances for the known source (GSM indices), the WGCNA procedure was used to construct the co-expression network and to identify the co-expressed modules in dataset GSE112676 (233 ALS and 508 control). We selected the GSE112676 as a training dataset because of the larger sample size.

First, the sample clustering was obtained using hierarchial clustering. The sample outliers were detected and removed by pruning. The sample dendrogram was plotted to all traits (Supplementary Fig. 3, Fig. 1a). When the power value was set to 8, the independence degree was up to 0.8 (Supplementary Fig. 4). By raising the matrix of expression to soft-thresholding (β) of 8, TOMs and consequently dissTOMs were obtained. The cluster dendrograms were created based on the TOMs with a minimum cluster size of 30 and modules with a high correlated eigengenes have been merged with a threshold value of 0.2 (Supplementary Fig. 5).

In order to identify modules that were significantly related to the evaluated clinical features, the expression profiles of each module were summarized as the eigenvector correlated to the first main component of the expression matrix using the Module Eigengene (ME). Values of gene significance (GS) were used to calculate the most important (central) elements in the modules were closely related to the trait and can also be considered as hub-genes.

In total, 16 distinct co-expression modules were identified; each module was assigned a unique number and a unique color (Fig. 1b). The irrelevant genes were allocated to module0 or gray module. The correlation analysis associated, the most important (central) elements in the modules were closely related to the trait and can also

https://doi.org/10.1038/s41598-022-24306-2
Figure 1. (a) Sample dendrogram with traits heatmap. Red is ALS in diagnosis, female in Sex, Spinal in site.onset, Yes in C9 state, the higher value in age.onset and survival. Gray is equal to NA. (b) Clustering dendrogram of genes, dissimilarity is based on topological overlap, together with assigned module colors. As a result, 20 co-expression modules were constructed and was shown in different color. Merging the modules with high correlated eigengenes, resulted in 16 modules. The number of genes in each module were listed in Table S2. (c) Module-trait associations (ALS vs Control). Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and P-value. The table is color-coded by correlation according to the color legend.
were correlated to spinal site of onset ($R = 0.18$, $P_{val} = 4e−06$; respectively). There was no significant correlation between the modules and the other clinical features.

**Preservation test of modules.** In order to evaluate the preservation of ALS-related modules in the validation dataset (ALS and control subjects of dataset GSE112680), we assessed the module preservation using z-summary. According to the results of module preservation analysis, all modules except green were strongly preserved (preservation $Z$-score > 10) between two datasets (Fig. 2a). Subsequently, we tested whether the ALS-related modules preserved using the module preservation method in datasets GSE112680 (ALS-mimic and control subjects) and GSE140830, respectively. The results indicated that green module is non-preserved in mimic diseases, whereas lightgreen, midnightblue, and purple modules are poorly preserved in FTD patients (Fig. 2b). Since green modules were non-preserved in validation dataset too, it could not be considered as ALS-mimic related module. While, Light green, midnightblue, and purple modules were considered as FTD-related modules.

**Enrichment analysis of the main modules.** Enrichment analysis was performed for the genes in the significant co-expression modules. Genes in the blue module were the most enriched in KEGG pathways: ‘pathways of neurodegeneration-multiple diseases’, ‘amyotrophic lateral sclerosis’, and ‘endocytosis’. However, the pathways related to ALS such as ‘apoptosis’, ‘cellular senescence’, ‘lysosome’ are of significant enriched terms in the blue module. The enrichment of inter-connected genes in black module resulted in terms related with ALS such as: ‘autophagy-animal’, ‘necroptosis’, and ‘forkO signaling pathway’ (Fig. 3a). Among the other ALS-related modules, salmon and brown modules were enriched in ‘amyotrophic lateral sclerosis’, too (Supplementary Table S3).

The genes in modules non-preserved in FTD disease, including lightgreen, midnightblue, and purple modules, were enriched in ‘prion disease’, and pathways related to metabolic disorders such as ‘diabetic cardiomyopathy’, ‘insulin signaling pathway’, and ‘aldoesterone synthesis and secretion’ (Fig. 3b).

**Identification of main modules and hub genes.** As mentioned above, blue and black were defined as ALS-related modules, while purple; lightgreen and midnightblue were considered as FTD-related modules. When the genes in each module met the criteria of MM > 0.8, they were considered as the module’s hub genes. The lists of hub genes according to modules have provided in Supplementary Table S4.

Subsequently, the DEGs obtained from previously studies were overlapped with the hub genes of the ALS-related modules and presented using a Venn diagram (Supplementary Fig. 6). Thirty overlapping genes were identified as follows: CNBP, AQP9, TAF7, TBK1, CRLS1, CSDKN1B, ANAPC13, SH3GLB1, ANSD1, MNDA, YPEL5, CLDN1D, CHUK, HBPI, MLKLN1, C1orf52, QPCT, SLC35A1, RGS18, GPBP1, STX3, VEZF1, TMEM71, CRBN, HIST1H2AC, PPP2R3C, TMEM126B, SAMP9, VNN2, and STXBP3.

**Protein–protein interaction network.** We used STRING to establish a PPI network with 264 nodes and 571 edges from 265 hub genes in blue and black modules. The PPI data were imported into Cytoscape software (version 3.9.1) and CytoHubba app was employed to predict important nodes or subnetworks based on the 571 edges from 265 hub genes in blue and black modules. The PPI data were imported into Cytoscape software to LASSO regression analysis. Based on the tenfold cross-validation, the remaining genes with nonzero LASSO and gene significance as hub genes related to ALS diagnosis $GC > 0.5$). Then, the top 30 hub genes were subjected to meta-analysis of multiple independent studies that increase the sample size and statistical power$^{27}$. Another result inconsistent across independent studies. One solution to address the problem of heterogeneity involves less the findings are strongly affected by the disease heterogeneity and sample-specific variation, making the leading to motor neuron degeneration and the clinical entity of this debilitating disease. The gene expression-based studies have improved our understanding about the genetic perturbation associated with ALS, nevertheless the findings are strongly affected by the disease heterogeneity and sample-specific variation, making the result inconsistent across independent studies. One solution to address the problem of heterogeneity involves meta-analysis of multiple independent studies that increase the sample size and statistical power$^{27}$. Another approach involves the study of the interaction among groups of genes, i.e. modules, which are investigated at the systems-level to identify genes associated with common function in a given biological state$^{28}$. By network analysis, i.e. WGCNA, genes with highly correlated expression are grouped into modules yet correlate with distinct biological pathways. Here, network analysis identified the blue module that were strongly correlated with ALS trait ($r = 0.58$, $P_{val} = 9e−58$). The genes related to ALS including C9orf72, TARDBP, CHMP2B, TBK1, VCP, DCTN1, OPTN, FIG4, ATXN2, BCL11B and PRPH were in the blue module. The blue module was enriched in pathways related to ALS, including: ‘Amyotrophic lateral sclerosis,’ ‘Pathways of neurodegeneration-multiple...
diseases', 'Endocytosis', 'Lysosome', 'Apoptosis', 'FoxO signaling pathway' and 'Autophagy-animal'. 'Amyotrophic lateral sclerosis' was also enriched in brown and salmon modules. Pathways such as 'Autophagy' and 'Lysosome' were the common terms in multiple modules.

Figure 2. The median rank and Z-summary statistics of the module preservation in validation dataset and (a) and the Z-summary of module preservation in ALS mimic diseases and FTD (b). The dashed blue and red lines in the graphs indicate the thresholds Z = 2 and Z = 10, respectively. These horizontal lines indicate the Z-summary thresholds for strong evidence of conservation (> 10) and for low to moderate evidence of conservation (> 2). All the modules are strongly preserved.
We also demonstrated a new 5-genes signature for diagnosis of ALS that was assigned through logistic LASSO regression analysis. The LASSO regression is a current widely used model for reducing the number of features, estimating of regression parameters, and selecting features especially in huge datasets. Furthermore,

Figure 3. KEGG enrichment analysis of main modules. (a) The top 15 terms of KEGG enrichment in ALS-related modules; blue and black, separately. (b) Enrichment analysis of FTD-related modules including purple, light green, and midnight blue.
the estimation of area under ROC curve (AUC) demonstrated the 5-genes set as an excellent predictive gene set for ALS diagnosis. All the genes including BCLAF1, GNA13, ARL6IP5, ARGLU1, and YPEL5 are involved in apoptosis. This is in accordance with the gene ontology term of ‘programmed cell death’ that was significantly enriched for genes differentially expressed between ALS and control subjects in the study of Van Rheenen15.

Bcl-2 associated factor1 (BCLAF1) is a binding protein that was originally identified to interact with anti-apoptotic Bcl-2 family member E1B19K, and induce apoptosis and suppress transcription29. Initial research demonstrated that BCLAF1 also plays a role in multiple biological processes that are not related to actions of Bcl2 family members, including RNA processing, skeletal muscle differentiation, DNA damage response, and autophagy. Notably, there is evidence that BCLAF1 could serve as a strong autophagy inducer by displacing beclin-1 from BCL2 in myeloma cells30. Another study reported that BCLAF1-mediated autophagy could prevent...

Figure 4. Protein–protein interaction (PPI) network analyzed by cytoHubba app of Sytoscape. (a) Top 20 key genes of the hub genes in ALS-related modules. (b) Top 10 key genes of the hub genes in FTD-related modules. The key genes explored according to MMC algorithm of cytoHubba. Darker red of nodes presents higher MCC score.

Figure 5. (a) The Lasso.model and cv.Lasso model including the selection of the optimal parameter lambda (λ). The optimal λ-value was 0.0021. (b) Five genes were finally selected. The ROC and optimum accuracy of 5-genes set for diagnosis of ALS. The values of AUC imply the 5-genes set is an excellent predictor for ALS disease.
cell apoptosis and increase cell proliferation in HCC cells. The role of BCLAF1 in ALS is not yet evaluated. However, in dopamin neurons of patients with Parkinson, it is indicated that BCLAF1 is downregulated which is associated with programmed cell death and mitochondrial function. A recent study showed that cardiac ischemia–reperfusion (I/R) injury can elicit BCLAF1 overexpression followed by upregulation of apoptosis-related proteins suggesting BCLAF1 is a detrimental factor in cardiac I/R injury.

Guanine nucleotide-binding protein subunit alpha-13 (GNA13/Gα13) is a member of the G protein family. In general, Guanine nucleotide-binding proteins (G proteins) are involved in transmembrane signal transducing and alpha subunits play a critical role in signal transducing by binding to GTP and transforming it to GDP.

GNA13 have been particularly associated with tumor invasion and metastasis in multiple types of cancer through RhoA/ROCK and PI3/AKT signaling pathways. GNA13 is overexpressed in tumors and closely associated with an aggressive phenotype and poor patient’s prognosis. GNA13, also negatively regulates BCL-2 and inhibits PI3K-AKT signaling pathway, whereas its deficits impairs apoptosis in germinal center B-cells. In neuronal cells, GNA13 is involved in neurite outgrowth and synaptic function through Rho signaling pathway. It is indicated that GNA13 is downregulated in prefrontal cortex of schizophrenia brain and may have negative effect on synaptic plasticity by suppression of GNA13-ERK pathway. A recent study showed that GNA13 is upregulated in Alzheimer’s disease and increases as the disease worsens. The dysregulation of GNA13 in ALS is reported previously, however the affected mechanisms is unknown.

ADP ribosylation factor-like GTPase 6 interacting protein 5 (ARL6IP5), also known as JWA gene encodes a microtubule-associated protein that contains a prenylated Rab acceptor motif. It regulates intracellular protein transport and oxidative stress. ARL6IP5, also has been shown to induce endoplasmic reticulum (ER) stress-mediated apoptosis in osteoblast cells, involve in TRAIL-triggered cell apoptosis in gastric cancer, and play a protective role against DNA damage induced by oxidative stress. The findings suggest that ARL6IP5/JWA play a critical role in oxidative stress-induced DNA damage. The role of ARL6IP5 in the pathogenesis of ALS has never been investigated. However, there is an increasing data that indicated ARL6IP5 has a neuroprotective role in DA neuronal cells via modulating intracellular redox and NF-xB signaling pathway, regulating Nrf2, and modulating GLT-1 expression and glutamate uptake in PD models.

Arginine and glutamate rich 1 (ARGLU1) is a highly evolutionarily conserved with a dual function. The glutamate rich C-terminal is able to co-activate glucocorticoid receptor (GR) and the arginine rich N-terminus modulates RNA splicing RNAs. Glucocorticoids are expressed in multiple regions of the brain. During chronic stress, glucocorticoids bind to GR and activate them. There is evidence that chronic stress is associated with neurodegenerative diseases. Clinical studies suggest that exposure to chronic stress followed by elevated GC levels can be a causal factor for neurodegenerative induction and progression. Magommedova et al. in a study have identified ARGLU1 as a GR coactivator and RNA binding protein that is involved in the both layers of gene regulation including transcription and alternative splicing in neuronal cells. Although, it has not been revealed the role of ARGLU1 in ALS, it has indicated in some pathological conditions that ARGLU1 serves as an anti-apoptotic factor. For example, targeting of ARGLU1 by miRNAs cause induction of apoptosis and inhibition of proliferation in hypertrophic scar, uterine leiomyomata and breast cancer.

YPEL5 is a member of the YPEL gene family that is highly conserved in the eukaryotic species and apparently involved in a certain cell division-related function. YPEL5 protein has been shown to be involved in the cell cycle progression. Human YPEL5 plays pro-apoptotic role and is also involved in DNA damage-induced apoptosis. Nonetheless, the dysregulation of the YPEL5 in ALS patients or models were not reported. YPEL5 were ubiquitously expressed in human tissues. YPEL5 is localized in the nucleus and centrosome, suggesting the conservation of a pro-apoptotic role of YPEL5 in DNA damage-induced apoptosis.

ALS and FTD with common genetic and clinical traits giving rise to an ALS/FTD spectrum that indicate some common pathological mechanisms leading to either of these neurological disorders. However, the pattern of two-disease onset is quite different. To identification of genes with different co-expression patterns in two diseases leading to distinct the different types of neurons involved, we analyzed the non-preserved modules between two expression datasets. Three modules with genes were non-preserved and considered as FTD-related modules. Enrichment analysis of these modules revealed the genes are involved in some metabolic impairments and prion disease. It has previously been shown that metabolic dysfunction, including the insulin–signaling pathway, is associated with cognitive impairment, including dementia. For example, there is a link between insulin resistance and cognitive impairment. In this regard, the hub genes of FTD-related modules including HEPB2, CAMK1G, ADRA2G, C22orf132 are associated with cognitive impairment. Increasing evidence points to closely relevant between insulin signaling impairment and pathologic mechanisms of cognitive dysfunction.

Insulin function in neurons is very important and can affect many biological functions including blood glucose and metabolism regulation. The association of ALS and insulin resistance was evidenced in past researches. Several studies have evaluated the association between insulin resistance and ALS and reported conflicting results from a protective effect up to a risk factor for developing ALS. A recent review study has concluded that metabolism changes occurred in FTD and ALS can be the result of neurodegeneration trend with a secondary effect of destroying the vital brain regions, especially in FTD patients. Giving the ALS-FTD continuum, the ALS patients without cognitive deficits have a different pattern of energy metabolism rather than FTD, whereas ALS patients with cognitive deficits have an increased BMI and eating changes (hyperphagia) compared with patients with ALS without cognitive deficits. The hyperphagia go to be sever in FTD patients.

Meanwhile, preservation analysis of ALS-mimic patients’ gene expression did not result in any non-preserved module that could be considered as MIM-related module. The result was consistent with the original study by van Rheenen et al. who did not find differentially expressed genes between ALS and ALS-mimic patients. This may be due to the large variety of patients who had participated in the study as ALS-mimic. There were 75 ALS-mimic patients with 25 different forms of diagnosis. Significant results were likely to be obtained if more participants had been selected with less variation in ALS-mimic disorders.
Conclusion
In conclusion, the current study combined WGCNA algorithm and LASSO logistic regression to screen potentially diagnostic biomarkers associated with ALS. We identified ARG1L1, CHMP2B, GNA13, SMAD4, VPS26A, and YPEL5 genes related to ALS. Moreover, we found autophagy may be involved in pathogenesis of ALS. The genes can be detected in blood and should be further studied as possible biomarkers of disease and drug development.

Data availability
The datasets analyzed during the current study are available in the European Nucleotide Archive with accession number GSE112676, GSE112680, and GSE140830.

Received: 14 May 2022; Accepted: 14 November 2022
Published online: 22 November 2022

References
1. Laferriere, F. & Polymenidou, M. Advances and challenges in understanding the multifaceted pathogenesis of amyotrophic lateral sclerosis. Swiss Med. Wkly. 145, 1–13 (2015).
2. Ferraro, P. M. et al. Multimodal structural MRI in the diagnosis of motor neuron diseases. NeuroImage Clin. 16, 240–247 (2017).
3. Qi, Y. et al. Neuroprotective effect of sonic hedgehog mediated PI3K/AKT pathway in amyotrophic lateral sclerosis model mice. Mol. Neurobiol. 59, 6971–6982 (2022).
4. Brenner, D. & Weishaupt, J. H. Update on amyotrophic lateral sclerosis genetics. Curr. Opin. Neurol. 32, 735–739 (2019).
5. Abramzon, Y. A., Fratta, P., Traynor, B. J. & Chia, R. The overlapping genetics of amyotrophic lateral sclerosis and frontotemporal dementia. Front. Neurosci. 14, 42 (2020).
6. Yang, X. et al. High-throughput transcriptome profiling in drug and biomarker discovery. Front. Genet. 11, 1–19 (2020).
7. Panahi, B. & Hejazi, M. A. Weighted gene co-expression network analysis of salt-responsive transcriptomes reveals novel hub genes in green halophytic microalgae Dunaliella salina. Sci. Rep. 11, 1–11 (2021).
8. Farhadian, M., Rafat, S. A., Panahi, B. & Mayack, C. Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process. Sci. Rep. 11, 1–15 (2021).
9. Ghaframani, N., Shodja, J., Rafat, S. A., Panahi, B. & Hasanpur, K. Integrative systems biology analysis elucidates mastitis disease underlying functional modules in dairy cattle. Front. Genet. 12, 1–12 (2021).
10. Panahi, B. & Hejazi, M. A. Integrative analysis of gene expression and alternative splicing in microalgae grown under heterotrophic condition. PLoS ONE 15, e0225767 (2020).
11. Farhadian, M., Rafat, S. A., Panahi, B. & Ebrahimie, E. Transcriptome signature of two lactation stages in Ghezel sheep identifies using RNA-sequencing. Anim. Biotechnol. 33, 223–233 (2022).
12. Panahi, B. & Farhadian, M. & Hejazi, M. A. Systems biology approach identifies functional modules and regulatory hubs related to secondary metabolites accumulation after transition from autotrophic to heterotrophic growth condition in microalgae. PLoS ONE 15, e0225767 (2020).
13. Panahi, B., Mohammadi, S. A., Ruzicka, K., Abbasi Holaso, H. & Zare Mehrjerdi, M. Genome-wide identification and co-expression network analysis of nuclear factor-Y in barley revealed potential functions in salt stress. Physiol. Mol. Biol. Plants 25, 485–495 (2019).
14. Panahi, B., Frahadian, M., Dums, J. T. & Hejazi, M. A. Integration of cross species RNA-Seq meta-analysis and machine-learning models identifies the most important salt–resistance pathways in microalgae Dunaliella. Front. Genet. 10, 752 (2019).
15. Van Rheenen, W. et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. PLoS ONE 13, e0198874 (2018).
16. Swindell, W. R., Kruse, C. P., List, E. O., Berryman, D. E. & Kopchick, J. J. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. J. Transl. Med. 17, 1–33 (2019).
17. Li, J. et al. Application of weighted gene co-expression network analysis for data from paired design. Sci. Rep. 8, 1–8 (2018).
18. Udyavar, A. R. et al. Co-expression network analysis identifies Spleen Tyrosine Kinase (SYK) as a candidate oncogenic driver in a subset of small-cell lung cancer. BMC Syst. Biol. 7, 1–16 (2013).
19. Baradaran, H., Heravi, Y., Van Beeckhoven, C. & van der Zee, I. Transgranule mediated protein aggregation and underlying gene defects in the FTD-ALS spectrum. Neurobiol. Dis. 134, 104639 (2020).
20. Bersano, E. et al. Decline of cognitive and behavioral functions in amyotrophic lateral sclerosis: A longitudinal study. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 21, 373–379 (2020).
21. Ahmed, R. M. et al. Amyotrophic lateral sclerosis and frontotemporal dementia: distinct and overlapping changes in eating behaviour and metabolism. Lancet Neurol. 15, 332–342 (2016).
22. Nachun, D. et al. Systems-level analysis of peripheral blood gene expression in patients reveals an innate immune response shared across multiple disorders. bioRxiv 13, 457 (2019).
23. Ogata, H. et al. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34 (1999).
24. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics J. Integrat. Biol. 16, 284–287 (2012).
25. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 31, 608–609 (2015).
26. Airola, A. et al. The spatial leave-pair-out cross-validation method for reliable AUC estimation of spatial classifiers. Data Min. Knowl. Disc. 33, 730–747 (2019).
27. Daneshfrooz, N. et al. Identification of let-7f and miR-338 as plasma-based biomarkers for sporadic amyotrophic lateral sclerosis using meta-analysis and empirical validation. Sci. Rep. 12, 1–12 (2022).
28. Guo, B., Kaufsh, A. & Nadeau, K. C. Novel meta-analysis pipeline of heterogeneous high-throughput gene expression datasets reveals dysregulated interactions and pathways in asthma. medRxiv 14, 29 (2019).
29. Kassof, G. M., Goyal, L. & White, E. Bf, a novel death-promoting transcriptional repressor that interacts with Bcl-2-related proteins. Mol. Cell Biol. 19, 4390–4404 (1999).
30. Lamy, L. et al. Control of autophagic cell death by caspase-10 in multiple myeloma. Cancer Cell 23, 435–449 (2013).
31. Yuan, S. et al. B-cell lymphoma-2 associated transcription factor 1 is overexpressed and contributes to sorafenib resistance in hepatocellular carcinoma. Hepatol. Res. 49, 1329–1340 (2019).
32. Simunovic, F. et al. Gene expression profiling of substantia nigra dopamine neurons: further insights into Parkinson's disease pathology. Brain 132, 1795–1809 (2009).
33. Zhang, Y. et al. The long noncoding RNA IncCIRRIIL disrupts the nuclear translocation of Bclaf1 alleviating cardiac ischemia–reperfusion injury. Nat. Commun. 12, 1–15 (2021).
34. Shimono, J. et al. Analysis of GNA13 protein in follicular lymphoma and its association with poor prognosis. *Am. J. Surg. Pathol.* **42**, 1466 (2018).
35. Teo, C. R., Casey, P. J. & Rashheed, S. A. K. The GNA13-RhoA signaling axis suppresses expression of tumor protective Kallikreins. *Cell. Signal.* **28**, 1479–1488 (2016).
36. Mu, G. et al. Gastrin stimulates pancreatic cancer cell directional migration by activating the Gα12/13-RhoA–ROCK signaling pathway. *Exp. Mol. Med.* **50**, 1–14 (2018).
37. Green, J. A. & Cyster, J. G. S1PR2 links germinal center confinement and growth regulation. *Immunol. Rev.* **247**, 36–51 (2012).
38. Zhang, J. X. et al. GNA13 as a prognostic factor and mediator of gastric cancer progression. *Oncotarget* **7**, 4414 (2016).
39. Healy, J. A. et al. GNA13 loss in germinal center B cells leads to impaired apoptosis and promotes lymphoma in vivo. *Blood* **127**, 2723–2731 (2016).
40. Xia, Z. et al. GNA13 regulates BCL2 expression and the sensitivity of GCB-DLBCL cells to BCL2 inhibitors in a palmitoylation-dependent manner. *Cell Death Dis.* **12**, 1–11 (2021).
41. Hirayama-Kurogi, M. et al. Downregulation of GNA13-ERK network in prefrontal cortex of schizophrenia brain identified by combined focused and targeted quantitative proteomics. *J. Proteomics* **158**, 31–42 (2017).
42. Zhang, T., Shen, Y., Guo, Y. & Yao, J. Identification of key transcriptome biomarkers based on a vital gene module associated with pathological changes in Alzheimer’s disease. *Aging* **13**, 14940 (2021).
43. Saris, C. G. et al. Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* **10**, 1–16 (2009).
44. Oliveira, G. P., Alves, C. J. & Chadi, G. Early gene expression changes in spinal cord from SOD1G93A Amyotrophic Lateral Sclerosis animal model. *Front. Cell. Neurosci.* **7**, 216 (2013).
45. Miao, S. H. et al. Astrocytic JWA expression is essential to dopaminergic neuron survival in the pathogenesis of Parkinson’s disease. *CNS Neurosci. Ther.* **20**, 754–762 (2014).
46. Wu, Y. et al. Deficiency of osteoblastic Arl6ip5 impaired osteoblast differentiation and enhanced osteoclastogenesis via disturbance of ER calcium homeostasis and induction of ER stress-mediated apoptosis. *Cell Death Dis.* **5**, e1464–e1464 (2014).
47. Wang, Q. et al. JWA regulates TRAIL-induced apoptosis via MARCH8-mediated DR4 ubiquitination in cisplatin-resistant gastric cancer cells. *Oncogenesis* **6**, e355–e355 (2017).
48. Chen, R. et al. Identification of JWA as a novel functional gene responsive to environmental oxidative stress induced by benzo[a]pyrene and hydrogen peroxide. *Free Radical Biol. Med.* **42**, 1704–1714 (2007).
49. Zhao, X. et al. JWA antagonizes paraquat-induced neurotoxicity via activation of Nrf2. *Toxicol. Lett.* **277**, 32–40 (2017).
50. Wang, R. et al. Astrocytic JWA deletion exacerbates dopaminergic neurodegeneration by decreasing glutamate transporters in mice. *Cell Death Dis.* **9**, 1–15 (2018).
51. Magomedova, L. et al. ARGLU1 is a transcriptional coactivator and splicing regulator important for stress hormone signaling and development. *Nucleic Acids Res.* **47**, 2856–2870 (2019).
52. Vyas, S. et al. Chronic stress and glucocorticoids: from neuronal plasticity to neurodegeneration. *Neural Plasticity* **36**, 30–42 (2016).
53. Li, L., Han, W., Chen, Y. & Chen, Y. MiR-3613-3p inhibits hypertrophic scar formation by down-regulating arginine and glutamate-rich 1. *Mol. Cell. Biochem.* **476**, 1025–1036 (2021).
54. Zhao, W., Zhao, Y., Chen, L., Sun, Y. & Fan, S. miR-335-5p inhibits progression of uterine leiomyoma by targeting ARGLU1. *Comput. Math. Methods Med.* **36**, 1–15 (2022).
55. Zhang, D., Jiang, P., Xu, Q. & Zhang, X. Arginine and glutamate-rich 1 (ARGLU1) interacts with mediator subunit 1 (MED1) and is required for estrogen receptor-mediated gene transcription and breast cancer cell growth. *J. Biol. Chem.* **286**, 17746–17754 (2011).
56. Hosono, K. et al. YPEL5 protein of the YPEL gene family is involved in the cell cycle progression by interacting with two distinct proteins RanBPM and RanBP10. *Genomics* **96**, 102–111 (2010).
57. Lee, J. Y. et al. Pro-apoptotic role of the human YPEL5 gene identified by functional complementation of a yeast moh1Δ mutation. *I. Microbiol. Biotechnol.* **27**, 633–643 (2017).
58. Ma, L., Wang, J. & Li, Y. Insulin resistance and cognitive dysfunction. *Clin. Chim. Acta* **444**, 18–23 (2015).
59. Yagensky, O. et al. Increased expression of heme-binding protein 1 early in Alzheimer’s disease is linked to neurotoxicity. *Elife* **8**, e47498 (2019).
60. Gomez Ravetti, M., Rosso, O. A., Berretta, R. & Moscato, P. Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus’ gene expression profiles in Alzheimer’s disease. *PLoS ONE* **5**, e10153 (2010).
61. Brocos-Mosquera, I. et al. Differential brain ADRA2A and ADRA2C gene expression and epigenetic regulation in schizophrenia. Effect of antipsychotic drug treatment. *Transl. Psychiatry* **11**, 1–9 (2021).
62. Potkin, S. G. et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer’s disease. *PLoS ONE* **4**, e6501 (2009).
