RESEARCH ARTICLE

Shannon entropy approach reveals relevant genes in Alzheimer's disease

Alfonso Monaco[1], **Nicola Amoroso**[1,2]*, **Loredana Bellantuono**[2], **Eufemia Lella**[2], **Angela Lombardi**[1], **Anna Monda**[2], **Andrea Tateo**[2], **Roberto Bellotti**[1,2], **Sabina Tangaro**[1]

1 Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Bari, Bari, Italy, **2** Department of Physics 'Michelangelo Merlin', University of Bari 'Aldo Moro', Bari, Italy

* nicola.amoroso@ba.infn.it

Abstract

Alzheimer's disease (AD) is the most common type of dementia and affects millions of people worldwide. Since complex diseases are often the result of combinations of gene interactions, microarray data and gene co-expression analysis can provide tools for addressing complexity. Our study aimed to find groups of interacting genes that are relevant in the development of AD. In this perspective, we implemented a method proposed in a previous work to detect gene communities linked to AD. Our strategy combined co-expression network analysis with the study of Shannon entropy of the betweenness. We analyzed the publicly available GSE1297 dataset, achieved from the GEO database in NCBI, containing hippocampal gene expression of 9 control and 22 AD human subjects. Co-expressed genes were clustered into different communities. Two communities of interest (composed by 72 and 39 genes) were found by calculating the correlation coefficient between communities and clinical features. The detected communities resulted stable, replicated on two independent datasets and mostly enriched in pathways closely associated with neuro-degenative diseases. A comparison between our findings and other module detection techniques showed that the detected communities were more related to AD phenotype. Lastly, the hub genes within the two communities of interest were identified by means of a centrality analysis and a bootstrap procedure. The communities of the hub genes presented even stronger correlation with clinical features. These findings and further explorations on the detected genes could shed light on the genetic aspects related with physiological aspects of Alzheimer's disease.

Citation: Monaco A, Amoroso N, Bellantuono L, Lella E, Lombardi A, Monda A, et al. (2019) Shannon entropy approach reveals relevant genes in Alzheimer's disease. PLoS ONE 14(12): e0226190. https://doi.org/10.1371/journal.pone.0226190

Editor: Habil Zare, University of Texas Health Science Center at San Antonio, UNITED STATES

Received: June 24, 2019

Accepted: November 19, 2019

Published: December 31, 2019

Data Availability Statement: The data underlying the results presented in the study are available from the GEO database in NCBI (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/). The data entry number of database is GSE1297.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Alzheimer's disease (AD) is the most common type of dementia in aging population (up to the 70% of dementia's cases) [1]. The World Alzheimer Report 2016 affirmed that 47 million people are affected by dementia and it is expected that over 131 million people will develop dementia by 2050 [2]. Pathological processes are involved in AD such as intraneuronal formation of NeuroFibrillary Tangles (NFTs) [3, 4], abnormal β-amyloid production [5–8], extracellular deposition of senile plaques, early loss of synapses [9], oxidative stress [10, 11], and inflammation [8, 12–14]. To date, the pathogenesis of AD remains largely unknown and there is no cure for this disease, but treatment can still help in reducing symptoms and providing a better quality of life [15]. Converging evidence suggests that complex diseases result from association of several interacting genes possibly merging in molecular process. This suggests the existence of genetic communities that may be relevant for AD. Hence, recognition of the genetic basis of the disease is absolutely required to understand the biology of AD and to discover novel pharmacological treatments. In particular, the study of specific gene communities could facilitate the identification of therapeutic targets or candidate biomarkers.

In this work, we investigated gene expression data from a "publicly accessible microarray database obtained from AD and control human hippocampus" to detect gene communities relevant for AD. In this context, we implemented a gene co-expression network analysis addressed to formalize and integrate information related to multiple genes. In this approach, we modeled data through a network whose edges model the correlation between gene expressions and genes are the nodes of the graph [16]. Several methods have been proposed to investigate gene co-expression networks [17–19]. For example Weighted Gene Co-expression Network Analysis (WGCNA) has become widely adopted to provide a network identification based on the correlation of gene expression of a microarray database [20]. In particular, WGCNA can be used for identifying clusters of co-expressed genes with highly correlated expression (communities). In this work we implemented a hard threshold analysis and a community detection method, proposed in a previous work [21] and used in an international competition [22], based on the study of information content of the network. In the original paper [21] we employed this procedure on gene expression data linked to schizophrenia, now we applied the method to AD data. We identified 127 gene communities; two of them resulted significantly correlated with Mini Mental State Examination (MMSE) and NFT value. We verified that the two communities are stable and mostly replicated on two independent datasets. We also compared the proposed method, in addition to WGCNA, to three traditional clustering methods and three network reconstruction techniques. We performed a gene set enrichment analysis and a study to identify the hub genes of the two communities of interest applying centrality metrics and a bootstrap procedure. The communities of the found hub genes appear highly correlated with AD phenotype, and they could represent targets for a future AD therapy.

1 Materials and methods

In this work we analyzed data achieved from the GEO database in NCBI (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/). The data entry number of database is GSE1297. This dataset was developed from the work of Blalock [14] and contains hippocampal gene expression of 31 human subjects: 9 control and 22 AD. Clinical data as the severity of the disease, NFT value, Braak stage, MMSE score, sex, age, and Post-Mortem Interval (PMI) values are also included in the GSE1297 dataset. Data analysis procedure is summarized in Fig 1 and it consists of five main steps:

1. firstly, the normalized data was downloaded and a preprocessing analysis was implemented to combine expression values of multiple probes for one gene and to select genes with larger standard deviations. Secondly, correlation measures were used to construct the co-expressed gene network;

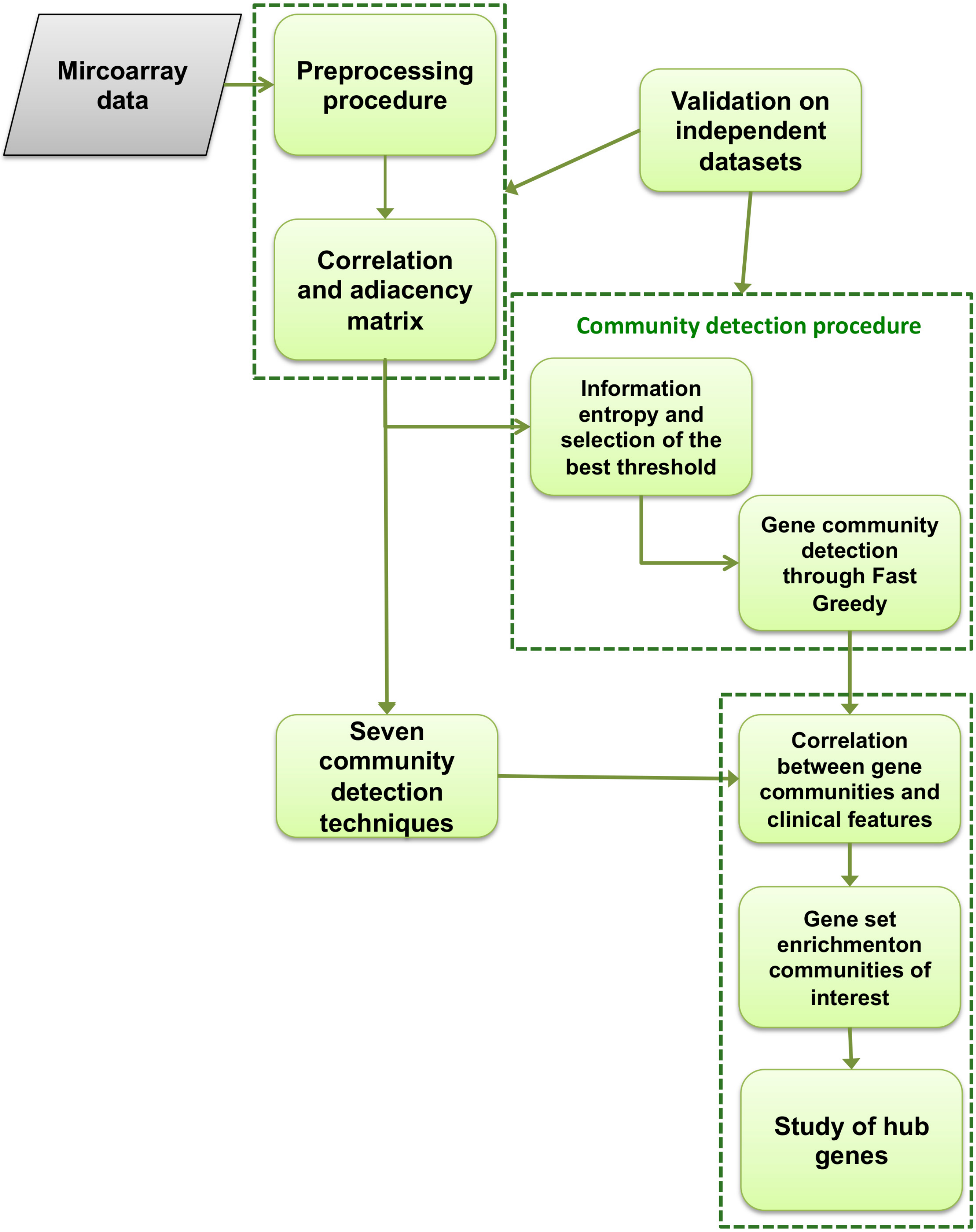
Mircoarray data
Preprocessing procedure
Correlation and adiacency matrix
Validation on independent datasets
Community detection procedure
Information entropy and selection of the best threshold
Gene community detection through Fast Greedy
Seven community detection techniques
Correlation between gene communities and clinical features
Gene set enrichmenton communities of interest
Study of hub genes

Fig 1. Flowchart of the methodology. After a preprocessing analysis, we implemented a community detection procedure based on hard threshold analysis and information theory proposed in a previous work [21]. We conduced a correlation analysis between gene communities and MMSE value

and NFT score. We repeated the analysis through different community detection techniques and we proposed a comparison with our method. Subsequently a gene set enrichment analysis and a hub gene study have been conduced. At last, we validated our procedure on two independent datasets.

https://doi.org/10.1371/journal.pone.0226190.g001

2. a hard threshold analysis was performed to compute the best threshold value, by using an information theory-based approach. Once the best threshold has been chosen, a community detection method was performed;
3. Person's correlation analysis between gene communities and clinical features (MMSE, NFT) was performed and communities of interest were identified; gene set enrichment analysis on communities of interest was applied and the hub genes of these modules were identified;
4. on the co-expressed gene network built in the step 1, community identification procedures by means of seven module detection algorithms were applied. Hence, correlation between gene communities and clinical features (MMSE, NFT) was computed. A comparison with results of our method was proposed;
5. steps 1 and 2 of the pipeline was repeated on two independent datasets to validate the proposed community detection method.

In the point 5 of the pipeline we analyzed hippocampal data of two other databases of GEO: GSE48350 and GSE29378. Both datasets contains gene expression of 38 and 16 control, 18 and 17 AD human subjects respectively.

1.1 Preprocessing data analysis

We downloaded the normalized data containing expression values of 22, 283 probes. The probes without corresponding annotation information were removed. In case of multiple probes for the same gene, probe with high Median Absolute Deviation (MAD) values was retained for further analysis. We chose MAD because it is a measure of dispersion, robust to outliers [23]. After the previous filtering, the standard deviation of the remaining gene expressions was calculated and values were sorted in decreasing order. Finally, we applied a selection criterion on the standard deviation to get a compromise between maximizing the information contained in gene expression data and minimizing the number of genes.

1.2 Correlation measures for a network of co-expressed genes

The network of co-expressed genes was built considering gene selected through the preprocessing procedure as nodes and analyzing their expressions for $N = 31$ subjects under investigation. In particular, given genes i and j, with expressions on the cohort $\{i_a\}_{a\,=\,1,...,N}$ and $\{j_b\}_{b\,=\,1,...,N}$ respectively, we computed the absolute value [24]:

$$d_{ij} = |r_{ij}| \tag{1}$$

where r_{ij} is the Pearson's pairwise correlation:

$$r_{ij} = \frac{\sum_{a=1}^{N}(i_a - \bar{i})\sum_{b=1}^{N}(j_b - \bar{j})}{\sqrt{\sum_{a=1}^{N}(i_a - \bar{i})^2}\sqrt{\sum_{b=1}^{N}(j_b - \bar{j})^2}} \tag{2}$$

with $\bar{i}$ and $\bar{j}$ mean values of the two expression distributions. The result is adjacency matrix C, in which each elements $c_{ij} = f(d_{ij})$ is a function of the correlation between expressions of genes i and j and characterizes the strength of the corresponding link in the network. Usually, two different thresholding methods to elaborate this matrix are considered: the soft and the hard thresholding. In the latter approach, the elements of C are defined as:

$$c_{ij} = signum(d_{ij}, th) = \begin{cases} 1 & \text{if } |d_{ij}| \geq th \\ 0 & \text{if } |d_{ij}| < th \end{cases} \tag{3}$$

where th is the selected threshold value. In this work, links were introduced in the network of co-expressed genes by means of a hard thresholding procedure [21], and the optimal threshold values was selected through information theory.

1.3 Information entropy based on betweenness to select the best threshold value

Betweenness b_i is a network centrality measure that evaluates the role of a node in connecting other pairs of nodes. For a complex network with M nodes the betweenness of the node i is defined as:

$$b_i = \sum_{j,k,j\neq k}^{M} \frac{n_{jk}(i)}{n_{jk}} \tag{4}$$

where n_{jk} indicates the number of geodesics between node j and k, and $n_{jk}(i)$ is the number of geodesics between the same genes, passing through node i. A geodesic between two nodes j and k is defined as the shortest path connecting them. Analyzing Eq 4 it is evident that betweenness has a crucial importance for graph characterization [25–27].

We implemented the method described in [21] and used in an international competition [22] to select the best threshold value with a hard threshold procedure based on information entropy [28] of the co-expression network betweenness. For a complex network with M nodes we defined the entropy based on betweenness as:

$$H_{bet} = -\sum_{i=1}^{M} b_i log_2[b_i] \tag{5}$$

where b_i is the betweenness of the i-th node defined by Eq 4. Since a system with maximum entropy value represents a system with maximum information content [29], we computed the entropy based on betweenness for the network of co-expressed genes, varying threshold values, and we chose the threshold value that maximized Eq 5.

1.4 Community detection through Fast Greedy algorithm

Once the best threshold was fixed, we applied the Fast Greedy [30] algorithm to split the whole network in communities. Fast Greedy is based on greedy optimization and it characterizes the community structure through the modularity. Modularity is based on the number of intra-community and inter-community links [31–35] and it allows the comparison of different partitions of the network. The modularity of a given partition is represented by the number of edges falling within groups minus the expected number in an equivalent random network. In particular we used the Fast Greedy algorithm described in [36]. Briefly, this method optimizes the modularity using three data structures:

- the matrix of modularity variation ΔQ_{ij} between communities i and j;

- a max-heap H composed of the largest element of each row of ΔQ_{ij} and of the labels, related to the communities i and j;
- an ordinary vector array which contains the sums of the elements of each row of the matrix e_{ij}. This matrix is the fraction of edges that joins vertices in community i to vertices in community j [36].

The use of max-heap H, which organizes the data in the form of binary trees, allows to update the matrix e_{ij} faster than the Newman's algorithm reported in [32]. In the present work, through an iterative procedure, we only selected communities that contained at least 3 and a maximum of 100 genes. Thus we referred to the concept of community, which was used in other international community detection studies as the DREAM challenge [22]. We implemented the hard threshold procedure described in the previous section, first on the whole network and then on all communities with more than 100 genes. As a matter of fact, communities comprising hundreds of genes are often too populated to gain meaningful biological insights [37].

1.5 Correlation analysis of gene modules with clinical phenotype

The association, between the communities found in the step 2 of the pipeline and clinical phenotype, was investigated by means of Pearson's correlation analysis. In particular, we computed the correlation of the module eigengene (first principal component, PC1, of the community) of each community with AD clinical features (MMSE, NFT). In other words, we synthetized the biological information of each community in one eigengene using principal component analysis [38–40]. Fig 2 shows a schematic view of the correlation procedure implemented.

To verify the robustness of significant correlations we implemented a procedure in which random subsets of 10 genes are removed from the communities of interest before to compute the first principal component. We repeated this procedure 100 times and for each sampling we calculated correlations with the clinical features through the method described in this section.

1.6 Gene set enrichment analysis

Only for the communities significantly correlated to clinical phenotype, we performed a gene set enrichment analysis using GSEA [41] web-tool. Through this tool we evaluated the overlap of the communities found in the step 3 of the pipeline, with the Molecular Signatures Database (MSigDB) [42]. Hence, we computed an estimate of the statistical significance to highlight common processes, pathways, and underlying biological themes. The overlap was measured by means of the hypergeometric distribution and p-value < 0.05 was considered to be significant enrichment. We applied a correction of hypergeometric p-value by means of multiple hypothesis testing according to Benjamini and Hochberg [43].

1.7 Hub genes identification

For the hub gene identification, the genes belonging to the communities found in the step 3 were analyzed using Kleinberg's centrality score (Kcs) [44]. The score of the vertices are defined as the principal eigenvector of CC^T where C is the adjacency matrix. In this work we considered hub of a community genes with a $Kcs > 0.8$ (the maximum value of Kcs is 1). To confirm the robustness of this selection criterion we implemented a bootstrap procedure [45, 46]. The data sample with 31 subjects was resampled 100 times and for each re-sampling we repeated the step 2 of the pipeline. For each resampling we computed the overlap between the found communities and we verified if hub genes of a given community were clusterized

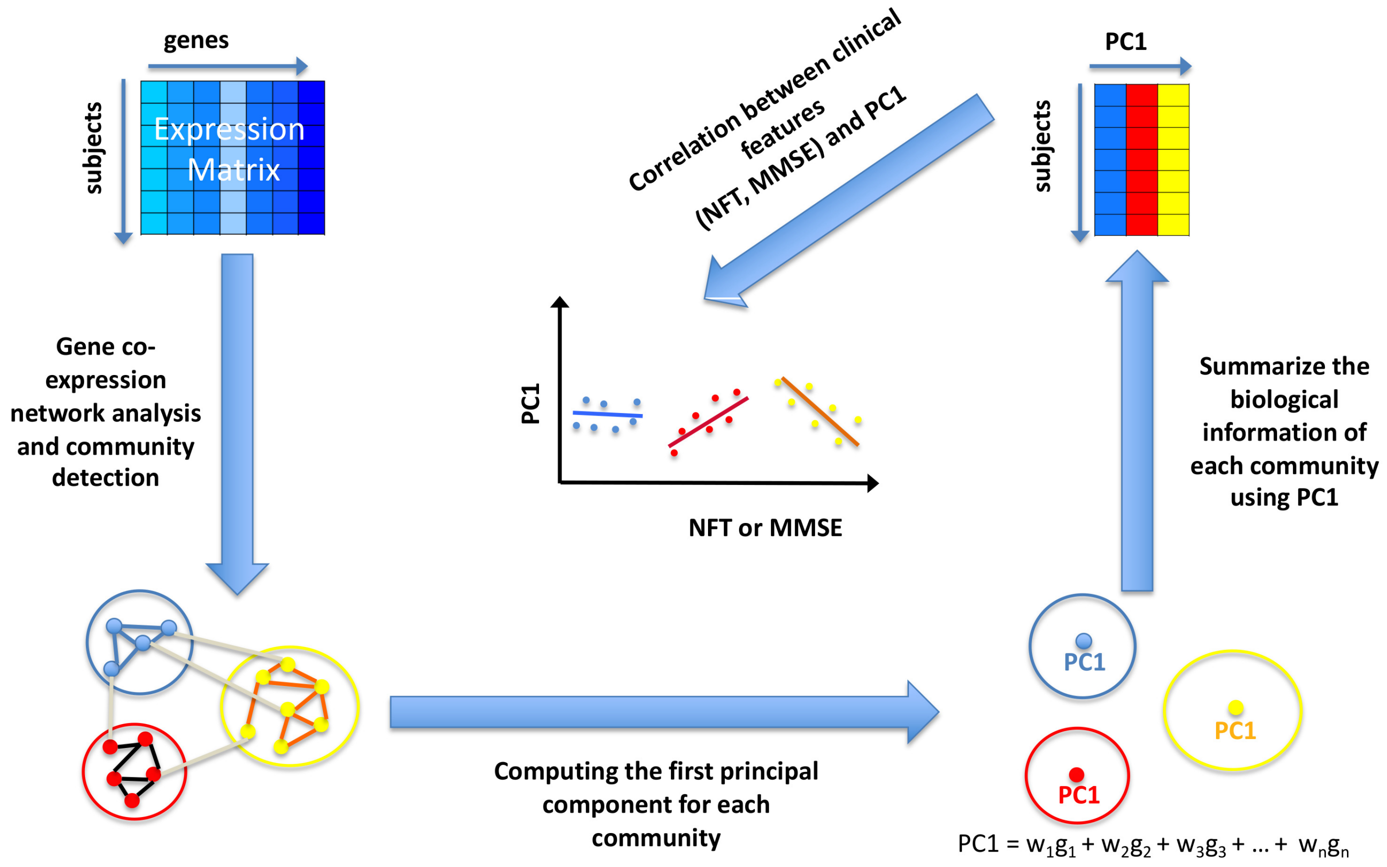
genes
subjects
Expression Matrix
Gene co-expression network analysis and community detection
Correlation between clinical features (NFT, MMSE) and PC1
PC1
NFT or MMSE
PC1
subjects
Summarize the biological information of each community using PC1
Computing the first principal component for each community
PC1
PC1
PC1
PC1 = w1g1 + w2g2 + w3g3 + … + wngn

Fig 2. Schematic view of the correlation procedure between module eigengene and clinical features. After the community detection procedure, we computed the module eigengene (first principal component, PC1, of the community) for each found community. Then we implemented a Pearson's correlation analysis between PC1 and clinical features (MMSE and NFT).

https://doi.org/10.1371/journal.pone.0226190.g002

together. In this way we also verified the existence of a pivotal and robust module of genes more connected to AD. To quantify the overlap between two different communities we estimated the overlap coefficient C_O [47]. It is defined as the maximum intersection between the target community and communities obtained by a different process (or a different dataset) divided by the smaller of the size of the two sets:

$$C_O(Q_t, Q_i) = \frac{max(Q_t \cap Q_i)}{min(|Q_t|, |Q_i|)} \tag{6}$$

where Q_t is the target community; i is an index between 1 and the number of communities obtained by a different process and Q_i is the related community.

1.8 Identification of gene communities through WGCNA algorithm

We applied WGCNA algorithm to the co-expressed gene network built in the step 2 of the pipeline and compared the outcome with results obtained by means of our procedure. At first, in order to assess the similarity of the gene expression profiles, the Pearson's correlation coefficient was calculated through Eq 2. Hence, the adjacency matrix C was obtained by applying a soft thresholding procedure [17]:

$$c_{ij} = power(d_{ij}, \beta) = (d_{ij})^{\beta} \tag{7}$$

where $\beta \geq 1$.

Zhang B and Horvath S. in [20] also proposed an another type of soft adjacency function, the sigmoid function:

$$c_{ij} = sigmoid(d_{ij}, \alpha, \mu) = \frac{1}{1 + e^{-\alpha(d_{ij}-\mu)}} \tag{8}$$

where α and μ are parameters to set. WGCNA uses hierarchical clustering to identify gene communities and their respective colors. In this work, different communities were detected using dynamic tree cut method, which is based on conversion of the adjacency matrix to a topology overlay matrix (TOM) and cluster analysis. Finally, we computed the correlation of the module eigengene of each module detected by WGCNA with AD clinical features by means of the procedure described in the section 1.5.

1.9 Identification of gene communities through traditional clustering methods

We compared results obtained though our procedure with findings of traditional clustering strategies as: agglomerative hierarchical clustering, Fuzzy c-means, and affinity propagation. The common property of clustering algorithms is that they distribute genes in groups based on similarity measures in gene expression [48]. We can classify the three proposed approaches according to the way used to determine the number of modules: explicit methods, such as agglomerative hierarchical clustering and fuzzy c-means, in which the number of clusters is imposed by the researcher; implicit methods, such as affinity propagation, where the number of modules is adapted on the dataset analyzed according to other information suggested by the researcher [48]. The agglomerative hierarchical clustering method merges iteratively clusters together if their similarity measure is sufficiently high [30]. This similarity is based on a specific metric that measures the distance between pairs of elements and a linkage criterion that computes how similar clusters are according to the chosen metric. In this work we implemented the Euclidean distance ($L_2 - norm$):

$$d_{E-AB} = \sum_{j=1}^{n} \sqrt{(a_j - b_j)^2} \tag{9}$$

and the Manhattan distance ($L_1 - norm$):

$$d_{M-AB} = \sum_{j=1}^{n} |a_j - b_j| \tag{10}$$

where a_j and b_j are elements of two data points respectively $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$.

We used the following linkage criteria:

- complete linkage: the maximum distance between clusters is calculated before merging;
- single linkage: the minimum distance between clusters is computed before merging;
- average linkage: the average distance between clusters is calculated before merging;
- centroid linkage: after finding centroids of clusters, the distance between them is calculated before merging.

To choose the metric and the optimal number of clusters we computed the silhouette coefficient. This coefficient quantifies how well an observation (in this work an observation is represented by a gene with a gene expression value for each considered subject) is clustered measuring the proximity of each point in a cluster to points in neighboring clusters. It's defined as [49]:

$$s(i) = \frac{z(i) - y(i)}{max(y(i), z(i))} \tag{11}$$

in which $y(i)$ is the average dissimilarity between *i-th* element and all other elements of the cluster to which belongs and $z(i)$ is the minimum average dissimilarity of *i-th* element to all observation of all other clusters D. The silhouette coefficient is a quantity between −1 and 1, where a value near 1 indicates that the element i is very well clustered whereas a value near -1 indicates that the point should be affected to another cluster.

The Fuzzy c-means clustering algorithm (FCM) operates by randomly assigning a degree of membership μ_{ij} to each element x_i with belongs to each cluster j that we want to derive [50]. Through an iterative process, the cluster centers c_j are dynamically moved towards the optimal localization that is going to minimize a objective function that represents the sum of the distances of each point from each cluster center, appropriately weighted with the correct degree of membership. This function is defined as [50]:

$$J = \sum_{i=1}^{D}\sum_{j=1}^{N}\mu_{ij}^{m} \| x_i - c_j \|^2 \tag{12}$$

where μ_{ij} is the degree to which an observation x_i belongs to a cluster j; c_j is the center of the cluster J; D is the number of observation; N is the desired number of clusters; m is the hyperparameter (fuzzifier) that controls how fuzzy the cluster will be. In general, the squared Euclidean distance metric is used with FCM in order to compute the distances between the cluster centers and each observation in the dataset. In this work we compared the performance of three different metrics: euclidean, Manhattan, Pearson correlation defined in Eqs (9), (10) and (2) respectively.

Affinity Propagation (AF) [51] is a cluster technique that does not require user to specify the number of clusters. This algorithm is based on the concept of message passing where all the data points send messages to all other points. These messages contains the willingness of the points being exemplars i.e. the points that best explain the other data points. Each cluster only has one exemplar. The messages according to their nature are inserted in two different matrices: the responsibility matrix R in which each element reflects how suitable a point k is to be an exemplar for a point i; the availability matrix A that quantify how appropriate it would be for a point i to choose point k as its exemplar. The values of matrix R are based on a similarity function. The standard similarity measure used in the papers of Frey and Dueck is the negative euclidian distance squared [51]. AF is an iterative algorithm in which iterations are performed until either the cluster boundaries remain unchanged over a number of iterations.

The exemplars are the points that at the last iteration satisfy the following condition:

$$R(i,i) + A(i,i) > 0 \tag{13}$$

1.10 Identification of gene communities through three other network reconstruction algorithms

We compared our findings with other three methods for constructing gene network: ARACNE, GENIE3, and SPACE. These methods can in some cases compete with clustering and decomposition algorithms. Here we give a brief summary of these techniques. The detailed methodology for each approach has been described in other papers [52–54]. Tools such as ARACNE and GENIE3 try to build regulatory networks from co-expression networks [55]. ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) eliminates indirect connections between genes, then partners of a gene having a stronger correlation with each other genes than with the gene itself [55]. Only those connections that should be regulatory are left. This tool is an information-theory-based method that uses mutual information instead of Pearson correlation with the advantage to identify the non-linear or irregular dependencies, which will be missed by Pearson correlation [56].

GENIE3 (GEne Network Inference with Ensemble of trees) is a network inference method based on variable selection with ensembles of regression trees [53]. This tool is able to divide the problem of prediction of a regulatory network between n genes into n different regression issues. In each of the regression problems, the expression pattern of the target gene is predicted from the expression patterns of all the other genes (input genes), by means of tree-based ensemble methods (for example Random Forests). A sign on a possible regulatory link is derived from the importance of an input gene in predicting a target gene. These possible links are then jointed over all genes to provide a ranking of interactions to build whole network [53].

SPACE (Sparse PArtial Correlation Estimation) [54] is a partial-correlation-based method. This technique points to estimate non-zero entries in the inverse of the covariance matrix, also known as the concentration matrix. In the SPACE algorithm the concentration matrix estimation problem is converted in a regression problem and the results are optimized with a symmetric constraint [56].

To compare the different proposed methods with our procedure, we computed the correlation of the module eigengene of each found communities with AD clinical features, as described in the section 1.5.

2 Results and discussion

2.1 Preprocessing procedure

Firstly, we calculated the standard deviation σ for each gene expression, then we determined the maximum value σ_{max}. In this analysis we considered only genes with standard deviation greater than 0.1 σ_{max}. We chose this selection criterion after a study to evaluate the optimal cut. We repeated steps 1-3 of the pipeline for different values. 0.1 σ_{max} is resulted the optimal cut that maximizes the average correlation between the communities of interest and clinical features MMSE and NFT, as shown in Fig 3. Furthermore selecting genes exceeding 0.1 σ_{max} would have resulted in no AD-related communities (step 3 of the pipeline).

Fig 4 illustrates the distribution of standard deviation of gene expression. Then we kept 4, 154 genes for further analysis.

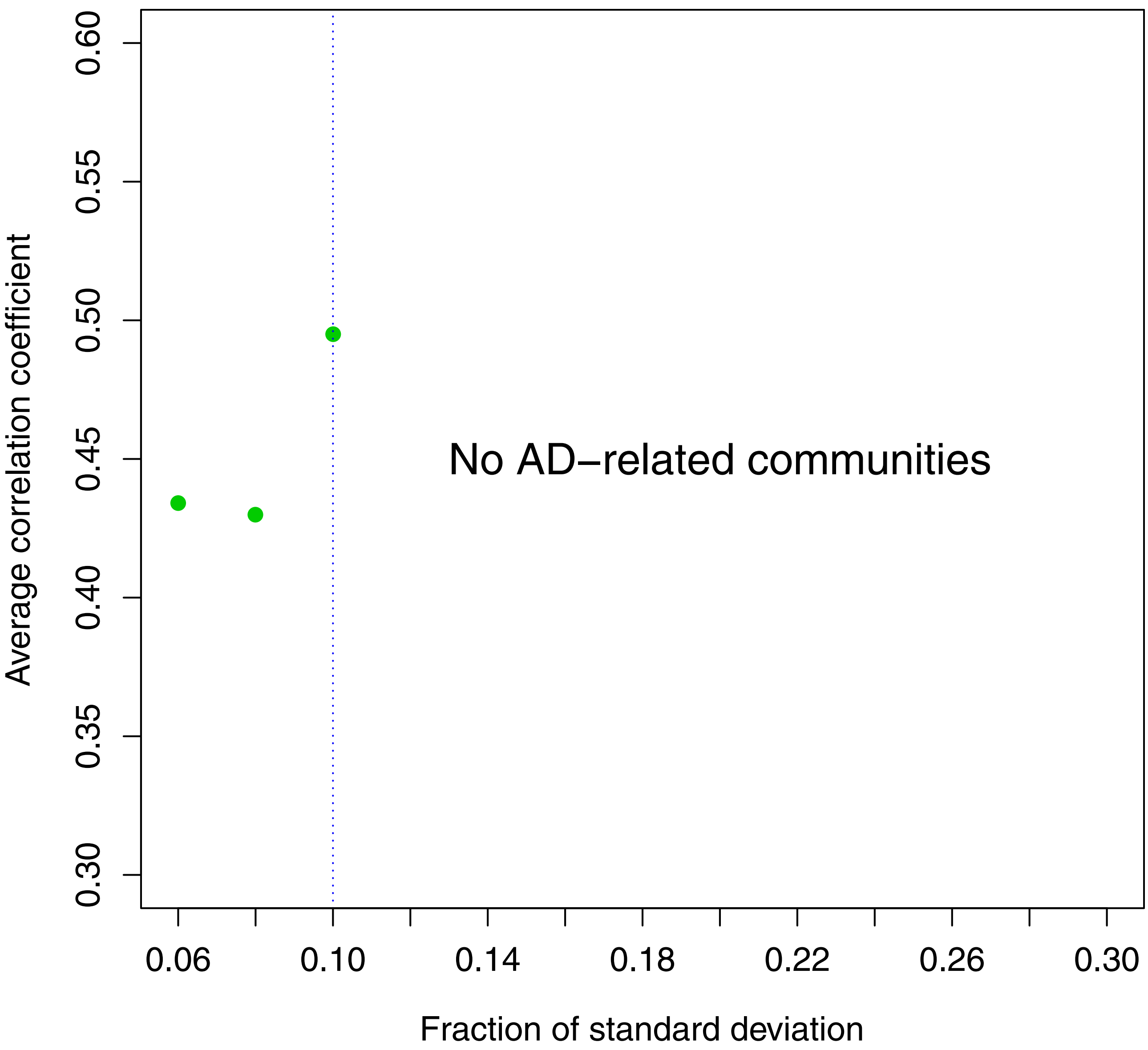

Fig 3. Average correlation coefficient with different cut. 0.1 σ_{max} is the optimal cut that maximizes the average correlation between the communities of interest and clinical features MMSE and NFT. In fact selecting genes exceeding 0.1 σ_{max} we did not find AD-related communities.

https://doi.org/10.1371/journal.pone.0226190.g003

2.2 Information entropy based on betweenness to select the best threshold value

With gene expressions selected in the preprocessing procedure, we built the network of co-expressed genes. Fig 5 shows entropy distribution based on betweenness, as a function of the threshold for the co-expression network. The distribution presents a maximum at the threshold value 0.74 which corresponds to the network configuration with the highest informative significance related to betweenness.

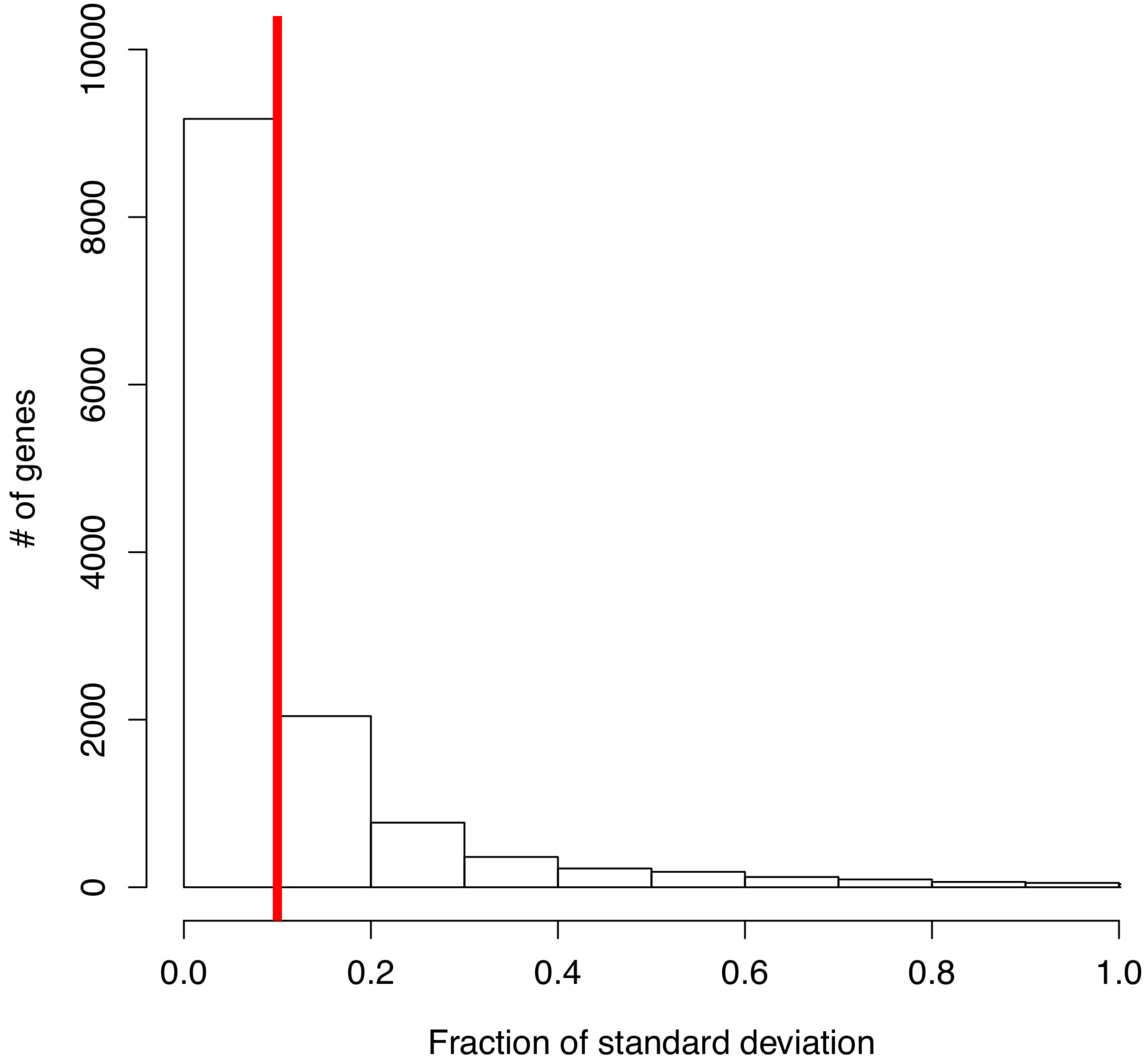

Fig 4. Distribution of standard deviation of gene expression. The vertical red line indicates the selection implemented on the data. Only the genes to right of the red line for further analysis have been selected.

https://doi.org/10.1371/journal.pone.0226190.g004

2.3 Community detection through Fast Greedy algorithm and correlation analysis with clinical phenotype

We applied the Fast Greedy community detection algorithm on the network at threshold equal to 0.74 resulting 127 gene communities. The Pearson's correlation coefficient between the principal component of communities and clinical phenotype was computed to identify

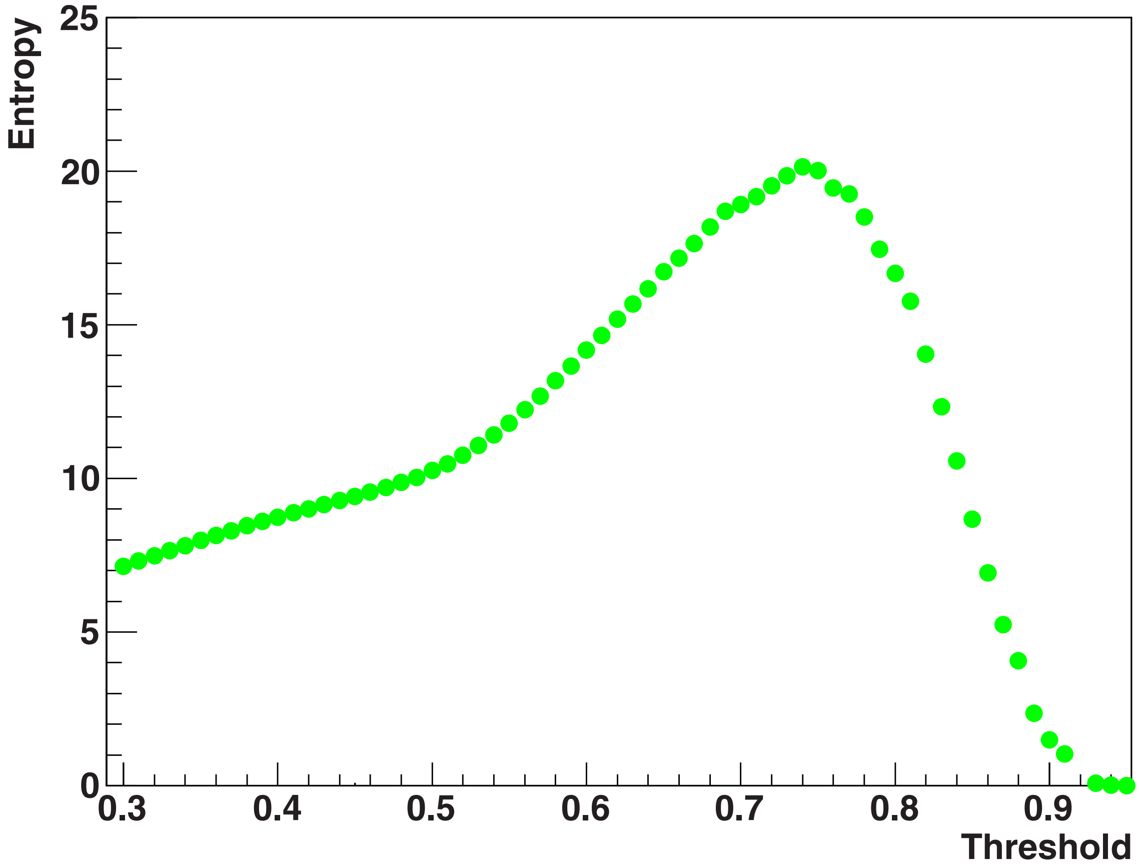

Fig 5. Betweenness entropy as a function of threshold. Information entropy distribution based on betweenness as a function of threshold for the whole network. The distribution presents a maximum at a threshold value equal to 0.74.

https://doi.org/10.1371/journal.pone.0226190.g005

communities significantly correlated with clinical features MMSE score and NFT value simultaneously. We found only two gene communities significantly associated with both clinical features:

- a community that included 72 genes (called C_1);
- a community that included 39 genes (called C_2).

Table 1 highlights the results obtained for C_1 and C_2 communities from the correlation test with MMSE and NFT. Fig 6 shows a schematization of C_1 and C_2 communities.

We computed the overlap coefficient through Eq 6 for different threshold values belonging to neighborhood 0.74 to evaluate the stability of C_1 and C_2 communities. Fig 7 displays the overlap coefficient computed in relation to C_1 and C_2 communities for different threshold values. The high values of overlap coefficient (> 0.6) certify the stability of the two communities over a wide range of the chosen threshold. To verify the robustness of results reported in Table 1 we applied the sampling procedure described in the section 1.5 on C_1 and C_2

Table 1. Results of correlation test between the first principal component of C_1 and C_2 communities and clinical features. The results show significant correlation (*p-value* < 0.05) of both gene communities with MMSE and NFT.

Community	MMSE	NFT
C_1	r = -0.46; p = 0.03	r = 0.59; p = 0.004
C_2	r = 0.47; p = 0.02	r = -0.46; p = 0.03

https://doi.org/10.1371/journal.pone.0226190.t001

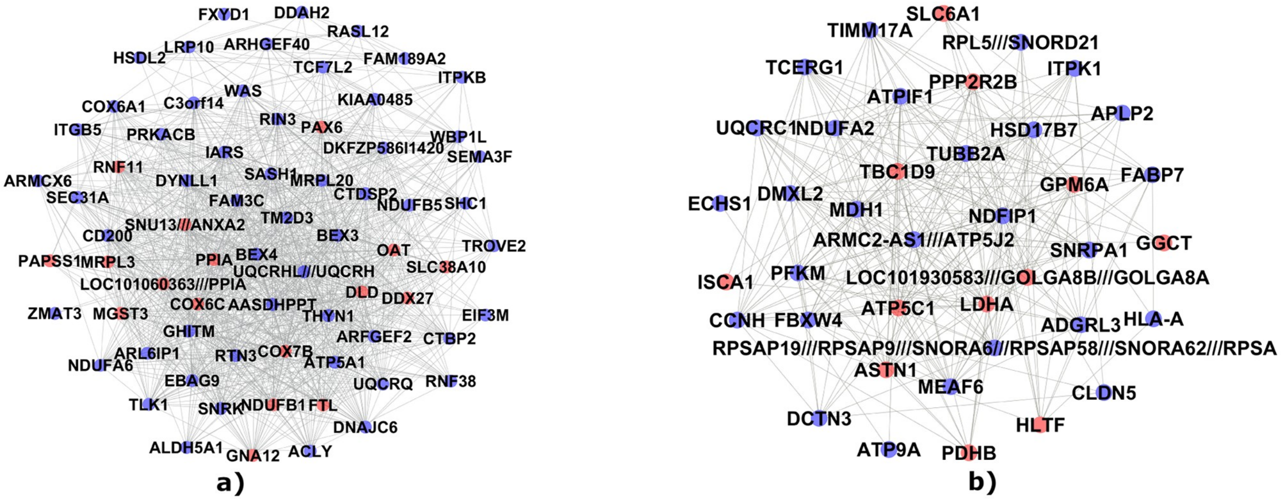

Fig 6. Schematization of C_1 and C_2 communities. Gene community C_1 (Panel a) and gene community C_2 (Panel b) composed by 72 and 39 genes respectively. In red we indicated genes significantly correlated with both clinical features simultaneously.

https://doi.org/10.1371/journal.pone.0226190.g006

communities and obtained distributions of correlation coefficients as shown in Fig 8. Distributions appear consistent with correlation coefficient values reported in Table 1.

The gene membership per community are reported in S4 Table. In S5 Table we listed genes of C_1 and C_2 communities with the weight in their contribution to the eigengene and their individual correlation with MMSE and NFT variables. Finally the list of detected communities with their correlation with clinical features is given in S6 Table.

2.4 Gene set enrichment analysis

C_1 and C_2 communities have been subjected to gene set enrichment analysis. The results for C_1 and C_2 communities are shown in Tables 2 and 3 respectively.

2.5 Hub genes identification

We identified 24 hub genes in C_1 community and 9 in C_2 community with $Kcs > 0.8$. Moreover we implemented a boostrap procedure, with 100 resamplings, to confirm these hub genes. In all resamplings, 28 genes belonging to C_1 community were clustered together as well as 13 of C_2 community. We called these modules sub-C_1 and sub-C_2. Hub genes of C_1 and C_2 communities are included in sub-C_1 and sub-C_2 modules, respectively. Hence, we implemented the correlation analysis described in section 1.5 for the hub genes of C_1 and C_2. The findings for MMSE score ($r_{hub-C_1} = -0.55, p_{hub-C_1} = 0.008; r_{hub-C_2} = 0.60, p_{hub-C_2} = 0.003$) and NFT value ($r_{hub-C_1} = 0.55, p_{hub-C_1} = 0.008; r_{hub-C_2} = -0.47, p_{hub-C_2} = 0.02$) show that the correlations between these genes and clinical features are improved compared to results reported in Table 1. Lists of hub genes for C_1 and C_2 communities are reported in S1 Table. Fig 9 displays overlap coefficient distributions for C_1 and C_2 communities and other communities found with $th = 0.74$ for each bootstrap resampling. C_1 and C_2 communities appear to be better preserved than other communities.

2.6 Identification of gene communities through WGCNA

We applied WGCNA algorithm to gene expressions selected in the section 2.1. We implemented the scale independence analysis and we found β value equal to 6 was the smallest threshold that resulted in a scale-free R^2 fit greater than 0.8, as shown in Fig 10.

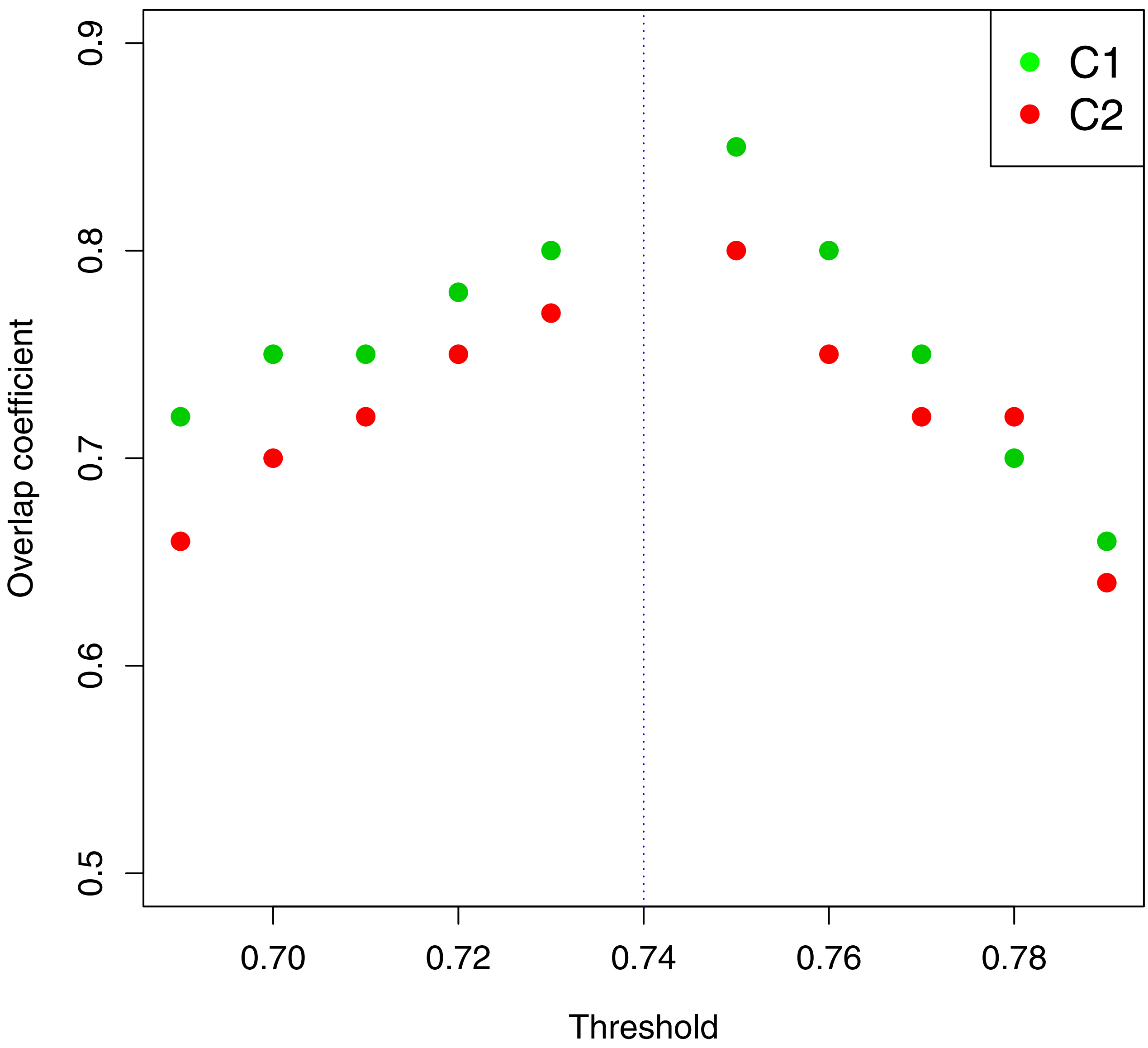

Fig 7. Overlap coefficient for C_1 and C_2 communities. The overlap coefficient as a function of threshold calculated in relation to the C_1 and C_2 communities. The dashed vertical line indicates the threshold value selected for our analysis.

https://doi.org/10.1371/journal.pone.0226190.g007

Through this configuration we obtained 21 gene communities. Fig 11 shows a dendrogram of 21 modules found by means of the WGCNA algorithm. We computed the linear correlation between communities and clinical features MMSE score and NFT value. Only one community, the blue module, was significantly associated with both clinical features ($r = -0.43$ with MMSE; $r = 0.43$ with NFT). This community was composed by 637 genes, and it contained the 78% of C_1 and C_2 communities (87 genes). All hub genes found in the previous section were included in the blue module. Fig 12 shows the overlap index between the blue module computed with $\beta = 6$ and other AD-related modules obtained for different threshold values. A strong module overlap emerges for different threshold values ($C_O > 0.75$) confirming the goodness of the choosen β. In S7 Table we reported for each threshold values and for different parameter values of sigmoid function the number of modules detected and their correlation

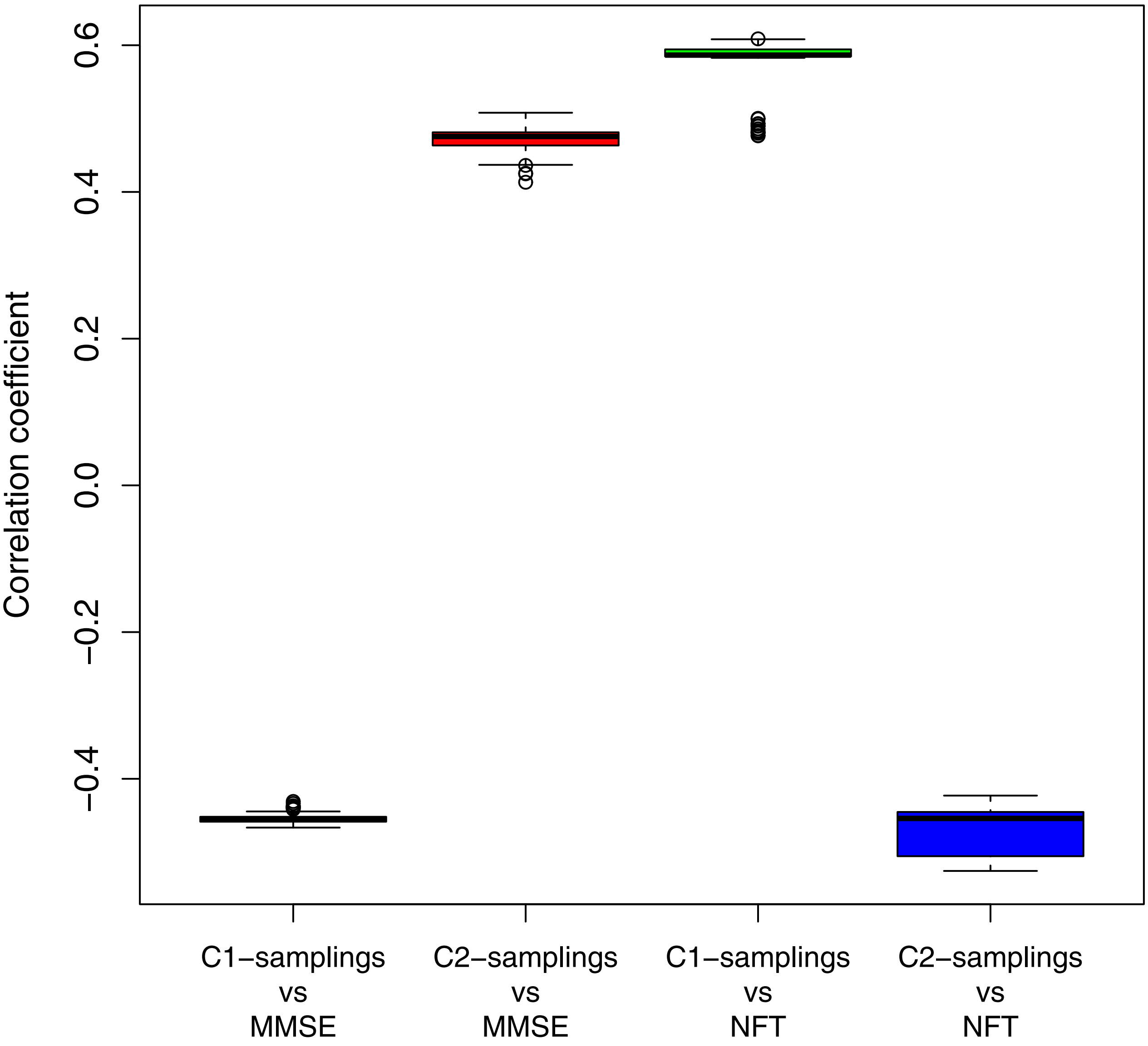

Fig 8. Distributions of correlation coefficient with the clinical features for 100 sampling of C_1 and C_2 communities. Random subsets of 10 genes are removed from C_1 and C_2 communities before to compute the first principal component.

https://doi.org/10.1371/journal.pone.0226190.g008

with the clinical features. Topological measures of co-expression network constructed through WGCNA and our method are reported in S2 and S3 Tables.

2.7 Identification of gene communities through traditional clustering methods and network reconstruction algorithms

On gene expressions selected in the section 2.1 we implemented other seven methods of community detection. Fig 13 displays the silhouette coefficient obtained by means of Eq 11 for agglomerative hierarchical clustering in many configurations. In particular in the panel A we presented distributions of the silhouette coefficient for the metrics and linkage criteria listed in the section 1.9. For each proposed methods we considered the number of clusters between 2 and 50. Instead Panel B shows the silhouette coefficient as a function of the number of clusters

Table 2. Results of gene set enrichment analysis for C_1 community. In the third column overlaps with gene set in the selected MSigDB gene set collection are reported. The fourth column displays the false discovery rate (FDR) analog of hypergeometric p-value after correction for multiple hypothesis testing according to Benjamini and Hochberg [43]. The table shows seven most significant enrichments.

Gene Set Name	Description	Genes in overlap	FDR p-value
BLALOCK ALZHEIMER'S DISEASE DN	Genes down-regulated in brain from patients with Alzheimer's disease—Homo sapiens	32	$1.25 \cdot 10^{-28}$
HALLMARK OXIDATIVE PHOSPHORYLATION	Genes encoding proteins involved in oxidative phosphorylation—Homo sapiens	11	$3.87 \cdot 10^{-11}$
GO ELECTRON TRANSPORT CHAIN	A process in which a series of electron carriers operate together to transfer electrons from donors to any of several different terminal electron acceptors to generate a transmembrane electrochemical gradient—Homo sapiens	9	$5.24 \cdot 10^{-11}$
GO CELLULAR RESPIRATION	The enzymatic release of energy from inorganic and organic compounds (especially carbohydrates and fats) which either requires oxygen (aerobic respiration) or does not (anaerobic respiration)—Homo sapiens	9	$1.24 \cdot 10^{-9}$
GO ORGANONITROGEN COMPOUND METABOLIC P C PROCESS	The chemical reactions and pathways involving organonitrogen compound—Homo sapiens	19	$6.23 \cdot 10^{-9}$
KEGG PARKINSONS DISEASE	Parkinson's disease—Homo sapiens	19	$1.38 \cdot 10^{-8}$
BLALOCK ALZHEIMER'S DISEASE UP	Genes up-regulated in brain from patients with Alzheimer's disease—Homo sapiens	18	$1.38 \cdot 10^{-8}$

https://doi.org/10.1371/journal.pone.0226190.t002

k for the chosen configuration (Manhattan-centroid). Based on the distribution of the silhouette coefficient, we chose *k* between 2 and 30 and we found no AD-related modules.

In Fig 14 we reported silhouette coefficient distributions computed for the Fuzzy c-means, for three different distance metrics and for a number of clusters between 2 and 50 (Panel A). Pick out Manhattan as metric, in panel B we reported the silhouette coefficient for a set of clusters number. Choosing *k* between 2 and 7 (since the average silhouette coefficient is greater than 0) no module was correlated with AD. In this work the default parameters were used for affinity propagation, ARACNE, GENIE3, and SPACE without additional tuning. Table 4 highlights the results obtained through the selected community detection methods.

2.8 Validation on independent datasets

Due to the relationship between aging and AD [57] we excluded the gene expression profiles of young (20-50 years) from GSE48350. In this way, we made comparable the three dataset age distributions, as displayed in Fig 15.

For GSE48350 database we found a first community composed by 35 genes with overlap of 23 with C_1 community (8 hub genes of C_1) and a second community of 28 genes, 23 members of C_2 community (8 hub genes of C_2). Instead applying our procedure on GSE29378 database we obtained a first community with 79 genes, 49 members of C_1 community (17 hub genes of C_1) and a second community containing 35 genes with overlap of 21 genes with C_2 community

Table 3. Results of gene set enrichment analysis for C_2 community. In the third column the genes in the overlapping gene sets are reported. The fourth column indicates the false discovery rate (FDR) analog of hypergeometric p-value after correction for multiple hypothesis testing according to Benjamini and Hochberg [43]. he table shows the top three significant enrichments.

Gene Set Name	Description	Genes in overlap	FDR p-value
BLALOCK ALZHEIMER'S DISEASE DN	Genes down-regulated in brain from patients with Alzheimer's disease—Homo sapiens	22	$1.07 \cdot 10^{-22}$
HALLMARK OXIDATIVE PHOSPHORYLATION	Genes encoding proteins involved in oxidative phosphorylation—Homo sapiens	9	$1.23 \cdot 10^{-10}$
KIM BIPOLAR DISORDER OLIGODENDROCYTE DENSITY CORR UP	Genes whose expression significantly and positively correlated with oligodendrocyte density in layer VI of BA9 brain region in patients with bipolar disorder—Homo sapiens	11	$5.49 \cdot 10^{-9}$

https://doi.org/10.1371/journal.pone.0226190.t003

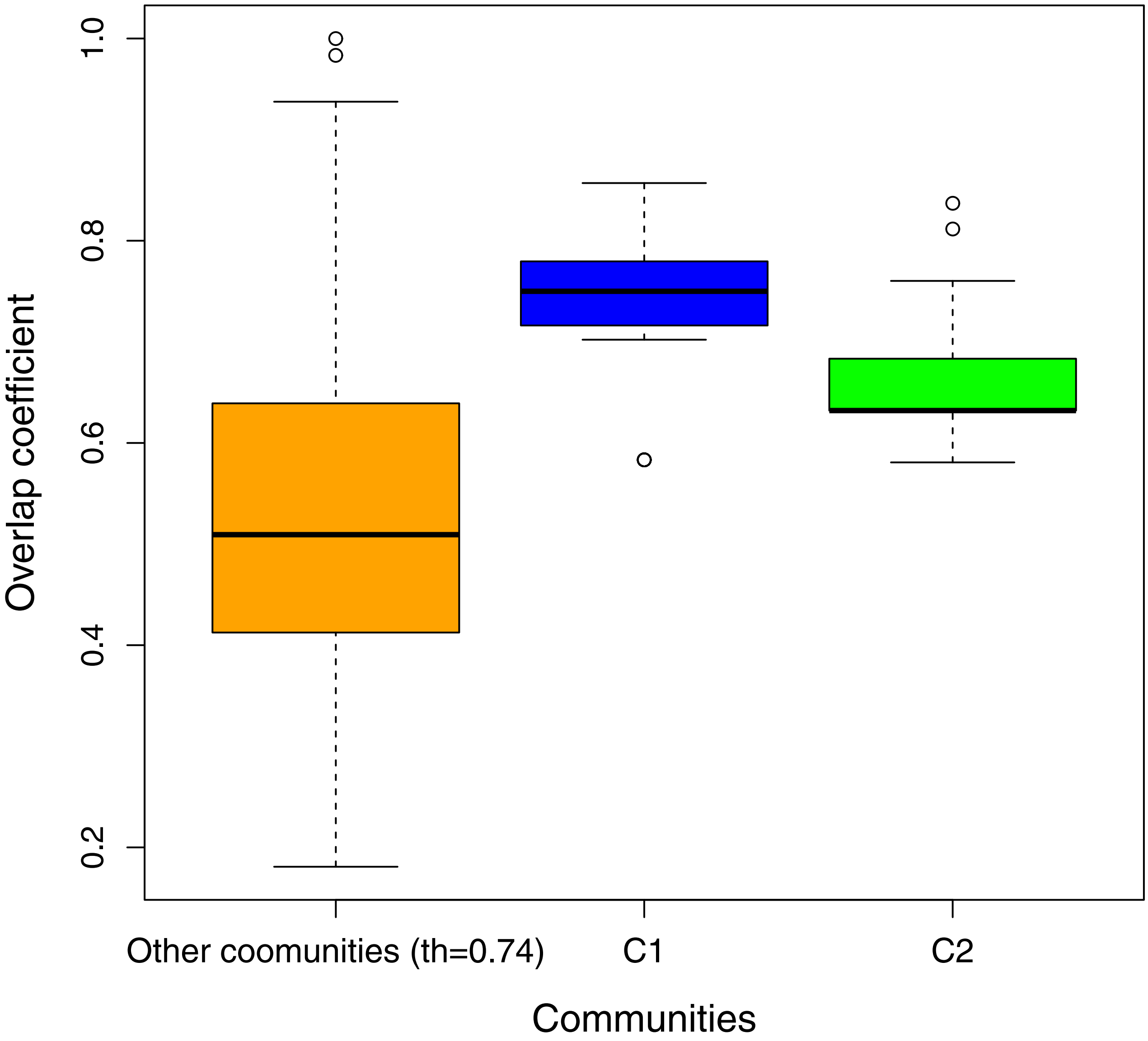

Fig 9. Overlap coefficient distributions for C_1 and C_2 communities and other communities found with *th* = 0.74. The overlap coefficient is computed for each bootstrap resampling.

https://doi.org/10.1371/journal.pone.0226190.g009

(5 hub genes of C_2). These findings highlight the replicability of our results on unseen datasets. Results of gene set enrichment analysis of these communities found for GSE48350 and GSE29378 are reported in S8 Table.

2.9 Discussion

In the present study, we implemented a co-expression network-based approach to analyze the whole genome expression data obtained from AD and control human hippocampus. The selected database GSE1297 was accompanied by the complete information of gene expression data and detailed clinical data. We applied a hard threshold analysis, proposed in a previous work [21], in which we analyzed the Shannon entropy based on betweenness of the network. We selected the network configuration with the highest informative significance. Through this

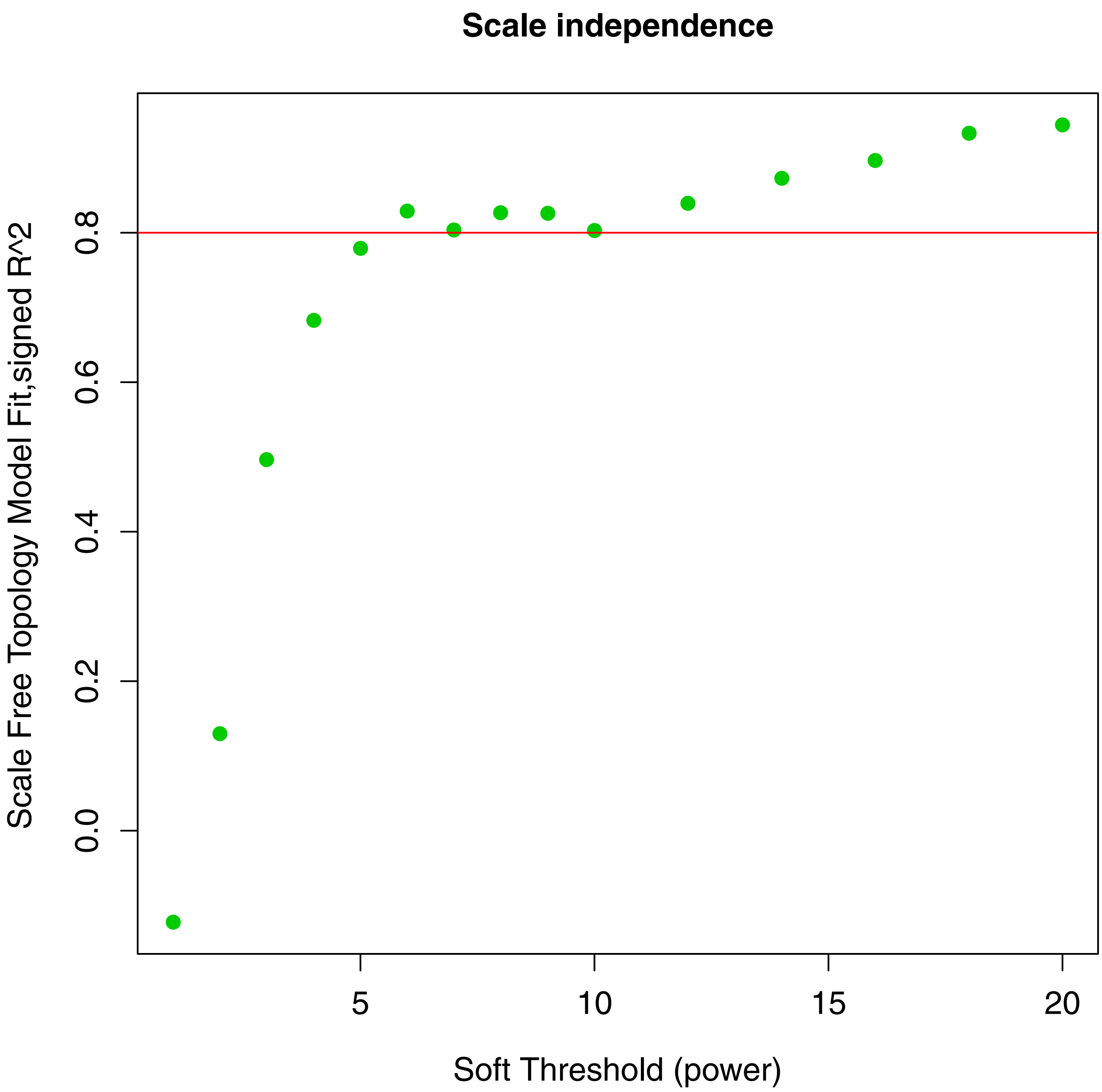

Fig 10. Scale free topology fitting index for different soft thresholds. β value equal to 6 was the smallest threshold that presented R^2 greater than 0.8.

https://doi.org/10.1371/journal.pone.0226190.g010

procedure we identified 127 gene communities. Due to the high threshold value selected (th = 0.74), genes in each community showed highly-related expression, indicating potential interaction and a common biological trend. By calculating the Pearson's correlation coefficient between the communities and two AD-related clinical features such as MMSE score and NFT value, we selected two communities (called C_1 and C_2) of interest. C_1 and C_2 communities were composed by 72 and 39 genes respectively and they showed significant but opposite correlations with the same variables (see Table 1). We compared our method with three traditional clustering algorithms: agglomerative hierarchical clustering, Fuzzy c-means, affinity propagation and with four network reconstruction approaches: WGCNA ARACNE, GENIE3,

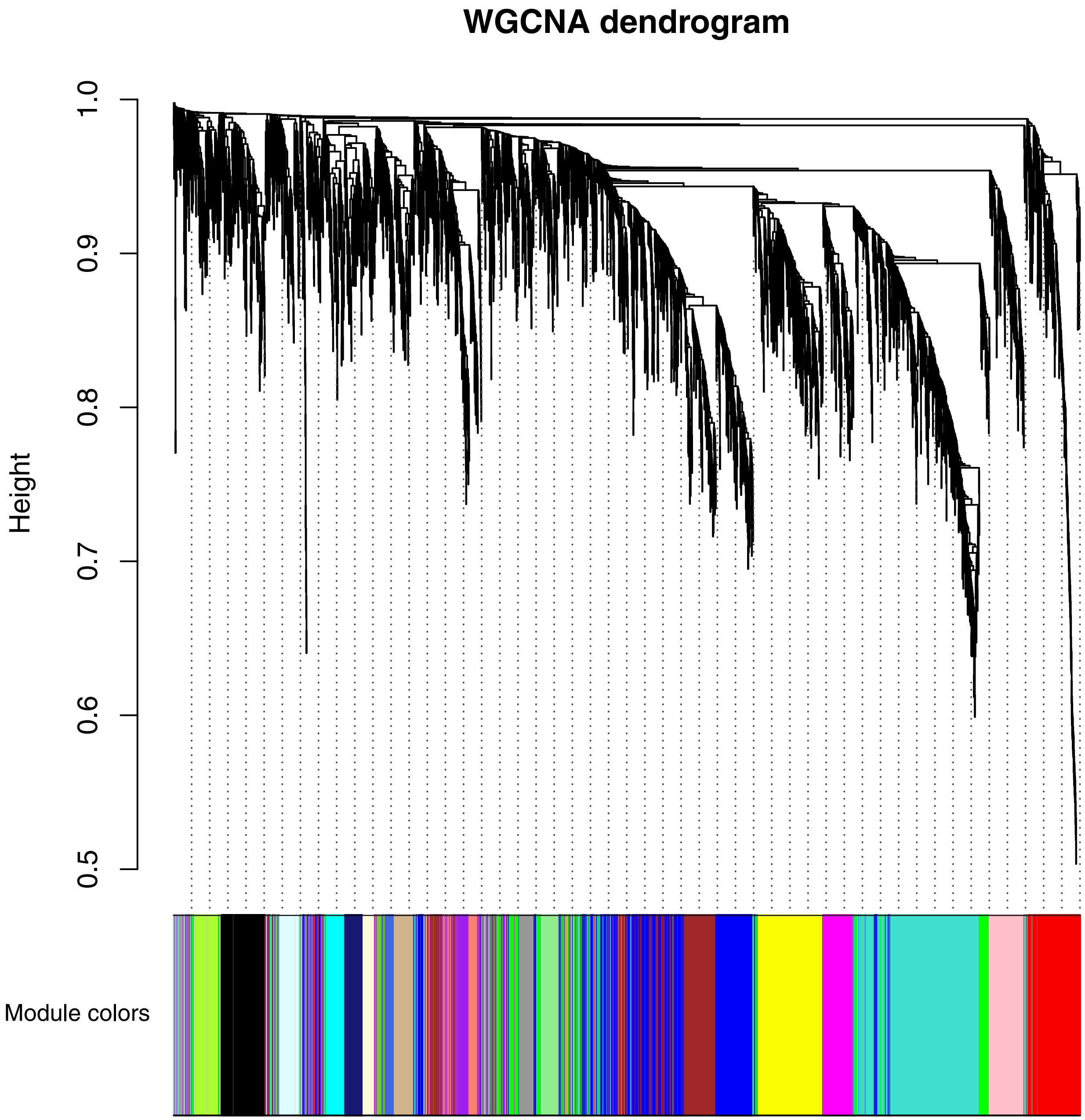

Fig 11. Dendrogram of the network with the modules indicated through the colors. This dendrogram was obtained through average linkage hierarchical clustering. The color spectrum underneath the plot indicates the module assignment determined by means of the Dynamic Tree Cut.

https://doi.org/10.1371/journal.pone.0226190.g011

and SPACE. As shown in Table 4 these other methods obtained communities with slightly lower correlations with clinical variables than C_1 and C_2 communities. In particular through WGCNA algorithm we found only one community of interest: the blue module. This community was quite large (637 genes) and the 78% of C_1 and C_2 communities was included in it. Our method preserved and condensed into two much smaller communities the relevant information highlighted by WGCNA. Furthermore, our analysis confirmed a limitation of the WGCNA approach, in which small gene communities may, in some cases, be incorporated into larger modules, with a loss of biological information [17]. Instead ARACNE and

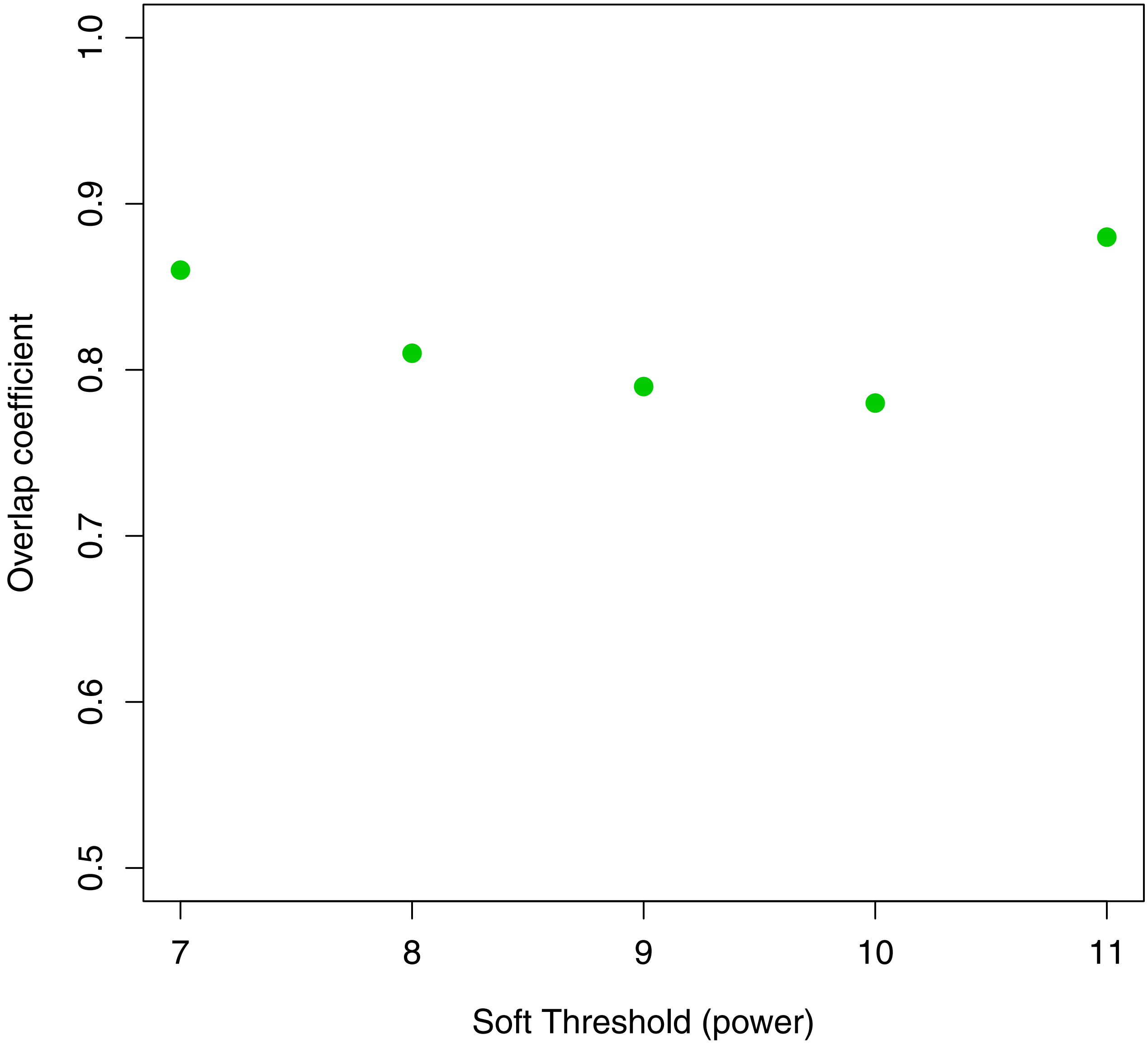

Fig 12. The overlap index between the blue module ($\beta = 6$) and other AD-related modules obtained for different threshold values. We obtained $C_0 > 0.75$ for all different threshold values.

https://doi.org/10.1371/journal.pone.0226190.g012

GENEIE3 algorithms suppress too many connections by creating a very large number of small communities. Compared to the remaining implemented techniques our findings are similar in terms of correlation to AD but C_1 and C_2 communities hold a different information content because they consist largely of different genes. The proposed approach outlines the relationships among genes and paves the way to further studies about the physiological interpretation of AD-related communities. To assess the reproducibility of our findings, we performed our community detection method on two independent datasets: GSE48350 and GSE29378. We obtained two communities for each datasets within at least the 65% of C_1 and C_2 genes. These results underline the importance of the found communities in relation to AD. In previous work a lot of genes belonging to C_1 and C_2 communities had already been highlighted as linked to AD. For example RIN3 was found to play key roles in the development of AD ad in

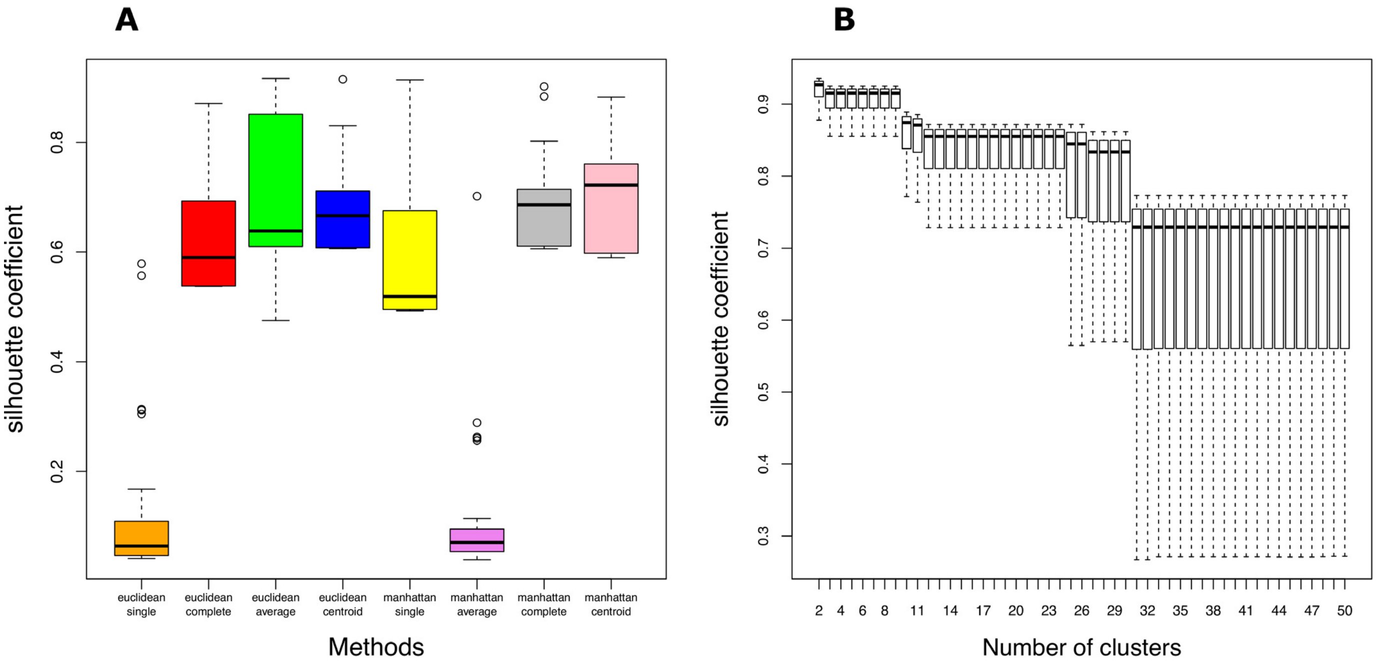

Fig 13. Silhouette coefficient distributions for agglomerative hierarchical clustering method. Silhouette coefficient distributions computed for different configurations (Panel A), and for different number of clusters (Panel B) only for the chosen configuration Manhattan-centroid.

https://doi.org/10.1371/journal.pone.0226190.g013

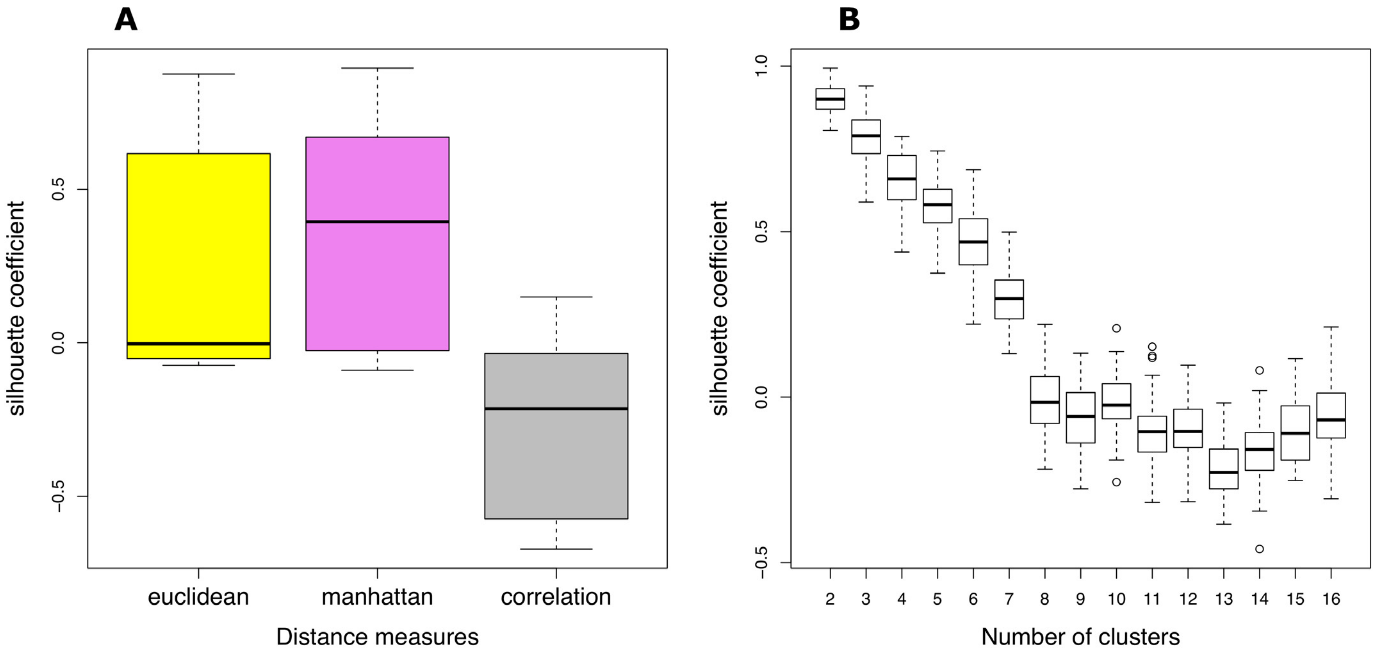

Fig 14. Silhouette coefficient distributions for fuzzy c-means method. Silhouette coefficient distributions computed for different distance measures (Panel A), and for different number of clusters (Panel B) only for the chosen metric Manhattan.

https://doi.org/10.1371/journal.pone.0226190.g014

Table 4. Summary table of the results obtained through the proposed community detection methods. In the fourth column the number of genes in common with C_1 and C_2 communities is shown in bold. We reported only significant correlations ($p\text{-}value < 0.05$).

Method	Number of communities	AD-related communities	Number of genes in AD-related communities	Correlation coefficient with NFT	Correlation coefficient with MMSE
Our method	127	C_1	71	0.59	−0.46
		C_2	39	−0.46	0.47
WGCNA	21	Blue module	637(**87**)	0.43	−0.43
Affinity propagation	233	A_1	31(**7**)	0.45	−0.43
		A_2	39(**23**)	0.45	−0.43
Agglomerative clustering	Between 2 and 30	None			
Fuzzy c-means	Between 2 and 7	None			
ARACNE	674	AR_1	8(**2**)	−0.45	0.44
		AR_2	7(**5**)	−0.43	0.43
GENIE3	675	G_1	9(**3**)	−0.44	0.44
		G_2	7(**2**)	−0.44	0.46
SPACE	24	S_1	25(**2**)	0.44	−0.44

https://doi.org/10.1371/journal.pone.0226190.t004

particular it resulted associated with AD risk [58, 59]. J.W. Liang et al. [40] reported an AD-related community containing RIN3 and GNA12, two genes belonging to C_1. Seyfried N.T. et al. [60] by means of a co-expression network analysis found 10 modules correlated with AD phenotype. In particular in M1 module (composed by 1294 genes) were contained 12 genes found in C_1 (PPIA, RTN3, COX6C, DNAJC6, AASDHPPT, COX7B, TM2D3, RNF11, BEX4, CD200, MRPL3, THYN1) and 9 found in C_2 (TUBB2A, GPM6A, MDH1, PFKM, ATP9A, UQCRC1, ASTN1, ATPIF1, GGCT) while in M5 module (composed by 775 genes) were clustered together 9 genes of C_1 (PRKACB, OAT, GHITM, PAPSS1, IARS, NDUFA6, ACLY, TLK1, C3orf14) and 4 of C_2 (LDHA, PDHB, ISCA1, TCERG1). Miller J.A. et al. [61] used WGCNA to identify 12 distinct modules related to synapticand metabolic processes of AD. Specifically we found a good overlap with the module enriched in mitochondrion pathway (containing 366 genes): 13 genes belonging to C_1 community (COX6C, ATP5A1, OAT, AASDHPPT, GHITM, COX7B, TM2D3, IARS, FAM3C, NDUFB5, DLD, CD200, MRPL3) and 9 to C_2 community (MDH1, ATP5C1, NDFIP1, PDHB, PFKM, UQCRC1, TIMM17A, CCNH, ATPIF1).

Further gene set enrichment analysis showed that both communities C_1 and C_2 were mostly enriched in AD, oxidative phosphorylation (OXPHOS), Krebs (TCA) cycle, Parkinson's disease and bipolar disorder pathways. Several works studied the pathway enrichments associated with AD [62–64]. For example, Naj et al. [65] provided a comprehensive review of genomic studies of AD. Moreover, many studies [66, 67] confirmed that oxidative phosphorylation and electron transfer defects were closely associated with neuro-degenerative diseases, such as AD. Oxidative phosphorylation rapresents the apex of a series of energy transformations indicated as cellular respiration or simply respiration in their entirety [68]. In this cellular process the electron transport chain constitutes a proton gradient across the inner mitochondrial membrane, in which the synthesis of ATP is driven through the chemiosmosis. The role of OXPHOS changes in the pathogenesis of AD is controversial. Abnormalities in cellular bioenergetics have been detected in a lot of people affected by AD and their links with dementia have been highlighted in several experiments in vivo and in vitro [69]. As reported in Table 2, C_1 community results also enriched for genes involved in Krebs (TCA) cycle. The TCA cycle is a mitochondrial metabolic process essential for generating the proton gradient across the inner membrane of the mitochondria that is used to produce ATP [70]. A connection between

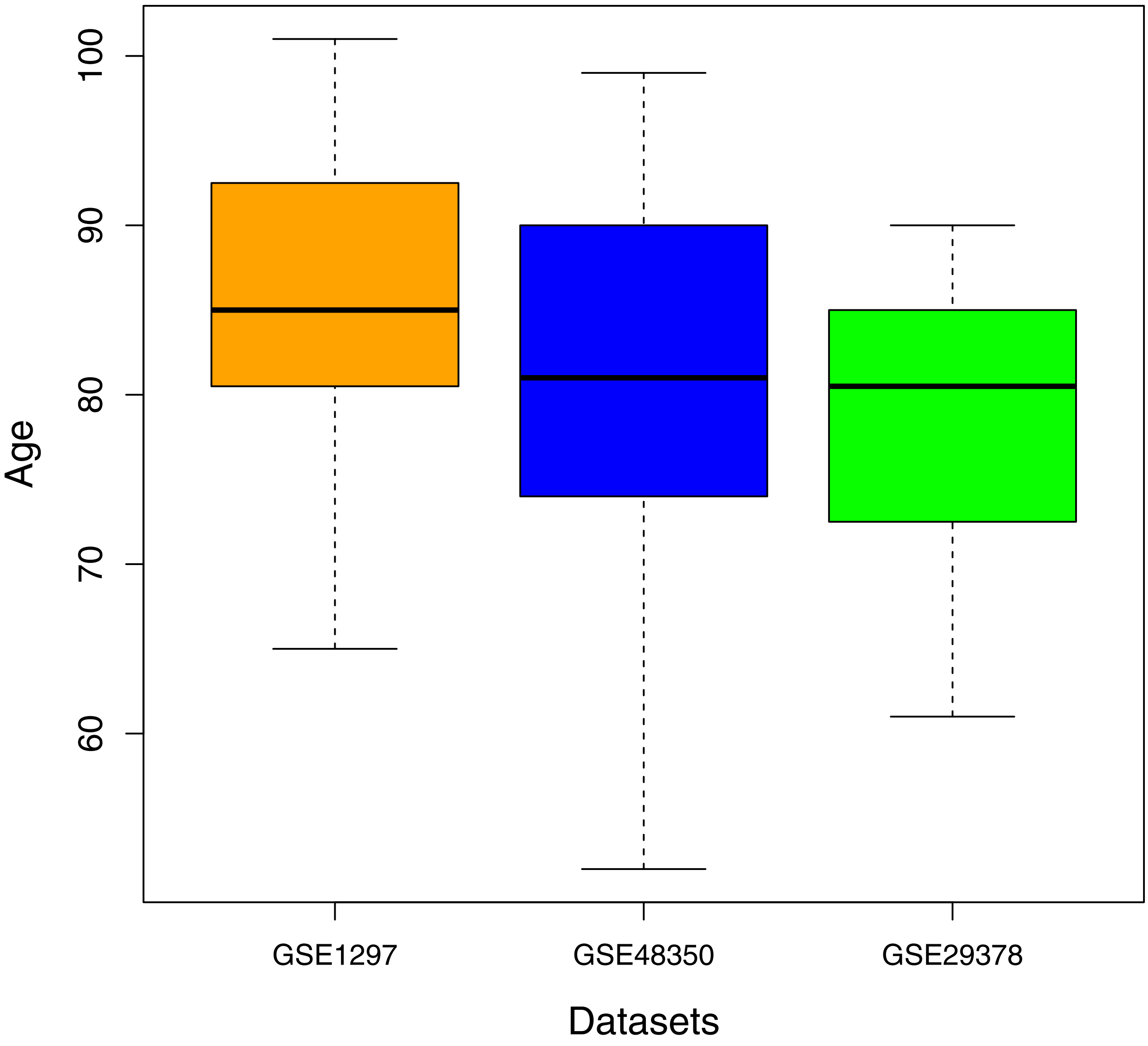

Fig 15. Age distributions of the three datasets analyzed. To make the three distributions comparable we did not consider 17 samples of young subjects in the database GSE48350.

https://doi.org/10.1371/journal.pone.0226190.g015

aberrations in TCA cycle and AD has been widely witnessed [70–73]. For example, Bubber et al. [74] stated that the metabolic activity of the TCA cycle decreases in AD mitochondria due to a decline in several of the enzymes of the cycle. As reported in S8 Table also the found communities within GSE48350 and GSE29378 were mostly enriched in AD, OXPHOS and energy cellular processes pathways. Finally, we implemented a study of hub genes by means of the Kleinberg's centrality and identified 24 hub genes in C_1 community and 9 in C_2 community. The robustness of our findings was confirmed through a bootstrap procedure. In fact, C_1 and $C2$ communities appear on average more stable than the other communities, as displayed in Fig 9. The communities of the hub genes presented a stronger correlation with the clinical phenotype than C_1 and C_2 communities. This could indicate the presence of a robust core component of genes within the detected communities more closely related to AD. As shown in

Fig 6 and reported in S5 Table the found hub nodes often do not coincide with the genes most correlated to AD phenotype. In fact our procedure was based on a centrality measure and a bootstrap verification independently by correlation with clinical features. The detected communities, C_1 and C_2, and in particular the communities of their hub genes could help to understand the mechanisms of Alzheimer's disease and they can be potential targets for a future AD therapy. Clearly, a more in-depth clinical validation is necessary to understand how these genes are implicated in the biological processes linked to AD.

3 Conclusions

In this paper we implemented a method proposed in a previous work and used in an international competition to detect gene communities linked to AD. Our strategy was based on a co-expression network analysis and a study of Shannon entropy of the betweenness. The pivotal role of co-expression networks consists in representing binary relationships between individual genes that may highlight obscure processes of cellular communication. As reported in Table 4 the communities found with our method are significantly smaller than the module obtained through WGCNA but much more populated than the communities detected by means of ARACNE and GENIE3 where gene connections appear suppressed. Compared to the remaining implemented techniques our results are similar in terms of correlation to AD but the two found communities have a different information content because they consist largely of different genes. These communities were stable and mostly replicated on two independent databases. Furthermore they contained some genes already known to be linked to AD as RIN3 and GNA12. The detected communities resulted mostly enriched in pathways closely associated with neuro-degenative diseases as energy cellular processes. A study of the hub genes of the two communities revealed even stronger correlation of hub genes communities to the clinical phenotype. Further exploration on the two AD-relevant communities and the detected hub genes, combined with analysis using different clinical tools such as neuroimaging [75], are needed to understand physiological mechanisms of AD.

Supporting information

S1 Table. Lists of hub genes of C_1 and C_2 communities.
(PDF)

S2 Table. Topological parameters of co-expression network constructed using WGCNA algorithm.
(PDF)

S3 Table. Topological parameters of co-expression network constructed using the proposed method [21].
(PDF)

S4 Table. The gene membership per community found by means of the proposed method.
(XLSX)

S5 Table. Lists of C_1 and C_2 genes containing the weight in their contribution to the eigengene and their individual correlation with MMSE and NFT variables.
(XLSX)

S6 Table. Lists of detected communities with their correlation with clinical features.
(XLSX)

S7 Table. Modules detected trough WGCNA and their correlation with the clinical features in function of threshold values.
(XLSX)

S8 Table. Results of gene set enrichment analysis of the detected communities in GSE48350 and GSE29378 databases.
(PDF)

Author Contributions

Conceptualization: Alfonso Monaco, Nicola Amoroso, Loredana Bellantuono, Eufemia Lella, Angela Lombardi, Anna Monda, Andrea Tateo, Roberto Bellotti, Sabina Tangaro.

Data curation: Alfonso Monaco.

Formal analysis: Alfonso Monaco.

Methodology: Alfonso Monaco, Nicola Amoroso, Andrea Tateo.

Supervision: Alfonso Monaco, Roberto Bellotti, Sabina Tangaro.

Validation: Nicola Amoroso, Loredana Bellantuono, Eufemia Lella, Angela Lombardi, Anna Monda, Andrea Tateo, Roberto Bellotti, Sabina Tangaro.

Writing – original draft: Alfonso Monaco.

Writing – review & editing: Alfonso Monaco, Nicola Amoroso, Loredana Bellantuono, Eufemia Lella, Angela Lombardi, Anna Monda, Andrea Tateo, Roberto Bellotti, Sabina Tangaro.

References

1. Brettschneider J, Tredici KD, Lee VM, Trojanowski JQ. Spreading of pathology in neurodegenerative diseases: a focus on human studies. Nature Reviews Neurology. 2015; 16(2): 109. https://doi.org/10.1038/nrn3887
2. Prince M, Comas-Herrera A, Knapp M, Guerchet M, Karagiannidou M. World Alzheimer report 2016: improving health- care for people living with dementia: coverage quality and costs now and in the future. Alzheimer's Disease International. 2016.
3. Noble W, Olm V, Takata K, Casey E, Mary O, Meyerson J, et al. Cdk5 is a key factor in tau aggregation and tangle formation in vivo. Neuron. 2003; 38(4): 555–565. https://doi.org/10.1016/s0896-6273(03)00259-9 PMID: 12765608
4. Johnson GV, Bailey CD. Tau, where are we now? Journal of Alzheimer's Disease. 2002; 4: 375–398. https://doi.org/10.3233/jad-2002-4505 PMID: 12446970
5. Tanzi RE, Bertram L. New frontiers in Alzheimer's disease genetics. Neuron. 2001; 32: 181–184. https://doi.org/10.1016/s0896-6273(01)00476-7 PMID: 11683989
6. Klein WL, Krafft GA, Finch CE. Targeting small Abeta oligomers: the solution to an Alzheimer's disease conundrum? Trends in Neurosciences. 2001; 24(4): 219–224. https://doi.org/10.1016/s0166-2236(00)01749-5 PMID: 11250006
7. Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. Science. 2012; 297(5580): 353–356. https://doi.org/10.1126/science.1072994
8. Mucke L, Masliah E, Yu GQ, Mallory M, Rockenstein EM, Tatsuno G. et al. High-level neuronal expression of abeta 1-42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. Journal of Neuroscience. 2000; 20: 4050–4058. https://doi.org/10.1523/JNEUROSCI.20-11-04050.2000 PMID: 10818140
9. Gouras GK, Tsai J, Naslund J, Vincent B, Edgar M, Checler F, et al. Intraneuronal Abeta42 accumulation in human brain. The American Journal of Pathology. 2000; 156: 15–20. https://doi.org/10.1016/s0002-9440(10)64700-1 PMID: 10623648

10. Aksenov MY, Aksenova MV, Butterfield DA, Geddes JW, Markesbery WR. Protein oxidation in the brain in Alzheimer's disease. Neuroscience. 2001; 103: 373–383. https://doi.org/10.1016/s0306-4522(00)00580-7 PMID: 11246152

11. Gibson GE. Intraneuronal Abeta42 accumulation in human brain. Free Radical Biology & Medicine. 2002; 32: 1061–1070.

12. Gemma C, Mesches MH, Sepesi B, Choo K, Holmes DB, Bickford PC. Diets Enriched in Foods with High Antioxidant Activity Reverse Age-Induced Decreases in Cerebellar β-Adrenergic Function and Increases in Proinflammatory Cytokines. Journal of Neuroscience. 2002; 22: 6114–6120. https://doi.org/10.1523/JNEUROSCI.22-14-06114.2002 PMID: 12122072

13. Mrak RE, Griffin WS. Interleukin-1, neuroinflammation, and Alzheimer's disease. Neurobiology of Aging. 2001; 22: 903–908. https://doi.org/10.1016/s0197-4580(01)00287-1 PMID: 11754997

14. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proceedings of the National Academy of Sciences. 2004; 101(7):2173–2178. https://doi.org/10.1073/pnas.0308512100

15. Monaco A, Sforza G, Amoroso N, Antonacci M, Bellotti R, de Tommaso M, et al. The PERSON project: a serious brain- computer interface game for treatment in cognitive impairment. Health and Technology. 2019; 9(2): 123–133. https://doi.org/10.1007/s12553-018-0258-y

16. Wilson RJ. Introduction to Graph theory. 4th ed. Addison Wesley, Longman; 1996.

17. Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9(1):1. https://doi.org/10.1186/1471-2105-9-559

18. Dawson JA, Kendziorski C. An empirical Bayesian approach for identifying differential coexpression in high-throughput experiments Biometrics. 2012; 68: 455–465. https://doi.org/10.1111/j.1541-0420.2011.01688.x PMID: 22004327

19. Liu BH, Yu H, Tu K, Li C, Li YX and Li YY. DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics. 2010; 26: 2637–2638. https://doi.org/10.1093/bioinformatics/btq471 PMID: 20801914

20. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4:Article17. https://doi.org/10.2202/1544-6115.1128 PMID: 16646834

21. Monaco A, Monda A, Amoroso N, Bertolino A, Blasi G, Di Carlo P, et al. A complex network approach reveals a pivotal substructure of genes linked to schizophrenia. PLoS ONE. 2018; 13(1): e0190110. https://doi.org/10.1371/journal.pone.0190110 PMID: 29304112

22. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, Li J, et al. Assessment of network module identification across complex diseases. Nature Methods. 2019; 16: 843–852. https://doi.org/10.1038/s41592-019-0509-5 PMID: 31471613

23. Chung N, Zhang XD, Kreamer A, Locco L, Kuan P, Bartz S, et al. Median Absolute Deviation to Improve Hit Selection for Genome-Scale RNAi Screens. Journal of Biomolecular Screening. 2008; 13: 149. https://doi.org/10.1177/1087057107312035 PMID: 18216396

24. Spitz A, Gimmler A, Stoeck T, Zweig KA, Horvat EA. Assessing Low-Intensity Relationships in Complex Networks. PLoS ONE. 2016; 11(4): e0152536. https://doi.org/10.1371/journal.pone.0152536 PMID: 27096435

25. Pinney JW, Westhead DR. Betweenness-based decomposition methods for social and biological networks. Interdisciplinary Statistics and Bioinformatics. 2006; 25: 87–90.

26. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC bioinformatics. 2005; 1: 6–39.

27. Brandes U. On variants of shortest-path betweenness centrality and their generic computation. Social Networks. 2008; 30(2): 136–145. https://doi.org/10.1016/j.socnet.2007.11.001

28. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal. 1958; 27: 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

29. West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Scientific Reports. 2012; 2: 802. https://doi.org/10.1038/srep00802 PMID: 23150773

30. Fortunato S. Community Detection in Graphs. Physics Report. 2010; 486: 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

31. Ravasz E. Detecting hierarchical modularity in biological networks. Computational Systems Biology. 2009; 541: 145–160. https://doi.org/10.1007/978-1-59745-243-4_7

32. Newman MEJ. Fast algorithm for detecting community structure in networks. Physical Review E. 2004; 69(6): 066133. https://doi.org/10.1103/PhysRevE.69.066133

33. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004; 69(2): 026113. https://doi.org/10.1103/PhysRevE.69.026113

34. Arenas A, Diaz-Guilera A. Synchronization and modularity in complex networks. European Physical Journal ST. 2007; 143: 19–25. https://doi.org/10.1140/epjst/e2007-00066-2

35. Lozano S, Duch J, Arenas A. Analysis of large social datasets by community detection. European Physical Journal ST. 2007; 143: 257–259. https://doi.org/10.1140/epjst/e2007-00098-6

36. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E. 2004; 70(6): 066111. https://doi.org/10.1103/PhysRevE.70.066111

37. Sieberts SK, Schadt EE. Moving toward a system genetics view of disease. Mammalian Genome. 2007; 18(6–7): 389–401. https://doi.org/10.1007/s00335-007-9040-6 PMID: 17653589

38. Foroushani A, Agrahari R, Docking R, Chang L, Duns G, Hudoba M, et al. Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: an introduction to the Pigengene package and its applications. BMC Medical Genomics. 2017; 10:16. https://doi.org/10.1186/s12920-017-0253-6 PMID: 28298217

39. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences. 2006; 103(47): 17973–17978. https://doi.org/10.1073/pnas.0605938103

40. Liang JW, Fang ZY, Huang Y, Liuyang ZY, Zhang XL, Wanga JL, et al. Application of Weighted Gene Co-Expression Network Analysis to Explore the Key Genes in Alzheimer's Disease Journal of Alzheimer's Disease. 2018; 65: 1353–1364. https://doi.org/10.3233/JAD-180400 PMID: 30124448

41. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: A desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007; 23(23): 3251–3253. https://doi.org/10.1093/bioinformatics/btm369 PMID: 17644558

42. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43): 15545–15550. https://doi.org/10.1073/pnas.0506580102

43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Royal Statistical Society. 1995; 57(1): 289–300.

44. Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM. 1999; 46(5): 604–632. https://doi.org/10.1145/324133.324140

45. Mooney CZ, Duval RD. Bootstrapping: a nonparametric approach to statistical inference. 1st ed. Newbury Park, CA: Sage University Paper; 1993.

46. Nankervis JC. Computational algorithms for double bootstrap confidence intervals. Computational Statistics & Data Analysis. 2005; 49(2): 461–474. https://doi.org/10.1016/j.csda.2004.05.023

47. Vijaymeena MK, Kavitha K. A Survey on Similarity Measures in Text Mining. Machine Learning and Applications: An International Journal. 2016: 3(1): 19–28. https://doi.org/10.5121/mlaij.2016.3103

48. Saelenes W, Cannoodt R, Saeys Y. A comprehensive evalutation of module detection methods for gene expression data. Nature Communications. 2005; 9:1090. https://doi.org/10.1038/s41467-018-03424-4

49. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

50. Bezdek JC. Cluster validity with fuzzy sets Journal of Cybernetics. 1974; 3: 58–73. https://doi.org/10.1080/01969727308546047

51. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007; 315: 972–976. https://doi.org/10.1126/science.1136800 PMID: 17218491

52. Chavez Montes RA, Coello G, Gonzalez-Aguilera KL, et al. ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks. BMC Plant Biology. 2014; 14: 97. https://doi.org/10.1186/1471-2229-14-97 PMID: 24739361

53. Huynh-Thu VA, Irrthum A, Wehenkel L, et al. Inferring regulatory networks from expression data using tree-based methods. PLoS One 2010; 5: e0012776. https://doi.org/10.1371/journal.pone.0012776

54. Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models. Journal of the American Statistical Association. 2009; 104(486): 735–746. https://doi.org/10.1198/jasa.2009.0126 PMID: 19881892

55. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics. 2018; 19(4): 575–592. https://doi.org/10.1093/bib/bbw139 PMID: 28077403

56. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing Statistical Methods for Constructing Large Scale Gene Networks. PLoS One. 2012; 7(1): e29348. https://doi.org/10.1371/journal.pone.0029348 PMID: 22272232

57. Lanke V, Moolamalla STR, Roy D, Vinod PK. Integrative Analysis of Hippocampus Gene Expression Profiles Identifies Network Alterations in Aging and Alzheimer's Disease. Frontiers in Aging Neuroscience. 2018; 23;10:153. https://doi.org/10.3389/fnagi.2018.00153

58. Moustafa AA, Hassan M, Hewedi DH, Hewedi I, Garami JK, Al Ashwalet H, et al. Genetic underpinnings in Alzheimer's disease—a review. Reviews in the Neurosciences. 2017; 29(1). https://doi.org/10.1515/revneuro-2017-0036

59. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's diseas. Nature Genetics. 2013; 45(12):1452–8. https://doi.org/10.1038/ng.2802 PMID: 24162737

60. Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, et al. A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. Cell Systems. 2017; 4(1): 60–72. https://doi.org/10.1016/j.cels.2016.11.006 PMID: 27989508

61. Miller AJ, Oldham MC, Geschwind DH. A Systems Level Analysis of Transcriptional Changes inAlzheimer's Disease and Normal Aging. The Journal of Neuroscience. 2008; 28(6): 1410–1420. https://doi.org/10.1523/JNEUROSCI.4098-07.2008 PMID: 18256261

62. Bird TD. Genetic aspects of Alzheimer disease. Genetics in Medicine. 2008; 10(4): 231–239. https://doi.org/10.1097/GIM.0b013e31816b64dc PMID: 18414205

63. Ertekin-Taner N. Genetics of Alzheimer's disease: a centennial review. Neurologic Clinics. 2007; 25(3): 611–667. https://doi.org/10.1016/j.ncl.2007.03.009 PMID: 17659183

64. Mattson MP. Pathways towards and away from Alzheimer's disease. Nature. 2004; 430(7000): 631–639. https://doi.org/10.1038/nature02621 PMID: 15295589

65. Naj CA, Schellenberg GD. Genomic Variants, Genes, and Pathways of Alzheimer's Disease: An Overview. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2017; 174(1): 5–26. https://doi.org/10.1002/ajmg.b.32499

66. Zabel C, Nguyen HP, Hin SC, Hartl D, Mao L and Klose J. Proteasome and oxidative phoshorylation changes may explain why aging is a risk factor for neurodegenerative disorders. Journal of Proteomics. 2010; 73: 2230–2238. https://doi.org/10.1016/j.jprot.2010.08.008 PMID: 20813214

67. Shoffner JM. Oxidative phosphorylation defects and Alzheimer's disease. Neurogenetics. 1997; 1(1): 13–9. https://doi.org/10.1007/s100480050002 PMID: 10735269

68. Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th ed. New York: W.H. Freeman; 2002.

69. Swerdlow RH. Mitochondria and Mitochondrial Cascades in Alzheimer's Disease. Journal of Alzheimer's Disease. 2018; 62(3): 1403–1416. https://doi.org/10.3233/JAD-170585 PMID: 29036828

70. Atamna H, Frey WH. Mechanisms of mitochondrial dysfunction and energy deficiency in Alzheimer's disease. Mitochondrion. 2007; 7(5): 297–310. https://doi.org/10.1016/j.mito.2007.06.001 PMID: 17625988

71. Atamna H, Newberry J, Erlitzki R, Schultz CS, Ames BN. Biotin deficiency inhibits heme synthesis and impairs mitochondria in human lung fibroblasts. Journal of Nutrition. 2007; 137: 25–30. https://doi.org/10.1093/jn/137.1.25 PMID: 17182796

72. Ponka P. Cell biology of heme. The American Journal of the Medical Sciences. 1999; 318: 241–256. https://doi.org/10.1097/00000441-199910000-00004 PMID: 10522552

73. Cadonic C, Sabbir MC, Albens BC. Mechanisms of Mitochondrial Dysfunction in Alzheimer's Disease. Molecular Neurobiology. 2016; 53(9): 6078–6090. https://doi.org/10.1007/s12035-015-9515-5 PMID: 26537901

74. Bubber P, Haroutunian V., Fisch G, Blass JP, Gibson GE: Mitochondrial abnormalities in Alzheimer brain: mechanistic implications. Annals of Neurology. 2005; 57: 695–703. https://doi.org/10.1002/ana.20474 PMID: 15852400

75. Amoroso N, La Rocca M, Bruno S, Maggipinto T, Monaco A, Bellotti R, et al. Multiplex Networks for Early Diagnosis of Alzheimer's Disease. Frontiers in Aging Neuroscience. 2018; 10: 365. https://doi.org/10.3389/fnagi.2018.00365 PMID: 30487745