Computational and Structural Biotechnology Journal 20 (2022) 3851–3863

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Gene expression data analysis using Hellinger correlation in weighted gene co-expression networks (WGCNA)

Tianjiao Zhang, Garry Wong *

Cancer Centre, Centre for Reproduction, Development and Aging, Department of Public Health and Medicinal Administration, Faculty of Health Sciences, University of Macau, Taipa 999078, Macau Special Administrative Region

ARTICLE INFO

Article history:
Received 13 March 2022
Received in revised form 9 July 2022
Accepted 9 July 2022
Available online 13 July 2022

Keywords:
WGCNA
Non-linear correlation
Alzheimer's disease
Hellinger correlation
GTEx
scRNA-seq

ABSTRACT

Weighted gene co-expression network analysis (WGCNA) is used to detect clusters with highly correlated genes. Measurements of correlation most typically rely on linear relationships. However, a linear relationship does not always model pairwise functional-related dependence between genes. In this paper, we first compared 6 different correlation methods in their ability to capture complex dependence between genes in three different tissues. Next, we compared their gene-pairwise coefficient results and corresponding WGCNA results. Finally, we applied a recently proposed correlation method, Hellinger correlation, as a more sensitive correlation measurement in WGCNA. To test this method, we constructed gene networks containing co-expression gene modules from RNA-seq data of human frontal cortex from Alzheimer's disease patients. To test the generality, we also used a microarray data set from human frontal cortex, single cell RNA-seq data from human prefrontal cortex, RNA-seq data from human temporal cortex, and GTEx data from heart. The Hellinger correlation method captures essentially similar results as other linear correlations in WGCNA, but provides additional new functional relationships as exemplified by uncovering a link between inflammation and mitochondria function. We validated the network constructed with the microarray and single cell sequencing data sets and a RNA-seq dataset of temporal cortex. We observed that this new correlation method enables the detection of non-linear biologically meaningful relationships among genes robustly and provides a complementary new approach to WGCNA. Thus, the application of Hellinger correlation to WGCNA provides a more flexible correlation approach to modelling networks in gene expression analysis that uncovers novel network relationships.

1. Introduction

Weighted gene co-expression network analysis (WGCNA) is a popular method for detecting and investigating highly correlated gene clusters based upon high-throughput sequencing expression datasets. These gene clusters may then act in coordination to facilitate specific biological processes [1–5]. The default correlation method of traditional WGCNA is Pearson's correlation, a standard measurement of linear correlation between two variables, measuring the extent of pairwise dependence between genes based on their expression level. However, linear correlation, or more generally monotone correlation, is not the only form of dependence (e.g. parabolic function, trigonometric function, etc.) [6]. Although dominated by several historical methods, such as Pearson's correlation, Spearman's correlation, and Biweight midcorrelation, new statistical methods quantifying dependence are constantly being explored. Mutual information (MI) quantifies complex dependence by measuring how much a random variable can be determined by knowing another [7]. Distance correlation is a new method that has the ability to detect both linear and non-linear dependence but has a higher priority for detecting linear dependence [8]. Hellinger correlation is a recently proposed estimator, which has better performance in characterizing general dependence. In contrast, Pearson's correlation and Biweight midcorrelation capture perfect linear dependence, and some rank-based measures such as Spearman's correlation only achieve their best performance in monotone dependence cases [9,10].

* Corresponding author at: Faculty of Health Sciences, E12-3005 – Avenida da Universidade, University of Macau, Taipa 999078, Macau Special Administrative Region

E-mail address: GarryGWong@um.edu.mo (G. Wong).

https://doi.org/10.1016/j.csbj.2022.07.018

Alzheimer's disease (AD) is the most common neurodegenerative disease that is pathologically characterized by β-amyloid (Aβ)-containing extracellular plaques and tau-containing intracellular neurofibrillary tangles, and clinically characterized by gradual cognitive decline and dementia [11–13]. Several hypotheses have been proposed to underlie AD including amyloidosis [14–17], tau pathology [18–20], mitochondria dysfunction [21–23], oxidative stress [23–25], and neuroinflammation [26–28], whereas the etiology remains elusive. Large-scale genome-wide association studies (GWAS) identified several risk genes of AD, such as mutations in *APP* (encoding amyloid precursor protein), *AGRN* (encoding agrin), *LILRB2* (encoding leukocyte immunoglobulin-like receptor B2), *GRN* (encoding granulin), and *APOE* (apolipoprotein E) [29]. However, no single gene has been identified to explain the mechanism of most AD cases. Therefore, network-based analysis, representing a more comprehensive approach to recapitulate and investigate the pathogenic mechanism at the molecular level, is popular in understanding pathogenesis and discovering therapeutic targets in human disease [30–33].

In the present study, we demonstrate the insensitivity and insufficiency of calculating linear dependence (Pearson's correlation, Spearman's correlation, and Biweight midcorrelation) of gene expression to predict corresponding functional dependence. Subsequently, we applied 6 WGCNA methods: Pearson WGCNA and Spearman WGCNA, Biweight WGCNA, Distance WGCNA, MI WGCNA and Hellinger WGCNA based on the corresponding correlation coefficients measuring gene dependence in RNA-seq human data sets including dorsolateral prefrontal cortex (DLPFC), temporal cortex (TC) and heart. The modules with highly correlated genes were compared. Finally, we focused on Hellinger WGCNA results of DLPFC, and investigated the modules selected by the significant correlation between module eigengene and pathological (neurofibrillary tangles and neuritic plaques) and clinical phenotypes (cognitive diagnosis). Our results demonstrate a notable advantage of utilizing a non-linear correlation measure to uncover novel gene co-expression networks in neurodegenerative disease. This approach may also be applicable to a wide range of gene expression data sets.

2. Materials and methods

2.1. 2.1Brief introduction of different correlation coefficients

Let X_1 and X_2 be two continuous random variables.

2.1.1. Pearson's correlation

In order to measure the dependence between X_1 and X_2, Pearson's correlation quantifies the similarity between covariance $cov(X_1, X_2)$ and the product of standard deviations $\sigma_{X1}\sigma_{X2}$, defined by:

$$\rho = \frac{cov(X_1, X_2)}{\sigma_{X1}\sigma_{X2}}$$

2.1.2. Spearman's rank correlation coefficient

The Spearman correlation is defined as the Pearson's correlation coefficient between the rank of two variables:

$$\rho = \frac{cov(Rank(X_1), Rank(X_2))}{\sigma_{Rank(X1)}\sigma_{Rank(X2)}}$$

Compared with Pearson's correlation, Spearman correlation is less sensitive to outliers. However, when variables are translated into rank, detailed information is lost. Therefore, Spearman's rank is less powerful than Pearson's correlation when the data is normally distributed.

2.1.3. Biweight midcorrelation

It is defined as:

$$bicor(x_1, x_2) = \frac{\sum_1^m (x_{1i} - med(x_1)) w_i^{(x_1)} (x_{2i} - med(x_2)) w_i^{(x_2)}}{\sqrt{\sum_{j=1}^m [(x_{1j} - med(x_1)) w_j^{(x_1)}]^2} \sqrt{\sum_{k=1}^m [(x_{2k} - med(x_2)) w_k^{(x_2)}]^2}}$$

Where.

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|)$$

$$u_i = \frac{x_i - med(x)}{9mad(x)}$$

$med(x)$ is the median of $\times$, and $mad(x)$ is the median absolute deviation of x.

As a median-based correlation method, biweight correlation is more robust to outliers than Pearson's correlation, and without losing excessive information compared with Spearman correlation.

2.1.4. Distance correlation

The representation of Distance correlation is analogous to methods mentioned above. But instead of sample moments (e.g., variance), distance correlation considers certain Euclidean distances between sample elements which are defined as $V_n^2(X) = \frac{1}{n^2}\sum_{k,l}^n A_{kl}^2$

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$$

$$a_{kl} = ||X_k - X_l||_p$$

And $V_n^2(X_1, X_2) = \frac{1}{n^2}\sum_{k,l}^n A_{1kl}^2 A_{2kl}^2$

Then, the distance correlation is defined as:

$$R_n^2(X_1, X_2) = \frac{V_n^2(X_1, X_2)}{\sqrt{V_n^2(X_1) V_n^2(X_2)}}$$

Rigorous proof shows that $R_n^2(X_1, X_2)$ ranges from 0 to 1, and $R_n^2(X_1, X_2) = 0$

if and only if X_1 and X_2 are independent while $R_n^2(X_1, X_2) = 1$ if and only if $X_1 = aX_2 + b$, which means perfect linear dependence (8). Therefore, distance correlation measures all types of possible relationships, including linear and non-linear dependence. But have higher priority for linear correlation.

2.1.5. Mutual information

$$MI(X_1, X_2) = \sum_{X_1, X_2} P_{XY}(x,\ y) log \frac{P_{XY}(x,\ y)}{P_X(x) P_Y(y)}$$
$$= \int\int_{\mathbb{R}^2} \phi\left(\frac{dF_{12}(x_1, x_2)}{dF_1(x_1) dF_2(x_2)}\right) dF_1(x_1) dF_2(x_2)$$

$\phi(t) = tlogt$

It is easy to notice that $MI(X_1, X_2)$ ranges from 0 to $+\infty$. However, previous proof indicates that there are cases in which MI goes to infinite without a strong dependence between X_1 and X_2 [34]. In practical applications, a normalization value from 0 to 1 for MI is usually used for comparison purposes. In this paper, we define MI as:

$$MI_{normalized}(X_1, X_2) = \frac{MI(X_1, X_2)}{\max(Entropy(X_1), Entropy(X_2))}$$

Where: $Entropy(X_1) = -\sum_{X_1} P(x_1) log P(x_1)$

MI reasonably and equally detects both linear and non-linear dependence. However, the power of MI is lower than other coefficients mentioned in this paper [35,36].

2.1.6. Hellinger correlation

In contrast, the Hellinger correlation (η) measures the dependence of X_1, X_2 by quantifying the similarity between their joint distribution F_{12} and marginal production F_1F_2. Given by:

$$\eta = \left(\frac{2}{\mathscr{B}^2}\right)\{\mathscr{B}^4 + (4 - 3\mathscr{B}^4)^{\frac{1}{2}} - 2\}^{\frac{1}{2}}$$

where:

$$\mathscr{B} = 1 - \frac{1}{2}\iint_{R^2}\left(\sqrt{\frac{dF_{12}(x_1, x_2)}{dF_1(x_1)dF_2(x_2)}} - 1\right)^2 dF_1(x_1)dF_2(x_2)$$

Detailed information and estimation procedures of η are shown in [9]. Hellinger correlation ranges from 0 to 1 (from completely independent to perfect dependence) and has comparable power compared to the above-mentioned methods. Compared with MI information, Hellinger coefficient has higher statistical power and achieves highest value if and only if there are perfect dependent cases. At the same time, unlike distance correlation, it does not prioritize linear correlation.

In this paper, we used WGCNA package [2] to estimate Pearson correlation (complexity: O(n)), Spearman correlation (complexity: O(n)), Biweight correlation (complexity: O(n)), and MI (complexity: $O(n^2)$); energy package [37,38] to estimate distance correlation (complexity: $O(n^2)$); HellCor package [9] to estimate Hellinger correlation (complexity: $O(n^2)$). Based on the properties of different coefficients, we named coefficients only measuring linear correlation as linear coefficients (Pearson, Spearman, and Biweight) while named coefficients measuring both linear and non-linear correlation as dependence-based coefficients. Since dependence-based coefficients identify dependence instead of linear correlation with direction, we built undirected WGCNA graphs. Therefore, we took the absolute value of linear coefficients during gene co-expression network construction.

2.2. Data collection and preprocessing

Table 1 shows the description of RNA datasets curated for the analysis.

RNA-seq DLPFC dataset. ROSMAP RNA-seq normalized data (syn3505720) and clinical file (syn3191087) were download from SYNAPSE [31]. We filtered out genes with FPKM less than 1 in more than 50 % of samples. A variance-stabilizing transformation from package DESeq2 [42] was applied. We clustered the whole samples on their Euclidean distance to test their similarity, and one sample was considered an outlier and removed from further analysis. Finally, the top 5000 genes with the highest standard deviation and the remaining 641 samples were selected for the rest of the analyses.

Microarray DLPFC dataset. Processed raw gene expression data and metadata were downloaded from GSE44772 [30]. The batch effect was removed by limma package [43]. Genes intersected with the 5000 genes selected from RNA-seq DLPFC dataset were chosen to validate the preservation of modules from RNA-seq DLPFC dataset WGCNA results.

RNA-seq Heart dataset. Heart RNA-seq normalized data was downloaded from GTEx database [41] and processed in the same manner as RNA-seq DLPFC dataset. The top 5000 genes with the highest standard deviation and 430 samples were selected for the rest of the analyses.

RNA-seq temporal cortex (TC) dataset. TC RNA-seq normalized data (syn8466815) and clinical file (syn8466814) were downloaded from SYNAPSE [40] and processed in the same manner as RNA-seq DLPFC dataset. Genes intersected with the 5000 genes selected from RNA-seq DLPFC dataset were chosen for further analyses.

Single-cell sequencing dataset. ROSMAP single-cell sequencing data (syn18485175) and metadata (syn3157322) were download from SYNAPSE [39]. Single-cell raw data was processed by the package Seurat [44]. Specifically, we kept genes detected in no less than 3 cells and kept all cells with at least 200 detected genes. Outlier cells in quality metrics, including unique gene counts, the ratio of mitochondrial relative to endogenous RNAs, total gene counts, were filtered out as they might represent dead (or unqualified) or doublets (or multiplets) cells. For the rest of cells that passed the quality control, we carried out log-normalization with a scale factor 10000. Louvain algorithm in Seurat was implemented to detect clusters. The cell type of each cluster was identified by the corresponding marker genes provided by a previous publication [39]. Considering the sparsity of single-cell sequencing data, we applied scWGCNA [45], a package aggregating k nearest neighboring cells (neighboring cells within a cell-type-specific cluster) as a newly constructed pseudo-cell. Finally, genes overlapped with the 5000 genes selected from RNA-seq DLPFC dataset were obtained to set up cell-type-specific gene expression matrices of pseudo-cells, which were included to validate the preservation of modules from RNA-seq DLPFC dataset WGCNA results.

2.3. WGCNA and identification of significant modules

Different coefficients were calculated to measure the pairwise dependence between genes. To comply with scale-free topology criterion and the recommendations of WGCNA use, we chose appropriate soft-thresholding powers to convert the gene expression matrices to adjacency matrices. Then topology overlap matrices (TOM) were calculated by adjacency matrices [46]. We then use hierarchical clustering and dynamic tree cut method to identify gene clusters [2,47]. All steps of WGCNA with different coefficients are the same except for the calculation to measure gene dependence.

2.4. Network validation

STRING database. We built Protein-Protein networks based on the annotation from the STRING database (version = 11.5) [48]. To increase the credibility of validation, high confidence (score threshold = 900) was referred to in determining the connection between two proteins.

Normalized Mutual Information (NMI) [49] and adjusted rand index (ARI) [50]. NMI and ARI are two common statistics measuring the similarity between two data clustering results. NMI ranges from 0 to 1 and ARI ranges from −1 to 1, which corresponds to completely different to exactly the same.

Module preservation. We mainly used z-summary [51], a network preservation statistic aggregating multiple preservation statistics (3 density-based statistics and 3 connectivity-based statistics), to quantify the conservation of the co-expression network in another dataset.

2.5. Modules-phenotype association analysis and Bayesian network construction

Module eigengene was defined as the first principal component of a module gene expression matrix [2]. We identified the module of interest by the correlation strength between module eigengene and phenotypes, including pathologic stages (braak stage, CERAD

score) and clinical behaviors (Clinical diagnosis of cognitive status (dcfdx_lv), Final consensus cognitive diagnosis (cogdx)).

Markov chain Monte Carlo (MCMC) methods for structure learning and sampling of Bayesian networks were demonstrated to have better performance [52]. We used order MCMC algorithm in BiDAG, and 20 million iterations were applied [53].

2.6. Gene set enrichment analysis and cell-type-specific expression analysis

We implemented gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis by g: Profiler [54], and cell-type-specific expression analysis by CSEA tool [55].

2.7. Hub genes identification

We used cytoHubba [56] in Cytoscape [57] to identify hub genes in the protein–protein network. For networks with linked genes larger than 300, we identified the top 200 hub genes and showed related subnetworks, otherwise we displayed the entire PPI network.

3. 3.Results

3.1. 3.1Comparison of different correlation coefficients

3.1.1. Coefficients measuring both linear and nonlinear association fit complex dependence

We hypothesized that there are complex relationships beyond linear correlation between genes in their expression in biological tissues. To test this hypothesis, we plotted several typical pairwise relationships of genes from the DLPFC RNA-seq dataset (Fig. 1A,C,E; Fig. 2A,C,E,G). We found that all correlation metrics are competent in accepting significant linear dependence and rejecting purely independent gene pairs, except that Pearson correlation is relatively more sensitive to outliers (Fig. 1B,D,F). They perform differently in detecting more complex dependence relationships (Fig. 2 B,D,F,H). In these cases, dependence-based coefficients have a higher correlation value. We subsampled the data and show the receiver operating characteristic (ROC) curves [58] to evaluate the performance and robustness of different correlation coefficients. The results indicate that only dependence-based coefficients can consistently detect all kinds of complex dependence (Fig. 2B,D,F,H). To further verify the authenticity of these complex dependencies, we provide 6 coefficients of these gene pairs in 53 GTEx tissues (Supplementary Table 1)[41]. We found that the dependence of *HBB* and *RCN2* (independent); *CRABP2* and *HS3ST2* (independent with outliers) are not significant in most of the tissues by most coefficients. The dependence of *CYTB* and *ND1* (linear dependent); *RPS28* and *RPL36* (threshold); *RAC2* and *CYBB* (complex dependent) are significant in most of the tissues by most coefficients. The dependence of *GSTM1* and *GSTM2* (dependent with outliers) is significant in most of the tissues by Hellinger correlation. However, in liver tissue, where *GSTM1* and *GSTM2* are highly expressed [59], the dependence is significant in all coefficients. The dependence of *MYC* and *COX2* (power function) is significant in most brain tissues and other tissues, especially in the kidney. Interestingly, *MYC* and *COX2* also show power function in those tissues.

3.1.2. Comparison of different coefficients

In Fig. 3, we show the edge-agreement results of the top 10 percent of gene pairs with the highest correlation values of each coefficient in the DLPFC dataset. As we expected from the statistical properties of each method, the results of dependence-based coefficients are more different from linear coefficients. Outlier insensitive methods have more overlap with each other (Spearman and Biweight). For dependence-based coefficients, MI and Hellinger are more likely to agree as they equally measure the linear and non-linear processes based on dependence. In contrast, distance correlation is more prone to give higher value to linear processes, so the results agree more with linear coefficients. Fig. 3B shows the number of gene pairs in Fig. 3A that can be found in the protein–protein interaction database (STRING database). It shows that many gene pairs with higher dependence-based coefficients have the potential to be functionally related. The corresponding figures of TC and heart are shown in Supplementary Fig. 1,2. These results highlight the necessity and advantage of applying a dependence-based coefficient to detect more comprehensive gene networks that linear coefficients might miss or ignore.

3.1.3. Comparison WGCNA results using different coefficients

To investigate the consistency and novelty of applying dependence-based coefficients in WGCNA, we compared their module results in cluster agreement, network preservation, and Module-wise comparison. The NMI and ARI suggest the cluster agreement level of WGCNA cluster results with different coefficients in DLPFC (Table 2), while the results of TC and heart are shown in Supplementary Tables 2 and 3, respectively. The results of dependence-based coefficients are more similar, which also happens in linear coefficient results. Network preservation results show that most modules are preserved in the network constructed by different coefficients (Supplementary Fig. 3,4,5). However, there are still few modules identified by dependent-based coefficients WGCNA that are not consistently preserved in networks built by other methods. For example, the turquoise module detected by Hellinger WGCNA appears to be a collection of several distant gene clusters in other WGCNA methods, and its density and connectivity are less preserved (Supplementary Fig. 3A, Supplementary Fig. 6). For DLPFC RNA-seq dataset, we also quantified module preservation in an independent microarray DLPFC dataset from another study (Fig. 4). We found significant preservation evidence for 9 in Pearson, 10 in Spearman, 10 in Biweight, 7 in Distance, 8 in MI, and 6 in Hellinger, indicating the modules identified by WGCNA with different coefficients are consistent in another DLPFC study. These results show the feasibility of applying dependence-based coefficients in WGCNA as a complementary approach to the standard approach (e.g. Pearson's correlation).

3.2. Investigation of Hellinger WGCNA modules

Here, we show the results of Hellinger WGCNA in DLPFC RNA-seq dataset, and investigate the novel network constructed by its dependence-based coefficients.

3.2.1. Modules related to AD pathologic and clinical phenotypes

We identified modules associated with pathologic and clinical features from two aspects: (1) correlation between module eigengenes and clinical phenotypes; (2) module-phenotype network by Bayesian network structure learning. We found the module blue, turquoise, green and red are highly related to braak stage (severity of neurofibrillary tangle), CERAD score (severity of neuritic plaques), and clinical cognitive diagnosis (Fig. 5A). Bayesian network indicated the relationship between these 4 selected modules and clinical phenotypes (Fig. 5B). GO functional and KEGG pathway enrichment analyses indicated that blue module related to oxidative phosphorylation and mitochondrial function (mainly genome-coding genes); turquoise module associated with mitochondrial function (mainly mitochondria-coding genes) and immunity; red and green modules are significant with synaptic vesicle cycle, neurotransmitter, and other neuron functions.

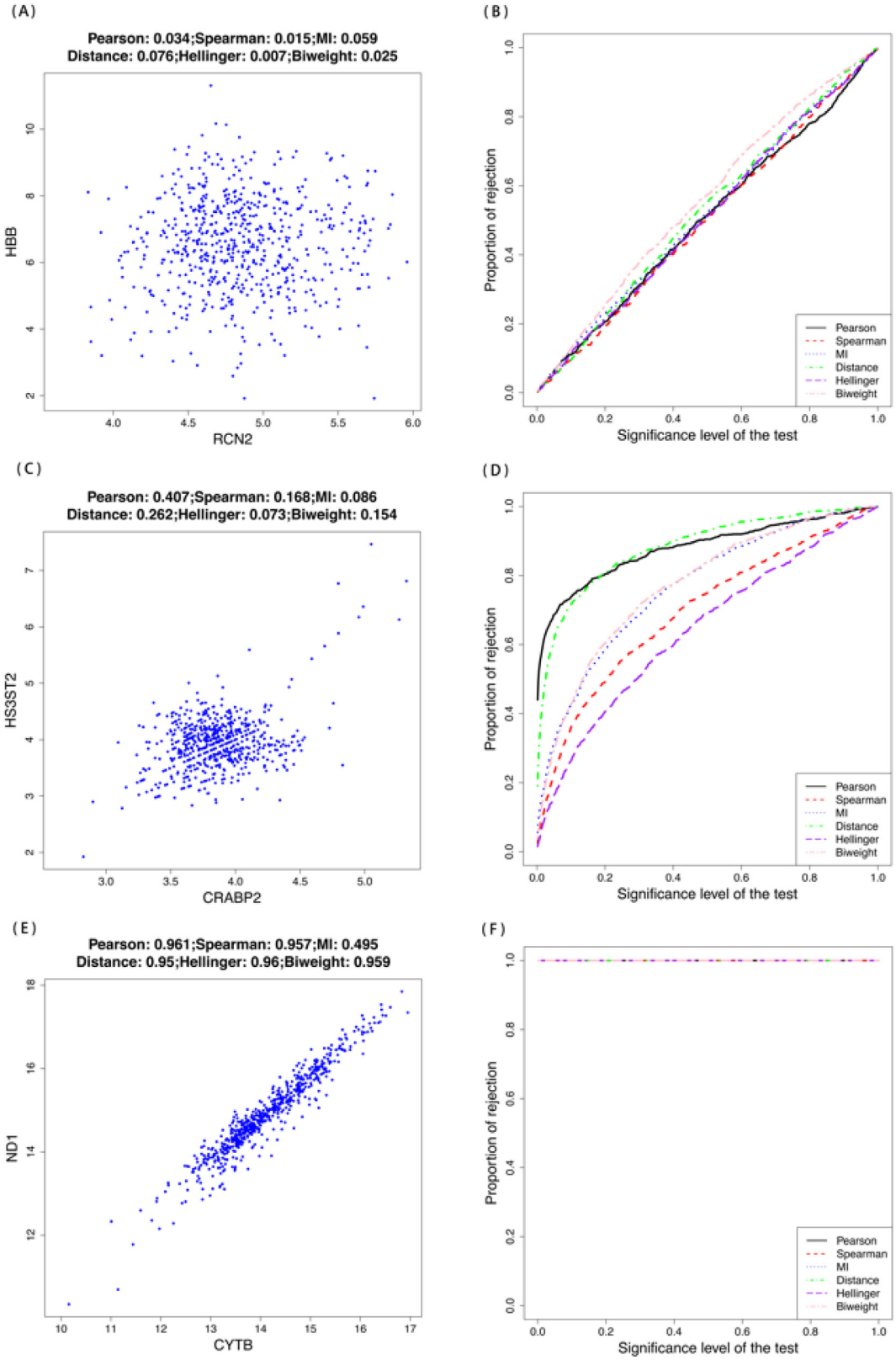

Fig. 1. Gene-wise relationships observed in DLPFC data and corresponding ROC curves. The scatter plots with axes represent related genes' expression levels (variance-stabilizing transformed FPKM). To construct the ROC curves, we calculated the gene pair coefficients of 50 samples sampling from DLPFC data sets. The sampling process was repeated 1000 times, and the proportion of rejecting the null hypotheses at different significance level are shown in the ROC. The null hypothesis indicates independence. The scatter plots and ROC curves of independent (A, B), independent with outliers (C, D), and linear dependent (E,F) gene-pairs are shown.

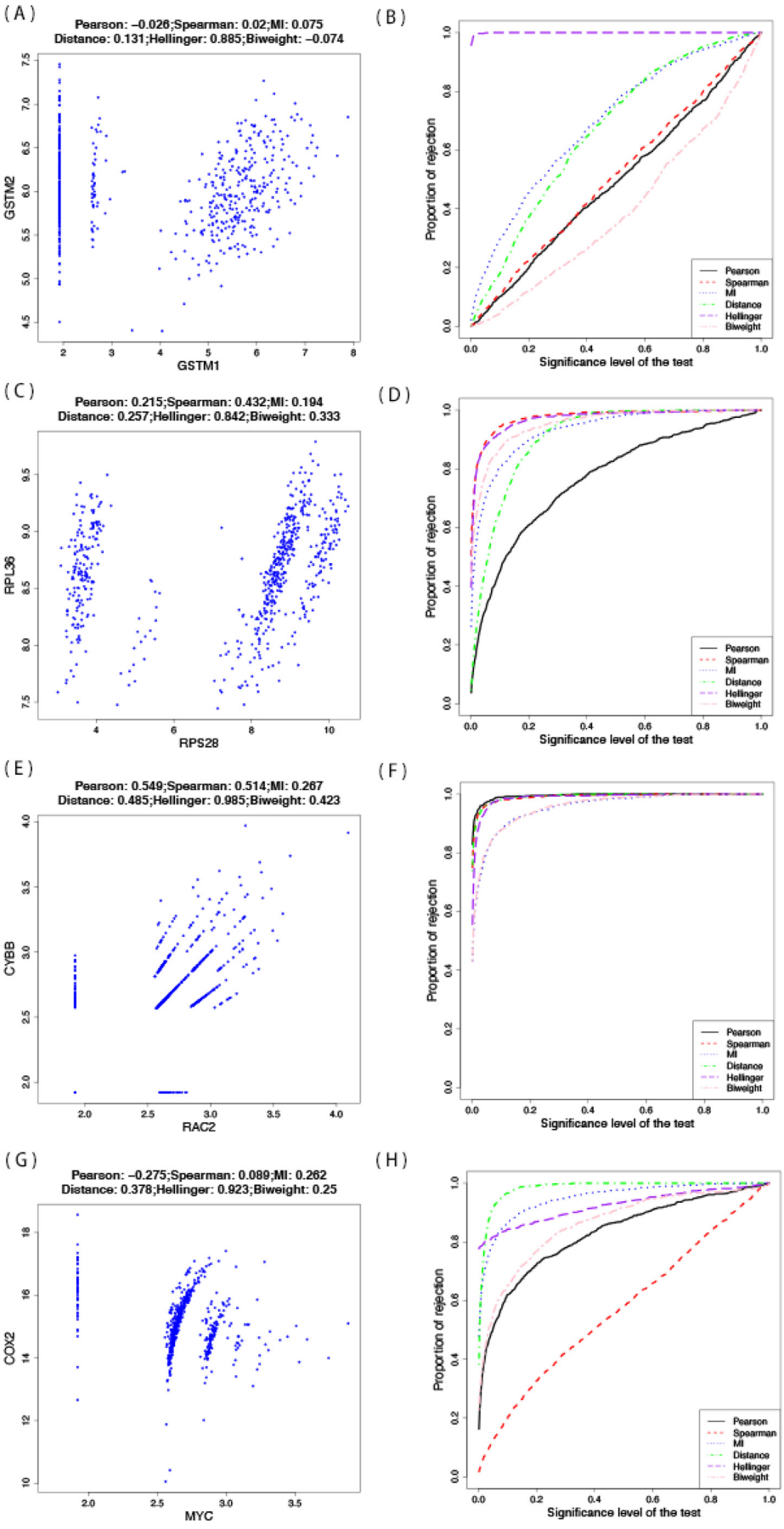
(A)
Pearson: −0.026;Spearman: 0.02;MI: 0.075
Distance: 0.131;Hellinger: 0.885;Biweight: −0.074
GSTM2
GSTM1
(B)
Proportion of rejection
Significance level of the test
Pearson
Spearman
MI
Distance
Hellinger
Biweight
(C)
Pearson: 0.215;Spearman: 0.432;MI: 0.194
Distance: 0.257;Hellinger: 0.842;Biweight: 0.333
RPL36
RPS28
(D)
Proportion of rejection
Significance level of the test
Pearson
Spearman
MI
Distance
Hellinger
Biweight
(E)
Pearson: 0.549;Spearman: 0.514;MI: 0.267
Distance: 0.485;Hellinger: 0.985;Biweight: 0.423
CYBB
RAC2
(F)
Proportion of rejection
Significance level of the test
Pearson
Spearman
MI
Distance
Hellinger
Biweight
(G)
Pearson: −0.275;Spearman: 0.089;MI: 0.262
Distance: 0.378;Hellinger: 0.923;Biweight: 0.25
COX2
MYC
(H)
Proportion of rejection
Significance level of the test
Pearson
Spearman
MI
Distance
Hellinger
Biweight

Table 1
Description of RNA datasets curated for the analysis. References for the datasets are shown in parentheses after the dataset. Abbreviations: DLPFC, dorsolateral prefrontal cortex; AD, Alzheimer's disease.

Dataset	Tissue	Sample Size
RNA-seq dataset (31)	DLPFC	Control: 235 AD: 406
Microarray dataset (30)	DLPFC	Control: 101 AD: 129
Single-cell sequencing dataset (39)	Prefrontal cortex (Brodmann area 10)	Control: 24 AD: 24
RNA-seq dataset (40)	Temporal cortex	AD: 80 Control: 71 Path age: 30 Progressive supranuclear palsy (PSP): 81
RNA-seq dataset (41)	Heart	430

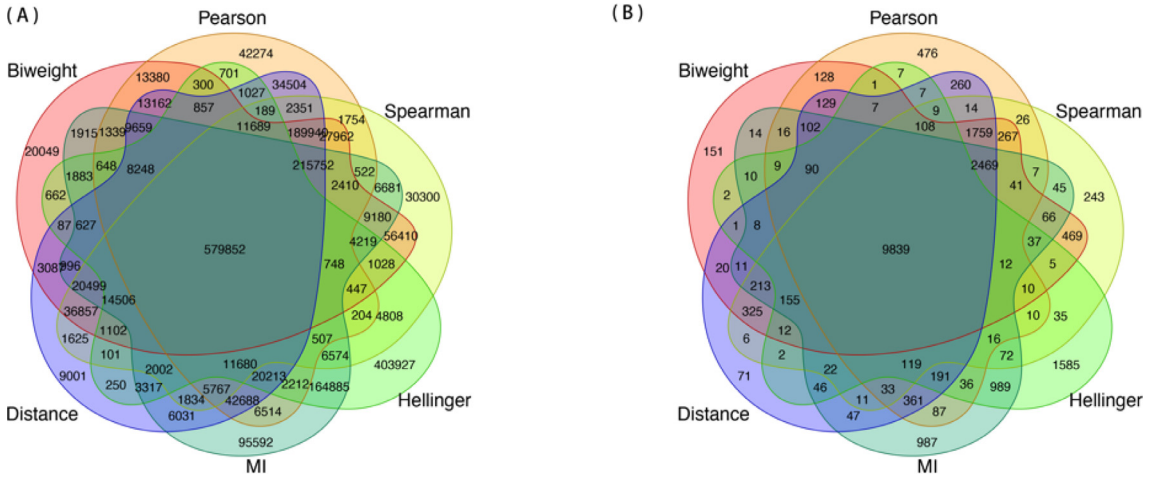

Fig. 3. Venn plot of gene pairs with top 10 % highest correlation values of each coefficient in DLPFC RNA-seq dataset. (A) Specifically, the gene-pairs with coefficients:$Biweight \geq 0.62, Pearson \geq 0.63, Spearman \geq 0.62, Hellinger \geq 0.74, MI \geq 0.16; Distance \geq 0.61$ were chosen and shown in Venn plots. (B) The number of gene pairs in panel (A) that can be found in the protein–protein interaction database (STRING database).

Table 2
NMI and ARI between the clustering results of WGCNA constructed by different coefficients in DLPFC. NMI and ARI measure the similarity between two data clustering results. NMI ranges from 0 to 1 and ARI ranges from −1 to 1, which corresponds to completely different to exactly same.

	Pearson	Spearman	Biweight	Distance	MI	Hellinger
ARI						
Pearson	1	0.38	0.54	0.63	0.37	0.21
Spearman	0.38	1	0.55	0.44	0.29	0.15
Biweight	0.54	0.55	1	0.58	0.38	0.19
Distance	0.63	0.44	0.58	1	0.39	0.23
MI	0.37	0.29	0.38	0.39	1	0.30
Hellinger	0.21	0.15	0.19	0.23	0.30	1
NMI						
Pearson	1	0.56	0.61	0.67	0.43	0.40
Spearman	0.56	1	0.67	0.59	0.40	0.38
Biweight	0.61	0.67	1	0.62	0.44	0.39
Distance	0.67	0.59	0.62	1	0.45	0.42
MI	0.43	0.40	0.44	0.45	1	0.47
Hellinger	0.40	0.38	0.39	0.42	0.47	1

Fig. 2. Gene-wise relationships observed in DLPFC data and corresponding ROC curves. The scatter plots with axes represent related genes' expression levels (variance-stabilizing transformed FPKM). To construct the ROC curves, we calculated the gene pair coefficients of 50 samples sampling from DLPFC data sets. The sampling process was repeated 1000 times, and the proportion of rejecting the null hypotheses at different significance level was shown in the ROC. The null hypothesis indicates independence. The scatter plots and ROC curves of dependent with outliers (A, B), threshold (C, D), complex dependent (E, F), and power function (G, H) are shown.

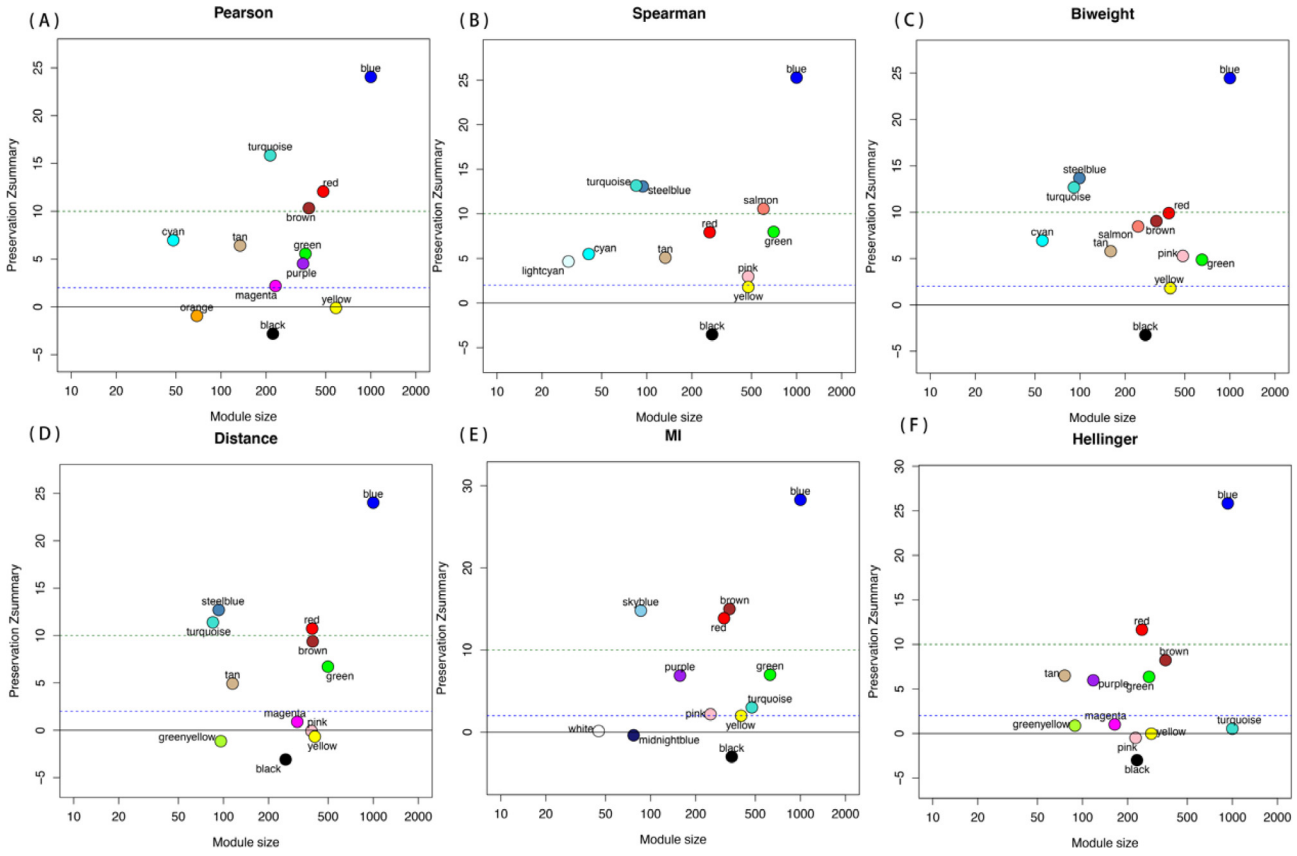

Fig. 4. DLPFC RNA-seq Module preservation assessed in the microarray dataset. The correlation methods used are indicated (A-F). The green dashed line (Z-summary = 10) marks the "strongly preserved" threshold and the blue dashed line (Z-summary = 2) marks the "moderately preserved" threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.2. Validation and investigation of module turquoise

As module turquoise is the main difference between Hellinger WGCNA and WGCNA with other coefficients, we further validated and investigated module turquoise's structural and functional construction. We performed Cell-type-specific expression analysis and found genes in the turquoise module are significantly enriched in oligodendrocytes, astrocytes, and immune cells. We investigated module preservation of module turquoise in an independent single-cell prefrontal cortex (Brodmann area 10) dataset, and Fig. 6 shows turquoise is highly preserved in oligodendrocytes and astrocytes. We ignored the module preservation test in immune cells because we did not detect enough immune cells in this single-cell dataset (Fig. 6).

The top 200 hub genes of module turquoise are summarized in Fig. 7.

Fig. 7 shows that Hellinger Correlation WGCNA constructs a network between mitochondria-coding genes and inflammation. It is worth noting that a similar network was also constructed in TC RNA-seq dataset (blue module), indicating this network might consistently exist in brain tissue (Supplementary Fig. 7). Interestingly, this module in TC is also not preserved in the network constructed by linear coefficients (Supplementary Fig. 4A). The interaction between inflammation and mitochondrial were proposed by many previous studies. Specifically, mitochondria plays an important role in immune pathways by releasing components, including mtRNA, mtROS, and related proteins, while many inflammatory processes affect mitochondria dynamics and functions. [60–63]. In addition, the dysfunction within the dependence between mitochondria and inflammation were hypothesized and demonstrated in neurodegenerative diseases [63–65]. Recently, new therapeutic approaches, targeting the crosstalk between mitochondria and inflammation, such as COX, PPAR-γ, NO synthases, were proposed for neurodegenerative diseases [66,67]. However, this mitochondria and inflammation relationship cannot be recapitulated by linear coefficients, as the dependence within some gene pairs beyond linear correlation (e.g. Fig. 2G,H in the DPFC RNA-seq dataset).

4. Discussion

This study illustrates the ability of Pearson's correlation and other linear coefficients to predict pairwise biological dependence between genes based on gene expression level. We identified a thousand pairs of genes, of which the STRING database annotated biological dependencies. However, the statistical dependence of these pairwise genes is easily detected by Hellinger correlation rather than linear coefficients, indicating the importance of including non-linear correlation statistics in gene-wise dependence tests.

Whole-transcriptome analyses, such as RNA-seq, microarray, and single-cell sequencing are conventional methods to explore the mechanisms of complex diseases [68]. Given the cost of sequencing experiments and inaccessibility of biopsy tissues, especially with large sample sizes of human tissues, more comprehensive analysis is needed to find disease mechanisms and therapeutic targets. We proposed integrating Hellinger Correlation as an alternative option of Pearson's correlation in WGCNA analysis. In Hellinger WGCNA, the turquoise module, including a connection between mitochondria-coding genes and inflammation, was uniquely constructed. We verified the preservation of Hellinger

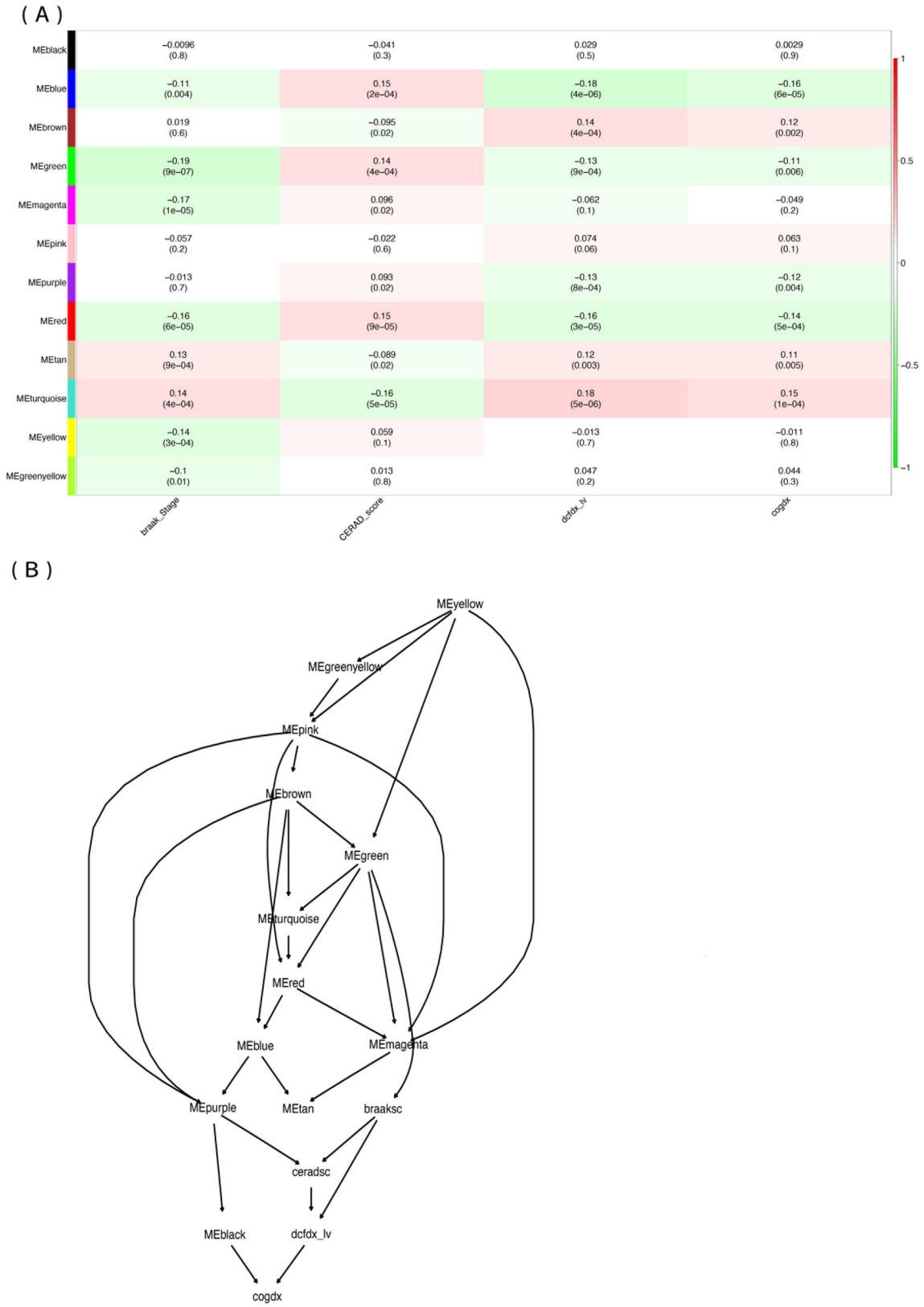

Fig. 5. Modules-phenotype association analysis. (A) Correlation between the module eigengene and phenotypes. Numbers indicate correlation coefficients and *p*-values. (B) Bayesian network including modules and phenotypes.

turquoise in an independent microarray study. The lack of preservation was later confirmed due to some turquoise hub genes (mainly mitochondrial coding genes) missing in the microarray dataset. Therefore, we further performed cell-type-specific expression analysis and found Hellinger turquoise module was enriched and preserved in astrocytes and oligodendrocytes. In fact, existing

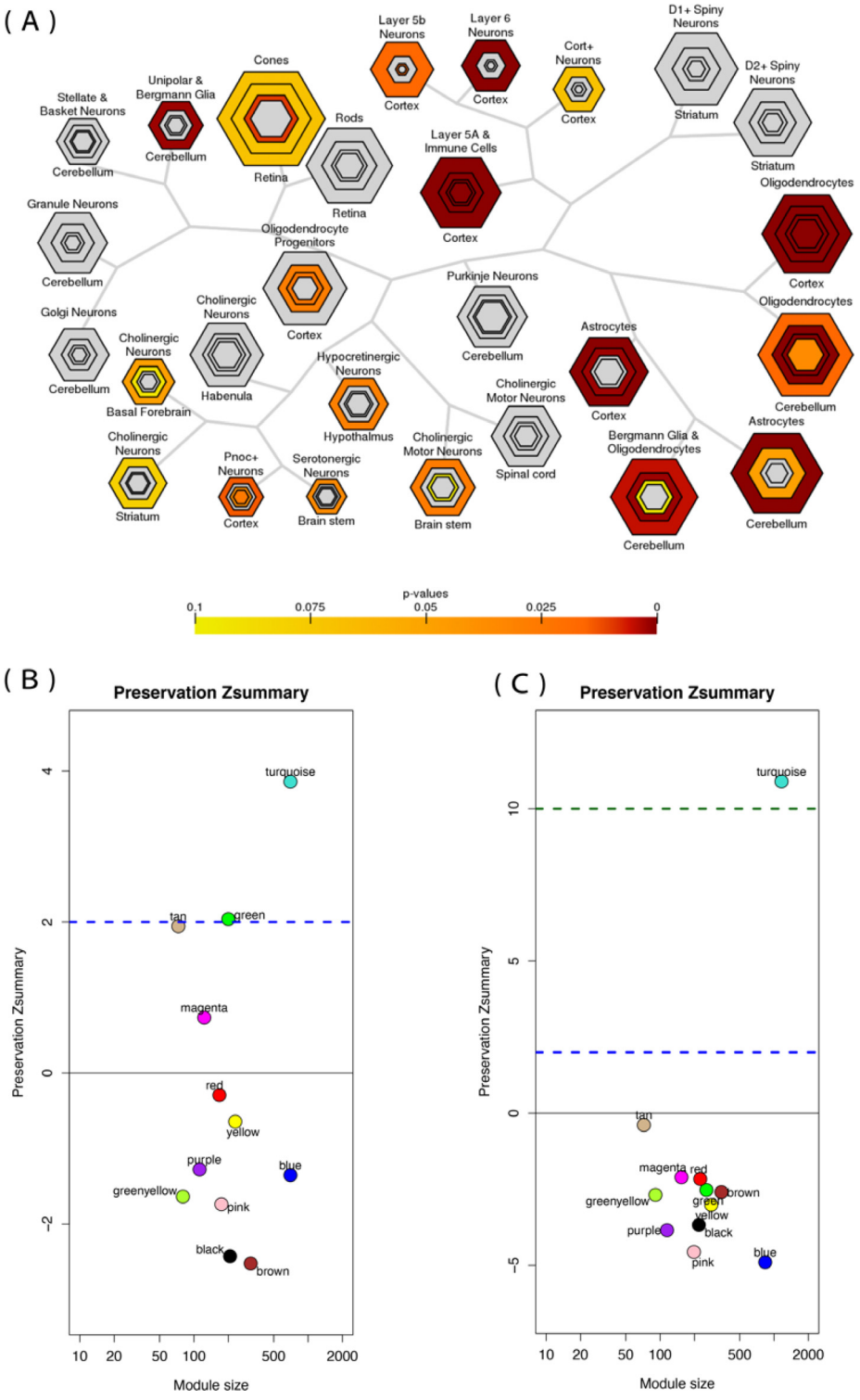

Fig. 6. Hellinger turquoise module preservation assessed in the single-cell sequencing data. (A) Results of cell-type-specific expression analysis of Hellinger turquoise preservation genes. Colors from yellow to red represent the enrichment significance, while the size of hexagon represents the cell-specific level (the smaller the hexagon, the more stringently cell type specific it is) (B) Hellinger modules preservation assessed in the astrocytes. The green dashed line (Z-summary = 10) marks the "strongly preserved" threshold and the blue dashed line (Z-summary = 2) marks the "moderately preserved" threshold. (C) Hellinger modules preservation assessed in the oligodendrocytes. The green dashed line (Z-summary = 10) marks the "strongly preserved" threshold and the blue dashed line (Z-summary = 2) marks the "moderately preserved" threshold. Z-summary quantifies the conservation of the co-expression network in another dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

studies suggest the cellular metabolism of the interaction between mitochondria function and inflammation pathway [69], while the related hypothesis and experimental demonstration was extended to neurodegenerative diseases [63–65]. Furthermore, we found *MYC* gene as an important hub gene linking mitochondrial coding genes and inflammation-related genes. As a proto-oncogene in

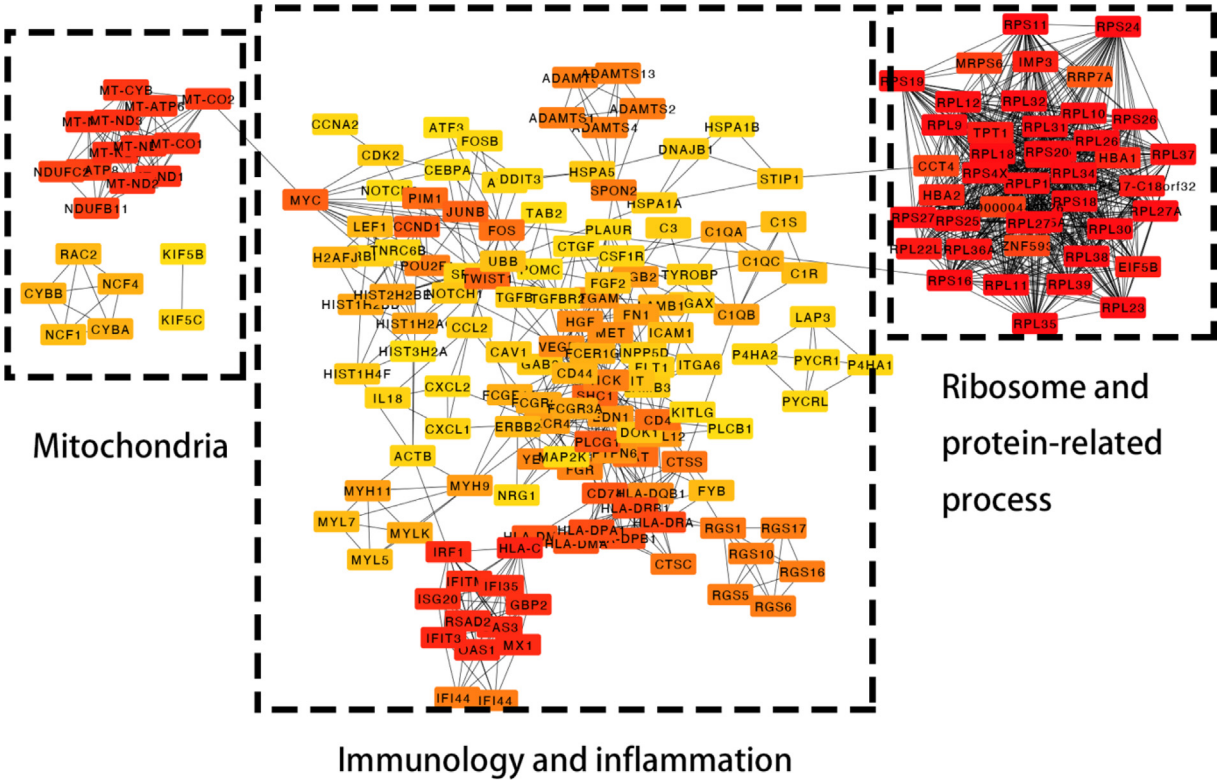

Fig. 7. Subnetwork of Hellinger turquoise module's 200 hub genes in DLPFC. The groups' within the dotted boxes share the functional annotation based on GO and KEGG enrichment analyses. Colors from light yellow to red mark the importance of the node in the turquoise module network measured by maximal clique centrality. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cancers, *MYC* controls the cell cycle, apoptosis, and metabolism [70]. In peripheral nervous system, axotomy elevates *MYC*, which promotes axon repair [71,72]. In the central nervous system, however, axotomy decreases *MYC*, and reduces axon repair [71]. This might be because maintaining and strengthening synaptic structure rather than cell regeneration is more necessary for the central nervous system during the evolution of mammals. Recently, many studies have reported the potential role of *MYC* in causing or accelerating AD through abnormal cell cycle re-entry [73], apoptosis induction [74], abnormal DNA synthesis [75], and other metabolism processes [76–78]. In addition, strong *MYC* expression was detected in AD astrocytes, where we demonstrated the preservation of the Hellinger turquoise module [79]. Therefore, the detailed biological function of the Hellinger turquoise module might provide an exciting opportunity to advance our knowledge of the AD mechanism. In the Bayesian network, 3 other Hellinger modules: blue, red, and green, were related to the AD progression from incipience to terminal. Correlation significance with phenotypes and enrichment analysis results substantiated their relationship with AD. However, since the minor differences between these Hellinger modules and the corresponding modules constructed by other coefficients, we did not discuss them in detail, but the significance of these modules is still worth noting.

Although the default function of WGCNA package is Pearson's correlation, other methods of calculating correlations can also be applied flexibly. The R [66] source code of combining Hellinger correlation to WGCNA is provided in the Supplementary text. The application of non-linear dependence rather than linear dependence in weighted gene co-expression network analysis has many advantages at the biological level. Non-linear dependence is better at modeling dosing-related saturation or threshold interaction in protein-binding processes and signaling pathways [67]. In addition, the existence of outliers or missing values in high-throughput sequencing data, especially for low expressed genes, affects the linear relationship between gene pairs. Therefore, non-linear correlation network is more appropriate to detect the complex biological gene-wise interaction and technical variation of high-throughput sequencing data, which provides a comprehensive basis for detecting hub genes as therapeutic targets or biomarkers. In this work, we only applied this non-linear dependence related WGCNA in the AD dataset, and more attempts in other datasets or diseases whose mechanisms are more clearly defined could better demonstrate the importance of this approach. In addition, we found Hellinger correlation is more sensitive than Pearson correlation and the mean connectivity of Hellinger network higher. Therefore, Hellinger network is easier to detect with less clusters than Pearson network with the premise of scale free topology.

While we were preparing this manuscript, Hou et al [80] published an approach that applies distance correlation in WGCNA and demonstrated the ability of distance correlation in non-linear dependence detection, robust to outliers in microarray (macrophage and liver) and RNA-seq (cervical cancer and pancreatic cancer). The advantages of Hellinger correlation we recommend in this paper over distance correlation is that Hellinger correlation provides a fairer measure of linear and non-linear relationship while distance correlation takes its maximum value only in the case of the perfect linear relationship [8,9]. As Pearson correlation, distance correlation is only defined if variables have finite second moment, which is not required by Hellinger correlation [8,9]. However, both Hellinger correlation and distance correlation are dedicated to detecting dependencies and cannot get positive or negative correlation information compared to Pearson correlation. Optimization by using different correlation measurements

depending upon specific datasets or problems would be worth pursuing in the future.

In summary, we applied Hellinger correlation in quantifying the dependence between gene pairs in WGCNA analysis to Alzheimer's disease data sets. The verification tests and downstream analyses results provide new insight into applying non-linear correlation statistics to construct gene networks. Such an application should complement current methods to obtain a more comprehensive understanding of biological processes underlying complex diseases.

CRediT authorship contribution statement

Tianjiao Zhang: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft. **Garry Wong:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank members of the Wong lab for discussion and critical reading and suggestions for the manuscript. Haibin Zhu is gratefully acknowledged for help and advice in running MATLAB source code. This work was supported in part by grant MYRG2020-00213-FHS from the Faculty of Health Sciences, University of Macau.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.018.

References

[1] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. In: Statistical applications in genetics and molecular biology. p. 4.
[2] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9:1–13.
[3] Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, et al. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci 2006;103:17402–7.
[4] Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 2008;4:e1000117.
[5] Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol 2007;1:1–17.
[6] Croxton, F.E. and Cowden, D.J. (1939) Applied general statistics.
[7] Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 2002;18(Suppl 2):S231–40.
[8] Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. The annals of statistics 2007;35:2769–94.
[9] Geenens G, Lafaye de Micheaux P. The Hellinger correlation. J Am Stat Assoc 2020:1–15.
[10] Wilcox RR. Introduction to robust estimation and hypothesis testing. Academic press; 2011.
[11] Knopman DS, Amieva H, Petersen RC, Chételat G, Holtzman DM, Hyman BT, et al. Alzheimer disease. Nature Reviews Disease Primers 2021;7:1–21.
[12] Goedert M, Spillantini MG. A century of Alzheimer's disease. *Science* 2006;314:777–81.
[13] Khachaturian ZS. Diagnosis of Alzheimer's disease. Arch Neurol 1985;42:1097–105.
[14] Merlini G, Bellotti V. Molecular mechanisms of amyloidosis. N Engl J Med 2003;349:583–96.
[15] Ghiso J, Frangione B. Amyloidosis and Alzheimer's disease. Adv Drug Deliv Rev 2002;54:1539–51.
[16] Price DL, Sisodia SS, Gandy SE. Amyloid beta amyloidosis in Alzheimer's disease. Curr Opin Neurol 1995;8:268–74.
[17] Hardy J, Selkoe DJ. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics 2002;science, 297:353–6.
[18] Iqbal, K., Alonso, A.d.C., Chen, S., Chohan, M.O., El-Akkad, E., Gong, C.-X., Khatoon, S., Li, B., Liu, F. and Rahman, A. (2005) Tau pathology in Alzheimer disease and other tauopathies. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1739**, 198-210.
[19] Pooler AM, Polydoro M, Wegmann S, Nicholls SB, Spires-Jones TL, Hyman BT. Propagation of tau pathology in Alzheimer's disease: identification of novel therapeutic targets. Alzheimer's research & therapy 2013;5:1–8.
[20] Attems J, Thal DR, Jellinger KA. The relationship between subcortical tau pathology and Alzheimer's disease. Biochem Soc Trans 2012;40:711–5.
[21] Morais VA, De Strooper B. Mitochondria dysfunction and neurodegenerative disorders: cause or consequence. J Alzheimers Dis 2010;20:S255–63.
[22] Cardoso SM, Santana I, Swerdlow RH, Oliveira CR. Mitochondria dysfunction of Alzheimer's disease cybrids enhances Aβ toxicity. J Neurochem 2004;89:1417–26.
[23] Wang, X., Wang, W., Li, L., Perry, G., Lee, H.-g. and Zhu, X. (2014) Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1842**, 1240-1247.
[24] Smith, M.A., Rottkamp, C.A., Nunomura, A., Raina, A.K. and Perry, G. (2000) Oxidative stress in Alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, **1502**, 139-144.
[25] Chen Z, Zhong C. Oxidative stress in Alzheimer's disease. Neuroscience bulletin 2014;30:271–81.
[26] Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, et al. Neuroinflammation in Alzheimer's disease. The Lancet Neurology 2015;14:388–405.
[27] Calsolaro V, Edison P. Neuroinflammation in Alzheimer's disease: current evidence and future directions. Alzheimer's & dementia 2016;12:719–32.
[28] Bronzuoli MR, Iacomino A, Steardo L, Scuderi C. Targeting neuroinflammation in Alzheimer's disease. Journal of inflammation research 2016;9:199.
[29] Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat Genet 2021;53:1276–82.
[30] Zhang B, Gaiteri C, Bodea L-G, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 2013;153:707–20.
[31] Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci 2018;21:811–9.
[32] Liang J-W, Fang Z-Y, Huang Y, Liuyang Z-Y, Zhang X-L, Wang J-L, et al. Application of weighted gene co-expression network analysis to explore the key genes in Alzheimer's disease. J Alzheimers Dis 2018;65:1353–64.
[33] Wang M, Li A, Sekiya M, Beckmann ND, Quan X, Schrode N, et al. Transformative network modeling of multi-omics data reveals detailed circuits, key regulators, and potential therapeutics for Alzheimer's disease. Neuron 2021;109(257–272):e214.
[34] Micheas AC, Zografos K. Measuring stochastic dependence using ϕ-divergence. Journal of Multivariate Analysis 2006;97:765–84.
[35] Thomas M, Joy AT. Elements of information theory. Wiley-Interscience; 2006.
[36] Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinf 2012;13:328.
[37] Székely GJ, Rizzo ML. The energy of data. Annu Rev Stat Appl 2017;4:447–79.
[38] Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. Journal of statistical planning and inference 2013;143:1249–72.
[39] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 2019;570:332–7.
[40] Allen M, Burgess JD, Ballard T, Serie D, Wang X, Younkin CS, et al. Gene expression, methylation and neuropathology correlations at progressive supranuclear palsy risk loci. Acta Neuropathol 2016;132:197–211.
[41] (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.
[42] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:1–21.
[43] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–e.
[44] Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 2015;33:495–502.
[45] Feregrino C, Tschopp P. Assessing evolutionary and developmental transcriptome dynamics in homologous cell types. Dev Dyn 2021.
[46] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *science*, **297**, 1551-1555.
[47] Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 2008;24:719–20.
[48] Mering CV, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 2003;31:258–61.
[49] Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans Neural Networks 2009;20:189–201.

[50] Santos JM, Embrechts M. International conference on artificial neural networks. Springer 2009:175–84.
[51] Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol 2011;7:e1001057.
[52] Tasaki S, Sauerwine B, Hoff B, Toyoshiba H, Gaiteri C, Chaibub Neto E. Bayesian network reconstruction using systems genetics data: comparison of MCMC methods. Genetics 2015;199:973–89.
[53] Suter, P., Kuipers, J., Moffa, G. and Beerenwinkel, N. (2021) Bayesian structure learning and sampling of Bayesian networks with the R package BiDAG. *arXiv preprint arXiv:2105.00488*.
[54] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8.
[55] Dougherty JD, Schmidt EF, Nakajima M, Heintz N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. Nucleic Acids Res 2010;38:4218–30.
[56] Chin C-H, Chen S-H, Wu H-H, Ho C-W, Ko M-T, Lin C-Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014;8:1–7.
[57] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
[58] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[59] Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 2009;10:1–8.
[60] West AP, Shadel GS. Mitochondrial DNA in innate immune responses and inflammatory pathology. Nat Rev Immunol 2017;17:363–75.
[61] Murphy, M.P. (2018). Nature Publishing Group.
[62] Zhong F, Liang S, Zhong Z. Emerging role of mitochondrial DNA as a major driver of inflammation and disease progression. Trends Immunol 2019;40:1120–33.
[63] Missiroli S, Genovese I, Perrone M, Vezzani B, Vitto VA, Giorgi C. The role of mitochondria in inflammation: from cancer to neurodegenerative disorders. Journal of clinical medicine 2020;9:740.
[64] Jassim AH, Inman DM, Mitchell CH. Crosstalk between dysfunctional mitochondria and inflammation in glaucomatous neurodegeneration. Front Pharmacol 2021;12.
[65] Beal MF. Mitochondria, oxidative damage, and inflammation in Parkinson's disease. ANNALS-NEW YORK ACADEMY OF SCIENCES 2003;991:120–31.
[66] Ajmone-Cat MA, Bernardo A, Greco A, Minghetti L. Non-steroidal anti-inflammatory drugs and brain inflammation: effects on microglial functions. Pharmaceuticals 2010;3:1949–65.
[67] Yu Y, Shen Q, Lai Y, Park SY, Ou X, Lin D, et al. Anti-inflammatory effects of curcumin in microglial cells. Front Pharmacol 2018;9:386.
[68] Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet 2013;21:134–42.
[69] Piantadosi, C.A. and Suliman, H.B. (2012) Transcriptional control of mitochondrial biogenesis and its interface with inflammatory processes. *Biochimica et Biophysica Acta (BBA)-General Subjects*, **1820**, 532-541.
[70] Dang CV. c-Myc target genes involved in cell growth, apoptosis, and metabolism. Mol Cell Biol 1999;19:1–11.
[71] Belin S, Nawabi H, Wang C, Tang S, Latremoliere A, Warren P, et al. Injury-induced decline of intrinsic regenerative ability revealed by quantitative proteomics. Neuron 2015;86:1000–14.
[72] Ma J-J, Ju X, Xu R-J, Wang W-H, Luo Z-P, Liu C-M, et al. Telomerase reverse transcriptase and p53 regulate mammalian peripheral nervous system and CNS axon regeneration downstream of c-Myc. J Neurosci 2019;39:9107–18.
[73] Bonda, D.J., Lee, H.-p., Kudo, W., Zhu, X., Smith, M.A. and Lee, H.-g. (2010) Pathological implications of cell cycle re-entry in Alzheimer disease. *Expert reviews in molecular medicine*, **12**.
[74] Ferrer I, Blanco R, Carmona M, Puig B. Phosphorylated c-MYC expression in Alzheimer disease, Pick's disease, progressive supranuclear palsy and corticobasal degeneration. Neuropathol Appl Neurobiol 2001;27:343–51.
[75] Yang Y, Geldmacher DS, Herrup K. DNA replication precedes neuronal cell death in Alzheimer's disease. J Neurosci 2001;21:2661–8.
[76] Majd S, Power J, Majd Z. Alzheimer's disease and cancer: when two monsters cannot be together. Front Neurosci 2019;13:155.
[77] Ariga H. Common mechanisms of onset of cancer and neurodegenerative diseases. Biol Pharm Bull 2015;38:795–808.
[78] Marinkovic T, Marinkovic D. Obscure Involvement of MYC in Neurodegenerative Diseases and Neuronal Repair. Mol Neurobiol 2021:1–9.
[79] Ferrer I, Blanco R. N-myc and c-myc expression in Alzheimer disease, Huntington disease and Parkinson disease. Mol Brain Res 2000;77:270–6.
[80] Hou J, Ye X, Feng W, Zhang Q, Han Y, Liu Y, et al. Distance correlation application to gene co-expression network analysis. BMC Bioinf 2022;23:1–24.