scientific reports

Check for updates

OPEN

Use of a graph neural network to the weighted gene co-expression network analysis of Korean native cattle

Hyo-Jun Lee[1], Yoonji Chung[2], Ki Yong Chung[3], Young-Kuk Kim[4], Jun Heon Lee[2], Yeong Jun Koh[4,5]✉ & Seung Hwan Lee[2,5]✉

In the general framework of the weighted gene co-expression network analysis (WGCNA), a hierarchical clustering algorithm is commonly used to module definition. However, hierarchical clustering depends strongly on the topological overlap measure. In other words, this algorithm may assign two genes with low topological overlap to different modules even though their expression patterns are similar. Here, a novel gene module clustering algorithm for WGCNA is proposed. We develop a gene module clustering network (gmcNet), which simultaneously addresses single-level expression and topological overlap measure. The proposed gmcNet includes a "co-expression pattern recognizer" (CEPR) and "module classifier". The CEPR incorporates expression features of single genes into the topological features of co-expressed ones. Given this CEPR-embedded feature, the module classifier computes module assignment probabilities. We validated gmcNet performance using 4,976 genes from 20 native Korean cattle. We observed that the CEPR generates more robust features than single-level expression or topological overlap measure. Given the CEPR-embedded feature, gmcNet achieved the best performance in terms of modularity (0.261) and the differentially expressed signal (27.739) compared with other clustering methods tested. Furthermore, gmcNet detected some interesting biological functionalities for carcass weight, backfat thickness, intramuscular fat, and beef tenderness of Korean native cattle. Therefore, gmcNet is a useful framework for WGCNA module clustering.

Weighted gene co-expression network analysis (WGCNA) is often used to explore the system-level functionality of gene sets. WGCNA groups thousands of genes into a number of modules, simplifying biological interpretation. The general framework of WGCNA[1] can be summarized as follows. First, the adjacencies of paired genes are calculated to define the gene co-expression network. The adjacencies are then incorporated into a topological overlap measure (TOM) to reveal gene-gene connections. Using the TOM, a clustering algorithm assigns intensively connected genes to the same modules. Finally, functional analyses are used to determine the biological meanings of the modules. This pipeline has been widely used in various fields. For example, recent biomedical studies used WGCNA to identify specific modules and hub genes related to human cancer[2] and arterial disease[3]. In animal and plant sciences, WGCNA has often been used to profile plant gene expression[4] and detect pathways responsible for complex animal traits[5,6]. The module definitions greatly affect the interpretations of the results. WGCNA commonly uses a hierarchical clustering (HC) algorithm. This unsupervised clustering method places adjacent genes into the same modules based on pairwise TOM data. However, a concern has been raised that transformations of gene expressions into a TOM results in loss of raw-level expression features. HC-based module assignment depends strongly on the TOM. This can degrade similarity of expression not only between modules, but also within modules. In other words, HC may assign two genes with low topological overlap to different modules even though their expression patterns are similar. Furthermore, once a gene is added to a specific module, HC can never reverse the decision. This poses challenges when clustering complicated networks with

[1]Department of Bio-AI Convergence, Chungnam National University, Daejeon 305-764, Korea. [2]Division of Animal and Dairy Science, Chungnam National University, Daejeon 305-764, Korea. [3]Department of Beef Science, Korea National College of Agriculture and Fisheries, Wanju 54874, Korea. [4]Department of Computer Science & Engineering, Chungnam National University, Daejeon 305-764, Korea. [5]These authors contributed equally: Yeong Jun Koh and Seung Hwan Lee. ✉email: yjkoh@cnu.ac.kr; slee46@cnu.ac.kr

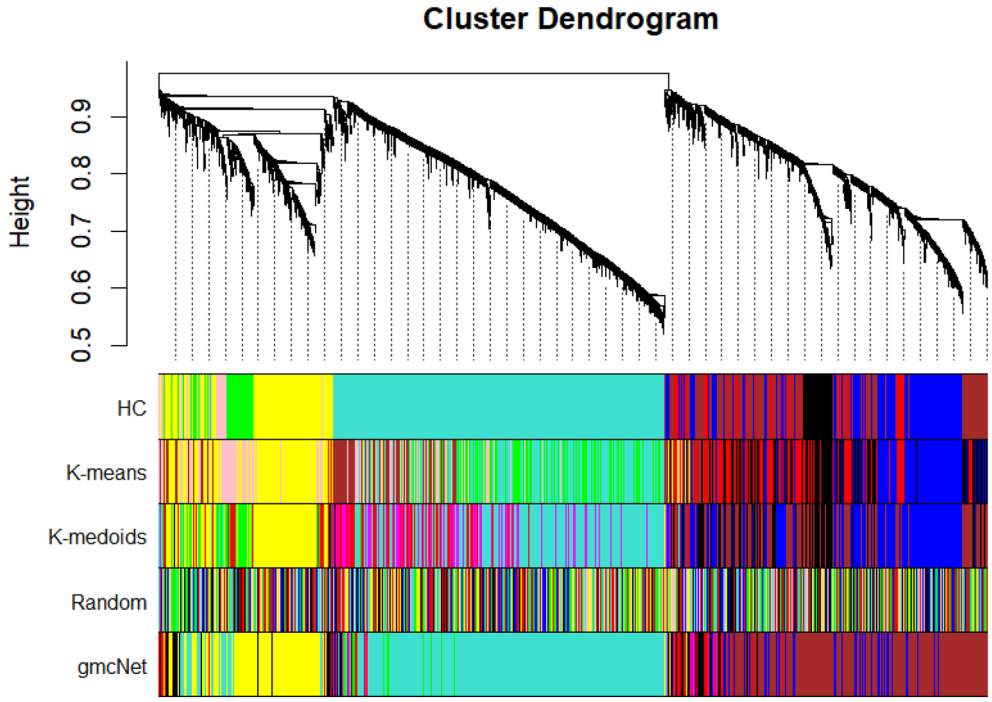

Figure 1. Module clustering results. The upper panel displays the hierarchical clustering dendrogram. In the lower panel, the colors show the module memberships determined by the methods on the left.

many interconnected gene pairs. Thus, a new algorithm is needed to more accurately identify WGCNA gene modules. Langfelder et al.[7] developed a "dynamic tree cut" technique that clusters gene modules based on the shapes of dendrogram branches, but this still depends on TOM. Botía et al.[8] employed a derivative of K-means processing to refine gene modules generated by standard HC. However, this algorithm requires more than four steps beginning with module clustering, centroid computation, distance measurements, and gene relocation. This complex pipeline requires significant computational time and is thus unsuitable for very large networks.

A graph neural network (GNN)[9–12] is a good alternative algorithm for module clustering. GNNs extend deep neural networks to learn a graph representation by finding stable features of nodes and its neighbors in graph-based data. Gilmer et al.[11] introduced a general framework termed message-passing neural network (MPNN), which effectively aggregates each node with its neighbors into embedding features. Many other studies for GNN have achieved impressive performance using this framework[12,13]. Given the recent successes of GNN, graph-based learning methods have been widely applied in bioinformatics. To predict drug-target interactions, recent studies employed various graphical convolutional networks[14,15]. For single-cell RNA-seq analysis, a GNN was used to model cell-cell relationships[16] and impute gene expression levels within single cells[17]. Yang et al.[18] developed a GNN that extracted protein features from graphical information. However, most studies on WGCNA did not use GNN for module clustering.

In this paper, we introduce a GNN-based clustering algorithm for WGCNA: the gene module clustering network (gmcNet). Our method clusters genes based on their co-expression topologies (genes in the same module should be strongly connected) and single-level expression (genes in the same module should exhibit similar expression patterns). The main innovation of gmcNet is incorporating the expression feature of single gene with co-expression feature of their neighbor genes. gmcNet includes a "co-expression pattern recognizer" (CEPR) and a module classifier. The CEPR has a message-passing (MP) operation similar to that of MPNN[11], except that the topological overlap matrix[1] is used as the input rather than the adjacency matrix. Using the former matrix, CEPR defines weighted relationships, consistent with the objective of WGCNA. The module classifier assigns genes to various modules using the CEPR-embedded features. We tested gmcNet using RNA-seq data for native Korean cattle, and compared the performance to that of other clustering algorithms. We also validated gmcNet performance on gene expression datasets of human, mouse, pig, and chicken which were downloaded from the Gene Expression Omnibus (GEO) repository[19]. As GNNs are not widely used for WGCNA, our findings will be of interest to computational biologists.

Results

Model performance on Korean native cattle. To validate gmcNet performance, it was compared to four baseline clustering algorithms including HC, K-means clustering, and K-medoids clustering (Fig. 1). We measured performance in terms of clustering strength and functional enrichment. We used graph modularity[20] to measure the clustering strength, and the differentially expressed module (DEM) signals to assess functional enrichment.

Method	HC	K-means	K-medoids	gmcNet
$\mathcal{Q}$	0.219	0.138	0.171	0.261
DEM-signal	18.618	22.723	18.236	27.739

Table 1. Model performance on Korean native cattle dataset in terms of graph modularity $\mathcal{Q}$ and DEM signaling.

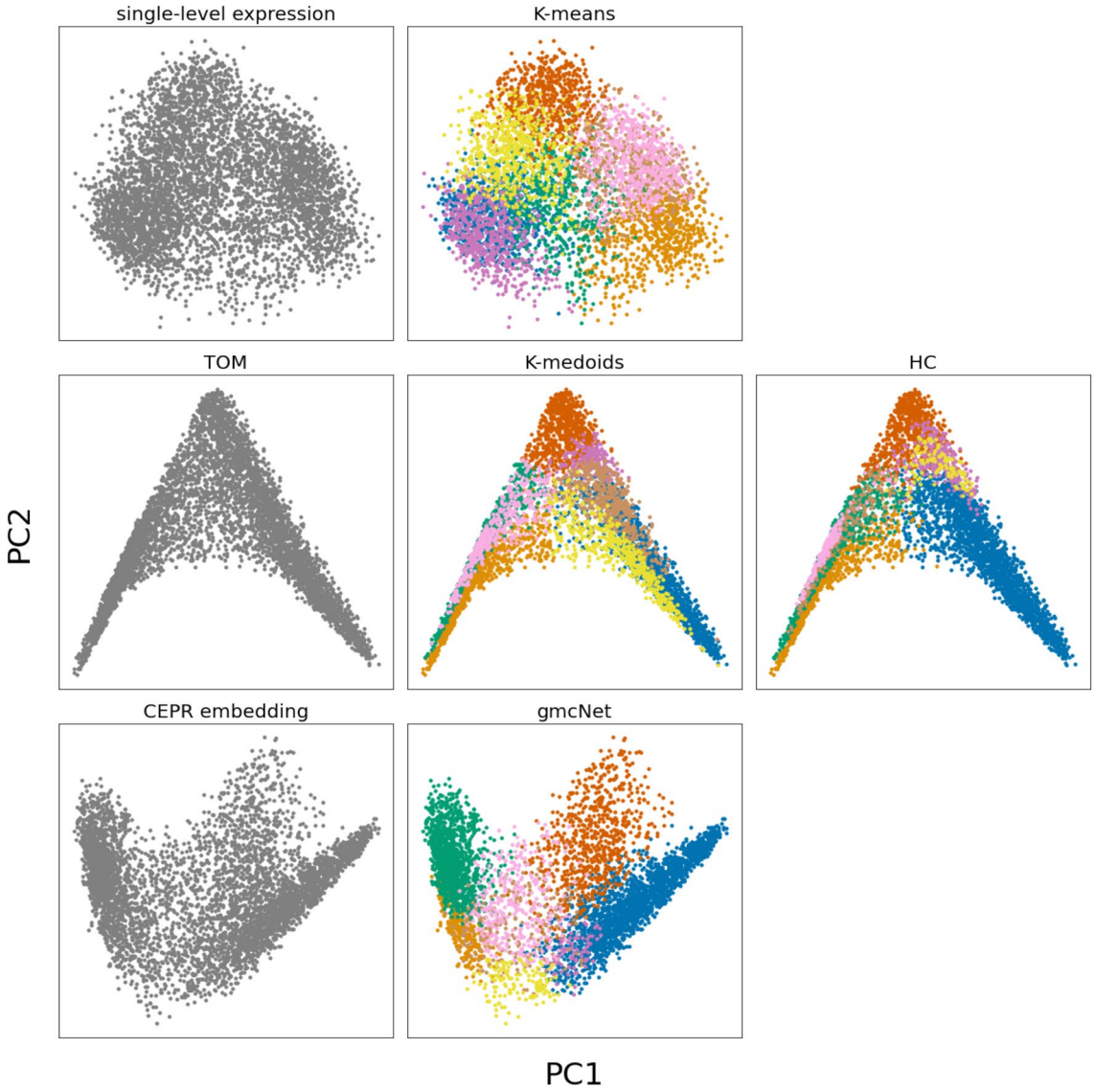

Figure 2. First and second principal components of three feature type of Korean native cattle dataset and clustering results of each method. The x-axis and y-axis are first and second principal component. The colors show the module memberships determined by the methods on the top.

Table 1 presents the performances of the various methods on Korean native cattle dataset. The single gene expression-based method (K-means) is robust to DEM signal capture, whereas the TOM-based methods (HC, K-medoids) provide higher modularity. On the other hand, gmcNet, which leverages both single gene expression and TOM, achieves the best DEM signal (27.739) and cluster modularity ($\mathcal{Q}$: 0.261). Comparison of gmcNet and HC revealed that gmcNet markedly increases modularity and the DEM signal by 0.042 and 9.121, respectively. Thus, gmcNet is more powerful than the other methods for revealing the apparent closeness of genes within the same module, and when making biological sense of the complex traits of native Korean cattle.

CEPR embedding. Figure 2 shows plots based on the first and second principal components of three feature types (single-level expression, TOM, and CEPR embedding) of Korean native cattle dataset. Single-level

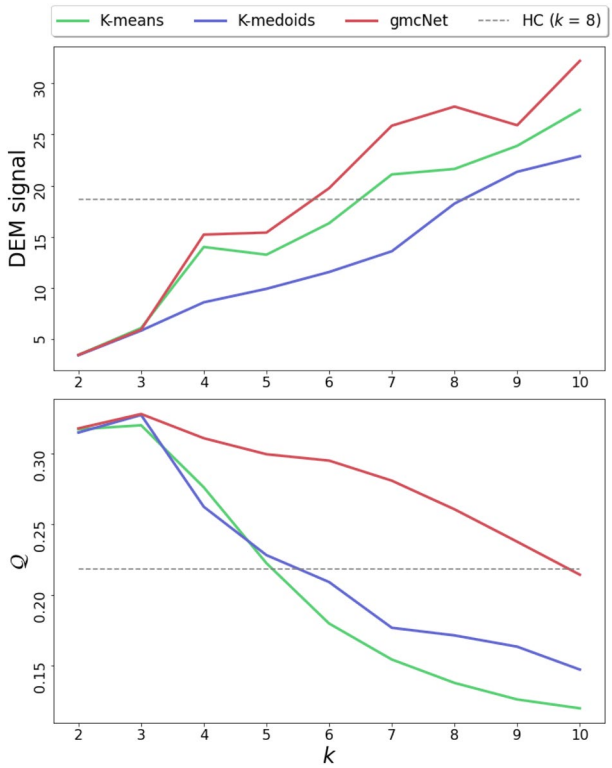

Figure 3. Optimal k searching considering DEM signaling and the modularity $\mathcal{Q}$.

expression fails to distinguish modules with ambiguous boundaries. This may reflect the low modularity of K-means, which uses single-level expression for clustering. The TOM provides stronger connections between genes than single-level expression. However, it also decreases the distances between different modules and genes. As shown in Fig. 2, K-medoids and HC, which use the TOM for clustering, do not clearly assign genes into different but closely related modules. Compared to the other types, CEPR embedding provides better separation, *i.e.* smaller distances between genes and larger ones between modules. With CEPR embedding, gmcNet defines gene modules more clearly and increases modularity.

Model performance at different *k* (number of clusters). Our current implementation of gmcNet requires the setting of an optimal k (number of clusters). The effects of the k-value on modularity $\mathcal{Q}$ and the DEM signal are summarized in Fig. 3. With an increasing k-value, the DEM signal increases while the $\mathcal{Q}$ decreases. In contrast, gmcNet yields a larger DEM signal than HC even at smaller k-values ($6 \leq k < 8$), and remains higher $\mathcal{Q}$ at larger k-values ($k = 9$). gmcNet outperforms K-means and K-medoids for all k-values. These results can demonstrate the superiority of gmcNet regardless of the k-value.

Functional enrichment analysis of native Korean cattle. To identify the DEMs, we performed linear regression analysis of the module eigengenes[1] for four complex traits, including carcass weight (CWT), backfat thickness (BF), intramuscular fat content (IMF), and the Warner-Bratzler shear force (WBSF). Figure 4 shows the results. In terms of the number of DEMs, IMF ranked first with four modules (K2, K3, K4, and K8) followed by BF (K2, K3, and K4), WBSF (K5 and K7), and CWT (K1 and K6). Interestingly, K5 and K7, which contain large numbers of genes, were significant to WBSF. This may reflect our mode of data collection; the RNA-seq data were from the *longissimus-dorsi* muscle and WBSF indicates the tenderness of beef muscle. Also, gmcNet detected 11 significant module-trait interactions. gmcNet found more DEMs than the other methods (HC: 9, K-means: 10, and K-medoids: 10) (Fig. S1).

We used Gene Ontology (GO) enrichment analysis[21] to annotate the biological processes of the modules defined by gmcNet. Three modules (K1, K5, and K7) were linked to significant processes (Fig. 5). K1, a CWT-related module, was enriched in "biosynthetic" and "metabolic" processes. Based on both the DEM analysis and the GO enrichment results, K1 seems to involve many genes associated with growth-related traits. Two WBSF-related modules (K5 and K7) were enriched in "immune system" and "protein catabolism" , respectively. Although several studies have suggested that the immune system plays a key role in cattle weight gain and feed efficiency[22,23], the association between beef tenderness and immune pathways is a novel finding. Various studies have reported an association between "protein catabolic process" and beef tenderness[24–26]. Therefore, the results suggest that K7 is a key module of beef tenderness in native Korean cattle.

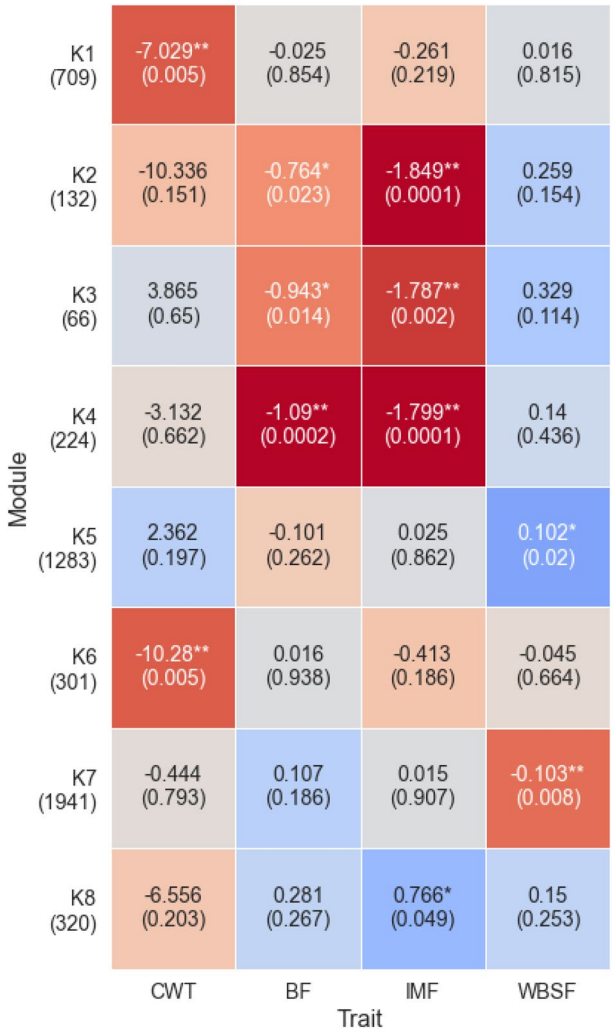

Figure 4. The DEM signals of modules defined by gmcNet. The y-axis shows the module names and numbers of genes within each module. The x-axis shows the complex traits. The numbers in each cell are regression coefficients (no parentheses) and the regression p-values (in parentheses). Red and blue indicate negative and positive coefficients, respectively. $^{*}p < 0.05$, $^{**}p < 0.01$.

Hub gene searches for modules of interest. Given the functional enrichment results, we selected the four modules, K1, K2, K4, and K7, as the principal modules of complex traits. Figure 6 shows the hub gene networks and Table 2 shows the related traits. The six hub genes of K1 are related to quantitative traits including growth (*LAMTOR5*[27] and *PAM16*[28]) and feed intake (*NDUFB1*[29], *NDUFB4*[30], *ATP5MF*[31], and *SEC61G*[32]). These findings support our suggestion that K1 is significant in terms of CWT. K2 and K4, associated with fat-related traits (BF and IMF) in DEM analysis, include eight (*ACSL3*[33], *NFKB1*[34], *CYP2R1*[35], *HSF2*[36], *TMEM135*[37], *PDCD4*[38], *HERPUD2*[39], and *NMRAL1*[40]) and seven (*SPNS1*[34], *MYOD1*[41], *PDXK*[42], *TMUB1*[37], *ARHGAP26*[43], *RAB15*[34], and *TP73*[44]) fat-related hub genes, respectively. Thus, future research should identify the relationships between fat metabolism and modules K2 and K4. Although K7 was associated with WBSF in DEM analysis, only four hub genes (*PARD3*[45], *EIF4G3*[46], *PAFAH1B1*[47], and *CAMTA2*[48]) were associated with growth-related traits; the other hub genes were all novel.

Gene Expression Omnibus (GEO) repository. We also performed our method on the NCBI GEO datasets[19]. The datasets include four different species (GDS6010: human, GDS5618: mouse, GDS4246: pig, and GDS3857: chicken). We measured DEM signals using the trait included in each dataset (human: virus infection, mouse: pancreatic islets, pig: blood, chicken: light pulse). The implementation details for GEO datasets can be shown in supproting information S2. Table 3 presents the performances of the various methods on GED datasets. For mouse and chicken gmcNet achieves the best cluster modularity, while for human and pig gmcNet show much lower modularity than other TOM-based method (HC and K-medoids). However, gmcNet outperforms all methods on DEM signal capture with reasonable modularity for all datasets. These results can prove the gmcNet is useful method to group thousands of gene according to their system-level functionality.

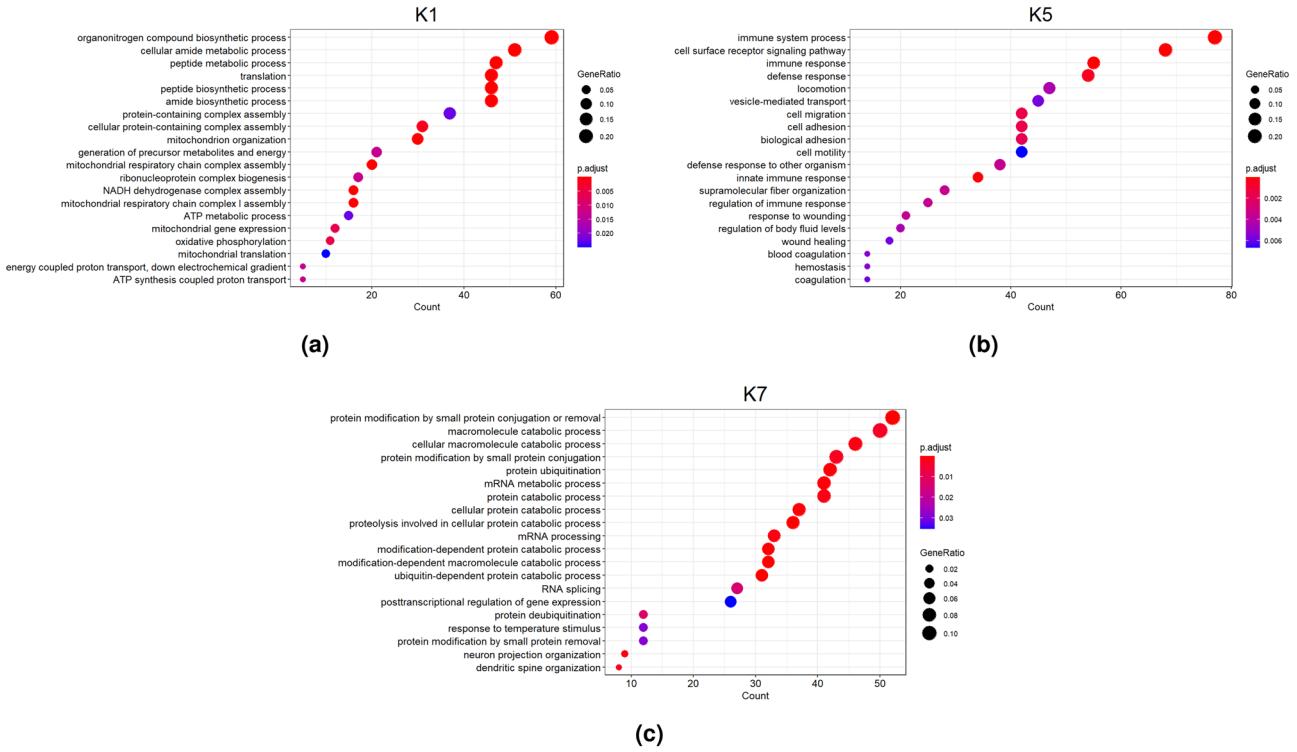

Figure 5. The biological processes of three significant modules: (**a**) K1, (**b**) K5, and (**c**) K7. p.adjust is a *p*-value adjusted by the Bonferroni method.

Discussion

Single-level expression is generally appropriate to identify trait-specific marker genes that are differentially expressed depending on the biological phenotype[63]. Here, we found that single-level expression also revealed trait-specific modules with strong DEM signals. However, most existing WGCNA methods address only the co-expression topology (including TOM); the DEM signals are weak. On the other hand, our gmcNet simultaneously addresses single-level expression and TOM. gmcNet thus yielded larger DEM signals than other clustering methods. Furthermore, gmcNet produced some novel and interesting results. Threfore, gmcNet can detect module functionality and improves our understanding of WGCNA system-level biology. Also, gmcNet yields strong adjacencies between genes in the same module. gmcNet exploits the learnable properties of CEPR, which aggregates single-gene expressions with the co-expression features of its first neighbors, embedding these features to reduced dimensions. As noted in the Results section, CEPR generates more robust features than single-gene expression data or TOM. Given the CEPR-embedded feature, gmcNet achieved the best WGCNA modularity of all clustering methods tested.

Many genes are uniformly expressed in all individuals. Such genes ("noise") are intimately connected with nested modules and exhibit no differential expression in complex trait analysis. Any attempt to cluster them disrupts module identification and obscures the biological implications. HC uses a dendrogram cut-off to exclude noisy genes. On the other hand, gmcNet assigns every gene to the most probable module. This may yield some meaningless assignments, because uniform expression may render the assignments to nested modules similar. Therefore, in future, it will be important to eliminate noise. We are exploring probability thresholding to this end. Specifically, genes with maximum probabilities lower than a given threshold will be excluded from module assignment. We will also add the optimal *k* search method to gmcNet; *k*-values can greatly increase model performance and may be modified depending on the characteristics of a dataset. Here, gmcNet used the optimal *k* of HC and performed better than other methods. In addition, gmcNet outperformed K-means and K-medoids at all *k*-values tested (2-10). Thus, the addition of an optimal *k* search would improve gmcNet performance in the context of WGCNA.

We derived a gene module clustering network, gmcNet, which simultaneously addresses single-level expression and TOM. We validated gmcNet performance using 4,976 genes from 20 native Korean cattle and four GEO datasets. gmcNet reliably assigned genes to modules exhibiting high modularity and DEM signals. gmcNet also detected some interesting biological functionalities. Therefore, gmcNet is a useful framework for WGCNA module clustering.

Materials and methods

Korean native cattle data. A total of 20 native Korean steers, born 2013 at Hanwoo Experiment Station, National Institute of Animal Science (NIAS), Rural Development Administration, South Korea, were used; all were humanely slaughtered at 30 months of age. The CWT (kg), and BF (mm) were measured after chilling for 24 hours. BF was measured at the junction of the 12th and 13th ribs. The WBSF and IMF were measured

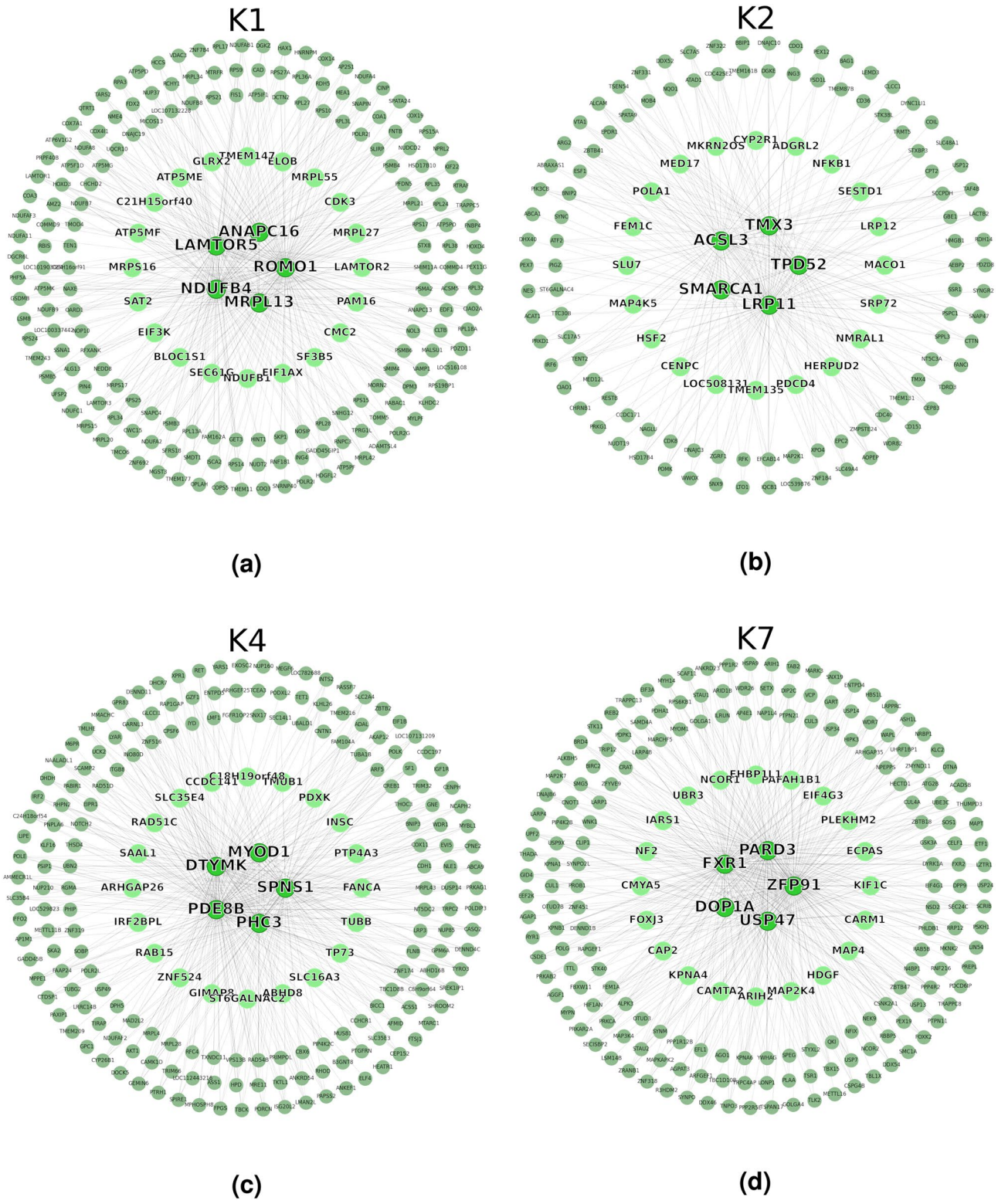

Figure 6. Hub gene networks of the four principal modules of native Korean cattle: (**a**) K1, (**b**) K2, (**c**) K4, (**d**) K7. From the outside in, the top 200, top 25, and top 5 hub genes are shown. The linkages of the top 5 hub genes are shown as the edges of the networks.

at the *longissimus-dorsi* muscle according to[64] and[65], respectively. RNA from the *longissimus-dorsi* muscle was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA). RNA quality and quantity were assessed by automated capillary gel electrophoresis performed using a Bioanalyzer 2100 running the RNA 6000 Nano Lab-

Module	Hub gene[1]	Significant trait[2]	Reported cattle traits affected
K1	*ROMO1*, *ANAPC16*, *LAMTOR5*, *NDUFB4*, *MRPL13*, *LAMTOR2*, *MRPL27*, *CDK3*, *MRPL55*, *ELOB*, *TMEM147*, *GLRX2*, *ATP5ME*, *C21H15orf40*, *ATP5MF*, *MRPS16*, *SAT2*, *EIF3K*, *BLOC1S1*, *SEC61G*, *NDUFB1*, *EIF1AX*, *SF3B5*, *CMC2*, *PAM16*	CWT** (0.005)	Growth[27,28]; Tenderness[7,49,50,50]; Feed intake[29,31,32]; Fat[34,51,52]
K2	*TPD52*, *TMX3*, *ACSL3*, *SMARCA1*, *LRP11*, *MACO1*, *LRP12*, *SESTD1*, *NFKB1*, *ADGRL2*, *CYP2R1*, *MKRN2OS*, *MED17*, *POLA1*, *FEM1C*, *SLU7*, *MAP4K5*, *HSF2*, *CENPC*, *LOC508131*, *TMEM135*, *PDCD4*, *HERPUD2*, *NMRAL1*, *SRP72*	BF* (0.023); IMF** (0.0001)	Fat[33–35]; Growth[53–55]
K4	*SPNS1*, *MYOD1*, *DTYMK*, *PDE8B*, *PHC3*, *FANCA*, *PTP4A3*, *INSC*, *PDXK*, *TMUB1*, *C18H19orf48*, *CCDC141*, *SLC35E4*, *RAD51C*, *SAAL1*, *ARH-GAP26*, *IRF2BPL*, *RAB15*, *ZNF524*, *GIMAP8*, *ST6GALNAC2*, *ABHD8*, *SLC16A3*, *TP73*, *TUBB*	BF** (0.0002); IMF** (0.0001)	Fat[34,41,42]; Growth[51,56,57]; Feed intake[5,58,59]; Tenderness[60,61]
K7	*ZFP91*, *PARD3*, *FXR1*, *DOP1A*, *USP47*, *KIF1C*, *ECPAS*, *PLEKHM2*, *EIF4G3*, *PAFAH1B1*, *EHBP1L1*, *NCOR1*, *UBR3*, *IARS1*, *NF2*, *CMYA5*, *FOXJ3*, *CAP2*, *KPNA4*, *CAMTA2*, *ARIH2*, *MAP2K4*, *HDGF*, *MAP4*, *CARM1*	WBSF** (0.008)	Fat[62]; Growth[45–47]

Table 2. Hub genes and associated traits of the main modules. Hub gene[1]: Top 25 hub genes; Significant trait[2]: Significant traits revealed by DEM analysis (p-values). The reported traits affected by each hub gene are listed in S4 Table.

	Method	HC	K-means	K-medoids	gmcNet
Human	$\mathcal{Q}$	0.255	0.155	**0.276**	0.231
	DEM-signal	21.620	24.975	22.082	**27.558**
Mouse	$\mathcal{Q}$	0.174	0.088	0.146	**0.186**
	DEM-signal	66.505	74.591	65.494	**76.680**
Pig	$\mathcal{Q}$	**0.181**	0.122	0.177	0.132
	DEM-signal	14.11	20.388	14.182	**21.295**
Chicken	$\mathcal{Q}$	0.328	0.334	0.265	**0.366**
	DEM-signal	25.95	31.709	23.65	**33.083**

Table 3. Model performance on GEO dataset in terms of graph modularity $\mathcal{Q}$ and DEM signaling. Significant values are in [bold].

Chip (Agilent Technologies Ireland, Dublin, Ireland). Only RNA samples with RNA integrity ≥ 7 were retained. Complementary DNA (cDNA) libraries were synthesized with Illumina TruSeq preparation Kit according to the manufacturer's instruction (Illumina, San Diego, CA, USA). The RNA sequencing was done using Hiseq 2000 Illumina platform to obtain paired-end reads. The quality of the raw RNA samples was confirmed using FastQC v0.11[66], and the reads with low quality were removed using Trimmomatic v0.36[67]. The reads were aligned to the reference genome Bos taurus (Ensemble UMD3.1) with TopHat v2.1[68]. The gene count of the reads was done with HTSeq v0.91[69]. Reads per kilobase per million (RPKM) were computed for each gene. We used Pearson correlation test to filter out uniformly expressed genes for the four traits (CWT, BF, IMF, and WBSF). Specifically, we calculated correlation coefficients between each gene and the traits. Then, the genes which show non-significant correlation (p-value > 0.1) for any of the traits, were excluded in further progresses. After deriving Pearson correlation test, we excluded 7,555 genes and subjected 4,976 genes in 20 samples to this study. Notice that the National Institute of Animal Science (NIAS) of the Rural Development Administration (RDA) of South Korea approved the experimental procedures (ethics committee approval number: 2015-150).

Co-expression network construction. To represent the co-expression network in matrix form, we used the topological overlap matrix of[1]. Briefly, the adjacency of each pair of genes i and j is given by $a_{ij} = |cor_{ij}|^{\beta}$ where β is a smoothing parameter and cor_{ij} is the correlation coefficient between the single-level expressions of the two genes. Given the adjacency values a_{ij}, the topological overlap matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ was created using a TOM[70], where n is the number of genes. TOM t_{ij}, which provides a similarity measure in the topological overlap matrix, is calculated as follows:

$$t_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} \tag{1}$$

where, $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_u a_{iu}$ is a node connectivity.

Also, we constructed two additional topological overlap matrices to train gmcNet (Fig. 7). $\mathbf{T}_{\mathrm{p}} \in \mathbb{R}^{n\times}$, representing the positive network, was created leaving only positive correlation coefficients, whereas $\mathbf{T}_{\mathrm{n}} \in \mathbb{R}^{n\times n}$, representing the negative network, was created leaving only negative correlation coefficients. After scale-free model fitting[1], we chose $\beta = 6$, $\beta = 9$, and $\beta = 10$ as the smoothing parameters for $\mathbf{T}$, $\mathbf{T}_{\mathrm{p}}$, and $\mathbf{T}_{\mathrm{n}}$, respectively.

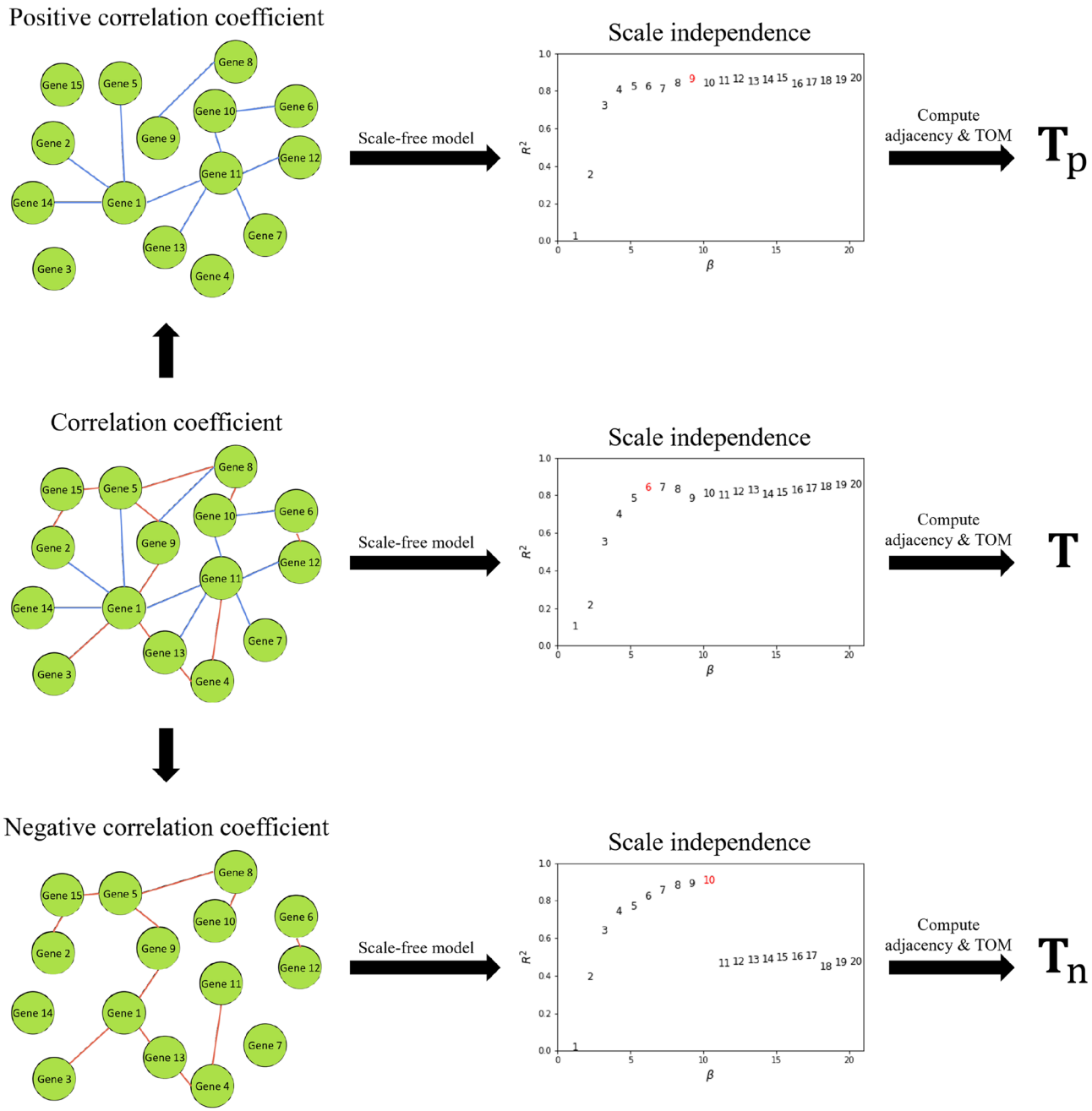

Figure 7. Construction of three topological overlap matrices. **T** is the topological overlap matrix of all relationships. $\mathbf{T}_p$ and $\mathbf{T}_n$ are the topological overlap matrices of positive and negative relationships respectively.

Gene module clustering network. We developed a gene module clustering network (gmcNet) that clusters genes according to their co-expression topologies (genes in the same module should be strongly connected) and their single-level expression (genes in the same module should exhibit similar expression patterns). Figure 8 shows an overview of gmcNet, which features a co-expression pattern recognizer (CEPR) and module classifier. The CEPR incorporates the expression features of single genes into the topological features of co-expressed ones. Given this CEPR-embedded feature, the module classifier computes module assignment probabilities.

Network structure. **CEPR:** The goal of CEPR is to integrate single-expression features with co-expression features. To achieve this, we used the MP operation of MPNN[11], but employed the topological overlap matrix rather than the adjacency matrix. We computed a new topological overlap matrix $\widetilde{\mathbf{T}}$ by zeroing the diagonal of **T** and applying degree normalization:

$$\mathbf{T}_z = \mathbf{T} - \mathbf{I}_n;\ \widetilde{\mathbf{T}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{T}_z\mathbf{D}^{-\frac{1}{2}} \tag{2}$$

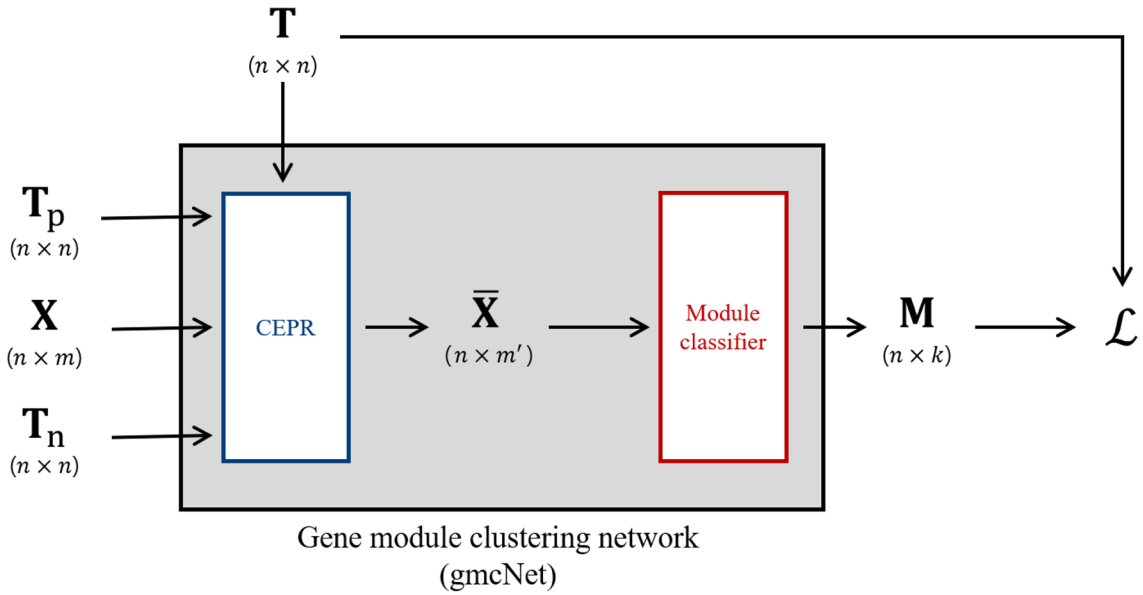

Figure 8. The architecture of gmcNet. $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the single-level expression of n genes in m samples. $\bar{\mathbf{X}} \in \mathbb{R}^{n \times m'}$ is CEPR-embedded feature with m' dimension. $\mathbf{M} \in \mathbb{R}^{n \times k}$ is assignment probability matrix of n genes to k modules. $\mathscr{L}$ is loss function.

where $\mathbf{D} = \mathrm{diag}(\mathbf{T}_z 1_n)$ is a degree matrix. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be the single-level expression of n genes in m samples. Then, single and co-expression can be simply combined via an MP operation:

$$\mathrm{MP}(\mathbf{X}, \widetilde{\mathbf{T}}) = \mathrm{ReLU}(\widetilde{\mathbf{T}}\mathbf{X}\mathbf{W}_{\mathrm{co}} + \mathbf{X}\mathbf{W}_{\mathrm{single}}) \tag{3}$$

where $\mathbf{W}_{\mathrm{co}}$ and $\mathbf{W}_{\mathrm{single}}$ are the trainable parameters of the co- and single-expression features. As $\widetilde{\mathbf{T}}$ includes the topological adjacencies between gene pairs, it is easy to see that $\widetilde{\mathbf{T}}\mathbf{X}$ can be interpreted as a co-expression feature.

A simple MP operation cannot separate positive and negative co-expressions, even when they differ in different biological pathways. Therefore, we refined a simple MP to become a CEPR, as follows:

$$\bar{\mathbf{X}} = \mathrm{CEPR}(\mathbf{X}, \widetilde{\mathbf{T}}, \widetilde{\mathbf{T}}_{\mathrm{p}}, \widetilde{\mathbf{T}}_{\mathrm{n}}) = \mathrm{ReLU}(\widetilde{\mathbf{T}}\mathbf{X}\mathbf{W}_{\mathrm{c}} + \widetilde{\mathbf{T}}_{\mathrm{p}}\mathbf{X}\mathbf{W}_{\mathrm{p}} + \widetilde{\mathbf{T}}_{\mathrm{n}}\mathbf{X}\mathbf{W}_{\mathrm{n}} + \mathbf{X}\mathbf{W}_{\mathrm{s}}) \tag{4}$$

where $\{\mathbf{W}_{\mathrm{c}}, \mathbf{W}_{\mathrm{p}}, \mathbf{W}_{\mathrm{n}}, \mathbf{W}_{\mathrm{s}}\} \in \mathbb{R}^{m \times m'}$ are the trainable weights of the co-expression, positive co-expression, negative co-expressions, and single-expression, respectively. m' is an embedding dimension (set to 8). As $\widetilde{\mathbf{T}}_{\mathrm{p}}\mathbf{X}\mathbf{W}_{\mathrm{p}}$ and $\widetilde{\mathbf{T}}_{\mathrm{n}}\mathbf{X}\mathbf{W}_{\mathrm{n}}$ are identical in terms of dimensionality, CEPR learns various co-expressions by simply adding them. By skip connections of single-expression $\mathbf{X}\mathbf{W}_{\mathrm{s}}$, CEPR generates the embedding feature $\bar{\mathbf{X}} \in \mathbb{R}^{n \times m'}$, which deals with single-expression and three different co-expressions in the m' dimension.

Module classifier: Given the CEPR-embedded feature $\bar{\mathbf{X}}$, the module classifier computes a module assignment probability using a multi-layer perceptron (MLP):

$$\mathbf{M} = \mathrm{softmax}(\bar{\mathbf{X}}\mathbf{W}_{\mathrm{m}}) \tag{5}$$

where $\mathbf{W}_{\mathrm{m}} \in \mathbb{R}^{m' \times k}$ are the trainable weights for clustering of k modules. As softmax activation guarantees that $m_{ij} \in [0, 1]$, the ith-row of $\mathbf{M} \in \mathbb{R}^{n \times k}$ corresponds to the module-assignment probability of gene i. In other words, gene i belongs to module c if m_{ic} is the maximum value of the ith-row of $\mathbf{M}$.

Loss function. For unsupervised clustering, we employed the cut and orthogonality loss terms of MinCutPool[71]. The loss function when training gmcNet was defined as:

$$\mathscr{L} = \lambda \mathscr{L}_{\mathrm{c}} + \mathscr{L}_{\mathrm{o}} = \lambda \underbrace{\left(-\frac{Tr(\mathbf{M}^{\mathrm{T}}\widetilde{\mathbf{T}}\mathbf{M})}{Tr(\mathbf{M}^{\mathrm{T}}\widetilde{\mathbf{D}}\mathbf{M})} \right)}_{\mathscr{L}_{\mathrm{c}}} + \underbrace{\left\| \frac{\mathbf{M}^{\mathrm{T}}\mathbf{M}}{\|\mathbf{M}^{\mathrm{T}}\mathbf{M}\|_F} + \frac{\mathbf{I}_k}{\sqrt{k}} \right\|_F}_{\mathscr{L}_{\mathrm{o}}} \tag{6}$$

where $\|\cdot\|_F$ indicates the Frobenius norm and Tr is the trace; λ is a balancing hyper-parameter, which is set to 2.6. The cut loss term, $\mathscr{L}_{\mathrm{c}}$, encourages clustering of strongly connected genes within the same module, and the orthogonality loss term, $\mathscr{L}_{\mathrm{o}}$, penalizes assignment to similarly sized modules.

Implementation Details. The model was iterated for 5,000 epochs using a GeForce RTX 2080ti. For the first 100 epochs, the balancing hyperparameter λ was set to 0 and the learning rate to 0.01. This prevented the creation of empty modules. After epoch 100, we set λ to 2.6 and the learning rate to 0.001. Model training was early stopped

at $\mathscr{L}_o > \tau$, where τ is the orthogonal threshold, which was set to 0.8. The Adam optimizer[72] was used to minimize the loss function. Finally, **M** at the end of training was used for module assignment.

Model performance. To validate gmcNet performance, HC[7], K-means[73] and K-medoids[74] were also used for module clustering and the results were compared to those of gmcNet. K-means uses single-expression feature **X** as input data; the HC and K-medoids use the topological distances $1 - \mathbf{T}$ as inputs. The optimal k for K-means, K-medoids, and gmcNet was set to 8, as suggested by application of the dynamic tree cut technique[7] to HC.

Metrics. We measured the model performance in terms of modularity and DEM signaling. Module modularity is a commonly used metric in graph clustering. In a fully random graph, gene i and j of degrees $c_i = \sum_u t_{iu}$ and $c_j = \sum_u t_{ju}$ are connected with a probability $c_i c_j / s$, where s is the total topological overlap $s = \sum_{ij} t_{ij}$. Modularity measures the divergence between intra-module connections as:

$$\mathscr{Q} = \frac{1}{s} \sum_{ij}^{n} (t_{ij} - \frac{c_i c_j}{s}) \delta[i,j] \tag{7}$$

where $\delta[i,j] = 1$ if i and j belong to the same module; otherwise, $\delta[i,j] = 0$.

To assess functional enrichment of clustering method, we introduce a novel metric, called DEM signal. Let $\rho[l,t] = 1$ if module l is significant (≤ 0.05) for trait t; otherwise, $\rho[l,t] = 0$. The final DEM signal was defined as:

$$\text{DEM signal} = \sum_{l}^{k} \sum_{t} -\log_{10}(p\text{-value}_{lt}) \rho[l,t] \tag{8}$$

where t is traits and p-value$_{lt}$ indicates the significance value of module l in terms of trait t. We employed linear regression analysis to the module eigengenes, *i.e.* the first principal components of the modules, for four complex traits: CWT, BF, IMF and WBSF.

Functional enrichment analysis. The Bioconductor R package "clusterProfiler"[75] was used for GO analysis. The adjusted p-value (obtained using the Bonferroni method) was employed to examine the significance (p.adjust< 0.05) of all GO terms. The top 20 biological processes were extracted if there were more than 20 significant results. To identify hub genes, we calculated the correlation coefficients between single-level expression of each gene and the ME of the module it belong to. The top 25 genes (in terms of correlation coefficients) were defined as hub genes.

Data availability

The gmcNet code and example data is available on GitHub at https://github.com/gywns6287/gmcNet. Request for full gene expression data of Korean native cattle can be made to Korea National Institute of Animal Science, Animal Genome & Bioinformatics Division (http://www.nias.go.kr/english/sub/boardHtml.do?boardId=depintro).

Received: 9 November 2021; Accepted: 27 May 2022
Published online: 14 June 2022

References

1. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. applications genetics molecular biology* **4** (2005).
2. Li, J. *et al.* Application of weighted gene co-expression network analysis for data from paired design. *Sci. Rep.* **8**, 1–8 (2018).
3. Zheng, P.-F., Chen, L.-Z., Guan, Y.-Z. & Liu, P. Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *Sci. Rep.* **11**, 1–13 (2021).
4. Rao, X. & Dixon, R. A. Co-expression networks for plant biology: why and how. Acta biochimica et biophysica Sinica 51, 981–988 (2019).
5. Salleh, M. *et al.* Rna-seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high-and low-residual feed intake in nordic dairy cattle. *BMC Genomics* **18**, 1–17 (2017).
6. Silva-Vignato, B. *et al.* Gene co-expression networks associated with carcass traits reveal new pathways for muscle and fat deposition in nelore cattle. *BMC Genomics* **20**, 1–13 (2019).
7. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. Bioinformatics 24, 719–720 (2008).
8. Botía, J. A. *et al.* An additional k-means clustering step improves the biological features of wgcna gene co-expression networks. *BMC Syst. Biol.* **11**, 1–16 (2017).
9. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *ICLR-17* (2017).
10. Xu, D., Zhu, Y., Choy, C. B. & Fei-Fei, L. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5419 (2017).
11. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *Int. Conf. Mach. Learn.* 1263–1272 (2017).
12. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1025–1035 (2017).
13. Wang, Y. *et al.* Dynamic graph cnn for learning on point clouds. *Acm. Trans. Graph. (tog)* **38**, 1–12 (2019).
14. Peng, J. *et al.* An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. *Brief. Bioinf.* (2021).
15. Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T. & Peng, J. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Brief. Bioinf.* **22**, 2141–2150 (2021).
16. Wang, J. *et al.* scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nat. Commun.* **12**, 1–11 (2021).

17. Rao, J., Zhou, X., Lu, Y., Zhao, H. & Yang, Y. Imputing single-cell rna-seq data by combining graph convolution and autoencoder neural networks. Iscience 24, 102393 (2021).
18. Yang, F., Fan, K., Song, D. & Lin, H. Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinf.* **21**, 1–16 (2020).
19. Database resources of the national center for biotechnology information. *Nucleic acids research* **46**, D8–D13 (2018).
20. Newman, M. E. Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**, 8577–8582 (2006).
21. Wu, T. *et al.* clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innov.* 100141 (2021).
22. Reynolds, J., Foote, A., Freetly, H., Oliver, W. & Lindholm-Perry, A. Relationships between inflammation-and immunity-related transcript abundance in the rumen and jejunum of beef steers with divergent average daily gain. *Anim. Gen.* **48**, 447–449 (2017).
23. Alexandre, P. A. *et al.* Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. *BMC Gen.* **16**, 1–13 (2015).
24. Zhao, C. et al. Functional proteomic and interactome analysis of proteins associated with beef tenderness in angus cattle. Livest. Sci. 161, 201–209 (2014).
25. Tian, X. *et al.* Quality and proteome changes of beef m. longissimus dorsi cooked using a water bath and ohmic heating process. *Innov. Food Sci. Emerg. Technol.* **34**, 259–266 (2016).
26. Li, Y. *et al.* Association of cast gene polymorphisms with carcass and meat quality traits in yanbian cattle of china. *Mol. Biol. Rep.* **40**, 1875–1881 (2013).
27. Ribeiro, V. M. P. et al. Genes underlying genetic correlation between growth, reproductive and parasite burden traits in beef cattle. Livest. Sci 244, 104332 (2021).
28. Kern, R. J. et al. Transcriptome differences in the rumen of beef steers with variation in feed intake and gain. Gene 586, 12–26 (2016).
29. Keogh, K., McKenna, C., Porter, R., Waters, S. & Kenny, D. Effect of dietary restriction and subsequent realimentation on hepatic oxidative phosphorylation in cattle. Animal 15, 100009 (2021).
30. Benedeti, P. D. B. *et al.* Nellore bulls (bos taurus indicus) with high residual feed intake have increased the expression of genes involved in oxidative phosphorylation in rumen epithelium. *Anim. Feed. Sci. Technol.* **235**, 77–86 (2018).
31. Nolte, W. *et al.* Identification and annotation of potential function of regulatory antisense long non-coding rnas related to feed efficiency in bos taurus bulls. *Int. J. Mol. Sci.* **21**, 3292 (2020).
32. Hardie, L. *et al.* The genetic and biological basis of feed efficiency in mid-lactation holstein dairy cows. *J. Dairy Sci.* **100**, 9061–9075 (2017).
33. Lv, Y. *et al.* Effect of acsl3 expression levels on preadipocyte differentiation in chinese red steppe cattle. *DNA Cell Biol.* **38**, 945–954 (2019).
34. Waters, S. M., Coyne, G. S., Kenny, D. A. & Morris, D. G. Effect of dietary n-3 polyunsaturated fatty acids on transcription factor regulation in the bovine endometrium. *Mol. Biol. Rep.* **41**, 2745–2755 (2014).
35. Li, Y. et al. Transcriptome profiling of longissimus lumborum in holstein bulls and steers with different beef qualities. PloS one 15, e0235218 (2020).
36. Baik, M., Vu, T., Piao, M. & Kang, H. Association of dna methylation levels with tissue-specific expression of adipogenic and lipogenic genes in longissimus dorsi muscle of korean cattle. *Asian-Australasian J. Anim. Sci.* **27**, 1493 (2014).
37. Seong, J., Yoon, H. & Kong, H. S. Identification of microrna and target gene associated with marbling score in korean cattle (hanwoo). *Gene. Gen.* **38**, 529–538 (2016).
38. Melnik, B. C., John, S. M. & Schmitz, G. Milk consumption during pregnancy increases birth weight, a risk factor for the development of diseases of civilization. *J. Transl. Med.* **13**, 1–11 (2015).
39. Yu, S.-L. *et al.* Identification of differentially expressed genes between preadipocytes and adipocytes using affymetrix bovine genome array. *J. Anim. Sci. Technol.* **51**, 443–452 (2009).
40. Engle, B., Masters, M., Boles, J. A. & Thomson, J. Gene expression and carcass traits are different between different quality grade groups in red-faced hereford steers. Animals 11, 1910 (2021).
41. Shao, T., McCann, J. C. & Shike, D. W. Effects of supplements differing in fatty acid profile to late gestational beef cows on steer progeny finishing phase growth performance, carcass characteristics, and mrna expression of myogenic and adipogenic genes. Animals 11, 1904 (2021).
42. Peletto, S. *et al.* Genetic basis of lipomatous myopathy in piedmontese beef cattle. *Livest. Sci.* **206**, 9–16 (2017).
43. Martins, R. et al. Genome-wide association study and pathway analysis for fat deposition traits in nellore cattle raised in pasture-based systems. J. Animal Breed. Genet 138, 360–378 (2021).
44. de Las Heras-Saldana, S. et al. Differential gene expression in longissimus dorsi muscle of hanwoo steers–new insight in genes involved in marbling development at younger ages. Genes 11, 1381 (2020).
45. Zhang, F. *et al.* Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: Feed efficiency and component traits. *BMC Gen.* **21**, 1–22 (2020).
46. Keogh, K. et al. Effect of dietary restriction and subsequent re-alimentation on the transcriptional profile of bovine ruminal epithelium. PloS one 12, e0177852 (2017).
47. Srivastava, S. et al. Haplotype-based genome-wide association study and identification of candidate genes associated with carcass traits in hanwoo cattle. Genes 11, 551 (2020).
48. Bazile, J. *et al.* Molecular signatures of muscle growth and composition deciphered by the meta-analysis of age-related public transcriptomics data. *Physiol. Geno.* **52**, 322–332 (2020).
49. Bernard, C. et al. New indicators of beef sensory quality revealed by expression of specific genes. J. Agric. Food Chem. 55, 5229–5237 (2007).
50. Muniz, M. M. M. *et al.* Identification of novel mrna isoforms associated with meat tenderness using rna sequencing data in beef cattle. *Meat Sci.* 108378 (2020).
51. de Lemos, M. V. A. *et al.* Association study between copy number variation and beef fatty acid profile of nellore cattle. *J. Appl. Gene.* **59**, 203–223 (2018).
52. Olivieri, B. F. *et al.* Differentially expressed genes identified through rna-seq with extreme values of principal components for beef fatty acid in nelore cattle. *J. Anim. Breed. Genet* **138**, 80–90 (2021).
53. de Almeida Santana, M. H. *et al.* Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. *J. Appl. Gene.* **57**, 495–504 (2016).
54. Anton, I. et al. Effect of single-nucleotide polymorphisms on the breeding value of fertility and breeding value of beef in hungarian simmental cattle. Acta Vet. Hungarica 66, 215–225 (2018).
55. Seabury, C. M. *et al.* Genome-wide association study for feed efficiency and growth traits in us beef cattle. *BMC Geno.* **18**, 1–25 (2017).
56. Manca, E. et al. Use of the multivariate discriminant analysis for genome-wide association studies in cattle. Animals 10, 1300 (2020).
57. Keel, B. N. *et al.* Rna-seq meta-analysis identifies genes in skeletal muscle associated with gain and intake across a multi-season study of crossbred beef steers. *BMC Geno.* **19**, 1–11 (2018).
58. Elolimy, A. A. *et al.* Skeletal muscle and liver gene expression profiles in finishing steers supplemented with amaize. *Anim. Sci. J.* **89**, 1107–1119 (2018).

59. Kong, R. S., Liang, G., Chen, Y. & Stothard, P. Transcriptome profiling of the rumen epithelium of beef cattle differing in residual feed intake. *BMC Geno.* **17**, 1–16 (2016).
60. Tizioto, P. *et al.* Variation in myogenic differentiation 1 mrna abundance is associated with beef tenderness in nelore cattle. *Anim. Gene.* **47**, 491–494 (2016).
61. Leal-Gutiérrez, J. D., Elzo, M. A., Johnson, D. D., Hamblen, H. & Mateescu, R. G. Genome wide association and gene enrichment analysis reveal membrane anchoring and structural proteins associated with meat quality in beef. *BMC Geno.* **20**, 1–18 (2019).
62. Ramayo-Caldas, Y. *et al.* A marker-derived gene network reveals the regulatory role of ppargc1a, hnf4g, and foxp3 in intramuscular fat deposition of beef cattle. *J. Anim. Sci* **92**, 2832–2845 (2014).
63. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Geno. Biol.* **15**, 1–21 (2014).
64. Wheeler, T., Shackelford, S. & Koohmaraie, M. Relationship of beef longissimus tenderness classes to tenderness of gluteus medius, semimembranosus, and biceps femoris. *J. Anim. Sci.* **78**, 2856–2861 (2000).
65. Feldsine, P., Abeyta, C. & Andrews, W. H. Aoac international methods committee guidelines for validation of qualitative and quantitative food microbiological official methods of analysis. J. AOAC Int. 85, 1187–1200 (2002).
66. Andrews, S. Fastqc: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
67. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
68. Trapnell, C., Pachter, L. & Salzberg, S. L. Tophat: Discovering splice junctions with rna-seq. *Bioinformatics* **25**, 1105–1111 (2009).
69. Anders, S., Pyl, P. T. & Huber, W. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
70. Li, A. & Horvath, S. Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics 23, 222–231 (2007).
71. Bianchi, F. M., Grattarola, D. & Alippi, C. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, 874–883 (PMLR, 2020).
72. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*, 1–15 (2015).
73. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. The.* **28**, 129–137 (1982).
74. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*, vol. 344 (John Wiley & Sons, 2009).
75. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterprofiler: An r package for comparing biological themes among gene clusters. *Omics: A J. Integr. Biol.* **16**, 284–287 (2012).

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No.2020-0-01441, Artificial Intelligence Convergence Research Center(Chungnam National University)).

Author contributions

Conceptualization: S.H.L., Y.J.K. Data Curation: K.Y.C. Formal Analysis: H.-J.L., J.H.L. Funding Acquisition: J.H.L., Y.-K.K. Methodology: H.-J.L., Y.-K.K. Software: H.-J.L., Y.J.K. Visualization: Y.C. Writing – Original Draft Preparation: H.-J.L., Y.C. Writing – Review & Editing: S.H.L., Y.J.K.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13796-9.

Correspondence and requests for materials should be addressed to Y.J.K. or S.H.L.