DOI: https://doi.org/10.15276/aait.08.2025.16
UDC 004.048+004.094

Post-clustering interpretation of gene expression data using functional enrichment and network analysis

Oleg R. Yarema[1)]
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
Denys O. Senchishen[2)]
ORCID: https://orcid.org/0000-0002-4311-7095; DSenchishen@ksu.ks.ua
Sergii A. Babichev[2,3)]
ORCID: https://orcid.org/0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
[1)] Ivan Franko National University of Lviv, 1, Universytetska Str. Lviv, 79000, Ukraine
[2)] Kherson State University, 14b Shevchenko Str, Sivka-Voynylivska, Ivano-Frankivsk Region, 77311, Ukraine
[3)] Jan Evangelista Purkyně University in Ústí nad Labem, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic

ABSTRACT

Clustering of gene expression profiles is a core technique used to reveal hidden biological structures and differentiate disease subtypes in high-dimensional biomedical datasets. Nevertheless, translating cluster structures into biologically meaningful insights requires integrative analytical strategies that go beyond unsupervised learning. In this work, we introduce a novel integrative computational approach that emphasizes post-clustering interpretation by combining statistical functional enrichment with network-based modeling. Clusters of gene expression profiles, previously identified in patients with distinct cancer types, were subjected to enrichment analysis using Gene Ontology, the Kyoto Encyclopedia of Genes and Genomes, and Reactome databases. The enrichment was performed with the g:Profiler tool, allowing the detection of significantly overrepresented biological processes, molecular functions, cellular components, and signaling pathways within each cluster. To visualize and further interpret the enriched functional categories, Cytoscape software was employed. Functional interaction networks were constructed using two key modules: *ClueGO*, which integrates Gene Ontology and pathway annotation into a functionally grouped network, and *CluePedia*, which expands these networks by showing relationships between genes and enriched terms. This network-based visualization enabled deeper biological interpretation and facilitated the identification of core functional themes. The analysis revealed that each gene cluster is associated with distinct biological processes, such as immune signaling, metabolic pathways, DNA repair, or cell cycle regulation. The novelty of the proposed approach lies in its systematic integration of enrichment statistics with graph-based visualization, ensuring both computational rigor and biological interpretability**.** These findings confirm that the method can extract biologically consistent knowledge from complex gene expression data. In summary, the study presents an innovative post-clustering interpretation strategy that bridges unsupervised machine learning and functional genomics. This approach advances the explainability of computational analysis and supports its application in disease subtyping, biomarker discovery, and personalized medicine research.

Keywords: Computational biology data analysis; bioinformatics; integrative analysis; gene expression data; post-clustering interpretation; functional enrichment; gene ontology; kyoto encyclopedia of genes and genomes; reactome; cytoscape; network-based analysis

For citation**: Yarema O., Senchishen D., Babichev S. "Post-clustering interpretation of gene expression data using functional enrichment and network analysis".** ***Applied Aspects of Information Technology.*** **2025; Vol.8 No.3: 249–262: DOI: https://doi.org/10.15276/aait.08.2025.16**

INTRODUCTION

Clustering of gene expression profiles is a widely used unsupervised learning technique for identifying genes with similar regulatory behavior across experimental conditions or patient groups [1], [2], [3]. Although clustering reveals underlying structure in high-dimensional biological data, the resulting clusters alone offer limited insight unless complemented by biological interpretation. A critical post-clustering step involves evaluating the functional coherence of genes within each cluster and identifying biologically relevant functional modules.

Genes grouped together often participate in common cellular pathways or molecular functions. Thus, functional enrichment analysis enables researchers to assess whether gene clusters are significantly associated with specific biological processes, molecular functions, cellular components, or signaling pathways. Among the most established resources for such analyses are Gene Ontology (GO) [4], [5], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [6], and the Reactome pathway database [7]. Kyoto Encyclopedia of Genes and Genomes provides detailed maps of metabolic and signaling pathways, while Reactome offers a comprehensive and curated database of biological reactions, spanning from signal transduction to metabolism. These resources support the contextual interpretation of clustered genes in terms of biological meaning.

Beyond enrichment statistics, network-based methods provide deeper insight by visualizing and analyzing interactions between genes and functional categories. In this study, we used an integrative approach that combines the R programming environment with Cytoscape – an open-source platform for network visualization and analysis [8]. The Cytoscape plugins *ClueGO* [9] and *CluePedia* [10] were employed to generate functionally grouped GO/pathway networks.

The interpretation pipeline includes the following key steps.

1. Preparing gene lists for each cluster based on the clustering output.
2. Importing gene lists into Cytoscape using the *ClueGO* plugin.
3. Performing enrichment analysis using GO (Biological Process, Molecular Function, Cellular Component), KEGG, or Reactome.
4. Visualizing enriched terms as functionally grouped GO networks.
5. Interpreting the biological role of each cluster using enriched functional terms and visualized relationships.

This integrative strategy bridges the gap between unsupervised clustering and biological relevance, enhancing the interpretability of gene expression patterns in the context of complex diseases such as cancer.

LITERATURE SURVEY

Biological interpretation of gene expression clustering results has become an essential task in transcriptomic studies, particularly for understanding the functional coherence of gene groups in relation to disease mechanisms. While clustering methods can identify genes with similar expression profiles, the biological relevance of these clusters must be established using domain-specific resources and tools developed in recent years.

Gene Ontology (GO), KEGG, and Reactome are the three primary knowledge bases widely used for functional enrichment analysis. GO offers a hierarchical structure of gene annotations spanning biological processes, molecular functions, and cellular components, and has evolved substantially over the past decade to reflect current biological knowledge [11]. KEGG provides detailed pathway maps for interpreting the biological context of genes, including signaling cascades and metabolic routes [6]. Similarly, Reactome presents curated pathways and reactions with a high level of biological granularity, making it suitable for comprehensive interpretation of gene clusters [7].

Recent methodological advancements focus on the integration of functional annotation and network-based visualization. The Cytoscape platform [8] and its plugins such as *ClueGO* [9] and *CluePedia* [10] have gained popularity for creating functionally grouped networks from enriched GO and pathway terms. These tools allow not only for the identification of overrepresented categories but also for the exploration of their interrelations, which is critical for elucidating higher-level biological mechanisms.

Furthermore, several studies emphasize the importance of combining clustering with ontology analysis to extract informative subsets of genes. For example, in [12], [13] the authors demonstrated the use of biclustering techniques in conjunction with GO-based enrichment to identify condition-specific gene modules in high-dimensional cancer datasets. Their work illustrates the potential of such hybrid frameworks for biomarker discovery and disease classification.

Advances in computational tools, such as *edgeR* [14] and *DESeq2* [15], have also enabled more accurate differential expression analysis, which precedes clustering and enrichment. Enrichment platforms like *g:Profiler*, integrated in many pipelines, support multi-database analysis for robust biological interpretation [16].

In summary, the literature analysis highlights a growing consensus around the integration of statistical enrichment and network visualization as key components in post-clustering gene expression analysis. These approaches bridge the gap between computational findings and biological interpretation, making them indispensable in modern bioinformatics and personalized medicine research.

The **goal of this study** is to develop and validate an integrative method for the biological interpretation of clustered gene expression profiles using functional enrichment techniques and network-based visualization.

To achieve this goal, the following tasks are addressed:

- comparison of the methods for the biological interpretation of clustering results of gene expression profiles using functional enrichment analysis (GO, KEGG, Reactome) and network-based modeling;
- implementation of the analytical pipeline combining R and Cytoscape for visualization of functional networks using *ClueGO* and *CluePedia* plugins;

- validation of the proposed method on real-world gene expression datasets from patients diagnosed with various cancer types.

GENE ONTOLOGY: CONCEPT AND IMPLEMENTATION

Gene Ontology is a widely adopted framework for the functional annotation of genes in the form of a structured ontology. GO encompasses three primary domains: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC), reflecting, respectively, the biological objectives a gene contributes to, the biochemical activity it performs, and the cellular location where the activity occurs. GO-based analysis serves as a foundation for the biological interpretation of gene lists derived from clustering results. It enables the assessment of whether a given gene set (e.g., genes in a cluster) is statistically enriched for specific GO categories beyond random expectation.

The main types of GO analysis include:

- Over-Representation Analysis (ORA): assesses whether certain GO terms occur more frequently in the input gene list compared to a reference background;
- Gene Set Enrichment Analysis (GSEA): identifies GO terms that are enriched in a ranked list of genes without requiring an explicit cutoff.

In this study, GO analysis was performed in the R environment using the *clusterProfiler* package [17]. This package supports the use of Entrez Gene IDs, offers both ORA and GSEA methodologies, and includes extensive options for visualization such as enrichment plots, dot plots, and term similarity networks. Since each GO domain captures a different aspect of gene function, a combined analysis across BP, MF, and CC was adopted to provide a comprehensive functional interpretation of gene clusters. After merging enrichment results from all domains, duplicates were removed and only the most statistically significant terms were retained. This strategy enables simultaneous consideration of gene function, molecular activity, and subcellular localization.

The step-by-step GO analysis procedure is as follows.

1. *Gene list preparation*: Extract Entrez Gene IDs for each expression profile cluster, treating each cluster independently.

2. *Library import*: Load R libraries for GO analysis, including the annotation database for Homo sapiens.

3. *GO enrichment*: Perform enrichment analysis separately for BP, MF, and CC, applying multiple testing correction. Merge the results across domains into a unified list, removing duplicates.

4. *Result simplification*: Reduce redundancy by eliminating highly similar GO terms based on semantic similarity, retaining the most representative ones.

5. *Visualization*: Display the results using dot plots or similarity-based term graphs to facilitate interpretation.

6. *Interpretation*: Focus on GO terms with the highest statistical significance and coverage, enabling inference about the dominant biological functions within each cluster.

Thus, GO analysis is a critical first step in the biological interpretation of gene expression clusters. Its integration into the clustering analysis pipeline provides a quantitative means of evaluating the functional relevance of each identified cluster.

KEGG: ANALYSIS OF METABOLIC AND SIGNALING PATHWAYS

The Kyoto Encyclopedia of Genes and Genomes [6] is one of the most comprehensive databases for understanding high-level functions and utilities of biological systems, particularly in terms of gene–gene, gene–protein, and gene–metabolite interactions. Unlike GO, which focuses on general biological functions, KEGG provides context-specific knowledge by representing genes as components of biochemical networks and molecular pathways. These pathways are directly associated with physiological and pathological processes, making KEGG a valuable resource for interpreting clusters of gene expression profiles.

The KEGG enrichment analysis was carried out using the *clusterProfiler* package in R, which enables pathway enrichment based on Entrez Gene IDs, incorporates multiple testing correction (e.g., Benjamini-Hochberg), and supports output formats suitable for downstream interpretation and visualization.

The step-by-step KEGG enrichment procedure is as follows.

1. *Input data preparation*: For each gene expression cluster, a list of Entrez Gene IDs is compiled to serve as the basis for enrichment analysis.

2. *Identification of enriched KEGG pathways*: Hypothesis testing is applied to determine whether specific KEGG pathways are statistically overrepresented in the gene list relative to a reference background. Correction for multiple testing is performed to control the false discovery rate.

3. *Aggregation of results*: When multiple clusters are analyzed, results can be aggregated to identify both cluster-specific and shared enriched pathways. This enables the construction of a functional landscape across all clusters.

4. *Visualization*: Results are visualized using dot plots, bar plots, or KEGG pathway networks, which illustrate the distribution of genes across pathways, their statistical significance, and coverage.

5. *Interpretation*: KEGG pathways with high statistical significance and dense gene coverage are considered biologically informative. These pathways often reflect key cellular processes such as metabolism, signal transduction, cell proliferation, apoptosis, and immune response.

Thus, KEGG analysis provides a functional complement to GO-based interpretation by revealing biochemical and regulatory contexts in which gene clusters are involved. The integration of KEGG enrichment results into the cluster interpretation workflow offers a deeper understanding of the potential roles that expression modules play in complex molecular systems.

REACTOME PATHWAY ANALYSIS: ONTOLOGY OF BIOLOGICAL REACTIONS

Reactome [7] is an open-access, manually curated knowledge base focused on modeling biological reactions and molecular interaction cascades that constitute key cellular processes such as signal transduction, transcription, metabolism, the cell cycle, and immune responses. Unlike GO, which classifies gene function based on descriptive categories, Reactome adopts an event-based approach, modeling biological knowledge as interconnected reactions – from elementary molecular events to complex signaling pathways. Owing to its hierarchical organization, Reactome enables the reconstruction of biological context in the form of ordered biochemical steps. Applying Reactome enrichment analysis to gene expression clusters allows us to determine which reactions or high-level processes the clustered genes are involved in, as well as to explore potential functional links across clusters.

The step-by-step Reactome analysis procedure is as follows.

1. *Input data preparatio*n: For each cluster, a gene list in Entrez ID format is generated to serve as input for identifying enriched pathways in the Reactome database.

2. *Identification of enriched pathways*: Statistical enrichment analysis is performed to detect biological reactions in which the cluster genes are significantly overrepresented compared to a background distribution. Correction for multiple testing (e.g., Benjamini-Hochberg method) is applied to reduce false discovery rates.

3. *Hierarchical grouping of results*: Due to Reactome’s hierarchical architecture, enriched pathways can be organized into broader functional categories (e.g., receptor signaling, lipid metabolism, T-cell activation), which facilitates biological interpretation.

4. *Visualization of enriched reactions*: Results can be presented as tree-structured diagrams, dot plots, or interactive pathway maps showing the number and proportion of genes involved in each enriched pathway.

5. *Result interpretation*: Key pathways with high statistical significance and broad gene coverage are analyzed to identify dominant cellular programs associated with each expression cluster.

Reactome pathway analysis serves as a powerful complement to GO and KEGG by not only identifying functional themes but also enabling the reconstruction of causal logic in cellular systems. Through its integration with external ontologies and curated evidence, Reactome improves the biological plausibility of functional interpretations and strengthens conclusions regarding the molecular roles of gene expression clusters.

FUNCTIONAL ENRICHMENT IN CYTOSCAPE: CLUEGO AND CLUEPEDIA

To enhance the interpretability of functional analysis results for clusters of gene expression profiles, visual representation of enrichment data plays a critical role. In this context, the Cytoscape platform serves as a powerful tool for the integration and visualization of bioinformatics data. Specifically, the *ClueGO* and *CluePedia* plugins enable the transformation of enrichment results into functionally grouped networks.

ClueGO is a Cytoscape plugin that integrates results from GO, KEGG, Reactome, and other biological databases into a unified graph-based structure. In these functional maps, each node represents a GO term or pathway, and edges indicate semantic similarity or gene set overlap between terms.

CluePedia extends *ClueGO* functionality by displaying gene–term and gene–gene interactions, and by integrating additional layers of information such as gene expression levels, fold changes, or correlation metrics. These are typically visualized through color-coded scales or heatmaps directly within the network.

The integration of *ClueGO* and *CluePedia* into the functional analysis pipeline overcomes the limitations of purely tabular representations by offering intuitive, interactive, and biologically coherent views. This facilitates comparison across clusters, identification of dominant biological themes, and the generation of hypotheses about gene coordination and functional modules.

The procedure for implementing functional enrichment analysis in Cytoscape is as follows.

1. *Genes ID list import*: For each gene expression cluster, a list of Entrez Gene IDs is generated and imported into Cytoscape through the *ClueGO* interface.

2. *Selection of functional databases*: The user specifies which annotation sources to include – GO (BP), MF, CC), KEGG, Reactome – providing a comprehensive view of gene function, localization, and involvement in signaling pathways.

3. *Enrichment parameters*: Statistical thresholds such as minimum gene count per term, p-value cutoffs, and multiple testing correction methods (e.g., Bonferroni or Benjamini-Hochberg) are configured. *ClueGO* also supports term grouping based on shared genes or biological themes.

4. *Network generation*: *ClueGO* automatically converts the enrichment results into a functional network, where nodes correspond to significant terms grouped by biological relevance (e.g., "mitotic cell cycle," "immune activation").

5. *Gene-to-term visualization*: *CluePedia* overlays gene-term relationships and optionally displays expression metrics, enabling dynamic exploration of gene-level contributions using color gradients or intensity coding.

6. *Interpretation*: Analysis of the resulting network allows for the identification of central functional nodes, clusters of related processes, and potentially regulatory genes. This facilitates a systems-level understanding of the functional architecture of each cluster.

In summary, the use of *ClueGO* and *CluePedia* significantly enhances biological interpretation by providing an interactive and structured visualization of the functional relationships among genes and biological terms. This approach is particularly useful for analyzing large clusters or when multi-parameter visualization is required for in-depth exploratory analysis.

EXPERIMENTAL DATA

This study utilized RNA-Seq gene expression data obtained from The Cancer Genome Atlas (TCGA) [18] via the Genomic Data Commons (GDC) data portal, using the *TCGAbiolinks* R package [19,20]. Gene expression quantification was performed using the workflow type “STAR– Counts”. The dataset included samples from 13 different cancer types. For each tumor type, samples were categorized as “Primary Tumor” or “Solid Tissue Normal” based on the sample_type metadata field.

Data preprocessing followed the methodology described in [4], which includes normalization, quality control, and removal of low-expression genes. In the final filtering stage, genes were selected based on Gene Ontology (GO) classification, retaining only unique genes associated with at least one of the GO domains: BP, MF, or CC. This ensured that only biologically interpretable genes were included in the downstream analysis.

As a result, 13 tumor-specific subsets were constructed, corresponding to the selected cancer types. Additionally, a fourteenth subset was generated by aggregating all non-cancerous samples into a unified class, representing "Solid Tissue Normal" controls. This yielded a 14-class classification structure, with each class corresponding to a specific tumor type or healthy control group.

Subsequent steps – such as variance-based gene filtering and quantile normalization – were performed using Bioconductor tools in R. The resulting gene expression matrix consisted of 6,310 biological samples (rows) and 18,564 genes (columns), representing a high-dimensional dataset for integrative machine learning and functional analysis. Fig. 1 presents the classification of the experimental data.

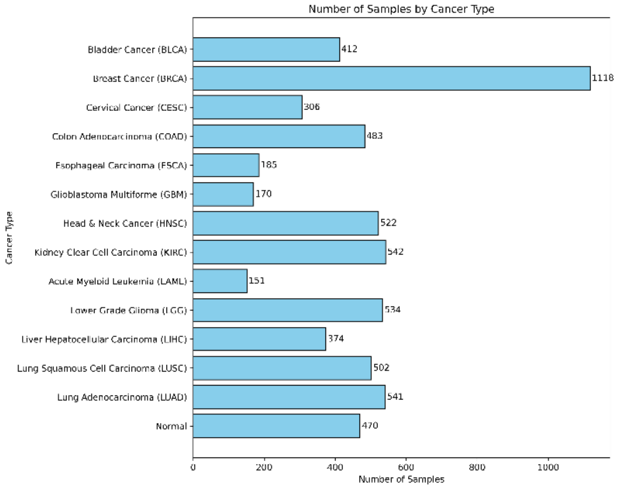

Fig. 1. **Clfssification of the experimental data**

Source: **compiled by the authors**

To explore the biological structure within this dataset, the gene expression matrix was clustered using a hybrid approach combining the Self-Organizing Tree Algorithm (SOTA) with consensus spectral clustering. This approach partitioned the expression profiles into four distinct clusters, each capturing a coherent gene expression pattern.

To validate the biological relevance of these clusters, a supervised classification was performed using a Random Forest algorithm within a stacked ensemble framework. Genes from each cluster were used as feature subsets to train the classifier. The model achieved perfect classification results across all evaluation metrics, correctly classifying 100% of the test samples, thereby confirming the discriminative power of the cluster-derived gene sets.

FUNCTIONAL COMPARISON OF ENRICHMENT STRATEGIES: GO, KEGG, AND REACTOME

To assess the biological validity of the identified gene expression clusters, we performed a comparative functional annotation using three complementary knowledge bases: GO, KEGG, and Reactome. Each of these resources offers a unique perspective – GO classifies gene function by process and localization, KEGG maps genes onto pathway diagrams, while Reactome represents biochemical events as interconnected reaction cascades. This integrated strategy enables cross-validation of functional coherence and improves interpretability for downstream applications such as classification models and feature selection in high-dimensional transcriptomic data.

Fig. 2–4 provide dot plot visualizations of the GO, KEGG, and Reactome enrichment results respectively, clearly demonstrating the functional divergence between clusters and the consistent enrichment across databases:

- Fig. 2: GO enrichment of clusters;
- Fig. 3: KEGG pathway enrichment;
- Fig. 4: Reactome reaction cascade enrichment.

The strong agreement among the GO, KEGG, and Reactome results validates the biological relevance of the clusters. Based on the analysis, the following functional consistencies were identified for each cluster:

Cluster 1 – Neural signaling and synaptic activity:

- *GO:* Enriched in terms such as regulation of synapse structure, postsynapse organization, and dendrite morphogenesis;
- *KEGG:* Highlighted calcium signaling, neuroactive ligand-receptor interaction, and glutamatergic/dopaminergic synapse signaling;
- *Reactome:* Confirmed relevance through neuronal system, transmission across chemical synapses, and synaptic protein interactions.

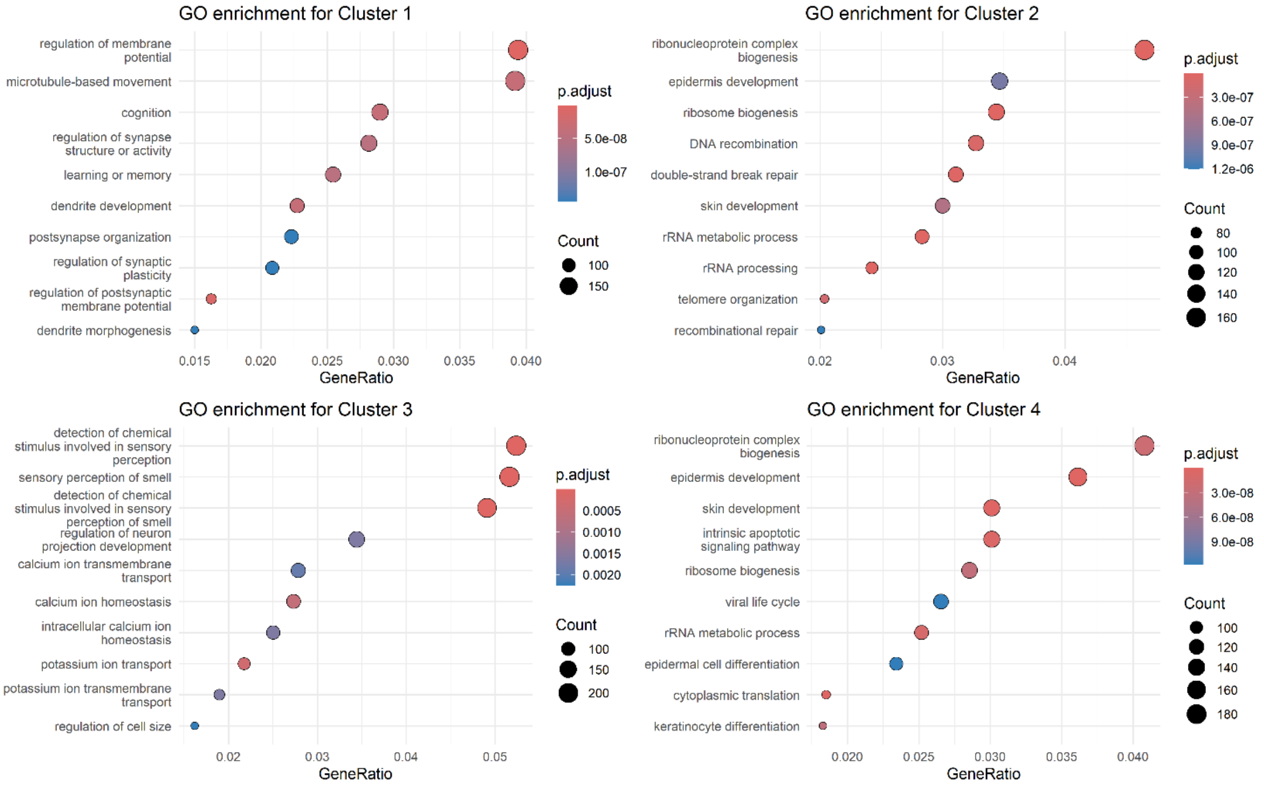

Fig. 2. **Comparative GO enrichment analysis across four gene expression clusters**
Source: **compiled by the authors**

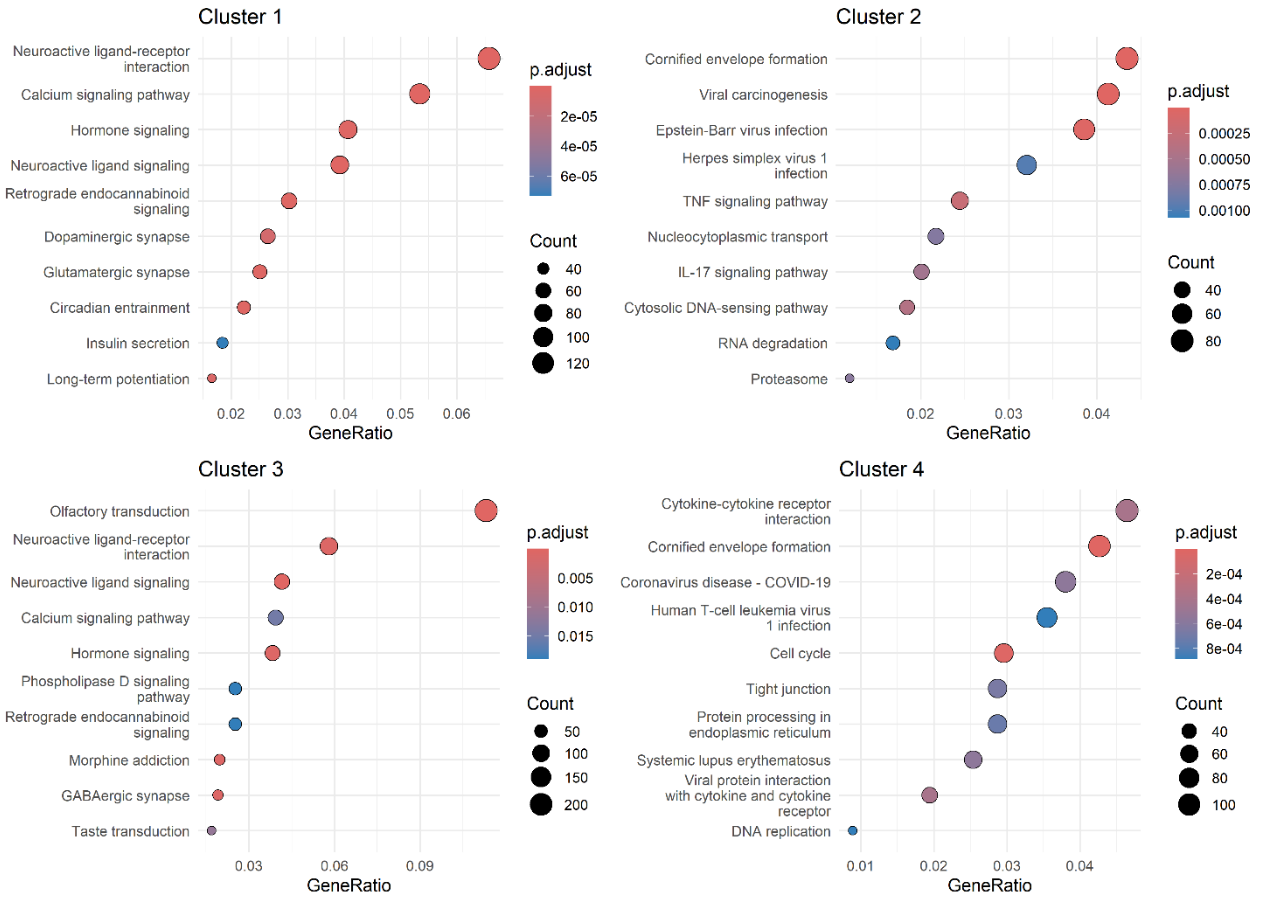

Fig. 3. **Kyoto Encyclopedia of Genes and Genomes enrichment of signaling and metabolic pathways across four gene expression clusters**

***Source:* compiled by the authors**

This consistent neural-specific signature suggests potential neuroendocrine reprogramming in certain tumor subtypes and underlines the diagnostic value of neural gene expression profiles.

Cluster 2 – Ribosomal biosynthesis and immune/viral response:

- *GO:* Strongly associated with ribosome biogenesis, ribonucleoprotein complex assembly, and telomere maintenance;
- *KEGG:* Revealed involvement in viral carcinogenesis, Epstein–Barr virus infection, and TNF signaling;
- *Reactome:* Indicated DNA replication, M phase progression, and rRNA processing.

These results reflect high transcriptional and translational activity and possible viral influence—common features in aggressive or immune-evasive cancers.

Cluster 3 – Olfactory and sensory perception:

- *GO:* Enriched in sensory perception of smell, detection of chemical stimulus, and ion transport;
- *KEGG:* Identified olfactory transduction, hormonal signaling, and neuroactive signaling;
- *Reactome:* Highlighted olfactory signaling pathway and olfactory receptor expression.

This cluster illustrates an unconventional but well-documented phenomenon of ectopic olfactory gene expression in epithelial tumors, particularly those of head and neck origin.

Cluster 4 – Protein synthesis, cytokine signaling, and antiviral response.

- *GO:* Captured keratinocyte and epidermal differentiation, and cytoplasmic translation;
- *KEGG:* Indicated cytokine-cytokine receptor interaction, cell cycle, and COVID-19-related immune pathways;
- *Reactome:* Emphasized translation, influenza infection, and rRNA processing.

This cluster aligns with highly proliferative tumors showing immune signaling involvement and elevated protein synthesis. Its features may serve as a predictive marker of immunogenicity and tumor aggressiveness.

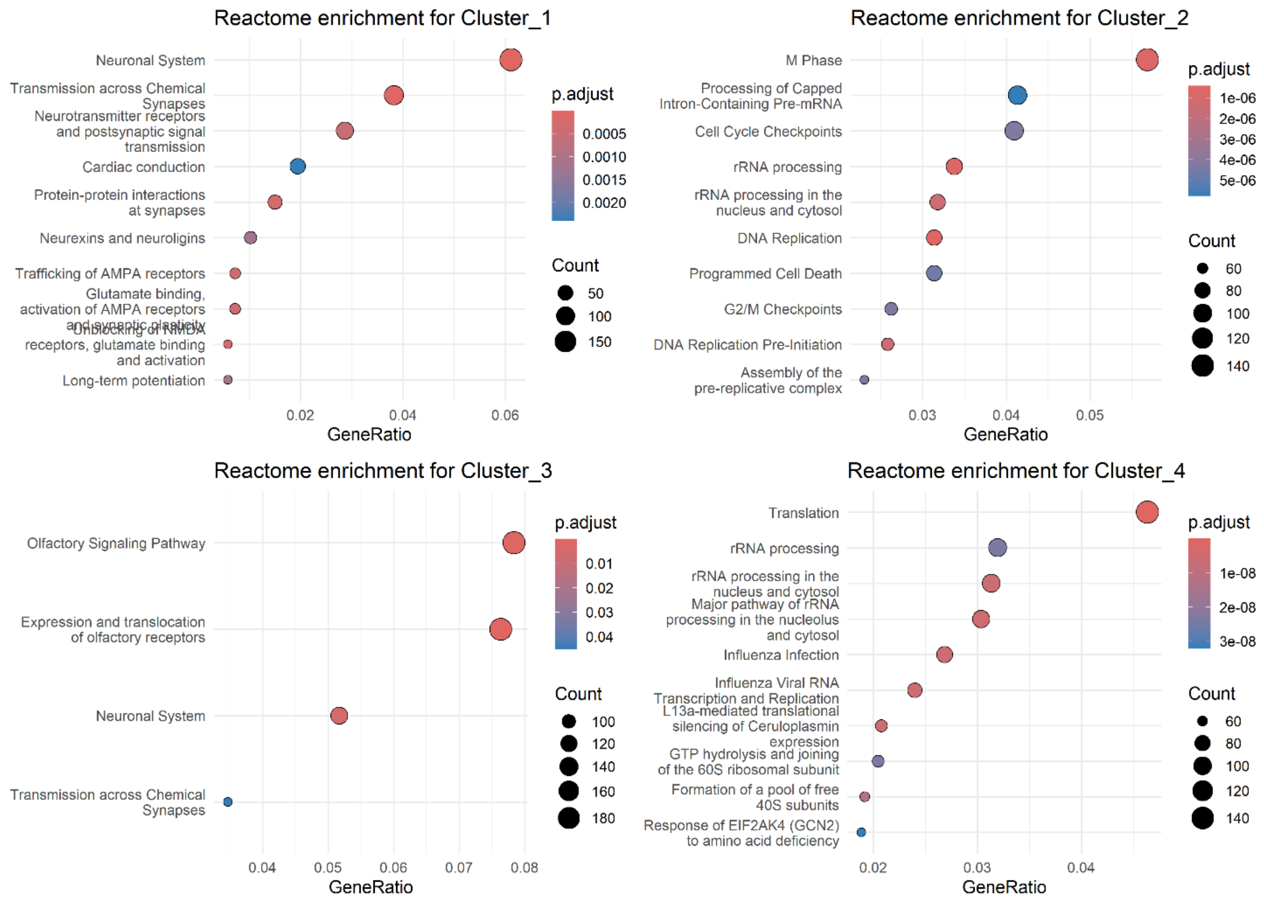

Fig. 4. **Reactome pathway enrichment of biological reactions across four gene expression clusters**

Source: **compiled by the authors**

It should be noted that this analysis confirms that each cluster represents a coherent and biologically distinct module, validated through multi-database functional annotation. The high degree of concordance across enrichment results supports the robustness of the applied SOTA combined with spectral consensus clustering method. Furthermore, the consistency of functional annotations facilitates the integration of biological knowledge into diagnostic rule generation, machine learning pipelines, and feature engineering for classification models – effectively linking transcriptomic insights with data-driven medical applications.

FUNCTIONAL PATHWAY NETWORKS OF GENE EXPRESSION CLUSTERS BASED ON KEGG: VISUAL MODELING IN CYTOSCAPE

Despite the informative value of GO and Reactome enrichment results, applying these databases to the full set of gene expression data within each cluster produced overly dense networks that hindered effective graphical interpretation. The excess complexity of the resulting networks obstructed the identification of functional modules, even when GO-term grouping was applied. To overcome this limitation, pathway enrichment analysis was performed separately for each gene expression cluster using KEGG as the primary knowledge base. This approach enabled the generation of more interpretable and compact network graphs, particularly in the context of signaling cascades and pathological processes. The modeling was performed using the *ClueGO* and *CluePedia* modules within the Cytoscape environment. Table 1 summarizes the configuration parameters used for KEGG-based network generation.

Based on the KEGG enrichment analysis, a separate functional network was constructed for each of the four clusters, as shown in Fig. 5–8.

Table 1. **ClueGO configuration parameters for KEGG-Based network generation**

Parameter	Value and Description
Knowledge Base	KEGG (335 pathways, updated: 25.05.2022). Only KEGG ontology was selected without overlap with other sources.
p-value Threshold	$p \leq 0.05$ (Benjamini-Hochberg correction with mid-P-values enabled). Only statistically significant pathways are displayed.
Minimum Gene Proportion	4.0% of genes per cluster to include KEGG pathway (filters weakly enriched terms).
Network Specificity	Medium level – balances detail and generalization.
Kappa Score	0.4 – minimum functional similarity threshold to define links between nodes.
Term Grouping	Enabled. Nodes are grouped by Kappa Score. The leading term is selected based on the lowest p-value.
Node Layout	Prefuse Force Directed Layout – an interactive physics-based model for intuitive spatial arrangement of nodes.

***Source:* compiled by the authors**

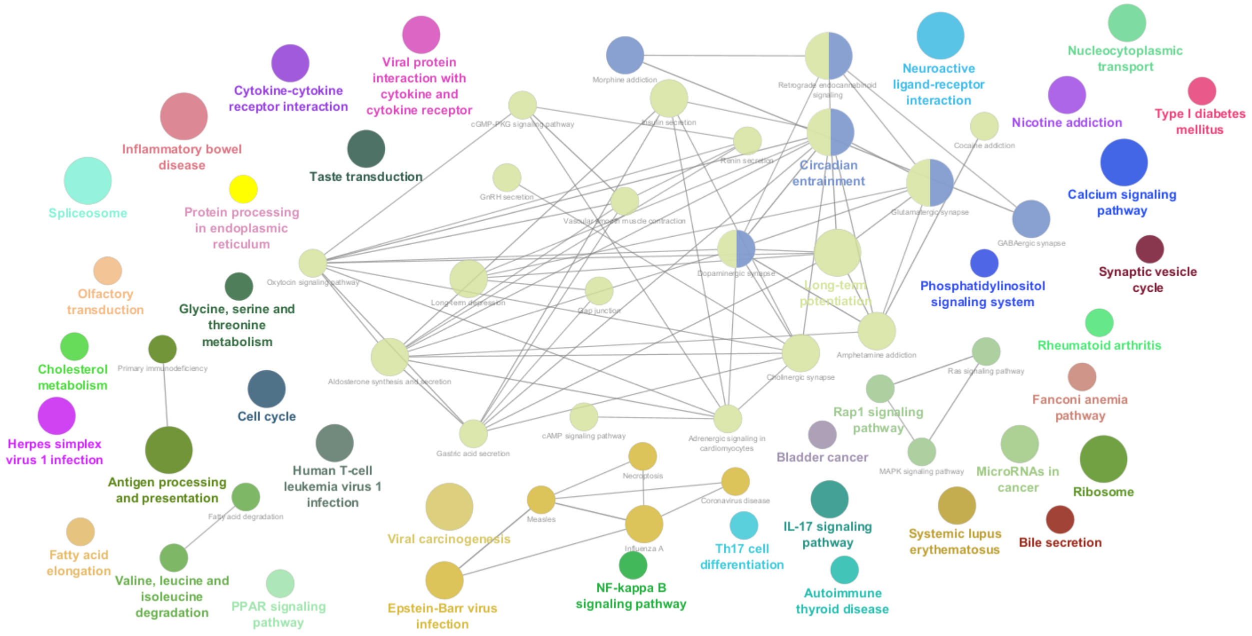
Viral protein interaction with cytokine and cytokine receptor
Cytokine-cytokine receptor interaction
Inflammatory bowel disease
Spliceosome
Protein processing in endoplasmic reticulum
Taste transduction
Olfactory transduction
Glycine, serine and threonine metabolism
Cholesterol metabolism
Primary immunodeficiency
Cell cycle
Herpes simplex virus 1 infection
Antigen processing and presentation
Human T-cell leukemia virus 1 infection
Fatty acid degradation
Fatty acid elongation
Valine, leucine and isoleucine degradation
PPAR signaling pathway
Viral carcinogenesis
Epstein-Barr virus infection
Morphine addiction
cGMP-PKG signaling pathway
Insulin secretion
Renin secretion
GnRH secretion
Vascular smooth muscle contraction
Oxytocin signaling pathway
Long-term depression
Gap junction
Aldosterone synthesis and secretion
Gastric acid secretion
cAMP signaling pathway
Adrenergic signaling in cardiomyocytes
Retrograde endocannabinoid signaling
Circadian entrainment
Glutamatergic synapse
Dopaminergic synapse
Long-term potentiation
Cholinergic synapse
Amphetamine addiction
Cocaine addiction
GABAergic synapse
Neuroactive ligand-receptor interaction
Nucleocytoplasmic transport
Nicotine addiction
Type I diabetes mellitus
Calcium signaling pathway
Phosphatidylinositol signaling system
Synaptic vesicle cycle
Rheumatoid arthritis
Ras signaling pathway
Fanconi anemia pathway
Rap1 signaling pathway
Bladder cancer
MAPK signaling pathway
MicroRNAs in cancer
Ribosome
Necroptosis
Coronavirus disease
Measles
Influenza A
NF-kappa B signaling pathway
Th17 cell differentiation
IL-17 signaling pathway
Autoimmune thyroid disease
Systemic lupus erythematosus
Bile secretion

Fig. 5. **KEGG-Based functional pathway network of gene expression cluster 1**

***Source:* compiled by the authors**

This cluster aligns with highly proliferative tumors showing immune signaling involvement and elevated protein synthesis. Its features may serve as a predictive marker of immunogenicity and tumor aggressiveness.

The analysis of KEGG-based functional networks reveals that each cluster forms a well-structured functional core. In the case of Cluster 1 (Fig. 5), the enriched pathways are associated with neurotransmission, circadian rhythms, and long-term potentiation. Key nodes include: Circadian entrainment, Dopaminergic synapse, and Long-term potentiation. Circadian entrainment holds a central position in the network, acting as a hub with numerous interactions. Its activation is crucial for regulating biological rhythms, and disruptions in circadian regulation – such as altered expression of core-clock genes – have been linked to tumor progression in hepatocellular carcinoma, breast cancer, and lung cancer. The presence of Dopaminergic synapse and Long-term potentiation nodes indicates involvement of neuroplasticity and neuromodulatory signaling mechanisms. In the oncological context, these may influence the tumor microenvironment, proliferation, and immune modulation. This functional hub is further supported by activation of Cholinergic synapse, Glutamatergic synapse, and cAMP signaling pathways, forming a neuro-signaling module relevant to neuro-oncological scenarios.

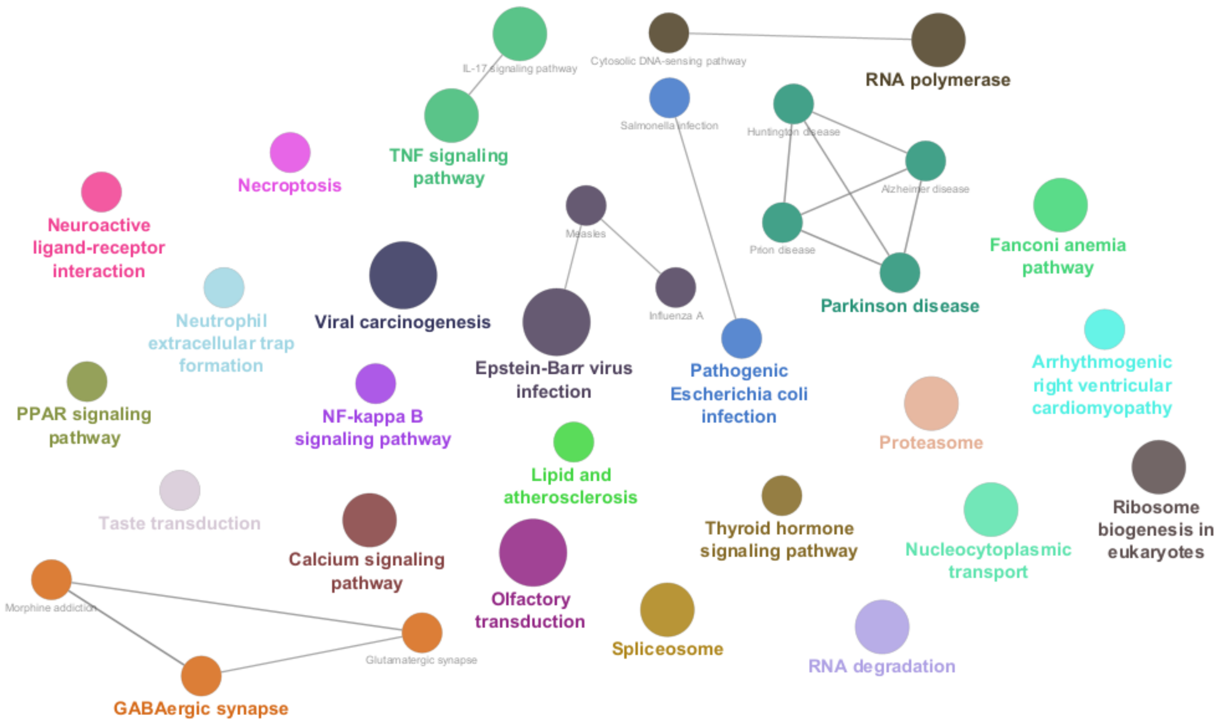
IL-17 signaling pathway
Cytosolic DNA-sensing pathway
RNA polymerase
TNF signaling pathway
Salmonella infection
Huntington disease
Necroptosis
Neuroactive ligand-receptor interaction
Alzheimer disease
Measles
Prion disease
Fanconi anemia pathway
Parkinson disease
Viral carcinogenesis
Neutrophil extracellular trap formation
Influenza A
Epstein-Barr virus infection
Pathogenic Escherichia coli infection
Arrhythmogenic right ventricular cardiomyopathy
PPAR signaling pathway
NF-kappa B signaling pathway
Proteasome
Lipid and atherosclerosis
Ribosome biogenesis in eukaryotes
Taste transduction
Thyroid hormone signaling pathway
Nucleocytoplasmic transport
Calcium signaling pathway
Olfactory transduction
Morphine addiction
Spliceosome
Glutamatergic synapse
RNA degradation
GABAergic synapse

Fig. 6. **KEGG-Based functional pathway network of gene expression cluster 2**

***Source:* compiled by the authors**

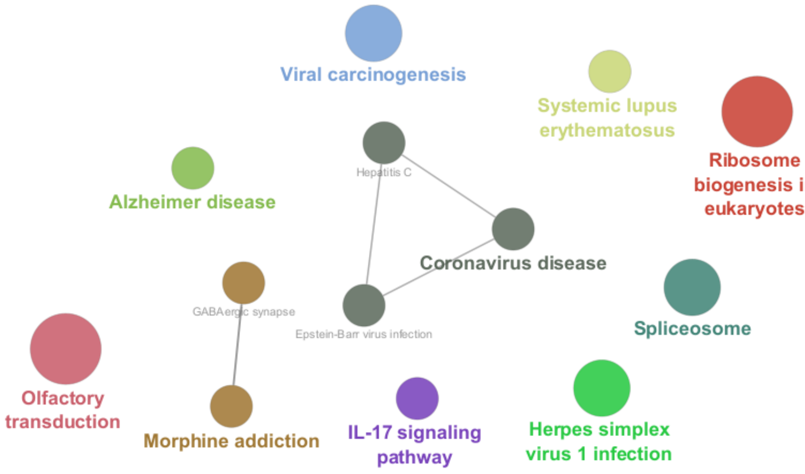

Fig. 7. **KEGG-Based functional pathway network of gene expression cluster 3**

***Source:* compiled by the authors**

The network derived from Cluster 2 (Fig. 6) highlights pathways associated with viral infection, inflammation, and DNA homeostasis disruption. Prominent nodes include: Epstein-Barr virus infection, NF-kappa B signaling pathway, and Parkinson disease. The presence of the Epstein-Barr virus infection pathway suggests activation of antiviral mechanisms and possible involvement in oncogenesis, particularly in lymphoproliferative disorders. NF-kappa B signaling plays a central role in immune response regulation, apoptosis, and cell survival, often activated in chronic inflammation driven by infection or mutations in cancer cells. The integration of Parkinson disease as a functional node may reflect mitochondrial dysfunction and oxidative stress – hallmarks of various tumor types. Altogether, this cluster's functional network points to immune-pathological and virus-induced signaling cascades potentially contributing to malignancy.

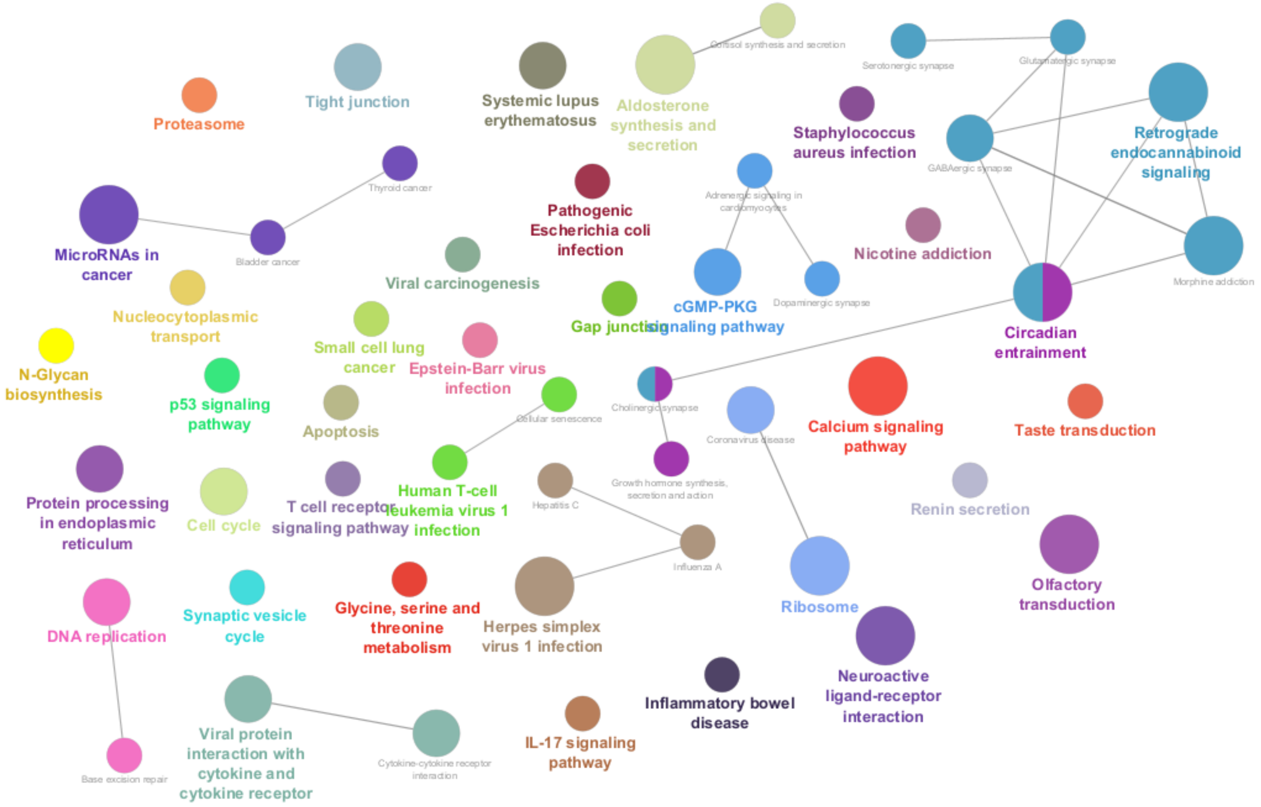
Proteasome
Tight junction
Systemic lupus erythematosus
Aldosterone synthesis and secretion
Cortisol synthesis and secretion
Serotonergic synapse
Glutamatergic synapse
Staphylococcus aureus infection
GABAergic synapse
Retrograde endocannabinoid signaling
Thyroid cancer
Pathogenic Escherichia coli infection
Adrenergic signaling in cardiomyocytes
MicroRNAs in cancer
Bladder cancer
Nicotine addiction
Morphine addiction
Viral carcinogenesis
Nucleocytoplasmic transport
Gap junction
cGMP-PKG signaling pathway
Dopaminergic synapse
Circadian entrainment
Small cell lung cancer
Epstein-Barr virus infection
N-Glycan biosynthesis
p53 signaling pathway
Apoptosis
Cellular senescence
Cholinergic synapse
Coronavirus disease
Calcium signaling pathway
Taste transduction
Human T-cell leukemia virus 1 infection
Hepatitis C
Growth hormone synthesis, secretion and action
Protein processing in endoplasmic reticulum
Cell cycle
T cell receptor signaling pathway
Renin secretion
Influenza A
Olfactory transduction
DNA replication
Synaptic vesicle cycle
Glycine, serine and threonine metabolism
Herpes simplex virus 1 infection
Ribosome
Neuroactive ligand-receptor interaction
Inflammatory bowel disease
IL-17 signaling pathway
Viral protein interaction with cytokine and cytokine receptor
Cytokine-cytokine receptor interaction
Base excision repair

Fig. 8. **KEGG-Based functional pathway network of gene expression cluster 4**

***Source:* compiled by the authors**

Cluster 3 (Fig. 7) is characterized by enrichment in infection-related, inflammatory, and metabolic pathways. Key functional nodes include: Coronavirus disease, IL-17 signaling pathway, and Ribosome biogenesis in eukaryotes. Coronavirus disease enrichment suggests activation of genes involved in antiviral defense, transcriptional activation, and cytokine regulation – features that align with immune alterations in tumor development. IL-17 signaling emphasizes the pro-inflammatory tumor microenvironment, stimulating cytokine and chemokine production to support angiogenesis and invasion. Ribosome biogenesis indicates high translational activity, typical of proliferating cancer cells. These processes together reflect the cluster's association with cellular activation, immune reactivity, and translational reprogramming – hallmarks of oncological progression.

The functional network for Cluster 4 (Fig. 8) reveals dominance of metabolic and signaling pathways strongly linked to cancer pathogenesis. Key nodes include: MicroRNAs in cancer, Calcium signaling pathway, and Pathogenic Escherichia coli infection. Enrichment in MicroRNAs in cancer points to the role of regulatory miRNAs in controlling oncogenes and tumor suppressors. Calcium signaling is critical for cell proliferation, apoptosis, and stress response, all pivotal in shaping the tumor microenvironment. The presence of Pathogenic Escherichia coli infection may reflect immune and inflammatory reactions triggered by compromised tissue barriers – common during oncogenic transformation. Overall, the cluster’s functional profile illustrates the interplay between metabolic, signaling, and immune mechanisms involved in malignant cell behavior.

The results of KEGG-based network analysis using *ClueGO* and *CluePedia* confirm the high quality of the identified gene expression cluster structure. Each cluster exhibits a functionally coherent core, with dominance of specific groups of signaling, metabolic, or immunological pathways. This supports the internal homogeneity and biological specialization of each gene group. Given that the dataset consists of transcriptomic profiles from patients with diverse cancer types, these distinct functional signatures are not random – they reflect key biological programs linked to specific pathophysiological conditions.

Thus, KEGG-based network analysis not only complements GO and Reactome enrichment but also confirms that gene expression clustering captures deeply rooted molecular mechanisms underlying tumor development. The resulting structure is both statistically sound and biologically relevant, demonstrating its potential utility in diagnostics, feature selection, and bioinformatic modeling of cancer heterogeneity.

CONCLUSIONS

This study presents an integrative approach for the post-clustering biological interpretation of gene expression data using statistical enrichment and network-based visualization. Gene expression profiles from patients with 13 types of cancer and a reference cohort were clustered using the Self-Organizing Tree Algorithm (SOTA) in combination with spectral consensus clustering. The resulting four-cluster structure was subsequently validated through multi-database functional enrichment analysis, including Gene Ontology, KEGG, and Reactome, as well as graph-based modeling in Cytoscape using ClueGO and CluePedia.

The findings demonstrate a high degree of concordance between enrichment sources, confirming that each cluster represents a biologically coherent module with distinct functional signatures. These signatures span neuronal signaling, immune response, translational regulation, and inflammation-associated pathways. Such clear functional specialization supports the reliability of the clustering method and underlines the internal consistency of the identified transcriptomic patterns.

Moreover, KEGG-based network analysis allowed for visual dissection of functional modules within each cluster, highlighting central pathway hubs such as Circadian entrainment, NF-kappa B signaling, and MicroRNAs in cancer. The functional coherence and specificity of the networks reinforce their potential utility for downstream applications in machine learning, diagnostic rule generation, and automated feature extraction.

From the perspective of information technologies, the proposed approach focuses on the interpretation stage, linking unsupervised clustering results with biological knowledge bases through statistical and network-based analysis. This contributes to interpretable modeling of transcriptomic heterogeneity and enhances the explainability of computational results.

Future research will concentrate on integrating these biologically validated clusters into classification pipelines, combining functional interpretation with predictive modeling to improve cancer subtype diagnosis and support personalized treatment strategies.

REFERENCES

1. Zhang, D. & Zhu, Y. “ECBN: Ensemble Clustering Based on Bayesian Network Inference for Single-Cell RNA-Seq Data”. *2020 39th Chinese Control Conference (CCC). IEEE*. 2020. p. 5884–5888. DOI: https://doi.org/10.23919/CCC50068.2020.9188589.

2. Burton, R. J., Cuff, S. M., Morgan, M. P., Artemiou, A. & Eberl, M. “GeoWaVe: geometric median clustering with weighted voting for ensemble clustering of cytometry data”. *Bioinformatics*. 2023; 39 (1): btac751. DOI: https://doi.org/10.1093/bioinformatics/btac751.

3. Petegrosso, R., Li, Z. & Kuang, R. “Machine learning and statistical methods for clustering single-cell RNA-sequencing data”. *Briefing in Bioinformatics*. 2020; 21 (4): 1209–1223. DOI: https://doi.org/10.1093/bib/bbz130.

4. Babichev, S., Yarema, O., Liakh, I. & Shumylo, N. “A Gene Ontology-Based Pipeline for Selecting Signicant Gene Subsets in Biomedical Applications”. *Applied Sciences*. 2025; 15: 4471. DOI: https://doi.org/10.3390/app15084471.

5. Saxena, R., Bishnoi, R. & Singla, D. “Gene Ontology: application and importance in functional annotation of the genomic data”. *Bioinformatics: Methods and Applications*. 2021. p. 145–157. DOI: https://doi.org/10.3233/978-1-60750-945-5-108.

6. Kanehisa, M., Furumichi, M., Sato, J., Matsuura, Y. &, Ishiguro-Watanabe, M. “KEGG: biological systems database as a model of the real world”. *Nucleic Acids Research*. 2025; 53 (D1): D672–D677. DOI: https://doi.org/10.1093/nar/gkae909.

7. Fabregat, A., Jupe, S., Matthews, L. et al. “The Reactome Pathway Knowledge base”. *Nucleic Acids Research*. 2018; 46 (D1): D649–D655. DOI: https://doi.org/10.1093/nar/gkx1132.

8. Shannon, P., Markiel, A., Ozier, O. et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. *Genome Research*. 2003; 13 (11): 2498–2504. DOI: https://doi.org/10.1101/gr.1239303.

9. Bindea, G., Mlecnik, B., Hackl, H. et al. “ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks”. *Bioinformatics*. 2009; 25 (8): 1091–1093. DOI: https://doi.org/10.1093/bioinformatics/btp101.

10. Bindea, G., Galon, J. & Mlecnik, B. "CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data". *Bioinformatics*. 2013; 29 (5): 661–663. DOI: https://doi.org/10.1093/bioinformatics/btt019.

11. Suzi, A., James, B., Seth, C. et al. "The Gene Ontology knowledge base in 2023". *Genetics*. 2023; 224 (1): iyad031. DOI: https://doi.org/10.1093/genetics/iyad031.

12. Babichev, S., Yarema, O., Liakh, I. & Honcharuk, A. "Integrative Approach to Gene Expression Data Analysis: Combining Biclustering Techniques with Gene Ontology". *Lecture Notes in Data Engineering, Computational Intelligence, and Decision-Making*. 2024; 1: 149–177. DOI: https://doi.org/10.1007/978-3-031-70959-3_8.

13. Babichev, S., Korobchynskyi, M., Rudenko, M. & Batenko, H. "Applying biclustering technique and gene ontology analysis for gene expression data processing". *CEUR Workshop Proceedings*. 2024; 3675: 14–28.

14. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics*. 2010; 26 (1): 139–140. DOI: https://doi.org/10.1093/bioinformatics/btp616.

15. Love, M. I., Huber ,W. & Anders, S. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome Biology*. 2014; 15: 550. DOI: https://doi.org/10.1186/s13059-014-0550-8.

16. Shin, M. G. & Pico, A. R. "Using published pathway gures in enrichment analysis and machine learning". *BMC Genomics*. 2023; 24: 713. DOI: https://doi.org/10.1186/s12864-023-09816-1.

17. Durinck, S., Spellman, P. T., Birney, E., Huber, W. "biomaRt: mapping identi_ers for the integration of genomic datasets with the R/Bioconductor package". *Bioinformatics*. 2009; 25 (4): 526–528. DOI: https://doi.org/10.1093/bioinformatics/btn505.

18. Xianyu, H., Zhenglin, W. & Qing, W. "Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: A multi-omics consensus ensemble clustering". *Biomedicine & pharmacotherapy*. 2021; 144: 112222. DOI: https://doi.org/10.1016/j.biopha.2021.112222.

19. Figueroa-Martinez, J., Saz-Navarro, D. M., Lopez-Fernandez, A. et al. "Computational ensemble gene co-expression networks for breast and prostate cancer biomarker identification". *Informatics*. 2024; 11: 14. DOI: https://doi.org/10.3390/informatics11020014.

20. Manganaro, L., Bianco, S., Bironzo, P. et al. "Consensus clustering methodology to improve molecular stratification in nsclc". *Scientific Reports*. 2023; 13: 7759. DOI: https://doi.org/10.1038/s41598-023-33954-x.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 30.07.2025
Received after revision 11.09.2025
Accepted 25.09.2025

DOI: https://doi.org/10.15276/aait.08.2025.16
УДК 004.048+004.094

Посткластерна інтерпретація даних експресії генів з використанням функціонального збагачення та мережевого аналізу

Ярема Олег Романович[1)]
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
Сенчишен Денис Олександрович[2)]
ORCID: https://orcid.org/0000-0002-4311-7095; oleh.yarema@lnu.edu.ua
Бабічев Сергій Анатолійович[2,3)]
ORCID: https://orcid.org/0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
[1)] Львівський національний університет імені Івана Франка, вул. Університетська, 1, Львів, 79000, Україна
[2)] Херсонський державний університет, вул. Шевченка, 14б, Сівка-Войнилівська, Івано-Франківська обл., 77311, Україна
[3)] Університет Яна Євангелісти Пуркіне в Усті-над-Лабем, Pasteurova 3632/15, 400 96 Усті-над-Лабем, Чеська Республіка

АНОТАЦІЯ

Кластеризація профілів експресії генів є ключовим методом для виявлення прихованих біологічних структур і диференціації підтипів захворювань у високорозмірних біомедичних наборах даних. Проте перехід від кластерних структур до біологічно значущих висновків потребує інтегративних аналітичних стратегій, що виходять за межі неконтрольованого навчання. У цьому дослідженні представлено новий інтегративний обчислювальний підхід, який робить акцент на посткластерній інтерпретації результатів шляхом поєднання статистичного функціонального збагачення з мережевим моделюванням**.** Кластери профілів експресії генів, попередньо виявлені у пацієнтів із різними типами раку, були проаналізовані з використанням баз даних Gene Ontology, Kyoto Encyclopedia of Genes and Genomes та Reactome. Збагачення виконувалося інструментом g:Profiler, що дозволило виявити статистично значущі біологічні процеси, молекулярні функції, клітинні компоненти та сигнальні шляхи в межах кожного кластера. Для візуалізації та поглибленої інтерпретації функціональних категорій застосовано програмне середовище Cytoscape з модулями *ClueGO* та *CluePedia*, які формують функціональні мережі та демонструють взаємозв’язки між генами і біологічними термінами. Новизна роботи полягає у системному поєднанні статистичних методів функціонального збагачення та графового представлення, що забезпечує як обчислювальну строгість, так і біологічну інтерпретованість**.** Запропонований підхід продемонстрував, що кожен кластер має відмінні функціональні підписи (імунна відповідь, метаболічні шляхи, репарація ДНК, регуляція клітинного циклу тощо), що підтверджує його здатність вилучати біологічно узгоджені знання зі складних транскриптомних даних. У підсумку дослідження пропонує інноваційну стратегію посткластерної інтерпретації, яка поєднує машинне навчання без учителя та функціональну геноміку. Такий підхід підвищує пояснюваність обчислювальних результатів і може бути застосований для стратифікації захворювань, відкриття біомаркерів та у персоналізованій медицині.

Ключові слова: обчислювальний аналіз біологічних даних; біоінформатика; інтегративний аналіз; дані експресії генів; посткластерна інтерпретація; функціональне збагачення; онтологія генів ; KEGG; reactome; cytoscape; мережевий аналіз

ABOUT THE AUTHORS

Oleg R. Yarema - Ph.D., Associate Professor,Department of Digital economics and Business Analytics, Ivan Franko National University of Lviv, 1, Universytetska Str. Lviv, 79000, Ukraine
ORCID: https://orcid.org/0000-0003-3736-4820; oleh.yarema@lnu.edu.ua. Scopus Author ID: 59250847800
***Research field*:** Deep Learning; data mining; machine learning; gene expression data processing; hybrid models; development of IT technologies

Ярема Олег Романович - кандидат економічних наук, доцент кафедри Цифрової економіки та бізнес-аналітики. Львівський національний університет імені Івана Франка, вул. Університетська, 1. Львів, 79000, Україна

Denys O. Senchyshen - PhD Student, Department of Computer Science and Software Engineering. Kherson State University, 14b, Shevchenko Str, Sivka-Voynylivska, Ivano-Frankivsk Oblast, 77311, Ukraine
ORCID: https://orcid.org/0000-0002-4311-7095; dsenchishen@ksu.ks.ua
***Research field*:** Data mining; machine learning; gene expression data processing; hybrid models

Сенчишен Денис Олександрович - аспірант кафедри Комп’ютерних наук та програмної інженерії. Херсонський державний університет, вул. Шевченка, 14б, Сівка-Войнилівська, Івано-Франківська обл., 77311, Україна

Sergii A. Babichev - Doctor of Engineering Sciences, Professor, Department of Informatics. Jan Evangelista Purkyně University in Ústí nad Labem, Pasteurova 3632/15, 400 96 Ústí nad Labem, Czech Republic
Professor of the Department of Physics, Kherson State University, 14b, Shevchenko Str, Sivka-Voynylivska, Ivano-Frankivsk Oblast, 77311, Ukraine
ORCID: https://orcid.org/0000-0001-6797-1467; sergii.babichev@ujep.cz. Scopus Author ID: 57189091127
***Research field*:** Deep Learning; data mining; machine learning; gene expression data processing; hybrid models; Internet of Things

Бабічев Сергій Анатолійович - доктор технічних наук, професор кафедри Інформатики. Університет Яна Євангелісти Пуркіне в Усті-над-Лабем, Pasteurova 3632/15, 400 96 Усті-над-Лабем, Чеська Республіка, професор кафедри Фізики. Херсонський державний університет, вул. Шевченка, 14б, Сівка-Войнилівська, Івано-Франківська обл., 77311, Україна