A Graphical Method for Identifying Gene Clusters from RNA Sequencing Data

Jake R. Patock
Data to Knowledge Lab
Rice University
Houston, USA
jp157@rice.edu

Rinki Ratnapriya
Department of Ophthalmology
Baylor College of Medicine
Houston, USA
rpriya@bcm.edu

Arko Barman
Data to Knowledge Lab
Rice University
Houston, USA
arko.barman@rice.edu

***Abstract*—The identification of disease-gene associations is instrumental in understanding the mechanisms of diseases and developing novel treatments. Besides identifying genes from RNA-Seq datasets, it is often necessary to identify gene clusters that have relationships with a disease. In this work, we propose a graph-based method for using an RNA-Seq dataset with known genes related to a disease and perform a robust clustering analysis to identify clusters of genes. Our method involves the construction of a gene co-expression network, followed by the computation of gene embeddings leveraging Node2Vec+, an algorithm applying weighted biased random walks and skipgram with negative sampling to compute node embeddings from undirected graphs with weighted edges. Finally, we perform spectral clustering to identify clusters of genes. All processes in our entire method are jointly optimized for stability, robustness, and optimality by applying Tree-structured Parzen Estimator. Our method was applied to an RNA-Seq dataset of known genes that have associations with Age-related Macular Degeneration (AMD). We also performed tests to validate and verify the robustness and statistical significance of our methods due to the stochastic nature of the involved processes. Our results show that our method is capable of generating consistent and robust clustering results. Our method can be seamlessly applied to other RNA-Seq datasets due to our process of joint optimization, ensuring the stability and optimality of the several steps in our method, including the construction of a gene co-expression network, computation of gene embeddings, and clustering of genes. Our work will aid in the discovery of natural structures in the RNA-Seq data, and understanding gene regulation and gene functions not just for AMD but for any disease in general.**

***Index Terms*—RNA-Seq, clustering, transcriptomics, gene co-expression networks, gene embeddings**

I. Introduction

Gene expression regulation has emerged as a key mechanism through which genetic variants mediate disease risk in humans [7]. Thus, studying gene expression data has resulted in the discovery of important links between diseases and genes in recent years [8], [14]. Technical advancements in recent years have allowed gene expression instrumentation to produce large amounts of data with lower costs on large samples [19]. Utilizing this gene expression data to discover relationships between genetic factors and phenotypic results across samples and populations has led to new insights on specific diseases and biochemical mechanisms [19]. However, given the size of the datasets and propensity for false positives in gene expression studies, using traditional methods like statistical hypothesis testing leads to noisy results, with additional work being needed to empirically validate the conclusions [2]. Given the current difficulties and the potential for translating gene expression findings into precision medicine, improvements in gene expression studies necessitate increased robustness [4].

Many of these difficulties pertaining to identifying the gene expression changes between controls and disease have been shown to be solvable by machine learning (ML) [19]. This potential, combined with increased computational resources, has allowed both classical ML and deep learning (DL) models to be used as effective tools for identifying relevant genes for diseases [19]. The ability of these models to capture complex interdependencies and abstract understandings (in the case of DL) has shown application in multiple gene expression studies in a range of populations [1], [14]. Model architectures employed include classic models such as linear regression, logistic regression, and random forest [14] and DL methods, such as graph-based neural networks (GNNs) and graph embedding techniques [2]. In particular, DL has the capability to learn interdependencies from genomic expression data and exhibit excellent performance [1].

A specific application area for utilizing DL in a gene expression application is in the field of Age-related Macular Degeneration (AMD), a neurodegenerative disease that can result in incurable blindness in patients [23]. It is a multifactorial disease caused by the cumulative impact of genetic factors, environmental factors, and advanced aging [9], [22]. While some genes related to AMD have already been discovered, a large proportion of heritability remains unexplained [22], suggesting many undiscovered relationships and dependencies between different genes and the disease. Finding new relationships between genes in AMD patients can lead to new drug targets, patient risk testing, and improved diagnosis for many patients [14]. Moreover, many known genes associated with AMD may contribute to the disease through distinct biological mechanisms. Grouping these known AMD-related genes into functionally similar clusters could help identify shared pathways or processes, thereby enabling the targeting of underlying disease mechanisms rather than focusing on individual genes. This clustering approach could also help identify previously unrecognized genes associated with AMD by uncovering similarities between undiscovered genes and

those already known to be involved in the disease [14], [25].

This study proposes a graph-based method to discover gene clusters among known genes for a disease from RNA sequencing (RNA-Seq) data. We applied our method to a bulk RNA-Seq dataset for AMD. First, we constructed a gene co-expression network and computed the gene embeddings from the network. Finally, the gene embeddings were clustered to produce a clustering of genes. The results of this workflow can be used to categorize already known genes into groups or clusters, possibly pertaining to similar mechanisms, and produce visual clustering results. Our approach of joint optimization of all involved processes ensures the stability, optimality, and robustness of the entire method.

In particular, we emphasize our novel approach of jointly optimizing all processes in our method, which enables us to generate the most optimal, robust, and stable clustering results as opposed to individually optimizing each step in most prior work. The final cost function of our joint optimization is based on the clustering performance in the overall process, which yields superior performance compared to separately optimizing co-expression network construction, embedding computation, and clustering that would guarantee optimal results for individual steps but not the overall process.

Our contributions can be summarized as follows:

- The design of a novel pipeline for clustering genes from RNA-Seq data, involving the construction of a co-expression network, computing gene embeddings from the network, and clustering genes based on their embeddings.
- The joint optimization of all involved steps in the pipeline for the best possible clustering results.
- The design of a set of tests to ensure the robustness, consistency, and statistical significance of the results.

II. Related Work

In recent years, ML models have been successfully applied to RNA-Seq datasets for different diseases, such as breast cancer, colorectal cancer, and other types of cancers, and AMD [5], [6], [11], [14], [15], [27]. Proposed methods and results show success in stratification of breast cancer into molecular subtypes using multiclass logistic regression [5], ML-based identification of differentially expressed genes and independently modulated gene sets [24], [29], identification of transcriptomic features that are commonly shared between cancer types [11], identification of diagnostic biomarkers for colorectal cancer [15], breast cancer prediction with transcriptome profiling [27], and prediction of gene regulatory networks [17]. A wide variety of both supervised and unsupervised classical ML algorithms, such as support vector machines (SVM), Extreme Gradient Boosting (XGBoost), self-organizing maps (SOM), different dimensionality reduction algorithms, such as t-distributed Stochastic Neighbor Embedding (tSNE) and PCA, and clustering algorithms, such as k-means and hierarchical clustering, have been used for the analysis of RNA-Seq datasets [6], [21]. The use of graph-based DL and ML algorithms, such as knowledge graphs [28] and Node2Vec [10], for transcriptomics is also being used in recent years for research in different diseases, including melanoma [30].

III. Methods

A. Overview

The purpose of this study was to perform a gene expression analysis to embed and cluster a group of 81 previously identified genes related to AMD [14]. These genes were identified by applying ML methods on an mRNAseq dataset, which contained the bulk gene expression levels for control and late-stage positive AMD patients. This set of AMD genes was experimentally validated to be statistically significant in separating the controls from late-stage AMD subjects [14].

An overview of our methods is shown in Figure 1. First, we constructed an undirected graph based on the co-expression data calculated from the dataset. To control for varying sequence depth errors, CS-CORE (cell-type-specific co-expressions), a statistical method, was used to derive the co-expression matrix from our subset of 81 genes [26]. The graph was then constructed between the genes (nodes) with edges being added to the graph where the co-expression coefficient's absolute value is greater than a specific threshold, τ, that can be tuned as a hyperparameter.

The resulting graph is then used for obtaining node embeddings (i.e., gene embeddings) in a latent space, by applying the `Node2Vec+` algorithm [13]. In `Node2Vec+`, the first step is to perform biased weighted random walks. The resulting anchor node inputs and neighboring nodes are then used as input and labels for Skip-Gram with Negative Sampling (SGNS) to produce the embedding for each node. These node embeddings were then clustered using spectral clustering, utilizing multimodality gap for optimizing the number of clusters [10].

Finally, the output clusters were evaluated to determine the statistical significance of the clustering of the 81 gene subset versus random samples of 81 genes from the entire set of genes. The entire pipeline, including the generation of the co-expression network, calculation of the gene embeddings in a latent space through `Node2Vec+`, and spectral clustering, was optimized jointly by applying Tree-structured Parzen Estimator [3]. We emphasize that our joint optimization strategy results in more consistent, robust, and optimal clustering results, as all processes are optimized jointly with a single cost function instead of optimizing each step separately with separate cost functions, which cannot guarantee an optimal final solution.

Since `Node2Vec+` (including biased weighted random walks) and spectral clustering are stochastic processes, we repeated these processes 100 times, leading to 100 different embeddings along with corresponding clusterings, to verify and validate the robustness and consistency of our methods.

B. Dataset

The dataset used in this work includes the complete mRNAseq data from $n_c = 105$ controls (Minnesota Grading System 1) and $n_s = 61$ subjects (Minnesota Grading System

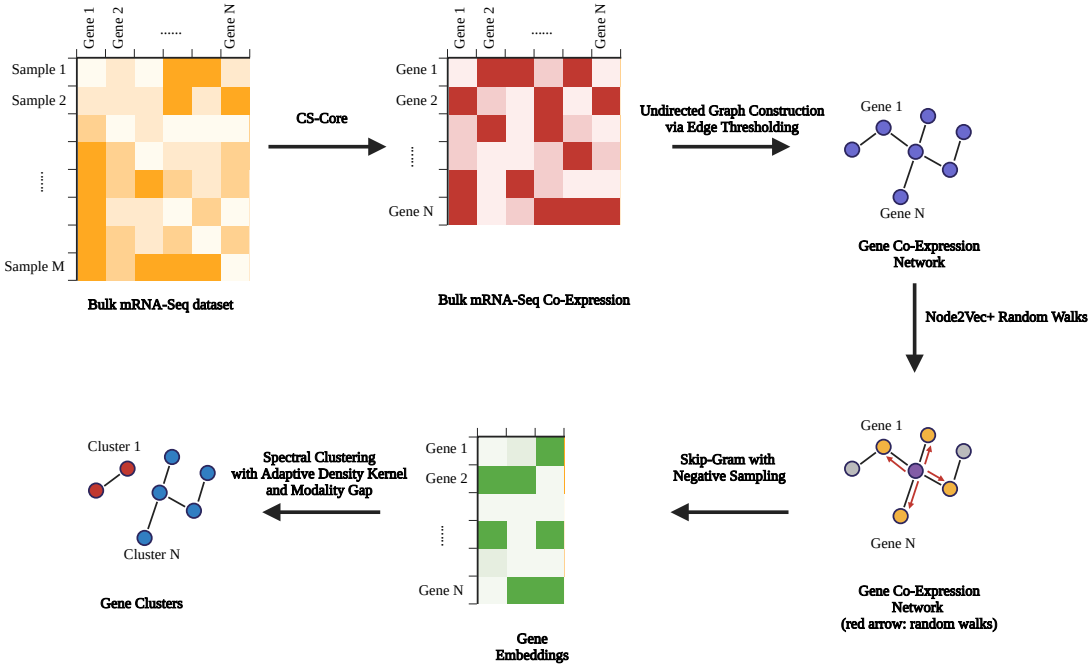

Fig. 1. Overview of our method used to process the subset of bulk mRNA-seq dataset comprising 81 AMD genes to generate a co-expression matrix using CS-CORE, generating a co-expression network by thresholding, embedding to a latent space using `Node2Vec+`, and finally applying spectral clustering.

4) with late-stage AMD [23]. In particular, for applying our entire pipeline, we used only a subset of 81 genes from this dataset that are related to AMD as previously identified using ML methods [14]. These 81 previously identified genes were validated to be statistically significant, and novel disease-gene associations were identified through the use of feature selection, XGBoost, and SHapley Additive exPlanations (SHAP) to create an interpretable ML pipeline to identify and discover disease-gene associations.

C. Gene Co-expression Network Construction from CS-Core

To calculate the co-expression levels within the dataset, CS-CORE was applied [26]. CS-CORE is a statistical method that accounts for varying sequence depth in genetic co-expression calculations within mRNAseq datasets [26]. This method leads to a more accurate estimation of co-expression between genes compared to other methods, such as the Pearson correlation coefficient [26]. CS-CORE was shown to have a much better ability to adjust for false positives, more accurate co-expression level estimation, and expression detection in both Alzheimer's disease and COVID-19 [26]. In our methods, we assumed that CS-CORE is applicable to bulk mRNAseq data in the absence of single-cell mRNAseq data.

From this co-expression matrix, a gene co-expression network was constructed, where the nodes of the graph were individual genes, with an undirected weighted edge added between a pair of nodes if the CS-CORE co-expression value for the pair was greater than a threshold, τ. This edge threshold hyperparameter, τ, was tuned in the optimization stage outlined in section III-F.

D. Graph Embedding Using Node2Vec+

From the gene co-expression network, we calculated node embeddings, i.e., gene embeddings, in a latent space using `Node2Vec+` [13]. `Node2Vec+` refines the classic `Node2Vec` algorithm for computing the node embeddings in a graph [10]. In the first stage of `Node2Vec`, an anchor node is chosen and random graph traversals are executed to select neighbor nodes that are accessible from the anchor node by performing weighted biased random walks (based on edge weights), and a list of visited nodes is recorded [10]. These anchor nodes are then used as inputs to a Skip-Gram with Negative Sampling (SGNS) model, with the corresponding neighborhood nodes being the output labels [10] to finally train the `Node2Vec` model to obtain node embeddings.

In `Node2Vec+`, edge weights are considered on this random walk through the hyperparameters – the return hyperparameter p, the in-out hyperparameter q, and a looseness hyperparameter γ [13]. These hyperparameters can bias the model towards moving down edges with higher weights or returning to previously traversed edges [13].

Overall, we optimized the following hyperparameters in `Node2Vec+` – return hyperparameter (p), in-out hyperparameter (q), length of each walk from an anchor node (W_L), embedding dimensionality (E), window size (W_S), and the number of negative samples per positive sample (N_s). The hyperparameter, γ, when set to 0, was shown to be robust and $\gamma = 0$ was used in our experiments [13] (refer to Section III-F for more details).

Overall, the refinement introduced by `Node2Vec+` improves the effectiveness of the second-order random walk of

`Node2Vec` by considering the edge weights of the graph and thus yields better node embeddings, especially for biological networks and datasets [13]. Thus, we chose `Node2Vec+` over the classic `Node2Vec` to embed the genes in the co-expression network into a latent node embedding space for the AMD gene dataset.

Three hyperparameters related to SGNS were optimized – the embedding dimension (E), window size (W_S), and number of negative sampling predictions (N_S) during the training of the SGNS. For fitting the SGNS model, it was trained for 100 epochs, and $32,768$ random walks were performed on the graph to generate the dataset for each model trial.

E. Spectral Clustering

Spectrum, a spectral clustering method for multi-omic datasets, was applied to the node embeddings computed using `Node2Vec+`. To calculate the affinity matrix of the embeddings, we applied the adaptive density-aware kernel designed in Spectrum [12]. The adaptive density-aware kernel defines the affinity between two nodes, s_i and s_j, as,

$$A_{ij} = \exp\left(-\frac{d^2(s_i, s_j)}{\sigma_i \sigma_j (CNN(s_i s_j) + 1}\right) \tag{1}$$

where

- $d(s_i, s_j) \in \mathbb{R}$ is the Euclidean distance between nodes s_i and s_j.
- $\sigma_i \in \mathbb{R}$ is the local scale parameter for the node s_i, calculated as the Euclidean distance between node s_i and its P-th nearest neighbor, $s_{i,P}$, i.e., $\sigma_i = d(s_i, s_{i,P})$,
- $\sigma_j \in \mathbb{R}$ is the local scale parameter for the node s_j, calculated as the Euclidean distance between node s_j and its P-th nearest neighbor, $s_{j,P}$, i.e., $\sigma_i = d(s_j, s_{j,P})$, and
- $CNN(s_i s_j)$ is the number of points in the intersection of the sets of S nearest neighbors of s_i and s_j.

The hyperparameters P and S were jointly optimized along with the other hyperparameters in our entire process. This method was found to enhance the similarities between two nodes in a graph if they shared similar connectivity.

To estimate the number of clusters, the last substantial multimodality gap heuristic was employed [12]. First a diagonal matrix, D, was constructed, where $D_{ii} = \sum_j A_{ij}$. The normalized graph Laplacian was then computed as,

$$L = D^{-1/2} A D^{-1/2} \tag{2}$$

We performed the eigendecomposition of L to obtain eigenvectors, $V = \{x_1, x_2, \ldots, x_N\}$, and eigenvalues, $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$. Multimodality was quantified by calculating the dip test statistics, $Z = \{z_1, z_2, \ldots, z_N\}$, for all eigenvectors. Finally, multimodality gaps were computed as $d_i = z_i - z_{i-1}$. The last substantial multimodality gap was calculated by iterating through d_i from the largest to the smallest corresponding eigenvalue using a search method that yields the optimal number of clusters, k^* [12].

For clustering, the top k^* eigenvectors were stacked to form a matrix, $X = [x_1\ x_2\ \ldots\ x_{k^*}] \in \mathbb{R}^{N \times k^*}$. A row-normalized matrix, $Y \in \mathbb{R}^{N \times k^*}$ was constructed, where,

$$Y_{ij} = \frac{X_{ij}}{\left(\sum_j X_{ij}^2\right)^{1/2}} \tag{3}$$

A Gaussian Mixture Model (GMM) was used to obtain k^* gene clusters, considering each row of Y as a data point.

F. Hyperparameter Optimization

All the steps in our methods – construction of the gene co-expression network from CS-CORE, Node2Vec+, and spectral clustering – involve hyperparameters that need to be tuned. Instead of optimizing each process individually, which might not result in the optimal clustering results, we chose to jointly optimize all the hyperparameters in our methods to maximize the density-based clustering validation index (DBCVI) [18]. DBCVI is suitable for non-convex clustering algorithms, such as spectral clustering, and has been shown to be effective for various applications and datasets [18].

The following processes and hyperparameters were jointly optimized in our experiments:

- Construction of the gene co-expression network from CS-CORE:
  - threshold of edge weights for connectivity between nodes, τ
- Computation of gene embeddings using `Node2Vec+`:
  - return hyperparameter, p
  - in-out hyperparameter, q
  - walk length, W_L
  - embedding dimensionality, E
  - window size, W_S
  - number of negative samples per positive sample, N_s
- Spectral Clustering:
  - nearest neighbor hyperparameter for calculating local scale parameters, P
  - number of nearest neighbors for calculating intersection, S

The complete search space for the hyperparameters is outlined in I. Tree-structured Parzen Estimator (TPE), a Bayesian optimization method, was used to optimize this set of hyperparameters using 256 hyperparameter samplings from the search space. Since the process of random walks in `Node2Vec+` is stochastic, every hyperparameter configuration was tried 8 times and averaged to produce the final DBCVI output, to average the effects of stochasticity on the results.

Our algorithm, including co-expression network construction, `Node2Vec+` for embedding calculation, and spectral clustering, along with our joint hyperparameter optimization strategy, is shown in Algorithm 1.

G. Testing Statistical Significance and Robustness

We performed experiments to validate the robustness and statistical significance of our clusterings for the AMD gene

Algorithm 1 Our Algorithm

Require: Bulk mRNA-seq dataset $X \in \mathbb{R}^{n \times g}$, hyperparameter search space $\mathcal{H}$, trials N

Ensure: Optimal hyperparameter configuration for clustering score

Initialize: TPE sampler; objective: maximize DBCVI
for trial $t = 1$ to N **do**
  Sample hyperparameters from TPE: $\{\tau, p, q, W_L, E, W_S, N_S, P, S\} \sim \text{TPE}(\mathcal{H})$
  Initialize DBCVI_Scores $\leftarrow [\,]$
  for repeat $r = 1$ to 8 **do**
    $\mathbf{C} \leftarrow$ CS-CORE(X)
    $G \leftarrow$ ConstructGraph$(\mathbf{C}, \tau)$
    $\mathbf{Z} \leftarrow$ `Node2Vec+`$(G, p, q, W_L, E, W_S, N_S)$
    Clusters $\leftarrow$ SpectralClustering$(\mathbf{Z}, P, S)$
    dbcvi $\leftarrow$ DBCVI(Clusters, $\mathbf{Z}$)
    Append dbcvi to DBCVI_scores
  end for
  $\overline{\text{DBCVI}}_t \leftarrow \frac{1}{8}\sum_{r=1}^{8}$ DBCVI_scores$[r]$
  Store $(\{\tau, p, q, W_L, E, W_S, N_h, S, P\}, \overline{\text{DBCVI}}_t)$
end for
Return: $\mathcal{H}^* \leftarrow \arg\max_{t \in [1,N]} \overline{\text{DBCVI}}_t$

TABLE I
HYPERPARAMETER SEARCH SPACE OPTIMIZED VIA TPE TO MAXIMIZE DBCVI OF OUTPUT CLUSTERINGS, WITH CORRESPONDING OPTIMAL VALUES.

Method	Hyperparameter	Search Space $\mathcal{H}$	Optimum
Graph Construction	τ	[0.3, 0.5]	0.4
Node2vec+	p	[0.01, 100.0]	0.35
	q	[0.01, 100.0]	11.66
	W_L	[10, 30]	20
	E	[2, 16]	10
	W_S	[4, 10]	10
	N_s	[5, 15]	7
Spectral Clustering	P	[3, 7]	7
	S	[P+2, P+4]	11

dataset. We show that our clusterings are significantly different from clusterings of 81 genes randomly sampled from the entire set of genes in the data. For this experiment, we generated 100 randomly sampled sets of 81 genes. Our methods were applied to these randomly-sampled sets – CS-CORE, co-expression network construction, `Node2Vec+`, and spectral clustering. Finally, to evaluate the clustering results, DBCVI was calculated for each of these 100 randomly-sampled sets. The distribution of DBCVI values from these randomly-sampled sets was then compared with the distribution of DBCVI values from 100 repetitions of clustering the known 81-gene AMD dataset using statistical tests. In particular, the Kolmogorov–Smirnov (K-S) test and the k-sample Anderson-Darling test were applied to test if the differences between these distributions were statistically significant.

Due to the inherent stochasticity in `Node2Vec+` and spectral clustering, it is imperative to test the robustness of our methods to ensure that the results are consistent to account for stochasticity. To test the robustness of our methods, we

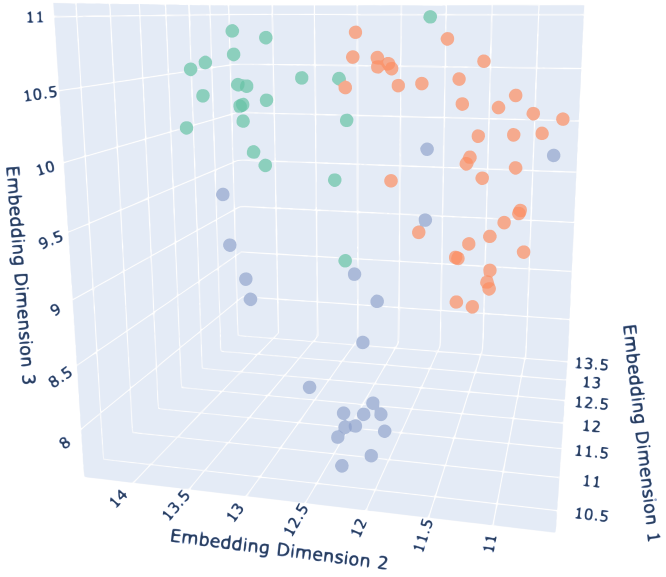

Fig. 2. Clustering results for the known 81-gene AMD subset. Embedding dimensionality is reduced from 10 to 3 using UMAP for visualization. (An interactive HTML visualization is available in our code repository.)

performed 100 repetitions of our entire method of clustering. Every pair of clustering results in these 100 repetitions was compared with each other using the adjusted mutual information (AMI) score to evaluate their agreement [20]. AMI has been shown to be the ideal metric in the presence of small clusters and unbalanced cluster sizes. The range of values for AMI is -1 to 1.

IV. RESULTS

Joint hyperparameter optimization was performed for all hyperparameters using 256 samplings from the hyperparameter space to maximize DBCVI for the clustering results. The optimal hyperparameters are shown in Table I. This optimal set of hyperparameters yielded a DBCVI score of 0.99 over the 8 clustering trials per hyperparameter configuration. We note that the optimal gene embedding dimensionality was 10.

After optimization, this optimal set of hyperparameters was then used to embed and cluster the dataset 100 times to perform statistical significance and robustness tests. The average DBCVI score for these 100 repetitions was 0.95, showing that each trial resulted in well-formed clusters. In our robustness testing, the average AMI between all pairs of clustering results was calculated to be 0.49, with a variance of 0.022. Hence, the 100 clustering repetitions show moderately to highly consistent agreement in the clustering of genes, especially for a small dataset with the possible presence of noise.

For visualization, the dimensional gene embeddings (of optimal dimensionality 10) were reduced to 3 dimensions using Uniform Manifold Approximation and Projection (UMAP) [16]. These low-dimensional embeddings were then plotted to produce interpretable results for AMD medical professionals

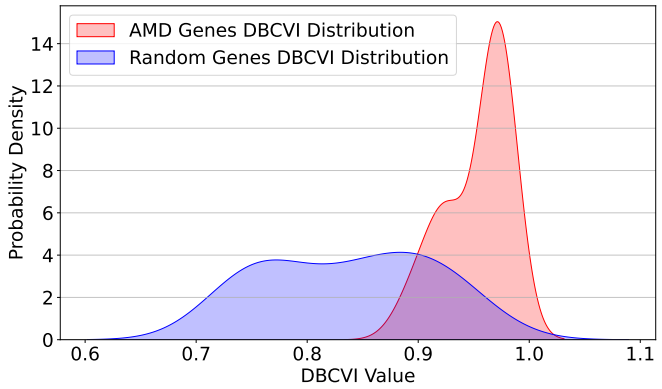

Fig. 3. Distributions of DBCVI values for 100 clustering repetitions for the known 81-gene subset for AMD and 100 samples of 81 randomly sampled genes.

to utilize in their research. The clustering visualization is shown in Figure 2.

In our statistical significance testing, the distribution of the DBCVI values for the 81-gene known subset of AMD genes was compared with that of 100 randomly sampled subsets of 81 genes from the entire set of genes. While the DBCVI for the known 81-gene subset yielded an average DBCVI of 0.95, the average DBCVI for the randomly sampled genes was 0.84. These distributions of the DBCVI values are shown in Figure 3. K-S test performed to validate that the differences between these distributions are statistically significant yielded $p = 2.68 \times 10^{-25}$. We also applied the k-sample Anderson-Darling test, a non-parametric test, to check if our clustering results were significantly different from the clusterings of the randomly sampled gene subsets, which yielded $p < 0.001$. Thus, our methods, when applied to the known 81-gene subset, yield significantly better clustering results than a randomly selected subset of 81 genes.

V. Discussion

Our method of embedding and clustering the bulk mRNAseq dataset showed robust, statistically significant results. The AMI score indicated a moderate to high level of agreement between clustering results. Considering the high level of noise typically present in transcriptomic datasets, the AMI scores show the robustness of our methods. We note that we employed our methods on a dataset of limited sample size, and using larger datasets is expected to lead to more consistent results and higher AMI.

We also noted that a higher number of random walks results in higher AMI. In the presence of more computing availability, the number of walks could continue to be scaled up to increase AMI further. Moreover, increasing the random walks performed on the graph will allow larger datasets as inputs to SGNS, leading to higher AMI. This hyperparameter can be dynamically adjusted based on the dataset on which this method is used.

The use of CS-CORE instead of Pearson correlation generated a more refined construction of gene co-expression network, resulting in better gene embeddings with the application of `Node2Vec+`, which in turn, resulted in more robust clustering results. Finally, spectral clustering leveraging the adaptive density kernel with modular gap metric showed robustness in consistently clustering these gene embeddings. Finding consistent clustering has been a long-standing challenge in various transcriptomics methods. Our method showed promise in providing a solution for our dataset and can possibly be extended to other transcriptomic analyses and datasets.

Our method is highly versatile due to the joint optimization of hyperparameters for all the steps involved in the entire process and can easily be applied to other datasets. This versatility and robustness allow our model to be applied to other datasets and diseases for discovering gene clusterings. Further, our method is computationally feasible without the use of GPUs or other computationally expensive hardware, making it widely accessible to researchers without any barriers related to computational capabilities and the availability of specialized hardware. Overall, our method is flexible, accessible, scalable, and generalizable.

Employing our method on more diverse mRNAseq datasets would be helpful for future research on other diseases. Our method can be applied to any available mRNAseq dataset. For future work, testing our method on a synthetic dataset with known ground-truth labels would allow for a more rigorous evaluation of its ability to recover the true information structure within the data. This approach would validate the results independently, complementing our experiments to test robustness and statistical significance.

VI. Conclusion

In this work, we proposed a comprehensive pipeline for discovering gene clusters by constructing a gene co-expression network, computing gene embeddings, and finally clustering genes in an mRNA-seq dataset for AMD. Our approach integrates multiple stages to ensure robust results. First, we apply CS-CORE to correct for expression errors and normalize the data. Next, we construct a gene co-expression network and calculate gene embeddings by applying `Node2Vec+`. Finally, we perform spectral clustering to cluster genes and identify gene clusters that have a potential relationship with the disease.

The workflow demonstrates high adaptability, as it can be tuned to the specific characteristics of any given dataset through joint hyperparameter optimization and consistently identifies cluster patterns across multiple repetitions, even though it has several stochastic components. Our results indicate that our method is flexible and robust in capturing meaningful biological structure in transcriptomic data. Notably, our method identifies gene clusters that experimental molecular geneticists can directly validate, interpret, and leverage for future research. These clusters can inform experimental design, guide hypothesis design, and be incorporated into existing laboratory workflows, ultimately supporting decision-making in studies of gene function and disease mechanisms.

Our code and results are available at https://github.com/arkobarman/graphical_gene_clustering_RNASeq.

REFERENCES

[1] Wardah S Alharbi and Mamoon Rashid. A review of deep learning applications in human genomics using next-generation sequencing data. *Human Genomics*, 16(1):26, 2022.

[2] Sezin Kircali Ata, Min Wu, Yuan Fang, Le Ou-Yang, Chee Keong Kwoh, and Xiao-Li Li. Recent advances in network-based methods for disease gene prediction. *Briefings in bioinformatics*, 22(4):bbaa303, 2021.

[3] James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123. PMLR, February 2013. ISSN: 1938-7228.

[4] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Paul R. Burton, Dan Davison, Peter Donnelly, Doug Easton, David Evans, Hin-Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C. A. Spencer, Martin D. Tobin, Lon R. Cardon, David G. Clayton, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S. Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, John A. Todd, Willem H. Ouwehand, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, Nick Craddock, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Paul R. Burton, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, Martin D. Tobin, John R. Thompson, Nilesh J. Samani, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, David G. Clayton, G. Mark Lathrop, John Connell, Anna Dominiczak, Nilesh J. Samani, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, , The Biologics in RA Genetics Genomics (BRAGGS), Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert, Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P. M. Symmons, Wendy Thomson, Jane Worthington, David G. Clayton, David B. Dunger, Sarah Nutland, Helen E. Stevens, Neil M. Walker, Barry Widmer, John A. Todd, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R. B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Mark I. McCarthy, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V. S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C. L. Gough, Sheila Seal, Breast Cancer Susceptibility Collaboration (UK), Michael R. Stratton, Nazneen Rahman, Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Melanie Newport, Giorgio Sirugo, Kirk A. Rockett, Dominic P. Kwiatkowski, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J. R. Ghori, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker, Claire Widden, David Withers, Panos Deloukas, Hin-Tak Leung, Sarah Nutland, Helen E. Stevens, Neil M. Walker, John A. Todd, Doug Easton, David G. Clayton, Paul R. Burton, Martin D. Tobin, Jeffrey C. Barrett, David Evans, Andrew P. Morris, Lon R. Cardon, Niall J. Cardin, Dan Davison, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Jonathan L. Marchini, Chris C. A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, Peter Donnelly, David Bentley, Matthew A. Brown, Lon R. Cardon, Mark Caulfield, David G. Clayton, Alistair Compston, Nick Craddock, Panos Deloukas, Peter Donnelly, Martin Farrall, Stephen C. L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V. S. Hill, Dominic P. Kwiatkowski, Christopher G. Mathew, Mark I. McCarthy, Willem H. Ouwehand, Miles Parkes, Marcus Pembrey, Nazneen Rahman, Nilesh J. Samani, Michael R. Stratton, John A. Todd, Jane Worthington, The Wellcome Trust Case Control Consortium, Management Committee, Data and Analysis Committee, UK Blood Services and University of Cambridge Controls, 1958 Birth Cohort Controls, Bipolar Disorder, Coronary Artery Disease, Crohn's Disease, Hypertension, Rheumatoid Arthritis, Type 1 Diabetes, Type 2 Diabetes, Tuberculosis, Ankylosing Spondylitis, Autoimmune Thyroid Disease, Breast Cancer, Multiple Sclerosis, Gambian Controls, Genotyping DNA, Data QC and Informatics, Statistics, and Primary Investigators. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.

[5] Silvia Cascianelli, Ivan Molineris, Claudio Isella, Marco Masseroli, and Enzo Medico. Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Scientific Reports*, 10(1):14071, August 2020. Publisher: Nature Publishing Group.

[6] Yuning Cheng, Si-Mei Xu, Kristina Santucci, Grace Lindner, and Michael Janitz. Machine learning and related approaches in transcriptomics. *Biochemical and Biophysical Research Communications*, 724:150225, 2024.

[7] The GTEx Consortium, François Aguet, Shankara Anand, Kristin G. Ardlie, Stacey Gabriel, Gad A. Getz, Aaron Graubert, Kane Hadley, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, Xiao Li, Daniel G. MacArthur, Samuel R. Meier, Jared L. Nedzel, Duyen T. Nguyen, Ayellet V. Segrè, Ellen Todres, Brunilda Balliu, Alvaro N. Barbeira, Alexis Battle, Rodrigo Bonazzola, Andrew Brown, Christopher D. Brown, Stephane E. Castel, Donald F. Conrad, Daniel J. Cotter, Nancy Cox, Sayantan Das, Olivia M. de Goede, Emmanouil T. Dermitzakis, Jonah Einson, Barbara E. Engelhardt, Eleazar Eskin, Tiffany Y. Eulalio, Nicole M. Ferraro, Elise D. Flynn, Laure Fresard, Eric R. Gamazon, Diego Garrido-Martín, Nicole R. Gay, Michael J. Gloudemans, Roderic Guigó, Andrew R. Hame, Yuan He, Paul J. Hoffman, Farhad Hormozdiari, Lei Hou, Hae Kyung Im, Brian Jo, Silva Kasela, Manolis Kellis, Sarah Kim-Hellmuth, Alan Kwong, Tuuli Lappalainen, Xin Li, Yanyu Liang, Serghei Mangul, Pejman Mohammadi, Stephen B. Montgomery, Manuel Muñoz-Aguirre, Daniel C. Nachun, Andrew B. Nobel, Meritxell Oliva, YoSon Park, Yongjin Park, Princy Parsana, Abhiram S. Rao, Ferran Reverter, John M. Rouhana, Chiara Sabatti, Ashis Saha, Matthew Stephens, Barbara E. Stranger, Benjamin J. Strober, Nicole A. Teran, Ana Viñuela, Gao Wang, Xiaoquan Wen, Fred Wright, Valentin Wucher, Yuxin Zou, Pedro G. Ferreira, Gen Li, Marta Melé, Esti Yeger-Lotem, Mary E. Barcus, Debra Bradbury, Tanya Krubit, Jeffrey A. McLean, Liqun Qi, Karna Robinson, Nancy V. Roche, Anna M. Smith, Leslie Sobin, David E. Tabor, Anita Undale, Jason Bridge, Lori E. Brigham, Barbara A. Foster, Bryan M. Gillard, Richard Hasz, Marcus Hunter, Christopher Johns, Mark Johnson, Ellen Karasik, Gene Kopen, William F. Leinweber, Alisa McDonald, Michael T. Moser, Kevin Myer, Kimberley D. Ramsey, Brian Roe, Saboor Shad, Jeffrey A. Thomas, Gary Walters, Michael Washington, Joseph Wheeler, Scott D. Jewell, Daniel C. Rohrer, Dana R. Valley, David A. Davis, Deborah C. Mash, Philip A. Branton, Laura K. Barker, Heather M. Gardiner, Maghboeba Mosavel, Laura A. Siminoff, Paul Flicek, Maximilian Haeussler, Thomas Juettemann, W. James Kent, Christopher M. Lee, Conner C. Powell, Kate R. Rosenbloom, Magali Ruffier, Dan Sheppard, Kieron Taylor, Stephen J. Trevanion, Daniel R. Zerbino, Nathan S. Abell, Joshua Akey, Lin Chen, Kathryn Demanelis, Jennifer A. Doherty, Andrew P. Feinberg, Kasper D. Hansen, Peter F. Hickey, Farzana Jasmine, Lihua Jiang, Rajinder Kaul, Muhammad G. Kibriya, Jin Billy Li, Qin Li, Shin Lin, Sandra E. Linder, Brandon L. Pierce, Lindsay F. Rizzardi, Andrew D. Skol, Kevin S. Smith, Michael Snyder, John Stamatoyannopoulos, Hua Tang, Meng Wang, Latarsha J. Carithers, Ping Guan, Susan E. Koester, A. Roger Little, Helen M. Moore, Concepcion R. Nierras, Abhi K. Rao, Jimmie B. Vaught, and Simona Volpi. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[8] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt,

and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, March 2009. Publisher: Nature Publishing Group.

[9] Bryan R. Gorman, Georgios Voloudakis, Robert P. Igo, Tyler Kinzy, Christopher W. Halladay, Tim B. Bigdeli, Biao Zeng, Sanan Venkatesh, Jessica N. Cooke Bailey, Dana C. Crawford, Kyriacos Markianos, Frederick Dong, Patrick A. Schreiner, Wen Zhang, Tamer Hadi, Matthew D. Anger, Amy Stockwell, Ronald B. Melles, Jie Yin, Hélène Choquet, Rebecca Kaye, Karina Patasova, Praveen J. Patel, Brian L. Yaspan, Eric Jorgenson, Pirro G. Hysi, Andrew J. Lotery, J. Michael Gaziano, Philip S. Tsao, Steven J. Fliesler, Jack M. Sullivan, Paul B. Greenberg, Wen-Chih Wu, Themistocles L. Assimes, Saiju Pyarajan, Panos Roussos, Neal S. Peachey, and Sudha K. Iyengar. Genome-wide association analyses identify distinct genetic architectures for age-related macular degeneration across ancestries. *Nature Genetics*, 56(12):2659–2671, December 2024. Publisher: Nature Publishing Group.

[10] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[11] Anupama Jha, Mathieu Quesnel-Vallières, David Wang, Andrei Thomas-Tikhonenko, Kristen W. Lynch, and Yoseph Barash. Identifying common transcriptome signatures of cancer by interpreting deep learning models. *Genome Biology*, 23(1):117, May 2022.

[12] Christopher R John, David Watson, Michael R Barnes, Costantino Pitzalis, and Myles J Lewis. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, 36(4):1159–1166, 2020.

[13] Renming Liu, Matthew Hirn, and Arjun Krishnan. Accurately modeling biased random walks on weighted networks using node2vec+. *Bioinformatics*, 39(1):btad047, 01 2023.

[14] Khang Ma, Hosei Nakajima, Nipa Basak, Arko Barman, and Rinki Ratnapriya. Integrating explainable machine learning and transcriptomics data reveals cell-type specific immune signatures underlying macular degeneration. *npj Genomic Medicine*, 10(1):48, 2025.

[15] Neha Shree Maurya, Sandeep Kushwaha, Aakash Chawade, and Ashutosh Mani. Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer. *Scientific Reports*, 11(1):14304, July 2021. Publisher: Nature Publishing Group.

[16] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[17] Keiichi Mochida, Satoru Koda, Komaki Inoue, and Ryuei Nishii. Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets. *Frontiers in Plant Science*, 9, November 2018. Publisher: Frontiers.

[18] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM, 2014.

[19] Giulia Muzio, Leslie O'Bray, and Karsten Borgwardt. Biological network analysis with deep learning. *Briefings in bioinformatics*, 22(2):1515–1530, 2021.

[20] Vinh Nguyen, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *International Conference on Machine Learning 2009*, pages 1073–1080. Association for Computing Machinery (ACM), 2009.

[21] Raphael Petegrosso, Zhuliu Li, and Rui Kuang. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Briefings in Bioinformatics*, 21(4):1209–1223, June 2019. _eprint: https://academic.oup.com/bib/article-pdf/21/4/1209/33584388/bbz063.pdf.

[22] R Ratnapriya and E Y Chew. Age-related macular degeneration—clinical review and genetics update. *Clinical Genetics*, 84(2):160–166, 2013.

[23] Rinki Ratnapriya, Olukayode A Sosina, Margaret R Starostik, Madeline Kwicklis, Rebecca J Kapphahn, Lars G Fritsche, Ashley Walton, Marios Arvanitis, Linn Gieser, Alexandra Pietraszkiewicz, et al. Retinal transcriptome and eqtl analyses identify genes associated with age-related macular degeneration. *Nature genetics*, 51(4):606–610, 2019.

[24] Kevin Rychel, Anand V. Sastry, and Bernhard O. Palsson. Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. *Nature Communications*, 11(1):6338, December 2020. Publisher: Nature Publishing Group.

[25] Ajeet Singh and Rinki Ratnapriya. Integration of multiomic data identifies core-module of inherited-retinal diseases. *Human Molecular Genetics*, 34(5):454–465, January 2025. _eprint: https://academic.oup.com/hmg/article-pdf/34/5/454/61405463/ddaf001.pdf.

[26] Chang Su, Zichun Xu, Xinning Shan, Biao Cai, Hongyu Zhao, and Jingfei Zhang. Cell-type-specific co-expression inference from single cell rna-sequencing data. *Nature Communications*, 14(1):4846, 2023.

[27] Eskandar Taghizadeh, Sahel Heydarheydari, Alihossein Saberi, Shabnam JafarpoorNesheli, and Seyed Masoud Rezaeijo. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinformatics*, 23(1):410, October 2022.

[28] Francesco Torgano, Mauricio Soto Gomez, Matteo Zignani, Jessica Gliozzo, Emanuele Cavalleri, Marco Mesiti, Elena Casiraghi, and Giorgio Valentini. RNA knowledge-graph analysis through homogeneous embedding methods. *Bioinformatics Advances*, 5(1):vbaf109, January 2025.

[29] Likai Wang, Yanpeng Xi, Sibum Sung, and Hong Qiao. RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC Genomics*, 19(1):546, July 2018.

[30] Liming Wang, Fangfang Liu, Longting Du, and Guimin Qin. Single-Cell Transcriptome Analysis in Melanoma Using Network Embedding. *Frontiers in Genetics*, 12, July 2021. Publisher: Frontiers.