Efficient Inference of Direct GeneGene Associations via High-Dimensional Precision Matrix with Rigorous FDR Control

Feifei Ran[1,†], Jing Li[2,†], Ying Liu[1], Bin Lian[1], Jie He[3,*], and Jialu Hu[1,*]

[1]School of Computer Science, Northwestern Polytechnical University, 1 Dong Xiang Rd., 710129, Shaanxi, China. Tel: 029-88431519.
[2]Xi'an Mingde Institute of Technology, School of Information Engineering, 1 Mingde Rd., 710124, Shaanxi, China
[3]School of Mathematics, Nanjing University of Aeronautics and Astronautics, 29 Jiangjunda Rd., 211106, Nanjing, China.
[†]These authors contributed equally.
[*]These authors are co-responding authors.

Abstract

Gene coexpression networks (GCNs) represent complex patterns of coordinated gene activity by modeling pairwise relationships between genes. Reconstructing GCNs from single-cell transcriptomic data is essential for the identification of gene modules, the inferring of cell-type-specific regulatory programs, and the exploration of dynamic transcriptional changes in diverse cellular states and developmental trajectories. However, predicting gene coexpression networks remains challenging due to confounding effects and the low inference accuracy of the graph structure under the high dimensional settings. We propose a novel method, dGGAPM (Direct Gene-Gene Associations via Precision Matrix), which reconstructs gene coexpression networks by statistically inferring direct gene-gene relationships. Unlike conventional approaches that rely on pairwise correlation measures and may capture indirect associations, dGGAPM leverages precision matrix to more accurately characterize the graph structure with direct interactions between genes by controlling for the influence of all other genes. The main contributions of this research are threefold. First, in contrast to existing methods, we construct the high-dimensional gene network using an estimated precision matrix, providing a rigorous and principled foundation for statistical inference of genegene relationships. Second, we recover the network structure through a high-dimensional multiple testing procedure that takes advantage of the asymptotic properties of the precision matrix estimator. Finally, we introduce a data-driven thresholding strategy that achieves strict false discovery rate (FDR) control, ensuring reliable identification of coexpression links in high-dimensional settings. The dGGAPM can also be used to identify functionally coherent gene modules and gain insight into their regulatory relationships. Although some nonlinear models can capture more complicated correlation relationships, a rigorous statistical inference method is challenging compared to the proposed dGGAPM. We evaluated dGGAPM on multiple publicly available single-cell transcriptomic datasets and benchmarked its performance against several state-of-the-art network inference algorithms. The results demonstrate that dGGAPM consistently achieves superior performance in terms of precision, specificity, and biological relevance, highlighting its potential as a robust tool for gene network reconstruction and functional module analysis. The source code of the dGGAPM is freely available at https://github.com/jhu99/DGGAPM.

Introduction

Gene coexpression networks (GCNs) can be leveraged to predict cell-type-specific transcriptional states and uncover a wide range of underlying biological processes in single-cell RNA sequencing (scRNA-seq) data (Delgado and Gómez-Vela, 2019; Hecker *et al.*, 2009). In these networks, nodes correspond to genes, while edges indicate interactions relationships among genes, typically inferred through correlation or mutual information measures (Specht and Li, 2015; Albert and Barabási, 2002). By capturing coordinated gene activity patterns, GCNs enable the identification of regulatory modules, inference of gene functions, and characterization of cell identity and heterogeneity across diverse cellular contexts. GCNs are also widely utilized to infer functions of previously uncharacterized genes by leveraging the known functions of co-expressed genes and their regulatory relationships (Parsana *et al.*, 2019). Through further analysis of network modules, more advanced tasks can be accomplished, including the identification of core genes within modules, association of gene modules with phenotypic traits, and modeling of metabolic and signaling pathways (Langfelder and Horvath, 2008). Thus, accurate estimation of coexpression relationships between genes is essential for reliable inference of gene coexpression networks. Recent studies have utilized functional networks derived from scRNA-seq data to address key biological questions, shedding light on transcriptional regulatory mechanisms in various contexts, including cancer progression (Zhang *et al.*, 2021b), immune system response (Wen *et al.*, 2020), and embryonic development (Shangguan *et al.*, 2020). However, current single-cell RNA sequencing (scRNA-seq) technologies are subject to several technical limitations that can introduce substantial artifacts into downstream analyzes. First, scRNA-seq data are characterized by a high prevalence of zero countscommonly referred to as dropoutsand relatively inaccurate low expression counts, which arise from a combination of technical noise and intrinsic biological variability (Kharchenko *et al.*, 2014; Li and Li, 2018; Azizi *et al.*, 2017). These artifacts can obscure true gene expression signals and distort inferred relationships. Second, variability in the depth of cell-specific sequencingoften caused by uneven loading of single-cell libraries into sequencing reactionsleads to inconsistent data quality. Third, the dynamic range of gene expression in single-cell datasets is typically compressed relative to bulk RNA-seq data, placing demands on the robustness and sensitivity of correlation-based inference methods.

Pearson correlation and Spearman correlation are often used to infer gene coexpression networks because of their computational efficiency and strong statistical properties. However, a significant drawback is that Pearson and Spearman correlations primarily focus on specific gene pairs during analysis. In reality, the correlation patterns within gene expression data are often intricate, influenced by numerous unmeasurable confounders. Consequently, analyzes based solely on these metrics can produce unconvincing results unless spurious inferences are adequately filtered out (Brown and Hendrix, 2005). To solve these problems, several computational approaches have been proposed, such as scLink (Li and Li, 2021), SCODE (Matsumoto *et al.*, 2017), GENIE3 (Huynh-Thu *et al.*, 2010; Greenfield *et al.*, 2010) etc, in the last two decade. ScLink is a method that adapts the Gaussian graphical model to infer gene coexpression networks. SCODE focuses on differentiated cells by integrating the transformation of linear ordinary differential equations (ODEs) with linear regression. It takes into account the temporal aspect of each cell when assessing regulatory relationships between genes. GENIE3 decomposes the task of inferring a gene regulatory network from bulk RNA-seq data among p genes into p individual regression problems, using tree-based ensemble methods to predict regulatory relationships and assign edge weights. It has also been integrated into an SCENIC framework (Aibar *et al.*, 2017) for network inference from scRNA-seq data. Later, GRNBoost2 Moerman *et al.* (2019) is proposed to solve the problem of GCN inference using Gradient Boosting Machine (GBM) regression and the GENIE3 framework. However,

their reliance on strong assumptions and limited capacity to capture indirect interactions or complex, non-linear dependencies may reduce their effectiveness in modeling intricate single-cell gene expression patterns.

These challenges highlight the need for analytical approaches that are capable of accommodating data sparsity, heterogeneity, and signal compression while preserving biologically meaningful coexpression patterns. To address the challenges, we propose a novel method, dGGAPM, which recovers the gene coexpression networks by statistical inference on the precision matrix with a FDR controlled threshold for single-cell RNA sequencing (scRNA-seq) data (Brown and Hendrix, 2005). In contrast to many existing approaches that rely on conventional correlation measures that capture only pairwise associations, precision matrix quantifies the direct linear relationship between two variables while conditioning on the influence of all other variables. This adjustment allows dGGAPM to effectively account for hidden or unmeasured confounding factors, thereby yielding a more accurate representation of gene-gene dependencies. By removing the indirect effects of the intervening variables, precision matrix offers a closer approximation to the true regulatory relationships embedded in gene expression data. In summary, we provide a more robust and biologically interpretable framework compared to existing methods, establishing a solid foundation for downstream analyzes and advancing our understanding of cellular function and regulation in the life sciences.

Results

Methods overview

Here, considering the problem of reconstructing gene coexpression network, we propose a novel method, dGGAPM (Direct GeneGene Associations via High-Dimensional Precision Matrix) by statistically inferring direct correlation relationships among genes. As shown in Fig. 1, it takes a normalized expression matrix as input and uses a sparse Lasso-based scaled linear regression to estimate the precision matrix of genes, along with an adaptive thresholding strategy, thus facilitating the construction of gene coexpression networks (GCNs) from sparse scRNA-seq data while effectively removing the influence of confounding factors. We evaluated the effectiveness and accuracy of dGGAPM by benchmarking it against several state-of-the-art algorithms using three real-world single-cell transcriptomic datasets. The correlation relationships estimated by dGGAPM were further employed in a variety of downstream analyzes, including the identification of functional modules, gene set enrichment analysis, and the exploration of cell type-specific coexpression patterns. These applications underscore the versatility and biological relevance of dGGAPM in extracting structured insights from high-dimensional and noisy single-cell RNA-seq data. The source code used in this study is freely accessible at https://github.com/jhu99/dGGAPM.

Identification of gene coexpression networks from human embryonic stem cells

To assess the efficiency of dGGAPM in capturing dynamic gene expression changes, we applied our algorithm alongside four widely used methods, including scLink, GENIE3, SCODE, and GRNBoost2, to a time-course single-cell RNA-seq dataset comprising 758 cells collected during the differentiation of human embryonic stem cells (hESCs) into definitive endoderm cells (Chu *et al.*, 2016). This dataset includes gene expression profiles measured at six distinct time points0, 12, 24, 36, 72, and 96 hoursthereby capturing the temporal transcriptional dynamics underlying early developmental processes. After that, we quantitatively evaluated the performance of each method using several established metrics, including the area under the receiver operating characteristic

curve (AUROC), area under the precision-recall curve (AUPRC), and biological pathway enrichment of the inferred gene modules.

As shown in Fig. 2a, we succesfully identified five clusters of genes from a correlation matrix calculated by dGGAPM. Each red dot in the figure represents positive correlations and purple represents negative correlations. To further elucidate the specific roles of the five gene clusters identified through dGGAPM, we conducted gene-set enrichment analysis. We selected the top six pathways of each gene set and ploted them into a heatmap as shown in (Fig. 2b). Genes grouped in Cluster 1 are predominantly associated with cell differentiation processes critical for kidney development. Cluster 2 comprises genes primarily involved in endodermal cell fate commitment and DNA biosynthetic processes. Genes in Cluster 3 are enriched for pathways regulating glucose homeostasis and the positive regulation of macroautophagy. Cluster 4 includes genes that play a central role in the differentiation of natural killer T (NK T) cells, while Cluster 5 contains genes implicated in cartilage morphogenesis. Results demonstrate that each gene cluster is involved in distinct biological processes during cell differentiation, which further validates the rationality of the gene correlation relationship and clustering results obtained through dGGAPM.

Next, we reconstructed gene coexpression networks by retaining only those edges that passed false discovery rate correction (FDR < 0.05). As shown in Fig. 2c, each vertex in the network represents a gene, color-coded according to its assigned cluster. Edges denote genegene interactions, with red indicating positive correlations and gray indicating negative correlations. The thickness of each edge reflects the interaction strength, represented by the absolute value of the correlation coefficient. Based on the results from the enrichment analysis, we identified several gene modules within the coexpression network, each of which plays a crucial role in specific biological processes. Specifically, the results show that:

- The module consisting of TFAP2A, NOTCH2, and STAT1 genes is primarily involved in kidney development (Takasato *et al.*, 2014), contributing to the formation and functional maturation of kidney structures,
- NANOG and SOX2 are key in the commitment of endodermal cell fate (Jaremko and Marikawa, 2013), regulating the differentiation and specification of the endoderm during early embryonic development,
- PCNA, KAT7, and DTL are pivotal in DNA biosynthesis, ensuring the accurate replication and transmission of genetic information,
- PRDM1 and ITPR1 play significant roles in late embryonic development (Palli, 2021), impacting the maturation and functionality of various tissues and organs,
- HAND2 and RSPO2 are associated with cartilage development (Quintana *et al.*, 2009), influencing the formation and morphological transformation of cartilage tissues through gene expression regulation and signaling pathways.

These findings provide potential drug targets for research on kidney development, embryonic stem cell differentiation, DNA replication, and chondrogenesis.

To evaluate the performance of dGGAPM, we used two metricsROC curve analysis and the number of true positive edgesto comprehensively evaluate the performance of our proposed method against four existing methods: scLink, GENIE3, SCODE, and GRNBoost2. Through ROC curve analysis, our method demonstrated superior performance with an Area Under the Curve (AUC) of 0.7917, significantly higher than scLink (0.5947), GENIE3 (0.6119), SCODE (0.5519), and GRNBoost2 (0.5689) (Fig. 3a). To assess sensitivity, we quantified the number of truepositive edges

among the top 1,000 ranked edges derived from a subset of 100 genes. As shown in Fig. 3b, our dGGAPM method identified 79 correct edges, compared with 53 for scLink, 81 for GENIE3, 62 for SCODE, and 72 for GRNBoost2. Although dGGAPM and GENIE3 recovered a similar total number of truepositive edges, dGGAPM demonstrated superior ability to capture highly correlated interactions, outperforming the other methods on this metric.

In the reconstructed gene coexpression network, the NANOG gene exhibits close connections with multiple other genes, indicating its central position within the network. This centrality suggests that NANOG may coordinate multiple biological processes and influence the differentiation potential of human embryonic stem cells (hESCs). We then focused on the NANOG gene within the dataset. Seven additional genes associated with NANOG were extracted from the gene coexpression network (Fig. 3c), and a thorough analysis was conducted on these eight genes. Considering the correlation between cells and time in the hESC data, we calculated the average expression levels of these genes at each time point (Fig. 3d). The line graph reveals that the expression level of NANOG gradually decreases as cell differentiation progresses, whereas the PRDM1 gene shows a negatively regulated relationship with NANOG over time. Other genes exhibited similar expression trends to NANOG, consistent with the inter-gene regulatory relationships shown in Fig. 3c. We performed enrichment analyzes for these eight genes and presented the results for the top 20 pathways in biological processes (Fig. 3e). For biological processes, we further refined the association between these genes and the top 10 most significant pathways (adjusted p-value $<$ 8.3e-03), presenting them in bar charts (Fig. 3f). Our analysis indicated that NANOG and SOX2 continue to play central roles during endodermal differentiation. This finding aligns with the study by Pan and Thomson (2007), which demonstrated that genes like NANOG and SOX2 act synergistically to co-regulate a set of target genes essential for embryonic stem cell pluripotency. Specifically, NANOG plays a key role in regulating the cell fate of the pluripotent inner cell mass (ICM), maintaining the pluripotent ectodermal state, and preventing differentiation into the primitive endoderm. Finally, to thoroughly validate the effectiveness of dGGAPM across different gene quantities, we conducted a more comprehensive validation. Specifically, we selected the top 200 and top 400 high-variance genes, ranked by degree, from the ground-truth network, as well as the top 400 genes based on degree ranking, to infer the gene coexpression network and calculate the ROC. As shown in Fig. 3g, dGGAPM obtained the highest AUC (0.5528), followed by SCODE (0.5155), GRNBoost2 (0.5129), GENIE3 (0.5075), scLink (0.5066). Overall, it can be concluded that dGGAPM demonstrated superior performance in capturing gene-gene interactions from dynamic gene expression changes in human embryonic stem cells.

Identification of gene coexpression networks from mouse embryonic stem cells

To validate the effectiveness and robustness of our proposed method across species, we employed an additional dataset comprising mouse embryonic stem cell (mESC) data. This dataset contains scRNA-seq expression measurements from 421 primitive endoderm (PrE) cells differentiated from mESCs, collected at five distinct time points (0, 12, 24, 48, and 72 hours). To evaluate the performance of our approach, we selected the top 100 highly connected genes for in-depth analysis. Initially, we extracted gene expression data for these 100 genes across the 421 cells and computed the precision matrix using our method (see Equation 3). Based on the estimated correlation coefficients, we applied the K-Means algorithm to cluster the genes into four distinct categories, containing 38, 25, 18, and 19 genes, respectively. The resulting correlation heatmap (Fig. 4a) illustrates positive correlations in red and negative correlations in purple. Notably, genes within the same category exhibited high correlation, whereas correlations between genes from different categories were relatively low. This pattern indicates strong intra-category cohesion and weak

inter-category connectivity. This visualization effectively highlights the modular structure of the gene coexpression network, providing a way for subsequent biological analyzes.

After identifying these gene sets, we performed enrichment analyzes, selecting the six most significantly enriched pathways from each category. We then calculated the negative logarithm (-log10) of the adjusted p-values for each pathway and visualized the results in a heatmap (Fig. 4b). The enrichment analysis revealed that different gene modules were significantly associated with various biological pathways. In particular, genes within the first and third modules appear to play key roles in maintaining stem cell populations. Additionally, the co-enrichment of the first and fourth modules in heterochromatin organization pathways suggests that genes within these modules may share overlapping functions in regulating stem cell-specific gene expression patterns. We highlight the corresponding pathways in Fig. 4b for intuitive observation and analysis. A partial representation of the gene coexpression network is depicted in Fig. 4c. In category 1 (red nodes), a submodule comprising five genes was identified, which, along with another five-gene submodule detected in category 3 (yellow nodes), is implicated in mechanisms that maintain stem cell populations and cell numbers (Yang *et al.*, 2022). In category 2 (orange nodes), a tightly knit coexpression module consisting of four genes was discovered, primarily responsible for the regulation of protein stability (Zhang *et al.*, 2021a), a critical process to maintain cell homeostasis and normal physiological functions. Additionally, a synergistic interaction between the genes Gata4 and Sox17 was observed in the morphogenesis of the embryonic foregut (Chin *et al.*, 2017). Finally, in category 4 (green nodes), a submodule composed of the genes Amdhd2 and Hexa was found to play a significant role in dynamic metabolic processes during early mammalian embryonic development (Zhao *et al.*, 2023). These discoveries not only enhance our understanding of the mechanistic roles of gene modules in specific biological pathways but also offer valuable insights for further exploration of gene interaction networks and potential therapeutic targets.

Next, we compared dGGAPM with scLink, GENIE3, SCODE, and GRNBoost2, evaluating their performance using Receiver Operating Characteristic (ROC) curves and the number of true positive edges identified among the strongest correlations. In terms of ROC curves, the area under the curve (AUC) for our method (AUC = 0.5877) surpassed that of the other four methods: scLink (0.5093), GENIE3 (0.5314), SCODE (0.5159), and GRNBoost2 (0.5305) (Fig. 5a). To further assess the statistical significance of dGGAPM, we conducted DeLongs test to compare its AUC with those of the other methods, yielding p-values of 0.0579, 0.1158, 0.0414, and 0.1093, respectively. These results suggest that dGGAPM has a comparative advantage over the alternative approaches. Regarding true positive edges, we quantified the number of correct edges among the top 100 and 200 edges derived from the 100 genes. As shown in the bar chart in Fig. 5b, among the top 100 edges, dGGAPM, scLink, GENIE3, SCODE, and GRNBoost2 identified 31, 20, 22, 18, and 19 true positive edges, respectively. For the top 200 edges, the five methods detected 54, 34, 44, 31, and 42 true positive edges, respectively. These findings demonstrate that, on this dataset, our method outperforms the other four approaches by identifying a greater number of true positive edges.

By sorting the nodes based on various attributes such as degree, betweenness centrality, and closeness centrality, we identified several central nodes, including Dppa2, Pou5f1, Trim28, and Sox2. These central nodes may play pivotal roles in the gene network, and any abnormal expression or functional loss of these genes could lead to significant biological consequences. We then focused on Sox2 and its associated genes for further investigation, specifically analyzing their interactions, as shown in Fig. 5c, where genes are color-coded by category. Given cells in the mESC dataset that were collected at multiple time points, we calculated the average expression of these genes at each time point to observe their dynamic expression changes. As depicted in the line chart in Fig. 5d, the expression of Sox2 gradually decreases as cell differentiation progresses. Furthermore, Sox2 exhibits a positive regulatory relationship with four other genes in the same category over time,

while the Gata4 gene, which negatively correlates with Sox2, shows an opposite expression pattern.

Then, we performed an enrichment analysis on these six genes and visualized the top 20 biological process pathways in Fig. 5e. We then focused on the ten most significant pathways (with an adjusted p-value $< 3.7e-04$) and examined the specific roles of individual genes within these pathways. The results are shown in a bar chart in Fig. 5f. The analysis highlights that Sox2, Nanog, and Gata4 collectively play a critical role in the formation and development of the endoderm. Furthermore, nine of the pathways involve both Sox2 and Nanog, which aligns with findings from human embryonic stem cell datasets. This reinforces the idea that Sox2 and Nanog work synergistically as key target genes in regulating embryonic stem cell pluripotency.

Identification of gene coexpression networks from mouse hematopoietic stem cells

Finally, we applied our method to mouse hematopoietic stem cell data to assess its adaptability and effectiveness across different cell types. In line with our previous analyzes, we first identified the top 100 highly variable genes with the greatest degrees in the real network. Using the expression profiles of these genes across 1,071 cells, we computed their precision matrix. We then performed K-means clustering, which grouped the genes into five categories containing 20, 6, 6, 14, and 54 genes, respectively. The correlation heatmap in Fig. 6a illustrates the distribution of gene correlations within and between categories. It is evident from the heatmap that genes in category 1 exhibit weak intra-category correlations and minimal associations with genes in other categories, suggesting these genes may play a reduced role in gene regulation. Next, we conducted enrichment analyzes on these gene modules. We selected the top five significantly enriched pathways for each category and visualized the negative logarithm of the adjusted p-values (-log10) for each pathway in a heatmap, as shown in Fig. 6b. The enrichment analysis results indicate that specific gene modules are significantly enriched in only a few corresponding biological pathways, while showing low enrichment scores (blue regions) in pathways that are strongly enriched in other modules. This confirms the precision of our method in calculating gene precision matrix and highlights its ability to identify gene sets with distinct biological functions effectively.

The partial topological structure of the gene coexpression network inferred by dGGAPM is shown in Fig. 6c, with the corresponding enrichment results depicted in Fig. 6b using color-coded frames that represent different gene categories. Notably, the inferred network excludes genes from category 1, which aligns with the prior analysis of the precision matrix. Additionally, dGGAPM identified several biologically significant gene submodules. For instance, a module from category 2, consisting of PCNA, MCM4, CDT1, and RAD51, plays a crucial role in DNA replication (Fortini *et al.*, 2012), facilitating DNA strand elongation and damage repair. Similarly, a four-gene module in category 3 (yellow nodes) is involved in chromosome segregation and spindle assembly, processes vital for cell division. Furthermore, dGGAPM detected overlapping three- and four-gene modules, including GATA2, JUN, and FOS, which positively regulate miRNA transcription and significantly contribute to hematopoiesis regulation (Liao *et al.*, 2022). Finally, a three-gene submodule within category 5 was notably enriched in pathways related to the positive regulation of protein localization to telomeres, a crucial mechanism for maintaining telomere integrity and genomic stability. These findings may provide insights into the molecular mechanisms underlying disease pathogenesis.

For the mHSC dataset, we conducted a comparative analysis of our method against the same four established methods used in previous experiments. In this evaluation, we applied two distinct gene selection strategies and computed the corresponding ROC curves. Initially, we selected 100 highly variable genes with the highest degrees in the real network for analysis. As shown in Fig. 7a, the Area Under the Curve (AUC) obtained by dGGAPM (AUC = 0.6037) surpassed that of

the other four methods: scLink (0.5550), GENIE3 (0.5484), SCODE (0.5461), and GRNBoost2 (0.5524). This result indicates that our method achieves superior accuracy in identifying gene regulatory relationships, particularly when addressing highly variable genes. Subsequently, we ranked the top 100 genes based solely on degree, and once again, dGGAPM consistently outperformed the other four methods in this dataset (Fig. 7b). These findings further highlight the enhanced capability of our approach in accurately inferring gene coexpression networks.

To evaluate the ability of each method to detect true positive edges when predicting strong gene correlations, we calculated the number of true positive edges among the top 50 and 100 most highly correlated edges derived from a set of 100 genes (Fig. 7c). Among the top 50 edges, dGGAPM, scLink, GENIE3, SCODE, and GRNBoost2 identified 7, 5, 2, 9, and 2 true positive edges, respectively. For the top 100 edges, these methods detected 15, 11, 2, 16, and 2 true positive edges, respectively. These results show that, while dGGAPM ranks slightly below SCODE in terms of detecting gene coexpression networks, it outperforms the other three methods in terms of both effectiveness and accuracy.

Our analysis successfully identified several central nodes critical to the gene coexpression network, including key genes such as JUN, GATA2, and HDAC. In subsequent investigations, we specifically focused on the GATA2 gene and its associated gene set, performing an in-depth analysis of their interactions (Fig. 7d). To clearly differentiate genes across categories, we employed unique color coding. Consistent with previous studies, we conducted enrichment analysis on these seven genes and presented the top 20 biological process pathways (Fig. 7e). We further identified the ten most significant pathways (adjusted p-value $<$ 7.9e-05), visualized in a bar chart in Fig. 7f. The results show that GATA2, JUN, and FOS consistently feature in the top ten significantly enriched pathways, primarily contributing to the positive regulation of miRNA transcription and metabolic processesessential functions for normal cellular activity and differentiation. GATA2, a well-known transcription factor, plays a pivotal role in several biological processes, including cell proliferation, differentiation, and apoptosis. Previous research has shown that GATA2 regulates early embryonic development and works in concert with other GATA transcription factors to control the proliferation and differentiation of granulocytic, erythroid, megakaryocytic, and mast cell lineages (Vicente *et al.*, 2012). Additionally, GATA2 is a key regulator in hematopoiesis, with its overexpression or mutation being associated with leukemogenesis, suggesting that further investigations into GATA2 could provide new insights into leukemia pathogenesis. JUN and FOS, members of the AP-1 transcription factor family, are integral to cellular stress responses, inflammation, and tumorigenesis (Eferl and Wagner, 2003). The persistent expression and activity of these transcription factors can profoundly affect cell survival and differentiation, influencing disease initiation and progression. Thus, understanding their roles within these pathways enhances our comprehension of the molecular mechanisms driving cellular differentiation and presents promising targets for disease treatment and prevention.

Methods and materials

Inferring gene coexpression networks from scRNA-seq data

To illustrate this point, consider a toy example involving a p-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$, where the goal is to recover the underlying network structure of $\mathbf{X}$ in a high dimensional setting with large p. A straightforward approach is to estimate the sample correlation matrix of $\mathbf{X}$, denoted by $\mathbf{\Sigma}$, where off-diagonal entry σ_{ij} corresponds to the Pearson correlation coefficient between X_i and X_j, for all $i \neq j \in \{1, \ldots, p\}$. However, Pearson correlation is easily influenced by multicollinearity within $\mathbf{X}$. Suppose that both (X_1, X_2) and (X_1, X_3) are highly

correlated, while (X_2, X_3) are independent. Due to the presence of both direct and indirect associations, the estimated Pearson correlation between X_2 and X_3 tends to be artificially inflated–an effect that becomes even more pronounced in high-dimensional settings. In contrast, the situation is entirely different for the precision matrix, which is the inverse of the correlation matrix. Each entry of the precision matrix, denoted by ω_{ij}, corresponds to the partial correlation between X_i and X_j, by removing the influence of all remaining variables $\mathbf{X}_{-(i,j)} = (X_k, k = 1, \ldots, p, k \neq i, j)^\top$. As a result, the estimated dependence structure is much more accurate. In essence, the precision matrix improves upon marginal correlation by filtering out spurious associations induced by other related variables. It has seen widely applied across various fields. For instance, Peng *et al.* (2009) introduced a nodewise regression approach to estimate the precision matrix in high-dimensional, low-sample-size scenarios. Their method consists of two main steps: first, estimating the precision matrix for high dimensional random vector by minimizing a joint sparse regression model based loss function; and second, inferring the network structure from the estimated precision matrix. Building on this foundation, Qiu and Zhou (2020) proposed a c-level high dimensional multiple testing procedure for precision matrix. In this framework, variable pairs whose corresponding estimated precision matrix entries exceed a predefined threshold in absolute value are connected by edges and subsequently tested for significance. This method was applied to the analysis of brain imaging data from patients with Alzheimers disease.

Motivated by these researches, we aim to extend the application of the statistical inference for the precision matrix to the analysis of gene coexpression networks based on single-cell data. Our research can be summarized by the following main steps: 1. Data Preprocessing. For the human embryonic stem cell (hESC) dataset, we conducted preprocessing using the Seurat workflow. Specifically, cells with fewer than 6,000 or more than 12,000 detected genes were filtered out, along with those exhibiting a mitochondrial gene proportion exceeding 10%. Additionally, genes expressed in fewer than 10% of cells were removed. The data were then normalized using Seurat's default LogNormalize method. After preprocessing, the final dataset contained 13,069 genes across 714 cells. For the mouse embryonic stem cell (mESC) dataset, we applied a log transformation to transcripts per kilobase million (TPM) or fragments per kilobase million (FPKM) values, using a pseudo-count of 1. The transformed values were treated as gene expression levels. As with the hESC dataset, genes expressed in fewer than 10% of the cells were filtered out. The resulting dataset included 18,385 genes across 421 cells. For the mouse hematopoietic stem cell (HSC) dataset, we focused on erythroid lineage data obtained from three hematopoietic lineages (erythroid, granulocyte-monocyte, and lymphoid), following a previous study (Pratapa *et al.*, 2020). The final processed dataset consisted of 4,762 genes with standardized expression values across 1,071 cells.

2. Estimating the precision matrix. Denote by $\mathbf{y} = (y_1, \ldots, y_p)^\top$ the population-level gene expression vector, where p represents the number of genes. More specifically, let $y_{i,j}$ be the expression value of the jth gene in the ith cell, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where n denotes the number of cells. By collecting all expression values of the jth gene across the n cells, we obtain an n-dimensional expression vector $\mathbf{y}_j = (y_{1,j}, \ldots, y_{n,j})^\top$, for $j = 1, \ldots, p$. Following the approach of Peng *et al.* (2009), the precision matrix can be estimated by fitting p node-wise regression models, each of the form:

$$y_j = \alpha_{j,0} + \boldsymbol{\alpha}_j^\top \mathbf{y}_{-j} + \epsilon_j \tag{1}$$

for $j = 1, \ldots, p$, where $\mathbf{y}_{-j} = (y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_p)^\top$ denotes the $p-1$-dimensional vector obtained by removing the jth element from $\mathbf{y}$. The vector $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \ldots, \alpha_{j,p-1})^\top$ represents the regression coefficients corresponding to $\mathbf{y}_{-j}$, and ϵ_j is the error term in the regression model (1). According to the results in (Peng *et al.*, 2009), the regression error ϵ_{j_1} is uncorrelated with $\mathbf{y}_{-j_1}$ if and only if $\alpha_{j_1,j_2} = -\omega_{j_1,j_2}/\omega_{j_1,j_1}$ for any $j_2 \neq j_1$, where ω_{j_1,j_2} denotes the (j_1, j_2)-th element of

the precision matrix. Let v_{j_1,j_2} denote the sample covariance between regression errors ϵ_{j_1} and ϵ_{j_2}, which can be estimated by

$$\hat{v}_{j_1,j_2} = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2} + \hat{\alpha}_{j_1,j_2}\hat{\epsilon}^2_{i,j_2} + \hat{\alpha}_{j_2,j_1}\hat{\epsilon}^2_{i,j_1}) & , j_1 \neq j_2, \\ \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2} & , j_1 = j_2. \end{cases} \tag{2}$$

Denote ρ_{j_1,j_2} as the partial correlation between genes y_{j_1} and y_{j_2}. From equation (2), we can derive the estimation of ρ_{j_1,j_2} as following

$$\begin{aligned} \hat{\rho}_{j_1,j_2} &= -\hat{v}_{j_1,j_2}(\hat{\omega}_{j_1,j_1}\hat{\omega}_{j_2,j_2})^{1/2} \\ &= \frac{-\hat{v}_{j_1,j_2}}{(\hat{v}_{j_1,j_1}\hat{v}_{j_2,j_2})^{1/2}} \quad \text{for} \quad \text{any} \quad j_1 \neq j_2. \end{aligned} \tag{3}$$

3. Building a Gene coexpression Network. Based on the estimated precision matrix and the associated variances for all entries of the precision matrix, we derived an adaptive threshold estimator for the precision matrix. To ensure the reliability of network construction, we determined the optimal threshold by rigorously controlling the false discovery rate (FDR). Specifically, two genes are considered connected if the absolute value of their estimated precision matrix exceeds the selected threshold. Under this criterion, an edge is introduced between the corresponding genes, thereby defining the structure of the resulting gene coexpression network.

Thresholds selection

According to large deviation theory, the bias-corrected estimators $\hat{v}_{j_1,j_2}$ and $\hat{v}_{j_1,j_1}$ can be approximated by the sample second moments of ϵ_{i,j_1} and ϵ_{i,j_2}, respectively. This implies that the estimated precision matrix $\hat{\rho}_{j_1,j_2}$ is also a function of the residuals of ϵ_{i,j_1} and ϵ_{i,j_2}. To incorporate statistical significance, we define a thresholded estimator as $\tilde{\rho}_{j_1j_2} = \hat{\rho}_{j_1j_2}\mathrm{I}\{|\hat{\rho}_{j_1,j_2}| > 2[\log(p)/\mathrm{n}]^{1/2}\}$, where $\mathbf{I}(\cdot)$ denotes the indicator function. This formulation filters out weak or spurious correlations, retaining only those precision matrix that exceed a theoretically justified threshold. The specific form of the asymptotic variance can then be derived as follows:

$$\begin{aligned} &\widehat{\mathrm{var}}(n^{1/2}\hat{\rho}_{j_1,j_2}) = \hat{\kappa}(1-\tilde{\rho}^2_{j_1,j_2})^2, \\ &where \quad \hat{\kappa} = \frac{n}{3p}\sum_{j=1}^{p}\frac{\sum_{i=1}^{n}\hat{\epsilon}^4_{i,j}}{(\sum_{i=1}^{n}\hat{\epsilon}^2_{i,j})^2}. \end{aligned} \tag{4}$$

Applying the standardized $\hat{\rho}_{j_1,j_2}$, the adaptive thresholding estimator for the precision matrix is

$$\begin{aligned} &\hat{\rho}_{j_1,j_2}(\tau) \\ &= \hat{\rho}_{j_1,j_2}\mathrm{I}\{|\hat{\rho}_{j_1,j_2}| > \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa}log(p)/n\}^{1/2}\}, \end{aligned} \tag{5}$$

and the corresponding graph structure is

$$\hat{\varepsilon} = [(j_1,j_2) : |\hat{\rho}_{j_1,j_2}| > \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa}log(p)/n\}^{1/2}]. \tag{6}$$

Qiu Qiu and Zhou (2020) also propose a novel method for determining the thresholding value τ, which ensures control of the false discovery rate (FDR) in graph estimator. For any set $\mathcal{A}$, let

$\#\{\mathcal{A}\}$ denote its cardinality, and let $\bar{\mathcal{A}}$ represents its complement. The false discovery proportion (FDP) is then defined as:

$$\begin{aligned} \mathrm{FDP}(\tau) &= \#\{\bar{\varepsilon}\}\mathrm{FPR}(\tau)/\max[1, \#\{\hat{\varepsilon}(\tau)\}], \quad \mathrm{w}here \\ \mathrm{FPR}(\tau) &= \frac{1}{\#\{\bar{\varepsilon}\}} \sum_{(j_1,j_2)\in\bar{\varepsilon}} \\ &\qquad \mathrm{I}\{|\hat{\rho}_{j_1,j_2}| > \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa} log(p)/n\}^{1/2}\}. \end{aligned} \tag{7}$$

From large deviation theorem, the false positive rate can be bounded by

$$G(\tau) = 2 - 2\Phi\{\tau\sqrt{log(p)}\} \tag{8}$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. To control the FDR at a desired level $\alpha \in (0, 1)$, the threshold parameter τ is selected as:

$$\tau_{\mathrm{FDP}} = \inf\{\tau \in (0, 2] : \frac{G(\tau)(p^2 - p)}{\max[1, \#\{\hat{\varepsilon}\}]} \leq \alpha\}. \tag{9}$$

Finally, we apply this data-driven threshold τ to update the precision matrix estimator (as given in Equation (5)), thereby obtaining an optimally thresholded gene coexpression network with controlled FDR.

Gene clustering

We apply the K-Means algorithm to perform clustering analysis on the studied genes, aiming to group genes with similar expression patterns together within the coexpression network. This clustering facilitates the visualization of network modularity, characterized by dense connections within clusters and sparse connections between them. First, the Euclidean distances between gene expression profiles are computed. To determine the optimal number of clusters, we plot the total within-cluster sum of squares (WSS) against varying numbers of clusters using the "fviz_nbclust" function. The elbow point of this plot provides an estimate of the optimal number of clusters based on the WSS method.

Calculation of p-value for each edge

After estimating the genegene correlations, we employed a permutation test to assess the statistical significance of each edge in the network. Specifically, we randomly shuffled the data and recalculated the correlation values for each edge across 1,000 permutations. For each permutation, the correlation value of the given edge was recorded. The p-value for each edge was then computed as the proportion of permutations in which the permuted correlation exceeded the original correlation estimated by dGGAPM. This procedure provides a robust, data-driven measure of significance for the inferred genegene associations.

Selection of central node

Gene nodes within the coexpression network were further analyzed using Cytoscape, where nodes were ranked based on various centrality metrics, including degree, betweenness, and closeness (Iacono *et al.*, 2019). Degree centrality, the most fundamental measure, represents the number of direct connections a node has. Betweenness centrality is computed by enumerating all shortest paths in the network and quantifying how often a node lies on these paths, thereby reflecting its

role in facilitating communication between other nodes. Closeness centrality assesses how close a node is to all other nodes in the network, defined as the reciprocal of the average shortest path length from the node to all others. By integrating these centrality measures, we identified key hub genes within each cluster, which may play critical regulatory roles.

Ground-truth networks

Creating artificial graphs or extracting subnetworks from large-scale transcriptional networks is a common strategy for establishing a ground trutha reference network of known regulatory interactions that govern the dynamics of genes of interest. In our study, the ground truth network was obtained from Pratapa *et al.* (2020), and was constructed by aggregating ChIP-seq data from multiple databases, including ENCODE (encodeproject.org), ChIP-Atlas (chip-atlas.org), and ESCAPE (PMC3689438), all derived from the same or similar cell types. To evaluate the performance of our method, we compared the inferred gene coexpression network with the ground truth. Each edge in our network was labeled based on its presence in the ground truth: an edge was assigned a label of "1" if it also appeared in the ground truth network, and "0" otherwise. This labeling enabled a rigorous, binary classification-based evaluation of edge prediction accuracy. We then assessed the performance of our method using these ground-truth labels, with the results presented below.

Variance analysis

After calculating the ROC curves for various methods, we applied the DeLong test to assess the differences in the Area Under the Curve (AUC) values. This statistical test was used to evaluate the significance of the differences in AUC between our method and the four other methods compared in this study.

Dynamic expression analysis of genes

The data for human and mouse embryonic stem cells were collected at different time points during the differentiation process. For each time point, we calculated the average gene expression across all cells to represent the gene expression profile at that stage.

GO enrichment analysis

Gene enrichment analysis (Falcon and Gentleman, 2008) is a method used to identify overrepresented functional categories (e.g., pathways) among a group of genes, based on genome annotation data. This approach helps identify genes that are actively involved in specific biological functions. Typically, the hypergeometric distribution is used to calculate the p-value for a given gene set in relation to a specific function or pathway. Let N represent the total set of genes, and M the number of genes in N that are associated with a particular function F. Denote x as the number of genes with function F in the differentially expressed gene set, and K as the total number of differentially expressed genes. The p-value for the enrichment of function F is calculated as follows:

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

Based on the results from dynamic expression analysis, we identified marker genes and selected gene sets for enrichment analysis using the "enrichGO" function from the "clusterProfiler" package. A smaller p-value indicates a more statistically significant enrichment of the pathway.

Preprocessing of real datasets

We applied our method, along with four existing algorithms, to analyze three single-cell RNA sequencing datasets. In the first application, we examined a human embryonic stem cell (hESC) gene expression dataset, with measurements taken at the following time points: 0, 12, 24, 36, 72, and 96 hours. This dataset is publicly available in the Gene Expression Omnibus (GEO) under accession number GSE75748 (Chu *et al.*, 2016). Based on the edges in the ground truth, we calculated the degree of each gene node. A higher degree indicates more frequent interactions of the gene within the network, suggesting a potentially pivotal regulatory role. To assess the performance of the proposed method, we selected the top 2000 highly variable genes obtained after preprocessing and sorted them according to their degree in the ground-truth network. From this list, we selected the top 100 genes for further analysis. Next, we extracted the dataset for the selected 100 genes from 714 cells and estimated the correlation structure of these genes using precision matrix through the node-wise regression model. Based on the estimated correlation values, we clustered the genes into five categories using the K-Means method. The categories contained 14, 11, 27, 5, and 43 genes, respectively. In the second application, we analyzed gene expression data from 421 primitive endoderm (PrE) cells derived from mouse embryonic stem cells, measured at five distinct time points: 0, 12, 24, 48, and 72 hours. This dataset is available in GEO under accession number GSE98664 (Hayashi *et al.*, 2018). For the third application, we utilized a preprocessed mouse hematopoietic stem cell (mHSC) gene expression dataset, which is accessible in GEO under accession number GSE81682 (Nestorowa *et al.*, 2016).

Compared methods

In this study, we compared our proposed method with four existing methods using the R software. The algorithms we compared are as follows:

(i) scLink (Li and Li, 2021). This widely used algorithm in gene coexpression network studies first normalizes the gene expression matrix, followed by a $\log_{10}$ transformation. It then calculates the paired correlation coefficients of different genes across various cells to generate the correlation matrix.

(ii) GENIE3 (Huynh-Thu *et al.*, 2010). GENIE3 decomposes the task of predicting a gene network between p genes into p separate regression problems. It uses tree-based ensemble methods to infer edges between genes and determines the weights of regulatory connections.

(iii) SCODE (Matsumoto *et al.*, 2017). SCODE integrates the transformation of linear Ordinary Differential Equations (ODEs) and linear regression to process differentiated cells. It considers the timing of each cell when inferring the regulatory relationships between genes.

(iv) GRNBoost2 (Moerman *et al.*, 2019). GRNBoost2 is an extension of GENIE3 that utilizes Gradient Boosting Machine (GBM) regression for gene regulatory network inference. It enhances computational efficiency and scalability while accurately identifying regulatory interactions. GRNBoost2 has been specifically applied to scRNA-seq data for efficient gene network inference.

Evaluation indicators

We used receiver operating characteristic (ROC) curves, the number of correctly identified edges, and enrichment analysis to compare the performance of different methods. For the same set of genes, we applied different methods to measure the correlations. Based on the estimated correlation values, we generated five ROC curves by calculating the true positive rate and false positive rate for each method, varying the thresholding parameter. The method with the largest area under the ROC curve (AUC) is considered to have the best performance. To quantify the accuracy of the different methods, we ranked the absolute values of the precision matrix calculated by each method in descending order. We then selected a predetermined number of gene pairs with true network edge connections from the top-ranked pairs. Afterward, we calculated the number of true positive edges with strong correlations. Additionally, we performed enrichment analysis using the "enrichGO" function to compare the performance of the methods. Specifically, we verified whether the original characteristics of the genes were consistent with the enriched pathways.

Discussion and conclusions

The high variability inherent in scRNA-seq data necessitates more accurate methods for inferring gene coexpression networks from single-cell gene expression data. In this study, we introduce dGGAPM, a novel method that leverages sparse regression and precision matrix to enhance the accuracy and reliability of inferring gene-gene and gene-module interactions. The contributions of dGGAPM are three-fold when compared to existing methods: (i). dGGAPM leverages precision matrix to effectively capture the linear relationships between genes, particularly in gene expression sequences with both time series data and missing values, such as ESC data, (ii). The dGGAPM method excels at identifying gene modules with distinct biological functions, enhancing our understanding of the underlying gene regulatory networks, (iii). dGGAPM outperforms other methods in handling gene expression data with small sample sizes, accurately reflecting the dynamic properties of gene coexpression networks despite limited data. Notably, independent benchmark studies have revealed that the accuracy of most gene network inference methods often approaches that of a random predictor (Pratapa *et al.*, 2020; Chen and Mar, 2018). The relative performance of coexpression network inference algorithms can vary significantly depending on the test dataset and the evaluation methods used (Chen and Mar, 2018). In general, existing methods tend to perform better on simulated datasets than on experimentally obtained scRNA-seq datasets (Pratapa *et al.*, 2020). Although dGGAPM does not incorporate regulatory directionality between genes, it still effectively identifies marker genes with significant regulatory changes based on the inferred gene coexpression network. One of the key strengths of dGGAPM is its ability to account for confounding factors, providing a more accurate reflection of gene interactions. Additionally, it enables the identification of specific functions for each gene module, which is particularly valuable in uncovering potential pathogenic factors in life science research. Furthermore, dGGAPM is not limited to ESC datasets and can be applied to other single-cell data for more accurate gene network inference. In future studies, we plan to include more genes for analysis and explore the correlation between gene modules and cell-level metadata, such as cell type. By improving the accuracy of gene network inference, particularly in reflecting the regulatory directionality between genes, dGGAPM could uncover further insights in subsequent studies.

Competing interests

There is NO Competing Interest.

Author contributions statement

JLH (Jialu HU) conceived the idea and provided funding support. JLH and Jie He (JH) jointly designed the experiments, FR developed the method, implemented the software, JLH, JH, JL, and FR wrote the manuscript. FR and YL performed applications on real data. BL, TW, and YW contributed to many discussions to improve the manuscript.

Funding

This work has been founded by National Natural Science Foundation of China (Grant No. 62572398, 12501407) and Natural Science Foundation of Jiangsu Province (BK20241363).

Competing interests

The authors declare that they have no competing interests.

References

Aibar, S. *et al.* (2017). Scenic: single-cell regulatory network inference and clustering. *Nature methods*, **14**(11), 1083–1086.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, **74**(1), 47.

Azizi, E. *et al.* (2017). Bayesian inference for single-cell clustering and imputing. *Genomics and Computational Biology*, **3**(1), e46–e46.

Brown, B. L. and Hendrix, S. B. (2005). Partial correlation coefficients. *Encyclopedia of statistics in behavioral science*.

Chen, S. and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, **19**, 1–21.

Chin, A. M. *et al.* (2017). Morphogenesis and maturation of the embryonic and postnatal intestine. In *Seminars in cell & developmental biology*, volume 66, pages 81–93. Elsevier.

Chu, L.-F. *et al.* (2016). Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, **17**(1), 1–20.

Delgado, F. M. and Gómez-Vela, F. (2019). Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial intelligence in medicine*, **95**, 133–145.

Eferl, R. and Wagner, E. F. (2003). Ap-1: a double-edged sword in tumorigenesis. *Nature Reviews Cancer*, **3**(11), 859–868.

Falcon, S. and Gentleman, R. (2008). Hypergeometric testing used for gene set enrichment analysis. In *Bioconductor case studies*, pages 207–220. Springer.

Fortini, P. *et al.* (2012). The plasticity of dna damage response during cell differentiation: Pathways and consequences.

Greenfield, A. *et al.* (2010). Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, **5**(10), e13397.

Hayashi, T. *et al.* (2018). Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, **9**(1), 1–16.

Hecker, M. *et al.* (2009). Gene regulatory network inference: data integration in dynamic modelsa review. *Biosystems*, **96**(1), 86–103.

Huynh-Thu, V. A. *et al.* (2010). Inferring regulatory networks from expression data using tree-based methods. *PloS one*, **5**(9), e12776.

Iacono, G. *et al.* (2019). Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome biology*, **20**, 1–20.

Jaremko, K. L. and Marikawa, Y. (2013). Regulation of developmental competence and commitment towards the definitive endoderm lineage in human embryonic stem cells. *Stem Cell Research*, **10**(3), 489–502.

Kharchenko, P. V. *et al.* (2014). Bayesian approach to single-cell differential expression analysis. *Nature methods*, **11**(7), 740–742.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, **9**, 1–13.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, **9**(1), 997.

Li, W. V. and Li, Y. (2021). sclink: Inferring sparse gene co-expression networks from single-cell expression data. *Genomics, proteomics & bioinformatics*, **19**(3), 475–492.

Liao, J. *et al.* (2022). The mouse anxa6/mir-9-5p/anxa2 axis modulates tgf-β1-induced mouse hepatic stellate cell (mhsc) activation and ccl4-caused liver fibrosis. *Toxicology Letters*, **362**, 38–49.

Matsumoto, H. *et al.* (2017). Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, **33**(15), 2314–2321.

Moerman, T. *et al.* (2019). Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, **35**(12), 2159–2161.

Nestorowa, S. *et al.* (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, **128**(8), e20–e31.

Palli, S. R. (2021). Epigenetic regulation of post-embryonic development. *Current opinion in insect science*, **43**, 63–69.

Pan, G. and Thomson, J. A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell research*, **17**(1), 42–49.

Parsana, P. *et al.* (2019). Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome biology*, **20**(1), 1–6.

Peng, J. *et al.* (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, **104**(486), 735–746.

Pratapa, A. *et al.* (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, **17**(2), 147–154.

Qiu, Y. and Zhou, X.-H. (2020). Estimating c-level partial correlation graphs with application to brain imaging. *Biostatistics*, **21**(4), 641–658.

Quintana, L. *et al.* (2009). Morphogenetic and regulatory mechanisms during developmental chondrogenesis: new paradigms for cartilage tissue engineering. *Tissue Engineering Part B: Reviews*, **15**(1), 29–41.

Shangguan, Y. *et al.* (2020). Application of single-cell rna sequencing in embryonic development. *Genomics*, **112**(6), 4547–4551.

Specht, A. T. and Li, J. (2015). Estimation of gene co-expression from rna-seq count data. *Statistics and Its Interface*, **8**(4), 507–515.

Takasato, M. *et al.* (2014). Directing human embryonic stem cell differentiation towards a renal lineage generates a self-organizing kidney. *Nature cell biology*, **16**(1), 118–126.

Vicente, C. *et al.* (2012). The role of the gata2 transcription factor in normal and malignant hematopoiesis. *Critical reviews in oncology/hematology*, **82**(1), 1–17.

Wen, W. *et al.* (2020). Immune cell profiling of covid-19 patients in the recovery stage by single-cell sequencing. *Cell discovery*, **6**(1), 31.

Yang, M. *et al.* (2022). Chemical-induced chromatin remodeling reprograms mouse escs to totipotent-like stem cells. *Cell Stem Cell*, **29**(3), 400–418.

Zhang, Y. *et al.* (2021a). Mk2 promotes tfcp2l1 degradation via β-trcp ubiquitin ligase to regulate mouse embryonic stem cell self-renewal. *Cell Reports*, **37**(5).

Zhang, Y. *et al.* (2021b). Single-cell rna sequencing in cancer research. *Journal of Experimental & Clinical Cancer Research*, **40**, 1–17.

Zhao, J. *et al.* (2023). Dynamic metabolism during early mammalian embryogenesis. *Development*, **150**(20), dev202148.

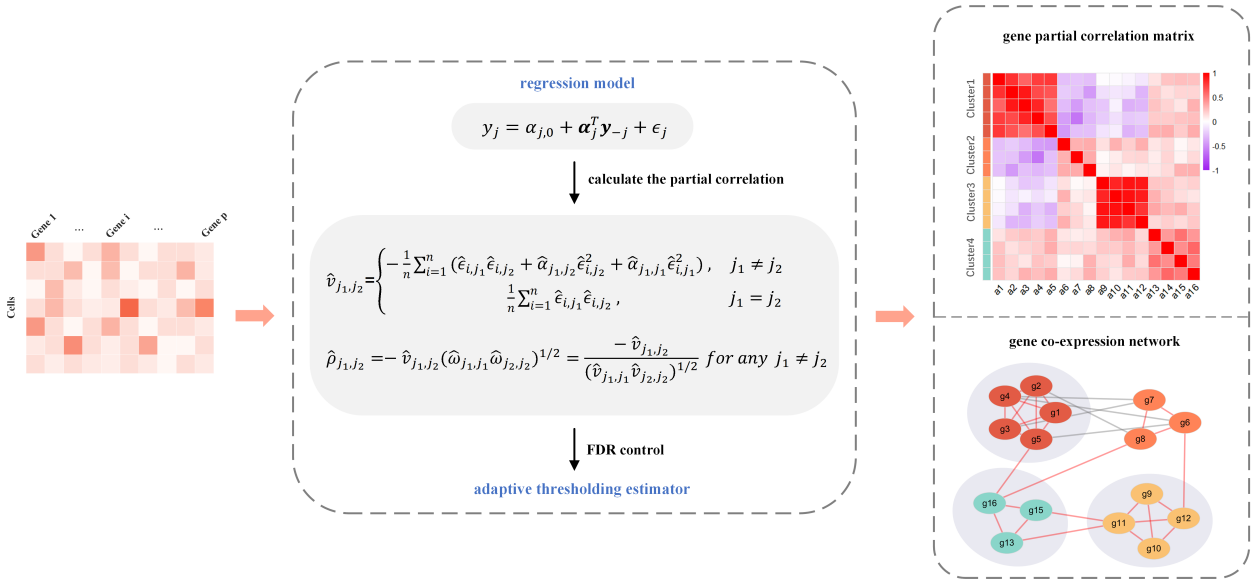

Figure 1: **dGGAPM schematic.** dGGAPM is a method to infer gene coexpression networks from single-cell gene expression data. The framework begins by utilizing a sparse linear regression model to compute regression coefficients and residuals based on the gene expression matrix. Next, the sample covariance of the residuals is calculated to derive precision matrix between genes. Subsequently, the K-Means algorithm is applied to cluster genes, resulting in a category-sorted gene correlation heatmap. Finally, an adaptive thresholding estimator is employed to determine the optimal threshold, constructing a gene network that captures positive and negative relationships among genes. This framework facilitates the exploration of gene interactions, providing a foundation for further studies on gene modules and their interrelationships. By uncovering these connections, dGGAPM offers valuable insights into the underlying mechanisms of gene regulation and coexpression.

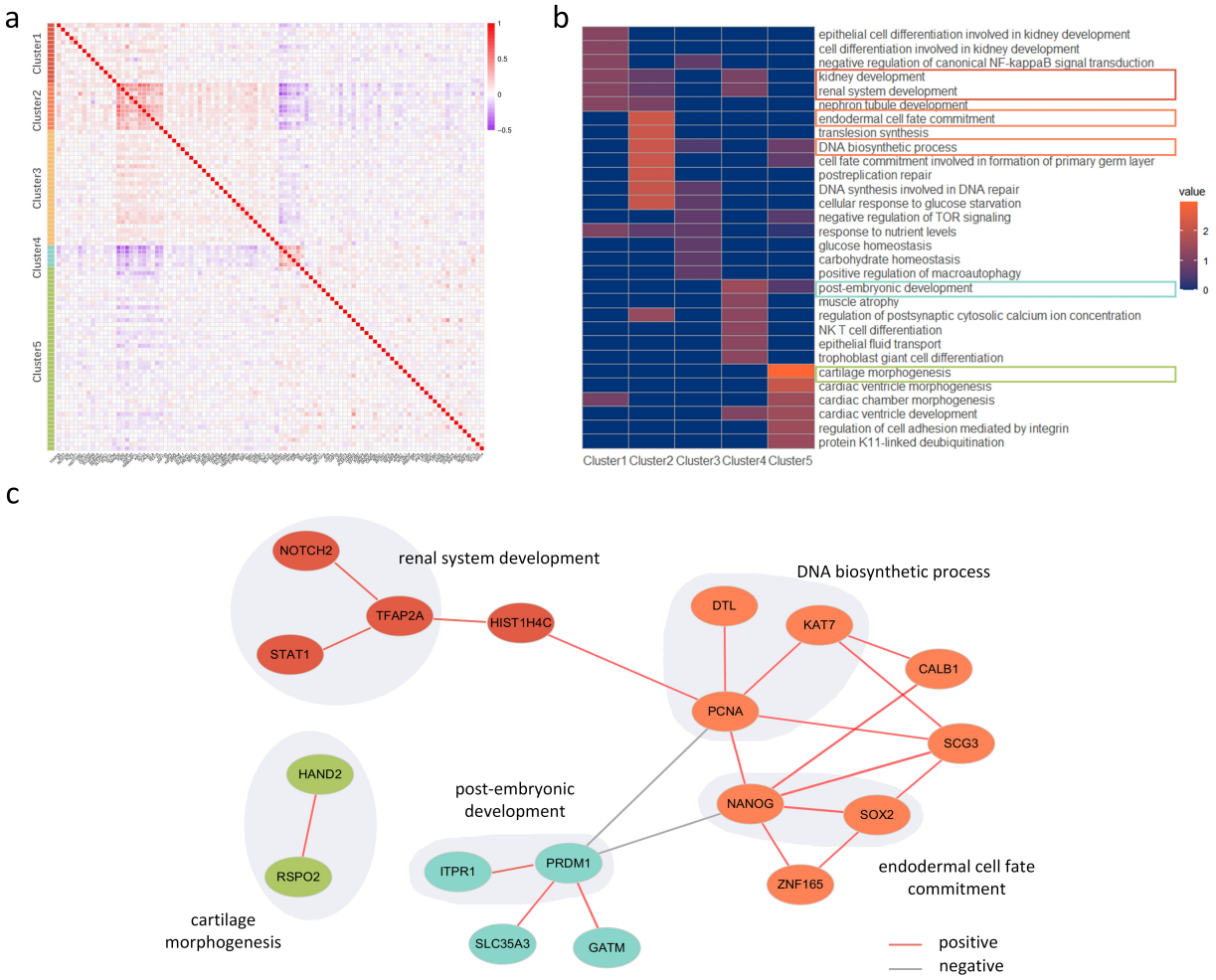

Figure 2: **Identification of gene coexpression networks from human embryonic stem cells.** **a**, Correlation heatmap of 100 genes in the hESCs data, where these genes are clustered into five groups. **b**, After performing enrichment analysis on the gene sets from the five categories, the top 6 pathways from each category were selected for merging, and the corrected p-values of the paths were represented as -log10. The larger the value, the more significant the pathway, and the brighter the color in the figure. **c**, A partial gene coexpression network colored by different categories, with red edges representing positive correlations and gray edges representing negative correlations. The shaded areas represent submodules composed of different genes.

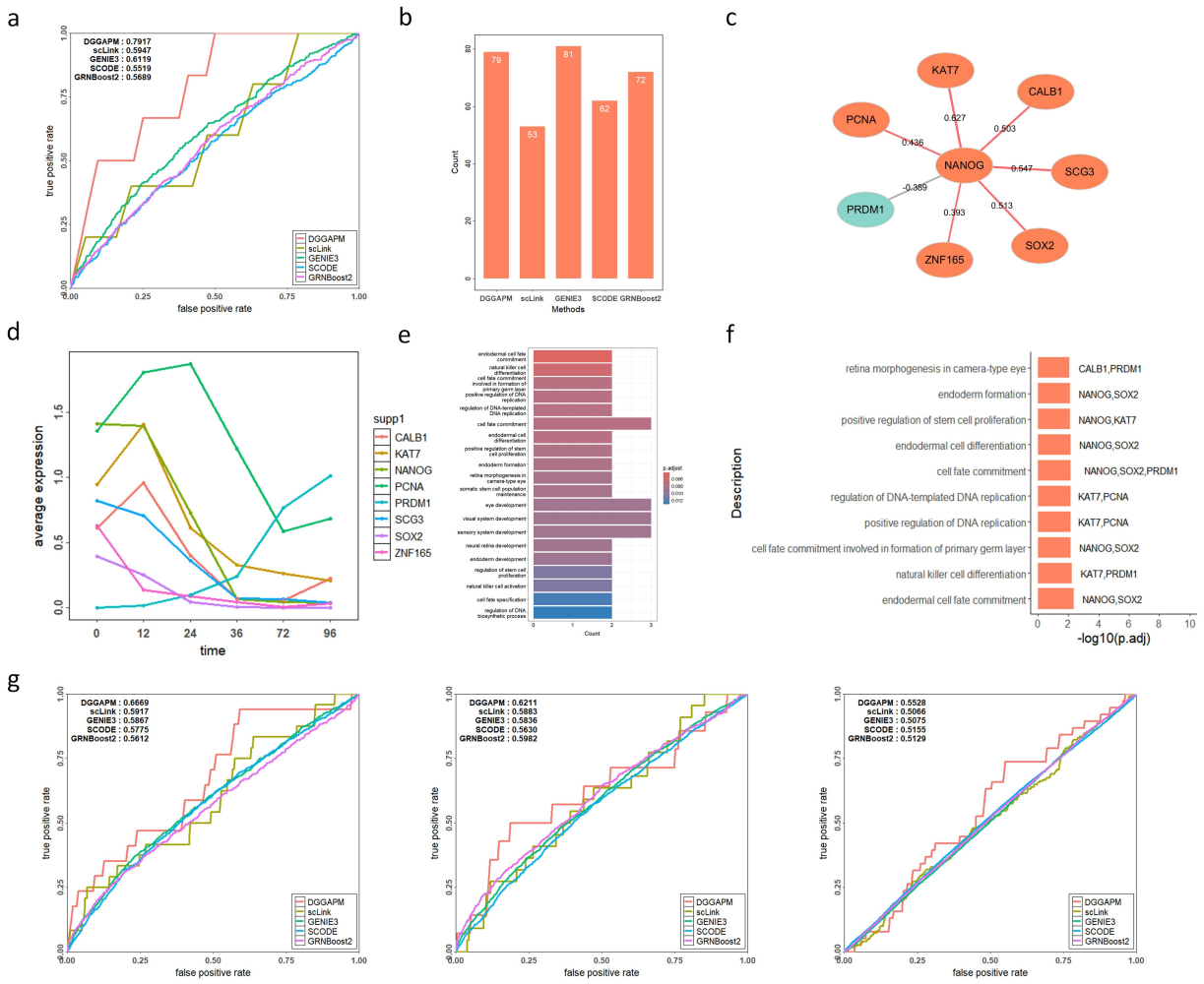

Figure 3: **Validation of the algorithm in human embryonic stem cells.** **a**, ROC curves and AUC values obtained by applying our algorithm, along with scLink, GENIE3, SCODE, and GRNBoost2 algorithms, on 714 cells and 100 genes. **b**, After comparison with the ground truth in the dataset, the number of correct edges identified by the five algorithms among the top 1000 edges is shown. **c**, The network diagram displays seven genes associated with the NANOG gene and the magnitude of their correlations. Node colors represent the categories to which the genes belong, with darker edge colors indicating stronger correlations. **d**, The line chart represents the average expression values of 8 genes at each time point (0, 12, 24, 36, 72, and 96 hours). **e**, The eight genes were analyzed for enrichment in terms of biological processes, and the top 20 pathways were selected for bar graphs. **f**, The bar chart displays the top ten enriched biological processes, with the y-axis representing the pathway names and the x-axis representing the size of the corrected p-values in -log10. Data labels represent the genes involved in each pathway. **g**, Gene expression data from different numbers of genes were selected for the experiment to evaluate the method further. Specifically, the top 200 and top 400 genes were selected among the highly variable genes sorted by their degree in the real network, and the top 400 genes were selected based on the degree of nodes in the real network.

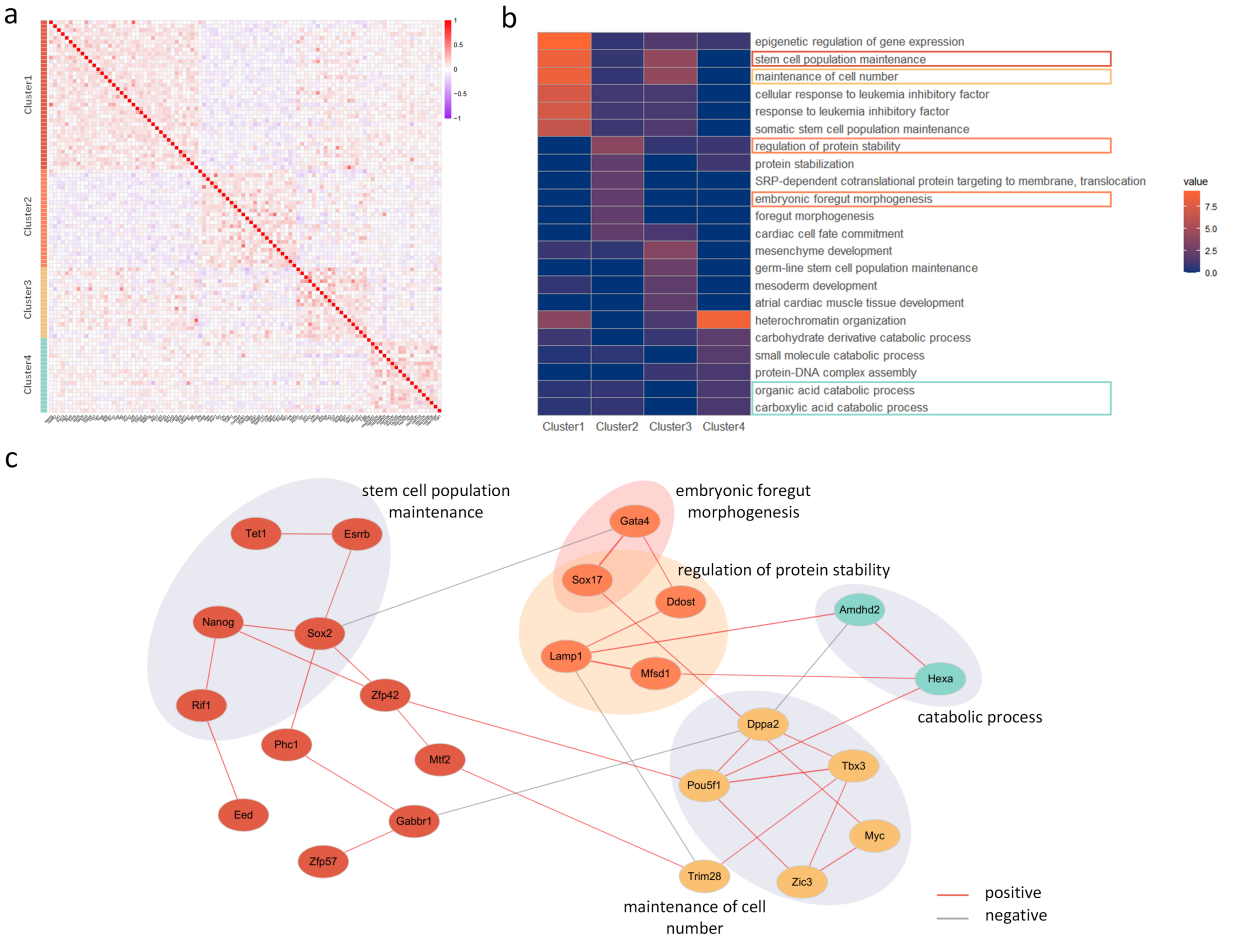

Figure 4: **Identification of gene co-expression networks from mouse embryonic stem cells. a**, Correlation heatmap of 100 genes divided into four groups in the mESCs dataset. **b**, After enrichment analysis of the four modules, the top 6 pathways from each module were merged. The corrected pathway p-value is represented as -log10. The larger the value, the more significant the pathway, with brighter colors indicating higher significance. **c**, A partial gene co-expression network with genes colored according to different categories. Red edges represent positive correlations, while gray edges indicate negative correlations. Shaded areas denote submodules composed of varying gene sets.

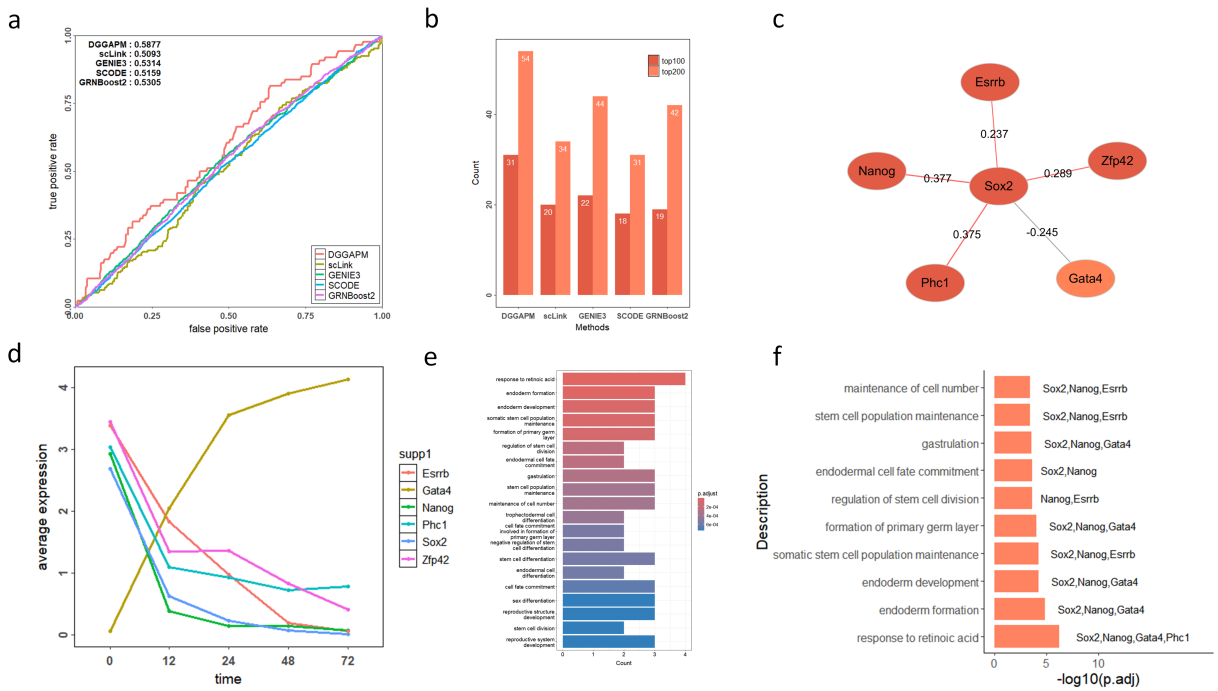

Figure 5: **Validation of the algorithm in mouse embryonic stem cells. a**, ROC curve and AUC analysis of 421 cells and 100 genes using our algorithm alongside scLink, GENIE3, SCODE, and GRNBoost2. **b**, Comparison with ground truth data from the dataset, showing the number of correctly identified edges among the top 100 and 200 predicted edges for each of the five algorithms. **c**, Network diagram illustrating five genes associated with Sox2 and their correlation strengths. Node colors represent the gene categories, while darker edge colors indicate stronger correlations. **d**, Line chart depicting the average expression levels of six genes at 0, 12, 24, 48, and 72 hours. **e**, Enrichment analysis of these six genes in biological processes, with the top 20 pathways visualized in a bar chart. **f**, Bar chart displaying the top 10 enriched biological pathways. The y-axis represents pathway names, while the x-axis indicates the adjusted p-value (-log10 transformation). Data labels denote the number of genes involved in each pathway.

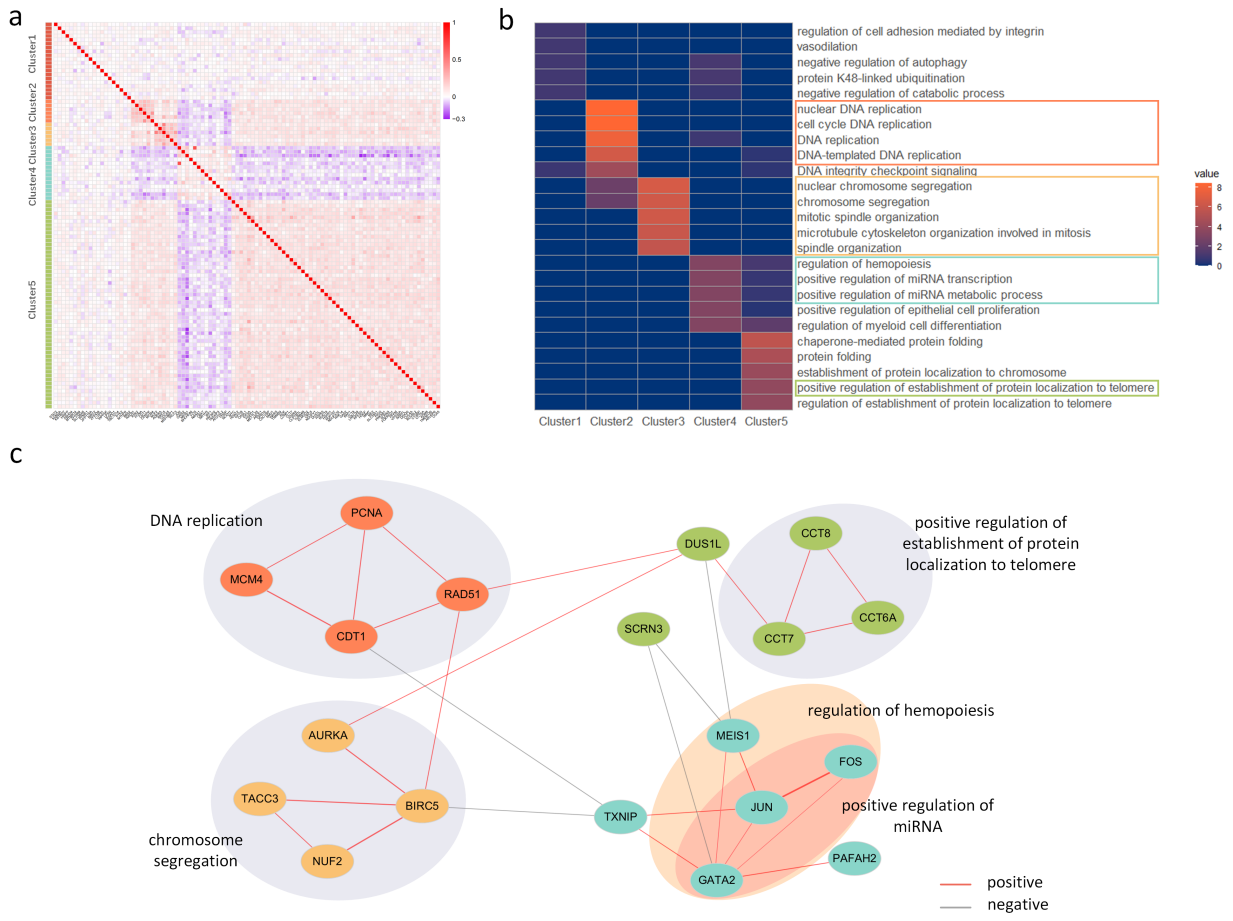

Figure 6: **Identification of gene co-expression networks from mouse hematopoietic stem cells.** **a**, Correlation heatmap of 100 genes divided into five groups within the mHSCs dataset. **b**, Enrichment analysis was performed on the five modules, and the top five pathways from each module were merged. The adjusted pathway p-values are represented as -log10. **c**, A partial gene co-expression network with nodes colored according to different categories.

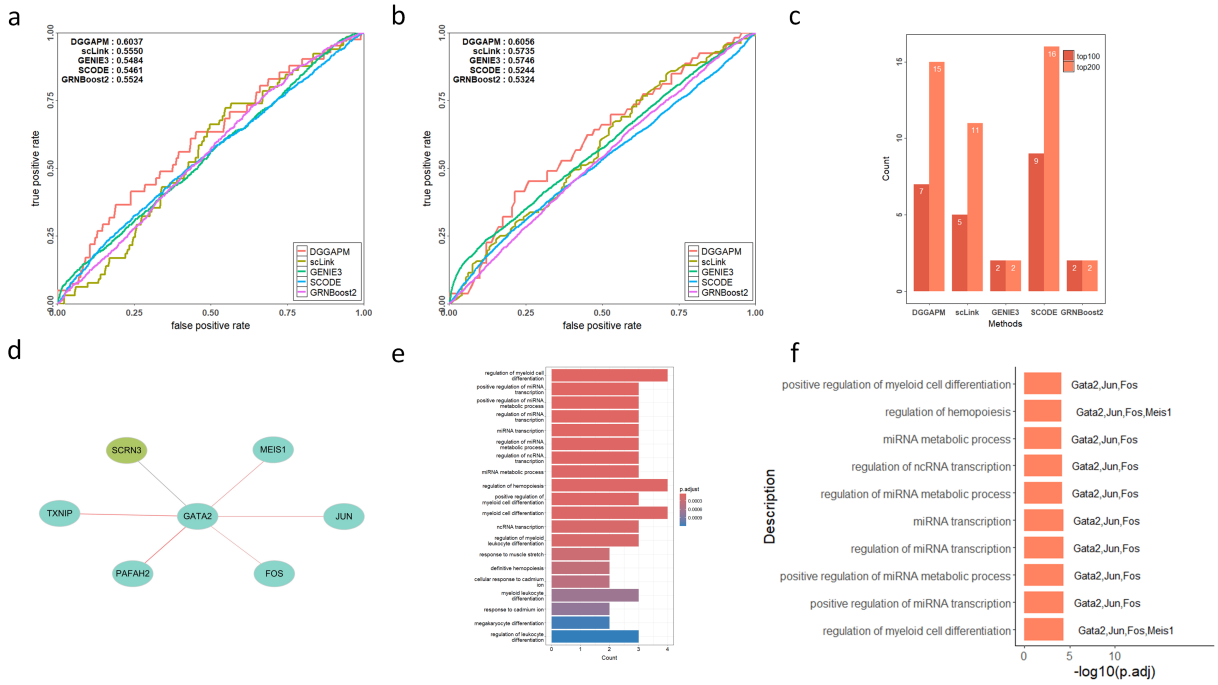

Figure 7: **Validation of the algorithm in mouse hematopoietic stem cells.** We compared our algorithm's performance against four existing algorithms by generating receiver operating characteristic (ROC) curves and calculating the area under the curve (AUC). **a** and **b**, We assessed two gene sets: the top 100 highly variable genes with a high degree in the true network, and the top 100 genes ranked by degree. After comparison with ground truth in the dataset, the correct number of edges in the first 1000 edges was found by the four algorithms. **c**, We compared the number of true positive edges identified by the five algorithms within the top 100 and 200 edges, using the ground truth from the dataset as a reference. **d**, A network graph was constructed to visualize the relationships of six genes associated with the GATA2 gene. **e**, Node colors represent gene categories, and edge colors indicate positive or negative correlations. Gene ontology (GO) enrichment analysis was performed on seven genes, and the top 20 pathways were selected for a bar plot. **f**, The bar plot illustrates the top 10 enriched pathways, with the y-axis representing pathway names and the x-axis representing the -log10 of the adjusted p-values. Data labels indicate the genes involved in each pathway.