Nucleic Acids Research, 2025, **53**, gkaf431
https://doi.org/10.1093/nar/gkaf431
Methods

SCIG: Machine learning uncovers cell identity genes in single cells by genetic sequence codes

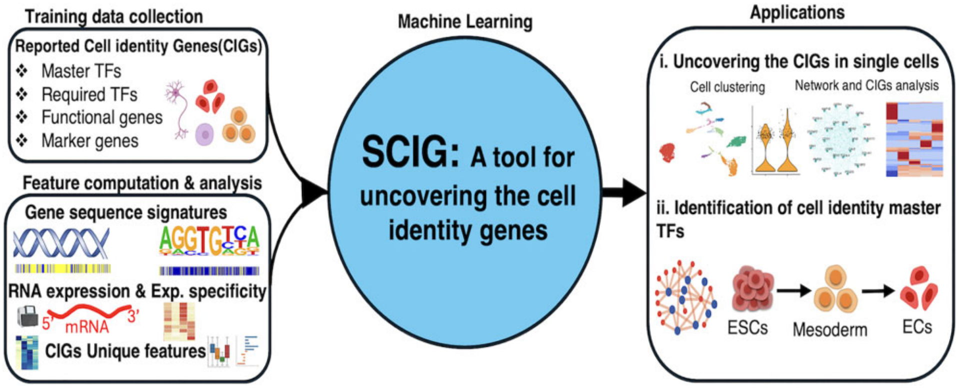
Kulandaisamy Arulsamy [1,2], Bo Xia [3], Yang Yu [1,2], Hong Chen [4], William T. Pu [1,2], Lili Zhang [1,2], Kaifu Chen [1,2,*]

[1] Basic and Translational Research Division, Department of Cardiology, Boston Children's Hospital, Boston, MA 02115, United States
[2] Department of Pediatrics, Harvard Medical School, Boston, MA 02115, United States
[3] Independent Researcher, Clemson, United States
[4] Vascular Biology Program, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, United States
[*] To whom correspondence should be addressed. Email: kaifu.chen@childrens.harvard.edu

Abstract

Deciphering cell identity genes is pivotal to understanding cell differentiation, development, and cell identity dysregulation involving diseases. Here, we introduce SCIG, a machine-learning method to uncover cell identity genes in single cells. In alignment with recent reports that cell identity genes (CIGs) are regulated with unique epigenetic signatures, we found CIGs exhibit distinctive genetic sequence signatures, e.g. unique enrichment patterns of *cis*-regulatory elements. Using these genetic sequence signatures, along with gene expression information from single-cell RNA-seq data, SCIG uncovers the identity genes of a cell without a need for comparison to other cells. CIG score defined by SCIG surpassed expression value in network analysis to reveal the master transcription factors (TFs) regulating cell identity. Applying SCIG to the human endothelial cell atlas revealed that the tissue microenvironment is a critical supplement to master TFs for cell identity refinement. SCIG is publicly available at https://doi.org/10.5281/zenodo.14726426 , offering a valuable tool for advancing cell differentiation, development, and regenerative medicine research.

Graphical abstract


Training data collection
Reported Cell identity Genes(CIGs)
Master TFs
Required TFs
Functional genes
Marker genes
Feature computation & analysis
Gene sequence signatures
AGGTGTCA
RNA expression & Exp. specificity
5'
mRNA
3'
CIGs Unique features
Machine Learning
SCIG: A tool for uncovering the cell identity genes
Applications
i. Uncovering the CIGs in single cells
Cell clustering
Network and CIGs analysis
ii. Identification of cell identity master TFs
ESCs
Mesoderm
ECs

Introduction

Every cell type possesses a distinct collection of several hundred or more cell identity genes (CIGs) governing its cellular characteristics. A recent curation work via thorough literature review suggested four categories of CIGs, including master transcription factors (TFs) whose expression is reported as sufficient to induce a cell type, required TFs whose depletion is reported to impair the differentiation toward a cell type, function genes reported to play a cell type-specific function, and marker genes that serve primarily to identify a specific cell type [1]. A cell might express both its CIGs and many other gene categories such as housekeeping genes and heat shock genes. Accurate expression of CIGs is essential for steering cell differentiation and the development of tissues and organs in living organisms. The CIGs of a cell comprise an intricate gene regulatory network (GRN), where several master TFs coordinate the expression of the CIGs [2–4]. For instance, the expression program of pluripotent stem cells can be established by the master TFs *Oct4*, *Sox2*, *Klf4*, and *c-Myc* [5, 6]. Cell differentiation from stem cells to somatic cell types, *trans*-differentiation between somatic cell types, and reprogramming from somatic cell types to stem cells are orchestrated by their CIG regulatory networks, enabling a source cell type to give rise to a specific target cell type [5]. This dynamic process encompasses silencing and activating the CIGs of the source and target cell types, respectively. Uncovering the CIGs of

Received: August 21, 2024. **Revised:** April 9, 2025. **Editorial Decision:** April 18, 2025. **Accepted:** May 9, 2025

individual cell types is pivotal in regenerative medicine to facilitate the precise generation of desired target cell types from a source cell type [7, 8].

It is increasingly recognized that the same type of epigenetic modification tends to display different patterns between CIGs and most of the other expressed genes in the same cell. For instance, the histone modification H3K4me3 tends to display a broad enrichment pattern of ∼5–100 kb covering both the promoter and the gene body of a CIG, whereas H3K4me3 tends to display a sharp enrichment in only ∼1 kb promoter region at each of the other expressed genes [9, 10]. It is also reported that CIGs tend to be regulated by super-enhancers [6, 11], which show a broad enrichment pattern of the histone modification H3K4me1 and H3K27ac covering a long stretch of enhancers, while most other genes are regulated by typical enhancers each displaying narrow enrichment peaks of these histone modifications [9, 10]. Moreover, CIGs tend to show the lowest RNA stability in the cell because the m6A methyltransferase preferentially modifies these RNAs in a co-transcriptional manner, likely guided by chromatin features, such as the super-enhancers and broad H3k4me3 modification [12]. Researchers have been able to uncover CIGs in a cell type using these epigenetic profiles [1]. However, generating these epigenetic profiles often entails a substantial investment of time and resources. Moreover, profiling multiple histone modifications in parallel from each single cell at the genome-wide scale is still a technological challenge. Therefore, uncovering CIGs by their epigenetic profiles, especially when applied to single-cell study, may be limited due to the difficulty of generating the epigenetic profile data.

Uncovering CIGs solely based on their expression in a cell also presents a great challenge since the expression level might resemble those of many other genes such as the housekeeping genes. Existing computation solutions often aim at defining cell type-specific marker genes by comparison between cell types at the bulk level or between cell clusters in a single-cell RNA sequencing (scRNA-seq) dataset. These solutions are not optimal because different CIGs display great differences in the degree of expression specificity. For instance, *c-Myc* is a well-known identity gene of embryonic stem cells (ESCs) [5] but also expressed to regulate the identity or play functions in some other cell types, such as fibroblasts [13, 14] and hematopoietic stem cell [15, 16]. Cell identity is often governed by the cell type-specific combination of identity genes, for which the combination is cell type-specific, but many CIGs might be not strictly cell type-specific. Therefore, some CIGs might be missed from the list of cell type-specific genes. Furthermore, the defined cell type-specific genes for a cell type may vary when compared to different sets of other cell types. It is also a challenge to distinguish between cell type-specific and experimental condition-specific genes, e.g. heat shock genes. Other bioinformatics methods aim at constructing the gene expression regulation network to define pivotal regulators [3]. This solution is yet not optimal for defining CIGs because the gene regulation network may include both a sub-network of CIGs and some sub-networks of other gene categories, e.g. cell cycle genes, stress response genes, and genes for DNA damage repair. It is yet a challenge to distinguish between these sub-networks. Therefore, despite the widespread application of scRNA-seq in recent years to define cell type-specific genes or expression networks, there is no existing method that employs a single-cell transcriptome to directly uncover CIGs.

Genes belonging to some functional categories were reported to show some unique genetic sequence features. For example, housekeeping genes tend to show lower promoter sequence conservation [17], sequences with less potential for nucleosome formation in the promoter region [18], and shorter gene bodies [19–21]. Recent studies demonstrated that genetic sequence signatures hold promise to be *cis*-regulatory codes for some aspects of biological function, e.g. RNA structure [22–24], protein-binding [25], RNA stability [26], and promoters and enhancers activities [27, 28]. Intriguingly, CIG tends to show a broad enrichment of TF-binding motifs across the gene body, in contrast to the narrow enrichment of these motifs in only the promoters of most other genes [9]. Exploring these genetic sequence signatures has provided valuable clues to understanding CIG expression regulation and might facilitate the identification of CIGs. Therefore, we reason that integrating gene expression information with genetic sequence codes to uncover CIGs is a promising approach offering significant advantages over prior methods.

In this study, we demonstrated the effectiveness of combining genetic sequence signatures with expression information for distinguishing between CIGs and the rest of the genes. Building upon this characterization, we developed a robust machine-learning algorithm, SCIG, for uncovering single-cell identity genes. Applying SCIG to diverse single-cell datasets, including individual subtypes of the endothelial cell (EC) lineage, our analysis generated new insights into the expression regulation of CIGs in cell differentiation. Notably, cell clustering results based on CIG scores more accurately reflected cell types and their differentiation trajectory when compared to the conventional strategy based on RNA expression values alone. The CIG catalog uncovered by SCIG holds great promise for advancing cell identity reprogramming experiments and facilitating the generation of desired cell types in regenerative medicine.

Materials and methods

Analyzing the genetic and expression features of cell identity genes

We obtained a set of CIGs and control genes manually curated for 10 human cell types in a recent work [1] (Supplementary Tables S1 and 2). We computed five categories of features from RNA expression and gene sequence data to capture the characteristics of CIGs.

i. Gene expression values: The bulk RNA-seq raw reads for 10 human cell types, including B cells, ECs, epithelial cells, ESCs, fibroblasts, hematopoietic stem cells, mesenchymal stem cells, neuronal cells, mid radial glial cells, and skeletal muscle myoblasts, were collected from the Encyclopedia of DNA Elements (ENCODE) project [29] and NCBI Sequence Read Archive (SRA) [30] repositories (Supplementary Table S3). The human reference genome version hg38.104 was downloaded from the Ensembl database [31], and the longest transcript for each protein-coding gene was selected for further analysis. We employed Trim Galore (version 0.6.6) for adaptor trimming and removal of low-quality sequencing reads, followed by mapping the high-quality reads to the human reference genome using the STAR alignment tool (version 2.7.9a) [32]. Quantification of the mapped reads was performed using featureCounts (v2.0.3) [33] for the entire gene and exon regions. The resulting read counts were then converted into log2-transformed transcript per million (TPM) values for model development and feature analysis. For the mouse genome, expression data for the aforementioned 10 cell types were gathered from the Mouse Cell Atlas (MCA) database [34].

ii. Gene expression specificity metrics: Gene expression values for human tissues and cell types were sourced from various RNA-seq projects, including the Genotype-Tissue Expression (GTEx) [35], Functional Annotation of Mammalian Genomes 5 (FANTOM5) [36], and Human Protein Atlas [37, 38]. These data were downloaded from The Human Protein Atlas website (https://www.proteinatlas.org/about/download) [37, 38]. For mouse tissues and cell types, the expression data were obtained from MCA [34], Mouse ENCODE [39], and mammalian transcriptomic database [40]. Next, the tspex [41] python package was used to calculate various gene expression specificity metrics, including tau, Gini coefficient, Simpson index, Shannon specificity, and Roku specificity, yielding specificity scores ranging from 0 to 1. Higher scores indicate that genes are expressed more specifically in certain tissues or cell types.

iii. Binding motifs and sites of TFs, RNA-binding proteins, and Micro-ribonucleic acid (miRNAs): The human and mouse TF-binding motifs were obtained from the "MotifDb [42] package and were used to scan the entire human and mouse genome sequences. The analysis focused on determining the number of TF-binding motifs within different genomic regions, including the promoter region with various window lengths (500 bp to 5 kb in both upstream and downstream sequences), gene bodies, 5′-UTR, exons, introns, and 3′-UTR. The binding sites of RNA-binding proteins were downloaded from the oRNAment database [43], and the number of binding sites within specific genomic regions (5′-UTR, exons, introns, and 3′-UTR) was determined using the bedtools intersect command [44]. Additionally, miRWalk [45] was employed to retrieve miRNA-binding sites. Thereafter, the number of miRNA-binding sites within the 5′-UTR, coding sequence (CDS), and 3′-UTR regions was calculated. The number of these features were normalized based on the length of the respective genomic regions.

iv. Evolutionary features: The phyloP100way sequence conservation score for the human and mouse genome was obtained from the UCSC database [46]. Using in-house scripts, we computed the mean and median conservation scores for a whole gene sequence and different genomic regions.

v. Generic gene-level features: The gene architecture features, including gene sequence length and the lengths of different genomic elements (exon, intron, CDS, 5′-UTR, and 3′-UTR), were calculated from the human and mouse reference genome [31]. Additionally, the number of introns and exons, as well as the ratio of introns to exons, were also computed. For each gene, we computed the frequencies of single-, di-, and tri-nucleotides in the gene sequence. Codon biases were determined by grouping triplet codons based on their coding amino acid and calculating their total proportion, mean, median, and coefficient of variation. Additionally, we included the transcription start site (TSS) distance, which represents the number of base pairs between the TSS of a gene and its closest TSS in the chromatin.

Features filtering and selection

Based on the feature extraction procedures described above, a total of 680 features were obtained for each gene. Initially, a Wilcoxon [47] nonparametric test was conducted to determine the significance of the difference in each feature between CIGs versus control or housekeeping genes. Features with a *P*-value >0.05 when comparing CIGs to control or housekeeping genes were subsequently removed. Next, we calculated the Pearson correlation coefficient between features. For a pair of features with a correlation coefficient >0.90, the feature with the greater *P*-value of its difference between CIGs and control genes was removed. These steps resulted in a final set of 73 nonredundant and significant features.

In our pursuit of identifying potential features for machine learning model development, we employed a systematic search followed by a forward feature selection algorithm [48]. The systematic search involved exploring all possible combinations of three features, resulting in a selection of the top 10 000 combinations based on the performance metric, Matthew's correlation coefficient, in 10-fold cross-validation. Subsequently, we implemented the forward selection algorithm to gradually incorporate additional features from the pool into the selected combinations. The forward feature selection process continued until no further enhancement in model performance was observed, resulting in a final set of selected features. To assess the efficacy of this feature selection method, we compared it with several other approaches, including the SK-best-mutual information, analysis of variance (ANOVA) F-classification, forward sequential selection (FSS), backward stepwise elimination (BSE), and recursive feature elimination based on Logistic Regression (RFE LR) with L1 regularization [49].

Developing the SCIG, a logistic regression model to uncover CIGs

After feature selection, the training dataset was standardized using Sklearn's preprocessing libraries [49]. A logistic regression algorithm with an L1 penalty parameter was applied to develop the model, aiming to prevent overfitting. The model selection involved a 10-fold cross-validation procedure, and further validation was conducted through bootstrapping with 500 iterations, using an 80% training and 20% testing split. Model performance was assessed using statistical metrics such as sensitivity, specificity, accuracy, Matthew's correlation coefficient, F1-score, and area under the receiver operating characteristic curve (AUROC).

Collecting reported master transcription factors of cell identity

The human TFs were obtained from the humantfs database (http://humantfs.ccbr.utoronto.ca/) [50]. We extracted the reported master TFs of cell identity manually curated for 10 cell types in recent work [1]. The remaining TFs from the human TF list were recognized as the control TFs in these 10 cell types. As the number of control TFs exceeds the number of master TFs, we addressed the class imbalance issue by utilizing the SMOTE algorithm [51], resulting in a balanced dataset of 221 randomly selected master and control TFs for model development.

Machine learning model to uncover master transcription factors in CIG networks

We compiled a reference GRN from seven different resources, including RegNetwork [52], DoRothEA [53], CellNet [4], GRNdb [54], ANANSE [55], PANDA [56], and Huang, J.K. *et al.* [57]. We focused on protein-coding gene interactions and determined the reliability of each interaction based on the number of votes from these sources, taking only the interactions that were present in at least two sources. For each known master TFs and control TF, we extracted their corresponding GRNs from the compiled reference GRN. Next, we computed several features including the number of children edges (indicating the number of target genes regulated by a specific TF), the number of parent edges (indicating the number of genes regulating a specific TF), cell identity gene score (CIG score) from SCIG, and RNA expression of the genes. Furthermore, we calculated the mean, median, and coefficient of variation for each feature of the TFs. This process yielded a total of 23 features for each master and control TF, which were subsequently used for constructing a logistic regression model to learn network features enriched in master TFs. Important features to use in the model were selected through systematic feature selection using 10-fold cross-validation. The model's performance was evaluated using statistical measures like accuracy, Matthew's correlation coefficient, and F1-score during both 10-fold cross-validation and bootstrapping procedures.

Exploring single-cell identity gene networks in human fetal heart using hdWGCNA

First, we employed the SCIG algorithm to calculate cell identity scores for individual genes in individual cells from human fetal hearts [58] (Fig. 4A). Then, we supplied the CIG score matrix and expression matrix independently to hdWGCNA [59] for gene network analysis at the single-cell level. Initially, we selected the genes that expressed in a minimum of 5% of cells in both the expression and CIGs score matrices. Subsequently, we generated metacells using the "MetacellsByGroups ($k = 20$, min_cells = 30)" function, followed by normalization using "NormalizeMetacells." Notably, normalization was omitted for the CIG score matrix as it was inherently in normalized form. Further, we constructed the network by considering all cells with default options. Each identified network module was visualized using the "ModuleNetworkPlot" function.

Using cell identity gene score for single-cell clustering analysis

The human forebrain glutamatergic neurogenesis [60] dataset consisted of 1720 scRNA-seq profiles, which were utilized to predict CIGs using the SCIG algorithm. This transformed the gene–cell expression matrix into a gene–cell CIG score matrix. We then analyzed the expression matrix and CIG score matrix independently using the Seurat [61] package for cell clustering and cell type annotation. Only protein-coding genes were analyzed, while mitochondrial and ribosomal genes were filtered out. For the expression matrix, expression values were log-normalized, and the top 2000 highly variable genes were selected using the "NormalizeData" and "FindVariableFeatures" functions, respectively, with default parameters. The normalized expression matrix was then scaled and centered using the "ScaleData" function, followed by running the RunPCA (npcs = 50), RunUMAP (reduction = pca, dims = 1:40), FindNeighbors (k.param = 20, nn.method = annoy, annoy.metric = euclidean), and FindClusters (resolution = 1) function to obtain cell clusters. We used the default values for each function unless otherwise specified. In the case of the CIG score matrix, the same clustering pipeline was employed except that the log-normalization step was skipped, as the CIG scores were already standardized. The identified cell clusters were annotated based on their RNA expression levels or CIG scores. The cell types identified based on each matrix were further analyzed using ScVelo [62] to determine their future state or transition direction based on spliced and unspliced mRNA expression values. The velocity information obtained was then utilized in CellRank [63] to identify the potential initial and terminal sites/cells within the given cell types. Additionally, CellRank quantified the transition probability for each terminal site from all other cell types, providing valuable insights into cell dynamics and potential cell state transitions.

CIG analysis in the process of endothelial differentiation

We obtained scRNA-seq profiles during endothelial differentiation from human ESCs (H9) at multiple time points, including days 0 (ESCs), 4 (mesoderm), 6 (mesenchymal), 8 (EC progenitors), and 12 (ECs) [64]. The unique molecular identifier (UMI) count matrix were preprocessed and transformed into pseudo-bulk datasets for each time point. These datasets were subsequently utilized in the SCIG model to predict CIGs and their master TFs.

Endothelial cell identity gene landscape analysis in 15 human tissue types

We obtained scRNA-seq data from 15 tissues, encompassing adipose, bladder, breast, gut, heart, intestine, kidney, liver, lung, ovary, skeletal muscle, skin, stomach, testis, and thymus, sourced from the DISCO [65] database. Endothelial subtypes, including arterial, capillary, venous, and lymphatic, were specifically extracted from each tissue. Subsequently, we aggregated gene UMI counts for each endothelial cell subtype from the count matrix. This consolidated dataset was then integrated into SCIG for the prediction of CIGs and their master TFs. To determine the specificity of each CIG and their master TFs across 15 tissues, we retrieved the SCIG-derived scores for the genes [with false discovery rate (FDR) < 0.05] in each endothelial subtype. Subsequently, these retrieved scores of CIGs and master TFs were utilized for Tau score calculation using tspex [41]. We conducted the Wilcoxon nonparametric one-sided test for assessing the statistical difference between the Tau specificity score of CIGs and CIG master TFs.

Pathway enrichment analysis

The significant CIGs (with FDR < 0.05), high-expression genes, and highly variable genes were subjected to a pathway enrichment analysis using the ClusterProfiler [66] R package. The enriched Gene Ontology (GO) pathways were determined based on the least adjusted P-values using the Benjamini method, with a significance threshold of q-value < 0.05. This pipeline helps identify the biological processes and functions associated with the identified CIGs and other genes. The fold enrichment for each pathway was computed by calculating the ratio between the gene ratio and the background ratio obtained from ClusterProfiler output.

Statistical analysis

For each genetic sequence and expression-derived feature, we used two-tailed Wilcoxon's test [47] for assessing the statistical difference (P-value) between CIGs and housekeeping, control genes. From SICG output, the significant CIGs and their master TFs were identified by FDR < 0.05. We utilized pROC [67] package for determining the P-values of ROC curves. During the pathway enrichment analysis, we used the q-value threshold of <0.05 for selecting the significant pathways.

Results

CIGs display unique genetic sequence signatures

Genetic sequence elements play important roles as *cis*-regulatory codes that dictate unique epigenetic patterns for the transcriptional regulation of CIGs. For example, super-enhancers tend to regulate CIGs and each comprises a long cluster of *cis*-regulatory elements that bind TFs [6, 68]. The broad H3K4me3 pattern is also enriched at CIGs and tends to be associated with a broad distribution of TF-binding motifs [9]. Therefore, we decided to perform a comprehensive survey of genetic sequence signatures together with expression features, expecting that some can be useful in uncovering CIGs.

Our meticulous analysis revealed 73 features enriched or depleted significantly in 247 CIGs, curated from literature [1], when compared to 245 control genes [1], 2142 housekeeping genes [69], or all human genes in the genome (Fig. 1A). Gene expression level is significantly higher for CIGs when compared to control genes and the entire human gene set, but intriguingly, shows little difference between CIGs and housekeeping genes (Fig. 1B). Gene expression specificity scores are significantly elevated in CIGs compared to housekeeping genes and other genes (Fig. 1C, and Supplementary Fig. S1A and B). This indicates that CIGs are overall more specific to certain cell types, consistent with their crucial role in maintaining unique cellular identities. Considering that expression specificity value for a gene will rely on the cell types used for comparison [70–72], and different CIGs display different degrees of expression specificity [1], we continued to explore genetic sequence signatures that might be useful to further improve accuracy for identifying CIGs. An investigation of the PhyloP100way score in the promoter regions of CIGs unveiled a higher degree of sequence conservation when compared to the other three gene categories, indicating stronger evolutionary conservation of CIG promoters across species (Fig. 1D, and Supplementary Fig. S1C and D). Moreover, CIGs exhibited stronger enrichment of binding motifs for TFs (Fig. 1E and Supplementary Fig. S1E), RNA-binding proteins (Fig. 1F and Supplementary Fig. S1E), and miRNAs (Fig. 1G and Supplementary Fig. S1E). CIGs displayed longer 3′-UTRs, potentially implicating greater involvement in post-transcriptional regulation, e.g. by miRNA-binding sites, which are known as often occurring in 3′-UTRs (Fig. 1H). This observation is consistent with the reported low RNA stability of CIGs [12], as miRNA mediates the decay of target RNAs in the cells [73, 74].

Unexpectedly, the CDSs of CIGs are longer than those of other gene categories, as shown in Supplementary Fig. S1F. We also found that there is a noticeable trend for the distance between the TSSs of CIGs and their nearest neighboring genes to be greater compared to other gene categories (Supplementary Fig. S1G). Additionally, a notable bias toward triplet codons encoding the amino acid Glycine was observed in CIGs (Fig. 1I), suggesting a preference for loop-forming in protein structures [75] that facilitate protein interactions [76] and catalytic activities [77]. Interestingly, CIGs exhibited a lower AT content in their coding regions (Fig. 1J and Supplementary Fig. S1H), potentially associated with the known low RNA stability of CIGs [78, 79]. Manual inspection confirmed these genetic sequence characteristics at many reported CIGs, e.g. the ESC identity genes MYC, NANOG, POU5F1, SOX2, and KLF4 but not at the housekeeping genes HPRT and EMC7 (Fig. 1K–N). It will be interesting to investigate the biological mechanisms underlying the association between these genetic signatures and CIGs.

SCIG uncovers CIGs accurately by integrating genetic signatures and expression information

Motivated by the observed significant difference in genetic signatures and expression patterns between the known CIGs versus the control gene categories, we developed SCIG, a logistic regression-based machine learning model to uncover new CIGs.

The significant features were obtained through the Wilcoxon test, followed by removing the multicollinearity among the features. This procedure yielded 73 features out of a comprehensive list of 680 candidate features (Supplementary Fig. S2A). Next, we performed forward feature selection to determine the optimal subset of these features for the logistic regression model (Supplementary Fig. S2A). Evaluation metrics such as Matthew's correlation coefficient and F1-score indicated the superior predictive performance of features selected by this pipeline when compared to conventional feature selection methods, including the Select K, ANOVA F1-score, mutual information, forward, backward, and RFE (Supplementary Fig. S2B and C). By increasing the feature number starting from 1, the model with more features shows better performance but requires up to 19 features to achieve superior performance (Fig. 2A), as a further increase in feature number does not significantly improve the performance (Supplementary Fig. S2D). The 19 key features include RNA expression level, expression specificity scores (Gini and Tau matrices), codon biases, sequence conservation, number of TF-binding motifs, number of protein-binding motifs in RNA, number of miRNA-binding motifs, nearby promoter distance, etc. (Fig. 2B). These features accurately distinguished between known CIGs and control genes with an AUROC of 0.95. Integration of these 19 features, SCIG successfully recaptured the known CIGs with an accuracy of 90.4%, sensitivity of 91.2%, and specificity of 89.6% (Supplementary Fig. S2E). RNA expression-based features, specifically the RNA expression-level and expression-specificity matrices, exhibited stronger feature coefficients when compared to the genetic sequence features (Fig. 2B). This observation is consistent with results from the leave-one-out feature test strategy (Supplementary Fig. S2F). However, an alternative model integrating only expression features and expression specificity matrices showed moderate performance (Fig. 2A). In contrast, the full model further integrating the genetic sequence signatures with the expression signatures for CIG prediction resulted in the best performance (Fig. 2A and Supplementary Table S4).

The default algorithm in SCIG is a logistics regression model and achieved remarkable performance, with a

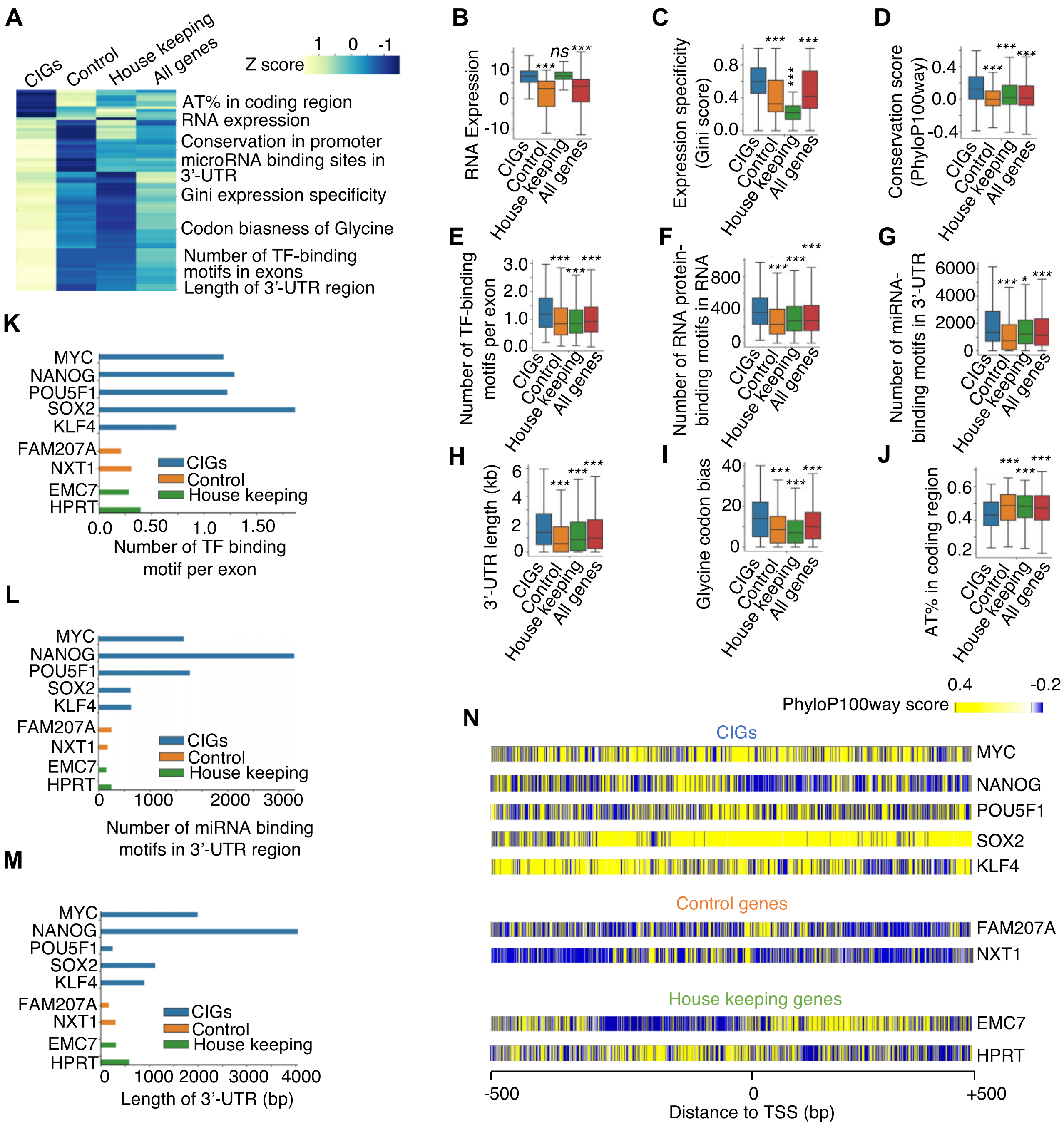


Figure 1. A systematic survey of genetic sequence signatures and RNA expression features for CIGs. (**A**) Heatmap illustrating the median values of 73 features that each displayed significant differences between CIGs and either control genes, housekeeping genes, or all human protein-coding genes. The median value of each feature was converted into *z*-score. (**B–J**) Boxplots showing values of represent features in individual gene categories. *P*-values determined by the two-tailed Wilcoxon test. **P*-value < 0.05, ****P*-value < 0.001, and ns: nonsignificant. (**K–M**) Bar plots showing genetic sequence feature values of individual embryonic stem CIGs and control or housekeeping genes. (**N**) Heatmap showing PhyloP100way scores around the TSS of individual embryonic stem CIGs and control or housekeeping genes.

Matthew's correlation coefficient of 0.81 (Supplementary Fig. S2B) and an F1-score of 0.90 (Supplementary Fig. S2C) in recapturing the known CIGs. For comparison, we assessed the predictive capacity of alternative machine-learning algorithms. The logistic regression model exhibited a lower error rate and a higher Matthew's correlation coefficient, underscoring its superior performance compared to the other models, including the Naive Bayes, support vector machines, and AdaBoost (Supplementary Fig. S2G). Applying SCIG to human umbilical vein endothelial cells (HUVECs) [80], well-known endothelial CIGs such as the NR2F2 [81, 82], MECOM [83], HOXA5 [84], and GATA6 [85] were top-ranked by SCIG, indicating the ability to recapitulate the endothelial CIGs (Fig. 2C). Additionally, we investigated the functional implications of the top 25 CIG candidates in HUVECs by reviewing the literature and categorized them into the four CIG categories defined based on literature curation in recent work [1]. Of these, six are known master TFs, three are required TFs, eight

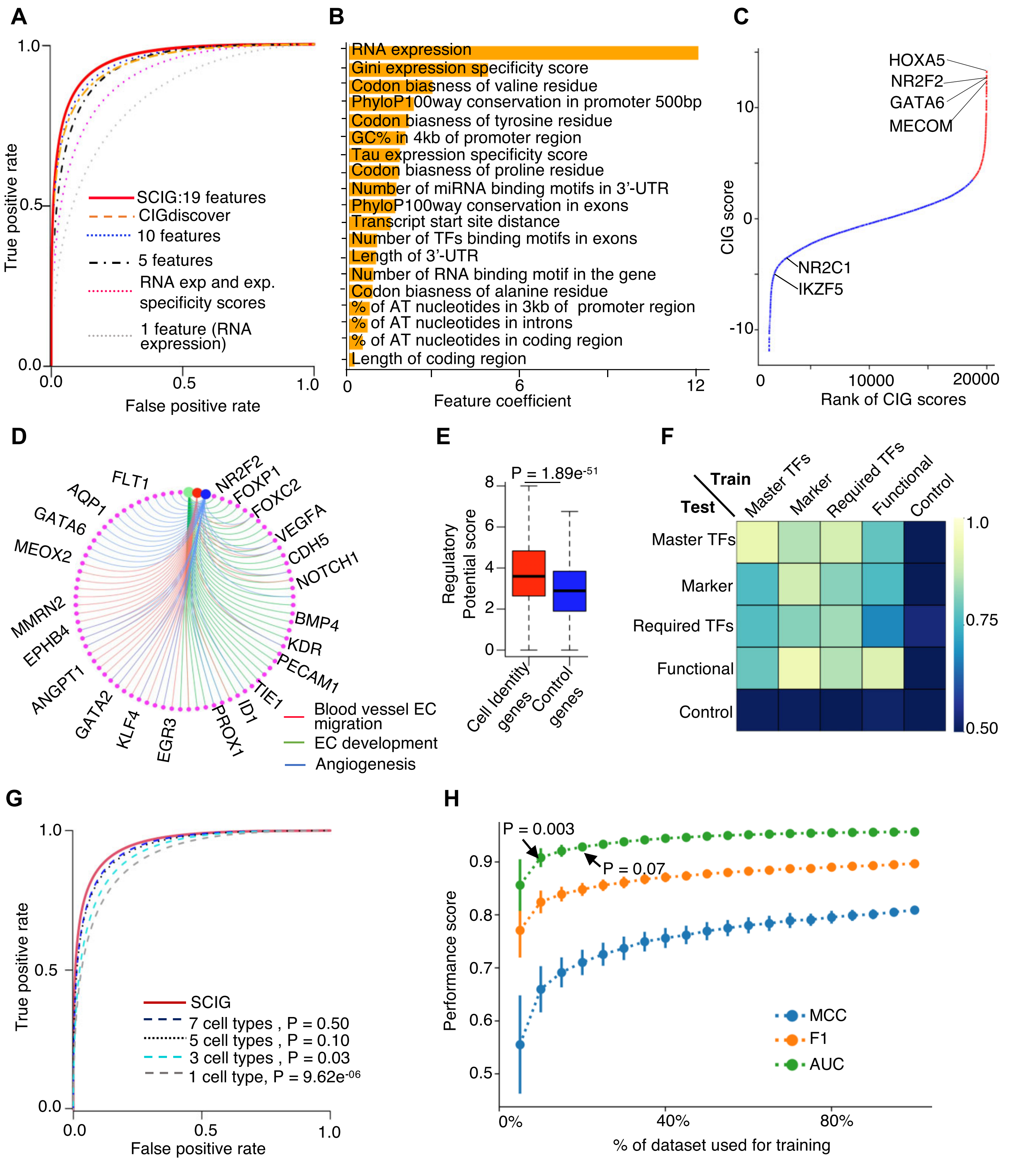


Figure 2. SCIG combines genetic sequence signatures with expression information to uncover CIGs in a cell. (**A**) ROC curve illustrating the performance of SCIG with varying numbers of top features. Performance of the CIGdiscover algorithm that uncovers CIGs by histone modification signatures, gene expression, and expression specificity information is also presented. (**B**) Bar plot showing feature coefficient of individual genetic sequence signatures or gene expression features used for machine learning in SCIG. (**C**) Rank plot presenting the CIG scores of individual genes in HUVEC. Red and blue colors indicate CIGs and other genes defined by SCIG, respectively. (**D**) The CIGs defined by SCIG in HUVEC are enriched with endothelial pathways. (**E**) Boxplot of regulatory potential scores demonstrating the regulation intensity of individual genes by TFs in HUVEC. (**F**) Heatmap showing AUROC of SCIG variants trained and tested by individual gene categories. (**G**) ROC curves depicting the performance of SCIG, and its variants trained with data from varying numbers of cell types. (**H**) Line plot illustrating the performance of SCIG, and its variants trained with varying subsets of the known CIGs.

are functional genes, four are marker genes, and four are not yet reported to be endothelial CIGs (Supplementary Table S5). The identified CIGs are enriched for endothelial pathways, reaffirming their association with EC identity (Fig. 2D and Supplementary Fig. S2H). As CIGs tend to be regulated by super-enhancers [6, 68], we interrogated ChIP-seq profiles of TFs from the Cistrome browser [86]. The result verified that the predicted CIGs exhibited a pronounced enrichment of TF-binding events compared to control genes (Fig. 2E). Additionally, we extended the SCIG model to the mouse genome, achieving an AUROC of 0.94 (Supplementary Fig. S3A) utilizing 16 distinct genetic and RNA expression features (Supplementary Fig. S3B). SCIG consistently identified pan-EC marker genes as CIGs in brain, lung, and liver ECs (Supplementary Fig. S3C–E). These results indicate that the algorithm in SCIG is optimal for accurately recapturing known CIGs based on the combination of genetic sequence signatures and RNA expression patterns.

Robust performance of SCIG with small and noisy training data

We assessed whether the performance of SCIG is consistent when applied to the four CIG categories defined based on literature curation in recent work [1]. These include master TFs, required TFs, key function genes, and marker genes. The model consistently demonstrated robust performance (AUROC > 0.83) when applied to each of these categories (Supplementary Fig. S4A). Impressively, a model trained by one CIG category can accurately recapture the other CIG categories but not the control genes, underscoring the similarity in characteristics of these CIG categories (Fig. 2F).

To test whether the size of the training data has been large enough to reach an optimal performance, we evaluated the algorithm trained with different subsets of the training data. The algorithm's performance improved when the known CIGs and control genes from more cell types were used to train the model, with the improvement saturated at up to 5 cell types (Fig. 2G). Also, the model requires up to 15% of the 247 known CIGs to achieve optimal performance (Fig. 2H). Considering the potential imbalance between the number of control genes and CIGs in a cell, we investigated the impact on SCIG performance. We systematically varied the number of non-identity genes used to train the model and analyzed its effect on SCIG accuracy. Despite increasing the number of non-identity genes up to 20-fold higher than the number of CIGs, we observed minimal changes in SCIG performance (Supplementary Fig. S4B).

We investigated whether the model trained by known CIGs from one cell type can accurately recapture the known CIGs of a different cell type. To this end, we performed two different tests. We first performed a parallel test, where the model was trained on 80% of the known CIGs from five cell types and tested on the remaining 20%, followed by a cross-test, where the model was trained on genes from one set of five cell types and tested on genes from another set. The results demonstrated that the algorithm performance is excellent in both the parallel test (AUROC: 0.93) and the cross-test (AUROC: 0.91) (Supplementary Fig. S4C). Further, we performed additional tests by randomly assigning the ten cell types into ranks from 1 to 10 and grouping the cell types into different subsets, with each group containing three cell types. For example, group 1 contains cell types 1–3, group 2 contains cell types 2–4, and so on. As a result, adjacent groups in the ranking overlap, while distant groups do not. Using these cell groups, we then trained the model on one group and tested it on the other groups with no (cross test) or partial overlap and also evaluated the model's performance within the same group (parallel test). Through this analysis, we show again that the performance was comparable between the cross-test and parallel tests (Supplementary Fig. S4D). These results suggest that SCIG performs well in identifying CIGs across independent cell types.

We extensively evaluated the robustness of SCIG against noise in the training data. To this end, the labels of a randomly picked subset of the known CIGs and control genes were swapped. As expected, the algorithm performance decreased along with the increased noise ratio in the training data. However, with 20% mislabeled CIGs or control genes, the AUROC only decreased from 0.95 to 0.75 (Supplementary Fig. S4E), suggesting that the algorithm is considerably resilient to noise in the training data.

SCIGNet uncovers master transcription factors of cell identity genes by machine-learning analysis of network features

It is well recognized that a cell identity can be established by a small cocktail of master TFs, such as the Oct4, Sox2, Klf4, and c-Myc for pluripotent stem cells [5, 6]. As a cell type often has several hundred or more CIGs, we next developed another logistic regression model, SCIGNet, to uncover the master TFs of a cell identity based on the network features of CIGs. Conventional network analysis often defines master TFs based on their number of downstream target genes. However, it is unclear if there are other network features that might be more useful to define the master TFs of a cell identity. To this end, we considered a set of 23 network features, including the number of downstream as well as upstream network edges connected to a TF, gene expression values, CIG scores, etc. (Fig. 3A). SCIGNet learns the features associated with the known master TFs of cell identity curated from literature [1]. It then optimizes the weight per feature to combine these features for predicting new master TFs.

Our feature analysis revealed that reported master TFs exhibit greater CIG scores calculated by the SCIG (Fig. 3B) and higher expression values (Fig. 3C) compared to other TFs. Additionally, when compared to other TFs in the CIG network, we observed a greater number of children edges connecting master TFs to downstream target genes (Fig. 3D). This observation is consistent with the role of these factors as a master regulator in the network. Intriguingly, we also observed a greater number of parental edges connecting the master TFs to their upstream regulators (Fig. 3E). This observation indicates that the master TFs themselves, when compared to other TFs in the CIG network, are under more regulations. These results were also observed by manually inspecting the feature values of individual master TFs in the embryonic CIG network (Fig. 3F–H).

The SCIGNet model employed nine features to achieve optimal performance with an AUROC of 0.91 (Fig. 3I and J). Unexpectedly, the mean CIG score and average RNA expression of the genes connected by the parental edges appeared to be the two most useful features for identifying the master TFs, followed by the RNA expression and CIG score of the master TFs themselves (Fig. 3J). When compared to simple methods

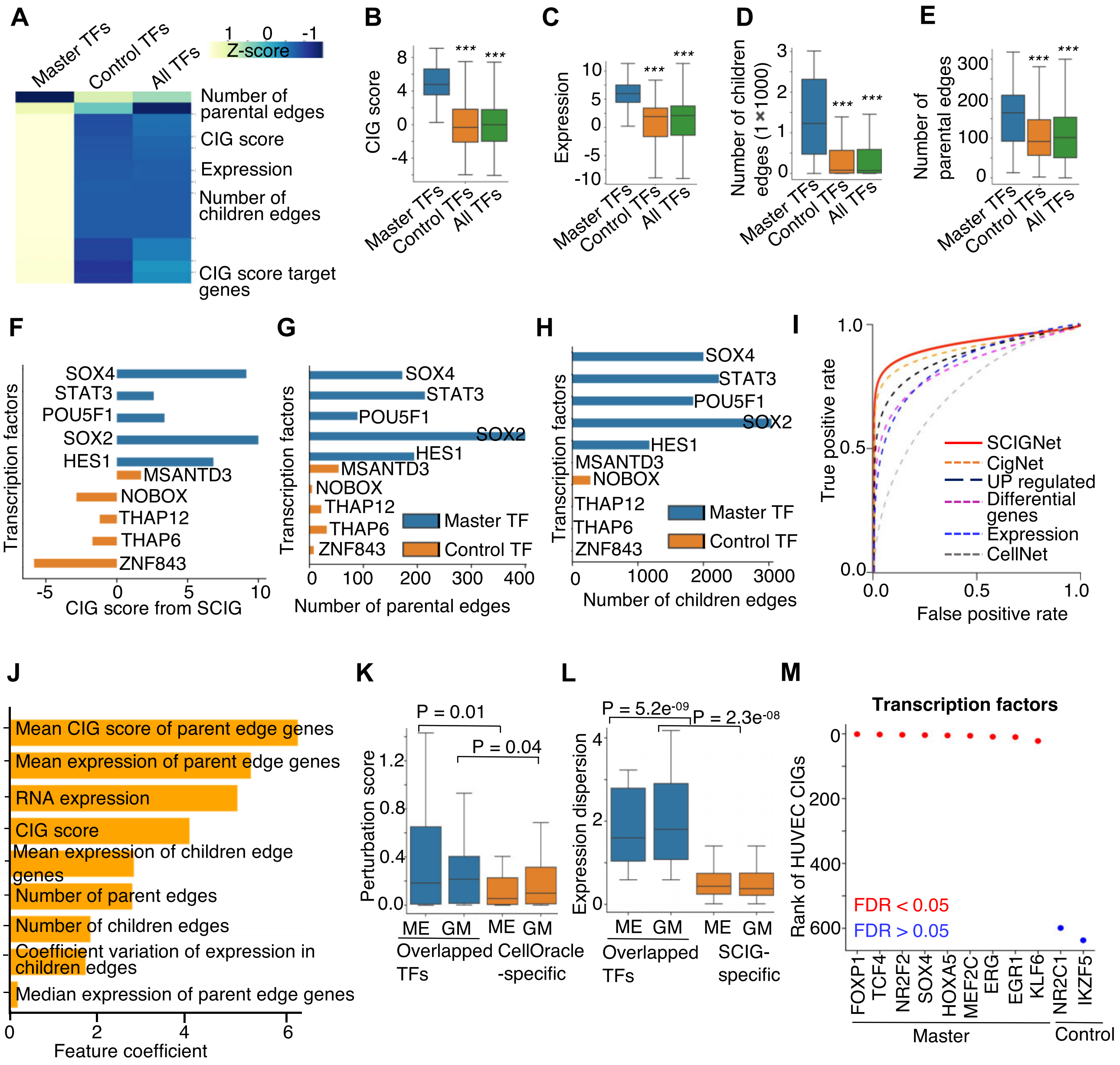


Figure 3. SCIGNet combines network features of CIGs to uncover master TFs of a cell identity. (**A**) Heatmap showing *Z* scores of individual network feature values for individual TF groups. (**B**–**E**) Box plot illustrating representative network feature values for individual TF groups. ****P*-value < 0.001. (**F**–**H**) Bar plot showing feature values of individual TFs. (**I**) ROC curves showing the performance of SCIGNet and other methods for uncovering master TFs of cell identity. (**J**) Bar plot showing feature coefficients of individual network features used by the machine learning models in SCIGNet for uncovering master TFs. (**K**) Comparison of perturbation scores derived from CellOracle for the cell identity TFs defined by SCIG and CellOracle. (**L**) Comparison of expression variability for cell identity TFs defined by SCIG and CellOracle. (**M**) Rank of individual TFs based on the scores calculated by SCIGNet in HUVEC.

based on analysis of gene expression level or the conventional network model CellNet, SCIGNet always showed a better performance in recapturing the known master TFs (Fig. 3I).

Additionally, we compared the cell identity TFs predicted by SCIG and CellOracle [87]. Although both our method and CellOracle can predict cell identity TFs, they are distinct in two major ways: (i) CellOracle requires both gene expression and chromatin accessibility data from the analyzed sample to predict TFs involved in cell identity. In contrast, SCIG only requires gene expression data from the analyzed sample, making it applicable to a broader range of single-cell datasets. (ii) CellOracle focuses on TFs identified as highly variable across single cells and computes perturbation scores for these TFs. SCIG considers all TFs in the sample and does not filter out genes in advance. As a result, the cell identity TFs defined by SCIG do not always have to be highly variable. This is important because it is known that different cell types can share some CIGs, although the specific combination of these genes is unique to each cell type [5, 88, 89]. We analyzed the mouse hematopoiesis differentiation dataset [90], which revealed two lineage differentiation trajectories, the ME (Megakaryocyte-Erythrocyte) and GM (Granulocyte-Monocyte), to compare the performance of SCIG and CellOracle. In these lineages, CellOracle identified 90 highly variable TFs, for which perturbation scores were generated. Because CellOracle results do not include *P*-values to define significant candidates, we included all these 90 candidates for comparison. SCIG identified 68 top-ranked cell identity TFs in the ME lineage and 64 in the GM lineage based on *P*-value < 0.05. We compared the cell identity TFs identified by both methods and observed significant overlap between the two (*P*-value < 0.05), with 28 TFs common to both SCIG and CellOracle in the ME lineage, and 25 common TFs in the GM lineage. This suggests that the two methods identify many common cell identity TFs, despite using different types of information in their predictions. To gain deeper insights and facilitate a comparison of their predictions, we next investigated the perturbation scores for the overlapping cell identity TFs recognized by both SCIG and CellOracle, compared with the TFs identified exclusively by CellOracle. We observed that the perturbation scores for the overlapping TFs were higher than those for the TFs identified solely by CellOracle. Such a result suggests that the TFs identified by SCIG were highly ranked in the CellOracle predictions (Fig. 3K). Further, we examined the expression variability of the overlapping cell identity TFs recognized by both methods, in comparison with the SCIG-specific TFs, and found that the overlapping TFs exhibited higher expression variability (Fig. 3L) than the SCIG-specific TFs. This result is consistent with expectation since SCIG considers all TFs as compared to CellOracle which considers only highly variable TFs.

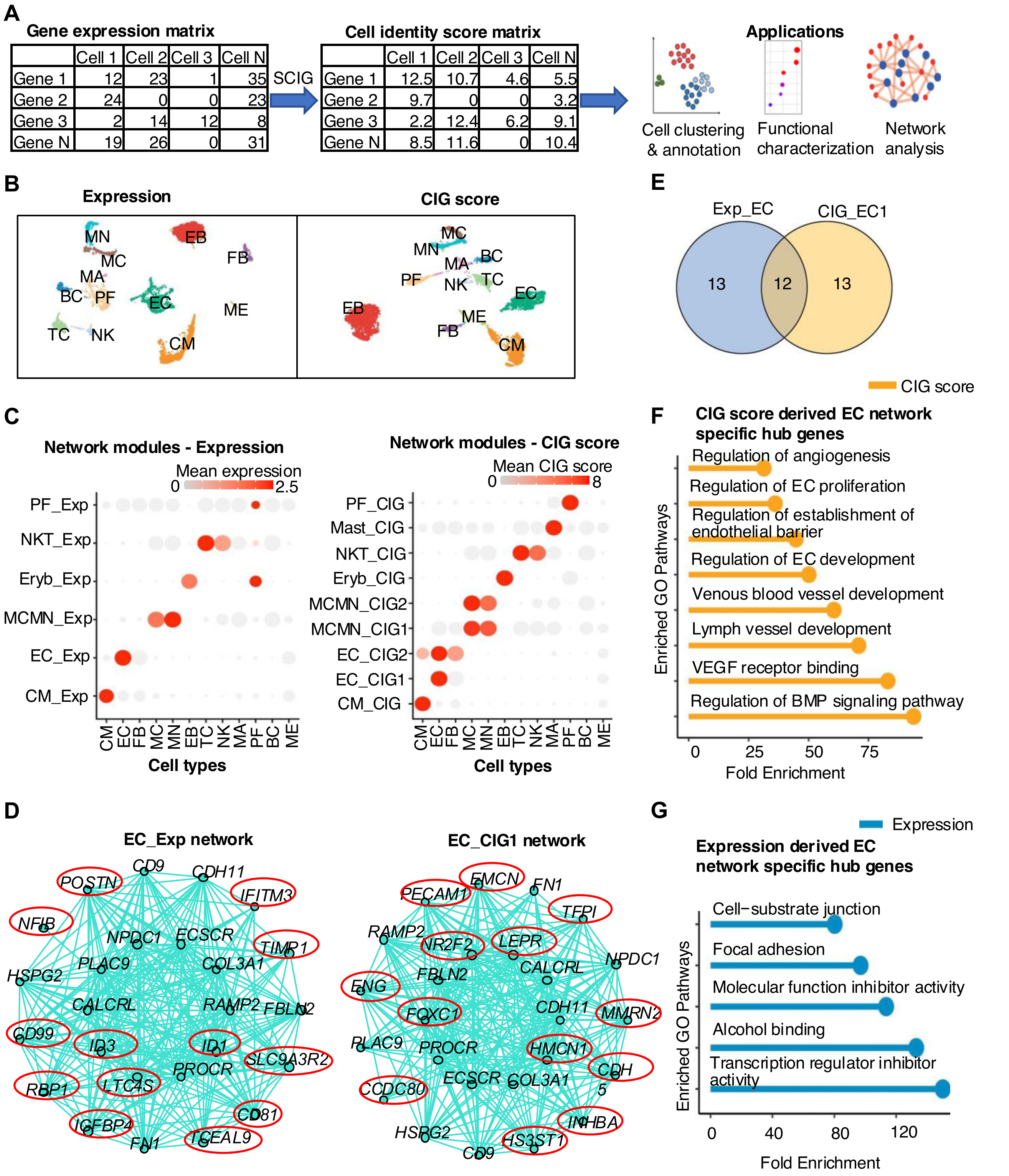
Furthermore, we employed SCIGNet in HUVEC cells to pinpoint master TFs regulating their identity. We uncovered that most of the top-ranked TFs, such as FOXP1 [91], NR2F2 [81], SOX4 [92], MEF2C [93], and ERG [94], have already been experimentally characterized for their role in the regulation of HUVEC identity (Fig. 3M). Also, applying the model to mice cells identifies the master TFs in the CIG network with an AUROC of 0.93 (Supplementary Fig. S3F). Applying SCIGNet to atrial (aCMs) and ventricular cardiomyocytes (vCMs) transcriptome data, SCIGNet identified six and eleven TFs as aCMs- and vCMs-specific master regulators (Supplementary Fig. S3G). Notably, most of these regulators were reported to play roles in the heart compartments (aCMs and vCMs). These include the Tbx5 and Esrra we have comprehensively validated in our recent works [95, 96] and the Foxp2 [97], Bhlhe40 [98], Pbx1 [99], Klf12 [100], Prdm16 [101], Hey2 [102], and Casz1 [103].

CIG score outperforms expression value in capturing cell identity in network analysis

We next tested if the CIG scores calculated by SCIG can be better than the gene expression values when aiming at uncovering CIG networks in single-cell RNA-seq analysis. We used SCIG to calculate a CIG score for each gene in each single cell based on single-cell transcriptomes of healthy human fetal hearts at 22 weeks [58] (Fig. 4A). We next clustered cells using either the CIG scores or the RNA expression values based on highly variable genes. Both methods effectively recaptured the 12 cell types, including cardiomyocytes, ECs, fibroblast, mesothelial cells, proliferating cells, mast cells, erythroblasts, etc., as reported in the original study (Fig. 4B). Thereafter, cell clusters from the two methods were independently used as input for network analysis based on either the expression values or CIG scores using the hdWGCNA package [59]. Employing gene expression data, we obtained six network modules (referred to as the expression network), while the CIG score yielded nine modules (referred to as the cell identity score network). The specificity of hub genes toward particular cell types was illustrated in a dot plot, showcasing their expression (Fig. 4C, left) or CIG score (Fig. 4C, right).

For a fair comparison between the expression network and cell identity score network for a cell type, we focused on the top 25 hub genes from each network. For ECs, a cell identity score network (EC_CIG1) and the expression network (EC_Exp) (Fig. 4D) shared 12 hub genes (Fig. 4E). Notably, pathway analysis revealed that EC cell identity score network-specific hub genes were associated with the regulation of EC and blood vessel development, VEGF signaling, and angiogenesis (Fig. 4F). In contrast, expression network-specific hub genes were associated with housekeeping-related cellular functions (Fig. 4G). This observation suggests that the CIG score is superior in revealing the EC identity network. Interestingly, based on CIG scores, hdWGCNA further constructed a network module shared by ECs, fibroblasts, and cardiomyocytes (EC_CIG2) (Fig. 4C, right and Supplementary Fig. S5A). Comparing this network with the EC_Exp module revealed only two common hub genes (Supplementary Fig. S5B). Pathway analysis indicates that the genes in the EC_CIG2 model are involved in the well-known epithelial (endothelial) to mesenchymal transition (Supplementary Fig. S5C), which plays critical roles in the development and many diseases, such as heart failure [104, 105].

For the cardiomyocyte network modules, the expression network (Supplementary Fig. S5D) and cell identity score network (Supplementary Fig. S5E) shared a set of 10 hub genes (Supplementary Fig. S5F). The hub genes specific to the cell identity score network demonstrated a more pronounced enrichment across pathways regulating the cellular processes and functions specific to cardiomyocytes (Supplementary Fig. S5G). The network modules derived from expression (Supplementary Fig. S6A) and cell identity score (Supplementary Fig. S6B) also uncovered different hub genes for the proliferating cell population. Upon pathway enrichment analysis, it became evident that the cell identity score network hub genes displayed a heightened association with cell cycle-related pathways (Supplementary Fig. S6D), contrasting with expression networks, which are only enriched with some housekeeping pathways (Supplementary Fig. S6C). Two network modules, MCMN_CIG1 and MCMN_CIG2, were identified as common for macrophages and monocytes using cell identity scores, while one module, MCMN_Exp, was identified based on expression values (Fig. 4c). We observed that 18 hub genes were shared between the MCMN_CIG2 and MCMN_Exp modules (Supplementary Fig. S6E). However, there were no overlapping hub genes between the MCMN_CIG1 and MCMN_Exp modules. Notably, the genes in the MCMN_CIG1 cell identity network (Supplementary Fig. S6F) were enriched with pathways related to inflammatory response and cellular response (Supplementary Fig. S6G). With the expression values, there was no network module detected for mast cells. Conversely, the unique cell identity score network of the mast cells unveiled an enrichment pattern aligning with mast cell functions, including regulation of chemokine ligand production and immune responses (Supplementary Fig. S6H) [106, 107]. Meanwhile, for erythroblast and natural killer T cells, less difference was observed between the expression- and CIG score-derived networks (Supplementary Fig. S6I and J), with no major



Gene expression matrix

	Cell 1	Cell 2	Cell 3	Cell N
Gene 1	12	23	1	35
Gene 2	24	0	0	23
Gene 3	2	14	12	8
Gene N	19	26	0	31

Cell identity score matrix

	Cell 1	Cell 2	Cell 3	Cell N
Gene 1	12.5	10.7	4.6	5.5
Gene 2	9.7	0	0	3.2
Gene 3	2.2	12.4	6.2	9.1
Gene N	8.5	11.6	0	10.4

Figure 4. CIG score outperforms expression value in capturing cell identity in network analysis. (**A**) Workflow to use SCIG for uncovering CIGs at the single-cell level and perform subsequent applications. (**B**) Single-cell clustering based on expression and CIG score. (**C**) Network modules defined using hdWGCNA based on expression and CIG score matrices. (**D**) Network plot showing the top 25 hub genes in EC identity score- and gene expression-derived network. Expression- and CIG score-specific network hub genes are marked in circles. (**E**) Ven diagram showing overlap of top 25 hub genes in CIG score- and gene expression-derived networks. (**F**) Pathway enrichment analysis of CIG score-specific network hub genes. (**G**) Pathway enrichment analysis of expression-specific network hub genes. CM, cardiomyocytes; EC, endothelial cells; FB, fibroblast; MC, macrophages; MN, monocytes; EB, erythroblast; TC, T cells; NK, natural killer cells; MA, mast cells; PF, proliferating cells; BC, B cells; ME, mesothelial cells.

distinction concerning their pathway associations. Overall, we posit that networks derived from CIG scores prove more advantageous in elucidating the gene network involving cell identity regulation.

Cell identity score improved single-cell trajectory analysis of neuronal differentiation

We harnessed the SCIG method to explore the landscape of CIGs in 1720 single-cells during human forebrain glutamatergic neuron differentiation. This dataset effectively captured the dynamic process of mature neuron formation starting from radial glial progenitors, exhibiting a linear trajectory [60, 108]. We performed single-cell clustering using either the RNA expression values or the CIG scores based on top 2000 high-variation genes. There was only a 24% overlap between the highly variable genes defined by these two methods (Fig. 5A). The highly variable genes identified through the CIG score exhibit a stronger enrichment in the forebrain and neuronal development pathways (Fig. 5B). Meanwhile, the clustering based on CIGs recapitulated all the cell types defined in previous studies based on gene expression [60, 108] (Fig. 5C and D, Supplementary Fig. S7A and B). However, the Silhouette coefficient, an internal cluster validation measure, is 1.5-folds smaller for the CIG-based than for the expression-based methods with the same cell clustering parameters (Supplementary Fig. S7C, left). This suggests that the CIG scores are better at capturing the relation between the cells in the differentiation process, as also can be observed in the single-cell clustering architecture derived from highly variable genes identified uniquely by expression or CIG score (Supplementary Fig. S7B). Meanwhile, we found substantial differences in the sizes of cell populations defined based on CIG score- versus gene expression-based clustering, e.g. population sizes of Neuroblast1, Immature neurons, and mature neurons (Fig. 5E). The greatest cell identity switch between the two methods happened to neuroblast1 and immature neuronal cells (Fig. 5F).

To assess the effects of cell type rearrangements between the two methods on differentiation trajectory analysis, we conducted RNA velocity analysis and quantified the transition probabilities between the cell populations using the algorithm CellRank [63]. Although the result indicated an overall consistent neuronal differentiation trajectory between the two methods (Fig. 5D), the probability of the expected transition from neuroblast 1 and neuroblast 2 toward immature neurons was increased by 2 to 5-fold by the CIG-based clustering compared to the expression-based one (Fig. 5G). Similarly, the transition probability from neuroblast 1 to neuroblast 2, neuroblast 2 to immature neuron, and neuron 1 to neuron 2 also increased in the CIG-based clustering result (Fig. 5G). In contrast, the probability of transition that is unknown or opposite to literature report, e.g. from neuroblast 2 to neuroblast 1 and neuron 2 to neuron 1, exhibited decreases (Fig. 5G). We next presented the transition probabilities of each cell using a circular projection plot, positioning naive, or intermediate cells at the center, while mature or fate-biased cells fell in the corners corresponding to their respective identities. This visualization highlighted the relationships between cell types during differentiation. From the cell types derived using CIG scores, we observed that intermediate cell types, such as immature neurons, tend to cluster closer to the middle of the plot, which is better than the gene expression method (Fig. 5H). In contrast, matured or terminated differentiated cell types, including radial glial cells and neurons, are situated at the corners corresponding to their respective identities. These results suggest that single-cell clustering based on CIG scores can better arrange cells in the differentiation trajectory.

Additionally, we performed optimization to achieve the same silhouette score between the pipelines based on gene expression value and CIG score (Supplementary Fig. S7C, right, and Supplementary Table S6). We observed that the cell clusters obtained using the expression values with the optimized clustering parameters, as well as the original CIG score-derived cell clusters, consistently captured all the cell types (Supplementary Figs S7D and 5D, right). The proportions of these cell types changed significantly, except for immature neurons and neuroblast1 (Supplementary Fig. S7E), with a few changes in a cell cluster label switching observed after optimization (Supplementary Fig. S7F). However, these changes did not alter our observation that cell identity transitions are more effectively captured by the CIG score than the expression value (Supplementary Fig. S7G).

Moreover, to evaluate the contribution of gene expression specificity matrices in SCIG prediction and single-cell trajectory analysis, we removed the expression specificity features (Gini and Tau scores) from the SCIG model (Fig. 2B) and predicted the CIG score for each gene in the single cells. This model variant failed to recapture all cell types, as one subpopulation of neuroblasts and radial glial cells was missed (Supplementary Fig. S8A). Additionally, the transition probabilities between cell types were significantly altered in the neurogenesis differentiation dataset. For instance, the transition probability from neuroblasts to immature neurons dropped from 0.58 to 0.15 (Supplementary Figs S8B and 5G, right).

We further extend the analysis by using the expression specificity features alone as a SCIG variant model. This model was unable to capture all cell types, and the clustering architecture differed substantially from the predictions of the full SCIG model (Supplementary Fig. S8C). Additionally, the predicted CIG scores for all genes were significantly lower (Supplementary Fig. S8D) compared to those obtained with the SCIG whole model that includes all features (Fig. 5C). This observation is expected because the gene expression specificity matrices used in SCIG were general values computed using the comprehensive atlas of many cell types in the public database, rather than the specific scRNA-seq dataset analyzed by SCIG. In a further extension of this analysis, we examined two additional variants of the SCIG model that either used gene expression alone or retained both gene expression and expression specificity. These SCIG variant models successfully recaptured all cell types in the neurogenesis dataset. However, we observed that the transition probabilities between neuroblast 1 and 2, as well as from immature neurons to mature neurons 1 and 2, were lower than those predicted by the complete SCIG model (Supplementary Fig. S8E–H). These tests highlight the importance of gene expression, expression specificity, and sequence features for accurate identification of CIGs in a given sample.

SCIG recapitulated the landscape of CIGs in the endothelial differentiation process

We further applied the SCIG algorithm to identify CIGs in the scRNA-seq profiles from various stages during the differentiation of human ESCs (H9) into ECs [64], including stem

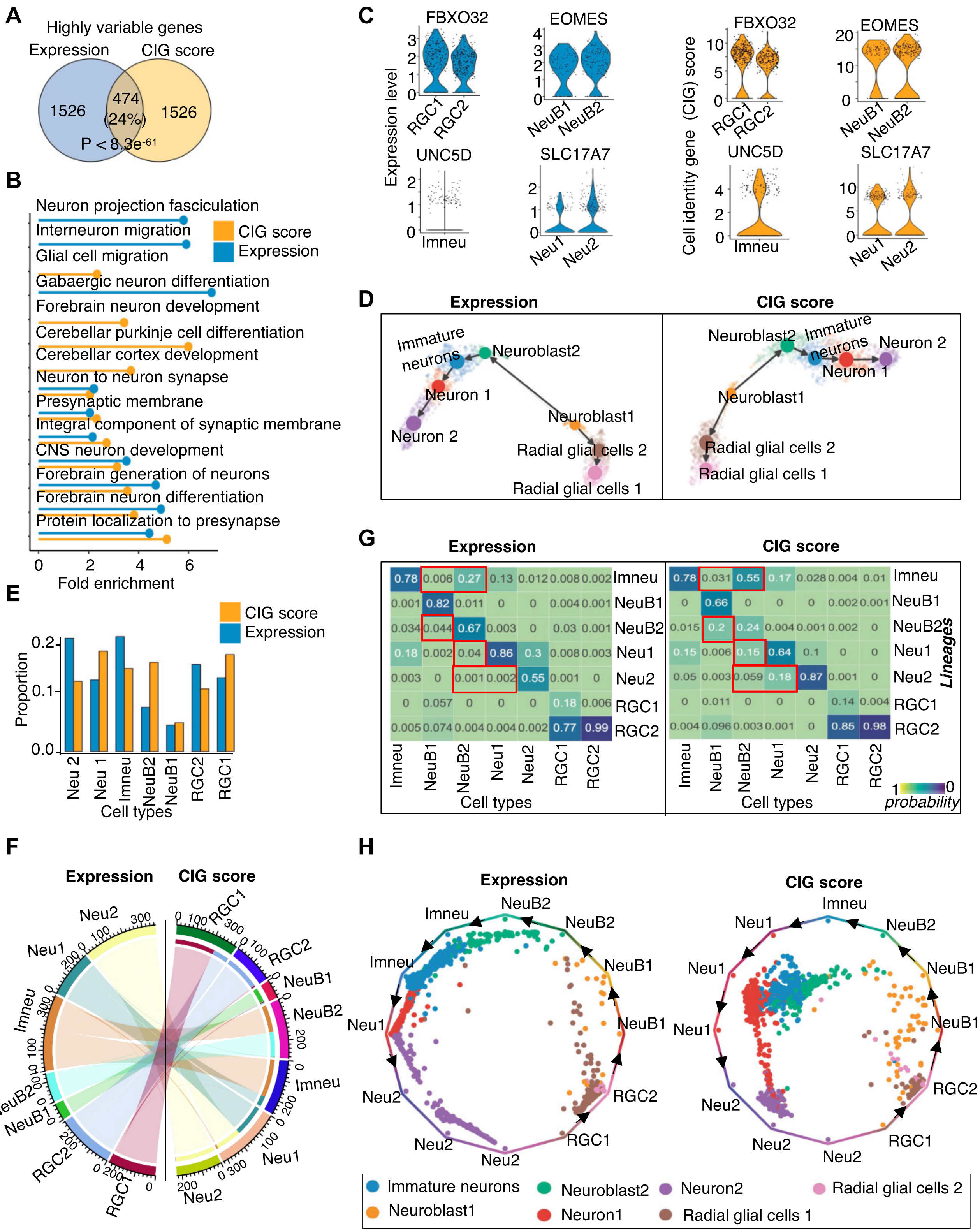


Figure 5. Cell identity score improved single-cell trajectory analysis of neuronal differentiation. (**A**) Venn diagram showing overlap between highly variable genes defined based on single-cell gene expression and CIG scores. (**B**) Pathway enrichment analysis of highly variable genes defined based on single-cell gene expression and CIG scores. (**C**) Expression level and cell identity score of marker genes that we used to define the cell types in this dataset. (**D**) UMAP displaying cell types and differentiation trajectories in the human forebrain glutamatergic neurogenesis dataset. (**E**) Barplot showing proportions of individual cell populations clustered based on CIG scores or expression values. (**F**) Chord diagram showing the cells that are switched between expression- and CIG score-based cell clustering. (**G**) Heatmaps depicting transition probabilities quantified using CellRank between cell populations clustered based on expression values (left) or CIG scores (right). (**H**) Projection plots showing the fate probabilities of each cell during the glutamatergic neuron genesis trajectory. RGC1, radial glial cells 1; RGC2, radial glial cells 2; NeuB1, Neuroblast1; NeuB2, Neuroblast2; imneu, immature neurons; Neu1, Neuron 1; Neu2, Neuron 2.

cells (day 0), mesoderm (day 4), EC-mesenchymal progenitors (days 6 and 8), and ECs (day 12) (Supplementary Fig. S9A). Subsequent pathway enrichment analysis of the identified CIGs at each stage revealed stage-related pathways (Supplementary Fig. S9B–E). To investigate the relationship between expressed genes and identified CIGs, we compared the top 10% highly expressed genes with the top 10% CIGs and observed only a 26% overlap (Fig. 6A). Pathway enrichment analysis showed that the unique high-scored CIGs are enriched with EC functional pathways, including VEGF signaling, BMP signaling, and endothelium development. In contrast, the highly expressed genes are involved in housekeeping-related functions such as RNA splicing and ribonucleoprotein complex processes (Fig. 6B). Pearson correlation analysis between the CIG scores and the expression levels of highly expressed genes revealed a minimal correlation, which is reasonable because the CIG score combines expression information and genetic sequence signatures (Supplementary Fig. S9F and G). Therefore, the CIG score is better than the expression level for enriching cell identity pathways.

SCIG exhibited the ability to identify CIGs that are specific to individual stages. These include the SOX2 [5] and POU5F1 (OCT-4) [5] for stem cells (day 0), the HAND1 [109], and MIXL1 [110] for mesoderm cells (day 4), FGF1 [111] and ZEB2 [112] for EC-mesenchymal progenitor cells (day 6 and 8), and the MECOM [83], and NR2F2 [81] for EC (day 12) (Fig. 6C). Meanwhile, notable overlap of CIGs between neighboring stages were also recapitulated, e.g. between stem cells (day 0) versus mesoderm cells (day 4) stages and EC-mesenchymal progenitors (day 6 and 8) versus ECs (day 12) stages (Supplementary Fig. S9A). Additionally, we used the SCIGNet to identify the network regulators of CIGs. Consistent with the literature report, we observed that the SOX2 [5], POU5F1 [5], and OTX2 [113], served as key regulators in stem cells, PAX2 [114], HOXC6 [115], and HAND1 [109] in mesoderm cells, while the NR2F2 [82], KLF6 [116], and EGR3 [117] are key regulators in ECs (Fig. 6D). Therefore, SCIGNet successfully recapitulated the network regulators governing the differentiation process of human ESCs toward the EC fate.

SCIG revealed new insight into EC identity refinement by tissue microenvironment

We obtained single-cell transcriptomes representing ECs from 15 tissue types in the DISCO database [65], including the adipose, bladder, breast, gut, heart, intestine, kidney, liver, lung, ovary, skeletal muscle, skin, stomach, testis, and thymus tissues. The EC in each tissue type comprises four EC subtypes, including the arterial, capillary, venous, and lymphatic ECs. We used SCIG to elucidate the identity gene landscape in each EC subtype from each tissue type. The algorithm identified a total of 2067 CIGs, including 86 CIG master regulators. For each gene, we computed a Tau score representing cell identity score specificity across all tissues, ranging from 0 to 1. A Tau score close to 0 and 1 indicates low and high tissue specificity, respectively.

Intriguingly, we found CIGs tend to exhibit significantly greater Tau scores than CIG master TFs (Fig. 6E). This difference between CIGs and their master TFs is consistent when we further analyze each EC subtype across the tissue types, with the greatest difference observed for venous EC followed by lymphatic EC (Supplementary Fig. S10A–D). The genes such as EGR3 [117, 118], KLF2 [119, 120], NR2F2 [81, 82], HOXB5 [121], ID1 [122], and KDR [123] possess moderate Tau scores, indicating conservation across tissue types, while AREG [124], SEMA3G [125], LEF1 [126], ILK [127], and ARHGEF15 [128] exhibit greater Tau scores (Fig. 6F). The conserved CIG genes are enriched with EC-related biological processes and functions (Fig. 6G). In contrast, the low-conservation CIGs demonstrate weaker enrichment in EC functions and a greater association with tissue-specific functions (Supplementary Fig. S10E). Expanding this analysis to EC master TFs, MEF2C [93], SOX17 [129], NR2F2 [130], EGR1 [131], and KLF6 [116], appeared conserved across tissue types, while MAFB [132], ETV1 [133], TCF15 [134], and HOXD8 [135] appeared to be less conserved (Fig. 6H). These results imply that the tissue microenvironment, as an additional factor to the master TFs in the cells, may play an important role in fine-tuning the EC identity.

Meanwhile, there are also a few differences in CIGs between endothelial subtypes (Fig. 6I). The conserved CIGs of each EC subtype across the tissue types include the GJA4 [136], CXCL12 [137, 138], and NOTCH4 [139] in arterial EC (Supplementary Fig. S10F), the NR2F2 [81], NRP2 [140], and EPHB4 [141] in venous EC (Supplementary Fig. S10G), the FABP5 [142], SPARC [143], and CD36 [144] in capillary EC (Supplementary Fig. S10H), and the PROX1 [145], LYVE1 [146], and CCL21 [147, 148] in lymphatic EC (Supplementary Fig. S10I). On the other hand, we also observed tissue-specific CIGs for each EC subtype. For example, the VEGFC [149], MYCN [150], and EFNB2 [151] in arterial EC (Supplementary Fig. S10F), the FZD5 [152], ARG1 [153], and LRG1 [154] in venous EC (Supplementary Fig. S10G), the VTN [155], RGCC [143], and PRX [156] in capillary EC (Supplementary Fig. S10H), and the CCL4 [157], FOXO3 [158], and MMRN2 [159] in lymphatic EC (Supplementary Fig. S10I). In the case of master TFs (Fig. 6J), the algorithm identified SOX17 [160] and KLF6 [116, 161] in arterial EC (Supplementary Fig. S10J), NR2F2 [81, 162], and MEF2C [163] in venous EC (Supplementary Fig. S10K), NFIB [164] and JUNB [165] in capillary EC (Supplementary Fig. S10L), and NR2F1 [166] and NR2F2 [167] in lymphatic EC (Supplementary Fig. S10M) as conserved across the tissue types. These analyses shed light on the heterogeneity of endothelial CIGs across tissue types, aiding our understanding of tissue-specific microenvironments in cell identity refinement.

Discussion

Advancements in high-throughput sequencing technologies, particularly scRNA-seq, have revolutionized the study of gene expressions at a large scale. This technique has enabled researchers to delve into cellular heterogeneity, cell fate conversion, and other cellular processes [168–170]. Investigating the role of genes in cellular identity regulation at the single-cell level is crucial for gaining insights into cell fate conversion and its potential applications in regenerative medicine.

The community has been analyzing CIGs by considering differences in epigenetic modification patterns between CIGs and other genes in the same cell type, or by considering differential expression between cell types. If one could define CIGs solely based on gene expression in a cell type without comparing to other cell types, it would offer great advantages.

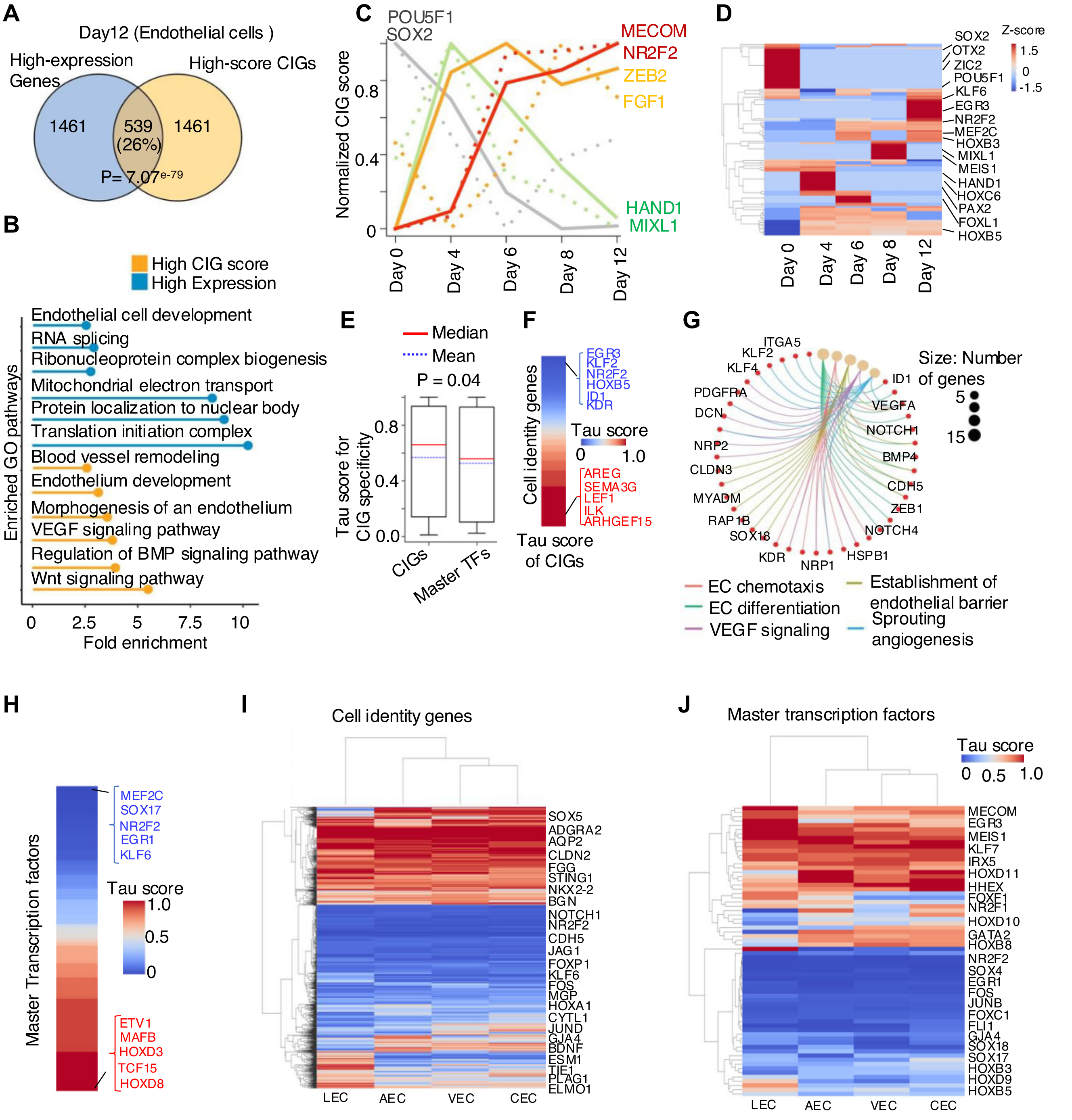


Figure 6. SCIG revealed new insight into EC identity fine-tuning by tissue microenvironment. (**A**) Venn diagram illustrating overlap between the top 10% highly expressed genes and top 10% high-score CIGs identified by SCIG in ECs. (**B**) Pathway enrichment analysis of the top 10% highly expressed-specific genes and top 10% high-score cell identity specific genes identified by SCIG in ECs. (**C**) CIG scores for known marker genes of ESC (SOX2, POU5F1), Mesoderm (MIXL1, HAND1), EC-mesenchymal progenitors (FGF1, ZEB2), and endothelial (MECOM, NR2F2) cells during the ESC to EC differentiation process. (**D**) Heatmap showcasing the identified master TFs of CIGs across different stages of ESC to EC differentiation. (**E**) Box plot showing Tau score of CIGs and their master TFs uncovered for ECs across 15 tissue types. (**F**) Heatmap showing the Tau score of endothelial CIGs. (**G**) Gene-concept network plot displaying pathways enriched in the endothelial CIGs conserved across 15 tissue types. (**H**) Heatmap showing the Tau score of endothelial master TFs of CIGs. (**I**) Heatmap showing Tau score of CIGs in each of the four EC subtypes across 15 tissue types. (**J**) Heatmap showing Tau score of CIG master TFs in each of the four EC subtypes. Data for arterial EC (AEC), venous EC (VEC), capillary EC (CEC) and lymphatic EC (LEC) were presented.

For example, it will be straightforward to implement, making it suitable for large-scale analyses using bulk as well as single-cell expression profiles. However, relying solely on gene expression-based analysis may not yield optimal results, as distinguishing between CIGs and other expressed genes, including housekeeping genes and genes expressed in response to some conditions, can be challenging. Epigenetic modification profiles are useful because CIGs tend to display unique modification patterns that do not enrich other gene categories. However, profiling epigenetic modifications at a single-cell level represents a current technological challenge. The abundance of available single-cell expression data greatly surpasses that of epigenetic data, making gene expression-based approaches more accessible. Thus, we proposed a solution that combines gene expression information with genetic sequence signatures, which overcomes the drawbacks of strategies that rely on expression data alone or in combination with epigenetic profiles.

The utilization of *cis*-regulatory codes in genetic sequence has proven valuable in predicting transcriptional regulation [27, 171], 3D genome organization [172], RNA structure [24], and more. In our study, we employed a comprehensive set of genetic sequence signatures, including the PhyloP100way conservation score, TF-binding motifs, protein-binding motifs in RNA, miRNA-binding motifs, codon biases, and gene architectural features, to characterize CIGs. Our analysis revealed intriguing genetic sequence patterns in CIGs compared to other genes. Specifically, CIGs exhibited greater sequence conservation in their promoters, demonstrating the importance of the conserved sequence in a tight regulation of transcription. Additionally, we observed that CIGs are enriched for binding sites of a large number of TFs and miRNAs, consistent with the reported frequent transcription elongation [9] and low RNA stability of CIGs [12]. Furthermore, the 3′-UTR regions of CIGs are longer than those of other genes, indicating a propensity to recruit regulatory factors, such as miRNA. These findings underscore the significance of *cis*-regulatory codes in DNA sequence for characterizing CIGs and shed light on the potential genetic codes that govern cell identity. Leveraging this information, we developed the SCIG algorithm, which identifies CIGs by combining RNA expression information with these genetic signatures. This novel strategy enables the prediction of CIGs at both the single-cell and bulk levels based on easily accessible RNA-seq and genetic sequence data. Utilizing the CIGs identified by SCIG and GRN information, we further developed the SCIGNet algorithm that analyzes the expression regulation networks of CIGs to define master TFs governing a cell identity. Applying SCIG and SCIGNet to diverse datasets, including HUVEC cells and atrial and ventricular cardiomyocytes, successfully identified experimentally validated CIGs and their master TFs. These results demonstrate SCIG's capability to uncover key cell identity factors involved in cell fate determination.

Utilizing human fetal heart single-cell data, the SCIG-derived CIG score effectively recapitulated cell types identified by gene expression analysis. The CIG score provides additional sequence signature information along with gene expression data, significantly enhancing the capture of cell identities for each cell type. Moreover, network analysis for each cell type using hdWGCNA revealed that CIG score-derived network hub genes align with cell identities. In contrast, expression-derived network hub genes, e.g. for ECs and proliferating cells, are more likely to be enriched for housekeeping-related functions. This result suggests the superiority of the CIG score for constructing CIG network. On the other hand, in the human neurogenesis dataset, the CIG score more effectively organizes cells during single-cell clustering, which reflects differentiation trajectory better when compared to expression-based analysis.

We further applied SCIG to explore the spectrum of CIGs during the differentiation from human ESCs to ECs, recapitulating stage-specific CIGs. For instance, SOX2 [5] and POU5F1 (OCT-4) [5] for ESCs, HAND1 [109] and MIXL1 [110] for mesoderm cells, and NR2F2 [81] and MECOM [83] for ECs were recapitulated. Similarly, exploring the endothelial CIG landscape across 15 tissue types revealed putative CIGs that are either conserved across tissue types or tissue-type specific. Conserved CIGs exhibited stronger enrichment in EC-related functions, while specific CIGs were weakly associated with EC-related pathways but significantly enriched in tissue-specific functions. Intriguingly, the tissue specificity appeared to be greater for CIGs than for their master TFs, suggesting that tissue microenvironment is an important factor in addition to the master TF in cell identity refinement. It will be interesting to investigate in future how microenvironment refines CIG expression program in a cell. We envision that one potential mechanism might be through cellular signaling involving cell–cell communications or environmental signals. In summary, SCIG is a powerful tool providing insights into cell fate determination and regulation.

Acknowledgements

Author contributions: Conceptualization: K.C.; Data curation and Formal analysis: K.A., B.X; Methodology: K.C., K.A; Software: K.A., B.X., K.C.; Validation: K.A., W.P., L.Z., K.C.; Writing – original draft: K.A., K.C.; Writing – review & editing: Y.Y., H.C., W.P., L.Z.;Supervision: K.C.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This project is supported in part by the National Institutes of Health [R01GM125632, R01GM138407, R01HL148338, R01HL133254, and R01CA278832 to K.C., R01HL155632 to L.Z.]; Additional Ventures (Single Ventricle Research Foundation) grant to L.Z. Funding to pay the Open Access publication charges for this article was provided by the NIH.

Data availability

Our curated repository of cell identity genes along with their associated cell type annotations is accessible at https://sites.google.com/view/cigdb/curated-db?authuser=0 and Supplementary Tables S1 and 2. The bulk RNA-seq datasets corresponding to the various cell types were obtained from the ENCODE project (https://www.encodeproject.org/) [29] and NCBI Sequence Read

Archive (SRA) https://www.ncbi.nlm.nih.gov/sra/ [30]. The accession numbers for each RNA-seq dataset are available in Supplementary Table S3. The SCIG software codes can be accessed via https://doi.org/10.5281/zenodo.14726426 or https://github.com/kaifuchenlab/SCIG.

References

1. Xia B, Zhao D, Wang G *et al.* Machine learning uncovers cell identity regulator by histone code. *Nat Commun* 2020;**11**:2696. https://doi.org/10.1038/s41467-020-16539-4
2. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science* 2006;**311**:796–800. https://doi.org/10.1126/science.1113832
3. Davidson EH. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic Press, 2006, https://doi.org/10.1016/B978-0-12-088563-3.X5018-4.
4. Cahan P, Li H, Morris SA *et al.* CellNet: network biology applied to stem cell engineering. *Cell* 2014;**158**:903–15. https://doi.org/10.1016/j.cell.2014.07.020
5. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;**126**:663–76. https://doi.org/10.1016/j.cell.2006.07.024
6. Whyte WA, Orlando DA, Hnisz D *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;**153**:307–19. https://doi.org/10.1016/j.cell.2013.03.035
7. Huang NF, Niiyama H, Peter C *et al.* Embryonic stem cell–derived endothelial cells engraft into the ischemic hindlimb and restore perfusion. *ATVB* 2010;**30**:984–91. https://doi.org/10.1161/ATVBAHA.110.202796
8. Sayed N, Wong WT, Ospino F *et al.* Transdifferentiation of human fibroblasts to endothelial cells: role of innate immunity. *Circulation* 2015;**131**:300–9. https://doi.org/10.1161/CIRCULATIONAHA.113.007394
9. Chen K, Chen Z, Wu D *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* 2015;**47**:1149–57. https://doi.org/10.1038/ng.3385
10. Benayoun BA, Pollina EA, Ucar D *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* 2015;**163**:1281–6. https://doi.org/10.1016/j.cell.2015.10.051
11. Wang G, Xia B, Zhou M *et al.* MACMIC reveals a dual role of CTCF in epigenetic regulation of cell identity genes. *Genomics Proteomics Bioinform* 2021;**19**:140–53. https://doi.org/10.1016/j.gpb.2020.10.008
12. Li Y, Yi Y, Lv J *et al.* Low RNA stability signifies increased post-transcriptional regulation of cell identity genes. *Nucleic Acids Res* 2023;**51**:6020–38. https://doi.org/10.1093/nar/gkad300
13. Wu KJ, Polack A, Dalla-Favera R. Coordinated regulation of iron-controlling genes, H-ferritin and IRP2, by c-MYC. *Science* 1999;**283**:676–9. https://doi.org/10.1126/science.283.5402.676
14. Yu M, Schreek S, Cerni C *et al.* PARP-10, a novel Myc-interacting protein with poly (ADP-ribose) polymerase activity, inhibits transformation. *Oncogene* 2005;**24**:1982–93. https://doi.org/10.1038/sj.onc.1208410
15. Wilson A, Murphy MJ, Oskarsson T *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev* 2004;**18**:2747–63. https://doi.org/10.1101/gad.313104
16. Casey SC, Baylot V, Felsher DW. The MYC oncogene is a global regulator of the immune response. *Blood* 2018;**131**:2007–15. https://doi.org/10.1182/blood-2017-11-742577
17. Farré D, Bellora N, Mularoni L *et al.* Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol* 2007;**8**:R140. https://doi.org/10.1186/gb-2007-8-7-r140
18. Ganapathi M, Srivastava P, Das Sutar SK *et al.* Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC Bioinf* 2005;**6**:126. https://doi.org/10.1186/1471-2105-6-126
19. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. *Trends Genet* 2013;**29**:569–74. https://doi.org/10.1016/j.tig.2013.05.010
20. Joshi CJ, Ke W, Drangowska-Way A *et al.* What are housekeeping genes? *PLoS Comput Biol* 2022;**18**:e1010295. https://doi.org/10.1371/journal.pcbi.1010295
21. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet* 2003;**19**:362–5. https://doi.org/10.1016/S0168-9525(03)00140-9
22. Kierzek E, Zhang X, Watson RM *et al.* Secondary structure prediction for RNA sequences including N6-methyladenosine. *Nat Commun* 2022;**13**:1271. https://doi.org/10.1038/s41467-022-28817-4
23. Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. *Methods Mol Biol* 2012;**905**:99–122.
24. Sato K, Akiyama M, Sakakibara Y. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat Commun* 2021;**12**:941. https://doi.org/10.1038/s41467-021-21194-4
25. Avsec Ž, Weilert M, Shrikumar A *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 2021;**53**:354–66. https://doi.org/10.1038/s41588-021-00782-6
26. Wayment-Steele HK, Kladwang W, Watkins AM *et al.* Deep learning models for predicting RNA degradation via dual crowdsourcing. *Nat Mach Intell* 2022;**4**:1174–84. https://doi.org/10.1038/s42256-022-00571-8
27. Chen KM, Wong AK, Troyanskaya OG *et al.* A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* 2022;**54**:940–9. https://doi.org/10.1038/s41588-022-01102-2
28. Kelley DR, Reshef YA, Bileschi M *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* 2018;**28**:739–50. https://doi.org/10.1101/gr.227819.117
29. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;**583**:699–710.
30. Katz K, Shutov O, Lapoint R *et al.* The sequence read Archive: a decade more of explosive growth. *Nucleic Acids Res* 2022;**50**:D387–90. https://doi.org/10.1093/nar/gkab1053
31. Cunningham F, Allen JE, Allen J *et al.* Ensembl 2022. *Nucleic Acids Res* 2022;**50**:D988–95. https://doi.org/10.1093/nar/gkab1049
32. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. https://doi.org/10.1093/bioinformatics/bts635
33. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. https://doi.org/10.1093/bioinformatics/btt656
34. Han X, Wang R, Zhou Y *et al.* Mapping the mouse cell atlas by Microwell-Seq. *Cell* 2018;**173**:1307. https://doi.org/10.1016/j.cell.2018.05.012
35. GTEx Consortium The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5. https://doi.org/10.1038/ng.2653
36. Abugessaisa I, Ramilowski JA, Lizio M *et al.* FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res* 2021;**49**:D892–8. https://doi.org/10.1093/nar/gkaa1054
37. Uhlén M, Fagerberg L, Hallström BM *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419. https://doi.org/10.1126/science.1260419

38. Sjöstedt E, Zhong W, Fagerberg L *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* 2020;**367**:eaay5947. https://doi.org/10.1126/science.aay5947
39. Shen Y, Yue F, McCleary DF *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* 2012;**488**:116–20. https://doi.org/10.1038/nature11243
40. Sheng X, Wu J, Sun Q *et al.* MTD: a mammalian transcriptomic database to explore gene expression and regulation. *Brief Bioinform* 2017;**18**:28–36. https://doi.org/10.1093/bib/bbv117
41. Camargo AP, Vasconcelos AA, Fiamenghi MB *et al.* Tspex: a tissue-specificity calculator for gene expression data. Research Square, https://doi.org/10.21203/rs.3.rs-51998/v1, 4 August 2020, preprint: not peer reviewed.
42. Shannon P, Richards M. MotifDb: an annotated collection of protein-DNA binding sequence motifs. 2023, https://bioconductor.org/packages/MotifDb
43. Benoit Bouvrette LP, Bovaird S, Blanchette M *et al.* oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res* 2020;**48**:D166–73. https://doi.org/10.1093/nar/gkz986
44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033
45. Sticht C, De La Torre C, Parveen A *et al.* miRWalk: an online resource for prediction of microRNA binding sites. *PLoS One* 2018;**13**:e0206239. https://doi.org/10.1371/journal.pone.0206239
46. Pollard KS, Hubisz MJ, Rosenbloom KR *et al.* Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**:110–21. https://doi.org/10.1101/gr.097857.109
47. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist* 1947;**18**:50–60. https://doi.org/10.1214/aoms/1177730491
48. Kulandaisamy A, Zaucha J, Frishman D *et al.* MPTherm-pred: analysis and prediction of thermal stability changes upon mutations in transmembrane proteins. *J Mol Biol* 2021;**433**:166646. https://doi.org/10.1016/j.jmb.2020.09.005
49. Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
50. Lambert SA, Jolma A, Campitelli LF *et al.* The Human transcription factors. *Cell* 2018;**172**:650–65. https://doi.org/10.1016/j.cell.2018.01.029
51. Chawla NV, Bowyer KW, Hall LO *et al.* SMOTE: synthetic minority over-sampling technique. *Jair* 2002;**16**:321–57. https://doi.org/10.1613/jair.953
52. Liu Z-P, Wu C, Miao H *et al.* RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015;**2015**:bav095. https://doi.org/10.1093/database/bav095
53. Garcia-Alonso L, Holland CH, Ibrahim MM *et al.* Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* 2019;**29**:1363–75. https://doi.org/10.1101/gr.240663.118
54. Fang L, Li Y, Ma L *et al.* GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Res* 2021;**49**:D97–D103. https://doi.org/10.1093/nar/gkaa995
55. Xu Q, Georgiou G, Frölich S *et al.* ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res* 2021;**49**:7966–85. https://doi.org/10.1093/nar/gkab598
56. Sonawane AR, Platig J, Fagny M *et al.* Understanding tissue-specific gene regulation. *Cell Rep* 2017;**21**:1077–88. https://doi.org/10.1016/j.celrep.2017.10.001
57. Huang JK, Carlin DE, Yu MK *et al.* Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst* 2018;**6**:484–95. https://doi.org/10.1016/j.cels.2018.03.001
58. Suryawanshi H, Clancy R, Morozov P *et al.* Cell atlas of the foetal human heart and implications for autoimmune-mediated congenital heart block. *Cardiovasc Res* 2020;**116**:1446–57. https://doi.org/10.1093/cvr/cvz257
59. Morabito S, Reese F, Rahimzadeh N *et al.* hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods* 2023;**3**:100498. https://doi.org/10.1016/j.crmeth.2023.100498
60. La Manno G, Soldatov R, Zeisel A *et al.* RNA velocity of single cells. *Nature* 2018;**560**:494–8. https://doi.org/10.1038/s41586-018-0414-6
61. Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87. https://doi.org/10.1016/j.cell.2021.04.048
62. Bergen V, Lange M, Peidli S *et al.* Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;**38**:1408–14. https://doi.org/10.1038/s41587-020-0591-3
63. Lange M, Bergen V, Klein M *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* 2022;**19**:159–70. https://doi.org/10.1038/s41592-021-01346-6
64. McCracken IR, Taylor RS, Kok FO *et al.* Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *Eur Heart J* 2020;**41**:1024–36. https://doi.org/10.1093/eurheartj/ehz351
65. Li M, Zhang X, Ang KS *et al.* DISCO: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res* 2022;**50**:D596–602. https://doi.org/10.1093/nar/gkab1020
66. Wu T, Hu E, Xu S *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov J* 2021;**2**:100141. https://doi.org/10.1016/j.xinn.2021.100141
67. Robin X, Turck N, Hainard A *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf* 2011;**12**:77. https://doi.org/10.1186/1471-2105-12-77
68. Hnisz D, Abraham BJ, Lee TI *et al.* Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–47. https://doi.org/10.1016/j.cell.2013.09.053
69. Hounkpe BW, Chenou F, de Lima F *et al.* HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Res* 2021;**49**:D947–55. https://doi.org/10.1093/nar/gkaa609
70. Wang D, Liu S, Warrell J *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* 2018;**362**:eaat8464. https://doi.org/10.1126/science.aat8464
71. Little P, Liu S, Zhabotynsky V *et al.* A computational method for cell type-specific expression quantitative trait loci mapping using bulk RNA-seq data. *Nat Commun* 2023;**14**:3030. https://doi.org/10.1038/s41467-023-38795-w
72. Kim J, Rothová MM, Madan E *et al.* Neighbor-specific gene expression revealed from physically interacting cells during mouse embryonic development. *Proc Natl Acad Sci USA* 2023;**120**:e2205371120. https://doi.org/10.1073/pnas.2205371120
73. Eisen TJ, Eichhorn SW, Subtelny AO *et al.* MicroRNAs cause accelerated decay of short-tailed target mRNAs. *Mol Cell* 2020;**77**:775–785.e8.e8. https://doi.org/10.1016/j.molcel.2019.12.004
74. Chekulaeva M. First demonstration of miRNA-dependent mRNA decay. *Nat Rev Mol Cell Biol* 2023;**24**:164–. https://doi.org/10.1038/s41580-022-00557-9
75. Krieger F, Möglich A, Kiefhaber T. Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. *J Am Chem Soc* 2005;**127**:3346–52. https://doi.org/10.1021/ja042798i
76. Kulandaisamy A, Lathi V, ViswaPoorani K *et al.* Important amino acid residues involved in folding and binding of protein–protein complexes. *Int J Biol Macromol* 2017;**94**:438–44. https://doi.org/10.1016/j.ijbiomac.2016.10.045
77. Yan BX, Sun YQ. Glycine residues provide flexibility for enzyme active sites. *J Biol Chem* 1997;**272**:3190–4. https://doi.org/10.1074/jbc.272.6.3190

78. Borisova OF, Shchyolkina AK, Chernov BK *et al.* Relative stability of AT and GC pairs in parallel DNA duplex formed by a natural sequence. *FEBS Lett* 1993;**322**:304–6. https://doi.org/10.1016/0014-5793(93)81591-M
79. Hia F, Yang SF, Shichino Y *et al.* Codon bias confers stability to human mRNAs. *EMBO Rep* 2019;**20**:e48220. https://doi.org/10.15252/embr.201948220
80. Nakato R, Wada Y, Nakaki R *et al.* Comprehensive epigenome characterization reveals diverse transcriptional regulation across human vascular endothelial cells. *Epigenetics Chromatin* 2019;**12**:77. https://doi.org/10.1186/s13072-019-0319-0
81. You L-R, Lin F-J, Lee CT *et al.* Suppression of Notch signalling by the COUP-TFII transcription factor regulates vein identity. *Nature* 2005;**435**:98–104. https://doi.org/10.1038/nature03511
82. Sissaoui S, Yu J, Yan A *et al.* Genomic characterization of endothelial enhancers reveals a multifunctional role for NR2F2 in regulation of arteriovenous gene expression. *Circ Res* 2020;**126**:875–88. https://doi.org/10.1161/CIRCRESAHA.119.316075
83. Lv J, Meng S, Gu Q *et al.* Epigenetic landscape reveals MECOM as an endothelial lineage regulator. *Nat Commun* 2023;**14**:2390. https://doi.org/10.1038/s41467-023-38002-w
84. Martínez-Ramos S, Rafael-Vidal C, Malvar-Fernández B *et al.* HOXA5 is a key regulator of class 3 semaphorins expression in the synovium of rheumatoid arthritis patients. *Rheumatology* 2023;**62**:2621–30. https://doi.org/10.1093/rheumatology/keac654
85. Froese N, Kattih B, Breitbart A *et al.* GATA6 promotes angiogenic function and survival in endothelial cells by suppression of autocrine transforming growth factor β/activin receptor-like kinase 5 signaling. *J Biol Chem* 2011;**286**:5680–90. https://doi.org/10.1074/jbc.M110.176925
86. Zheng R, Wan C, Mei S *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* 2019;**47**:D729–35. https://doi.org/10.1093/nar/gky1094
87. Kamimoto K, Stringa B, Hoffmann CM *et al.* Dissecting cell identity via network inference and *in silico* gene perturbation. *Nature* 2023;**614**:742–51. https://doi.org/10.1038/s41586-022-05688-9
88. Yamamizu K, Piao Y, Sharov AA *et al.* Identification of transcription factors for lineage-specific ESC differentiation. *Stem Cell Rep* 2013;**1**:545–59. https://doi.org/10.1016/j.stemcr.2013.10.006
89. Guerrero-Ramirez G, Valdez-Cordoba C, Islas-Cisneros J *et al.* Computational approaches for predicting key transcription factors in targeted cell reprogramming (Review). *Mol Med Rep* 2018;**18**:1225–37. https://doi.org/10.3892/mmr.2018.9092
90. Paul F, Arkin Y, Giladi A *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* 2015;**163**:1663–77. https://doi.org/10.1016/j.cell.2015.11.013
91. Zhou Y, Xuan Y, Liu Y *et al.* Transcription factor FOXP1 mediates vascular endothelial dysfunction in diabetic retinopathy. *Graefes Arch Clin Exp Ophthalmol* 2022;**260**:3857–67. https://doi.org/10.1007/s00417-022-05698-3
92. Cheng CK, Lin X, Pu Y *et al.* SOX4 is a novel phenotypic regulator of endothelial cells in atherosclerosis revealed by single-cell analysis. *J Adv Res* 2023;**43**:187–203. https://doi.org/10.1016/j.jare.2022.02.017
93. Maiti D, Xu Z, Duh EJ. Vascular endothelial growth factor induces MEF2C and MEF2-dependent activity in endothelial cells. *Invest Ophthalmol Vis Sci* 2008;**49**:3640–8. https://doi.org/10.1167/iovs.08-1760
94. Kalna V, Yang Y, Peghaire CR *et al.* The transcription factor ERG regulates super-enhancers associated with an endothelial-specific gene expression program. *Circ Res* 2019;**124**:1337–49. https://doi.org/10.1161/CIRCRESAHA.118.313788
95. Cao Y, Zhang X, Akerberg BN *et al.* *In vivo* dissection of chamber-selective enhancers reveals estrogen-related receptor as a regulator of ventricular cardiomyocyte identity. *Circulation* 2023;**147**:881–96. https://doi.org/10.1161/CIRCULATIONAHA.122.061955
96. Sweat ME, Cao Y, Zhang X *et al.* Tbx5 maintains atrial identity in postnatal cardiomyocytes by regulating an atrial-specific enhancer network. *Nat Cardiovasc Res* 2023;**2**:881–98. https://doi.org/10.1038/s44161-023-00334-7
97. Shu W, Yang H, Zhang L *et al.* Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. *J Biol Chem* 2001;**276**:27488–97. https://doi.org/10.1074/jbc.M100636200
98. Ren K-W, Yu X-H, Gu Y-H *et al.* Cardiac-specific knockdown of Bhlhe40 attenuates angiotensin II (Ang II)-induced atrial fibrillation in mice. *Front Cardiovasc Med* 2022;**9**:957903. https://doi.org/10.3389/fcvm.2022.957903
99. Chang C-P, Stankunas K, Shang C *et al.* Pbx1 functions in distinct regulatory networks to pattern the great arteries and cardiac outflow tract. *Development* 2008;**135**:3577–86. https://doi.org/10.1242/dev.022350
100. Sotoodehnia N, Isaacs A, de Bakker PIW *et al.* Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat Genet* 2010;**42**:1068–76. https://doi.org/10.1038/ng.716
101. Wu T, Liang Z, Zhang Z *et al.* PRDM16 is a compact myocardium-enriched transcription factor required to maintain compact myocardial cardiomyocyte identity in left ventricle. *Circulation* 2022;**145**:586–602. https://doi.org/10.1161/CIRCULATIONAHA.121.056666
102. Seya D, Ihara D, Shirai M *et al.* A role of Hey2 transcription factor for right ventricle development through regulation of Tbx2-mycn pathway during cardiac morphogenesis. *Dev Growth Differ* 2021;**63**:82–92. https://doi.org/10.1111/dgd.12707
103. Liu Z, Li W, Ma X *et al.* Essential role of the zinc finger transcription factor Casz1 for mammalian cardiac morphogenesis and development. *J Biol Chem* 2014;**289**:29801–16. https://doi.org/10.1074/jbc.M114.570416
104. Zeisberg EM, Tarnavski O, Zeisberg M *et al.* Endothelial-to-mesenchymal transition contributes to cardiac fibrosis. *Nat Med* 2007;**13**:952–61. https://doi.org/10.1038/nm1613
105. Kovacic JC, Dimmeler S, Harvey RP *et al.* Endothelial to mesenchymal transition in cardiovascular disease: JACC state-of-the-art review. *J Am Coll Cardiol* 2019;**73**:190–209. https://doi.org/10.1016/j.jacc.2018.09.089
106. Colin-York H, Li D, Korobchevskaya K *et al.* Cytoskeletal actin patterns shape mast cell activation. *Commun Biol* 2019;**2**:93. https://doi.org/10.1038/s42003-019-0322-9
107. Pastwińska J, Żelechowska P, Walczak-Drzewiecka A *et al.* The art of mast cell adhesion. *Cells* 2020;**9**:2664.
108. Zhang Z, Zhang X. Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity. *Cell Rep Methods* 2021;**1**:100095. https://doi.org/10.1016/j.crmeth.2021.100095
109. Barnes RM, Firulli BA, Conway SJ *et al.* Analysis of the Hand1 cell lineage reveals novel contributions to cardiovascular, neural crest, extra-embryonic, and lateral mesoderm derivatives. *Dev Dyn* 2010;**239**:3086–97. https://doi.org/10.1002/dvdy.22428
110. Zhang H, Fraser ST, Papazoglu C *et al.* Transcriptional activation by the Mixl1 homeodomain protein in differentiating mouse embryonic stem cells. *Stem Cells* 2009;**27**:2884–95. https://doi.org/10.1002/stem.203
111. van den Bos C, Mosca JD, Winkles J *et al.* Human mesenchymal stem cells respond to fibroblast growth factors. *Hum Cell* 1997;**10**:45–50.
112. DaSilva-Arnold SC, Kuo C-Y, Davra V *et al.* ZEB2, a master regulator of the epithelial–mesenchymal transition, mediates trophoblast differentiation. *Mol Hum Reprod* 2019;**25**:61–75. https://doi.org/10.1093/molehr/gay053

113. Acampora D, Di Giovannantonio LG, Simeone A. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development* 2013;**140**:43–55. https://doi.org/10.1242/dev.085290
114. Patel SR, Dressler GR. Expression of Pax2 in the intermediate mesoderm is regulated by YY1. *Dev Biol* 2004;**267**:505–16. https://doi.org/10.1016/j.ydbio.2003.11.002
115. Larsen BM, Hrycaj SM, Newman M *et al.* Mesenchymal Hox6 function is required for mouse pancreatic endocrine cell differentiation. *Development* 2015;**142**:3859–68.
116. Gallardo-Vara E, Blanco FJ, Roqué M *et al.* Transcription factor KLF6 upregulates expression of metalloprotease MMP14 and subsequent release of soluble endoglin during vascular injury. *Angiogenesis* 2016;**19**:155–71. https://doi.org/10.1007/s10456-016-9495-8
117. Liu D, Evans I, Britton G *et al.* The zinc-finger transcription factor, early growth response 3, mediates VEGF-induced angiogenesis. *Oncogene* 2008;**27**:2989–98. https://doi.org/10.1038/sj.onc.1210959
118. Suehiro J-I, Hamakubo T, Kodama T *et al.* Vascular endothelial growth factor activation of endothelial cells is mediated by early growth response-3. *Blood* 2010;**115**:2520–32. https://doi.org/10.1182/blood-2009-07-233478
119. Lin Z, Natesan V, Shi H *et al.* Kruppel-like factor 2 regulates endothelial barrier function. *ATVB* 2010;**30**:1952–9. https://doi.org/10.1161/ATVBAHA.110.211474
120. Sangwung P, Zhou G, Nayak L *et al.* KLF2 and KLF4 control endothelial identity and vascular integrity. *JCI Insight* 2017;**2**:e91700. https://doi.org/10.1172/jci.insight.91700
121. Wu Y, Moser M, Bautch VL *et al.* HoxB5 is an upstream transcriptional switch for differentiation of the vascular endothelium from precursor cells. *Mol Cell Biol* 2003;**23**:5680–91. https://doi.org/10.1128/MCB.23.16.5680-5691.2003
122. Gadomski S, Singh SK, Singh S *et al.* Id1 and Id3 maintain steady-State hematopoiesis by promoting sinusoidal endothelial cell survival and regeneration. *Cell Rep* 2020;**31**:107572. https://doi.org/10.1016/j.celrep.2020.107572
123. Terman BI, Dougher-Vermazen M, Carrion ME *et al.* Identification of the KDR tyrosine kinase as a receptor for vascular endothelial cell growth factor. *Biochem Biophys Res Commun* 1992;**187**:1579–86. https://doi.org/10.1016/0006-291X(92)90483-2
124. Florentin J, Zhao J, Tai Y-Y *et al.* Loss of amphiregulin drives inflammation and endothelial apoptosis in pulmonary hypertension. *Life Sci Alliance* 2022;**5**:e202101264. https://doi.org/10.26508/lsa.202101264
125. Chen D-Y, Sun N-H, Chen X *et al.* Endothelium-derived semaphorin 3G attenuates ischemic retinopathy by coordinating β-catenin–dependent vascular remodeling. *J Clin Invest* 2021;**131**:e135296. https://doi.org/10.1172/JCI135296
126. Planutiene M, Planutis K, Holcombe RF. Lymphoid enhancer-binding factor 1, a representative of vertebrate-specific Lef1/Tcf1 sub-family, is a Wnt-beta-catenin pathway target gene in human endothelial cells which regulates matrix metalloproteinase-2 expression and promotes endothelial cell invasion. *Vasc Cell* 2011;**3**:28.
127. Friedrich EB, Liu E, Sinha S *et al.* Integrin-linked kinase regulates endothelial cell survival and vascular development. *Mol Cell Biol* 2004;**24**:8134–44. https://doi.org/10.1128/MCB.24.18.8134-8144.2004
128. Kusuhara S, Fukushima Y, Fukuhara S *et al.* Arhgef15 promotes retinal angiogenesis by mediating VEGF-induced Cdc42 activation and potentiating RhoJ inactivation in endothelial cells. *PLoS One* 2012;**7**:e45858. https://doi.org/10.1371/journal.pone.0045858
129. Liu M, Zhang L, Marsboom G *et al.* Sox17 is required for endothelial regeneration following inflammation-induced vascular injury. *Nat Commun* 2019;**10**:2126. https://doi.org/10.1038/s41467-019-10134-y
130. Pereira FA, Qiu Y, Zhou G *et al.* The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes Dev* 1999;**13**:1037–49. https://doi.org/10.1101/gad.13.8.1037
131. Fu M, Zhu X, Zhang J *et al.* Egr-1 target genes in human endothelial cells identified by microarray analysis. *Gene* 2003;**315**:33–41. https://doi.org/10.1016/S0378-1119(03)00730-3
132. Jeong H-W, Hernández-Rodríguez B, Kim J *et al.* Transcriptional regulation of endothelial cell behavior during sprouting angiogenesis. *Nat Commun* 2017;**8**:726. https://doi.org/10.1038/s41467-017-00738-7
133. Meadows SM, Myers CT, Krieg PA. Regulation of endothelial cell development by ETS transcription factors. *Semin Cell Dev Biol* 2011;**22**:976–84. https://doi.org/10.1016/j.semcdb.2011.09.009
134. Coppiello G, Collantes M, Sirerol-Piquer MS *et al.* Meox2/Tcf15 heterodimers program the heart capillary endothelium for cardiac fatty acid uptake. *Circulation* 2015;**131**:815–26. https://doi.org/10.1161/CIRCULATIONAHA.114.013721
135. Toshner M, Dunmore BJ, McKinney EF *et al.* Transcript analysis reveals a specific HOX signature associated with positional identity of human endothelial cells. *PLoS One* 2014;**9**:e91334. https://doi.org/10.1371/journal.pone.0091334
136. Fang JS, Coon BG, Gillis N *et al.* Shear-induced notch-Cx37-p27 axis arrests endothelial cell cycle to enable arterial specification. *Nat Commun* 2017;**8**:2149. https://doi.org/10.1038/s41467-017-01742-7
137. Ivins S, Chappell J, Vernay B *et al.* The CXCL12/CXCR4 axis plays a critical role in coronary artery development. *Dev Cell* 2015;**33**:455–68. https://doi.org/10.1016/j.devcel.2015.03.026
138. Ara T, Tokoyoda K, Okamoto R *et al.* The role of CXCL12 in the organ-specific process of artery formation. *Blood* 2005;**105**:3155–61. https://doi.org/10.1182/blood-2004-07-2563
139. Villa N, Walker L, Lindsell CE *et al.* Vascular expression of Notch pathway receptors and ligands is restricted to arterial vessels. *Mech Dev* 2001;**108**:161–4. https://doi.org/10.1016/S0925-4773(01)00469-5
140. Benwell CJ, Taylor JAGE, Robinson SD. Endothelial neuropilin-2 influences angiogenesis by regulating actin pattern development and α5-integrin-p-FAK complex recruitment to assembling adhesion sites. *FASEB j* 2021;**35**:e21679. https://doi.org/10.1096/fj.202100286R
141. Dulak J, Józkowicz A, Łoboda A. *Angiogenesis and vascularisation: cellular and molecular mechanisms in health and diseases*. Springer Science & Business Media, 2014, https://doi.org/10.1007/978-3-7091-1428-5
142. Iso T, Maeda K, Hanaoka H *et al.* Capillary endothelial fatty acid binding proteins 4 and 5 play a critical role in fatty acid uptake in heart and skeletal muscle. *ATVB* 2013;**33**:2549–57. https://doi.org/10.1161/ATVBAHA.113.301588
143. Kalucka J, de Rooij LPMH, Goveia J *et al.* Single-cell transcriptome atlas of murine endothelial cells. *Cell* 2020;**180**:764–779.e20. https://doi.org/10.1016/j.cell.2020.01.015
144. Watanabe K, Ohta Y, Toba K *et al.* Myocardial CD36 expression and fatty acid accumulation in patients with type I and II CD36 deficiency. *Ann Nucl Med* 1998;**12**:261–6. https://doi.org/10.1007/BF03164911
145. Wigle JT, Harvey N, Detmar M *et al.* An essential role for Prox1 in the induction of the lymphatic endothelial cell phenotype. *EMBO J* 2002;**21**:1505–13. https://doi.org/10.1093/emboj/21.7.1505
146. Podgrabinska S, Braun P, Velasco P *et al.* Molecular characterization of lymphatic endothelial cells. *Proc Natl Acad Sci USA* 2002;**99**:16069–74. https://doi.org/10.1073/pnas.242401399

147. Fatima A, Wang Y, Uchida Y *et al.* Foxc1 and Foxc2 deletion causes abnormal lymphangiogenesis and correlates with ERK hyperactivation. *J Clin Invest* 2016;**126**:2437–51. https://doi.org/10.1172/JCI80465
148. Norden PR, Sabine A, Wang Y *et al.* Shear stimulation of FOXC1 and FOXC2 differentially regulates cytoskeletal activity during lymphatic valve maturation. *eLife* 2020;**9**:e53814. https://doi.org/10.7554/eLife.53814
149. Witzenbichler B, Asahara T, Murohara T *et al.* Vascular endothelial growth factor-C (VEGF-C/VEGF-2) promotes angiogenesis in the setting of tissue ischemia. *Am J Pathol* 1998;**153**:381–94. https://doi.org/10.1016/S0002-9440(10)65582-4
150. Miller AZ, Satchie A, Tannenbaum AP *et al.* Expandable arterial endothelial precursors from human CD34+ cells differ in their proclivity to undergo an endothelial-to-mesenchymal transition. *Stem Cell Rep* 2018;**10**:73–86. https://doi.org/10.1016/j.stemcr.2017.12.011
151. Shin D, Garcia-Cardena G, Hayashi S *et al.* Expression of ephrinB2 identifies a stable genetic difference between arterial and venous vascular smooth muscle as well as endothelial cells, and marks subsets of microvessels at sites of adult neovascularization. *Dev Biol* 2001;**230**:139–50. https://doi.org/10.1006/dbio.2000.9957
152. Brandt MM, van Dijk CGM, Chrifi I *et al.* Endothelial loss of Fzd5 stimulates PKC/Ets1-mediated transcription of Angpt2 and Flt1. *Angiogenesis* 2018;**21**:805–21. https://doi.org/10.1007/s10456-018-9625-6
153. Zhu C, Yu Y, Montani J-P *et al.* Arginase-I enhances vascular endothelial inflammation and senescence through eNOS-uncoupling. *BMC Res Notes* 2017;**10**:82. https://doi.org/10.1186/s13104-017-2399-x
154. Wang X, Abraham S, McKenzie JAG *et al.* LRG1 promotes angiogenesis by modulating endothelial TGF-β signalling. *Nature* 2013;**499**:306–11. https://doi.org/10.1038/nature12345
155. Ayloo S, Lazo CG, Sun S *et al.* Pericyte-to-endothelial cell signaling via vitronectin-integrin regulates blood-CNS barrier. *Neuron* 2022;**110**:1641–55. https://doi.org/10.1016/j.neuron.2022.02.017
156. Wang MM, Zhang X, Lee SJ *et al.* Expression of periaxin (PRX) specifically in the human cerebrovascular system: PDZ domain-mediated strengthening of endothelial barrier function. *Sci Rep* 2018;**8**:10042. https://doi.org/10.1038/s41598-018-28190-7
157. Lien M-Y, Tsai H-C, Chang A-C *et al.* Chemokine CCL4 induces vascular endothelial growth factor C expression and lymphangiogenesis by miR-195-3p in oral Squamous cell carcinoma. *Front Immunol* 2018;**9**:412. https://doi.org/10.3389/fimmu.2018.00412
158. Luu TT, Søndergaard JN, Peña-Pérez L *et al.* FOXO1 and FOXO3 cooperatively regulate innate lymphoid cell development. *Front Immunol* 2022;**13**:854312. https://doi.org/10.3389/fimmu.2022.854312
159. Xiang M, Grosso RA, Takeda A *et al.* A single-cell transcriptional roadmap of the mouse and Human lymph node lymphatic vasculature. *Front Cardiovasc Med* 2020;**7**:52. https://doi.org/10.3389/fcvm.2020.00052
160. Corada M, Orsenigo F, Morini MF *et al.* Sox17 is indispensable for acquisition and maintenance of arterial identity. *Nat Commun* 2013;**4**:2609. https://doi.org/10.1038/ncomms3609
161. Seki T, Yun J, Oh SP. Arterial endothelium-specific activin receptor-like kinase 1 expression suggests its role in arterialization and vascular remodeling. *Circ Res* 2003;**93**:682–9. https://doi.org/10.1161/01.RES.0000095246.40391.3B
162. Trimm E, Red-Horse K. Vascular endothelial cell development and diversity. *Nat Rev Cardiol* 2023;**20**:197–210. https://doi.org/10.1038/s41569-022-00770-1
163. De Val S, Anderson JP, Heidt AB *et al.* Mef2c is activated directly by Ets transcription factors through an evolutionarily conserved endothelial cell-specific enhancer. *Dev Biol* 2004;**275**:424–34. https://doi.org/10.1016/j.ydbio.2004.08.016
164. Stoltz RA, Abraham NG, Laniado-Schwartzman M. The role of NF-kappaB in the angiogenic response of coronary microvessel endothelial cells. *Proc Natl Acad Sci USA* 1996;**93**:2832–7. https://doi.org/10.1073/pnas.93.7.2832
165. Licht AH, Pein OT, Florin L *et al.* JunB is required for endothelial cell morphogenesis by regulating core-binding factor beta. *J Cell Biol* 2006;**175**:981–91. https://doi.org/10.1083/jcb.200605149
166. Wang T, Wang Z, de Fabritus L *et al.* 1-Deoxysphingolipids bind to COUP-TF to modulate lymphatic and cardiac cell development. *Dev Cell* 2021;**56**:3128–45. https://doi.org/10.1016/j.devcel.2021.10.018
167. Watabe T. Roles of transcriptional network during the formation of lymphatic vessels. *J Biochem* 2012;**152**:213–20. https://doi.org/10.1093/jb/mvs081
168. Haque A, Engel J, Teichmann SA *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**:75. https://doi.org/10.1186/s13073-017-0467-4
169. Van de Sande B, Lee JS, Mutasa-Gottgens E *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov* 2023;**22**:496–520. https://doi.org/10.1038/s41573-023-00688-4
170. Wen L, Li G, Huang T *et al.* Single-cell technologies: from research to application. *Innovation* 2022;**3**:100342.
171. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat Biotechnol* 2024;**42**:927–35. https://doi.org/10.1038/s41587-023-01905-6
172. Zhang Y, Boninsegna L, Yang M *et al.* Computational methods for analysing multiscale 3D genome organization. *Nat Rev Genet* 2024;**25**:123–41. https://doi.org/10.1038/s41576-023-00638-1

Received: August 21, 2024. **Revised:** April 9, 2025. **Editorial Decision:** April 18, 2025. **Accepted:** May 9, 2025