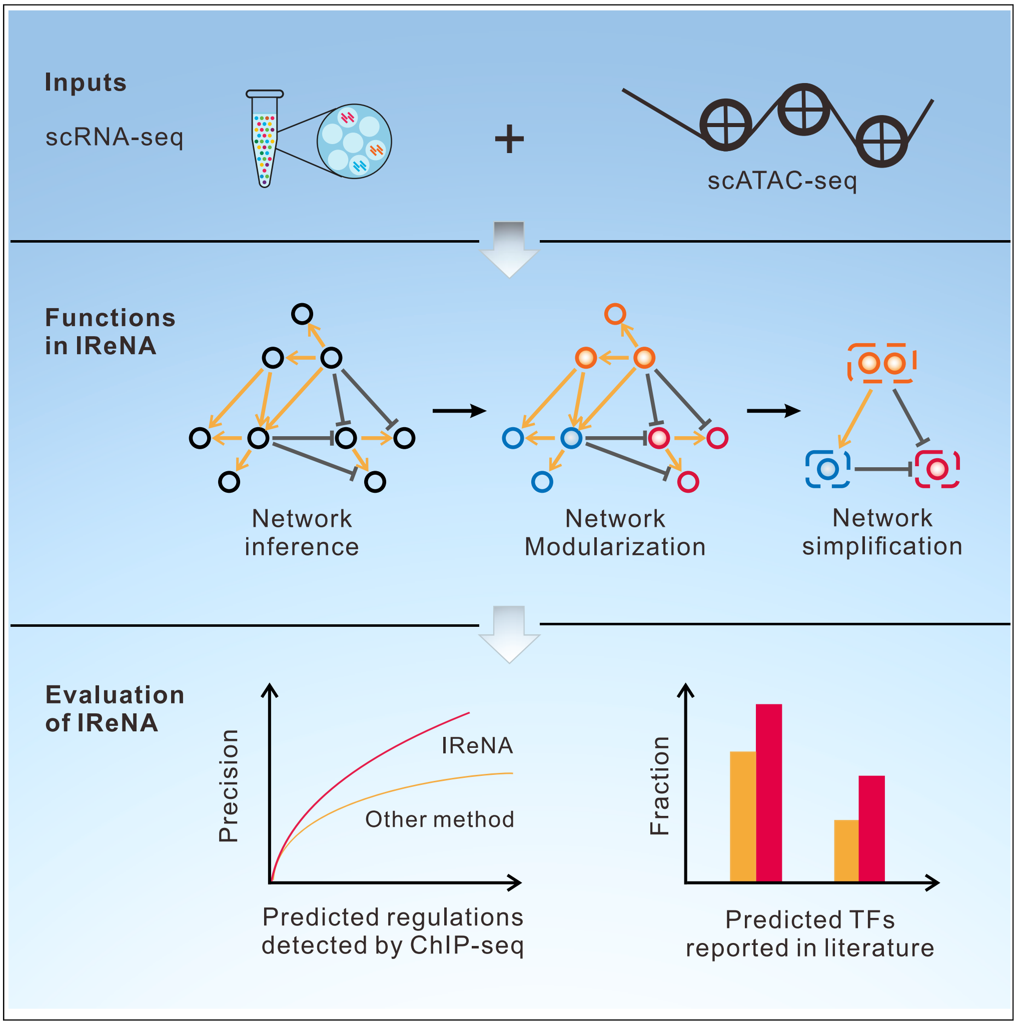
iScience

CellPress
OPEN ACCESS

Article

IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles



Junyao Jiang, Pin Lyu, Jinlian Li, ..., Seth Blackshaw, Jiang Qian, Jie Wang

wang_jie01@gibh.ac.cn

Highlights

IReNA infers regulatory networks using single-cell RNA-seq and ATAC-seq data

IReNA establishes modular regulatory networks to identify key regulators

IReNA specifically constructs simplified regulatory networks among modules

Applying to public datasets, IReNA shows a better performance on network analysis

Jiang et al., iScience *25*, 105359
November 18, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.105359

iScience

CellPress OPEN ACCESS

Article

IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles

Junyao Jiang,[1,6,7] Pin Lyu,[2,6] Jinlian Li,[1,6] Sunan Huang,[1] Jiawang Tao,[1] Seth Blackshaw,[3] Jiang Qian,[2] and Jie Wang[1,4,5,8,*]

SUMMARY

Recently, single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) have been developed to separately measure transcriptomes and chromatin accessibility profiles at the single-cell resolution. However, few methods can reliably integrate these data to perform regulatory network analysis. Here, we developed integrated regulatory network analysis (IReNA) for network inference through the integrated analysis of scRNA-seq and scATAC-seq data, network modularization, transcription factor enrichment, and construction of simplified intermodular regulatory networks. Using public datasets, we showed that integrated network analysis of scRNA-seq data with scATAC-seq data is more precise to identify known regulators than scRNA-seq data analysis alone. Moreover, IReNA outperformed currently available methods in identifying known regulators. IReNA facilitates the systems-level understanding of biological regulatory mechanisms and is available at https://github.com/jiang-junyao/IReNA.

INTRODUCTION

Dynamic changes of *trans*-regulators (e.g., transcription factors) and *cis*-regulatory elements (e.g., promoters) control gene expression in biological systems (Thompson et al., 2015). This fact makes it possible to infer gene regulatory networks using transcriptomic and epigenomic profiles. Recent advances in single-cell sequencing technologies provide new opportunities to reconstruct cell-type-specific regulatory networks (Macosko et al., 2015). Single-cell transcriptomes have been widely detected through single-cell RNA sequencing (scRNA-seq). Currently, dozens of methods have used scRNA-seq data to infer regulatory networks, including top-performing methods GENIE3 and PIDC (Pratapa et al., 2020; Huynh-Thu et al., 2010; Chan et al., 2017). Software SCODE was also developed to infer regulatory networks from scRNA-seq data based on ordinary differential equations (Matsumoto et al., 2017). Complementary to scRNA-seq, epigenomic profiling technique of assay for transposase-accessible chromatin using sequencing (ATAC-seq), including the latest single-cell ATAC-seq (scATAC-seq), measures accessible states of *cis*-regulatory elements to *trans*-regulators facilitating regulatory network inference (Buenrostro et al., 2013). Using scRNA-seq and bulk ATAC-seq data, we have developed a method to infer regulatory networks controlling retinal regeneration (Hoang et al., 2020). Recently, several methods have been developed to integrate scATAC-seq and scRNA-seq data for regulatory network inference, e.g., SOMatic and DIRECT-NET (Jansen et al., 2019; Zhang et al., 2022). However, few methods comparatively assess the performance of network analysis by integrating scRNA-seq and scATAC-seq data relative to using scRNA-seq data alone.

Besides network inference, dissecting regulatory networks to detect modules and to identify key regulators is another major challenge in network biology. Clustering and decomposition are two frequently used methods for module detection (Saelens et al., 2018). Weighted correlation network analysis (WGCNA) performed hierarchical clustering of expression profiles to detect gene modules after correlation-based network inference (Langfelder and Horvath, 2008). Using single-cell transcriptomes, the SCENIC software combined GENIE3 and Rcistarget separately for network inference and identification of key transcription factors (Aibar et al., 2017). Although current methods can infer regulatory networks to identify gene

[1] CAS Key Laboratory of Regenerative Biology, Guangdong Provincial Key Laboratory of Biocomputing, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China

[2] Department of Ophthalmology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[3] Solomon H. Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[4] State Key Laboratory of Respiratory Disease, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China

[5] China-New Zealand Joint Laboratory on Biomedicine and Health, Guangzhou 510530, China

[6] These authors contributed equally

[7] Present address: Westlake University, Hangzhou 310030, China

[8] Lead contact

*Correspondence: wang_jie01@gibh.ac.cn

https://doi.org/10.1016/j.isci.2022.105359

modules and key individual regulatory genes, they do not construct a simple and statistically robust regulatory network among modules to provide biological insights.

Here, we developed IReNA to perform regulatory network analysis by integrating scRNA-seq and scATAC-seq data. As only scRNA-seq data are available for many biological samples, IReNA can also provide network analysis using only scRNA-seq data. Using IReNA, we analyzed published single-cell RNA-seq and ATAC-seq profiles, reconstructed regulatory networks, identified key regulators, and revealed simplified regulatory networks among modules. Integrated analysis of scRNA-seq and scATAC-seq data showed obviously improved performance on regulatory network inference. In comparison with Rcistarget from SCENIC software, one of the most frequently used existing methods, IReNA had a better overall performance at identifying known regulators.

RESULTS

The framework of integrated regulatory network analysis

To perform regulatory network analysis using single-cell RNA-seq data or integrating with ATAC-seq data, we developed IReNA which consists of two components: network inference and network decoding (Figures 1 and S1). Two pipelines of network inference were developed separately for analyzing scRNA-seq data alone or integrating scRNA-seq data with bulk or single-cell ATAC-seq data. We compared top-performing methods and chose the tree-based ensemble method GENIE3 as the default method for network inference using scRNA-seq data in IReNA (Figure S2A). After potential regulatory relationships were inferred by GENIE3, transcription factor binding motifs were used to refine regulatory relationships. If bulk or single-cell ATAC-seq data are available, both transcription factor binding motifs and footprints are used to refine regulatory relationships. For scATAC-seq data, peaks were linked to genes and used to identify transcription factor binding motifs and footprints. Network decoding in IReNA included network modularization, identification of enriched transcription factors, and a unique function for the construction of simplified regulatory networks among modules. Network modularization was based on K-means clustering of gene expression. Two statistical tests were applied to the modularized regulatory networks to separately identify enriched transcription factors and significant regulatory relationships among modules.

To illustrate the features and application of IReNA, we used public scRNA-seq and scATAC-seq data of hepatocytes from a mouse model of hepatectomy in the study of liver regeneration (Seidman et al., 2020). Although scRNA-seq and scATAC-seq data were obtained, the original study didn't perform regulatory network analysis.

Network analysis only using single-cell RNA sequencing data

Here, IReNA was applied to infer regulatory networks using scRNA-seq data alone. We firstly analyzed scRNA-seq data to identify genes used for network inference. 2,815 hepatocytes from the control and 48 h liver tissues after hepatectomy were used to construct the trajectory based on their single-cell expression profiles (Figure 2A). In the trajectory, we observed three branches, including the rest, activation, and proliferation. Based on the trajectory, we calculated the pseudotime and identified 4,014 differentially expressed genes (DEGs) including 165 transcription factors changed during the pseudotime. Meanwhile, we identified 45 transcription factors that are expressed in more than 5% hepatocytes but not statistically differential during the pseudotime. Next, all 4,059 DEGs and expressed transcription factors were divided into five modules through K-means clustering of the smoothed expression profiles (Figure 2B). Genes in each module showed specific expression profiles. For instance, genes in the first and fifth modules were specifically expressed in the rest and proliferating hepatocytes, respectively. Function enrichment analysis revealed relevant biological functions enriched in each module of genes, including fatty acid metabolism, organic acid catabolism, cytoplasmic translation, autophagy, and cell cycle (Figure 2C).

We then applied GENIE3 to infer regulatory relationships of all DEGs and expressed transcription factors in hepatocytes. We identified 434,218 potential regulatory relationships, each of which has >0.0001 wt and contains at least one transcription factor. For each regulatory relationship, Pearson's correlation was calculated to determine the positive or negative regulation. To refine 434,218 potential regulatory relationships, we further analyzed transcription factor binding motifs in the promoter regions of genes. Totally, 180,322 regulatory relationships with binding motifs were used to reconstruct regulatory networks (Figure 2D).

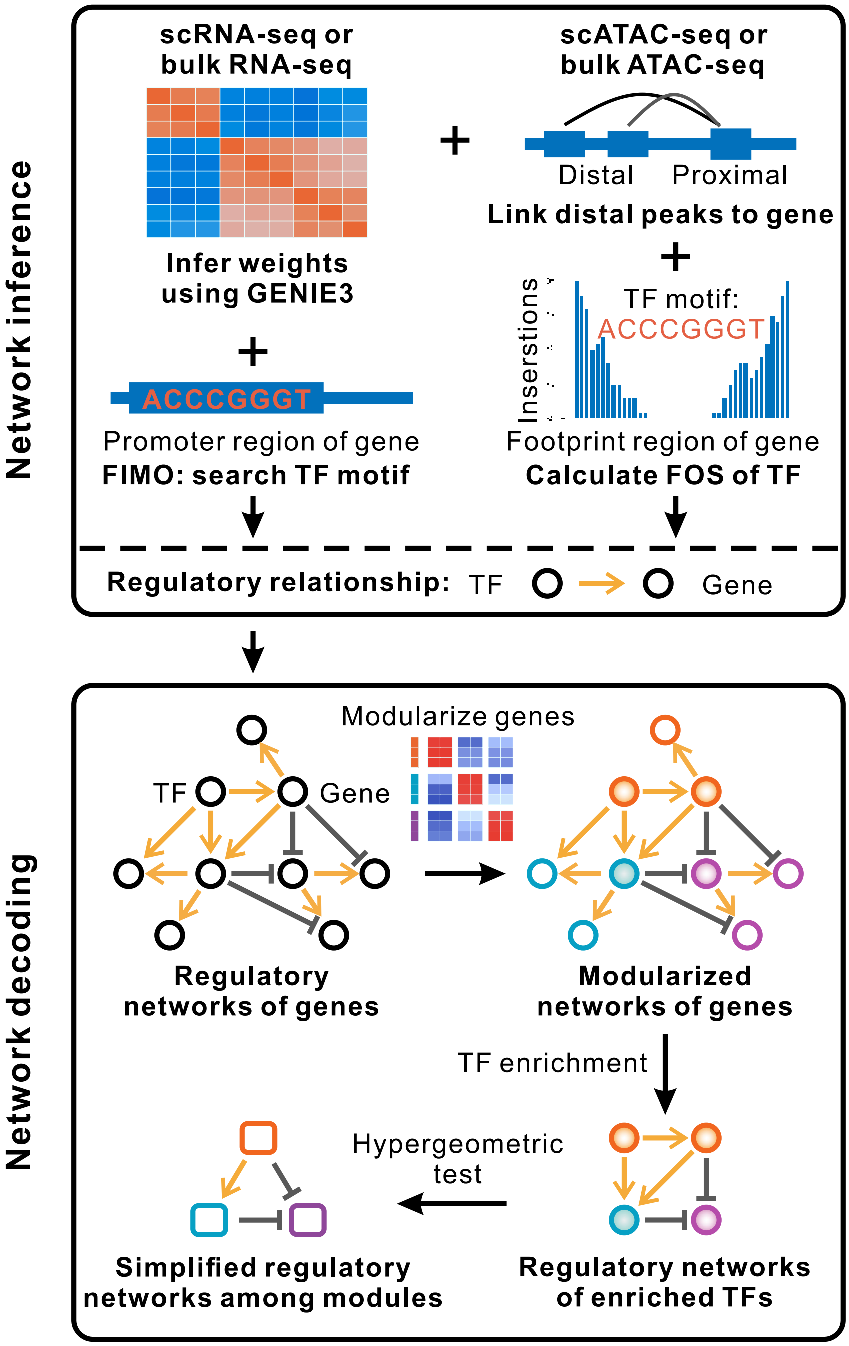


Figure 1. Flowchart of IReNA

IReNA consists of two components: network inference and network decoding. Network inference is performed according to weights calculated by GENIE3. FIMO is then used to identify binding motifs of transcription factors and refine regulatory relationships if only gene expression profiles (single-cell or bulk RNA-seq data) are used. If bulk or single-cell ATAC-seq data are available, binding motifs and footprint occupancy score (FOS) are used to refine regulatory relationships inferred from gene expression analysis. Especially for single-cell ATAC-seq data, ArchR is used to link peaks and genes. Next, inferred networks are modularized and used to enrich transcription factors (TFs). Regulatory networks of enriched TFs are further decoded by establishing simplified regulatory networks among modules.

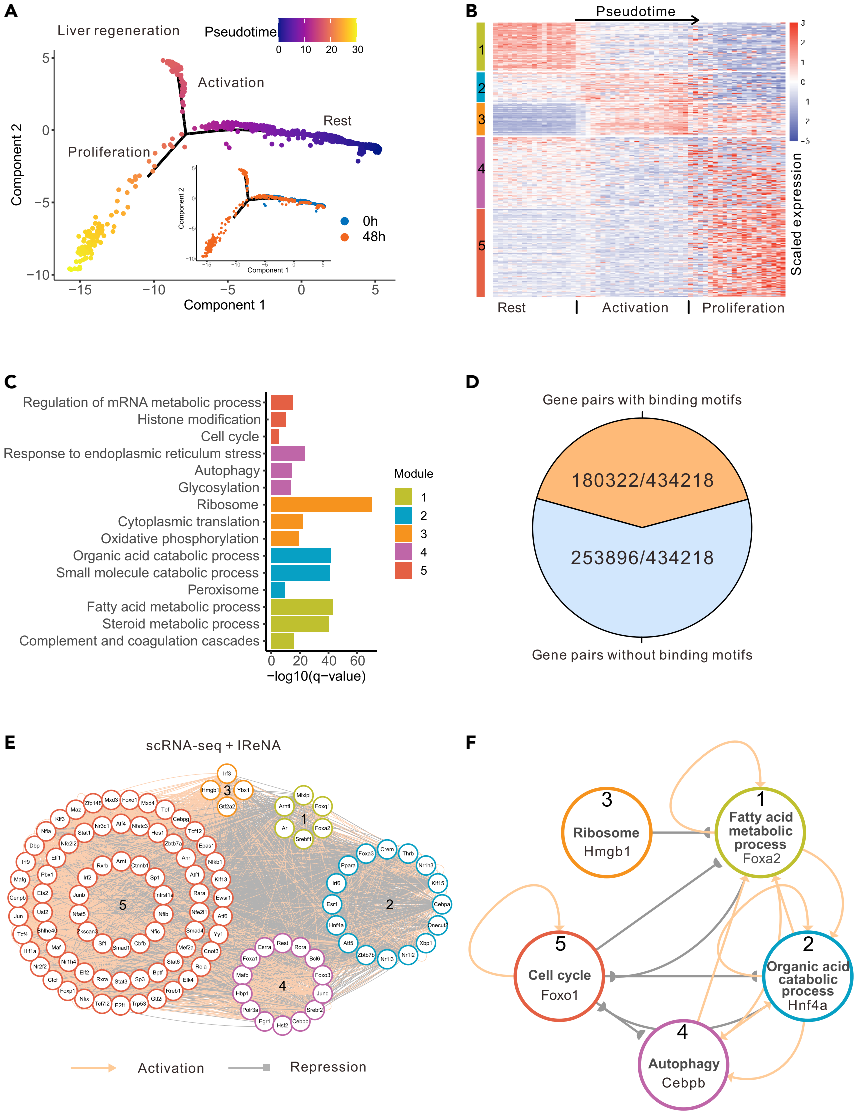


Figure 2. Regulatory network analysis of hepatocytes from liver regeneration

(A) Trajectory of 2,815 hepatocytes from liver tissues at 0 and 48 h after partial hepatectomy. Two trajectories are separately colored by the pseudotime and samples.

(B) Heatmap of scRNA-seq expression profiles of 4,059 differentially expressed genes (DEGs) and expressed transcription factors. K-means clustering was used to divide genes into five modules. Rows represent genes, and columns indicate intervals of the pseudotime separated by three branches.

(C) Enriched functions of five modules of genes.

(D) Fraction of regulatory gene pairs inferred by GENIE3 which contain transcription factor binding motifs in the promoter regions of target genes.

(E) Regulatory networks of 41 enriched transcription factors obtained from analyzing scRNA-seq data alone in IReNA (scRNA-seq + IReNA). Color of each circle indicates the module. Gray edge represents negative regulation, and the yellow edge represents positive regulation.

(F) Simplified regulatory networks among five modules obtained from analyzing scRNA-seq data alone. Representative biological functions and transcription factors were selected from (C and E) to label each module. The edges represent statistically significant regulations among modules.

Meanwhile, regulatory networks were modularized according to five modules of genes identified through K-means clustering in Figure 2B. Statistically analyzing modular regulatory networks, we identified 115 transcription factors that significantly regulated each module of genes. We reconstructed regulatory networks of these 115 enriched transcription factors, which were also divided into five modules (Figure 2E).

To obtain a simple regulatory network among modules for providing biological regulation insights, we performed the hypergeometric test to analyze modular regulatory networks of enriched transcription factors. We separated positive regulations and negative regulations to identify significant activations and repressions among modules. Fifteen significant regulatory relationships among modules were identified and used to establish simplified regulatory networks among modules (FDR < 0.005, Figure 2F). We showed representative biological function and transcription factor in each module. Simplified regulatory networks indicate that transcription factors related to fatty acid metabolism, organic acid catabolism, and autophagy significantly activated each other. In return, these factors significantly repressed transcription factors controlling cell cycle regulation. These suggest that the inhibition of transcription factors related to fatty acid metabolism, organic acid catabolism, and autophagy may activate cell cycle progression.

These predictions obtained from simplified regulatory network analysis were supported by previous studies. Hepatic nuclear factor 4 alpha (*Hnf4a*) as a well-known marker gene of hepatocytes, coordinated organic acid metabolism and activated the transcription of autophagy-related gene *Ulk1* in the liver (Martinez-Jimenez et al., 2010; Lee et al., 2021). According to gene ontology, *Hnf4a* negatively regulates the mitotic cell cycle. In simplified regulatory networks, we also observed that transcription factors regulating the ribosome may repress transcription factors of fatty acid metabolism. This is consistent with that hepatic rRNA transcriptional repression is essential for energy storage and lipid metabolism (Oie et al., 2014).

Network analysis by integrating single-cell RNA sequencing data and assay for transposase-accessible chromatin using sequencing data

Next, IReNA was used to infer regulatory networks by integrating scRNA-seq and scATAC-seq data from the study of liver regeneration. We analyzed scATAC-seq data of 7,004 hepatocytes after hepatectomy using ArchR (Granja et al., 2021). We identified 94,595 significant peak-to-gene links, each of which had a high peak-to-gene correlation, e.g., *Rora* and *Mlx* (Figure 3A). We further uncovered the binding sites of transcription factors to peaks and identified 386,597 regulatory relationships of transcription factors to genes. Among 386,597 regulatory relationships, 154,601 regulatory relationships had high footprint occupancy scores. Overlapping 154,601 regulatory relationships with 434,218 potential regulatory relationships inferred by GENIE3, we refined 47,721 regulatory relationships to reconstruct regulatory networks which consisted of 3,185 genes.

Analyzing modular regulatory networks of 3,185 genes, we identified 47 transcription factors that significantly regulate genes in each module. We performed functional enrichment analysis on target genes of 47 enriched transcription factors and observed a significant enrichment of cell type-specific functions, such as hepatocyte proliferation (q value = 6.27×10^{-3}). We then reconstructed modular regulatory networks of 47 enriched transcription factors (Figure 3B). Six significant regulation relationships among modules were identified to reconstruct simplified regulatory networks among modules (FDR < 0.05, Figure 3C). We found that intermodular regulatory networks from the integrated analysis of scRNA-seq and scATAC-seq data were consistent with intermodular regulatory networks obtained from analyzing scRNA-seq data alone. We also observed that the module of transcription factors related to organic acid catabolism of hepatocytes, e.g., *Hnf4a*, significantly repressed the cell cycle in liver regeneration. This suggests that the inhibition of hepatocyte metabolism may promote liver regeneration.

Performance of integrated regulatory network analysis

To directly compare IReNA with other methods of network analysis, we analyzed scRNA-seq data to reconstruct regulatory networks and identified key transcription factors using GENIE3 and Rcistarget package from the SCENIC software (Aibar et al., 2017). Using Rcistarget, we refined 39,192 regulatory relationships from 434,218 potential regulatory relationships inferred by GENIE3. Next, we identified 108 transcription factors whose binding motifs were overrepresented in the promoter regions of 4,059 DEGs and expressed transcription factors identified by scRNA-seq data analysis. We obtained 1,761 significant regulatory relationships for 108 enriched transcription factors and then reconstructed modular regulatory networks of enriched transcription factors (Figure 3D).

Using chromatin immunoprecipitation followed by sequencing (ChIP-seq) data and genetic perturbation data of transcription factors from liver samples obtained from public databases, we assessed the performance of IReNA and Rcistarget on regulatory network inference. Regulatory relationships identified by GENIE3 analysis of scRNA-seq data were refined separately by integrating scATAC-seq data in IReNA

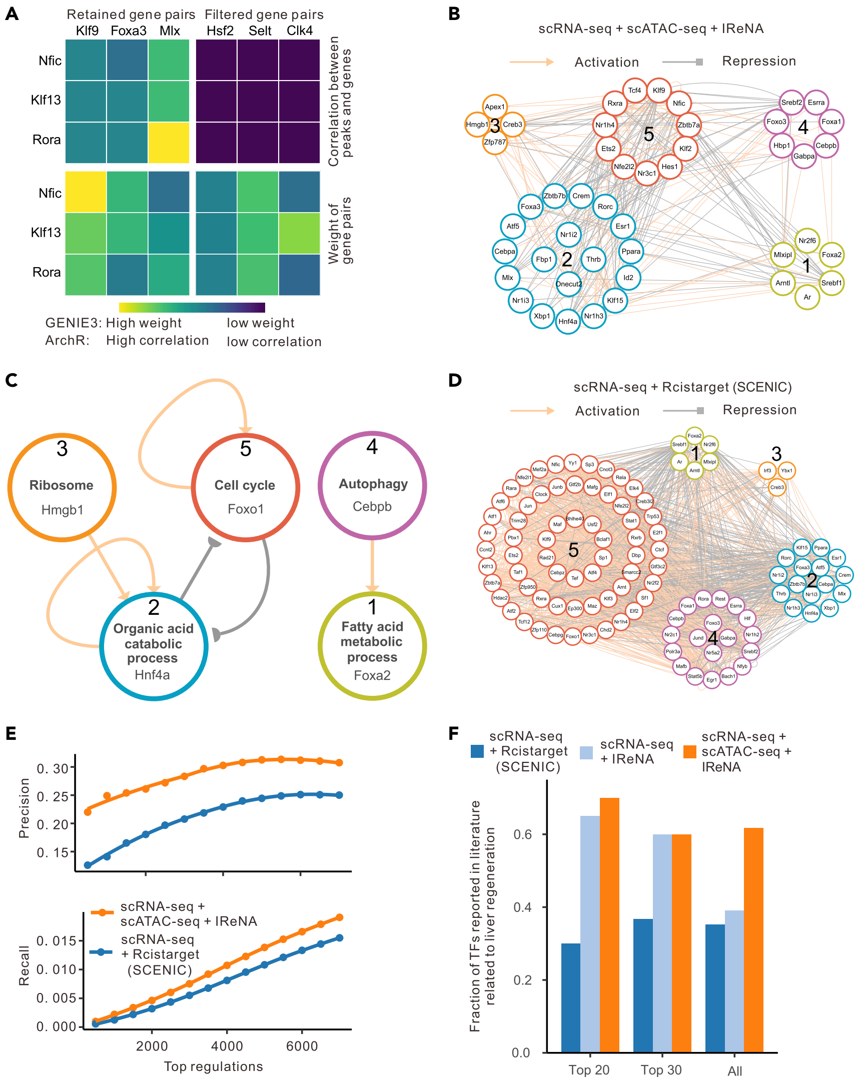


Figure 3. Comparison of regulatory networks related to liver regeneration

(A) Heatmap of peak-to-gene correlations (top panel) calculated by ArchR and weights of gene pairs (bottom panel) calculated by GENIE3. In the top panel, each gene in the column represents a transcription factor that has the binding motif in the peak. Gene pairs with both high correlations and high weights were used to infer regulatory networks of integrating scRNA-seq and scATAC-seq data.

(B) Regulatory networks of 47 enriched transcription factors obtained through the integrated analysis of scRNA-seq and scATAC-seq data using IReNA (ATAC-seq + scRNA-seq + IReNA). Color of the circle indicates the module. Gray edge represents negative regulation, and the yellow edge represents positive regulation.

(C) Simplified regulatory networks among modules obtained through the integrated analysis of scRNA-seq and scATAC-seq data.

(D) Regulatory networks for 108 enriched transcription factors from scRNA-seq data analysis using Rcistarget from SCENIC software named scRNA-seq + Rcistarget (SCENIC).

(E) Precision and recall of regulatory networks predicted by the analysis of ATAC-seq + scRNA-seq + IReNA and scRNA-seq + Rcistarget (SCENIC). ChIP-seq data from liver samples were used as the ground truth.

(F) Fraction of enriched transcription factors reported in the literature related to liver regeneration. Top 20, top 30 and all enriched transcription factors were compared for all three types of regulatory networks, including networks from scRNA-seq data analysis in IReNA (scRNA-seq + IReNA). Enriched transcription factors were ranked according to FDR or normalized enrichment score.

and overlapping DNA motifs by Rcistarget. Using ChIP-seq data alone, we observed that integrative network analysis of scRNA-seq and scATAC-seq data using IReNA showed higher precision and recall of regulatory relationships than Rcistarget analysis (Figure 3E). Similar results were observed when both ChIP-seq data and genetic perturbation data were used (Figure S2B). These results indicated that the integrated analysis of scRNA-seq and scATAC-seq data in IReNA had overall better performance on regulatory network inference than did Rcistarget analysis.

We then compared regulatory networks of enriched transcription factors inferred through three different approaches described above. Among 47 transcription factors identified by IReNA using the integrated analysis of scRNA-seq and scATAC-seq data, 36 (76.60%) transcription factors were also present in regulatory networks inferred using IReNA analysis of scRNA-seq data alone (Figure S2C). By comparing gene regulatory networks obtained by analyzing only scRNA-seq data separately using IReNA and Rcistarget, we observed an overlap of 66.96% (77 in 115) of all enriched transcription factors. These results indicated that a large fraction of transcription factors was identified by two methods of network analysis.

To assess the significance of transcription factors identified using IReNA or Rcistarget, we manually examined whether these factors had been previously reported in the literature related to liver regeneration. To compare different methods, we ranked the enriched transcription factors according to the significance in statistics (FDR for IReNA or normalized enrichment score for Rcistarget). We found that regulatory networks from the integrated analysis of scRNA-seq and ATAC-seq data had the best performance, 70.0% of transcription factors in the top 20, 60.0% in the top 30 and 61.7% of all enriched transcription factors were previously reported in liver regeneration-related literature (Figure 3F and Table S1). For regulatory networks inferred from analyzing scRNA-seq data alone, 65.0% of transcription factors in the top 20, 60.0% in the top 30 and 39.1% of all transcription factors were reported in liver regeneration-related literature. For regulatory networks inferred by Rcistarget from SCENIC software, 30.0% of transcription factors were in the top 20, 36.7% in the top 30 and 35.3% of all transcription factors were reported in liver regeneration-related literature. Among three types of regulatory networks, the highest fraction of transcription factors was reported for regulatory networks from the integrated analysis of scRNA-seq and scATAC-seq data. We also used the software CoCiter to assess the co-citation significance of transcription factors with liver regeneration-related terms in the literature. Similarly, we observed that scRNA-seq and scATAC-seq data analysis using IReNA identified the highest fraction of transcription factors (Figure S2D). These results indicate that the integrated network analysis of scRNA-seq and scATAC-seq data using IReNA improved the precision of identifying known transcription factors. Moreover, IReNA shows a better performance at identifying known regulators than the Rcistarget method from SCENIC software.

To further demonstrate the performance of IReNA, we conducted regulatory network analysis on another two datasets from heart regeneration and NASH (Cui et al., 2020; Seidman et al., 2020). Prior to inferring gene regulatory networks controlling heart regeneration, we reconstructed the trajectory of 4,884 cardiomyocytes from neonatal heart tissues (Figure S3A). We found that cardiomyocytes formed two distinct branches (named the activation branch and the proliferation branch) following myocardial infarction. We further identified and divided 4,340 DEGs and expressed transcription factors into 4 modules (Figure S3B). Genes expressed in the activation branch (module 2 and 3) were related to oxidative phosphorylation and muscle cell differentiation, whereas genes specifically expressed in the proliferation branch (module 4) are enriched for the cell cycle (Figure S3C). Then, regulatory networks were inferred using the same three methods described above (Figures 4A, 4B, S3D, and S3E). In comparison with regulatory networks inferred by Rcistarget, regulatory networks reconstructed by IReNA contained a higher fraction of transcription factors previously reported in the literature related to heart regeneration (Figure 4C and Table S1). The precision of identifying known regulators among the top 20 transcription factors in heart regeneration were 45.0%, 30.0 and 20.0% respectively for scRNA-seq + scATAC-seq + IReNA, scRNA-seq + IReNA and scRNA-seq + Rcistarget. We also observed that scRNA-seq + scATAC-seq + IReNA identified the highest fraction of known regulators for both the top 30 transcription factors and all enriched transcription factors.

For the study of NASH, we used 2,748 Kupffer cells to construct the trajectory (Figure S4A). 2,742 DEGs and expressed transcription factors were identified and divided into three modules, which were separately enriched for myeloid cell differentiation, ribosome, and oxidative phosphorylation (Figures S4B and S4C). In NASH, only bulk ATAC-seq data were available and used to refine regulatory relationships inferred from scRNA-seq data analysis (Figure S4D). We reconstructed three types of regulatory networks and a

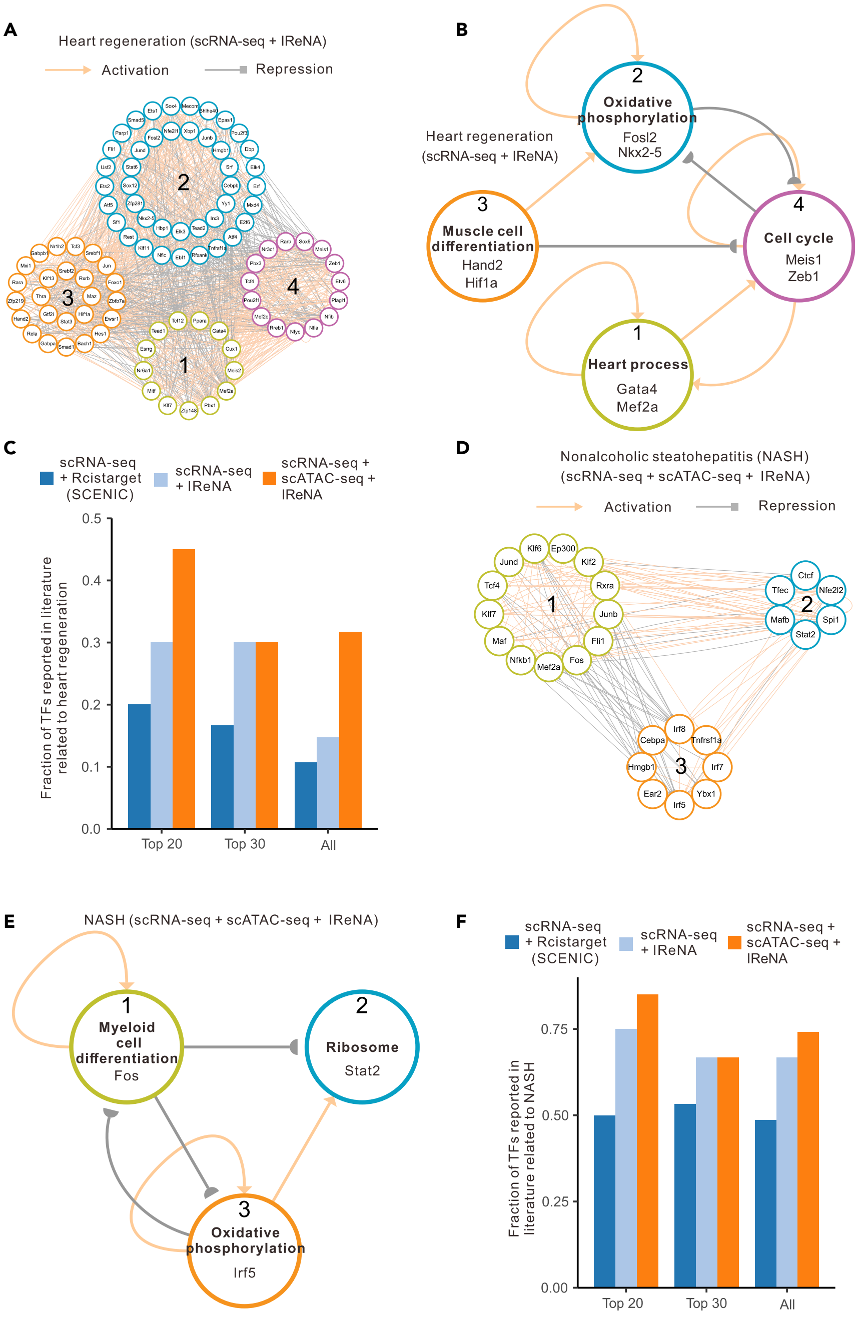


Figure 4. Regulatory network analysis for two studies of heart regeneration and NASH

(A) Regulatory networks of 95 enriched transcription factors obtained by analyzing scRNA-seq data alone from heart regeneration. Modules are represented by the color of the circles. Gray edge represents negative regulation, and the yellow edge represents positive regulation.

(B) Simplified regulatory networks among modules obtained by analyzing scRNA-seq data alone from heart regeneration.

(C) Fraction of enriched transcription factors reported in the literature related to heart regeneration.

(D) Regulatory networks of 27 enriched transcription factors obtained by integrating scRNA-seq and bulk ATAC-seq data from nonalcoholic steatohepatitis (NASH).

(E) Simplified regulatory networks among modules obtained by integrating scRNA-seq and ATAC-seq data from NASH.

(F) Fraction of transcription factors reported in the literature related to NASH.

simplified regulatory network among modules (Figures 4D, 4E, S4E, and S4F). Regulatory network comparison of NASH study has a similar trend with studies in liver regeneration and heart regeneration. IReNA analysis using scRNA-seq and bulk ATAC-seq data identified the most transcription factors (85.0% of transcription factors in the top 20, 66.7% in the top 30 and 74.1% of all enriched transcription factors) which were reported to associate with NASH, followed by IReNA analysis using scRNA-seq data alone (75.0% in the top 20, 66.7% in the top 30 and 66.7% of all enriched transcription factors), and finally Rcistarget analysis using scRNA-seq data alone (50.0% in the top 20, 53.3% in the top 30 and 48.7% of all enriched transcription factors) (Figure 4F and Table S1). The integrated analysis of single-cell or bulk ATAC-seq data with scRNA-seq data overall substantially improved the reconstruction of gene regulatory networks, and had higher precision of identifying known transcription factors. In addition, transcription factors enriched through analyzing scRNA-seq data alone using IReNA showed improved accuracy relative to those identified using Rcistarget from SCENIC software.

DISCUSSION

In the study, we developed IReNA to perform regulatory network analysis, including network inference and network decoding. In IReNA, gene regulatory networks are inferred by analyzing either scRNA-seq data alone or by integrating scRNA-seq and scATAC-seq data. The regulatory relationships between transcription factors and target genes are firstly inferred according to the weights calculated by GENIE3 using scRNA-seq data. Then, transcription factor binding motifs are identified to refine regulatory relationships if only scRNA-seq data are available. When both scRNA-seq and ATAC-seq data are used, transcription factor binding motifs and footprints are applied to further refine transcriptional regulatory relationships used for network inference. IReNA also provides functions to decode inferred regulatory networks, including the modularization of regulatory networks, the enrichment of transcription factors, and the construction of simplified regulatory networks among modules.

In IReNA, we developed several specific functions for network analysis. First, unlike the methods using scRNA-seq data for network analysis, IReNA integrates scRNA-seq data with single-cell or bulk ATAC-seq data to reconstruct regulatory networks. Analysis of three independent datasets consistently indicates that the integrated analysis of scRNA-seq and ATAC-seq data could more precisely identify known regulators than network analysis using scRNA-seq data alone. Second, IReNA could provide cell state-specific regulatory networks through modularizing regulatory networks. Regulatory networks are modularized according to the clustering results of gene expression profiles which represent different cell states and specific biological functions. Third, IReNA statistically analyzes modular regulatory networks and identifies reliable transcription factors including known regulators. Applied to multiple datasets, the method used in IReNA showed a consistently better performance on the identification of known regulators than Rcistarget, which conducts transcription factor enrichment analysis based on the rank of all genes for each motif (Imrichová et al., 2015). Fourth, we created a unique function in IReNA to construct the simplified regulatory networks among modules that reveal key regulatory modules and factors, facilitating the interpretation of dynamic biological regulations. Consistent results from IReNA analysis were obtained in three independent datasets, suggesting IReNA could provide robust network analysis by combining modularization analysis and statistical tests.

In the light of the sparsity of scRNA-seq data, we smoothed gene expressions to calculate correlations used for the signs of regulatory relationships in IReNA. IReNA could directly calculate correlations using original expression data independent of the pseudotime. However, the smoothed gene expressions improved regulatory network inference relative to original gene expressions (Figure S2A). Gene expressions were smoothed according to the pseudotime which was calculated by Monocle in IReNA. We compared other methods for calculating the pseudotime and found that Monocle-based pseudotime showed a better performance than Slingshot-based pseudotime for network inference (Figure S2A). We used IReNA to infer regulatory networks for the same cell type, or for several cell types which are related by the lineage during the development or disease progression. Correspondingly, cell type-specific or lineage-specific transcriptomes and epigenomes should be extracted prior to network inference. Transcriptomes could be measured by bulk RNA-seq or scRNA-seq, whereas epigenomes may be detected using either bulk ATAC-seq or scATAC-seq. Transcriptomes and epigenomes used in IReNA should be matched for the same cell type or the same condition. Currently, IReNA has used scRNA-seq and scATAC-seq data of unpaired cells from the same condition to perform network analysis. However, parallel scRNA-seq and scATAC-seq profiles from the same cells are emerging, and updates to IReNA will accommodate these new data.

In previous studies, we demonstrated that IReNA can be used to integrate scRNA-seq and bulk ATAC-seq to reconstruct modular gene regulatory networks controlling retinal regeneration and retinal development (Hoang et al., 2020; Lyu et al., 2021). We also identified key transcription factors and constructed the simplified regulatory networks regulating different cell states in retinal Müller glia in zebrafish and mice. In zebrafish, we predicted that reactivity-related transcription factors such as *hmga1* and *yap1* promoted the proliferation of Müller glia. In mice, network analysis indicated that nuclear factors *Nfia/b/x*, which were found in regulatory modules that maintain or restore the resting glial state, repressed the reactivity of Müller glia. Using the genetic loss of function analysis, we confirmed that several transcription factors identified by IReNA were critical for retinal regeneration, including *hmga1a* and *yap1* in zebrafish and *Nfia/b/x* in mice (Hoang et al., 2020). These results indicate that IReNA analysis can provide valuable regulatory network insights and reveal key regulators in distinct biological processes including retinal regeneration.

Using public scRNA-seq and ATAC-seq data from three studies, we further performed regulatory network analysis through IReNA, which provides meaningful biological insights. According to simplified intermodular regulatory networks in liver regeneration, organic acid catabolism-related modules and transcription factors are needed to be repressed to activate cell cycle progression in hepatocytes, e.g., hepatocyte nuclear factor 4 alpha (*Hnf4a*) (Figures 2F and 3C). The study in mouse liver regeneration demonstrated that the deletion of *Hnf4a* leads to sustained proliferation (Huck et al., 2019). In heart regeneration, simplified regulatory networks among modules imply that regulatory modules and transcription factors controlling the heart process promote the cell cycle of cardiomyocytes (Figure 4B). It has been reported that one of such factors *Gata4* could activate heart regeneration in zebrafish and mice (Kikuchi et al., 2010; Malek Mohammadi et al., 2017). In regulatory networks of NASH, we identified *IRF5* which was reported to promote hepatic fibrosis in Kupffer cells in nonalcoholic fatty liver disease (Alzaid et al., 2016). These results indicate that IReNA could be applied to identify key modules and transcription factors that regulate a range of different biological processes, although further functional validation of putative key regulators is required.

In summary, we have developed IReNA to perform regulatory network analysis, including network inference, network modularization, transcription factor enrichment, and the construction of simplified regulatory networks among modules. IReNA showed a consistently better performance at identifying known regulators when integrating scRNA-seq data with scATAC-seq data to reconstruct gene regulatory networks. IReNA also outperformed the existing method Rcistarget from SCENIC software in identifying known regulators. Key transcription factors and regulatory relationships identified by IReNA are potential targets controlling tissue regeneration, diseases, and other dynamic biological processes. Through the construction of modular regulatory networks and simplified regulatory networks among modules, IReNA facilitates the understanding of regulatory mechanisms and provides meaningful biological insights.

Limitations of the study

IReNA depends on the network inference method GENIE3 and K-means clustering method for gene modularization which require application-specific analysis of single-cell sequencing data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Description of scRNA-seq, bulk ATAC-seq and scATAC-seq data
  - ○ Regulatory network inference
  - ○ Identify transcription factor binding motifs in regulatory regions of targeted genes
  - ○ Analyze bulk or single-cell ATAC-seq data to refine regulatory relationships
  - ○ Network modularization based on gene co-expression

- ○ Transcription factor enrichment for modular regulatory networks
- ○ Construct simplified regulatory networks among modules
- ○ Comparison of regulatory networks

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105359.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [No. T2222003, 32170849], and the Guangdong Province Science and Technology Program [No. 2016B030301007, 2020B1212060052].

AUTHOR CONTRIBUTIONS

J.W. and J.Q. conceived the project. J.W., J.Q., and S.B. supervised the research. J.J., P.L., J.L., and J.W. developed IReNA and performed sequencing data analysis to construct regulatory networks. J.J., J.L., S.H., and J.T. checked transcription factors reported in the literature.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 29, 2022
Revised: September 19, 2022
Accepted: October 12, 2022
Published: November 18, 2022

REFERENCES

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.

Alzaid, F., Lagadec, F., Albuquerque, M., Ballaire, R., Orliaguet, L., Hainault, I., Blugeon, C., Lemoine, S., Lehuen, A., Saliba, D.G., et al. (2016). IRF5 governs liver macrophage activation that promotes hepatic fibrosis in mice and humans. JCI Insight *1*, e88689.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. Nucleic Acids Res. *43*, W39–W49.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Chan, T.E., Stumpf, M.P.H., and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. *5*, 251–267.e3.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Chen, T., Oh, S., Gregory, S., Shen, X., and Diehl, A.M. (2020). Single-cell omics analysis reveals functional diversification of hepatocytes during liver regeneration. JCI Insight *5*, 141024. https://doi.org/10.1172/jci.insight.141024.

Cui, M., Wang, Z., Chen, K., Shah, A.M., Tan, W., Duan, L., Sanchez-Ortiz, E., Li, H., Xu, L., Liu, N., et al. (2020). Dynamic transcriptional responses to injury of regenerative and non-regenerative cardiomyocytes revealed by single-nucleus RNA sequencing. Dev. Cell *55*, 665–667.

Feng, C., Song, C., Liu, Y., Qian, F., Gao, Y., Ning, Z., Wang, Q., Jiang, Y., Li, Y., Li, M., et al. (2020). KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. Nucleic Acids Res. *48*, D93–D100.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411.

Hoang, T., Wang, J., Boyd, P., Wang, F., Santiago, C., Jiang, L., Yoo, S., Lahne, M., Todd, L.J., Jia, M., et al. (2020). Gene regulatory networks controlling vertebrate retinal regeneration. Science *370*, eabb8598. https://doi.org/10.1126/science.abb8598.

Huck, I., Gunewardena, S., Espanol-Suner, R., Willenbring, H., and Apte, U. (2019). Hepatocyte nuclear factor 4 alpha activation is essential for termination of liver regeneration in mice. Hepatology *70*, 666–681.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, e12776. https://doi.org/10.1371/journal.pone.0012776.

Imrichová, H., Hulselmans, G., Atak, Z.K., Potier, D., and Aerts, S. (2015). i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. Nucleic Acids Res. *43*, W57–W64.

Jansen, C., Ramirez, R.N., El-Ali, N.C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., Conesa, A., and Mortazavi, A. (2019). Building gene regulatory networks from scATAC-seq and scRNA-seq using linked Self Organizing Maps. PLoS Comput. Biol. *15*, e1006555.

Kikuchi, K., Holdway, J.E., Werdich, A.A., Anderson, R.M., Fang, Y., Egnaczyk, G.F., Evans, T., Macrae, C.A., Stainier, D.Y.R., and Poss, K.D. (2010). Primary contribution to zebrafish heart regeneration by gata4(+) cardiomyocytes. Nature *464*, 601–605.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. *9*, 559.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, D.-H., Park, S.-H., Ahn, J., Hong, S.P., Lee, E., Jang, Y.-J., Ha, T.-Y., Huh, Y.H., Ha, S.-Y., Jeon, T.-I., and Jung, C.H. (2021). Mir214-3p and Hnf4a/Hnf4α reciprocally regulate Ulk1 expression and autophagy in nonalcoholic hepatic steatosis. Autophagy *17*, 2415–2431.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing

Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019). Identification of transcription factor binding sites using ATAC-seq. Genome Biol. *20*, 45.

Lyu, P., Hoang, T., Santiago, C.P., Thomas, E.D., Timms, A.E., Appel, H., Gimmen, M., Le, N., Jiang, L., Kim, D.W., et al. (2021). Gene regulatory networks controlling temporal patterning, neurogenesis, and cell-fate specification in mammalian retina. Cell Rep. *37*, 109994.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Malek Mohammadi, M., Kattih, B., and Grund, A. (2017). The transcription factor GATA 4 promotes myocardial regeneration in neonatal mice. Mol. Med. *9*, 265–279.

Martinez-Jimenez, C.P., Kyrmizi, I., Cardot, P., Gonzalez, F.J., and Talianidis, I. (2010). Hepatocyte nuclear factor 4alpha coordinates a transcription factor network regulating hepatic fatty acid metabolism. Mol. Cell Biol. *30*, 565–577.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T., and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics *33*, 2314–2321.

Oie, S., Matsuzaki, K., Yokoyama, W., Tokunaga, S., Waku, T., Han, S.-I., Iwasaki, N., Mikogai, A., Yasuzawa-Tanaka, K., Kishimoto, H., et al. (2014). Hepatic rRNA transcription regulates high-fat-diet-induced obesity. Cell Rep. *7*, 807–820.

Ou, J., Liu, H., Yu, J., Kelliher, M.A., Castilla, L.H., Lawson, N.D., and Zhu, L.J. (2018). ATACseqQC: a bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genom. *19*, 169.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.

Qiao, N., Huang, Y., Naveed, H., Green, C.D., and Han, J.-D.J. (2013). CoCiter: an efficient tool to infer gene function by assessing the significance of literature co-citation. PLoS One *8*, e74074.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nat. Commun. *9*, 1090.

Seidman, J.S., Troutman, T.D., Sakai, M., Gola, A., Spann, N.J., Bennett, H., Bruni, C.M., Ouyang, Z., Li, R.Z., Sun, X., et al. (2020). Niche-specific reprogramming of epigenetic landscapes drives myeloid cell diversity in nonalcoholic steatohepatitis. Immunity *52*, 1057–1074.e7.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genom. *19*, 477.

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu. Rev. Cell Dev. Biol. *31*, 399–428.

Wang, J., Zibetti, C., Shang, P., Sripathi, S.R., Zhang, P., Cano, M., Hoang, T., Xia, S., Ji, H., Merbs, S.L., et al. (2018). ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. Nat. Commun. *9*, 1364.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383.

Zhang, L., Zhang, J., and Nie, Q. (2022). DIRECT-NET: an efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. Sci. Adv. *8*, eabl7393.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biol. *9*, R137.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zou, Z., Ohta, T., Miura, F., and Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res. *50*, W175–W182. https://doi.org/10.1093/nar/gkac199.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
scRNA-seq data of liver regeneration	GEO	GSE158866
scATAC-seq data of liver regeneration	GEO	GSE158873
scRNA-seq data of heart regeneration	GEO	GSE130699
scATAC-seq data of heart regeneration	GEO	GSE142365
scRNA-seq data of NASH	GEO	GSE128334
ATAC-seq data of NASH	GEO	GSE128335
ChIP-seq data used for evaluating regulatory relationships	Zou et al. (2022)	https://chip-atlas.org/
Genetic perturbation data for evaluating regulatory relationships	Feng et al. (2020)	http://www.licpathway.net/KnockTFv2/index.php
Software and algorithms		
R	R Foundation for Statistical Computing	https://www.r-project.org/
IReNA package	This paper	https://github.com/jiang-junyao/IReNA
Monocle package	Qiu et al. (2017)	http://cole-trapnell-lab.github.io/monocle-release/
Slingshot package	Street et al. (2018)	https://github.com/kstreet13/slingshot
GENIE3 package	Huynh-Thu et al. (2010)	https://github.com/aertslab/GENIE3
PIDC Julia module	Chan et al. (2017)	https://github.com/Tchanders/InformationMeasures.jl
BEELINE framework	Pratapa et al. (2020)	https://github.com/Murali-group/Beeline
Fimo	Bailey et al. (2015)	https://meme-suite.org/meme/doc/fimo.html
Fastp	Chen et al. (2018)	https://github.com/OpenGene/fastp
Bowtie2	Langmead and Salzberg (2012)	https://github.com/BenLangmead/bowtie2
ATACseqQC package	Ou et al. (2018)	https://github.com/jianhong/ATACseqQC
Samtools	Li et al. (2009)	http://www.htslib.org/
Picard	Broad institute	http://broadinstitute.github.io/picard/
Macs2	Zhang et al. (2008)	https://github.com/macs3-project/MACS
HTseq	Anders et al. (2015)	https://htseq.readthedocs.io/en/master/
EdgeR package	Robinson et al. (2010)	https://bioconductor.org/packages/release/bioc/html/edgeR.html
CellRanger	10X Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
ArchR package	Granja et al. (2021)	https://www.archrproject.com/index.html
HINT	Li et al. (2019)	https://reg-gen.readthedocs.io/en/latest/hint/introduction.html
ChIPseeker	Yu et al. (2015)	https://github.com/YuLab-SMU/ChIPseeker
Rsamtools	Bioconductor	https://github.com/Bioconductor/Rsamtools
ClusterProfile	Yu et al. (2012)	https://github.com/YuLab-SMU/clusterProfiler
Cytoscape	Shannon et al. (2003)	https://cytoscape.org/
SCENIC package	Aibar et al. (2017)	https://scenic.aertslab.org/tutorials/
Rcistarget package	Bioconductor	https://github.com/aertslab/RcisTarget
Cociter	Qiao et al. (2013)	https://picb.ac.cn/hanlab/cociter21/gene-term/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Jie Wang (wang_jie01@gibh.ac.cn) upon request.

Materials availability

This study did not generate new unique reagents.

Data and code availability

- IReNA is an open software and available online at https://github.com/jiang-junyao/IReNA.
- Public data and corresponding accession numbers used in this paper are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Description of scRNA-seq, bulk ATAC-seq and scATAC-seq data

To demonstrate analysis flow of IReNA, we used public scRNA-seq and ATAC-seq data from three studies, which analyzed datasets obtained from models of liver regeneration, heart regeneration, and nonalcoholic steatohepatitis (NASH), respectively (Seidman et al., 2020; Cui et al., 2020; Chen et al., 2020).

For liver regeneration, partial hepatectomy (PHx) was performed in adult mice (Chen et al., 2020). The study conducted both scRNA-seq and scATAC-seq on liver tissues at 0 and 48 hours after PHx (accession number GSE158866 and GSE158873, available at gene expression omnibus database https://www.ncbi.nlm.nih.gov/geo/). According to the original annotation of cell types in the study, there are 2,815 and 7,004 hepatocytes measured separately by scRNA-seq and scATAC-seq.

In the study of heart regeneration, scRNA-seq profiles were measured on 4,884 cardiomyocytes at 1 day, 3 days after myocardial infarction, and 1 day after sham surgery in neonatal mice (accession number GSE130699) (Cui et al., 2020). Meanwhile, scATAC-seq was performed on 755 cardiomyocytes at 3 days after myocardial infarction on postnatal day one (accession number GSE142365).

The scRNA-seq and bulk ATAC-seq data from the study of NASH are available through the accession numbers GSE128334 and GSE128335 (Seidman et al., 2020). In this study, scRNA-seq profiles were measured on 6,184 non-parenchymal cells, including 2,748 Kupffer cells, from liver tissues of healthy and NASH mice. Bulk ATAC-seq was conducted on Kupffer cells from two healthy and two NASH samples.

Regulatory network inference

If no specific parameters were reported in IReNA, default parameters were used for existing software. Prior to network inference, we identified differentially expressed genes (DEGs) and the expressed transcription factors as the potential genes in regulatory networks. We compared Monocle (version 2.1.8) and Slingshot (version 2.5.2), which are two frequently used methods for calculating the pseudotime of single cells (Qiu et al., 2017; Street et al., 2018). In IReNA, Monocle was used as the default method to construct the trajectory and to infer the pseudotime of individual cells from scRNA-seq data. The smoothed expression profiles were calculated according to the pseudotime and branches on the trajectory. If there is a branch in the trajectory, pseudotime in each branch is divided into 20 equal intervals. Otherwise, pseudotime is divided into 50 equal intervals. Then, we calculated the average expression profile of single cells in each interval and obtained the smoothed expression profiles.

DEGs were further identified according to the pseudotime of individual cells. To obtain reliable DEGs in pseudotime analysis, we used rigorous statistical thresholds, including q-value < 0.005, fraction of expressed cells $>10\%$ and single-cell expression difference >0.1. Single-cell expression difference was defined as previously described (Hoang et al., 2020). The formula for single-cell expression difference was as follows: single-cell expression difference = Q95-expression - Q5-expression, where Q95-expression and Q5-expression represent 95% quantile and 5% quantile of expression values across all intervals of the pseudotime, respectively. Given that some key transcription factors may not be DEGs, we also included transcription factors which expressed in >5% cells for network analysis.

According to the previous evaluation of the methods for network inference, we compared two top performing methods GENIE3 (version 1.16) and PIDC (version 0.1.1) (Chan et al., 2017; Huynh-Thu et al., 2010). GENIE3 was selected as the default method for network inference in IReNA. GENIE3 infers regulatory relationships of transcription factors to target genes based on random forest regression (Huynh-Thu et al., 2010). We used GENIE3 to calculate the weight of the regulation for each gene pair based on scRNA-seq data. Several cutoffs for the weight of the regulation were assessed according to the number of regulations and degree distribution of genes after pruning. The default weight >0.0001 was used in liver regeneration, whereas the weight >0.0003 was used in heart regeneration and NASH. Meanwhile, only gene pairs which contain at least one transcription factor from the TRANSFAC database (version 2018.3) were chosen as potential regulatory relationships.

To determine activating and repressive regulatory relationships, we calculated Pearson's correlations of all gene pairs using the smoothed expression profiles. The regulation types of gene pairs were defined as activation and repression separately for positive correlations and negative correlations. Original expression profiles were also used to calculate gene correlations which were compared with the correlations from analyzing the smoothed expression profiles. To compare different methods for network inference, we used the relevant evaluation framework BEELINE (Pratapa et al., 2020).

Identify transcription factor binding motifs in regulatory regions of targeted genes

If bulk or single-cell ATAC-seq data is not available, gene regulatory relationships inferred from scRNA-seq data analysis were further refined according to transcription factor binding motifs present in the promoter regions of genes. Fimo (version 5.4.1, parse-genomic-coord, max-stored-scores = 2,000,000) was used to identify transcription factor binding motifs in the promoter regions (ranging from 1,000 bp upstream to 500 bp downstream of the transcription start sites) of the genes (Bailey et al., 2015). Uniform background model was used and contained in the motif file. Position weight matrices of binding motifs were from TRANSFAC database (version 2018.3). Regulatory relationships are selected for further network analysis if the binding motif of transcription factor occurs in the promoter region of the target gene.

Analyze bulk or single-cell ATAC-seq data to refine regulatory relationships

If bulk ATAC-seq data is available, we use the following six steps to preprocess raw data in fastq format. (I) Remove adaptors of pair-end raw reads using fastp software (version 0.21.0) (Chen et al., 2018). (II) Align reads the GRCm38/mm10 genome using bowtie2 (version 2.4.1) with default parameters (Langmead and Salzberg, 2012). To precisely locate the center on the Tn5 cut site, we shifted reads on the forward strand by +4 bp and reads on the reverse strand by −5 bp from the bam-files using the function shiftGAlignmentsList from R package ATACseqQC (Ou et al., 2018). (III) Filter low-mapping-quality reads (MAPQ < 10) and exclude duplicated reads separately using Samtools (version 1.3.1) and Picard (http://broadinstitute.github.io/picard/) (Li et al., 2009). (IV) Call peaks through MACS2 (version 2.1.0) with the parameter extsize = 200 and shift = 100 (Zhang et al., 2008). (V) Use HTseq (version 0.12.4) to calculate the count number of each peak (Anders et al., 2015). (VI) Combine the peaks across all samples to obtain the union peaks and identify differentially accessible peaks using EdgeR (version 3.32.0) (Robinson et al., 2010).

Different from bulk ATAC-seq data, scATAC-seq data was preprocessed through following steps. (I) Map raw sequencing data in fastq format to the reference genome (GRCm38/mm10) with cellranger (version 2.0.0) (Zheng et al., 2017). Reads were shifted on the forward strand by +4 bp and on the reverse strand by −5 bp. (II) Use ArchR (version 1.0.1, minTSS = 4, minFrags = 1,000, dimsToUse = 2:30, knnIteration = 1,500) to integrate scATAC-seq and scRNA-seq data with unconstrained integration methods (Granja et al., 2021). We identified peak-to-gene links by calculating the correlation between peak accessibility and gene expression across individual cells, and retained peak-to-gene links with absolute value of correlation >0.2, FDR < 1E-6, varCutOffATAC (variance of peak accessibility) > 0.7 and varCutOffRNA (variance of gene expression) > 0.3. Here, strict parameters were set to control the number of significant peak-to-gene links.

After processing bulk or single-cell ATAC-seq data, the following steps were used to identify footprints and to refine regulatory relationships. (I) Identify the footprints of peaks through HINT (version 0.13.2) and select high-quality footprints (tag-count score >80th percentile) for downstream analysis (Li et al., 2019). (II) Select footprints which are covered by differentially accessible peaks. (III) Run Fimo to find binding motifs in the footprints according to the position weight matrices of motifs from TRANSFAC database (Bailey et al., 2015). (IV) Identify footprint-related genes. For bulk ATAC-seq data, ChIPseeker (version 1.26.2, tssRegion = from upstream 3,000 to downstream 3,000) was used to annotate footprint regions. Genes related to footprint regions were considered as footprint-related genes (Yu et al., 2015). For scATAC-seq data, genes linked to peaks by ArchR were considered as footprint-related genes. (V) Use Rsamtools (version 2.6.0) to obtain the sequencing depth of the mapped reads which was used to calculate the number of insertions at each position of footprints (https://bioconductor.org/packages/Rsamtools). (VI) Use the number of insertions to calculate footprint occupancy score (FOS), and then select regulatory relationships which have high FOS (FOS > 0.1) to reconstruct regulatory networks. FOS was calculated using the formula defined as previously described (Wang et al., 2018).

$$FOS = \min\left(-\log\frac{N_C+1}{N_L+1}, -\log\frac{N_C+1}{N_R+1}\right) \qquad \text{(Equation 1)}$$

where N_L, N_C and N_R are numbers of insertions separately in the left, center and right regions of the motif.

Network modularization based on gene co-expression

Given the sparsity of scRNA-seq data, we used the smoothed expression profiles to perform gene co-expression analysis. DEGs and the expressed transcription factors were divided into different modules using the K-means clustering of the smoothed expression profiles. The optimal number of modules was determined by the silhouette coefficient calculated by R package 'cluster'. Based on the modules of genes, the inferred regulatory networks were modularized. For each module, we used ClusterProfile (version 3.18.1) to perform functional enrichment analysis which is based on gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) databases (Yu et al., 2012).

Transcription factor enrichment for modular regulatory networks

Refined regulatory relationships were used to reconstruct modular regulatory networks of DEGs and the expressed transcription factors. Cytoscape (version 3.8.2) was used to display regulatory networks (Shannon et al., 2003). We performed the hypergeometric test to calculate the probability P that an individual transcription factor regulates a module, and then adjusted p values to false discovery rate (FDR) values. Enriched transcription factors which significantly regulate each module of genes were used to reconstruct regulatory networks. For the hypergeometric test, the probability was calculated as follows.

$$P(x = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \qquad \text{(Equation 2)}$$

Here, $\binom{a}{b}$ is a binomial coefficient. For identifying the significant transcription factor regulating module A, N and n represent numbers of all regulations and regulations targeting module A, respectively. K and k separately indicate the number of regulations from transcription factor and the number of regulations targeting module A from transcription factor.

Construct simplified regulatory networks among modules

Based on modular regulatory networks of enriched transcription factors, we carried out another hypergeometric test to determine significant regulatory relationships among modules and reconstructed simplified regulatory networks among modules. For positive and negative regulations, we separately performed hypergeometric tests to identify significant regulatory relationships among modules. If positive (or negative) regulations among modules are statistically significant, activating (or repressive) connections are present among modules in the simplified regulatory networks.

The same formula as (Equation 2) was used to calculate the probability of the regulation of module A to module B. In the formula, N and n represent numbers of all regulations and regulations from module A, respectively. K and k separately indicate the number of regulations from module B and the number of regulations from module A to module B.

The p values of regulations among modules were adjusted to FDR values. Regulatory relationships with FDR < 0.05 were regarded as significant regulations, and used to construct the simplified regulatory networks among modules. The FDR < 0.05 was used for the simplified regulatory network in heart regeneration and NASH. In liver regeneration, FDR < 0.005 was used to obtain more reliable regulatory relationships between modules. Given that biological functions are enriched for genes in each module, simplified regulatory networks suggest that enriched biological functions in one module may regulate biological functions associated with other modules.

Comparison of regulatory networks

To compare IReNA with existing methods for network inference and transcription factor enrichment, we used Rcistarget from SCENIC software to identify key regulators analyzing the same scRNA-seq data (Aibar et al., 2017). In network inference, we used candidate regulatory regions identified by i-cisTarget to refine regulatory relationships inferred from scRNA-seq data analysis (Imrichová et al., 2015). To measure the enrichment of transcription factor binding motifs in the promoter regions of DEGs, Rcistarget calculates the normalized enrichment score (NES). We reconstructed regulatory networks for transcription factors which have >3 NES. For the comparison of Rcistarget, we ranked transcription factors according to NES.

To assessed the accuracy of regulatory relationships inferred by IReNA and Rcistarget, we used ChIP-seq data and genetic perturbation data of transcription factors separately from ChIP-Atlas (https://chip-atlas.org/) and KnockTF (http://www.licpathway.net/KnockTF/index.php) databases (Zou et al., 2022; Feng et al., 2020). We used regulatory relationships of transcription factors reported in literature of liver regeneration and their gene targets to assess the performance of IReNA and Rcistarget on network inference. Liver-specific ChIP-seq and genetic perturbation data were used to assess regulatory relationships inferred in liver regeneration. We obtained 368,599 regulatory relationships from liver ChIP-seq data which are obviously a larger number than 2,500 regulatory relationships obtained from genetic perturbation data in liver samples. Given this, ChIP-seq data alone, and the combination of ChIP-seq and genetic perturbation data were separately used as the ground truth to evaluate the inferred regulatory networks.

To compare regulatory networks inferred from the integrated analysis of both scRNA-seq and ATAC-seq data, and from scRNA-seq data alone, we examined whether top and all enriched transcription factors in network analysis had been previously reported in literature. Enriched transcription factors were ranked by FDR or NES. For the study of nonalcoholic steatohepatitis, we used biological terms ‘nonalcoholic steatohepatitis’, ‘steatosis’ and ‘fatty liver disease’ to search the Google Scholar and PubMed databases. We searched biological terms ‘liver regeneration’, ‘hepatic regeneration’ and ‘hepatocyte regeneration’ for liver regeneration. The terms ‘heart regeneration’, ‘cardiac regeneration’ and ‘myocardial regeneration’ were used for heart regeneration. We confirmed if the gene symbol and/or common gene name for individual transcription factors were present in literature. We also searched for gene aliases according to the NCBI database (https://www.ncbi.nlm.nih.gov/gene/). If the gene symbol/gene name and biological term are both present in the title or the same sentence in the abstract, the biological function of the enriched transcription factor is regarded to have been reported in literature. Two researchers independently searched literature and checked the results. We also calculated the significance of the co-citation between each transcription factor and all specific terms using CoCiter (version 2.1) which is based on literature from the PubMed database (Qiao et al., 2013). p values from CoCiter were used to assess the association of transcription factors with specific terms.