bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

IReNA: integrated regulatory network analysis of single-cell transcriptomes

Junyao Jiang[1], Seth Blackshaw[2], Jiang Qian[3], Jie Wang[1*]

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

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

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

[*]Corresponding author. Email: wang_jie01@gibh.ac.cn

Abstract

While single-cell RNA sequencing (scRNA-seq) is widely used to profile gene expression, few methods are available to infer gene regulatory networks using scRNA-seq data. Here, we developed and extended IReNA (Integrated Regulatory Network Analysis) to perform regulatory network analysis using scRNA-seq profiles. Four features are developed for IReNA. First, regulatory networks are divided into different modules which represent distinct biological functions. Second, transcription factors significantly regulating each gene module can be identified. Third, regulatory relationships among modules can be inferred. Fourth, IReNA can integrate ATAC-seq data into regulatory network analysis. If ATAC-seq data is available, both transcription factor footprints and binding motifs are used to refine regulatory relationships among co-expressed genes. Using public datasets, we showed that integrated network analysis of scRNA-seq data with ATAC-seq data identified a higher fraction of known regulators than scRNA-seq data alone. Moreover, IReNA provided a better performance of network analysis than currently available methods. Beyond the reconstruction of regulatory networks, IReNA can modularize regulatory networks, and identify key regulators and significant regulatory relationships for modules, facilitating the systems-level understanding of biological regulatory mechanisms. The R package IReNA is available at https://github.com/jiang-junyao/IReNA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Key points:

1. We have developed IReNA, a new method for reconstructing regulatory networks using either scRNA-seq and ATAC-seq data or scRNA-seq data alone.
2. IReNA can establish modular regulatory networks to identify key regulatory factors and statistically significant regulatory relationships among modules.
3. Through analyzing three public scRNA-seq datasets, IReNA shows better performance on identifying known regulators than Rcistarget, the most widely used alternative method.

Keywords: gene regulatory networks, single-cell RNA sequencing, chromatin accessibility

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Dynamic expression of transcription factors and changes in chromatin accessibility regulate gene expression in biological systems[1]. This fact makes it possible to infer gene regulatory networks using transcriptomic and epigenomic profiles[2]. Transcriptomes are conventionally measured through bulk RNA-seq, which only provides average mRNA levels of diverse cell types present in the sample. Recently, however, the development of single-cell RNA sequencing (scRNA-seq) has made it possible to measure the transcriptomes of thousands of individual cells[3]. ScRNA-seq provides new opportunities to reconstruct cell type-specific regulatory networks. Several features of regulatory networks could be identified by analysis of transcriptomes. For instance, genes regulated by the same transcription factors are likely co-expressed[4]. Given that biological systems are organized as modular networks, genes in the same functional module tend to share similar expression profiles[5]. Complementary to transcriptomics, epigenomic analysis identifies dynamic chromatin states that allow to identify key features of regulatory networks, including changes in chromatin accessibility and transcription factor binding sites. These features can be measured using Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq)[6]. Binding motifs in regulatory regions indicate transcription factor binding and regulation of nearby genes. Transcription factor binding motifs can be used to refine gene regulatory relationships inferred from co-expression analysis.

Several methods have been developed to reconstruct regulatory networks using bulk or single-cell transcriptomes. Through calculating weighted co-expression, WGCNA performs network analysis using bulk transcriptomes[7]. For single-cell transcriptomes, the software SCENIC has been developed for gene co-expression and cis-regulatory analysis[8]. Although current methods can infer regulatory networks to identify key regulators, they do not reveal regulatory relationships among modules of co-expressed genes. Methods for reconstructing regulatory networks by integrating single-cell transcriptomes with ATAC-seq profiles are very limited. Recently, we have developed IReNA to integrate both bulk and single-cell RNA-seq data with bulk ATAC-seq data to reconstruct modular regulatory networks which provide key transcription factors and intermodular regulations on retinal regeneration[9]. Here, we further extended the function of IReNA to regulatory network analysis only using single-cell transcriptomes, since only scRNA-seq data are available for many biological samples. Using IReNA, we analyzed published single-cell transcriptomes and ATAC-seq

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

profiles, reconstructed regulatory networks, identified key regulators, and revealed regulatory relationships among modules. In comparison with Rcistarget parckage from SCENIC, one of the most frequently used existing methods, IReNA shows better performance on identifying known regulators through modular network analysis.

MATERIALS AND METHODS

Depending on the availability of ATAC-seq data, two methods were developed to reconstruct regulatory networks using IReNA. If ATAC-seq is not available, only transcription factor binding motifs are used to refine correlation relationships obtained through scRNA-seq data analysis. If ATAC-seq data is available, transcription factor footprints and binding motifs are identified by analyzing ATAC-seq data, and then used to refine correlation relationships obtained from scRNA-seq data analysis.

Description of scRNA-seq and ATAC-seq data

To demonstrate analysis flow of IReNA, we used public scRNA-seq and ATAC-seq data from three studies, which respectively analyzed datasets obtained from models of nonalcoholic steatohepatitis (NASH), liver regeneration, and spermatogonial stem cell (SSC) development[10–12].

The scRNA-seq and ATAC-seq data from the study of NASH are available through the accession numbers GSE128334 and GSE128335 at Gene Expression Omnibus (GEO) database https://www.ncbi.nlm.nih.gov/geo/. In this study, liver tissues from mice fed a control or NASH diet were analyzed [10]. ScRNA-seq profiles were measured on 6184 non-parenchymal cells, including Kupffer cells, from healthy and NASH mice. ATAC-seq was conducted on Kupffer cells from two healthy and two NASH samples.

For the study of liver regeneration, partial hepatectomy (PHx) was performed in mice[11]. ScRNA-seq profiles were measured on 5010 hepatocytes at 0 and 48 hours after PHx (accession number GSE158866). ATAC-seq was conducted on hepatocytes with two replicates at each of 0 and 48 hours after PHx (accession number GSE158865).

The scRNA-seq and ATAC-seq data for spermatogonial stem cell development are separately available through GSE92276 and GSE92279. In the study, SSEA4 and KIT were

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

separately used as markers of spermatogonial stem cells (SSCs) and spermatogonia[12]. SSEA4$^+$ SSCs and KIT$^+$ spermatogonia from adult human testis were isolated for genome-wide sequencing. ScRNA-seq profiles were measured on 101 SSEA4$^+$ SSCs and 74 KIT$^+$ spermatogonia. ATAC-seq was conducted on one biological replicate of embryonic stem cells (ESCs) and two replicates of pooled SSEA4$^+$ SSCs.

Identify regulatory relationships using scRNA-seq data

IReNA includes four steps for analyzing scRNA-seq data to obtain regulatory relationships. (I) Use Monocle (version 2.1.8) to construct the trajectory and calculate the pseudotime of single cells[13]. (II) Identify differentially expressed genes (DEGs) during pseudotime (q-value < 0.05, fraction of expressed cells > 10% and single-cell expression difference > 0.1). Single-cell expression difference was defined as described previously[9]. (III) Calculate the smoothed expression profiles based on pseudotime and divide DEGs into different modules using the K-means clustering of the smoothed expression profiles. For each module of DEGs, ClusterProfile (version 3.18.1) was used to perform functional enrichment analysis[14]. (IV) Calculate expression correlation (Pearson's correlation) for each pair of DEGs and select highly correlated gene pairs which contain at least one transcription factor from the TRANSFAC database (version 2018.3) as potential regulatory relationships. IReNA used the R packages pheatmap to visualize expression profiles.

Identify binding motifs of transcription factors in regulatory regions of targeted genes

If ATAC-seq data is not available, IReNA uses transcription factor binding motifs present in the promoter regions of DEGs to refine gene regulatory relationships inferred from scRNA-seq data analysis. Fimo was used to identify transcription factor binding motifs in the promoter regions (ranging from 1000bp upstream to 500bp downstream of the transcription start sites) of the genes[15]. Regulatory relationships were selected for further analysis if the binding motif of transcription factor occurs in the promoter region of the target gene.

Analyze ATAC-seq data to refine identified regulatory relationships

To process ATAC-seq data for IReNA, we first removed adaptors of pair-end raw reads using fastp software (version 0.21.0)[16]. Reads were aligned to GRCh38/hg38 genome for the human or GRCm38/mm10 genome for the mouse using bowtie2 (version 2.4.1) with default parameters[17]. Then, we filtered low-mapping-quality reads (MAPQ < 10) and

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

excluded duplicated reads separately using Samtools (version 1.3.1) and Picard (http://broadinstitute.github.io/picard/)[18]. Peak calling was performed through MACS2 (version 2.1.0) with the parameter extsize = 200 and shift = 100[19]. HTseq (version 0.12.4) was used to calculate the counts of each peak[20]. Then we combined the peaks across all samples to obtain the union peaks and identified differentially accessible peaks using EdgeR[21]. Finally, we identified the footprints from ATAC-seq data through HINT (version 0.13.2) and selected high-quality footprints (tag-count score > 80th percentile) for downstream analysis[22].

After processing ATAC-seq data, we used IReNA to analyze footprints and to refine regulatory relationships in the following five steps. (I) Select footprints which are covered by differentially accessible peaks. (II) Run Fimo to find motifs in the footprints according to the position weight matrices of motifs from TRANSFAC database[15]. (III) Use ChIPseeker (version 1.26.2) to identify footprint-related genes and select footprints which contain DEGs from expression analysis[23]. (IV) Use Rsamtools (version 2.6.0) to obtain the sequencing depth of the mapped reads which is used to calculate the number of insertions at each position of footprints (https://bioconductor.org/packages/Rsamtools). (V) Use the number of insertions to calculate footprint occupancy score (FOS), and then select regulatory relationships which have high FOS to reconstruct regulatory networks. FOS was calculated using the formula defined as previously described[9].

$$FOS = min\left(-log_2 \frac{N_C + 1}{N_L + 1}, -log_2 \frac{N_C + 1}{N_R + 1}\right)$$

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

Reconstruct regulatory networks

The refined regulatory relationships were used to reconstruct regulatory networks of DEGs. Cytoscape was used to display regulatory networks[24]. We performed the hypergeometric test to calculate the probability that an individual transcription factor regulates each module, and identified the enriched transcription factors which significantly regulate each module of DEGs. We then reconstructed regulatory networks using enriched transcription factors. Next, we carried out another hypergeometric test to determine significant regulatory relationships among modules and reconstructed intermodular

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

regulatory networks. For above two hypergeometric tests, the probability was calculated as previously described[9].

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

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.

For identifying the significant regulatory relationship from module A to module B, *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.

Comparison of 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 checked whether the top 25% (ranked by degree) of transcription factors in the network had been previously reported in the literature. For the three studies of nonalcoholic steatohepatitis, liver regeneration, and spermatogonial stem cell development, we separately used the biological terms ‘nonalcoholic steatohepatitis or fatty liver disease’, ‘regeneration’ and 'stem cell' to search the literature in the Google Scholar and PubMed databases. We also confirmed if the gene symbol and/or common gene name for individual transcription factors were present. If the gene symbol/gene name and biological term were both present in the title or the same sentence in the abstract, the biological function of the enriched transcription factor was regarded to have been reported in the literature.

To compare IReNA with existing methods for inferring gene regulatory networks, we also reconstructed gene regulatory networks through SCENIC using the same scRNA-seq data[8]. R package Rcistarget from SCENIC was used to calculate normalized enrichment score (NES) which measure the enrichment of transcription factor binding motifs in the promoter regions of DEGs. We reconstructed regulatory networks for transcription factors

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

which have > 3 NES. Then, we checked whether the top 25% of transcription factors identified through Rcistarget were reported in the literature.

RESULTS

To illustrate the features and application of IReNA, we firstly used public scRNA-seq and ATAC-seq data of Kupffer cells from a mouse model of nonalcoholic steatohepatitis (NASH)[10]. Although scRNA-seq and ATAC-seq data were obtained, the original study didn't perform regulatory network analysis. We used IReNA to reconstruct two types of regulatory networks. The first was inferred using scRNA-seq data alone. The second was obtained from integrating scRNA-seq with ATAC-seq profiles (Figure 1). After reconstruction, regulatory networks of differentially expressed genes (DEGs) were modularized to identify enriched transcription factors. Regulatory networks among modules were further inferred according to regulatory networks of enriched transcription factors.

Network analysis only using scRNA-seq data

First, IReNA was used to infer regulatory networks by analyzing scRNA-seq data alone. We used 2748 Kupffer cells from healthy and NASH liver tissues to construct the trajectory and calculate the pseudotime of cells (Figure 2A). We then identified 2523 DEGs changed during the pseudotime from the health to NASH. Given the sparsity of scRNA-seq data, we used the smoothed expression profiles to perform downstream analysis. The smoothed expression profiles were obtained through dividing pseudotime into 50 equal intervals and calculating the average expression profile of single cells in each interval. Through K-means clustering of the smoothed expression profiles, 2523 DEGs were separated into three modules (Figure 2B). DEGs in three modules showed specific expression profiles and distinct enriched functions, such as myeloid cell differentiation, cytoplasmic translation, and oxidative phosphorylation (Figure 2C). We then identified potential regulatory relationships based on expression correlations. The correlation was calculated for each pair of 2523 DEGs. We selected 46,729 gene pairs as the potential regulatory relationships, each of which has > 0.4 absolute correlation and contains at least one transcription factor.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To refine 46,729 potential regulatory relationships, we further analyzed transcription factor binding motifs in the promoter regions of DEGs. We obtained 16,923 regulatory relationships and reconstructed regulatory networks (Figure 2D). For regulatory networks of DEGs, we performed the hypergeometric test and identified 41 enriched transcription factors which significantly regulate modules of DEGs. Regulatory networks of enriched transcription factors were further reconstructed (Figure 2E). To identify significant regulatory relationships among modules, we performed the hypergeometric test on regulatory networks of enriched transcription factors. Five significant regulations among modules were identified and used to establish intermodular regulatory networks (Figure 2F). In intermodular regulatory networks, we observed that transcription factors from the module of myeloid cell differentiation significantly repress transcription factors related to oxidative phosphorylation.

Network analysis through integrating scRNA-seq and ATAC-seq data

Next, we use IReNA to infer regulatory networks through integrating scRNA-seq and ATAC-seq data. To refine 46,729 potential regulatory relationships from co-expression analysis, we analyzed ATAC-seq data and obtained 73,282 peaks from healthy and NASH Kupffer cells (Figure 3A). Within the peaks, we identified 31,551 footprints which have > 1 footprint occupancy score. Genes related to footprinted regions were identified and linked to binding motifs of individual transcription factors. Comparing footprints with 46,729 correlated gene pairs, we refined 2982 regulations and reconstructed regulatory networks which contain 994 DEGs.

Analyzing modular regulatory networks of 994 DEGs, we identified 28 transcription factors which significantly regulate the expression of each module of genes (p-value < 0.01). We then reconstructed regulatory networks of 28 enriched transcription factors (Figure 3B). Four significant regulations among three modules were also identified to reconstruct intermodular regulatory networks (FDR < 0.05, Figure 3C). We observed that intermodular regulatory networks from the integrated analysis of scRNA-seq and ATAC-seq data are consistent with intermodular regulatory networks obtained from analyzing scRNA-seq data alone.

Performance of IReNA on regulatory network analysis

To directly compare the performance of other methods with IReNA, we analyzed scRNA-seq data to reconstruct regulatory networks using the Rcistarget package from

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SCENIC software[8]. Using Rcistarget, we identified 42 transcription factors whose binding motifs are overrepresented in the promoter regions of 46,729 potential regulations inferred from scRNA-seq data analysis. We obtained 152 significant regulatory relationships for 42 enriched transcription factors, and then reconstructed gene regulatory networks (Figure 3D).

We then compared gene regulatory networks inferred through three different approaches described above. Among 41 transcription factors identified using IReNA analysis of scRNA-seq data alone, 20 (48.78%) transcription factors are also present in the networks inferred by IReNA using integrated analysis of scRNA-seq and ATAC-seq data (Figure 3E). By comparing gene regulatory networks obtained by analyzing only scRNA-seq data separately using IReNA and Rcistarget, we observed an overlap of 46.34% (19 in 41) of all transcription factors. These indicated that about half transcription factors were identified by two methods.

To further assess the significance of transcription factors identified by IReNA or Rcistarget, we manually checked whether these factors had been previously reported in the literature. We ranked and checked top 25% enriched transcription factors in regulatory networks. We found that in regulatory networks from the integrated analysis of scRNA-seq and ATAC-seq data, 57.14% transcription factors are reported in NASH-related literature, including Fos, Egr1 and Fosb (Figure 3F and Table S1). In regulatory networks inferred from analyzing scRNA-seq data alone, 50.00% transcription factors are reported in NASH-related literature. 45.45% transcription factors in regulatory networks inferred by Rcistarget are reported in NASH-related literature. Among three types of regulatory networks, the highest fraction of transcription factors is reported for regulatory networks from the integrated analysis of ATAC-seq data with scRNA-seq data. These results indicate that integrated network analysis of ATAC-seq with scRNA-seq data using IReNA improved the identification of known transcription factors. Moreover, IReNA shows better performance at identifying known regulators than the Rcistarget method.

To further demonstrate the performance of IReNA, we performed regulatory network analysis on another two datasets from liver regeneration and spermatogonial stem cell development. Prior to identifying gene regulatory networks controlling liver regeneration, we reconstructed the trajectory of 5010 hepatocytes from liver tissues at 0 h and 48 h after hepatectomy. We found that hepatocytes formed two distinct branches (named the activation branch and the proliferative branch) in the trajectory at 48 h after hepatectomy (Figure S1A). We further identified and divided 4641 DEGs into 4 modules (Figure S1B). DEGs in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

activation branch (module 3 and 4) are enriched in genes involved in fatty acid metabolic process and ribosome, whereas DEGs in the proliferative branch (module 1) are enriched for focal adhesion (Figure S1C). Then, regulatory networks were inferred using the same three methods described above (Figure 4A-4B and Figure S1D-S1E). In comparison with regulatory networks inferred by Rcistarget, regulatory networks reconstructed by IReNA contain a higher fraction of transcription factors previously reported in the literature related to liver regeneration (75.00% for scRNA-seq + ATAC-seq + IReNA, 55.56% for scRNA-seq + IReNA and 37.50% for scRNA-seq + Rcistarget) (Figure 4C and Table S1).

For the study of spermatogonial stem cell development, we used the 2684 DEGs and four modules reported by the original paper to reconstruct three types of regulatory networks (Figure 4D-4E and Figure S2A-S2B). IReNA analysis using scRNA-seq and ATAC-seq data, IReNA analysis using scRNA-seq data alone, and Rcistarget analysis using scRNA-seq data alone respectively identified 77.78%, 70.59% and 66.67% transcription factors which are reported to associate with spermatogonial stem cell development (Figure 4F and Table S1). The integrated analysis of ATAC-seq with scRNA-seq data overall substantially improved the reconstruction of gene regulatory networks and identified a higher fraction of known transcriptional regulators. In addition, regulatory networks reconstructed from scRNA-seq data alone using IReNA show improved accuracy relative to those identified using Rcistarget.

DISCUSSION

In the study, we developed the R package IReNA to perform regulatory network analysis. In IReNA, gene regulatory networks could be reconstructed using either scRNA-seq data alone or the integrated analysis of scRNA-seq and ATAC-seq data. When only scRNA-seq data are used, expression correlation and transcription factor binding motifs are calculated to determine regulatory relationships. When both scRNA-seq and ATAC-seq data are available, expression correlation, transcription factor binding motifs and footprints are used to identify regulatory relationships. Beyond the reconstruction of gene regulatory networks, IReNA provides additional insights, including the modularization of regulatory networks, the enrichment of transcription factors in each module, and the inference of regulatory relationships among modules. IReNA not only provides cell type-specific

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

regulatory networks, but also identifies modular regulatory networks which are associated with distinct cell states of the same cell type, such as healthy and NASH Kupffer cells.

In previous work, we demonstrated that IReNA can be used to integrate bulk RNA-seq, scRNA-seq and ATAC-seq to reconstruct gene regulatory networks controlling retinal regeneration[9]. Through network analysis using IReNA, we identified modular gene regulatory networks and key transcription factors regulating different cell states in retinal Müller glia in zebrafish and mice. Using genetic loss of function analysis, we confirmed that several transcription factors identified by IReNA are critical for retinal regeneration, including *hmga1a* and *yap1* in zebrafish and *Nfia/b/x* in mice. These results indicate that IReNA can provide reliable regulatory networks and reveal key regulators during biological processes including retinal regeneration.

Using public scRNA-seq and ATAC-seq data from three studies, we further performed regulatory network analysis through IReNA, using either integrated analysis of both datasets or scRNA-seq data alone. When both scRNA-seq and ATAC-seq data are used, a much higher fraction of known transcription factors is identified by IReNA. For instance, in analysis of NASH samples, we identified *Fos* and *Klf6*, which regulate TH17/Treg cells in NASH patients and are up-regulated in the context of steatohepatitis, respectively[25,26]. IReNA also identified *Cebpa* and *Hnf4a* as regulators of liver regeneration[27]. In regulatory networks associated with SSC development, IReNA identified transcription factors *DMRT1* and *NR6A1*, which were reported as regulating SSC development in the original study[12]. The results indicate that IReNA could be applied to identify key regulators in different biological processes and outperforms Rcistarget, the most widely used alternative method.

SUPPLEMENTARY DATA

Supplementary data are available online.

ACKNOWLEDGEMENT

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

This work was supported by the National Natural Science Foundation of China [32170849], and the Guangdong Province Science and Technology Program [2020B1212060052].

AUTHOR CONTRIBUTION

J.W., and J.Q. conceived the project. J.W., J.Q., and S.B. supervised the research. J.Y., and J.W. developed IReNA and performed sequencing data analysis to construct regulatory networks.

REFERENCES

1. Thompson D, Regev A, Roy S. Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu. Rev. Cell Dev. Biol. 2015; 31:399–428
2. Fiers MWEJ, Minnoye L, Aibar S, et al. Mapping gene regulatory networks from single-cell omics data. Brief. Funct. Genomics 2018; 17:246–254
3. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015; 161:1202–1214
4. Saha A, Kim Y, Gewirtz ADH, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. 2017; 27:1843–1858
5. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 2003; 34:166–176
6. Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 2013; 10:1213–1218
7. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9:559
8. Aibar S, González-Blas CB, Moerman T, et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 2017; 14:1083–1086
9. Hoang T, Wang J, Boyd P, et al. Gene regulatory networks controlling vertebrate retinal regeneration. Science 2020; 370:
10. Seidman JS, Troutman TD, Sakai M, et al. Niche-Specific Reprogramming of Epigenetic Landscapes Drives Myeloid Cell Diversity in Nonalcoholic Steatohepatitis. Immunity 2020; 52:1057–1074.e7

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

11. Chen T, Oh S, Gregory S, et al. Single-cell omics analysis reveals functional diversification of hepatocytes during liver regeneration. JCI Insight 2020; 5:

12. Guo J, Grow EJ, Yi C, et al. Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development. Cell Stem Cell 2017; 21:533–546.e6

13. Qiu X, Mao Q, Tang Y, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 2017; 14:979–982

14. Yu G, Wang L-G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012; 16:284–287

15. Bailey TL, Johnson J, Grant CE, et al. The MEME Suite. Nucleic Acids Res. 2015; 43:W39–49

16. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018; 34:i884–i890

17. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods 2012; 9:357–359

18. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078–2079

19. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137

20. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 2015; 31:166–169

21. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26:139–140

22. Li Z, Schulz MH, Look T, et al. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019; 20:45

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

24. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504

25. Cai C, Chen D-Z, Tu H-X, et al. MicroRNA-29c acting on FOS plays a significant role in nonalcoholic steatohepatitis through the interleukin-17 signaling pathway. Front. Physiol. 2021; 12:597449

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. Stärkel P, Sempoux C, Leclercq I, et al. Oxidative stress, KLF6 and transforming growth factor-β up-regulation differentiate non-alcoholic steatohepatitis progressing to fibrosis from uncomplicated steatosis in rats. J. Hepatol. 2003; 39:538–546

27. Jakobsen JS, Waage J, Rapin N, et al. Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. Genome Res. 2013; 23:592–603

FIGURE AND TABLE LEGENDS

Figure 1. Workflow of IReNA (integrated regulatory network analysis). (A) Workflow of IReNA. (B) Diagram of IReNA. Regulatory networks could be reconstructed by integrating gene expression profiles (scRNA-seq or bulk RNA-seq data) in combination with chromatin accessibility profiles (ATAC-seq data), or by using only gene expression profiles (scRNA-seq or bulk RNA-seq data). In cases where only gene expression profiles are used, regulatory relationships are inferred from expression correlation and binding motifs. If ATAC-seq data is available, expression correlation, binding motifs and footprints are used to reconstruct regulatory networks of differentially expressed genes (DEGs). Enriched transcription factors (TFs) are identified by performing the hypergeometric test on modular regulatory networks of DEGs. Regulatory networks among modules are further inferred by analyzing regulatory networks of enriched TFs.

Figure 2. Regulatory network analysis of Kupffer cells from healthy and NASH liver tissues. (A) Trajectory of 2748 Kupffer cells separately colored by pseudotime and samples. (B) ScRNA-seq expression profiles of 2523 differentially expressed genes (DEGs). K-means clustering was used to divide DEGs into three modules. Each row represents one gene, and each column indicates one interval of pseudotime. (C) Enriched functions of three modules of DEGs. (D) Fraction of correlated gene pairs 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. Color of each circle indicates the module. Grey edge represents negative regulation, and the yellow edge represents positive regulation. (F) Regulatory networks among three modules obtained from analyzing scRNA-seq data alone.

Figure 3. Comparison of regulatory networks related to NASH. (A) Genome-wide chromatin accessibility of healthy and NASH Kupffer cells measured by ATAC-seq. (B) Regulatory networks of 28 enriched transcription factors obtained through the integrated analysis of scRNA-seq and ATAC-seq data. Color of the circle indicates the module. Grey edge represents negative regulation, and the yellow edge represents positive regulation. (C) Regulatory networks among modules obtained through the integrated analysis of scRNA-seq and ATAC-seq data. (D) Regulatory networks for 42 enriched transcription factors from scRNA-seq data analysis using Rcistarget. (E) Venn diagram of enriched transcription factors from three types of regulatory networks, one from the integrated analysis of scRNA-seq and ATAC-seq data using IReNA (ATAC-seq + scRNA-seq + IReNA), one from scRNA-seq data analysis using IReNA (scRNA-seq + IReNA) and one from scRNA-seq data analysis

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

using Rcistarget (scRNA-seq + Rcistarget). (F) Fraction of transcription factors reported in the literature related to NASH. Top 25% enriched transcription factors were compared for all three types of regulatory networks.

Figure 4. Regulatory network analysis for two studies of liver regeneration and spermatogonial stem cell (SSC) development. (A) Regulatory networks of 15 enriched transcription factors obtained by integrating scRNA-seq and ATAC-seq data from liver regeneration. Module is represented by the color of circles. Grey edge represents negative regulation, and the yellow edge represents positive regulation. (B) Regulatory networks among modules obtained by integrating of scRNA-seq and ATAC-seq data from liver regeneration. (C) Fraction of transcription factors reported in the literature related to liver regeneration. (D) Regulatory networks of 34 enriched transcription factors obtained by integrating scRNA-seq and ATAC-seq data from SSC development. (E) Regulatory networks among modules obtained by integrating scRNA-seq and ATAC-seq data from SSC development. (F) Fraction of transcription factors reported in the literature related to SSC development.

Figure S1. Regulatory network analysis of liver regeneration. (A) Trajectory of 5010 hepatocytes at 0 and 48h after hepatectomy. (B) 4641 DEGs were divided into 4 modules according to K-means clustering of scRNA-seq expression profiles. Each row represents one gene, and each column indicates one interval of pseudotime. (C) Enriched functions of four modules of DEGs. Colors represent modules. (D) Regulatory networks of 73 enriched transcription obtained by analyzing scRNA-seq data alone using IReNA. Grey edge represents negative regulation, and the yellow edge represents positive regulation. (E) Regulatory networks of 33 enriched transcription factors obtained by Rcistarget analysis of scRNA-seq data.

Figure S2. Regulatory network analysis of spermatogonial stem cell (SSC) development. (A) Regulatory networks of 67 enriched transcription factors obtained from scRNA-seq data analysis using IReNA. Grey edge represents negative regulation, and the yellow edge represents positive regulation. (B) Regulatory networks for 24 enriched transcription factors obtained from scRNA-seq data analysis using Rcistarget.

Table S1. The list of transcription factors reported in the literature related to nonalcoholic steatohepatitis, liver regeneration and spermatogonial stem cell development. Transcription factors were identified through the integrated analysis of scRNA-seq and ATAC-seq data using IReNA.

Study	Transcription factor	Related literature	Report of transcription factor in the title or abstract of literature
Nonalcoholic steatohepatitis	Fos	Cai,C., Chen,D.-Z., Tu,H.-X., et al. (2021) Front. Physiol., 12, 597449.	MicroRNA-29c Acting on FOS Plays a Significant Role in Nonalcoholic Steatohepatitis Through the Interleukin-17 Signaling Pathway.
Nonalcoholic steatohepatitis	Fosb	Hasenfuss,S.C., Bakiri,L., Thomsen,M.K., et al. (2014) Cell Metab., 19, 84–95.	Regulation of Steatohepatitis and PPARγ Signaling by Distinct AP-1 Dimers. Note: activator protein 1 (AP-1) is a dimeric transcription factor consisting of Fos (c-

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

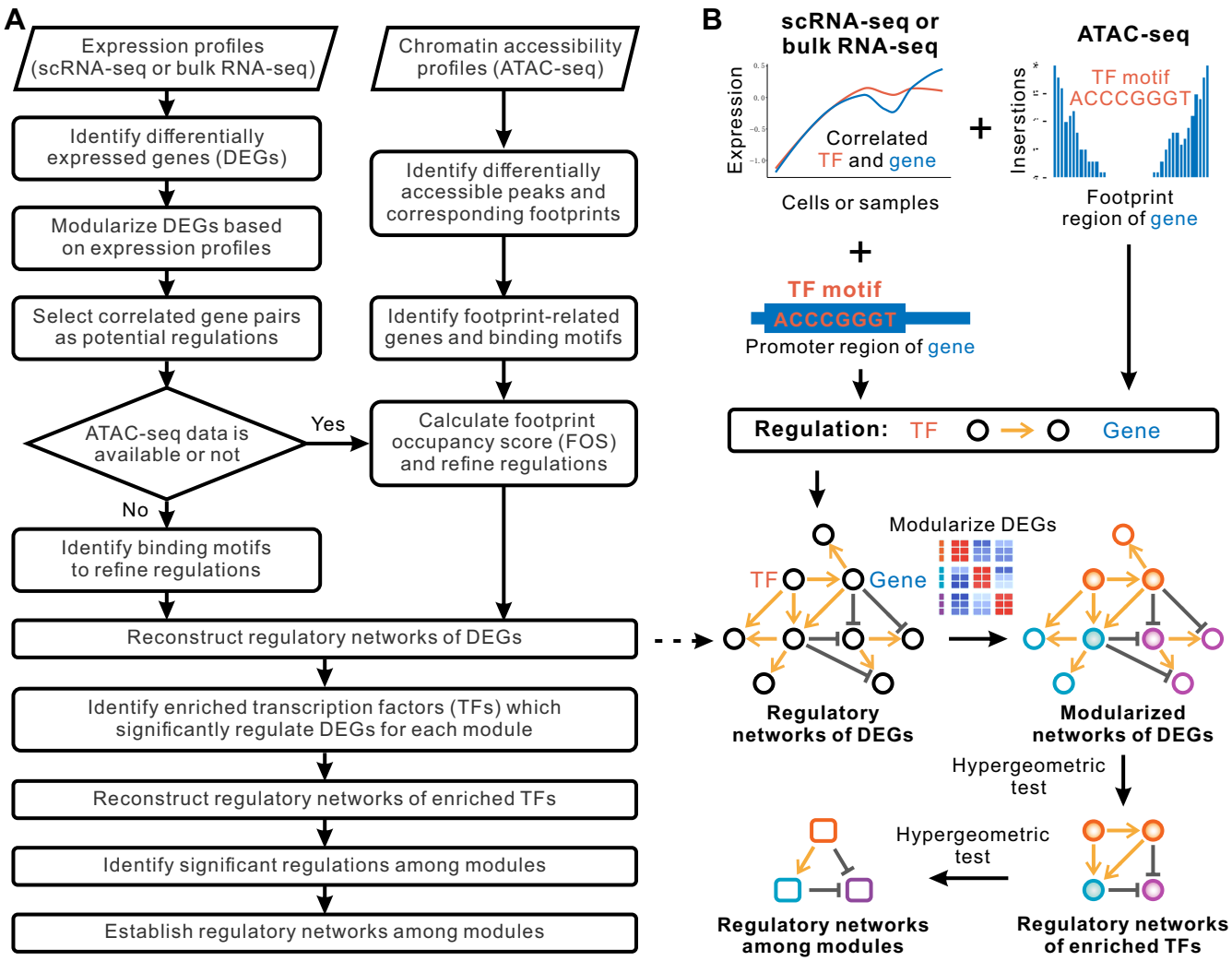
			Fos, FosB, Fra-1 and Fra-2) and other proteins.
Nonalcoholic steatohepatitis	Egr1	Li, Z., Yu, P., Wu, J., et al. (2019) Medical science monitor: international medical journal of experimental and clinical research., 25, 2993.	Transcriptional Regulation of Early Growth Response Gene-1 (EGR1) is Associated with Progression of Nonalcoholic Fatty Liver Disease (NAFLD) in Patients with Insulin Resistance.
Nonalcoholic steatohepatitis	Klf6	Stärkel, P., Sempoux, C., Leclercq, I., et al. (2003) Journal of hepatology., 39(4), 538-546.	Oxidative stress, KLF6 and transforming growth factor-β up-regulation differentiate non-alcoholic steatohepatitis progressing to fibrosis from uncomplicated steatosis in rats.
Nonalcoholic steatohepatitis	Ets2	Unreported	Unreported
Nonalcoholic steatohepatitis	Zeb2	Unreported	Unreported
Nonalcoholic steatohepatitis	Klf2	Unreported	Unreported
Liver regeneration	Zbtb7a	Unreported	Unreported
Liver regeneration	Hnf4a	Huck, I., Gunewardena, S., Espanol-Suner, R., et al. (2019) Hepatology., 70(2), 666-681.	Hepatocyte Nuclear Factor 4 Alpha Activation Is Essential for Termination of Liver Regeneration in Mice.
Liver regeneration	Klf13	Chen, L., Yao, F., Qin, Y., et al. (2020) Gene., 735, 144407.	The potential role of Krüppel-like factor 13 (Aj-klf13) in the intestine regeneration of sea cucumber Apostichopus japonicus.
Liver regeneration	Cebpa	Jakobsen, J. S., Waage, J., Rapin, N., et al. (2013) Genome research., 23(4), 592-603.	CEBPA and CEBPB (also known as C/EBPalpha and C/EBPbeta), at multiple time points during the highly dynamic process of liver regeneration.
Spermatogonial stem cell development	DMRTB1	Zhang, T., Oatley, J., Bardwell, V. J., et al. (2016) PLoS genetics., 12(9), e1006293.	Here we investigate the role of Dmrt1 in mouse spermatogonial stem cell (SSC) homeostasis.
Spermatogonial stem cell development	TEAD1	Pagliari, S., Vinarsky, V., Martino, F., et al. (2021) Cell Death & Differentiation., 28(4), 1193-1207.	YAP–TEAD1 control of cytoskeleton dynamics and intracellular tension guides human pluripotent stem cell mesoderm specification.
Spermatogonial stem cell development	KLF13	Hu, Y., Zhang, M., Tian, N., et al. (2019) The Journal of clinical investigation., 129(8), 3072-3085.	The antibiotic clofoctol suppresses glioma stem cell proliferation by activating KLF13.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spermatogonial stem cell development	KLF6	Unreported	Unreported
Spermatogonial stem cell development	FOXG1	Bulstrode, H., Johnstone, E., Marques-Torrejon, M. A., et al. (2017) Genes & development., 31(8), 757-773.	Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators.
Spermatogonial stem cell development	KLF3	Hao, J., Yang, X., Zhang, C., Zhang, X. T., et al. (2020) Cell proliferation., 53(11), e12914.	KLF3 promotes the 8-cell-like transcriptional state in pluripotent stem cells.
Spermatogonial stem cell development	RAD21	Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., et al. (2011) PloS one., 6(5), e19470.	RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity.
Spermatogonial stem cell development	TCF4	Mardaryev, A. N., Meier, N., Poterlowicz, K., et al. (2011) Development., 138(22), 4843-4852.	Lhx2 differentially regulates Sox9, Tcf4 and Lgr5 in hair follicle stem cells to promote epidermal regeneration after injury.
Spermatogonial stem cell development	AEBP2	Unreported	Unreported

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1

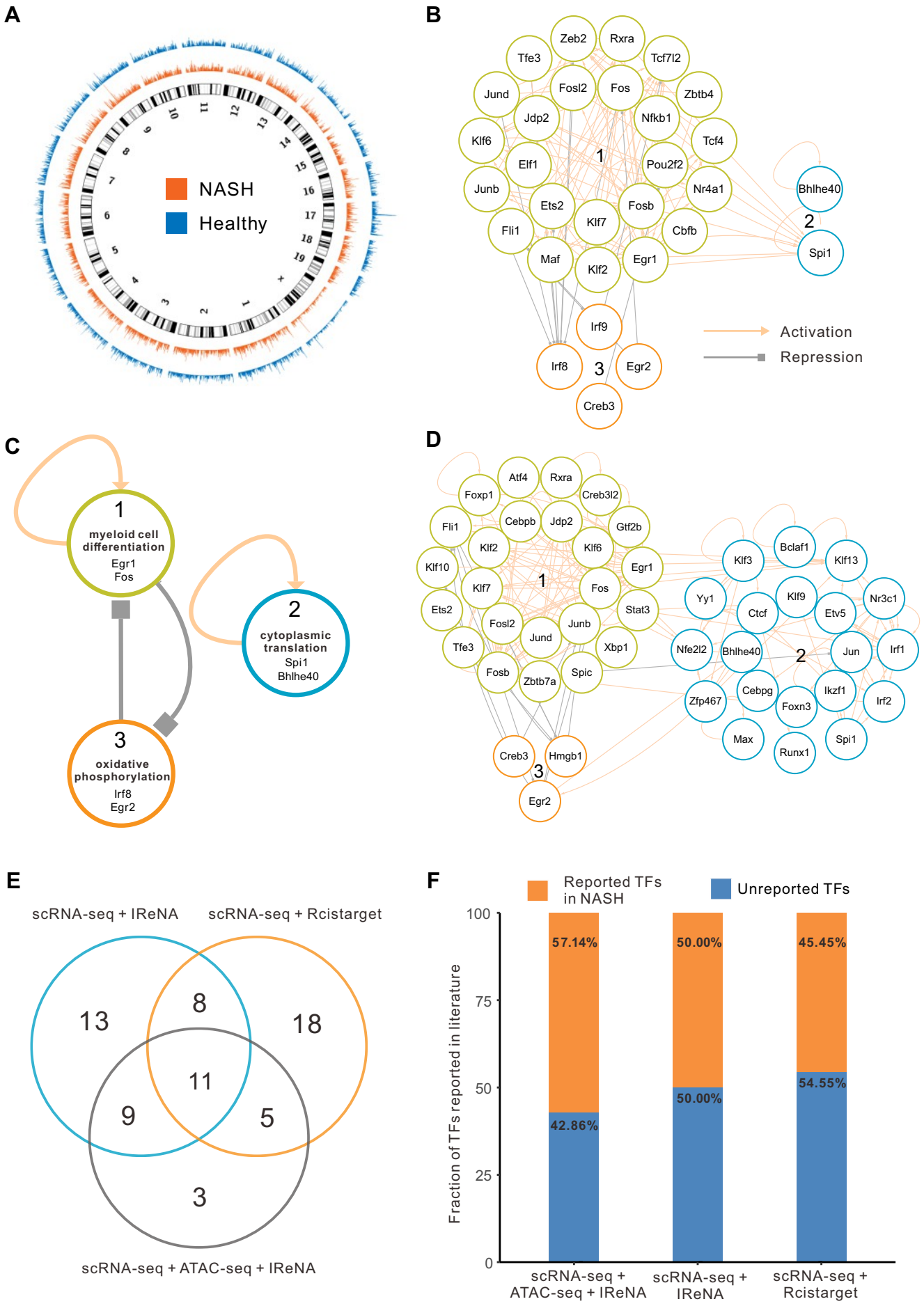


bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2

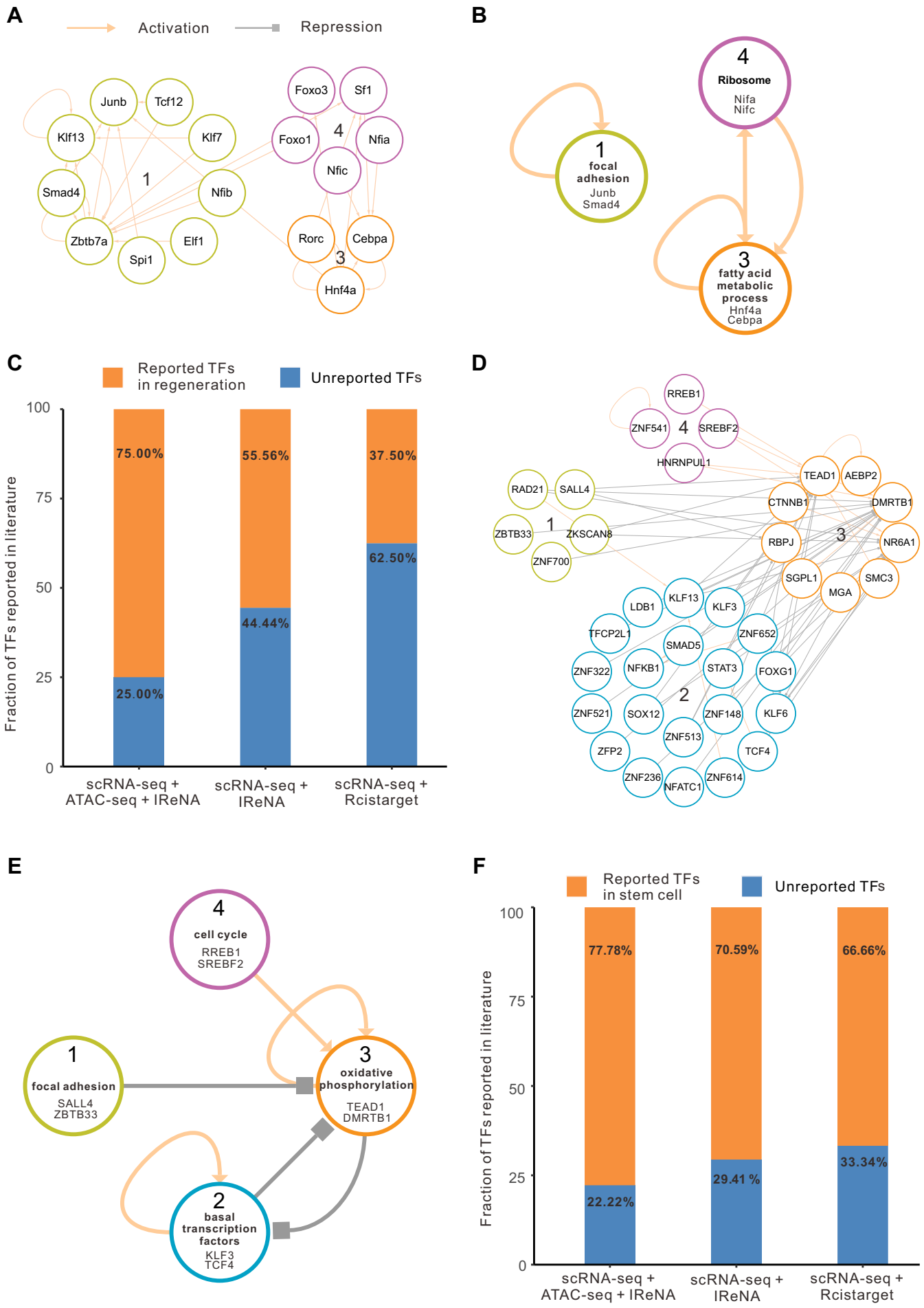
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3



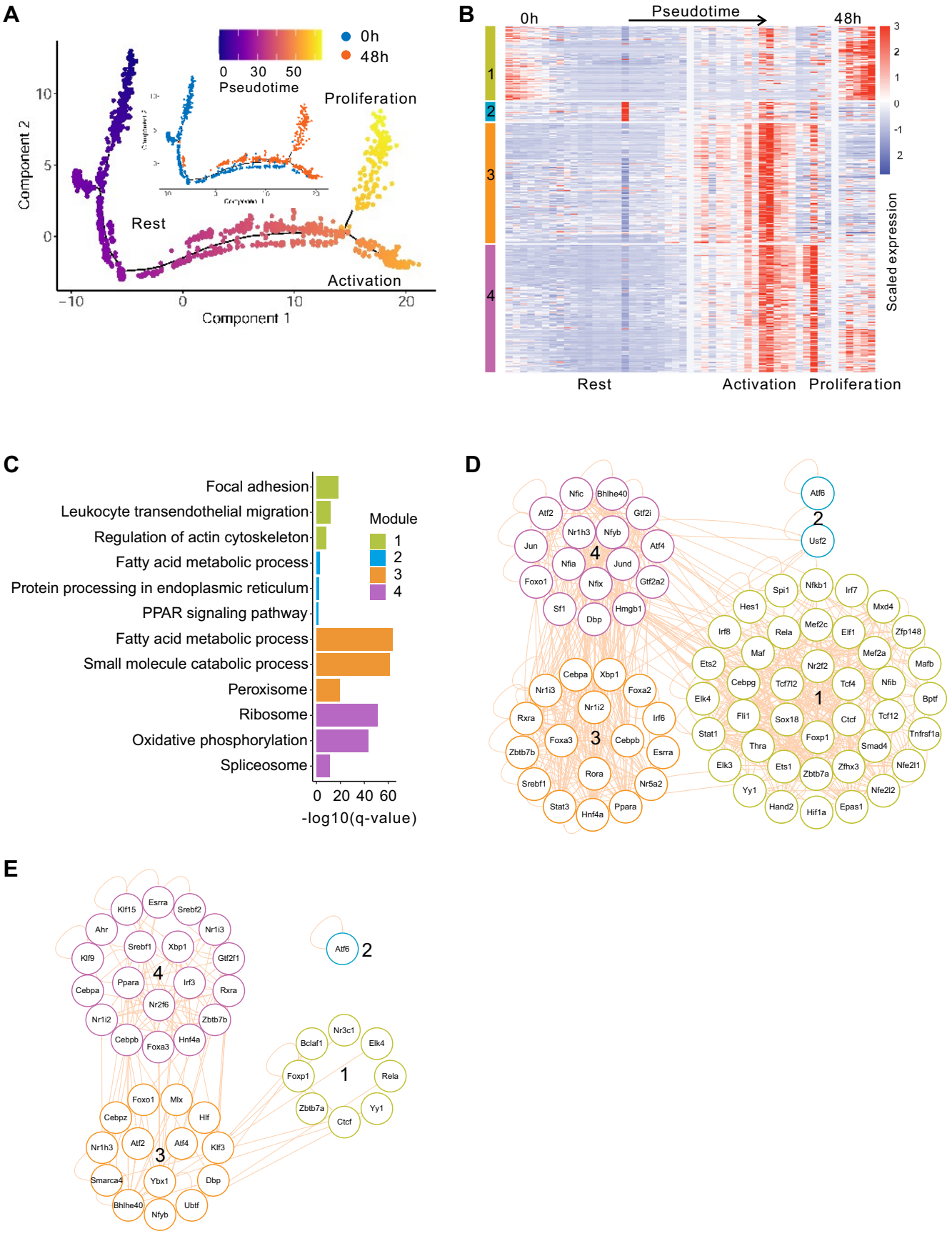
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S1



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.22.469628; this version posted November 23, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

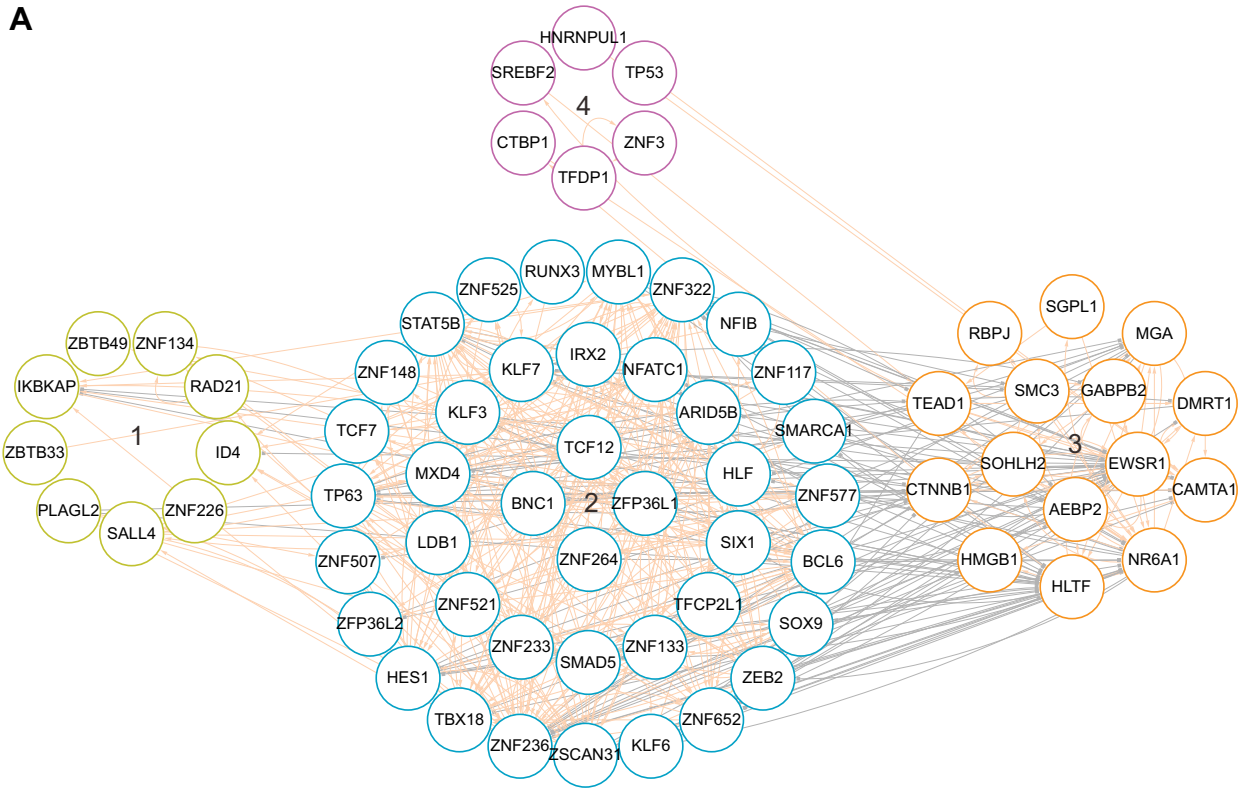
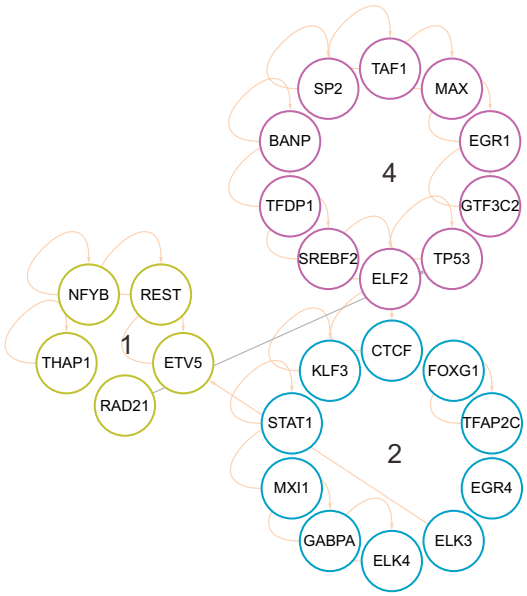



Figure S2