Published online 8 November 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 4 **1**
https://doi.org/10.1093/nargab/lqab100

Taiji-reprogram: a framework to uncover cell-type specific regulators and predict cellular reprogramming cocktails

Jun Wang[1,†], Cong Liu[1,†], Yue Chen[1,†] and Wei Wang[1,2,*]

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0359, USA and [2]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093-0359, USA

Received July 27, 2021; Revised September 29, 2021; Editorial Decision October 04, 2021; Accepted October 05, 2021

ABSTRACT

Cellular reprogramming is a promising technology to develop disease models and cell-based therapies. Identification of the key regulators defining the cell type specificity is pivotal to devising reprogramming cocktails for successful cell conversion but remains a great challenge. Here, we present a systems biology approach called Taiji-reprogram to efficiently uncover transcription factor (TF) combinations for conversion between 154 diverse cell types or tissues. This method integrates the transcriptomic and epigenomic data to construct cell-type specific genetic networks and assess the global importance of TFs in the network. Comparative analysis across cell types revealed TFs that are specifically important in a particular cell type and often tightly associated with cell-type specific functions. A systematic search of TFs with differential importance in the source and target cell types uncovered TF combinations for desired cell conversion. We have shown that Taiji-reprogram outperformed the existing methods to better recover the TFs in the experimentally validated reprogramming cocktails. This work not only provides a comprehensive catalog of TFs defining cell specialization but also suggests TF combinations for direct cell conversion.

INTRODUCTION

Transcription factors (TFs) play pivotal roles during development, cell type specification and aging (1). Identification of the key TFs in each cell type would facilitate understanding the regulatory mechanisms that decide cell fate and cell identity, and provide clues and intervention strategies for cell fate determination. The best known example for cellular reprogramming is the generation of induced pluripotent stem cells (iPSCs) from somatic cells by the introduction of four TFs (Oct4, Sox2, Klf4 and c-Myc) (2), which demonstrated that using TFs as reprogramming factors can induce drastic cell conversion. Subsequently, transdifferentiation between different pairs of terminally differentiated cell types without going through a pluripotent state has been achieved (3,4).

Cellular reprogramming opens new doors towards understanding the mechanisms underlying development as well as developing new cell therapy (5–8). A major roadblock toward achieving cell conversion is to develop effective reprogramming cocktails. The main challenges include how to identify key regulators in the source and target cell types/tissues that can serve as the candidate reprogramming factors and how to efficiently consider the exponentially increasing combinations given a set of candidate TFs. Furthermore, as epigenetic state is crucial in deciding cell state and cell type specificity (9–11), how to consider the epigenomes of the source and target cell types/tissues is also pivotal to developing reprogramming cocktails.

Efforts have been devoted to addressing these challenges. High expression level is useful to select important TFs in a particular cell type while its limitation is also well acknowledged as the activity of a TF can be regulated through post-translational modifications and other non-transcriptional mechanisms. Methods such as Schacht *et al.* (12) and Arrieta-Ortiz *et al.* (13) have been developed to consider the target genes of a TF whose expression levels reflect the TF's regulatory activity. Other methods including CellNet (14), Mogrify (15) and PANDA (16) predict the key TFs by reconstructing genetic network based on expression or protein–protein interaction data and how the combinations of TFs would regulate the differentially expressed genes in the source and target cell types in cellular conversion. While such an approach has helped to identify key TFs that are not found by only considering their own expressions, it does not fully consider the regulatory effect of a TF propagating through the genetic network, i.e. a TF impacts not only its direct targets but also their descendants and the feedback from the descendants to the TF is also crucial to

*To whom correspondence should be addressed. Tel: +1 858 822 4240; Fax: +1 858 822 4236; Email: wei-wang@ucsd.edu
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors

affect the phenotypic outcome. Therefore, considering the complexity of the genetic network and assessing the global importance of a TF in the network is critical to identify the key regulators for cell type/tissue specification.

Previously we developed a systems biology method called Taiji (17) that integrates transcriptomic and epigenomic data to construct a cell-type specific genetic network, based on which the global importance of each TF is evaluated by a personalized PageRank algorithm. We have shown that Taiji is robust and resistant to noise that is unavoidable in constructing genetic networks. The effectiveness of identifying key regulators by Taiji has been confirmed using simulated data and experimental validations (17,18).

Here, we leverage the power of Taiji to develop a systematic approach called Taiji-reprogram for efficiently developing reprogramming cocktails. Taking advantage of the vast amount of epigenomic data generated by the ENCODE (19,20) and the NIH Epigenomics Roadmap (21) projects, we have applied Taiji to identify key TFs in diverse cell types and tissues. Because the epigenomic data are highly cell-type specific, the genetic network constructed by Taiji captures the regulatory interactions specifically present in a particular cell type and the key regulators with the most global importance in the network are expected to be tightly associated with cell-type specific functions. Using the PageRank scores of the top TFs, we can efficiently evaluate the TF combinations and find the most promising cocktails for a defined reprogramming task. We showed the superior performance of our approach in comparison with the existing methods on identifying reprogramming factors in the experimentally achieved cocktails.

MATERIALS AND METHODS

Data acquisition

Taiji integrates gene expression and open chromatin (ATAC-Seq or DNase-seq) or H3K27ac ChIP-Seq data to identify key regulators. By 1 February 2019, there were 154 cell-types with matched open chromatin/H3K27ac (indicating active promoters/enhancers) and gene expression data in the same cell type/tissue from ENCODE and the NIH Epigenomics Roadmap projects. Bam files for open-chromatin and text files for expression were downloaded from the project portals (Supplementary Table S1).

Taiji-reprogram prediction and evaluation

The Taiji PageRank score S_{tf} reflects the global importance of the TF. We first calculated PageRank score ratio.abs for each TF as the larger ratio between its PageRanks in the target and source cell types: if $\frac{S^t_{tf}}{S^s_{tf}} > 1$, ratio.abs $= \frac{S^t_{tf}}{S^s_{tf}}$; otherwise, ratio.abs $= \frac{S^s_{tf}}{S^t_{tf}}$. A higher ratio.abs suggests that the TF has distinct importance in the target and source cell types and its perturbation is likely to have a significant contribution to cell conversion. We selected the top 30 TFs based on ratio.abs as candidate TFs to search for reprogramming cocktails. We next calculated the product of PageRank score ratio.abs of any three candidate TFs. These values for all the possible combinations (cocktails) of the top 30 TFs were transformed to z-scores, and *P*-value of 0.001 was used as a cut-off to select as candidate cocktails.

To compare the performance of Taiji-reprogram with the other methods, we ranked the TFs by their frequencies appearing in the candidate cocktails. To have the same number of TFs predicted by other methods for assessing the performance, we selected the top 8 TFs with the highest frequency. A score $\sqrt{\frac{N/R}{N_0/R_0}} \times 100$ was computed to evaluate each method, where N is the number of correctly predicted TFs by the method, R is the average rank of the correctly predicted TFs, $N_0$is the number of TFs in experimentally validated cocktails (ground truth), and $R_0$is the average rank of TFs in experimentally validated cocktails. The optimum prediction will get a score equal to 100.

RESULTS

Identification of key TFs in human cells and tissues

We collected 154 matched RNA-seq and open-chromatin or H3K27ac datasets in human cell lines, primary cells and tissues from the ENCODE and the NIH Roadmap Epigenomics projects in diverse cell types/tissues, including 54 in the embryonic and 100 postnatal stages (Supplementary Table S1). We further divided the 100 postnatal tissues into 5 newborn, 18 child and 77 adult tissues, following the ENCODE definition (20). We analyzed these data using Taiji to define the regulatory roles of a total 745 TFs that have experimentally determined motifs.

Taiji is a method integrating transcriptomic and epigenomic data to identify the global importance of TFs in the genetic network (Figure 1A). Specifically, Taiji first identifies active regulatory regions, including active promoters and active enhancers, defined by ATAC-seq, DNase or H3K27ac ChIP-seq peaks. Enhancers are then linked to their interacting promoters using chromatin interactions predicted by EpiTensor (22). In order to construct transcriptional regulatory networks, Taiji scans all active regulatory regions to identify putative TF binding sites based on motifs from the CIS-BP database (23). TFs with putative binding sites in active promoters or enhancers are then linked to the target genes. The global importance of the TFs are evaluated by applying the personalized PageRank algorithm. In this study, we used the node and edge weights to personalize the ranking algorithm. The node weights were determined by the *z*-scores of gene expression levels, which assign higher ranks to TFs regulating more differentially expressed genes. The edge weights were set to be proportional to TFs' expression levels, which help to remove TFs that are not expressed or with low-expression levels. Our previous studies have confirmed that Taiji can effectively identify important TFs using simulated data and experimental validations (17,24).

Using the Taiji pipeline, we calculated the PageRank scores for 745 TFs in the 154 cell types/tissues (Figure 1B). Based on the TF PageRank scores, we clustered the tissues/cell types and similar ones were largely close to each other in the clustering tree, obviously better than the clusters generated using the TF expression profiles (Supplementary Figure S1). For example, embryonic stem cell lines (H1, H7) were clustered with iPSC GM23338; embryonic mus-

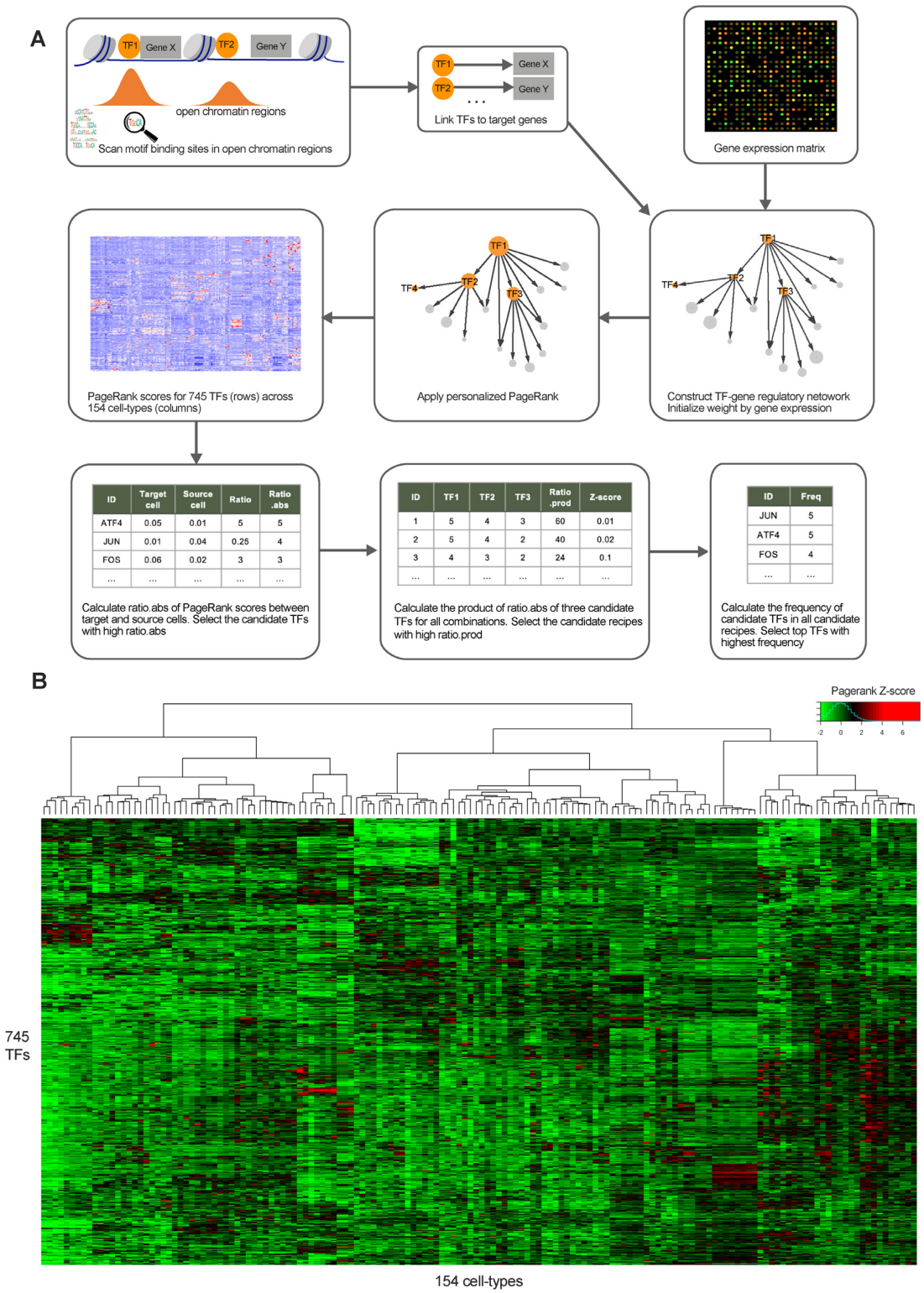

Figure 1. (**A**) The overview of the Taiji-reprogram framework. Taking the open-chromatin data and RNA-Seq data as input, the Taiji pipeline generates cell-specific PageRank scores and cell-specific regulatory networks. To predict the reprogramming cocktails, the Taiji-reprogram first calculates the ratios of PageRank scores between the target and source cell types and selects the TFs with significant ratios as candidates. Next, the candidate cocktails are selected based on the product of the PageRank ratios of the composing TFs. (**B**) PageRank z-scores of the 745 TFs across 154 cell-types.

cles of arm, back, trunk, leg, forelimb and hindlimb were clustered together. There were several tissues that share high similarity between embryonic and postnatal stages, such as testis and ovary, while most embryonic and postnatal tissues were clustered with the other tissues in the same developmental stage, such as heart and lung tissues. Based on this observation, we analyzed the embryonic and postnatal tissues separately and comparatively.

Identification of TFs important in development and differentiation

The PageRank scores represent the TFs' global importance. Starting from the PageRank score matrix with 745 TFs and 154 cell-types/tissues, we identified the important TFs in development and differentiation. We found 43 and 27 TFs that are constitutively active in the embryonic and postnatal stages, respectively, with an averaged PageRank score ranked within 10% in all TFs and coefficient of variance (CV) <0.5. These TFs are involved in general biological functions, such as metabolic process, biological regulation and cellular process, suggesting the basic roles for these TFs in embryogenesis (Figure 2A). Twenty-two of them are common in the two developmental stages, including proteins such as CTCF and YY1 that have broad functions.

We next identified TFs that are specifically important in either embryonic or postnatal stage using Mann–Whitney *U* test with a *Q*-value cutoff of 0.01 (Figure 2B). The 71 embryo-specific TFs are largely involved in development with enriched GO terms of cellular developmental process, anatomical structure morphogenesis and pattern specification process. For example, Gli2 and Gli3 play essential roles in the development of lung, trachea and esophagus during embryo development (25). In contrast, the 21 postnatal-specific TFs are associated with specialized functions, such as immune system development, response to stress, and response to the chemical. For example, RUNX3 plays important roles in the B-cell proliferation (26,27) and T-cell development (28).

As three germ-layers (ectoderm, endoderm and mesoderm) form during embryonic development and give rise to all the tissues, it is important to identify the germ-layer specific TFs. Each of the 54 embryonic cell types was assigned to a germ layer. We compared the PageRank scores of TFs from one germ layer to the other two using Mann–Whitney *U* test, and defined the germ-layer specific TFs (Figure 2C). Many well-known germ-layer specific TFs were identified, including PAX6 (29,30), ZIC2 (31), ZIC5 (32), SOX2 (33), SOX3 (34), SOX21 (35) and POU3F1 (36) for ectoderm, SNAI1 (37) and MEF2D (38) for mesoderm, and FOXA2 (39), RFX6 (40) for endoderm. Besides the well studied germ-layer specific TFs, we also identified some new TFs whose roles in the germ-layer remain unreported. For instance, FOXF1, expressed in the splanchnic mesoderm, is involved in mesenchymal-epithelial signaling required for development of organs derived from foregut endoderm such as lung, liver, gallbladder and pancreas (41,42). The GO analysis showed that the ectoderm-specific TFs are enriched in the development of tissues derived from ectoderm, such as neural tissues, brain, glial cell, spinal cord and epidermis (Supplementary Table S2). The mesoderm-specific TFs are involved in the development of mesoderm derived tissues, such as kidney, muscle and cardiac ventricle. The endoderm-specific TFs are associated with the development of epithelium and lung. Similarly, we identified the germ-layer specific TFs based on the postnatal cell types in Supplementary Figure S2.

Identification of cell-type and tissue-specific TFs

To identify the tissue/cell-type specific TFs, we first selected the variable TFs with the coefficient of variance (CV) across all cell-types larger than 1, which resulted in 231 TFs for the embryonic group and 289 TFs for the postnatal group. There are 169 TFs in common between the embryonic and postnatal groups. The 231 variable TFs in the embryonic stage are enriched at the central nervous system development, anterior/posterior pattern specification and embryonic organ morphogenesis (Supplementary Table S3). The 289 variable TFs in the postnatal stage are over-represented in the anterior/posterior pattern specification, embryonic organ morphogenesis and epithelium development. The variable TFs in either stage contain many key regulators in the tissue development and morphogenesis.

To identify the cell-type and tissue-specific TFs, we further calculated the *z*-score and *P*-value in each cell-type/tissue assuming each TF's PageRank score across all cell-types following a log-normal distribution. The cell-type specific TFs were identified using a *P*-value cut-off of 0.05. We list the identified specific TFs in 7 common cell-types/tissues in Figure 3 and Supplementary Table S4. We manually searched the literature and found that, 33% of the identified TFs were reported to play key roles in the corresponding cell-type, and 26% of them were associated with corresponding diseases, while the remaining 41% were playing unknown roles. Taking heart as an example, we found 8 heart-specific TFs in the 3 embryonic (Figure 3A) and 20 in the 8 postnatal heart-related cell-types (Figure 3C). Five of them [TBX5 (43), GATA4 (44), HAND2 (45), TBX20 (46) and NKX2-5 (47)] are pivotal to heart development and they were indeed identified in both embryonic and postnatal stages. Among the remaining three TFs (NR4A3, NKX2-6 and RXRG), NKX2-6 and RXRG are specific to the cardiac muscle cells while NR4A3 was found important in the embryonic heart tissue analyzed by ENCODE. In fact, NKX2-6 is related to congenital heart defects (48). NR4A3, also known as NOR-1, can modulate vascular smooth muscle cell proliferation (49) and is related to coronary artery disease (50). There are 398 regulatees of NR4A3 in the regulatory network constructed by Taiji and they are enriched in the regulation of vasculogenesis and cardiac muscle cell action potential (Figure 3B). CACNA1C, a regulatee of NR4A3, encodes cardiac L-type calcium channel (Cav1.2) which is essential for cardiomyocyte action potential duration. Mutations of CACNA1C may cause cardiac arrhythmia syndromes (51). RXRG is not known to function in heart development but it is highly expressed in the embryonic cardiac muscle cells in the ENCODE dataset. Retinoid X receptors (RXRs) play a key role in the formation of the heart (52). As a member of RXRs family, RXRG's importance in cardiac muscle cells is not surprising.

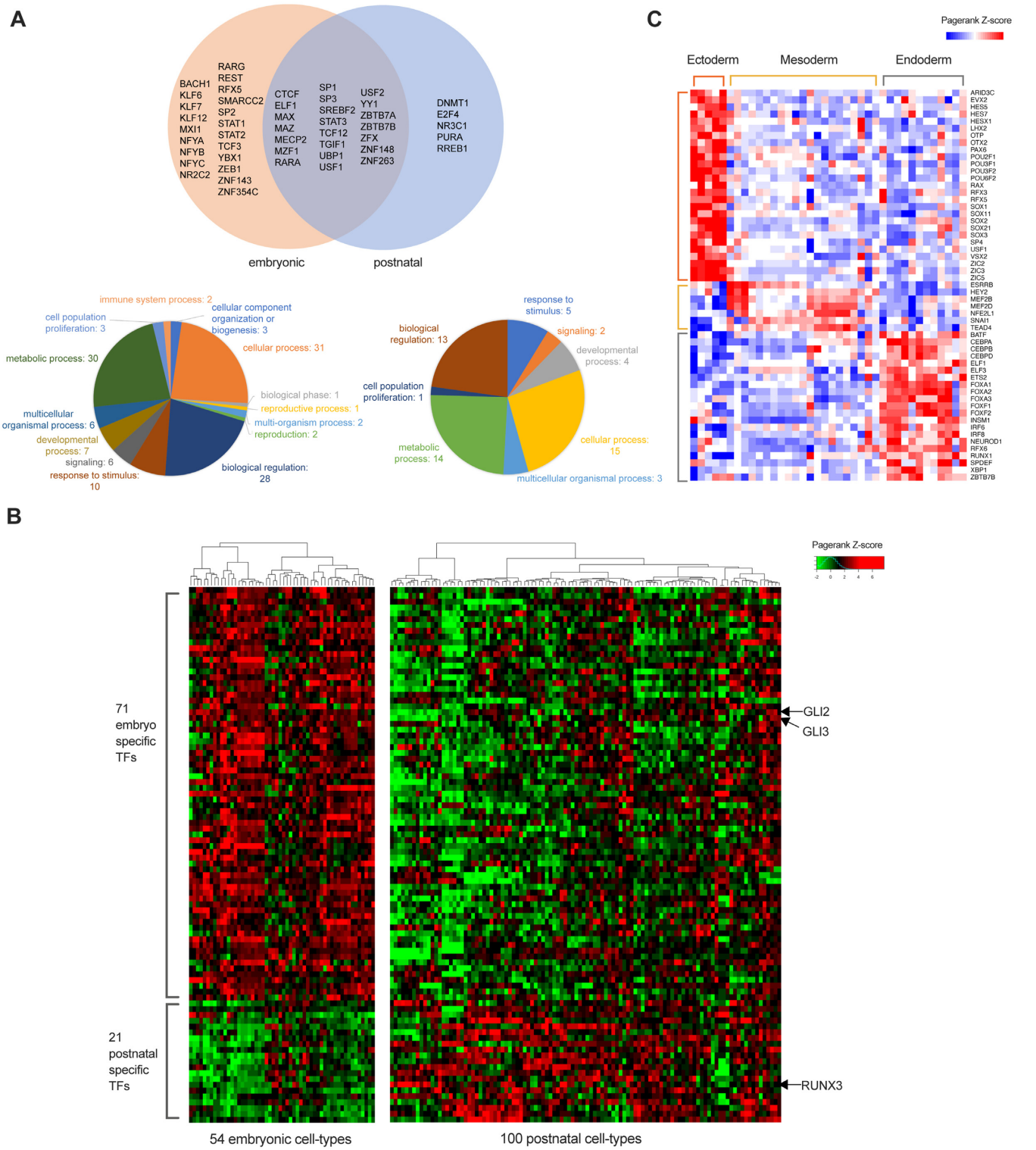

Figure 2. (**A**) There are 43 housekeeping TFs in the embryonic stage, and 27 in the postnatal stage, while 22 of them are in common. The identified housekeeping TFs are enriched mainly in the metabolic process, biological regulation and cellular process for both stages. (**B**) The 71 embryo-specific TFs and 21 postnatal-specific TFs are identified by the Mann–Whitney *U* test. The embryo-specific TFs are over-represented in the biological processes related to development such as cellular developmental process, anatomical structure morphogenesis, pattern specification process. In contrast, the postnatal specific TFs are enriched at the cytokine production, response to stress and response to the chemical. (**C**) Cell-types from the embryonic stage are split by their corresponding germ layers. The germ-layer specific TFs, which were identified for each germ-layer by comparing it with the remaining germ-layers, include many well-known ones, such as PAX6, ZIC2, ZIC5, SOX2, SOX21, SOX3, SOX11 and POU3F1 for ectoderm, SNAI1 and MEF2D for mesoderm, and FOXA2, RFX6 for endoderm.

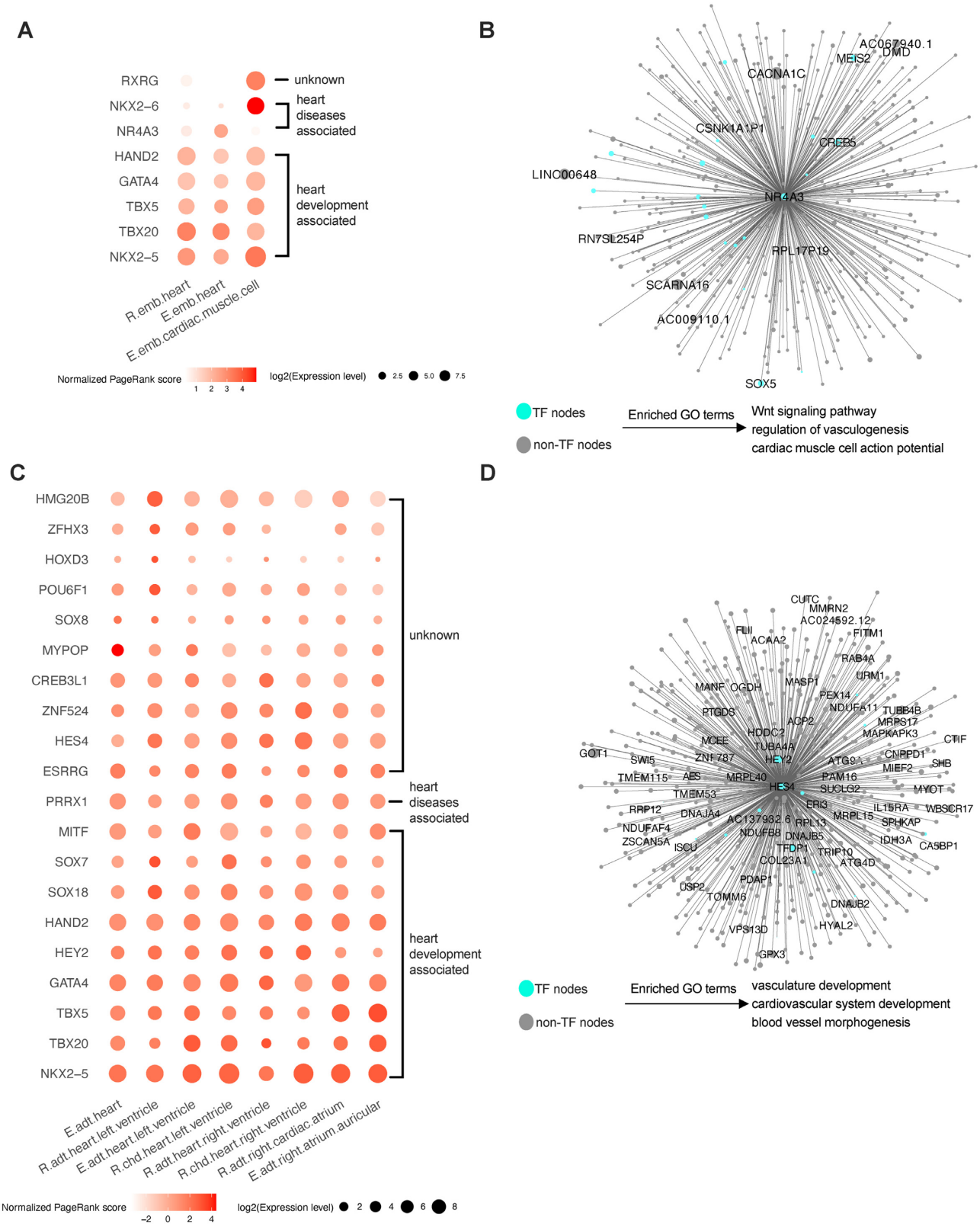

Figure 3. Specifically important TFs in heart-related cell types. (**A**) Eight embryonic-stage-specific TFs. The color codes represent the normalized PageRank scores, while the node size represents the logarithmic expression level (transcripts per million, TPM). Five out of the eight TFs are associated with heart development, and two of them are heart-disease associated. (**B**) The subnetwork for NR4A3, a TF related to heart disease, and its 398 regulatees in the embryonic heart cell types. Larger nodes represent TFs with higher PageRanks. The bottom shows the three enriched Gene Ontology (GO) terms for NR4A3's regulatees. (**C**) Twenty postnatal-stage-specific TFs. Nine of them are known to be related to heart development, and five TFs are also found important in the embryonic stage. (**D**) The subnetwork for HES4. HES4 has not been reported to regulate heart function. Among its regulatees, HEY2 plays an essential role in heart development.

For the 20 identified TFs specific in the postnatal heart-related cell-types (Figure 3C), nine of them [NKX2-5 (47), TBX20 (46), TBX5 (43), GATA4 (44), HEY2 (53), HAND2 (45), SOX18 (54), SOX7 (54) and MITF (55)] have been reported to be important in heart development. NKX2-5, TBX20, TBX5, GATA4 and HAND2 are also identified specific in the embryonic heart cell-types. SOX18 and SOX7 are important regulators of heart muscle differentiation (54). Knockout of SOX17 (closely related to SOX18 and SOX7) in mouse significantly increased the ventricular internal dimension (56). For the PRRX1, the suppression of its expression is associated with increased risk of atrial fibrillation and shortening of the cardiac action potential (57). ESRRG, highly expressed in heart muscle, is an essential transcriptional coordinator of cardiac energy production and consumption (58). Although the role of HES4 in ventricles development has not been reported, the 142 regulatees of HES4 from Taiji's network are enriched for functions of vasculature development, cardiovascular system development and blood vessel morphogenesis (Figure 3D). HES4 also interacts with HEY2 in the genetic network, a TF involved in the cardiovascular system development. HES/HEY genes encode a family of basic helix-loop-helix (bHLH) transcription factors, and they are also the direct targets of the Notch signaling pathway (59). The Notch2, Hey1, and Hey2 initiate a signaling cascade that delimits the non-chamber atrioventricular canal and inner curvature regions (60). The remaining eight TFs (ZNF524, CREB3L1, MYPOP, SOX8, POU6F1, HOXD3, ZFHX3 and HMG20B) had unknown regulatory roles in heart functions but there are some suggestive evidence to support their importance. For example, the orthologous gene of Pou6f1 is pouC in zebrafish whose knockdown impairs cardiac morphogenesis and affects cardiovascular function (61).

Identifying TF cocktails for cell reprogramming

We developed a method based on PageRank scores to predict reprogramming cocktails that can convert one cell type to another. Since the Taiji PageRank score reflects the global importance of a TF in a particular cell type, we first identified the TFs with significantly different PageRank scores in the target and source cells. We calculated the ratio between the larger and the smaller PageRank scores in the target and source cells, i.e. this ratio is always ≥ 1 and a large value indicates that the TFs are much more important in the target (or source) than in the source (or target) cells (see Materials and Methods). We selected the top 30 TFs with largest ratio values as candidates for cell reprogramming factors. We then calculated the product of PageRank ratios of three candidate TFs for all possible combinations. The products were transformed to z-scores and P-value of 0.001 was used as a cut-off to select candidate reprogramming cocktails. Because many 3-TF combinations may have very similar product scores, we considered the TFs that occurred most often in the candidate cocktails as the most promising reprogramming factors. To compare the reprogramming result with other methods, the top 8 TFs with highest PageRank ratios were selected since other methods predicted eight candidate TFs.

We first assessed the prediction performance of our method (Taiji-reprogram) by comparing the predicted reprogramming cocktails with the established ones. Reprogramming from terminally differentiated fibroblast cells to the pluripotent state to generate iPSC has been widely studied and many different reprogramming cocktails have been identified, including the original Yamanaka factors (Pou5f1, Sox2, Klf4 and cMyc) and its variations (Pou5f1, Sox2, Nanog, and Lin28) (62). Pou5f1 (also known as Oct4), Nanog and Sox2 are critical regulators of embryonic stem cells (ESCs). The other factors such as Klf4, cMyc and Lin28 affect the reprogramming efficiency. Among the eight candidate TFs, POU5F1 and SOX2 were the top 2 TFs, while NANOG ranked the 5th. ZSCAN10 was ranked the third in our prediction list, which has been reported to recover genomic stability by normalizing the homeostatic balance of ROS (reactive oxygen species)–glutathione and the DNA damage response (63), supporting its roles in generating iPSC.

We compared Taiji-reprogram with other three popular methods including Mogrify (15), CellNet (14) and D-Alessio *et al*. method (64) on 13 experimentally validated cocktails that convert between cell types included in our analysis. We compared the number of correctly identified candidate TFs and the average rank of the published TFs. For doing this, we first computed the average rank by summing the ranks of all correctly identified TFs and then divided by the total number of correctly identified TFs. The optimal case would be that all published TFs were correctly identified and had the highest ranks in the predicted cocktail. Cocktail score was then calculated by first multiplying the number of correctly predicted TFs with the inverse of average rank and then normalized by setting the optimal case as 100 (see Materials and Methods, Supplementary Figure S3). Using this score, Taiji-reprogram achieved the best performance on 9 out of 13 cocktails while Mogrify performed best on the other four cocktails (Figure 4A).

In one of the validated transdifferentiation cocktails from fibroblast cell to heart cell, both Taiji-reprogram and Mogrify predicted three out of four TFs. Specifically, Taiji-reprogram took the right atrium auricular region tissue from ENCODE as the target cell-type. Taiji-reprogram performed better when comparing the average rank of three TFs (Figure 4B, top). Four (TBX5, HAND2, GATA4, ESRRG) out of eight TFs in the Taiji-reprogram cocktail have been identified as key regulators during the transdifferentiation from fibroblast to heart cells. KLF15 is a new TF identified in our analysis that plays a potential role in the transdifferentiation from fibroblast to heart cells. We extracted the regulatees of the KLF15 in the regulatory network of embryonic heart cells and performed the GO term and KEGG pathway analysis on the 378 TF regulatees. Figure 4E shows the top 30 most significant GO terms based on the P-values, including heart morphogenesis, heart looping and Notch signaling involved in heart development. Among all the identified GO terms, heart development is the most significant one with nearly 30 genes involved, suggesting that a non-negligible portion of regulatees of KLF15 have actively participated in the heart development and differentiation.

Figure 4. (**A**) Performance comparison of the reprogramming cocktails predicted by Taiji and Mogrify. The scores in the axis take account of both count and average ranks of correctly predicted candidate TFs. (**B**) Two examples of predicted reprogramming cocktails of various methods. (Top) cocktail from fibroblast to heart; (bottom) cocktail from fibroblast to liver. (**C**) Fourteen predicted TFs in two reprogramming cocktails from fibroblast to heart and liver. (**D**) Enriched GO term and KEGG pathways of TF regulatees of HNF4G in E.emb.liver. (**E**) Enriched GO term and KEGG pathways of TF regulatees of KLF15 in E.emb.heart.

Another example in which the Taiji-reprogram performed the best is the 'liver_3' cocktail, where liver sample was taken from the right lobe of the liver in ENCODE. While all other three methods only predicted one correct TF HNF4A, Taiji-reprogram identified three TFs: HNF1A, HNF4A and GATA4 (Figure 4B, bottom). HNF4G is also one of the eight candidates in the cocktail from fibroblast to liver cell, which remained poorly investigated before. We performed similar functional analyses of HNF4G's regulatees in the embryonic liver cells and found two items related to liver: liver regeneration and liver development (Figure 4D). Interestingly, heart development also showed up as the enriched GO term in HNF4G's regulatees in the embryonic liver cells. Similarly, liver regeneration is one of the enriched GO terms in the regulatees of KLF15 in the embryonic heart cells. The bubble plot in Figure 4C shows the candidate TFs in the two aforementioned cocktails predicted by Taiji-reprogram with size representing the gene expression level and color representing PageRank score. PROX1 and GATA4 were predicted as candidate TFs in both 'heart_1' and 'liver_3' cocktails and both of them display relatively high expression levels and PageRank scores in the heart cells and liver cells compared to that in the fibroblast cells. The overlap of two groups of candidate TFs suggests that some downstream regulatees of the candidate regulators are in common, which corresponds to the shared GO terms shown in Figure 4D and E.

In addition to the 13 experimental validated cocktails, we also evaluated whether Taiji-reprogram is able to predict somatic barriers identified using shRNA screening in (65). Qin *et al*. combined the genome-wide RNAi screening and Yamanaka factors together in the reprogramming from fibroblast to iPSCs. They reported the log odds and *P*-value of 6072 genes, and 956 of them were identified as potential barriers with P-value < 0.05 (FDR < 0.07). The 6072 reported genes include 249 TFs overlapping with our analyzed TFs (i.e. their motifs are known), and 35 of them with a FDR < 0.07 (i.e. they are considered as somatic barriers). We found that the average PageRank ratio (hESCs versus fibroblast cells) within the FDR < 0.07 group of 35 TFs was 0.85, which is significantly lower than the average PageRank ratio of 1.34 in the FDR ≥ 0.07 group of the remaining 214 TFs (P-value $= 0.013$). A lower PageRank ratio means the TF plays a more important role in the fibroblast cell, which is consistent with the experimental findings.

Furthermore, in addition to the assessment of the somatics barriers, we also validated the Taiji-reprogram on the prediction of TFs which can induce the human pluripotent stem cells (hPSCs) to differentiate and decrease the pluripotency. Ng *et al*. (66) screened 1,564 TFs and found 290 TFs that can induce differentiation of hPSCs. Among the 290 TFs, 125 have known motifs and were included in our Taiji analysis. We calculated the 745 TFs' PageRank ratios in all possible conversions between the 3 hESCs and 151 non-hESC cell-types collected in this study, and selected the top 20 TFs with the highest PageRank ratio (non-hESC/hESC) in each conversion, i.e. these 20 TFs had higher PageRank scores in non-hESC cells than in hESCs. Over-expression of these TFs in hESCs is expected to induce the differentiation of hESCs to the target cells. We pooled together all the top 20 TFs in each conversion and ranked them by their frequency. We selected the top 125 TF with the higher frequency as our prediction of inducing TFs, among which 29 were found in the Ng *et al*. study. The chi-square test showed a P-value of 0.048, which suggests a significant overlap between our predictions and those found by Ng *et al*. (Supplementary Figure S4). Note that Ng *et al*. identified TFs inducing loss of pluripotency but not direct differentiation into a specific cell type, while our analysis found TFs for direct conversion between hESC and a particular differentiated cell type, which is the best mimic of losing pluripotency. Considering this difference, we argue the overlap between the predicted and identified TFs is satisfactory.

DISCUSSION

We present here a comprehensive identification of TFs that are key regulators of deciding cell specificity in diverse human cell types and tissues using a systems biology approach called Taiji. Taiji integrates gene expression and epigenomic data (open chromatin or histone modification) to construct a genetic network. Compared to using protein–protein interaction networks that are not cell-type specific in other methods, Taiji analyzes expression and epigenomic data in the cell type or tissue under consideration, which better represents cell-type/tissue-specific regulatory interactions. Importantly, the PageRank score of a TF naturally considers the impacts of its upstream regulators and downstream regulatees in the network including the feedback from its regulatees. Therefore, the PageRank score represents the global importance of a TF in the genetic network, i.e. the top ranked TFs by PageRank score are master regulators and any perturbation on them would have significant impact on the network.

By comparing the key TFs found in each cell type/tissue, we successfully identified lineage-specific and cell-type/tissue-specific regulators including many well-known key regulators. In particular, the previously unknown TFs were uncovered to be responsible for tissue development and differentiation in the human embryonic and postnatal stages, which can serve as a valuable reference for future mechanistic studies to better understand the regulatory mechanisms of development. Furthermore, given the fast advancement of mapping human tissues using RNA-seq, ATAC-seq and other epigenetic assays, our analysis is readily applicable to these data and can provide a powerful way to integrate gene expression and epigenomic data at the systems level. Importantly, Taiji is applicable to individual data sets and thus expansion to include additional data is straightforward. Comparative analysis on more diverse cell types and tissues would also better define cell/tissue-specific regulatory roles of TFs and further our understanding of the mechanisms underlying cell/tissue specification.

By leveraging the TF PageRank scores' measurement of global importance in individual cell-types, we developed a systematic approach to identifying TF cocktails that can convert one cell type to another. This new approach is straightforward and computationally efficient. By considering the TFs that have the most differential PageRank scores in the target and source cells, we find the candidate regulators whose perturbation would likely facilitate cell con-

version. Given the PageRanks scores, we can easily consider a large number of TF combinations to select the most promising reprogramming cocktails. In particular, because the PageRank scores reflect the global importance of the TFs, we can directly use the product of PageRank scores of the TFs in a cocktail to rank the candidate TF combinations. This way, we avoid the heuristic process to select a set of differentially expressed genes between the target and source cells to score the reprogramming TF combinations. We demonstrated the superior power of this approach on predicting reprogramming cocktails compared to the existing popular methods. The functional analysis of novel TF candidates in reprogramming cocktails indicates the potential roles in the conversion between different cell types, which can guide further experimental investigations. As the ongoing efforts such as Human Cell Atlas aim to measure transcriptome and epigenome in all the cell types of the human body, our approach will be readily applicable to identify key regulators defining the cell types and develop reprogramming cocktails for cell conversion, which will greatly facilitate devising disease models and new cell-based therapeutics.

DATA AVAILABILITY

Codes for the analysis and related source data are available from https://github.com/Wang-lab-UCSD/Taiji-reprogram. Reprogramming cocktails found by Taiji-reprogram can be found at http://wanglab.ucsd.edu:8080/taiji-reprogram/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

NIH [R01AI150282, R01HG009626 (in part)].
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
3. Merrell,A.J. and Stanger,B.Z. (2016) Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. *Nat. Rev. Mol. Cell Biol.*, **17**, 413–425.
4. Jopling,C., Boue,S. and Izpisua Belmonte,J.C. (2011) Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat. Rev. Mol. Cell Biol.*, **12**, 79–89.
5. Takahashi,K. and Yamanaka,S. (2016) A decade of transcription factor-mediated reprogramming to pluripotency. *Nat. Rev. Mol. Cell Biol.*, **17**, 183–193.
6. Robinton,D.A. and Daley,G.Q. (2012) The promise of induced pluripotent stem cells in research and therapy. *Nature*, **481**, 295–305.
7. Li,M. and Izpisua Belmonte,J.C. (2016) Looking to the future following 10 years of induced pluripotent stem cell technologies. *Nat. Protoc.*, **11**, 1579–1585.
8. Wang,H., Yang,Y., Liu,J. and Qian,L. (2021) Direct cell reprogramming: approaches, mechanisms and progress. *Nat. Rev. Mol. Cell Biol.*, **22**, 410–424.
9. Tang,W.W.C., Kobayashi,T., Irie,N., Dietmann,S. and Surani,M.A. (2016) Specification and epigenetic programming of the human germ line. *Nat. Rev. Genet.*, **17**, 585–600.
10. Cantone,I. and Fisher,A.G. (2013) Epigenetic programming and reprogramming during development. *Nat. Struct. Mol. Biol.*, **20**, 282–289.
11. Meissner,A. (2010) Epigenetic modifications in pluripotent and differentiated cells. *Nat. Biotechnol.*, **28**, 1079–1088.
12. Schacht,T., Oswald,M., Eils,R., Eichmüller,S.B. and König,R. (2014) Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*, **30**, i401–7.
13. Arrieta-Ortiz,M.L., Hafemeister,C., Bate,A.R., Chu,T., Greenfield,A., Shuster,B., Barry,S.N., Gallitto,M., Liu,B., Kacmarczyk,T. *et al.* (2015) An experimentally supported model of the bacillus subtilis global transcriptional regulatory network. *Mol. Syst. Biol.*, **11**, 839.
14. Cahan,P., Li,H., Morris,S.A., Lummertz da Rocha,E., Daley,G.Q. and Collins,J.J. (2014) CellNet: network biology applied to stem cell engineering. *Cell*, **158**, 903–915.
15. Rackham,O.J.L., Firas,J., Fang,H., Oates,M.E., Holmes,M.L., Knaupp,A.S. and FANTOM ConsortiumFANTOM Consortium, Suzuki,H., Nefzger,C.M., Daub,C.O. *et al.* (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet.*, **48**, 331–335.
16. Sonawane,A.R., Platig,J., Fagny,M., Chen,C.-Y., Paulson,J.N., Lopes-Ramos,C.M., DeMeo,D.L., Quackenbush,J., Glass,K. and Kuijjer,M.L. (2017) Understanding Tissue-specific Gene Regulatio. *Cell reports*, **21**, 1077–1088.
17. Zhang,K., Wang,M., Zhao,Y. and Wang,W. (2019) Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development. *Sci Adv*, **5**, eaav3262.
18. Yu,B., Zhang,K., Milner,J.J., Toma,C., Chen,R., Scott-Browne,J.P., Pereira,R.M., Crotty,S., Chang,J.T., Pipkin,M.E. *et al.* (2017) Erratum: epigenetic landscapes reveal transcription factors that regulate CD8 t cell differentiation. *Nat. Immunol.*, **18**, 705.
19. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
20. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
21. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
22. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
23. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
24. Yu,B., Zhang,K., Milner,J.J., Toma,C., Chen,R., Scott-Browne,J.P., Pereira,R.M., Crotty,S., Chang,J.T., Pipkin,M.E. *et al.* (2017) Epigenetic landscapes reveal transcription factors that regulate CD8 t cell differentiation. *Nat. Immunol.*, **18**, 573–582.
25. Motoyama,J., Liu,J., Mo,R., Ding,Q., Post,M. and Hui,C.C. (1998) Essential function of gli2 and gli3 in the formation of lung, trachea and oesophagus. *Nat. Genet.*, **20**, 54–57.
26. Whiteman,H.J. and Farrell,P.J. (2006) RUNX expression and function in human b cells. *Crit. Rev.*, **16**, 31–44.
27. Boto,P., Csuth,T.I. and Szatmari,I. (2018) RUNX3-Mediated immune cell development and maturation. *Crit. Rev. Immunol.*, **38**, 63–78.
28. Milner,J.J., Toma,C., Yu,B., Zhang,K., Omilusik,K., Phan,A.T., Wang,D., Getzler,A.J., Nguyen,T., Crotty,S. *et al.* (2018) Erratum: runx3 programs CD8 t cell residency in non-lymphoid tissues and tumours. *Nature*, **554**, 392.
29. Zhang,X., Huang,C.T., Chen,J., Pankratz,M.T., Xi,J., Li,J., Yang,Y., Lavaute,T.M., Li,X.-J., Ayala,M. *et al.* (2010) Pax6 is a human neuroectoderm cell fate determinant. *Cell Stem Cell*, **7**, 90–100.

30. Dimanlig,P.V., Faber,S.C., Auerbach,W., Makarenkova,H.P. and Lang,R.A. (2001) The upstream ectoderm enhancer in pax6 has an important role in lens induction. *Development*, **128**, 4415–4424.
31. Elms,P., Siggers,P., Napper,D., Greenfield,A. and Arkell,R. (2003) Zic2 is required for neural crest formation and hindbrain patterning during mouse development. *Dev. Biol.*, **264**, 391–406.
32. Nakata,K., Koyabu,Y., Aruga,J. and Mikoshiba,K. (2000) A novel member of the xenopus zic family, zic5, mediates neural crest development. *Mech. Dev.*, **99**, 83–91.
33. Zhang,S. and Cui,W. (2014) Sox2, a key factor in the regulation of pluripotency and neural differentiation. *World J. Stem Cells*, **6**, 305–311.
34. Dee,C.T., Hirst,C.S., Shih,Y.-H., Tripathi,V.B., Patient,R.K. and Scotting,P.J. (2008) Sox3 regulates both neural fate and differentiation in the zebrafish ectoderm. *Dev. Biol.*, **320**, 289–301.
35. Whittington,N., Cunningham,D., Le,T.-K., De Maria,D. and Silva,E.M. (2015) Sox21 regulates the progression of neuronal differentiation in a dose-dependent manner. *Dev. Biol.*, **397**, 237–247.
36. Zhu,Q., Song,L., Peng,G., Sun,N., Chen,J., Zhang,T., Sheng,N., Tang,W., Qian,C., Qiao,Y. *et al.* (2014) The transcription factor pou3f1 promotes neural fate commitment via activation of neural lineage genes and inhibition of external signaling pathways. *Elife*, **3**, e02224.
37. Carver,E.A., Jiang,R., Lan,Y., Oram,K.F. and Gridley,T. (2001) The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. *Mol. Cell. Biol.*, **21**, 8184–8188.
38. Kolpakova,A., Katz,S., Keren,A., Rojtblat,A. and Bengal,E. (2013) Transcriptional regulation of mesoderm genes by MEF2D during early xenopus development. *PLoS One*, **8**, e69693.
39. Burtscher,I. and Lickert,H. (2009) Foxa2 regulates polarity and epithelialization in the endoderm germ layer of the mouse embryo. *Development*, **136**, 1029–1038.
40. Pearl,E.J., Jarikji,Z. and Horb,M.E. (2011) Functional analysis of rfx6 and mutant variants associated with neonatal diabetes. *Dev. Biol.*, **351**, 135–145.
41. Mahlapuu,M., Enerbäck,S. and Carlsson,P. (2001) Haploinsufficiency of the forkhead gene foxf1, a target for sonic hedgehog signaling, causes lung and foregut malformations. *Development*, **128**, 2397–2406.
42. Kalinichenko,V.V., Zhou,Y., Bhattacharyya,D., Kim,W., Shin,B., Bambal,K. and Costa,R.H. (2002) Haploinsufficiency of the mouse forkhead box f1 gene causes defects in gall bladder development. *J. Biol. Chem.*, **277**, 12369–12374.
43. Steimle,J.D. and Moskowitz,I.P. (2017) TBX5: a key regulator of heart development. *Curr. Top. Dev. Biol.*, **122**, 195–221.
44. Garg,V., Kathiriya,I.S., Barnes,R., Schluterman,M.K., King,I.N., Butler,C.A., Rothrock,C.R., Eapen,R.S., Hirayama-Yamada,K., Joo,K. *et al.* (2003) GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature*, **424**, 443–447.
45. McFadden,D.G. (2004) The hand1 and hand2 transcription factors regulate expansion of the embryonic cardiac ventricles in a gene dosage-dependent manner. *Development*, **132**, 189–201.
46. Brown,D.D. (2005) Tbx5 and tbx20 act synergistically to control vertebrate heart morphogenesis. *Development*, **132**, 553–563.
47. Zhang,L., Nomura-Kitabayashi,A., Sultana,N., Cai,W., Cai,X., Moon,A.M. and Cai,C.-L. (2014) Mesodermal nkx2.5 is necessary and sufficient for early second heart field development. *Dev. Biol.*, **390**, 68–79.
48. Li,T., Liu,C., Xu,Y., Guo,Q., Chen,S., Sun,K. and Xu,R. (2016) Identification of candidate genes for congenital heart defects on proximal chromosome 8p. *Sci. Rep.*, **6**, 36133.
49. Martínez-González,J., Rius,J., Castelló,A., Cases-Langhoff,C. and Badimon,L. (2003) Neuron-derived orphan receptor-1 (NOR-1) modulates vascular smooth muscle cell proliferation. *Circ. Res.*, **92**, 96–103.
50. Rodríguez-Calvo,R., Guadall,A., Calvayrac,O., Navarro,M.A., Alonso,J., Ferrán,B., de Diego,A., Muniesa,P., Osada,J., Rodríguez,C. *et al.* (2013) Over-expression of neuron-derived orphan receptor-1 (NOR-1) exacerbates neointimal hyperplasia after vascular injury. *Hum. Mol. Genet.*, **22**, 1949–1959.
51. Betzenhauser,M., Pitt,G. and Antzelevitch,C. (2015) Calcium channel mutations in cardiac arrhythmia syndromes. *Curr. Mol. Pharmacol.*, **8**, 133–142.
52. Stefanovic,S. and Zaffran,S. (2017) Mechanisms of retinoic acid signaling during cardiogenesis. *Mech. Dev.*, **143**, 9–19.
53. Koibuchi,N. and Chin,M.T. (2007) CHF1/Hey2 plays a pivotal role in left ventricular maturation through suppression of ectopic atrial gene expression. *Circ. Res.*, **100**, 850–855.
54. Afouda,B.A., Lynch,A.T., de Paiva Alves,E. and Hoppler,S. (2018) Genome-wide transcriptomics analysis identifies sox7 and sox18 as specifically regulated by gata4 in cardiomyogenesis. *Dev. Biol.*, **434**, 108–120.
55. Tshori,S., Gilon,D., Beeri,R., Nechushtan,H., Kaluzhny,D., Pikarsky,E. and Razin,E. (2006) Transcription factor MITF regulates cardiac growth and hypertrophy. *J. Clin. Invest.*, **116**, 2673–2681.
56. Lange,A.W., Haitchi,H.M., LeCras,T.D., Sridharan,A., Xu,Y., Wert,S.E., James,J., Udell,N., Thurner,P.J. and Whitsett,J.A. (2014) Sox17 is required for normal pulmonary vascular morphogenesis. *Dev. Biol.*, **387**, 109–120.
57. Tucker,N.R., Dolmatova,E.V., Lin,H., Cooper,R.R., Ye,J., Hucker,W.J., Jameson,H.S., Parsons,V.A., Weng,L.-C., Mills,R.W. *et al.* (2017) Diminished PRRX1 expression is associated with increased risk of atrial fibrillation and shortening of the cardiac action potential. *Circulation*, **10**, e001902.
58. Wang,T., McDonald,C., Petrenko,N.B., Leblanc,M., Wang,T., Giguere,V., Evans,R.M., Patel,V.V. and Pei,L. (2015) Estrogen-related receptor α (ERRα) and ERRγ are essential coordinators of cardiac metabolism and function. *Mol. Cell. Biol.*, **35**, 1281–1298.
59. Zhou,M., Yan,J., Ma,Z., Zhou,Y., Abbood,N.N., Liu,J., Su,L., Jia,H. and Guo,A.-Y. (2012) Comparative and evolutionary analysis of the HES/HEY gene family reveal exon/intron loss and teleost specific duplication events. *PLoS One*, **7**, e40649.
60. Rutenberg,J.B., Fischer,A., Jia,H., Gessler,M., Zhong,T.P. and Mercola,M. (2006) Developmental patterning of the cardiac atrioventricular canal by notch and Hairy-related transcription factors. *Development*, **133**, 4381–4390.
61. Bhakta,M., Padanad,M.S., Harris,J.P., Lubczyk,C., Amatruda,J.F. and Munshi,N.V. (2018) pouC regulates expression of bmp4 during atrioventricular canal formation in zebrafish. *Dev. Dyn.*, **248**, 173–188.
62. Yu,J. (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science*, **318**, 1917–1920.
63. Skamagki,M., Correia,C., Yeung,P., Baslan,T., Beck,S., Zhang,C., Ross,C.A., Dang,L., Liu,Z., Giunta,S. *et al.* (2017) ZSCAN10 expression corrects the genomic instability of iPSCs from aged donors. *Nat. Cell Biol.*, **19**, 1037–1048.
64. D'Alessio,A.C., Fan,Z.P., Wert,K.J., Baranov,P., Cohen,M.A., Saini,J.S., Cohick,E., Charniga,C., Dadon,D., Hannett,N.M. *et al.* (2015) A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep.*, **5**, 763–775.
65. Qin,H., Diaz,A., Blouin,L., Lebbink,R.J., Patena,W., Tanbun,P., LeProust,E.M., McManus,M.T., Song,J.S. and Ramalho-Santos,M. (2014) Systematic identification of barriers to human iPSC generation. *Cell*, **158**, 449–461.
66. Ng,A.H.M., Khoshakhlagh,P., Rojo Arias,J.E., Pasquini,G., Wang,K., Swiersy,A., Shipman,S.L., Appleton,E., Kiaee,K., Kohman,R.E. *et al.* (2021) A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.*, **39**, 510–519.