Stem Cell Reports
Article
ISSCR
OPEN ACCESS

Gene regulatory network reconfiguration in direct lineage reprogramming

Kenji Kamimoto,[1,2,3] Mohd Tayyab Adil,[1,2,3] Kunal Jindal,[1,2,3] Christy M. Hoffmann,[1,2,3] Wenjun Kong,[1,2,3,4] Xue Yang,[1,2,3] and Samantha A. Morris[1,2,3,*]

[1]Department of Developmental Biology, Washington University School of Medicine in St. Louis, 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA
[2]Department of Genetics, Washington University School of Medicine in St. Louis, 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA
[3]Center of Regenerative Medicine, Washington University School of Medicine in St. Louis, 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA
[4]Present address: Calico Life Sciences LLC, South San Francisco, CA, 94080, USA
*Correspondence: s.morris@wustl.edu
https://doi.org/10.1016/j.stemcr.2022.11.010

SUMMARY

In direct lineage conversion, transcription factor (TF) overexpression reconfigures gene regulatory networks (GRNs) to reprogram cell identity. We previously developed CellOracle, a computational method to infer GRNs from single-cell transcriptome and epigenome data. Using inferred GRNs, CellOracle simulates gene expression changes in response to TF perturbation, enabling *in silico* interrogation of network reconfiguration. Here, we combine CellOracle analysis with lineage tracing of fibroblast to induced endoderm progenitor (iEP) conversion, a prototypical direct reprogramming paradigm. By linking early network state to reprogramming outcome, we reveal distinct network configurations underlying successful and failed fate conversion. Via *in silico* simulation of TF perturbation, we identify new factors to coax cells into successfully converting their identity, uncovering a central role for the AP-1 subunit Fos with the Hippo signaling effector, Yap1. Together, these results demonstrate the efficacy of CellOracle to infer and interpret cell-type-specific GRN configurations, providing new mechanistic insights into lineage reprogramming.

INTRODUCTION

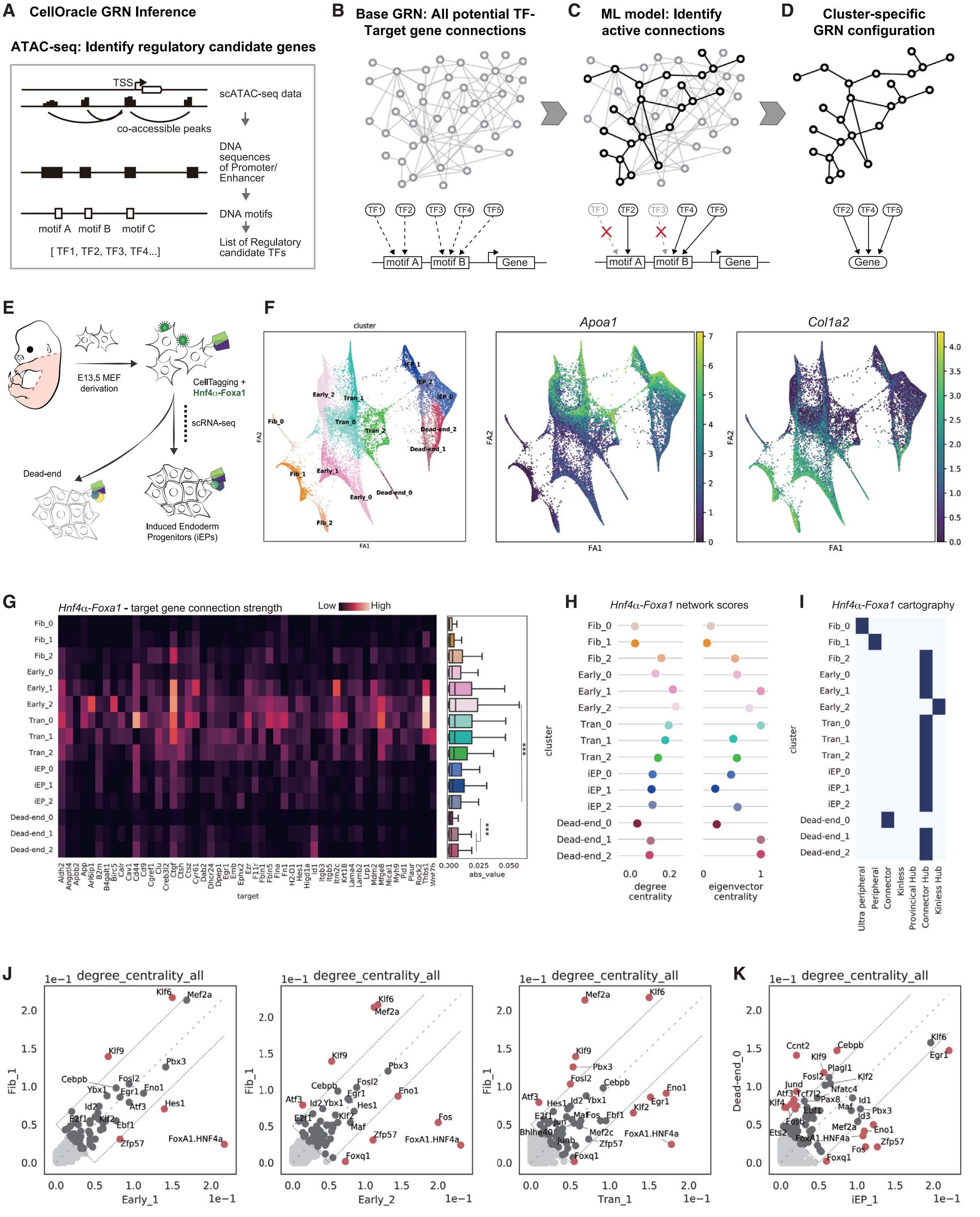
Direct lineage reprogramming aims to transform cell identity between fully differentiated somatic states via the forced expression of select transcription factors (TFs). Using this approach, fibroblasts have been directly converted into many clinically valuable cell types (Cohen and Melton, 2011). These protocols are currently limited, though, because only a fraction of cells convert to the target cell type and remain developmentally immature or incompletely specified (Morris and Daley, 2013). Therefore, the resulting cells are generally unsuitable for therapeutic application and have limited utility for disease modeling and drug screening *in vitro*.

A comprehensive characterization of cell identity is crucial to improve reprogramming methods. Gene regulatory networks (GRNs) represent the complex, dynamic molecular interactions that act as critical determinants of cell identity. These networks describe the intricate interplay between transcriptional regulators and multiple *cis*-regulatory DNA sequences, resulting in the precise spatial and temporal regulation of gene expression (Davidson and Erwin, 2006). Systematically delineating GRN structures enables a logic map of regulatory factor cause-effect relationships to be mapped. In turn, this knowledge supports a better understanding of how cell identity is determined and maintained, informing new strategies for cellular reprogramming.

We previously described CellOracle, a computational pipeline for GRN inference via integrating different single-cell data modalities (Kamimoto et al., 2020). CellOracle overcomes current challenges in GRN inference by using single-cell transcriptomic and chromatin accessibility profiles, integrating prior biological knowledge via regulatory sequence analysis to infer TF-target gene interactions. We designed CellOracle to apply inferred GRNs to simulate gene expression changes in response to TF perturbation. This unique feature enables inferred GRN configurations to be interrogated *in silico*, facilitating their interpretation. We have benchmarked CellOracle against ground-truth TF-gene interactions, demonstrating its efficacy to recapitulate known regulatory changes across hematopoiesis (Kamimoto et al., 2020). Further, we have applied CellOracle to predict TFs regulating medium spiny neuron maturation in human fetal striatum development (Bocchi et al., 2021). Other groups have successfully used the method to investigate mouse and human T cell differentiation (Chopp et al., 2020; Nie et al., 2022), T cell dysfunction in glioblastoma (Ravi et al., 2022), and pharyngeal organ development (Magaletta et al., 2022).

Here, we apply CellOracle to interrogate GRN reconfiguration during direct lineage reprogramming of fibroblasts to induced endoderm progenitors (iEPs), a prototypical TF-mediated fate conversion. Via single-cell lineage tracing, we previously demonstrated that this protocol comprises two distinct trajectories leading to reprogrammed and dead-end fates (Biddy et al., 2018). We expand on this lineage tracing strategy to experimentally define state-fate relationships, supporting the inference of early network states

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).



(legend on next page)

associated with defined reprogramming outcomes. These analyses reveal the early GRN configurations associated with the successful conversion of cell identity. Using principles of graph theory to identify critical regulatory nodes in conjunction with *in silico* simulation predicts several novel regulators of reprogramming, which we experimentally validate. We also demonstrate that one of these TFs, *Fos*, plays roles in both iEP reprogramming and maintenance, where interrogation of inferred *Fos* targets reveals a role for AP1-Yap1. We validate these findings to demonstrate that Fos and Yap1 overexpression significantly enhances reprogramming efficiency. Together, these results demonstrate the efficacy of CellOracle to infer and interpret cell-type-specific GRN configurations at high resolution, enabling new mechanistic insights into reprogramming. CellOracle code and documentation are available at https://github.com/morris-lab/CellOracle.

RESULTS

CellOracle GRN inference applied to direct lineage reprogramming

CellOracle is designed to infer GRN configurations, revealing how networks are rewired during the establishment of defined cellular identities and states, highlighting known and putative regulatory factors of fate commitment (Kamimoto et al., 2020). In the first step of the CellOracle pipeline, single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) is used to assemble a "base" GRN structure, representing a list of all potential regulatory genes associated with each defined DNA sequence (Figures 1A and 1B). The second step in the CellOracle pipeline uses single-cell RNA sequencing (scRNA-seq) data to convert the base GRN into context-dependent GRN configurations for each defined cell cluster. Removal of inactive connections refines the base GRN structure, selecting the active edges that represent regulatory connections associated with a specific cell type or state (Figures 1C, 1D, and S1A). Here, we apply CellOracle to infer GRN reconfiguration during TF-mediated direct lineage reprogramming.

The generation of induced endoderm progenitors (iEPs) from mouse embryonic fibroblasts (MEFs) represents a prototypical lineage reprogramming protocol, which, like most conversion strategies, is inefficient and lacks fidelity. Initially reported as hepatocyte-like cells that functionally engraft the liver (Sekiya and Suzuki, 2011), we demonstrated that these cells also harbor intestinal identity and can functionally engraft the colon, prompting their re-designation as iEPs (Guo et al., 2019; Morris et al., 2014). More recently, we have shown that iEPs transcriptionally resemble injured biliary epithelial cells (BECs) and exhibit BEC-like behavior in 3D-culture models (Kong et al., 2022). Building on these findings, our single-cell lineage tracing revealed two distinct trajectories: one to a successfully reprogrammed iEP state, and one to a dead-end, mesenchymal-like state (Figure 1E; Biddy et al., 2018).

Our previously published MEF to iEP reprogramming scRNA-seq dataset consists of eight time points collected over 28 days (n = 27,663 cells) (Biddy et al., 2018). We reprocessed this dataset using partition-based graph abstraction (PAGA; Wolf et al., 2019), manually annotating 15 clusters based on marker gene expression, identifying the expected trajectories (Figures 1F and S1B–S1D). Relative to reprogrammed cells, dead-end cells only weakly express iEP markers, *Cdh1* and *Apoa1*, accompanied by higher expression levels of fibroblast markers, such as *Col1a2* (Figures 1F, S1B, and S1C). Using CellOracle with a base GRN generated using a mouse scATAC-seq atlas (Cusanovich et al., 2018), we inferred GRN configurations for each cluster, calculating network connectivity scores to analyze GRN dynamics during reprogramming.

Analysis of network reconfiguration during reprogramming

We initially assess the network configuration associated with the exogenous reprogramming TFs, *Hnf4α* and *Foxa1*,

Figure 1. Application of CellOracle to assess reprogramming GRN dynamics

(A and B) Overview of CellOracle. (A) First, CellOracle uses scATAC-seq data to identify accessible regulatory elements, which are scanned for TF binding motifs, generating a Base GRN—a list of potential regulatory connections between a TF and its target genes (B).

(C) Using single-cell expression data, active connections are identified from all potential connections in the base GRN.

(D) Cell type- and state-specific GRN configurations are constructed by pruning insignificant or weak connections.

(E) Hnf4α and Foxa1-mediated fibroblast to iEP reprogramming.

(F) (Left) Force-directed graph: 15 clusters of cells are grouped into five cell types; fibroblasts (Fib), early transition (Early), transition (Tran), dead-end, and reprogrammed iEPs (iEP). (Right) Projection of *Apoa1* (iEP marker) and *Col1a2* (fibroblast marker) expression.

(G) CellOracle analysis. Heatmap (left) and boxplot (right) of network edge strength between *Hnf4α-Foxa1* and its target genes. ***$p < 0.001$.

(H) Degree and eigenvector centrality scores for *Hnf4α-Foxa1*.

(I) *Hnf4α-Foxa1* network cartography terms for each cluster.

(J and K) Scatterplots of degree centrality scores between specific clusters.

(J) Degree centrality score comparison between Fib_1 cluster GRN and other early and transition reprogramming cluster GRNs.

(K) Degree centrality score comparison between iEP_1 and Dead-end_0 cluster GRNs.

focusing on the strength of their connections to target genes. *Hnf4α* and *Foxa1* receive a combined score in these analyses since they are expressed as a single transcript that produces two independent factors via 2A-peptide-mediated cleavage. Network strength scores show significantly stronger connectivity of *Hnf4α-Foxa1* to its inferred target genes in early reprogramming, followed by decreasing connection strength (Early_2 versus iEP_2: $p < 0.001$, Wilcoxon test; Figure 1G). We next evaluated the inferred GRN structures using traditional graph theory methods. We examined (1) degree centrality of each gene, a straightforward measure reporting how many edges are directly connected to a node; and (2) eigenvector centrality, a measure of influence via connectivity to other well-connected genes (Klein et al., 2012). *Hnf4α-Foxa1* receives high degree and eigenvector centrality scores in the early conversion stages, gradually decreasing as reprogramming progress (Figure 1H). In agreement with a central role for the transgenes early in reprogramming, network cartography analysis (Guimerà and Amaral, 2005) classified *Hnf4α-Foxa1* as a prominent "connector hub" in the early_2 cluster network configuration (Figures 1I and S1E). Together, these analyses show that *Hnf4α-Foxa1* network configuration connectivity and strength peak in early reprogramming phases.

Next, we analyzed the *Hnf4α-Foxa1* network configuration in later conversion, following bifurcation into reprogrammed and dead-end trajectories (Figures 1F and S1B–S1D). The reprogrammed clusters (iEP_0, iEP_1, iEP_2) exhibit stronger network connectivity scores relative to the dead-end clusters 1 and 2 (Figure 1G; iEP versus dead-end; $p < 0.001$, Wilcoxon test). We also identify a smaller dead-end cluster (Dead-end_0); cells within this cluster only weakly initiate reprogramming, retaining robust fibroblast gene expression signatures and expressing significantly lower levels of reprogramming initiation markers such as *Apoa1* (Figure S1C; $p < 0.001$, permutation test). This cluster also exhibits significantly lower *Hnf4α-Foxa1* connectivity scores relative to Dead-end_1 and 2 (Figure 1G; $p < 0.001$, Wilcoxon test), accompanied by lower degree centrality and eigenvector centrality scores (Figure 1H). However, CellTag lineage data reveal that most cells (93% of tracked cells) on this unique path derive from a single clone, representing a rare reprogramming event captured due to clonal expansion (Figure S1F).

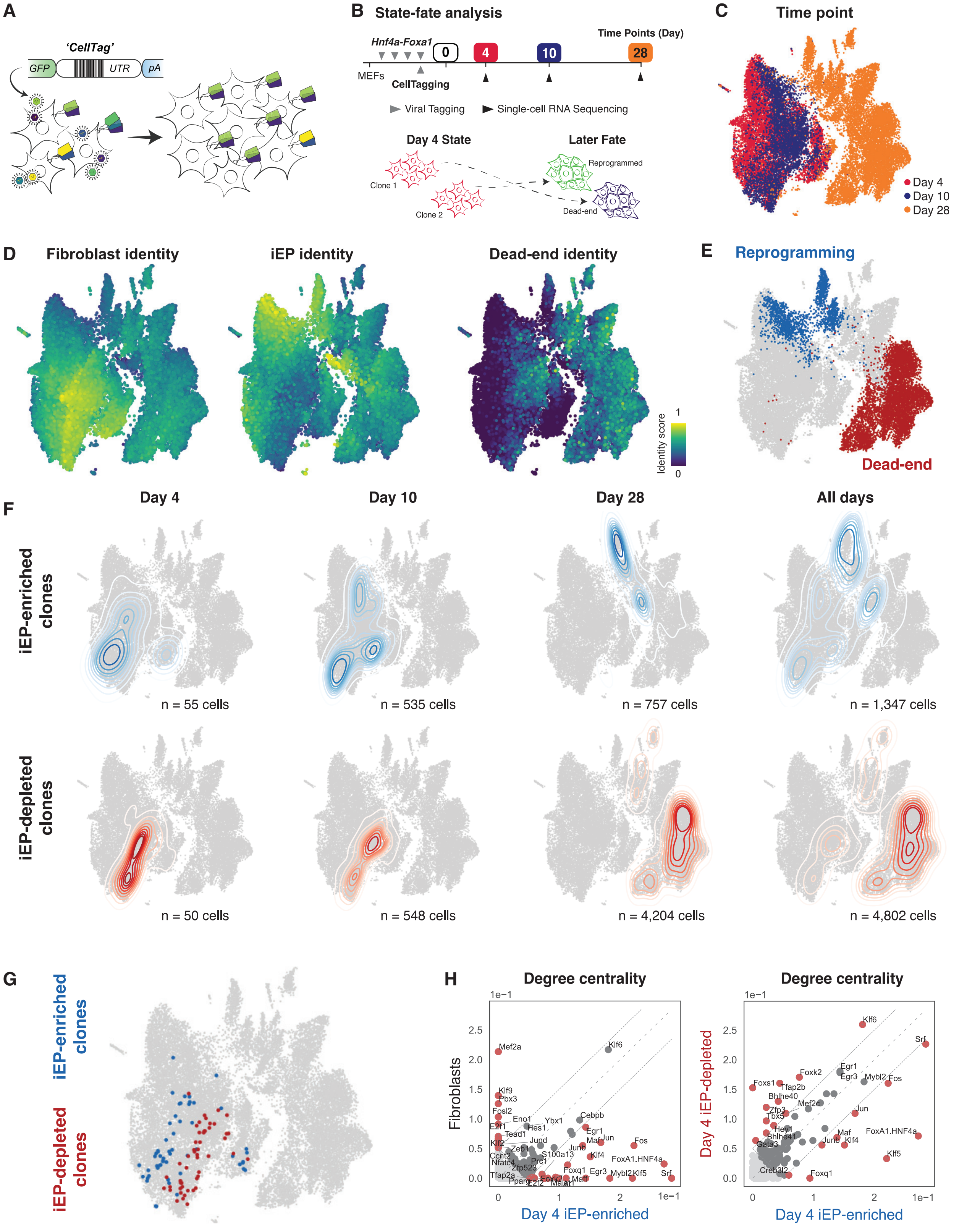
We next turned to global GRN reconfiguration to identify candidate TFs initiating reprogramming. Comparing degree centrality scores between fibroblast and early reprogramming clusters reveals differential connectivity of a handful of key TFs. For example, *Hes1*, *Eno1*, *Fos*, *Foxq1*, and *Zfp57* receive relatively high degree centrality scores in the early reprogramming clusters, whereas *Klf2* and *Egr1* degree centrality increases in later transition stages (Figure 1J). These factors remain highly connected on the reprogramming trajectory relative to the dead-end (Figure 1K), suggesting that the GRN configurations controlling reprogramming outcome are remodeled at the initiation of fate conversion.

Altogether, reprogramming network analysis suggests that *Hnf4α-Foxa1* function peaks at conversion initiation. These early, critical changes in GRN configuration determine reprogramming outcome, with dysregulation or loss of this program leading to dead-ends, where cells either do not successfully initiate or complete reprogramming. This hypothesis is consistent with our previous CellTag lineage tracing, showing the establishment of reprogramming outcomes from early stages of the conversion process (Biddy et al., 2018). We next performed new experimental lineage tracing targeting cells at reprogramming initiation to further investigate how early GRN configuration relates to the successful generation of iEPs.

Clonal tracing links early network state to reprogramming fate

Barcoding and tracking cells via scRNA-seq represents a powerful method to investigate how the early molecular state of a cell relates to its eventual fate (Biddy et al., 2018; Jindal et al., 2022; Weinreb et al., 2020). Cells are labeled with combinations of heritable random barcodes, CellTags, delivered using lentivirus, enabling cells to be uniquely labeled and tracked over time; cells sharing identical barcodes are identified as clonal relatives; thus, early cell state can be directly linked to reprogramming outcome (Biddy et al., 2018; Kong et al., 2020; Figure 2A). However, our previous lineage tracing study was not designed to maximize the capture of clones early in reprogramming; thus, we did not meet the minimum cell number required for accurate GRN inference (50 cells; Kamimoto et al., 2020). Here, we performed new lineage tracing experiments to associate early-stage cells with reprogramming outcome.

Cells were reprogrammed with *Hnf4α-Foxa1*, as above, and CellTagged at the end of the reprogramming TF transduction period. After 4 days of expansion (reprogramming day 4), we collected 25% of the cell population for scRNA-seq, reseeding the remaining cells. A total of 24,799 cells were sequenced: 8,440 on day 4, 4,836 on day 10, and 11,523 on day 28 (Figures 2B and 2C). Using our previous method to score cell identity along with established marker gene expression (Biddy et al., 2018), we identify reprogrammed and dead-end fates (reprogrammed n = 1,895; dead-end n = 6,324; Figures 2D, S2A, and S2B). Next, using clonal information, we identify the day 4 clones whose day 10 and day 28 descendants are significantly enriched or depleted of successfully reprogrammed cells. From CellTag processing (supplemental experimental procedures), we recovered 1,158 clones, containing a total of



(legend on next page)

10,927 cells across all time points. Using randomized testing, we identified two groups of day 4 cells: iEP-enriched (55 cells in nine clones) and iEP-depleted (50 cells in 43 clones), from which reprogramming and dead-end trajectories stem (Figures 2F and 2G), reproducing our earlier observations (Biddy et al., 2018).

Pooling the day 4 clones by outcome, we meet the minimum number of cells required for GRN inference (Figure S2C). We first compared the global GRN configurations for each of these states relative to MEFs to assess early GRN reconfiguration on each trajectory. For example, comparing degree centrality between day 4 cells destined to reprogram and native fibroblasts agrees with our above analysis comparing early transition to fibroblast states (Figure 1J), showing high connectivity of similar factors, such as *Klf9*, and *Mef2a*, in fibroblasts and *Fos* and *Foxq1* in day 4 reprogrammed-destined clones (Figure 2H, left). Additional highly connected TFs also emerge in this reprogramming group, including the known induced pluripotency factor, *Klf4* (Takahashi and Yamanaka, 2006), and *Klf5*, *Mybl2*, and *Foxk2*. The appearance of several additional factors here is likely due to assessing the early cells with known reprogramming descendants rather than the early reprogramming cluster as a whole, in which many cells will not successfully reprogram, highlighting how these state-fate experiments can further dissect population heterogeneity.

Indeed, the state-fate experimental design allows us to compare those early cells destined to reprogram versus early cells that fail to reprogram, for which clonal information is essential. A comparison of these two groups reveals subtle differences in GRN configuration, with *Klf6*, *Tbx5*, *Tfapb2*, and *Foxs1* demonstrating higher connectivity in cells failing to reprogram, in contrast to *Fos*, *Klf5*, and *Junb* in cells destined to attain full iEP identity (Figure 2H, right). Differential expression analysis between day 4 reprogramming and dead-end groups did not identify these TFs (Table S3). CoSpar, a computational tool designed to identify lineage-specific gene markers based on single-cell lineage tracing data (Wang et al., 2022b), identified only *Foxs1* and *Junb*. Overall, this new experimental state-fate analysis reveals the highly connected fibroblast TFs decoupled upon reprogramming initiation, representing potential targets to extinguish fibroblast identity. Further, we identify many TFs that are highly connected early on the successful reprogramming trajectory, representing potential candidates to improve iEP yield. We next use CellOracle's *in silico* perturbation function to identify putative regulators of reprogramming in a systematic, unbiased manner.

Systematic *in silico* simulation of TF knockout to identify novel regulators of iEP reprogramming

While network structure can point to how gene regulation changes during reprogramming, it offers a static picture that does not necessarily provide functional insight. CellOracle bridges this gap by using its unique GRN inference model to interrogate networks to gain mechanistic insight into how specific TFs regulate cell identity (Kamimoto et al., 2020). CellOracle simulates the transition of cell identity following candidate TF perturbation (knockout [KO] or overexpression), using cluster-specific GRNs to model subsequent expression changes in regulated genes. The simulated values are then converted into a transition vector map and visualized in the dimensional reduction space, enabling an intuitive interpretation of how a candidate TF regulates cell identity (Kamimoto et al., 2020); Figures 3A–3C and S3A–S3C; supplemental experimental procedures).

In silico TF perturbation comprises four steps: (1) GRN configurations are constructed (as in Figure 1A). (2) Using these GRN models, shifts in target gene expression in response to TF perturbation are calculated. This step applies the GRN model as a function to propagate the shift in gene expression rather than the absolute gene expression value,

Figure 2. Lineage tracing links early network state to reprogramming outcome

(A) Overview of CellTag-based clonal tracking. Cells are transduced with the random CellTag lentiviral library so that each cell expresses three to four CellTags, resulting in a unique, heritable barcode signature. CellTags are transcribed and captured during single-cell profiling, enabling clonally related cells to be tracked throughout an experiment.

(B) Experimental strategy to capture state-fate relationships. MEFs are transduced with Hnf4α-Foxa1 for 48 h, then transduced with CellTags. The end of this period is considered reprogramming day 0. Cells are expanded, and 25% of the population is profiled at day 4; this is termed the state population. The remaining cells are reseeded and profiled again on days 10 and 28 to capture reprogramming fate.

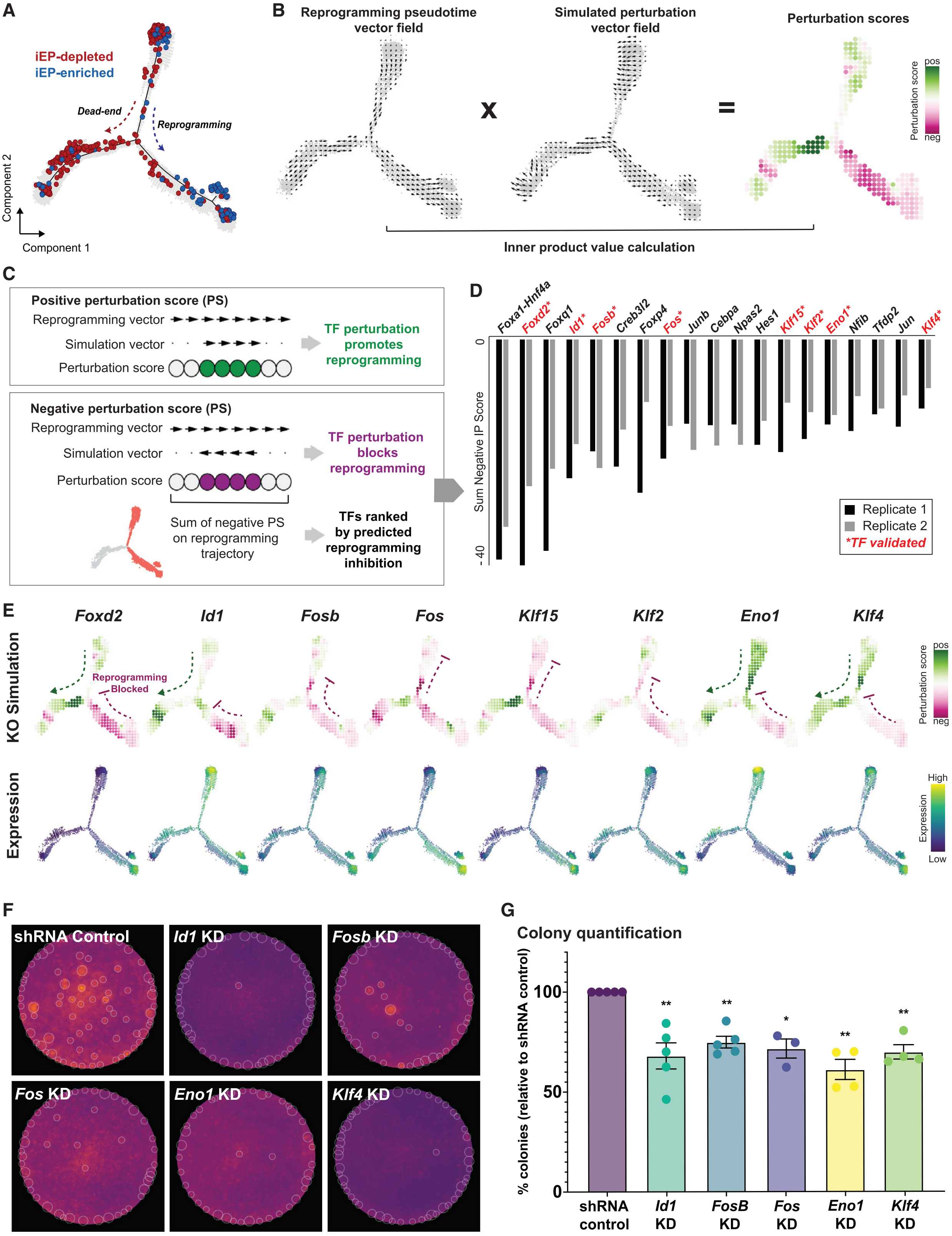
(C) Captured state-fate cells. Time point information projected onto the Uniform Manifold Approximation and Projection (UMAP) embedding. A total of 24,799 cells were sequenced: 8,440 on day 4, 4,836 on day 10, and 11,523 on day 28.

(D) Projection of fibroblast, iEP, and dead-end identity scores and (E) fate annotations onto the UMAP embedding.

(F) A randomized test identified day 4 state clones whose day 10 and 28 fate sisters were iEP-enriched or iEP-depleted. (Top) Kernel density estimation of iEP-enriched day 4 state clones and their day 10 and 28 fates, outlining the reprogramming trajectory (n = 1,347 cells). (Bottom) iEP-depleted state-fate cells outlining the dead-end trajectory (n = 4,802 cells).

(G) Projection of iEP-enriched and iEP-depleted clones onto the UMAP embedding.

(H) Comparison of degree centrality scores between native fibroblasts and day 4 reprogrammed-destined cells (left) and day 4 reprogrammed- and dead-end-destined cells (right).



(legend on next page)

representing TF-to-target gene signal flow. This signal is propagated iteratively to calculate the broad, downstream effects of TF perturbation, allowing the global transcriptional shift to be estimated (Figures S3A and S3B). (3) The probability of a cell identity transition is estimated by comparing this gene expression shift with the gene expression of local neighbors (Figure S3C). (4) The transition probability is converted into a weighted local average vector to represent the simulated directionality of cell state transition for each cell upon candidate TF perturbation. This final step converts the simulation results into a 2D vector map, enabling robust predictions by mitigating the effect of errors or noise derived from scRNA-seq data and the preceding simulation (Figures 3B middle; S3C). The resulting small-length vectors allow the directionality of cell identity transitions to be feasibly predicted rather than interpreting long-ranging terminal effects from initial states.

To enable the simulation results to be assessed systematically and unbiasedly, we consider the changes in cell identity induced by reprogramming, together with the predicted effects from the perturbation. Taking the relatively densely sampled time course from Biddy et al. (2018), we use semi-supervised Monocle analysis (Trapnell et al., 2014) to order cells in pseudotime based on the expression of the fibroblast marker *Col1a2* and the iEP marker *Apoa1*, capturing the distinctive reprogramming and dead-end trajectories as distinguished by their respective lineage-restricted clones (n = 48,515 cells, two independent biological replicates; Figures 3A and S3D). We use the pseudotime information to calculate a vector gradient, representing the direction of reprogramming as a vector field (Figures 3B, left; S3E; supplemental experimental procedures). We then quantify the similarity between the reprogramming and perturbation simulation vector fields by calculating their inner-product value, which we term perturbation score (Figure 3B). A negative perturbation score implies that the TF perturbation blocks reprogramming (Figure 3C, magenta). Conversely, a positive perturbation score indicates that reprogramming is promoted following TF perturbation (Figure 3C, green). By calculating the sum of the negative perturbation scores, we rank TFs by their potential to regulate the reprogramming process, where a greater negative score indicates that reprogramming is impaired upon KO of the candidate TF. Using these metrics, we can interpret perturbation effects on cell fate quantitatively and objectively.

We used this approach to perform a systematic *in silico* simulation of TF KOs during iEP generation to identify novel reprogramming regulators (Figure S3F). Following GRN inference for each of the seven Monocle states identified (Figure S3D), we performed KO simulations for all TFs with inferred connections to at least one other gene (“active” TFs, n = 180), calculating the sum of the negative perturbation scores to rank TFs by the predicted inhibition of reprogramming following their KO. This *in silico* screen prioritizes factors for experimental validation. In the top-ranked TFs, many factors are shared between independent biological replicates ((Figure 3D; Pearson’s, *r* = 0.72). The *Hnf4α-Foxa1* transgene is ranked top, as expected since these factors are driving the reprogramming process. Only half of the remaining top-ranked factors are differentially expressed in reprogrammed cells (Table S1). Further, only three of these prioritized TFs (*Jun*, *Junb*, *Hes1*) were identified by orthogonal analysis using CoSpar (Wang et al., 2022b) (Table S3), highlighting the utility of CellOracle to recover novel candidate regulators.

For experimental validation, we further prioritized candidate genes based on GRN degree centrality, enrichment of gene expression along the entire reprogramming trajectory, and ranking agreement across biological replicates, yielding eight candidates: *Eno1*, *Fos*, *Fosb*, *Foxd2*, *Id1*, *Klf2*, *Klf4*, and *Klf15* (Figure 3E). For all TFs, CellOracle predicts impaired reprogramming following their KO. We performed an initial screen for all eight TFs, using a short hairpin RNA (shRNA)-based strategy to knock down each TF during reprogramming (confirmed by qRT-PCR;

Figure 3. Systematic *in silico* simulation of TF KO to identify novel regulators of iEP reprogramming

(A) Monocle-based pseudotemporal ordering of 48,515 cells from Biddy et al. (2018), two independent biological replicates.

(B) Schematic for perturbation score calculations. CellOracle calculates a perturbation score by comparing the direction of the simulated cell state transition with the direction of cell differentiation. First, the pseudotime data is summarized by grid points and converted into a 2D gradient vector field. The results of the perturbation simulation are converted into the same vector field format, and the inner product of these vectors is calculated to produce a perturbation score.

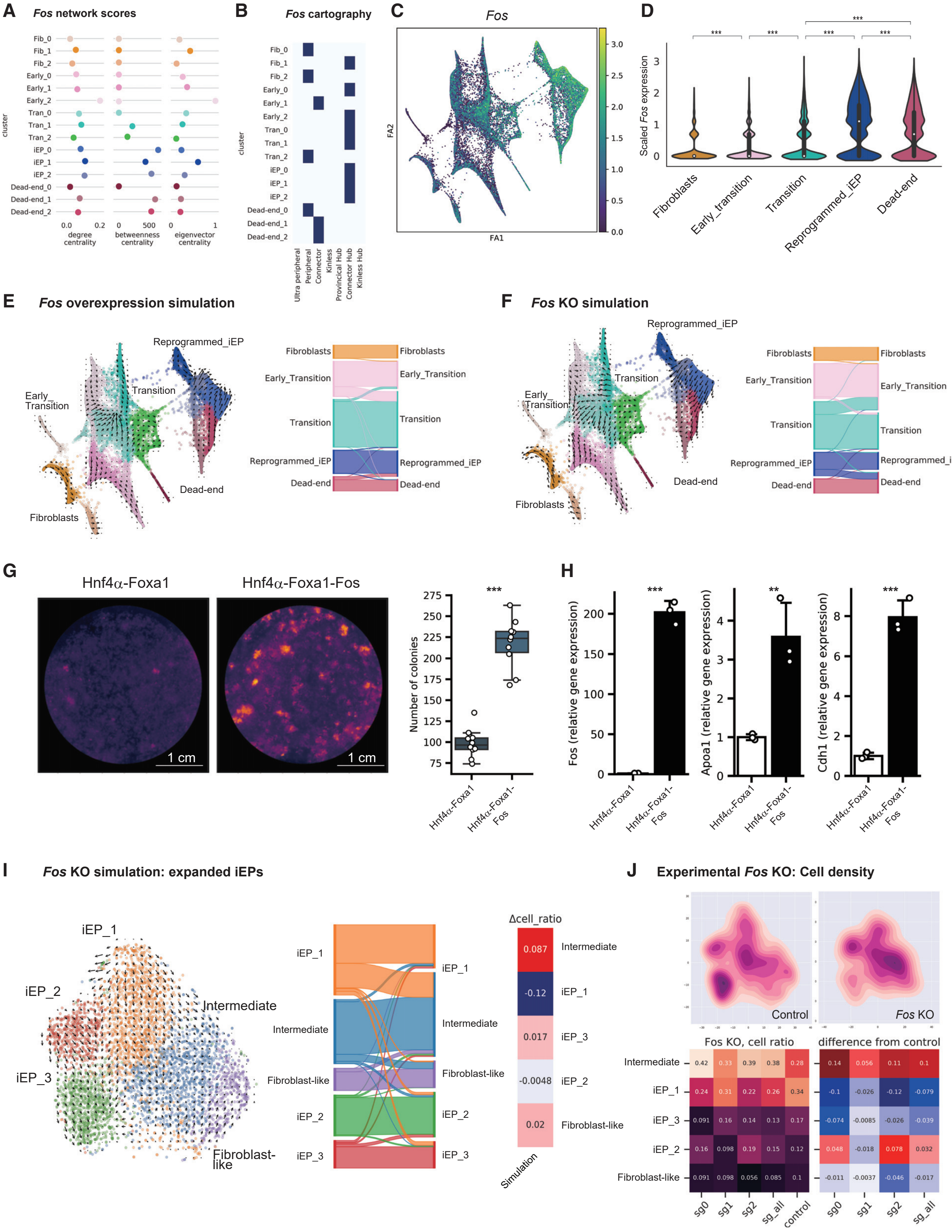
(C) A positive perturbation score (green) suggests that the perturbation is predicted to promote reprogramming. In contrast, the negative perturbation score (magenta) represents impaired reprogramming.

(D) Ranked list of TFs based on the sum of the negative perturbation score.

(E) Representative examples of TF KO simulation (top row). Expression of respective genes (bottom row).

(F) Experimental validation of candidate TFs: colony-formation assay.

(G) Colony quantification. n = 5 independent biological replicates for non-targeting scramble shRNA control, *Fosb*, *Id1*; n = 4 independent biological replicates for *Eno1*, *Klf4*; n = 3 independent biological replicates for Fos; unpaired t test with Welch’s correction, two-tailed; *p < 0.05, **p < 0.01.



(legend on next page)

Figure S3G), followed by colony-formation assay to quantify clusters of successfully reprogrammed cells based on E-cadherin expression. From this initial screen, reprogramming was impaired following the knockdown of six of the eight TFs (*Eno1*, *Fos*, *Fosb*, *Id1*, *Klf4*, and *Klf15*), with 20%–50% fewer colonies formed (Figures S3H and S3I). We selected *Eno1*, *Fos*, *Fosb*, *Id1*, and *Klf4* for additional colony-formation assays, confirming that their knockdown significantly reduces reprogramming efficiency (n = 5 independent biological replicates for non-targeting scramble shRNA control, *Fosb*, *Id1*; n = 4 for *Eno1*, *Klf4*; n = 3 for *Fos*; unpaired t test with Welch's correction, two-tailed; *p < 0.05, **p < 0.01; Figures 3F and 3G).

Overall, our systematic perturbation simulation and experimental validation revealed several novel regulators of MEF to iEP reprogramming. Of these TFs, we identified Fos as a positive regulator of reprogramming. Further, our above state-fate analysis identified *Fos* as a highly connected factor in day 4 reprogrammed-destined clones, suggesting a role for this TF from the early stages of cell fate conversion. Indeed, we noted an enrichment of genes associated with the activator protein-1 TF (AP-1), a dimeric complex primarily containing members of the FOS and JUN factor families (Eferl and Wagner, 2003). AP-1 establishes cell-type-specific enhancers and gene expression programs (Vierbuchen et al., 2017) and reconfigures enhancers during reprogramming to pluripotency (Knaupp et al., 2017). As part of the AP-1 complex, Fos plays broad roles in proliferation, differentiation, and apoptosis, both in development and tumorigenesis (Eferl and Wagner, 2003; Jochum et al., 2001). We next focused on further *in silico* simulation and experimental validation of *Fos*, a core component of AP-1.

The AP-1 TF subunit Fos is central to reprogramming initiation and maintenance of iEP identity

Comparing degree centrality scores between fibroblast and early reprogramming clusters, *Fos* receives relatively high degree and eigenvector centrality scores, along with connector hub classification (Figures 1J, 4A, 4B, and S4A). Clonal analysis of early ancestors destined to reprogram successfully agrees with a central role for *Fos* (Figure 2H). Indeed, perturbation simulation and reduced reprogramming efficiency following experimental knockdown (Figures 3 and S3) led us to select *Fos* for deeper mechanistic investigation as a candidate gene playing a critical role in initiating iEP conversion.

During MEF to iEP reprogramming, *Fos* is gradually and significantly upregulated (Figures 4C and 4D; p < 0.001, permutation test, one sided). Several *Jun* AP-1 subunits are also expressed in iEPs, classifying as connectors and connector hubs across various reprogramming stages (Figures S4C–S4E). *Fos* and *Jun* are among a battery of genes reported to be upregulated in a cell-subpopulation-specific manner in response to cell dissociation-induced stress, potentially leading to experimental artifacts (van den Brink et al., 2017). Considering this report, we performed qRT-PCR for *Fos* on dissociated and undissociated cells. This orthogonal validation confirms an 8-fold upregulation (p < 0.01, t test, one sided) of *Fos* in iEPs, relative to MEFs, revealing no significant changes in gene expression in cells that are dissociated and lysed versus cells lysed directly on the plate (Figure S4F). Further, analysis of unspliced and spliced *Fos* mRNA levels reveals an accumulation of spliced *Fos* transcripts in reprogrammed cells (la Manno et al., 2018). This observation suggests that these transcripts accumulated over time rather than by rapid induction of expression by cell dissociation (Figure S4G).

To further investigate the role of *Fos*, we simulated its overexpression. In these analyses, to assess the *in silico* perturbation of a specific candidate, we use a Markov simulation to predict how cell identity shifts within the overall cell population, visualizing the results as a Sankey diagram. Overexpression simulation for *Fos* predicts a major cell state shift from the early transition to transition clusters,

Figure 4. CellOracle analysis and experimental validation of Fos in establishing and maintaining iEP identity

(A) Degree centrality, betweenness centrality, and eigenvector centrality of *Fos* for each cluster.

(B) Network cartography terms of *Fos* for each cluster.

(C) *Fos* expression projected onto the force-directed graph.

(D) Violin plot of *Fos* expression across reprogramming stages. ***p < 0.001.

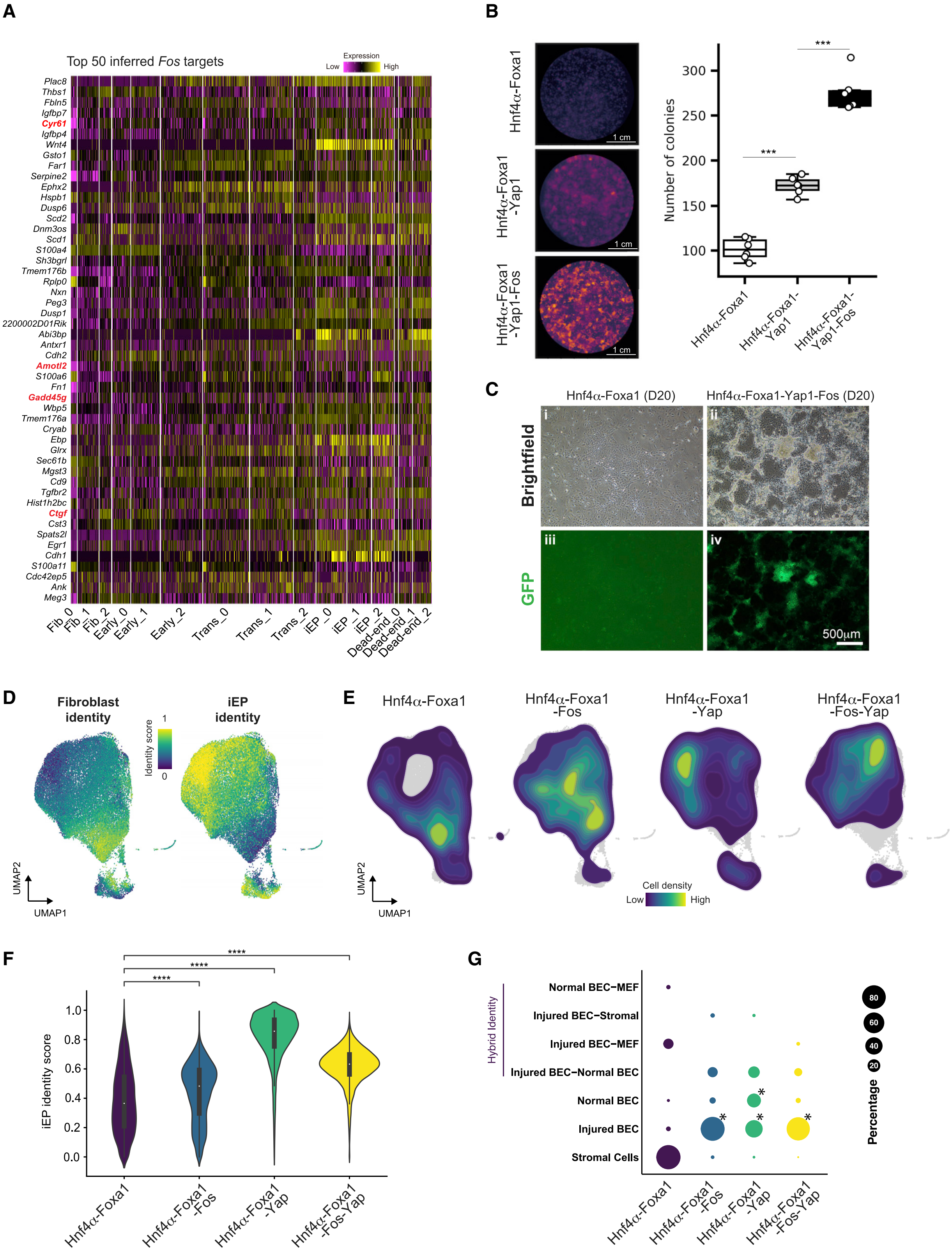
(E and F) (E) *Fos* gene overexpression simulation with reprogramming GRN configurations. (Left) The projection of simulated cell transitions onto the force-directed graph. The Sankey diagram summarizes the simulation of cell transitions between cell clusters. For overexpression simulation, *Fos* expression was set to 1.476, representing its maximum value in the imputed gene expression matrix (F) *Fos* gene KO simulation.

(G) Colony-formation assay with addition of Fos to Hnf4α-Foxa1. (Left) E-cadherin immunohistochemistry. (Right) Boxplot of colony numbers (n = 6 technical replicates, two independent biological replicates; ***p < 0.001, t test, one sided).

(H) qPCR for *Fos* and iEP marker expression (*Apoa1* and *Chd1*) following addition of Fos to Hnf4α-Foxa1 (n = 3 independent biological replicates; ***p < 0.001, **p < 0.01, t test, one sided).

(I) *Fos* gene KO simulation in expanded, long-term cultured iEPs.

(J) CRISPR-Cas9 *Fos* KO in expanded iEP cells. (Left) Kernel density estimation was applied with the t-SNE (t-distributed stochastic neighbor embedding) to compare cell density between control guide RNAs and guide RNAs targeting *Fos*. (Right) Quantification of changes in cell ratio following *Fos* KO.



(legend on next page)

in addition to predicting shifts in identity from dead-end to reprogrammed clusters (Figure 4E). In contrast, the simulation of *Fos* KO produces the opposite results. (Figure 4F). We experimentally validated this simulation by adding Fos to the iEP reprogramming cocktail. As expected, we see a significant increase in the number of iEP colonies formed (n = 10, p < 0.001, t test, one sided; Figure 4G), increasing reprogramming efficiency more than 2-fold, accompanied by significant increases in iEP marker expression as measured by qRT-PCR (n = 3, p < 0.001, t test, one sided; Figure 4H).

Turning our attention to the later stages of reprogramming, *Fos* continues to receive relatively high network scores in the iEP GRN configurations (Figure 4A). *Fos* also classifies as a connector hub (Figure 4B) in iEPs, suggesting a role for Fos in the stabilization and maintenance of the reprogrammed fate. To test this hypothesis, we use CellOracle to perform KO simulation, followed by experimental KO validation in an established iEP cell line. Here, we leverage the ability to culture iEPs, long term, where they retain a range of phenotypes (from fibroblast-like to iEP states; Figure S4H) and functional engraftment potential (Guo et al., 2019; Morris et al., 2014). Simulation of *Fos* KO using these long-term cultured iEP GRN configurations predicts the loss of iEP identity upon *Fos* KO (Figure 4I). To test this prediction, we used CRISPR-Cas9 to knock out *Fos* in established iEPs. Quantitative comparison of the cell proportions between control and KO groups confirms that fully reprogrammed iEPs regress toward an intermediate state upon *Fos* KO, confirming a role for this factor in maintaining iEP identity (Figure 4J), in addition to the establishment of iEPs, as we demonstrate in our systematic simulation and experimental validation in Figure 3.

Fos target inference uncovers a role for the Hippo signaling effector Yap1 in reprogramming

To gain further insight into Fos regulation of reprogramming, we interrogated a list of the top 50 inferred *Fos* targets (Figure 5A; Table S2). We also assembled a list of genes predicted to be downregulated following *Fos* KO simulation (Figure S5A). From this analysis, we noted the presence of direct targets of YAP1, a central downstream transducer of the Hippo signaling pathway (Ramos and Camargo, 2012). These targets include *Cyr61*, *Amotl2*, *Gadd45g*, and *Ctgf*. Previous associations between Yap1 and Fos support these observations; YAP1 is recruited to the same genomic regions as FOS via complex formation with AP-1 (Zanconato et al., 2015). Moreover, AP-1 is required for YAP1-regulated gene expression and liver overgrowth caused by Yap overexpression, where FOS induction contributes to the expression of YAP/TAZ downstream target genes (Koo et al., 2020).

Together, this evidence suggests that Fos may play a role in reprogramming via an AP-1-Yap1-mediated mechanism. Since Yap1 does not directly bind to DNA, we cannot deploy CellOracle to perform perturbation simulations, highlighting a limitation of our approach. However, in lieu of this analysis, we again turn to our previous reprogramming data (Biddy et al., 2018). Using an established active signature of Yap1 (Dong et al., 2007), we find significant enrichment of this signature as reprogramming progresses (Figures S5B and S5C; p < 0.001, permutation test, one-sided). Together, these results suggest a role for the Hippo signaling component Yap1 in reprogramming, potentially affected via its interactions with Fos/AP-1. Indeed, the Hippo signaling axis plays a role in liver regeneration (Pepe-Mooney et al., 2019) and regeneration of the colonic epithelium (Yui et al., 2018), in line with the known potential of iEPs to functionally engraft the liver and intestine (Guo et al., 2019; Morris et al., 2014; Sekiya and Suzuki, 2011). Further, we have recently demonstrated that iEPs transcriptionally resemble injured BECs (Kong et al., 2022), the target of YAP signaling in the context of liver regeneration (Pepe-Mooney et al., 2019).

To test the role of Yap1 in iEP reprogramming, we first performed colony-formation assays. We find that the addition of Yap1 to the Hnf4α-Foxa1 cocktail significantly enhances reprogramming efficiency, where the addition of Fos and Yap1 together increase colony formation almost 3-fold, accompanied by significant increases in iEP marker

Figure 5. Inferred Fos targets reveal a role for the Hippo signaling effector, Yap1, in reprogramming

(A) Heatmap of expression of the top 50 inferred *Fos* targets across reprogramming. Established YAP1 targets are highlighted in red.

(B) Colony-formation assay with the addition of Yap1 and Fos to Hnf4α-Foxa1. (Left) E-cadherin immunohistochemistry. (Right) Boxplot of colony numbers (n = 6 independent biological replicates; ***p < 0.001, t-test, one sided).

(C) Brightfield and epifluorescence images of cells reprogrammed with Hnf4α-Foxa1 or Hnf4α-Foxa1-Fos-Yap1. Scale bar, 500 μm.

(D) scRNA-seq of cells reprogrammed with Hnf4α-Foxa1 (n = 7,414 cells), Hnf4α-Foxa1-Fos (n = 8,771 cells), Hnf4α-Foxa1-Yap1 (n = 8,549 cells), and Hnf4α-Foxa1-Fos-Yap1 (n = 10,507 cells), profiled at day 20. Projection of fibroblast and iEP identity scores onto the UMAP embedding.

(E) Kernel density estimation of cell density for each reprogramming cocktail from (D).

(F) Violin plot of iEP identity scores for each reprogramming cocktail. ****p < 0.0001, Wilcoxon test.

(G) Unsupervised cell type classification for each reprogramming cocktail, using normal and injured mouse liver as a reference. BEC, biliary epithelial cells. *p < 0.0001, randomized test.

expression (Figures 5B, S5D, and S5E, $p < 0.001$, t test, one sided). Further, we note the formation of extremely dense colonies (Figure 5C). To further characterize this distinctive phenotype, we performed scRNA-seq on cells reprogrammed with Hnf4α-Foxa1 (n = 7,414 cells), Hnf4α-Foxa1-Yap1 (n = 8,549 cells), Hnf4α-Foxa1-Fos (n = 8,771 cells), and Hnf4α-Foxa1-Yap1-Fos (n = 10,507 cells), profiled at day 20 (Figure S5F).

We scored cells using established markers of MEFs and iEPs (Biddy et al., 2018), revealing a significant increase in reprogramming efficiency, particularly following the addition of Yap1 ($p < 0.0001$, Wilcoxon test; Figures 5F and S5F), which is also accompanied by a reduction in fibroblast marker expression (Figure S5G). We further classify cell identity using our unsupervised method for cell-type classification, Capybara (Kong et al., 2022). In agreement with our previous reports, using a healthy and regenerating liver atlas, iEPs generated with Hnf4α-Foxa1 alone classify mainly as stromal cells (Figure 5G). However, following the addition of Fos and Yap1, a significant population ($p < 0.0001$, randomized test) of injured BECs emerges, in similar proportions to those observed in long-term cultured iEPs (Kong et al., 2022). We also observe a significant expansion of a normal BEC population, from ~4% to ~12%–35%, upon the addition of Yap1 to the reprogramming cocktail ($p < 0.0001$, randomized test), where endogenous *Fos* expression is also upregulated (Figure S5G). We observed a similar expansion of the normal BEC population when long-term iEPs were cultured in a 3D Matrigel sandwich culture (Kong et al., 2022). Here, our results are consistent with these previous observations and point to the molecular regulation driving changes in cell identity. In summary, CellOracle analysis and *in silico* prediction, combined with experimental validation, have revealed several new factors and putative regulatory mechanisms to enhance the efficiency and fidelity of reprogramming.

DISCUSSION

Our application of CellOracle to iEP reprogramming has revealed new insight into this lineage conversion paradigm. Using CellTag-based lineage tracing, we had previously demonstrated the existence of distinct conversion trajectories: one path leading to successfully reprogrammed cells and a route to a dead-end state, accompanied by fibroblast gene re-expression (Biddy et al., 2018). From lineage analysis, we found that sister cells follow the same reprogramming trajectories, suggesting that conversion outcome is established shortly after overexpression of the reprogramming TFs. The network analysis we present in this study, powered by CellOracle, supports these earlier observations, revealing GRN reconfiguration within the first few days of reprogramming.

From our analysis of early GRN rewiring, we find that *Mef2a* and *Klf6* are highly connected in fibroblasts and that these connections are largely decommissioned in successfully converting cells. Although better known as a cardiac factor (Filomena and Bang, 2018), *Mef2a* expression is enriched in the dead-end population, whereas *Klf6* is enriched in early transition states, followed by its downregulation as reprogramming progresses. In this study, we have mainly focused on the TFs associated with installing new cell identities. From our clonal analysis of GRN reconfiguration in reprogrammed-destined cells, we find many previously unreported regulators of iEP reprogramming, such as *Klf5*, *Mybl2*, *Fox*q1, *Fos*, and *Junb*. The recovery of these factors is likely due to the clonal analysis, which further breaks down population heterogeneity to target those rare cells that successfully reprogram.

To explore the role of these factors in reprogramming, we leverage the unique feature of CellOracle: simulation of cell identity transition following candidate TF perturbation (KO or overexpression). From systematic *in silico* KO simulation and experimental validation, we identified five new regulators of iEP reprogramming: *Id1*, *Fosb*, *Fos*, *Eno1*, and *Klf4*. Klf4 is one of the previously described core pluripotency reprogramming factors (Takahashi and Yamanaka, 2006). The reduction of iEP reprogramming efficiency following its knockdown also suggests that Klf4 plays a role in this direct lineage conversion paradigm. Similarly, Id1 has also been shown to play a positive role in reprogramming to pluripotency (Hayashi et al., 2016), suggesting parallels with direct lineage conversion. We also noted the involvement of several AP-1 factors, both from our network analyses and *in silico* simulations, including *Fos*, *Fosb*, *Fosl2*, and *Junb*. The FOS-JUN-AP1 complex has been reported to regulate reprogramming to pluripotency (Xing et al., 2020) and direct reprogramming to cardiomyocytes (Wang et al., 2022a); thus, we selected Fos for further investigation.

The CellOracle analyses presented here provide new mechanistic insight into the reprogramming process, revealing a role for the Fos-Yap1 axis, which we experimentally validated. In a parallel study, we found that iEPs resemble post-injury BECs (Kong et al., 2022). Considering that Yap1 plays a central role in liver regeneration (Pepe-Mooney et al., 2019), these results raise the possibility that iEPs represent a regenerative cell type, explaining their Yap1 activity, self-renewal *in vitro*, and capacity to functionally engraft the liver (Sekiya and Suzuki, 2011) and intestine (Guo et al., 2019; Morris et al., 2014). Indeed, our unsupervised cell type classification of iEPs reprogrammed with the addition of Fos and Yap to the Hnf4α-Foxa1 reprogramming cocktail suggests that these factors can directly expand the injured and normal BEC populations, supporting the notion that iEPs may resemble a regenerative population. Altogether, these new mechanistic insights have been enabled by CellOracle analysis, placing it as a powerful tool for the dissection of cell identity, aiding improvements in reprogramming efficiency and fidelity.

EXPERIMENTAL PROCEDURES

Detailed experimental procedures can be found in the supplemental information.

Resource availability

Corresponding author

Samantha A. Morris, s.morris@wustl.edu.

Materials availability

Pooled CellTag libraries have been deposited at Addgene: https://www.addgene.org/pooled-library/morris-lab-celltag/

Data and code availability

All source data, including sequencing reads and single-cell expression matrices, are available from the Gene Expression Omnibus (GEO) under accession codes GSE99915 (Biddy et al., 2018) and GSE217675 for the new scRNA-seq data presented in this manuscript. CellOracle code, documentation, and tutorials are available on GitHub: (https://github.com/morris-lab/CellOracle).

Computational methods

CellOracle code is open source and available on GitHub: (https://github.com/morris-lab/CellOracle). For alignment, digital gene expression matrix generation, the Cell Ranger v6.0.1 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process data generated using the 10x Chromium platform. For clone calling, we used our CellTag analysis pipeline: https://github.com/morris-lab/newCloneCalling. Cell type classification was performed using Capybara: https://github.com/morris-lab/Capybara.

Experimental methods

MEFs were derived from E13.5 C57BL/6J embryos (the Jackson laboratory: 000664). Retroviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pGCDN-Sam construct containing Hnf4α-t2a-Foxa1/Fos/Yap1, along with packaging construct pCL-Eco (Imgenex). Lentiviral particles were produced with the envelope construct pCMV-VSV-G (Addgene plasmid 8454), the packaging construct pCMV-dR8.2 dvpr (Addgene plasmid 8455), and the shRNA expression vector for the respective candidate TF to be knocked down. For generation of the complex CellTag library, lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct, along with packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455) and pCMV-VSVG (Addgene plasmid 8454). For iEP reprogramming, MEFs (< passage 6) were converted to iEPs as in Biddy et al. (2018), modified from (Sekiya and Suzuki, 2011). Colony-formation assays were performed as in Biddy et al. (2018). Perturb-seq was performed as previously described (Adamson et al., 2016). Single-cell libraries were prepared using the 10x Genomics Chromium platform.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stemcr.2022.11.010.

AUTHOR CONTRIBUTIONS

Conceptualization and methodology, K.K. and S.A.M.; software, K.K.; formal analysis, K.K., M.T.A., K.J., C.M.H., and S.A.M; investigation, K.K., M.T.A., K.J., C.M.H., X.Y., and S.A.M.; data curation, K.K., M.T.A., and K.J.; writing – original draft, K.K. and S.A.M.; writing – review & editing, K.K., M.T.A., K.J., C.M.H., and S.A.M.; visualization, K.K. and S.A.M.; funding acquisition and resources, Supervision, S.A.M.

ACKNOWLEDGMENTS

We thank members of the Morris laboratory for critical feedback. This work was funded by the National Institute of General Medical Sciences (R01 GM126112) and Silicon Valley Community Foundation, Chan Zuckerberg Initiative grant HCA2-A-1708-02799, both to S.A.M. S.A.M. is supported by an Allen Distinguished Investigator Award (through the Paul G. Allen Frontiers Group), a Vallee Scholar Award, a Sloan Research Fellowship, and a New York Stem Cell Foundation Robertson Investigator Award. K.K. is supported by a Japan Society for the Promotion of Science Postdoctoral Fellowship. C.M.H. is supported by a National Science Foundation Graduate Research Fellowship (DGE-1745038).

CONFLICT OF INTERESTS

S.A.M. is a co-founder of CapyBio LLC.

Received: April 14, 2022
Revised: November 11, 2022
Accepted: November 14, 2022
Published: December 29, 2022

REFERENCES

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell *167*, 1867–1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219–224. https://doi.org/10.1038/s41586-018-0744-4.

Bocchi, V.D., Conforti, P., Vezzoli, E., Besusso, D., Cappadona, C., Lischetti, T., Galimberti, M., Ranzani, V., Bonnal, R.J.P., De Simone, M., et al. (2021). The coding and long noncoding single-cell atlas of the developing human fetal striatum. Science *372*, eabf5759. https://doi.org/10.1126/science.abf5759.

van den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and van Oudenaarden, A.

(2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936. https://doi.org/10.1038/nmeth.4437.

Chopp, L.B., Gopalan, V., Ciucci, T., Ruchinskas, A., Rae, Z., Lagarde, M., Gao, Y., Li, C., Bosticardo, M., Pala, F., et al. (2020). An integrated epigenomic and transcriptomic map of mouse and human αβ T cell development. Immunity *53*, 1182–1201.e8. https://doi.org/10.1016/J.IMMUNI.2020.10.024.

Cohen, D.E., and Melton, D. (2011). Turning straw into gold: directing cell fate for regenerative medicine. Nat. Rev. Genet. *12*, 243–252. https://doi.org/10.1038/nrg2938.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A single-cell atlas of in vivo mammalian chromatin accessibility. Cell *174*, 1309–1324.e18. https://doi.org/10.1016/J.CELL.2018.06.052.

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796–800. https://doi.org/10.1126/science.1113832.

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S.A., Gayyed, M.F., Anders, R.A., Maitra, A., and Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. Cell *130*, 1120–1133. https://doi.org/10.1016/J.CELL.2007.07.019.

Eferl, R., and Wagner, E.F. (2003). AP-1: a double-edged sword in tumorigenesis. Nat. Rev. Cancer *3*, 859–868. https://doi.org/10.1038/nrc1209.

Filomena, M.C., and Bang, M.L. (2018). In the heart of the MEF2 transcription network: novel downstream effectors as potential targets for the treatment of cardiovascular disease. Cardiovasc. Res. *114*, 1425–1427. https://doi.org/10.1093/CVR/CVY123.

Guimerà, R., and Amaral, L.A.N. (2005). Cartography of complex networks: modules and universal roles. J. Stat. Mech. *2005*, P02001-1–P02001-13. https://doi.org/10.1088/1742-5468/2005/02/P02001.

Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G.C., Yang, X., Kirita, Y., and Morris, S.A. (2019). CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biol. *20*, 90. https://doi.org/10.1186/s13059-019-1699-y.

Hayashi, Y., Hsiao, E.C., Sami, S., Lancero, M., Schlieve, C.R., Nguyen, T., Yano, K., Nagahashi, A., Ikeya, M., Matsumoto, Y., et al. (2016). BMP-SMAD-ID promotes reprogramming to pluripotency by inhibiting p16/INK4A-dependent senescence. Proc. Natl. Acad. Sci. USA *113*, 13057–13062. https://doi.org/10.1073/PNAS.1603668113.

Jochum, W., Passegué, E., and Wagner, E.F. (2001). AP-1 in mouse development and tumorigenesis. Oncogene *20*, 2401–2412. https://doi.org/10.1038/sj.onc.1204389.

Jindal, K., Adil, M.T., Yamaguchi, N., Wang, H.C., Yang, X., Kamimoto, K., Rivera-Gonzalez, G.C., and Morris, S.A. (2022). Multiomic single-cell lineage tracing to dissect fate-specific gene regulatory programs. Preprint at bioRxiv 2022.10.23.512790 *20*. https://doi.org/10.1101/2022.10.23.512790.

Kamimoto, K., Hoffmann, C.M., and Morris, S.A. (2020). CellOracle: dissecting cell identity via network inference and in silico gene perturbation. Preprint at bioRxiv. https://doi.org/10.1101/2020.02.17.947416.

Klein, C., Marino, A., Sagot, M.-F., Vieira Milreu, P., and Brilli, M. (2012). Structural and dynamical analysis of biological networks. Brief. Funct. Genomics *11*, 420–433. https://doi.org/10.1093/bfgp/els030.

Knaupp, A.S., Buckberry, S., Pflueger, J., Lim, S.M., Ford, E., Larcombe, M.R., Rossello, F.J., de Mendoza, A., Alaei, S., Firas, J., et al. (2017). Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. Cell Stem Cell *21*, 834–845.e6. https://doi.org/10.1016/J.STEM.2017.11.007.

Kong, W., Biddy, B.A., Kamimoto, K., Amrute, J.M., Butka, E.G., and Morris, S.A. (2020). CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. Nat. Protoc. *15*, 750–772. https://doi.org/10.1038/s41596-019-0247-2.

Kong, W., Fu, Y.C., Holloway, E.M., Garipler, G., Yang, X., Mazzoni, E.O., and Morris, S.A. (2022). Capybara: a computational tool to measure cell identity and fate transitions. Cell Stem Cell *29*, 635–649.e11. https://doi.org/10.1016/J.STEM.2022.03.001.

Koo, J.H., Plouffe, S.W., Meng, Z., Lee, D.-H., Yang, D., Lim, D.-S., Wang, C.-Y., and Guan, K.-L. (2020). Induction of AP-1 by YAP/TAZ contributes to cell proliferation and organ growth. Genes Dev. *34*, 72–86. https://doi.org/10.1101/gad.331546.119.

Magaletta, M.E., Lobo, M., Kernfeld, E.M., Aliee, H., Huey, J.D., Parsons, T.J., Theis, F.J., and Maehr, R. (2022). Integration of single-cell transcriptomes and chromatin landscapes reveals regulatory programs driving pharyngeal organ development. Nat. Commun. *13*, 457. https://doi.org/10.1038/s41467-022-28067-4.

la Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. https://doi.org/10.1038/s41586-018-0414-6.

Morris, S.A., and Daley, G.Q. (2013). A blueprint for engineering cell fate: current technologies to reprogram cell identity. Cell Res. *23*, 33–48. https://doi.org/10.1038/cr.2013.1.

Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., and Daley, G.Q. (2014). Dissecting engineered cell types and enhancing cell fate conversion via CellNet. Cell *158*, 889–902. https://doi.org/10.1016/j.cell.2014.07.021.

Nie, J., Carpenter, A.C., Chopp, L.B., Chen, T., Balmaceno-Criss, M., Ciucci, T., Xiao, Q., Kelly, M.C., McGavern, D.B., Belkaid, Y., et al. (2022). The transcription factor LRF promotes integrin β7 expression by and gut homing of CD8αα+ intraepithelial lymphocyte precursors. Nat. Immunol. *23*, 594–604. https://doi.org/10.1038/s41590-022-01161-x.

Pepe-Mooney, B.J., Dill, M.T., Alemany, A., Ordovas-Montanes, J., Matsushita, Y., Rao, A., Sen, A., Miyazaki, M., Anakk, S., Dawson, P.A., et al. (2019). Single-cell analysis of the liver epithelium reveals dynamic heterogeneity and an essential role for YAP in homeostasis and regeneration. Cell Stem Cell *25*, 23–38.e8. https://doi.org/10.1016/J.STEM.2019.04.004.

Ramos, A., and Camargo, F.D. (2012). The Hippo signaling pathway and stem cell biology. Trends Cell Biol. *22*, 339–346. https://doi.org/10.1016/J.TCB.2012.04.006.

Ravi, V.M., Neidert, N., Will, P., Joseph, K., Maier, J.P., Kückelhaus, J., Vollmer, L., Goeldner, J.M., Behringer, S.P., Scherer, F., et al. (2022). T-cell dysfunction in the glioblastoma microenvironment is mediated by myeloid cells releasing interleukin-10. Nat. Commun. *13*, 925. https://doi.org/10.1038/s41467-022-28523-1.

Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature *475*, 390–393. https://doi.org/10.1038/nature10263.

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. https://doi.org/10.1016/j.cell.2006.07.024.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. https://doi.org/10.1038/nbt.2859.

Vierbuchen, T., Ling, E., Cowley, C.J., Couch, C.H., Wang, X., Harmin, D.A., Roberts, C.W.M., and Greenberg, M.E. (2017). AP-1 transcription factors and the BAF complex mediate signal-dependent enhancer selection. Mol. Cell *68*, 1067–1082.e12. https://doi.org/10.1016/J.MOLCEL.2017.11.026.

Wang, H., Yang, Y., Qian, Y., Liu, J., and Qian, L. (2022a). Delineating chromatin accessibility re-patterning at single cell level during early stage of direct cardiac reprogramming. J. Mol. Cell. Cardiol. *162*, 62–71. https://doi.org/10.1016/J.YJMCC.2021.09.002.

Wang, S.W., Herriges, M.J., Hurley, K., Kotton, D.N., and Klein, A.M. (2022b). CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. Nat. Biotechnol. *40*, 1066–1074. https://doi.org/10.1038/s41587-022-01209-1.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381. https://doi.org/10.1126/SCIENCE.AAW3381.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. *20*, 59. https://doi.org/10.1186/s13059-019-1663-x.

Xing, Q.R., el Farran, C.A., Gautam, P., Chuah, Y.S., Warrier, T., Toh, C.X.D., Kang, N.Y., Sugii, S., Chang, Y.T., Xu, J., et al. (2020). Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. Sci. Adv. *6*, 18. https://doi.org/10.1126/SCIADV.ABA1190.

Yui, S., Azzolin, L., Maimets, M., Pedersen, M.T., Fordham, R.P., Hansen, S.L., Larsen, H.L., Guiu, J., Alves, M.R.P., Rundsten, C.F., et al. (2018). YAP/TAZ-Dependent reprogramming of colonic epithelium links ECM remodeling to tissue regeneration. Cell Stem Cell *22*, 35–49.e7. https://doi.org/10.1016/j.stem.2017.11.001.

Zanconato, F., Forcato, M., Battilana, G., Azzolin, L., Quaranta, E., Bodega, B., Rosato, A., Bicciato, S., Cordenonsi, M., and Piccolo, S. (2015). Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. Nat. Cell Biol. *17*, 1218–1227. https://doi.org/10.1038/ncb3216.

Stem Cell Reports, Volume *18*

Supplemental Information

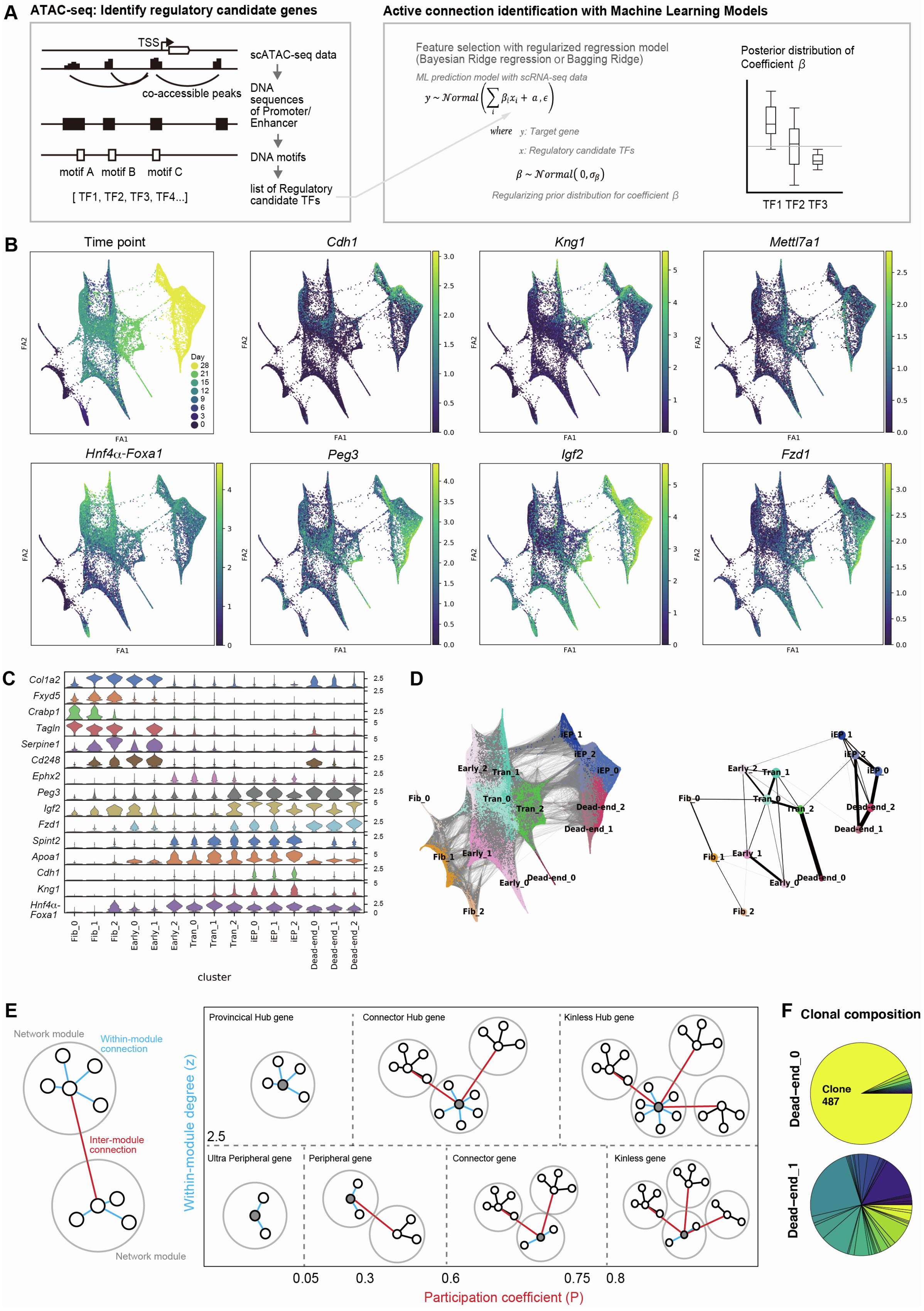
Gene regulatory network reconfiguration in direct lineage reprogramming

Kenji Kamimoto, Mohd Tayyab Adil, Kunal Jindal, Christy M. Hoffmann, Wenjun Kong, Xue Yang, and Samantha A. Morris

Supplemental Figures and Methods

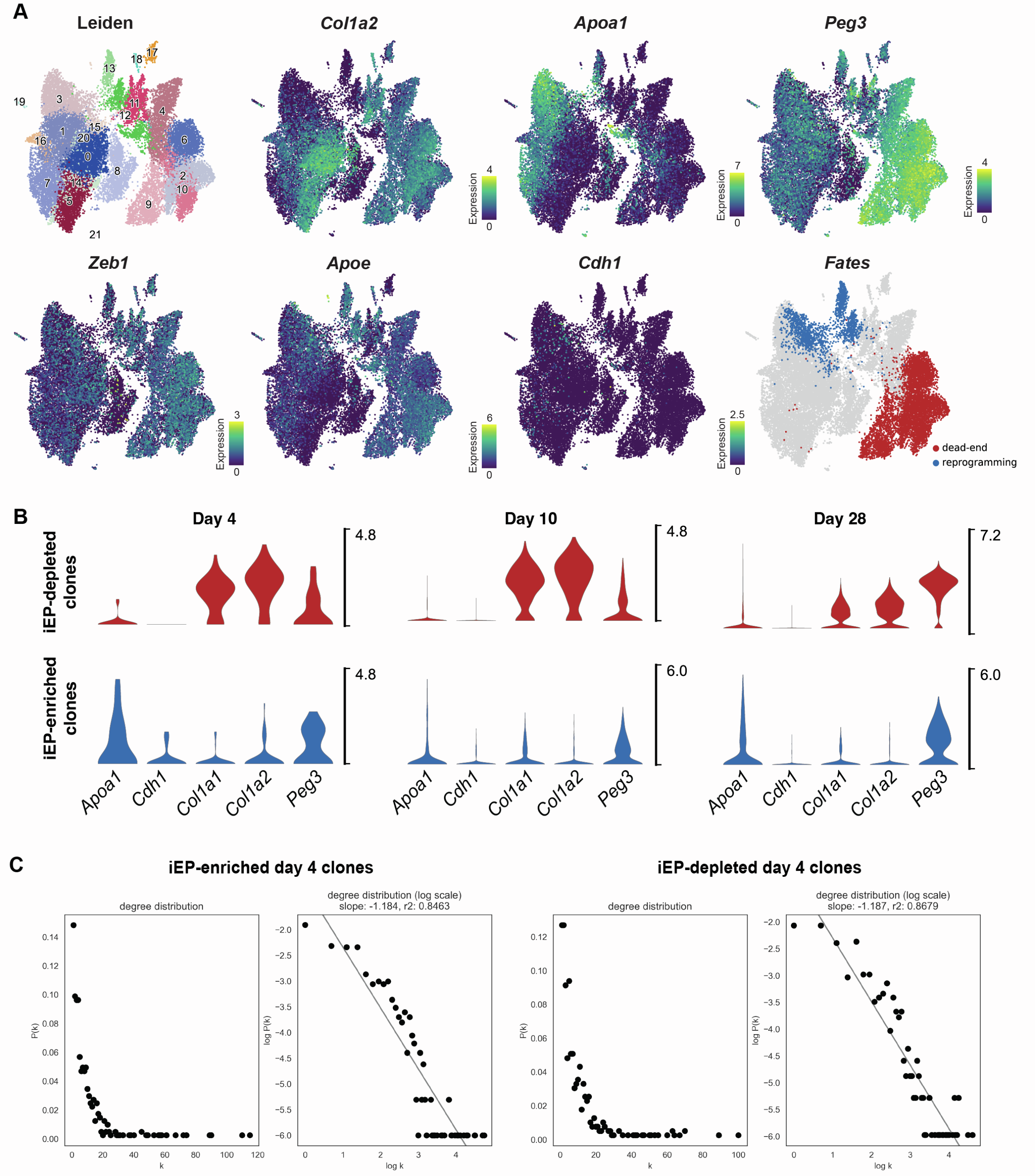
Supplemental Figure 1 (Related to Figure 1). GRN analysis of fibroblast to iEP reprogramming. (A) After base GRN construction (left), single-cell expression data is used to identify active TF-target gene connections for defined cell identities and states. To achieve this, we build a machine learning (ML) model that predicts the relationship between the TF and the target gene. ML model fitting results present the certainty of connection as a distribution, enabling the identification of GRN configurations by removing inactive connections from the base GRN structure. **(B)** Force-directed graph of iEP reprogramming scRNA-seq data (n = 27,663 cells). Reprogramming time point information is projected onto the force-directed graph. There are eight time points; day 0, 3, 6, 9, 12, 15, 21, and 28. *Hnf4α-t2a-Foxa1* (*Hnf4α-Foxa1)* transgene expression levels, and marker gene expression for key iEP states are projected onto the graph. Reprogrammed iEP marker genes: *Cdh1, Apoa1,* and *Kng1*. Fibroblast marker gene: *Col1a2*. Transition marker gene: *Mettl7a1*. Dead-end marker genes: *Peg3, Igf2,* and *Fzd1.* **(C)** Violin plots of marker gene expression in each cluster. **(D)** PAGA connectivity analysis across the reprogramming time course. **(E)** Illustration of the cartography analysis method. The cartography method classifies genes into seven groups according to two network scores: within-module degree and participation coefficient (Guimerà and Amaral, 2005). In complex networks, high-degree nodes (hubs) play the most significant roles in maintaining network structure. **(F)** Pie charts depicting the clonal composition of Dead-end cluster 0 and Dead-end cluster 1. Clone and trajectory information is derived from our previous CellTagging study (Biddy et al., 2018).

Supplemental Figure 1



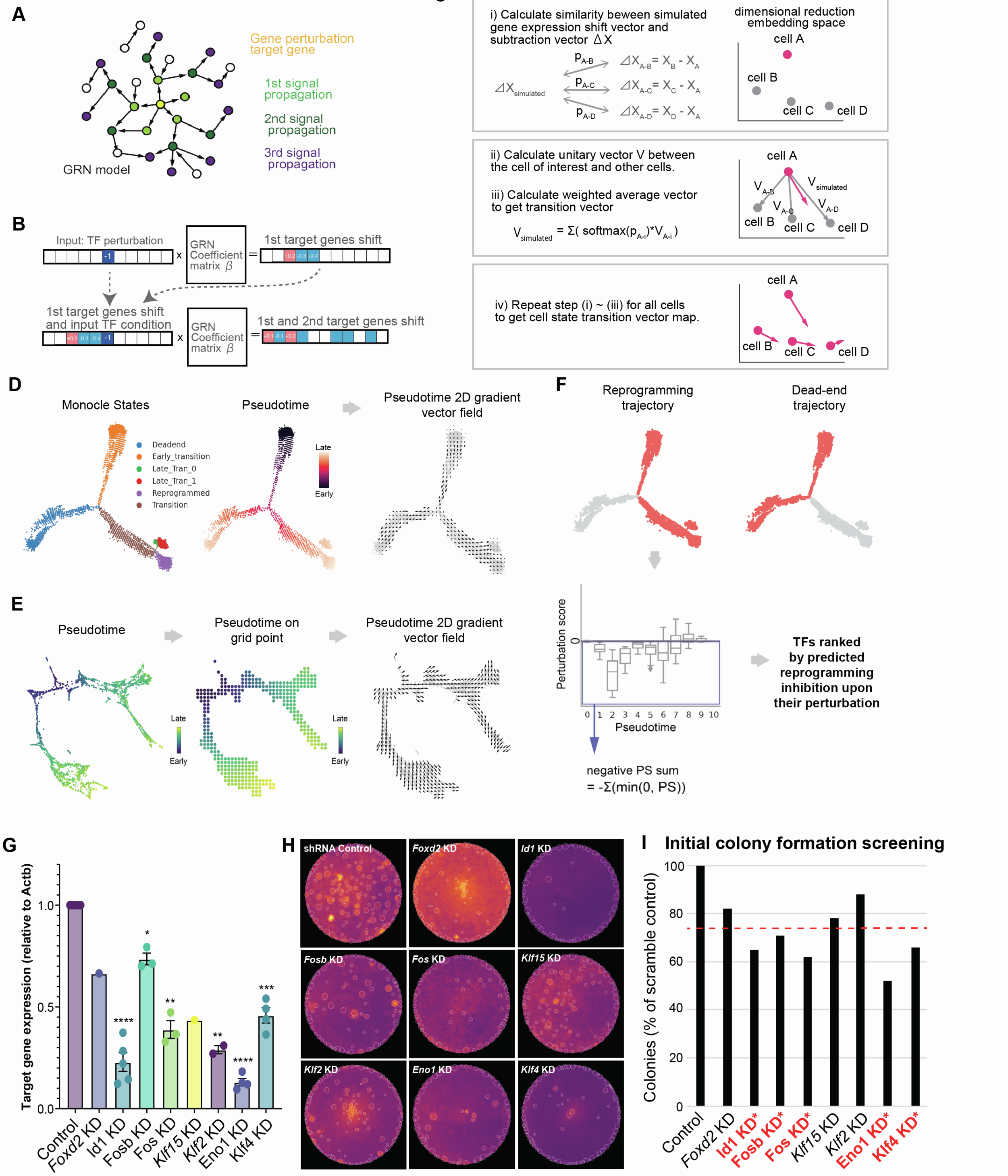
Supplemental Figure 2 (Related to Figure 2). CellOracle network analysis of cells destined to reprogrammed or dead-end fates. (A) Projection of Leiden cluster and gene expression information onto the state-fate UMAP embedding (from **Figure 2C-F**) to identify reprogrammed and dead-end fates. **(B)** Violin plots of reprogrammed (*Apoa1, Cdh1*), fibroblast (*Col1a1, Col1a2*), and dead-end (*Peg3*) marker expression along the iEP-enriched and iEP-depleted trajectories. **(C)** To assess the quality of the inferred networks, we calculate the degree distribution for each GRN configuration after pruning weak network edges based on the p-value and strength. We count the network degree (k), representing the number of network edges for each gene. P(k) is the frequency of network degree k, visualized in scatter plots. We also visualize the relationship between k and P(k) after log transformation, showing that these are scale-free networks, demonstrating successful network inference from these relatively small cell populations.

Supplementary Figure 2



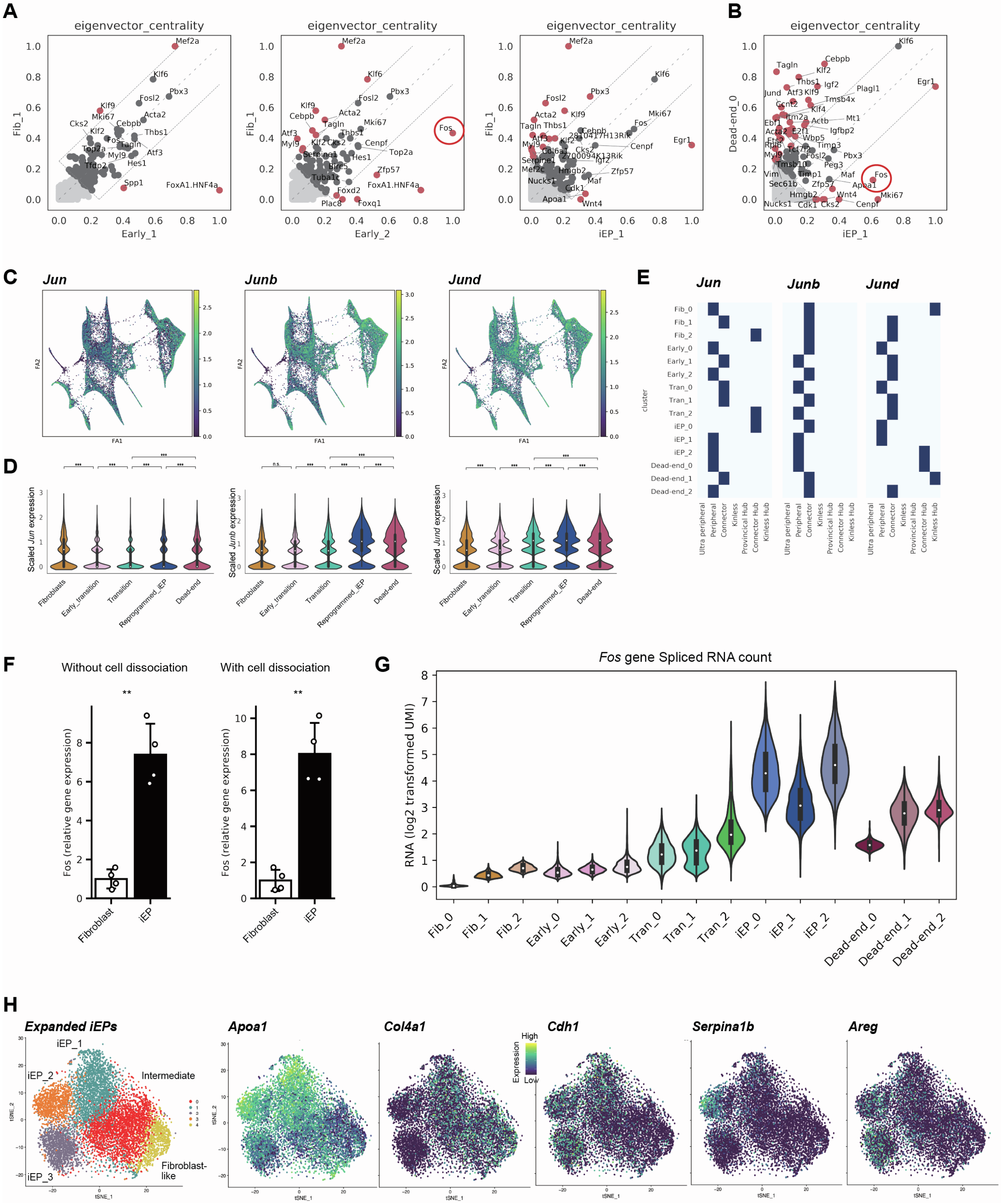
Supplemental Figure 3 (Related to Figure 3). Systematic *in silico* simulation of TF knockout. (A) Overview of signal propagation simulation. CellOracle leverages an inferred GRN model to simulate how target gene expression changes in response to the changes in regulatory gene (TF) expression. The input TF perturbation (yellow) is propagated side-by-side within the network model. **(B)** Leveraging the linear predictive ML algorithm features, CellOracle uses the GRN model as a function to perform the signal propagation calculation. Iterative matrix multiplication steps enable the estimation of indirect and global downstream effects resulting from the perturbation of a single TF. **(C)** After signal propagation, the simulated gene expression shift vector is converted into a 2D vector and projected onto the dimensional reduction space. **(D)** Left: Monocle states identified and used for GRN inference. Right: Calculated pseudotime projected on the Monocle embedding and converted to a 2D gradient vector field. **(E)** Schematic of the method to convert pseudotime to a 2D gradient vector field: First, the pseudotime data is summarized by grid points, then CellOracle calculates a 2D gradient vector of the pseudotime data that represents the directionality of reprogramming pseudotime. **(F)** Outline of reprogramming and dead-end trajectories projected onto the Monocle embedding. The sum of the negative perturbation score was calculated only for reprogramming trajectory clusters in this study. **(G)** Quantitative RT-PCR (qRT-PCR) to validate knockdown efficiency for each shRNA. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$, **** = $p<0.0001$; unpaired t-test with Welch's correction, two-tailed. **(H)** Colony formation assay (E-cadherin immunohistochemistry) to test iEP reprogramming efficiency following the knockdown of each candidate factor. **(I)** Quantification of colonies formed in the initial screen. Factors marked red and * were selected for further experimental validation.

Supplemental Figure 3



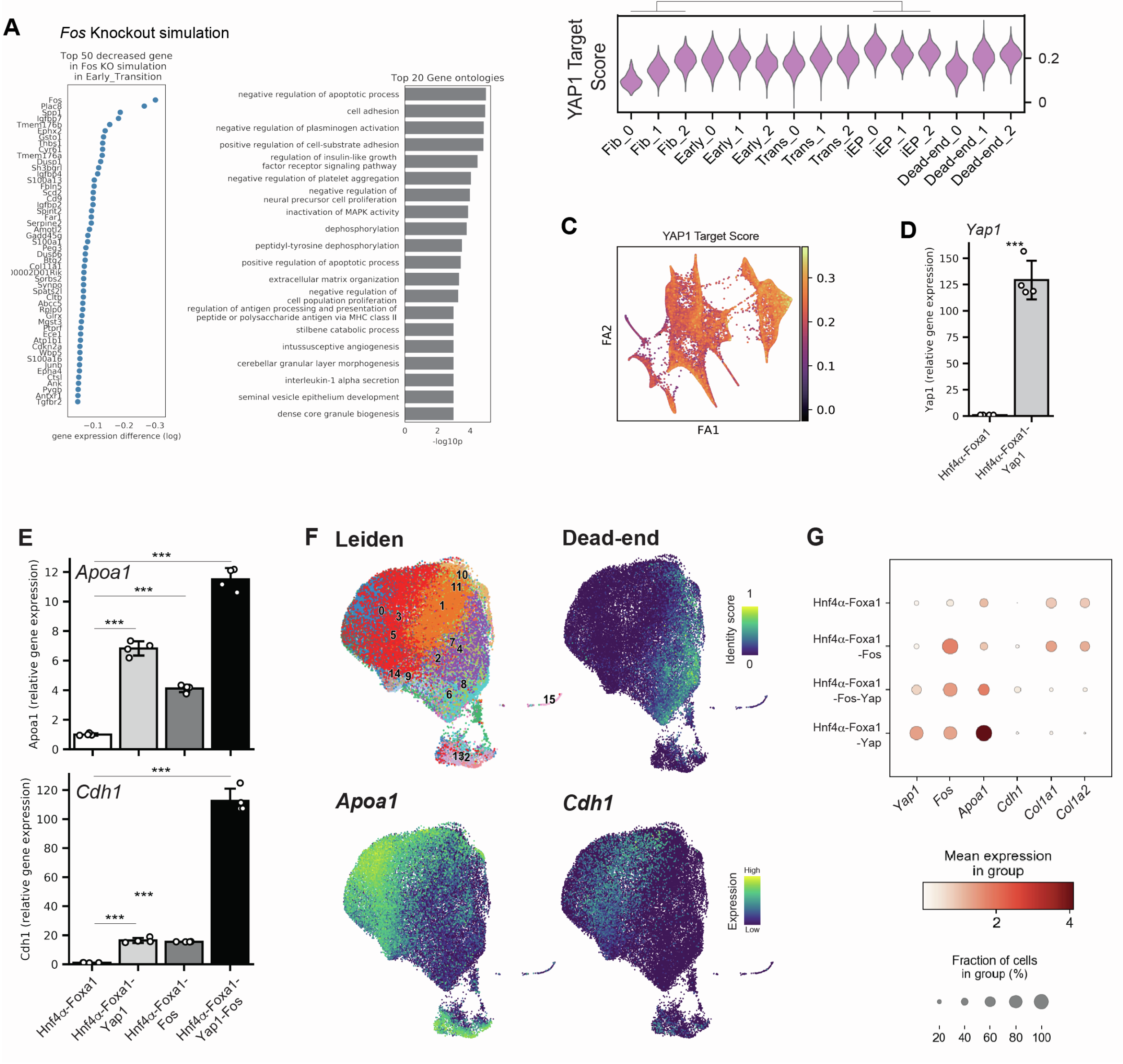
Supplemental Figure 4 (Related to Figure 4). CellOracle analysis of the role of Fos in fibroblast to iEP reprogramming. (A) Comparison of eigenvector centrality scores between the Fib_1 cluster GRN configuration and the GRN configurations of other clusters in relatively early stages of reprogramming. **(B)** Comparison of eigenvector centrality scores between iEP_1 and Dead-end_0 cluster GRN configurations. **(C-E)** Expression and network cartography of Jun family members, *Jun, Junb,* and *Jund*. **(F)** qRT-PCR of *Fos* expression in fibroblasts and iEPs, with and without cell dissociation prior to the assay, ** = $P < 0.01$, *t*-test, one-sided. **(G)** Analysis of *Fos* mRNA splicing state in the scRNA-seq data of iEP reprogramming to investigate the *Fos* mRNA maturation state: Violin plot for spliced *Fos* mRNA counts. **(H)** *t*-SNE plots of 9,914 expanded iEPs, cultured long-term, revealing fibroblast-like, intermediate, and three iEP subpopulations. Expression levels of *Apoa1* (marking typical iEPs)*, Col4a1* (fibroblast-like cells)*, Cdh1, Serpina1b* (hepatic-like iEPs)*,* and *Areg* (intestine-like iEPs) projected onto the *t*-SNE plot.

Supplemental Figure 4



Supplemental Figure 5 (Related to Figure 5). The role of Fos and Yap1 in fibroblast to iEP reprogramming. (A) Top 50 decreased genes in *Fos* knockout simulation in the early reprogramming transition (left) and GO analysis based on these genes (right). **(B)** Violin plot of YAP1 target gene scores across reprogramming, which are significantly enriched as reprogramming progresses (*** = $P < 0.001$, permutation test, one-sided). **(C)** Projection of YAP1 target gene scores onto the force-directed graph of reprogramming. **(D)** qRT-PCR assay for *Yap1* expression following addition of Yap1 and Fos to the Hnf4α-Foxa1 reprogramming cocktail (n = 4 independent biological replicates; *** = $P < 0.001$, ** = $P < 0.01$, *t*-test, one-sided), confirming Yap1 overexpression. **(E)** qRT-PCR assay for iEP marker expression (*Apoa1* and *Cdh1*) following addition of Yap1 and Fos to the Hnf4α-Foxa1 reprogramming cocktail (n = 4 independent biological replicates; *** = $P < 0.001$, ** = $P < 0.01$, *t*-test, one-sided). **(F)** Projection of Leiden cluster, dead-end identity scores, and gene expression information onto the state-fate UMAP embedding (from **Figure 5D, E**). **(G)** Expression of key marker genes for each reprogramming cocktail.

Supplemental Figure 5



Supplemental Table 1. Differentially expressed iEP markers from (Biddy et al., 2018). Top-ranked genes from CellOracle *in silico* perturbation are marked in red.

Supplemental Table 2. Top 50 CellOracle-inferred *Fos* targets across all reprogramming clusters. Confirmed YAP1 targets are highlighted in red.

Supplemental Table 3. Differential expression analysis of day 4 reprogrammed and dead-end destined clones. Genes in bold are also identified by CoSpar analysis. The right column shows TFs prioritized by CellOracle analysis. Genes in bold are also identified by CoSpar analysis.

Supplemental Methods

CellOracle. CellOracle is an integrative tool for GRN inference and network analysis. It consists of several steps: (1) base GRN construction using scATAC-seq data, (2) context-dependent GRN inference using scRNA-seq data, (3) network analysis, and (4) simulation of cell identity after perturbation. We created the algorithm in Python and designed it for use in the Jupyter notebook environment. CellOracle code is open source and available on GitHub (https://github.com/morris-lab/CellOracle), along with detailed function descriptions and tutorials. Further details can be found in the original preprint (Kamimoto et al., 2020).

Alignment and digital gene expression matrix generation. The Cell Ranger v6.0.1 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process data generated using the 10x Chromium platform. Cell Ranger processes, filters, and aligns reads generated with the Chromium single-cell RNA sequencing platform. This pipeline was used with a custom reference genome, created by concatenating the sequences corresponding to the *Hnf4α-t2a-Foxa1* transgene as a new chromosome to the mm10 genome. The unique UTRs in the *Hnf4α-t2a-Foxa1* transgene construct allowed us to monitor transgene expression. To create Cell Ranger compatible reference genomes, the references were rebuilt according to instructions from 10x (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references). To achieve this, we first created a custom gene transfer format (GTF) file, containing our transgenes, followed by indexing of the FASTA and GTF files, using Cell Ranger 'mkgtf' and 'mkref' functions. Following this step, the default Cell Ranger pipeline was implemented, then the filtered output data was used for downstream analyses.

CellTag clone calling

Reads containing the CellTag sequence were extracted from the processed and filtered BAM files produced by the 10x Genomics pipeline using our CellTagR pipeline: https://github.com/morris-lab/CellTagR. The resulting filtered CellTag UMI count matrix was then used for all downstream clonal and lineage analyses. The CellTag matrix was initially filtered by removing CellTags that do not appear on the allowlist generated for each CellTag plasmid library. Cells expressing more than 20 CellTags (likely corresponding to cell multiplets) and less than 2 CellTags per cell were filtered out. To identify clonally related cells, Jaccard analysis using the R package Proxy was used to calculate the similarity of CellTag signatures between cells. Clones were defined as groups of 2 or more related cells. Clones were called on cells pre-filtered for numbers of genes, UMIs, and mitochondrial RNA content.

Cell type classification with Capybara

Cells reprogrammed with Hnf4α-Foxa1, Hnf4α-Foxa1-Fos, Hnf4α-Foxa1-Yap1, and Hnf4α-Foxa1-Fos-Yap1 were classified using Capybara (Kong et al., 2022). Briefly, the single-cell datasets were processed, filtered, and clustered using Seurat, resulting in 35,241 cells (7,414 HF, 8,771 HF-Fos, 8,549 HF-Yap, 10,507 HF-Fos-Yap1). To construct a reference for cell-type classification, we obtained scRNA-seq data of biliary epithelial cells (BECs) and hepatocytes, before and after injury, from GSE125688 (Pepe-Mooney et al., 2019). We built a custom high-resolution reference by incorporating additional tissues from the MCA: fetal liver, MEFs, and embryonic mesenchyme. Following the construction of a high-resolution reference, we performed preprocessing on the reference and the samples, on which we then applied quadratic programming to generate the identity score matrices. Further, we categorized cells into discrete, hybrid, and unknown, calculated the empirical p-value matrices and performed binarization and classification. We calculated the percent composition of each cell type. Cells with hybrid identities were filtered and refined based on their identity scores and representation by more than 0.5% cells of the population. Code and documentation are available at:
https://github.com/morris-lab/Capybara.

Differential Gene Expression analysis. Genes differentially expressed between Day 4 reprogramming and dead-end destined cells were identified using Seurat *FindMarkers* command and subsetted to retain hits with an adjusted p-value of less than 0.05 (Bonferroni Correction).

Experimental Methods

Mice and derivation of mouse embryonic fibroblasts. Mouse Embryonic Fibroblasts were derived from E13.5 C57BL/6J embryos. (The Jackson laboratory: 000664). Heads and visceral organs were removed from E13.5 embryos. The remaining tissue was minced with a razor blade and then dissociated in a mixture of 0.05% Trypsin and 0.25% Collagenase IV (Life Technologies) at 37°C for 15 minutes. After passing the cell slurry through a 70μM filter to remove debris, cells were washed and then plated on 0.1% gelatin-coated plates in DMEM supplemented with 10% FBS (Sigma-Aldrich), 2mM L-glutamine and 50mM β-mercaptoethanol (Life Technologies). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee.

Retrovirus Production. Retroviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pGCDN-Sam construct containing Hnf4α-t2a-Foxa1/Fos/Yap1, along with packaging construct pCL-Eco (Imgenex). Virus was harvested 48hr and 72hr after transfection and applied to cells immediately following filtering through a low-protein binding 0.45μM filter.

Lentiviral constructs and lentivirus production. Lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the envelope construct pCMV-VSV-G (Addgene plasmid 8454), the packaging construct pCMV-dR8.2 dvpr (Addgene plasmid 8455), and the shRNA expression vector for the respective candidate TF to be knocked down. The shRNA expression vectors (with the TRC2 pLKO.5 backbone) were obtained directly from Millipore-Sigma or cloned into the empty backbone using oligonucleotides (Integrated DNA Technologies). The sequences of shRNA used are SHC202 (non-target shRNA control) CAACAAGATGAAGAGCACCAA; *Eno1* GGCACAGAGAATAAATCTAAA; *Fos* ATCCGAAGGGAACGGAATAAG; *FosB* ATGACGGAAGGACCTCCTTTG; *Foxd2* AGATCATGTCCTCCGAGAGCT *Id1* GAGCTGAACTCGGAGTCTGAA; *Klf2* GACCGATTGTATTTCTATAAG *Klf4* CATGTTCTAACAGCCTAAATG; *Klf15* CTACCCTGGAGGAGATTGAAG. Virus was harvested 48hr and 72hr after transfection and applied to cells following filtering through a low-protein binding 0.45μm filter. For the generation of the complex CellTag library, lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct, along with packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455), and pCMV-VSVG (Addgene plasmid 8454), as in (Biddy et al., 2018; Guo et al., 2019; Jindal et al., 2022).

Generation and collection of iEPs. Mouse embryonic fibroblasts (< passage 6) were converted to iEPs as in (Biddy et al., 2018), modified from (Sekiya and Suzuki, 2011). Briefly, we transduced cells every 12hr for two days, with fresh Hnf4α-t2a-Foxa1 retrovirus, in the presence of 4mg/ml Protamine Sulfate (Sigma-Aldrich), followed by culture on 0.1% gelatin-treated plates for one week in hepato-medium (DMEM: F-12, supplemented with 10% FBS, 1 mg/ml insulin (Sigma-Aldrich), dexamethasone (Sigma-Aldrich), 10mM nicotinamide (Sigma-Aldrich), 2mM L-glutamine, 50mM β-mercaptoethanol (Life Technologies), and penicillin/streptomycin, containing 20 ng/ml epidermal growth factor (Sigma-Aldrich). After seven days of culture, the cells were transferred onto plates coated with 5μg/cm^2 Type I rat collagen (Gibco, A1048301). For single-cell processing, 30,000 reprogrammed, expanded iEPs were collected and fixed in methanol, as previously described in (Alles et al., 2017). Briefly, cells were collected and washed in Phosphate

Buffered Saline (PBS), followed by resuspension in ice-cold 80% Methanol in PBS, with gentle vortexing. These cells were stored at -80°C for up to three months and processed on the 10x platform (below). For the state-fate experiments, we followed the above protocol with some slight modifications. We transduced cells every 12hr for two days, with fresh Hnf4α-t2a-Foxa1 retrovirus and added CellTagging lentivirus on the final round of transduction. After 12hr, cells were washed and expanded in hepato-medium for four days, at which point the cells were dissociated and 25% of the population profiled by scRNA-seq. The remaining population was replated, and additional samples were profiled on days 10 and 28.

Colony formation assays. Mouse *Fos* and *Yap1* were cloned from iEPs into the retroviral vector, pGCDNSam (Sekiya and Suzuki, 2011), and retrovirus produced as above. For comparative reprogramming experiments, mouse embryonic fibroblasts (2x10^5/well of a 6-well plate) were serially transduced over 72hr (as above). In control experiments, virus produced from an empty vector control expressing only GFP was added to the Hnf4α-Foxa1 reprogramming cocktail. Virus produced from the *Fos* and *Yap1* IRES-GFP constructs was added to the standard Hnf4α and Foxa1 cocktail. Cells underwent reprogramming for two weeks and were processed for colony formation assays: cells were fixed on the plate with 4% PFA, permeabilized in 0.1% Triton-X100 then blocked with the Mouse on Mouse Elite Peroxidase Kit (Vector PK-2200). Primary antibody, mouse anti-E-Cadherin (1:100, BD Biosciences), was applied for 30 min before washing and processing with the VECTOR VIP Peroxidase Substrate Kit (Vector SK-4600). Colonies were visualized on a flatbed scanner, adding heavy cream to each well to increase image contrast. Colonies were counted using our automated colony counting tool: https://github.com/morris-lab/Colony-counter. *Fos and Yap1* overexpression was confirmed by harvesting RNA from Hnf4α-Foxa1 and Hnf4α-Foxa1-Fos/Yap1-transduced cells (RNeasy kit, Qiagen). Following cDNA synthesis (Maxima cDNA synthesis kit, Life Tech), qRT-PCR was performed to quantify *Fos/Yap1* overexpression (TaqMan Probes: Gapdh Mm99999915_g1; *Cdh1* Mm01247357_m1; *Apoa1* Mm00437569_m1; *Fos* Mm00487425_m1; *Yap1* Mm01143263_m1; TaqMan qPCR Mastermix, Applied Biosystems).

Colony formation assays for TF knockdowns were conducted similarly, with the following modifications. To initiate reprogramming, mouse embryonic fibroblasts (75x10^3/well of a 6-well plate) were serially transduced over 72hr (as above). Lentivirus produced from the non-target shRNA control and the respective TF knockdown shRNA constructs was then added at 84hr and 96hr (only added at 96hr for the initial screen). At 120hr, cells were seeded for colony formation assays (40x10^3cells/well of a 6-well plate), which were then processed for colony formation on day 14 as above. The remaining cells from each sample were seeded for harvesting RNA for qPCR on day 14, as above. In the initial screen, cells from each sample were split equally and seeded in 6 well plates for colony formation and RNA extraction at 15 days following reprogramming initiation. For *Fos* and *Fosb* knockdowns, mouse embryonic fibroblasts (120x10^3 in a 6-cm dish) were transduced with the respective shRNA lentivirus at 24hr and 36hr post-seeding. qPCR confirmation was performed at 72hr post-seeding. TaqMan Probes used: *Actb* Mm02619580_g1; *Eno1* Mm01619597_g1; *Fos* Mm00487425_m1; *Fosb* Mm00500401_m1; *Foxd2* Mm00500529_s1; *Id1* Mm00775963_g1; *Klf2* Mm00500486_g1; *Klf4* Mm00516104_m1; *Klf15* Mm00517792_m1.

CRISPR/Cas9 *Fos* Knockout

The *Fos* knockouts were performed as part of a larger screen, using Perturb-seq as previously described (Dixit et al., 2016). The protocol was modified, as outlined below, to apply the strategy to our experimental system:

(1) Vector backbone and gene barcode pool construction: For Perturb-seq experiments, we used a lentivirus vector to express guide RNAs and gene barcodes (GBC). The lentivirus vector backbone contains an antiparallel cassette containing a guide RNA and GBC. In the original perturb-seq paper, the authors used pPS and pBA439 to construct the guide RNA-GBC vector pool. Here, we modified pPS and pBA439 to generate the pPS2 vector, in which the Blasticidin-t2a-BFP gene replaced the Puromycin-t2a-BFP gene. We constructed the guide RNA-GBC vector using a multi-step cloning strategy: First, we synthesized dsDNA, via PCR, for a random GBC pool. We purified the PCR product with AMPure XP SPRI beads. We inserted the purified GBC pool into the pPS2 vector at the EcoRI site in the 3’ UTR of the Blasticidin-t2a-BFP gene. We used the product of Gibson assembly for transformation into DH5α competent cells (NEB: C2987H). Transformed cells were cultured directly in LB. We extracted plasmid DNA to yield the pPS2-GBC pool.

(2) Guide RNA cloning. We designed guide RNAs using https://zlab.bio/guide-design-resources. We synthesized oligo DNA for each guide RNA. Oligo DNA pairs were annealed and inserted into the pPS2-GBC vector following BsmB1 digestion. After isolation and growth of single colonies, plasmid DNA was extracted and sanger DNA sequenced; sequences of the guide RNA inserted site and GBC site were used to construct a gRNA/GBC reference table:

Fos_sg0	CAGCCGACTGAACGCGTTATTC
Fos_sg1	CATATATCAAAGATGAACATTG
Fos_sg2	TCAAGGCTGTAATTTCTTGGGC
empty0	TTGATGAACTGCGCTAGCGAGG
empty1	AAGAGCGGCTCGCAAGGGAAAA
empty2	AGTAGGATACGTGGAGTTAATA

(3) Lentivirus guide RNA pool generation. An equal amount of DNA for each pPS2-guide RNA vector was mixed to generate the plasmid pool. Three control vectors were also mixed with this plasmid vector pool; the weight ratio of each pPS2-guide vector to each control vector was 1:4. We used this mixed DNA pool for lentivirus production. Lentiviral particles were produced by transfecting 293T-17 cells (ATT: CRL-11268) with the pPS-guide RNA-GBC constructs, along with the packaging plasmid, psPAX2 (https://www.addgene.org/12260/), and pMD2.G (https://www.addgene.org/12259/).

(4) Cell culture for Perturb-seq. We transduced reprogrammed iEP cells with retrovirus carrying Cas9 (MSCV-Cas9-Puro). The cells were treated with Puromycin (4 μg/ml) for four days to eliminate non-transduced cells. iEP-Cas9 cells were transduced with the lentivirus guide RNA pool for 24 hours. The concentration of lentivirus was pre-determined to target 10~20% transduction efficiency. After four days of cell culture, we flow sorted BFP-positive cells to purify transduced cells. Cells were cultured for a further 72 hours and fixed with methanol, as previously described (Alles et al., 2017).

(5) GBC amplification and sequencing. Following library preparation on the 10x Chromium platform (below), we PCR amplified the GBC. The amplification was performed according to the original perturb-seq paper (Dixit et al., 2016), but we modified the PCR primer sequence for the Chromium single-cell library v2 kit:

P7_ind_R2_BFP_primer:
CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTC
CGATCTTAGCAAACTGGGGCACAAGC
P5_partial_primer: AATGATACGGCGACCACCGA
GBG_Amp_F: GCTGATCAGCGGGTTTAAACGGGCCCTCTAGG

GBG_Amp_R: CGCGTCGTGACTGGGAAAACCCTGGCGAATTG

GBC_Oligo:

TTAAACGGGCCCTCTAGGNNNNNNNNNNNNNNNNNNNNNNNNNCAATTCGCCAGGGTTTTCCC

Following amplification, we purified the PCR product with AMPure XP SPRI beads. The purified sample was sequenced on the Illumina Mi-seq platform.

(6) Alignment of cell barcode/GBC. For preprocessing of Perturb-seq metadata, we used MIMOSCA, a computational pipeline to analyze perturb-seq data (https://github.com/asncd/MIMOSCA). First, the reference table for the cell barcode/GBC pair was generated from Fastq files. The data table was converted into the guide RNA/cell barcode table using the guide RNA-GBC reference table. This metadata was integrated into the scRNA-seq data. The guide metadata was processed with an EM-like algorithm in MIMOSCA to filter out unperturbed cells computationally, as previously described (Dixit et al., 2016).

10x procedure. For single-cell library preparation on the 10x Genomics platform, we used: the Chromium Single Cell 3′ Library & Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236), and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer’s instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Prior to cell capture, methanol-fixed cells were placed on ice, then spun at 3000rpm for 5 minutes at 4°C, followed by resuspension and rehydration in PBS, according to (Alles et al., 2017). 17,000 cells were loaded per lane of the chip, aiming to capture 10,000 single-cell transcriptomes. The resulting cDNA libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 2500.

References

Alles, J., Karaiskos, N., Praktiknjo, S.D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., et al. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol *15*, 44. https://doi.org/10.1186/s12915-017-0383-5.

Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219–224. https://doi.org/10.1038/s41586-018-0744-4.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with

Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.

Guimerà, R., and Amaral, L.A.N. (2005). Cartography of complex networks: modules and universal roles. Journal of Statistical Mechanics: Theory and Experiment *2005*, P02001. https://doi.org/10.1088/1742-5468/2005/02/P02001.

Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G.C., Yang, X., Kirita, Y., and Morris, S.A. (2019). CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. Genome Biol *20*, 90. https://doi.org/10.1186/s13059-019-1699-y.

Jindal, K., Adil, M.T., Yamaguchi, N., Wang, H.C., Yang, X., Kamimoto, K., Rivera-Gonzalez, G.C., and Morris, S.A. (2022). Multiomic single-cell lineage tracing to dissect fate-specific gene regulatory programs. BioRxiv 2022.10.23.512790. https://doi.org/10.1101/2022.10.23.512790.

Kamimoto, K., Hoffmann, C.M., and Morris, S.A. (2020). CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. BioRxiv 2020.02.17.947416. https://doi.org/10.1101/2020.02.17.947416.

Kong, W., Fu, Y.C., Holloway, E.M., Garipler, G., Yang, X., Mazzoni, E.O., and Morris, S.A. (2022). Capybara: A computational tool to measure cell identity and fate transitions. Cell Stem Cell *29*, 635-649.e11. https://doi.org/10.1016/J.STEM.2022.03.001.

Pepe-Mooney, B.J., Dill, M.T., Alemany, A., Ordovas-Montanes, J., Matsushita, Y., Rao, A., Sen, A., Miyazaki, M., Anakk, S., Dawson, P.A., et al. (2019). Single-Cell Analysis of the Liver Epithelium Reveals Dynamic Heterogeneity and an Essential Role for YAP in Homeostasis and Regeneration. Cell Stem Cell *25*, 23-38.e8. https://doi.org/10.1016/J.STEM.2019.04.004.

Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature *475*, 390–393. https://doi.org/10.1038/nature10263.