1 **Gene Regulatory Network Reconfiguration in Direct Lineage Reprogramming**

2

3 Kenji Kamimoto[1,2,3], Mohd Tayyab Adil[1,2,3], Kunal Jindal[1,2,3], Christy M. Hoffmann[1,2,3], Wenjun

4 Kong[1,2,3,4], Xue Yang[1,2,3], and Samantha A. Morris[1,2,3,*].

5

6 [1]Department of Developmental Biology; [2]Department of Genetics; [3]Center of Regenerative

7 Medicine. Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus

8 Box 8103, St. Louis, MO 63110, USA. [4]Current address: Calico Life Sciences

9 LLC. *Correspondence: s.morris@wustl.edu

10

11 **Summary**

12 In direct lineage reprogramming, transcription factor (TF) overexpression reconfigures Gene

13 Regulatory Networks (GRNs) to convert cell identities between fully differentiated cell types. We

14 previously developed CellOracle, a computational pipeline that integrates single-cell

15 transcriptome and epigenome profiles to infer GRNs. CellOracle leverages these inferred GRNs

16 to simulate gene expression changes in response to TF perturbation, enabling network re-

17 configuration during reprogramming to be interrogated *in silico.* Here, we integrate CellOracle

18 analysis with lineage tracing of fibroblast to induced endoderm progenitor (iEP) conversion, a

19 prototypical direct lineage reprogramming paradigm. By linking early network state to

20 reprogramming success or failure, we reveal distinct network configurations underlying different

21 reprogramming outcomes. Using these network analyses and *in silico* simulation of TF

22 perturbation, we identify new factors to coax cells into successfully converting cell identity,

23 uncovering a central role for the AP-1 subunit Fos with the Hippo signaling effector, Yap1.

24 Together, these results demonstrate the efficacy of CellOracle to infer and interpret cell-type-

25 specific GRN configurations at high resolution, providing new mechanistic insights into the

26 regulation and reprogramming of cell identity.

27

28 **Keywords:** Gene perturbation simulation; cell fate prediction; gene regulatory networks;

29 machine learning; direct lineage reprogramming; single-cell analysis

30

31 **Introduction**

32   Advances over the past half-century, such as nuclear transfer (Gurdon et al., 1958) and

33 factor-mediated reprogramming (Takahashi and Yamanaka, 2006), have revealed the remarkable

34 plasticity of cell identity. Cells reprogrammed to pluripotency can be directed to differentiate

35 toward desired target populations by recapitulating embryonic development *in vitro,* although this

36 approach is inefficient and produces heterogeneous populations of developmentally immature

37 cells. "Direct lineage reprogramming" aims to directly transform cell identity between fully

38 differentiated somatic states via the forced expression of select transcription factors (TFs). Using

39 this approach, fibroblasts have been directly converted into many clinically valuable cell types

40 (Cohen and Melton, 2011). These protocols are currently limited because only a fraction of cells

41 convert to the target cell type and remain developmentally immature or incompletely specified

42 (Morris and Daley, 2013). Therefore, the resulting cells are generally unsuitable for therapeutic

43 application and have limited utility for disease modeling and drug screening *in vitro*, where fully

44 differentiated and functional cells are highly sought-after.

45   Gene Regulatory Networks (GRNs) represent the complex, dynamic molecular

46 interactions that act as critical determinants of cell identity. These networks describe the intricate

47 interplay between transcriptional regulators and multiple cis-regulatory DNA sequences, resulting

48 in the precise spatial and temporal regulation of gene expression (Davidson and Erwin, 2006).

49 Systematically delineating GRN structures enables a logic map of regulatory factor cause-effect

50 relationships to be mapped (Materna and Davidson, 2007). In turn, this knowledge supports a

51 better understanding of how cell identity is determined and maintained, informing new strategies

52 for cellular reprogramming to support disease modeling or cell-based therapeutic approaches.

53   We previously described CellOracle, a computational pipeline for GRN inference via the

54 integration of different single-cell data modalities (Kamimoto et al., 2020). CellOracle overcomes

55 current challenges in GRN inference by using single-cell transcriptomic and chromatin

56 accessibility profiles, integrating prior biological knowledge via regulatory sequence analysis to

57 infer transcription factor (TF)-target gene interactions. Moreover, we designed CellOracle to apply

58 inferred GRNs to simulate gene expression changes in response to TF perturbation. This unique

59 feature enables inferred GRN configurations to be interrogated *in silico*, facilitating their

60 interpretation. We have benchmarked CellOracle against ground-truth TF-gene interactions,

61 demonstrating its efficacy to recapitulate known regulatory changes across hematopoiesis

62 (Kamimoto et al., 2020). Further, we have applied CellOracle to predict TFs regulating medium

63 spiny neuron maturation in human fetal striatum development (Bocchi et al., 2021). Other groups

64 have successfully used the method to investigate mouse and human T-cell differentiation (Chopp

65 et al., 2020; Nie et al., 2022), T-cell dysfunction in glioblastoma (Ravi et al., 2022), and pharyngeal

66 organ development (Magaletta et al., 2022).

67   Here, we apply CellOracle to interrogate GRN reconfiguration during the direct lineage

68 reprogramming of fibroblasts to induced endoderm progenitors (iEPs), a prototypical TF-mediated

69   fate conversion protocol. Via single-cell resolution lineage tracing, we previously demonstrated

70   that this protocol comprises two distinct trajectories leading to reprogrammed and dead-end

71   states (Biddy et al., 2018). In this study, we expand on this lineage tracing strategy to

72   experimentally define state-fate relationships, supporting the inference of early network states

73   associated with defined reprogramming outcomes. These analyses reveal the early GRN

74   configurations associated with the successful conversion of cell identity. Using principles of graph

75   theory to identify critical nodes in conjunction with *in silico* simulation predicts several novel

76   regulators of reprogramming. We experimentally validate these predictions via experimental TF

77   perturbation: knockdown, overexpression, and Perturb-seq-based knockout. We also

78   demonstrate that one of these TFs, *Fos*, plays roles in both iEP reprogramming and maintenance,

79   where interrogation of inferred *Fos* targets reveals a putative role for AP1-Yap1 in fibroblast to

80   iEP conversion. We experimentally validate these findings to demonstrate that Fos and Yap1

81   overexpression significantly enhances reprogramming efficiency. Together, these results

82   demonstrate the efficacy of CellOracle to infer and interpret cell-type-specific GRN configurations

83   at high resolution, enabling new mechanistic insights into the regulation and reprogramming of

84   cell identity. CellOracle code and documentation are available at

85   https://github.com/morris-lab/CellOracle.

86

87   **Results**

88   **CellOracle GRN Inference applied to direct lineage reprogramming**

89   CellOracle is designed to infer GRN configurations to reveal how networks are rewired

90   during the establishment of defined cellular identities and states, highlighting known and putative

91   regulatory factors of fate commitment. CellOracle overcomes population heterogeneity by

92   leveraging single-cell genomic data, enabling accurate inference of the GRN dynamics underlying

93   complex biological processes (Kamimoto et al., 2020). In the first step of the CellOracle pipeline,

94   single-cell chromatin accessibility data (scATAC-seq) is used to assemble a 'base' GRN structure,

95   representing a list of all potential regulatory genes associated with each defined DNA sequence.

96   This   step   leverages   the   transcriptional   start   site   (TSS)   database

97   (http://homer.ucsd.edu/homer/ngs/annotation.html) and Cicero, an algorithm that identifies co-

98   accessible scATAC-seq peaks (Pliner et al., 2018), to identify accessible promoters/enhancers.

99   The DNA sequence of these regulatory elements is then scanned for TF binding motifs, repeating

100   this task for all regulatory sequences, to generate a base GRN structure of all potential regulatory

101   interactions (**Figure 1A, B**).

102   The second step in the CellOracle pipeline uses scRNA-seq data to convert the base GRN

103   into context-dependent GRN configurations for each defined cell cluster. Removal of inactive

104   connections refines the base GRN structure, selecting the active edges representing regulatory

105   connections associated with a specific cell type or state (**Figure 1C**). For this process, we

106   leverage regularized machine learning regression models (Camacho et al., 2018), primarily to

107   select active regulatory genes and to obtain their connection strength (**Figure S1A**). CellOracle

108   builds a machine learning model that predicts target gene expression from the expression levels

109   of the regulatory genes identified in the prior base GRN refinement step. After fitting models to

110   sample data, CellOracle extracts gene-gene connection information by analyzing model

111   variables. With these values, CellOracle prunes insignificant or weak connections, resulting in a

112   cell-type/state-specific GRN configuration (**Figure 1D**). Here, we apply CellOracle to infer GRN

113   reconfiguration during TF-mediated direct lineage reprogramming.

114   We previously investigated mouse embryonic fibroblast (MEF) to induced endoderm

115   progenitor (iEP) reprogramming, induced via the forced expression of two TFs: Hnf4α and Foxa1

116   (**Figure 1E;** (Biddy et al., 2018; Morris et al., 2014)). iEP generation represents a prototypical

117   lineage reprogramming protocol, which, like most conversion strategies, is inefficient and lacks

118   fidelity. Initially reported as hepatocyte-like cells, the resulting cells can functionally engraft the

119   liver (Sekiya and Suzuki, 2011). However, we demonstrated that these cells also harbor intestinal

120   identity and can functionally engraft the colon in a mouse model of acute colitis, prompting their

121   re-designation as iEPs (Guo et al., 2019; Morris et al., 2014). More recently, we have shown that

122   iEPs transcriptionally resemble injured biliary epithelial cells (BECs) and exhibit BEC-like behavior

123   in 3D-culture models (Kong et al., 2022). Building on these studies, our single-cell lineage tracing

124   of this protocol revealed two distinct trajectories arising during MEF to iEP conversion: one to a

125   successfully reprogrammed state, and one to a dead-end state, where cells fail to fully convert to

126   iEPs (Biddy et al., 2018). Although we identified factors to improve the efficiency of

127   reprogramming, mechanisms of cell fate conversion from the viewpoint of GRN reconfiguration

128   remain unknown.

129   Our previously published MEF to iEP reprogramming scRNA-seq dataset consists of eight

130   time points collected over 28 days (*n* = 27,663 cells) (Biddy et al., 2018). We reprocessed this

131   dataset using partition-based graph abstraction (PAGA; (Wolf et al., 2019)), manually annotating

132   15 clusters based on marker gene expression and PAGA connectivity (**Figure 1F; S1B-D**). After

133   successfully initiating conversion, cells diverge down one of two trajectories: one leading to a

134   successfully reprogrammed state, and one to a dead-end state. Relative to reprogrammed cells,

135   dead-end cells only weakly express iEP markers, *Cdh1,* and *Apoa1*, accompanied by higher

136 expression levels of fibroblast marker genes, such as *Col1a2* (**Figure 1F; S1B, C**). Using

137 CellOracle, we inferred GRN configurations for each cluster, calculating network connectivity

138 scores to analyze GRN dynamics during lineage reprogramming.

139

140 **Analysis of network reconfiguration during reprogramming**

141 We initially assess the network configuration associated with the exogenous

142 reprogramming TFs, Hnf4α and Foxa1, focusing on the strength of their connections to target

143 genes. Hnf4α and Foxa1 receive a combined score in these analyses since they are expressed

144 as a single transcript that produces two independent factors via 2A-peptide-mediated cleavage

145 (Liu et al., 2017). Network strength scores show significantly stronger connectivity of *Hnf4α-Foxa1*

146 to its inferred target genes in the early stages of reprogramming, followed by decreasing

147 connection strength in later conversion stages (Early_2 vs. iEP_2: $P < 0.001$, Wilcoxon Test;

148 **Figure 1G**). We next evaluated the inferred GRN structures using traditional graph theory

149 methods. We examined: 1) Degree centrality of each gene, a straightforward measure reporting

150 how many edges are connected to a node directly; 2) Eigenvector centrality, a measure of

151 influence via connectivity to other well-connected genes (Klein et al., 2012). *Hnf4α-Foxa1*

152 receives high degree centrality and eigenvector centrality scores in the early phases of lineage

153 conversion, gradually decreasing as reprogramming progresses (**Figure 1H**). In agreement with

154 a central role for the transgenes early in reprogramming, network cartography analysis (Guimerà

155 and Amaral, 2005) classified *Hnf4α-Foxa1* as a prominent "connector hub" in the early_2 cluster

156 network configuration (**Figure 1I; S1E**). Together, these analyses reveal that *Hnf4α-Foxa1*

157 network configuration connectivity and strength peak in early reprogramming phases.

158 Next, we analyzed the *Hnf4α-Foxa1* network configuration in later conversion stages,

159 following bifurcation into reprogrammed and dead-end trajectories (**Figure 1F; S1B-D**). The

160 reprogrammed clusters (iEP_0, iEP_1, iEP_2) exhibit stronger network connectivity scores,

161 relative to the dead-end clusters 1 and 2 (**Figure 1G;** iEP vs. Dead-end; $P < 0.001$, Wilcoxon

162 Test). We also identify a smaller dead-end cluster (Dead-end_0); cells within this cluster only

163 weakly initiate reprogramming, retaining robust fibroblast gene expression signatures and

164 expressing significantly lower levels of reprogramming initiation markers such as *Apoa1* (**Figure**

165 **S1C;** $P < 0.001$, permutation test). This cluster also exhibits significantly lower *Hnf4α-Foxa1*

166 connectivity scores relative to Dead-end_1 and 2 (**Figure 1G;** $P < 0.001$, Wilcoxon Test;),

167 accompanied by lower degree centrality and eigenvector centrality scores (**Figure 1H**). However,

168 CellTag lineage data reveals that the majority of the cells (93% of tracked cells) on this unique

169 path derive from a single clone, representing a rare reprogramming event captured due to clonal

170 expansion (**Figure S1F**).

171 We next turned to global GRN reconfiguration to identify candidate TFs reprogramming

172 initiation. Comparing degree centrality scores between fibroblast and early reprogramming

173 clusters reveals differential connectivity of a handful of key TFs. For example, *Hes1*, *Eno1, Fos,*

174 *Foxq1,* and *Zfp57* receive relatively high degree centrality scores in the early reprogramming

175 clusters, whereas *Klf2* and *Egr1* degree centrality increases in later transition stages (**Figure 1J**).

176 These factors remain highly connected on the reprogramming trajectory relative to the dead-end

177 (**Figure 1K**), suggesting that the GRN configurations controlling reprogramming outcome are

178 remodeled at initiation.

179 Altogether, the MEF to iEP reprogramming network analysis presented here suggests that

180 *Hnf4α-Foxa1* function peaks at conversion initiation. These early, critical changes in GRN

181 configuration determine reprogramming outcome, with dysregulation or loss of this program

182 leading to dead-ends, where cells either do not successfully initiate or complete reprogramming.

183 This hypothesis is consistent with our previous CellTag lineage tracing, showing the

184 establishment of reprogramming outcomes from early stages of the conversion process (Biddy et

185 al., 2018). We next performed new experimental lineage tracing to capture cells at reprogramming

186 initiation to investigate further how early GRN configuration relates to the successful generation

187 of iEPs.

188

189 **Clonal tracing links early network state to reprogramming fate**

190 Barcoding and tracking cells via scRNA-seq represents a powerful method to investigate

191 how the early molecular state of a cell relates to its eventual fate (Biddy et al., 2018; Weinreb et

192 al., 2020). Cells are labeled with combinations of heritable random barcodes, CellTags, delivered

193 using lentivirus, enabling cells to be uniquely labeled and tracked over time; cells sharing identical

194 barcodes are identified as clonal relatives; thus, early cell state can be directly linked to

195 reprogramming outcome (Biddy et al., 2018; Kong et al., 2020; **Figure 2A**). However, our previous

196 lineage tracing study was not designed to maximize the capture of clones early in reprogramming;

197 thus, the 30-cell minimum requirement of CellOracle for GRN inference was not met. Here, we

198 performed new lineage tracing experiments to associate early-stage cells with reprogramming

199 outcome.

200 Cells were reprogrammed with *Hnf4α-Foxa1,* as above, and CellTagged at the end of the

201 reprogramming TF transduction period. After four days of expansion (reprogramming day 4), we

202 collected 25% of the cell population for scRNA-seq, reseeding the remaining cells. A total of

203   24,799 cells were sequenced: 8,440 at day 4, 4,836 at day 10, and 11,523 at day 28 (**Figure 2B,**

204   **C**). Using our previous method to score cell identity along with established marker gene

205   expression (Biddy et al., 2018), we identify reprogrammed and dead-end reprogramming

206   outcomes (reprogrammed n = 1,895; dead-end n = 6,492; **Figure 2D; S2A, B**). Next, using clonal

207   information, we identify the day 4 clones whose day 10 and day 28 descendants are significantly

208   enriched or depleted of successfully reprogrammed cells. From CellTag processing (**Methods**),

209   we recovered 1,158 clones, containing a total of 10,927 cells across all time points. Using

210   randomized testing, we identified two groups of day 4 iEPs: iEP-enriched (64 cells in 7 clones)

211   and iEP depleted (59 cells in 39 clones), from which reprogramming and dead-end trajectories

212   stem (**Figure 2F**), reproducing our earlier observations (Biddy et al., 2018).

213   Pooling the day 4 clones by outcome, we meet the minimum number of cells required for

214   GRN inference (**Figure S2C**). We first compared the global GRN configurations for each of these

215   states relative to MEFs, to assess early GRN reconfiguration on each trajectory. For example,

216   comparing degree centrality between day 4 cells destined to reprogram and native fibroblasts

217   agrees with our above analysis comparing early transition to fibroblast states (**Figure 1J**),

218   showing high connectivity of similar factors, such as *Klf6, Klf9,* and *Mef2a*, in fibroblasts and *Fos*,

219   *Egr1*, and *Foxq1* in day 4 reprogrammed destined clones (**Figure 2G, left**). Additional highly-

220   connected TFs, receiving relatively high degree centrality scores, also emerge in this

221   reprogramming group, including the known induced pluripotency factor, *Klf4* (Takahashi and

222   Yamanaka, 2006) in addition to *Klf5, Cebpb, Mybl2,* and *Foxk2*, amongst other TFs. The

223   appearance of several additional factors here is likely due to assessing the early cells with known

224   reprogramming descendants rather than the early reprogramming cluster as a whole, in which

225   many cells will not successfully reprogram, highlighting how these state-fate experiments can

226   further dissect population heterogeneity.

227   Indeed, the state-fate experimental design allows us to compare those early cells destined

228   to reprogram vs. early cells that fail to reprogram, for which clonal information is essential. A

229   comparison of these two groups reveals subtle differences in GRN configuration that lead to

230   different reprogramming outcomes, with *Klf6, Egr3, Tfapb2,* and *Foxs1* demonstrating higher

231   connectivity in cells failing to fully reprogram, in contrast to *Fos, Cebpb, Klf5,* and *Junb* in cells

232   destined to attain full iEP identity (**Figure 2G, right**). Overall, the new experimental state-fate

233   analysis presented here supports the network analysis of our previous time course, revealing the

234   highly connected fibroblasts TFs that are decoupled upon reprogramming initiation. These factors

235   represent potential targets to extinguish fibroblast identity. Further, we identify many TFs that are

236   highly connected from early stages on the successful reprogramming trajectory, representing

237   potential candidates to improve iEP yield and understand how cell identity is maintained and

238   respecified more broadly. We next use CellOracle's *in silico* perturbation function to identify

239   putative regulators of reprogramming in a systematic, unbiased manner.

240

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

242   **reprogramming**

243   While network structure can point to how gene regulation changes during reprogramming,

244   it offers a static picture that does not necessarily provide functional insight. CellOracle bridges

245   this gap by using its unique GRN inference model to interrogate networks to gain mechanistic

246   insight into how specific TFs regulate cell identity (Kamimoto et al., 2020). CellOracle simulates

247   the transition of cell identity following candidate TF perturbation (knockout or overexpression),

248   using cluster-specific GRNs to model subsequent expression changes in regulated genes. The

249   simulated values are then converted into a transition vector map and visualized in the dimensional

250   reduction space, enabling an intuitive interpretation of how a candidate TF regulates cell identity

251   (Kamimoto et al., 2020); **Figure 3A-C; S3A-C; Methods**). This approach allows factors to be

252   ranked for further experimental investigation, as detailed below.

253   *In silico* TF perturbation comprises four steps: 1) GRN configurations are constructed (as

254   in **Figure 1A**); 2) Using these GRN models, shifts in target gene expression in response to TF

255   perturbation are calculated. This step applies the GRN model as a function to propagate the shift

256   in gene expression rather than the absolute gene expression value, representing TF-to-target

257   gene signal flow. This signal is propagated iteratively to calculate the broad, downstream effects

258   of TF perturbation, allowing the global transcriptional 'shift' to be estimated (**Figure S3A, B**); 3)

259   The probability of a cell identity transition is estimated by comparing this gene expression shift to

260   the gene expression of local neighbors (**Figure S3C**); 4) The transition probability is converted

261   into a weighted local average vector to represent the simulated directionality of cell state transition

262   for each cell upon candidate TF perturbation. This final step converts the simulation results into a

263   2D vector map, enabling robust predictions by mitigating the effect of errors or noise derived from

264   scRNA-seq data and the preceding simulation (**Figure 3B middle; S3C**). The resulting small-

265   length vectors allow the directionality of cell identity transitions to be feasibly predicted, rather

266   than interpreting long-ranging terminal effects from initial states.

267   To enable the simulation results to be assessed in a systematic and unbiased manner, we

268   consider the changes in cell identity induced by reprogramming, together with the predicted

269   effects from the perturbation. Taking the relatively densely sampled time course from Biddy et al.,

270   2018, we use semi-supervised Monocle analysis (Trapnell et al., 2014) to order cells in

271   pseudotime based on the expression of the fibroblast marker *Col1a2* and the iEP marker *Apoa1*,
272   capturing the distinctive reprogramming and dead-end trajectories as distinguished by their
273   respective lineage restricted-clones (n = 48,515 cells, 2 independent biological replicates; **Figure
274   3A; S3D**). We use the pseudotime information to calculate a vector gradient, representing the
275   direction of reprogramming as a vector field (**Figure 3B, left; S3E; Methods**). We then quantify
276   the similarity between the reprogramming and perturbation simulation vector fields by calculating
277   their inner-product value, which we term 'perturbation score' (**Figure 3B**). A negative perturbation
278   score implies that the TF perturbation blocks reprogramming (**Figure 3C,** shown in magenta).
279   Conversely, a positive perturbation score indicates that reprogramming is promoted following TF
280   perturbation (**Figure 3C,** shown in green). By calculating the sum of the negative perturbation
281   scores, we can rank TFs by their potential to regulate the reprogramming process, where a greater
282   negative score indicates that reprogramming is impaired upon perturbation of the candidate TF.
283   Using these metrics, we can interpret perturbation effects on cell fate quantitatively and
284   objectively.

285   Via this approach, we performed a systematic *in silico* simulation of TF knockouts (KOs)
286   during iEP generation to identify novel regulators of reprogramming, specifically along the
287   reprogramming trajectory (**Figure S3F**). Following GRN inference for each of the 7 Monocle
288   states identified (**Figure S3D**), we performed KO simulations for all TFs with inferred connections
289   to at least one other gene ('active' TFs, n = 180; **Methods**), calculating the sum of the negative
290   perturbation scores to rank TFs by the predicted inhibition of reprogramming following their KO.
291   This *in silico* screen allows us to quickly screen 180 candidate TFs, prioritizing factors for
292   experimental validation. In the top-ranked TFs, many factors are shared between independent
293   biological replicates, demonstrating the consistency of reprogramming and our analysis ((**Figure
294   3D**; Pearson's, r = 0.72). The *Hnf4α-Foxa1* transgene is ranked top, as expected, since these
295   factors are driving the reprogramming process. Of the remaining top-ranked factors, only half are
296   differentially expressed in reprogrammed cells (**Table S1**), highlighting the utility of CellOracle to
297   recover novel candidate regulators.

298   For experimental validation, we further prioritized candidate genes based on GRN degree
299   centrality, enrichment of gene expression along the entire reprogramming trajectory, and ranking
300   agreement across biological replicates. Following this selection step, eight TFs remained: *Eno1,
301   Fos, Fosb, Foxd2, Id1, Klf2, Klf4, Klf15* (**Figure 3E**). For all TFs, CellOracle predicts impaired
302   reprogramming following their KO. We performed an initial screen for all eight TFs, using a short
303   hairpin RNA (shRNA)-based strategy to knock down each TF during reprogramming (Confirmed
304   by qRT-PCR; **Figure S3G**), followed by colony formation assay to quantify clusters of successfully

305 reprogrammed cells based on E-Cadherin expression. From this initial screen, reprogramming

306 was impaired following the knockdown of 6 of the 8 TFs, with 25-50% fewer colonies formed

307 (**Figure S3H, I**). We selected *Eno1, Fos, Fosb, Id1, and Klf4* for additional colony formation

308 assays, confirming that their knockdown significantly reduces reprogramming efficiency (n = 5

309 independent biological replicates for scramble shRNA control, Fosb, Id1; n = 4 for Eno1, Klf4; n

310 = 3 for Fos; paired t-test, two-tailed; * = p<0.05; ** = p <0.01; **Figure 3F, G**).

311   Overall, our systematic perturbation simulation and experimental validation revealed

312 several novel regulators of MEF to iEP reprogramming. Of these TFs, Fos appears across

313 orthogonal analyses and independent datasets as a putative regulator of iEP reprogramming.

314 Indeed, we noted an enrichment of genes associated with the activator protein-1 TF (AP-1), a

315 dimeric complex primarily containing members of the Fos and Jun factor families (Eferl and

316 Wagner, 2003). AP-1 functions to establish cell-type-specific enhancers and gene expression

317 programs (Heinz et al., 2010; Vierbuchen et al., 2017) and to reconfigure enhancers during

318 reprogramming to pluripotency (Knaupp et al., 2017; Madrigal and Alasoo, 2018). As part of the

319 AP-1 complex, Fos plays broad roles in proliferation, differentiation, and apoptosis, both in

320 development and tumorigenesis (Eferl and Wagner, 2003; Jochum et al., 2001; Velazquez et al.,

321 2015). We next focused on further *in silico* simulation and experimental validation of *Fos*, a core

322 component of AP-1.

323

324 **The AP-1 transcription factor subunit Fos is central to reprogramming initiation and**

325 **maintenance of iEP identity**

326   Comparing degree centrality scores between fibroblast and early reprogramming

327 clusters, *Fos* receives relatively high degree and eigenvector centrality scores, along with

328 connector hub classification in the early reprogramming clusters (**Figure 1A; 4A, B; S4A**). Clonal

329 analysis of early ancestors destined to reprogram successfully agrees with a central role for *Fos*

330 (**Figure 2; S2**)*.* Indeed, perturbation simulation and reduced reprogramming efficiency following

331 experimental knockdown (**Figure 3; S3**) lead us to select *Fos* for deeper mechanistic investigation

332 as a candidate gene playing a critical role in initiating iEP conversion.

333   During MEF to iEP reprogramming, *Fos* is gradually and significantly upregulated (**Figure**

334 **4C, D**; *P* < 0.001, permutation test, one-sided). Several Jun AP-1 subunits are also expressed in

335 iEPs, classifying as connectors and connector hubs across various reprogramming stages

336 (**Figure S4C-E**). *Fos* and *Jun* are among a battery of genes reported to be upregulated in a cell-

337 subpopulation-specific manner in response to cell dissociation-induced stress, potentially leading

338 to experimental artifacts (van den Brink et al., 2017). Considering this report, we performed qPCR

10

339   for *Fos* on dissociated and undissociated cells. This orthogonal validation confirms an 8-fold

340   upregulation (*P* <0.01, *t*-test, one-sided) of *Fos* in iEPs, relative to MEFs, revealing no significant

341   changes in gene expression in cells that are dissociated and lysed versus cells lysed directly on

342   the plate (**Figure S4F**). Furthermore, analysis of unspliced and spliced *Fos* mRNA levels reveals

343   an accumulation of spliced *Fos* transcripts in reprogrammed cells. This observation suggests that

344   these transcripts accumulated over time rather than by rapid induction of expression in the five-

345   minute cell dissociation and methanol fixation in our single-cell preparation protocol (**Figure S4G**)

346   (la Manno et al., 2018).

347   To further investigate the role of *Fos* across reprogramming, we simulated its

348   overexpression, using MEF to iEP reprogramming time course GRN configurations inferred by

349   CellOracle (**Figure 1**). In these analyses, to assess the *in silico* perturbation of a specific

350   candidate, we use a Markov simulation to predict how cell identity shifts within the overall cell

351   population, visualizing the results as a Sankey diagram (**Methods**). Overexpression simulation

352   for *Fos* predicts a major cell state shift from the early transition to transition clusters, in addition

353   to predicting shifts in identity from dead-end to reprogrammed clusters (**Figure 4E**). In contrast,

354   the simulation of *Fos* KO produces the opposite results. (**Figure 4F**)**.** We experimentally validated

355   this simulation by adding Fos to the iEP reprogramming cocktail. As expected, we see a significant

356   increase in the number of iEP colonies formed (n = 10, *P* < 0.001, *t*-test, one-sided; **Figure 4G**),

357   increasing reprogramming efficiency more than two-fold, accompanied by significant increases in

358   iEP marker expression as measured by qPCR (n = 3, *P* < 0.001, *t*-test, one-sided; **Figure 4H**).

359   Turning our attention to the later stages of reprogramming, *Fos* continues to receive

360   relatively high network scores, particularly for betweenness centrality, in the iEP GRN

361   configurations **(Figure 4A)**. *Fos* also classifies as a Connector Hub **(Figure 4B)** in iEPs,

362   suggesting a role for Fos in the stabilization and maintenance of the reprogrammed state. To test

363   this hypothesis, we use CellOracle to perform knockout simulations, followed by experimental

364   knockout validation in an established iEP cell line. Here, we leverage the ability to culture iEPs,

365   long-term, where they retain a range of phenotypes (from fibroblast-like to iEP states; **Figure**

366   **S4H**) and functional engraftment potential (Guo et al., 2019; Morris et al., 2014). Simulation of

367   *Fos* knockout using these long-term cultured iEP GRN configurations predicts the loss of iEP

368   identity upon factor knockout (**Figure 4I**). To test this prediction, we used a CRISPR-Cas9 based

369   approach to knock out *Fos* in established iEPs. Quantitative comparison of the cell proportions

370   between control and knockout groups confirms that fully reprogrammed iEPs regress toward an

371   intermediate state upon *Fos* knockout, confirming a role for this factor in maintaining iEP identity

372   (**Figure 4J**), in addition to the establishment of iEPs, as we demonstrate in our systematic

373   simulation and experimental validation, in Figure 3.

374

375   *Fos* **target inference uncovers a role for the hippo signaling effector Yap1 in**

376   **reprogramming**

377   To gain further insight into the mechanism of how Fos regulates reprogramming, we

378   interrogated a list of the top 50 inferred *Fos* targets across all stages of reprogramming (**Figure**

379   **5A; Table S2**). We also assembled a list of genes predicted to be downregulated following *Fos*

380   knockout simulation for the reprogramming time course (**Figure S5A**). From this analysis, we

381   noted the presence of direct targets of YAP1, a central downstream transducer of the Hippo

382   signaling pathway (Galli et al., 2015; Ramos and Camargo, 2012; Stein et al., 2015). These

383   targets include *Cyr61*, *Amotl2, Gadd45g,* and *Ctgf*. Previous associations between Yap1 and Fos

384   support these observations; for example, YAP1 is recruited to the same genomic regions as FOS

385   via complex formation with AP-1 (Zanconato et al., 2015). Moreover, AP-1 is required for YAP1-

386   regulated gene expression and the liver overgrowth caused by Yap overexpression, where FOS

387   induction contributes to the expression of YAP/TAZ downstream target genes (Koo et al., 2020).

388   Together, this evidence suggests that Fos may play a role in reprogramming via an AP-1-

389   Yap1-mediated mechanism. Since Yap1 does not directly bind to DNA, we cannot deploy

390   CellOracle here to perform network analysis or perturbation simulations, highlighting a limitation

391   of our approach. However, in lieu of these analyses, we again turn to our rich single-cell time

392   course of iEP reprogramming (Biddy et al., 2018). Using a well-established active signature of

393   Yap1 (Dong et al., 2007), we find significant enrichment of this signature as reprogramming

394   progresses (**Figure S5B, C;** *P* < 0.001, permutation test, one-sided). Together, these results

395   suggest a role for the Hippo signaling component Yap1 in reprogramming, potentially effected via

396   its interactions with Fos/AP-1. Indeed, the hippo signaling axis plays a role in liver regeneration

397   (Pepe-Mooney et al., 2019; Yimlamai et al., 2014) and regeneration of the colonic epithelium (Yui

398   et al., 2018), in line with the known potential of iEPs to functionally engraft the liver and intestine

399   (Guo et al., 2019; Morris et al., 2014; Sekiya and Suzuki, 2011). Further, we have recently

400   demonstrated that iEPs transcriptionally resemble injured biliary epithelial cells (BECs) (Kong et

401   al., 2022), the target of YAP signaling in the context of liver regeneration (Pepe-Mooney et al.,

402   2019).

403   To test the role of Yap1 in iEP reprogramming, we first performed colony formation assays.

404   We find that the addition of Yap1 to the Hnf4α-Foxa1 cocktail significantly enhances

405   reprogramming efficiency, where the addition of Fos and Yap1 together increase colony formation

406 by almost three-fold, accompanied by significant increases in iEP marker expression (**Figure 5B;**

407 **Figure S5D, E,** P < 0.001, t-test, one-sided). Further, we note the emergence of a unique cell

408 morphology when Fos and Yap1 are added to the reprogramming cocktail, characterized by the

409 formation of extremely dense colonies (**Figure 5C**). To further characterize this distinctive

410 phenotype, we performed scRNA-seq on cells reprogrammed with Hnf4α-Foxa1 (n= 7,414 cells),

411 Hnf4α-Foxa1-Yap1 (n= 8,549 cells), Hnf4α-Foxa1-Fos (n= 8,771 cells), Hnf4α-Foxa1-Yap1-Fos

412 (n= 10,507 cells) and collected at day 20. Cells were clustered using the Leiden clustering

413 algorithm. Integration was performed using Seurat, and cells were visualized in 2D using UMAP

414 (**Figure S5F**).

415 We scored cells using established markers of MEFs and iEPs (Biddy et al., 2018),

416 revealing a significant increase in reprogramming efficiency, particularly following the addition of

417 Yap1 (p<0.0001, Wilcoxon test, **Figure 5F; S5F**), which is also accompanied by a reduction in

418 fibroblast marker expression (**Figure S5G**). We further classify cell identity using our

419 unsupervised method for cell-type classification, *Capybara* (Kong et al., 2022). In agreement with

420 our previous reports, using a healthy and regenerating liver atlas, iEPs generated with Hnf4α-

421 Foxa1 alone classify mainly as stromal cells (**Figure 5G**). However, following the addition of Fos

422 and Yap1, a significant population (p<0.0001, randomized test) of injured BECs emerges, in

423 similar proportions as observed in long-term cultured iEPs (Kong et al., 2022). In addition to

424 several hybrid cell types that we previously reported, we also observe a significant expansion of

425 a normal BEC population, from ~4% to ~12-35%, particularly upon the addition of Yap1 to the

426 reprogramming cocktail (p<0.0001, randomized test), where endogenous *Fos* expression is also

427 upregulated (**Figure S5G**). We observed a similar expansion of the normal BEC population when

428 long-term iEPs were cultured in a 3D matrigel sandwich culture (Kong et al., 2022). Here, our

429 results are consistent with these previous observations and point to the molecular regulation

430 driving changes in cell identity. In summary, CellOracle analysis and *in silico* prediction, combined

431 with experimental validation, have revealed several new factors and putative regulatory

432 mechanisms to enhance the efficiency and fidelity of reprogramming.

433

434 **Discussion**

435 Here, our application of CellOracle to the direct reprogramming of MEF to iEPs revealed

436 many new insights into this lineage conversion paradigm. Using CellTag-based lineage tracing,

437 we had previously demonstrated the existence of distinct conversion trajectories: one path leading

438 to successfully reprogrammed cells and a route to a dead-end state, accompanied by fibroblast

439 gene re-expression (Biddy et al., 2018). From lineage analysis, we found that sister cells follow

13

440　the same reprogramming trajectories, suggesting that conversion outcome is established shortly

441　after overexpression of the reprogramming TFs. The network analysis we present in this study,

442　powered by CellOracle, supports these earlier observations, revealing GRN reconfiguration within

443　the first few days of reprogramming. Further, the new clonal tracking we present here confirms

444　this early GRN configuration and that key wiring differences between reprogrammed and dead-

445　end outcomes can be identified from early stages.

446　　　　From our analysis of early GRN reconfiguration, we find that *Mef2a* and *Klf6* are highly

447　connected in fibroblasts and that these connections are largely decommissioned in successfully

448　converting cells. Although better known as a cardiac factor (Filomena and Bang, 2018), *Mef2a*

449　expression is enriched in the dead-end population, whereas *Klf6* is enriched in early transition

450　states, followed by its downregulation as reprogramming progresses (Supplemental data; Biddy

451　et al., 2018). Considering that relatively few iEPs successfully reprogram, a broad hallmark of

452　many lineage conversion protocols, targeting such TFs that are highly connected in the starting

453　population may represent one approach to enhance reprogramming efficiency by promoting the

454　erasure of starting cell identity.

455　　　　In this study, we have focused on the TFs associated with installing new cell identity. From

456　our clonal analysis of GRN reconfiguration in reprogrammed-destined cells, we find many

457　previously unreported regulators of iEP reprogramming. Indeed, our previous time-course

458　analysis did not identify many candidate regulators in the early stages, as the gene expression

459　differences were relatively subtle. Here, our network-based analysis recovers several novel early

460　factors, such as *Klf5, Cebpb, Mybl2, Foxk2, Fos,* and *Junb.* The recovery of additional factors is

461　also likely due to the clonal analysis, which further breaks down population heterogeneity to target

462　those rare cells that successfully reprogram.

463　　　　Indeed, from our GRN network configuration analysis, we identify several factors that may

464　regulate the reprogramming process. At this point, we would typically prioritize these factors for

465　further experimental validation, often basing the prioritization on previous literature, gene

466　expression patterns, or other available data. Here, we leverage the unique feature of CellOracle:

467　simulation of cell identity transition following candidate TF perturbation (knockout or

468　overexpression), using cluster-specific GRNs to model subsequent expression changes in

469　regulated genes. In a series of analyses complementary to the network analyses, we perform a

470　systematic *in silico* simulation of 180 TF knockouts to test which factors are required for successful

471　iEP reprogramming. This analysis revealed many putative reprogramming regulators, and from a

472　shortlist of eight candidates, we experimentally validated a role for six. Future *in silico* studies

473 could be designed to identify factors to block the entry of cells onto the dead-end trajectory or
474 factors to accelerate cells down the reprogramming trajectory.

475 From the systematic *in silico* knockout simulation and experimental validation, we
476 identified five new regulators of iEP reprogramming: *Id1, Fosb, Fos, Eno1*, and *Klf4.* Klf4 is one
477 of the previously described core pluripotency reprogramming factors (Takahashi and Yamanaka,
478 2006). The reduction of iEP reprogramming efficiency following its knockdown also suggests that
479 Klf4 plays a role in this direct lineage conversion paradigm. Similarly, Id1 has also been shown to
480 play a positive role in reprogramming to pluripotency (Hayashi et al., 2016), suggesting parallels
481 with direct lineage conversion. We also noted the involvement of several AP-1 factors, both from
482 our network analyses and *in silico* simulations, including *Fos, Fosb, Fosl2,* and *Junb*. The FOS-
483 JUN-AP1 complex has been reported to regulate reprogramming to pluripotency (Xing et al.,
484 2020) and direct reprogramming to cardiomyocytes (Wang et al., 2022); thus, we selected Fos
485 for further investigation.

486 The CellOracle analyses presented here provide new mechanistic insight into the
487 reprogramming process. Network connectivity scores and cartography analyses support a role
488 for the AP-1 subunit *Fos* as a putative reprogramming regulator. Indeed, our simulated
489 perturbations of Fos support its role in generating and maintaining iEPs. We confirmed these
490 simulations experimentally, where the addition of *Fos* to the reprogramming cocktail significantly
491 increases the yield of iEPs. Conversely, iEP identity is attenuated upon *Fos* knockout. Further
492 investigation of inferred *Fos* targets implicates a role for Yap1, a Hippo signaling effector, in
493 reprogramming. This observation is supported by our finding that a well-established signature of
494 active Yap1 is enriched as reprogramming progresses, which suggested a role for Yap1,
495 potentially effected via its interactions with Fos/AP-1. Indeed, the addition of Fos or Yap1 to the
496 reprogramming cocktail resulted in a significant increase in reprogramming efficiency, where the
497 addition of both factors yielded a three-fold increase in iEP colony formation.

498 In a parallel study, we have found that iEPs resemble post-injury biliary epithelial cells
499 (BECs) (Kong et al., 2022). Considering that Yap1 plays a central role in liver regeneration (Pepe-
500 Mooney et al., 2019; Yimlamai et al., 2014), these results raise the possibility that iEPs represent
501 a regenerative cell type, explaining their Yap1 activity, self-renewal *in vitro*, and capacity to
502 functionally engraft liver (Sekiya and Suzuki, 2011), and intestine (Guo et al., 2019; Morris et al.,
503 2014). Indeed, our unsupervised cell type classification of iEPs reprogrammed with the addition
504 of Fos and Yap to the Hnf4α-Foxa1 reprogramming cocktail suggests that these factors can
505 directly expand both the injured and normal BEC population, supporting the notion that iEPs may
506 resemble a regenerative population. Altogether, these new mechanistic insights have been

15

507   enabled by CellOracle analysis, placing it as a powerful tool for the dissection of cell identity,

508   aiding improvements in reprogramming efficiency and fidelity.

509

510   **Limitations of the study**

511   Here, we have presented an analysis of network reconfiguration during fibroblast to iEP

512   reprogramming, revealing several novel regulators of direct conversion that we further investigate

513   via *in silico* perturbation and experimental validation. As we have demonstrated, these factors can

514   be used to increase reprogramming efficiency and fidelity. One limitation of CellOracle is that it

515   cannot be used to make 'out-of-network' predictions and, due to its use of a linear model, is not

516   suited to simulating the effects of perturbing several factors in parallel. Moreover, the model is not

517   designed to simulate the effects of non-physiological levels of factor expression. For these

518   reasons, CellOracle is not designed to discover *de novo* reprogramming cocktails. Instead, it is

519   best applied to dissecting the mechanisms of existing reprogramming strategies to enhance their

520   fidelity and efficiency. Finally, based on the GRN model used for *in silico* simulation, only TF

521   perturbation can be simulated at present. However, as we have demonstrated with Yap1, the

522   inferred gene targets of TFs can be scrutinized to provide mechanistic insight.

523

524   **Code availability**

525   CellOracle code, documentation, and tutorials are available on GitHub

526   (https://github.com/morris-lab/CellOracle).

527

528   **Data availability**

529   All source data, including sequencing reads and single-cell expression matrices, are available

530   from the Gene Expression Omnibus (GEO) under accession codes GSE99915 (Biddy et al., 2018)

531   and GSE145298 for the new scRNA-seq data presented in this manuscript.

532

533   **Acknowledgments**

543   **Author Contributions**

544   Conceptualization, Methodology, K.K., S.A.M.; Software, K.K.; Formal Analysis, K.K., M.A.T.,

545   K.J., C.M.H., S.A.M; Investigation, K.K., M.A.T., K.J., C.M.H., X.Y., S.A.M.; Data Curation, K.K.,

546   M.A.T., K.J.; Writing – Original Draft, K.K., S.A.M.; Writing – Review & Editing, K.K., M.A.T., K.J.,

547   C.M.H., S.A.M.; Visualization, K.K., S.A.M.; Funding Acquisition, Resources, Supervision, S.A.M.

548

549   **Competing Interests**

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

551   **Correspondence and requests for materials** should be addressed to S.A.M.

552

553   **Figure Legends**

554   **Figure 1. Application of CellOracle to assess GRN dynamics direct lineage reprogramming.**

555   Overview of the CellOracle pipeline to infer cell type- and state-specific GRN configurations. **(A)**

556   First, CellOracle uses scATAC-seq data to identify accessible promoter/enhancer DNA

557   sequences. The DNA sequence of regulatory elements is scanned for TF binding motifs,

558   generating a list of potential regulatory connections between a TF and its target genes to generate

559   a 'Base GRN' **(B)**. **(C)** Using single-cell expression data, active connections are identified from

560   all potential connections in the base GRN. **(D)** Cell type- and state-specific GRN configurations

561   are constructed by pruning insignificant or weak connections. **(E)** Schematic of Hnf4α and Foxa1-

562   mediated fibroblast to iEP reprogramming. Our previous CellTag lineage tracing revealed two

563   conversion trajectories; reprogramming and dead-end (Biddy et al., 2018). **(F)** Left panel: Force-

564   directed graph of fibroblast to iEP reprogramming: from Louvain clustering, 15 clusters of cells

565   were annotated manually, using marker gene expression, and grouped into five cell types;

566   Fibroblasts, Early_Transition, Transition, Dead-end, and Reprogrammed iEPs. Right panels:

567   Projection of *Apoa1* (iEP marker) and *Col1a2* (fibroblast marker) expression onto the force-

568   directed graph. **(G)** CellOracle analysis: The strength of network edges between *Hnf4α-Foxa1*

569   and its target genes, visualized as a heatmap (left panel), and plotted as a boxplot (right panel).

570   **(H)** Degree and Eigenvector centrality scores for the *Hnf4α-Foxa1* transgene. **(I)** *Hnf4α-Foxa1*

571   network cartography terms for each cluster. **(J, K)** Scatter plots showing a comparison of degree

572   centrality scores between specific clusters. **(J)** Comparison of degree centrality scores between

573 the Fib_1 cluster GRN configuration and the GRN configurations of other clusters in relatively

574 early stages of reprogramming. **(K)** Comparison of degree centrality scores between iEP_1 and

575 Dead-end_0 cluster GRN configurations.

576

577 **Figure 2. Lineage tracing reveals how early network state shapes reprogramming outcome**

578 **(A)** Overview of CellTag-based clonal tracking. The CellTag construct contains a random

579 'CellTag' barcode in the 3′ UTR of GFP, followed by an SV40 polyadenylation signal. Cells are

580 transduced with the CellTag lentiviral library (produced via transfection of HEK293T cells with the

581 complex plasmid library) so that each cell expresses ~3–4 CellTags, resulting in a unique,

582 heritable signature. CellTags are transcribed and captured during single-cell profiling, enabling

583 clonally related cells to be tracked throughout an experiment. **(B)** Experimental strategy to capture

584 'state-fate' relationships. MEFs are first transduced with Hnf4α-Foxa1, delivered via four rounds

585 of retrovirus in a 48 hr period. The complex CellTag lentivirus library is introduced on the last

586 round of transduction. The end of this period, with transgene expression at a maximum, is

587 considered reprogramming day 0. Cells are expanded, and 25% of the population is profiled at

588 day 4, to maximize the capture of clones in early stages – this is referred to as the 'state'

589 population. The remaining population is reseeded and profiled again on days 10 and 28 to capture

590 reprogramming outcome, referred to as 'fate'. **(C)** Cells captured in the state-fate experiment.

591 Timepoint information is projected onto the UMAP embedding. A total of 24,799 cells were

592 sequenced: 8,440 on day 4, 4,836 on day 10, and 11,523 on day 28. **(D)** Projection of fibroblast,

593 iEP, and dead-end identity scores onto the UMAP embedding to reveal reprogrammed and dead-

594 end cell fates **(E)**. **(F)** A randomized test identified day 4 state clones whose day 10 and 28 fate

595 sisters were iEP enriched or iEP depleted. Top: Kernel density estimation of iEP-enriched day 4

596 state clones and their day 10 and 28 fates, outlining the 'reprogramming' trajectory (n = 879 cells).

597 Bottom: Kernel density estimation of iEP-depleted day 4 state clones and their day 10 and 28

598 fates, outlining the 'dead-end' trajectory (n = 4,955 cells). **(H)** Comparison of degree centrality

599 scores between native fibroblasts and day 4 reprogrammed-destined cells (left) and day 4

600 reprogrammed- and dead-end-destined cells (right).

601

602 **Figure 3. Systematic *in silico* simulation of TF knockout to identify novel regulators of iEP**

603 **reprogramming. (A)** Monocle-based pseudotemporal ordering of 48,515 cells from the Biddy

604 2018 reprogramming dataset, 2 independent biological replicates. **(B)** Schematic for perturbation

605 score calculations. CellOracle calculates a perturbation score by comparing the direction of the

606 simulated cell state transition with the direction of cell differentiation. First, the pseudotime data

607   is summarized by grid points and converted into a 2D gradient vector field. The results of the

608   perturbation simulation are converted into the same vector field format, and the inner product of

609   these vectors is calculated to produce a perturbation score. **(C)** A positive perturbation score

610   (green suggests the perturbation is predicted to promote differentiation. In contrast, the negative

611   perturbation score (magenta) represents impaired differentiation. **(D)** Ranked list of TFs based on

612   the sum of the negative perturbation score. **(E)** Representative example of a TF KO simulation.

613   **(F)** Experimental validation of candidate TFs: Colony formation assay. **(G)** Colony quantification.

614   n = 5 indpendent biological replicates for scramble shRNA control, Fosb, Id1; n = 4 indpendent

615   biological replicates for Eno1, Klf4; n = 3 indpendent biological replicates for Fos; paired t-test,

616   two-tailed; * = p<0.05; ** = p <0.01

617

618   **Figure 4. CellOracle analysis and experimental validation of Fos in the establishment and**

619   **maintenance of iEP identity. (A)** Degree centrality, betweenness centrality, and eigenvector

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

621   expression projected onto the force-directed graph of the 2018 reprogramming time course. **(D)**

622   Violin plot of *Fos* expression across reprogramming stages. **(E)** *Fos* gene overexpression

623   simulation with reprogramming GRN configurations. The left panel is the projection of simulated

624   cell transitions onto the force-directed graph. The Sankey diagram summarizes the simulation of

625   cell transitions between cell clusters. For overexpression simulation, *Fos* expression was set to a

626   value of 1.476, representing its maximum value in the imputed gene expression matrix **(F)** *Fos*

627   gene knockout simulation. **(G)** Colony formation assay with addition of Fos to the Hnf4α-Foxa1

628   reprogramming cocktail. Left panel: E-cadherin immunohistochemistry. Right panel: box plot of

629   colony numbers (n = 6 technical replicates, 2 independent biological replicates; *** = $P < 0.001$,

630   *t*-test, one-sided). **(H)** qPCR assay for *Fos* and iEP marker expression (*Apoa1* and *Chd1*)

631   following addition of Fos to the Hnf4α-Foxa1 reprogramming cocktail (n = 3 independent biological

632   replicates; *** = $P < 0.001$, ** = $P < 0.01$, *t*-test, one-sided). **(I)** *Fos* gene knockout simulation in

633   expanded, long-term cultured iEPs. **(J)** CRISPR/Cas9 knockout of *Fos* using CRISPR/Cas9 in

634   expanded iEP cells. We designed 3 guide RNAs to target *Fos,* and transduced Cas9-expressing

635   iEP cells with this guide RNA lentivirus pool. Left panels: Kernel density estimation method was

636   applied with the t-SNE embedding to compare cell density between control guide RNAs and guide

637   RNAs targeting *Fos.* Right panels: Quantification of changes in cell ratio following *Fos* knockout.

638

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

640   **reprogramming. (A)** Heatmap of expression of the top 50 inferred *Fos* targets across all stages

641 of reprogramming. Established targets of YAP1 are highlighted in red. **(B)** Colony formation assay
642 with the addition of Yap1 and Fos to the Hnf4α-Foxa1 reprogramming cocktail. Left panels: E-
643 cadherin immunohistochemistry. Right panel: box plot of colony numbers (n = 6 independent
644 biological replicates; *** = $P < 0.001$, t-test, one-sided). **(C)** Brightfield and epifluorescence images
645 of cells reprogrammed with Hnf4α-Foxa1 or Hnf4α-Foxa1-Fos-Yap1 cocktails. Scale bar = 500
646 μM. **(D)** scRNA-seq analysis of cells reprogrammed with Hnf4α-Foxa1 (n= 7,414 cells), Hnf4α-
647 Foxa1-Fos (n= 8,771 cells), Hnf4α-Foxa1-Yap1 (n= 8,549 cells), and Hnf4α-Foxa1-Fos-Yap1 (n=
648 10,507 cells) cocktails and collected at day 20. Projection of fibroblast and iEP identity scores
649 onto the UMAP embedding. **(E)** Kernel density estimation of cell density for each reprogramming
650 cocktail from **(D)**. **(F)** Violin plot of iEP identity scores for each reprogramming cocktail. **** =
651 p<0.0001, Wilcoxon test. **(G)** Unsupervised cell type classification for each reprogramming
652 cocktail, using normal and injured mouse liver as a reference. BEC: Biliary epithelial cells. * = p =
653 0, randomized test.

654

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

663

664 **10x alignment, digital gene expression matrix generation.** The Cell Ranger v6.0.1 pipeline
665 (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was
666 used to process data generated using the 10x Chromium platform. Cell Ranger processes, filters,
667 and aligns reads generated with the Chromium single-cell RNA sequencing platform. This pipeline
668 was used in conjunction with a custom reference genome, created by concatenating the
669 sequences corresponding to the *Hnf4α-t2a-Foxa1* transgene as a new chromosome to the mm10
670 genome. The unique UTRs in the *Hnf4α-t2a-Foxa1* transgene construct allowed us to monitor
671 transgene expression. To create Cell Ranger compatible reference genomes, the references were
672 rebuilt according to instructions from 10x (https://support.10xgenomics.com/single-cell-gene-
673 expression/software/pipelines/latest/advanced/references). To achieve this, we first created a
674 custom gene transfer format (GTF) file, containing our transgenes, followed by indexing of the

20

675 FASTA and GTF files, using Cell Ranger 'mkgtf' and 'mkref' functions. Following this step, the

676 default Cell Ranger pipeline was implemented, then the filtered output data was used for

677 downstream analyses.

678

679 **CellTag clone calling**

680 Reads containing the CellTag sequence were extracted from the processed and filtered BAM files

681 produced by the 10x Genomics pipeline, using our CellTagR pipeline:

682 https://github.com/morris-lab/CellTagR. The resulting filtered CellTag UMI count matrix was then

683 used for all downstream clone and lineage analysis. The CellTag matrix was initially filtered by

684 removing CellTags that do not appear on the allowlist generated for each CellTag plasmid library

685 Cells expressing more than 20 CellTags (likely corresponding to cell multiplets) and less than 2

686 CellTags per cell were filtered out. To identify clonally related cells, Jaccard analysis using the R

687 package Proxy was used to calculate the similarity of CellTag signatures between cells. Clones

688 were defined as groups of 2 or more related cells. Clones were called on cells pre-filtered for

689 numbers of genes, UMIs, and mitochondrial RNA content.

690

691 **Cell type classification with Capybara**

692 Cells reprogramed with Hnf4α-Foxa1, Hnf4α-Foxa1-Fos, Hnf4α-Foxa1-Yap1, and Hnf4α-Foxa1-

693 Fos-Yap1 were classified using Capybara (Kong et al., 2022). Briefly, the single-cell datasets

694 were processed, filtered, and clustered using Seurat, resulting in 35,241 cells (7,414 HF, 8,771

695 HF-Fos, 8,549 HF-Yap, 10,507 HF-Fos-Yap1). To construct a reference for cell-type

696 classification, we obtained scRNA-seq data of biliary epithelial cells (BECs) and hepatocytes,

697 before and after injury, from GSE125688 (Pepe-Mooney et al., 2019). We built a custom high-

698 resolution reference by incorporating additional tissues from the MCA: fetal liver, MEFs, and

699 embryonic mesenchyme. Following the construction of a high-resolution reference, we performed

700 preprocessing on the reference and the samples, on which we then applied quadratic

701 programming to generate the identity score matrices. Further, we categorized cells into discrete,

702 hybrid, and unknown, calculated the empirical p-value matrices, and performed binarization and

703 classification. We calculated the percent composition of each cell type. Cells with hybrid identities

704 were filtered and refined based on their identity scores as well as representation by more than

705 0.5% cells of the population. Code and documentation are available at:

706 https://github.com/morris-lab/Capybara.

707

708 **Experimental Methods**

709 **Mice and derivation of mouse embryonic fibroblasts.** Mouse Embryonic Fibroblasts were
710 derived from E13.5 C57BL/6J embryos. (The Jackson laboratory: 000664). Heads and visceral
711 organs were removed from E13.5 embryos. The remaining tissue was minced with a razor blade
712 and then dissociated in a mixture of 0.05% Trypsin and 0.25% Collagenase IV (Life Technologies)
713 at 37°C for 15 minutes. After passing the cell slurry through a 70μM filter to remove debris, cells
714 were washed and then plated on 0.1% gelatin-coated plates, in DMEM supplemented with 10%
715 FBS (Sigma-Aldrich), 2mM L-glutamine, and 50mM β-mercaptoethanol (Life Technologies). All
716 animal procedures were based on animal care guidelines approved by the Institutional Animal
717 Care and Use Committee.
718
719 **Retrovirus Production.** Retroviral particles were produced by transfecting 293T-17 cells (ATCC:
720 CRL-11268) with the pGCDN-Sam construct containing Hnf4α-t2a-Foxa1/Fos/Yap1, along with
721 packaging construct pCL-Eco (Imgenex). Virus was harvested 48hr and 72hr after transfection
722 and applied to cells immediately following filtering through a low-protein binding 0.45μM filter.
723
724 **Lentiviral constructs and lentivirus production.** Lentiviral particles were produced by
725 transfecting 293T-17 cells (ATCC: CRL-11268) with the envelope construct pCMV-VSV-G
726 (Addgene plasmid 8454), the packaging construct pCMV-dR8.2 dvpr (Addgene plasmid 8455),
727 and the shRNA expression vector for the respective candidate TF to be knocked down. The
728 shRNA expression vectors (with the TRC2 pLKO.5 backbone) were obtained directly from
729 Millipore-Sigma or cloned into the empty backbone using oligonucleotides (Integrated DNA
730 Technologies). Sequences of shRNA used: SHC202 (non-target shRNA control)
731 CAACAAGATGAAGAGCACCAA; *Eno1* GGCACAGAGAATAAATCTAAA; *Fos*
732 ATCCGAAGGGAACGGAATAAG; *FosB* ATGACGGAAGGACCTCCTTTG; *Foxd2*
733 AGATCATGTCCTCCGAGAGCT *Id1* GAGCTGAACTCGGAGTCTGAA; *Klf2*
734 GACCGATTGTATTTCTATAAG *Klf4* CATGTTCTAACAGCCTAAATG; *Klf15*
735 CTACCCTGGAGGAGATTGAAG. Virus was harvested 48hr and 72hr after transfection and
736 applied to cells following filtering through a low-protein binding 0.45μm filter.
737 For generation of the complex CellTag library, lentiviral particles were produced by
738 transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct, along with
739 packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455), and pCMV-VSVG (Addgene
740 plasmid 8454).
741

742 **Generation and collection of iEPs.** Mouse embryonic fibroblasts (< passage 6) were converted

743 to iEPs as in (Biddy et al., 2018), modified from (Sekiya and Suzuki, 2011). Briefly, we transduced

744 cells every 12hr for 3 days, with fresh Hnf4α-t2a-Foxa1 retrovirus, in the presence of 4mg/ml

745 Protamine Sulfate (Sigma-Aldrich), followed by culture on 0.1% gelatin-treated plates for 1 week

746 in hepato-medium (DMEM:F-12, supplemented with 10% FBS, 1 mg/ml insulin (Sigma-Aldrich),

747 dexamethasone (Sigma-Aldrich), 10mM nicotinamide (Sigma-Aldrich), 2mM L-glutamine, 50mM

748 β-mercaptoethanol (Life Technologies), and penicillin/streptomycin, containing 20 ng/ml

749 hepatocyte growth factor (Sigma-Aldrich), and 20 ng/ml epidermal growth factor (Sigma-Aldrich).

750 After the seven days of culture, the cells were transferred onto plates coated with 5μg/cm^2 Type

751 I rat collagen (Gibco, A1048301). For single-cell processing, 30,000 reprogrammed, expanded

752 iEPs were collected and fixed in methanol, as previously described in (Alles et al., 2017). Briefly,

753 cells were collected and washed in Phosphate Buffered Saline (PBS), followed by resuspension

754 in ice-cold 80% Methanol in PBS, with gentle vortexing. These cells were stored at -80°C for up

755 to three months, and processed on the 10x platform (below).

756   For the state-fate experiments, we followed the above protocol with some slight

757 modifications. We transduced cells every 12hr for 2 days, with fresh Hnf4α-t2a-Foxa1 retrovirus,

758 and added CellTagging lentivirus on the final round of transduction. After 12hr, cells were washed

759 and expanded in hepato-medium for 4 days, at which point the cells were dissociated and 25%

760 of the population profiled by scRNA-seq. The remaining population was replated and additional

761 samples were profiled at days 10 and 28.

762

763 **Colony formation assays.** Mouse *Fos* and *Yap1* were cloned from iEPs into the retroviral vector,

764 pGCDNSam (Sekiya and Suzuki, 2011), and retrovirus produced as above. For comparative

765 reprogramming experiments, mouse embryonic fibroblasts (2x10^5/well of a 6-well plate) were

766 serially transduced over 72hr (as above). In control experiments, virus produced from an empty

767 vector control expressing only GFP was added to the Hnf4α-Foxa1 reprogramming cocktail. Virus

768 produced from the *Fos* and *Yap1* IRES-GFP constructs was added to the standard Hnf4α and

769 Foxa1 cocktail. Cells underwent reprogramming for two weeks and were processed for colony

770 formation assays: cells were fixed on the plate with 4% PFA, permeabilized in 0.1% Triton-X100

771 then blocked with Mouse on Mouse Elite Peroxidase Kit (Vector PK-2200). Primary antibody,

772 mouse anti-E-Cadherin (1:100, BD Biosciences) was applied for 30 min before washing and

773 processing with the VECTOR VIP Peroxidase Substrate Kit (Vector SK-4600). Colonies were

774 visualized on a flatbed scanner, adding heavy cream to each well in order to increase image

775 contrast. Colonies were counted, using our automated colony counting tool:

23

776   https://github.com/morris-lab/Colony-counter. *Fos and Yap1* overexpression was confirmed by

777   harvesting RNA from Hnf4α -Foxa1 and Hnf4α -Foxa1-Fos/Yap1-transduced cells (RNeasy kit,

778   Qiagen). Following cDNA synthesis (Maxima cDNA synthesis kit, Life Tech), qPCR was

779   performed to quantify *Fos/Yap1* overexpression (TaqMan Probes: Gapdh Mm99999915_g1;

780   *Cdh1* Mm01247357_m1; *Apoa1* Mm00437569_m1; *Fos* Mm00487425_m1; *Yap1*

781   Mm01143263_m1; TaqMan qPCR Mastermix, Applied Biosystems).

782

783   Colony formation assays for TF knockdowns were conducted similarly, with the following

784   modifications. To initiate reprogramming, mouse embryonic fibroblasts (75×10^3/well of a 6-well

785   plate) were serially transduced over 72hr (as above). Lentivirus produced from the non-target

786   shRNA control and the respective TF knockdown shRNA constructs was then added at 84hr and

787   96 hr (only added at 96hr for initial screen). At 120hr, cells were seeded for colony formation

788   assays (40×10^3 cells/well of a 6-well plate), which were then processed for colony formation on

789   day 14 as above. Remaining cells from each sample were seeded for harvesting RNA for qPCR

790   on day 14 as above. In the initial screen, cells from each sample were split equally, and seeded

791   in 6 well plates for colony formation and RNA extraction at D15 from reprogramming initiation.

792   For *Fos* and *FosB* knockdowns, mouse embryonic fibroblasts (120×10^3 in a 6-cm dish) were

793   transduced with the respective shRNA lentivirus at 24hr and 36hr post-seeding. qPCR

794   confirmation was done on RNA harvested from cells at 72hr post-seeding. TaqMan Probes used:

795   *Actb* Mm02619580_g1; *Eno1* Mm01619597_g1; *Fos* Mm00487425_m1; *Fosb* Mm00500401_m1;

796   *Foxd2* Mm00500529_s1; *Id1* Mm00775963_g1; *Klf2* Mm00500486_g1; *Klf4* Mm00516104_m1;

797   *Klf15* Mm00517792_m1.

798

799   **CRISPR/Cas9 *Fos* Knockout**

800   The Fos knockouts were performed as part of a larger screen, using Perturb-seq as previously

801   described (Adamson et al., 2016). The protocol was modified, as outlined below, to apply the

802   strategy to our experimental system:

803

804   *(1) Vector backbone and gene barcode pool construction*: For Perturb-seq experiments, we used

805   a lentivirus vector to express guide RNAs and gene barcodes (GBC). The lentivirus vector

806   backbone contains an antiparallel cassette containing a guide RNA and GBC. In the original

807   perturb-seq paper, the authors used pPS and pBA439 to construct the guide RNA-GBC vector

808   pool. Here, we modified pPS and pBA439 to generate the pPS2 vector, in which the Puromycin-

809   t2a-BFP gene was replaced by the Blasticidin-t2a-BFP gene. We constructed the guide RNA-

24

810   GBC vector using a multi-step cloning strategy: First, we synthesized dsDNA, via PCR, for a

811   random GBC pool. We purified the PCR product with AMPure XP SPRI beads. We then inserted

812   the purified GBC pool into the pPS2 vector at the EcoRI site in the 3' UTR of the Blasticidn-t2a-

813   BFP gene. We used the product of Gibson assembly for transformation into DH5α competent

814   cells (NEB: C2987H). Transformed cells were cultured directly in LB liquid. We extracted plasmid

815   DNA to yield the pPS2-GBC pool.

816

817   *(2) Guide RNA cloning.* We designed guide RNAs using https://zlab.bio/guide-design-resources.

818   We synthesized oligo DNA for each guide RNA. Oligo DNA pairs were annealed and inserted into

819   the pPS2-GBC vector, following BsmB1 digestion. After isolation and growth of single colonies,

820   plasmid DNA was extracted and sanger DNA sequenced; sequences of the guide RNA inserted

821   site and GBC site were used to construct a gRNA/GBC reference table:

822

823   Fos_sg0   CAGCCGACTGAACGCGTTATTC

824   Fos_sg1   CATATATCAAAGATGAACATTG

825   Fos_sg2   TCAAGGCTGTAATTTCTTGGGC

826   empty0   TTGATGAACTGCGCTAGCGAGG

827   empty1   AAGAGCGGCTCGCAAGGGAAAA

828   empty2   AGTAGGATACGTGGAGTTAATA

829

830   *(3) Lentivirus guide RNA pool generation.* An equal amount of DNA for each pPS2-guide RNA

831   vector was mixed together to generate the plasmid pool. Three control vectors were also mixed

832   with this plasmid vector pool; the weight ratio of each pPS2-guide vector to each control vector

833   was 1:4. We used this mixed DNA pool for lentivirus production. Lentiviral particles were produced

834   by transfecting 293T-17 cells (ATT: CRL-11268) with the pPS-guide RNA-GBC constructs, along

835   with the packaging plasmid, psPAX2 (https://www.addgene.org/12260/), and pMD2.G

836   (https://www.addgene.org/12259/).

837

838   *(4) Cell culture for Perturb-seq.* We transduced reprogrammed iEP cells with retrovirus carrying

839   Cas9 (MSCV-Cas9-Puro). The cells were treated with Puromycin (4 μg/ml) for four days to

840   eliminate non-transduced cells. iEP-Cas9 cells were transduced with the lentivirus guide RNA

841   pool for 24 hours. The concentration of lentivirus was pre-determined to target 10~20%

842   transduction efficiency. After four days of cell culturing, we sorted BFP positive cells to purify

843 transduced cells. Cells were cultured for a further 72 hours and fixed with methanol as previously

844 described (Alles et al., 2017).

845

846 *(5) GBC amplification and sequencing.* Following library preparation on the 10x chromium

847 platform (below), we PCR amplified the GBC. The amplification was performed largely according

848 the original perturb-seq paper (Adamson et al., 2016), but we modified the PCR primer sequence

849 for the Chromium single cell library v2 kit:

850

851 P7_ind_R2_BFP_primer:

852 CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTC

853 CGATCTTAGCAAACTGGGGCACAAGC

854 P5_partial_primer: AATGATACGGCGACCACCGA

855 GBG_Amp_F: GCTGATCAGCGGGTTTAAACGGGCCCTCTAGG

856 GBG_Amp_R: CGCGTCGTGACTGGGAAAACCCTGGCGAATTG

857 GBC_Oligo:

858 TTAAACGGGCCCTCTAGGNNNNNNNNNNNNNNNNNNNNNNNCAATTCGCCAGGGTTTTCCC

859   Following amplification, we purified the PCR product with AMPure XP SPRI beads. The

860 purified sample was sequenced on the Illumina Mi-seq platform.

861

862 *(6) Alignment of cell barcode/GBC.* For preprocessing of Perturb-seq metadata, we used

863 MIMOSCA, a computational pipeline for the analysis of perturb-seq data

864 (https://github.com/asncd/MIMOSCA). First, the reference table for the cell barcode/GBC pair was

865 generated from Fastq files. The data table was converted into the guide RNA/cell barcode table

866 using the guide RNA-GBC reference table. This metadata was integrated into the scRNA-seq

867 data. The guide metadata was processed with an EM-like algorithm in MIMOSCA to filter out

868 unperturbed cells computationally, as previously described (Adamson et al., 2016).

869

870 **10x procedure.** For single-cell library preparation on the 10x Genomics platform, we used: the

871 Chromium Single Cell 3′ Library & Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip

872 kit v2 (PN-120236), and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer's

873 instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Prior to cell capture,

874 methanol-fixed cells were placed on ice, then spun at 3000rpm for 5 minutes at 4°C, followed by

875 resuspension and rehydration in PBS, according to (Alles et al., 2017). 17,000 cells were loaded

26

876   per lane of the chip, aiming to capture 10,000 single-cell transcriptomes. The resulting cDNA

877   libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 2500.

878

879   **Supplemental Figure Legends**

880   **Supplemental Figure 1 (Related to Figure 1). GRN analysis of fibroblast to iEP**

881   **reprogramming. (A)** After base GRN construction (left panel) using single-cell expression data,

882   an active connection between the TF and the target gene is identified for defined cell identities

883   and states by building a machine learning (ML) model that predicts the relationship between the

884   TF and the target gene. ML model fitting results present the certainty of connection as a

885   distribution, enabling the identification of GRN configurations by removing inactive connections

886   from the base GRN structure. **(B)** Force-directed graph of iEP reprogramming scRNA-seq data

887   (n = 27,663 cells). Projection of: Reprogramming time point information onto the force-directed

888   graph. There are 8 time points; day 0, 3, 6, 9, 12, 15, 21, and 28; *Hnf4α-t2a-Foxa1* (*Hnf4α-Foxa1)*

889   transgene expression levels; marker gene expression for key iEP states. Reprogrammed iEP cell

890   cluster marker genes: *Cdh1, Apoa1,* and *Kng1*. Fibroblast marker gene: *Col1a2*. Transition

891   marker gene: *Mettl7a1*. Dead-end marker genes: *Peg3, Igf2,* and *Fzd1.* **(C)** Violin plots of marker

892   gene expression in each cluster. **(D)** PAGA connectivity analysis across the reprogramming time

893   course. **(E)** Illustration of the cartography analysis method. The cartography method classifies

894   genes into seven groups according to two network scores: within-module degree and participation

895   coefficient (Guimerà and Amaral, 2005). In complex networks, high degree nodes (hubs) play the

896   most significant roles in maintaining network structure. **(F)** Pie charts depicting the clonal

897   composition of Dead-end cluster 0 and Dead-end cluster 1. Clone and trajectory information is

898   derived from our previous CellTagging study (Biddy et al., 2018).

899

900   **Supplemental Figure 2 (Related to Figure 2). CellOracle network analysis of cells destined**

901   **to reprogrammed or dead-end states. (A)** Projection of Leiden cluster and gene expression

902   information onto the state-fate UMAP embedding (from **Figure 2C-F**) to identify reprogrammed

903   and dead-end fates. **(B)** Violin plots of reprogrammed (*Apoa1, Cdh1*), fibroblast (*Col1a1, Col1a2*),

904   and dead-end (*Peg3*) marker expression along the iEP-enriched and iEP-depleted trajectories.

905   **(C)** To assess the quality of the inferred networks, we calculated the degree distribution for each

906   GRN configuration after pruning weak network edges, based on the p-value and strength. We

907   counted the network degree (k), representing the number of network edges for each gene. P(k)

908   is the frequency of network degree k, visualized in scatter plots. We also visualized the

909   relationship between k and P(k) after log-transformation shows that these are scale-free networks,

910   demonstrating successful network inference from these relatively small cell populations.

911

912   **Supplemental Figure 3 (Related to Figure 3). Systematic *in silico* simulation of TF**

913   **knockout. (A)** Overview of signal propagation simulation. CellOracle leverages an inferred GRN

914   model to simulate how target gene expression changes in response to the changes in regulatory

915   gene expression. The input TF perturbation (shown in yellow) is propagated side-by-side within

916   the network model. **(B)** Leveraging the linear predictive ML algorithm features, CellOracle uses

917   the GRN model as a function to perform the signal propagation calculation. Iterative matrix

918   multiplication steps enable the estimation of indirect and global downstream effects resulting from

919   the perturbation of a single TF. **(C)** After signal propagation, the simulated gene expression shift-

920   vector is converted into a 2D vector and projected onto the dimensional reduction space. Details

921   are described in the methods section. **(D)** Left: Monocle states identified and used for GRN

922   inference. Right: Calculated pseudotime projected on the Monocle embedding and converted to

923   a 2D gradient vector field. **(E)** Schematic of the method to convert pseudotime to a 2D gradient

924   vector field: First, the pseudotime data is summarized by grid points, then CellOracle calculates

925   a 2D gradient vector of the pseudotime data that represents the directionality of reprogramming

926   pseudotime. **(F)** Outline of reprogramming and dead-end trajectories projected onto the Monocle

927   embedding. The sum of the negative perturbation score was calculated only for reprogramming

928   trajectory clusters in this study. **(G)** Quantitative RT-PCR to validate knockdown efficiency for

929   each shRNA. * = p<0.05, ** = p<0.01, *** = p<0.001, **** = p<0.0001; unpaired t-test with Welch's

930   correction, two-tailed. **(H)** Colony formation assay (E-cadherin immunohistochemistry) to test iEP

931   reprogramming efficiency following the knockdown of each candidate factor. **(I)** Quantification of

932   colonies formed in the initial screen. Factors marked red and * were selected for further

933   experimental validation.

934

935   **Supplemental Figure 4 (Related to Figure 4). CellOracle analysis of the role of Fos in**

936   **fibroblast to iEP reprogramming. (A)** Comparison of eigenvector centrality scores between the

937   Fib_1 cluster GRN configuration and the GRN configurations of other clusters in relatively early

938   stages of reprogramming. **(B)** Comparison of eigenvector centrality scores between iEP_1 and

939   Dead-end_0 cluster GRN configurations. **(C-E)** Expression and network cartography of Jun family

940   members, *Jun, Junb,* and *Jund*. **(F)** qPCR of *Fos* expression in fibroblasts and iEPs, with and

941   without cell dissociation prior to the assay, ** = $P < 0.01$, *t*-test, one-sided. **(G)** Analysis of *Fos*

942   mRNA splicing state in the scRNA-seq data of iEP reprogramming to investigate the *Fos* mRNA

943   maturation state: Violin plot for spliced *Fos* mRNA counts. **(H)** *t*-SNE plots of 9,914 expanded

944   iEPs, cultured long-term, revealing fibroblast-like, intermediate, and three iEP subpopulations.

945   Expression levels of *Apoa1* (marking typical iEPs)*, Col4a1* (fibroblast-like cells)*, Cdh1, Serpina1b*

946   (hepatic-like iEPs)*,* and *Areg* (intestine-like iEPs) projected onto the *t*-SNE plot.

947

948   **Supplemental Figure 5 (Related to Figure 5). The role of Fos and Yap1 in fibroblast to iEP**

949   **reprogramming. (A)** Top 50 decreased genes in *Fos* knockout simulation in the early

950   reprogramming transition (left) and GO analysis based on these genes (right). **(B)** Violin plot of

951   YAP1 target gene scores across reprogramming, which are significantly enriched as

952   reprogramming progresses (*** = $P < 0.001$, permutation test, one-sided). **(C)** Projection of YAP1

953   target gene scores onto the force-directed graph of reprogramming. **(D)** qPCR assay for *Yap1*

954   expression following addition of Yap1 and Fos to the Hnf4α-Foxa1 reprogramming cocktail (n =

955   4 independent biological replicates; *** = $P < 0.001$, ** = $P < 0.01$, *t*-test, one-sided), confirming

956   Yap1 overexpression. **(E)** qPCR assay for iEP marker expression (*Apoa1* and *Chd1*) following

957   addition of Yap1 and Fos to the Hnf4α-Foxa1 reprogramming cocktail (n = 4 independent

958   biological replicates; *** = $P < 0.001$, ** = $P < 0.01$, *t*-test, one-sided). **(F)** Projection of Leiden

959   cluster, dead-end identity scores, and gene expression information onto the state-fate UMAP

960   embedding (from **Figure 5D, E**). **(G)** Expression of key marker genes for each reprogramming

961   cocktail.

962

963   **Supplemental Table 1.** Differentially expressed iEP markers from (Biddy et al., 2018). Top-

964   ranked genes from CellOracle *in silico* perturbation are marked in red.

965

966   **Supplemental Table 2.** Top 50 CellOracle-inferred *Fos* targets, across all reprogramming

967   clusters. Confirmed YAP1 targets are highlighted in red.

968   **References**

969   Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert,

970   L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening

971   Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867-

972   1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

973   Alles, J., Karaiskos, N., Praktiknjo, S.D., Grosswendt, S., Wahle, P., Ruffault, P.-L., Ayoub, S.,

974   Schreyer, L., Boltengagen, A., Birchmeier, C., et al. (2017). Cell fixation and preservation for

975   droplet-based single-cell transcriptomics. BMC Biology *15*, 44. https://doi.org/10.1186/s12915-

976   017-0383-5.

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

980   Bocchi, V.D., Conforti, P., Vezzoli, E., Besusso, D., Cappadona, C., Lischetti, T., Galimberti, M.,
981   Ranzani, V., Bonnal, R.J.P., Simone, M. de, et al. (2021). The coding and long noncoding single-
982   cell   atlas   of   the   developing   human   fetal   striatum.   Science   (1979)   *372*.
983   https://doi.org/10.1126/science.abf5759.

984   van den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S.,
985   Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced
986   gene   expression   in   tissue   subpopulations.   Nat   Methods   *14*,   935–936.
987   https://doi.org/10.1038/nmeth.4437.

988   Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-
989   Generation   Machine   Learning   for   Biological   Networks.   Cell   *173*,   1581–1592.
990   https://doi.org/10.1016/J.CELL.2018.05.015.

991   Chopp, L.B., Gopalan, V., Ciucci, T., Ruchinskas, A., Rae, Z., Lagarde, M., Gao, Y., Li, C.,
992   Bosticardo, M., Pala, F., et al. (2020). An Integrated Epigenomic and Transcriptomic Map of
993   Mouse   and   Human   αβ   T   Cell   Development.   Immunity   *53*,   1182-1201.e8.
994   https://doi.org/10.1016/J.IMMUNI.2020.10.024/ATTACHMENT/EAD3FF5E-B3B9-46C4-95E3-
995   E303CC056184/MMC8.XLSX.

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

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

1000   Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S.A., Gayyed, M.F., Anders,
1001   R.A., Maitra, A., and Pan, D. (2007). Elucidation of a Universal Size-Control Mechanism in
1002   Drosophila and Mammals. Cell *130*, 1120–1133. https://doi.org/10.1016/J.CELL.2007.07.019.

1003   Eferl, R., and Wagner, E.F. (2003). AP-1: a double-edged sword in tumorigenesis. Nature
1004   Reviews Cancer *3*, 859–868. https://doi.org/10.1038/nrc1209.

1005   Filomena, M.C., and Bang, M.L. (2018). In the heart of the MEF2 transcription network: novel
1006   downstream effectors as potential targets for the treatment of cardiovascular disease.
1007   Cardiovascular Research *114*, 1425–1427. https://doi.org/10.1093/CVR/CVY123.

1008   Galli, G.G., Carrara, M., Yuan, W.-C., Valdes-Quezada, C., Gurung, B., Pepe-Mooney, B., Zhang,
1009   T., Geeven, G., Gray, N.S., de Laat, W., et al. (2015). YAP Drives Growth by Controlling

1010 Transcriptional Pause Release from Dynamic Enhancers. Molecular Cell *60*, 328–337.

1011 https://doi.org/10.1016/J.MOLCEL.2015.09.001.

1012 Guimerà, R., and Amaral, L.A.N. (2005). Cartography of complex networks: modules and

1013 universal roles. Journal of Statistical Mechanics: Theory and Experiment *2005*, P02001.

1014 https://doi.org/10.1088/1742-5468/2005/02/P02001.

1015 Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G.C., Yang, X., Kirita, Y., and Morris, S.A.

1016 (2019). CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics.

1017 Genome Biology *20*, 90. https://doi.org/10.1186/s13059-019-1699-y.

1018 Gurdon, J.B., Elsdale, T.R., and Fischberg, M. (1958). Sexually mature individuals of Xenopus

1019 laevis from the transplantation of single somatic nuclei. Nature *182*, 64–65. .

1020 Hayashi, Y., Hsiao, E.C., Sami, S., Lancero, M., Schlieve, C.R., Nguyen, T., Yano, K., Nagahashi,

1021 A., Ikeya, M., Matsumoto, Y., et al. (2016). BMP-SMAD-ID promotes reprogramming to

1022 pluripotency by inhibiting p16/INK4A-dependent senescence. Proc Natl Acad Sci U S A *113*,

1023 13057–13062.

1024 https://doi.org/10.1073/PNAS.1603668113/SUPPL_FILE/PNAS.201603668SI.PDF.

1025 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh,

1026 H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors

1027 prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576–

1028 589. https://doi.org/10.1016/j.molcel.2010.05.004.

1029 Jochum, W., Passegué, E., and Wagner, E.F. (2001). AP-1 in mouse development and

1030 tumorigenesis. Oncogene *20*, 2401–2412. https://doi.org/10.1038/sj.onc.1204389.

1031 Kamimoto, K., Hoffmann, C.M., and Morris, S.A. (2020). CellOracle: Dissecting cell identity via

1032 network inference and in silico gene perturbation. BioRxiv 2020.02.17.947416.

1033 https://doi.org/10.1101/2020.02.17.947416.

1034 Klein, C., Marino, A., Sagot, M.-F., Vieira Milreu, P., and Brilli, M. (2012). Structural and dynamical

1035 analysis of biological networks. Briefings in Functional Genomics *11*, 420–433.

1036 https://doi.org/10.1093/bfgp/els030.

1037 Knaupp, A.S., Buckberry, S., Pflueger, J., Lim, S.M., Ford, E., Larcombe, M.R., Rossello, F.J., de

1038 Mendoza, A., Alaei, S., Firas, J., et al. (2017). Transient and Permanent Reconfiguration of

1039 Chromatin and Transcription Factor Occupancy Drive Reprogramming. Cell Stem Cell *21*, 834-

1040 845.e6. https://doi.org/10.1016/J.STEM.2017.11.007.

1041 Kong, W., Biddy, B.A., Kamimoto, K., Amrute, J.M., Butka, E.G., and Morris, S.A. (2020).

1042 CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell

1043 resolution. Nature Protocols 1–23. https://doi.org/10.1038/s41596-019-0247-2.

1044   Kong, W., Fu, Y.C., Holloway, E.M., Garipler, G., Yang, X., Mazzoni, E.O., and Morris, S.A.
1045   (2022). Capybara: A computational tool to measure cell identity and fate transitions. Cell Stem
1046   Cell *29*, 635-649.e11. https://doi.org/10.1016/J.STEM.2022.03.001.

1047   Koo, J.H., Plouffe, S.W., Meng, Z., Lee, D.-H., Yang, D., Lim, D.-S., Wang, C.-Y., and Guan, K.-
1048   L. (2020). Induction of AP-1 by YAP/TAZ contributes to cell proliferation and organ growth. Genes
1049   & Development *34*, 72–86. https://doi.org/10.1101/gad.331546.119.

1050   Liu, Z., Chen, O., Wall, J.B.J., Zheng, M., Zhou, Y., Wang, L., Ruth Vaseghi, H., Qian, L., and Liu,
1051   J. (2017). Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector.
1052   Sci Rep *7*, 2193. https://doi.org/10.1038/s41598-017-02460-2.

1053   Madrigal, P., and Alasoo, K. (2018). AP-1 Takes Centre Stage in Enhancer Chromatin Dynamics.
1054   Trends Cell Biol *28*, 509–511. https://doi.org/10.1016/j.tcb.2018.04.009.

1055   Magaletta, M.E., Lobo, M., Kernfeld, E.M., Aliee, H., Huey, J.D., Parsons, T.J., Theis, F.J., and
1056   Maehr, R. (2022). Integration of single-cell transcriptomes and chromatin landscapes reveals
1057   regulatory programs driving pharyngeal organ development. Nature Communications 2022 13:1
1058   *13*, 1–16. https://doi.org/10.1038/s41467-022-28067-4.

1059   la Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K.,
1060   Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*,
1061   494–498. https://doi.org/10.1038/s41586-018-0414-6.

1062   Materna, S.C., and Davidson, E.H. (2007). Logic of gene regulatory networks. Current Opinion in
1063   Biotechnology *18*, 351–354. https://doi.org/10.1016/J.COPBIO.2007.07.008.

1064   Morris, S.A., and Daley, G.Q. (2013). A blueprint for engineering cell fate: current technologies to
1065   reprogram cell identity. Cell Res *23*, 33–48. https://doi.org/10.1038/cr.2013.1.

1066   Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., and
1067   Daley, G.Q. (2014). Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via
1068   CellNet. Cell *158*, 889–902. https://doi.org/10.1016/j.cell.2014.07.021.

1069   Nie, J., Carpenter, A.C., Chopp, L.B., Chen, T., Balmaceno-Criss, M., Ciucci, T., Xiao, Q., Kelly,
1070   M.C., McGavern, D.B., Belkaid, Y., et al. (2022). The transcription factor LRF promotes integrin
1071   β7 expression by and gut homing of CD8αα+ intraepithelial lymphocyte precursors. Nature
1072   Immunology 2022 23:4 *23*, 594–604. https://doi.org/10.1038/s41590-022-01161-x.

1073   Pepe-Mooney, B.J., Dill, M.T., Alemany, A., Ordovas-Montanes, J., Matsushita, Y., Rao, A., Sen,
1074   A., Miyazaki, M., Anakk, S., Dawson, P.A., et al. (2019). Single-Cell Analysis of the Liver
1075   Epithelium Reveals Dynamic Heterogeneity and an Essential Role for YAP in Homeostasis and
1076   Regeneration. Cell Stem Cell *25*, 23-38.e8. https://doi.org/10.1016/J.STEM.2019.04.004.

1077   Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie,
1078   D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero Predicts cis-Regulatory
1079   DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol Cell *71*, 858-871.e8.
1080   https://doi.org/10.1016/j.molcel.2018.06.044.

1081   Ramos, A., and Camargo, F.D. (2012). The Hippo signaling pathway and stem cell biology.
1082   Trends in Cell Biology *22*, 339–346. https://doi.org/10.1016/J.TCB.2012.04.006.

1083   Ravi, V.M., Neidert, N., Will, P., Joseph, K., Maier, J.P., Kückelhaus, J., Vollmer, L., Goeldner,
1084   J.M., Behringer, S.P., Scherer, F., et al. (2022). T-cell dysfunction in the glioblastoma
1085   microenvironment is mediated by myeloid cells releasing interleukin-10. Nature Communications
1086   2022 13:1 *13*, 1–16. https://doi.org/10.1038/s41467-022-28523-1.

1087   Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells
1088   by defined factors. Nature *475*, 390–393. https://doi.org/10.1038/nature10263.

1089   Stein, C., Bardet, A.F., Roma, G., Bergling, S., Clay, I., Ruchti, A., Agarinis, C., Schmelzle, T.,
1090   Bouwmeester, T., Schübeler, D., et al. (2015). YAP1 Exerts Its Transcriptional Control via TEAD-
1091   Mediated Activation of Enhancers. PLOS Genetics *11*, e1005465.
1092   https://doi.org/10.1371/journal.pgen.1005465.

1093   Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse
1094   Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.
1095   https://doi.org/10.1016/j.cell.2006.07.024.

1096   Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak,
1097   K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions
1098   are revealed by pseudotemporal ordering of single cells. Nat Biotechnol *32*, 381–386.
1099   https://doi.org/10.1038/nbt.2859.

1100   Velazquez, F.N., Caputto, B.L., and Boussin, F.D. (2015). c-Fos importance for brain
1101   development. Aging *7*, 1028–1029. https://doi.org/10.18632/aging.100862.

1102   Vierbuchen, T., Ling, E., Cowley, C.J., Couch, C.H., Wang, X., Harmin, D.A., Roberts, C.W.M.,
1103   and Greenberg, M.E. (2017). AP-1 Transcription Factors and the BAF Complex Mediate Signal-
1104   Dependent Enhancer Selection. Molecular Cell *68*, 1067-1082.e12.
1105   https://doi.org/10.1016/J.MOLCEL.2017.11.026.

1106   Wang, H., Yang, Y., Qian, Y., Liu, J., and Qian, L. (2022). Delineating chromatin accessibility re-
1107   patterning at single cell level during early stage of direct cardiac reprogramming. Journal of
1108   Molecular and Cellular Cardiology *162*, 62–71. https://doi.org/10.1016/J.YJMCC.2021.09.002.

1109　Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on

1110　transcriptional landscapes links state to fate during differentiation. Science (1979) *367*.

1111　https://doi.org/10.1126/SCIENCE.AAW3381.

1112　Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon,

1113　L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference

1114　through a topology preserving map of single cells. Genome Biology *20*, 59.

1115　https://doi.org/10.1186/s13059-019-1663-x.

1116　Xing, Q.R., el Farran, C.A., Gautam, P., Chuah, Y.S., Warrier, T., Toh, C.X.D., Kang, N.Y., Sugii,

1117　S., Chang, Y.T., Xu, J., et al. (2020). Diversification of reprogramming trajectories revealed by

1118　parallel single-cell transcriptome and chromatin accessibility sequencing. Science Advances *6*,

1119　18. https://doi.org/10.1126/SCIADV.ABA1190/SUPPL_FILE/ABA1190_SM.PDF.

1120　Yimlamai, D., Christodoulou, C., Galli, G.G., Yanger, K., Pepe-Mooney, B., Gurung, B., Shrestha,

1121　K., Cahan, P., Stanger, B.Z., and Camargo, F.D. (2014). Hippo pathway activity influences liver

1122　cell fate. Cell *157*, 1324–1338. https://doi.org/10.1016/j.cell.2014.03.060.

1123　Yui, S., Azzolin, L., Maimets, M., Pedersen, M.T., Fordham, R.P., Hansen, S.L., Larsen, H.L.,

1124　Guiu, J., Alves, M.R.P., Rundsten, C.F., et al. (2018). YAP/TAZ-Dependent Reprogramming of

1125　Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. Cell Stem Cell *22*, 35-49.e7.

1126　https://doi.org/10.1016/j.stem.2017.11.001.

1127　Zanconato, F., Forcato, M., Battilana, G., Azzolin, L., Quaranta, E., Bodega, B., Rosato, A.,

1128　Bicciato, S., Cordenonsi, M., and Piccolo, S. (2015). Genome-wide association between

1129　YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. Nature Cell Biology *17*, 1218–

1130　1227. https://doi.org/10.1038/ncb3216.
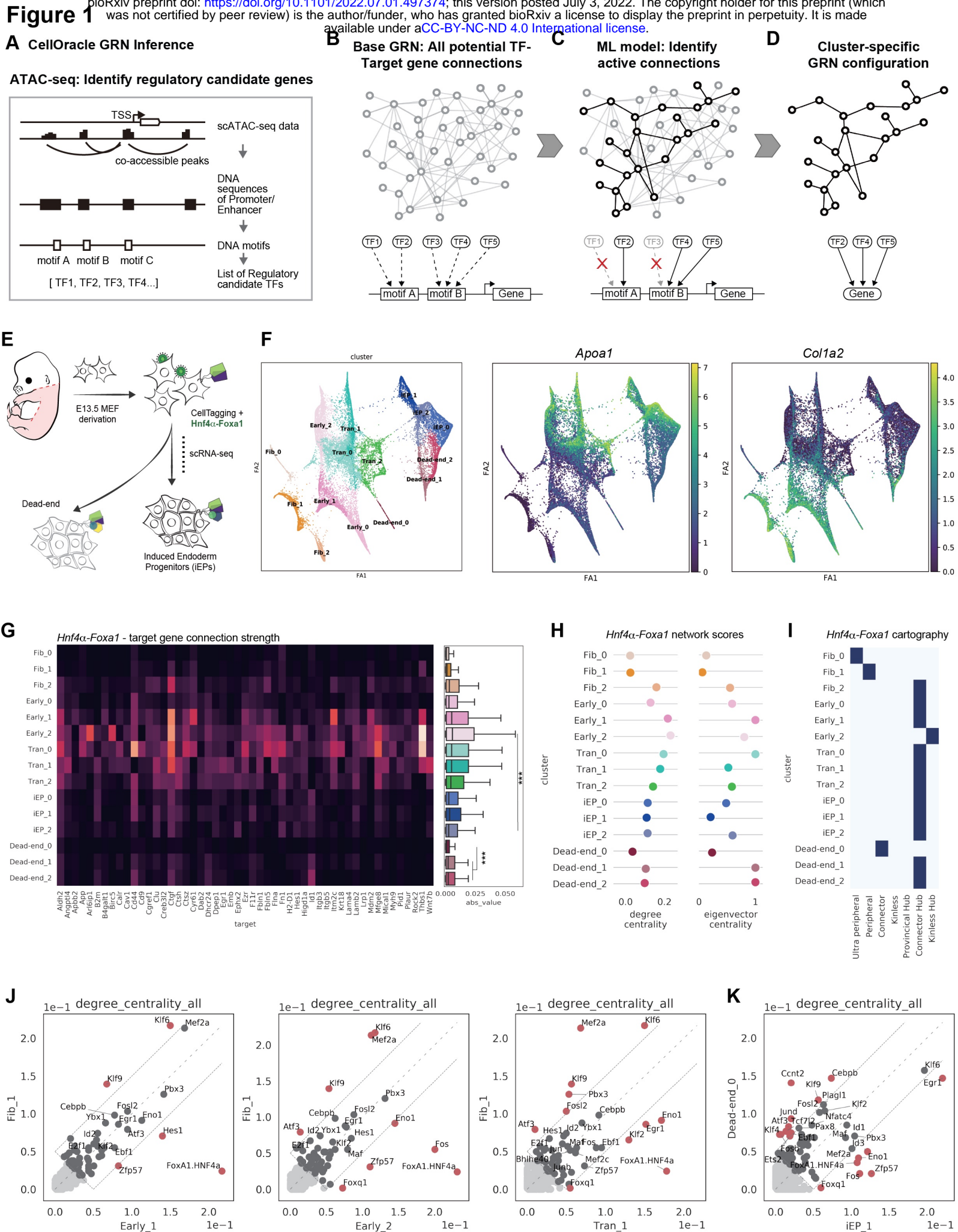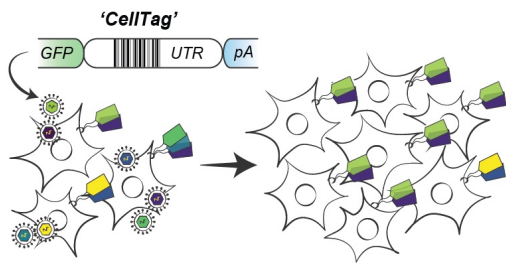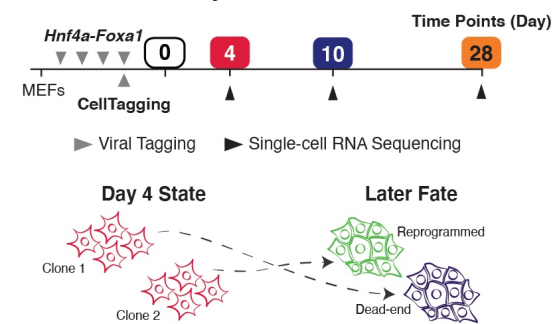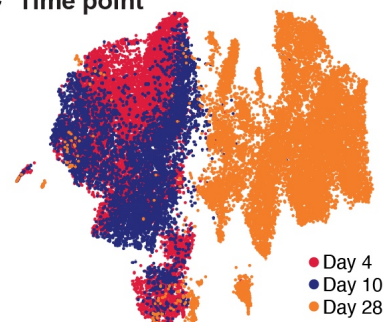
1131

Figure 1

A CellOracle GRN Inference

ATAC-seq: Identify regulatory candidate genes

B Base GRN: All potential TF-Target gene connections

C ML model: Identify active connections

D Cluster-specific GRN configuration

E

F

G *Hnf4α-Foxa1* - target gene connection strength

H *Hnf4α-Foxa1* network scores

I *Hnf4α-Foxa1* cartography

J degree_centrality_all

K degree_centrality_all

Figure 2

Figure 3

Figure 4

A: Fos network scores

B: Fos cartography

C: *Fos*

D: Scaled *Fos* expression

E: Fos overexpression simulation

F: Fos KO simulation

G: Hnf4α-Foxa1 / Hnf4α-Foxa1-Fos

H

I: Fos KO simulation: expanded iEPs

J: Experimental Fos KO: Cell density
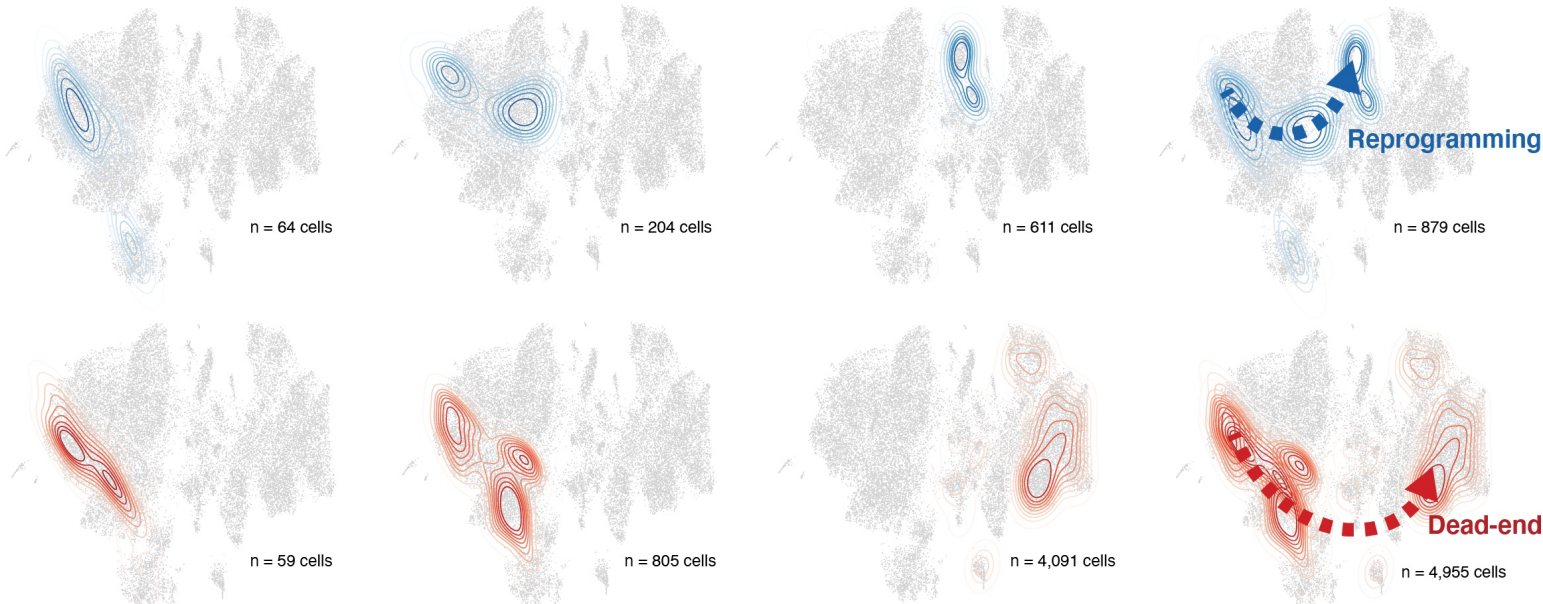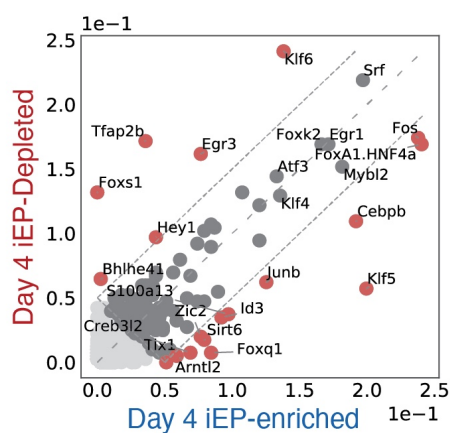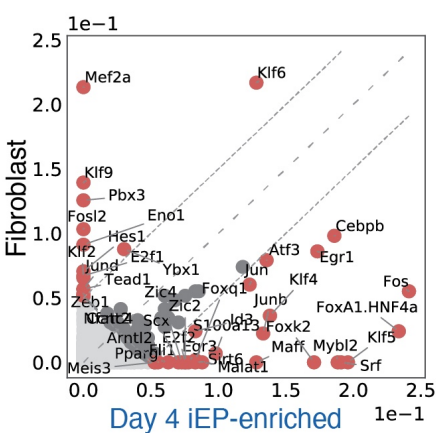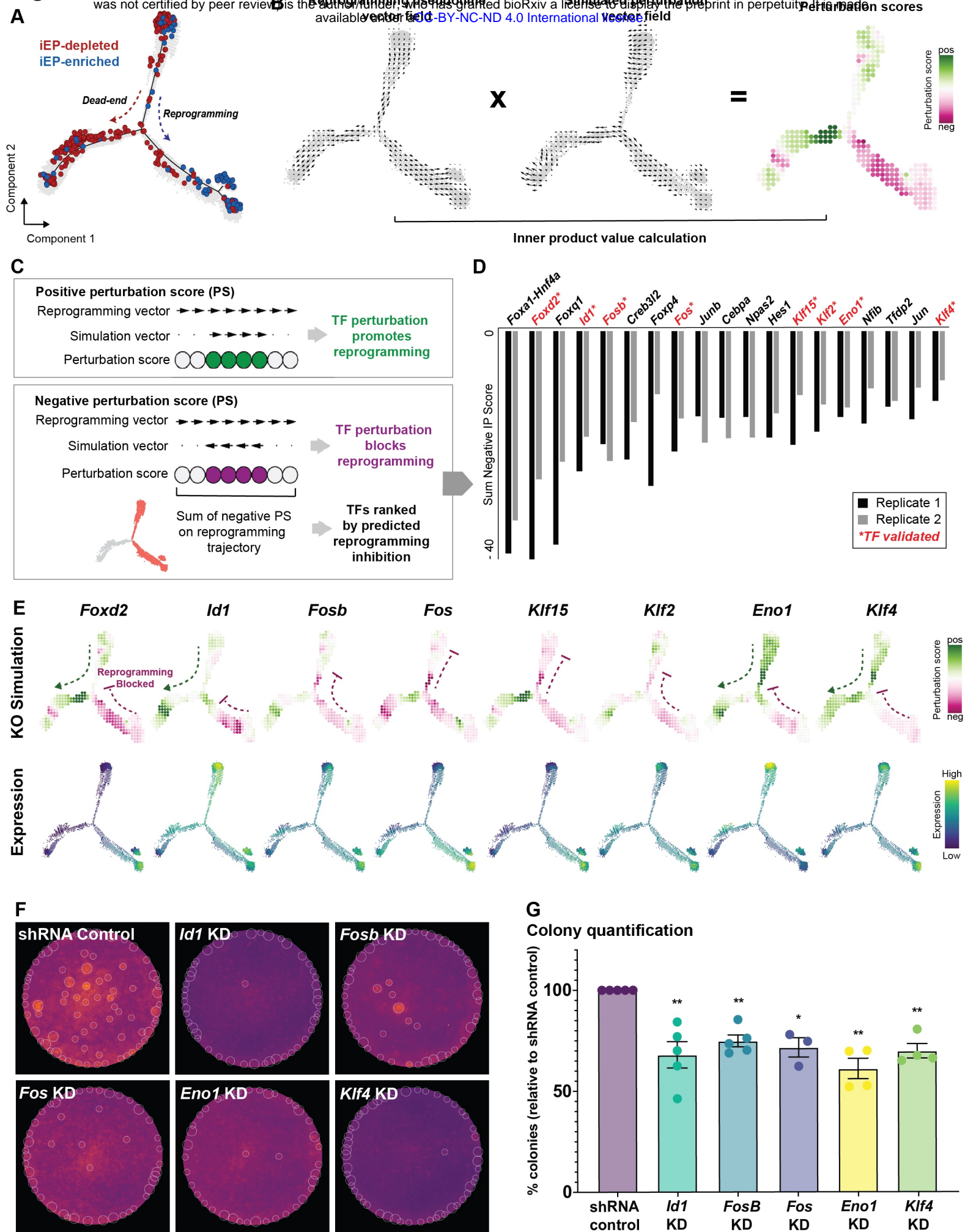
Figure 5
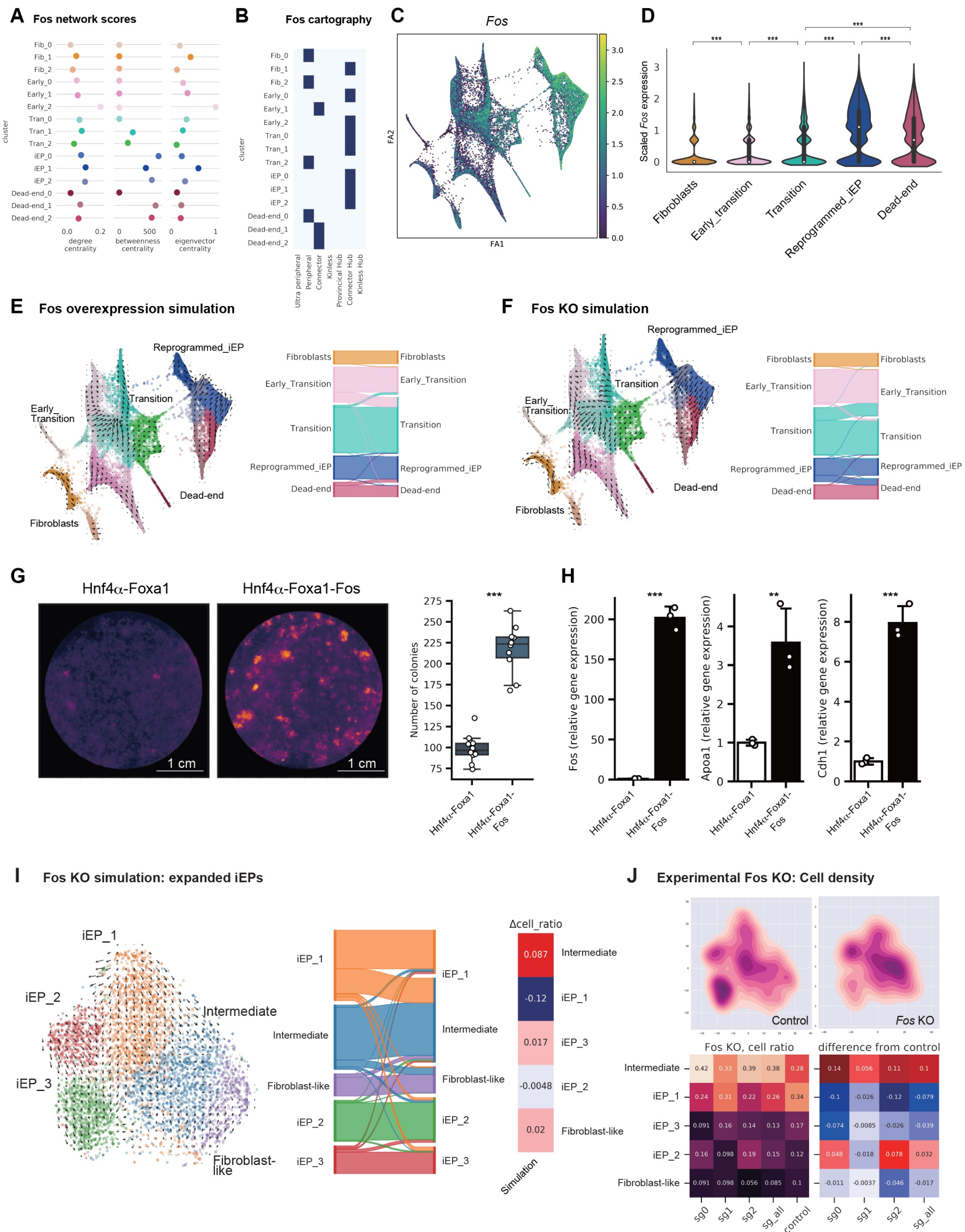
Supplemental Figure 1
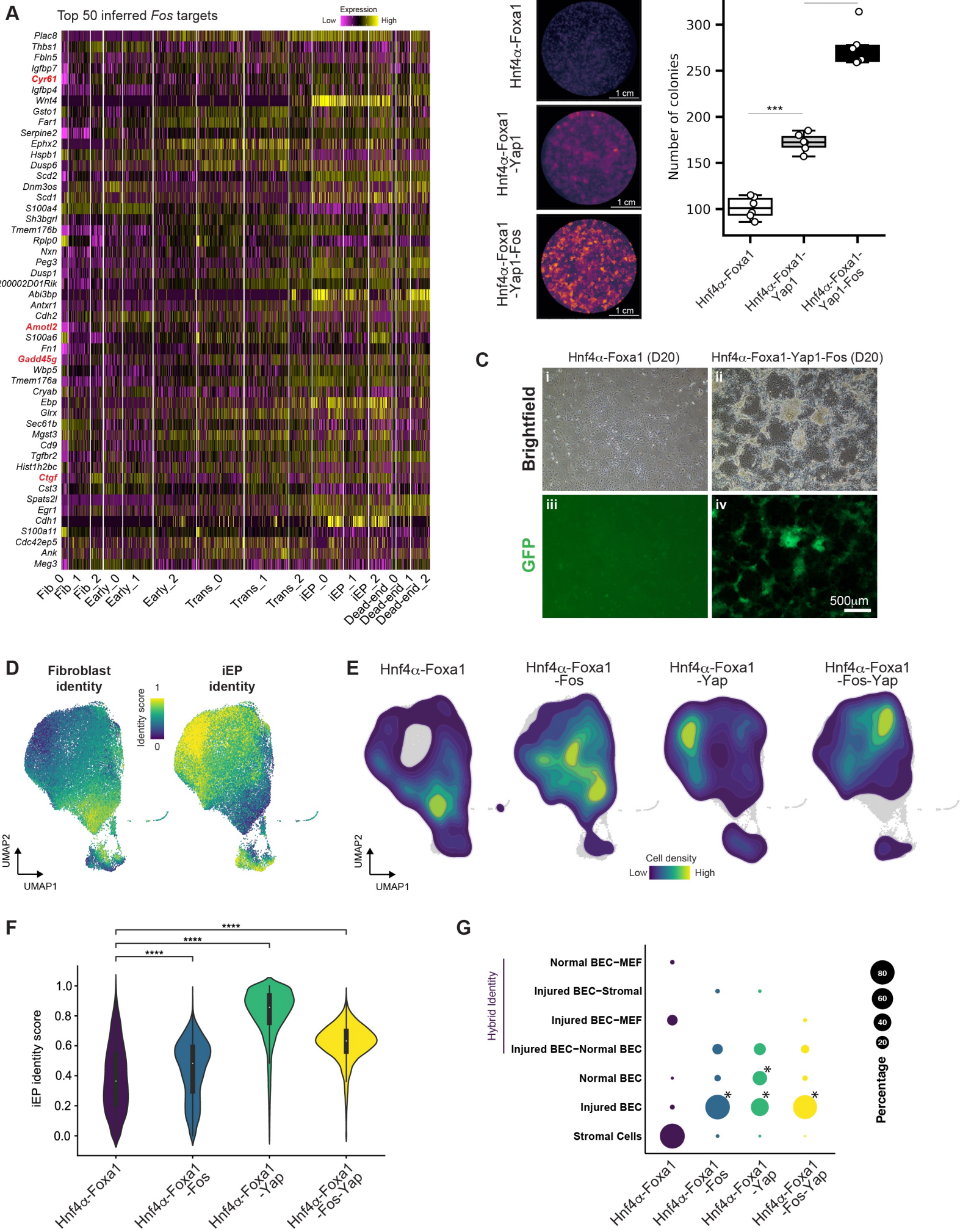
A **ATAC-seq: Identify regulatory candidate genes** **Active connection identification with Machine Learning Models**
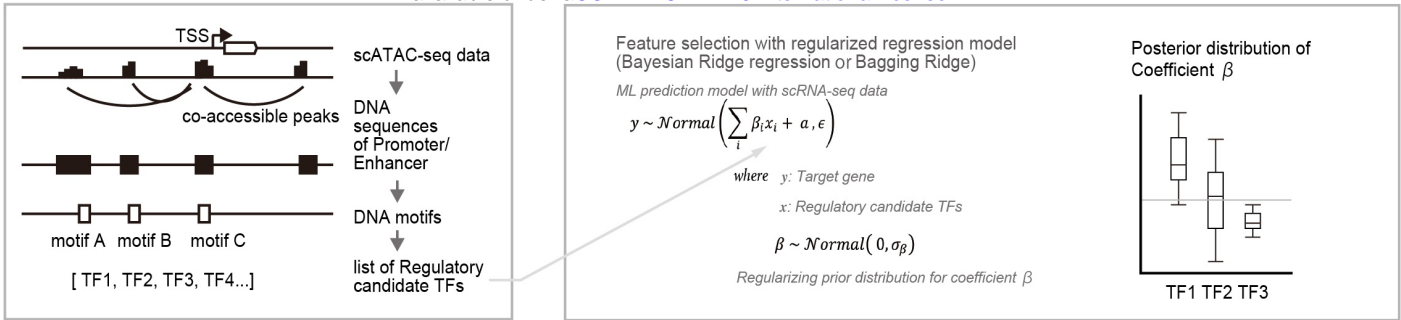
B
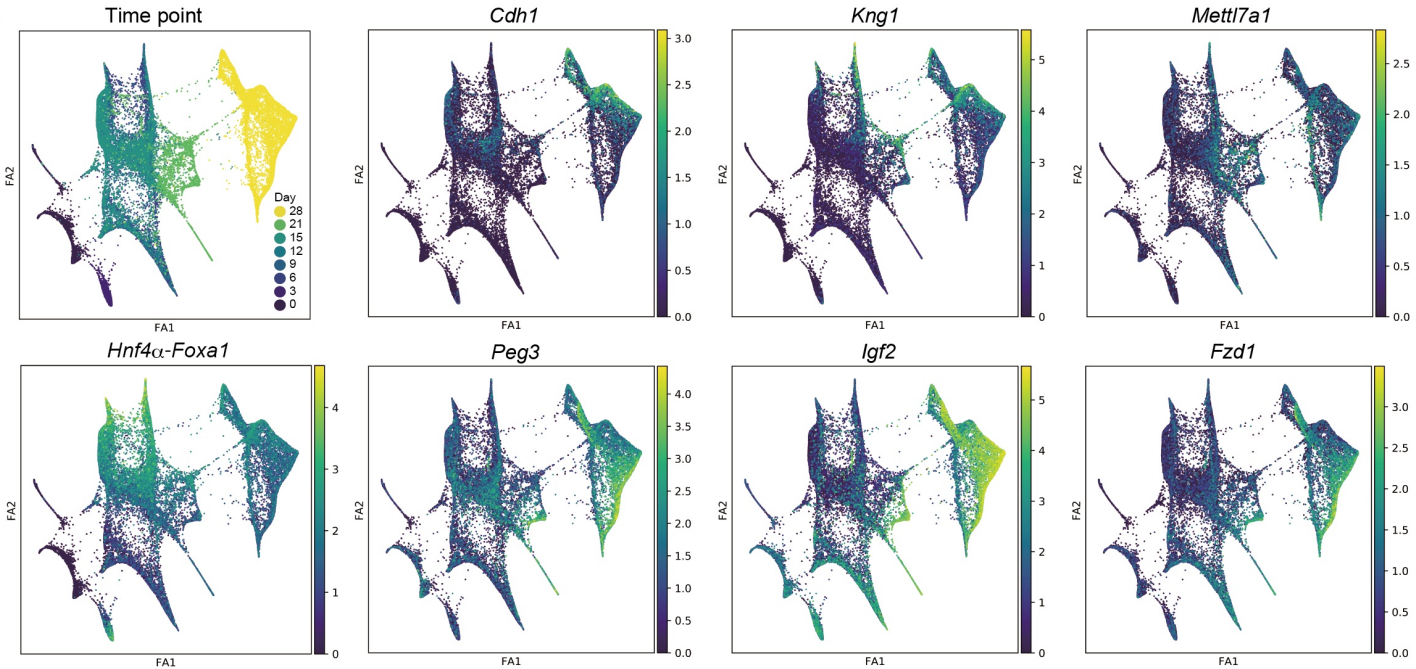
C
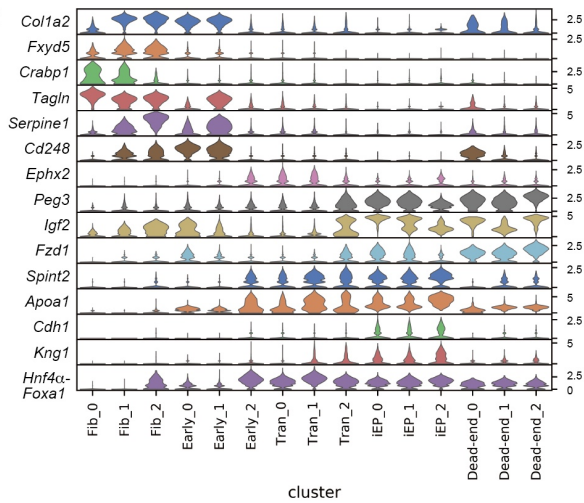
D
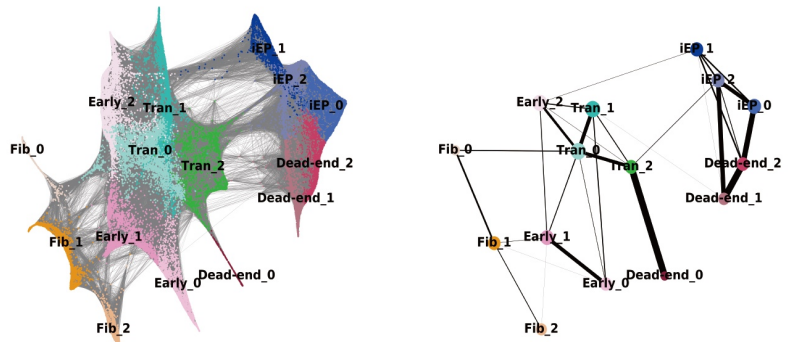
E
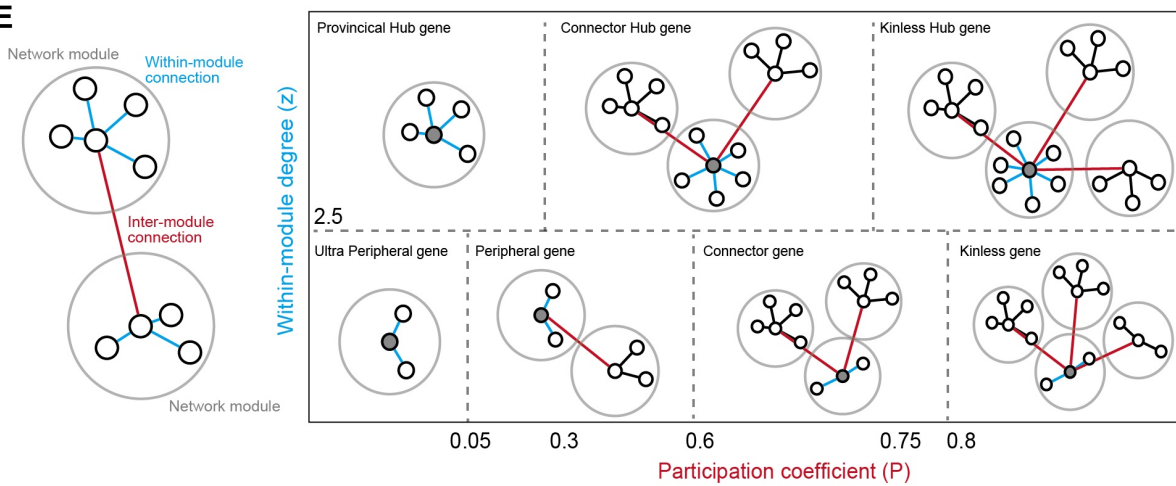
F **Clonal composition**
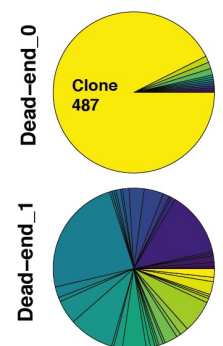
Supplementary Figure 2
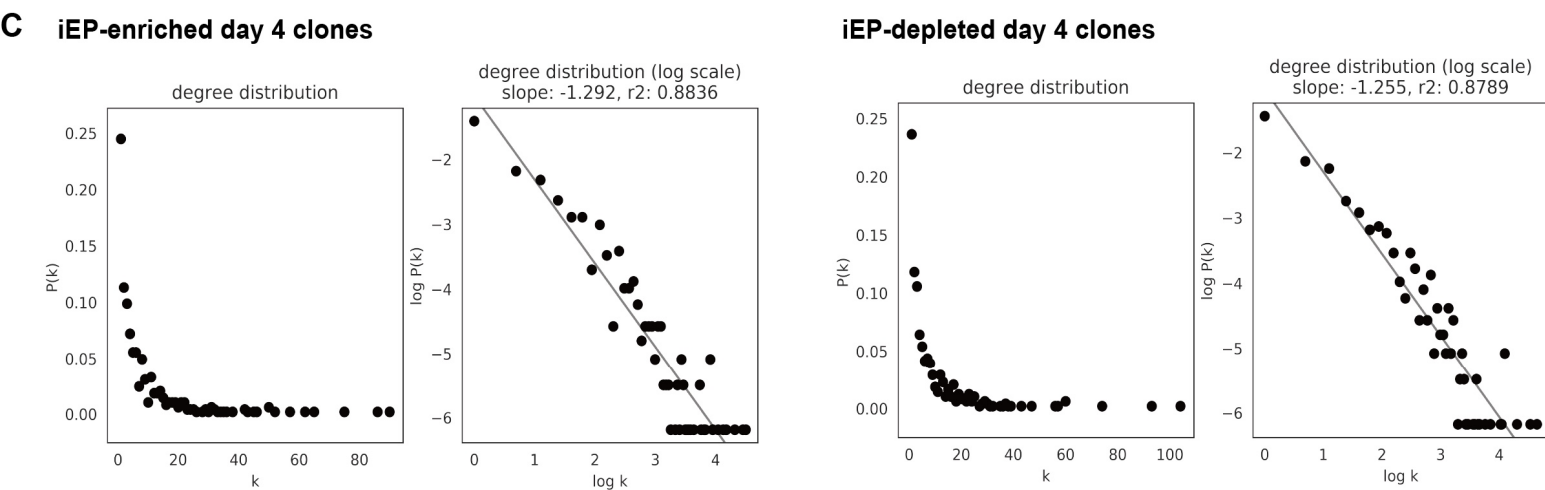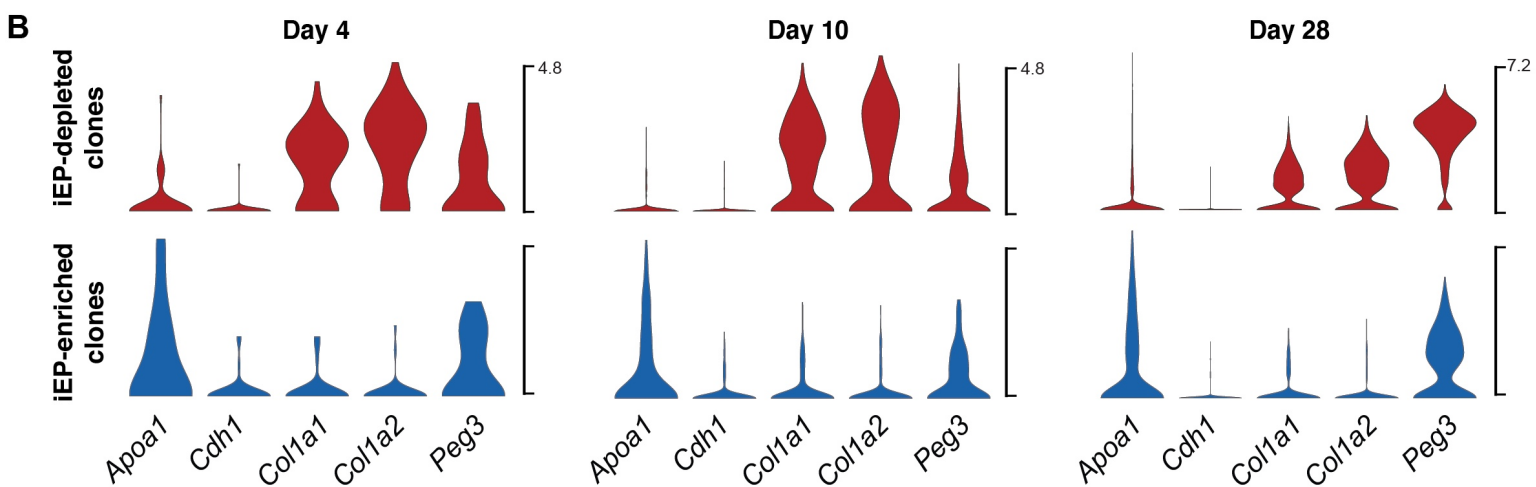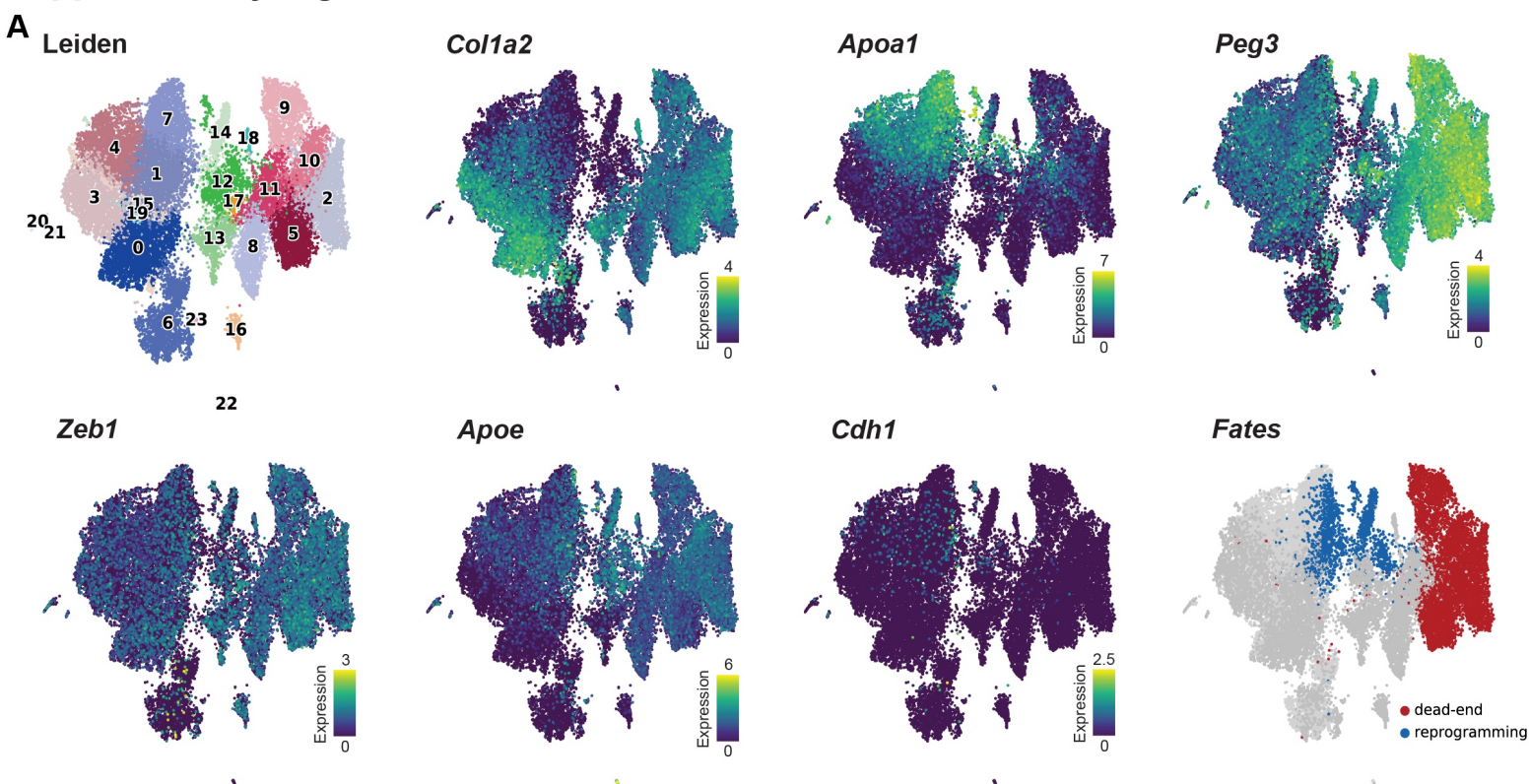
Supplemental Figure 3
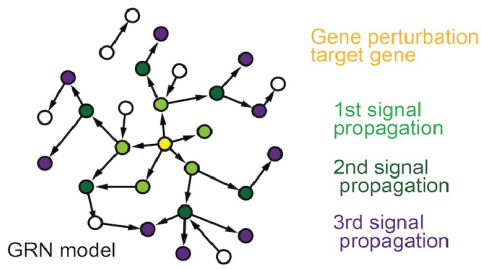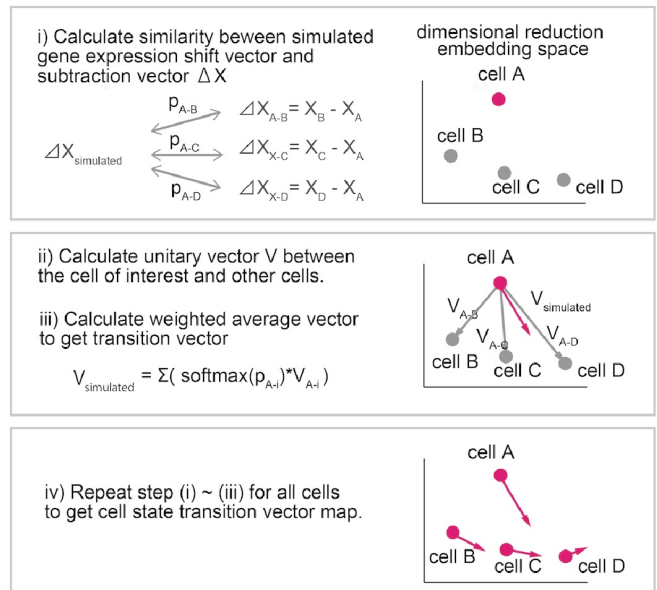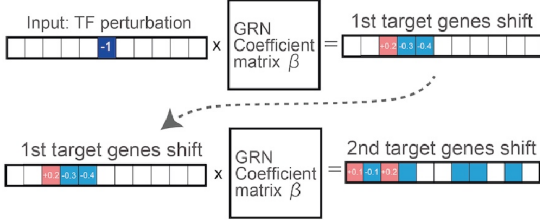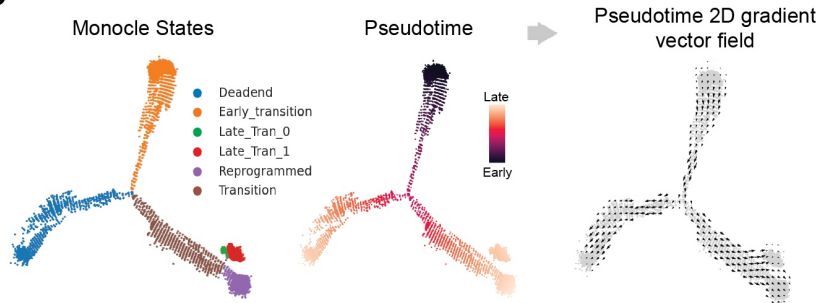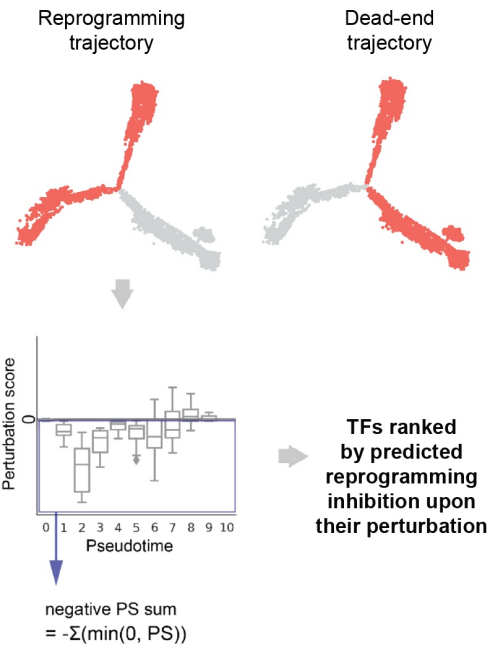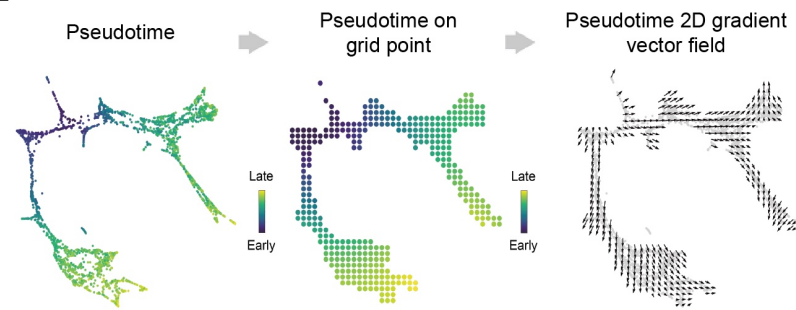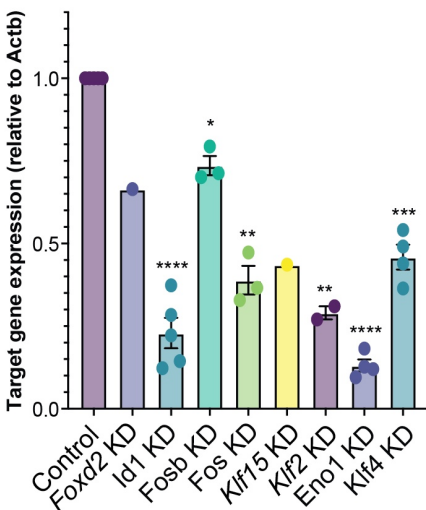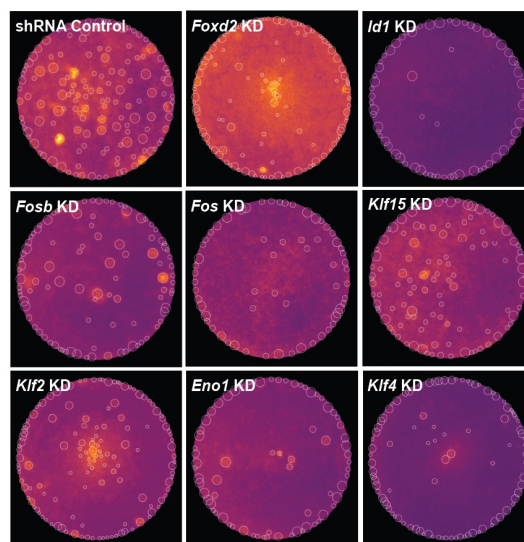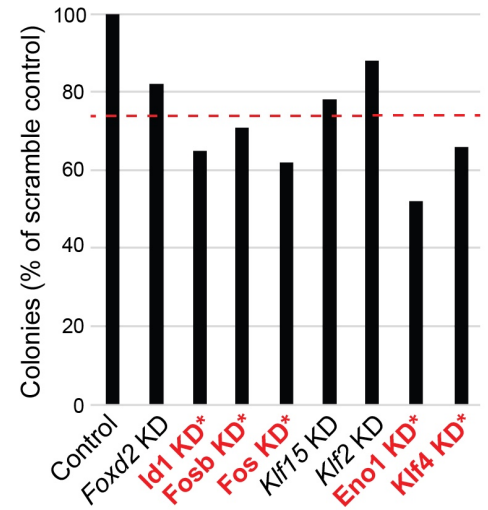
Supplemental Figure 4

Supplemental Figure 5