

Reviewed Preprint

v1 • August 6, 2026

Not revised

✉ For correspondence:

Nikhil.Munshi@UTSouthwestern.edu

Gary.Hon@UTSouthwestern.edu

* These authors contributed equally

Competing interests: No competing interests declared

Funding: See page 29

Reviewing editor: Paschalis Kratsios, University of Chicago, United States

© 2026, Duan et al. This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Combinatorial transcription factor interactions drive modular gene regulatory networks

Jialei Duan^{1,*}, Boxun Li^{1,*}, Kartik Kulkarni^{2,*}, Gabriela Orquera-Tornakian², Daryl Barth¹, Lei Wang¹, Vrushali Pandit², Joel Liou¹, Nikhil V Munshi^{2,3}✉, Gary C Hon^{1,3}✉

¹Laboratory of Regulatory Genomics, Cecil H. and Ida Green Center for Reproductive Biology Sciences, Division of Basic Reproductive Biology Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, United States • ²Department of Molecular Biology, Department of Internal Medicine, Division of Cardiology, McDermott Center for Human Growth and Development, UT Southwestern Medical Center, Dallas, United States • ³Hamon Center for Regenerative Science and Medicine, Dallas, United States

eLife Assessment

This study offers a large-scale resource that maps the transcriptional landscape following combinatorial transcription factor overexpression. The authors propose a modular framework for understanding gene regulatory networks and identify transcription factor cocktails for cellular reprogramming. While the dataset and proposed principles are **valuable**, the strength of the evidence is **incomplete**, in part due to the complexity of the computational analyses and limited orthogonal experimental validation. The study will be of interest to the fields of gene regulation and developmental biology.

<https://doi.org/10.7554/eLife.111977.1.sa3>

Abstract

Transcription factors (TFs) cooperatively drive gene regulatory networks (GRNs) to establish transcriptional states. Forced induction of TFs in combination can reprogram cell state by supplanting existing GRNs. Thus, TFs and GRNs are the building blocks to engineering transcriptional state. However, one key challenge is that the relationship between TF combinations and GRNs remains largely uncharacterized and difficult to accurately predict. Here, we apply single-cell overexpression screens to map the combinatorial activities of ~100 TFs to gene expression states. Our analysis identifies diverse TF combinations driving cell-type specific regulatory programs. Notably, different TF combinations induce shared gene sets with cell-type specific functions, suggesting a modular regulatory architecture of the transcriptome. Furthermore, we define pairwise TF interactions and show that cooperative interactions improve transcriptional reprogramming. Finally, we developed tools to predict combinatorial TF phenotypes. These findings improve our understanding of cell state and how to manipulate it for biomedical applications.

Highlights

- Combinatorial over-expression screens for ~100 transcription factors (TFs).
- Diverse TF combinations drive cell-type specific regulatory programs.
- TF regulatory networks reveal a modular regulatory architecture of the transcriptome.
- TF-TF interactions and predictive models enhance reprogramming cocktails.

Introduction

Single-cell genomics is rapidly defining transcriptional states across a spectrum of cell types and diseases ^{1–3}, providing rigorous evidence that transcriptional states define functional cell states. However, despite these extensive maps, the drivers of gene expression states remain less characterized.

Transcription factors (TFs) act cooperatively to drive gene regulatory networks (GRNs) and establish a cell's transcriptional state ^{4–7}. Although developmental GRNs are hierarchical and often unidirectional ^{8,9}, direct reprogramming by TFs can rewire GRNs to ectopically express lineage-specific genes and potentially create alternative cell types. For example, Oct4, Sox2, and Nanog coordinately regulate pluripotency gene expression in embryonic stem cells ⁷, but forced induction of TF combinations can ectopically activate GRNs in other cells ^{10,11}. In extreme cases, this results in functional reprogramming, for example in the conversion of fibroblasts to iPS cells with Oct4, Sox2, Klf4, and Myc ^{12,13}. Similarly, in cardiac reprogramming, GHMT (Gata4, Hand2, Mef2c, Tbx5) synergistically activate cardiac enhancers and regulatory programs ¹¹. Despite these successes, identifying reprogramming TFs remains challenging for several reasons: the difficulty of engineering reporters of successful reprogramming, the lack of candidate reprogramming factors rooted in developmental biology ^{12,14}, and the slow, labor intensive, and low throughput nature of traditional approaches. To address these problems, we and others have developed single-cell screens to identify TF cocktails for transcriptional reprogramming ^{15–18}. Our screening method was named Reprogram-Seq (1.0), which we continue to improve here ¹⁶.

While diverse studies rigorously support TF-mediated reprogramming ^{14,19–21}, *functional* reprogramming (full cell state conversion) can be inefficient. In contrast, *transcriptional* reprogramming is more readily attainable and represents a prerequisite for functional reprogramming. In turn, transcriptional reprogramming can be distilled into combinations of GRNs that are activated by combinations of TFs. Although individual TFs drive transcription, most reprogramming cocktails consist of 2 or more TFs ²² that often interact non-linearly to produce emergent properties ^{23,24}. For example, cooperativity between TFs can change the DNA binding specificities of paired TFs ²⁵, and specific TF pairs reprogram distinct neuronal subtypes, while individual TFs do not ²⁶. These observations indicate that TF pairs represent the basic unit of meaningful GRN reprogramming. Thus, empirical testing of TF cooperativity and understanding these interactions will improve our ability to accurately identify the building blocks necessary for predicting reprogramming cocktails.

While combinatorial TF interactions are important for reprogramming, their nonlinear interactions remain difficult to predict. One common approach to predict *individual* TF activity is to construct GRNs by examining transcriptomic datasets ^{27–30}. However, GRN inference remains a difficult computational problem because important parameters are either missing or incomplete, such as the association of TFs, binding motifs, and promoters ³¹. Aside from a handful of successes ^{32,33}, the problem of predicting *combinatorial* TF activities is more complex. Despite these known pitfalls, contemporary methods critically rely on predicted GRNs to predict TF cocktails ^{29,30,34,35}. Deconvoluting the relationship between TFs and their target genes could be aided by direct perturbation of TF combinations.

Recent studies have comprehensively cataloged the functional activities of ~1800 human TFs individually ^{17,36}, but the space of combinatorial TF activities is too immense to be comprehensively sampled (~1.6 million combinations of 2 TFs, ~970 million combinations of 3 TFs). Therefore, systematic cataloging of all possible higher-order combinatorial perturbations is not possible given current technologies. Here, we perform combinatorial perturbation studies of ~100 TFs in mouse embryonic fibroblasts (MEFs) for applications in reprogramming. By measuring the transcriptional outcomes of TF over-expression, these experiments directly measure the gene regulatory networks driven by TFs, define the interactions of TF pairs in the context of reprogramming, and reveal a shared regulatory architecture of common gene modules. An initial study of fitness phenotypes in a loss-of-function paradigm focused analyses on a specific subset of genetic interactions ³⁷. In the current study, we implement and extend this framework to

examine combinatorial interactions between TFs in a gain-of-function, lineage reprogramming model system. Altogether, these approaches represent a generalizable framework for transcriptional, and ultimately functional, reprogramming.

Results

Combinatorial over-expression screens for 105 transcription factors

Reprogram-Seq couples combinatorial over-expression of a library of TF open reading frames (ORFs) and single-cell RNA-Seq, for high throughput measurements of TF perturbations and resulting transcriptional states in individual cells (Figure 1A) ¹⁶. We introduced TF-specific barcodes immediately downstream of the TF transgene to minimize TF/barcode recombination during lentivirus production ^{38,39}. A PCR handle upstream of the TF barcode facilitates construction of perturbation libraries. We synthesized codon optimized TFs and barcodes, and validated synthesis by sequencing. To reduce the heterogeneity of lentiviral titers across ORFs, we developed an iterative sequencing-based strategy to normalize relative viral titers across pools of TFs, and we incorporated a GFP control to control absolute viral titers (see Methods) ^{40–42}. In benchmarking experiments, we confirm the ability to precisely control and evaluate viral titer prior to sequencing (Figure S1A–D).

To apply Reprogram-Seq in a cost-effective way, we sought to minimize the number of TFs synthesized while maximizing the cell types where these TFs had reprogramming potential. Using Tabula Muris as a reference, we identified 105 cell-type specifically expressed TFs spanning 7 diverse target cell types (Figure 1B): cardiomyocytes (CM, 29 TFs), epicardial cells (EC, 10 TFs), lung endothelial cells (LEC, 26 TFs), mesenchymal stem cells (MSC, 29 TFs), neuroendocrine cells (NC, 15 TFs), skeletal muscle satellite cells (SMSC, 27 TFs), and stromal cells (SC, 22 TFs) (Supplemental Table 1). This synthesized set of TFs represents ~7% of TFs encoded in the mouse genome.

For each of the 7 target cell types, we applied Reprogram-Seq to combinatorially overexpress a subset of the 105 TFs relevant to its transcriptional state (Supplemental Table 1) in mouse embryonic fibroblasts (MEFs). We refer to each of these 7 experiments as a batch. We infected MEFs across a range of lentiviral titers (low titer averaging 1–2 TFs per cell; high titer averaging 5–6 TFs per cell), to enable different downstream applications (Figure 1C–E). After 14 days of lentiviral infection, we performed single-cell RNA-Seq (scRNA-Seq) to detect the transcriptomes and TF perturbations in individual cells. Overall, we recovered ~200k high-quality MEF-derived cells, with a median of 4 TFs per cell (Supplemental Table 2, Figure S1C). This dataset spans >100k unique TF combinations, with 1335 TF combinations having at least 10 cells (Figure 1D). As a negative control, we also sequenced MEFs infected with lentiviruses expressing fluorescent proteins instead of TFs (Figure S1D–E).

Combinatorial TF over-expression reprograms MEFs to diverse states

To examine how combinatorial over-expression of TFs alters transcriptional states, we compared MEF-derived cells to reference mouse atlases ^{16,43} (Figure 1E). By co-embedding the cells, we observed that reprogrammed MEFs (Figure 1F) transcriptionally cluster with diverse *in vivo* cell states (Figure 1G). 13 out of 41 clusters contained significant enrichment of perturbed cells, with an average of 4-fold enrichment over unperturbed MEFs (Figure 1H–I). These clusters contained 41% of all perturbed MEFs. To highlight the single-cell resolution of the TF perturbations, we examined known reprogramming cocktails. For example, we previously identified Atf3, Hand2, and Gata6 (3F) as an epicardial-reprogramming cocktail. Using Reprogram-Seq, we find that cells perturbed with 3F are significantly enriched in the epicardial cluster ($p < 2.2 \times 10^{-16}$) (Figure 1J). Notably, consistent with the observation that known reprogramming cocktails often contain

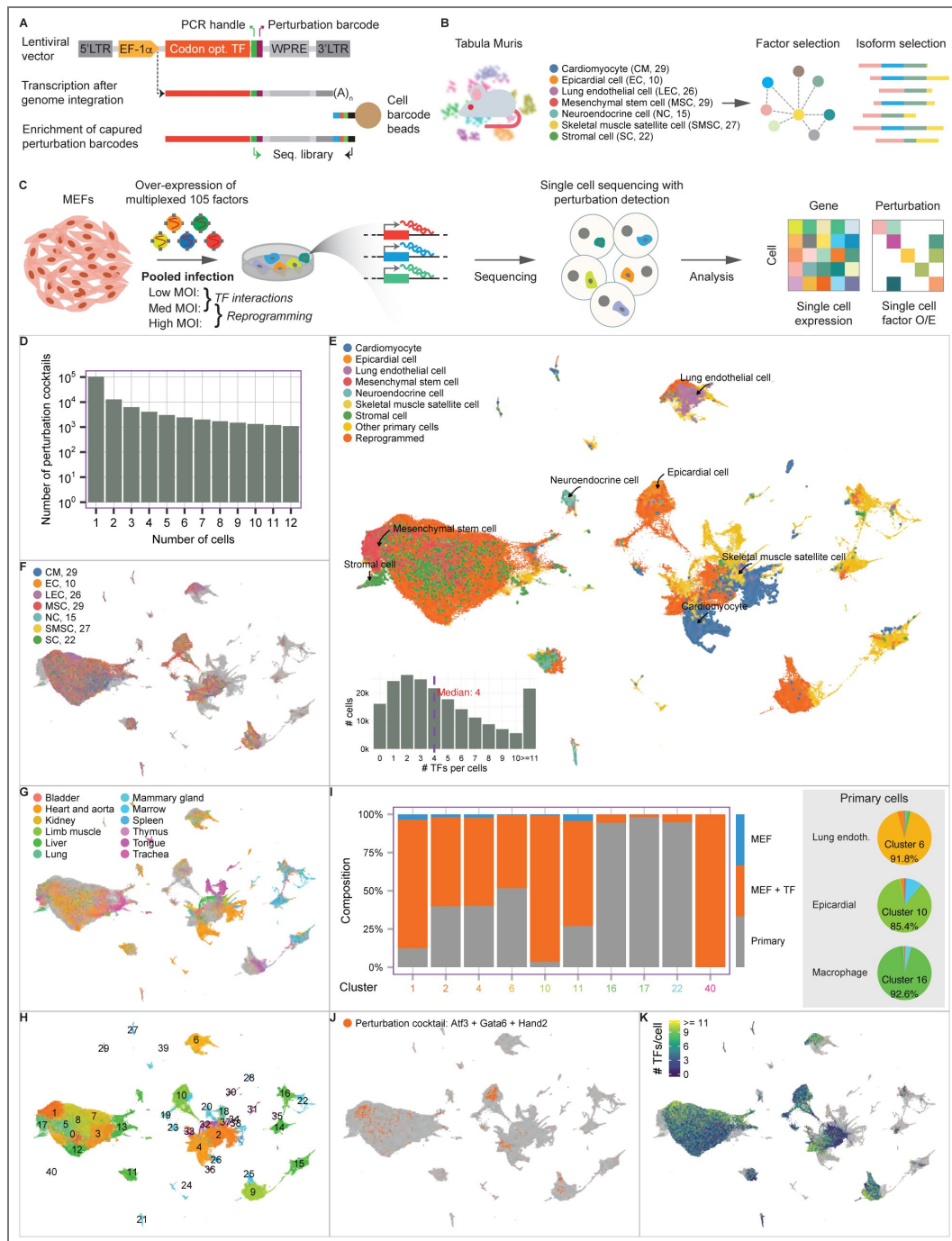


Figure 1. Diverse transcriptional states from combinatorial TF reprogramming.

A. Schematic of Reprogram-Seq lentiviral TF ORF barcoding system compatible with single-cell RNA-Seq detection. B. Schematic of Reprogram-Seq applied to Tabula Muris cell types. For 7 cell types, 10 to 29 candidate TFs were prioritized for combinatorial over-expression studies. C. For each of the 7 cell type batches, MEFs were infected with lentiviral libraries for the candidate TFs, followed by single-cell RNA sequencing and analysis. D. Distribution of the number of cells recovered for distinct TF cocktails. E. UMAP visualization of reprogrammed MEFs and reference cell types. Reference cells are colored by cell type. Reference cell types (excluding “Other primary cells”) are plotted as larger dots to improve visibility. (insert) Distribution of the number of TFs identified in sequenced cells. F. UMAP feature plot, with cells colored by reprogramming batch. G. UMAP feature plot, with reference cells colored by origin tissue. MEF-derived cells are grey. H. UMAP feature plot, with cells numbered and colored by cluster. I. (left) For each cell cluster indicated, shown is the origin of cells in the cluster (control MEF, reprogrammed MEF, primary cell). The “Primary” label on the bottom indicates the most frequent primary cell type in the cluster. (right) For each primary cell type indicated, the pie chart shows the distribution of the primary cells across the clusters. J. UMAP feature plot, where cells receiving the epicardial TF combination *Atf3* + *Gata6* + *Hand2* are colored orange. K. UMAP feature plot, with cells colored by the number of distinct exogenous TFs detected.

multiple TFs, the number of TFs induced in MEFs correlates with the frequency of transcriptional reprogramming (Figure 1K [↗](#)). Overall, these results suggest that our strategy effectively identifies TF combinations that perturb global transcriptional states.

We next analyzed the reprogramming outcome of each batch. First, for each cluster that represents a primary cell type, we examined the batch specificity of the reprogrammed cells in the cluster. That is, are all reprogrammed cells derived from the expected batch (e.g., reprogrammed CMs from the CM batch) or do some of them come from ‘non-specific’ batches (e.g., reprogrammed CMs from non-CM batches)? On one extreme, Cluster 6 represents cells derived from the LEC Reprogram-Seq experiment (which contained 26 TFs targeting lung endothelial cells) and 91.8% of all primary cells endothelial cells are in this cluster, demonstrating high batch specificity (Figure 1I [↗](#)). Unexpectedly, other clusters had almost even representation of cells from multiple batches of Reprogram-Seq experiments, even though each batch contains different TFs (Figure 1G [↗](#)). Second, we considered the specificity of the reprogramming outcome of each batch. Consistent with the results above, we find that one reprogramming batch often yields cells in multiple transcriptional states (Figure S1G [↗](#)) (19 transcriptional states per batch on average). For example, while MEFs induced with CM TFs were significantly enriched in the cardiomyocyte cluster, they were also enriched in clusters representing skeletal muscle satellite cells (Figure S1F-G [↗](#)). Third, we also surprisingly find that perturbed MEFs are significantly enriched for unexpected cell states that were not intended targets for transcriptional reprogramming (Figure 1I [↗](#)). Notable examples include Clusters 16 and 22, which contain distinct hematopoietic cell types including macrophages (Cluster 16, $p = 8.15E-102$) and granulocytes (Cluster 22, $p = 4.69E-42$). These clusters are 5.5-fold more enriched in perturbed cells over control MEFs.

Overall, these results indicate that combinatorial TF induction reprograms MEFs towards diverse transcriptional states, many of which have similarity to in vivo cell states. Our observations highlight that diverse TF combinations can reprogram MEFs to similar transcriptional states. For instance, 16 out of 26 predicted TFs for lung endothelial cell reprogramming are enriched in cluster 6. These results also demonstrate that some TF combinations can reprogram towards unintended cell fates, which is consistent with documented examples of TFs with promiscuous activation of cellular regulatory programs [44–47](#).

TF combinations drive cell-type specific regulatory programs

While clustering analysis can identify global transcriptional changes, it does not distinguish the gene regulatory programs driven by different TF combinations. To address this, we first identified 1335 distinct TF combinations (Figure 2A [↗](#), Supplemental Table 3) that met stringent quality filters. This set consists of 104 TF singles, 801 TF doubles, 322 TF triples, and 107 combinations with 4 or more TFs (Figure 2B [↗](#)). We performed MDE embedding with parameters tuned to distinguish subtle transcriptional changes (Figure 2C [↗](#)) [48](#). In this embedding, different TF combinations occupied distinct transcriptional spaces. Many TF combinations were represented by multiple reprogramming batches (e.g., TF Lmo2 in CM, LEC, NC, SC, EC batches), suggesting that batch effects are limited (Figure 2D-E [↗](#)). We then used the density-based HDBScan method to identify 51 perturbation clusters, each consisting of multiple TF combinations (Figure 2F [↗](#), Figure S2A [↗](#)). Clusters also contained distinct TF combinations from the same reprogramming batch (Figure 2D [↗](#)). For example, in Perturbation Cluster 6 (Figure S2A [↗](#)), 81.8% of TF combinations are derived from the MSC reprogramming batch. In contrast, Perturbation Cluster 1 represents experiments from almost all reprogramming batches.

Perturbation clusters represent TF combinations that share similar transcriptional impacts (Figure 2F [↗](#)). Since TFs drive gene regulatory networks (GRNs), the set of genes induced and repressed by a given combination represents an experimentally estimated GRN. To establish these GRNs, we identified the genes induced and repressed in perturbation clusters (Supplemental Table 4). To evaluate the biological relevance of these GRNs, we annotated them with several functional databases including Reactome, Gene Ontology, and KEGG (Figure 2F-G [↗](#), Supplemental Table 5). One notable example is Perturbation Cluster 1, which induced genes highly enriched for cardiac function, including muscle contraction (Reactome ID: R-MMU-397014, $p = 2.01E-28$) and skeletal

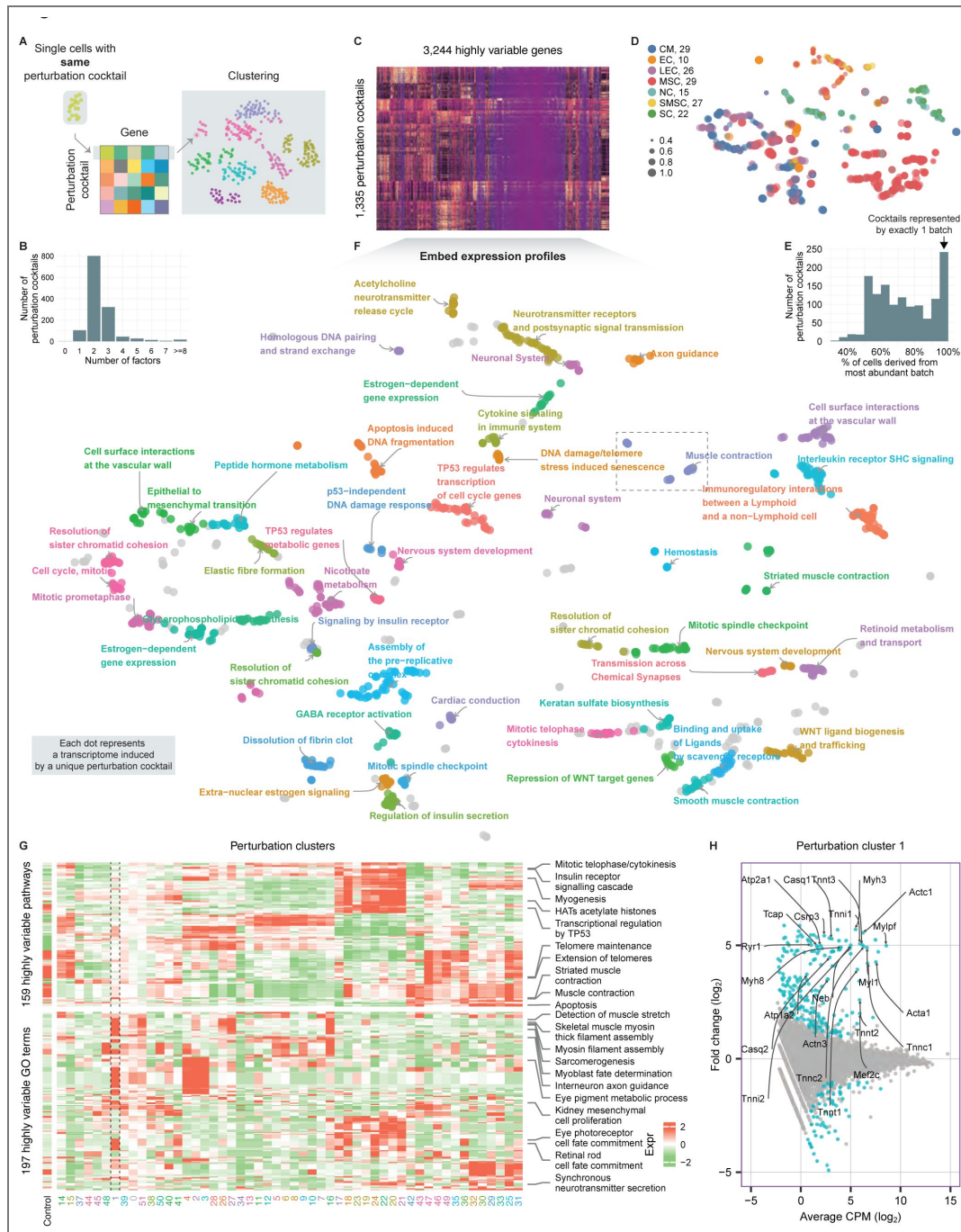


Figure 2. TF combinations drive cell-type specific regulatory programs.

A. Schematic to visualize transcriptional phenotypes from combinatorial TF over-expression. Cells induced with identical TF combinations were combined. B. For the 1335 TF combinations analyzed, shown is the distribution of the number of TFs in each combination. C. The heatmap indicates the expression of 3244 variably expressed genes across the 1335 TF combinations. D. MDE embedding of transcriptomes for 1335 TF combinations. Each dot indicates a TF combination, colored by experimental batch. When a given TF combination is recovered by multiple batches, the size of each dot indicates the fraction of cells derived from the more abundant batch. E. Across the 1335 TF combinations, shown is the fraction of cells derived from the most abundant batch. F. MDE embedding, with TF combinations colored by perturbation cluster and annotated with enriched Gene Ontology terms. G. Heatmap illustrating the enrichment of Gene Ontology terms and pathways across the 51 perturbation clusters. H. The MA plot indicates the differentially expressed genes in perturbation cluster 1 (teal). Several genes with cardiac roles are indicated.

muscle myosin thick filament assembly (GO ID: GO:0030241, $p = 3.8E-05$) (Figure 2G [↗](#)). Consistently, well-known cardiac genes were highly induced in Cluster 1. For example, we observed a 3.8-fold induction of *Tnnt2* ($p = 1.13E-06$) and a 15.7-fold induction of *Tnnt3* ($p = 3.87E-12$) (Figure 2H [↗](#)). Thus, TFs in Perturbation Cluster 1 induce a GRN representing a muscle gene expression program. We also identify other candidate cell-type specific GRNs. These include GRNs that have enrichment for: nervous system development (Cluster 6, $p = 2.04E-03$), axon guidance (Cluster 18, $p = 4.48E-04$), Estrogen-dependent gene expression (Cluster 22, $p = 2.52E-03$), striated muscle contraction (Cluster 16, $p = 1.17E-22$), signaling by insulin receptor (Cluster 37, $p = 5.32E-03$), and cardiac conduction (Cluster 34, $p = 2.20E-02$). These analyses demonstrate that TF combinations can drive diverse cell-type specific gene programs (Figure 2F [↗](#), Figure S2B [↗](#)).

TF combinations drive modular gene regulatory networks

The previous analysis focused on TF-induced GRNs at the level of perturbation clusters (containing multiple TF combinations) and pathways (containing multiple genes). To gain more granular knowledge of the GRNs, we next performed analysis at the level of distinct TF combinations and genes (Figure 3A [↗](#)).

First, to gain insights into individual genes, we examined the set of commonly induced and repressed genes in each perturbation cluster. On average, each perturbation cluster significantly induced 834 genes and repressed 607 genes (Figure 3B [↗](#)). The union across all perturbation clusters spans a total of 7751 differentially expressed genes. Surprisingly, despite the sparse sampling of the complete perturbation space ($\sim 7\%$ of all TFs, $\sim 7\%$ of all single TF perturbations, and $\sim 0.04\%$ of all double TF perturbations), we find that the TF-induced GRNs span $\sim 23\%$ of all genes in the transcriptome. This observation suggests that sparse subsampling of combinatorial TF perturbations could be used to span transcriptome-wide gene regulation.

Next, to examine specificity, we assessed how often a given gene was recovered across GRNs. We observed that TF-induced GRNs highly overlap: 77.6% of genes were induced in multiple perturbation clusters (Figure 3C [↗](#), Figure S3A [↗](#)). For example, *Mylpf* was highly induced in both perturbation clusters 1 and 16 (Figure 3D [↗](#)). Overall, each induced gene was found in a median of ~ 5 perturbation clusters (0.90 specificity score; where 0 indicates a gene found in all modules). Genes with the lowest specificity were significantly enriched for Gene Ontology terms including Regulation of multicellular organismal development ($p = 7.9e-06$) and Response to external stimulus ($p = 3.3e-05$) (Figure 3E [↗](#)), indicating that the common features of GRN rewiring are similar to developmental responses. In contrast, genes with the highest specificity and induced in a small number of perturbation clusters are significantly enriched in DNA-binding TF activity ($p = 6.7e-10$), suggesting the activation of cell-type specific genes. Overall, these observations imply a shared regulatory structure of GRNs, and that multiple TF perturbations can be used to induce a given gene.

To examine the diversity of TF perturbations, we compared the composition of TFs across perturbation clusters (Figure 3F [↗](#), Figure S3B [↗](#)). Supporting our observation that the transcriptomes induced by different clusters are distinct, we found that TF perturbations were also generally distinct across perturbation clusters. Notably, 25.0% of all perturbation cluster pairs share no TFs, and most pairs (61.6%) have very low TF overlap (Jaccard similarity < 0.2) (Figure 3F [↗](#)). These results suggest that perturbation clusters span diverse sets of TFs.

The observations above indicate a striking modularity of TF-driven GRNs at the level of individual TF cocktails and genes. Figure 3G [↗](#) illustrates an example (Figure S3C [↗](#)). Perturbation cluster 1 induces a specific module of cardiac and muscle genes. These genes are rarely induced by other perturbation clusters. Perturbation cluster 1 has 27 distinct TF cocktails spanning 38 unique TFs evaluated in 6 batches of Reprogram-Seq experiments. In contrast, Perturbation cluster 2 induces a module of endothelial marker genes, including *Lilr4b* and *Lilrb4a*, that is distinct from Perturbation cluster 1. Notably, the TF combinations in Perturbation cluster 2 are highly enriched for the ETS transcription factor *Erg* (*Erg*-205, 61.5%; *Erg*-203, 20.5%) or *Flt1* (25.6%), suggesting that these TFs induce an endothelial gene program. Interestingly, a small set of TF combinations in perturbation clusters 1 induce both cardiac and endothelial gene programs. Highlighting the

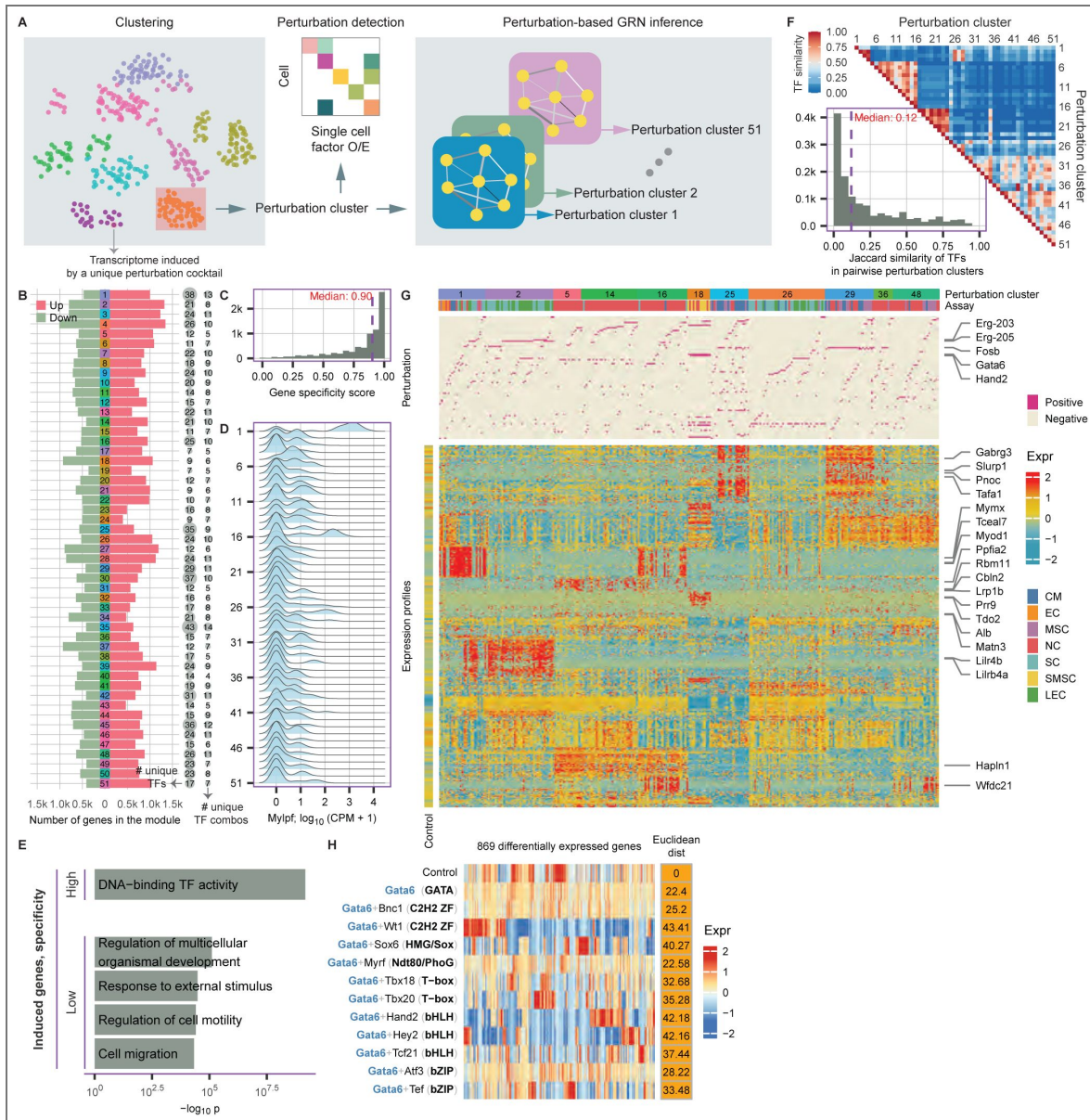


Figure 3. TF combinations drive modular gene regulatory networks.

A. Schematic to estimate gene regulatory networks driven by perturbation clusters. Cells in each perturbation cluster are grouped, followed by differential expression analysis. B. For each perturbation cluster, shown is the number of differentially expressed genes induced and repressed. (left) The number of unique TFs in the perturbation cluster. (right) The number of unique TF combinations. C. Across all differentially expressed genes identified in B, shown is the specificity across modules (1 - fraction of occurrence across modules). A specificity score of 0 indicates a gene is found in all modules. D. Expression of Myl2 in cells across the 51 perturbation clusters. E. Gene ontology enrichment for high specificity and low specificity genes. F. The heatmap indicates the overlap fraction of TFs across perturbation clusters. The histogram quantifies Jaccard similarity of TFs across perturbation clusters. G. Heatmap indicating gene expression patterns (bottom) across perturbation clusters. Each column indicates a distinct TF combination (middle), grouped by batch/assay and perturbation cluster (top). H. The heatmap illustrates the expression of 869 differentially expressed genes in cells induced with Gata6 alone or Gata6 + TF_x compared to control MEFs. TF_x spans 11 different TFs from different TF families (bold). For each Gata6 + TF_x pair, the values on the right column indicate Euclidean distance relative to control cells.

modularity of these gene expression programs, these TF combinations are also enriched for Erg and Flt1. Consistent with these observations, the ETS family of TFs are well-established lineage-specifiers of endothelial cell fate, and lineage tracing studies of Isl1⁺ cardiac progenitors demonstrate a common origin of cardiomyocytes and endothelial cells ^{49,50}, and Nkx2-5 transactivates Etv2 in multi-potent progenitors cells to regulate endothelial/endocardial cell fate ⁵¹.

The analyses above indicate that TF combinations exhibit highly modular activities. To illustrate an example, we focused on the epicardial TF Gata6 and its interactions with 11 other TFs. Several TFs (Bnc1, Myrf, Atf3), when paired with Gata6, exhibit similar effects on the transcriptome compared to Gata6 induction alone (Figure 3H [↗](#), Figure S3D [↗](#)). In contrast, pairing the remaining 8 TFs with Gata6 yields transcriptional responses that are distinct from Gata6 alone. Notably, each of these pairs Gata6 + TF_X induces a unique module of genes. For example, Gata6 + Wt1 induces an expression program most similar to primary epicardial cells, including the epicardial genes Bnc1, Gpm6a, Myrf and Upk1b. Interestingly, we observe distinct behaviors of Gata6 + TF_X pairs even when TF_X belong to the same family of transcription factors, suggesting diverse functions for TFs that share structural similarities at the level of protein domains, and often binding motifs (Figure S3E [↗](#)).

Top-down analysis nominates TF cocktails that reprogram transcriptional state

One goal of Reprogram-Seq analysis is to identify TF cocktails that reprogram transcriptional state. In a previous study, we used pseudotime analysis to nominate a cocktail of 3 transcription factors (Atf3, Gata6, Hand2; AGH) for epicardial reprogramming. Since the current study has greater resolution of cellular perturbations, we first revisited epicardial reprogramming with the goal of systematically ranking TF combinations. Pseudotime analysis illustrates a cellular trajectory from control MEFs to primary epicardial (Krt19⁺) cells that is spanned by reprogrammed MEFs (Figure 4A [↗](#), Figure S4 [↗](#)). Across the pseudotime trajectory, reprogrammed MEFs induce different signature genes, culminating in the endogenous activation of epicardial genes including Gpm6a, Wt1, and Krt19 (Figure 4B-C [↗](#)).

Consistent with our observations above (Figure 1 [↗](#)), we find that cells receiving more TFs generally exhibited better performance in transcriptional reprogramming compared to cells with fewer TFs (Figure 4D [↗](#), Figure S4B-C [↗](#)). However, this trend does not appear to be additive, as cells with the highest number TFs (>6) only exhibited moderate performance. These observations indicate the utility of cells with moderate to high titer in identifying novel reprogramming cocktails. In agreement with our previous study, we also observed that cells co-induced with AGH are highly enriched towards the end of pseudotime (93.3th percentile). Notably, we identify several TF combinations with pseudotime performance that exceeds that of AGH, including a 4-TF combination AGHW (AGH + WT1) (97.5th percentile). Consistent with this role in reprogramming, WT1 (Wilms' tumor 1) is a highly expressed transcription factor in developing epicardial cells and is an essential regulator of epicardial cell development ^{52,53}. Independent qPCR experiments validated newly identified TF cocktails for transcriptional reprogramming towards epicardial states (Figure 4E [↗](#)). Finally, to investigate reprogramming phenotypes outside of transcription, we examined the localization of the tight junction protein ZO1, which is a feature of epicardial cells. Confirming our approach, we observe that MEFs induced with the newly identified TF combinations exhibit strong ZO1 localization at cell-cell junctions (Figure 4F-G [↗](#)).

To test the generality of this approach, we extended this analysis to the remaining batches of Reprogram-Seq experiments (Figure 4H-I [↗](#), Figure S4D-E [↗](#)). By anchoring primary cells as a destination transcriptional state, we ranked TF cocktails for transcriptional reprogramming. For example, in the LEC experiment, this analysis recovers Aryl hydrocarbon receptor (Ahr) as a strong TF in several high-ranking TF cocktails. Ahr has an essential role in maintaining the function of lung endothelial cells during viral infection ⁵⁴. In the smooth muscle stem cell (SMSC) experiment, one TF appearing in several high-ranking combinations is Pax7, which has been documented as an important reprogramming factor for satellite cells ⁵⁵. Finally, while we do not

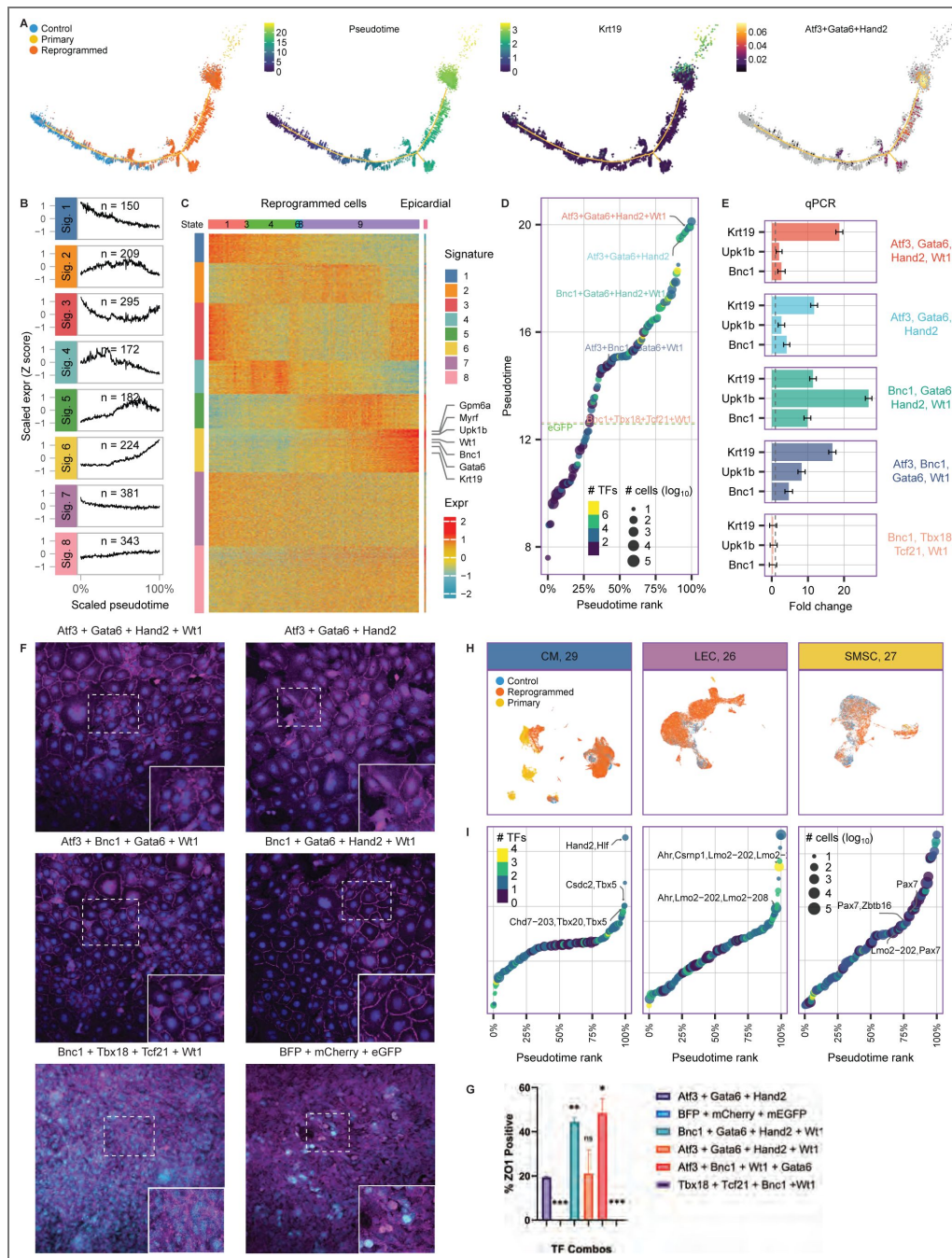


Figure 4. Prediction of TF combinations for transcriptional reprogramming.

A. Pseudotime trajectory of epicardial reprogramming. (left) Indicated are control MEFs (blue), TF-induced MEFs (orange), and primary epicardial cells (yellow). (middle left) Feature plot colored by pseudotime. (middle right) Feature plot colored by the expression of epicardial marker gene *Krt19*. (right) Feature plot colored by cells induced with *Atf3* + *Gata6* + *Hand2*. B. Clusters of distinct gene expression patterns across epicardial transcriptional reprogramming. Pseudotime is indicated by the x-axis. C. The heatmap shows the expression of the gene clusters in C across all MEFs induced with epicardial TFs (ordered by pseudotime). D. Distinct TF combinations are ordered by transcriptional reprogramming performance according to pseudotime. Several TF combinations are indicated. *Atf3* + *Gata6* + *Hand2* was previously identified as an epicardial reprogramming cocktail. E. Bulk qPCR validation of putative TFs for transcriptional reprogramming of epicardial cells. F. Immunostaining of tight junction protein ZO1 (a feature of epicardial cells), in MEFs after over-expression of putative TFs for epicardial reprogramming. G. Quantification of ZO1 positive cells in panel F. H. Across all reprogramming batches and targeted primary cells, shown is the clustering of control MEFs (blue), MEF-derived cells (orange), and primary cells (yellow). I. Pseudotime analysis to rank putative TFs for neuroendocrine and stromal cell reprogramming.

recover cells with simultaneous expression of the well-known cardiac reprogramming factors such as Gata4, Hand2, Mef2c, and Tbx5 in the CM experiment (because the combinatorial space was too large), we did recover enrichment of individual factors in high-scoring cocktails ⁵⁶.

TF-TF interactions influence reprogramming potential

The basic unit of combinatorial activity is a pair of TFs. Several modes of TF interaction include independence, dominance, and cooperativity ^{37,57–59}. To mathematically define how TF pairs functionally interact in the context of transcriptional reprogramming, we used a previously established framework to model interactions by linear regression ³⁷. Briefly, we modeled changes in gene expression observed in doubly perturbed cells (AB) as a linear combination of singly perturbed cells (A and B alone) (Figure 5A [↗](#)):

$$\delta AB = c_0 + c_1 \delta A + c_2 \delta B + \epsilon$$

Here, $c_1 \delta A$ and $c_2 \delta B$ represent linear effects and ϵ is an error term that encapsulates non-linear effects. We applied this linear model to 801 TF pairs. 14 pairs were excluded from the downstream analysis (F-test, $p \geq 0.05$, suggests this model doesn't work). On average, 595 cells were used for modeling (median, $10^{2.775}$) (Figure 5B [↗](#)). Overall, the linear model performed well, with most tested TF pairs yielding R^2 values above 0.5 (median 0.571) (Figure 5C [↗](#), Figure S5A [↗](#)). This good linear fit was also supported by the distance correlation metric (dcor, Figure S5B [↗](#)) ³⁷. Visual embedding of these linear model fit parameters indicate that TF interactions are diverse (Figure 5F–G [↗](#), Figure S5A–I [↗](#)). To gain more insights into TF interactions, we developed statistical tests to identify several classes of interactions (see Methods). Notably, the most frequent mode of TF-TF interaction is dominance (24.5%) (Figure 5E [↗](#), Figure S5C–F [↗](#)), where one TF masks the impact of another TF (e.g., Hand2 and Wt1; Figure 5G–H [↗](#)). Redundant interactions are the next most frequent (13.7%), where two TFs have similar transcriptional impacts, alone or in combination (e.g., Hey2 and Nkx2.5; Figure 5G–H [↗](#), Figure S5G [↗](#)).

Published studies indicate that reprogramming TFs can interact nonlinearly ^{25,37}. To represent one type of non-linear cooperative interaction between A and B, we augmented our linear model by incorporating a cooperative component ($c_3 \delta A \times B$), similar to a previous study ¹⁷ (Figure 5A [↗](#), Supplemental Table 6):

$$\delta AB = c_0 + c_1 \delta A + c_2 \delta B + c_3 \delta A \times B + \epsilon$$

Using this augmented linear model, we identified 46 TFs that behaved cooperatively ($c_3, p < 0.05$), where TFs are multiplicatively cooperative (e.g., Atf3, Gata6; Figure 5G–H [↗](#)). To characterize other types of non-linear interactions, we also identified 79 neomorphic interactions (large mean absolute error). (e.g., Etv1 and Mlip; Figure 5G [↗](#), Figure S5H [↗](#)).

TF-mediated reprogramming is the outcome of multiple TFs interacting to reconfigure the transcriptome. We wondered if linear (redundant, dominant) or non-linear (cooperative, neomorphic) interactions were more likely to lead to transcriptional reprogramming. To address this question, we examined the enrichment of linear and non-linear interactions in TF cocktails as a function of transcriptional reprogramming (as defined by pseudotime). We observed that TF-TF interactions from successfully reprogrammed cells are significantly enriched in cooperative interactions (one-tailed t -test, p -value = 3.44×10^{-14}) (Figure 5I [↗](#), Figure S5J [↗](#)). One example of a cooperative TF cocktail for epicardial reprogramming is Bnc1 and Gata6 ($c_3 = 0.14$, p -value = 7.05×10^{-8}). To test if this TF pair exhibits non-linear behavior, we compared with two other TF pairs from the epicardial batch (Gata6 + Tbx20; Atf3 + Tbx18; Figure 5J [↗](#)). qPCR experiments confirm that, while Bnc1 or Gata6 alone do not induce the epicardial genes Gpm6a and Upk1b, the combination of both TFs induces non-linear induction of these target genes (Figure 5K [↗](#)).

Predicting TFs driving single-cell transcriptional phenotypes

The analysis in Figure 4 [↗](#) represents a top-down approach using cells that have reprogrammed to a transcriptional state resembling *in vivo* cells to gain insights into reprogramming TFs. However, this approach requires reprogrammed transcriptional states to be well-represented by perturbed

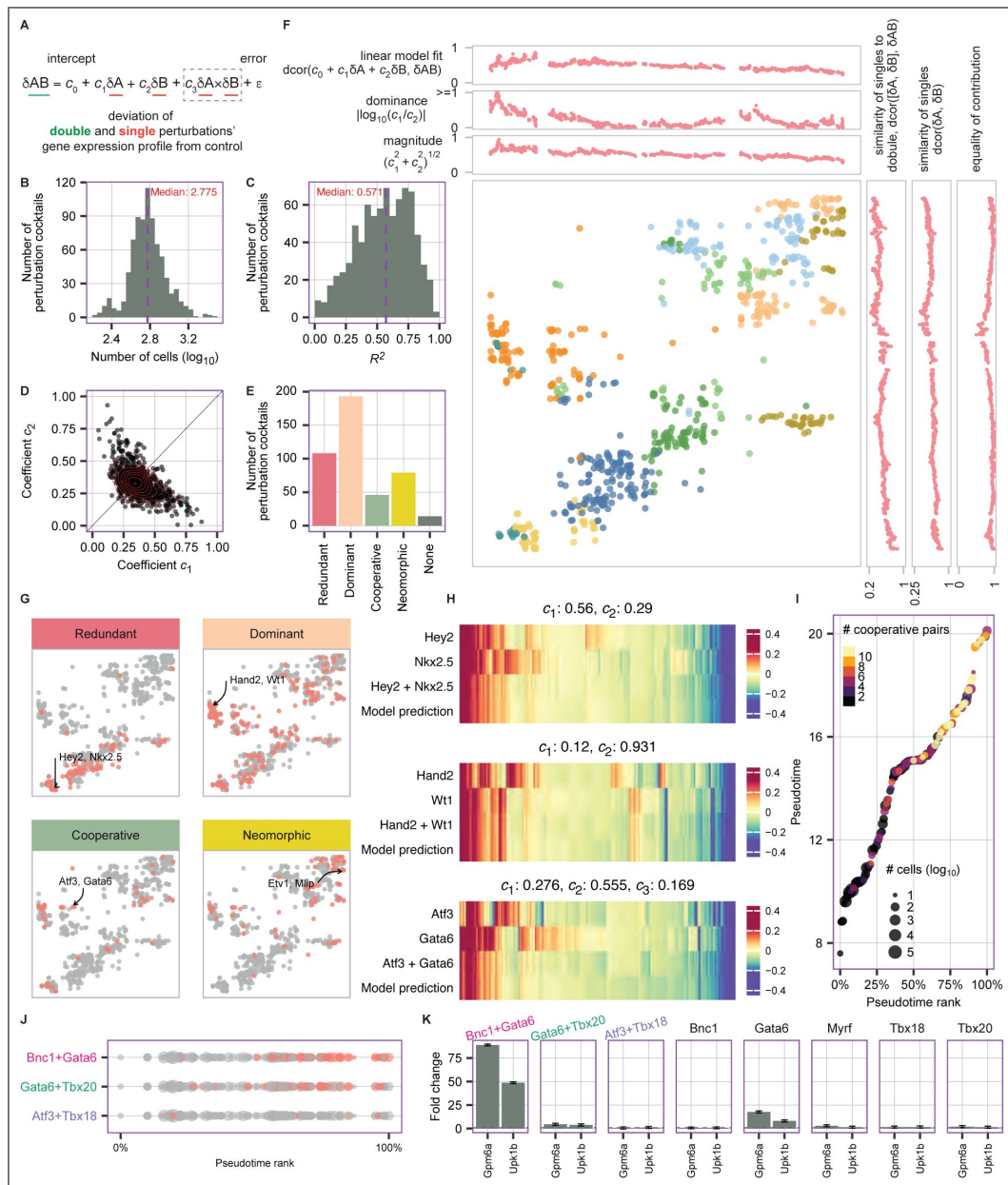


Figure 5. Transcription factor cooperativity in transcriptional reprogramming.

A. A linear model to decompose combinatorial TF perturbations from single TF data. The $c_1 \delta A$ and $c_2 \delta B$ terms model independent contributions of individual TFs A and B, while the $c_3 \times \delta A \times \delta B$ term models cooperativity. B. Distribution of R^2 across linear models across 801 TF pairs. C. Across all 801 TF pairs, shown is the distribution of c_1 and c_2 values in the linear model. D. Distribution of the number of cells used in each application of the linear model. E. Visual embedding of TF pairs based on parameters of the linear model. F. The bar chart indicates the number of TF pairs with the following relationship between constituent TFs: redundant, dominant, cooperative, neomorphic, and none. G. Feature plots indicating TF pairs with redundant, dominant, cooperative, and neomorphic relationships between constituent TFs. Example TF pairs are indicated. H. Example TF pairs with redundant (Hey2 + Nkx2.5), dominant (Hand2 + Wt1), and cooperative (Atf3 + Gata6) relationships. For each of the three plots, shown is the transcriptome expression pattern for: cells induced with each individual TF (top 2 rows), cells induced with both TFs (third row), and the prediction from the linear model (bottom row). The model parameters are indicated above each plot. I. TF interaction relationship in the context of epicardial reprogramming. J. For the epicardial reprogramming batch, cells were ranked by pseudotime and colored by whether they contain the indicated TF pairs. K. qPCR analysis of Gpm6a and Upk1b expression for pairwise and individual TFs.

MEF-derived cells. As an alternative approach, we next developed a bottom-up approach to predict higher-order TF cocktails from simpler TF combinations.

To test the feasibility of this approach, we first focus on the simpler problem of predicting the identity of single TF perturbations from single-cell transcriptomes (Figure 6A). We used two approaches, a linear model with stochastic gradient descent (SGD) learning and the XGBoost algorithm, to train the models using a filtered set of 66 single-TF perturbations well-represented by high quality sequenced cells (≥ 100). To assess the accuracy of the models, we used a computational cross-validation approach, with 80% of the dataset used for training and 20% withheld for testing. Since the results of the two approaches are comparable (Figure S6A), we only refer to the results of the linear model in the main text. Overall, we observe that single-cell transcriptomes have enough information to predict individual TF perturbations (Figure 6B). On average, 24.5% of cells can be correctly classified into the 66 TF categories, compared to an expectation of 1.5% by random guessing (Figure 6D) (odds ratio 16.3). Supporting the robustness of this approach, we observed similar trends on a smaller set of 36 single-TF perturbations with more sequenced cells ≥ 200 cells, with an odds ratio of 14.8 (Figure S6B).

Several examples are illustrative. One class consists of predictions specific to one TF. For example 58.0% of cells induced with the T-box transcription factor Tbx2 are correctly predicted to be perturbed by the TF using each cell's transcriptional profile alone (Figure 6B). A second class contains predictions specific to a small set of related TFs. For example, we tested two isoforms of Erg (denoted as Erg-203 and Erg-205), and the model occasionally predicts the incorrect TF isoform. For example, 25.4% of Erg-203 perturbed cells are predicted as Erg-203, while 3.66% of Erg-205 perturbed cells are predicted as Erg-203. Supporting the idea that these two isoforms likely induce similar transcriptional phenotypes, embedding of single-TF perturbed cells shows that Erg-203 and Erg-205 co-cluster while being distinct from other TFs like Tbx2 (Figure 6C, Figure S6C with all feature plots). Interestingly, Fli1-induced transcriptomes share similarity with those induced by Erg isoforms. Similarly, the misclassification of cells induced with Wt1 and Gata6 suggest similar transcriptional responses for these genes (Figure 6B, Figure S6C). In addition, we observe that while most cells with Nkx2-5 are correctly classified, a subset is also misclassified to Hey2. These observations could suggest similar transcriptional responses between these TFs. Consistent with this possibility, the TF interaction analysis indicates that Hey2 and Nkx2.5 have a redundant relationship (Figure 5G-H). This redundancy of transcriptional responses between distinct TFs complicates the computational task of correctly classifying cells by TF perturbations.

Next, we explored several parameters of the prediction model. At one extreme, we found that retraining the model on the top 5 performing TFs alone yielded 59% accuracy, compared with the bottom 5 performing TFs at 32% accuracy. We asked why some TFs were more difficult to predict than others. One possibility was that poorly predicted TFs have smaller gene expression changes. However, we observed no correlation between the number of differentially expressed genes induced by a TF and the ability to predict the identity of perturbed cells (Figure S6D). We also considered the possibility that the number of sequenced cells correlated with prediction performance. But we observed instances where highly sequenced TFs had poor performance and lowly sequenced TFs had high performance (Figure S6E, Supplemental Table 7). Motivated by the redundancy of TF effects observed above, we also considered the second-best or third-best predictions for the full set of 66 TFs (Figure 6B,D). For example, considering the top 3 predictions, performance increased to 41.3% accuracy, compared to 7.18% expected by chance. We next asked if removing redundant TFs improves prediction performance. To systematically test the impact of redundancy, we dropped each TF one at a time, retrained the model, and re-assessed performance. Globally, we only noticed a marginal improvement (24.7%) in prediction performance (Supplemental Table 7). However, some predictions significantly improved. For example, removing the transcription factor Fosb-201 yielded a 2.56-fold increase in prediction performance for Fosb-206. Overall, these results suggest that multiple factors will need to be considered to improve prediction performance in the future.

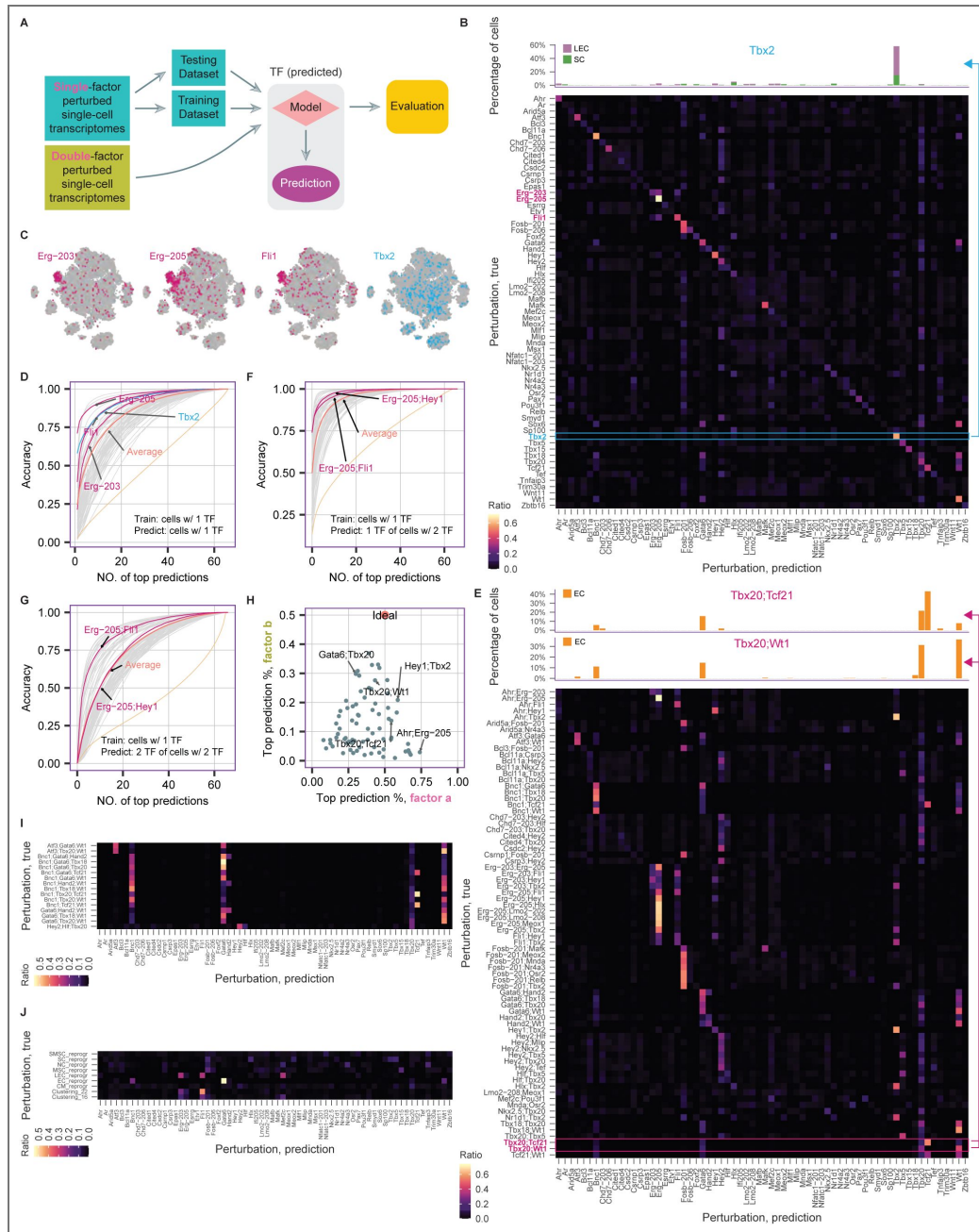


Figure 6.

A. Schematic of approach to model transcriptional responses from single-TF experiments, and to use this model to predict TF perturbations from multi-TF experiments. B. (bottom) Heatmap illustrating the performance of a linear model to predict TF perturbations from transcriptional responses in cells with exactly 1 TF. Columns indicate predictions and rows indicate observations. Heatmap units are the fraction of cells in each row (which sums to 1). (top) TF predictions for cells with Tbx2. C. UMAP embedding of cells with exactly 1 TF. Feature plots indicate TF perturbations. D. Receiver operating characteristic (ROC)-like curves illustrating the performance of predicting single TF perturbations from single-cell transcriptomes, after training on cells perturbed with 1 TF. Dotted line indicates performance where TF labels are randomly shuffled. E. (bottom) Heatmap illustrating the performance of a linear model to predict TF perturbations from transcriptional responses in cells with exactly 2 TFs. (top) F. As in D, but illustrating the performance of predicting one of TFs of the double TF perturbations from single-cell transcriptomes, after training on cells perturbed with 1 TF. G. As in D, but illustrating the performance of predicting two of the TFs of double TF perturbations from single-cell transcriptomes, after training on cells perturbed with 1 TF. H. When predicting the perturbations of cells with 2 TFs (a and b), shown is the frequency that the top prediction is either a or b. I. As in (E), but predicting perturbations in cells that have exactly 3 TFs. J. As in (E), but predicting perturbations in cells that cluster near primary mouse cell types.

The analysis above predicted single TF perturbations from single-cell transcriptomes. We next asked whether this computational model could be applied to predict perturbations in cells receiving multiple TFs. We focused on 75 high quality TF pairs represented by > 50 cells (Figure 6E–F). Several examples are illustrative. First, some TFs (Bnc1, Erg-205, Fox-201, Hey2) are frequently found among the 75 TF pairs, and these are represented as vertical stripes of enrichment in the heatmap. Second, TFs with accurate prediction performance in the singlet TF analysis generally performed well in the doublet analysis. For example, Tbx20 and Tcf21 scored highly compared to Tbx18 in the singlet analysis. Likewise, prediction performance for doublets Tbx20-Wt1 and Tcf21-Wt1 was higher than Tbx18-Wt1. Third, some well-predicted singlet TFs are poorly predicted as doublets. For example, the singlet model poorly predicts Ahr when it is part of a TF doublet. Fourth, when other TFs like Fosb-201 are paired with another TF, only predictions for Fosb-201 are made. These last two examples may highlight dominance relationships between TF pairs (Figure 6H). Next, we asked whether both perturbations in a TF pair were represented in the top predictions (Figure 6G). On average, the model recovers 50.7% of TF pairs in the top 10 predictions, compared to 17.1% expected by chance.

Finally, we extended this analysis to cells with higher-order combinations of TFs. Although we only had sufficient cell numbers for 16 distinct TF triplets, the singlet model performed well to predict the TFs in these cells (Figure 6I). On average, we observed that top 3 TF predictions from the model matched the observed TFs 35.4% of the time. To further extend this analysis to predict TFs driving cellular reprogramming, we examined MEF-derived cells that co-clustered with *in vivo* cells (Figure 1). Consistent with our results above, the model identified Bnc1, Gata6, and Wt1 as TFs for reprogrammed epicardial cells (EC). It also highly scores Erg, Flt1, and Mef2C in endothelial cell reprogramming, consistent with the roles of these genes in maintaining endothelial cell fate and in angiogenesis^{60–62}. Finally, our reprogramming experiments in Figure 1 also recovered cells that unexpectedly co-clustered with immune cells from Tabula Muris (clusters 16 and 22). Our computational model indicates that Erg, Flt1, and Mef2C are among the highest scoring TFs for these immune clusters. This result is consistent with the known roles of these TFs in maintaining homeostasis in hematopoietic lineages and blood cell reprogramming^{63–65}.

Discussion

One key finding of our study is that, despite a vast combinatorial space of TF combinations, the space of transcriptional outputs is much more limited. After overexpressing combinations of 105 TFs, clustering of single-cell transcriptomes shows that many cells induced with TFs cluster with *in vivo* cell types. Notably, the generation of *de novo* cell states was rare, at least on a transcriptome-wide level. At the finer resolution of gene modules, we identified ~50 perturbation clusters. Surprisingly, while each cluster contains distinct TF combinations, they independently activate common gene modules. For example, Perturbation Cluster 1 contains diverse TF combinations that specifically induce a cardiac regulatory program (Figure 3). This striking modularity suggests a shared regulatory architecture of TFs on gene expression states. Together, these results support a many-to-few relationship between TFs and cell states. This simplifying organizational principle of TF-driven gene modules and gene regulatory networks has implications for cell state engineering and the topology of developmental regulatory programs.

TFs do not function in isolation. Traditionally, testing the cooperativity of a given pair of TFs required low throughput individual experiments⁶⁶, and recent technologies including single-cell screens have increased scale to moderate throughput^{25,37,67}. Motivated by a recent study using CRISPRi Perturb-Seq to knock down gene expression³⁷, our approach is to systematically define TF cooperativity through high throughput TF over-expression. Our observations highlight that, on a transcriptome-wide level, most TF pairs interact independently or redundantly. While comparatively fewer TF pairs have cooperative or neomorphic interactions, these interactions are over-represented in TFs that can successfully reprogram transcriptional state. These results suggest that complete sampling of all TF interactions may not be unnecessary and that focused sampling strategies could be used to map the interactions in an economical way.

GRNs comprise all regulatory inputs and transcriptional outputs for sets of functionally related genes⁹. During embryo development, GRNs possess a strictly hierarchical structure. Upper level GRNs (“kernels”) establish the overall body plan through highly recursive and interconnected networks that do not tolerate change⁸. Lower level GRNs, such as specification and differentiation GRNs, are much more adaptable and less recursively wired. During embryogenesis, sequential activation of specification and differentiation GRNs is coupled with canalization, the process by which alternative cell fates become restricted. Thus, direct reprogramming requires simultaneous de-canalization and alternative cell fate specification. In our framework, gene modules can be viewed as GRN sub-circuits that are re-deployed in multiple cellular contexts. Identification of which TFs activate a particular gene module may thus suggest how GRN topology is wired (or rewired). For example, the “muscle contraction” module is activated by many different TFs across several Reprogram-Seq batches. In contrast, the “blood vessel” module is dominantly activated by ETS family TFs. This striking difference exemplifies the unique GRN architectures that can exist. For the “muscle contraction” module, a horizontal or distributed topology likely reflects numerous cis-regulatory inputs capable of activation by a variety of TF combinations. For the “blood vessel” module, however, the vertical topology likely represents the dominant action of a single TF to specify cell fate, which may in turn reflect the distributed nature of endothelial cells within vertebrate organisms. In this way, analysis of Reprogram-Seq datasets enables reconstruction of GRNs using a top-down approach, which is distinct from traditional bottom-up dissection of regulatory element evolution.

The development of generalizable tools for the rational manipulation of cell type-specific transcriptional states will improve our understanding of the molecular basis of transcriptional states. Here, we focused on overexpressing the open reading frames of TFs, the most common approach for successful cellular reprogramming. The advantage of TF overexpression is the robustness of induction, which may be required to overcome the cell’s existing transcriptional state. Disadvantages include being exogenous, cellular integration by lentivirus, and non-physiological levels of induction. Similar approaches using CRISPRa have also recently been used for cellular reprogramming screens, which induce TFs endogenously. However, one disadvantage of this approach is that the level of TF induction is lower, and could be influenced by locus-specific epigenetic effects and/or chromatin accessibility. Alternatively, small molecule approaches have also been successfully applied to iPSC reprogramming, though they are currently less amenable to rational design and more difficult to customize gene targets^{68,69}. Another focus of our study is on the direct reprogramming of differentiated fibroblasts to other differentiated cell states. In contrast, parallel studies have focused on the “programming” of pluripotent stem cells to differentiated cell states, with similarities to differentiation during development. Both approaches have advantages and disadvantages. Advantages of direct reprogramming is the long history of studies using MEFs to derive diverse cell types and the ease of genetically manipulating MEFs. Advantages of programming human pluripotent stem cells is a more plastic epigenetic state and the ability to directly translate results for regenerative medicine. In sum, all of these resources will provide critical information needed to engineer transcriptomes for *in vitro* cell state modeling^{70–72}, with the aspirational goal of providing engineered cellular substrates for research and medicine⁷³.

Methods

Experimental methods

Cell lines

Commercial Mouse Embryonic Fibroblasts (cMEFs) were purchased frozen from Cell Biologicals, defrosted as needed, and grown in growth media (DMEM high-glucose + 10% FBS + 1% penicillin/streptomycin + 1% 100X GlutaMAX).

Construction of lentiviral vectors

TFs were codon optimized, synthesized, and cloned by Twist Bioscience into a lentiviral vector driven by EF1a and modified to contain 10X Genomics Capture Sequence 2 and a PCR handle (Supplemental Table S8). Each insert is composed of (5' → 3'): Kozak sequence (GCCACC), TF open reading frame (ORF), a common PCR handle, and an ORF-specific barcode. The eGFP-bc2 v1.3 plasmid map annotates elements in the vector (Supplemental Table S8).

Viral packaging, infection, and single cell reprogramming

A packaging cell line HEK 293Ts were used to produce lentivirus for reprogramming. Packaging cells were transfected using LT-1 per manufacturer's instruction with lentiviral packaging vectors psPAX2 and pMD2.G in ratio of 2:1:3 with lentiviral vectors expressing specific transcription factors. Approximately 14-16 hours after transfection, the viral supernatant was aspirated and replaced with new media containing caffeine at a concentration of 2mM. 48 hours after transfection, virus was harvested, filtered (0.45µm pore) and concentrated via ultracentrifugation (insert protocol reference here). The viral pellet was then resuspended in growth media (DMEM high-glucose + 10% FBS + 1% penicillin/streptomycin + 1% 100X GlutaMAX) plus 8µg/mL of polybrene (Sigma-Aldrich, St. Louis, MO) and applied to MEFs. The cells were incubated in growth media for 14 days and subsequently washed with DPBS and trypsinized into a single cell suspension. Cells were spun at 500 xg for 5 minutes at 4°C and resuspended in cold growth media.

Viral titering

To determine the relative titer of each TF in cell type pools, bulk RNA-seq libraries amplifying only our exogenous perturbations were constructed. After extracting RNA using the Zymo Quick RNA prep kit, Reverse Transcription was performed using an oligo containing the CS2 sequence from the 10X platform. After a PCR purification using the Qiagen Minielute kit, the cDNA is amplified using the 10X feature primer, then size selected using a 0.6X SPRIselect single sided selection. Another cDNA amplification on the PCR handle was performed, purified, and then amplified one last time to affix Nextera indices. A final 1.2X SPRIselect single sided selection was performed, then the libraries were tested for Quality Control metrics and sent off for next generation sequencing.

Flow cytometry

For flow cytometry, cells were trypsinized and resuspended in growth media. Cells were then spun at 500 xg for 5 minutes at 4°C and washed with room temperature DPBS. Cells were spun again at 500 xg for 5 minutes at 4°C and resuspended into a single-cell suspension in DPBS with 0.04% BSA. Flow cytometry was carried out on a Beckman Coulter CytoFLEX flow cytometer.

Cell hashing

We adopted the Cell Hashing protocol as described in:

https://citeseq.files.wordpress.com/2019/02/cell_hashing_protocol_190213.pdf⁷⁴. Briefly, for the single-cell RNA-seq library of multi-time point FP-control MEFs, MEFs that were transduced with a pool of nine fluorescent proteins for different lengths of time (0 days - uninfected, 5 days, 15 days, and 21 days, respectively) were prepared separately into cell suspension. Additionally, MEFs 5 days after transduced with another three-fluorescent protein pool were also prepared into cell suspension. Cell hashing was performed on these five populations of cells as described previously⁷⁴. Briefly, cells were blocked with TruStain FcX™ (anti-mouse CD16/32) Antibody (Biolegend), and were then stained with TotalSeq™-A anti-mouse with Hashtag (Biolegend). Each of the five cell populations was stained with a different Hashtag. Cells were washed twice with PBS + 2% BSA + 0.01% Tween, before checked viability with TrypanBlue and counted. Based on the counts, the five populations were pooled in equal numbers, spun down, and resuspended in PBS + 0.04% BSA to reach the concentration of 1600 cells/µL. Finally, the pooled population was loaded for 10X Genomics scRNA-seq at a 2.3x overload (i.e., adding 10X recommended volumes of cells and water for a cell suspension of 700 cells/µL aiming for recovering 10 thousand cells, while in reality having a cell suspension of 1600 cells/µL).

To prepare gene expression as well as hashtag libraries for the control MEFs, we followed standard 10X Genomics protocol, with modifications to amplify the hashtags as described previously⁷⁴. Briefly, HTO PCR additive was spiked into the cDNA amplification PCR of the 10X protocol. During post-cDNA amplification PCR cleanup with SPRI beads (0.6x), whereas the supernatant after beads incubation was normally discarded, it was kept instead this time with another round of 2x SPRI cleanup to purify the hashtag PCR product. The purified product was amplified again with 2x KAPA Hifi PCR Master Mix, TruSeq DNA D7xx_s primer (containing i7 index) and SI-PCR primer (Supplementary Table S8). Finally, the PCR product was purified using a 1.6x SPRI cleanup and sequenced on a Nextseq sequencer.

Single cell RNA sequencing library preparation

14 days post-infection, infected MEFs were trypsinized and prepared for single cell RNA-seq following the standard 10X cell preparation protocol. The concentration of single cell suspension was adjusted to 500-1000 cells/ μ L and was loaded on the 10x Genomics Chromium system (10x Genomics, Pleasanton, CA) with the aim of generating 6000 to 10000 transcriptomes per channel (Chromium Single Cell 3' Library & Gel Bead Kit v2, catalog number 120237). Single cell RNA sequencing libraries were constructed following the 10X Chromium Next GEM Single Cell 3' Protocol with some modifications, which enables customized TF perturbation enrichment⁷⁵. (Supplementary Table S8)

To prepare TF perturbation libraries: Aliquot 50-100 ng of purified cDNA amplification product (usually 2-4 μ L) from the 10X Genomics 3' single-cell RNA-seq protocol. The cDNA is amplified for 13 cycles with primers (SI-PCR Primer from 10X Genomics 3' single-cell RNA-seq protocol and Modified Nextera N7 Adapter) and 2x Kapa Hifi Mastermix (Roche #7958935001) (Supplementary Table S8). The PCR product is about 1 kb-long. It is purified with a 0.6x SPRIselect beads clean-up, following standard protocol, and sequenced on a Nextseq sequencer.

Computational methods

Reprogramming TF prediction

The raw fastq files from the Tabula Muris project were downloaded and reprocessed the same way as in-house datasets to minimize technical differences⁴³. Reprogramming TFs were predicted for every cell type as described in¹⁶ with some modifications. In brief, entropy of each TF was calculated and used to access TF cell type specificity. In house primary datasets were used for cardiac cell type predictions^{16,76,77}. With the aim of sampling 10% of total TFs (~200), we sought to maximize the number of cell types targeted. In the end, cardiomyocyte, lung endothelial cell, mesenchymal stem cell, neuroendocrine cell, skeletal muscle satellite cell, and stromal cell were chosen.

RSEM (version 1.3.3) was used to select the most abundant isoforms of the predicted reprogramming TFs in target primary cells.

Sequencing, basecalling and demultiplexing

Libraries were sequenced using the Illumina NextSeq 500/550 and NovaSeq 6000 (Illumina). The generated binary base call (BCL) files were demultiplexed and converted to standard FASTQ files using the mkfastq function from the Cell Ranger pipeline (version 5.0.1).

Read alignment and generation of gene expression matrix

Cell Ranger pipeline (version 5.0.1) was used to preprocess the 10X Genomics single-cell data with default parameters except the number of cells expected was set to 10,000 to generate the expression matrix. Scrublet was used to detect multiplets (version 0.2.3)⁷⁸.

Cell hashing demultiplexing

Perturbation detection

The detection and quantification of perturbations in single cells was performed using FBA ⁷⁹ (version 0.0.12). The mismatch thresholds for cell and feature barcodes were set to 1 and 2, respectively. Demultiplexing was performed using the gaussian mixture model. Demultiplexing of barcoded antibody-labeled MEFs (cell hashing) results was performed using FBA.

Dimensionality reduction

Prior to normalization, non-informative cells and genes were filtered. Cells with less than 200 genes detected were excluded in the downstream analysis. For non-cardiac reprogramming assays, cells with more than 20% of UMIs from mitochondria were filtered and for cardiac reprogramming, the cutoff is 50%. Genes detected in fewer than 30 cells or with fewer than 60 UMIs in total across all single cells were filtered. Total UMI counts of single cells were normalized to the median UMI counts per cell and natural-logarithmized with a pseudocount of 1. The median-normalized log-transformed matrix was standardized (mean 0, standard deviation 1).

Principal component analysis (PCA) was performed using the Scanpy package (version 1.7.2) ⁸⁰ with the median-normalized log-transformed and standardized matrix as input. Batch effect removal was performed using the harmony algorithm with default parameters (<https://github.com/Slowkow/harmonypy> ⁸¹, version 0.0.5) ⁸¹. Each 10X Genomics RNA-seq library was treated as a different batch.

Uniform Manifold Approximation and Projection (UMAP) visualization was performed using UMAP with harmony-corrected PCA result as input (version 0.5.3, metric="euclidean", min_dist=0.1, spread=1.0) ⁸².

Minimum-distortion embedding (MDE) of perturbation cocktails was performed using PyMDE (version 0.1.18) ⁸³. Expression matrix of perturbation cocktails was median-normalized and natural-logarithmized with a pseudocount of 1. Variable expression detection and regression out effects of total UMIs per cell was performed using Scanpy with default parameters. PCA was performed and used as input for pyMDE (n_neighbors=2, repulsive_fraction=5). The initial iterate was a 2-dimensional matrix generated by PHATE using PCA result (version 1.0.10) (knn=2, decay=40, n_landmark=2000, t="auto", gamma=1, mds_solver="smacof", knn_dist="euclidean", mds_dist="euclidean").

Clustering

Single cells were clustered using the leiden algorithm (leidenalg, version 0.9.1) implemented in the Scanpy package (version 1.7.2) ⁸⁰. In brief, the harmony-corrected matrix was used to compute a neighborhood graph of observations (n_neighbours=30, method = "umap", metric="euclidean"). Then, the leiden algorithm was used to cluster single cells (resolution=None, flavor="igraph").

Hierarchical DBSCAN clustering of perturbation cocktails was performed using the hdbscan method implemented in dbscan with minPts setting to 10 (version 1.1.1) ⁸⁴.

Differential expression analysis

Binomial test

The binomial test based method used for testing if a gene is expressed more frequently between different groups of cells was described in Duan et al. 2019 ¹⁶ with small modifications.

Pseudo-bulk

A pseudo-bulk approach was used to detect differential expression between two groups of single cells. Single cells belonging to the same groups were randomly combined (minimal cells: 30, maximum cells: 500). edgeR lrt method was used to detect differentially expressed genes between the pseudo-bulk replicates (version 3.38.4) ⁸⁵.

Functional annotation and enrichment analysis

GO term enrichment was performed using Fisher's exact test implemented in topGO (version 2.48.0; org.Mm.eg.db version 3.15.0). Reactome pathway enrichment was performed using ReactomePA (version 1.40.0; org.Mm.eg.db version 3.15.0; reactome.db version 1.81.0) ⁸⁶.

Pseudotime construction

Monocle (version 2.24.1) was used for pseudotime analysis with differentially expressed genes between target primary cell types and control cells used for pseudotime ordering ⁸⁷.

TF Interaction analysis

Genes detected in less than 30 cells or with less than 60 UMIs in total across all single cells were excluded for downstream analysis. UMI count matrix of perturbed cells were normalized to fluorescent protein infected control cells. For pairwise TF interactions, we required two-TF perturbed cells and their singlet constituents to have at least 10 cells. The median number of cells used for TF interaction interference is 595. Theil-Sen regression implemented in scikit-learn (version 1.2.0) was used for linear regression (max_subpopulation=1e5, n_subsamples=None, max_iter=1000, tol=0.001).

$$\delta AB = c_0 + c_1 \delta A + c_2 \delta B + \varepsilon \quad (1)$$

dcor package was used for distance correlation calculation (version 0.6). F-statistics was calculated to test the significance of regression coefficients in models. If the p-value associated with the F-statistics is less than 0.05, the interaction of the TF pair was characterized as follows:

Redundant

Variance Inflation Factors (VIF) was calculated to detect multicollinearity. If VIF is higher than 5, this pair is characterized as redundant.

Dominant

To determine the dominant effect of TF pairs (AB), the results of linear models (1, 2, 3) were taken into consideration.

$$\delta AB = c_0 + c_1 \delta A + \varepsilon \quad (2)$$

$$\delta AB = c_0 + c_1 \delta B + \varepsilon \quad (3)$$

For a given TF pair (AB), if the R^2 (or the coefficient of determination) of (2) is higher than that of (3) and c_1 is higher than c_2 in (1), then this TF Pair is characterized as A dominant; If the R^2 of (3) is higher than that of (2) and c_2 is higher than c_1 in (1), B is characterized as dominant.

Interactive

TF pairs (AB) were modeled using (4), and t-statistics of c_1 , c_2 and c_3 were calculated. If the p-value associated with the t-statistic of c_3 is less than 0.05, this pair of TFs was characterized as interactive.

$$\delta AB = c_0 + c_1 \delta A + c_2 \delta B + c_3 \delta A \times B + \varepsilon \quad (4)$$

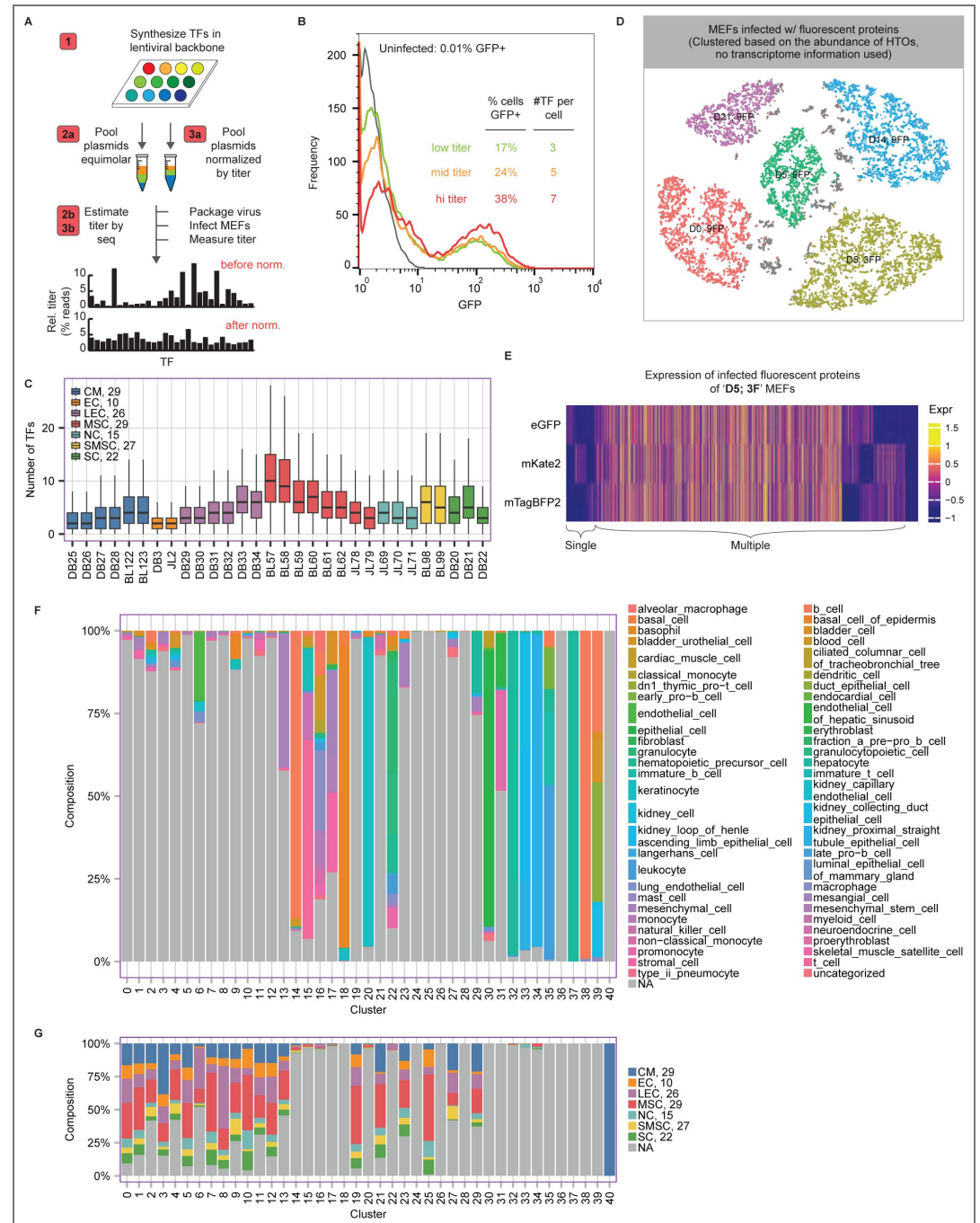
Neomorphic

Mean absolute error (MAE) from (1) was calculated for each TF pair. The largest 10% of TF pairs' MAE were characterized as neomorphic.

Prediction

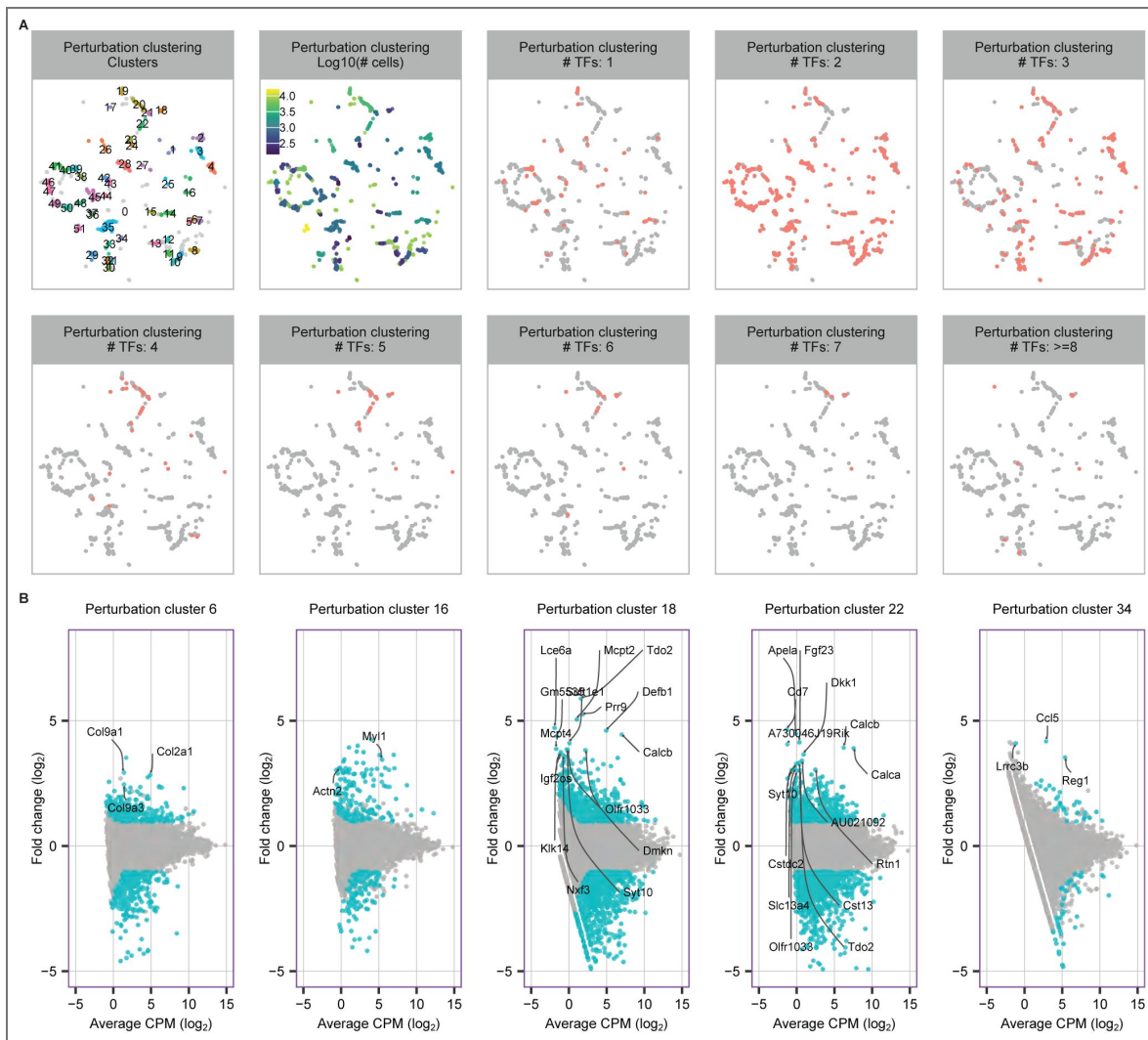
Prior to fitting classification models, low informative genes of single cells were filtered as described above. Then single cells were normalized and log-transformed. For each training (total 100 times), 20% of each group of single TF perturbed cells were randomly left out for validation. Two strategies were used. One is stochastic gradient descent classifier (SGDClassifier implemented in scikit-learn, version 1.3.1, LOSS="log_loss", PENALTY="l2"), and the other is extreme gradient boosting (XGBoost, python implementation version 1.7.6, learning_rate=0.2, objective="multi:softprob", tree_method="hist").

Supplemental figures



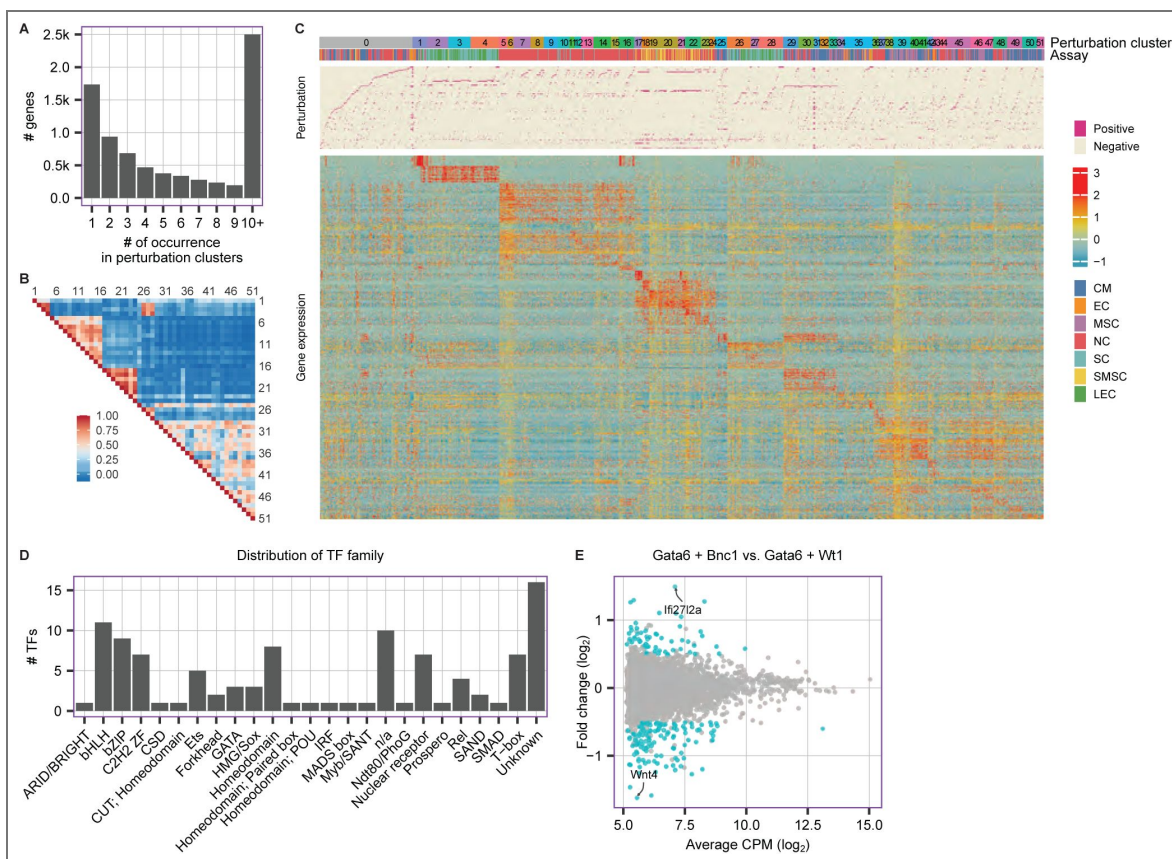
Supplemental Figure 1. A. Schematic of iterative strategy to normalize lentiviral TF titers. As the lentiviral titer of each TF may be different, normalization of relative titers will be important for downstream analysis. To assess the relative titers of TFs in each reprogramming batch, we will create an equimolar pool of plasmids, package pooled lentivirus, infect MEFs for 7 days, and measure the titer for each TF by next-generation sequencing of TF barcodes from mRNA. We then adjust the concentration of each plasmid and iterate this analysis to confirm normalization of relative titers. B. Controlling perturbation complexity by controlling titer. We infected MEFs with a pool of 30 normalized TFs, quantified GFP+ cells by flow cytometry as a measure of absolute titer, and confirmed these measurements by scRNA-Seq. C. The median number of TFs in each batch of perturbation assay. D. Visualization of MEFs demultiplexing. Four batches of MEFs were labeled and pooled together prior to sequencing (cell hashing). The clustering of visualization was performed using the abundance of hashtag oligos. Colors indicate cell groups after demultiplexing. Grey dots indicate multiplets. E. Expression heatmap of infected fluorescent

proteins of 'D5; 3F' MEFs (infected with 3 fluorescent proteins and cultured for 5 days). F. Cell type compositions of each cluster. Only cell type annotations from Tabula Muris were shown (non-grey colored). G. Cell type compositions of each cluster. Only cell type annotations from reprogrammed cells were shown (non-grey colored).



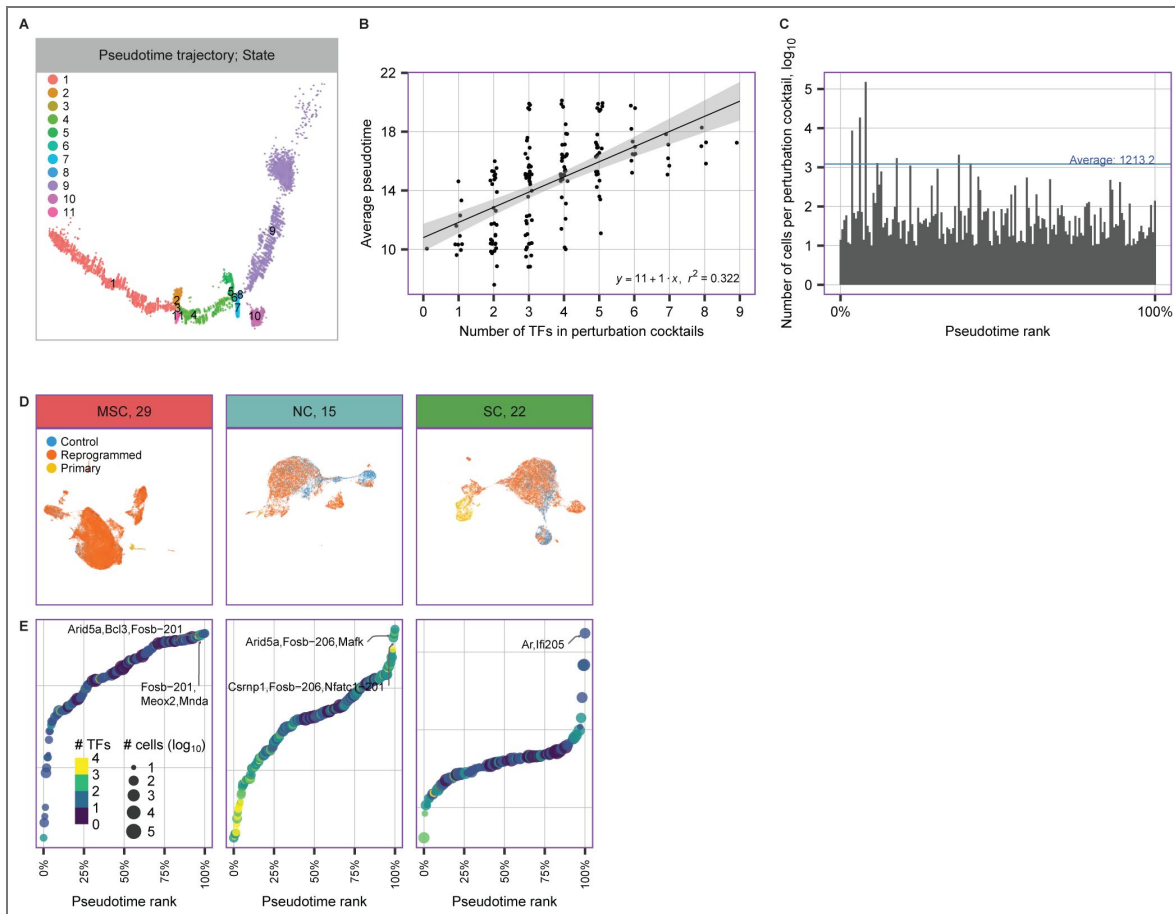
Supplemental Figure 2.

A. MDE visualization of perturbation clusters. From top left to bottom right are perturbation clusters, number of cells for each perturbation cluster, and number of TFs in each perturbation cluster. B. Differentially expressed genes (DEGs) for selected perturbation clusters. Teal indicates DEGs.



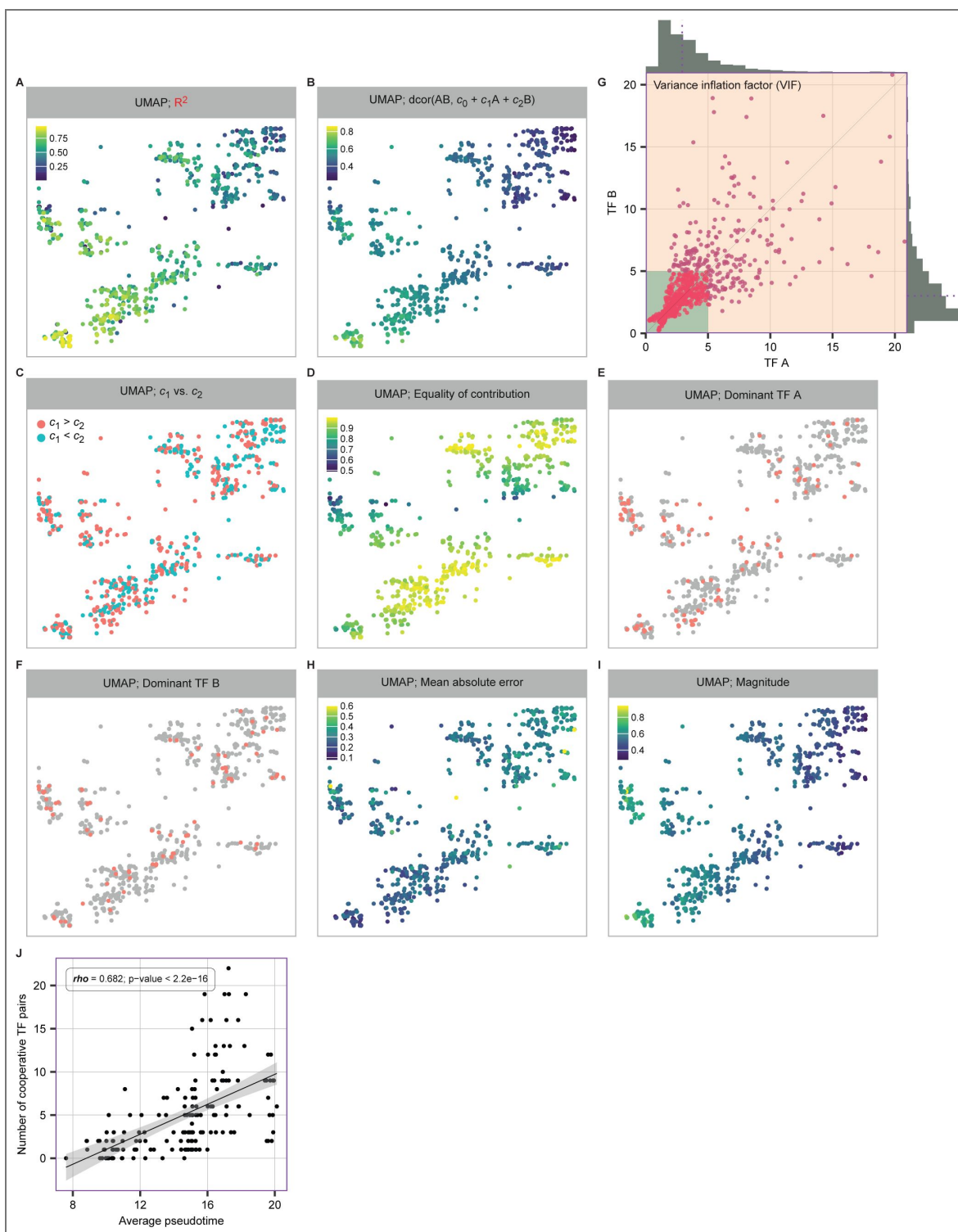
Supplemental Figure 3.

A. Across all differentially expressed genes identified in Fig. 3B, shown is the occurrence across modules). 1 indicates a gene is found only in one module. B. The heatmap indicates the correlation of TFs across perturbation clusters. C. Heatmap indicating gene expression patterns (bottom) across perturbation clusters. Each column indicates a distinct TF combination (middle), grouped by batch/assay and perturbation cluster (top). D. Distribution of the TF family annotations for the 105 TFs. E. The MA plot indicates the differentially expressed genes between TF pairs Gata6 + Bnc1 and Gata6 + Wt1 (teal). Several genes are indicated.



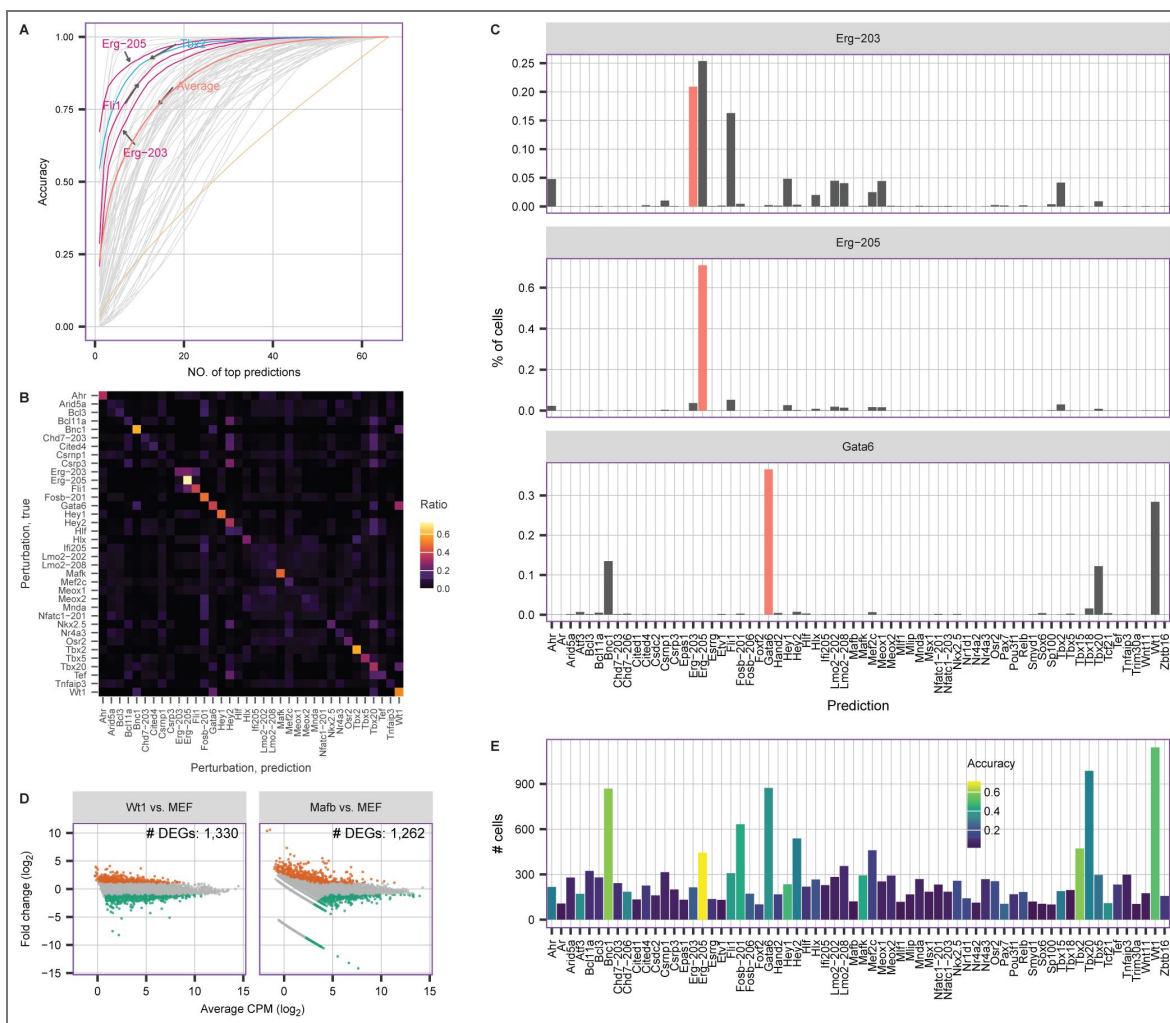
Supplemental Figure 4.

A. Pseudotime trajectory of epicardial reprogramming. Colors indicate different cellular states. B. The distribution of the numbers of TFs in perturbation cocktails and the average pseudotimes. C. The distribution of pseudotime ranking and the numbers of cells of these perturbation cocktails. D. Across all reprogramming batches and targeted primary cells, shown is the clustering of control MEFs (blue), MEF-derived cells (orange), and primary cells (yellow). E. Pseudotime analysis to rank putative TFs for cardiomyocytes, lung endothelial cells and mesenchymal stem cells reprogramming.



Supplemental Figure 5. Visual embedding of TF pairs based on parameters of the linear model, the colors indicating R^2 values (A), dcor correlation (B), relationship of c_1 and c_2 (C), equality of contribution (D), TF A dominant (E), TF B dominant (F), mean absolute error (H), magnitude (I) and variance inflation factor (G).

J. The correlation of average pseudotimes and the number of cooperative TF pairs in the perturbation cocktails.



Supplemental Figure 6.

A. Prediction accuracy for 66 TFs using XGBoost algorithm with different cutoffs (0.2266). B. Heatmap visualization of the distribution of predictions (the sum of row is 1). C. The distributions of predictions for Erg-203, Erg-205, and Gata6. D. MA-plots of Wt1 and Mafb perturbed cells. Colors indicate differentially expressed genes. E. Number of cells for each single TF perturbation. Color indicates the prediction accuracy of each TFs.

Data availability

Data availability: The single cell RNA sequencing data reported in this paper were deposited in NCBI Gene Expression Omnibus (GEO) under the accession number GSE225838. To review, go to <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225838> and enter token unnmjwmaibbwpxpkd into the box.

Acknowledgements

We acknowledge the BioHPC computational infrastructure at UT Southwestern for providing HPC and storage resources that have contributed to the research results reported within this paper. G.C.H is supported by CPRIT (RP190451), NIH (DP2GM128203, UM1HG011996, 1R35GM145235), the Burroughs Wellcome Fund (1019804), the Welch Foundation (I-2103-20220331), and the Green Center for Reproductive Biology. N.V.M. was supported by the NIH (HL136604, HL151650, UM1HG011996, and HG012768), the Burroughs Wellcome Fund (1009838), and the Department of Defense (PR172060).

Additional files

Supplemental tables. [Supplemental Table 1](#): Transcription factor prioritization for each reprogramming assay. **Supplemental Table 2**: Sequencing statistics for each reprogramming assay. **Supplemental Table 3**: Sequencing statistics for TF combinations. **Supplemental Table 4**: Differentially expressed genes in each perturbation cluster. **Supplemental Table 5**: Reactome pathways and Gene Ontology terms used in Fig. 2G. **Supplemental Table 6**: Statistics for linear models. **Supplemental Table 7**: Summary of perturbation prediction accuracy. **Supplemental Table 8**: Primers, TF ORFs, and constructs used in this study

Additional information

Funding

Funder	Grant reference number	Author
HHS NIH National Institute of General Medical Sciences (NIGMS)	DP2GM128203	Gary C Hon
HHS NIH National Institute of General Medical Sciences (NIGMS)	1R35GM145235	Gary C Hon
Welch Foundation (The Welch Foundation)	I-2103-20250403	Gary C Hon
HHS NIH National Heart, Lung, and Blood Institute (NHLBI)	HL136604	Nikhil V Munshi
HHS NIH National Heart, Lung, and Blood Institute (NHLBI)	HL151650	Nikhil V Munshi
HHS NIH National Human Genome Research Institute (NHGRI)	HG012768	Nikhil V Munshi Gary C Hon
Burroughs Wellcome Fund (BWF)	1009838	Nikhil V Munshi
U.S. Department of Defense (DOD)	PR172060	Nikhil V Munshi

Author ORCID iDs

Gary C Hon:  <https://orcid.org/0000-0002-1615-0391>

References

1. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA (2017) The Human Cell Atlas: from vision to reality. *Nature* **550**:451-453 <https://doi.org/10.1038/550451a> | [PubMed](#)

2. **Tabula Muris Consortium** (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**:367-372 <https://doi.org/10.1038/s41586-018-0590-4> | [PubMed](#)
3. **Snyder MP**, Lin S, Posgai A, Atkinson M, Regev A, Rood J, Rosen O, Gaffney L, Hupalowska A, Satija R, *et al.* (2019) Mapping the Human Body at Cellular Resolution--The NIH Common Fund Human BioMolecular Atlas Program. *arXiv* <https://doi.org/10.48550/arxiv.1903.07231>
4. **Small S**, Blair A, Levine M (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* **11**:4047-4057 <https://doi.org/10.1002/j.1460-2075.1992.tb05498.x> | [PubMed](#)
5. **Davidson EH**, Rast JP, Oliveri P, Ransick A, Caestani C, Yuh C.-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, *et al.* (2002) A genomic regulatory network for development. *Science* **295**:1669-1678 <https://doi.org/10.1126/science.1069883> | [PubMed](#)
6. **Neph S**, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**:1274-1286 <https://doi.org/10.1016/j.cell.2012.04.040> | [PubMed](#)
7. **Boyer LA**, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**:947-956 <https://doi.org/10.1016/j.cell.2005.08.020> | [PubMed](#)
8. **Davidson EH**, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* **311**:796-800 <https://doi.org/10.1126/science.1113832> | [PubMed](#)
9. **Erwin DH**, Davidson EH (2009) The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genet* **10**:141-148 <https://doi.org/10.1038/nrg2499> | [PubMed](#)
10. **Mikkelsen TS**, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**:49-55 <https://doi.org/10.1038/nature07056> | [PubMed](#)
11. **Hashimoto H**, Wang Z, Garry GA, Malladi VS, Botten GA, Ye W, Zhou H, Osterwalder M, Dickel DE, Visel A, *et al.* (2019) Cardiac Reprogramming Factors Synergistically Activate Genome-wide Cardiogenic Stage-Specific Enhancers. *Cell Stem Cell* **25**:69-86. <https://doi.org/10.1016/j.stem.2019.03.022> | [PubMed](#)
12. **Takahashi K**, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**:663-676 <https://doi.org/10.1016/j.cell.2006.07.024> | [PubMed](#)
13. **Takahashi K**, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**:861-872 <https://doi.org/10.1016/j.cell.2007.11.019> | [PubMed](#)
14. **Ieda M**, Fu J.-D, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, Srivastava D (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**:375-386 <https://doi.org/10.1016/j.cell.2010.07.002> | [PubMed](#)
15. **Parekh U**, Wu Y, Zhao D, Worlikar A, Shah N, Zhang K, Mali P (2018) Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single-Cell RNA-Sequencing Readout. *Cell systems* <https://doi.org/10.1016/j.cels.2018.10.008> | [PubMed](#)
16. **Duan J**, Li B, Bhakta M, Xie S, Zhou P, Munshi NV, Hon GC (2019) Rational Reprogramming of Cellular States by Combinatorial Perturbation. *Cell Rep* **27**:3486-3499. <https://doi.org/10.1016/j.celrep.2019.05.079> | [PubMed](#)
17. **Joung J**, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, Guo B, Arias-Garcia MA, Allen WE, Singh A, *et al.* (2023) A transcription factor atlas of directed differentiation. *Cell* **186**:209-229. <https://doi.org/10.1016/j.cell.2022.11.026> | [PubMed](#)
18. **Luginbühl J**, Kouno T, Nakano R, Chater TE, Sivaraman DM, Kishima M, Roudnicki F, Carninci P, Plessy C, Shin JW (2021) Decoding Neuronal Diversification by Multiplexed Single-cell RNA-Seq. *Stem Cell Reports* **16**:810-824 <https://doi.org/10.1016/j.stemcr.2021.02.006> | [PubMed](#)

19. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, Wernig M (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**:1035-1041 <https://doi.org/10.1038/nature08797> | PubMed
20. Sekiya S, Suzuki A (2011) Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**:390-393 <https://doi.org/10.1038/nature10263> | PubMed
21. Huang P, He Z, Ji S, Sun H, Xiang D, Liu C, Hu Y, Wang X, Hui L (2011) Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**:386-389 <https://doi.org/10.1038/nature10116> | PubMed
22. Xu J, Du Y, Deng H (2015) Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell* **16**:119-134 <https://doi.org/10.1016/j.stem.2015.01.013> | PubMed
23. Grossman SR, Zhang X, Wang L, Engreitz J, Melnikov A, Rogov P, Tewhey R, Isakova A, Deplancke B, Bernstein BE, et al. (2017) Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A* **114**:E1291-E1300 <https://doi.org/10.1073/pnas.1621150114> | PubMed
24. Porter W, Saville B, Hoivik D, Safe S (1997) Functional synergy between the transcription factor Sp1 and the estrogen receptor. *Mol. Endocrinol* **11**:1569-1580 <https://doi.org/10.1210/mend.11.11.9916> | PubMed
25. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**:384-388 <https://doi.org/10.1038/nature15518> | PubMed
26. Tsunemoto R, Lee S, Szűcs A, Chubukov P, Sokolova I, Blanchard JW, Eade KT, Bruggemann J, Wu C, Torkamani A, et al. (2018) Diverse reprogramming codes for neuronal identity. *Nature* <https://doi.org/10.1038/s41586-018-0103-5> | PubMed
27. Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z, Bergmann S (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods* **13**:366-370 <https://doi.org/10.1038/nmeth.3799> | PubMed
28. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**:2314-2321 <https://doi.org/10.1093/bioinformatics/btx194> | PubMed
29. Cahan P, Li H, Morris SA, Lummertz da Rocha E, Daley GQ, Collins JJ (2014) CellNet: network biology applied to stem cell engineering. *Cell* **158**:903-915 <https://doi.org/10.1016/j.cell.2014.07.020> | PubMed
30. Morris SA, Cahan P, Li H, Zhao AM, San Roman AK, Shivdasani RA, Collins JJ, Daley GQ (2014) Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**:889-902 <https://doi.org/10.1016/j.cell.2014.07.021> | PubMed
31. Pratapa A, Jaliha AP, Law JN, Bharadwaj A, Murali TM (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**:147-154 <https://doi.org/10.1038/s41592-019-0690-6> | PubMed
32. Bonnaïffoux A, Herbach U, Richard A, Guillemin A, Gonin-Giraud S, Gros P.-A, Gandrillon O (2019) WASABI: a dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics* **20**:220 <https://doi.org/10.1186/s12859-019-2798-1> | PubMed
33. Delgado FM, Gómez-Vela F (2019) Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artif. Intell. Med* **95**:133-145 <https://doi.org/10.1016/j.artmed.2018.10.006> | PubMed
34. Rackham OJL, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, FANTOM Consortium, Suzuki H, Nefzger CM, Daub CO, et al. (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet* **48**:331-335 <https://doi.org/10.1038/ng.3487> | PubMed

35. **The FANTOM Consortium**, Riken Omics Science Center (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics* **41**:553-562 <https://doi.org/10.1038/ng.375> | PubMed
36. **Ng AHM**, Khoshakhlagh P, Rojo Arias JE, Pasquini G, Wang K, Swiersy A, Shipman SL, Appleton E, Kiaee K, Kohman RE, *et al.* (2021) A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol* **39**:510-519 <https://doi.org/10.1038/s41587-020-0742-6> | PubMed
37. **Norman TM**, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, Gilbert LA, Weissman JS (2019) Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365**:786-793 <https://doi.org/10.1126/science.aax4438> | PubMed
38. **Xie S**, Cooley A, Armendariz D, Zhou P, Hon GC (2018) Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS One* **13**:e0198635 <https://doi.org/10.1371/journal.pone.0198635> | PubMed
39. **Hill AJ**, McFaline-Figueroa JL, Starita LM, Gasperini MJ, Matreyek KA, Packer J, Jackson D, Shendure J, Trapnell C (2018) On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* <https://doi.org/10.1038/nmeth.4604> | PubMed
40. **Ellis BL**, Potts PR, Porteus MH (2011) Creating higher titer lentivirus with caffeine. *Hum. Gene Ther* **22**:93-100 <https://doi.org/10.1089/hum.2010.068> | PubMed
41. **Segura MM**, Kamen AA, Garnier A (2011) Overview of current scalable methods for purification of viral vectors. *Methods Mol. Biol* **737**:89-116 https://doi.org/10.1007/978-1-61779-095-9_4 | PubMed
42. **Zufferey R**, Donello JE, Trono D, Hope TJ (1999) Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. *J. Virol* **73**:2886-2892 <https://doi.org/10.1128/jvi.73.4.2886-2892.1999> | PubMed
43. **The Tabula Muris Consortium**, Quake SR, Wyss-Coray T, Darmanis S (2018) Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. *bioRxiv* <https://doi.org/10.1101/237446>
44. **Wang H**, Keepers B, Qian Y, Xie Y, Colon M, Liu J, Qian L (2022) Cross-lineage potential of Ascl1 uncovered by comparing diverse reprogramming regulatomes. *Cell Stem Cell* **29**:1491-1504. <https://doi.org/10.1016/j.stem.2022.09.006> | PubMed
45. **Treutlein B**, Lee QY, Camp JG, Mall M, Koh W, Shariati SAM, Sim S, Neff NF, Skotheim JM, Wernig M, *et al.* (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**:391-395 <https://doi.org/10.1038/nature18323> | PubMed
46. **Fong AP**, Yao Z, Zhong JW, Cao Y, Ruzzo WL, Gentleman RC, Tapscott SJ (2012) Genetic and epigenetic determinants of neurogenesis and myogenesis. *Dev. Cell* **22**:721-735 <https://doi.org/10.1016/j.devcel.2012.01.015> | PubMed
47. **Lee QY**, Mall M, Chanda S, Zhou B, Sharma KS, Schaukowitch K, Adrian-Segarra JM, Grieder SD, Kareta MS, Wapinski OL, *et al.* (2020) Pro-neuronal activity of Myod1 due to promiscuous binding to neuronal genes. *Nat. Cell Biol* **22**:401-411 <https://doi.org/10.1038/s41556-020-0490-3> | PubMed
48. **Replogle JM**, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, Mascibroda L, Wagner EJ, Adelman K, Lithwick-Yanai G, *et al.* (2022) Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**:2559-2575. <https://doi.org/10.1016/j.cell.2022.05.013> | PubMed
49. **Moretti A**, Caron L, Nakano A, Lam JT, Bernshausen A, Chen Y, Qyang Y, Bu L, Sasaki M, Martin-Puig S, *et al.* (2006) Multipotent embryonic isl1+ progenitor cells lead to cardiac, smooth muscle, and endothelial cell diversification. *Cell* **127**:1151-1165 <https://doi.org/10.1016/j.cell.2006.10.029> | PubMed
50. **Lee S**, Park C, Han JW, Kim JY, Cho K, Kim EJ, Kim S, Lee S.-J, Oh SY, Tanaka Y, *et al.* (2017) Direct Reprogramming of Human Dermal Fibroblasts Into Endothelial Cells Using ER71/ETV2. *Circ. Res* **120**:848-861 <https://doi.org/10.1161/circresaha.116.309833> | PubMed

51. **Ferdous A**, Caprioli A, Iacovino M, Martin CM, Morris J, Richardson JA, Latif S, Hammer RE, Harvey RP, Olson EN, *et al.* (2009) Nkx2-5 transactivates the Ets-related protein 71 gene and specifies an endothelial/endocardial fate in the developing embryo. *Proc. Natl. Acad. Sci. U. S. A* **106**:814-819 <https://doi.org/10.1073/pnas.0807583106> | PubMed
52. **Moore AW**, McInnes L, Kreidberg J, Hastie ND, Schedl A (1999) YAC complementation shows a requirement for Wt1 in the development of epicardium, adrenal gland and throughout nephrogenesis. *Development* **126**:1845-1857 <https://doi.org/10.1242/dev.126.9.1845> | PubMed
53. **Zeng B**, Ren X.-F, Cao F, Zhou X.-Y, Zhang J (2011) Developmental patterns and characteristics of epicardial cell markers Tbx18 and Wt1 in murine embryonic heart. *J. Biomed. Sci* **18**:67 <https://doi.org/10.1186/1423-0127-18-67> | PubMed
54. **Major J**, Crotta S, Finsterbusch K, Chakravarty P, Shah K, Frederico B, D'Antuono R, Green M, Meader L, Suarez-Bonnet A, *et al.* (2023) Endothelial AHR activity prevents lung barrier disruption in viral infection. *Nature* **621**:813-820 <https://doi.org/10.1038/s41586-023-06287-y> | PubMed
55. **Ito N**, Kii I, Shimizu N, Tanaka H, Takeda S (2017) Direct reprogramming of fibroblasts into skeletal muscle progenitor cells by transcription factors enriched in undifferentiated subpopulation of satellite cells. *Sci. Rep* **7**:8097 <https://doi.org/10.1038/s41598-017-08232-2> | PubMed
56. **Keepers B**, Liu J, Qian L (2020) What's in a cardiomyocyte – And how do we make one through reprogramming?. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1867**:118464 <https://doi.org/10.1016/j.bbamcr.2019.03.011>
57. **Mall M**, Kareta MS, Chanda S, Ahlenius H, Perotti N, Zhou B, Grieder SD, Ge X, Drake S, Euong Ang C, *et al.* (2017) Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. *Nature* **544**:245-249 <https://doi.org/10.1038/nature21722> | PubMed
58. **Doppler W**, Geymayer S, Weirich HG (2002) Synergistic and Antagonistic Interactions of Transcription Factors in the Regulation of Milk Protein Gene Expression. In: Mol JA, Clegg RA (Eds). *Biology of the Mammary Gland* Springer US. pp. 139-146 https://doi.org/10.1007/0-306-46832-8_17 | PubMed
59. **Chi T**, Lieberman P, Ellwood K, Carey M (1995) A general mechanism for transcriptional synergy by eukaryotic activators. *Nature* **377**:254-257 <https://doi.org/10.1038/377254a0> | PubMed
60. **Gomez-Salinerio JM**, Itkin T, Houghton S, Badwe C, Lin Y, Kalna V, Dufton N, Peghaire CR, Yokoyama M, Wingo M, *et al.* (2022) Cooperative ETS Transcription Factors Enforce Adult Endothelial Cell Fate and Cardiovascular Homeostasis. *Nat Cardiovasc Res* **1**:882-899 <https://doi.org/10.1038/s44161-022-00128-3> | PubMed
61. **Li T**, Conroy KL, Kim AM, Halmai J, Gao K, Moreno E, Wang A, Passerini AG, Nolta JA, Zhou P (2023) Role of MEF2C in the Endothelial Cells Derived from Human Induced Pluripotent Stem Cells. *Stem Cells* **41**:341-353 <https://doi.org/10.1093/stmcls/sxad005> | PubMed
62. **Sacilotto N**, Chouliaras KM, Nikitenko LL, Lu YW, Fritzsche M, Wallace MD, Nornes S, García-Moreno F, Payne S, Bridges E, *et al.* (2022) Corrigendum: MEF2 transcription factors are key regulators of sprouting angiogenesis. *Genes Dev* **36**:1096 <https://doi.org/10.1101/gad.350249.122> | PubMed
63. **Kruse EA**, Loughran SJ, Baldwin TM, Josefsson EC, Ellis S, Watson DK, Nurden P, Metcalf D, Hilton DJ, Alexander WS, *et al.* (2009) Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. *Proc. Natl. Acad. Sci. U. S. A* **106**:13814-13819 <https://doi.org/10.1073/pnas.0906556106> | PubMed
64. **Gekas C**, Rhodes KE, Gereige LM, Helgadottir H, Ferrari R, Kurdistani SK, Montecino-Rodriguez E, Bassel-Duby R, Olson E, Krivtsov AV, *et al.* (2009) Mef2C is a lineage-restricted target of Scl/Tal1 and regulates megakaryopoiesis and B-cell homeostasis. *Blood* **113**:3461-3471 <https://doi.org/10.1182/blood-2008-07-167577> | PubMed
65. **Batta K**, Florkowska M, Kouskoff V, Lacaud G (2014) Direct reprogramming of murine fibroblasts to hematopoietic progenitor cells. *Cell Rep* **9**:1871-1884 <https://doi.org/10.1016/j.celrep.2014.11.002> | PubMed

66. Bell SP, Learned RM, Jantzen HM, Tjian R (1988) Functional cooperativity between transcription factors UBF1 and SL1 mediates human ribosomal RNA synthesis. *Science* **241**:1192-1197 <https://doi.org/10.1126/science.3413483> | PubMed
67. Mohaghegh N, Bray D, Keenan J, Penvose A, Andrienas KK, Ramlall V, Siggers T (2019) NextPBM: a platform to study cell-specific transcription factor binding and cooperativity. *Nucleic Acids Res* **47**:e31 <https://doi.org/10.1093/nar/gkz020> | PubMed
68. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, Pliner HA, Jackson DL, Daza RM, Christiansen L, et al. (2020) Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**:45-51 <https://doi.org/10.1126/science.aax6234> | PubMed
69. Ye C, Ho DJ, Neri M, Yang C, Kulkarni T, Randhawa R, Henault M, Mostacci N, Farmer P, Renner S, et al. (2018) DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun* **9**:4307 <https://doi.org/10.1038/s41467-018-06500-x> | PubMed
70. Camp JG, Wollny D, Treutlein B (2018) Single-cell genomics to guide human stem cell and tissue engineering. *Nat. Methods* **15**:661-667 <https://doi.org/10.1038/s41592-018-0113-0> | PubMed
71. Grath A, Dai G (2019) Direct cell reprogramming for tissue engineering and regenerative medicine. *J. Biol. Eng* **13**:14 <https://doi.org/10.1186/s13036-019-0144-9> | PubMed
72. Kelaini S, Cochrane A, Margariti A (2014) Direct reprogramming of adult cells: avoiding the pluripotent state. *Stem Cells Cloning* **7**:19-29 <https://doi.org/10.2147/sccaa.s38006> | PubMed
73. Horvath P, Aulner N, Bickle M, Davies AM, Nery ED, Ebner D, Montoya MC, Östling P, Pietiäinen V, Price LS, et al. (2016) Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov* **15**:751-769 <https://doi.org/10.1038/nrd.2016.175> | PubMed
74. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, Smibert P, Satija R (2018) Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**:224 <https://doi.org/10.1186/s13059-018-1603-1> | PubMed
75. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun* **8**:14049 <https://doi.org/10.1038/ncomms14049> | PubMed
76. Bhattacharyya S, Duan J, Wang L, Li B, Bhakta M, Fernandez-Perez A, Hon GC, Munshi NV (2020) Using Gjd3-CreEGFP mice to examine atrioventricular node morphology and composition (vol 45, pg 1291, 2020). *Sci. Rep.* **10** <https://doi.org/10.1038/s41598-020-60915-5> | PubMed
77. Bhattacharyya S, Kollipara RK, Orquera-Tornakian G, Goetsch S, Zhang M, Perry C, Li B, Shelton JM, Bhakta M, Duan J, et al. (2023) Global chromatin landscapes identify candidate noncoding modifiers of cardiac rhythm. *J. Clin. Invest* **133** <https://doi.org/10.1172/JCI153635> | PubMed
78. Wolock SL, Lopez R, Klein AM (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**:281-291. <https://doi.org/10.1016/j.cels.2018.11.005> | PubMed
79. Duan J, Hon GC (2021) FBA: feature barcoding analysis for single cell RNA-Seq. *Bioinformatics* **37**:4266-4268 <https://doi.org/10.1093/bioinformatics/btab375> | PubMed
80. Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**:15 <https://doi.org/10.1186/s13059-017-1382-0> | PubMed
81. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P.-R, Raychaudhuri S (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**:1289-1296 <https://doi.org/10.1038/s41592-019-0619-0> | PubMed
82. McInnes L, Healy J, Melville J (2018) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software* **3**:861 <https://doi.org/10.21105/joss.00861>
83. Agrawal A, Ali A, Boyd S (2021) Minimum-Distortion Embedding. *arXiv* <https://doi.org/10.48550/arXiv.2103.02559>
84. Hahsler M, Piekenbrock M, Doran D (2019) dbSCAN: Fast Density-Based Clustering with R. *J. Stat. Softw* **91**:1-30 <https://doi.org/10.18637/jss.v091.i01>

85. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139-140 <https://doi.org/10.1093/bioinformatics/btp616> | PubMed
 86. Yu G, He Q.-Y (2016) ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst* **12**:477-479 <https://doi.org/10.1039/c5mb00663e> | PubMed
 87. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**:979-982 <https://doi.org/10.1038/nmeth.4402> | PubMed
 88. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT (2018) The Human Transcription Factors. *Cell* **172**:650-665 <https://doi.org/10.1016/j.cell.2018.01.029> | PubMed
- Duan et al (2023) Modular regulatory architecture of combinatorial transcription factor networks. NCBI Gene Expression Omnibus. ID GSE225838 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE225838>

Peer reviews

Reviewer #1 (Public review):

Summary:

Duan, Li, Kulkarni et al. apply a multiplexed single-cell overexpression screen (Reprogram-Seq) to combinatorially perturb 105 transcription factors across 7 target cell types in mouse embryonic fibroblasts, generating a resource of ~200,000 single-cell transcriptomes spanning over 1,300 TF combinations. They develop a framework for classifying pairwise TF-TF interactions, identify a modular, shared architecture of gene regulatory programs across diverse TF combinations, and use these tools to nominate and partially validate new reprogramming cocktails.

Strengths:

The scale of the combinatorial screen is substantial, and the resulting dataset is a genuine resource for the field. The TF-TF interaction typing framework is a useful conceptual extension of prior genetic-interaction approaches to an overexpression/reprogramming context, and the modularity finding that diverse TF combinations converge on shared gene programs is a compelling organizing principle. The authors are, for the most part, careful and appropriately hedged in their claims; the overclaiming we flag below is the exception, not the rule. We also note that the core Reprogram-Seq assay itself builds directly on the authors' own prior work; the novelty here rests on scale, the interaction framework, and the modularity analysis.

Weaknesses:

Most of the concerns raised below relate to how existing data are quantified, cited, and reconciled with the text, rather than to the underlying experiments themselves. Several quantitative and comparative claims in the Results are not fully supported by the figures cited, and some conclusions are in tension with the authors' own data. Key methodological details relevant to interpreting the central TF-TF interaction framework, including TF expression dosage and within-combination transcriptional variability, are not reported or controlled for, which limits confidence in the resulting interaction classifications. The relationship between TF number and reprogramming efficiency is not clearly distinguished from a simple combinatorial coverage effect and does not consistently generalize across batches. Experimental validation of predicted cocktails is limited to a single target cell type. The manuscript would also benefit from addressing whether TF overexpression in fibroblasts

can fully capture a factor's endogenous regulatory network, given that pioneer activity and chromatin accessibility are not addressed.

<https://doi.org/10.7554/eLife.111977.1.sa2>

Reviewer #2 (Public review):

Summary:

This manuscript presents a large-scale combinatorial transcription factor overexpression screen in mouse embryonic fibroblasts coupled with single-cell RNA-seq to systematically map the relationship between TF combinations, gene regulatory networks, and transcriptional reprogramming. Using approximately 100 transcription factors, the authors identify TF combinations that shift cells toward diverse transcriptional states, organize TF combinations into "perturbation clusters" with shared transcriptional outputs, infer modular gene regulatory networks, model pairwise TF interactions, and use pseudotime analyses to nominate candidate reprogramming TF combinations. The resulting dataset represents a potentially valuable resource for studying combinatorial TF activity and transcriptional reprogramming.

Strengths:

The primary strength of the study is its experimental scale and the breadth of the generated dataset. The Reprogram-Seq platform enables systematic interrogation of thousands of TF combinations that would be difficult to test individually, and the authors develop several computational frameworks to organize these data and generate biological hypotheses. The epicardial reprogramming analyses, including independent qPCR and protein localization experiments, provide proof-of-principle that the platform can recover biologically relevant TF combinations.

Weaknesses:

Many of the manuscript's central conclusions rely on a complex computational pipeline that is not sufficiently justified or independently validated. Identification of transcriptionally reprogrammed cells depends on co-embedding with reference atlases, yet the robustness of this analysis and the interpretation of cells occupying primary-cell clusters are not explored in depth. Similarly, the conclusions regarding modular gene regulatory networks depend critically on the perturbation clusters defined by MDE embedding and HDBSCAN clustering. These perturbation clusters form the basis for nearly all downstream analyses, including differential expression, gene module identification, gene specificity, and TF modularity, yet little evidence is provided that the clusters are robust to alternative embedding strategies, clustering parameters, or resampling approaches.

The manuscript also provides relatively limited orthogonal validation of its computational predictions. Although the epicardial analyses are validated experimentally, comparable validation is not performed for most other predicted cell fates, TF interaction classes, or perturbation modules. Consequently, many conclusions regarding the generality of modular TF activity, TF cooperativity, and the predicted reprogramming cocktails remain supported primarily by computational inference.

In addition, several aspects of the analytical workflow—including the quality filtering of TF combinations, interpretation of unclustered perturbations, selection of genes for downstream visualization, and robustness of pseudotime analyses across lineages—would benefit from greater methodological transparency.

Overall, this work provides a valuable dataset and introduces analytical approaches that will likely be useful to the community. However, in its current form, I believe the strongest

biological conclusions are insufficiently validated. Additional analyses demonstrating the robustness of the computational framework, together with broader orthogonal validation of representative predictions, would substantially strengthen confidence in the proposed principles governing combinatorial transcription factor activity and transcriptional reprogramming.

<https://doi.org/10.7554/eLife.111977.1.sa1>

Reviewer #3 (Public review):

Summary:

In this manuscript, Duan et al perform a combinatorial TF overexpression screen combined with single-cell RNA-seq (Reprogram-Seq) to extract general principles of how combinatorial TF interactions drive distinct gene regulatory networks in reprogrammed cells. Using a library of 105 TFs, they induce different cell fates, many of which resemble in vivo cell identities. They observe that combinations of TFs have better reprogramming results than inductions driven by a single TF. By looking at gene expression enrichment/depletion in different reprogrammed cell clusters, they infer functional GRNs induced by specific TF combinations and identify GRNs specific for certain cell types. They also identify TFs that could improve known TF cocktails for the induction of certain cell fates. They observe that TFs with cooperative interactions regarding the regulation of gene expression may lead to better reprogramming results, and finally, they build a bottom-up approach that utilizes the single-cell transcriptomes to predict TFs driving certain reprogrammed fates.

Strengths:

Reprogram-Seq is not new, but the strength of the study lies in the fact that the authors assess the induction outcome from a large number of different TF combinations. The authors are thus able to make broad observations, such as the modularity of GRNs and TF cooperativity, as well as propose new TFs and TF interactions to be tested for the induction of certain cell fates. The manuscript is well written, and the conclusions are, in general, supported well by the authors' analyses and data.

Weaknesses:

The study would benefit from some further analysis and discussion to better tighten the conclusions:

While both expression enrichment and depletion were used to define perturbation clusters, the authors then focused on analyzing functional gene groups only for the enriched genes. Are there any functional relations between the repressed genes within a perturbation cluster? Do the authors observe the same modularity (in terms of regulation by TFs) for repressed genes as they do for induced genes?

Can the authors give a description of how neomorphic TF interactions work? How would gene expression be affected in single vs double perturbation in those cases?

What does it mean functionally when a TF pair shows more than one type of interaction (as shown in Supplementary Table 6), and how do such interactions correlate with successful transcriptional reprogramming?

In the last Results section, the authors are able to use the transcriptome to predict the TF that was used for the induction. Could the authors discuss some plausible applications of this TF prediction method? For example, could they use it on in vivo single-cell RNA-seq of a certain cell type to predict candidate TFs for the induction of that cell type?

It would help the reader if, at the end of each Results section, the authors add a concluding paragraph, highlighting the most important conclusions and findings (same as they have done in the section titled "Combinatorial TF over-expression reprograms MEFs to diverse states").

<https://doi.org/10.7554/eLife.111977.1.sa0>