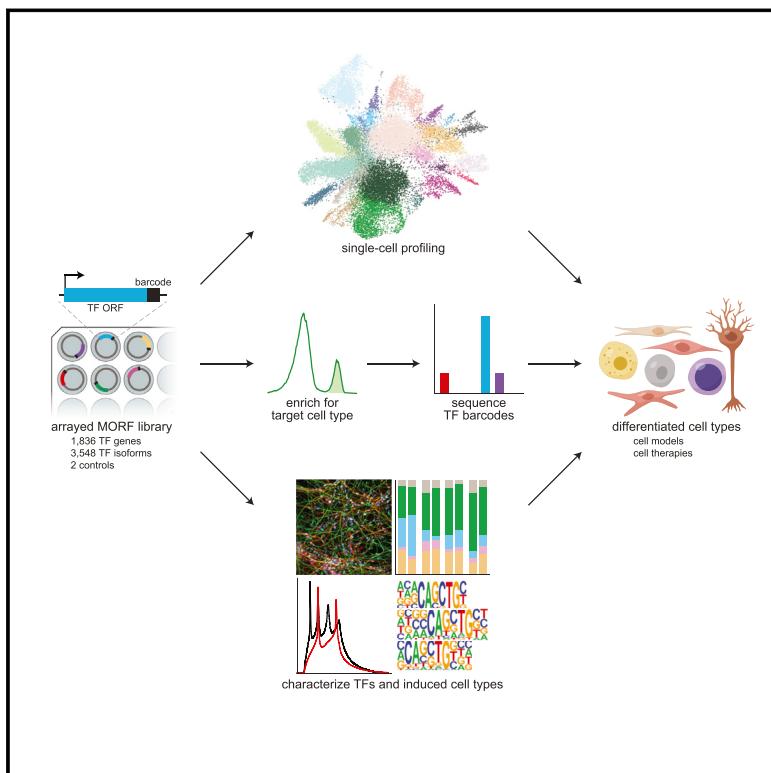


A transcription factor atlas of directed differentiation

Graphical abstract



Authors

Julia Joung, Sai Ma, Tristan Tay, ...,
Jason D. Buenrostro, Aviv Regev,
Feng Zhang

Correspondence

zhang@broadinstitute.org

In brief

Generation of a comprehensive human transcription factor (TF) barcoded ORF library and further application to embryonic stem cells to build a TF Atlas of resulting expression profiles, which enabled identification of individual and combinations of TFs that produce target cell types and, thus, accelerate cellular engineering efforts.

Highlights

- Barcoded ORF library of all 3,548 human TF splice isoforms
- Applied to human stem cells to build a TF Atlas of resulting expression profiles
- Mapping of TFs that produce cell types from all three germ layers and trophoblasts
- Prediction of TF combinations to produce target cell types



Resource

A transcription factor atlas of directed differentiation

Julia Joung,^{1,2,3,4,5,12} Sai Ma,^{6,7,8,9,14} Tristan Tay,^{8,9,14} Kathryn R. Geiger-Schuller,^{6,7,13,14} Paul C. Kirchgatterer,^{1,2,3,4,5} Vanessa K. Verdine,^{1,2,3,4,5} Baolin Guo,^{4,10} Mario A. Arias-Garcia,^{4,10} William E. Allen,^{1,2,3,4,5,11} Ankita Singh,^{1,2,3,4,5} Olena Kuksenko,^{6,7} Omar O. Abudayyeh,^{1,2,3,4,5} Jonathan S. Gootenberg,^{1,2,3,4,5} Zhanyan Fu,^{4,10} Rhiannon K. Macrae,^{1,2,3,4,5} Jason D. Buenrostro,^{8,9} Aviv Regev,^{6,7,13} and Feng Zhang^{1,2,3,4,5,15,*}

¹Department of Biological Engineering, MIT, Cambridge, MA 02139, USA

²Department of Brain and Cognitive Science, MIT, Cambridge, MA 02139, USA

³Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

⁴McGovern Institute for Brain Research at MIT, Cambridge, MA 02139, USA

⁵Howard Hughes Medical Institute, MIT, Cambridge, MA 02139, USA

⁶Department of Biology, MIT, Cambridge, MA 02139, USA

⁷Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

⁸Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA

⁹Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

¹⁰Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

¹¹Society of Fellows, Harvard University, Cambridge, MA, USA

¹²Present address: Whitehead Institute for Biomedical Research, 455 Main St, Cambridge, MA 02142, USA

¹³Present address: Genentech, 1 DNA Way, South San Francisco, CA 94080, USA

¹⁴These authors contributed equally

¹⁵Lead contact

*Correspondence: zhang@broadinstitute.org

<https://doi.org/10.1016/j.cell.2022.11.026>

SUMMARY

Transcription factors (TFs) regulate gene programs, thereby controlling diverse cellular processes and cell states. To comprehensively understand TFs and the programs they control, we created a barcoded library of all annotated human TF splice isoforms (>3,500) and applied it to build a TF Atlas charting expression profiles of human embryonic stem cells (hESCs) overexpressing each TF at single-cell resolution. We mapped TF-induced expression profiles to reference cell types and validated candidate TFs for generation of diverse cell types, spanning all three germ layers and trophoblasts. Targeted screens with subsets of the library allowed us to create a tailored cellular disease model and integrate mRNA expression and chromatin accessibility data to identify downstream regulators. Finally, we characterized the effects of combinatorial TF overexpression by developing and validating a strategy for predicting combinations of TFs that produce target expression profiles matching reference cell types to accelerate cellular engineering efforts.

INTRODUCTION

Achieving a comprehensive understanding of the gene regulatory networks (GRNs) governing cell states is a fundamental goal in molecular biology. Transcription factors (TFs) bind to specific sequences in the genome to alter gene expression and specify cell states.^{1–3} Overexpression of single TFs can drive profound changes in cell fate. For instance, single TFs have been shown to direct differentiation of pluripotent stem cells toward many different cell types,^{4,5} including muscle⁶ and neurons.⁷ Overexpression of TF combinations can produce even larger changes in GRNs,^{8–10} such as overexpression of the four “Yamanaka factors” (Oct4, Sox2, Klf4, and c-Myc) to reprogram fibroblasts into stem cells.^{11–13} These findings underscore the power of TFs to drive changes in cell state and highlight the utility of TF overex-

pression for understanding the gene expression programs that control cell fates.¹⁴

The human genome contains >1,800 TF loci encoding >3,500 isoforms, creating a vast landscape of possible regulatory outcomes. Previous studies have explored aspects of this landscape through observational studies, such as quantitative trait locus mapping to associate TFs with phenotype, and perturbation studies that overexpress or inhibit TFs in model systems. However, perturbation studies have typically had to choose between breadth of perturbation and phenotypic content, either conducting large screens with a simple readout or small focused screens with detailed readout. For example, in the context of TF overexpression screens, a recent study screened a large library of 1,732 TF isoforms for the focused readout of pluripotency marker expression⁴; conversely, a smaller analysis of 61 TFs



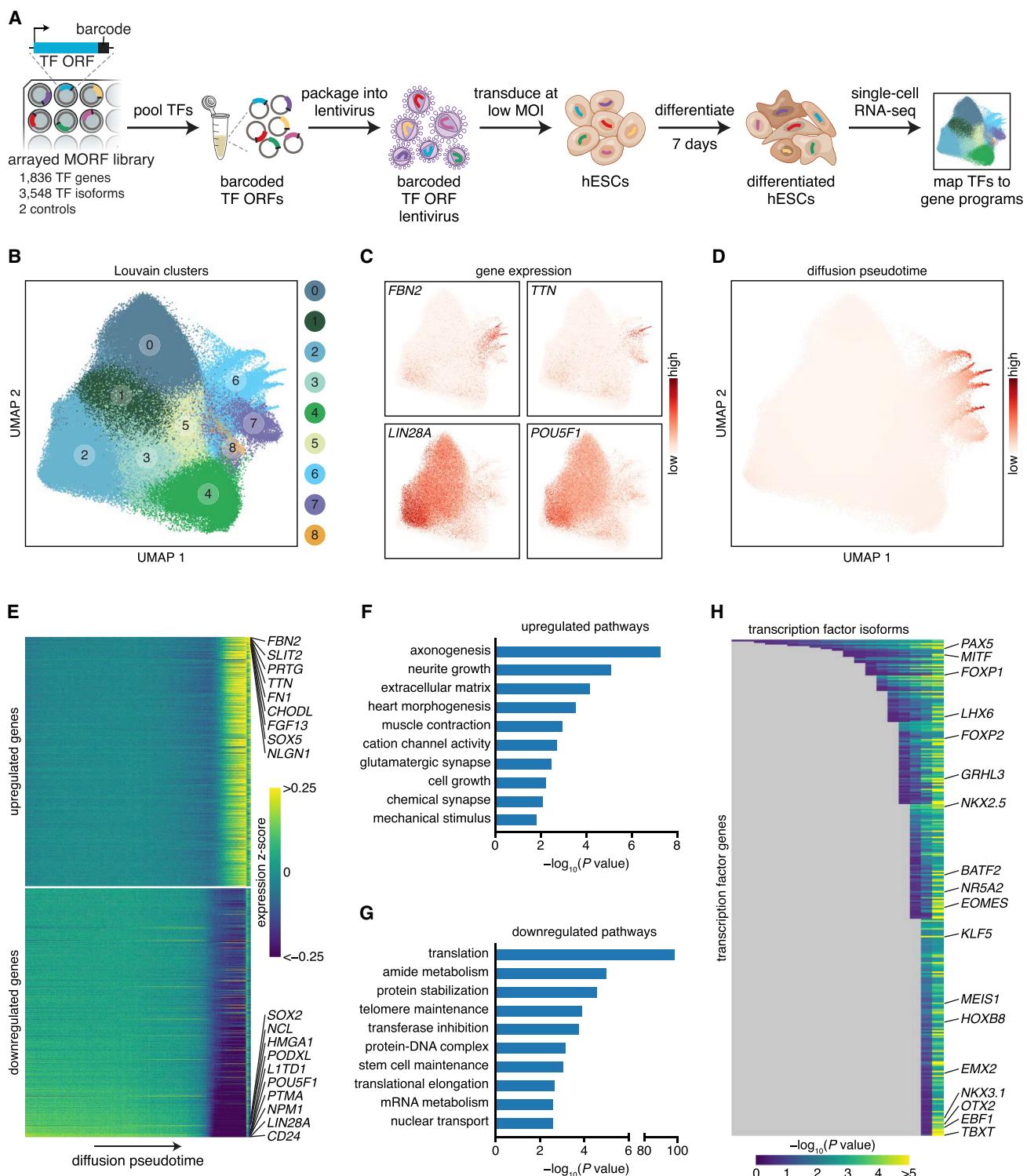


Figure 1. Building a TF Atlas of directed differentiation

(A) Schematic of TF Atlas setup. MOI, multiplicity of infection.

(B–D) Uniform manifold approximation and projection (t-SNE) of scRNA-seq data from 671,453 cells overexpressing 3,266 TF isoforms. Colors indicate Louvain clusters (B), gene expression (C), and diffusion pseudotime (D).

(legend continued on next page)

assessed the comprehensive impact of each through single-cell profiling.⁵

Fully deciphering regulatory circuits, however, requires a systematic approach that integrates both the breadth of a comprehensive screen and the depth of a rich readout, especially of the transcriptomic changes induced by each TF. Here, we systematically mapped expression changes driven by all human TF isoforms at single-cell resolution and used these data to identify TFs and their combinations that direct differentiation in human embryonic stem cells (hESCs). We created a barcoded ORF library of 3,548 TF splice isoforms and developed a screening platform to build a TF Atlas of >1 million cells charting expression profile changes induced by TF overexpression in hESCs. Our comprehensive TF Atlas enables both a systematic identification of TFs that drive changes in cell state and generalized observations such as classification of orphan TFs by gene program. Furthermore, we applied our TF Atlas to predict and validate TF combinations for target reference cell types. Our TF library and Atlas thus provide a valuable resource for systematic elucidation of TF gene programs toward a complete understanding of GRNs that govern cell states.

RESULTS

Development of multiplexed overexpression of regulatory factors library

We first established the most effective method for TF upregulation by comparing CRISPR activation (CRISPRa)¹⁵ and ORF overexpression. ORF overexpression of either *NEUROD1* or *NEUROG2* in hESCs efficiently induced neuronal differentiation,⁷ but CRISPRa and ORFs with endogenous untranslated regions (UTRs) did not (Figures S1A–S1D). This may indicate that hESCs have post-transcriptional regulatory mechanisms encoded in UTRs that buffer against protein expression. We therefore proceeded with ORF overexpression.

To enable pooled screening, we created a barcoded human TF ORF library, multiplexed overexpression of regulatory factors (MORFs; Figures 1A and S1E; Table S1A). The MORF library consists of all 3,548 annotated splice isoforms encoded by 1,836 genes, including histone modifiers. Each isoform is associated with a unique barcode that facilitates TF identification and minimizes ORF length-dependent PCR bias.¹⁶ As an arrayed library that could be selectively pooled for targeted screening, MORF is a generalizable resource for discovering TFs that induce phenotypes of interest.

Construction of a TF Atlas of directed differentiation

We first applied MORF to comprehensively profile gene programs regulated by each TF. As existing cellular differentiation protocols often use different culture media, we tested 7 culture media to identify an optimal media condition that could capture the broadest range of effects of TFs. To this end, we pooled the

MORF library TFs, transduced H1 hESCs with the library, and cultured cells for 7 days in each of 7 media conditions (Figures S1F and S1G; STAR Methods). We sorted cells into two populations (top and bottom 10%) based on the expression level of pluripotency markers (TRA-1-60 and SSEA4), as a proxy for differentiation, and sequenced the TF barcodes. Despite even TF distributions in the initial libraries (skew = 5), the distributions became very skewed across all media conditions (skew = 105–115; Figure S1H). As the distributions were remarkably consistent across replicates and media conditions (Pearson $r > 0.94$; Figures S1I and S1J; Table S1A), the increase in skew is likely a result of TF-dependent effects, rather than media-derived factors. TFs that promote pluripotency (e.g., *IDs* 1, 3, and 4¹⁷; and *YAF2*¹⁸) increased cell fitness, whereas TFs involved in DNA damage (e.g., *BRCA1* and *TP53BP1*¹⁹) decreased cell fitness (Figure S1I). TF representation negatively correlated with TF length (Figures S1K and S1L). We found that TF distributions in the sorted populations were relatively consistent across media conditions (Figures S1M–S1O; Table S1A), suggesting that culture media does not strongly influence differentiation potential under TF overexpression. Of the 7 media conditions, we selected STEMdiff APEL because it produced the highest enrichment and most even distribution for known developmentally critical TFs⁵ (Figures S1N and S1P).

To build an expression atlas of all TF overexpression effects, we transduced hESCs with the MORF library, cultured cells in STEMdiff APEL media for 7 days, and profiled the cells by single-cell RNA sequencing (scRNA-seq) using a combinatorial indexing protocol based on simultaneous high-throughput ATAC and RNA expression with sequencing²⁰ (SHARE-seq; Figure 1A; STAR Methods). We obtained >1.1 million cell profiles (3,761 UMIs per cell on average; Figures S2A and S2B) and down-sampled the data by TF ORF to 671,453 cells covering 3,266 TFs (92% of the MORF library; Figure S2C; STAR Methods) to ensure even representation (3–1,000 cells, with an average of 206 cells per TF ORF). Expression level of TF ORFs did not correlate with TF length or expression of the respective endogenous TF (Figures S2D and S2E). Cluster analysis showed that clusters 6–8 expressed higher levels of differentiation genes (*FBN2* and *TTN*) and lower levels of pluripotency genes (*LIN28A* and *OCT4* [*POU5F1*]; Figures 1B and 1C). Cells overexpressing developmentally critical TFs such as Brachyury (*TBXT*) and *KLF4* had distinct transcriptomes and clustered together (Figures S2F and S2G). Together, these results show that our approach for profiling TF effects by scRNA-seq maximizes the range of possible TF-induced cell states and lays the groundwork for detailed characterization of the impact of TF overexpression.

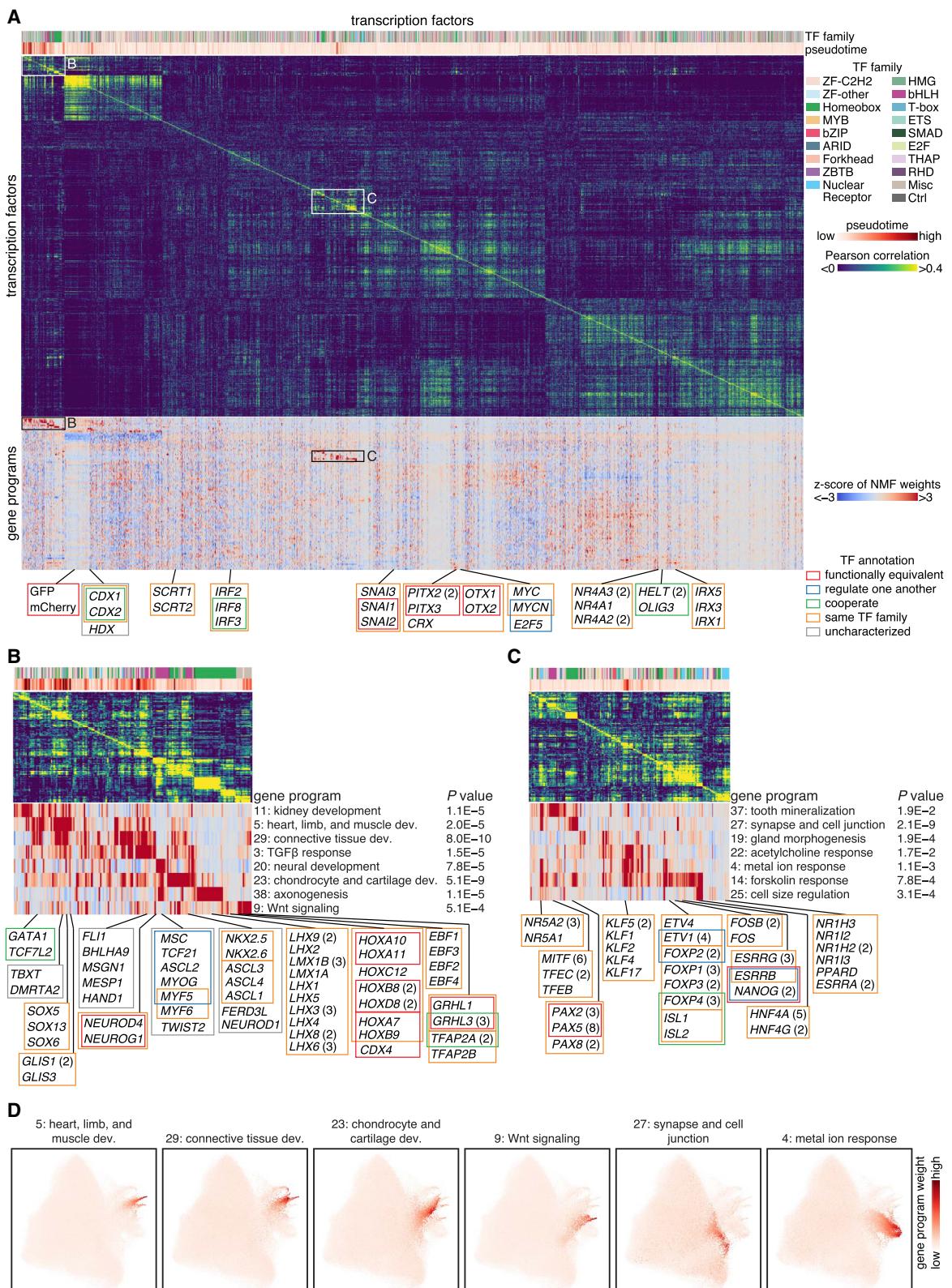
Over a quarter of TFs direct differentiation of hESCs

To study the effects of TF overexpression on hESC differentiation, we computationally inferred differentiation trajectories

(E) Smoothed heatmap of the top 1,000 upregulated and downregulated genes over diffusion pseudotime. Genes are ordered by change over pseudotime. (F and G) Most enriched pathways among the top 100 upregulated (F) and downregulated (G) genes.

(H) Heatmap showing the significance of the pseudotime difference between cells expressing each TF isoform and those expressing controls. Only 320 TF genes with multiple isoforms, at least one of which is significant, are included.

See also Figures S1 and S2.



(legend on next page)

from the TF Atlas. We ordered TF-overexpressing cells in pseudotime using two approaches, diffusion,²¹ and RNA velocity,²² based on expression profile similarity to cells expressing GFP or mCherry controls (Figures 1D and S2H–S2M; STAR Methods). Inferred pseudotimes were comparable between the two methods and did not correlate with quality control variables (Figures S2N–S2Q).

Examination of the expression profiles across pseudotime showed that expression of genes that drive differentiation (e.g., *FBN2*, *TTN*, and *SOX5*) increased over pseudotime, whereas those that maintain pluripotency (e.g., *CD24*, *LIN28A*, and *OCT4* [*POU5F1*]) decreased (Figures 1E and S2R; Table S2A). Further confirming our pseudotime inference, differentiation pathways such as axonogenesis and heart morphogenesis were enriched in pseudotime-upregulated genes (Figure 1F). Pluripotency maintenance pathways such as telomere and stem cell maintenance were enriched in pseudotime-downregulated genes (Figure 1G). Remarkably, translation (i.e., ribosomal subunits and translation factors) was by far the top pseudotime-downregulated pathway (Figure 1G), suggesting its importance during differentiation.²³

Using pseudotime as a measurement of differentiation, we evaluated the ability of each TF ORF to direct differentiation. Some TF ORFs that increased pseudotimes were also identified in the marker-based screen (Brachyury [*TBXT*], *MSGN1*, *RFX4*, and *EOMES*), whereas others (*SOX6*, *KLF4*, and *TOX3*) were not, potentially because scRNA-seq captures the full expression profile (Figure S2S). Surprisingly, 496 (27%) TFs encoding 694 (20%) isoforms could significantly alter pseudotime (Wilcoxon rank-sum test; FDR < 0.05), suggesting a high percentage of TFs could act as master regulators, perhaps because of the relatively open chromatin in hESCs²⁴ (Table S2B). Notably, differentiation efficiencies were often drastically different between splice isoforms of the same TF gene (Figures 1H and S2T), and these differences could not be predicted based on nominal protein domain annotations, length, or consensus sequence (Table S2C), highlighting the need for experimental testing. Together, our pseudotime analyses identified universal pathways in differentiation and a high percentage of TFs that can drive profound changes in cell state, often in an isoform-dependent manner.

Co-functional TF modules annotate uncharacterized TFs

We next leveraged the comprehensive scope of our TF Atlas to group co-functional modules of TFs that impact the same programs and classify unknown, orphan TFs. We first inferred gene programs using non-negative matrix factorization and then clustered the TFs by their effects across the programs (cluster Pearson correlation p value < 10⁻⁷; Figure 2; Table S3). Clustering TFs using pairwise correlation of their mean expression

profiles produced similar groupings (Table S3; Data S1.1). Our analysis grouped together splice isoforms and functionally equivalent TFs (e.g., *NEUROD4* and *NEUROG1*,²⁵ *PAX2* and *PAX5*,²⁶ and *ESRRB* and *NANOG*²⁷), as well as TFs from the same family, including 18 LIM homeodomain TFs (*LHX*s 1–6, 8, 9; *LMX*s 1A and 1B), 9 posterior *HOX* genes (*HOX*s *A7*, *B8*, *D8*, *B9*, *A10*, *A11*, and *C12*), and 8 nuclear receptors (*NR1H2*, *NR1H3*, *NR1I2*, *NR1I3*, *PPARD*, and *ESRR*A) (Figures 2B and 2C).

This analysis helps annotate relatively uncharacterized TFs by their functional association with well-characterized TFs. For instance, *KLF17* has little functional information and is distantly related to other *KLF* TF family members, which can act as activators or repressors.²⁸ As *KLF17* induces a similar gene program to *KLF* activators (*KLF*s 1, 2, 4, and 5), it is likely an activator (Figure 2C). As another example, we found that overexpression of *HDX*, which was previously assigned to the POU homeobox class based on phylogenetic analyses despite the lack of a POU-specific domain,²⁹ produces similar gene programs as *CDX1* and *CDX2*, suggesting that *HDX* likely belongs to the caudal-related homeobox family instead (Figure 2A). TFs from different families also group together based on similarities in gene programs, such as *DMRTA2* and *TBX7* in mesoderm development and *FERD3L* and *NEUROD1* in neural development (Figure 2B).

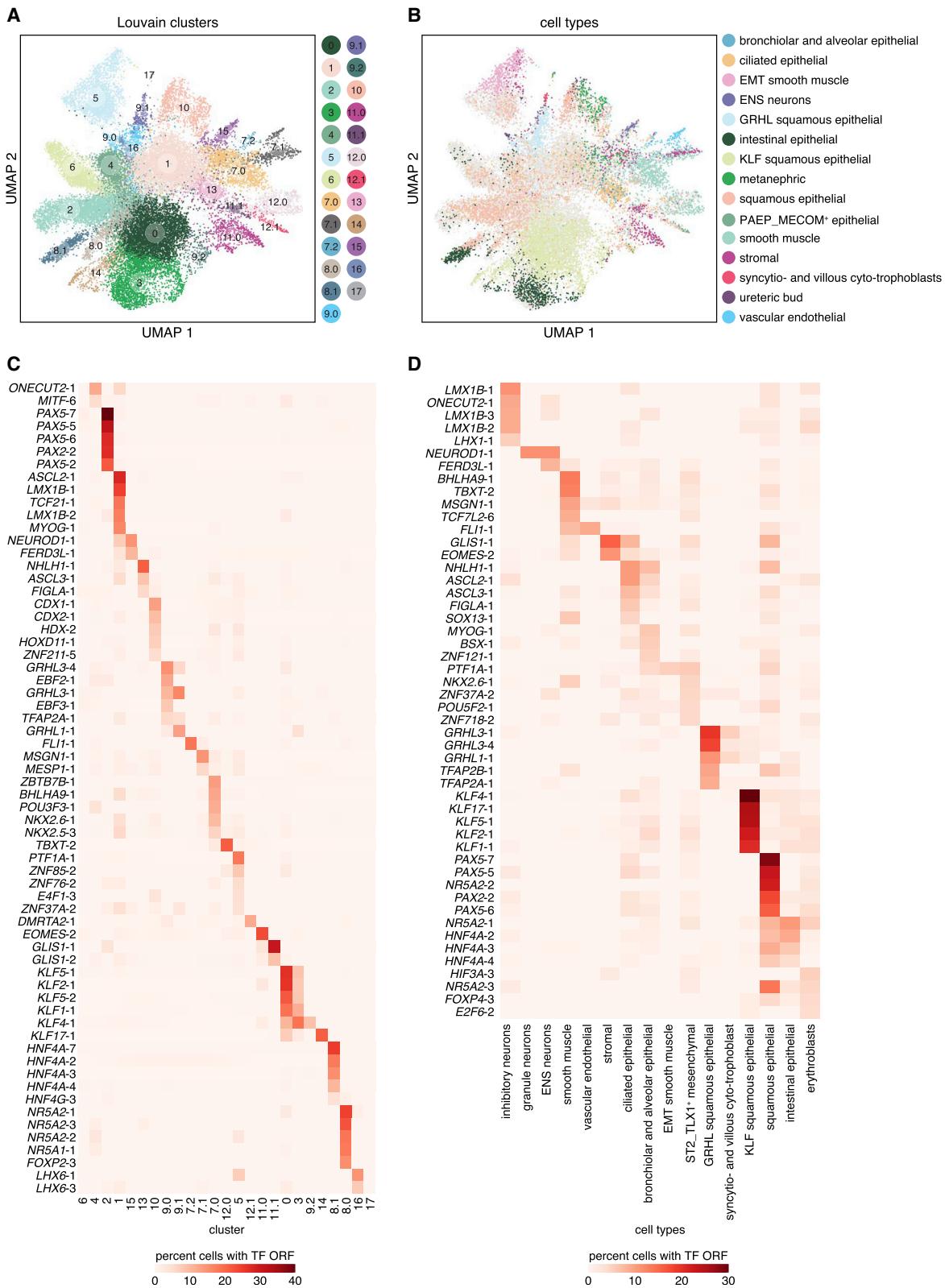
To identify differences between TFs that regulate similar gene programs, we defined a “TF potential” vector for each TF that characterizes its differentiation potential using a linear regression model to fit the TF-induced expression changes against pseudotime, defining TF potential as the slope (STAR Methods). We applied this analysis to distinguish between TFs within the same family that share similar DNA-binding domains. For example, in the *KLF* TF family, TF potential analysis showed that the two *KLF5* isoforms are nearly identical and more similar to *KLF1*, whereas *KLF17* is more distinct (Figure S2U). Similarly, for the LIM homeodomain TF family, *LHX1* and *LHX3* are functionally closer to *LMX1A* and *LMX1B*, whereas *ISL1* and *ISL2* are more like each other than the rest of the TF family (Figure S2V). Together, these results demonstrate that our TF Atlas enables systematic functional annotation of uncharacterized TFs, as well as in depth investigation of TF differentiation potential, providing guidance for further study.

Mapping TF effects to reference cell types

To characterize the ability of TFs to drive differentiation to particular endpoints, we mapped TF-induced expression profiles to those of reference cell types from the human fetal expression atlas.³⁰ We subclustered the TF Atlas differentiated cells (clusters 6–8 from Figure 1B) and annotated cells by label transfer³¹ (Figures 3A and 3B; STAR Methods). Dataset integration³² and random forest classifiers^{33,34} produced similar annotations

Figure 2. Unbiased grouping of TFs based on gene programs

(A) Heatmaps showing pairwise Pearson correlation (top) and enrichment of 100 gene programs (bottom) identified using non-negative matrix factorization (NMF) on mean expression profiles of 3,266 TF ORFs. TFs are ordered by hierarchical clustering. Relations between some groups of TFs are annotated. Numbers in parentheses indicate the number of TF isoforms in the same group.
 (B and C) Zoomed in subsets of (A) with top enriched pathway for each gene program.
 (D) UMAP of TF Atlas scRNA-seq data highlighting enrichment of each gene program.



(legend on next page)

(Figures S3A–S3E; STAR Methods). The mapping results suggest that we generated cells resembling types from each of the three germ layers, such as (1) squamous epithelial and neurons (ectoderm), (2) smooth muscle and metanephric (mesoderm), and (3) intestinal epithelial and bronchiolar and alveolar epithelial (endoderm), as well as from the extraembryonic lineage (syncytiotrophoblast and villous cytotrophoblast) (Figure 3B). Each cluster comprised cells with distinct groups of TF ORFs (adjusted mutual information score of 0.43 for TFs with >5% cells) and differentially expressed genes (DEGs), indicating the diversity and specificity of TF-induced differentiation states (Figure 3C; Table S4A). The biological pathways enriched in each cluster were consistent with cell type annotations (Figure S3F).

Matching TF ORFs to cell types nominated candidate TFs that could induce differentiation of each cell type (Figure 3D). Notably, several candidate TFs are known to be important for specifying the target cell type during development. For instance, *FERD3L* is important for neurogenesis,³⁵ *FLI1* for endothelial development,³⁶ and *KLF4* for intestinal epithelial homeostasis.³⁷

For reference cell types with insufficient numbers of cells mapped by label-transfer, we applied our TF potential analysis to identify TFs that drive differentiation toward those cell types (STAR Methods). We first verified that TF potential analysis could recover TFs nominated by label-transfer (Figures 3D and S3G). For unmapped reference cell types, our analysis nominated several known TFs, such as *HNF4A* and *HNF1A* for reprogramming toward hepatoblasts^{38,39} (Figure S3H). Other nominated TFs were essential for generating the target cell type during development. Examples included *PTF1A* for acinar cells,⁴⁰ *KLF2* for erythrocytes,⁴¹ *ATOH7* for retinal ganglion cells (RGCs),⁴² and *BHLHE22* for retinal and peripheral neurons.^{43,44}

Furthermore, we could predict cell fate bottlenecks by using the distances between expression profiles of TF Atlas cells and reference cell types as a proxy for differentiation complexity. Reference cell types that were farther from TF Atlas cells over-expressing a single TF may require induction by multiple TFs and have bottlenecks in cell fate specification. Our analysis identified bottlenecks in known developmental trajectories, supporting this approach (Figures S3I–S3K). For instance, lymphatic endothelial cells were farther from TF Atlas cells than vascular endothelial cells, reflecting the specification of lymphatic vasculature from blood vasculature during development⁴⁵ (Figure S3I). Similarly, microglia were more distant than blood cell types and acinar cells were more distant than islet endocrine cells, following their respective developmental trajectories^{46,47} (Figures S3I and S3K). In addition, we identified potential bottlenecks for specifying cell types with more specialized functions, such as cardiomyocytes (Figure S3J) and acinar cells (Figure S3K). Our findings demonstrate that

the TF Atlas enables discovery of cell fate specification bottlenecks that were not apparent from examining the reference cell atlas alone.³⁰ Together, these results show that our TF Atlas could nominate TFs that drive hESCs toward specific cell fates and predict bottlenecks, underscoring its utility for cellular engineering.

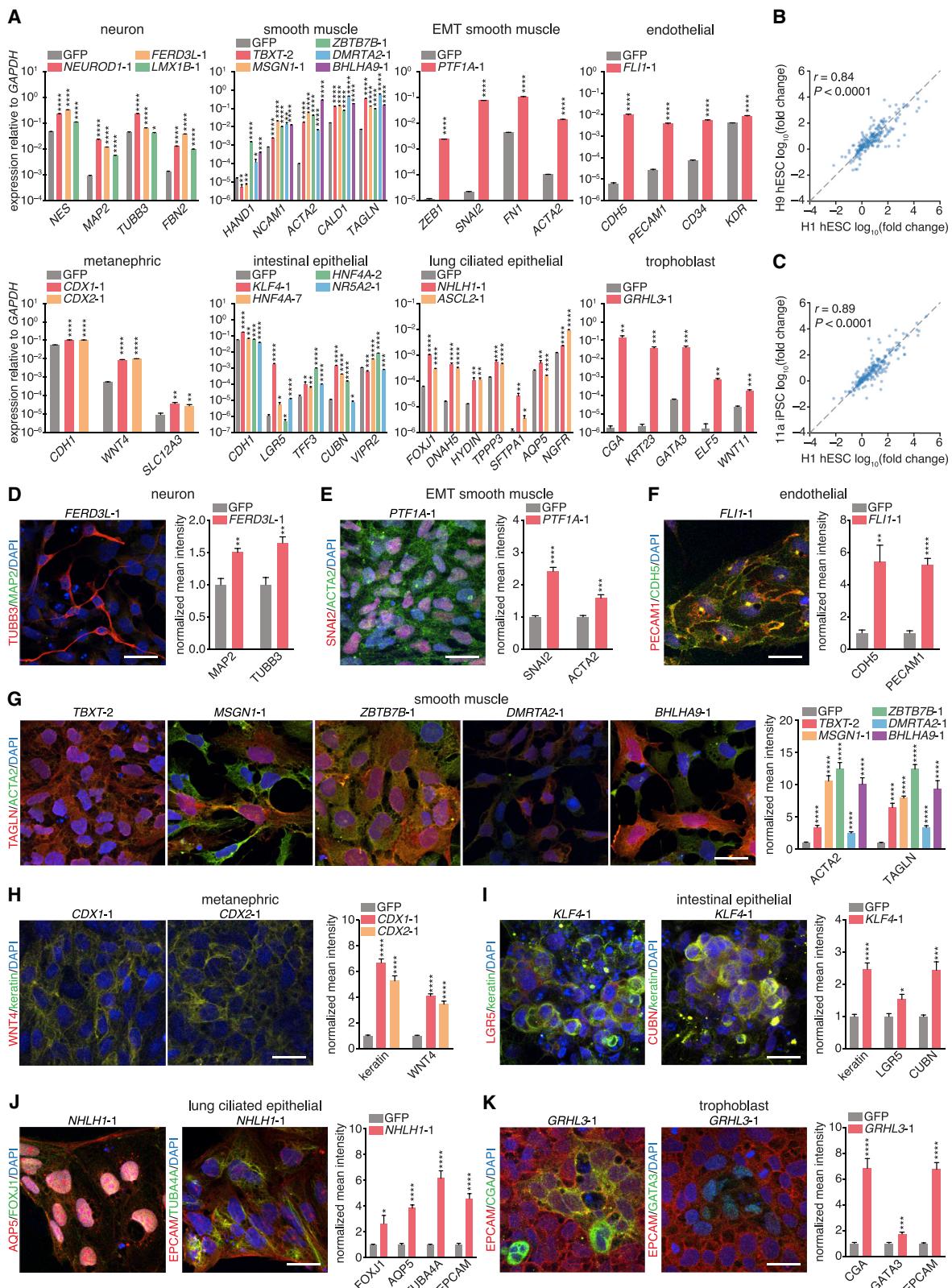
Validation of differentiation-directing TFs

To validate the cell type mapping results, we selected 24 candidate TFs that were predicted to generate 10 distinct cell types, including *NEUROD1* as a positive control.⁷ After 7 days of TF overexpression, most candidates (22 out of 24) induced expression of known marker genes that delineate each predicted cell type (Figures 4A, S3L, and S3M; Table S4B). For instance, *NEUROD1*, *FERD3L*, and *LMX1B* produced peripheral neuron-like cells; *FLI1* produced vascular endothelial-like cells; *KLF4*, *HNF4A*, and *NR5A2* produced intestinal epithelial-like cells; and *NHLH1* and *ASCL2* produced lung ciliated epithelial-like cells. Within each cell type, different TFs sometimes generated distinct expression profiles of marker genes, indicating differences in differentiation efficiencies or trajectories. For example, *ZBTB7B*, *DMRTA2*, and *BHLHA9* induced smooth muscle cells (SMMs) expressing *HAND1*, suggesting a lateral plate mesoderm origin,⁴⁸ whereas *TBXT* and *MSGN1* did not, consistent with their roles in paraxial mesoderm specification^{49,50} (Figure 4A). *GRHL3*, which was predicted to induce both trophoblasts and ureteric bud cells, only generated trophoblast-like cells (Figures 4A and S3M). *EOMES* and *GLIS1* induced expression of general stromal cell markers *LUM* and *COL1A1*, but not the subpopulation marker *ENG* (Figure S3M). *PAX2* and *PAX5* were the only candidate TFs that did not produce the predicted cell type (Figure S3M). TF effects on expression were remarkably consistent in H9 hESCs and 11a induced pluripotent stem cells (iPSCs; Pearson $r = 0.84$ and 0.89 , respectively), suggesting that the results extend beyond the screening cell line (Figures 4B and 4C; Data S1.2A and S1.2B).

We then further characterized 8 cell types generated by a subset of 17 candidate TFs by immunostaining. Of the 17 TFs, 15 significantly upregulated protein expression of marker genes and induced morphologies that resembled those of reference cell types (Figures 4D–4K and S3N–S3P). For instance, *FERD3L* cells had neuron-like axonal projections, *FLI1* cells established vascular endothelial-like tight junctions, *KLF4* cells organized into clumps resembling intestinal crypts, and *GRHL3* cells formed syncytia typical of trophoblasts. *HNF4A* and *ASCL2* did not significantly upregulate protein expression of marker genes due to low differentiation efficiencies (Figures S3O and S3P). These results indicate that our analysis robustly predicts candidates that direct specific cell fates, some of which were missed by the bulk TF screen due to choice of pluripotency markers⁵¹.

Figure 3. Mapping TF ORFs in differentiated cells to reference cell types

(A and B) UMAP of scRNA-seq data from 28,825 differentiated cells (clusters 6–8 in Figure 1B). Colors indicate Louvain clusters (A) and nominated reference cell types³⁰ with score >0.3 (B). (C and D) Heatmaps showing percentage of cells with the indicated TF ORF for each cluster (C) or nominated cell type (D). Numbers after TF gene names indicate the isoform. Percentages are normalized to the total number of cells with the indicated TF ORF in the TF Atlas. For each cluster, only the 5 most enriched TF ORFs >5% are shown. EMT, epithelial-mesenchymal transition; ENS, enteric nervous system. See also Figure S3.



(legend on next page)

(Data S1.2C), highlighting the importance of unbiased expression profiling.

Targeted TF screening to create tailored cellular disease models

Cellular disease models are a tractable system that can be perturbed, genetically or chemically, to assess effects in a cell type-specific context.^{52,53} However, it remains challenging or impossible to generate many cell types. The best differentiation methods are often labor intensive and can require months to produce even heterogeneous or immature cell populations. We sought to address this challenge through targeted TF screening.

As a generalizable approach for constructing targeted TF libraries, we used available expression data to select 90 TF isoforms specifically expressed in a target cell type-induced neural progenitors (iNPs; Table S1B; STAR Methods). Although we could use our TF Atlas to predict TFs that generate iNPs (Figure S4A), we sought to establish a universal approach for producing any cell type of interest. We introduced the pooled, targeted TF library into hESCs and differentiated the cells for 7 days (Figure 5A). We explored three methods for selecting iNPs that can simultaneously assay different numbers of marker genes: reporter cell line (1 gene), flow-FISH (2–10 genes), and scRNA-seq (up to ~2,000 genes; Figure 5A; STAR Methods). We obtained concordant screening results (Spearman correlation p value < 10⁻³) with overlapping sets of candidate TFs (Figures 5B and S4B–S4K; Table S1B), some of which are critical for neural development (e.g., *NFIB*,⁵⁴ *PAX6*,⁵⁵ and *ASCL1*⁵⁶).

For downstream analysis, we focused on 9 TFs (*NFIB*, *RFX4*, *NFIC*, *EOMES*, *OTX1*, *LHX2*, *PAX6*, *FOS*, and *ASCL1*) with the highest average screen ranking. Although all 9 TFs induced expression of VIM, a screening selection marker, 4 (*RFX4*, *NFIB*, *PAX6*, and *ASCL1*) produced multipotent iNPs, like NPs, could spontaneously differentiate into neurons and astrocytes (Figures 5C and S4L–S4N; Table S5A; Data S1.3). scRNA-seq profiles of spontaneously differentiated cells revealed a broad range of cell types that was distinct between TFs, with *RFX4*-iNPs producing more central nervous system (CNS) cell types (Figures S4O–S4R; Data S1.4). Chromatin immunoprecipitation with sequencing (ChIP-seq) identified TF motifs, transcriptional co-regulators, and candidate target genes (Table S5C; Data S1.5).

We further optimized *RFX4*-iNPs by combining *RFX4* overexpression with dual SMAD (DS) inhibition (Data S1.6A–S1.6D) and compared our optimized protocol (*RFX4*-DS) with two previous NP differentiation methods.^{57,58} scRNA-seq profiling showed that *RFX4*-DS-iNPs were most consistent within and between replicates (Figures 5D–5H and S4S–S4V). Moreover,

spontaneously differentiated cells from *RFX4*-DS-iNPs were remarkably reproducible and predominantly consisted of radial glia (RG) and neurons, with only 2%–6% meningeal cells (Figures 5I–5K and S5A–S5C; Data S1.6E–S1.6G). The propensity for *RFX4*-DS-iNPs to spontaneously differentiate into GABAergic neurons (Figure S5D), rather than glutamatergic neurons like iNPs produced by alternative methods,^{57,59} may stem from *RFX4* target genes, *NR2F1* and *NR2F2*, which mark GABAergic neurons^{60–62} (Figures 5L, 5M, and S4V; Tables S5A and S5C). Similarly, in human brain spheroids, *RFX4* expression was associated with GABAergic neurons.⁶³

Intriguingly, the candidate TF *EOMES* generated induced cardiomyocytes (iCMs; Figures S5E–S5M; Data S1.6H and S1.6I). We capitalized on this serendipitous finding to demonstrate sequential TF screening, which mirrors the successive upregulation of TFs during development toward more mature cell types. We used flow-FISH to screen 54 TFs in *RFX4*-iNPs for astrocytes and 80 TFs in *EOMES*-iCMs for atrial, ventricular, and mature iCMs (STAR Methods). Candidate TFs included those known to produce the target cell type (e.g., *NOTCH2* for astrocytes⁶⁴ and *TBX5* and *GATA4* for ventricular iCMs^{65,66}), as well as developmentally critical TFs (e.g., *SHOX2* and *NR2F2* for atrial iCMs^{67,68} and *GATA6* for ventricular and mature iCMs⁶⁹), supporting the screening results (Figures S5N and S5O; Tables S1C and S1D). Furthermore, the screens nominated TFs for further study, including *TRPS1* for astrocytes, *FOS* for ventricular and mature iCMs, and *MITF* for mature iCMs.

To explore the utility of *RFX4*-iNPs for modeling neurological disorders, we evaluated the effects of *DYRK1A* perturbation in this model. *DYRK1A* knockout and overexpression have been implicated in autism spectrum disorder^{70,71} and Down syndrome,⁷² respectively. Bulk RNA-seq identified 42 genes, including those involved in neuronal migration and synapse formation, that were expressed in a *DYRK1A* dosage-dependent manner (Figures S5P–S5U; Table S5B). During spontaneous differentiation, *DYRK1A* knockout increased, whereas *DYRK1A* overexpression decreased, the proportion of proliferating iNPs (Figures 5N and 5O). Interestingly, *DYRK1A* perturbations ultimately reduced neurogenesis: when knocked out, increased iNP proliferation deterred neurogenesis (Figure 5P; Data S1.6J), whereas when overexpressed, there were fewer iNPs due to lower initial proliferation (Figure 5Q). Electrophysiological characterization of spontaneously differentiated neurons showed that both *DYRK1A* knockout and overexpression reduced the proportion of mature neurons (Figures S5V–S5X; Data S1.6K). Our findings are consistent with previous *DYRK1A* studies in other model systems^{73–76} and provide additional

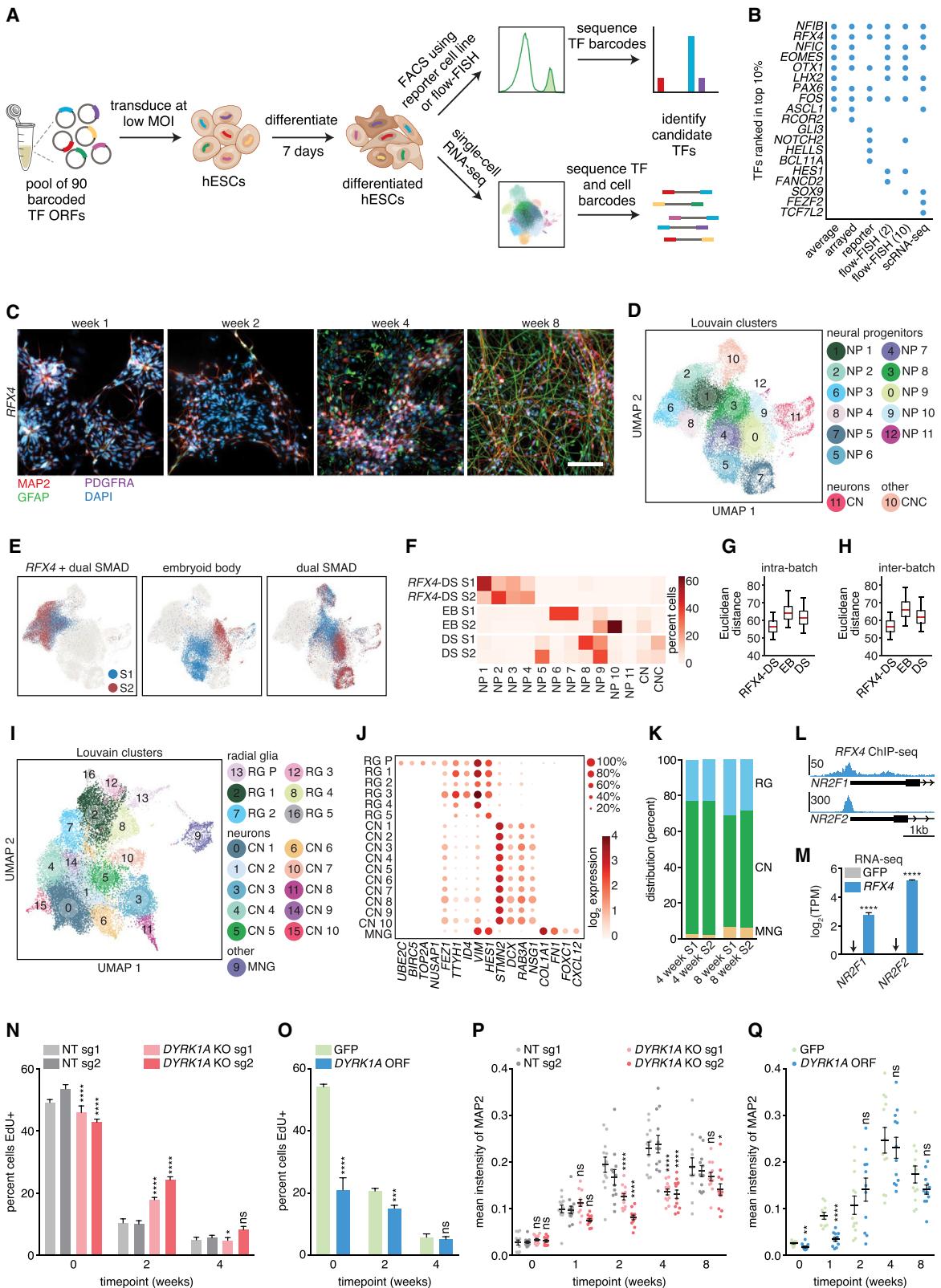
Figure 4. Validation of candidate TFs for differentiation toward nominated cell types

(A) Expression of marker genes for each cell type measured by quantitative PCR in H1 hESCs after 7 days of TF ORF or GFP overexpression. Numbers after TF gene names indicate the isoform. N = 4.

(B and C) Scatterplot comparing expression of all marker genes (205 from Figures 4A, S3L, and S3M) in H1 hESCs to H9 hESCs (B) or 11a iPSCs (C). Expression is measured relative to GFP control. Pearson correlation coefficients (r) and associated p values are shown.

(D–K) Left, expression of marker genes measured by immunostaining in H1 hESCs after 7 days of TF ORF overexpression. Right, intensity of marker gene staining normalized to GFP control from n = 6 images. Scale bars, 25 μm. Marker genes for neuron (D), EMT smooth muscle (E), endothelial (F), smooth muscle (G), metanephric (H), intestinal epithelial (I), lung ciliated epithelial (J), and trophoblast (K) cells are shown. EMT, epithelial-mesenchymal transition. Values represent mean ± SEM. ***p < 0.0001; **p < 0.001; *p < 0.01; *p < 0.05.

See also Figure S3.



(legend on next page)

insight. By combining cellular and genome engineering, we have outlined a versatile approach using the MORF library to create various cell types for studying development and disease.

Discovery of TF regulatory networks by joint profiling of chromatin accessibility and expression

As TFs often alter chromatin state to regulate gene expression, integrating single-cell chromatin accessibility with expression using SHARE-seq²⁰ provides an additional layer of information for deciphering the complex TF regulatory network. We selected 198 TFs that induced representative gene programs (Figure 2A), introduced the TF library into hESCs and, after 4 and 7 days, performed SHARE-seq (average of 3,317 and 2,384 UMIs per cell for scATAC-seq and scRNA-seq, respectively; Figures S6A–S6D). We constructed a weighted nearest neighbor (WNN) graph that integrated the ATAC and RNA profiles into a single representation for joint cluster analysis⁷⁷ (Figures 6A, 6B, and S6E–S6H; STAR Methods).

Based on the joint profiles, we identified GRNs by nominating TFs downstream of each TF ORF. Specifically, for each cluster, we matched top TF ORFs with putative downstream TFs that had enriched expression and motif accessibility (STAR Methods). GRNs of TF ORFs were consistent with their induced cell types (e.g., *GRHL1* and *GRHL3* targeted *TFAP2C* and *TEAD* family TFs to induce trophoblasts,⁷⁸ and *FLI1* targeted AP-1 family TFs [*JUN* and *FOS*] and *ETV2* to induce vascular endothelial cells⁷⁹), as well as their roles in development (e.g., *CDX1*, *CDX2*, and *HOXD11* targeted posterior *HOX* genes to specify the anterior-posterior axis⁸⁰) (Figure 6C). For 18 TF ORFs, such as *KLF5*, *FIGLA*, *MSGN1*, and *ATOH7*, the endogenous TF itself was nominated as the putative downstream TF, suggesting a positive feedback mechanism that enhances TF expression. The MORF library design allows this distinction because the ORF sequence is too distant (>1 kb) from the 3' end of the transcript to be captured by 3' scRNA-seq (Figures S1E and S2E). A complementary approach to identify downstream TFs based on motif enrichment in functional ATAC peaks yielded similar relationships (Figure S6I; STAR Methods).

We further investigated three GRNs: GRN 8 (*CDX1*, *CDX2*, and *HOXD11* regulation of *CDX4* and posterior *HOX* genes), GRN 3

(*ASCL2*-4 and *NHLH1* regulation of *FOX* genes), and GRN 6 (*MSGN1*, *MESP1*, and *MESP2* regulation of *GATA2*-6, *MEIS2*, *TWIST1*, and *SNAI1*-3). In the corresponding clusters of each GRN, the promoter regions of the respective downstream TFs were more accessible (Figure 6D). To validate the TF relationships, we examined DEGs of each TF ORF using our TF Atlas. In theory, the set of DEGs of an upstream TF should include those of downstream TFs, but not vice versa (Figure 6E). Our DEG analysis confirmed the GRN relationships (Figure 6F), highlighting how TF regulatory networks can be deciphered with joint profiling, providing a systems-level understanding of gene regulation.

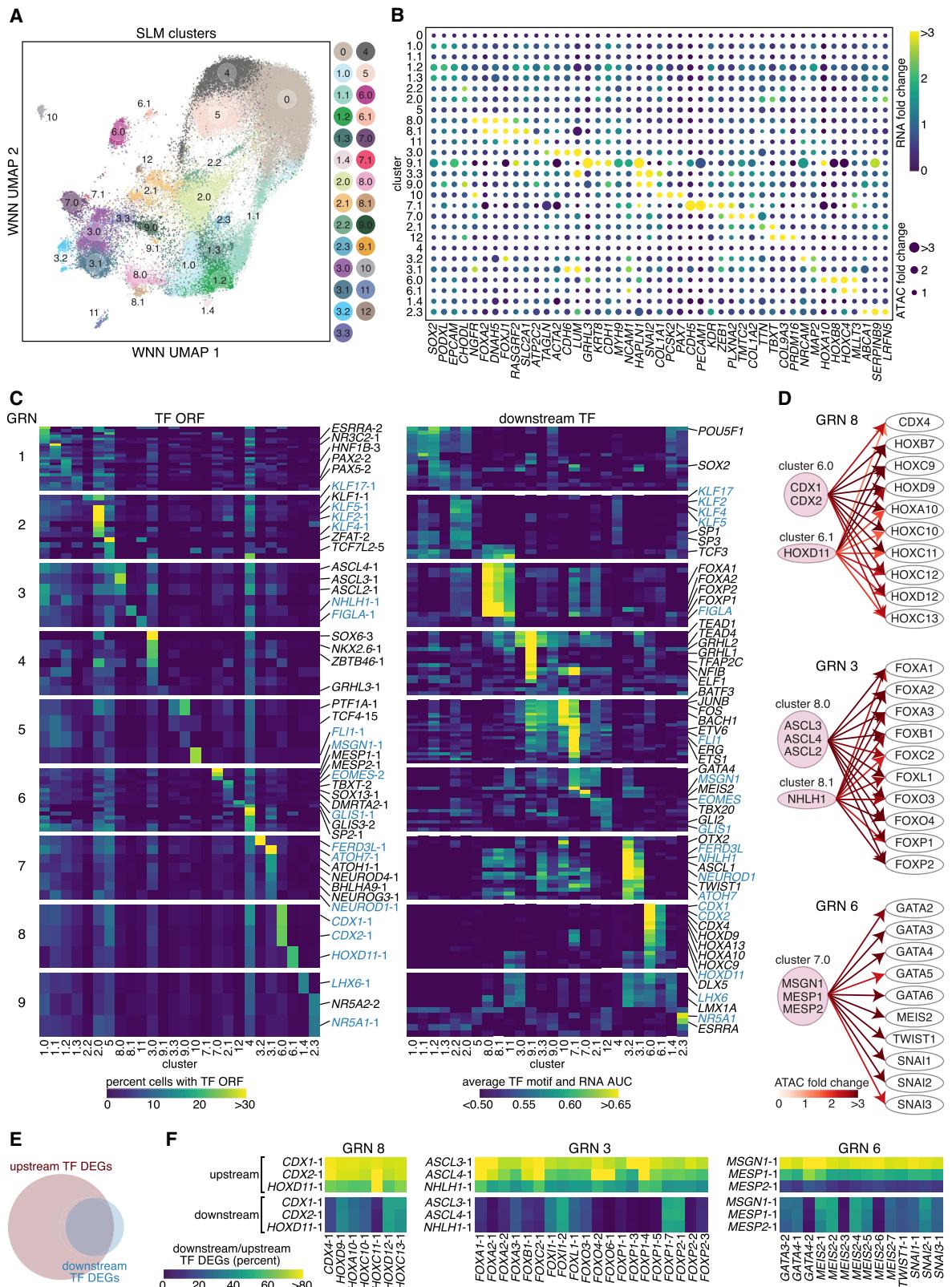
Combinatorial TF screening and prediction

Finally, as TFs often act in combination, we explored how TF ORFs combine to produce the resulting expression state. To model this, we first generated a scRNA-seq dataset for 10 TF ORFs in combinations, including 44 doubles and 3 triples, as well as 10 singles. Low dimensionality embedding and cluster analysis showed that expression profiles of combinations with similar TFs often grouped together, sometimes with the single TF profile of one member of the respective pair (e.g., *CDX1*, *FLI1*, and *KLF4*; Figures 7A and S7A; Data S1.7A and S1.7B). In other cases, two TFs generated a continuum (e.g., *FERD3L* and *NR5A2*). We quantitatively modeled TF interactions by a linear regression model ($ab = c_1 * a + c_2 * b + c_3 * a * b$) that fits the double TF expression profiles (*ab*) as a linear combination of the respective single TF profiles (*a* and *b*) along with an interaction term (*a * b*)^{81,82} (Figures S7B–S7D). The model suggested that although most TF combinations were overall additive, some TFs tended to be buffering (*PTF1A*), synergistic (*FLI1*), or dominant (*CDX1*; Figures S7E and S7F).

We then used the combinatorial dataset to nominate TF combinations that could produce a measured combinatorial expression profile. We ranked TF combinations based on how well their respective single profiles combine to fit the measured combinatorial profile (STAR Methods). We tested different approaches for combining, including averaging or using linear and nonlinear regression methods. As the baseline, we randomly selected TF combinations from the same set of possible combinations (i.e.,

Figure 5. Targeted TF overexpression screening platform for directed differentiation

- (A) Schematic of targeted TF screening. MOI, multiplicity of infection.
 - (B) Comparison of TF ranks from 5 iNP differentiation screens.
 - (C) Expression of markers for neurons (MAP2), astrocytes (GFAP), and oligodendrocyte precursor cells (PDGFRA) after spontaneous differentiation from *RFX4*-iNPs. Scale bars, 100 μm.
 - (D–H) scRNA-seq profiling of iNPs differentiated using different methods. EB, embryoid body; DS, dual SMAD; NP, neural progenitor; CN, CNS neuron; CNC, cranial neural crest. Data represent n = 2 batch replicates with 15,211 *RFX4*-DS, 11,148 EB, and 16,421 DS. (D and E) UMAP of scRNA-seq data with colors indicating Louvain clusters (D) or batch replicates (E). (F) Heatmap showing the percentage of cells from each replicate in each cluster. (G and H) Box plots showing intra- (G) or inter- (H) batch Euclidean distances between cells. Whiskers indicate the 5th and 95th percentiles.
 - (I–K) scRNA-seq data from 26,111 cells spontaneously differentiated from *RFX4*-DS-iNPs for 4 or 8 weeks. Data represent n = 2 biological replicates. RG, radial glia; CN, CNS neuron; MNG, meninges; and P, proliferating cells. (I) UMAP with colors indicating Louvain clusters. (J) Dot plot showing marker genes for each cluster. Circle size and color indicate percentage and expression level, respectively. (K) Distribution of cell types produced by each replicate.
 - (L) *RFX4* ChIP-seq reads at *NR2F1* and *NR2F2* promoter regions.
 - (M) Expression of *NR2F1* and *NR2F2* measured by bulk RNA-seq after 7 days of *RFX4* or GFP overexpression.
 - (N–Q) Modeling effects of *DYRK1A* perturbation in *RFX4*-iNPs derived from 11a iPSCs. (N and O) Percentage of EdU labeled cells after spontaneous differentiation for *DYRK1A* knockout (N) or overexpression (O). N = 3. (P and Q) Intensity of MAP2 staining for neurons after spontaneous differentiation for *DYRK1A* knockout (P) or overexpression (Q). N = 12 images. Values represent mean ± SEM. KO, knockout; NT, non-targeting; sg, single guide RNA. ***p < 0.0001; **p < 0.001; *p < 0.01; ns, not significant.
- See also Figures S4 and S5.



(legend on next page)

45 combinations for doubles and 120 combinations for triples). Surprisingly, simply computing the average outperformed regression-based approaches (Figures 7B and 7C), potentially because most TF combinations were additive (Figure S7E). For double TF profiles, averaging achieved an accuracy of 81% when evaluating only the top TF combination and 91% when evaluating the top 10% of possible combinations (i.e., top 4 of the 45 total combinations; Figure 7B). For triple TF profiles, averaging correctly predicted all 3 sets of TFs when evaluating the top ~2% of possible combinations (Figure 7C). Furthermore, we could still predict TF combinations by using single TF profiles from the TF Atlas, though with lower accuracy (Figures S7G and S7H; Data S1.7C–S1.7F).

We applied our findings from the combinatorial dataset to develop an approach for nominating TF combinations that could differentiate hESCs into more mature cell types. First, we averaged single TF profiles from the TF Atlas to estimate combinatorial TF profiles. We then scored each combinatorial TF profile for enrichment of reference cell type expression signatures^{30,83} and ranked potential TF combinations. We confirmed that our approach enriched for experimentally validated double TF combinations for the respective cell types, including hepatoblasts (*HNF4A* and *FOXA1*⁸⁴), astrocytes (*SOX9* and *NFIB*⁸⁵), and inhibitory neurons (*ASCL1* and *DLX2*⁸⁶) (Figure 7D). Moreover, the top predicted combinations included TFs that have been individually validated. For instance, combinations included *NHLH1* for bronchiolar and alveolar epithelial cells, *CDX1* for metanephric cells, and *GRHL3* for trophoblasts and suggested that other TFs that could improve differentiation efficiency and fidelity (Figures 7E–7I; Tables S6A and S6B). Furthermore, predicted combinations included TFs that are crucial for establishing the respective cell type during development, supporting the prediction results. Examples include *KLF6* for hepatocytes⁸⁷ (Figure 7E) and *ERG* for vascular endothelial cells⁸⁸ (Figure 7G). Similarly, our approach enriched for triple TF combinations that were experimentally validated and developmentally relevant (Figures 7J and S7I–S7M; Tables S6C and S6D).

To validate our prediction approach, we experimentally tested 12 predicted TF combinations for 3 cell types. Most combinations (11 of 12) induced expression of known marker genes for the target cell type (Figures 7K and 7L). Of these, 9 induced higher expression of at least 2 marker genes compared with the single TF, suggesting that the TF combination may produce

target cell types with increased fidelity or efficiency. Combining TFs also provides an opportunity to fine-tune induced cell types. Of the 4 TF combinations that included *ERG* and produced vascular endothelial-like cells expressing canonical marker genes (*CDH5*, *PECAM1*, and *KDR*), the combination with *MIXL1* did not express *CD34*, suggesting that *MIXL1* directs differentiation toward non-hemogenic endothelium⁸⁹ (Figure 7K). Thus, our results illustrate an approach to reduce the exponentially vast search space of combinatorial TF effects for follow-up empirical experimentation, accelerating the pace of cellular engineering.

DISCUSSION

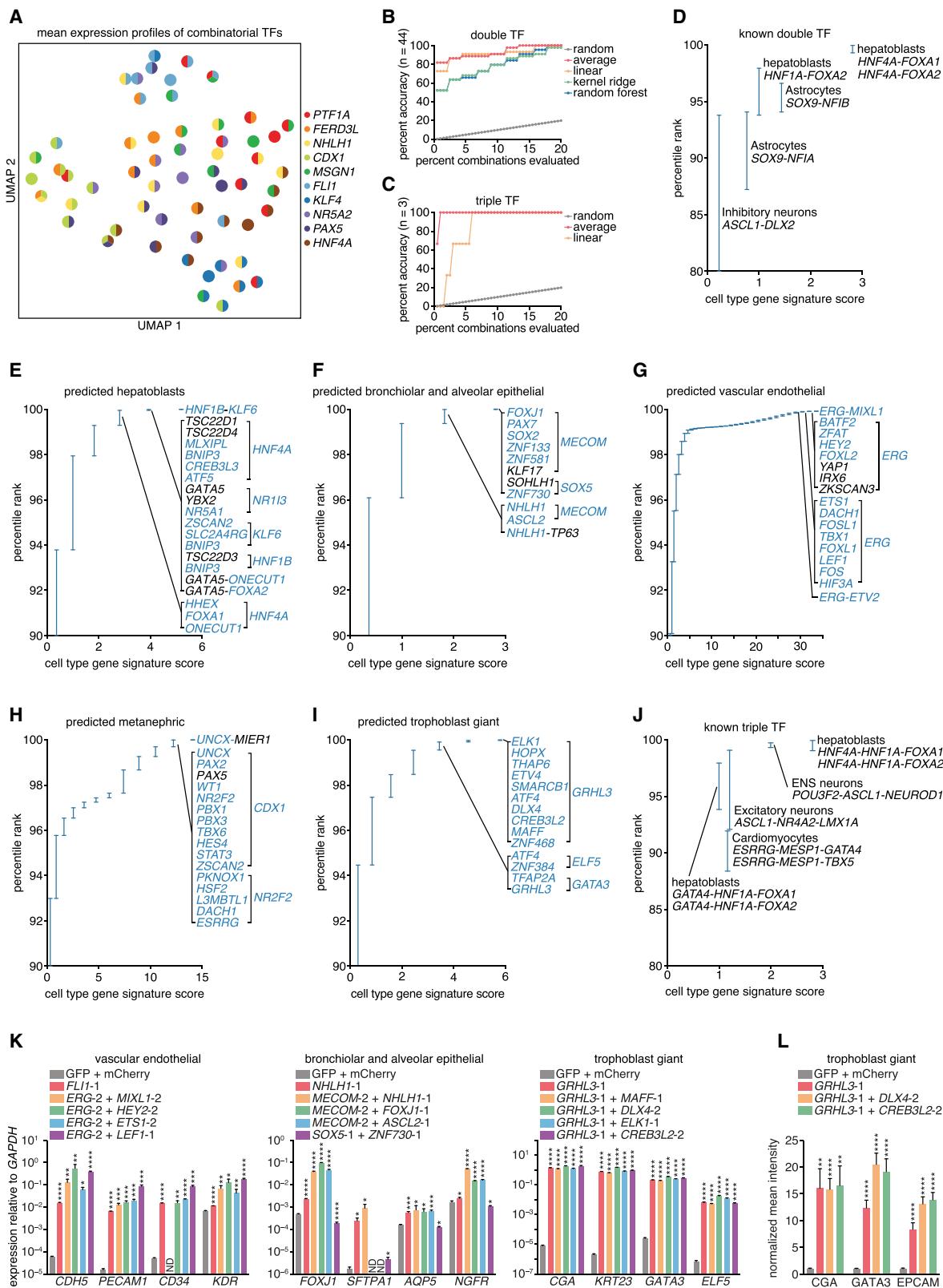
To achieve a comprehensive understanding of gene regulation, we developed a systematic, flexible approach for TF overexpression and characterization. We created a library of all human TF splice isoforms and built a TF Atlas that maps TF overexpression to corresponding expression changes. The TF Atlas enabled systematic identification of TFs that drive profound cell state changes, including production of cell types from all three germ layers and trophoblasts, as well as broad-spectrum findings, such as classification of orphan TFs. We used our library for targeted and sequential TF screens to establish cellular disease models. As only a small number of expressed TFs induced differentiation (e.g., 4 out of 90 for iNPs), TF screening facilitates efficient identification of TFs. We extensively profiled *RFX4*-iNPs and *EOMES*-iCMs to show that TF overexpression could lock cells in defined fates. We also integrated expression and chromatin accessibility data to characterize TF regulatory networks. Finally, we modeled the effects of TF ORF combinations using a combinatorial dataset and developed an approach for predicting TF combinations for reference cell types, which can help reduce the combinatorial search space.

The accessibility and flexibility of our screening approach lends itself to scalable extensions of the technology to additional contexts. For example, increasing the screening multiplicity of infection (MOI) to simultaneously overexpress multiple TFs could identify combinations of TFs that drive specific phenotypes. Incorporating TF target analysis and combinatorial prediction with TF screening may further facilitate derivation of complex cell types. Furthermore, TF screening can be applied to study

Figure 6. Discovery of TF regulatory networks by joint profiling of chromatin accessibility and expression

- (A) Weighted nearest neighbor (WNN) UMAP of joint chromatin accessibility and expression profiles from 69,085 cells overexpressing 198 TFs for 4 or 7 days. Colors indicate smart local moving (SLM) clusters.
- (B) Dot plot showing marker genes for each cluster. Color and circle size indicate expression level and chromatin accessibility, respectively, relative to other clusters.
- (C) Heatmaps showing gene regulatory networks (GRNs) containing the top TF ORFs (left) and nominated downstream TFs (right) for each cluster. Left, percentage of cells with the indicated TF ORF. Numbers after TF gene names indicate the isoform. Percentages are normalized to the total number of cells with the TF ORF. Only the 6 most enriched TF ORFs >5% are shown for each cluster. Right, average area under the ROC curve (AUC) of TF motif enrichment and RNA expression is shown for significantly enriched (FDR < 0.05) TFs. TFs that were identified as top ORFs and downstream TFs are labeled in blue.
- (D) Examples of GRNs identified by matching the top TF ORFs (left) with nominated downstream TFs (right). Color of arrows indicates chromatin accessibility of the downstream TF promoter region relative to cluster 0.
- (E) Schematic showing the relation between differentially expressed genes (DEGs) induced by upstream and downstream TFs.
- (F) Heatmaps showing the percentage of downstream TF DEGs included in upstream TF DEGs using TFs in each row as upstream (top) or downstream (bottom) relative to TFs in each column.

See also Figure S6.



(legend on next page)

trans-differentiation,^{8,66} aging,⁹⁰ and cancer.^{91,92} Moreover, our ORF barcoding approach allows for a variety of screening selection methods and could be extended to pooled ORF screening of other protein families.

Future applications of our MORF library in other contexts will illuminate factors driving nearly any cellular phenotype of interest. Similarly, as single-cell profiling becomes more affordable, we anticipate that the resolution of this TF Atlas will increase. Our MORF library and TF Atlas provide valuable resources that lay the foundation for deciphering TF circuits toward a comprehensive understanding of GRNs that govern cell states.

Limitations of the study

Some conclusions of our study were limited by technical aspects of our experimental design. As we were constrained by sequencing cost, we chose one starting cell type, time point, and media condition for our TF Atlas. Expression changes induced by TF overexpression may depend on the cell type. We chose a time point of 7 days to obtain more differentiated cells for mapping to reference cell types, which precluded identification of immediate TF target genes. Future studies at earlier time points will enable mapping of direct TF target genes. Although we showed that cell culture media did not strongly alter TF-induced differentiation outcome at 7 days, exogenous factors and media conditions may influence differentiation of more mature cell types. For the TF Atlas cell types that we mapped and validated, further functional characterization, as we have done for *RFX4*-iNPs, is necessary. Finally, our approach for predicting TF combinations assumes that TF effects are additive, which is often but not necessarily always the case. More complex integration approaches will increase the precision of TF combination modeling.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS

- Cell culture and differentiation
- Lentivirus production
- Lentivirus transduction

● METHOD DETAILS

- Sequences and cloning
- qPCR quantification of transcript expression
- Western blot
- Immunofluorescence and imaging
- Design and cloning of TF ORF libraries
- Reporter cell line screen
- Flow-FISH screen
- 10X single cell RNA sequencing (scRNA-seq) screen
- 10X scRNA-seq library preparation and sequencing
- Arrayed screen
- Bulk MORF library screen
- Bulk RNA sequencing (RNA-seq)
- ChIP-seq analysis with sequencing (ChIP-seq)
- Indel sequencing
- Flow cytometry assays
- Electrophysiology
- TF Atlas SHARE-seq library preparation

● QUANTIFICATION AND STATISTICAL ANALYSIS

- Image quantification
- 10X scRNA-seq analysis
- Bulk RNA-seq analysis
- ChIP-seq analysis
- Indel analysis
- Electrophysiology analysis
- SHARE-seq data preprocessing
- Pseudotime analysis
- Non-negative matrix factorization
- Pathway enrichment analysis
- TF potential analysis
- Cell type mapping
- Joint chromatin accessibility and gene expression analysis
- Combinatorial TF prediction
- Statistics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cell.2022.11.026>.

Figure 7. Combinatorial TF screening and prediction

- (A) UMAP of scRNA-seq profiles from the combinatorial screen of 10 TF ORFs in combinations, including 44 doubles and 3 triples, as well as 10 singles. Each circle represents the mean expression profile of cells with the indicated TF ORF(s).
- (B and C) Percent accuracy for different approaches to predict TFs for measured double (B) or triple (C) TF expression profiles.
- (D–I) Cell type prediction results for double TF profiles. Known combinations (D) or predicted combinations for hepatoblasts (E), bronchiolar and alveolar epithelial cells (F), metanephric cells (G), vascular endothelial cells (H), and trophoblast giant cells (I) are shown. As gene signature scores were discrete, the percentile ranks were reported as ranges. TFs that are part of known combinations, developmentally critical, or specifically expressed in the target cell types are indicated in blue.
- (J) Prediction results for known combinations of triple TF profiles. To expand the number of combinations, parts of known combinations with >3 TFs were included.
- (K and L) Marker gene expression for each cell type measured by quantitative PCR (K; n = 4) or immunostaining (L; n = 6 images) after 7 days of TF ORF or control overexpression. Numbers after TF gene names indicate the isoform. Intensity is normalized to GFP control. Values represent mean ± SEM. ***p < 0.0001; **p < 0.001; *p < 0.01; *p < 0.05; ND, not detected.

See also Figure S7.

ACKNOWLEDGMENTS

We thank E. Habibi, T. Kamath, E.Z. Macosko, and F. Chen for scRNA-seq analysis advice; K. Waddell, L. Amir-Zilberman, L. Nguyen, M.S. Cuoco, and D. Dionne for scRNA-seq library generation; O. Rozenblatt-Rosen for help with large-scale scRNA-seq design; J.G. Doench and D.E. Root for discussions on MORF library design; J.M. Engreitz and C.P. Fulco for flow-FISH advice; M.C. Marchetto and F.H. Gage for embryoid body (EB) protocols; R. Belliveau for overall research support; A. Tang for illustrations; and the entire Zhang laboratory for support and advice. J.J. is supported by fellowships from the McGovern Institute and NIH F31-MH117886. F.Z. is supported by NIH grants R01-HG009761 and DP1-HL141201, the Howard Hughes Medical Institute (HHMI), the Poitras Center for Psychiatric Disorders Research at MIT, the Hock E. Tan and K. Lisa Yang Center for Autism Research at MIT, the Yang-Tan Center for Molecular Therapeutics at MIT, and the Phillips family and J. and P. Poitras. A.R. was an HHMI Investigator while conducting this work and work was supported by the Klarman Cell Observatory (KCO), NHGRI Center of Excellence for Genome Science, and HHMI. J.D.B. is supported by an NIH grant DP2-HL151353, Gene Regulation Observatory, and Harvard Stem Cell Institute (BA-0002-21-00). Z.F. is supported by the Stanley Center for Psychiatric Research and an NIH grant R01-MH127085. The MORF library synthesis was funded by the Stanley Center for Psychiatric Research. TF Atlas scRNA-seq was funded by the KCO. Reagents are available through Addgene; support forums and computational tools are available via the Zhang lab website (<https://zlab.bio/>).

AUTHOR CONTRIBUTIONS

J.J. and F.Z. conceived the study. J.J. designed all experiments. J.J., P.C.K., V.K.V., and A.S. performed the experiments. S.M., T.T., K.R.G.-S., and O.K. prepared SHARE-seq libraries under the supervision of A.R. and J.D.B. B.G. and M.A.A.-G. performed and analyzed the electrophysiological recordings under the supervision of Z.F. J.J. analyzed data with assistance from S.M., T.T., and W.E.A. and guidance from A.R. O.O.A. performed ChIP-seq motif analysis. J.J., P.C.K., and J.S.G. quantified immunofluorescence images. F.Z. supervised the research and experimental design with support from R.K.M. J.J., R.K.M., A.R., and F.Z. wrote the paper with input from all authors.

DECLARATION OF INTERESTS

J.J. and F.Z. are inventors listed on an International PCT Application related to this work. F.Z. is a scientific advisor and co-founder of Editas Medicine, Beam Therapeutics, Pairwise Plants, Arbor Biotechnologies, and Proof Diagnostics. F.Z. is a scientific advisor for Octant. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov until 31 July 2020. Since 1 August 2020, A.R. has been an employee of Genentech and has equity in Roche. A.R. is an inventor on patents and patent applications filed at the Board related to single-cell genomics. J.D.B. holds patents related to ATAC-seq and scATAC-seq and serves on the scientific advisory boards of CAMP4 Therapeutics, seqWell, and CelSee. J.S.G. and O.O.A. are cofounders of Sherlock Biosciences, Proof Diagnostics, Moment Biosciences, and Tome Biosciences. Since 16 November 2020, K.R.G.-S. has been an employee of Genentech.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research.

Received: February 3, 2022

Revised: August 4, 2022

Accepted: November 23, 2022

Published: January 5, 2023

Corrected online: April 29, 2024

REFERENCES

- Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. *Cell* 172, 650–665. <https://doi.org/10.1016/j.cell.2018.01.029>.
- Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326, 257–263. <https://doi.org/10.1126/science.1179050>.
- Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752. <https://doi.org/10.1016/j.cell.2010.01.044>.
- Ng, A.H.M., Khoshakhlagh, P., Rojo Arias, J.E., Pasquini, G., Wang, K., Swiersy, A., Shipman, S.L., Appleton, E., Kiaee, K., Kohman, R.E., et al. (2021). A comprehensive library of human transcription factors for cell fate engineering. *Nat. Biotechnol.* 39, 510–519. <https://doi.org/10.1038/s41587-020-0742-6>.
- Parekh, U., Wu, Y., Zhao, D., Worlikar, A., Shah, N., Zhang, K., and Mali, P. (2018). Mapping cellular reprogramming via pooled overexpression screens with paired fitness and single-cell RNA-sequencing readout. *Cell Syst.* 7, 548–555.e8. <https://doi.org/10.1016/j.cels.2018.10.008>.
- Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B., and Miller, A.D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc. Natl. Acad. Sci. USA* 86, 5434–5438.
- Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* 78, 785–798. <https://doi.org/10.1016/j.neuron.2013.05.029>.
- Pang, Z.P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D.R., Yang, T.Q., Citri, A., Sebastian, V., Marro, S., Südhof, T.C., and Wernig, M. (2011). Induction of human neuronal cells by defined transcription factors. *Nature* 476, 220–223. <https://doi.org/10.1038/nature10202>.
- Perrier, A.L., Tabar, V., Barberi, T., Rubio, M.E., Bruses, J., Topf, N., Harrison, N.L., and Studer, L. (2004). Derivation of midbrain dopamine neurons from human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* 101, 12543–12548. <https://doi.org/10.1073/pnas.0404700101>.
- Sugimura, R., Jha, D.K., Han, A., Soria-Valles, C., da Rocha, E.L., Lu, Y.F., Goettel, J.A., Serrao, E., Rowe, R.G., Malleshaiah, M., et al. (2017). Haematopoietic stem and progenitor cells from human pluripotent stem cells. *Nature* 545, 432–438. <https://doi.org/10.1038/nature22370>.
- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and 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>.
- Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318, 1917–1920. <https://doi.org/10.1126/science.1151526>.
- Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676. <https://doi.org/10.1016/j.cell.2006.07.024>.
- Yao, Z., Fong, A.P., Cao, Y., Ruzzo, W.L., Gentleman, R.C., and Tapscott, S.J. (2013). Comparison of endogenous and overexpressed MyoD shows enhanced binding of physiologically bound sites. *Skelet. Muscle* 3, 8. <https://doi.org/10.1186/2044-5040-3-8>.
- Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., and Church, G.M. (2015). Multiplex genome editing in human cells. *Nature* 523, 471–476. <https://doi.org/10.1038/nature14641>.

- H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* 517, 583–588. <https://doi.org/10.1038/nature14136>.
16. Dabney, J., and Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques* 52, 87–94. <https://doi.org/10.2144/000113809>.
 17. Hayashi, Y., Hsiao, E.C., Sami, S., Lancero, M., Schlieve, C.R., Nguyen, T., Yano, K., Nagahashi, A., Ikeya, M., Matsumoto, Y., et al. (2016). BMP-SMAD-ID promotes reprogramming to pluripotency by inhibiting p16/INK4A-dependent senescence. *Proc. Natl. Acad. Sci. USA* 113, 13057–13062. <https://doi.org/10.1073/pnas.1603668113>.
 18. Zhao, W., Liu, M., Ji, H., Zhu, Y., Wang, C., Huang, Y., Ma, X., Xing, G., Xia, Y., Jiang, Q., and Qin, J. (2018). The polycomb group protein Yaf2 regulates the pluripotency of embryonic stem cells in a phosphorylation-dependent manner. *J. Biol. Chem.* 293, 12793–12804. <https://doi.org/10.1074/jbc.RA118.003299>.
 19. Daley, J.M., and Sung, P. (2014). 53BP1, BRCA1, and the choice between recombination and end joining at DNA double-strand breaks. *Mol. Cell. Biol.* 34, 1380–1388. <https://doi.org/10.1128/MCB.01639-13>.
 20. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 183, 1103–1116.e20. <https://doi.org/10.1016/j.cell.2020.09.056>.
 21. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. <https://doi.org/10.1038/nmeth.3971>.
 22. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerer, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. <https://doi.org/10.1038/s41586-018-0414-6>.
 23. Saba, J.A., Liakath-Ali, K., Green, R., and Watt, F.M. (2021). Translational control of stem cell function. *Nat. Rev. Mol. Cell Biol.* 22, 671–690. <https://doi.org/10.1038/s41580-021-00386-2>.
 24. Gaspar-Maia, A., Alajem, A., Meshorer, E., and Ramalho-Santos, M. (2011). Open chromatin in pluripotency and reprogramming. *Nat. Rev. Mol. Cell Biol.* 12, 36–47. <https://doi.org/10.1038/nrm3036>.
 25. Halluin, C., Madelaine, R., Naye, F., Peers, B., Roussigné, M., and Blader, P. (2016). Habenular neurogenesis in zebrafish is regulated by a hedgehog, Pax6 proneural gene cascade. *PLoS One* 11, e0158210. <https://doi.org/10.1371/journal.pone.0158210>.
 26. Bouchard, M., Pfeffer, P., and Busslinger, M. (2000). Functional equivalence of the transcription factors Pax2 and Pax5 in mouse development. *Development* 127, 3703–3713.
 27. Festuccia, N., Osorno, R., Halbritter, F., Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F., Yates, A., Tomlinson, S.R., and Chambers, I. (2012). Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. *Cell Stem Cell* 11, 477–490. <https://doi.org/10.1016/j.stem.2012.08.002>.
 28. McConnell, B.B., and Yang, V.W. (2010). Mammalian Kruppel-like factors in health and diseases. *Physiol. Rev.* 90, 1337–1381. <https://doi.org/10.1152/physrev.00058.2009>.
 29. Holland, P.W., Booth, H.A., and Bruford, E.A. (2007). Classification and nomenclature of all human homeobox genes. *BMC Biol.* 5, 47. <https://doi.org/10.1186/1741-7007-5-47>.
 30. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. *Science* 370, eaaba7721. <https://doi.org/10.1126/science.aba7721>.
 31. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>.
 32. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and 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>.
 33. Tan, Y., and Cahan, P. (2019). SingleCellNet: A computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* 9, 207–213.e2. <https://doi.org/10.1016/j.cels.2019.06.004>.
 34. Cahan, P., Li, H., Morris, S.A., Lumertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. *Cell* 158, 903–915. <https://doi.org/10.1016/j.cell.2014.07.020>.
 35. Ono, Y., Nakatani, T., Minaki, Y., and Kumai, M. (2010). The basic helix-loop-helix transcription factor Nato3 controls neurogenic activity in mesencephalic floor plate cells. *Development* 137, 1897–1906. <https://doi.org/10.1242/dev.042572>.
 36. Li, Y., Luo, H., Liu, T., Zackenhaus, E., and Ben-David, Y. (2015). The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene* 34, 2022–2031. <https://doi.org/10.1038/onc.2014.162>.
 37. Ghaleb, A.M., McConnell, B.B., Kaestner, K.H., and Yang, V.W. (2011). Altered intestinal epithelial homeostasis in mice with intestine-specific deletion of the Kruppel-like factor 4 gene. *Dev. Biol.* 349, 310–320. <https://doi.org/10.1016/j.ydbio.2010.11.001>.
 38. Du, Y., Wang, J., Jia, J., Song, N., Xiang, C., Xu, J., Hou, Z., Su, X., Liu, B., Jiang, T., et al. (2014). Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* 14, 394–403. <https://doi.org/10.1016/j.stem.2014.01.008>.
 39. Huang, P., Zhang, L., Gao, Y., He, Z., Yao, D., Wu, Z., Cen, J., Chen, X., Liu, C., Hu, Y., et al. (2014). Direct reprogramming of human fibroblasts to functional and expandable hepatocytes. *Cell Stem Cell* 14, 370–384. <https://doi.org/10.1016/j.stem.2014.01.003>.
 40. Hoang, C.Q., Hale, M.A., Azevedo-Pouly, A.C., Elsässer, H.P., Deering, T.G., Willet, S.G., Pan, F.C., Magnuson, M.A., Wright, C.V., Swift, G.H., and MacDonald, R.J. (2016). Transcriptional maintenance of pancreatic acinar identity, differentiation, and homeostasis by PTF1A. *Mol. Cell. Biol.* 36, 3033–3047. <https://doi.org/10.1128/MCB.00358-16>.
 41. Basu, P., Morris, P.E., Haar, J.L., Wani, M.A., Lingrel, J.B., Gaensler, K.M., and Lloyd, J.A. (2005). KLF2 is essential for primitive erythropoiesis and regulates the human and murine embryonic beta-like globin genes in vivo. *Blood* 106, 2566–2571. <https://doi.org/10.1182/blood-2005-02-0674>.
 42. Brown, N.L., Patel, S., Brzezinski, J., and Glaser, T. (2001). Math5 is required for retinal ganglion cell and optic nerve formation. *Development* 128, 2497–2508. <https://doi.org/10.1242/dev.128.13.2497>.
 43. Feng, L., Xie, X., Joshi, P.S., Yang, Z., Shibasaki, K., Chow, R.L., and Gan, L. (2006). Requirement for Bhlhb5 in the specification of amacrine and cone bipolar subtypes in mouse retina. *Development* 133, 4815–4825. <https://doi.org/10.1242/dev.02664>.
 44. Ross, S.E., Mardini, A.R., McCord, A.E., Zurawski, J., Cohen, S., Jung, C., Hu, L., Mok, S.I., Shah, A., Savner, E.M., et al. (2010). Loss of inhibitory interneurons in the dorsal spinal cord and elevated itch in Bhlhb5 mutant mice. *Neuron* 65, 886–898. <https://doi.org/10.1016/j.neuron.2010.02.025>.
 45. Oliver, G. (2004). Lymphatic vasculature development. *Nat. Rev. Immunol.* 4, 35–45. <https://doi.org/10.1038/nri1258>.
 46. Ginhoux, F., Lim, S., Hoeffel, G., Low, D., and Huber, T. (2013). Origin and differentiation of microglia. *Front. Cell. Neurosci.* 7, 45. <https://doi.org/10.3389/fncel.2013.00045>.
 47. Pan, F.C., and Brissova, M. (2014). Pancreas development in humans. *Curr. Opin. Endocrinol. Diabetes Obes.* 21, 77–82. <https://doi.org/10.1097/MED.0000000000000047>.
 48. Barnes, R.M., Firulli, B.A., Conway, S.J., Vincentz, J.W., and Firulli, A.B. (2010). Analysis of the Hand1 cell lineage reveals novel contributions to cardiovascular, neural crest, extra-embryonic, and lateral mesoderm

- derivatives. *Dev. Dyn.* 239, 3086–3097. <https://doi.org/10.1002/dvdy.22428>.
49. Yamaguchi, T.P., Takada, S., Yoshikawa, Y., Wu, N., and McMahon, A.P. (1999). *T*(brachyury) is a direct target of *Wnt3a* during paraxial mesoderm specification. *Genes Dev.* 13, 3185–3190. <https://doi.org/10.1101/gad.13.24.3185>.
 50. Chalamalasetty, R.B., Garriock, R.J., Dunty, W.C., Jr., Kennedy, M.W., Jainwala, P., Si, H., and Yamaguchi, T.P. (2014). Mesogenin 1 is a master regulator of paraxial presomitic mesoderm differentiation. *Development* 141, 4285–4297. <https://doi.org/10.1242/dev.110908>.
 51. Ramirez, J.M., Gerbal-Chaloin, S., Milhavet, O., Qiang, B., Becker, F., Assou, S., Lemaître, J.M., Hamamah, S., and De Vos, J. (2011). Brief report: benchmarking human pluripotent stem cell markers during differentiation into the three germ layers unveils a striking heterogeneity: all markers are not equal. *Stem Cells* 29, 1469–1474. <https://doi.org/10.1002/stem.681>.
 52. Robinton, D.A., and Daley, G.Q. (2012). The promise of induced pluripotent stem cells in research and therapy. *Nature* 481, 295–305. <https://doi.org/10.1038/nature10761>.
 53. Wu, S.M., and Hochedlinger, K. (2011). Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nat. Cell Biol.* 13, 497–505. <https://doi.org/10.1038/ncb0511-497>.
 54. Steele-Perkins, G., Plachez, C., Butz, K.G., Yang, G., Bachurski, C.J., Kinsman, S.L., Litwack, E.D., Richards, L.J., and Gronostajski, R.M. (2005). The transcription factor gene *Nfib* is essential for both lung maturation and brain development. *Mol. Cell. Biol.* 25, 685–698. <https://doi.org/10.1128/MCB.25.2.685-698.2005>.
 55. Götz, M., Stoykova, A., and Gruss, P. (1998). *Pax6* controls radial glia differentiation in the cerebral cortex. *Neuron* 21, 1031–1044.
 56. Casarosa, S., Fode, C., and Guillemot, F. (1999). *Mash1* regulates neurogenesis in the ventral telencephalon. *Development* 126, 525–534.
 57. Schafer, S.T., Paquola, A.C.M., Stern, S., Gosselin, D., Ku, M., Pena, M., Kuret, T.J.M., Liyanage, M., Mansour, A.A., Jaeger, B.N., et al. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. *Nat. Neurosci.* 22, 243–255. <https://doi.org/10.1038/s41593-018-0295-x>.
 58. Shi, Y., Kirwan, P., and Livesey, F.J. (2012). Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat. Protoc.* 7, 1836–1846. <https://doi.org/10.1038/nprot.2012.116>.
 59. Shi, Y., Kirwan, P., Smith, J., Robinson, H.P., and Livesey, F.J. (2012). Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nat. Neurosci.* 15, 477–486. <https://doi.org/10.1038/nn.3041>.
 60. Delgado, R.N., Allen, D.E., Keefe, M.G., Mancia Leon, W.R., Ziffra, R.S., Crouch, E.E., Alvarez-Buylla, A., and Nowakowski, T.J. (2022). Individual human cortical progenitors can produce excitatory and inhibitory neurons. *Nature* 601, 397–403. <https://doi.org/10.1038/s41586-021-04230-7>.
 61. Mayer, C., Hafemeister, C., Bandler, R.C., Machold, R., Batista Brito, R., Jaglin, X., Allaway, K., Butler, A., Fishell, G., and Satija, R. (2018). Developmental diversification of cortical inhibitory interneurons. *Nature* 555, 457–462. <https://doi.org/10.1038/nature25999>.
 62. Reinchisi, G., Ijichi, K., Glidden, N., Jakovcevski, I., and Zecevic, N. (2012). COUP-TFII expressing interneurons in human fetal forebrain. *Cereb. Cortex* 22, 2820–2830. <https://doi.org/10.1093/cercor/bhr359>.
 63. Trevino, A.E., Sinnott-Armstrong, N., Andersen, J., Yoon, S.J., Huber, N., Pritchard, J.K., Chang, H.Y., Greenleaf, W.J., and Pasca, S.P. (2020). Chromatin accessibility dynamics in a model of human forebrain development. *Science* 367, eaay1645. <https://doi.org/10.1126/science.aay1645>.
 64. Tchorz, J.S., Tome, M., Cloëtta, D., Sivasankaran, B., Grzmil, M., Huber, R.M., Rutz-Schatzmann, F., Kirchhoff, F., Schaeren-Wiemers, N., Gassmann, M., et al. (2012). Constitutive Notch2 signaling in neural stem cells promotes tumorigenic features and astroglial lineage entry. *Cell Death Dis.* 3, e325. <https://doi.org/10.1038/cddis.2012.65>.
 65. Ieda, M., Fu, J.D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Brunneau, B.G., and 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>.
 66. Song, K., Nam, Y.J., Luo, X., Qi, X., Tan, W., Huang, G.N., Acharya, A., Smith, C.L., Tallquist, M.D., Neilson, E.G., et al. (2012). Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature* 485, 599–604. <https://doi.org/10.1038/nature11139>.
 67. Espinoza-Lewis, R.A., Yu, L., He, F., Liu, H., Tang, R., Shi, J., Sun, X., Martin, J.F., Wang, D., Yang, J., and Chen, Y. (2009). *Shox2* is essential for the differentiation of cardiac pacemaker cells by repressing *Nkx2.5*. *Dev. Biol.* 327, 376–385. <https://doi.org/10.1016/j.ydbio.2008.12.028>.
 68. Wu, S.P., Cheng, C.M., Lanz, R.B., Wang, T., Respress, J.L., Ather, S., Chen, W., Tsai, S.J., Wehrens, X.H., Tsai, M.J., and Tsai, S.Y. (2013). Atrial identity is determined by a COUP-TFII regulatory network. *Dev. Cell* 25, 417–426. <https://doi.org/10.1016/j.devcel.2013.04.017>.
 69. Xin, M., Davis, C.A., Molkentin, J.D., Lien, C.L., Duncan, S.A., Richardson, J.A., and Olson, E.N. (2006). A threshold of GATA4 and GATA6 expression is required for cardiovascular development. *Proc. Natl. Acad. Sci. USA* 103, 11189–11194. <https://doi.org/10.1073/pnas.0604103>.
 70. De Rubeis, S., He, X., Goldberg, A.P., Poultnay, C.S., Samocha, K., Ciccek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209–215. <https://doi.org/10.1038/nature13772>.
 71. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Steessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216–221. <https://doi.org/10.1038/nature13908>.
 72. Smith, D.J., Stevens, M.E., Sudaganguta, S.P., Bronson, R.T., Makinson, M., Watabe, A.M., O'Dell, T.J., Fung, J., Weier, H.U., Cheng, J.F., and Rubin, E.M. (1997). Functional screening of 2 Mb of human chromosome 21q22.2 in transgenic mice implicates minibrain in learning defects associated with Down syndrome. *Nat. Genet.* 16, 28–36. <https://doi.org/10.1038/ng0597-28>.
 73. Fotaki, V., Dierssen, M., Alcántara, S., Martínez, S., Martí, E., Casas, C., Visa, J., Soriano, E., Estivill, X., and Arbonés, M.L. (2002). Dyrk1A haploinsufficiency affects viability and causes developmental delay and abnormal brain morphology in mice. *Mol. Cell. Biol.* 22, 6636–6647.
 74. Hämerle, B., Ulin, E., Guimera, J., Becker, W., Guillermot, F., and Tejedor, F.J. (2011). Transient expression of Mnb/Dyrk1a couples cell cycle exit and differentiation of neuronal precursors by inducing p27kip1 expression and suppressing NOTCH signaling. *Development* 138, 2543–2554. <https://doi.org/10.1242/dev.066167>.
 75. Park, J., Oh, Y., Yoo, L., Jung, M.S., Song, W.J., Lee, S.H., Seo, H., and Chung, K.C. (2010). Dyrk1A phosphorylates p53 and inhibits proliferation of embryonic neuronal cells. *J. Biol. Chem.* 285, 31895–31906. <https://doi.org/10.1074/jbc.M110.147520>.
 76. Yabut, O., Domogauer, J., and D'Arcangelo, G. (2010). Dyrk1A overexpression inhibits proliferation and induces premature neuronal differentiation of neural progenitor cells. *J. Neurosci.* 30, 4004–4014. <https://doi.org/10.1523/JNEUROSCI.4711-09.2010>.
 77. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. <https://doi.org/10.1016/j.cell.2021.04.048>.
 78. Krendl, C., Shaposhnikov, D., Rishko, V., Ori, C., Ziegenhain, C., Sass, S., Simon, L., Müller, N.S., Straub, T., Brooks, K.E., et al. (2017). GATA2/3-TFAP2A/C transcription factor network couples human pluripotent stem cell differentiation to trophectoderm with repression of

- pluripotency. *Proc. Natl. Acad. Sci. USA.* 114, E9579–E9588. <https://doi.org/10.1073/pnas.1708341114>.
79. Dejana, E., Taddei, A., and Randi, A.M. (2007). Foxs and ets in the transcriptional regulation of endothelial cell differentiation and angiogenesis. *Biochim. Biophys. Acta* 1775, 298–312. <https://doi.org/10.1016/j.bbcan.2007.05.003>.
 80. Neijts, R., Amin, S., van Rooijen, C., and Deschamps, J. (2017). Cdx is crucial for the timing mechanism driving colinear Hox activation and defines a trunk segment in the Hox cluster topology. *Dev. Biol.* 422, 146–154. <https://doi.org/10.1016/j.ydbio.2016.12.024>.
 81. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17. <https://doi.org/10.1016/j.cell.2016.11.038>.
 82. Capaldi, A.P., Kaplan, T., Liu, Y., Habib, N., Regev, A., Friedman, N., and O’Shea, E.K. (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.* 40, 1300–1306. <https://doi.org/10.1038/ng.235>.
 83. Cortal, A., Martignetti, L., Six, E., and Rausell, A. (2021). Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nat. Biotechnol.* 39, 1095–1102. <https://doi.org/10.1038/s41587-021-00896-6>.
 84. Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* 475, 390–393. <https://doi.org/10.1038/nature10263>.
 85. Canals, I., Ginisty, A., Quist, E., Timmerman, R., Fritze, J., Miskinyte, G., Monni, E., Hansen, M.G., Hidalgo, I., Bryder, D., et al. (2018). Rapid and efficient induction of functional astrocytes from human pluripotent stem cells. *Nat. Methods* 15, 693–696. <https://doi.org/10.1038/s41592-018-0103-2>.
 86. Yang, N., Chanda, S., Marro, S., Ng, Y.H., Janas, J.A., Haag, D., Ang, C.E., Tang, Y., Flores, Q., Mall, M., et al. (2017). Generation of pure GABAergic neurons by transcription factor programming. *Nat. Methods* 14, 621–628. <https://doi.org/10.1038/nmeth.4291>.
 87. Zhao, X., Monson, C., Gao, C., Gouon-Evans, V., Matsumoto, N., Sadler, K.C., and Friedman, S.L. (2010). Klf6/coopeb is required for hepatic outgrowth in zebrafish and for hepatocyte specification in mouse ES cells. *Dev. Biol.* 344, 79–93. <https://doi.org/10.1016/j.ydbio.2010.04.018>.
 88. Birdsey, G.M., Shah, A.V., Dufton, N., Reynolds, L.E., Osuna Almagro, L., Yang, Y., Aspalter, I.M., Khan, S.T., Mason, J.C., Dejana, E., et al. (2015). The endothelial transcription factor ERG promotes vascular stability and growth through Wnt/beta-catenin signaling. *Dev. Cell* 32, 82–96. <https://doi.org/10.1016/j.devcel.2014.11.016>.
 89. Gritz, E., and Hirschi, K.K. (2016). Specification and function of hemogenic endothelium during embryogenesis. *Cell. Mol. Life Sci.* 73, 1547–1567. <https://doi.org/10.1007/s00018-016-2134-0>.
 90. Ocampo, A., Reddy, P., Martinez-Redondo, P., Platero-Luengo, A., Hatanaka, F., Hishida, T., Li, M., Lam, D., Kurita, M., Beyret, E., et al. (2016). In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell* 167, 1719–1733.e12. <https://doi.org/10.1016/j.cell.2016.11.052>.
 91. Darnell, J.E., Jr. (2002). Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* 2, 740–749. <https://doi.org/10.1038/nrc906>.
 92. Suvà, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill, D.P., Nahed, B.V., et al. (2014). Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* 157, 580–594. <https://doi.org/10.1016/j.cell.2014.02.030>.
 93. Joung, J., Konermann, S., Gootenberg, J.S., Abudayyeh, O.O., Platt, R.J., Brigham, M.D., Sanjana, N.E., and Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat. Protoc.* 12, 828–863. <https://doi.org/10.1038/nprot.2017.016>.
 94. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* 11, 783–784. <https://doi.org/10.1038/nmeth.3047>.
 95. Yoo, A.S., Sun, A.X., Li, L., Shcheglovitov, A., Portmann, T., Li, Y., Lee-Messer, C., Dolmetsch, R.E., Tsien, R.W., and Crabtree, G.R. (2011). MicroRNA-mediated conversion of human fibroblasts to neurons. *Nature* 476, 228–231. <https://doi.org/10.1038/nature10323>.
 96. Norrander, J., Kempe, T., and Messing, J. (1983). Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* 26, 101–106. [https://doi.org/10.1016/0378-1119\(83\)90040-9](https://doi.org/10.1016/0378-1119(83)90040-9).
 97. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7, R100. <https://doi.org/10.1186/gb-2006-7-10-r100>.
 98. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15. <https://doi.org/10.1186/s13059-017-1382-0>.
 99. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.
 100. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323. <https://doi.org/10.1186/1471-2105-12-323>.
 101. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. <https://doi.org/10.1186/gb-2008-9-9-r137>.
 102. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>.
 103. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* 38, 1408–1414. <https://doi.org/10.1038/s41587-020-0591-3>.
 104. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). gprofiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198. <https://doi.org/10.1093/nar/gkz369>.
 105. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
 106. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. <https://doi.org/10.1093/nar/gkw377>.
 107. Korsunsky, I., Nathan, A., Millard, N., and Raychaudhuri, S. (2019). Presto scales Wilcoxon and auROC analyses to millions of observations <https://doi.org/10.1101/653253>.
 108. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* 14, 975–978. <https://doi.org/10.1038/nmeth.4401>.
 109. Lian, X., Zhang, J., Azarin, S.M., Zhu, K., Hazeltine, L.B., Bao, X., Hsiao, C., Kamp, T.J., and Palecek, S.P. (2013). Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat. Protoc.* 8, 162–175. <https://doi.org/10.1038/nprot.2012.150>.

110. Li, X., Tao, Y., Bradley, R., Du, Z., Tao, Y., Kong, L., Dong, Y., Jones, J., Yan, Y., Harder, C.R.K., et al. (2018). Fast generation of functional sub-type astrocytes from human pluripotent stem cells. *Stem Cell Rep.* 11, 998–1008. <https://doi.org/10.1016/j.stemcr.2018.08.019>.
111. Jin, Y., Liu, Y., Li, Z., Santostefano, K., Shi, J., Zhang, X., Wu, D., Cheng, Z., Wu, W., Terada, N., et al. (2018). Enhanced differentiation of human pluripotent stem cells into cardiomyocytes by bacteria-mediated transcription factors delivery. *PLoS One* 13, e0194895. <https://doi.org/10.1371/journal.pone.0194895>.
112. Hester, M.E., Murtha, M.J., Song, S., Rao, M., Miranda, C.J., Meyer, K., Tian, J., Boulting, G., Schaffer, D.V., Zhu, M.X., et al. (2011). Rapid and efficient generation of functional motor neurons from human pluripotent stem cells using gene delivered transcription factor codes. *Mol. Ther.* 19, 1905–1912. <https://doi.org/10.1038/mt.2011.135>.
113. Takayama, K., Inamura, M., Kawabata, K., Katayama, K., Higuchi, M., Tashiro, K., Nonaka, A., Sakurai, F., Hayakawa, T., Furue, M.K., and Mizuguchi, H. (2012). Efficient generation of functional hepatocytes from human embryonic stem cells and induced pluripotent stem cells by HNF4alpha transduction. *Mol. Ther.* 20, 127–137. <https://doi.org/10.1038/mt.2011.234>.
114. Takayama, K., Inamura, M., Kawabata, K., Sugawara, M., Kikuchi, K., Higuchi, M., Nagamoto, Y., Watanabe, H., Tashiro, K., Sakurai, F., et al. (2012). Generation of metabolically functioning hepatocytes from human pluripotent stem cells by FOXA2 and HNF1alpha transduction. *J. Hepatol.* 57, 628–636. <https://doi.org/10.1016/j.jhep.2012.04.038>.
115. Tanaka, A., Woltjen, K., Miyake, K., Hotta, A., Ikeya, M., Yamamoto, T., Nishino, T., Shoji, E., Sehara-Fujisawa, A., Manabe, Y., et al. (2013). Efficient and reproducible myogenic differentiation from human iPS cells: prospects for modeling Miyoshi myopathy in vitro. *PLoS One* 8, e61540. <https://doi.org/10.1371/journal.pone.0061540>.
116. Elcheva, I., Brok-Volchanskaya, V., Kumar, A., Liu, P., Lee, J.H., Tong, L., Vodyanik, M., Swanson, S., Stewart, R., Kyba, M., et al. (2014). Direct induction of haematoendothelial programs in human pluripotent stem cells by transcriptional regulators. *Nat. Commun.* 5, 4372. <https://doi.org/10.1038/ncomms5372>.
117. Tomizawa, M., Shinozaki, F., Motoyoshi, Y., Sugiyama, T., Yamamoto, S., and Ishige, N. (2016). Transcription factors and medium suitable for initiating the differentiation of human-induced pluripotent stem cells to the hepatocyte lineage. *J. Cell. Biochem.* 117, 2001–2009. <https://doi.org/10.1002/jcb.25494>.
118. Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* 43, D76–D81. <https://doi.org/10.1093/nar/gku887>.
119. UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212. <https://doi.org/10.1093/nar/gku989>.
120. Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular identity of human outer radial glia during cortical development. *Cell* 163, 55–67. <https://doi.org/10.1016/j.cell.2015.09.004>.
121. Camp, J.G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Brauninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M., et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proc. Natl. Acad. Sci. USA* 112, 15672–15677. <https://doi.org/10.1073/pnas.1520760112>.
122. Johnson, M.B., Wang, P.P., Atabay, K.D., Murphy, E.A., Doan, R.N., Hecht, J.L., and Walsh, C.A. (2015). Single-cell analysis reveals transcriptional heterogeneity of neural progenitors in human cortex. *Nat. Neurosci.* 18, 637–646. <https://doi.org/10.1038/nn.3980>.
123. Llorens-Bobadilla, E., Zhao, S., Baser, A., Saiz-Castro, G., Zwadlo, K., and Martin-Villalba, A. (2015). Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell* 17, 329–340. <https://doi.org/10.1016/j.stem.2015.07.002>.
124. Shin, J., Berg, D.A., Zhu, Y., Shin, J.Y., Song, J., Bonaguidi, M.A., Enikolopov, G., Nauen, D.W., Christian, K.M., Ming, G.L., and Song, H. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* 17, 360–372. <https://doi.org/10.1016/j.stem.2015.07.013>.
125. Thomsen, E.R., Mich, J.K., Yao, Z., Hodge, R.D., Doyle, A.M., Jang, S., Shehata, S.I., Nelson, A.M., Shapovalova, N.V., Levi, B.P., and Ramamathan, S. (2016). Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods* 13, 87–93. <https://doi.org/10.1038/nmeth.3629>.
126. Wu, J.Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., Raha, D., Egholm, M., Lin, H., Weissman, S., et al. (2010). Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. USA* 107, 5254–5259. <https://doi.org/10.1073/pnas.0914114107>.
127. Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S., Li, G., et al. (2016). Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* 89, 37–53. <https://doi.org/10.1016/j.neuron.2015.11.013>.
128. Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., et al. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555, 524–528. <https://doi.org/10.1038/nature25980>.
129. Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarneri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* 34, 11929–11947. <https://doi.org/10.1523/JNEUROSCI.1860-14.2014>.
130. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnérberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. <https://doi.org/10.1126/science.aaa1934>.
131. Sloan, S.A., Darmanis, S., Huber, N., Khan, T.A., Birey, F., Caneda, C., Reimer, R., Quake, S.R., Barres, B.A., and Pașca, S.P. (2017). Human astrocyte maturation captured in 3D cerebral cortical spheroids derived from pluripotent stem cells. *Neuron* 95, 779–790.e6. <https://doi.org/10.1016/j.neuron.2017.07.035>.
132. Albar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichová, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. <https://doi.org/10.1038/nmeth.4463>.
133. DeLaughter, D.M., Bick, A.G., Wakimoto, H., McKean, D., Gorham, J.M., Kathiriyva, I.S., Hinson, J.T., Homsky, J., Gray, J., Pu, W., et al. (2016). Single-cell resolution of temporal gene expression during heart development. *Dev. Cell* 39, 480–490. <https://doi.org/10.1016/j.devcel.2016.10.001>.
134. Li, G., Xu, A., Sim, S., Priest, J.R., Tian, X., Khan, T., Quertermous, T., Zhou, B., Tsao, P.S., Quake, S.R., and Wu, S.M. (2016). Transcriptomic profiling maps anatomically patterned subpopulations among single embryonic cardiac cells. *Dev. Cell* 39, 491–507. <https://doi.org/10.1016/j.devcel.2016.10.014>.
135. Lescroart, F., Wang, X., Lin, X., Swedlund, B., Gargouri, S., Sánchez-Dáñez, A., Moignard, V., Dubois, C., Paulissen, C., Kinston, S., et al. (2018). Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* 359, 1177–1181. <https://doi.org/10.1126/science.aao4174>.
136. Friedman, C.E., Nguyen, Q., Lukowski, S.W., Helfer, A., Chiu, H.S., Miklas, J., Levy, S., Suo, S., Han, J.J., Osteil, P., et al. (2018). Single-cell transcriptomic analysis of cardiac differentiation from human PSCs

- reveals HOPX-dependent cardiomyocyte maturation. *Cell Stem Cell* 23, 586–598.e8. <https://doi.org/10.1016/j.stem.2018.09.009>.
137. Sereti, K.I., Nguyen, N.B., Kamran, P., Zhao, P., Ranjbarvaziri, S., Park, S., Sabri, S., Engel, J.L., Sung, K., Kulkarni, R.P., et al. (2018). Analysis of cardiomyocyte clonal expansion during mouse heart development and injury. *Nat. Commun.* 9, 754. <https://doi.org/10.1038/s41467-018-02891-z>.
 138. Churko, J.M., Garg, P., Treutlein, B., Venkatasubramanian, M., Wu, H., Lee, J., Wessells, Q.N., Chen, S.Y., Chen, W.Y., Chetal, K., et al. (2018). Defining human cardiac transcription factor hierarchies using integrated single-cell heterogeneity analysis. *Nat. Commun.* 9, 4906. <https://doi.org/10.1038/s41467-018-07333-4>.
 139. Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., et al. (2019). Single-cell transcriptome analysis maps the developmental track of the human heart. *Cell Rep.* 26, 1934–1950.e5. <https://doi.org/10.1016/j.celrep.2019.01.079>.
 140. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502. <https://doi.org/10.1038/s41586-019-0969-x>.
 141. Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566, 490–495. <https://doi.org/10.1038/s41586-019-0933-9>.
 142. Sahara, M., Santoro, F., Sohlmér, J., Zhou, C., Witman, N., Leung, C.Y., Mononen, M., Bylund, K., Gruber, P., and Chien, K.R. (2019). Population and single-cell analysis of human cardiogenesis reveals unique LGR5 ventricular progenitors in embryonic outflow tract. *Dev. Cell* 48, 475–490.e7. <https://doi.org/10.1016/j.devcel.2019.01.005>.
 143. Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* 151, 206–220. <https://doi.org/10.1016/j.cell.2012.07.035>.
 144. Pervolaraki, E., Dachtler, J., Anderson, R.A., and Holden, A.V. (2018). The developmental transcriptome of the human heart. *Sci. Rep.* 8, 15362. <https://doi.org/10.1038/s41598-018-33837-6>.
 145. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51, 1664–1669. <https://doi.org/10.1038/s41588-019-0538-0>.
 146. ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640. <https://doi.org/10.1126/science.1105136>.
 147. Palpant, N.J., Pabon, L., Friedman, C.E., Roberts, M., Hadland, B., Zaunbrecher, R.J., Bernstein, I., Zheng, Y., and Murry, C.E. (2017). Generating high-purity cardiac and endothelial derivatives from patterned mesoderm using human pluripotent stem cells. *Nat. Protoc.* 12, 15–31. <https://doi.org/10.1038/nprot.2016.153>.
 148. Nehme, R., Zuccaro, E., Ghosh, S.D., Li, C., Sherwood, J.L., Pietilainen, O., Barrett, L.E., Limone, F., Worringer, K.A., Kommineni, S., et al. (2018). Combining NGN2 programming with developmental patterning generates human excitatory neurons with NMDAR-mediated synaptic transmission. *Cell Rep.* 23, 2509–2523. <https://doi.org/10.1016/j.celrep.2018.04.066>.
 149. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. <https://doi.org/10.1038/ncomms14049>.
 150. Eze, U.C., Bhaduri, A., Haeussler, M., Nowakowski, T.J., and Kriegstein, A.R. (2021). Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* 24, 584–594. <https://doi.org/10.1038/s41593-020-00794-1>.
 151. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 358, 1318–1323. <https://doi.org/10.1126/science.aap8809>.
 152. Quadrato, G., Nguyen, T., Macosko, E.Z., Sherwood, J.L., Min Yang, S., Berger, D.R., Maria, N., Scholvin, J., Goldman, M., Kinney, J.P., et al. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* 545, 48–53. <https://doi.org/10.1038/nature22047>.
 153. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
 154. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. <https://doi.org/10.1093/bioinformatics/btt656>.
 155. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* 27, 491–499. <https://doi.org/10.1101/gr.209601.116>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
NEUROD1	Abcam	ab60704 (1:1000)
GAPDH	Cell Signaling Technologies	2118L (1:1000)
V5	Cell Signaling Technologies	13202S (1:1000, Western blot)
ACTB	MilliporeSigma	A5441 (1:5000)
IRDye 680RD Donkey anti-Mouse IgG	LiCOR	925-68072 (1:20000)
IRDye 800CW Donkey anti-Rabbit IgG	LiCOR	925-32213 (1:20000)
DYRK1A	Novus Biologicals	H00001859-M01 (1:250)
ACTB	Cell Signaling Technologies	4967L (1:1000)
anti-mouse IgG, HRP-linked	Cell Signaling Technologies	7076S (1:5000)
anti-rabbit IgG, HRP-linked	Cell Signaling Technologies	7074S (1:5000)
MAP2	MilliporeSigma	M1406 (1:500)
PAX6	Abcam	ab5790 (1:500)
NES	MilliporeSigma	MAB5326 (1:200)
VIM	Proteintech	10366-1-AP (1:200)
GFAP	Abcam	ab4674 (1:500)
NG2	MilliporeSigma	AB5320 (1:200)
PDGFRA	Cell Signaling Technologies	3164S (1:200)
TNNT2	Abcam	ab8295 (1:500, immunofluorescence)
NKX2.5	R&D Systems	AF2444 (1:200)
ACTA2	MilliporeSigma	A2547 (1:200)
AQP5	Abcam	ab78486 (1:200)
CDH5	Santa Cruz Biotechnology	sc-9989 (1:100)
CGA	Thermo Fisher Scientific	MA1-82895 (1:200)
COL1A1	Abcam	ab6308 (1:100)
CUBN	Abcam	ab191073 (1:200)
cytokeratin	Dako	M351529-2 (1:200)
EPCAM	Abcam	ab71916 (1:100)
FOXJ1	Thermo Fisher Scientific	14-9965-82 (1:100)
GATA3	Santa Cruz Biotechnology	sc-268 (1:200)
LGR5	Abcam	ab75850 (1:100)
LUM	Abcam	ab168348 (1:200)
PECAM1	Abcam	ab28364 (1:100)
SNAI2	Cell Signaling Technologies	9585S (1:200)
TAGLN	Abcam	ab14106 (1:200)
acetylated TUBA4A	MilliporeSigma	T6793-100UL (1:500)
TUBB3	Abcam	ab18207 (1:500)
WNT4	R&D Systems	AF475 (1:200)
Alexa Fluor 568 goat anti-mouse IgG	Thermo Fisher Scientific	A-11031 (1:1000)
Alexa Fluor 488 goat anti-chicken IgY	Thermo Fisher Scientific	A-11039 (1:1000)
Alexa Fluor 647 goat anti-rabbit IgG	Thermo Fisher Scientific	A-21244 (1:1000)
Alexa Fluor 488 goat anti-rabbit IgG	Thermo Fisher Scientific	A-11008 (1:1000)
Alexa Fluor 647 donkey anti-goat IgG	Thermo Fisher Scientific	A-21447 (1:1000)
V5	Thermo Fisher Scientific	R960-25 (1:1000, chromatin immunoprecipitation)

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Alexa Fluor 488 SSEA4	Thermo Fisher Scientific	53-8843-42 (1:50)
PE TRA-1-60	Thermo Fisher Scientific	12-8863-82 (1:50)
TNNT2	Thermo Fisher Scientific	MS-295-P1 (1:200, flow cytometry)
Bacterial and virus strains		
Stbl3	Thermo Fisher Scientific	C737303
Chemicals, peptides, and recombinant proteins		
high-glucose DMEM with GlutaMax and pyruvate	Thermo Fisher Scientific	10569010
fetal bovine serum	VWR	97068-085
penicillin/streptomycin	Thermo Fisher Scientific	15140122
TrypLE Express	Thermo Fisher Scientific	12604021
Geltrex	Thermo Fisher Scientific	A1413202
mTeSR1	STEMCELL Technologies	85850
ReLeSR	STEMCELL Technologies	05873
Accutase	STEMCELL Technologies	07920
ROCK Inhibitor Y27632	Enzo Life Sciences	ALX-270-333-M025
Normocin	Invivogen	ant-nr-1
Neurobasal	Thermo Fisher Scientific	21103049
B-27	Thermo Fisher Scientific	17504044
GlutaMAX	Thermo Fisher Scientific	35050061
DMEM/F-12 with HEPES	Thermo Fisher Scientific	11330057
EGF	MilliporeSigma	E9644
bFGF	STEMCELL Technologies	78003
heparin	STEMCELL Technologies	07980
doxycycline	MilliporeSigma	D9891
N-2	Thermo Fisher Scientific	17502048
insulin	Millipore Sigma	19278
nonessential amino acids	Thermo Fisher Scientific	11140050
2-mercaptoethanol	Millipore Sigma	M6250
B-27 minus vitamin A	Thermo Fisher Scientific	12587010
dorsomorphin	Millipore Sigma	P5499
SB-431542	R&D Systems	1614
RPMI 1640 with GlutaMax	Thermo Fisher Scientific	A1895601
B-27 minus insulin	Thermo Fisher Scientific	17504044
Ascorbic acid	Millipore Sigma	A4403
CHIR99021	Selleckchem	S1263
IWP4	Stemgent	04-0036
StemPro-34 SFM	Thermo Fisher Scientific	10639011
STEMdiff APEL 2	STEMCELL Technologies	05275
KnockOut DMEM	Thermo Fisher Scientific	10829018
KnockOut Serum Replacement	Thermo Fisher Scientific	10828010
Lipofectamine 3000	Thermo Fisher Scientific	L3000150
P3000 Enhancer	Thermo Fisher Scientific	L3000150
Opti-MEM	Thermo Fisher Scientific	31985070
Hygromycin	Thermo Fisher Scientific	10687010
Blasticidin	Thermo Fisher Scientific	A1113903
Puromycin	Thermo Fisher Scientific	A1113803
RIPA lysis buffer	Cell Signaling Technologies	9806S
protease inhibitor cocktail	MilliporeSigma	05892791001
Odyssey Blocking Buffer (TBS)	LiCOR	927-50000
BLOT-QuickBlocker	G Biosciences	786-011

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pierce ECL Western Blotting Substrate	Thermo Fisher Scientific	32209
paraformaldehyde	VWR	15710
goat serum	Cell Signaling Technologies	5425S
Triton X-100	MilliporeSigma	93443
ProLong Gold Antifade Mountant with DAPI	Thermo Fisher Scientific	P36941
Formaldehyde	MilliporeSigma	252549
glycine	MilliporeSigma	G7126
Protein A/G Magnetic Beads	Thermo Fisher Scientific	88802
Proteinase K	Qiagen	19133
RNase A	Qiagen	19101
SPRIselect Reagent	Beckman Coulter	B23318
QuickExtract DNA Extraction Solution	Lucigen	QE09050
BSA	MilliporeSigma	A9418
Critical commercial assays		
Pierce BCA protein assay	VWR	23227
QIAamp Viral RNA Mini Kit	Qiagen	52906
qScript Flex cDNA Kit	VWR	95049-100
P3 Primary Cell 4D-Nucleofector X Kit	Lonza	V4XP-3024
PrimeFlow RNA assay kit	Thermo Fisher Scientific	88-18005-204
Chromium Single Cell 3' Library & Gel Bead Kit v2	10x Genomics	120237
Chromium Single Cell 3' Library & Gel Bead Kit v3	10x Genomics	1000075
RNeasy Plus Mini Kit	Qiagen	74134
NEBNext Ultra RNA Library Prep Kit for Illumina	NEB	E7530
NEBNext Ultra II DNA Library Prep Kit for Illumina	NEB	E7645
Click-iT EdU Alexa Fluor 488 Flow Cytometry Assay Kit	Thermo Fisher Scientific	C10420
Deposited data		
Raw sequencing data	This paper	GEO: GSE216481
Processed data	This paper	GEO: GSE216481
Original Western blots	This paper	Mendeley Data: https://doi.org/10.17632/8zhjwbz254
Human fetal expression atlas	Cao et al. ³⁰	GEO: GSE156793
Experimental models: Cell lines		
HEK293FT	Thermo Fisher Scientific	R70007
H1 hESCs	WiCell	WA01
HUES66 hESCs	Harvard Stem Cell Institute iPS Core Facility	NA
H9 hESCs	WiCell	WA09
11a iPSCs	Paola Arlotta laboratory, Harvard University	NA
Oligonucleotides		
Oligo sequences	See Table S7	N/A
Recombinant DNA		
lentiMPHv2	Joung et al. ⁹³	Addgene 89308
lentiSAMv2	Joung et al. ⁹³	Addgene 75112
LentiCRISPRv2	Sanjana et al. ⁹⁴	Addgene 52961

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pUltra-puro-RTTA3	Yildirim Dogan and Kitai Kim	Addgene 58750
pLX_TRC209	Broad Genetic Perturbation Platform	pLX_TRC209
pLX_TRC317	Broad Genetic Perturbation Platform	pLX_TRC317
pTight promoter	Yoo et al. ⁹⁵	Addgene 31877
psPAX2	Didier Trono	Addgene 12260
pMD2.G	Didier Trono	Addgene 12259
pUC19	Norrander et al. ⁹⁶	Addgene 50005
NEUROD1 sgRNA 1	This paper	CATGCGCCATATGGTCTTCC
NEUROD1 sgRNA 2	This paper	ATACAAATGGGCAGGTCACG
NEUROG2 sgRNA 1	This paper	GAAAAGAATAAGCCAGAGGA
NEUROG2 sgRNA 2	This paper	CTGACAGGAGGAGGAGGCAGG
Non-targeting sgRNA 1	This paper	CTGAAAAAGGAAGGAGTTGA
Non-targeting sgRNA 2	This paper	AAGATGAAAGGAAAGGCCTT
SLC1A3 sgRNA	This paper	ATGGAGAACGCCAAGA
VIM sgRNA	This paper	TCCGCAGCCATGTCCACC
DYRK1A sgRNA 1	This paper	TCAGCAACCTCTAACTAAC
DYRK1A sgRNA 2	This paper	TCATTGGCACCACTGAACAG
Software and algorithms		
CellProfiler (v3.1.8, v4.2.1)	Carpenter et al. ⁹⁷	https://cellprofiler.org/
Cell Ranger Single cell Software Suite (v3.1.0)	10x Genomics	https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome
Scanpy (v1.7.2)	Wolf et al. ⁹⁸	https://github.com/scverse/scanpy
Bowtie (v1.2.3)	Langmead et al. ⁹⁹	https://github.com/BenLangmead/bowtie
RSEM (v1.3.1)	Li and Dewey ¹⁰⁰	https://github.com/deweylab/RSEM
MACS (v1.4.2)	Zhang et al. ¹⁰¹	https://github.com/macs3-project/MACS
HOMER (v4.10.3)	Heinz et al. ¹⁰²	http://homer.ucsd.edu/homer/
FlowJo (v10.8.1)	BD Biosciences	https://www.flowjo.com/
Clampex (v10.7)	Molecular Devices	https://www.moleculardevices.com/products/axon-patch-clamp-system/acquisition-and-analysis-software/pclamp-software-suite
SHARE-seq preprocessing	Ma et al. ²⁰	https://github.com/masai1116/SHARE-seq-alignment
SHARE-seq TF alignment	This paper	https://github.com/fengzhanglab/Joung_TFAtlas_manuscript ; Zenodo: https://doi.org/10.5281/zenodo.7466207
Harmony (Scanpy)	Korsunsky et al. ³²	https://github.com/immunogenomics/harmony
scVelo (v0.2.3)	Bergen et al. ¹⁰³	https://github.com/theislab/scvelo
g:Profiler	Raudvere et al. ¹⁰⁴	https://biit.cs.ut.ee/gprofiler/gost
scikit-learn (v0.24.2)	Pedregosa et al. ¹⁰⁵	http://scikit-learn.sourceforge.net
GSEAp (Enrichr) (v0.10.4)	Kuleshov et al. ¹⁰⁶	https://github.com/zqfang/GSEAp
Seurat (v4)	Hao et al. ⁷⁷	https://github.com/satijalab/seurat
SingleCellNet	Tan and Cahan ³³	https://github.com/pcahan1/singleCellNet
Presto	Korsunsky et al. ¹⁰⁷	https://github.com/immunogenomics/presto
chromVAR	Schep et al. ¹⁰⁸	https://github.com/GreenleafLab/chromVAR
CellID	Cortel et al. ⁸³	https://github.com/RausellLab/CellID
PRISM (v9.1.0)	GraphPad	https://www.graphpad.com/scientific-software/prism/

RESOURCE AVAILABILITY

Lead contact

Requests for further information should be directed to and will be fulfilled by the lead contact, Feng Zhang (zhang@broadinstitute.org).

Materials availability

The pooled and arrayed versions of the MORF library have been deposited to Addgene.

Data and code availability

- All raw and processed sequencing data generated as part of this study have been deposited at GEO and are publicly available as of the date of publication. Accession number is listed in the [key resources table](#). Original Western blot images have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the [key resources table](#). Microscopy data reported in this paper will be shared by the [lead contact](#) upon request.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOI is listed in the [key resources table](#).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture and differentiation

HEK293FT cells were maintained in high-glucose DMEM with GlutaMax and pyruvate, 10% fetal bovine serum, and 1% penicillin/streptomycin. Cells were passaged every other day at a ratio of 1:4 or 1:5 using TrypLE Express.

Unless otherwise specified, human embryonic stem cells (hESCs) used in these experiments were from H1 hESCs. HUES66 hESCs were used for the induced neural progenitor (iNP) screens, iNP candidate TF validation, induced cardiomyocyte (iCM) characterization, and iCM scRNA-seq. Sequential TF screens were performed in iNPs and iCMs derived from H1 hESCs. Other stem cell lines used in this study include H9 hESCs and 11a human induced pluripotent stem cell (iPSCs). hESCs and iPSCs were maintained in cell culture dishes coated with 1% Geltrex membrane matrix in mTeSR1 medium. For routine maintenance, stem cells were passaged 1:10-1:20 using ReLeSR. For lentivirus transduction and differentiation, cells were dissociated using Accutase and seeded in mTeSR1 with 10 µM ROCK Inhibitor Y27632. All stem cells were maintained below passage 30 and confirmed to be karyotypically normal and negative for mycoplasma every 5-10 passages. Normocin was used as an antibiotic for stem cell culture and differentiation.

During neuronal differentiation, stem cell media was incrementally shifted towards neuronal media (Neurobasal medium, B-27, and GlutaMAX) in 25% increments starting from day 2. On day 5, media was changed to 100% neuronal media.

During TF-iNP differentiation, stem cell media was gradually shifted towards NP media (DMEM/F-12 with HEPES, B-27, 20 ng/mL EGF, 20 ng/mL bFGF, and 2 µg/mL heparin) in 25% increments as described above for neuronal differentiation. Cells were passaged at day 4. For spontaneous differentiation, 2 µg/mL doxycycline was added to the media starting on day 0 for 7 days to induce TF expression. After 7 days, cells were maintained in NP media for 3 days before media was changed to differentiation media (DMEM/F-12 with HEPES, B-27, and 2 µg/mL heparin). Half of the media was refreshed every other day during spontaneous differentiation.

For *RFX4*-iNP protocol optimization, base media from the dual SMAD inhibition (DS)⁵⁸ and embryoid body (EB)⁵⁷ protocols were tested. DS media is a 1:1 mix of N-2 (DMEM/F12 with HEPES, N-2, 5 µg/mL insulin, 100 µM nonessential amino acids, and 100 µM 2-mercaptoethanol) and neuronal media. EB media (DMEM/F12 with HEPES, N-2, and B-27 minus vitamin A) was also tested. SMAD inhibitors dorsomorphin and SB-431542 were added where indicated. To provide the best comparison between *RFX4*-iNP, DS, and EB methods, the differentiation timelines were aligned such that the iNPs produced by the three methods were dissociated for scRNA-seq at the same time.

For *EOMES*-iCM differentiation, hESCs were seeded in mTeSR. After 2 days, when cells have reached confluence (day 0), 2 µg/mL doxycycline was added to the media for 2 days unless otherwise indicated. On day 1, media was switched to CM differentiation media (RPMI 1640 with GlutaMax, B-27 minus insulin, and 10 mg/mL Ascorbic acid). Media was refreshed on day 2 and every other day afterwards. On day 7, half of the media was replaced with CM maintenance media (RPMI 1640 with GlutaMax and B-27). On day 8, all of the media was replaced with CM maintenance media. For CM differentiation using GSK and Wnt inhibitors, 10 µM CHIR99021 and 5 mM IWP4 were used as described previously.¹⁰⁹

To select the optimal media condition for the TF Atlas, stem cell media was gradually shifted towards 7 medias in 25% increments starting from day 2 as described above. Medias tested include M1 (DMEM/F-12 with HEPES, N-2, B-27, and 100 µM nonessential amino acids), M2 (1:1 mix of neuronal media and DMEM/F-12 with HEPES, N-2, and 100 µM nonessential amino acids), M3 (StemPro-34 SFM and GlutaMAX), M4 (STEMdiff APEL 2), M5 (CM maintenance media), M6 (KnockOut DMEM, KnockOut Serum Replacement, GlutaMAX, and 100 µM nonessential amino acids), and M7 (mTeSR). M4 was selected for the TF Atlas and validation.

We chose 7 days of TF overexpression for media testing, as it is the median duration of established TF-driven differentiation protocols.^{4,7,85,86,110–117}

Lentivirus production

HEK293FT cells were cultured as described above. 1 day prior to transfection, cells were seeded at ~40% confluence in T25, T75, or T225 flasks. Cells were transfected the next day at ~90–99% confluence. For each T25 flask, 3.4 µg of plasmid containing the vector of interest, 2.6 µg of psPAX2, and 1.7 µg of pMD2.G were transfected using 17.5 µL of Lipofectamine 3000, 15 µL of P3000 Enhancer, and 1.25 mL of Opti-MEM. Transfection parameters were scaled up linearly with flask area for T75 and T225 flasks. Media was changed 5 h after transfection. Virus supernatant was harvested 48 h post-transfection, filtered with a 0.45 µm PVDF filter, aliquoted, and stored at –80 °C.

Lentivirus transduction

For transduction, 3×10^6 hESCs or iPSCs were seeded in 10-cm cell culture dishes with an appropriate volume of lentivirus. After 24 h, media was refreshed with the appropriate antibiotic. For 5 days, media with the appropriate antibiotic was refreshed every day, and cells were passaged after 3 days of selection. Concentrations for selection agents were determined using a kill curve: 150 µg/mL Hygromycin, 3 µg/mL Blasticidin, and 1 µg/mL Puromycin. Lentiviral titers were calculated by transducing cells with 5 different volumes of lentivirus and determining viability after a complete selection of 3 days.⁹³

METHOD DETAILS

Sequences and cloning

The plasmids lentiIMPHv2 and lentiSAMv2 were used for CRISPR activation. LentiCRISPRv2 was used for CRISPR-Cas9 mediated homology-directed repair (HDR). The Puromycin resistance gene in lentiCRISPRv2 was replaced with the lentiSAMv2 Blasticidin resistance gene for CRISPR-Cas9 knockout of *DYRK1A*. Single guide RNAs (sgRNAs) were cloned into the respective vectors as previously described.⁹³ For dox-inducible gene expression, the plasmid pUltra-puro-RTTA3 was used for rtTA. The dox-inducible ORF vector was cloned by replacing the EF1a promoter in pLX_TRC209 with the pTight promoter. For *DYRK1A* overexpression, the codon-optimized *DYRK1A* sequence (NM_001396) was cloned into pLX_TRC209 for expression under EF1a and the Hygromycin resistance gene was replaced with the lentiSAMv2 Blasticidin resistance gene.

qPCR quantification of transcript expression

Cells were seeded in 96-well plates and grown to 60–90% confluence before RNA was reverse transcribed for qPCR as described previously.⁹³ TaqMan qPCR was performed with custom or ready-made probes (Table S7).

Western blot

Protein lysates were harvested with RIPA lysis buffer containing protease inhibitor cocktail. Samples were standardized for protein concentration using the Pierce BCA protein assay and incubated at 70°C for 10 mins under reducing conditions. After denaturation, samples were separated by Bolt 4–12% Bis-Tris Plus Gels (Thermo Fisher Scientific NW04125BOX) and transferred onto a PVDF membrane using iBlot Transfer Stacks (Thermo Fisher Scientific IB401001).

For NEUROD1 and V5, blots were blocked with Odyssey Blocking Buffer (TBS) for 1 h at room temperature. Blots were then probed with different primary antibodies in Odyssey Blocking Buffer overnight at 4°C. Blots were washed with TBST before incubation with secondary antibodies in Odyssey Blocking Buffer for 1 h at room temperature. Blots were washed with TBST and imaged using the Odyssey CLx (LiCor).

For *DYRK1A*, blots were blocked with 5% BLOT-QuickBlocker in TBST for 1 h at room temperature. Blots were then probed with different primary antibodies in 2.5% BLOT-QuickBlocker in TBST overnight at 4°C. Blots were washed with TBST before incubation with secondary antibodies in 2.5% BLOT-QuickBlocker in TBST for 1 h at room temperature. Blots were washed with TBST and imaged using the Pierce ECL Western Blotting Substrate on the ChemiDox XRS+ (Bio-Rad).

Immunofluorescence and imaging

Cells were cultured on poly-D-lysine/laminin coated glass coverslips (VWR 354087) in 24-well plates as described above. Prior to staining, cells were washed with 1 mL PBS and fixed with 4% paraformaldehyde in PBS for 30 mins at room temperature. Cells were washed with PBS and blocked in PBS with 2.5% goat serum and 0.1% Triton X-100 for 1 h at room temperature. Cells were then stained with different primary antibodies in PBS with 1.25% goat serum and 0.1% Triton X-100 overnight at 4°C. Cells were washed in PBS with 0.1% Triton X-100 before staining with the appropriate secondary antibodies in PBS with 1.25% goat serum and 0.1% Triton X-100 for 1 h at room temperature. Cells were washed in PBS with 0.1% Triton X-100, mounted onto slides using ProLong Gold Antifade Mountant with DAPI, and nail polished. Immunostained coverslips for NPs were imaged on a Zeiss Axio Observer with a Hamatsu Camera using a Plan-Apochromat 20x objective and a 1.6x Optovar. Immunostained coverslips for TF Atlas validation were imaged on a Leica Stellaris 5 confocal microscope using a 20x objective. Images were taken from randomly selected regions using fixed exposure times.

Design and cloning of TF ORF libraries

The MORF barcoded human TF library consisted of 1,836 genes that were selected based on AnimalTFDB¹¹⁸ and Uniprot¹¹⁹ annotations and included histone modifiers (Table S1A). The library included all 3,548 splice isoforms that overlapped between RefSeq and Gencode annotations, as well as 2 control vectors expressing GFP and mCherry. Each TF ORF isoform has a unique 24-bp barcode with a Hamming distance of at least 3 compared to all other barcodes. 593 of the 3,548 isoforms were obtained from the Broad Genomic Perturbation Platform. As ORF libraries generated from cDNA libraries often contain missense mutations that can result in screening artifacts, we individually synthesized the rest of the isoforms (Genewiz). TF ORFs were cloned in an arrayed format into pLX_TRC317 for expression under the EF1a promoter. All constructs in the MORF library have been sequence verified.

To create targeted TF ORF libraries, TFs specifically expressed in target cell types were selected using published single cell or bulk RNA-seq datasets. TFs that were identified in 2 or more datasets were included. For NPs, 70 TF genes were selected using 8 datasets from radial glia, neural stem cells, differentiated neural progenitors, and fetal astrocytes.^{120–127} Then, isoforms that comprised >25% of the transcripts for the respective gene were selected using bulk RNA-seq data of human fetal astrocytes,¹²⁷ resulting in 90 TF isoforms (Table S1B). The targeted NP library was cloned into pLX_TRC209. For astrocytes, 44 TF genes were selected using 6 datasets from purified and differentiated astrocytes.^{127–132} Isoforms were selected using bulk RNA-seq data of human fetal astrocytes¹²⁷ as described for NPs, resulting in 54 TF isoforms (Table S1C). For CM, 49 TF genes were selected using 11 datasets from purified and differentiated cardiomyocytes.^{133–143} Isoforms that comprised >5% of the transcripts were selected using bulk RNA-seq of the whole fetal human heart,¹⁴⁴ resulting in 80 TF isoforms (Table S1D). A lower threshold was chosen for CMs because isoform prevalence in the whole fetal heart may not be representative of ventricular, atrial, and mature CMs.

To assess TF distribution, TF barcodes were amplified (TF NGS Fwd 1–10 and Rev 1–12 primers in Table S7) and sequenced on the Illumina MiSeq or NextSeq platforms as previously described.⁹³ For the pooled lentiviral library, lentiviral RNA was harvested using the QIAamp Viral RNA Mini Kit and reverse transcribed using the qScript Flex cDNA Kit with gene-specific priming (TF RT) before barcode amplification. NGS reads that perfectly matched each barcode were counted and normalized to the total number of perfectly matched NGS reads for each condition. Skew ratio was calculated as the normalized count for the 10th percentile divided by the 90th percentile.

Reporter cell line screen

To generate reporter cell lines, EGFP from pLX_TRC209 followed by a T2A (GGCAGTGGAGAGGGCAGAGGAAGTCTGCTAACAT GCGGTGACGTGAGGAGAACCTGGCCCC) self-cleaving peptide was inserted at the N-terminus of endogenous *SLC1A3* and *VIM* genomic sequences. *SLC1A3* and *VIM* were selected as NP marker genes based on convergence across published RNA-seq datasets and high expression levels.^{120–127} Clonal reporter cell lines were generated using CRISPR-Cas9 mediated HDR. To construct the HDR plasmids for each gene, the HDR templates that consisted of the 850–1,000 bp genomic regions flanking the sgRNA cleavage sites were PCR amplified from HUES66 genomic DNA. Then EGFP-T2A flanked by HDR templates were cloned into pUC19. HUES66 hESCs were nucleofected with 10 µg of LentiCRISPRv2 plasmid and 6 µg of HDR plasmid using the P3 Primary Cell 4D-Nucleofector X Kit according to the manufacturer's instructions. Cells were then seeded sparsely (2 electroporation reactions per 10-cm cell culture dish) to form single cell clones. After 18 h, cells were selected for Cas9 expression with 0.5 µg/mL Puromycin for 2 days and expanded until colonies can be picked (~1 week).

Cell colonies were detached by replacing the media with PBS and incubating at room temperature for 15 mins. Each cell colony was removed from the Petri dish using a 200 µL pipette tip and transferred a well in a 96-well plate for expansion. Clones with EGFP insertions were identified by 2-round PCR amplification, first with primers amplifying outside of the HDR template (HDR Fwd 1 and Rev primers in Table S7, 15 cycles) and then with primers amplifying the region of insertion (HDR Fwd 2 and Rev primers in Table S7, 15 cycles) to avoid detecting the HDR template plasmid as a false positive. Products were run on a gel to identify clones with insertions and Sanger sequencing confirmed that EGFP had been inserted at the intended site without mutations. For each reporter cell line, 3 clones with EGFP inserted into one of the two alleles were selected for further expansion and characterization.

For TF screening, *SLC1A3* or *VIM* reporter HUES66 hESC lines were transduced with the pooled TF ORF library at MOI <0.3 and differentiated into iNPs as described above. After 7 days, 5–10 × 10⁶ cells were sorted for EGFP expression using the Sony SH800S Cell Sorter. For each clonal line, the percentage of cells sorted for the control condition was matched to those expressing EGFP (~15–20%). After sorting, TF barcodes from each population were sequenced as described above for the MORF library. Enrichment of each TF was calculated as the normalized barcode count in the high population divided by the count in the low population.

Flow-FISH screen

Marker genes were selected based on convergence across published RNA-seq datasets as described above for targeted TF ORF libraries. HUES66 hESCs were transduced with the pooled TF ORF library at MOI <0.3 and differentiated into iNPs as described above. After 7 days, cells were labeled with the appropriate FISH probes (Table S7) using the PrimeFlow RNA assay kit with 20 million cells per biological replicate. FISH probes targeting transcripts with similar expression levels were pooled together. Once the cells were labeled, the entire cell population was sorted for high or low fluorescence (15% of cells per bin), indicating an aggregate expression level of the transcripts labeled with the pooled FISH probes for the particular wavelength. After sorting, TF barcodes from each population were amplified using a modified ChIP reverse cross-linking protocol as described previously.¹⁴⁵ Enrichment of each TF was determined as described above for the reporter cell line screen.

10X single cell RNA sequencing (scRNA-seq) screen

HUES66 hESCs were transduced with the pooled TF ORF library at MOI <0.3 and differentiated into iNPs. Then, iNPs were dissociated for scRNA-seq analysis as described above. To pair TF barcodes with cell barcodes, TF and cell barcodes were PCR amplified from cDNA retained following the whole transcriptome amplification step of the 10x Genomics scRNA-seq library preparation protocol (10X TF Fwd and Rev 1-8 primers in [Table S7](#)). The resulting amplicon was sequenced on the Illumina NextSeq platform, aiming for a minimum coverage of 20,000 reads per single cell (paired-end; read 1: 16 cycles; read 2: 72 cycles).

10X scRNA-seq library preparation and sequencing

Cells were dissociated with Accutase for 10 mins (NP) or 50 mins (spontaneously differentiated cells) at 37°C and filtered using a 70 µm cell strainer to obtain single cells. Cells were loaded in the 10x Genomics Chromium Controller with 10,000 cells per channel. For cells from the scRNA-seq pooled screen and spontaneous differentiation of four candidate TFs, scRNA-seq libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead Kit v2 according to the manufacturer's instructions. Libraries were sequenced on the NextSeq platform, aiming for a minimum coverage of 20,000 reads per single cell (paired-end; read 1: 26 cycles; i7 index: 8 cycles, i5 index: 0 cycles; read 2: 55 cycles). For cells from the NP method comparison and spontaneous differentiation of RFX4-DS-iNPs, scRNA-seq libraries were prepared using the Chromium Single Cell 3' Library & Gel Bead Kit v3 and sequenced on the HiSeq X platform (paired-end; read 1: 28 cycles; i7 index: 8 cycles, i5 index: 0 cycles; read 2: 96 cycles).

Arrayed screen

TF ORFs were packaged individually into lentivirus for delivery into HUES66 hESCs at MOI <0.5. After 7 days, cells were differentiated into NP and harvested for qPCR as described above to measure expression of *SLC1A3* and *VIM*.

Bulk MORF library screen

H1 hESCs were transduced with the pooled MORF library at MOI <0.3 and differentiated for 7 days in different culture media as described above. Cells were stained for pluripotency markers, SSEA4 and TRA-1-60 and sorted for high or low fluorescence (10% of cells per bin). After sorting, TF barcodes from each population were sequenced. Enrichment of each TF was calculated as the normalized barcode count in the low population divided by the count in the high population.

Bulk RNA sequencing (RNA-seq)

RNA from cells plated in 24-well plates and grown to 60–90% confluence was harvested using the RNeasy Plus Mini Kit. RNA-seq libraries were prepared using NEBNext Ultra RNA Library Prep Kit for Illumina and sequenced on the Illumina NextSeq platform (>9 million reads per biological replicate).

Chromatin immunoprecipitation with sequencing (ChIP-seq)

Cells were plated in 10-cm cell culture dishes and grown to 60–80% confluence. For each condition, two biological replicates were harvested for ChIP-seq. Formaldehyde was added directly to the growth media for a final concentration of 1% and cells were incubated at 37°C for 10 mins to initiate chromatin fixation. Fixation was quenched by adding 2.5 M glycine in PBS for a final concentration of 125 mM glycine and incubated at room temperature for 5 mins. Cells were then washed with ice-cold PBS, scraped, and pelleted at 1,000×g for 5 mins.

Cell pellets were prepared for ChIP-seq using the Epigenomics Alternative Mag Bead ChIP Protocol v2.0.¹⁴⁶ Briefly, cell pellets were resuspended in 100 µL of lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.1) containing protease inhibitor cocktail and incubated for 10 mins at 4°C. Then 400 µL of dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.1, and 167 mM NaCl) containing protease inhibitor cocktail was added. Samples were pulse sonicated with 2 rounds of 10 mins (30s on-off cycles, high frequency) in a rotating water bath sonicator (Diagenode Bioruptor) with 5 mins on ice between each round. 10 µL of sonicated sample was set aside as input control. Then 500 µL of dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.1, and 167 mM NaCl) containing protease inhibitor cocktail and 1 µL of anti-V5 was added to the sonicated sample. ChIP samples were rotated end over end overnight at 4°C.

For each ChIP, 50 µL of Protein A/G Magnetic Beads was washed with 1 mL of blocking buffer (0.5% TWEEN and 0.5% BSA in PBS) containing protease inhibitor cocktail twice before resuspending in 100 µL of blocking buffer. ChIP samples were transferred to the beads and rotated end over end for 1 h at 4°C. ChIP supernatant was then removed and the beads were washed twice with 200 µL of RIPA low salt buffer (0.1% SDS, 1% Triton x-100, 1 mM EDTA, 20 mM Tris-HCl pH 8.1, 140 mM NaCl, 0.1% DOC), twice with 200 µL of RIPA high salt buffer (0.1% SDS, 1% Triton x-100, 1 mM EDTA, 20 mM Tris-HCl pH 8.1, 500 mM NaCl, 0.1% DOC), twice with 200 µL of LiCl wash buffer (250 mM LiCl, 1% NP40, 1% DOC, 1 mM EDTA, 10 mM Tris-HCl pH 8.1), and twice with 200 µL of TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0). ChIP samples were eluted with 50 µL of elution buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.1% SDS). 40 µL of water was added to the input control samples. 8 µL of reverse cross-linking buffer (250 mM Tris-HCl pH 6.5, 62.5 mM EDTA pH 8.0, 1.25 M NaCl, 5 mg/ml Proteinase K, 62.5 µg/ml RNase A) was added to the ChIP and input control samples and then incubated at 65°C for 5 h. After reverse crosslinking, samples were purified using 116 µL of SPRIselect Reagent. ChIP-seq libraries were prepared with NEBNext Ultra II DNA Library Prep Kit for Illumina and sequenced on the Illumina NextSeq platform (>60 million reads per condition).

Indel sequencing

Cells plated in 96-well plates were grown to 60–80% confluence and assessed for indel rates as previously described.⁹³ Genomic DNA was harvested from cells using QuickExtract DNA Extraction Solution. The genomic region flanking the site of interest was amplified, first with region-specific primers (DYRK1A KO sgRNA Fwd and Rev primers in [Table S7](#)) for 15 cycles and then with bar-coded primers for 15 cycles. PCR products were sequenced on the Illumina MiSeq platform (>10,000 reads per condition).

Flow cytometry assays

TNNT2 immunostaining was performed using TNNT2 antibodies as described previously.¹⁴⁷ For the EdU assay, cells plated in 24-well plates were differentiated and EdU incorporation was measured using the Click-iT EdU Alexa Fluor 488 Flow Cytometry Assay Kit according to a modified version of the manufacturer's instructions. EdU was added to the culture medium to a final concentration of 10 µM for 2 h before cells were dissociated with Accutase for 15–45 mins at 37°C. Cells were transferred to a 96-well plate, pelleted at 200×g for 5 mins, and washed once with 200 µL of 1% BSA in PBS. Cells were resuspended in 100 µL of Click-iT fixative and incubated for 15 mins at room temperature in the dark. After fixing, cells were washed with 200 µL of 1% BSA in PBS twice, resuspended in 100 µL of Click-iT saponin-based permeabilization and wash reagent, and incubated for 15 mins in the dark. To each sample, 500 µL of Click-iT reaction cocktail was added, and the reaction mixture was incubated for 30 mins at room temperature in the dark. Cells were washed with 200 µL of Click-iT saponin-based permeabilization and wash reagent twice and resuspended in 200 µL of 1% BSA in PBS. For each sample, 10,000 cells were analyzed on a CytoFLEX Flow Cytometer (Beckman Coulter) and quantified with FlowJo.

Electrophysiology

Whole-cell patch-clamp recordings were performed as previously described.¹⁴⁸ Recording pipettes were pulled from thin-walled borosilicate glass capillary tubing (King Precision Glass KG33) on a P-97 puller (Sutter Instrument) and had resistances of 3–5 MΩ when filled with internal solution (128 mM K-gluconate, 10 mM HEPES, 10 mM phosphocreatine sodium salt, 1.1 mM EGTA, 5 mM ATP magnesium salt and 0.4 mM GTP sodium salt, pH 7.3, 300–305 mOsm). The cultured cells were constantly perfused at a speed of 3 mL/min with the extracellular solution (119 mM NaCl, 2.3 mM KCl, 2 mM CaCl₂, 1 mM MgCl₂, 15 mM HEPES, 5 mM glucose, pH 7.3–7.4; Osmolarity was adjusted to 325 mOsm with sucrose). All the experiments were performed at room temperature unless otherwise specified.

Cells were visualized with a 40X water-immersion objective on an upright microscope (Olympus) equipped with IR-DIC. Recordings were made using a Multiclamp 700B amplifier (Molecular Devices) and Clampex 10.7 software. In current clamp mode, membrane potential was held at -65 mV with a Multiclamp 700B amplifier, and step currents were then injected to elicit action potentials. The spontaneous AMPA receptor mediated excitatory postsynaptic currents (sEPSCs) were recorded after entering whole-cell patch clamp recording mode for at least 3 min.

TF Atlas SHARE-seq library preparation

For single TF over-expression, H1 hESCs were transduced with the pooled MORF library at MOI <0.3. We chose H1 hESCs for the TF Atlas because it is a commonly used stem cell line. For combinatorial TF overexpression, H1 hESCs were transduced with combinations of 2 or 3 TFs at MOI <1 in an arrayed format and multiple antibiotics were used to select for expression of TF combinations. Cells were pooled during passaging at day 4. Cells were differentiated for 7 days in STEMdiff APEL 2 as described above. Cells were dissociated with Accutase for 10 mins at 37°C and filtered using a 70 µm cell strainer to obtain single cells.

SHARE-seq libraries were prepared as previously described.²⁰ Briefly, cells were fixed and permeabilized. For joint measurements of single cell chromatin accessibility and expression (scATAC- and scRNA-seq), cells were first transposed by Tn5 transposase to mark regions of open chromatin. The mRNA was reverse transcribed using a poly(T) primer containing a unique molecular identifier (UMI) and a biotin tag. Permeabilized cells were distributed in a 96-well plate to hybridize well-specific barcoded oligonucleotides to transposed chromatin fragments and poly(T) cDNA. Hybridization was repeated three times to expand the barcoding space and ligate cell barcodes to cDNA and chromatin fragments. Reverse crosslinking was performed to release barcoded molecules. cDNA was separated from chromatin using streptavidin beads, and each library was prepared separately for sequencing. Libraries were sequenced on the Illumina NovaSeq platform, aiming for a minimum coverage of 20,000 reads per single cell (for scRNA-seq only, read 1: 100 cycles, read 2: 10 cycles, index 1: 99 cycles, index 2: 8 cycles; for scATAC- and RNA-seq, read 1: 50 cycles, read 2: 50 cycles, index 1: 99 cycles, index 2: 8 cycles). To pair TF barcodes with cell barcodes, TF and cell barcodes were PCR amplified from cDNA retained following the whole transcriptome amplification step and before tagmentation (SHARE-seq TF Fwd 1–8 and Rev primers in [Table S7](#)). The resulting amplicon was sequenced on the Illumina NovaSeq platform, aiming for a minimum coverage of 10,000 reads per single cell (read 1: 65 cycles; index 1: 99 cycles).

QUANTIFICATION AND STATISTICAL ANALYSIS

Image quantification

For quantification of MAP2 staining ([Figures 5P and 5Q](#)), the MeasureImageIntensity module in CellProfiler 3.1.8⁹⁷ was used to measure mean intensity on grayscale MAP2 420 µm × 420 µm images. The IdentifyPrimaryObjects module in CellProfiler was used to

identify and count nuclei in grayscale DAPI images with the following settings modified from default: Typical diameter of objects, in pixel units (Min, Max) = 25, 70; Threshold strategy = Adaptive; Threshold smoothing scale = 1.5; Lower and upper bounds on threshold = 0.06, 1.0. For quantification of marker gene staining (Figures 4D–4K and S3N–S3P), the MeasureImageIntensity module in CellProfiler 4.2.1⁹⁷ was used to measure mean intensity on grayscale 580 µm × 580 µm images. The IdentifyPrimaryObjects module in CellProfiler was used to identify and count nuclei in grayscale DAPI images with the following settings modified from default: Typical diameter of objects, in pixel units (Min, Max): 25, 100; Threshold method = Otsu; Three-class thresholding; Assign pixels in the middle intensity class to the foreground; Threshold smoothing scale = 5; Threshold correction factor = 0.9; Lower and upper bounds on threshold = 0.02, 1.0; Size of smoothing filter = 10; Suppress local maxima that are closer than this minimum allowed distance = 15; speed up by using lower-resolution image to find local maxima = no.

10X scRNA-seq analysis

Sequencing data were aligned and quantified using the Cell Ranger Single cell Software Suite v3.1.0¹⁴⁹ against the GRCh38 human reference genome provided by Cell Ranger. Scanpy v1.7.2⁹⁸ was used to cluster and visualize cells. Cells with 400–7,000 detected genes and less than 10% total mitochondrial gene expression were retained for analysis. Genes that were detected in fewer than 3 cells were removed. Scanpy was used to log normalize, scale, and center the data and unwanted variation was removed by regressing out the number of UMIs and percent mitochondrial reads. Next, highly variable genes were identified and used as input for dimensionality reduction via principal component analysis (PCA). The resulting principal components were then used to cluster the cells, which were visualized using Uniform manifold approximation and projection (UMAP). Clusters were identified using Louvain by fitting the top 50 principal components to compute a neighborhood graph of observations with local neighborhood number of 20 using the scanpy.pp.neighbors function. Cells were then clustered into subgroups using the Louvain algorithm implemented as the scanpy.tl.louvain function. Cluster marker genes and associated p-values were identified using the scanpy.tl.rank_gene_groups function.

To map TF perturbations to expression profiles, for each cell, the TF whose corresponding barcode had the highest number of perfectly matching NGS reads was paired with the cell if the TF barcode had at least 2 reads and >25% more reads than the second highest TF. Otherwise, the cell was excluded from the scRNA-seq analysis. To identify TFs that produced similar expression profiles to radial glia, TF scRNA-seq signatures were correlated to available human fetal cortex or brain organoid scRNA-seq datasets.^{120,128,150–152} The 1,121 most variable genes identified using the scanpy.pp.highly_variable_genes function with the parameters “min_mean=0.0125, max_mean=3 and min_disp=0.5” were used. Candidate TFs were ranked based on Pearson correlations between mean expression profiles of each TF ORF and radial glia from reference datasets.

To compare iNP differentiation methods, the cluster of spontaneously differentiated neurons was excluded. Intra- and inter-batch Euclidean distances were calculated on the 2,305 variable genes using the spatial.distance.pdist and spatial.distance.cdist functions, respectively, from SciPy. Wasserstein distances were determined using the wasserstein_distance function from SciPy.

Bulk RNA-seq analysis

Bowtie⁹⁹ index was created based on the hg38 genome and RefSeq transcriptome. Next, RSEM v1.3.1¹⁰⁰ was run with command line options “–estimate-rspd –bowtie-chunkmbs 512 –paired-end” to align paired-end reads directly to this index using Bowtie and estimate expression levels in transcripts per million (TPM) based on the alignments. TFs with similar RNA-seq signatures to reference cell types from human fetal cortex or brain organoid^{120,151,152} were identified using Pearson correlation between expression profiles. For each TF ORF, the expression signature was defined as the top 2,000 genes with the highest fold change relative to the GFP control condition. For each reference cell type, the average expression profile in TPM was used. To identify genes that were differentially expressed, TPM values were log-transformed ($\log_2(\text{TPM}+1)$) and filtered for genes that were detectable (above or equal to 1) in either condition. TF overexpression conditions were compared to control conditions using the Student’s t-test. Only genes that were significant (FDR < 0.05) were reported.

ChIP-seq analysis

Bowtie⁹⁹ was used to align paired-end reads to the hg38 genome with command line options “-q -X 300 –sam –chunkmbs 512”. Next, biological replicates were merged and Model-based Analysis of ChIP-seq (MACS)¹⁰¹ was run with command line options “-g hs -B -S –mfold 6,30” to identify TF peaks. HOMER¹⁰² was used to discover motifs in the TF peak regions identified by MACS. The findMotifsGenome.pl program from HOMER was run with the command line options “-size 200 -mask” and the top 3 known and *de novo* motifs were presented. TFs were considered potential regulators of a candidate gene if the TF peak region identified by MACS overlapped with the 20kb region centered around the transcriptional start site of the candidate gene RefSeq annotations.

Indel analysis

Indel analysis was performed using a custom Python script as previously described.⁹³

Electrophysiology analysis

Analysis was performed using Clampfit 10.7. Cells in which the series resistance (Rs) changed by >20% were excluded from data analysis. In addition, cells with Rs more than 20 MΩ at any time during the recordings were discarded.

SHARE-seq data preprocessing

SHARE-seq libraries were aligned as previously described.²⁰ Briefly, SHARE-ATAC-seq reads were trimmed and aligned to the hg38 genome using bowtie2. Reads were demultiplexed using four sets of 8-bp barcodes in the index reads, tolerating one mismatched base per barcode. Reads mapping to the mitochondria and chrY were discarded. Duplicates were removed using Picard tools (<http://broadinstitute.github.io/picard/>). Open chromatin region peaks were called on individual samples using MACS2 peak caller.¹⁰¹ Peaks from all samples were merged and peaks overlapping with ENCODE blacklisted regions (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) were filtered out. Peak summits were extended by 150bp on each side and defined as accessible regions. The fragment counts in peaks and TF scores were calculated using chromVAR.¹⁰⁸

SHARE-RNA-seq reads were trimmed and filtered for reads that contain TTTTTT at the 11–16 bases of read 2 allowing for one mismatch. Reads were aligned to hg38 genome using STAR.¹⁵³ Reads were demultiplexed as described above for SHARE-ATAC-seq. Aligned reads were annotated to both exons and introns using featurecounts.¹⁵⁴ UMI-Tools¹⁵⁵ was used to collapse UMIs that were within one mismatch of another UMI. UMIs with only one read were removed as potential ambient RNA contamination. A matrix of gene counts by cell was created with UMI-Tools. Cells that expressed <7,000 genes and <5% mitochondrial reads were retained. The minimum number of genes per cell was selected based on the distribution for each dataset (TF Atlas, >700 genes; joint scRNA- and ATAC-seq, >400 genes; combinatorial TF screen, >500 genes) Scanpy v1.7.2⁹⁸ was used to preprocess the count matrix and cluster cells as described above for 10X scRNA-seq libraries. Harmony³² as implemented by Scanpy (max_iter_harmony = 30, max_iter_kmeans = 50) was used for batch correction. To map TF ORFs to single cells, TF barcodes were ranked by number of perfectly matching NGS reads and filtered for >10 reads. For single TF overexpression, the TF barcode with the highest number of NGS reads and >50% more reads than the second highest TF were mapped. For combinatorial TF overexpression, the top two or three TF barcodes were mapped. Cells without mapped TF ORFs were excluded from downstream analyses. Scanpy's sc.pp.subsample was used to subsample datasets by TF ORF.

Pseudotime analysis

Two approaches were used to order cells along pseudotime: diffusion²¹ and RNA velocity.²² Diffusion pseudotime was determined using Scanpy's sc.tl.diffmap (n_comps = 15) and sc.tl.dpt functions. RNA velocity pseudotime was determined using scVelo.¹⁰³ The top 5,000 most dispersed genes were used to estimate velocity (mode = 'stochastic') and the velocity_pseudotime function was used to determine pseudotime. For each approach, pseudotimes were computed using each cell expressing GFP or mCherry controls as the root cells and averaged. Genes that were differentially expressed over pseudotime were identified by fitting a linear regression on the raw counts against pseudotime using scipy.stats.linregress. Genes with slopes that were significantly different than 0 (FDR < 0.05) were considered differentially expressed.

Non-negative matrix factorization

To identify gene programs in the scRNA-seq data, non-negative matrix factorization (NMF) was performed using scikit-learn¹⁰⁵ (tol = 1e-5, max_iter = 10000). The analysis was performed on log-normalized, centered expression data for the set of variable genes. Negative values were converted to zero to identify enriched gene programs. Positive values were converted to zero and the data multiplied by -1 to identify depleted gene programs. The optimal number of NMF programs for enriched and depleted gene programs was determined by performing NMF analysis over a range of K values (20, 30, 50, 100, 200). The average NMF program weights for each TF ORF were ordered by hierarchical clustering using 1 – correlation coefficient as the distance and Ward's linkage. Clustering results were examined for groups of TFs with known similarities to select the best value of K. We chose 50 NMF programs each for enriched and depleted gene programs.

Pathway enrichment analysis

Pathway enrichment analysis was performed using g:Profiler.¹⁰⁴ The top 100 differentially expressed genes over diffusion pseudotime or genes with the highest NMF gene program weights were provided as a ranked list for input. GO:BP pathways with between 5 and 500 genes that were significantly enriched (FDR < 0.05) were included. To identify non-overlapping pathways, the enriched pathways were sorted by FDR and any pathway that had more than 50% genes overlapping was excluded. For Figure S3F, Enrichr¹⁰⁶ implemented by GSEApY was used to evaluate the enrichment of each pathway in the set of differentially expressed genes for each cluster.

TF potential analysis

We defined a "TF potential" vector for each TF by the relation of TF-induced expression changes to the differentiation trajectory. Specifically, we subclustered cells overexpressing TFs of interest or controls and reordered cells in diffusion pseudotime to attain higher resolution of TF-induced expression changes. For each TF, we used a linear regression model to fit the corresponding expression profiles against pseudotime, defining TF potential as the slope of the linear regression.

Cell type mapping

Expression profiles of differentiated cells were mapped to those of reference cell types from the human fetal cell transcriptome atlas.³⁰ The set of common variable genes between the differentiated cells and reference dataset were used for mapping. For

mapping by label-transfer, the FindTransferAnchors function from Seurat³¹ was run with parameters: dims = 1:50, k.anchor = 15, k.filter = 100, k.score = 50, and max.features = 300. Cells with a maximum prediction score > 0.2 were mapped to the respective cell type. For mapping by dataset integration, the two datasets were integrated using Harmony³² as described above. To annotate each differentiated cell, the cell type labels of the 10 nearest reference cells, as measured by Euclidean distance in the latent space, were evaluated. Differentiated cells were assigned the most common cell type if 8 (80%) or more of the reference cells shared the same cell type label. For mapping by Random Forest classification, SingleCellNet³³ was used to classify differentiated cells. The classifier was trained using the scn_train function from SingleCellNet with parameters: nRand = 200, nTrees = 2000, nTopGenePairs = 200. Cells with a maximum score > 0.1 were mapped to the respective cell type.

Joint chromatin accessibility and gene expression analysis

Seurat v4⁷⁷ was used for the joint chromatin accessibility and gene expression (scATAC- and scRNA-seq) multimodal analysis. Dimensionality reduction was performed on each dataset separately. The scRNA-seq data was normalized and variable features were retained for scaling and PCA. The scATAC-seq data was normalized using term-frequency inverse-document-frequency and the top 250,000 most accessible regions were retained for latent semantic indexing (LSI). Weighted nearest neighbor analysis from Seurat (dims.list = list(1:50, 2:50), prune.SNN = 1/40) was performed using the scRNA-seq PCA and scATAC-seq LSI to simultaneously cluster scATAC- and scRNA-seq data. Marker genes for each cluster were identified using Presto.¹⁰⁷

Two approaches were used to identify top regulators in each cluster. First, chromVAR¹⁰⁸ computed accessibility scores for known motifs at the single cell level. Presto¹⁰⁷ identified TFs whose expression and motif accessibility were both significantly enriched (*FDR* < 0.05) in each cluster. These TFs were ranked by the average of the Presto AUC statistic to identify top regulators. Second, the ATAC peaks were filtered for those whose accessibility was significantly correlated with neighboring genes (41,376 peaks with *FDR* < 0.25) with background correction, as described previously.²⁰ Presto identified ATAC peaks that were significantly enriched (*FDR* < 0.05) in each cluster. Enrichment of known TF motifs was determined by a Kolmogorov-Smirnov test of the position weight matrix (PWM) scores in the cluster compared to PWM scores in GC- and accessibility-matched peaks.

Combinatorial TF prediction

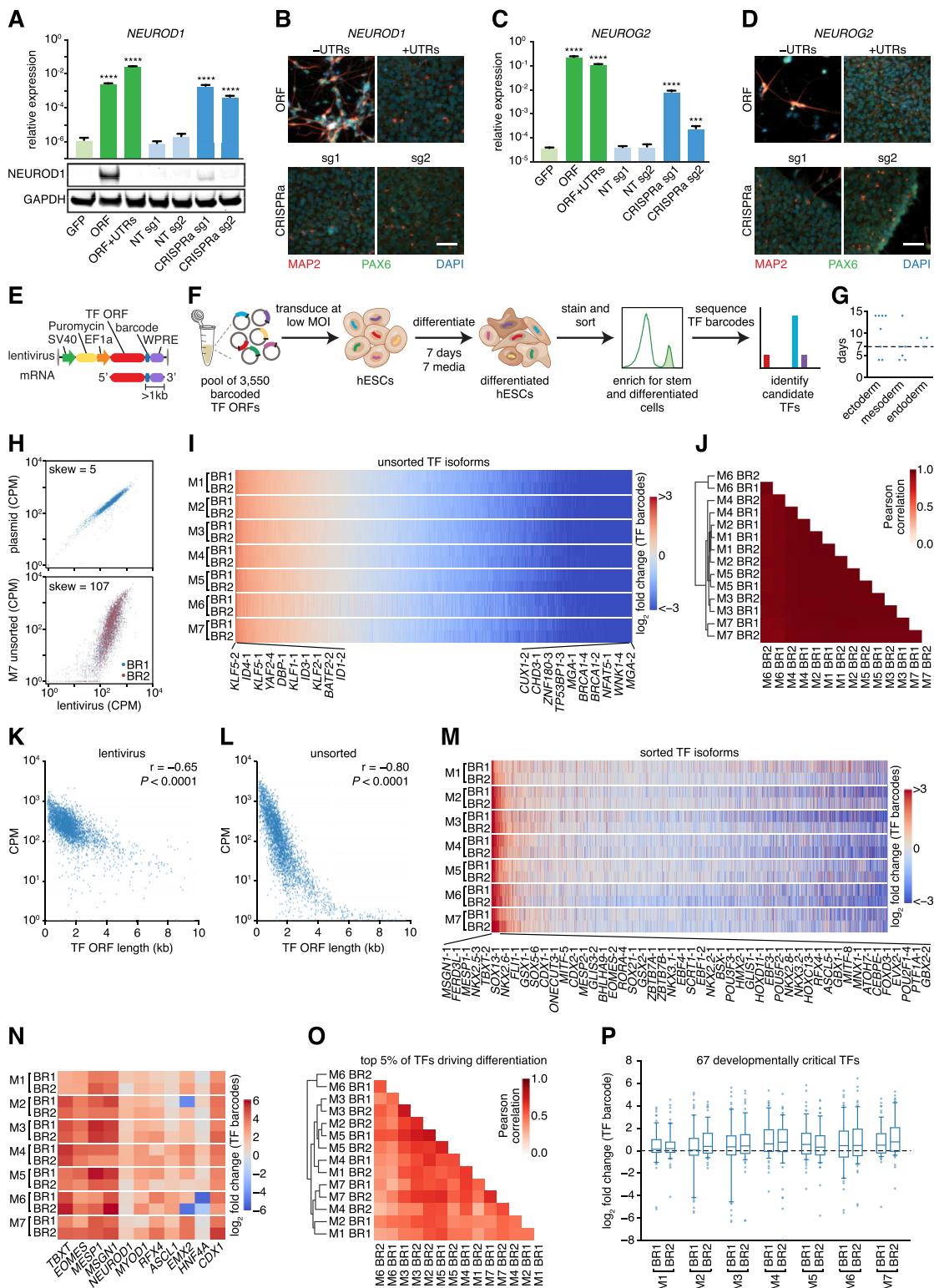
The average expression profiles of TF combinations were used for prediction. All possible combinations of single TF expression profiles were fitted against each measured double or triple TF profile to select the TF combination with the best fit. Linear regression (fit_intercept = False, positive = True), kernel ridge regression (alpha = 1), and random forest regression (max_depth = 4, n_estimators = 200) from scikit-learn¹⁰⁵ were evaluated and scored based on the coefficient of determination. Average expression profiles were scored based on Pearson correlation. To predict double and triple TF profiles using single TF profiles from the TF Atlas, the two datasets were integrated using Harmony.³² The average expression profiles from the TF Atlas differentiated cells were fitted against each double or triple TF profile. To reduce the number of possible combinations, average expression profiles from the TF Atlas were grouped by hierarchical clustering using 1 – correlation coefficient as the distance and Ward's linkage where indicated.

To predict TF combinations for reference cell types, expression profiles of double or triple TFs were estimated using the mean expression profiles from the TF Atlas differentiated cells. Individual TF profiles were grouped by hierarchical clustering as described above into 365 clusters for double TF profiles and 151 clusters for triple TF profiles to reduce the number of combinations. Group gene signatures for reference cell types from the human fetal cell atlas³⁰ were extracted using CellID⁸³ with default parameters. Cell type-specific gene signature scores were computed on all possible estimated expression profiles for multiple TFs. Predicted TF combinations were ranked by cell type-specific gene signature scores. Combinations that did not include any cell type-specific TFs (approximately 10%–50% of combinations depending on the cell type) were eliminated. For each cell type, up to 100 TFs that were significantly enriched (*FDR* < 0.05) based on the human fetal cell atlas³⁰ analysis were considered specific for that cell type. In cases with more than 100 significantly enriched TFs, the top 100 TFs with the highest expression relative to other cell types were included.

Statistics

Statistical tests were applied with the sample size listed in the text and figure legends. Sample size represents the number of independent biological replicates. Data supporting main conclusions represents results from at least two independent experiments. All graphs with error bars report mean ± s.e.m. values. Two-tailed *t*-tests were performed unless otherwise indicated. PRISM was used for basic statistical analysis and plotting (<http://www.graphpad.com>), and the R language and programming environment (<https://www.r-project.org>) was used for the remainder of the statistical analysis. Multiple hypothesis testing correction was applied where indicated.

Supplemental figures



(legend on next page)

Figure S1. Bulk TF screening in different cell culture media, related to Figure 1

(A–D) Comparison of ORF and CRISPRa. HUES66 hESCs were transduced with ORF, ORF with untranslated regions (UTRs), or SAM CRISPRa to upregulate *NEUROD1* (A and B) or *NEUROG2* (C and D) for 7 days. n = 4. Values represent mean \pm SEM. Scale bars, 50 μ m. ****p < 0.0001; ***p < 0.001. NT, non-targeting; sg, single guide RNA. (A and C) Expression of *NEUROD1* mRNA and protein (A) or *NEUROG2* mRNA (C). (B and D) Immunostaining of marker genes for neurons (MAP2) and neural progenitors (PAX6).

(E) MORF vector design. WPRE, Woodchuck hepatitis virus post-transcriptional regulatory element.

(F) Schematic of bulk TF screening in H1 hESCs. MOI, multiplicity of infection.

(G) Duration of established TF-driven differentiation protocols. Dashed line indicates the median.

(H) Scatterplots comparing the TF barcode distribution for the plasmid library, lentivirus library, and unsorted cells cultured in stem cell media (M7) after 7 days of TF ORF overexpression. CPM, counts per million; BR1 and BR2, biological replicates; skew, ratio of the 90th and 10th percentile CPM.

(I) Heatmap showing TF barcode counts in each media condition (M1–M7; [STAR Methods](#)) relative to the lentivirus library. The 10 most enriched and depleted TFs are labeled. Numbers after the TF gene name indicate the isoform.

(J) Heatmap showing pairwise Pearson correlation between each of the conditions in (I), ordered by hierarchical clustering.

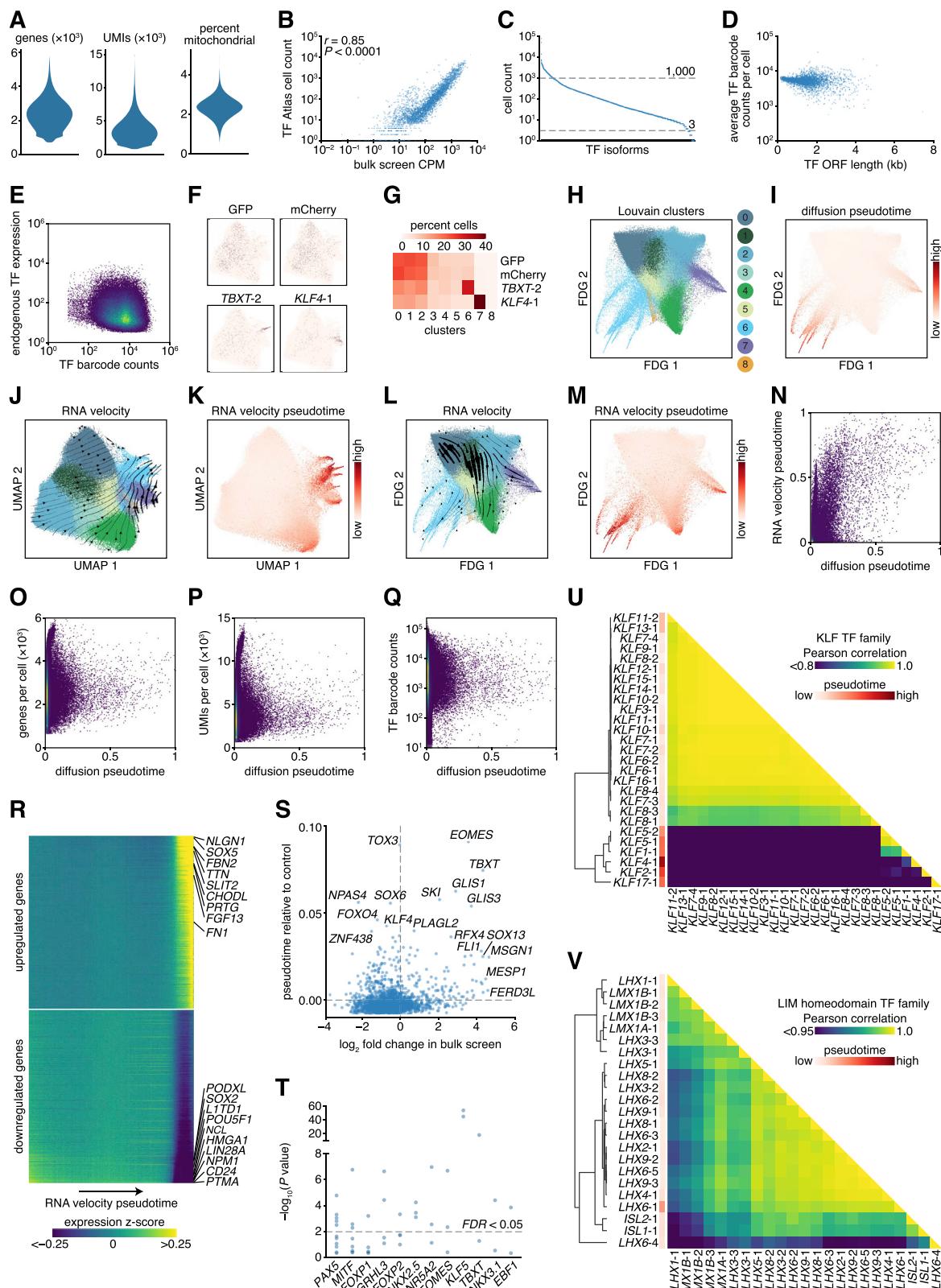
(K and L) Scatterplots showing the relation between TF barcode counts and ORF length for the lentivirus library (K) and unsorted cells (L), averaged across 7 media conditions. Pearson correlation coefficients (r) and associated p values are shown.

(M) Heatmap showing TF barcode counts in the sorted differentiated cells relative to stem cells. The 50 most enriched TFs are labeled.

(N) Subset of data in (M), highlighting TFs with known roles in development or differentiation.

(O) Heatmap showing the pairwise Pearson correlation between each of the conditions in (M), ordered by hierarchical clustering. The top 5% of TFs with the highest average fold change were evaluated.

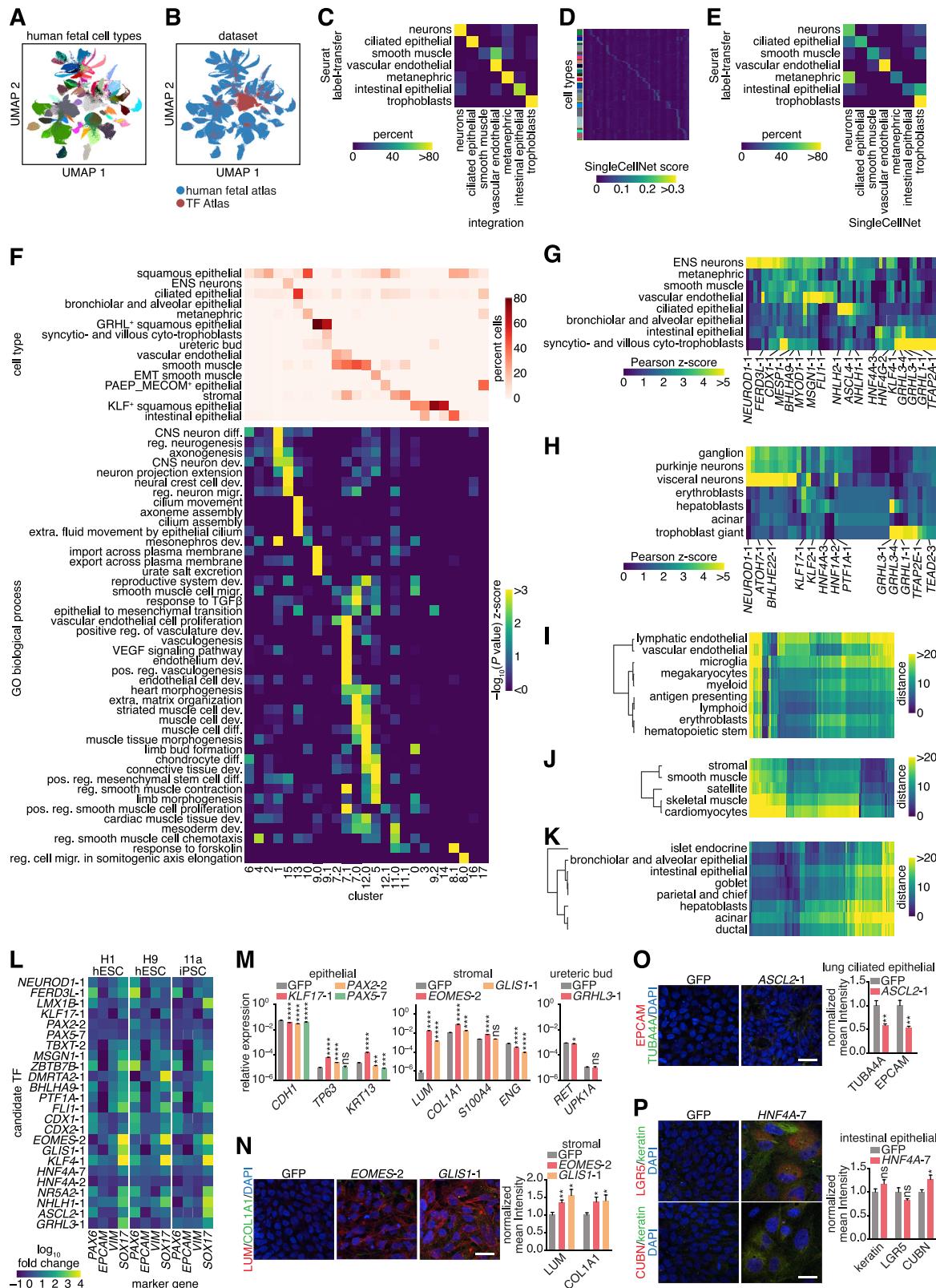
(P) Box plots showing enrichment of 67 developmentally critical TFs from Parekh et al.⁵ and (N). Whiskers indicate the 10th and 90th percentiles.



(legend on next page)

Figure S2. TF Atlas data quality control and pseudotime analysis, related to Figure 1

- (A) Violin plots showing distribution of genes, unique molecular identifiers (UMIs), and percent mitochondrial counts per cell in the TF Atlas.
- (B) Comparison of TF ORF representation between the bulk TF screen and the TF Atlas scRNA-seq.CPM, counts per million. Pearson correlation coefficients (r) and associated p values are shown.
- (C) Distribution of cells overexpressing each TF isoform. Dashed lines indicate subsampling thresholds.
- (D) Scatterplot showing the relation between average TF ORF expression per cell and TF ORF length.
- (E) Density scatterplot showing, for each cell, expression of the exogenous TF ORF measured via TF barcode counts and the corresponding endogenous TF.
- (F) UMAPs of TF Atlas scRNA-seq data highlighting cells with indicated ORF. Numbers after TF gene names indicate the isoform.
- (G) Heatmaps showing percentage of cells with the indicated TF ORF that were assigned to each Louvain cluster from Figure 1B.
- (H and I) Force-directed graph (FDG) representation of TF Atlas scRNA-seq data. Colors indicate Louvain clusters (H) and diffusion pseudotime (I).
- (J and K) UMAP of TF Atlas scRNA-seq data. (J) Stream plot of RNA velocities. Line weights indicate speed and colors indicate Louvain clusters. (K) Colors indicate RNA velocity pseudotimes.
- (L and M) FDG representations of (J) and (K), respectively.
- (N-Q) Density scatterplots showing RNA velocity pseudotime (N), genes (O), UMIs (P), and TF barcode counts (Q) over diffusion pseudotime for each cell.
- (R) Smoothened heatmap of the top 1,000 upregulated and downregulated genes over RNA velocity pseudotime. Genes are ordered by change over pseudotime.
- (S) Scatterplot comparing the differentiation results of the scRNA-seq pseudotime analysis to the bulk TF screen.
- (T) Significance of the pseudotime difference between cells expressing each TF isoform and those expressing controls. Subset of Figure 1H. Dashed line indicates FDR < 0.05.
- (U and V) Heatmaps showing pairwise Pearson correlation between TF potential vectors for the KLF (U) and LIM homeodomain (V) TF families. TFs are ordered by hierarchical clustering.



(legend on next page)

Figure S3. Mapping and validating differentiated cells against reference cell types, related to Figures 3 and 4

(A–C) Integration of TF Atlas differentiated cells and human fetal expression atlas³⁰ datasets. (A and B) UMAPs of scRNA-seq data with colors indicating reference cell types (A) and source dataset (B). (C) Heatmap showing the percentage of major cell types annotated by Seurat label-transfer that were mapped by dataset integration.

(D and E) SingleCellNet³³ annotation of TF Atlas differentiated cells.

(D) Heatmap showing the scores for TF Atlas cells (columns) against reference cell types (rows).

(E) Heatmap showing the percentage of major cell types annotated by Seurat label-transfer that were mapped by SingleCellNet.

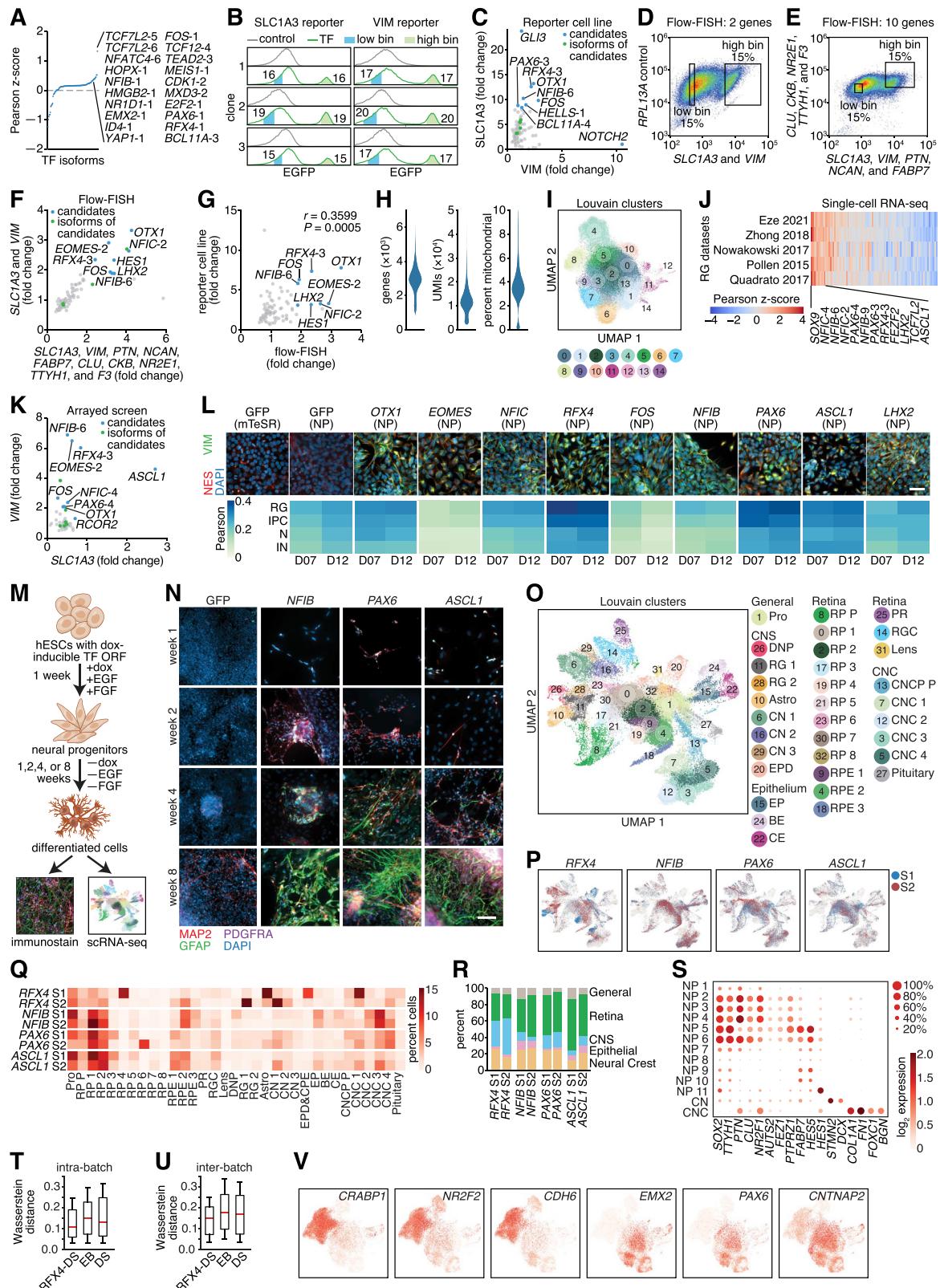
(F) Heatmaps showing percentage of cells from each cluster that mapped to the indicated reference cell type (top) and enrichment of gene ontology (GO) terms in differentially expressed genes (bottom). EMT, epithelial-mesenchymal transition; ENS, enteric nervous system. CNS, central nervous system; diff., differentiation; pos, positive; reg, regulation of; dev, development; migr, migration.

(G and H) Heatmaps showing correlation between TF potential vectors and expression profiles of reference cell types that matched (G) or did not match (H) by label-transfer. The top 10 TFs with the highest correlation for each cell type are included. Numbers after TF gene names indicate the isoform.

(I–K) Heatmaps showing distances between expression profiles of TF Atlas differentiated cells (columns) and reference cell types (rows). Values represent Euclidean distance in the latent space for the integrated datasets (Figures S3A and S3B). Cell types are ordered by hierarchical clustering. Cell types derived from the circulatory system (I), myocytes (J), and endoderm (K) are shown.

(L and M) Expression of marker genes measured by quantitative PCR after 7 days of candidate TF or GFP overexpression. n = 4. (L) Heatmap showing expression relative to GFP control. (M) Expression of marker genes in H1 hESCs.

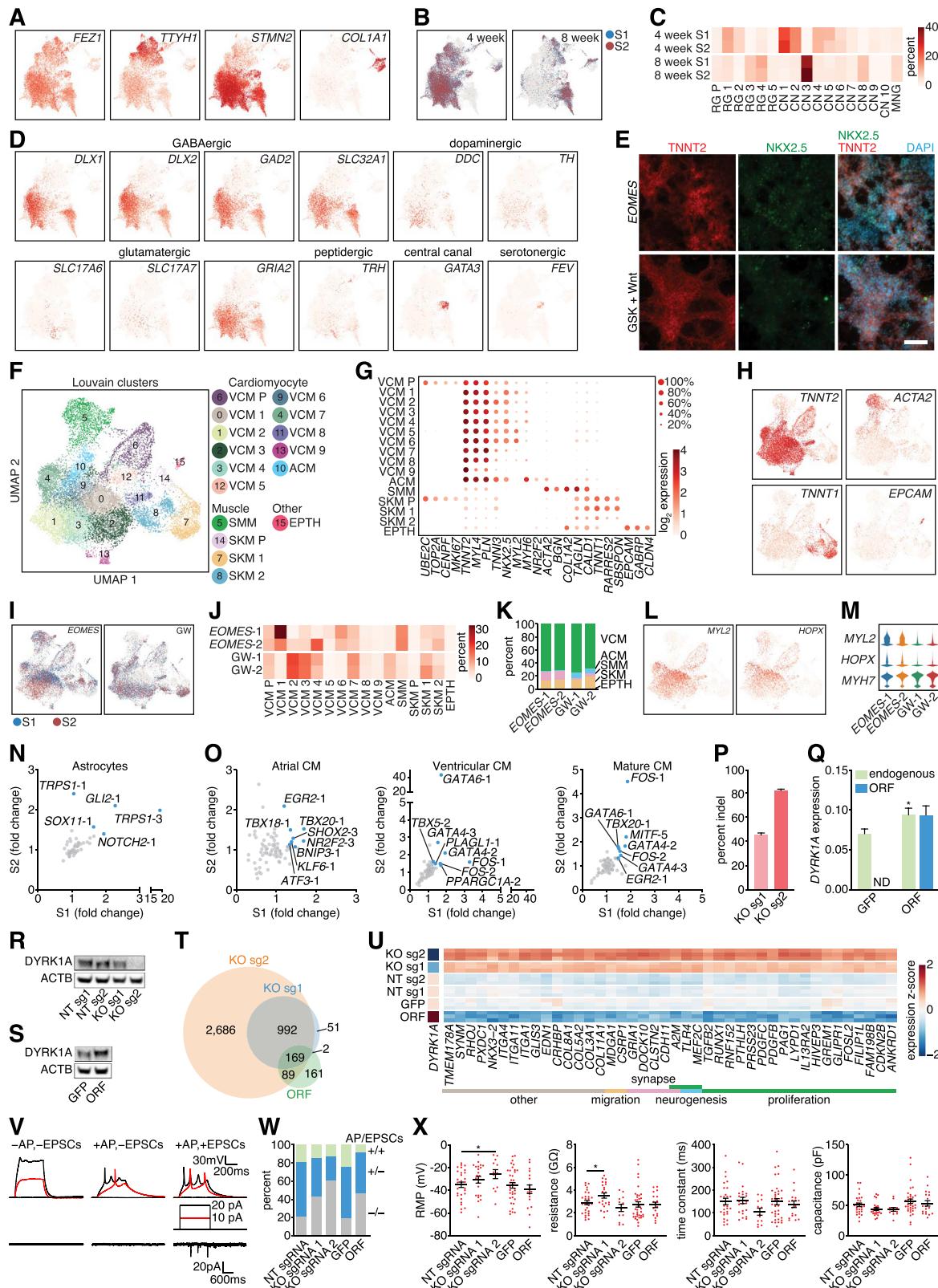
(N–P) Left, expression of marker genes measured by immunostaining in H1 hESCs after 7 days of TF ORF overexpression. Right, intensity of marker gene staining normalized to GFP control from n = 6 images. Scale bars, 25 μ m. Marker genes for stromal (N), lung ciliated epithelial (O), and intestinal epithelial (P) cells are shown. Values represent mean \pm SEM. ***p < 0.0001; **p < 0.001; *p < 0.01; ns, not significant.



(legend on next page)

Figure S4. Developing a targeted TF ORF screening platform for generating cell types of interest, related to Figure 5

- (A) Prediction of TFs for iNP differentiation based on correlation between TF potential vectors in the TF Atlas and reference iNP expression datasets. The top 20 TFs are listed. Numbers after TF gene names indicate the isoform.
- (B) FACS histograms showing EGFP expression in reporter cell lines with or without the targeted TF library. High and low bins indicate populations sorted for sequencing of TF barcodes.
- (C) Scatterplot showing TF enrichment in reporter cell line screens. n = 3.
- (D and E) Representative FACS plots showing expression of 2 (D) or 10 (E) NP marker genes labeled by pooled FISH probes. High and low bins indicate populations sorted for sequencing of TF barcodes.
- (F) Scatterplot showing TF enrichment in flow-FISH screens targeting 2 or 10 NP marker genes. n = 3.
- (G) Comparison of TF enrichment in reporter cell line and flow-FISH screens.
- (H–J) TF screening using scRNA-seq for 60,997 cells. (H) Violin plots showing distribution of genes, UMIs, and percent mitochondrial counts per cell. (I) UMAP of scRNA-seq data with colors indicating Louvain clusters. (J) Heatmap showing correlation between mean TF-induced expression profiles and human radial glia (RG) from reference datasets.^{120,128,150–152}
- (K) Scatterplot showing TF enrichment in the arrayed screen. Expression was measured by quantitative PCR. N = 3.
- (L) Top, expression of NP markers VIM and NES measured by immunostaining after 7 days of TF overexpression. Bottom, heatmap showing correlation between bulk RNA-seq expression profiles of iNPs and human fetal cortex cell types.¹²⁰ Cell culture media is indicated in parentheses. Scale bars, 50 µm. D07 and D12 indicate number of days of ORF overexpression. RG, radial glia; IPC, intermediate progenitor cell; N, neuron; IN, interneuron.
- (M) Schematic of spontaneous differentiation. dox, doxycycline; EGF, epidermal growth factor; FGF, fetal growth factor.
- (N) Expression of marker genes for neurons (MAP2), astrocytes (GFAP), and oligodendrocyte precursor cells (PDGFRA) in cells spontaneously differentiated for 1, 2, 4, or 8 weeks from iNPs produced by candidate TFs. Scale bars, 100 µm.
- (O–R) ScRNA-seq profiling of 53,113 cells that have been spontaneously differentiated from iNPs for 8 weeks. iNPs were derived using *RFX4*, *NFIB*, *ASCL1*, or *PAX6* with n = 2 biological replicates. Pro, uncommitted progenitors; RPs, retinal progenitors; RPE, retinal pigment epithelium; PRs, photoreceptors; RGCs, retinal ganglion cells; DNPs, dorsal neural progenitors; RG, radial glia; Astro, astrocytes; CNs, CNS neurons; EPD, ependyma; EPs, epithelial progenitors; BE, bronchial epithelium; CE, cranial epithelium; CNC, cranial neural crest; CNCPs, cranial neural crest progenitors; P, proliferative cells. (O and P) UMAP of scRNA-seq data with colors indicating Louvain clusters (O) and replicates from each TF (P), S1 and S2. (Q) heatmap showing the percentage of cells from each replicate that were grouped into each cluster. (R) Distribution of general cell types produced by each replicate.
- (S–V) ScRNA-seq profiling of iNPs differentiated using different methods. EB, embryoid body; DS, dual SMAD; NPs, neural progenitors; CNs, CNS neurons; CNC, cranial neural crest. Data represent n = 2 batch replicates with 15,211 *RFX4*-DS, 11,148 EB, and 16,421 DS. (S) Dot plot showing marker genes for each cluster. Circle size and color indicate percentage and expression level, respectively. (T and U) Box plots showing intra- (T) or inter- (U) replicate Wasserstein distances between cells. Whiskers indicate the 5th and 95th percentiles. (V) UMAP of scRNA-seq data highlighting genes that differentiate *RFX4*-DS-iNPs from EB- and DS-iNPs. Color indicates expression.



(legend on next page)

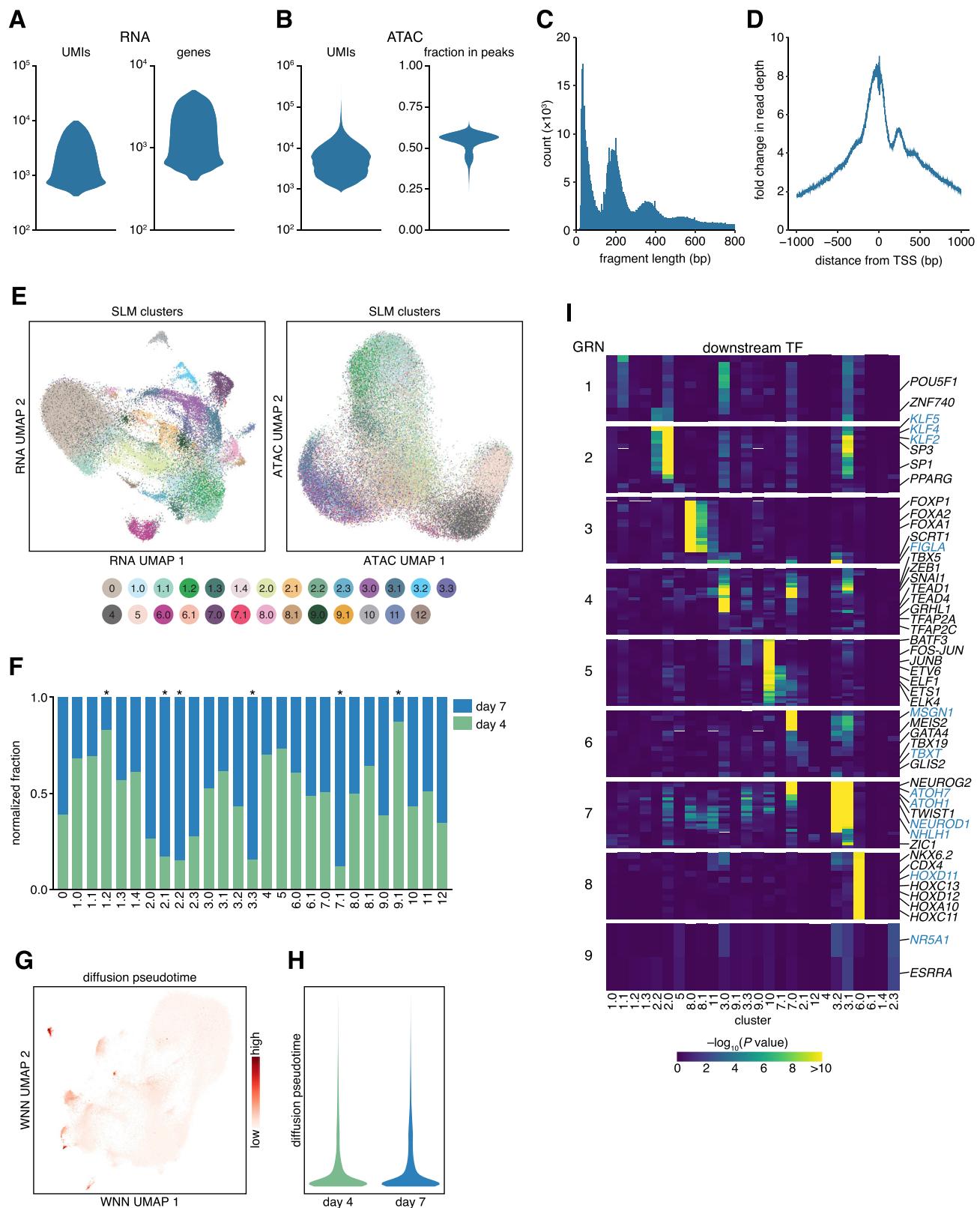
Figure S5. Characterization of *RFX4*-iNPs and *EOMES*-iCMs for sequential TF screening and disease modeling, related to Figure 5

(A–D) scRNA-seq profiling of 26,111 cells that have been spontaneously differentiated from *RFX4*-DS-iNPs for 4 or 8 weeks. Data represent n = 2 biological replicates. (A and B) UMAP of scRNA-seq data. Colors indicate expression of marker genes for major cell types (A) and replicates from each time point (B), S1 and S2. (C) heatmap showing the percentage of cells from each biological replicate that were grouped into each cluster. CNs, CNS neurons; RG, radial glia; MNGs, meninges; P, proliferative cells. (D) UMAP of scRNA-seq data showing marker genes for neuronal subtypes. Colors indicate gene expression.

(E–M) Characterization of *EOMES*-iCMs. (E) Expression of cardiomyocyte markers TNNT2 and NKX2.5 by immunostaining at day 30 after 2 days of *EOMES* induction or GSK and Wnt inhibition (GW). Scale bars, 100 μ m. (F) UMAP of scRNA-seq data from 16,698 cells that have been spontaneously differentiated for 4 weeks after *EOMES* induction or GW. Data represent n = 2 biological replicates. Colors indicate Louvain clusters. VCMs, ventricular cardiomyocytes; ACMs, atrial cardiomyocytes; SMMs, smooth muscle cells; SKMs, skeletal muscle cells; EPTHs, epithelial cells; P, proliferating cells. (G) Dot plot showing marker genes for each cluster. Circle size and color indicate percentage and expression level, respectively. (H and I) UMAP of scRNA-seq data with colors indicating marker gene expression (H) and replicates from each differentiation method (I), S1 and S2. (J) Heatmap showing the percentage of cells from each replicate that were grouped into each cluster. (K) Distribution of general cell types produced by each replicate. (L and M) Expression of marker genes for mature cardiomyocytes shown on UMAP (L) and violin plots (M).

(N and O) Scatterplots showing TF enrichment in sequential TF screens using flow-FISH. N = 2 replicates, S1 and S2. Numbers after TF gene names indicate the isoform. (N) Screen of 54 TFs in *RFX4*-iNPs for differentiation of astrocytes. After 7 days, cells were sorted for expression of astrocyte marker genes (NCAN, AQP4, FAM107A, and GFAP). (O) Screen of 80 TFs in *EOMES*-iCMs for production of atrial, ventricular, or mature cardiomyocytes. After 4 weeks, cells were sorted for expression of atrial (*MYBPHL*, *GJA5*, *MYH6*), ventricular (*MYL2*, *GJA1*, and *PLN*), and mature (*MYL2*, *TNNI3*, and *TNNT2*) cardiomyocyte marker genes.

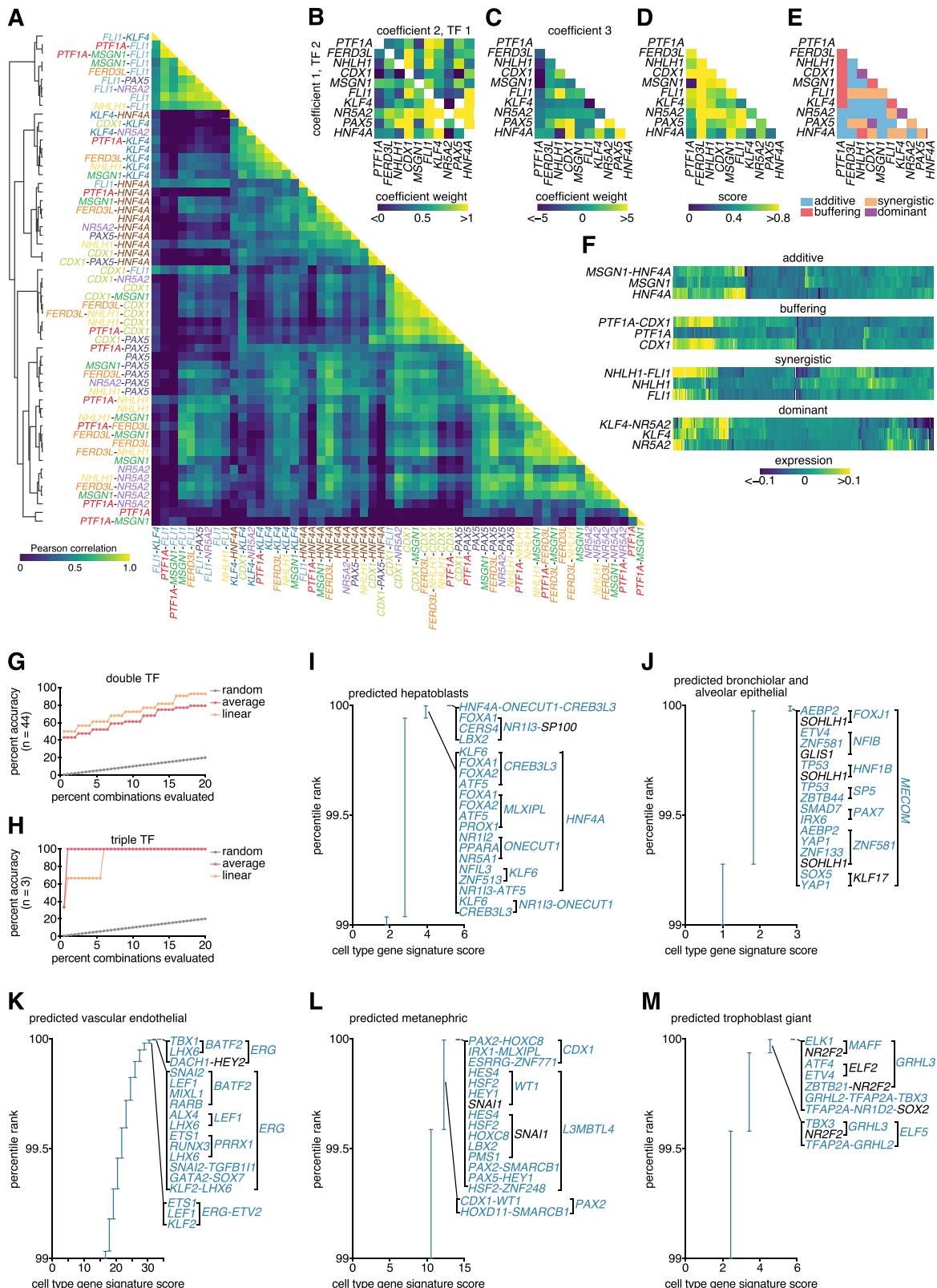
(P–X) Modeling effects of *DYRK1A* perturbation in *RFX4*-iNPs derived from 11a iPSCs. sg, single guide RNA; KO, knockout; NT, non-targeting. (P) Percent indels in *RFX4*-iNPs transduced with *DYRK1A* KO sgRNAs. n = 3. (Q) *DYRK1A* expression measured using quantitative PCR probes targeting the endogenous sequence or the ORF sequence. n = 4. (R and S) Western blots of *DYRK1A* at 7 days after transduction with Cas9 and *DYRK1A* KO sgRNAs (R) or *DYRK1A* ORF (S). (T and U) Bulk RNA-seq of *DYRK1A* perturbations to identify significantly differentially expressed genes (DEGs; FDR < 0.05). n = 3. (T) Venn diagram summarizing the DEGs. (U) Heatmap showing *DYRK1A* dosage-dependent DEGs. (V–X) Characterization of *DYRK1A* perturbations by electrophysiology. (V) Representative traces for neurons with or without evoked action potentials (APs) and spontaneous excitatory postsynaptic currents (EPSCs). (W) Proportion of neurons with or without AP and EPSCs from n = 31–45 neurons per condition. (X) Intrinsic membrane properties from n = 12–36 neurons with evoked AP per condition. Values represent mean \pm SEM. *p < 0.05; ND, not detected.



(legend on next page)

Figure S6. Joint profiling of chromatin accessibility and gene expression on a subset of TF ORFs, related to Figure 6

- (A and B) Violin plots showing distribution of UMIs and genes for scRNA-seq (A), as well as UMIs and fraction of reads in the top 500,000 peaks for scATAC-seq (B) from the joint profiling dataset. Data included 69,085 cells overexpressing 198 TF ORFs for 4 or 7 days.
- (C) Representative fragment histogram for scATAC-seq data using the first two megabases of chromosome 1.
- (D) Transcriptional start site (TSS) enrichment score for scATAC-seq data.
- (E) RNA (left) and ATAC (right) UMAP of joint profiling data. Colors indicate smart local moving (SLM) clusters.
- (F) Distribution of cells from each time point across clusters from Figure 6A. Asterisks indicate clusters with >75% cells from either time point.
- (G) Weighted nearest neighbor (WNN) UMAP of joint profiling data from Figure 6A, colored by diffusion pseudotime.
- (H) Violin plots comparing diffusion pseudotimes of each time point.
- (I) Heatmaps showing gene regulatory networks (GRNs) for downstream TFs nominated by evaluating motif enrichment in ATAC peaks with significant peak-gene associations. Only the 10 most significantly enriched TF ORFs are shown for each cluster. TF ORFs that were identified as top ORFs and downstream TFs are labeled in blue.



(legend on next page)

Figure S7. Combinatorial TF screening, modeling, and prediction, related to Figure 7

(A) Heatmap showing pairwise Pearson correlation between mean expression profiles from the combinatorial screen of 10 TF ORFs in combinations, including 44 doubles and 3 triples, as well as 10 singles. TF combinations are ordered by hierarchical clustering.

(B–F) Modeling effects of TF combinations with linear regression. (B–D) Heatmaps showing the coefficient weights (B and C) and score (D) for linear regression. (E) Annotated relationships for each TF combination based on the linear regression coefficients. (F) Heatmaps showing average expression profile of double TFs with those of respective single TFs for example combinations with annotated relationships.

(G–M) Predicting TF combinations using the TF Atlas. (G and H) Percent accuracy for different approaches to predict TFs for measured double (G) or triple (H) TF expression profiles. For comparison to the combinatorial TF screen dataset, profiles of the 10 corresponding TFs from the TF Atlas were used. (I–M) Cell type prediction results for triple TF profiles. Predicted combinations for hepatoblasts (I), bronchiolar and alveolar epithelial cells (J), metanephric cells (K), vascular endothelial cells (L), and trophoblast giant cells (M) are shown. As gene signature scores were discrete, the percentile ranks were reported as ranges. TFs that are part of known combinations, developmentally critical, or specifically expressed in the target cell types are indicated in blue.

Update

Cell

Volume 187, Issue 12, 6 June 2024, Page 3161

DOI: <https://doi.org/10.1016/j.cell.2024.04.038>

Correction

A transcription factor atlas of directed differentiation

Julia Joung, Sai Ma, Tristan Tay, Kathryn R. Geiger-Schuller, Paul C. Kirchgatterer, Vanessa K. Verdine, Baolin Guo, Mario A. Arias-Garcia, William E. Allen, Ankita Singh, Olena Kuksenko, Omar O. Abudayyeh, Jonathan S. Gootenberg, Zhanyan Fu, Rhiannon K. Macrae, Jason D. Buenrostro, Aviv Regev, and Feng Zhang*

*Correspondence: zhang@broadinstitute.org

<https://doi.org/10.1016/j.cell.2024.04.038>

(Cell 186, 209–229.e1–e26; January 5, 2023)

In our Resource article, the primer sequences for TF NGS Fwd 1–10 in Table S7 were incorrect. This error does not affect the results or conclusions of our paper, and the supplemental table has now been corrected in the online version. We apologize for any confusion this error may have caused.

