Stem Cell Reports
Article

ISSCR

OPEN ACCESS

Decoding Neuronal Diversification by Multiplexed Single-cell RNA-Seq

Joachim Luginbühl,[1,2] Tsukasa Kouno,[1,2] Rei Nakano,[1,3] Thomas E. Chater,[4] Divya M. Sivaraman,[1,2,5] Mami Kishima,[1,2] Filip Roudnicky,[6] Piero Carninci,[1,2] Charles Plessy,[1,2] and Jay W. Shin[1,2,*]

[1]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan
[2]RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Kanagawa 230-0045, Japan
[3]Nihon University, College of Bioresource Sciences, Laboratory of Veterinary Radiology, Fujisawa, Kanagawa 252-0880, Japan
[4]RIKEN Center for Brain Science, Wako-Shi, Saitama 351-0198, Japan
[5]Sree Chitra Tirunal Institute for Medical Sciences and Technology, Department of Pathology, Thiruvananthapuram 695-011, Kerala, India
[6]ETH Zurich, Institute of Pharmaceutical Sciences, 8057 Zurich, Switzerland
*Correspondence: jay.shin@riken.jp
https://doi.org/10.1016/j.stemcr.2021.02.006

SUMMARY

Cellular reprogramming is driven by a defined set of transcription factors; however, the regulatory logic that underlies cell-type specification and diversification remains elusive. Single-cell RNA-seq provides unprecedented coverage to measure dynamic molecular changes at the single-cell resolution. Here, we multiplex and ectopically express 20 pro-neuronal transcription factors in human dermal fibroblasts and demonstrate a widespread diversification of neurons based on cell morphology and canonical neuronal marker expressions. Single-cell RNA-seq analysis reveals diverse and distinct neuronal subtypes, including reprogramming processes that strongly correlate with the developing brain. Gene mapping of 20 exogenous pro-neuronal transcription factors further unveiled key determinants responsible for neuronal lineage specification and a regulatory logic dictating neuronal diversification, including glutamatergic and cholinergic neurons. The multiplex scRNA-seq approach is a robust and scalable approach to elucidate lineage and cellular specification across various biological systems.

INTRODUCTION

The brain consists of a wide range of neuronal subtypes. Understanding the mechanism that underlies neuronal diversification is critical to the study of brain development and neurodegenerative disease. The recent discoveries that fully differentiated somatic cells can be reprogrammed into alternative cell fates opened up exciting avenues to study cellular specification and regenerative medicine (Chanda et al., 2014; Liu et al., 2012; Smith et al., 2016; Takahashi et al., 2007; Vierbuchen et al., 2010; Shin et al., 2012). Despite the rapid development of cell reprogramming, the possibility to identify the gene regulatory logic that underlies neuronal reprogramming has remained elusive, which has limited our understanding of the cellular plasticity of stem cells and cellular diversification during development and reprogramming.

Several studies have aimed to identify neurogenic transcription factors (TFs) that allow the reprogramming of major neuronal subtypes, where distinct combinations of TFs have been shown to create dopaminergic neurons and cholinergic motor neurons, which are selective targets of degeneration in patients with Parkinson disease and amyotrophic lateral sclerosis, respectively (An et al., 2016; Caiazzo et al., 2011; Kim et al., 2011; Liu et al., 2013; Mazzoni et al., 2013; Pfisterer et al., 2011; Son et al., 2011). Moreover, *in silico* predictive models (Rackham et al., 2016) and combinatorial screens (Chen et al., 2015; Liu et al., 2018) provide high-throughput approaches to identify various combination of TFs involved in cell reprogramming. However, these approaches often require complex and costly experimental setups, which is further exacerbated by the time and labor-intensive retesting and validation of newly identified candidates.

Multiplexing offers a more efficient and scalable strategy, where a collection of genes can be perturbed either by small hairpin RNA or CRISPR in a single experiment. This is followed by a phenotypic selection and targeted DNA sequencing to reveal candidate genes that attribute to a pre-defined cellular phenotype (Chen et al., 2015; Hsu et al., 2014; Liu et al., 2018; Shalem et al., 2015). While the approach is highly scalable, it relies on a limited number of phenotypic readouts or requires specific markers to enrich for target cell populations, which are often not available or limited to a single cell type. The advancement in single-cell RNA sequencing (scRNA-seq) has opened up new opportunities to profile gene expression profiles of heterogeneous tissues at the single-cell resolution without the need for cell selection (Adamson et al., 2016; Dixit et al., 2016; Nowakowski et al., 2017; Picelli et al., 2014). Moreover, adaptation of scRNA-seq with CRISPR/sgRNA led to reconstruction of regulatory networks involved in immune activation (Shifrut et al., 2018), cholesterol biogenesis (Replogle et al., 2020), and zygotic genome activation (Alda-Catalinas et al., 2020). However, how TFs work in concert to drive neuronal diversification and profiling scRNA-seq with multiplexed TFs remain unexplored.

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

To address these challenges, we introduce a strategy that combines multiplexing ectopic expression of full-length TFs with scRNA-seq to retrospectively assign reprogrammed neurons with exogenous and endogenous expression profiles. We identify key determinants and regulatory processes governing the classification of neuronal subtypes in a single experiment. Single-cell RNA-seq of induced neurons generated by the multiplexed transduction of 20 pro-neuronal TFs reveals unique combinations of key TFs and genetic programs orchestrating the plasticity of human fibroblasts into various neuronal subtypes, including glutamatergic and cholinergic neurons. We demonstrate that the method efficiently gains deep insights into biological processes governing cell plasticity and cell fate decisions, and to systematically dissect regulatory processes controlling neuronal diversification during cell reprogramming.

RESULTS

Inducing a Heterogeneous Population of Human Neuronal Subtypes

We established a system to identify TFs governing direct cell reprogramming toward diverse neuronal subtypes by infecting multiplexed TFs followed by scRNA-seq (Figure 1A). We first generated a pool of lentiviruses encoding TFs previously implicated in neuronal reprogramming, either as pioneering factors that can access closed chromatin (*ASCL1*, *NEUROG2*) (Smith et al., 2016; Wapinski et al., 2013), as factors that increase the efficiency of reprogramming (*POU3F2*, *ZIC1*, *OLIG2*, *NEUROD1*) (Pang et al., 2011; Vierbuchen et al., 2010) or as factors that convert human fibroblasts into GABAergic (*DLX1*, *DLX2*) (Victor et al., 2014), cholinergic (*ISL1*) (Son et al., 2011), serotonergic (*FEV*) (Xu et al., 2016), or dopaminergic (*FOXA2*, *NR4A2*, *PITX3*) (Liu et al., 2012; Pfisterer et al., 2011) neurons when co-expressed with other TFs (Figure S1A). These 13 TFs were also upregulated during human induced pluripotent stem cell (iPSC) differentiation into neuronal progenitor cells (NPCs) (Figure S1A). We also selected seven TFs that were upregulated during the differentiation but were not previously implicated in direct neuronal reprogramming. Altogether 20 TFs, hereafter termed the "TF-pool," were individually tuned for multiplicity of infection (MOI) to allow for ~85% of fibroblasts to express mixed combinations of 2–6 TFs after pooled infection (Figures S1B and S1C).

Transduction of the TF-pool in human dermal fibroblasts followed by a serial application of neuronal induction medium for 14 days and neuronal maturation medium for the subsequent 7 days led to heterogeneous neuron-like morphologies, exhibiting multiple elongated dendrites from the cell body and, in some cases, star shape-like features mimicking astrocytes (Figure 1B). At 9 days post-infection (dpi), 78.6% of TF-pool-induced (TFi) fibroblasts stained positive for the immature neuronal marker TUBB3 and 34.7% expressed MAP2, a microtubule-associated protein expressed specifically in neurons (Gelles et al., 1988) (Figures 1B and 1C). By 21 dpi, TUBB3+ and MAP2+ populations increased to 93.5% and 42.4%, respectively, indicating progressive transdifferentiation toward the neuronal lineage. Consistent with TUBB3 and MAP2 protein expressions, qPCR revealed upregulation of neuronal marker transcripts (*MAP2*, *NRCAM*, *NEUN*, *SYN1*) and neuronal subtype marker transcripts (*SLC17A7* [glutamatergic neurons], *GABRA1* [GABAergic neurons], *TH* [dopaminergic neurons], and *CHAT* [cholinergic neurons]) as well as significant downregulation of fibroblast marker expression (*VIM*, *SNAI1*) starting at 7 dpi (Figure 1D).

Previous studies have shown that mouse and human fibroblasts can be directly converted to induced neurons solely by chemical cocktails of small molecules (Hu et al., 2015; Li et al., 2015). We also found that application of neuronal induction medium to fibroblasts (hereafter termed chemical-induced [Chi] cells) generated TUBB3+ neurites, upregulated the RNA expression of *NEUN*, *SYN1*, *SLC17A7*, and *TUBB3*, and downregulated the expression of *VIM* and *SNAI1* (Figures 1B–1D). However, CHi failed to express *MAP2* at any time point analyzed, indicating that small molecules/growth factors, in the absence of key TFs, were insufficient to elicit neuronal maturation in human dermal fibroblasts. Consistently, neuronal complexity, measured as number of branch points and average neurite length, was significantly lower at 9 dpi (8.5- and 1.8-fold, respectively) and at 21 dpi (5.6- and 4.8-fold, respectively) when fibroblasts were treated with the neuronal induction medium alone compared with TF-pool induction (Figures S1E and S1F). Immunofluorescence at 21 dpi indicated that TFi exhibited a heterogeneous population of cells expressing glutamatergic (VGlut1+; ~29%), GABAergic (GABA+; ~23%), cholinergic (CHAT+; ~19%), and dopaminergic (TH+; ~4%) subtype-specific genes (Figures 1E and 1F). In contrast, only ~15% of CHi expressed VGlut1 at 21 dpi and none of the other subtype-specific genes were induced by small molecules alone. Together, pooled overexpression of pro-neuronal TFs generated a heterogeneous population of both immature and mature neuronal subtypes, while small molecules partially reprogrammed fibroblasts toward the glutamatergic fate at low efficiency.

Molecular Characterization of Induced Neurons Using scRNA-Seq

To gain further insights into the cellular heterogeneity of TFi and CHi fibroblasts, we used droplet-based massively parallel scRNA-seq (Zheng et al., 2017) to profile 2,092

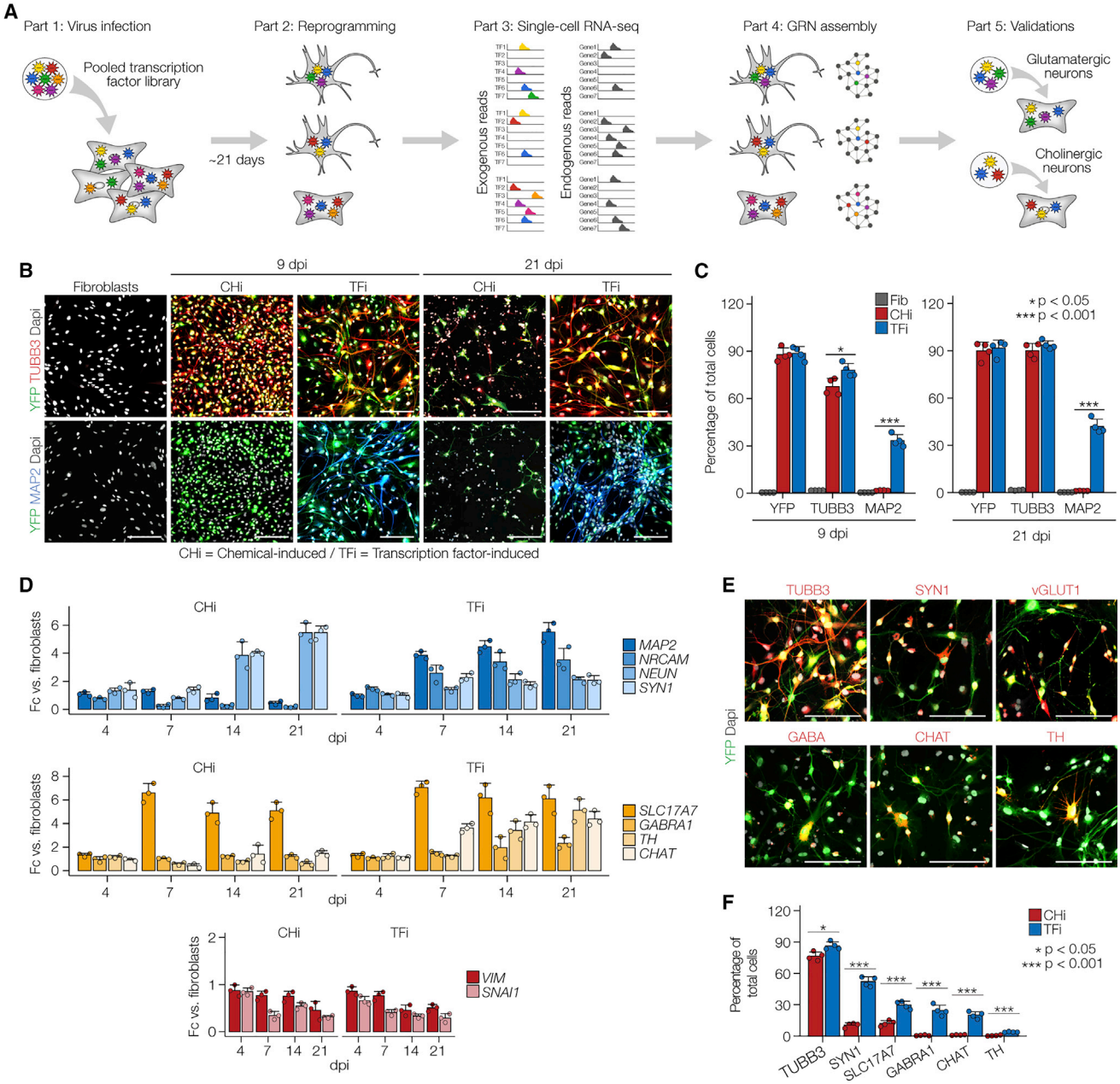


Figure 1. Generation of a Heterogeneous Population of Human Induced Neurons

(A) Overview of the single-cell TF multiplex pipeline.

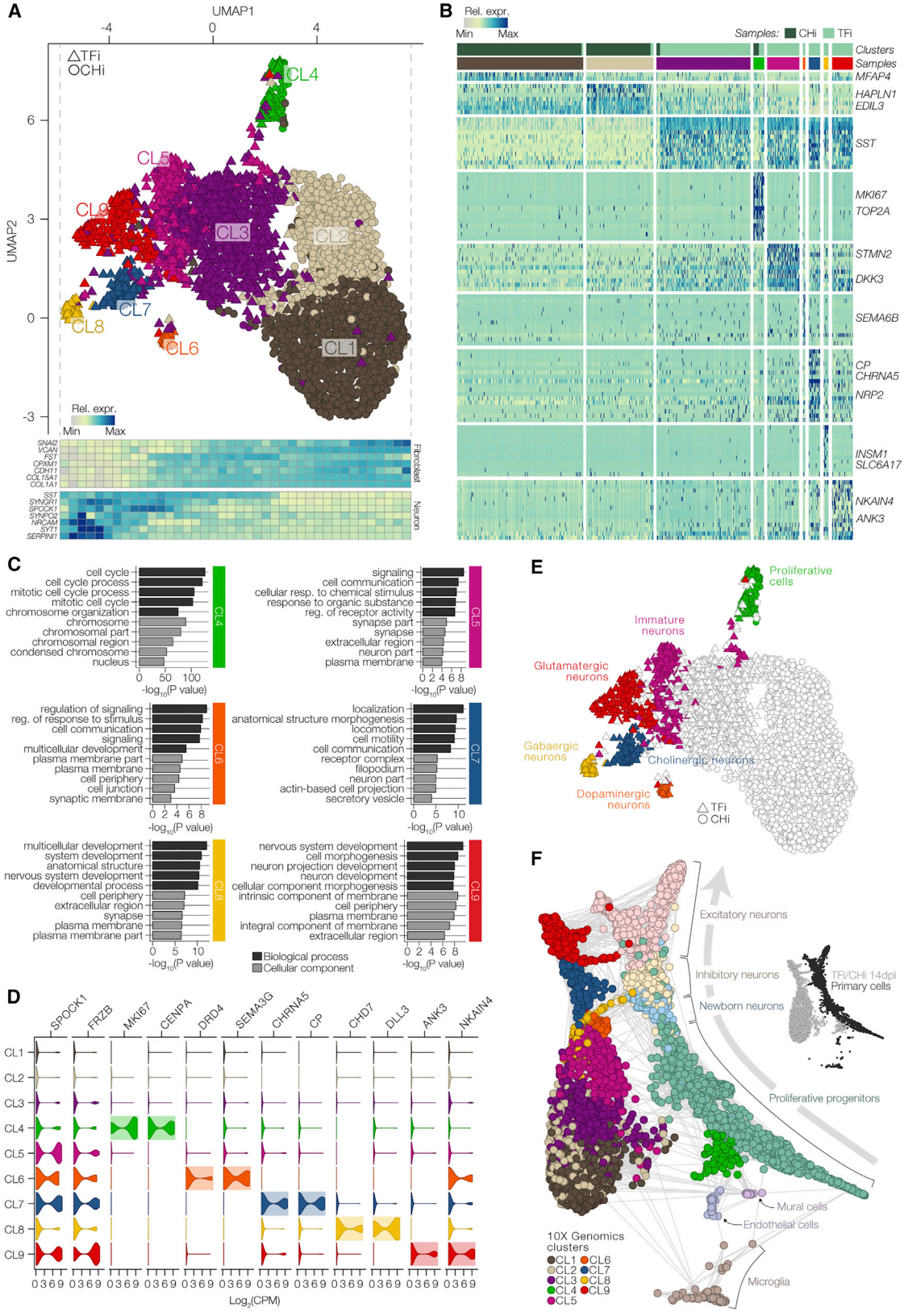
(B) Immunostaining for canonical neuronal marker genes of fibroblasts at day 0 and CHi and TFi at 9 and 21 dpi. Scale bars, 100 μm. YFP (green) marks infected cells and cell nuclei were visualized using DAPI nuclear stain (gray).

(C) Quantification of immunostainings in (B). n = 4 independent experiments, unpaired Student's t test. Error bars represent mean + SD.

(D) qPCR for pan-neuronal marker genes (*MAP2*, *NRCAM*, *NEUN*, *SYN1*; top), canonical neuronal subtype markers (*SLC17A7*, *GABRA1*, *TH*, *CHAT*; middle), and canonical fibroblast markers (*VIM*, *SNAI1*; bottom). n = 3 independent experiments, unpaired Student's t test. Error bars represent mean + SD.

(E) Immunostainings for canonical neuronal subtype markers (red) of TFi at 21 dpi. Scale bars, 100 μm. YFP (green) marks infected cells and cell nuclei were visualized using DAPI nuclear stain (gray).

(F) Quantification of immunostainings in (E). n = 4 independent experiments, unpaired Student's t test. Error bars represent mean + SD.



(legend on next page)

CHi and 1,900 TFi at 14 dpi. Uniform Manifold Approximation and Projection (UMAP) (Mcinnes et al., 2018) distinctively separated CHi and TFi fibroblasts along the first dimension, UMAP1 (Figures 2A and S2A–S2D), and revealed activation of neuronal differentiation genes (*NRCAM*, *STMN2*, *SST*, *DKK3*) and synapse formation genes (*SYT1*, *SERPINI1*, *SYNGR1*, *SYNPO2*), which was accompanied by a decrease in fibroblast-specific genes (*SNAI2*, *THY1*) (Figure 2A). Other suppressed genes included extracellular matrix genes (*COL3A1*, *COL15A1*, *EDIL3*, *HAPLN1*, *CPXM1*, *MFAP4*), reflective of the morphological changes that occur during the cell transformation toward the neurons. UMAP analysis further partitioned TFi cells into several transcriptionally distinct clusters (clusters 4–9 [CL4–CL9]). Cells in CL4 dominantly expressed cell-cycle-related genes, including *MIKI67* and *TOP2A*, and showed enrichment of gene ontology (GO) terms related to cell division, suggesting that CL4 cells did not successfully exit the cell cycle and failed to initiate the reprogramming process (Table S1). Interestingly, CL5–CL9 showed enrichment of genes and GO terms related to nervous system development and neurogenesis (Figures 2B, 2C, S2E, and S2F).

To further annotate the remaining clusters, from CL5 to CL9, we first interrogated top differentially expressed genes (Seurat [Satija et al., 2015]; $p < 10^{-20}$) for known neuronal subtype marker genes. The analysis revealed CL5 to associate with immature neurons, CL6 with the dopaminergic neuron program (*DRD4*, *SEMA3G*), CL7 with the cholinergic neuron program (*CHRNA5*, *CP*), CL8 with the GABAergic neuron program (*CHD7*, *DLL3*), and CL9 with the glutamatergic neuron program (*ANK3*, *NKAIN4*) (Figures 2D, 2E, and S2E). To support these findings, we used a publicly available scRNA-seq dataset of human primary cortical and medial ganglionic eminence brain (Nowakowski et al., 2017) to infer cellular relationships among the cells in a force-directed k-nearest neighbors graph (Weinreb et al., 2018) (Figure 2F). The comparative analysis aligned the primary brain cells along a developmental progression from proliferative progenitor cells, immature neurons, and finally to mature inhibitory and excitatory neurons. Consistent with our previous results, CL4 cells expressing cell-cycle genes positioned near proliferative progenitors and non-neuronal cells, whereas all other clusters positively aligned with primary cells along the same developmental trajectory. Importantly, clusters CL6–CL9 exhibiting neuronal subtype markers aligned along the cellular trajectory with newborn neurons, and mature inhibitory and excitatory neurons. Cluster CL3, which represented almost half of the TF-induced data, showed lower expression of neuronal genes as compared with CL4 and lower expression of fibroblast genes as compared with CL1–CL2. Furthermore, the placement of CL3 in both UMAP and force-directed k-nearest neighbor graphs indicated that these cells were still in transition or partially reprogrammed toward neurons (Figures 2A and 2F). Notably, CHi cells, largely represented by CL1 and CL2, also showed neuronal GO terms, which is consistent with our notion that the chemical cocktail alone upregulated neuronal genes (Figure S2F). However, high expression levels of fibroblast-specific genes, including CHD11 and SNAI2, suggest a failure to suppress the fibroblastic network in the absence of exogenous TF expression. Together, these data revealed that transduction of the TF-pool converted fibroblasts into a heterogeneous population of cells exhibiting distinct neuronal subtype-specific molecular signatures and congruent progression of neuronal development.

Distinguished Detection of Exogenous and Endogenous Transcripts along Developmental Trajectories

Next, to identify which TF in the TF-pool was responsible for inducing distinct neuronal subtype specification, we performed single-cell Smart-seq2 to achieve high gene coverage across the transcripts. We relied on full-length reads to extract nucleotides at the 5′ and 3′ junctions of exogenous open reading frames (ORFs) where distinctive alignment of reads to the specific 5′ and 3′ junctions allows to discriminate between exogenous (ORFs with the *att*B Gateway cloning sequences) and endogenous (ORFs without *att*B) gene expression (Figure 3A).

First, to benchmark the accuracy and sensitivity of detecting exogenous transcripts, we infected fibroblasts with single TFs, two combinations of 10 TFs, and the

Figure 2. Molecular Characterization of Induced Neurons using scRNA-Seq

(A) Top: visualization of droplet-based scRNA-seq data from CHi and TFi at 14 dpi using UMAP (n = 3,865 cells). The detected clusters are indicated by different colors. Bottom: heatmap of the relative expression of canonical fibroblast and neuron markers along UMAP1.

(B) Heatmap of the relative expression of top marker genes for each cluster in the UMAP plot in (A).

(C) GO analysis of cluster-specific marker genes in clusters CL4–CL9. Shown are the top 5 GO terms related to biological process (dark gray) and cellular component (light gray) for each cluster; colors as in (A).

(D) Violin plots of $\log_2$-transformed counts per million (CPM) values of marker genes in all clusters.

(E) Annotation of TFi clusters CL4–CL9 based on genes differentially expressed between each cluster.

(F) Relationship of CHi and TFi to primary human brain cells in a force-directed k-nearest neighbors graph created using SPRING. Primary cells are colored in light colors, induced cells are colored by cluster as in (A).

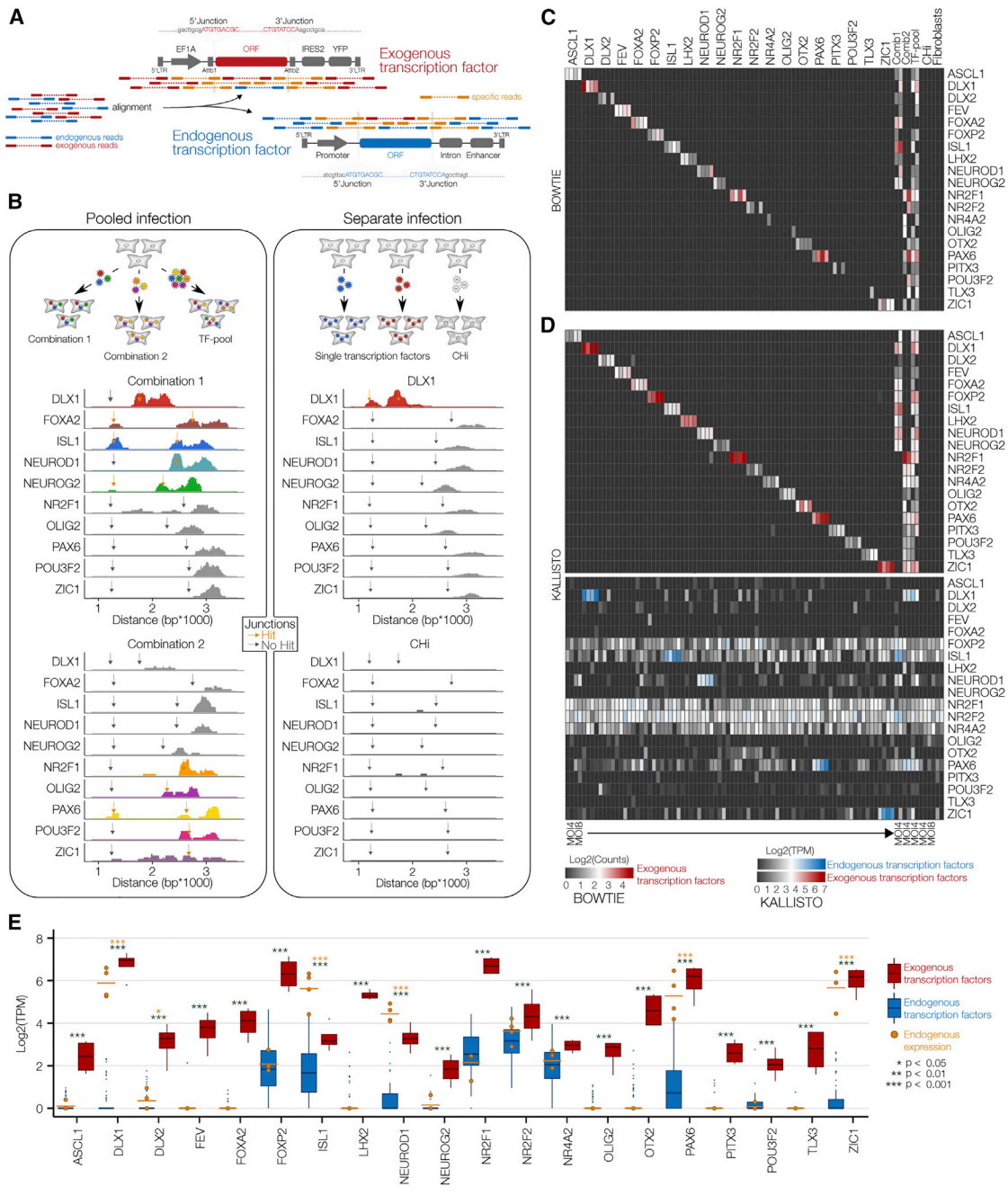


Figure 3. Distinguished Detection of Exogenous and Endogenous Transcripts

(A) Schematic depicting the strategy to distinguish exogenous and endogenous sequencing reads.

(B) Bulk RNA-seq on pooled and separately infected fibroblasts. Horizontal dimension, distance from the 5′ end of the EF1A promoter; vertical dimension, number of aligned paired-end reads. Gray arrows (no overlap) and golden arrows (overlap) mark 5′ and 3′ junctions of exogenous ORFs.

(C and D) Heatmaps showing $\log_2$-transformed count values of exogenous TFs after alignment using Bowtie (C) and $\log_2$-transformed tags per million (TPM) values of exogenous and endogenous TF pairs after trimming junction sequences to ~100 base pairs and alignment using Kallisto (D). For individually infected fibroblasts and CHi, two replicates at an MOI of 4 and two replicates at an MOI of 8 were included. For pooled infected fibroblasts, two replicates at an MOI of 4 were included.

(legend continued on next page)

complete (20) TF-pool at two MOIs, low and high (Figures 3B and S3B). Sequence alignment to the exogenous transcript model using Bowtie (Langmead et al., 2009) resulted in 24.4% false-negative events, possibly due to the exclusion of multi-mapping reads (Figure 3C). To reduce the number of false-negative events, we implemented the pseudo-alignment tool Kallisto, providing the advantage of assigning multi-mapping reads to transcripts without pinpointing exactly how the sequences of the reads and transcripts align (Bray et al., 2016). This approach reduced the occurrence of false-negative events to 6.3% and yielded highly specific and sensitive detection efficiencies of 97.5% and 87.5% in individual infection and pooled infection samples, respectively (Figures 3D and S3C). Specific detection of lentivirus-mediated expression was further confirmed by revealing significantly higher mean expression of exogenous TFs compared with the mean expression of corresponding endogenous TFs ($p < 0.05$; Figure 3E).

Subsequently, we performed Smart-seq2 on TFi and CHi cells at an early (9 dpi) and late (21 dpi) time point (Figure 4A). Cells were sequenced at an average depth of ~2.6 million reads and a median of 10,143 genes per cell. We used unsupervised hierarchical clustering and UMAP (Figures S4A–S4D) to reveal that cells clustered largely by type rather than batch and that we could assign individual cells with exogenous TF expression (Figures 4B, 4C, S4E, and S4F). Most notably, 30%–40% of TFi 9 dpi and TFi 21 dpi cells clustered as CL5 (Figure 4B). CL5 showed the highest degree of neuronal reprogramming as indicated by GO terms associated to nervous system development, synaptic transmission, cholinergic regulation of neuron differentiation; CL3 and CL4 showed GO terms related to cytokine-mediated signaling pathway and blood vessel development, respectively (Figure S5A). This suggests that cells acquired a neuronal program fairly early in the reprogramming process and only a fixed number of cells undergo neuronal reprogramming. Based on the TF enrichment in CL5, it is plausible that only the cells that acquired the necessary TF combinations induced neuronal conversion. The exogenous genes detected in day 9 and 21 were similar, whereas the exogenous factors were significantly different between clusters CL3 versus CL5 or CL4 versus CL5. This suggests that cells that acquired a neuronal program, in this case CL5, harbor a distinct set of TF combinations. Indeed, we detected over-representation of *PAX6*, *NEUROG2, POU3F2, ZIC1*, and *FEV* in CL5 as compared with other clusters (Figure S5B). Overall, these data indicate that a distinct set of TF combinations was necessary, as opposed to the presence of pre-determined cells, to initiate cell conversion toward neurons as early as day 9.

To investigate the dynamic progression of neuronal reprogramming, we placed the cells in pseudo-temporal order (Trapnell et al., 2014) based on differentially expressed genes between 9 and 21 dpi (qval < 0.01) (Figure S4G). We found that genes along pseudo-time were enriched for GO terms related to several developmental lineages (Figures S6H and S6I; Table S2), including neurogenic genes (*NRCAM*, *SFRP1*, *SNAP25*, and *SYT1*) and genes regulating the development of mesodermal tissues: bone (*BMP4*), kidney (*FAT4*), and endothelial cells (*PGF* and *VEGFA*) (Figures S6J–S6L). Because genes involved in cell reprogramming generally overlap with developmental genes (Masserdotti et al., 2016), we reordered the cells only using genes implicated in the developmental processes (~3,000 genes, Figure 4D; Table S3). This secondary pseudo-temporal ordering revealed clear bifurcation of cells into two main trajectories: branch 1 associated with non-neuronal developmental fates (Figures 4E and 4F; Table S4) and branch 2 associated with the neuronal lineage based on gene program (fold-change > 4; $p < 0.05$).

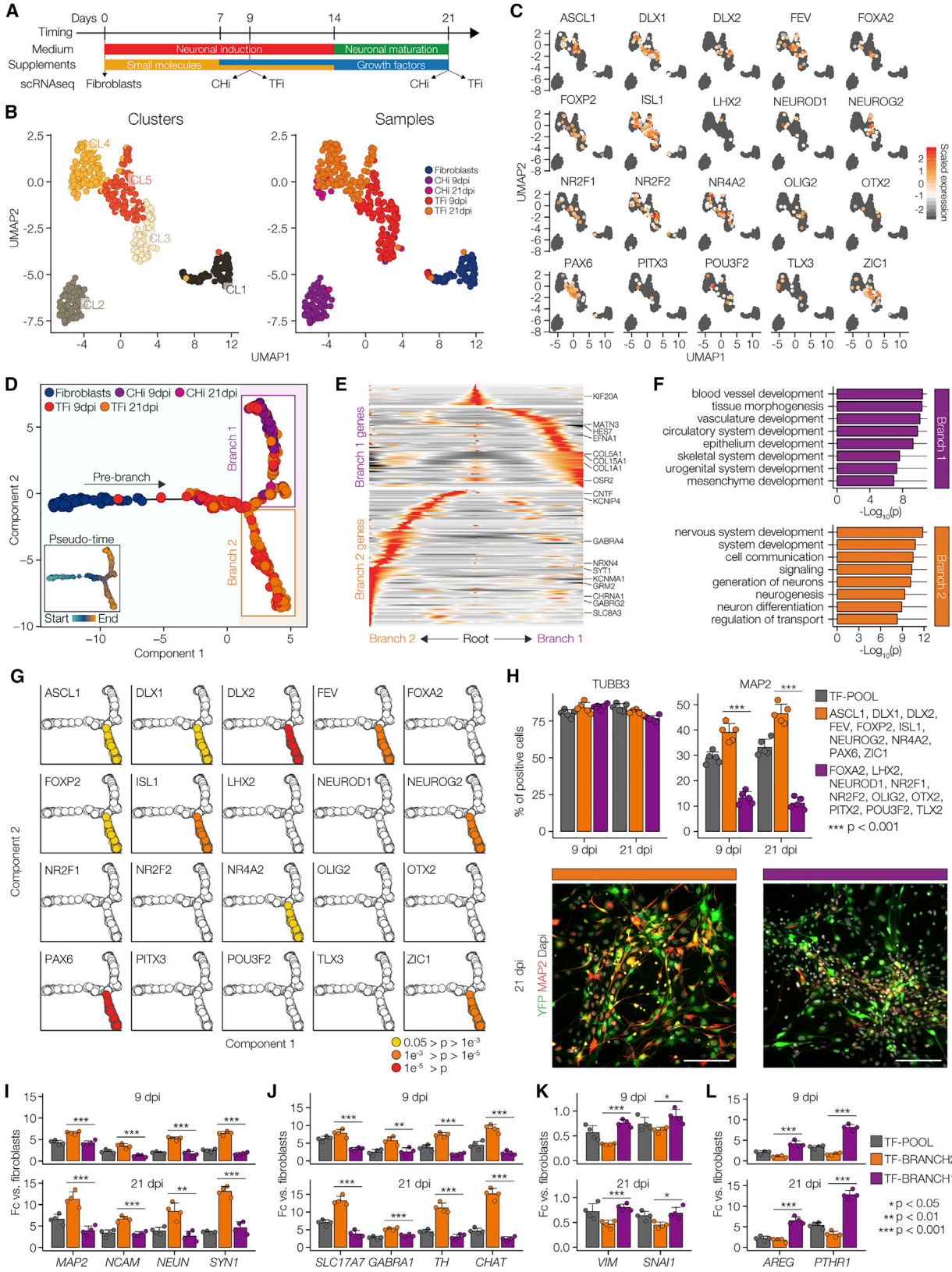
Mapping the 20 exogenous genes to each branch, we found significant enrichment of 10 TFs in branch 2 ($p < 10^{-2}$), while no exogenous TF was significantly enriched in branch 1 (Fisher's exact test; $p > 0.05$) (Figures 4G and S4M).

To confirm whether these 10 TFs derived from branch 2 can induce neuronal phenotype, we independently transduced fibroblasts with the 20 TF-pool, the 10 TFs enriched in branch 2 (neuron branch) and the remaining 10 TFs showing no enrichment in either branch (control). Profiling of neuronal markers revealed that TFs from the neuron branch markedly increased the efficiency of neuronal expression and morphological features as compared with both the complete 20 TF-pool and control TFs (Figures 4H–4K). In fact, infection with control TFs reduced the efficiency of neuronal conversion as compared with the complete TF-pool and induced upregulation of genes that were functionally relevant in breast and kidney, such as *AREG* and *PTHR1*, respectively (Figures 4L).

Identification of Novel TF Combinations Controlling Neuronal Subtype Specification

To further interrogate the 10 TFs enriched in the neuron branch, we leveraged the droplet-based scRNA-seq UMAP to annotate the cells derived from the Smart-seq2 time course experiment (Figure 5A). Consistent with our previous results (Figure 2), TFi cells mapped to mature neuronal

(E) Boxplots showing increased exogenous (red) versus endogenous (blue) expression of all TFs across all individually infected fibroblasts. Golden dots show endogenous expression in samples infected with the corresponding exogenous TFs. Unpaired Student's t test. Error bars represent mean + SD.



(legend on next page)

subtypes, whereas chemically induced cells mapped to clusters containing partially reprogrammed cells and immature neurons (Figures S5A–S5C). The majority of TFi clusters were assigned to either the glutamatergic or cholinergic cluster, suggesting that these two neuronal subtypes prevailed in the Smart-seq2 time course experiment.

To identify the major determinants controlling glutamatergic and cholinergic reprogramming, we performed co-expression module analysis (Fisher's exact tests; $p < 0.05$) and subsequently attributed genes in each subtype with combination scores (CSs) (Figure 5C). The CS represents the significance of enrichment of marker gene expression in cells containing at least three of the predicted exogenous TFs. Based on this approach, we identified *DLX2*, *ZIC1*, *NEUROG2*, and *PAX6* controlling the reprogramming of glutamatergic-like neurons and *DLX1*, *ISL1*, *NEUROG2*, and *PAX6* controlling the reprogramming of cholinergic-like neurons (Figure 5D). The analysis revealed that most gene modules found in the glutamatergic and cholinergic network are indirect targets of exogenous TFs, with few TFs directly linked to marker genes (Figure 5C, Table S5).

Vesicular glutamate transporter 1, VGlut1, mediates glutamate uptake into synaptic vesicles and is a hallmark of glutamatergic neurons (Zhou and Danbolt, 2014). To independently validate whether TFs derived from the gene module analysis can induce the expression of VGlut1, we infected fibroblasts with *DLX2*, *ZIC1*, *NEUROG2*, and *PAX6* and revealed significant induction of VGlut1 when compared with fibroblasts infected with the complete TF-pool or CHi only (Figure 5D). Similarly, we observed a significant induction of choline acetyltransferase (CHAT), a transferase enzyme responsible for the synthesis of the neurotransmitter acetylcholine (Lin et al., 2005), upon infection of *DLX1*, *ISL1*, *NEUROG2*, and *PAX6* in fibroblasts. We further characterized the electrophysiological properties and found that both glutamatergic-like and cholinergic-like neurons generated repetitive but distinct action potentials upon current injection at 21 dpi (Figures 5E, 5F, and S6F) and both action potentials were inhibited by withdrawal of extracellular Na^+ ion in the medium (Figures 5E, 5F, and S6G–S6J).

Finally, *NEUROG2* acts as a pioneering factor during direct fibroblast-to-neuron reprogramming (Smith et al., 2016). To identify whether other pioneering factors present in the TF-pool are involved in direct fibroblast-to-neuron reprogramming, we compared our gene module analyses for each subtype and revealed that, in addition to *NEUROG2*, *PAX6* exhibited hallmarks of a common regulator of direct cell reprogramming toward the neurons based on its gene association analysis. During brain development, *PAX6*-expressing neuroectodermal cells can be readily patterned to region-specific neuro-progenitor cells that give rise to various neuronal subtypes, including cholinergic neurons (Li et al., 2005), dopaminergic neurons (Liu et al., 2012), and GABAergic neurons (Kallur et al., 2008); however, its role in neuronal direct reprogramming has not been described. Thus, we hypothesized that cells expressing exogenous *PAX6* should exhibit increased neuronal molecular phenotype as compared with cells lacking exogenous *PAX6*. To test this, we grouped cells with and without exogenous *PAX6* and assessed expression signatures of neurogenic, neuronal subtype-specific, and fibroblast genes (Figures 5G and 5H). Indeed, we found that cells expressing *PAX6* consistently showed significantly increased expression of neurogenic and neuronal subtype-specific genes and decreased expression of

Figure 4. Association of Developmental Trajectories with Exogenous Expression Profiles

(A) Diagram of the differentiation protocol of TFi and CHi, depicting the samples used for the time course Smart-seq experiment.

(B) UMAP 2D cell maps of the time course data. Left: cells were colored by cluster identity. Right: cells were colored by sample identity. Fibroblasts (n = 78 cells), CHi at 9 dpi (n = 87 cells), TFi at 9 dpi (n = 129), CHi at 21 dpi (n = 15 cells), and TFi at 21 dpi (n = 137 cells).

(C) Visualization of relative expression values of exogenous TFs on UMAP plots.

(D) Pseudo-temporal ordering of the Smart-seq time course based on the expression of 2,925 developmental genes (n = 446 cells). Small inset shows the same plot colored by pseudo-temporal values.

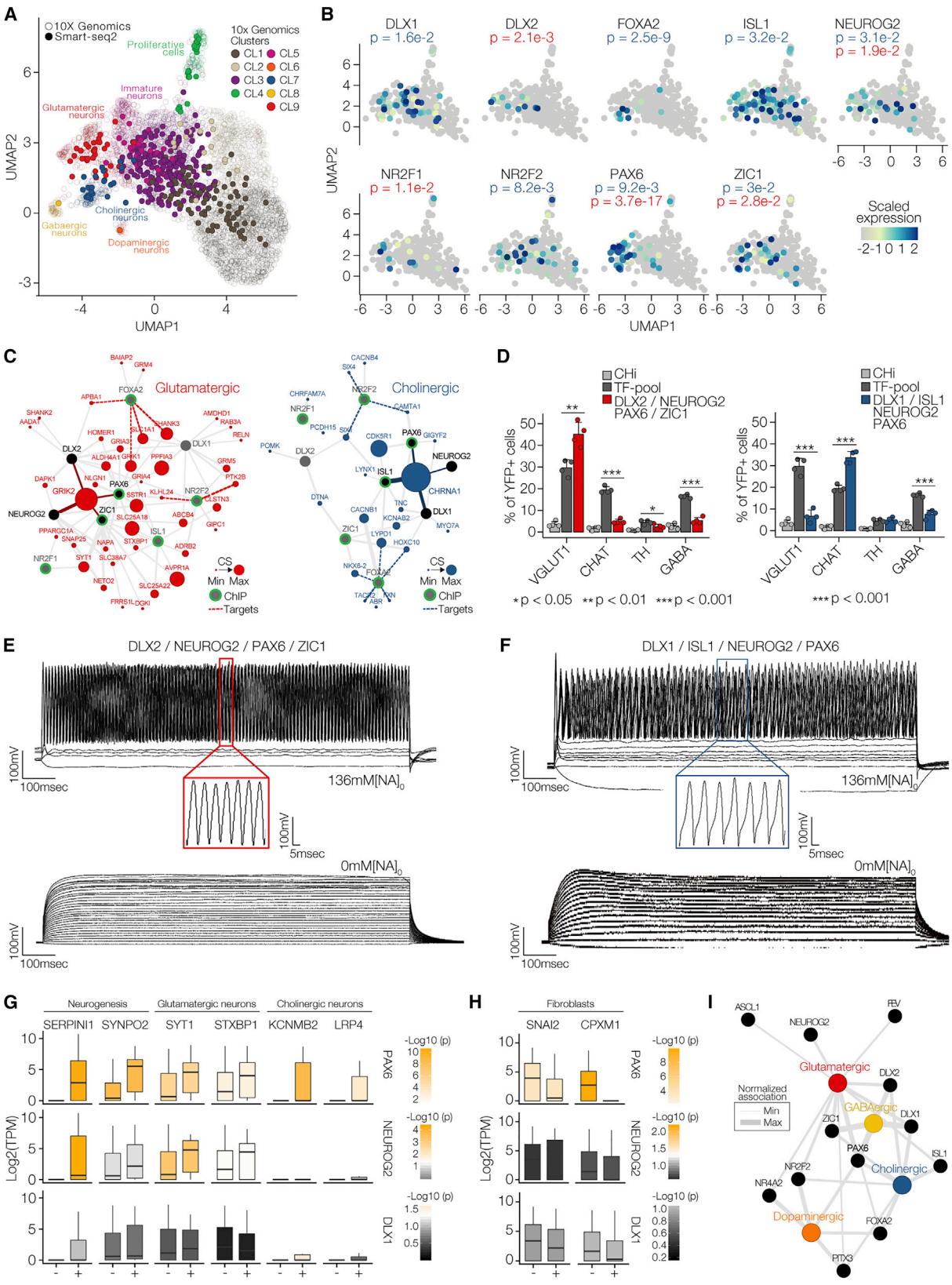
(E) Heatmap showing ~200 genes with branch-specific differential expression as determined by BEAM (R Package "Monocle"). In this heatmap, columns are points in pseudo-time, rows are genes, and the middle (Root) is the beginning of pseudo-time. Branch 1 goes from the middle of the heatmap to the right, while branch 2 goes to the left.

(F) GO analysis of genes differentially expressed in branch 1 (top panel) and branch 2 (bottom panel).

(G) Identification of exogenous TFs with branch-specific enrichment based on Fisher's exact tests.

(H) Top panels: quantification of TUBB3+ and MAP2+ cells in fibroblasts infected with the complete TF-pool (gray), branch 2-enriched TFs (orange), and unenriched TFs (purple) at 9 and 21 dpi. n = 5 independent experiments, unpaired Student's t test. Error bars represent mean + SD. Bottom panels: representative images of immunostainings for MAP2 (red) at 21 dpi; colors as in top panels. YFP (green) marks infected cells and cell nuclei were visualized using DAPI nuclear stain (gray). Scale bars, 100 μm.

(I–L) Neuronal differentiation, loss of fibroblast characteristics, and acquisition of alternative developmental fates as revealed by qPCR for pan-neuronal marker genes (MAP2, NRCAM, NEUN, SYN1) (I), canonical neuronal subtype markers (VGlut1, GABA, TH, CHAT) (J), canonical fibroblast markers (VIM, SNAI2) (K), and branch 1-enriched genes (AREG, PTHR1) (L) at 9 and 21 dpi; colors as in (H). n = 4 independent experiments, unpaired Student's t test. Error bars represent mean +SD.



(legend on next page)

fibroblast genes compared with cells lacking exogenous *PAX6*. Moreover, the same extraction analysis for other exogenous genes, such as *NEUROG2*, *DLX1*, *DLX2*, *ISL1*, and *ZIC1*, suggests that *PAX6* is the most dominant factor in inducing neurons (Figures 5G, 5H, S6K, and S6L). In line with these findings, expanding our analysis to additional neuronal subtypes placed *PAX6* at the core of the association network, possibly as a pioneering factor during direct neuronal reprogramming, while TFs that induce neuronal subtype specification were placed at its periphery for a defined specification (Figure 5I).

DISCUSSION

This study adapts retrospective identification of vector-based gene expression in single cells for the scale of massively parallel scRNA-seq. Our results demonstrate reliable distinction of endogenous and exogenous gene expression by a deep full-length scRNA-seq approach and further identify the most influential pro-neuronal TFs and alternative combinations of TFs from a large pool of candidate genes that govern the reprogramming of multiple and distinct neuronal subtypes.

This approach is set apart from earlier studies testing various possible combinations of TFs in a one-by-one approach (An et al., 2016; Caiazzo et al., 2011; Kim et al., 2011; Liu et al., 2012; Pang et al., 2011; Vierbuchen et al., 2010). Unlike other perturbation single-cell approaches (Adamson et al., 2016; Dixit et al., 2016), the multiplex overexpression method combined with scRNA-seq is compatible with virtually any vector-based expression system without the need for barcoding, avoiding possible barcode swapping and reconstruction of existing vectors (Griffiths et al., 2018); however, profiling of integrated barcodes in a vector-based expression system may allow the inclusion of higher cell numbers using either the 5′ or 3′ droplet-based single-cell RNA-seq.

Based on single-cell expression profiles of perturbed cells, we identified distinct developmental trajectories emerging during neuronal lineage specification. Specifically, we could associate *NEUROG2* with glutamatergic (Winpenny et al., 2011), *DLX1/DLX2* with GABAergic (Pla et al., 2018), *ISL1* with cholinergic (Cho et al., 2014), and *NR4A2* with dopaminergic neuron development (Caiazzo et al., 2011). Collectively this study strongly advocates that: (1) the multiplex scRNA-seq efficiently identifies key determinants of transdifferentiation in a single experiment, (2) genes identified by this approach are not random but specifically enriched in a defined neuronal program and function, (3) neuronal induction medium alone could not induce reprogramming, possibly due to the epigenetic gridlock of the original cell type, (4) multiple factors work in concert to drive cell fate specifications and diversification, and finally (5) our findings support a hierarchical reprogramming model which predicts that replacement of only a few factors can alter the fate of generated cells (Wapinski et al., 2013). However, retrospective identification of exogenous gene expression necessitates continuous exogenous gene expression over the whole time period analyzed. Here, we did not address temporal aspects of exogenous gene expression, which is still largely unexplored. Future studies that aim to explore temporal aspects of exogenous gene expression during cell reprogramming will likely require temporally controlled activation and suppression of gene sets and sequential

Figure 5. *PAX6* Acts as Master Regulator to Control Reprogramming of Glutamatergic and Cholinergic Neurons

(A) Computational mapping of the Smart-seq time course onto the 10X Genomics UMAP. Smart-seq cells are colored based on 10X Genomics cluster membership and positioned based on the five nearest neighbors.

(B) Visualization of scaled expression values of exogenous TFs that showed significant enrichment (Fisher's exact test, $p < 0.05$) in glutamatergic and/or cholinergic clusters on 2D UMAPs.

(C) Neuronal subtype-specific co-expression modules on the basis of significant associations (Fisher's exact test, $p < 0.05$) with exogenous TFs shown in (B). Exogenous TFs associated with genes showing highest CS in each module are shown in black, all other exogenous TFs are shown in gray. Neuronal subtype-specific genes are colored. Direct targets of exogenous TF are indicated with dashed lines based on TF with chromatin immunoprecipitation sequencing (ChIP-seq) evidence, highlighted with green borders.

(D) Validation of novel combinations of exogenous TFs by quantification of immunostainings for VGlut1, CHAT, TH, and GABA of CHi (light gray), fibroblasts infected with the complete TF-pool (dark gray) and fibroblasts infected with novel combinations (color) at 21 dpi. $n = 4$ independent experiments, unpaired Student's t test. Error bars represent mean + SD.

(E and F) The generation of repetitive action potentials in induced neurons infected with DLX2, NEUROG2, PAX6, ZIC1 (E) or DLX1, ISL1, NEUROG2, or PAX6 (F). Representative traces in the presence (upper panel) or absence (lower panel) of extracellular Na^+ were recorded using the current-clamp protocol.

(G and H) Boxplots showing the log2-transformed TPM values of neurogenic and neuronal subtype-specific genes (G) and fibroblast-specific genes (H) in cells with (+) or without (−) exogenous *PAX6* (top), *NEUROG2* (middle), and *DLX1* (bottom). Boxplots are colored based on −log10-transformed p values.

(I) Edge-normalized network summarizing the associations of exogenous TFs with glutamatergic, cholinergic, GABAergic, and dopaminergic modules.

gene delivery methods. In addition, due to the massively increased complexity, a much larger number of single cells will need to be analyzed.

Our results also suggest that *PAX6* acts as a key driver of direct neuronal reprogramming. *PAX6* acts as a major regulator of mammalian nervous system development and is expressed in a region-specific manner in NPCs and uniformly in neuroectodermal cells differentiated from embryonic stem cells and iPSCs (Chapouton et al., 1999; Stoykova et al., 2000; Yun et al., 2001; Zhang et al., 2010). In line with this, combination of *PAX6* with *NEUROG2*, *DLX2*, and *ZIC1* generates mainly glutamatergic neurons, whereas *DLX1* and *ISL1* generate mainly cholinergic neurons, suggesting that combinations of TFs are involved in both activation and repression toward cell specification. Previous reports implicated that *DLX1* and *DLX2* act as repressors to glia lineage tilting the cell fate toward the neuronal lineage (Petryniak et al., 2007). In another study, *ZIC1* represses dopaminergic specification (Tiveron et al., 2017) and *PAX6*/*NEUROG2* direct the cells toward glutamatergic fate (Winpenny et al., 2011), while Isl1 further pushes the cells toward cholinergic development (Cho et al., 2014). While precise dynamics will require systematic perturbation analysis, our approach demonstrates that a defined set of TFs important for cell-type switching and lineage specification can be precisely extrapolated.

In conclusion, our data corroborate previous studies on neuronal reprogramming, predict novel combinations of pro-neuronal TFs that allow the generation of neuronal subtypes, and yield the regulatory logic of neuronal reprogramming. We envision that this approach is an effective strategy to identify transcriptional codes controlling cell fate conversions, speaking to its potential to become a standard strategy to unravel molecular mechanisms governing other cell reprogramming pathways.

EXPERIMENTAL PROCEDURES

Cell Culture and Generation of Induced Neurons

Human neonatal dermal fibroblasts were purchased from Lonza (C-2509; passages 4–8) and cultured in DMEM containing 10% FBS/L-glutamine/10% penicillin and streptomycin at 37°C 5% CO_2 until virus infection. For TFi generation, a total of six MOI (~0.3 MOI per virus) were pooled into culture medium containing 8 μg mL^{-1} polybrene (Sigma) to increase infection efficiency. For CHi cell generation, dermal fibroblasts were infected with a virus containing multiple cloning sites (e.g., no TF) at an MOI of 6 expressing YFP. The virus pool was infected and incubated overnight at 37°C 5% CO_2. On the following day, the virus-containing medium was removed and cells were incubated for an additional day in the fresh culture medium at 37°C 5% CO_2. Henceforth, CHi and TFi were cultured under identical culture conditions: cells were split onto poly-D-lysine (100 μg mL^{-1}; Sigma)/laminin (50 μg mL^{-1}; Sigma)-coated culture plates and incubated overnight in culture medium at 37°C 5% CO_2. On the following day, the medium was changed to neuronal induction medium containing DMEM/F12 and Neurobasal-A (Thermo Fisher Scientific) mixed at a 1:1 ratio, 2% (vol/vol) B27 supplement, and 0.5% N2 supplement (Gibco), 1× nonessential amino acids (Thermo Fisher Scientific), 1% GlutaMAX supplement, VPA (0.5 mM, Wako), L-ascorbic acid (200 nM, Sigma), Y-27632 (10 μM, Wako), dcAMP (0.5 mM, Sigma), 10% FBS, and 10% penicillin and streptomycin (Wako). The concentrations of small molecules used were as follows: CHIR99021 (2 μM, Abcam), SB-431542 (10 μM, Sigma), LDN-193189 (0.5 μM, Stemgent), and Noggin (100 ng mL^{-1}, Sigma). Neuronal induction medium was changed every third day, during which the concentration of FBS was gradually reduced from 10% to 0%. After 2 weeks, neuronal induction medium was replaced with neuronal maturation medium without small molecules, but containing 10 ng mL^{-1} BDNF (Gibco), 10 ng mL^{-1} NT3 (R&D Systems), and 10 ng mL^{-1} GDNF (Thermo Fisher Scientific) and the medium was changed every third day until further analysis.

Electrophysiology

Whole-cell current-clamp recordings were performed as described (Ichikawa et al., 2012). All experiments were conducted at 25°C. Patch pipettes with resistances ranging from 3 to 7 MΩ were pulled from capillary tubes using a DMZ-Universal Puller (Zeitz Instruments, Martinsried, Germany) and then backfilled with intracellular solution. Action potentials were recorded using a patch-clamp amplifier (Axopatch 200B; Axon Instruments, Foster City, CA) with a series of current steps from 0 to 200 pA with a 2,000-ms duration. The action potentials were monitored and stored using pCLAMP software (Molecular Devices, CA) after digitizing the analog signals at 5 kHz (DigiData 1322A; Axon Instruments). For patch-clamp recordings, the extracellular solution (ECS) consisted of the following: 137 mM NaCl, 5 mM KCl, 0.44 mM KH_2PO_4, 0.33 mM Na_2HPO_4, 10 mM glucose, 12 mM $NaHCO_3$, 0.5 mM $MgCl_2$, and 10 mM HEPES, adjusted to pH 7.4 with tris(hydroxymethyl)aminomethane. To examine the Na^+ selectivity, extracellular 136 mM NaCl was substituted with equimolar extracellular LiCl (Na^+-free ECS). To record ionic currents under physiological conditions, intracellular solution containing 150 mM KCl, 10 mM HEPES, and 2 mM magnesium adenosine triphosphate (pH 7.2 by tris(hydroxymethyl)aminomethane) was used.

Computational Methods for scRNA-Seq Data

Artificial Transcript Model

To determine the exact nucleotide sequences flanking the ORF of each exogenous transcript, we sequenced recombined plasmids using a 3730/3730xl DNA Analyzer (Applied Biosystems) following the manufacturer's protocol. In brief, we first amplified templates by PCR using primers annealing to the EF1A promoter sequence near the 5′ end of each ORF to amplify the 5′ junction sequences, and primers annealing to the IRES2 sequence near the 3′ end of each ORF to amplify the 3′ junction sequences. After gel purification, we sequenced templates using BigDye Terminator v.3.1 Cycle Sequencing Kits (Applied Biosystems). We integrated results derived from three primers (three replicates each) at both the 5′ and 3′ junction of each ORF and combined the resulting 5′ and

3′ junction sequences with known sequences of the ORFs of each TF and the CSII-EF-RfA-IRES2-VENUS pENTR lentivirus vector to compile the artificial exogenous transcript model. Finally, we combined our artificial exogenous transcript model with the human transcriptome (version GRCh38.p5) to obtain the final artificial transcript model.

Mapping of the Smart-Seq Time Course Data onto the 10X Genomics UMAP

To integrate the Smart-seq time course data with the 10X Genomics data, we computed a pairwise correlation matrix (Spearman correlation) of all cells based on the expression of ~1,300 highly variable genes in cells of clusters CL4–CL9. Smart-seq cells were mapped onto the 10X Genomics UMAP by averaging x and y coordinates of the five 10X Genomics cells showing the highest correlation values and assigned to their most likely 10X Genomics cluster by determining to which cluster the majority of these five 10X Genomics cells belonged to. If two or more clusters tied for majority, clusters were assigned at random. To identify exogenous TFs with enrichment in any cluster, we performed Fisher's exact tests to calculate the significance of association of a given exogenous TF with each cluster. Exogenous TFs with $p < 0.05$ were considered significantly enriched.

Gene Co-expression Module Assembly

To construct neuronal subtype-specific gene co-expression modules, we calculated the significance of association of each exogenous TF with all other genes using Fisher's exact tests. Exogenous TFs were attributed to neuronal subtype-specific module based on the following criteria: (1) $p < 0.05$ in Fisher's exact test and (2) at least three edges to three neuronal subtype-specific genes. Gene co-expression modules were visualized with Cytoscape using the organic layout.

Combination Score

The CS represents the −log10-transformed p value of the significance (Mann-Whitney U test) of increased gene expression in cells containing at least two of the predicted exogenous TFs in a network compared with all other cells.

Data and Code Availability

All analysis code used in this study is available upon request. Custom code for the main analytical steps can be found in: https://github.com/JoachimLu/Decoding-neuronal-diversification-by-multiplexed-single-cell-RNA-seq. All sequence data are accessible with accession number GEO: GSE117075.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.stemcr.2021.02.006.

AUTHOR CONTRIBUTIONS

Formulation of research goals and aims: J.L. and J.W.S.; data curation: J.L., T.K., and C.P.; formal analysis: J.L., T.K., R.N., and J.W.S.; funding acquisition: J.L., P.C., and J.W.S.; performing experiments: J.L., T.K., R.N., T.E.C., D.M.S., M.K., and F.R.; project supervision: P.C. and J.W.S.; writing – original draft: J.L. and J.W.S.; writing – review & editing: J.L. and J.W.S.

ACKNOWLEDGMENTS

This work was supported by a Research Grant for RIKEN Center for Life Science Technology, Division of Genomic Technologies (CLST DGT) and RIKEN Center for Integrative Medical Sciences (IMS) from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) and a Postdoctoral Fellowship for Research in Japan by the Japan Society for the Promotion of Science (JSPS). The authors wish to acknowledge RIKEN GeNAS for the sequencing of the libraries.

Received: August 7, 2020
Revised: February 9, 2021
Accepted: February 9, 2021
Published: March 11, 2021

REFERENCES

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

Alda-Catalinas, C., Bredikhin, D., Hernando-Herraez, I., Eckersley-Maslin, A., Stegle, O., and Reik, W. (2020). A single-cell transcriptomics CRISPR-activation screen identifies epigenetic regulators of the zygotic genome activation program. Cell. Syst. *11*, 25–41.

An, N., Xu, H., Gao, W.Q., and Yang, H. (2016). Direct conversion of somatic cells into induced neurons. Mol. Neurobiol. *55*, 642–651.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Caiazzo, M., Dell'Anno, M.T., Dvoretskova, E., Lazarevic, D., Taverna, S., Leo, D., Sotnikova, T.D., Menegon, A., Roncaglia, P., Colciago, G., et al. (2011). Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. Nature *476*, 224–227.

Chanda, S., Ang, C.E., Davila, J., Pak, C., Mall, M., Lee, Q.Y., Ahlenius, H., Jung, S.W., Sudhof, T.C., and Wernig, M. (2014). Generation of induced neuronal cells by the single reprogramming factor ASCL1. Stem Cell Reports *3*, 282–296.

Chapouton, P., Gartner, A., and Gotz, M. (1999). The role of Pax6 in restricting cell migration between developing cortex and basal ganglia. Development *126*, 5569–5579.

Chen, S., Sanjana, N.E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D.A., Song, J., Pan, J.Q., Weissleder, R., et al. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell *160*, 1246–1260.

Cho, H., Cargnin, F., Kim, Y., Lee, B., Kwon, R., Nam, H., Shen, R., Barnes, A.P., Lee, J.W., Lee, S., et al. (2014). Isl1 directly controls a cholinergic neuronal identity in the developing forebrain and spinal cord by forming cell type-specific complexes. PLoS Genet. *10*, e1004280.

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.

Gelles, J., Schnapp, B.J., and Sheetz, M.P. (1988). Tracking kinesin-driven movements with nanometer-scale precision. Nature *331*, 450–453.

Griffiths, J.A., Richard, A.C., Bach, K., Lun, A.T.L., and Marioni, J.C. (2018). Detection and removal of barcode swapping in single-cell RNA-seq data. Nat. Commun. *9*, 1–6.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.

Hu, W., Qiu, B., Guan, W., Wang, Q., Wang, M., Li, W., Gao, L., Shen, L., Huang, Y., Xie, G., et al. (2015). Direct conversion of normal and Alzheimer's disease human fibroblasts into neuronal cells by small molecules. Cell Stem Cell *17*, 204–212.

Ichikawa, H., Kim, H., Shuprisha, A., Shikano, T., Tsumura, M., Shibukawa, Y., and Tazaki, M. (2012). Voltage-dependent sodium channels and calcium-activated potassium channels in human odontoblasts in vitro. J. Endod. *38*, 1355–1362.

Kallur, T., Gisler, R., Lindvall, O., and Kokaia, Z. (2008). Pax6 promotes neurogenesis in human neural stem cells. Mol. Cell Neurosci. *38*, 616–628.

Kim, J., Su, S.C., Wang, H., Cheng, A.W., Cassady, J.P., Lodato, M.A., Lengner, C.J., Chung, C.Y., Dawlaty, M.M., Tsai, L.H., et al. (2011). Functional integration of dopaminergic neurons directly converted from mouse fibroblasts. Cell Stem Cell *9*, 413–419.

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.

Li, X.J., Du, Z.W., Zarnowska, E.D., Pankratz, M., Hansen, L.O., Pearce, R.A., and Zhang, S.C. (2005). Specification of motoneurons from human embryonic stem cells. Nat. Biotechnol. *23*, 215–221.

Li, X., Zuo, X., Jing, J., Ma, Y., Wang, J., Liu, D., Zhu, J., Du, X., Xiong, L., Du, Y., et al. (2015). Small-molecule-driven direct reprogramming of mouse fibroblasts into functional neurons. Cell Stem Cell *17*, 195–203.

Lin, W., Dominguez, B., Yang, J., Aryal, P., Brandon, E.P., Gage, F.H., Lee, K., and Jolla, L. (2005). Neurotransmitter acetylcholine negatively regulates neuromuscular synapse formation by a Cdk5-dependent mechanism*46*, 569–579.

Liu, X., Li, F., Stubblefield, E.A., Blanchard, B., Richards, T.L., Larson, G.A., He, Y., Huang, Q., Tan, A.C., Zhang, D., et al. (2012). Direct reprogramming of human fibroblasts into dopaminergic neuron-like cells. Cell Res. *22*, 321–332.

Liu, M.L., Zang, T., Zou, Y., Chang, J.C., Gibson, J.R., Huber, K.M., and Zhang, C.L. (2013). Small molecules enable neurogenin 2 to efficiently convert human fibroblasts into cholinergic neurons. Nat. Commun. *4*, 2183.

Liu, Y., Yu, C., Daley, T.P., Wong, W.H., Wernig, M., and Qi, L.S. (2018). CRISPR activation screens systematically identify factors that drive neuronal fate and reprogramming. Stem Cell *23*, 758–771.e8.

Masserdotti, G., Gascon, S., and Gotz, M. (2016). Direct neuronal reprogramming: learning from and for development. Development *143*, 2494–2510.

Mazzoni, E.O., Mahony, S., Closser, M., Morrison, C.A., Nedelec, S., Williams, D.J., An, D., Gifford, D.K., and Wichterle, H. (2013). Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. Nat. Neurosci. *16*, 1219–1227.

Mcinnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, 1802.03426v2.

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.

Pang, Z.P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D.R., Yang, T.Q., Citri, A., Sebastiano, V., Marro, S., Sudhof, T.C., et al. (2011). Induction of human neuronal cells by defined transcription factors. Nature *476*, 220–223.

Petryniak, M.A., Potter, G.B., Rowitch, D.H., and Rubenstein, J.L.R. (2007). Dlx1 and Dlx2 control neuronal versus oligodendroglial cell fate acquisition in the developing forebrain. Neuron *55*, 417–433.

Pfisterer, U., Kirkeby, A., Torper, O., Wood, J., Nelander, J., Dufour, A., Bjorklund, A., Lindvall, O., Jakobsson, J., and Parmar, M. (2011). Direct conversion of human fibroblasts to dopaminergic neurons. Proc. Natl. Acad. Sci. U S A *108*, 10343–10348.

Picelli, S., Faridani, O.R., Bjorklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pla, R., Stanco, A., Howard, M.A., Rubin, A.N., Vogt, D., Mortimer, N., Cobos, I., Potter, G.B., Lindtner, S., Price, J.D., et al. (2018). Dlx1 and Dlx2 promote interneuron GABA synthesis, synaptogenesis, and dendritogenesis. Cereb. Cortex *28*, 3797–3815.

Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Consortium, F., Suzuki, H., Nefzger, C.M., Daub, C.O., et al. (2016). A predictive computational framework for direct reprogramming between human cell types. Nat. Genet. *48*, 331–335.

Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol. *38*, 954–961.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. Nat. Rev. Genet. *16*, 299–311.

Shifrut, E., Carnevale, J., Tobin, V., Diolaiti, M.E., Ashworth, A., Marson, A., Shifrut, E., Carnevale, J., Tobin, V., Roth, T.L., et al. (2018). Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function resource. Cell *175*, 1958–1971.

Shin, J.W., Suzuki, T., Ninomiya, N., Kishima, M., Hasegawa, Y., Kubosaki, A., Yabukami, H., Hayashizaki, Y., and Suzuki, H. (2012). Establishment of single-cell screening system for the rapid identification of transcriptional modulators involved in direct cell reprogramming. Nucleic Acids Res. *40*, e165.

Smith, D.K., Yang, J., Liu, M.L., and Zhang, C.L. (2016). Small molecules modulate chromatin accessibility to promote NEUROG2-mediated fibroblast-to-neuron reprogramming. Stem Cell Reports *7*, 955–969.

Son, E.Y., Ichida, J.K., Wainger, B.J., Toma, J.S., Rafuse, V.F., Woolf, C.J., and Eggan, K. (2011). Conversion of mouse and human fibroblasts into functional spinal motor neurons. Cell Stem Cell *9*, 205–218.

Stoykova, A., Treichel, D., Hallonet, M., and Gruss, P. (2000). Pax6 modulates the dorsoventral patterning of the mammalian telencephalon. J. Neurosci. *20*, 8042–8050.

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.

Tiveron, M., Beclin, C., Murgan, X.S., Wild, S., Angelova, A., Marc, J., Core, N., De Chevigny, X.A., Herrera, X.E., Bosio, A., et al. (2017). Zic-proteins are repressors of dopaminergic forebrain fate in mice and *C. elegans*. J. Neurosci. *37*, 10611–10623.

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

Victor, M.B., Richner, M., Hermanstyne, T.O., Ransdell, J.L., Sobieski, C., Deng, P.Y., Klyachko, V.A., Nerbonne, J.M., and Yoo, A.S. (2014). Generation of human striatal neurons by microRNA-dependent direct conversion of fibroblasts. Neuron *84*, 311–323.

Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035–1041.

Wapinski, O.L., Vierbuchen, T., Qu, K., Lee, Q.Y., Chanda, S., Fuentes, D.R., Giresi, P.G., Ng, Y.H., Marro, S., Neff, N.F., et al. (2013). Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell *155*, 621–635.

Weinreb, C., Wolock, S., and Klein, A.M. (2018). SPRING: a kinetic interface for visualizing high dimensional single-cell expression data*34*, 1246–1248.

Winpenny, E., Lebel-Potter, M., Fernandez, M.E., Brill, M.S., Götz, M., Guillemot, F., and Raineteau, O. (2011). Sequential generation of olfactory bulb glutamatergic neurons by Neurog2-expressing precursor cells. Neural Dev. *6*, 1–18.

Xu, Z., Jiang, H., Zhong, P., Yan, Z., Chen, S., and Feng, J. (2016). Direct conversion of human fibroblasts to induced serotonergic neurons. Mol. Psychiatry *21*, 62–70.

Yun, K., Potter, S., and Rubenstein, J.L. (2001). Gsh2 and Pax6 play complementary roles in dorsoventral patterning of the mammalian telencephalon. Development *128*, 193–205.

Zhang, X., Huang, C.T., Chen, J., Pankratz, M.T., Xi, J., Li, J., Yang, Y., Lavaute, T.M., Li, X.J., Ayala, M., et al. (2010). Pax6 is a human neuroectoderm cell fate determinant. Cell Stem Cell *7*, 90–100.

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.

Zhou, Y., and Danbolt, N.C. (2014). Glutamate as a neurotransmitter in the healthy brain. J. Neural Transm. (Vienna) *121*, 799–817.

Stem Cell Reports, Volume *16*

Supplemental Information

Decoding Neuronal Diversification by Multiplexed Single-cell RNA-Seq

Joachim Luginbühl, Tsukasa Kouno, Rei Nakano, Thomas E. Chater, Divya M. Sivaraman, Mami Kishima, Filip Roudnicky, Piero Carninci, Charles Plessy, and Jay W. Shin

Decoding neuronal diversification by multiplexed single-cell RNA-seq

Joachim Luginbühl, Tsukasa Kouno, Rei Nakano, Thomas E Chater, Divya M Sivaraman, Mami Kishima, Filip Roudnicky, Piero Carninci, Charles Plessy and Jay W Shin

Supplementary Experimental Procedures

Complementary DNA and virus generation

Complementary DNA (cDNA) and viruses were generated as described previously (Shin et al., 2012). Briefly, we recombined Gateway-compatible human full-length cDNA entry clones derived from RIKEN BRC clone bank (**http://www.brc.riken.jp/**) into the pENTR lentivirus vector CSII-EF-RfA-IRES2-VENUS using Gateway LR clonase II enzyme mix (Invitrogen). After Proteinase K treatment, recombinant plasmids were transformed into competent *Escherichia coli* and plasmids derived from single colonies were expanded and purified using PureYield Plasmid Midiprep System (Promega). Plasmids, HIV-gp and VSV envelope genes were co-transfected into 293T cells using FuGeneHD (Roche). Supernatant-containing viruses were collected, centrifuged by ultracentrifugation and dissolved in 100μl HBSS buffer (WAKO) and stored at −80°C for later use.

Immunocytochemistry and quantitative RT-PCR

For immunocytochemistry, cells were fixed in 4% paraformaldehyde for 20 min at room temperature and permeabilized using 0.2% Triton X-100 (SIGMA) for 10 min at room temperature. Following permeabilization, cells were pre-incubated with blocking solution (2% BSA, 0.2% Triton X-100) to block non-specific sites for 1 h. Primary antibodies were diluted in

blocking solution and applied to cells overnight at 4°C. Secondary antibodies were diluted in blocking solution and applied to cells at room temperature for 1 h. Imaging and quantification was performed using the INCell Analyzer 6000 (GE Healthcare). For each condition, 40 fields containing 100-500 cells/field were measured. The following primary antibodies and dilutions were used: mouse anti-TUBB3 (Covance, MMS-435P, 1:1000), mouse anti-MAP2 (Abcam, ab11267, 1:500), rabbit anti-SYNAPSIN 1 (Abcam, ab64581, 1:200), rabbit anti-VGLUT1 (Synaptic Systems, 135303, 1:100), mouse anti-GABA (Abcam, ab86186, 1:200), sheep anti-CHAT (Abcam, ab18736, 1:100), rabbit anti-TH (Abcam, ab112, 1:500). The following secondary antibodies and dilutions were used: goat anti-mouse IgG1 (GIBCO, A-21121, 1:200), goat anti-mouse IgG2a (Thermo Fisher Scientific, A-21131, 1:200), goat anti-rabbit IgG (Thermo Fisher Scientific, A-11008, 1:200), donkey anti-sheep IgG (Thermo Fisher Scientific, A-11015, 1:200). Human neonatal dermal fibroblasts were used as negative controls. Quantification of immunostainings was performed using the INCell Investigator Developer Toolbox. For quantitative RT-PCR (qRT-PCR), total RNA was purified using the RNeasy Mini Kit (QIAGEN) according to the manufacturer's specification. Quality and quantity of RNA was determined using a DropSense96 (Trinean). Equal amounts of RNA were reverse-transcribed using the One-Step SYBR PrimeScript RT PCR Kit II, and cDNAs were normalized to equal amounts using primers against *GAPDH*. qRT-PCR was performed on a 7900HT Fast Real-Time PCR system (Applied Biosystems).

Fluidigm C1 reversed loading protocol (backloading) for bulk RNA-seq

To perform bulk RNAseq of a total of 96 samples, we used the Fluidigm Script Builder™ to design a reversed protocol that allows to load each sample into a separate chamber, where RT and cDNA amplification is performed. After priming the chips, 25 ng of RNA of each sample was loaded into

the output wells on a medium size C1 Single-cell Open App IFC and the IFC was sealed using a C1 Porous Barrier Tape kit (Fluidigm). RT and cDNA amplification was performed following the manufacturer's protocol (P100-7168L1). We ran the backloading script for 15 min at 4°C and switched to the mRNA seq RT and Amp script (1772x), which harvested cDNA back into the output wells. To remove remaining RNA, we added Rnase One Ribonuclease (Promega) at room temperature. To quantify the cDNA, we used the Quant-iT PicoGreen dsDNA Assay kit. Library preparation was performed using the Nextera XT DNA Library Preparation kit (Illumina), the Nextera XT Index Kit v2 (Illumina) and Ampure XP beads (Beckman Coulter). Libraries were quantified using the High Sensitivity DNA Reagents (Agilent Technologies) and the KAPA Library Quantification kit (KAPA BIOSYSTEMS). Libraries were sequenced on the Illumina Hiseq 2500 platform in rapid mode (100bp paired end).

Droplet-based scRNA-seq

Library preparation and sequencing: Droplet-based scRNA-seq libraries were generated using the Chromium™ Single Cell 3' Reagent kits V1 (CG00026, 10x Genomics). Briefly, cell number and cell viability were assessed using the Countess II Automated Cell Counter (ThermoFisher). Thereafter, cells were mixed with the Single Cell Master Mix and loaded together with Single Cell 3' Gel beads and Partitioning Oil into a Single Cell 3' Chip. RNA transcripts were uniquely barcoded and reverse-transcribed in droplets. cDNAs were pooled and amplified according to the manufacturer's protocol. Libraries were quantified by High Sensitivity DNA Reagents (Agilent Technologies) and the KAPA Library Quantification kit (KAPA BIOSYSTEMS). Libraries then were sequenced by Illumina Hiseq 2500 in rapid mode.

Read alignment and gene quantification: Initial read alignment to hg19 human reference genome, filtering and UMI counting was performed by the CellRanger Software ver 1.1.0 using default parameters. This software implements STAR as an alignment tool. Data from TFi and CHi were normalized to the same sequencing depth and aggregated into a single gene-barcode matrix. The expression values were quantified as count per million (CPM) and transformed to $\log_2$ (CPM+1).

scRNAseq using the Fluidigm C1 platform

Single cell RNA-seq analysis was performed following the manufacturer's protocol (P100-7168L1, Fluidigm). Briefly, cell number and cell viability were assessed using the Countess II Automated Cell Counter (ThermoFisher). After priming medium size C1 Single-cell Open App IFCs, 250 cells/μL were loaded and capture efficiency and cell morphology was assessed using the IN Cell Analyzer 6000 (GE Healthcare). To exclude chambers loaded with no cells, more than one cell (cell doublets) or dead cells for downstream analysis, we took 11 z-stacking images per chamber. Next, the cells were lysed with 20,000-fold diluted ERCC RNA Spike-In Mix1 (Thermo Fisher Scientific) and reverse transcription (RT) and cDNA amplification were performed using the SMARTer Ultra Low RNA Kit for the Fluidigm C1™ System (Clontech). The amplified cDNAs were harvested into 96 well plates and quantified with Quant-iT™ PicoGreen dsDNA Assay kit. Library preparation was performed with the Nextera XT DNA Library Preparation kit (Illumina), Nextera XT Index Kit v2 (Illumina) and AMpure XP beads (Beckman Coulter). Libraries were quantified by High Sensitivity DNA Reagents (Agilent Technologies) and KAPA Library Quantification kit (KAPA BIOSYSTEMS). Each of the libraries were sequenced by Illumina Hiseq 2500 in high output mode (100bp paired end). Reads were aligned to the trimmed artificial transcript model using Kallisto with the default parameter settings for paired-end reads.

The expression values were quantified as transcripts per million (TPM) and transformed to $\log_2$ (TPM+1).

Computational methods for scRNA-seq data

Quality control, cell clustering and UMAP visualization: All analyzes and visualization of data were conducted in the R environment. For droplet-based 10X Genomics scRNA-seq data, clustering and UMAP visualization was performed using the R package 'Seurat' (Satija et al., 2015) (v2.3.4). Genes expressed in less than 3 cells and cells expressing less than 1000 genes or more than 4500 genes were removed. In addition, we removed cells expressing more than 2% mitochondrial genes, indicative of dead cells. PCA was performed on the z-transformed expression levels of the identified ~1000 highly variable genes after regressing out the number of UMI and the percentage of mitochondrial genes. Using the 20 most significant principal components (PCs), we projected individual cells based on their PC scores onto a single two-dimensional map using UMAP. Gene expression heat map along UMAP1 was obtained by dividing cells into 40 groups based on their UMAP1 scores, averaging gene expression within each group and scaling expression values by column. For the Smart-seq time-course data, we excluded chambers containing no cells, multiple cells or cells exhibiting morphological features of cell death based on visual inspection using the IN Cell Analyzer 6000 (GE Healthcare). Additionally, cells not expressing either of the two housekeeping genes *ACTB* and *GAPDH* (encoding β-actin and glyceraldehyde-3-phosphate dehydrogenase, respectively), or expressing them at less than three standard deviations below the mean, were scored as unhealthy and removed. After applying these filters, 78 fibroblasts, 216 cells for the time-point 9 dpi (87 CHi and 129 TFi) and 152 cells for the time-point 21 dpi (15 CHi and 137 TFi) remained, yielding 446 cells in total. Genes expressed in less than 3 cells were removed. PCA was performed on the ~5000 most variable genes. Using the

9 most significant principal components (PCs), we projected individual cells based on their PC scores onto a single two-dimensional map using UMAP. Hierarchical clustering was performed on cells and on PCA scores using Euclidean distance metric.

Read alignment with Bowtie: Reads were aligned to the artificial transcript model using Bowtie v1.2.2 with the default parameter settings for paired-end reads. After retrieving BED12 files using samtools and bedtools, we intersected all reads using a custom GFF file in which 5' and 3' junctions of all exogenous sequences were defined. Only reads overlapping the junction sequences by at least 5 bp were counted as specific reads. The expression values of all exogenous TFs were quantified as count per million (CPM) and transformed to $\log_2$ (CPM+1).

Read alignment with Kallisto: For alignment using Kallisto (v0.42.4), alignment to the full artificial transcript model yielded many false-positive hits (Supplementary Fig. 3c). Therefore, we trimmed the 5' and 3' junction sequences to ~100 bp on either side, which markedly reduced the number of false positive hits (Fig. 3d). Reads were aligned with the default parameter settings for paired-end reads. Custom R scripts were used to merge transcript isoforms and compile a single expression matrix.

Construction of the force-directed k-nearest neighbors graph: The force-directed k-nearest neighbors graph was constructed based on the expression of ~ 1300 highly variable genes using the online tool SPRING (Weinreb et al., 2018) with the following parameters: Gene variability percentile: 90.0, Number of PCs: 20, Number of nearest neighbors: 20, Number of force layout iterations: 500.

Differential expression test and GO analysis: Marker genes of each cluster were determined using a likelihood ratio test based on zero-inflated data ($p < 1e\text{-}4$) considering only genes that show a

minimum log fold expression change of 0.25 in at least a fraction of 0.25 of cells in the clusters using the non-integrated expression values. For GO analysis, we used marker genes which showed, on average, at least 3-fold enrichment in a cluster compared to all other clusters. GO analysis was performed using the PANTHER database (**http://www.pantherdb.org/**) which uses Fisher's Exact tests with FDR multiple test correction.

Pseudotemporal ordering: Pseudotemporal ordering of cells was performed using the R package 'Monocle' (Trapnell et al., 2014) (v2.2.0). For unsupervised ordering, we used genes differentially expressed between cells at day 0 (fibroblasts) and CHi and TFi at day 9 and day 21 (qval < 0.1; ~10'000 genes). To determine genes that are significantly branch-dependent ($p < 10^{-4}$), we applied the BEAM algorithm. GO analysis for branch-dependent genes was performed using genes that met the following criteria: 1) $p < 0.01$ in a likelihood ratio test based on zero-inflated data; 2) absolute $\log_2$ fold changes between the branch under consideration and others were larger than 2. GO analysis for genes that changed significantly as a function of pseudotime was performed using genes that met the following criteria: 1) $p < 10^{-4}$ of differentialGeneTest; 2) among the top 1000 genes showing positive or negative correlation with pseudotime values. For semi-supervised ordering, we used ~3000 unique genes previously implicated in nervous system development (GO:0007399), circulatory system development (GO:0072359), urogenital system development (GO:0001655), heart development (GO:0007507), mesenchyme development (GO:0060485), ear development (GO:0043583), muscle structure development (GO:0061061), stem cell development (GO:0048864), pancreas development (GO:0031016) and skeletal system development (GO:0001501) (Supplementary Table 3). GO analysis was performed using genes that met the following criteria: 1) $p < 0.01$ in a likelihood ratio test based on zero-inflated data; 2) absolute $\log_2$ fold changes between the branch under consideration and others were larger than 2. To determine

exogenous TFs that are significantly branch dependent, expression values were binarized (0 = not expressed, 1 = expressed). Then we performed Fisher's exact tests to calculate the significance of association of a given exogenous TF with each branch. Exogenous TFs with $p < 0.05$ were considered significantly enriched.

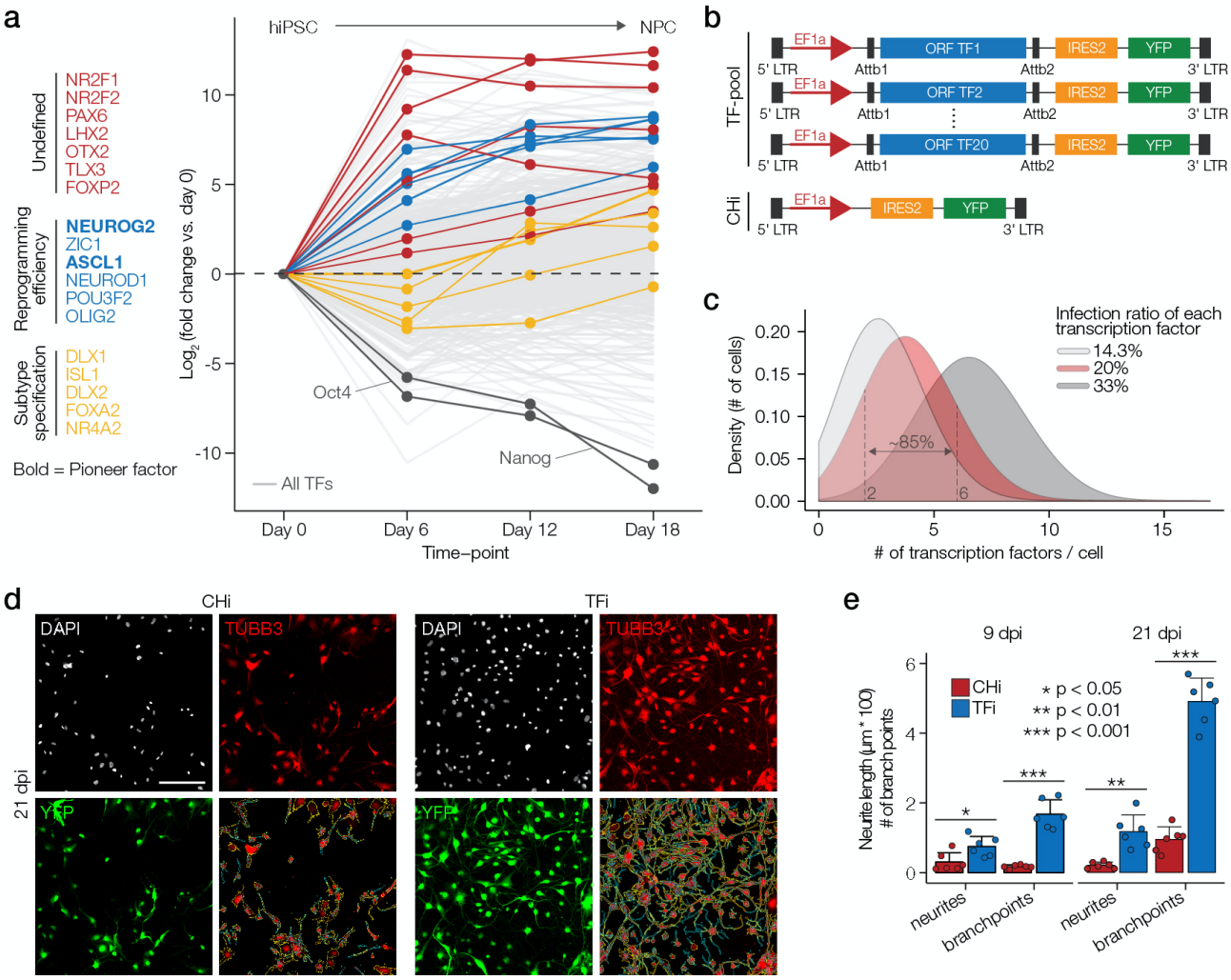
CHIP-seq analysis

To distinguish direct and indirect targets of exogenous transcription factors, we downloaded CHIP-seq datasets from the CHIP-Atlas public repository (https://chip-atlas.org/) and intersected all matching exogenous TFs. Genes within 1 kilobase of the transcription start site and with a combined score greater than 10 were considered direct targets of exogenous TFs (Supplementary Table 1).

Statistics

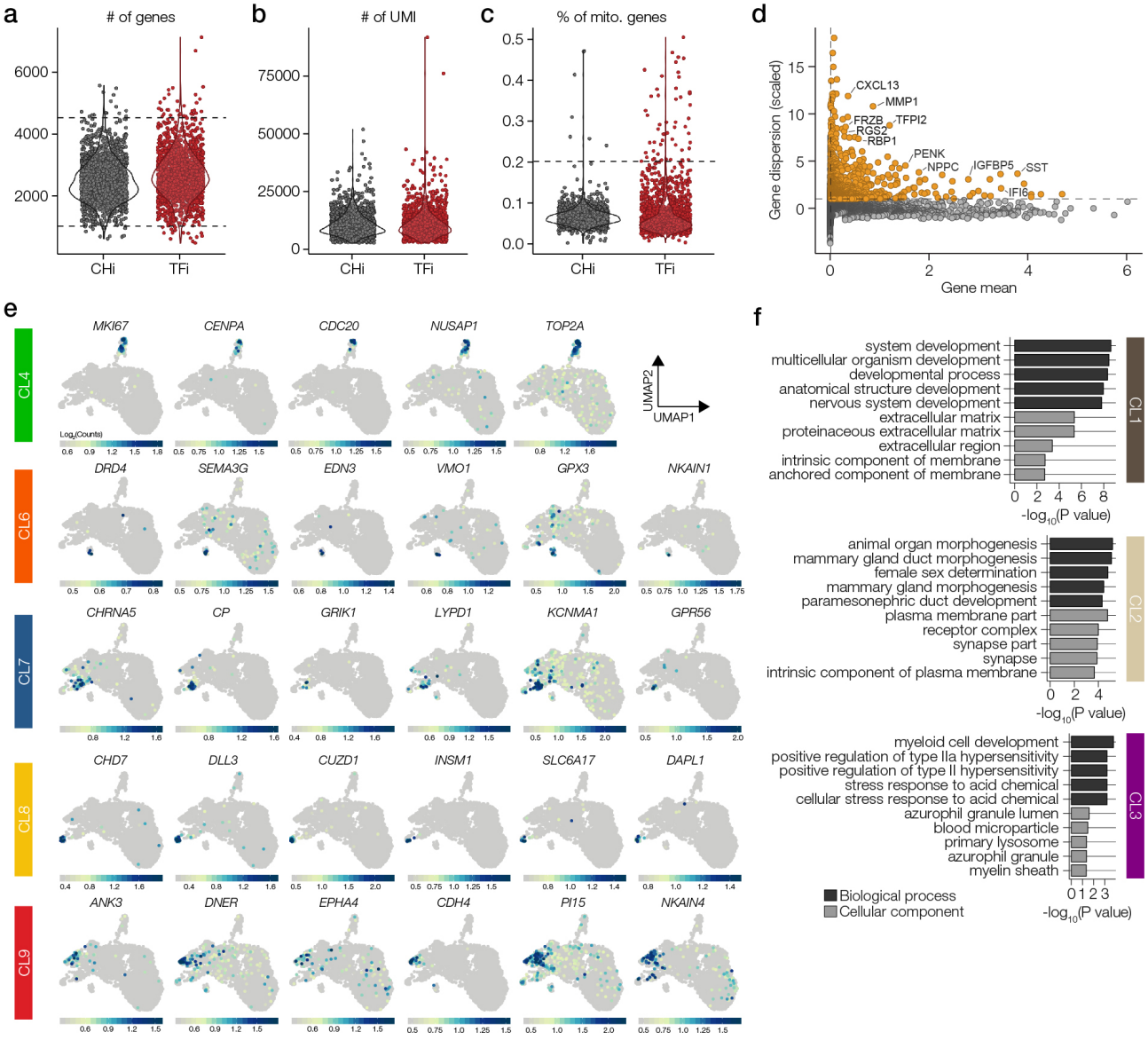
Statistical analyses were performed using R and detailed in the corresponding figure legends. All Student's *t*-tests are two-sided.

Supplementary Figures



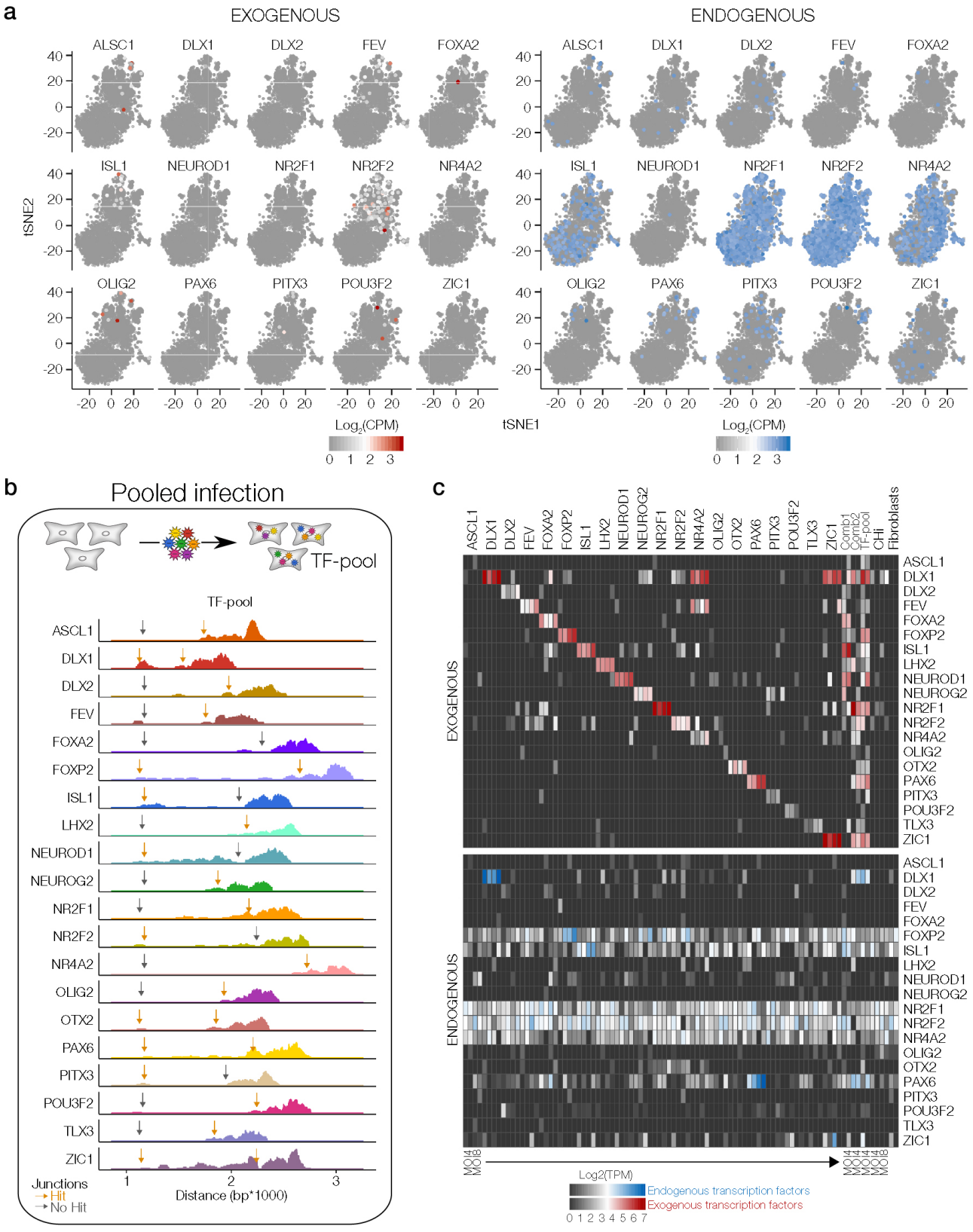
Supplementary Figure 1: Candidate TF expression during iPSC-to-NPC differentiation and neuronal profiling of TFi and CHi. a, Expression fold changes of 18 out of 20 candidate neurogenic TFs (colored) during differentiation of human induced pluripotent stem cells (hiPSC; day 0) into early neuronal progenitor cells (NPC; day 18). Pluripotency markers *OCT4* and *NANOG* are shown in dark grey, all other TFs are shown in light gray. **b**, Schematic overview of the expression vectors of the TF-pool (top) and CHi (bottom). **c**, Theoretical prediction of the number of TFs each cell will be infected with, assuming that each TF infects 14.3% (light gray), 20% (red) and 33% (dark gray) of cells. **d**, Neuronal profiling of CHi and TFi was performed on

pictures of immunostainings for TUBB3 (red). **e**, Quantifications of the length of neurites, scaled by a factor of 100, and the number of branch points of CHi and TFi at 7 dpi and 21 dpi are independently shown on the x-axis. $n = 6$ independent experiments, unpaired Student's t-test. Error bars represent mean + SD.

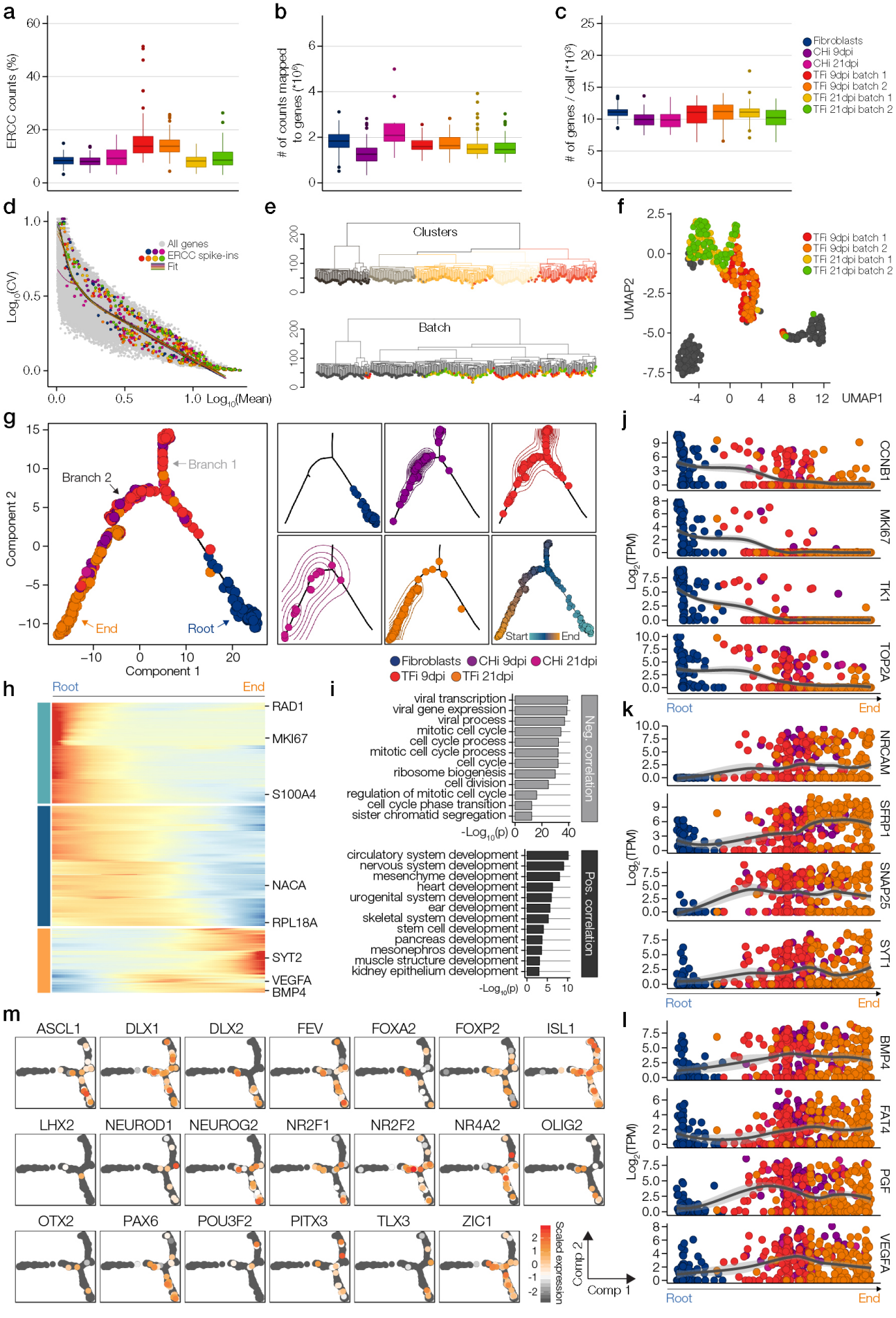


Supplementary Fig. 2: Quality control and marker gene expression of droplet-based scRNA-seq data. a-c, Violin plots of the number of detected genes (a), the number of UMI (b) and the percentage of mitochondrial genes (c) of sequenced CHi (gray) and TFi (red) (n = 3865 cells).

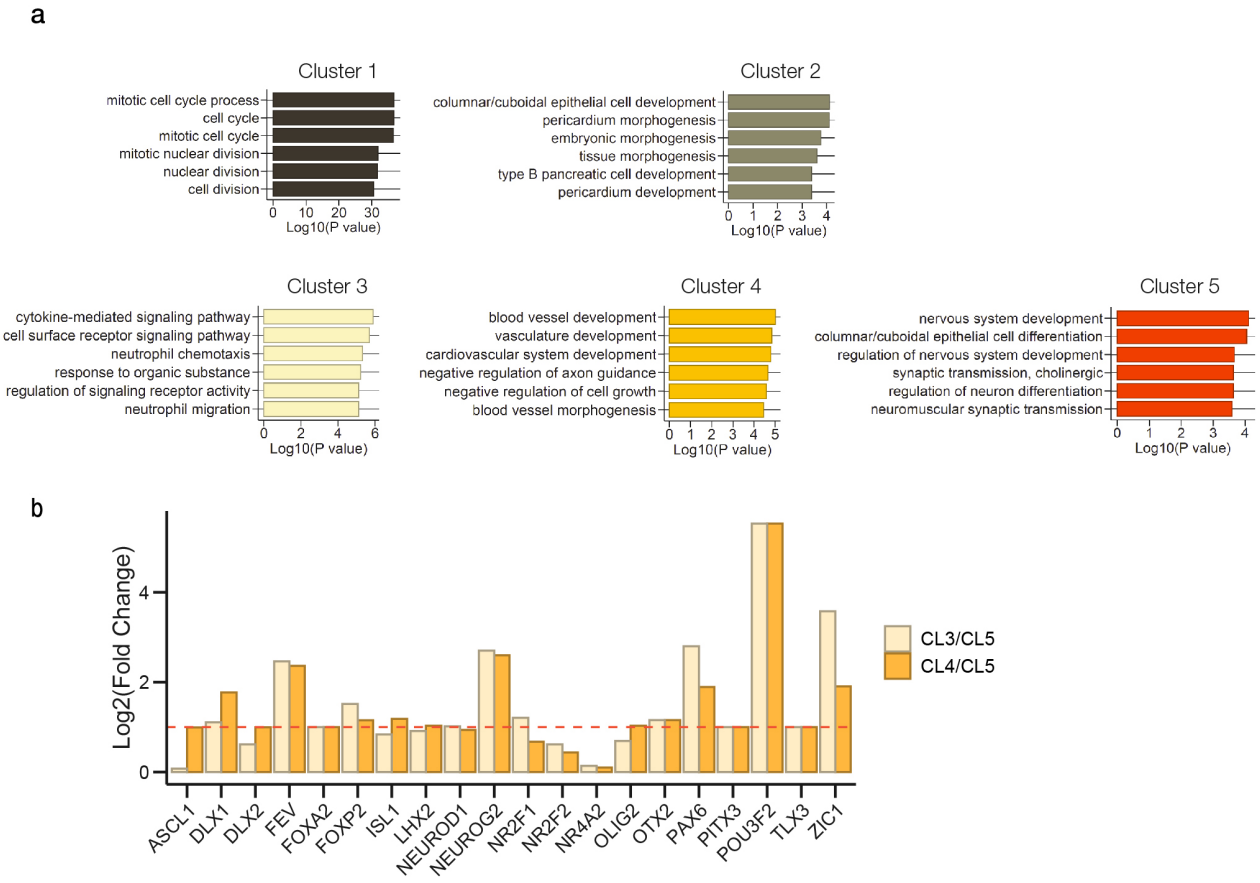
Dashed lines show thresholds applied for quality control. **d**, Mean variation plot of all genes after quality control. Variable genes used for PCA and UMAP are shown in red. **e**, Visualization of the cluster-specific relative expression levels of marker genes using UMAP; cluster colors as in Fig. 2a. The full list of differentially expressed genes of all clusters can be found in Supplementary Table 1. **f**, GO analysis of cluster-specific marker genes in clusters CL1 - CL3. Shown are top 5 GO terms related to biological process (dark grey) and cellular component (light grey) for each cluster.


a
EXOGENOUS
ENDOGENOUS
ALSC1
DLX1
DLX2
FEV
FOXA2
ISL1
NEUROD1
NR2F1
NR2F2
NR4A2
OLIG2
PAX6
PITX3
POU3F2
ZIC1
tSNE2
tSNE1
Log2(CPM)
b
Pooled infection
TF-pool
Junctions
Hit
No Hit
Distance (bp*1000)
c
EXOGENOUS
ENDOGENOUS
Comb1
Comb2
TF-pool
Chi
Fibroblasts
MOI4
MOI8
Log2(TPM)
Endogenous transcription factors
Exogenous transcription factors

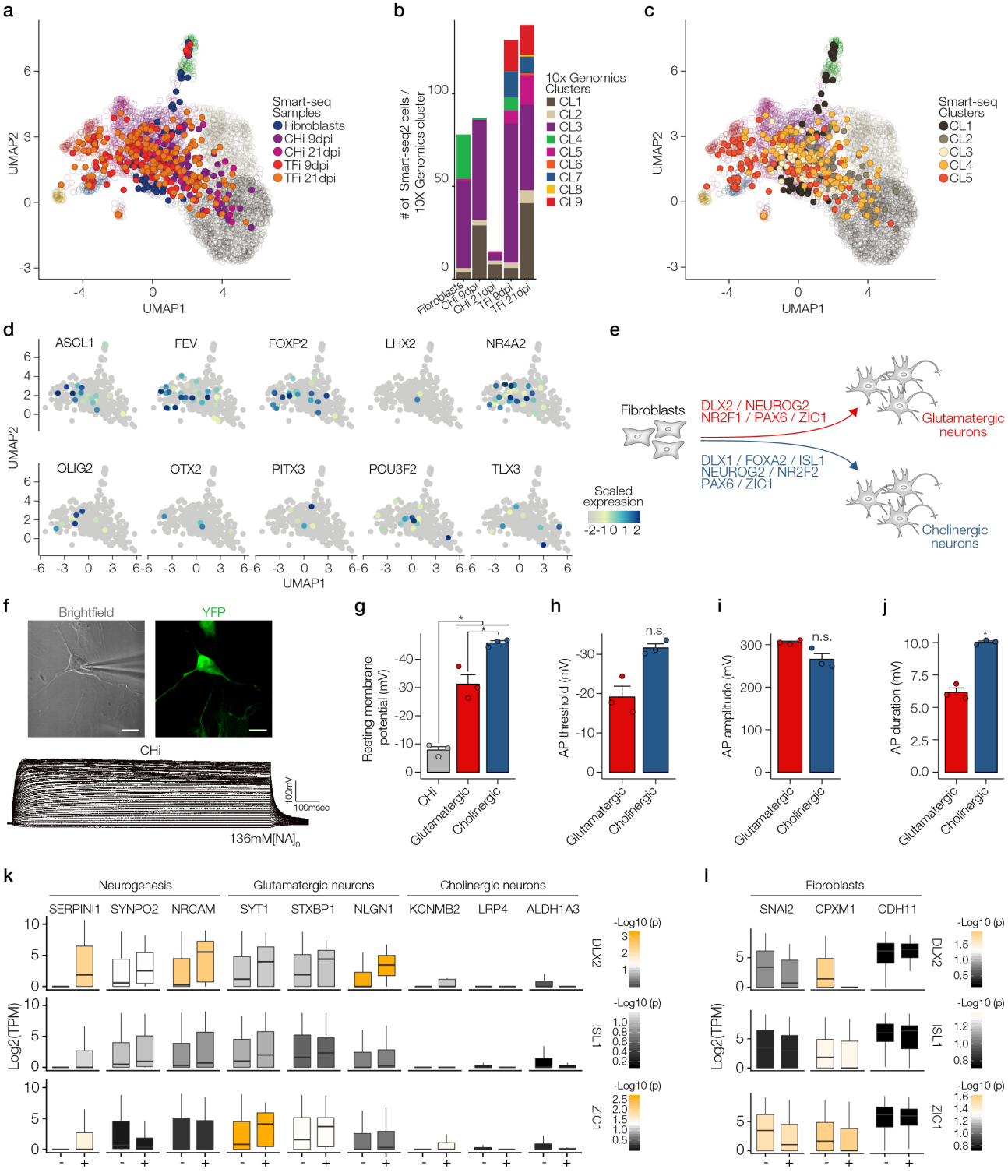
Supplementary Fig. 3: Benchmark of distinguished detection of exogenous and endogenous TFs in bulk and droplet-based single-cell RNA-seq. a, Visualization of log_2-transformed CPM expression values of exogenous (red) and endogenous (blue) TFs on two-dimensional UMAP projections reveal inefficient detection of exogenous TFs in droplet-based scRNA-seq data (n = 3865 cells). **b**, Bulk RNA-seq on pooled infected fibroblasts. Horizontal dimension; distance from the 5' end of the EF1A promoter, vertical dimension; number of aligned paired-end reads. Gray arrows (no overlap) and golden arrows (overlap) mark 5' and 3' junctions of exogenous ORFs. **c**, Heat map showing log_2-transformed TPM values of exogenous (red) and endogenous (blue) TF pairs after alignment using Kallisto without trimming junction sequences. For individually infected fibroblasts and CHi, 2 replicates at an MOI of 4 and 2 replicates at an MOI of 8 were included. For pooled infected fibroblasts, 2 replicates at an MOI of 4 were included.


a
ERCC counts (%)
b
# of counts mapped to genes (*10^6)
c
# of genes / cell (*10^3)
Fibroblasts
CHi 9dpi
CHi 21dpi
TFi 9dpi batch 1
TFi 9dpi batch 2
TFi 21dpi batch 1
TFi 21dpi batch 2
d
Log10(CV)
All genes
ERCC spike-ins
Fit
Log10(Mean)
e
Clusters
Batch
f
UMAP2
UMAP1
g
Branch 1
Branch 2
End
Root
Component 1
Component 2
Start
End
Fibroblasts
CHi 9dpi
CHi 21dpi
TFi 9dpi
TFi 21dpi
h
Root
End
RAD1
MKI67
S100A4
NACA
RPL18A
SYT2
VEGFA
BMP4
i
viral transcription
viral gene expression
viral process
mitotic cell cycle
cell cycle process
mitotic cell cycle process
cell cycle
ribosome biogenesis
cell division
regulation of mitotic cell cycle
cell cycle phase transition
sister chromatid segregation
Neg. correlation
-Log10(p)
circulatory system development
nervous system development
mesenchyme development
heart development
urogenital system development
ear development
skeletal system development
stem cell development
pancreas development
mesonephros development
muscle structure development
kidney epithelium development
Pos. correlation
-Log10(p)
j
CCNB1
MKI67
TK1
TOP2A
Log2(TPM)
Root
End
k
NRCAM
SFRP1
SNAP25
SYT1
Log2(TPM)
Root
End
l
BMP4
FAT4
PGF
VEGFA
Log2(TPM)
Root
End
m
ASCL1
DLX1
DLX2
FEV
FOXA2
FOXP2
ISL1
LHX2
NEUROD1
NEUROG2
NR2F1
NR2F2
NR4A2
OLIG2
OTX2
PAX6
POU3F2
PITX3
TLX3
ZIC1
Scaled expression
Comp 2
Comp 1

Supplementary Fig. 4: Quality control and unsupervised pseudo-temporal ordering of the Smart-seq time-course. a-c, Box plots showing the percentage of ERCC spike-ins (a), the number of counts mapped to genes (b) and the number of detected genes (c) in all 7 runs sequenced for the time-course experiment (n = 446 cells). **d**, Coefficient of variation is plotted against mean TPM, all genes are shown in gray, lines indicate the fit for each run and ERCC spike-ins are shown in colored dots; colors as in A. **e**, Hierarchical clustering of fibroblasts, CHi and TFi at 9 dpi and 21 dpi recapitulates 2-dimensional visualization by UMAP shown in Fig. 4b (top) and reveals no batch effects (bottom). **f**, UMAP projection of transcriptomic data (Fluidigm C1) where TFi batches are colored and all other cells are gray. Clustering of TFi is not batch-dependent. **g**, Pseudo-temporal ordering of time-course data based on genes differentially expressed between 9 dpi and 21 dpi (n = 446 cells). Small squares show the same plot colored by pseudo-temporal values and separate density plots for each sample. **h**, Heat map showing ~1000 genes whose relative expression changes as a function of pseudo-time. **i**, Top GO terms enriched in the top 2000 genes showing negative (top panel) and positive (bottom panel) Pearson correlation with pseudo-temporal values. **j-l**, Dot plots and fit (gray) of $\log_2$-transformed TPM expression values of cell cycle-related genes (*CCNB1*, *MKI67*, *TK1*, *TOP2A*; j), canonical neuronal genes (*NRCAM*, *SFRP1*, *SNAP25*, *SYT1*; k) and genes associated with alternative developmental fates (*BMP4*, *FAT4*, *PGF*, *VEGFA*; l) along pseudo-time; colors as in A. **m**, Visualization of relative expression values of exogenous TFs along pseudo-time where cells are ordered based on the expression of developmental genes.



Supplementary Fig. 5: Functional classification of cluster 5. a, Top six most significant gene ontology terms defined in Figure 4b (left panel) clusters. **b**, Ratios of exogenous TFs in cluster 3 and cluster 4 as compared to cluster 5.



Supplementary Fig. 6: Validation of novel combinations of exogenous TFs. a, Same plot as in Fig. 5a, but cells are colored based on Smart-seq cell identity. **b**, Quantification of the percentage of Smart-seq cells mapping to 10x Genomics clusters. **c**, Same plot as in Fig. 5a, but cells are

colored based on Smart-seq cluster identity. **d**, Visualization of scaled expression values of exogenous TFs that showed no significant enrichment (Fisher's exact test, $p > 0.05$) in any cluster on two-dimensional UMAPs. **e**, Schematic summary of exogenous TFs showing enrichment in glutamatergic and/or cholinergic clusters. **f**, Top: Recording electrode patched onto a YFP^+ cell with a stimulation electrode. Scale bars, 20μm. Bottom: The generation of the action potential in control cells. Representative traces in the presence of extracellular Na^+ were recorded using the current-clamp protocol. **g-j**, Electrophysiological properties of control cells (CHi), iN infected with *DLX2*, *NEUROG2*, *PAX6*, *ZIC1* (glutamatergic) or *DLX1*, *ISL1*, *NEUROG2*, *PAX6* (cholinergic). $n = 3$ independent experiments, unpaired Student's *t*-test. Error bars represent mean + SD. **k-l**, Box plots showing the Log2-transformed TPM values of neurogenic and neuronal subtype-specific genes (k) and fibroblast-specific genes (l) in cells with (+) or without (-) exogenous *DLX2* (top), *ISL1* (middle) and *ZIC1* (bottom). Box plots are colored based on -Log10-transformed p-values.