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^{TM} 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^{TM} 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$^{\text{TM}}$ System (Clontech). The amplified cDNAs were harvested into 96 well plates and quantified with Quant-iT$^{\text{TM}}$ 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 ($p < 1e-4$) 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.

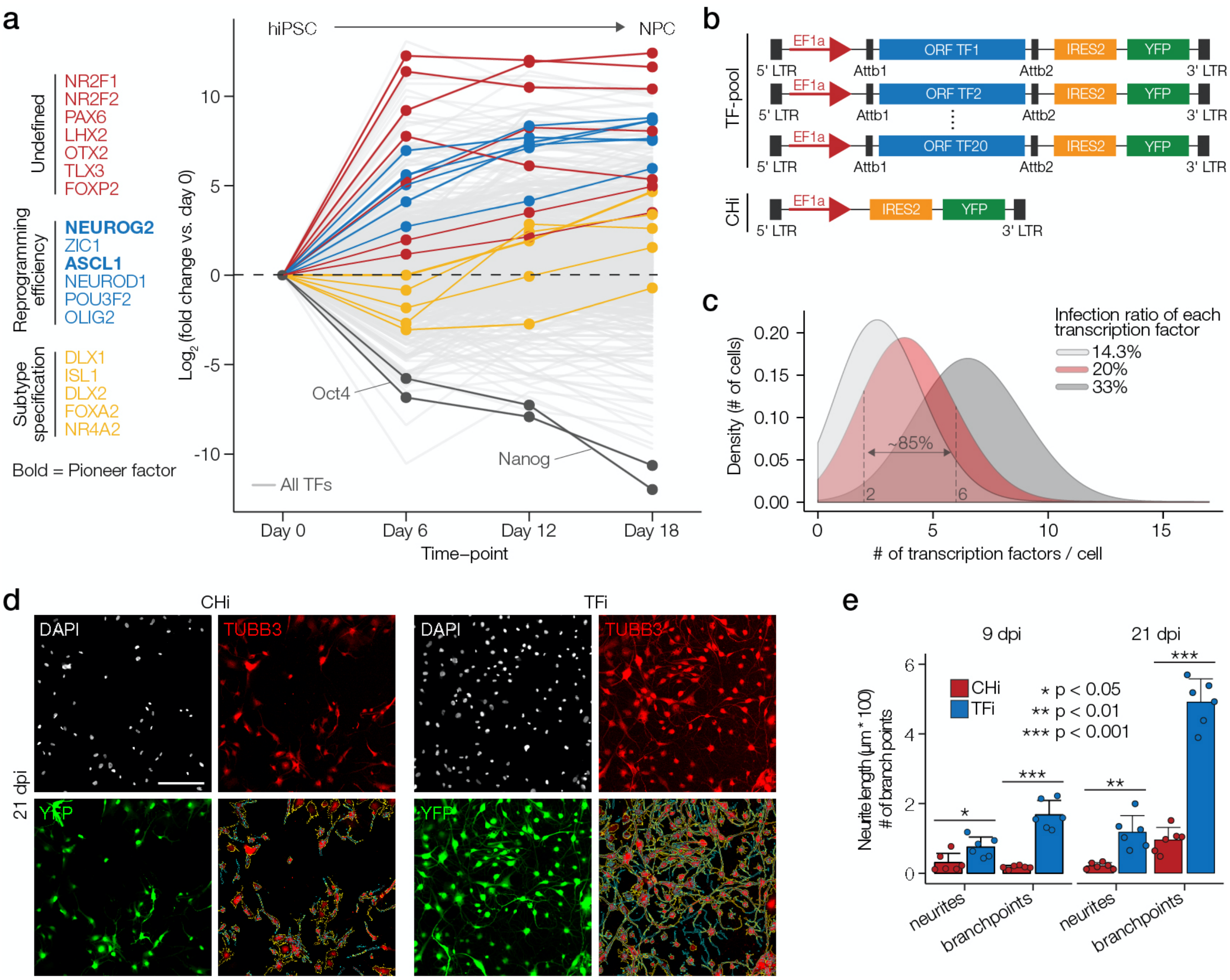
CHIP-seq analysis

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).

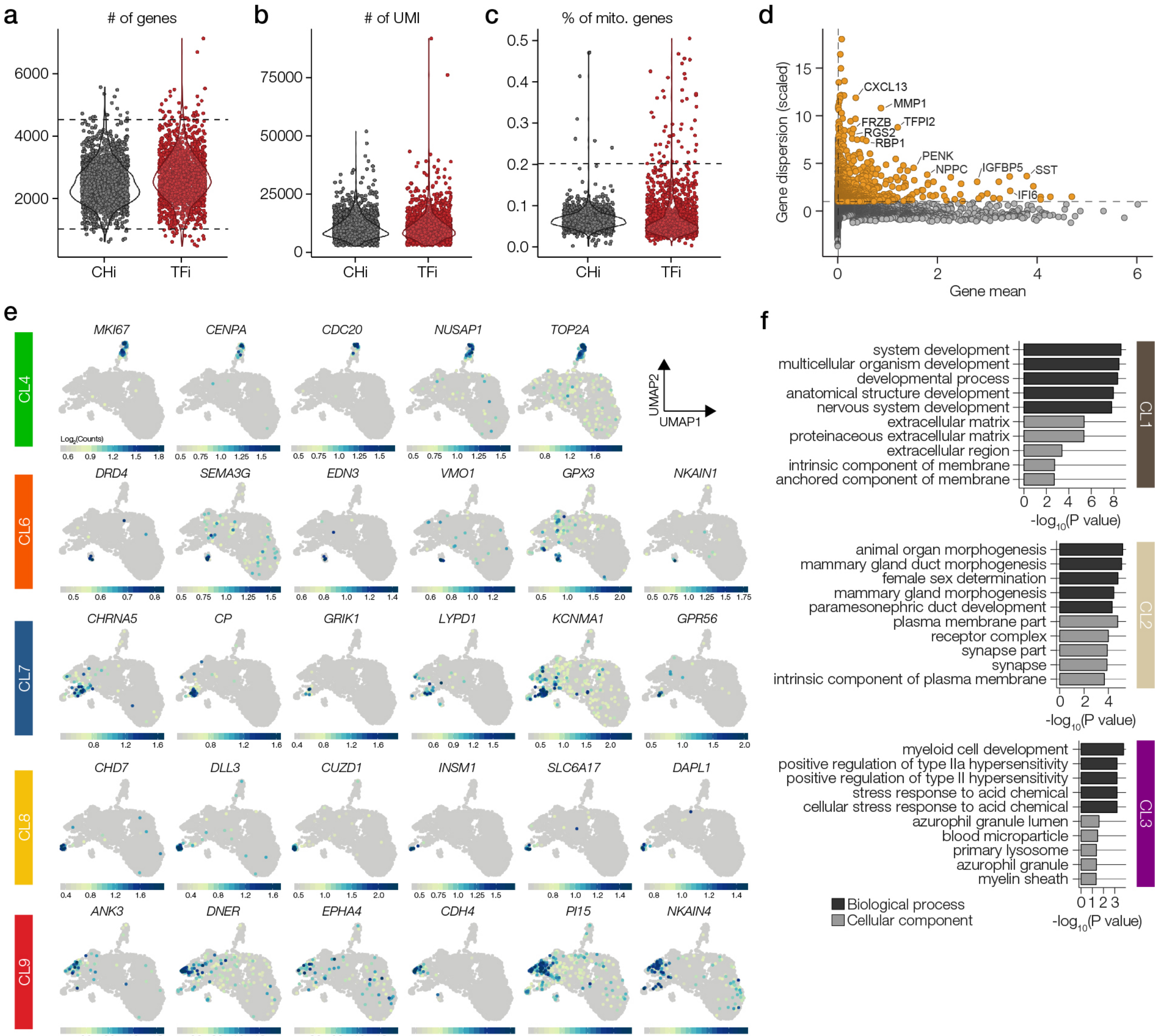
Statistics

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

Supplementary Figures

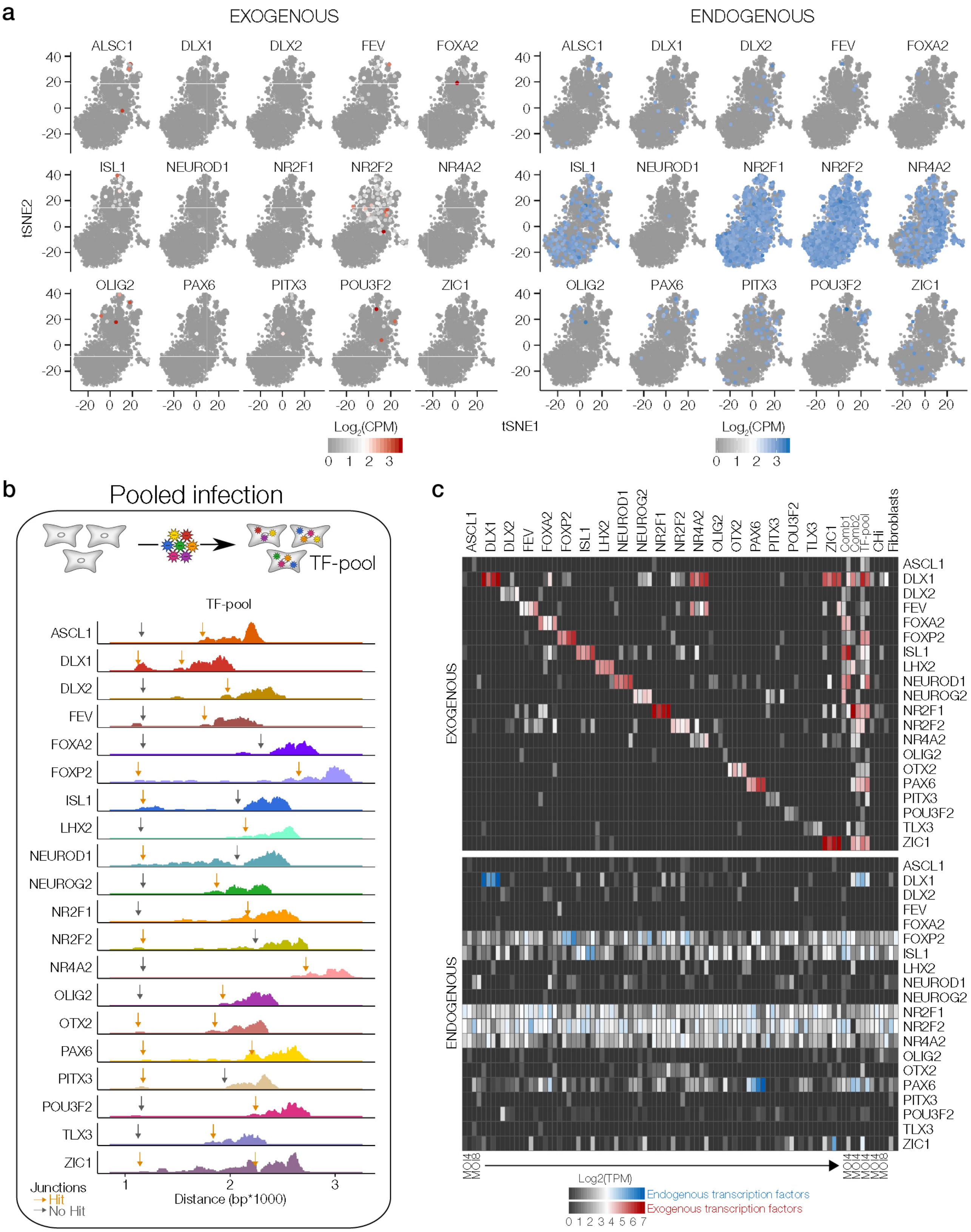


Supplementary Figure 1: Candidate TF expression during iPSC-to-NPC differentiation and neuronal profiling of TFi and CHi. a, Expression fold changes of 18 out of 20 candidate neurogenic TFs (colored) during differentiation of human induced pluripotent stem cells (hiPSC; day 0) into early neuronal progenitor 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.

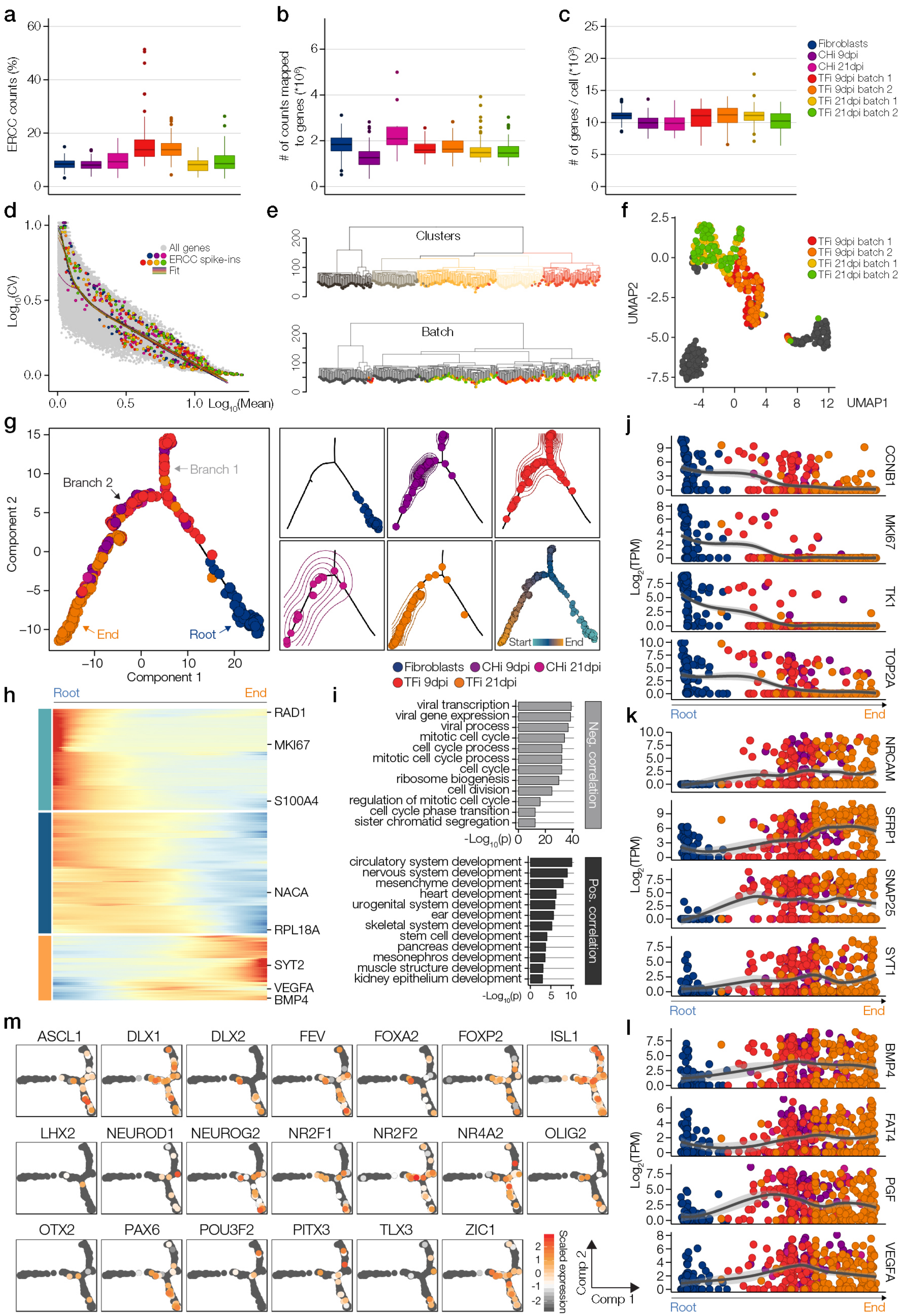


Supplementary Fig. 2: Quality control and marker gene expression of droplet-based scRNA-seq data. a-c, Violin plots of the number of detected genes (a), the number of UMI (b) and the percentage of mitochondrial genes (c) of sequenced CHi (gray) and TFi (red) (*n* = 3865 cells).

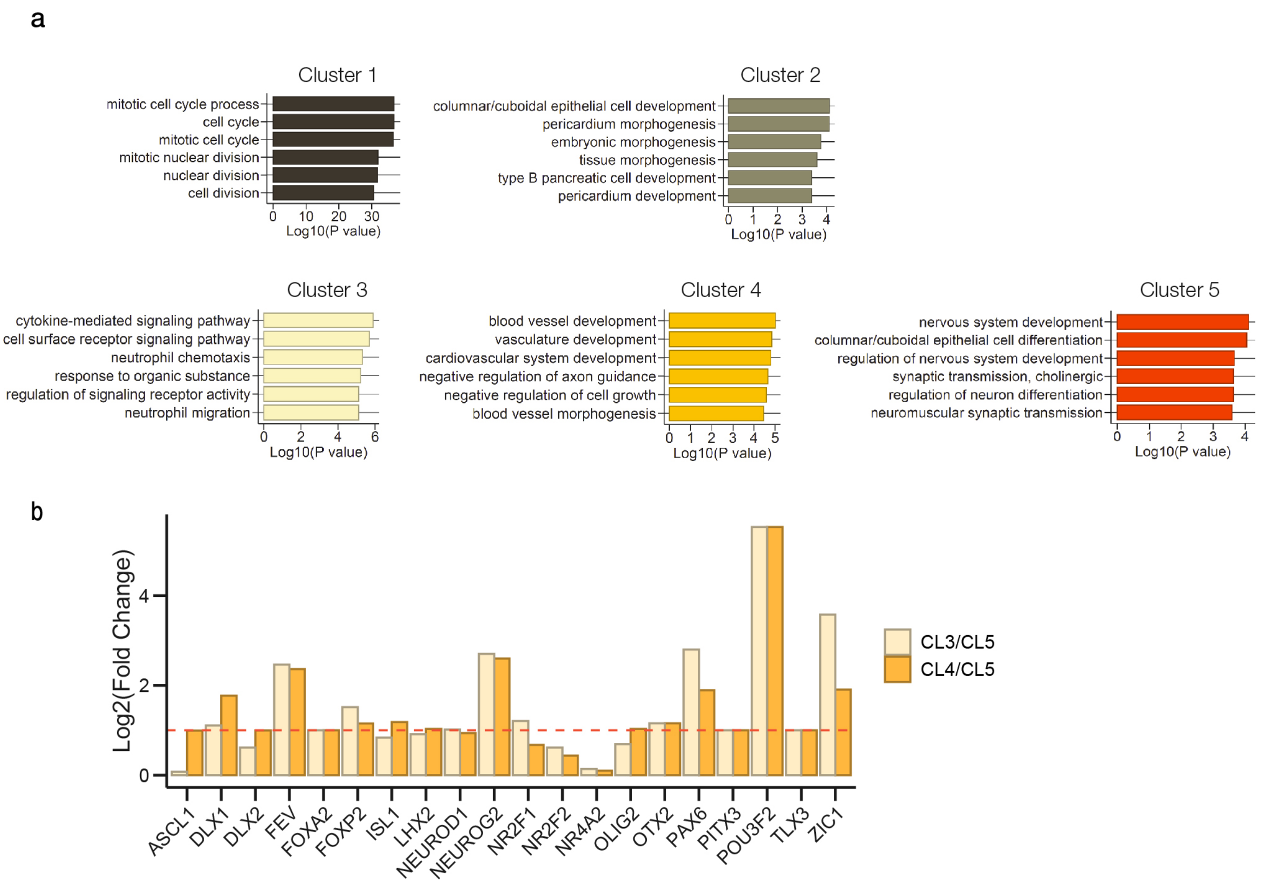
Dashed lines show thresholds applied for quality control. **d**, Mean variation plot of all genes after quality control. Variable genes used for PCA and UMAP are shown in red. **e**, Visualization of the cluster-specific relative expression levels of marker genes using UMAP; cluster colors as in Fig. 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
40
20
0
-20
Log2(CPM)
0 1 2 3
b
Pooled infection
TF-pool
ASCL1
DLX1
DLX2
FEV
FOXA2
FOXP2
ISL1
LHX2
NEUROD1
NEUROG2
NR2F1
NR2F2
NR4A2
OLIG2
OTX2
PAX6
PITX3
POU3F2
TLX3
ZIC1
Junctions
→ Hit
→ No Hit
1
2
3
Distance (bp*1000)
c
ASCL1
DLX1
DLX2
FEV
FOXA2
FOXP2
ISL1
LHX2
NEUROD1
NEUROG2
NR2F1
NR2F2
NR4A2
OLIG2
OTX2
PAX6
PITX3
POU3F2
TLX3
ZIC1
Comb1
Comb2
TF-pool
CHi
Fibroblasts
EXOGENOUS
ENDOGENOUS
MOI4
MOI8
MOI4
MOI4
MOI4
MOI8
Log2(TPM)
Endogenous transcription factors
Exogenous transcription factors
0 1 2 3 4 5 6 7

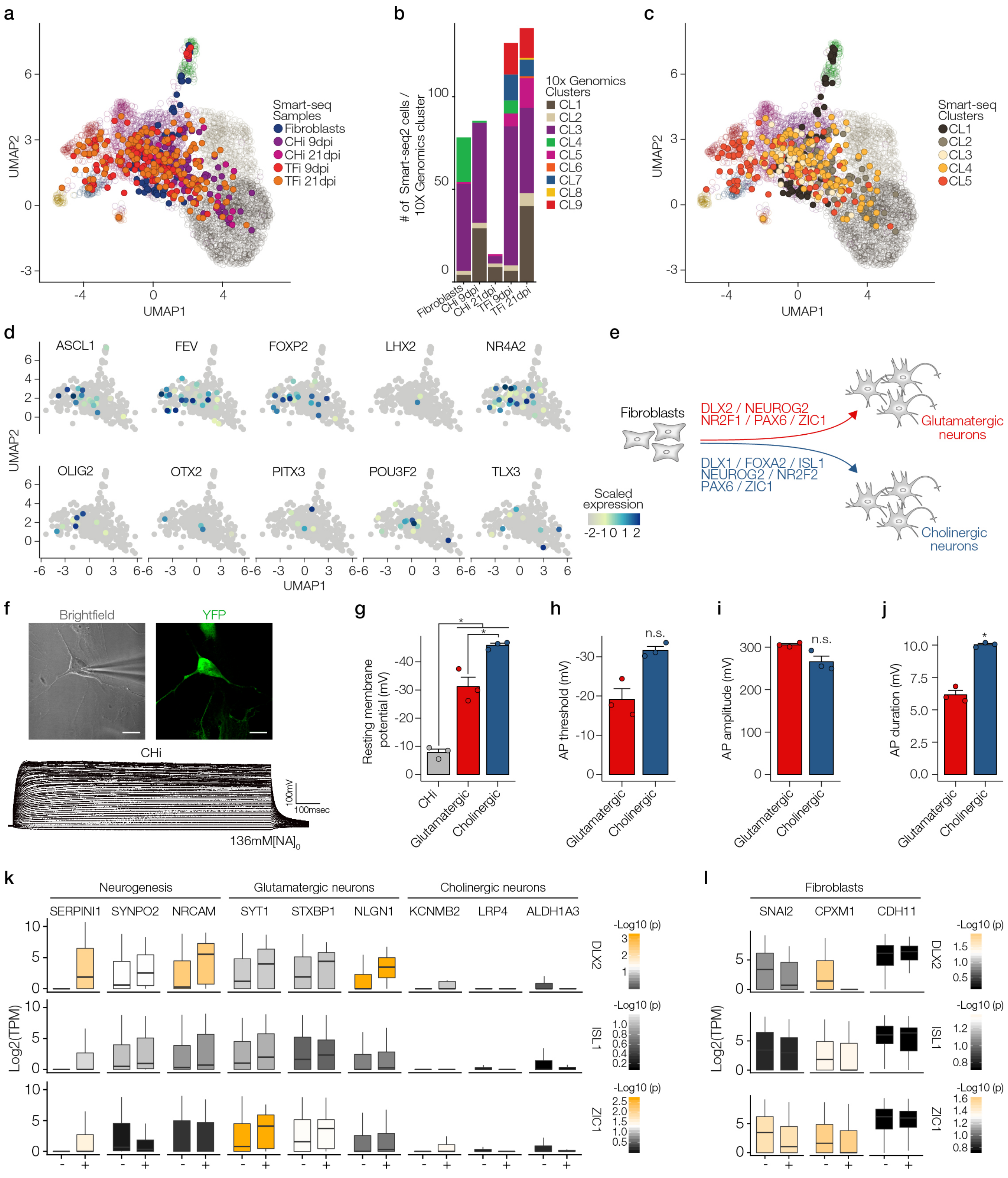
Supplementary Fig. 3: Benchmark of distinguished detection of exogenous and endogenous TFs in bulk and droplet-based single-cell RNA-seq. a, Visualization of $\log_2$-transformed CPM expression values of exogenous (red) and endogenous (blue) TFs on two-dimensional UMAP projections reveal inefficient 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 (%)
60
40
20
0
b
# of counts mapped to genes (*10⁶)
6
4
2
0
c
# of genes / cell (*10³)
25
20
15
10
5
0
Fibroblasts
CHi 9dpi
CHi 21dpi
TFi 9dpi batch 1
TFi 9dpi batch 2
TFi 21dpi batch 1
TFi 21dpi batch 2
d
Log10(CV)
1.0
0.5
0.0
All genes
ERCC spike-ins
Fit
0.0
0.5
1.0
Log10(Mean)
e
Clusters
200
100
0
Batch
200
100
0
f
UMAP2
2.5
0.0
-2.5
-5.0
-7.5
-4
0
4
8
12
UMAP1
TFi 9dpi batch 1
TFi 9dpi batch 2
TFi 21dpi batch 1
TFi 21dpi batch 2
g
Component 2
15
10
5
0
−5
−10
Branch 1
Branch 2
End
Root
−10
0
10
20
Component 1
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
-Log10(p)
0
20
40
Neg. correlation
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
-Log10(p)
0
5
10
Pos. correlation
j
Log2(TPM)
CCNB1
MKI67
TK1
TOP2A
Root
End
k
Log2(TPM)
NRCAM
SFRP1
SNAP25
SYT1
Root
End
l
Log2(TPM)
BMP4
FAT4
PGF
VEGFA
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
2
1
0
-1
-2
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.