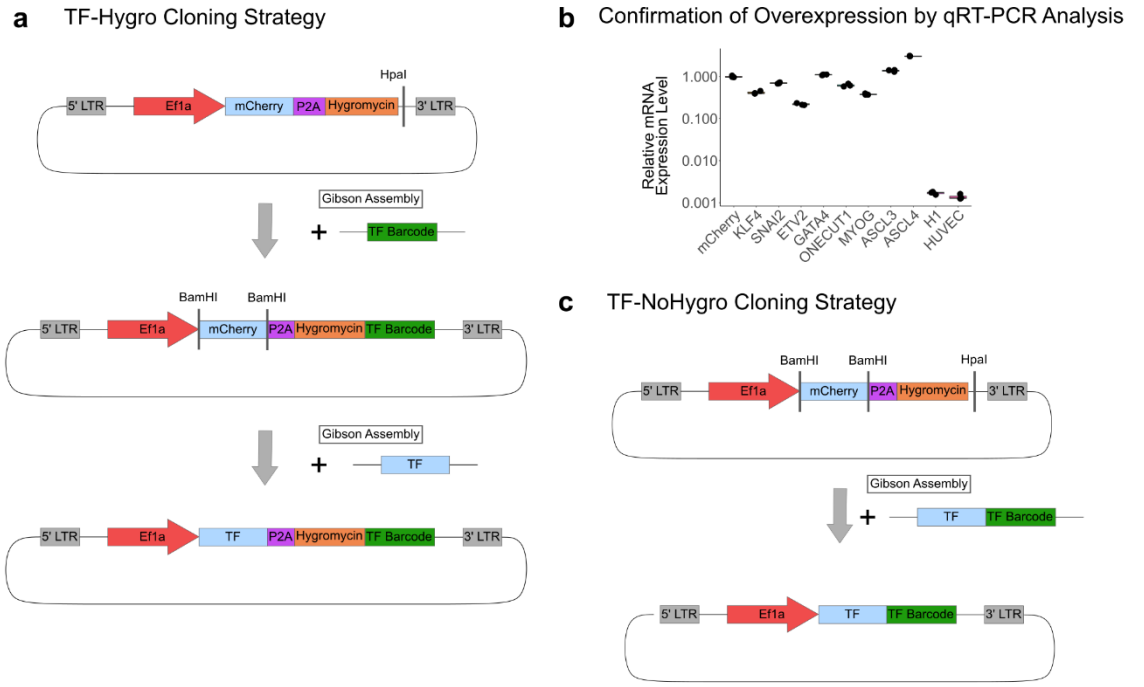
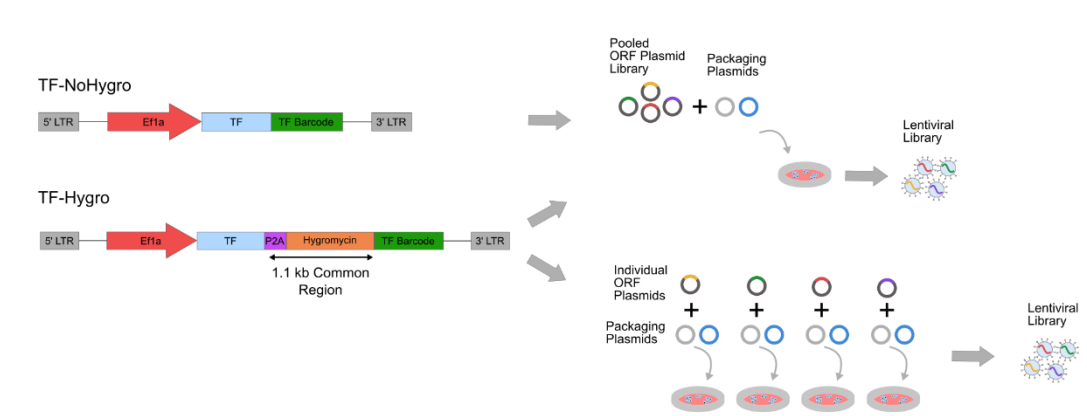
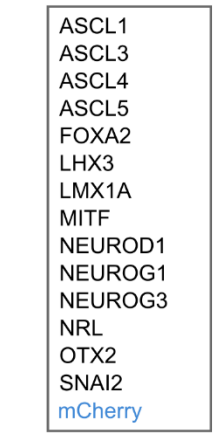
Supplemental Information

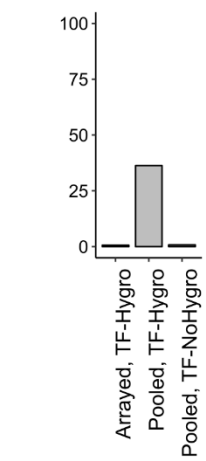
Mapping Cellular Reprogramming via Pooled Overexpression Screens with Paired Fitness and Single Cell RNA-Sequencing Readout

Udit Parekh, Yan Wu, Dongxin Zhao, Atharv Worlikar, Neha Shah, Kun Zhang, Prashant Mali


a
TF-Hygro Cloning Strategy
HpaI
5' LTR
Ef1a
mCherry
P2A
Hygromycin
3' LTR
Gibson Assembly
TF Barcode
BamHI
BamHI
TF
b
Confirmation of Overexpression by qRT-PCR Analysis
Relative mRNA Expression Level
1.000
0.100
0.010
0.001
mCherry
KLF4
SNAI2
ETV2
GATA4
ONECUT1
MYOG
ASCL3
ASCL4
H1
HUVEC
c
TF-NoHygro Cloning Strategy


d
Schematic of different lentiviral ORF overexpression vector design and packaging methods
TF-NoHygro
TF-Hygro
1.1 kb Common Region
Pooled ORF Plasmid Library
Packaging Plasmids
Lentiviral Library
Individual ORF Plasmids


e
Neural TF Sub-Library
ASCL1
ASCL3
ASCL4
ASCL5
FOXA2
LHX3
LMX1A
MITF
NEUROD1
NEUROG1
NEUROG3
NRL
OTX2
SNAI2
mCherry


f
Barcode Shuffling Rates
100
75
50
25
0
Arrayed, TF-Hygro
Pooled, TF-Hygro
Pooled, TF-NoHygro

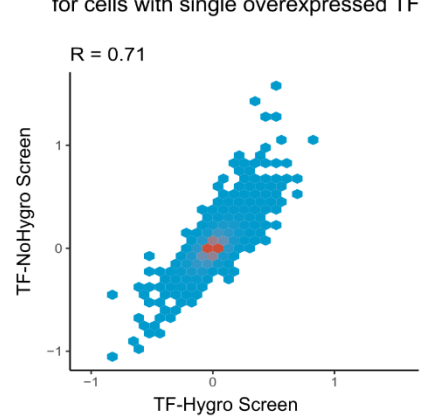

g
Correlation of Regression Coefficients for cells with single overexpressed TF
R = 0.71
TF-NoHygro Screen
TF-Hygro Screen
1
0
−1

Figure S1: Vector design, cloning and characterization, and identification of significant TFs. Related to Figure 1 and STAR Methods sections "Library Preparation", "Quantification of Barcode Shuffling" and "Replicate Correlation". (**a**) The construction of TF-Hygro involved two steps: (i) insertion of a pool of barcodes into the backbone after digestion with HpaI, (ii) individually substituting mCherry with TFs after digestion with BamHI. (**b**) Confirmation of exogenous expression of select overexpressed TFs by qRT-PCR analysis. Data for all assays were normalized to *GAPDH* and expressed relative to control mCherry-transduced cells. Untransduced H1 hESCs and HUVECs were used as negative controls. Primers were chosen such that they amplified a portion of the transcript in the hygromycin resistance region. This was done to avoid amplification of any endogenous transcripts, and since the overexpression is driven by a single promoter the TF, P2A peptide and the hygromycin resistance are on a single transcript. (**c**) The construction of TF-NoHygro involved a single step, individually substituting mCherry with TFs after digestion with BamHI and HpaI. Barcodes were inserted via the TF amplification primers, using which TFs were amplified out of TF-Hygro plasmids. (**d**) Schematic of different lentiviral ORF overexpression vector designs and pooled vs arrayed packaging methods. (**e**) 14-element Neural TF sub-library. (**f**) Quantification of shuffling rates for individually packaged virus using the TF-Hygro design, pooled virus packaging using the TF-Hygro design, and pooled virus packaging using the TF-NoHygro design. (**g**) Correlation between single TF regression coefficients for scRNA-seq screens on the Neural TF library using the original and updated vectors.

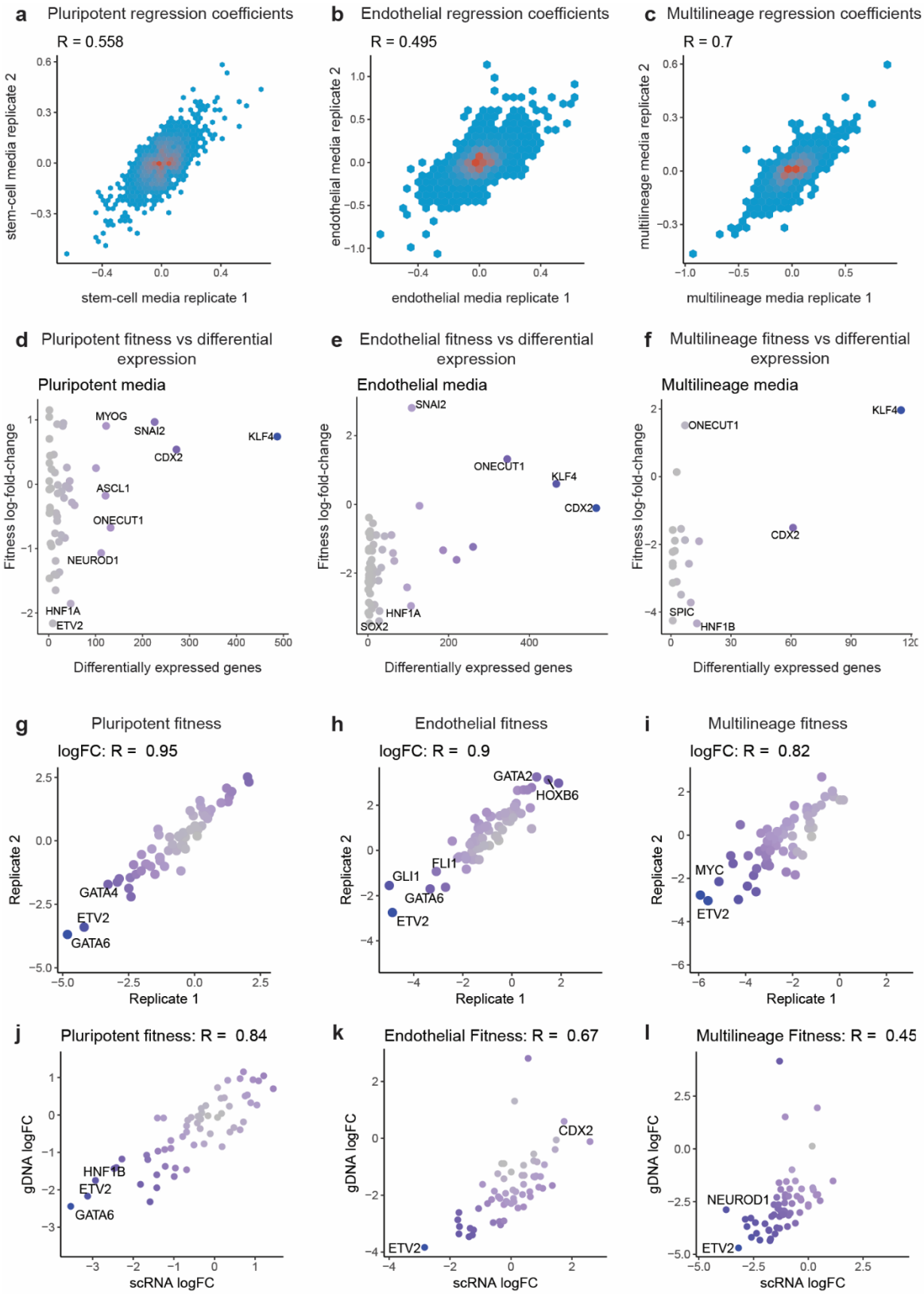

a
Pluripotent regression coefficients
R = 0.558
stem-cell media replicate 2
stem-cell media replicate 1
b
Endothelial regression coefficients
R = 0.495
endothelial media replicate 2
endothelial media replicate 1
c
Multilineage regression coefficients
R = 0.7
multilineage media replicate 2
multilineage media replicate 1
d
Pluripotent fitness vs differential expression
Pluripotent media
MYOG
SNAI2
KLF4
CDX2
ASCL1
ONECUT1
NEUROD1
HNF1A
ETV2
Fitness log-fold-change
Differentially expressed genes
e
Endothelial fitness vs differential expression
Endothelial media
SNAI2
ONECUT1
KLF4
CDX2
HNF1A
SOX2
Fitness log-fold-change
Differentially expressed genes
f
Multilineage fitness vs differential expression
Multilineage media
KLF4
ONECUT1
CDX2
SPIC
HNF1B
Fitness log-fold-change
Differentially expressed genes
g
Pluripotent fitness
logFC: R = 0.95
GATA4
ETV2
GATA6
Replicate 2
Replicate 1
h
Endothelial fitness
logFC: R = 0.9
GATA2
HOXB6
FLI1
GLI1
GATA6
ETV2
Replicate 2
Replicate 1
i
Multilineage fitness
logFC: R = 0.82
MYC
ETV2
Replicate 2
Replicate 1
j
Pluripotent fitness: R = 0.84
HNF1B
ETV2
GATA6
gDNA logFC
scRNA logFC
k
Endothelial Fitness: R = 0.67
CDX2
ETV2
gDNA logFC
scRNA logFC
l
Multilineage Fitness: R = 0.45
NEUROD1
ETV2
gDNA logFC
scRNA logFC

Figure S2: Correlation of scRNA-seq data between replicates, correlation between transcriptomic and fitness effects, correlation of fitness estimates from scRNA-seq genotyped cell counts between replicates, and correlation of fitness estimates from scRNA-seq genotyped cell counts vs bulk fitness from genomic DNA. Related to Figure 1 and STAR Methods sections "Replicate Correlation" and "Fitness Effects Analysis". (**a**) Correlation between coefficients in the pluripotent medium screens. (**b**) Correlation between coefficients in the unilineage medium screens. (**c**) Correlation between coefficients in the multilineage medium screens. For (**a**)-(**c**), correlation was between regression coefficients, with each coefficient representing the effect of a TF on an individual gene. We subset to coefficients that are nonzero with an adjusted p-value (FDR) of less than 0.5 in *either* replicate to filter out coefficients that are zero in both replicates. (**d**)-(**f**) Correlation of the number of differentially expressed genes for each TF vs the fitness effect (log-FC) for: **(d)** Pluripotent medium. **(e)** Unilineage medium (EGM). **(f)** Multilineage medium. (**g**)-(**i**) Correlation between replicates of fitness estimates from scRNA-seq genotyped cell counts for: **(g)** Pluripotent medium. **(h)** Unilineage medium (EGM). **(i)** Multilineage medium. (**j**)-(**l**) Correlation between log fold change of TF counts vs plasmid library control for genomic DNA reads vs cell counts fitness for: **(j)** Pluripotent medium **(k)** Unilineage medium (EGM) **(l)** Multilineage medium

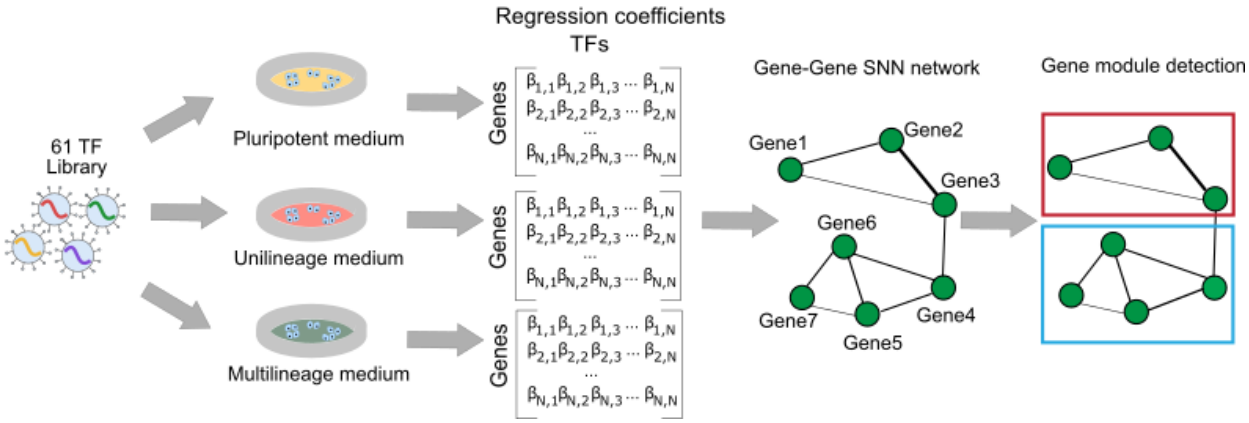


Figure S3: Schematic for gene-gene co-perturbation network analysis, related to Figure 1 and STAR Methods section "Gene Co-perturbation Network and Module Detection". A SNN network is built from the linear model coefficients and the network is then segmented into gene modules. Genes have a highly weighted edge between them if they respond similarly to TF overexpression.

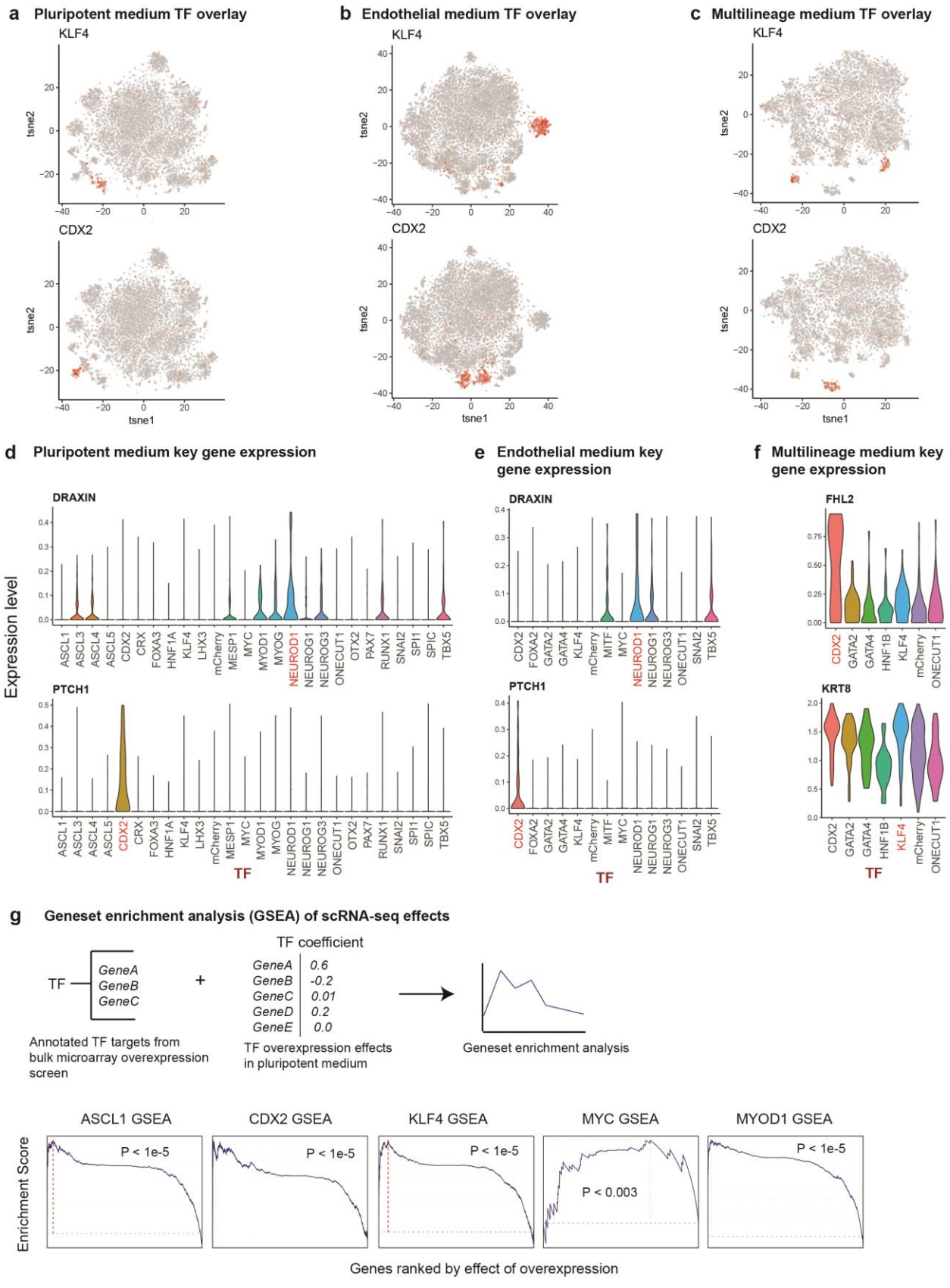

a Pluripotent medium TF overlay
KLF4
CDX2
tsne1
tsne2
b Endothelial medium TF overlay
KLF4
CDX2
tsne1
tsne2
c Multilineage medium TF overlay
KLF4
CDX2
tsne1
tsne2
d Pluripotent medium key gene expression
DRAXIN
PTCH1
Expression level
ASCL1 ASCL3 ASCL4 ASCL5 CDX2 CRX FOXA3 HNF1A KLF4 LHX3 mCherry MESP1 MYC MYOD1 MYOG NEUROD1 NEUROG1 NEUROG3 ONECUT1 OTX2 PAX7 RUNX1 SNAI2 SPI1 SPIC TBX5
TF
e Endothelial medium key gene expression
DRAXIN
PTCH1
CDX2 FOXA2 GATA2 GATA4 KLF4 mCherry MITF MYC NEUROD1 NEUROG1 NEUROG3 ONECUT1 SNAI2 TBX5
TF
f Multilineage medium key gene expression
FHL2
KRT8
CDX2 GATA2 GATA4 HNF1B KLF4 mCherry ONECUT1
TF
g Geneset enrichment analysis (GSEA) of scRNA-seq effects
TF
GeneA
GeneB
GeneC
Annotated TF targets from bulk microarray overexpression screen
TF coefficient
GeneA 0.6
GeneB -0.2
GeneC 0.01
GeneD 0.2
GeneE 0.0
TF overexpression effects in pluripotent medium
Geneset enrichment analysis
ASCL1 GSEA
CDX2 GSEA
KLF4 GSEA
MYC GSEA
MYOD1 GSEA
P < 1e-5
P < 0.003
Enrichment Score
Genes ranked by effect of overexpression

Figure S4: Analysis and validation of significant TFs. Related to Figure 1 and STAR Methods section "Replicate Correlation". (**a**) tSNE plot of screens in pluripotent medium, color coded by select TFs (*KLF4*, *CDX2*) (**b**) tSNE plot of screens in unilineage medium, color coded by select TFs (*KLF4*, *CDX2*) (**c**) tSNE plot of screens in multilineage medium, color coded by select TFs (*KLF4*, *CDX2*) (**d**) Expression level of *DRAXIN* and *PTCH1* in screens in pluripotent medium with the TFs expected to upregulate expression of each highlighted in red (**e**) Expression level of *DRAXIN* and *PTCH1* in screens in unilineage medium with the TFs expected to upregulate expression of each highlighted in red (**f**) Expression level of *FHL2* and *KRT8* in screens in unilineage medium with the TFs expected to upregulate expression of each highlighted in red. (**g**) Geneset enrichment analysis for homologous genes in mESCs upon overexpression of TFs[1]. TFs present in both datasets – *ASCL1*, *CDX2*, *MYC, KLF4* and *MYOD1* – display a highly significant degree of overlap in their effects.

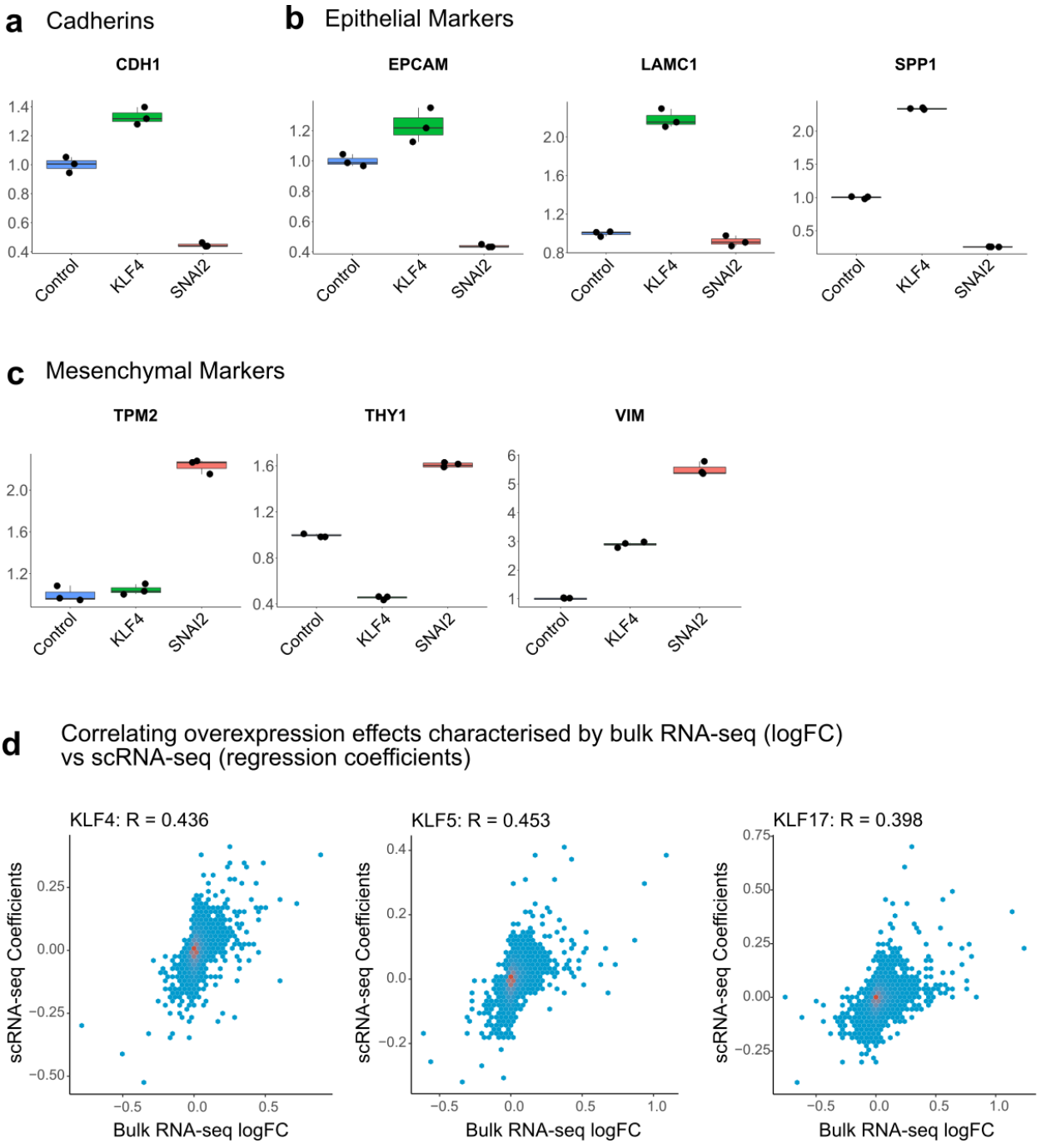


Figure S5: Confirmatory assays for effects of *KLF4* and *SNAI2* on key genes involved in EMT. Related to Figure 2 and STAR Methods section "Bulk RNA-seq Analysis and Correlation". (**a**) qRT-PCR analysis of signature cadherin during EMT: *CDH1* at day 5 post-transduction in pluripotent stem cell medium. (**b**) qRT-PCR analysis of signature epithelial marker genes during EMT: *EPCAM*, *LAMC1* and *SPP1* at day 5 post-transduction in pluripotent stem cell medium. **(c)** qRT-PCR analysis of signature

mesenchymal marker genes during EMT: *TPM2*, *THY1* and *VIM* at day 5 post-transduction in pluripotent stem cell medium. Data for all assays were normalized to *GAPDH* and expressed relative to control cells. (**d**) Correlation of scRNA-seq regression coefficients versus log fold change in bulk RNA-seq for *KLF4*, *KLF5* and *KLF17*. Log fold change was calculated from bulk RNA-seq data for cells transduced with a TF versus cells transduced with control mCherry virus.

Gene	Forward Primer (5' -> 3')	Reverse Primer (5' -> 3')
CDH5	AGACCACGCCTCTGTCATGTACCAAATC	CACGATCTCATACCTGGCCTGCTTC
PECAM1	GGTCAGCAGCATCGTGGTCAACATAAC	TGGAGCAGGACAGGTTCAGTCTTTCA
VWF	TCTCCGTGGTCCTGAAGCAGACATA	AGGTTGCTGCTGGTGAGGTCATT
KDR	AGCCATGTGGTCTCTCTGGTTGTGTATG	GTTTGAGTGGTGCCGTACTGGTAGGA
NANOG	TTTGTGGGCCTGAAGAAAACT	AGGGCTGTCCTGAATAAGCAG
POU5F1	CTTGAATCCCGAATGGAAAGGG	GTGTATATCCCAGGGTGATCCTC
SOX2	TACAGCATGTCCTACTCGCAG	GAGGAAGAGGTAACCACAGGG
DNMT3B	GAGTCCATTGCTGTTGGAACCG	ATGTCCCTCTTGTCGCCAACCT
SALL2	CAGCGGAAACCCCAACAGTTA	GAGGGTCAGTAGAACATGCGT
DPPA4	GACCTCCACAGAGAAGTCGAG	TGCCTTTTTCTTAGGGCAGAG
VIM	AGTCCACTGAGTACCGGAGAC	CATTTCACGCATCTGGCGTTC
CDH1	CGAGAGCTACACGTTCACGG	GGGTGTCGAGGGAAAAATAGG
CDH2	AGCCAACCTTAACTGAGGAGT	GGCAAGTTGATTGGAGGGATG
EPCAM	TGATCCTGACTGCGATGAGAG	CTTGTCTGTTCTTCTGACCCC
LAMC1	GGCAACGTGGCCTTTTCTAC	AGTGGCAGTTACCCATTCCTG
SPP1	GAAGTTTCGCAGACCTGACAT	GTATGCACCATTCAACTCCTCG
THY1	ATCGCTCTCCTGCTAACAGTC	CTCGTACTGGATGGGTGAACT
TPM2	CTGAGACCCGAGCAGAGTTTG	TGAATCTCGACGTTCTCCTCC
TF Overexpression	TTAGCCAGACGAGCGGGTTC	GTCGTCCATCACAGTTTGCCAG

Table S5: qRT-PCR primers, Related to Figure 2, Figure S1, Figure S5, STAR Methods