

1 **Diversification of Reprogramming Trajectories Revealed by Parallel Single-cell**
2 **Transcriptome and Chromatin Accessibility Sequencing**

3 Qiao Rui Xing^{1,2,19}, Chadi El Farran^{1,3,19}, Pradeep Gautam^{1,3}, Yu Song Chuah¹, Tushar Warrier^{1,3},
4 Cheng-Xu Delon Toh¹, Nam-Young Kang^{4,5}, Shigeki Sugii^{6,7}, Young-Tae Chang^{4,5}, Jian Xu³,
5 James J. Collins^{8,9,10,11}, George Q. Daley^{8,12,13,14,15}, Hu Li^{16,*}, Li-Feng Zhang^{2,*}, Yuin-Han
6 Loh^{1,3,17,18*}

7 ¹ Epigenetics and Cell Fates Laboratory, Programme in Stem Cell, Regenerative Medicine and
8 Aging, A*STAR Institute of Molecular and Cell Biology, Singapore 138673, Singapore

9 ² School of Biological Sciences, Nanyang Technological University, 637551, Singapore

10 ³ Department of Biological Sciences, National University of Singapore, Singapore 117543,
11 Singapore

12 ⁴ Laboratory of Bioimaging Probe Development, Singapore Bioimaging Consortium, A*STAR,
13 Singapore 138667, Singapore

14 ⁵ Department of Chemistry, National University of Singapore, Singapore 117543, Singapore

15 ⁶ Fat Metabolism and Stem Cell Group, Laboratory of Metabolic Medicine, Singapore Bioimaging
16 Consortium, A*STAR, 11 Biopolis Way, 138667 Singapore

17 ⁷ Cardiovascular and Metabolic Disorders Program, Duke-NUS Medical School, 8 College Road,
18 169857 Singapore

19 ⁸ Howard Hughes Medical Institute, Boston, MA 02114, USA

20 ⁹ Institute for Medical Engineering and Science Department of Biological Engineering, and
21 Synthetic Biology Center, Massachusetts Institute of Technology, Cambridge, MA 02114, USA

22 ¹⁰ Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA

23 ¹¹ Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA

24 ¹² Stem Cell Transplantation Program, Division of Pediatric Hematology/Oncology, Boston
25 Children's Hospital and Dana-Farber Cancer Institute, Boston, MA 02115, USA

26 ¹³ Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School,
27 Boston, MA 02115, USA

28 ¹⁴ Harvard Stem Cell Institute, Boston, MA 02115, USA

29 ¹⁵ Manton Center for Orphan Disease Research, Boston, MA 02115, USA

30 ¹⁶ Center for Individualized Medicine, Department of Molecular Pharmacology & Experimental
31 Therapeutics, Mayo Clinic, Rochester, Minnesota 55905, USA

32 ¹⁷ NUS Graduate School for Integrative Sciences and Engineering, National University of
33 Singapore, 28 Medical Drive, Singapore 117456, Singapore

34 ¹⁸ Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore,
35 Singapore 117593, Singapore

36 ¹⁹These authors contributed equally to this work

37

38

39

40

41 *Correspondence:

42 Yuin-Han Loh, Epigenetics and Cell Fates Laboratory, A*STAR, Institute of Molecular and Cell
43 Biology, 61 Biopolis Drive Proteos, Singapore 138673,
44 Singapore; e-mail: yhloh@imcb.a-star.edu.sg

45 Li-Feng Zhang, School of Biological Sciences, Nanyang Technological University, 60 Nanyang
46 Drive, Singapore, 637551, Singapore; e-mail: zhanglf@ntu.edu.sg

47 Hu Li, Center for Individualized Medicine, Department of Molecular Pharmacology &
48 Experimental Therapeutics, Mayo Clinic, 200 First St. SW Rochester, MN 55905, USA; e-mail:
49 Li.Hu@mayo.edu

50 Keywords: human cellular reprogramming, single-cell ATAC-Seq (scATAC-Seq), single-cell
51 RNA-Seq (scRNA-Seq), *GDF3*, BDD2-C8, CD13, CD44, CD201, stage-specific regulatory
52 networks, lineage specifiers, chromatin accessibility, FOSL1, TEAD4.

53 Running title: Parallel Single-cell Transcriptome and Chromatin Sequencing on Human Cellular
54 Reprogramming

55

56 **SUMMARY**

57 To unravel the mechanism of human cellular reprogramming process at single-cell
58 resolution, we performed parallel scRNA-Seq and scATAC-Seq analysis. Our analysis reveals that
59 the cells undergoing reprogramming proceed in an asynchronous trajectory and diversify into
60 heterogeneous sub-populations. BDD2-C8 fluorescent probe staining and negative staining for
61 CD13, CD44 and CD201 markers, could enrich for the *GDF3*+ early reprogrammed cells.
62 Combinatory usage of the surface markers enables the fine segregation of the early-intermediate
63 cells with diverse reprogramming propensities. scATAC-Seq analysis further uncovered the
64 genomic partitions and transcription factors responsible for the regulatory phasing of
65 reprogramming process. Binary choice between a FOSL1 or a TEAD4-centric regulatory network
66 determines the outcome of a successful reprogramming. Altogether, our study illuminates the
67 multitude of diverse routes transversed by individual reprogramming cells and presents an
68 integrative roadmap for identifying the mechanistic part-list of the reprogramming machinery.

69

70 **INTRODUCTION**

71 Somatic cells can be reverted to pluripotency by inducing the expression of four
72 transcription factors namely OCT4, SOX2, KLF4 and MYC in a process known as cellular
73 reprogramming¹⁻⁶. Discovery of this phenomenon has raised the hopes for advancing the field of
74 regenerative medicine⁷. However, cellular reprogramming suffers from extremely low efficiency
75 especially for the human cells, resulting in a heterogeneous population in which few cells can be
76 characterized as pluripotent⁸⁻¹⁰. Although a handful of studies analyzed bulk population to

77 understand the reprogramming mechanisms¹¹⁻¹⁷, ensemble measurement of the heterogeneous
78 population impedes the discerning of transcriptomic and epigenetic changes taking place in the
79 minority of cells undergoing the route towards successful reprogramming. Single-cell sequencing
80 technologies provide tools to decipher the types of cells present in a heterogeneous mixture¹⁸. In the
81 present study, we adopted the parallel genome-wide single-cell assays of scRNA-Seq and scATAC-
82 Seq¹⁹⁻²¹ to profile transcriptome and chromatin accessibility of human reprogramming cells across
83 various stages. We identified cellular diversification and trajectories where individual cell displays
84 different dynamics and potential for reprogramming. Moreover, with a set of cell surface markers
85 and a fluorescent probe BDD2-C8, we were able to enrich for the early-intermediary cells
86 undergoing route towards successful reprogramming. In addition, we identified the modulators
87 driving the changes in gene regulatory network and chromatin accessibility, as cells advanced
88 towards the diverse reprogramming trajectories. Of note, the pivot from FOSL1 to TEAD4-centric
89 regulatory networks is essential for the acquisition of the pluripotent state.

90

91 **RESULTS**

92 **Single-cell Profiling of Cell-Fate Reprogramming**

93 To study the heterogeneity of human reprogramming, we analyzed a total of 33468 scRNA-
94 Seq and scATAC-Seq libraries with good quality, including day 0 (BJ), day 2 (D2), day 8 (D8),
95 day12 (D12) and day 16 (D16) OSKM-induced reprogramming cells (Figure 1a and Supplementary
96 Table 1). On D16, cells were sorted using the TRA-1-60 marker to distinguish the successfully
97 reprogrammed (D16+) from the non-reprogrammed (D16-) cells (Figures 1a and S1a-b). The
98 generated iPSCs were characterized with immunostaining, DNA methylation, and teratoma assay
99 (Figures S1c-e). Two distinct approaches were used for scRNA-Seq library preparation.
100 Microfluidic cell capture-based assay (Fluidigm C1) reads the full-length transcripts from hundreds
101 of cells with high resolution, whereas the droplet-based assay (10X Genomics) probes the 3' end of

102 the transcripts from thousands of cells albeit at a relatively low genomic coverage. In addition, we
103 have also screened for fluorescence probes to distinguish early-intermediate cells poised for
104 successful reprogramming (Figure 1a). The cumulative data enable us to characterize the sub-
105 populations in-depth and to construct a trajectory map of human reprogramming.

106 Majority of the capture-based scRNA-Seq libraries showed high exon mapping percentage
107 ($>=75\%$) and gene detection rate ($>=20\%$) and displayed even distribution over the gene-bodies
108 (Figure 1b-c, and Supplementary Table 1). Furthermore, epithelial and pluripotency genes were
109 progressively expressed with the advancement of reprogramming, as opposed to the mesenchymal
110 and fibroblast genes (Figure S1f). Similarly, majority of the 10X libraries were of good quality
111 (Figures S1g-h). Notably, t-SNE plot revealed a dynamic transcriptomic transition from the parental
112 BJ to D16+ cells (Figure 1d). Expectedly, *ZEB1* (mesenchymal) and *COL1A1* (somatic) were
113 abundantly expressed in the early time-points and non-reprogrammed cells (Figures 1e-f). On the
114 contrary, *EPCAM* (epithelial), and *NANOG* and *LIN28A* (pluripotent) were expressed highly in the
115 successfully reprogrammed cells (Figures 1e-f and S1i). Likewise, most of the scATAC-Seq
116 libraries passed the previously reported QC indices¹⁹, and exhibited enrichment over TSS regions
117 and nucleosomal distributions (Figure 1g-i and S1j-k, and Supplementary Table 1). Collectively, we
118 generated reliable libraries for tens of thousands of reprogramming cells, providing a rich resource
119 to decipher its deep molecular mechanisms.

120 **Deciphering the Heterogeneous Subgroups with Diverse Reprogramming Potentials**

121 Due to its high sensitivity, capture-based scRNA-Seq libraries were analyzed first to
122 decipher the heterogeneity. CellNet²² revealed the dynamics of reduced fibroblast similarity and
123 increased ESC correlation (Figure S2a). To determine the diverse populations present at each
124 reprogramming time-point, we clustered scRNA-Seq libraries using Reference Component Analysis
125 (RCA)²³. Interestingly, BJ cells correlated significantly with the smooth muscle lineage, which was
126 also detected across the published BJ libraries (Figures S2b-c). D2 cells were marked by four

127 distinct subgroups (G1-G4), among which G1-3 cells displayed lower correlation to the fibroblasts
128 and mesenchymal stem cells (MSCs) (Figures 2a and S2d). D8 cells were distributed among three
129 discrete sub-populations (Figures 2b and S2d). D8 G1 cells corresponded to the fibroblasts, smooth
130 muscles, myocytes, and MSCs lineages, whilst G3 cells displayed substantial similarity to the
131 pluripotent stem cells (PSCs). Interestingly, D8 G2 cells represented the intermediate state. Two
132 sub-populations were present in the D16+ cells (Figures 2c and S2d). D16+ G2 cells were highly
133 associated with PSCs, while G1 cells maintained detectable correlation to the MSCs, adipose cells,
134 and endothelial cells, other than PSCs. In summary, RCA analysis strongly indicates that the
135 reprogramming cells are highly diverse, some of which may deviate from the route to pluripotency
136 and acquire alternative lineage cell-fates.

137 We then performed differential gene expression (DGE) analysis to evaluate the subgroup
138 specific genes (Figure 2d and Supplementary Table 2). Among D2 subgroups, G3 cells had the
139 most distinct transcriptomic profile with exclusive expression of a remarkable number of genes,
140 including *ERBB3* (Figures 2d-e). Majority of D2 G1-2 genes, for example *CDK1*, were expressed
141 highly in D16+ G2, suggestive of their higher reprogramming propensity (Figures 2d-f). Among D8
142 subgroups, G1-2 specific genes were expressed highly in BJ and D16- but not the D16+ cells, such
143 as *JUNB*, *LUM*, *COLIA1*, and *COL6A3* (Figure 2d). The opposite trend was observed for D8 G2-3
144 genes. *RFC3* was vastly expressed in D8 G2-3 cells, whereas *GDF3* was specifically expressed in
145 the G3 (Figures 2d-f). In agreement with the correlation to the differentiated lineages, D16+ G1
146 specific genes, including *MMP2*, were also expressed highly in the D16- cells, suggesting that
147 D16+ G1 cells were at most partially reprogrammed (Figures 2d-f). On the other hand, epithelial
148 genes and pluripotent genes including *CDH1*, *NANOG* and *LIN28A*, as well as genes associated
149 with mRNA splicing and transcription of the small RNAs including *WBP11* and *POLR3K*, were
150 specifically expressed in D16+ G2 cells. Intriguingly, depletion of *POLR3K* and *WBP11* resulted in
151 ablated reprogramming efficiency, indicating the functional importance of the D16+ G2 specific
152 genes for reprogramming (Figure S2e). Additionally, D8 G3 and D16+ G2 specific genes exhibited

153 high stemness score for PSCs, whereas D8 G1-2 and D16+ G1 specific genes significantly
154 associated with the differentiated lineages (Figures S2f-g).

155 To test the notion of the diverse reprogramming potentials, D8 subgroups were correlated to
156 the subgroups of various time-points (Figure S2h). Interestingly, G3 cells highly correlated with
157 D16+ G2 cells, whereas G2 cells represented an intermediate state in which the cells were
158 moderately correlated with all D16 cells. G1 cells, on the other hand, strongly correlated with D16-
159 and cells from the early time-points (BJ and D2). In-house devised classifier displayed a similar
160 correlation trend of the reprogramming time-points to the D8 subgroups (Figure 2g).

161 **Pseudotemporal trajectory of reprogramming cells**

162 We next analyzed 10X libraries to construct pseudotemporal map of cellular reprogramming.
163 CellNet and RCA of 10X libraries reproduced the reprogramming dynamics and diverse subgroups
164 with variable lineage correlations (Figures S3a-c). Resultant pseudotemporal trajectories^{24,25}
165 consisted of 9 states and 4 branching events (Figures 3a and S3d). Notably, pseudotime highly
166 correlated with the reprogramming time-points (Figures 3a-c). We then traced the trajectory of RCA
167 subgroups. Interestingly, majority of the D2 G1-3 cells were found in state 3 (95% of G1, 65% of
168 G2, and 80% of G3), whereas G4 cells scattered across the early states (Figure S3e). Joint RCA
169 analysis demonstrated that D2 G1-2 cells clustered closer to the D8 G2 cells and correlated stronger
170 to the ESC fate, indicating their higher reprogramming propensity (Figures S3f-g). Of note, D8 G1
171 cells enriched across the different states other than state 9 (successfully reprogrammed) (Figures 3d-
172 e). On the contrary, D8 G3 cells were mostly found in state 9. D8 G2 cells distributed across the
173 intermediate (3-5) and late states (7-9). Intriguingly, state 4 comprised almost entirely of D8 G2
174 cells (544 vs. 627) (Figures 3d-e). Expectedly, D16+ G2 cells were the major constituents of state 9,
175 whereas cells of D16+ G1 were enriched in both state 8 (non-reprogrammed) and state 9,
176 corroborating their partially- and non-reprogrammed identities (Figures 3d and S3h). Further,
177 subgroup specific markers were expressed differentially along the pseudotime axis (Figure 3f).

178 *RFC3* (D8 G2-3) and *GDF3* (D8 G3), and *NANOG* and *LIN28A* (D16+ G2) were expressed highly
179 in the cells on the successful reprogramming trajectory, whereas *MMP2* (D16+ G1) showed the
180 opposite trend. Next, we determined the gene expression patterns and significant biological
181 processes defining the trajectory (Figure S3i and Supplementary Table 3). At branching event 1,
182 cells with high levels of lineage genes advanced towards successful reprogramming, whereas cells
183 with abundant DNA replication genes deviated from the path (Figures 3g-h and S3i). Interestingly,
184 successfully reprogrammed cells at branching point 4 were highly active for DNA replication and
185 mRNA splicing, which were implicated to be essential for reprogramming^{26,27}. On the other hand,
186 collagen and extracellular matrix related genes emerged to be detrimental for reprogramming at the
187 branching point 3 and 4, and lineage processes such as angiogenesis and epidermis development
188 contributed to the unsuccessful reprogramming at branching point 4. Together, these analyses
189 provide a plethora of data for identifying the novel processes and modulators affecting the
190 reprogramming trajectory at an unprecedented single-cell resolution.

191 **Toolkits to enrich for early-intermediate cells with diverse reprogramming potentials**

192 To enrich for the intermediate cells with high reprogramming potential, we conducted a
193 screen for a library of 34 DOLFA^{28,29} fluorescent dyes (Figure S4a). Candidate dyes differentially
194 stain the intermediate reprogramming cells with the accelerated dynamics upon treatment with a
195 TGF β inhibitor^{30,31}, A83-01. BDD1-A2, BDD2-A6, and BDD2-C8 were identified as the top hits,
196 which showed co-localized signal with TRA-1-60 (Figures S4a-c). Importantly, D8 cells enriched
197 with the candidate probes gave rise to a significantly higher number of TRA-1-60+ colonies
198 (Figures 4a and S4d). Among them, BDD2-C8 displayed the best performance in an alternative
199 reprogramming of MRC5 fibroblasts (Figure S4e). BDD2-C8 also precisely captured changes in
200 reprogramming efficiency upon depletion of the key modulators¹⁵ (Figure S4f). Single-cell qPCR
201 showed that BDD2-C8+ cells expressed higher levels of epithelial and pluripotent genes, and lower
202 levels of mesenchymal and somatic genes (Figure 4b). To further characterize, we prepared 192

203 scRNA-Seq libraries for D8 cells stained highly ($D8^{BDD2-C8^+}$) and lowly ($D8^{BDD2-C8^-}$) for BDD2-C8.
204 In the ensuing RCA analysis, $D8^{BDD2-C8^+}$ and $D8^{BDD2-C8^-}$ cells clustered close to the D8 G2-3 and G1
205 respectively, which were substantiated by the similar expression profiles for the subgroup specific
206 genes (Figures 4c-d and S4g-h). GO enriched for $D8^{BDD2-C8^+}$ specific genes were related to cell
207 cycle, embryo development and stem cell population maintenance, whereas $D8^{BDD2-C8^-}$ genes were
208 predominantly represented by epithelial to mesenchymal transition, extracellular matrix
209 organization, and development processes (Figure 4e and Supplementary Table 4). In terms of
210 structural complexes, $D8^{BDD2-C8^-}$ genes were specifically enriched in the endoplasmic reticulum (ER)
211 lumen and Golgi (Figure S4i). Interestingly, BDD2-C8 were localized in the ER and Golgi (Figure
212 S4j). In addition, higher expression of secretory genes in the $D8^{BDD2-C8^-}$ cells implicated its active
213 ER-Golgi secretion pathway (Figure S4k). Depletion of these genes indeed resulted in the retention
214 of BDD2-C8 (Figures S4l-m). This indicates that BDD2-C8 may be actively effluxed from BJ and
215 the non-reprogrammed cells (D8 G1) but retained in the pluripotent cells and intermediate cells with
216 high reprogramming potential, due to the differential ER-Golgi secretion activities.

217 We next identified surface markers with differential expression among D8 subgroups to
218 enrich for the intermediate cells with varying reprogramming potentials (Supplementary Table 4).
219 Shortlisted surface markers displayed higher expression in D8 G1/G2 and D16- cells than D8 G3
220 and D16+ cells respectively (Figures 4f-g). Majority of them were abundantly enriched in the
221 parental BJ cells, except for *CD201*. Expression dynamics of the surface markers were validated by
222 time-course FACS analysis (Figure 4h). Further, D8 cells stained negatively for the surface markers
223 exhibited lower expression of mesenchymal markers but higher epithelial and pluripotency genes,
224 including the D8 G3 marker *GDF3* (Figure 4i). Noteworthy, negatively stained D8 populations
225 gave rise to more TRA-1-60+ colonies, indicating the capacity of surface markers to isolate early
226 reprogrammed cells with high stemness feature (Figure 4j). We then examined the similarity of
227 cells sorted by BDD2-C8 and the identified surface markers. Indeed, $D8^{BDD2-C8^+}$ cells demonstrated
228 lower level of CD13, CD44, and CD201 (Figure 4k-l). Of note, difference in the CD201 protein

229 amount was subtler, which could be due to its inconsistent expression across the D8 subgroups (G1-
230 2 like) with BDD2-C8 (G2-3 like). Of note, co-staining of CD13 and CD44 markers showed
231 extensive overlaps across the time-points (Figures 4m and S4n). Surface markers were verified in
232 an alternative reprogramming of MSCs induced by Sendai viruses (Figure S4o-p). D8 sorted MSC
233 reprogramming cells demonstrated the similar expression trends as BJ reprogramming (Figures 4n
234 and S4q). Importantly, CD13- MSCs resulted in a remarkably higher reprogramming efficiency
235 with or without the aid of TGF-beta inhibition (Figure 4o). Collectively, the fluorescent dye and
236 surface markers established the technological platforms to enrich for the minority of intermediate
237 cells primed for successful reprogramming.

238 **Refined classification of the intermediate population**

239 To decipher the intermediate reprogramming cells, we prepared 10X libraries on D8 cells
240 sorted with CD13. Majority of the libraries passed the QC thresholds (Figure S5a). Expectedly,
241 libraries showed 2 distinct groups with differential *CD13* expression (Figures 5a and S5b). Notably,
242 majority of the genes that were expressed highly in CD13+ cells were correspondingly expressed in
243 the D8^{BDD2-C8-} cells and vice versa (Figure S5c). The libraries demonstrated 8 clusters. Clusters 5-8
244 were mainly comprised of CD13+ cells (Figures 5b and S5d). Consistently, CellNet analysis
245 showed that clusters 1-4 correlated with ESCs, whereas the other clusters to the fibroblast state
246 (Figure S5e). RCA corroborated the association of CD13+ cells to the MSCs and Fibroblast
247 lineages (resembling G1) (Figure 5c). Interestingly, among the CD13- cells, cluster 1 showed the
248 highest correlation score to the PSCs, indicating similarity to the D8 G3 cells (Figure 5c). Whereas,
249 cells of cluster 3 and 4 showed resemblance to both the MSCs and ESCs, albeit at a lower
250 significance than cluster 1 in terms of the ESCs (Figure 5c). On the other hand, cells of cluster 2
251 and 7 exhibited a transitional intermediate profile (Figure 5c). We next used MAGIC³² for pairwise
252 comparison between *CD13* and *GDF3*. Interestingly, cells with lower *CD13* levels exhibited higher
253 expression of *GDF3* and *NANOG*, among which cluster 1 and 4 exhibited the highest *GDF3*

254 expression (Figures 5d-e and S5f). Next, to trace the trajectory of the CD13 sub-clusters, we
255 aggregated the CD13 sorted 10X libraries with the other timepoints and performed pseudotemporal
256 analysis. Remarkably, D8 CD13+ and CD13- cells were enriched at the distinct trajectory states
257 (Figure 5f). Cells of cluster 1 and 4 concentrated in the branch shared by D16+ cells. Notably, cells
258 of cluster 3 were found at the unsuccessful reprogramming branch, whereas cluster 2 cells were
259 scattered.

260 DGE analysis demonstrated that cluster 7 had unique expression trends among the CD13+
261 clusters, which associated with epidermis development, and I-KB/NF-KB signaling (Figures 5g-h).
262 On the other hand, cluster 5, 6 and 8 specific genes related to extracellular matrix organization and
263 collagen catabolic process (Figure 5h). cluster 1 and 4 specific genes related to cell division,
264 heterochromatin assembly, embryonic development, organ development of multiple lineages, and
265 MAPK cascade, suggesting that reprogramming cells of cluster 1 and 4 might adopt an open
266 chromatin structure, especially at genes crucial for multi-lineage development and differentiation.
267 Mathematical imputation showed an extensive correlation between *CD13* and *CD44* (Figures S5g-
268 h). Interestingly, we detected a group of cells which were low in *CD13* but high in *CD201*,
269 representing the D8 G2-like intermediate cells belonging mostly to cluster 2 and 7 (Figures 5i, and
270 S5i). To substantiate their existence, we performed time-course co-staining using CD13 and CD201
271 antibodies. A significant number of intermediate cells were CD13-CD201+, and their proportion
272 increased as reprogramming advanced (Figures 5j and S5j). Furthermore, CD13-CD201+ cells
273 corresponded to higher BDD2-C8 staining signals than CD13+CD201+ cells (Figure S5k). Hence,
274 we hypothesized that dual sorting with CD13 and CD201 markers would allow us to enrich for
275 successfully reprogrammed early-intermediary cells with higher purity.

276 We thus categorized D8 cells into double negative (CD13-CD201-), double positive
277 (CD13+CD201+) and intermediate (CD13-CD201+) cells, which were then subjected to gene
278 measurement (Supplementary Table 5). CD13+CD201+ cells expressed fibroblast and

279 mesenchymal genes and genes associated with extracellular matrix and cell adhesion (Figures 5k-l).
280 On the contrary, CD13-CD201- cells expressed genes related to pluripotency, epithelial lineage, cell
281 division, neuronal differentiation, and stem cell population maintenance. CD13-CD201+ cells
282 exhibited an intermediate transcription profile (Figures 5k-l). In addition, genes highly expressed in
283 CD13+CD201+ cells were mostly enriched in the D16- but not D16+ cells, whereas the opposite
284 was observed for CD13-CD201- specific genes (Figure S5l). Notably, depletion of genes highly
285 expressed in the CD13+CD201+ population resulted in more reprogrammed colonies (Figure 5m).
286 Importantly, enrichment of CD13-CD201- population gave rise to the highest reprogramming
287 efficiency, in comparison with the CD13-CD201+ (intermediate) and CD13+CD201+ cells (lowest)
288 (Figure 5n). We validated the presence of these distinct D8 populations in an alternative
289 reprogramming of BJ cells using Sendai viruses (Figures S5m-o). Altogether, concurrent use of
290 CD13 and CD201 antibodies enable us to dissect the precise populations differentially poised for
291 successful reprogramming.

292 **Regulatory networks of Transcription Factors in cellular reprogramming**

293 Transcription Factors (TFs) define the cell-selective regulatory network underlying the
294 cellular identity and function³³. However, the stage-specific core regulatory networks of the human
295 cellular reprogramming remain elusive. To this end, we performed DGE analysis for TFs across the
296 pseudotemporal states, which were then categorized as Early Silenced, Late Silenced, Transient,
297 Early Expressed and Late Expressed (Figures 6a and S6a, and Supplementary Table 6). Notably,
298 many TFs exhibited similar expression trends in the D8 RCA subgroups and the D8 sorted libraries.
299 Specifically, Early Silenced TFs (e.g. *FOSL1*, *CREB3L1*, *AHRR*, *DRAP1* and *ELL2*) showed higher
300 expression in the D8 BDD2-C8- and CD13+CD201+ populations, whereas Late Silenced TFs
301 exhibited the opposite trends (Figures 6a and S6b). Majority of Transient TFs exhibited higher
302 expression in CD13+CD201+ and CD13-CD201+ populations, including lineage associated factors
303 namely *HAND1* (mesoderm), *ASXL3* and *NEUROG2* (neuroectoderm) (Figures 6a and S6c).

304 Contrastingly, Early Expressed TFs adopted higher expression in the cells of CD13-CD201- and
305 CD13-CD201+ (Figures 6a and S6d). CD13-CD201- cells expressed the highest level of Late
306 Expressed TFs, such as *PRDM14*, *DNMT3B* and *LHX6*.

307 To investigate how the regulatory TFs accessed its genomic targets, we then analyzed time-
308 course scATAC-Seq libraries of reprogramming cells (Figure 1a). Both batches of libraries showed
309 similar promoter accessibility profiles (Figure S6e). We also observed diminished accessibility on
310 the fibroblast-specific DHS and increased accessibility on the pluripotency-specific DHS as
311 reprogramming progressed (Figure S6f). Next, we utilized chromVAR³⁴ to identify the TFs
312 determining the variable epigenomes accessibility. Correlation of scATAC-Seq libraries among
313 themselves resulted in three major clusters. The early cells consisting mostly of BJ and D2 cells
314 clustered together (Cluster II), while D8, D16+ and H1 cells shared similar accessibility profile
315 (Cluster I). A third cluster composed mainly of D8 and D16- cells (Cluster III) (Figure 6b). of note,
316 FOSL1 and its partners, CEBPA, ZEB1, PAX6, SOX8, SOX10, POU5F1 and TEAD4 were found
317 to be the TFs contributing to the reprogramming heterogeneity (Figure 6c and Supplementary Table
318 6). TFs were then categorized to open starting from the intermediate stage (Early open) or the late
319 stage (Late open), open transiently at the intermediate stage (Transient open), and close starting
320 from the intermediate stages of reprogramming (Close) which were either homogenously or
321 heterogeneously open in the early cells (Figure 6d and 6e). Particularly, Close TFs belonged mostly
322 to the FOS-JUN-AP1 complex, such as FOSL1 and JDP2 (Figures 6f and S6g). Intriguingly, motifs
323 of lineage specifiers (Mesendoderm: GATA1, SOX8 and HNF4G; Ectoderm: PAX6, NRL and
324 RXR) were found to be accessible only in the intermediate reprogramming cells, but not the H1
325 hESCs (Figures 6d, 6g and S6h). This observation corroborated with the model of counteracting
326 lineage specification networks underlying the induction of pluripotency^{35,36}. On the contrary, Early
327 Open TFs (e.g. TEAD4 and POU5F1) exhibited accessibility starting from the D8 reprogramming
328 cells, whilst Late Open TFs (e.g. FOXL1, TCF4, and YY2) demonstrated accessibility in D16+
329 cells and ESCs exclusively (Figures 6d, 6h-i, S6i-k). Of note, TFs of FOX family and YY1 were

330 previously reported to be important for reprogramming³⁷⁻³⁹. Independently, we used SCENIC
331 analysis to infer the interaction between TFs and their target genes. Corroborating with our earlier
332 results, Silenced and Close TFs showed decrease in their regulon activity as reprogramming
333 progressed, whereas majority of Expressed and Open TFs displayed the opposite dynamics (Figure
334 S6l). Likewise, regulon activity of the Transient TFs was observed only in the intermediate cells.
335 Together, these data represent the compendium of TFs which regulate the networks of downstream
336 key modulators in the cellular reprogramming process.

337 **Identification of key regulators for the intermediate stage of reprogramming**

338 In order to deduce TFs essential for the intermediate cells to acquire pluripotency, we
339 further sub-clustered the D8 scATAC-Seq libraries (Figures S7a-b). Intriguingly, the most variable
340 motifs of D8 cells belonged to the FOS-JUN-AP1 and TEAD families (Figure 7a and
341 Supplementary Table 7). Strikingly, D8 cells were either accessible for *FOSL1*-JUN-AP1 or
342 TEAD4 motif (Figures 7b-c). Of note, *FOSL1* and *TEAD4* displayed a contrasting expression
343 pattern and regulon activity during reprogramming (Figures 7d-g). In D8, *FOSL1* and *TEAD4* were
344 exclusively expressed in the CD13+ and CD13- cells (Figure 7h).

345 Given its expression and motif accessibility at the early stage of reprogramming, *FOSL1*
346 was hypothesized to act as a roadblock. To verify, reprogramming efficiency was assessed upon the
347 depletion of *FOSL1* by siRNA, which showed no effect on cell proliferation and lasted for about 4
348 days (Figures S7c-e). Remarkably, depletion of *FOSL1* at day 1, 2, 3, or 5 post-OSKM induction
349 resulted in an increased reprogramming efficiency (Figure 7i). Specificity of the *FOSL1* siRNA
350 construct was validated using a mutant construct which showed no effect (Figures S7f-g). Depletion
351 of *FOSL1* in an alternative Sendai virus reprogramming reproduced the phenotypic change (Figures
352 S7h-i). Conversely, ectopic expression induced a drastic reduction in the number of reprogrammed
353 colonies (Figure 7j). Noteworthy, high reprogramming propensity of CD13- cells was negated by
354 the elevated *FOSL1* level (Figure 7k).

355 Contrastingly, we posit that TEAD4 could serve as an effector. Indeed, depletion of *TEAD4*
356 resulted in the reduced number of reprogrammed colonies (Figure 7l). Specificity of si*TEAD4* was
357 affirmed by the mutant construct with no phenotypic change (Figures S7j-k). Knock-down effect of
358 si*TEAD4* lasted for around 5 days (Figure S7l). Interestingly, when knock-down was performed on
359 D5 cells, *TEAD4* expression did not restore, which could be due to the perpetuations of the non-
360 reprogrammed state introduced by si*TEAD4* (Figure S7m). Notably, depletion of *TEAD4* only from
361 day 5 onwards induced the decreased reprogramming efficiency (Figure 7l). This established the
362 vital role of TEAD4 at the intermediate-late stages of reprogramming. In contrast, overexpression
363 of *TEAD4* resulted in a significant increase in the number of reprogrammed colonies (Figure 7m).
364 Importantly, depletion of *TEAD4* revoked the reprogramming potential of D8 CD13- cells (Figure
365 7n). Collectively, these analyses illustrate that the pivot from FOSL1 to TEAD4-directed regulatory
366 network is essential for successful reprogramming.

367 **Mechanistic roles of FOSL1 and TEAD4 in cellular reprogramming**

368 To find the direct targets of FOSL1, we prepared ChIP-Seq libraries for D8 cells, which
369 demonstrated the expected genomic distribution profile and enriched with FOSL1 motif and its
370 regulatory partners (Figures S8a-c). We also investigated the genomic binding profile of TEAD4 in
371 D8 cells and CD13 sorted D8 cells (Figures S8d-e). Majority of the D8 TEAD4 bound sites were
372 shared by both CD13+ and CD13- cells (Figures S8f). 1.7-fold higher numbers of TEAD4 bound
373 loci were detected in the CD13- than CD13+ cells (18550 vs. 10986 sites). In addition, CD13-
374 specific sites showed greater enrichment of TEAD4 motif and exclusive enrichment of pluripotency
375 associated OCT4-SOX2-TCF-NANOG motif, indicating the differential regulatory role of TEAD4
376 (Figures S8g).

377 Consistent with their contrasting roles, majority of loci highly enriched with FOSL1 were
378 distinctive from that with TEAD4 (Figure 8a). We next clustered BJ, D16- and D16+ scATAC-Seq
379 libraries, based on the accessibility of FOSL1 and TEAD4 bound sites (Figure 8b). Notably, FOSL1

380 specific sites exhibited higher accessibility in BJ and D16- cells, whereas CD13- specific and CD13
381 common TEAD4 bound sites were mostly accessible in D16+ cells (Figures 8b-c and S8h-i). these
382 differential accessible sites were annotated as the functional FOSL1 and TEAD4 targets
383 (Supplementary Table 8). Similar to the motif enrichment pattern, most of the D8 cells were either
384 accessible for FOSL1 (“FOSL1 ChIP only” cells) or CD13- TEAD4 bound sites (“TEAD4 ChIP
385 only” cells) (Figures 8d-e and S8k-l). We next asked if their binding has any consequence to the
386 transcription of target genes. To this end, we analyzed the expression of functional FOSL1 and
387 TEAD4 targets across the D8 CD13-CD201 sorted cells. Remarkably, majority of the D8 FOSL1
388 targets (e.g. *TGFBR2*, *SMAD3*, and *COL5/7/21A1*) were expressed highly in the CD13+CD201+
389 cells (Figure 8f). In contrast, D8 CD13-CD201- cells demonstrated high expression of the TEAD4
390 targets, including key pluripotent genes such as *DNMT3B*, *LIN28A*, *SOX2*, and *PODXL* which
391 codes for TRA-1-60 (Figure 8g). This was further substantiated at single-cell resolution using the
392 coupled NMF analysis of D8 scRNA-Seq and scATAC-Seq libraries, by which regulatory
393 connectivity between the accessible chromatin regions and expression of target genes were
394 established (Figure S8m). Notably, cluster 1 composed of “TEAD4 ChIP only” cells (scATAC-Seq)
395 and D8 RCA G3 cells (scRNA-Seq), whilst cluster 2 comprised of “FOSL1 ChIP only” cells
396 (scATAC-Seq) and D8 RCA G1 (scRNA-Seq) cells. Differential analysis demonstrated that
397 TEAD4 and its targets, such as *TET1* and *CDH1*, were highly accessible and expressed in the
398 cluster 2 cells, whereas FOSL1 and its targets, such as *MMP2* and *SMAD3*, displayed the opposite
399 trends (Figures S8m-n). Besides, interactome analysis revealed that cluster 1 genes associated with
400 splicing process, whereas cluster 2 genes related to ER-Golgi transport and extracellular matrix
401 organization, including FOSL1 targets *MMP2* and several collagen genes (Figure S8o). More
402 importantly, downstream targets of FOSL1 and TEAD4 were themselves modulators of
403 reprogramming. For instance, knock-down of FOSL1 bound genes, *MMP2* and *SMAD3*, resulted in
404 higher numbers of the reprogrammed colonies (Figure 8h). Conversely, depletion of TEAD4 targets,
405 *PRDM14* and *SOX2*, showed the opposite effect (Figure 8i).

406 Taken together, we present the single-cell roadmap of the human cellular reprogramming
407 process, which reveals the diverse cell-fate trajectory of individual reprogramming cells (Figure 8j).
408 D2 cells consist of four subgroups, out of which *CDK1*+ cells showed greater propensity for
409 reprogramming. Among the three subgroups of D8 cells, G1 represents the unsuccessful
410 reprogramming cells, whereas G3 cells with *GDF3* expression are highly primed for successful
411 reprogramming. G2 is the intermediary group between G1 and G3. In-house developed fluorescent
412 probe (BDD2-C8) and the identified surface markers (CD13, CD44 and CD201), enables the
413 segregation of the heterogeneous population based on their reprogramming potential. Moreover,
414 TFs analysis reveals the stage-specific regulatory networks of reprogramming. Importantly, we
415 describe the crucial switch from a *FOSL1* to a *TEAD4*-centric expression which collectively
416 regulate genomic accessibility, cell-lineage transcription program, and network of functional
417 downstream modulators favoring the acquisition of the pluripotent state.

418

419 **DISCUSSION**

420 **Heterogeneity of Cellular Reprogramming**

421 Due to the inherent heterogeneity, Single-cell NGS techniques are more suited to
422 characterize the molecular and epigenetics signatures of reprogramming cells with diverse
423 trajectories. Several studies described the mouse reprogramming process at single-cell resolution^{40–}
424 ⁴². However, there have been reports for extensive differences between the mouse and human
425 reprogramming in terms of kinetics, modulators involved, and the molecular nature of the generated
426 iPSCs. In addition, this is the first study presenting a comprehensive human reprogramming
427 roadmap by integrating the transcriptomic changes and the alteration of accessible epigenetic
428 regions at single-cell resolution.

429 Our analyses also identify an early reprogramming marker *GDF3*, which marks the
430 intermediate cells primed for successful reprogramming. *GDF3* was previously shown to be
431 expressed in pluripotent stem cells and played a role in regulating cell fate via BMP signaling
432 inhibition⁴³. We also develop a fluorescent probe, BDD2-C8, and identify a panel of cell surface
433 markers, namely CD13, CD44 and CD201, to distinguish and characterize the diverse sub-
434 populations of the intermediate reprogramming cells. Interestingly, CD44 sorting was previously
435 reported as a mean to isolate reprogrammed cells in the mouse system⁴⁴. Noteworthy, combinatorial
436 use of the surface markers enables a more refined segregation. The toolkits to decipher the
437 intermediate cells with different stemness capacity, will help deepen our understanding of the
438 mechanisms of reprogramming process. Additionally, the ability to enrich for early reprogramming
439 cells will help increase the success rate of iPSC generation from the cell lines or patient-derived
440 primary cells which are refractory to reprogramming.

441 **Transcription Factors Contribute to the Heterogeneity of Cellular Reprogramming**

442 Facilitated by the integrative analysis of transcriptomic and chromatin accessibility profiles,
443 TFs contributing to the heterogeneous reprogramming trajectories are identified. Accessible regions
444 with the motifs of FOS-JUN-AP1 and CEBPA are rapidly closed upon induction, which were
445 reported to act as repressors in mouse reprogramming^{45,46} (Figure 6). An earlier report showed that
446 FOSL1 lost many of its binding as early as day 2 of mouse reprogramming⁴⁶. Our study reveals that,
447 in the human system, FOSL1 regulates myriad of genes which serve as barriers of reprogramming,
448 including *MMP2*, *SMAD3*, *TGFB2*, and collagen genes. Strikingly, chromatin regions with the
449 motifs of lineage TFs are open in the intermediate cells, which might be due to the induction of ME
450 and ECT lineages driven by POU5F1 and SOX2⁴⁷. This is further supported by the replacement of
451 POU5F1 and SOX2 with the ME and ECT lineage specifiers, for both mouse and human
452 reprogramming^{35,36}. We unravel for the first time the transitory epigenetic accessibility directed by

453 the lineage TFs, which contribute extensively to the diverse cell fate potentials observed during
454 cellular reprogramming.

455

456 **REFERENCE**

- 457 1. Loh, Y.-H. *et al.* Reprogramming of T Cells from Human Peripheral Blood. *Cell Stem Cell* **7**,
458 15–19 (2010).
- 459 2. Park, I. H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
- 460 3. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by
461 Defined Factors. *Cell* **131**, 861–872 (2007).
- 462 4. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic
463 and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
- 464 5. Tan, H.-K. *et al.* Human Finger-Prick Induced Pluripotent Stem Cells Facilitate the
465 Development of Stem Cell Banking. *Stem Cells Transl. Med.* **3**, 586–598 (2014).
- 466 6. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science*
467 **318**, 1917–20 (2007).
- 468 7. Seah, Y. F. S., El Farran, C. A., Warrier, T., Xu, J. & Loh, Y. H. Induced pluripotency and
469 gene editing in disease modelling: Perspectives and challenges. *Int. J. Mol. Sci.* **16**, 28614–
470 28634 (2015).
- 471 8. Cheow, L. F. *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity.
472 *Nat. Methods* **13**, 833–836 (2016).
- 473 9. Polo, J. M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell*
474 **151**, 1617–1632 (2012).
- 475 10. Stadtfeld, M. & Hochedlinger, K. Induced pluripotency: History, mechanisms, and
476 applications. *Genes and Development* **24**, 2239–2263 (2010).
- 477 11. Maherali, N. *et al.* A High-Efficiency System for the Generation and Study of Human
478 Induced Pluripotent Stem Cells. *Cell Stem Cell* (2008). doi:10.1016/j.stem.2008.08.003
- 479 12. Merkl, C. *et al.* Efficient Generation of Rat Induced Pluripotent Stem Cells Using a Non-
480 Viral Inducible Vector. *PLoS One* **8**, (2013).
- 481 13. Borkent, M. *et al.* A Serial shRNA Screen for Roadblocks to Reprogramming Identifies the
482 Protein Modifier SUMO2. *Stem Cell Reports* **6**, 704–716 (2016).
- 483 14. Qin, H. *et al.* Systematic identification of barriers to human Ipse generation. *Cell* **158**, 449–
484 461 (2014).

486 15. Toh, C. X. D. *et al.* RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell
487 Reprogramming. *Cell Rep.* **15**, 2597–2607 (2016).

488 16. Yang, C. S., Chang, K. Y. & Rana, T. M. Genome-wide Functional Analysis Reveals Factors
489 Needed at the Transition Steps of Induced Reprogramming. *Cell Rep.* **8**, 327–337 (2014).

490 17. Fang, H. T. *et al.* Global H3.3 dynamic deposition defines its bimodal role in cell fate
491 transition. *Nat. Commun.* **9**, (2018).

492 18. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the
493 science. *Nature Reviews Genetics* **17**, 175–188 (2016).

494 19. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory
495 variation. *Nature* **523**, 486–490 (2015).

496 20. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly
497 multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).

498 21. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual
499 circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

500 22. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* **158**, 903–
501 915 (2014).

502 23. Li, H. *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular
503 heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).

504 24. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by
505 pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

506 25. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat.*
507 *Methods* **14**, 979–982 (2017).

508 26. Tsubouchi, T. *et al.* DNA synthesis is required for reprogramming mediated by stem cell
509 fusion. *Cell* **152**, 873–883 (2013).

510 27. Lu, Y. *et al.* Alternative splicing of MBD2 supports self-renewal in human pluripotent stem
511 cells. *Cell Stem Cell* (2014). doi:10.1016/j.stem.2014.04.002

512 28. Jeong, Y. M. *et al.* CDy6, a Photostable Probe for Long-Term Real-Time Visualization of
513 Mitosis and Proliferating Cells. *Chem. Biol.* (2015). doi:10.1016/j.chembiol.2014.11.018

514 29. Yun, S.-W. *et al.* Neural stem cell specific fluorescent chemical probe binding to FABP7.
515 *Proc. Natl. Acad. Sci.* (2012). doi:10.1073/pnas.1200817109

516 30. Tan, F., Qian, C., Tang, K., Abd-Allah, S. M. & Jing, N. Inhibition of transforming growth
517 factor β (TGF- β) signaling can substitute for Oct4 protein in reprogramming and maintain
518 pluripotency. *J. Biol. Chem.* (2015). doi:10.1074/jbc.M114.609016

519 31. Ruetz, T. *et al.* Constitutively Active SMAD2/3 Are Broad-Scope Potentiators of
520 Transcription-Factor-Mediated Cellular Reprogramming. *Cell Stem Cell* (2017).
521 doi:10.1016/j.stem.2017.10.013

522 32. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.
523 *Cell* (2018). doi:10.1016/j.cell.2018.05.061

524 33. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks.
525 *Cell* (2012). doi:10.1016/j.cell.2012.04.040

526 34. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring
527 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*
528 **14**, 975–978 (2017).

529 35. Shu, J. *et al.* XIInduction of pluripotency in mouse somatic cells with lineage specifiers. *Cell*
530 (2013). doi:10.1016/j.cell.2013.05.001

531 36. Montserrat, N. *et al.* Reprogramming of human fibroblasts to pluripotency with lineage
532 specifiers. *Cell Stem Cell* (2013). doi:10.1016/j.stem.2013.06.019

533 37. Koga, M. *et al.* Foxd1 is a mediator and indicator of the cell reprogramming process. *Nat.*
534 *Commun.* (2014). doi:10.1038/ncomms4197

535 38. Takahashi, K. *et al.* Induction of pluripotency in human somatic cells via a transient state
536 resembling primitive streak-like mesendoderm. *Nat. Commun.* (2014).
537 doi:10.1038/ncomms4678

538 39. Gabut, M. *et al.* An alternative splicing switch regulates embryonic stem cell pluripotency
539 and reprogramming. *Cell* (2011). doi:10.1016/j.cell.2011.08.023

540 40. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an
541 early stochastic and a late hierachic phase. *Cell* **150**, 1209–1222 (2012).

542 41. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies
543 Developmental Trajectories in Reprogramming. *Cell* (2019). doi:10.1016/j.cell.2019.01.006

544 42. Guo, L. *et al.* Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-
545 Cell RNA-Seq. *Mol. Cell* (2019). doi:10.1016/j.molcel.2019.01.042

546 43. Levine, A. J. & Brivanlou, A. H. GDF3, a BMP inhibitor, regulates cell fate in stem cells and
547 early embryos. *Development* **133**, 209–16 (2006).

548 44. O’Malley, J. *et al.* High-resolution analysis with novel cell-surface markers identifies routes
549 to iPS cells. *Nature* **499**, 88–91 (2013).

550 45. Liu, J. *et al.* The oncogene c-Jun impedes somatic cell reprogramming. *Nat. Cell Biol.* **17**,
551 856–867 (2015).

552 46. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming.
553 *Cell* **168**, 442–459.e20 (2017).

554 47. Loh, K. M. & Lim, B. A precarious balance: Pluripotency factors as lineage specifiers. *Cell*
555 *Stem Cell* (2011). doi:10.1016/j.stem.2011.03.013

556

557 **ACKNOWLEDGMENTS**

558 We are grateful for Ahad Khalilnezhad for technical assistance and Lin Yang and Samantha Seah
559 for helpful discussion. H.L. is supported by Glenn Foundation for Medical Research, Mayo Clinic
560 Center for Biomedical Discovery and Mayo Clinic Cancer Center. Y-H.L. is supported by the [NRF
561 Investigatorship award -NRFI2018-02] and [JCO Development Programme Grant - 1534n00153]
562 grants. Y-H.L. and L-F.Z. are supported by the Singapore National Research Foundation under its
563 Cooperative Basic Research Grant administered by the Singapore Ministry of Health's National
564 Medical Research Council [NMRC/CBRG/0092/2015]. We are grateful to the Biomedical Research
565 Council, Agency for Science, Technology and Research, Singapore for research funding.

566

567 **AUTHOR CONTRIBUTIONS**

568 Contribution: Q.R.X, C.E.F designed and performed research, analyzed data, and wrote the
569 paper; P.G, Y.S.C, C.X.D.T, T.W designed and conducted research; N.Y.K, S.S, Y.T.C, J.X, J.C,
570 H.L, L.F.Z analyzed data; and Y.H.L. designed research, analyzed data, and wrote the paper.

571

572 **AUTHOR INFORMATION**

573 The authors declare no competing financial interests.

574

575 **FIGURE LEGENDS**

576 **Figure 1. Schematic of the single-cell systems used for de-convoluting the heterogeneity in**
577 **human cellular reprogramming**

578 (a) Overview of the prepared single-cell NGS libraries at the indicated time-points of the human
579 cellular reprogramming. Two single-cell platforms were utilized for the study. Microfluidics
580 platform was used to prepare 439 single cell RNA-Seq and 891 single cell ATAC-Seq libraries

581 (Duplicates). 10x Genomics platform provided us with an additional 32138 cells for scRNA-Seq
582 analysis. (b) Quality control of Microfluidics-capture based scRNA-Seq libraries. Dotplot
583 demonstrates the exon mapping percentage (X-axis) of each scRNA-Seq libraries, along with its
584 corresponding detected gene rates (Y-axis). Axis crosses each other at the cutoff values that filter
585 the libraries. Blue dots represent libraries that passed the QC filters. (c) Average enrichment of
586 scRNA-Seq libraries over genebodies. (d) t-SNE plot of the prepared 10x scRNA-Seq libraries
587 based on total cellular transcriptomes. (e-f) Super-imposition of the single-cell expression levels of
588 MET genes: *ZEB1* and *EPCAM* (e), and fibroblast and pluripotent genes: *COL1A1* and *NANOG* (f),
589 on the tSNE plot. The expression ranges from no expression (grey) to high expression (red). (g)
590 Quality control of scATAC-Seq libraries. Dotplot demonstrating the library size (X-axis) of each
591 scATAC-Seq library, along with its corresponding contribution to its respective time-point's HARs
592 (Y-axis). Red dots represent the cells that pass the QC filters. (h) Average enrichment profile of a
593 D16+ scATAC-Seq library around Transcription Start Sites (TSS) of the genome with a window of
594 -3K to 3K. Y-axis denotes the average normalized read counts of the library over the indicated
595 region in the genome (x-axis). (i) Histograms of insert size metric of a D16+ scATAC-Seq library
596 revealing a nucleosomal pattern, which is characteristic of a good ATAC-Seq library.

597 **Figure 2. Identification of diverse reprogramming subgroups**

598 (a-c) RCA clustering of D2 (a), D8 (b), and D16+ (c) cells based on the profile of the expressed
599 genes. The PCAs show subgroups of reprogramming cells at individual time points. Each color
600 represents a subgroup. (d) Heatmap showing dynamics of the genes differentially expressed among
601 4 subgroups of D2, 3 subgroups of D8, and 2 subgroups of D16+ cells, across the reprogramming
602 process. Color represents the expression level, ranging from dark blue (low) to dark red (high). The
603 expression values are $\text{Log}_{10}(\text{FPKM}+1)$. The color code on top represents the time points (above)
604 and their respective subgroups (below). (e) Bar charts demonstrating the expression pattern of D2
605 (left), D8 (middle), and D16+ (right) subgroup specific genes. Color represents the expression level.
606 X-axis represents the percentage of cells with the expression within the indicated range. Red dotted

607 boxes highlight the subgroups with high expression of the respective gene. (f) Box plots showing
608 the single cell expression of the differentially expressed genes *CDK1* (D2), *GDF3* (D8), *MMP2*
609 (D16+), and *LIN28A* (D16+) across all the time points and their respective subgroups. Lines in the
610 box represent the median expression in the respective subgroups or time points. The expression
611 values are Log_{10} (FPKM+1). (g) Confusion matrix of the scRNA-seq libraries of all time points,
612 using Random Forest algorithm. Color represents the number of libraries from each time point
613 predicted to be similar to the indicated subgroups of D8. Scale ranges from dark blue (low number)
614 to dark red (high number).

615 **Figure 3. Trajectories of cellular reprogramming**

616 (a) Trajectory of reprogramming cells identified from the 10x scRNA-Seq libraries based on
617 DDRTree dimension reduction. Color represents the time points. (b) Pseudotime calculation based
618 on the DDRTree trajectory. Color indicates pseudotime, ranging from dark blue (early) to light blue
619 (late). The plot shows four branching events. “+succ” and “-succ” represents the successful and
620 unsuccessful branches. (c) Stacked columns indicating the percentage of cells belonging to the
621 respective time-points in the reprogramming trajectory states. Color represents the time points. (d)
622 Super-imposition of D8 subgroups (left) and D16+ subgroups (right) on the trajectory of
623 reprogramming. Color represents the subgroups. (e) Stacked columns revealing the percentage of
624 cells of the respective D8 subgroups in the indicated reprogramming trajectory states. Colors
625 represent the subgroups of D8 and grey color indicates cells of the other time points. (f) Super-
626 imposition of the expression of D8 subgroup genes (*RFC3*, *GDF3*) and D16+ subgroup genes
627 (*NANOG*, *MMP2* and *LIN28A*) on the reprogramming trajectories. Expression ranges from purple
628 (low) to yellow (high). (g-h) GO analysis of the differentially expressed genes associated with the
629 unsuccessful (g) and successful (h) reprogramming branches. The enrichment score ranges from
630 white (no) to red (high). The branches are labelled in Fig.3b.

631 **Figure 4. Identification of chemical dyes and surface markers for early-intermediate**
632 **reprogramming cells**

633 (a) Quantification of TRA-1-60+ colonies yielded from D8 cells sorted with BDD2-C8 (top).
634 Representative images are shown below. n=3; error bar indicates SD. (b) Gene expression in the D8
635 BDD2-C8+ and BDD2-C8- cells, measured by single cell qRT-PCR. Student t-test (2-tails) was
636 applied (p<0.05). (c) RCA plot of D8, D8^{BDD2-C8+} and D8^{BDD2-C8-} cells. (d) Expression heatmap of
637 the D8 differential genes in D8^{BDD2-C8+} and D8^{BDD2-C8-} cells. Color code on top represents the time
638 points (above) and the subgroups or sorted cells (below). (e) GO terms enriched by the differential
639 genes of D8^{BDD2-C8+} and D8^{BDD2-C8-} cells. (f) Heatmap showing the expression dynamics of the
640 surface markers. Color code on top represents the time point and their respective subgroups. (g)
641 Boxplots showing the surface markers expression across the time points and their respective
642 subgroups. Lines in the box represent the median expression. (h) Stacked histograms (top) showing
643 the fluorescence intensity (X-axis) of the surface markers. Red dotted boxes highlight the positively
644 stained cells. Quantifications are shown below. (i) Barchart exhibiting the relative expression levels
645 in the D8 sorted cells, normalized to that of D8 CD44+ cells. n=2; error bar indicates SD. (j)
646 Quantification of TRA-1-60+ colonies yielded from the D8 sorted cells. Representative images are
647 shown below. n=2; error bar indicates SD. (k) Boxplots showing the surface marker expression in
648 the D8^{BDD2-C8+} and D8^{BDD2-C8-} cells. Lines represent the median expression. (l) Overlaid histograms
649 showing the surface marker staining intensity in the BDD2-C8+ and BDD2-C8- cells. Red dotted
650 boxes highlight the positively stained cells. The numbers on top indicate the percentage of
651 positively stained cells. (m) Barchart showing the distribution of co-staining signals of CD13 and
652 CD44 in the cells of various reprogramming time points and cell lines. (n) Barcharts exhibiting the
653 relative expression of the collagen/ mesenchymal genes (top) and pluripotent genes (bottom) in the
654 D8 CD13 sorted cells. n=2; error bar indicates SD. (o) Quantification of TRA-1-60+ colonies
655 yielded from D8 CD13 sorted cells (top). Representative images are shown below. n=2; error bar
656 indicates SD.

657 **Figure 5. Refined classification and enrichment of early-intermediate reprogramming cells**

658 (a) t-SNE plot generated by 10X libraries of D8 CD13 sorted cells. (b) t-SNE plot showing the
659 clusters among the CD13+ and CD13- cells. (c) RCA heatmap demonstrating the clustering of
660 CD13 sorted cells (column) based on their correlation to the cells of different lineages (row). Color
661 code on top indicates the CD13 antigen profile (below) and clusters (above). (d) MAGIC plot
662 showing the correlative expression of *CD13* and *GDF3* in the D8 CD13-sorted 10x libraries. Color
663 represents the expression level of *NANOG*. (e) Violin plot demonstrating the expression of *GDF3*
664 across the identified clusters. (f) Top: Reprogramming trajectory constructed by the 10X scRNA-
665 Seq libraries of various time points and the D8 CD13 sorted cells. Bottom: Super-imposition of
666 CD13 clusters on the reprogramming trajectory. (g) Heatmap showing the differentially expressed
667 genes of CD13 clusters. Genes highlighted in orange were expressed higher in D8 G2 and G3 than
668 G1. (h) Barcharts showing the GO terms enriched by the genes highly expressed in the indicated
669 CD13 clusters. (i) MAGIC plot showing the correlative expression of *CD13* and *CD201* in the
670 CD13 10X libraries. Color represents the expression level of *GDF3*. (j) Barchart showing the
671 distribution of co-staining signals across the various reprogramming time points and cell lines. (k)
672 Barchart exhibiting the relative gene expression in the D8 dual antibody sorted cells, normalized to
673 that in CD13+CD201+ cells. n=2; error bar indicates SD. (l) Left: Heatmap showing the
674 differentially expressed genes among the D8 dual antibody sorted cells. GO terms (Middle)
675 enriched by the differentially expressed genes, with the enrichment values and the related genes
676 indicated on the right. (m) Bar chart demonstrating the number of normalized TRA-1-60+ colonies
677 (Y-axis) upon the knock-down of genes highly expressed in CD13+ CD201+ or CD13-CD201+
678 cells, at day 5 of reprogramming. Representative images are shown above. n=3. Error bar indicates
679 SD. (n) Quantification of TRA-1-60+ colonies yielded from the D8 dual antibody sorted cells.
680 Representative images are shown above. n=2; error bar indicates SD.

681 **Figure 6. Transcription factors critical for reprogramming**

682 (a) Heatmap showing the expression dynamics of the transcription factors across the pseudotime
683 states. Color represents the expression level, ranging from purple (low) to yellow (high). Color code
684 on top represents the pseudotime states. Transcription factors were classified to 5 categories
685 according to the states when they were silenced or expressed. The representative TFs of each
686 category are listed on the right. (b) Correlation heatmap of scATAC-Seq libraries based on the
687 calculated JASPAR motifs deviations in the HARs peaks. Color represents the correlation value,
688 ranging from blue (no) to red (high). Side color bar indicates the time-point to which each scATAC-
689 Seq library belongs. (c) Plot indicating the significantly variable motifs in terms of accessibility
690 from the scATAC-Seq libraries. Y-axis represents the variability score assigned to each JASPAR
691 motif whereas the X-axis represents the motif rank. (d) Heatmap of scATAC-Seq libraries based on
692 the deviation scores of the significantly variable JASPAR motif. Color indicates the accessibility
693 level, ranging from dark blue (no enrichment) to dark red (high enrichment). Color code on top
694 represents the time points. Motifs were classified to 3 major types according to the dynamics of
695 accessibility across the time points. (e) t-SNE plot of scATAC-Seq libraries based on the deviation
696 scores of JASPAR motifs. (f-i) Super-imposition of motif enrichment scores for Close motif-
697 FOSL1 and CEBPA (f), Transient motif-GATA1:TAL1 (g), Open motif: Early Open-TEAD4 (h);
698 Late Open- FOXL1 (i) on the scATAC-Seq tSNE plot. Color indicates the accessibility level and
699 ranges from dark blue (no enrichment) to dark red (high enrichment).

700 **Figure 7. Transcription factors contributing to the heterogeneity in chromatin accessibility of**
701 **the intermediate reprogramming cells**

702 (a) Plot indicating the significantly variable motifs in terms of accessibility among D8 cells. Y-axis
703 represents the variability score assigned to each JASPAR motif whereas X-axis represents the motif
704 rank. (b) Heatmap showing clustering of D8 scATAC-Seq libraries based on the deviations scores
705 of the significantly variable JASPAR motif. (c) Super-imposition of motif enrichment scores for
706 FOSL1 and TEAD4 on the D8 scATAC-Seq tSNE plot. (d) Super-imposition of *FOSL1* and *TEAD4*
707 expression on the reprogramming trajectory displayed in Figure 3a. (e) Dotplots indicating the

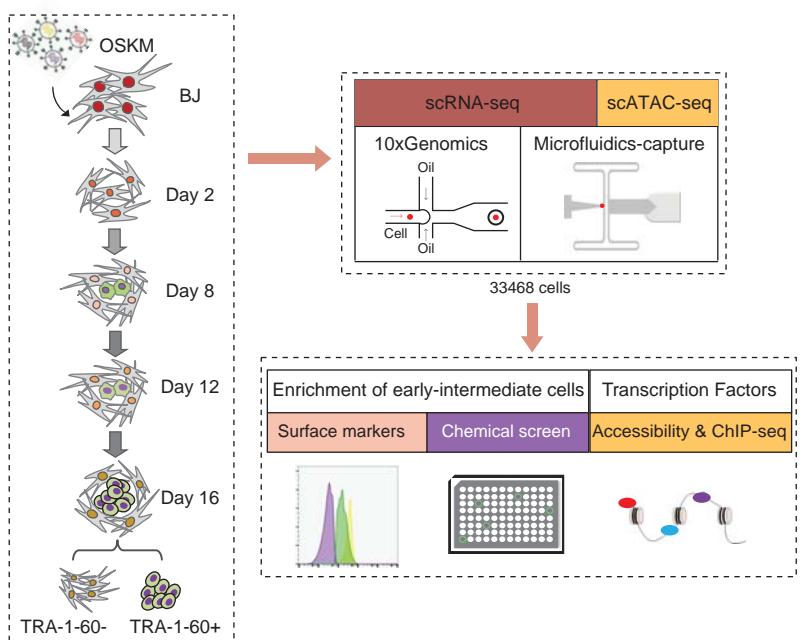
708 expression of *FOSL1* and *TEAD4* along the pseudotime. Smooth lines represent the mean
709 expression level at each pseudotime, regardless of the state. (f) Top: MAGIC plot showing
710 correlative expression of *FOSL1* and *TEAD4* in the 10X libraries. Bottom: Super-imposition of
711 *GDF3* expression on the magic plot. (g) Top: tSNE plot based on the regulon activity matrix
712 generated by SCENIC. Bottom: Activity of *FOSL1* and *TEAD4* regulons across the time-points. (h)
713 Super-imposition of *FOSL1* and *TEAD4* expression on the D8 CD13 sorted tSNE plot. (i) Bar chart
714 demonstrating the number of normalized TRA-1-60+ colonies upon the knock-down of *FOSL1* at
715 the indicated time-points of reprogramming. Representative images are shown below, n=3. Error
716 bar indicates SD. (j) Bar chart demonstrating the quantification of TRA-1-60+ colonies upon the
717 overexpression of *FOSL1* at D5. Representative images are shown below, n=3. Error bar indicates
718 SD. (k) Bar chart demonstrating the quantification of TRA-1-60+ colonies yielded from D8 CD13-
719 cells with *FOSL1* overexpression. Representative images are shown below, n=3. Error bar indicates
720 SD. (l) Bar chart demonstrating the number of normalized TRA-1-60+ colonies upon the knock-
721 down of *TEAD4* at the indicated time-points of reprogramming. Representative images are shown
722 below, n=2. Error bar indicates SD. (m) Bar chart demonstrating the quantification of TRA-1-60+
723 colonies upon the overexpression of *TEAD4*. Representative images are shown below, n=3. Error
724 bar indicates SD. (n) Bar chart demonstrating the number of TRA-1-60+ colonies yielded from D8
725 CD13- cells with *TEAD4* depletion. Representative images are shown below, n=3. Error bar
726 indicates SD.

727 **Figure 8. Mechanistic role of FOSL1 and TEAD4 during reprogramming**

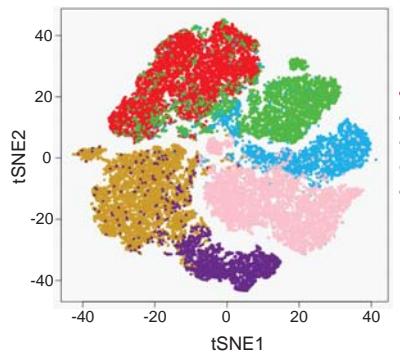
728 (a) Heatmaps exhibiting the D8 FOSL1 and D8 TEAD4 ChIP-Seq enrichment over merged binding
729 loci of both these factors. (b) Left: t-SNE clustering of BJ, D16+, and D16- scATAC-Seq libraries
730 based on the deviation scores of common and specific sites identified from D8 FOSL1 ChIP-seq,
731 D8 CD13- TEAD4 ChIP-seq, and D8 CD13+ TEAD4 ChIP-seq. Right: Super-imposition of
732 deviation score for FOSL1 specific bound sites and TEAD4 CD13- specific bound sites on the
733 scATAC-Seq t-SNE plot. Color indicates the accessibility level. (c) MA plots of scATAC-Seq

734 revealing the differentially accessible FOSL1 bound sites (left) and CD13- specific TEAD4 bound
735 sites (right) between BJ and D16+ cells. X-axis denotes the mean of normalized counts in BJ cells.
736 Red dots denote the sites with significant accessibility changes. (d) Super-imposition of enrichment
737 scores for FOSL1 specific bound sites (left) and CD13- TEAD4 specific bound sites (right) on the
738 D8 scATAC-Seq tSNE plot. Color indicates the accessibility leve. (e) UCSC screenshots
739 illustrating the chromatin accessibility of FOSL1bound genes (MMP2 and SMAD3) and TEAD4
740 CD13- bound sites (TET1 and RFC3) in D8 “FOSL1 ChIP only” and D8 “TEAD4 ChIP only” cells.
741 The differentially accessible sites with highlighted with dotted rectangles. (f-g) Expression heatmap
742 of the differentially accessible genes bound by FOSL1 (f) in D8 and TEAD4 in D8 CD13- cells (g),
743 among the D8 dual antibody sorted cells. (h-i) Bar chart demonstrating the number of TRA-1-60+
744 colonies upon knock-down of FOSL1 targets (h) and TEAD4 targets (i) at D5. Representative
745 images are shown below. n=3. Error bar indicates SD. (j) Proposed model of the study.

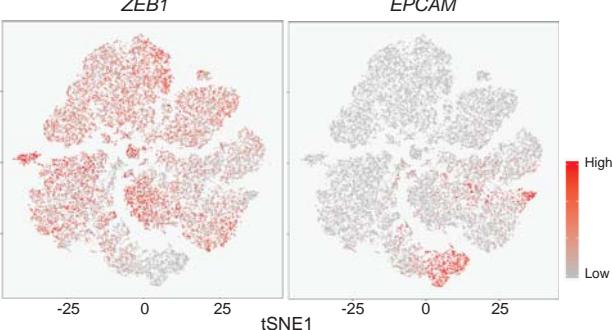
a



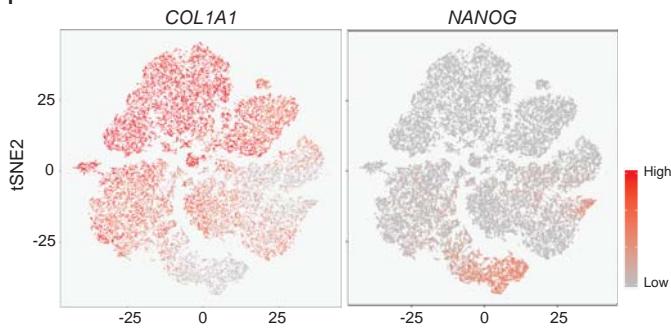
d



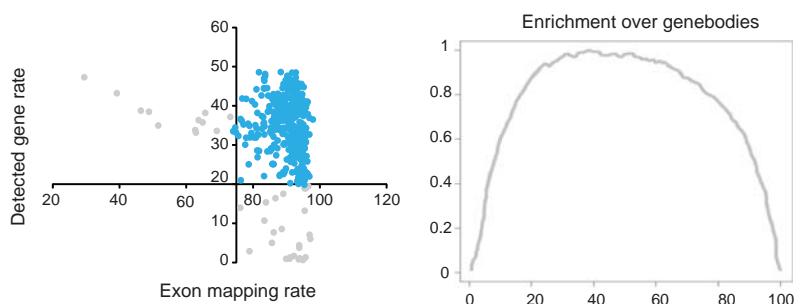
e



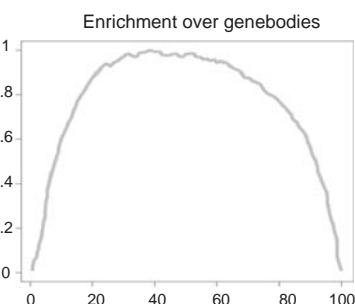
f



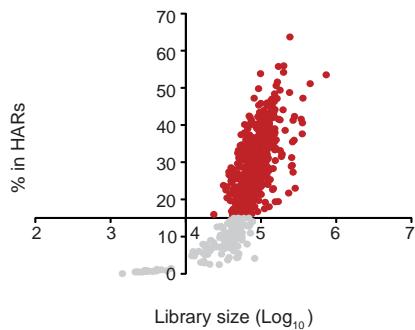
b



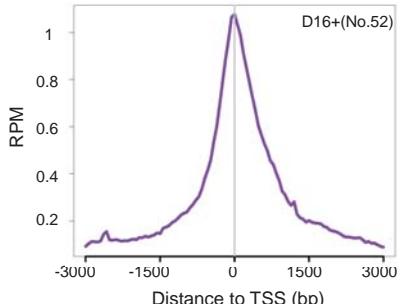
c



g



h



i

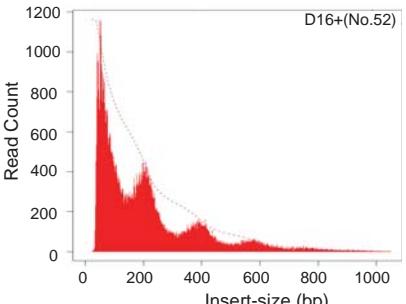


Figure 1

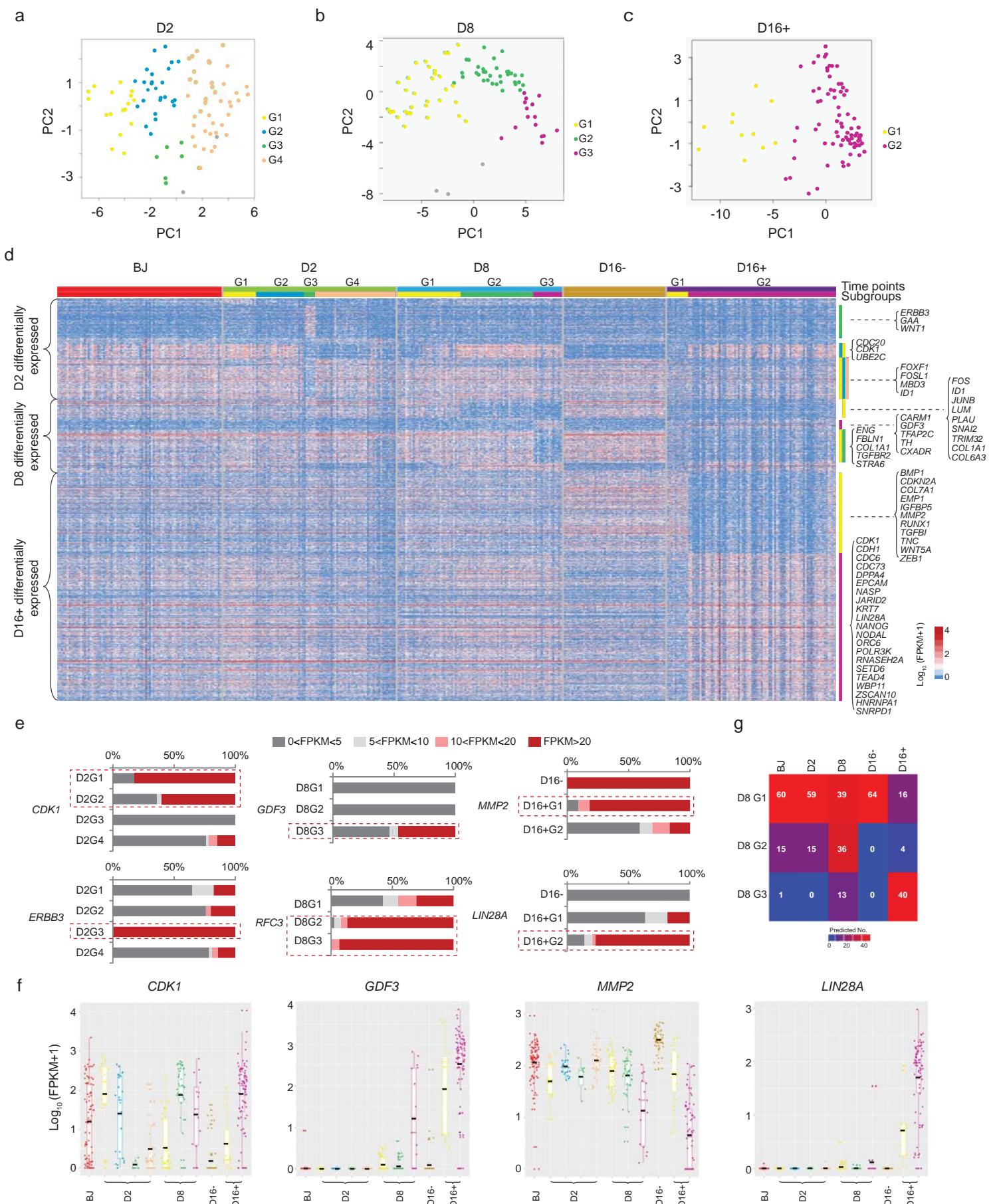


Figure 2

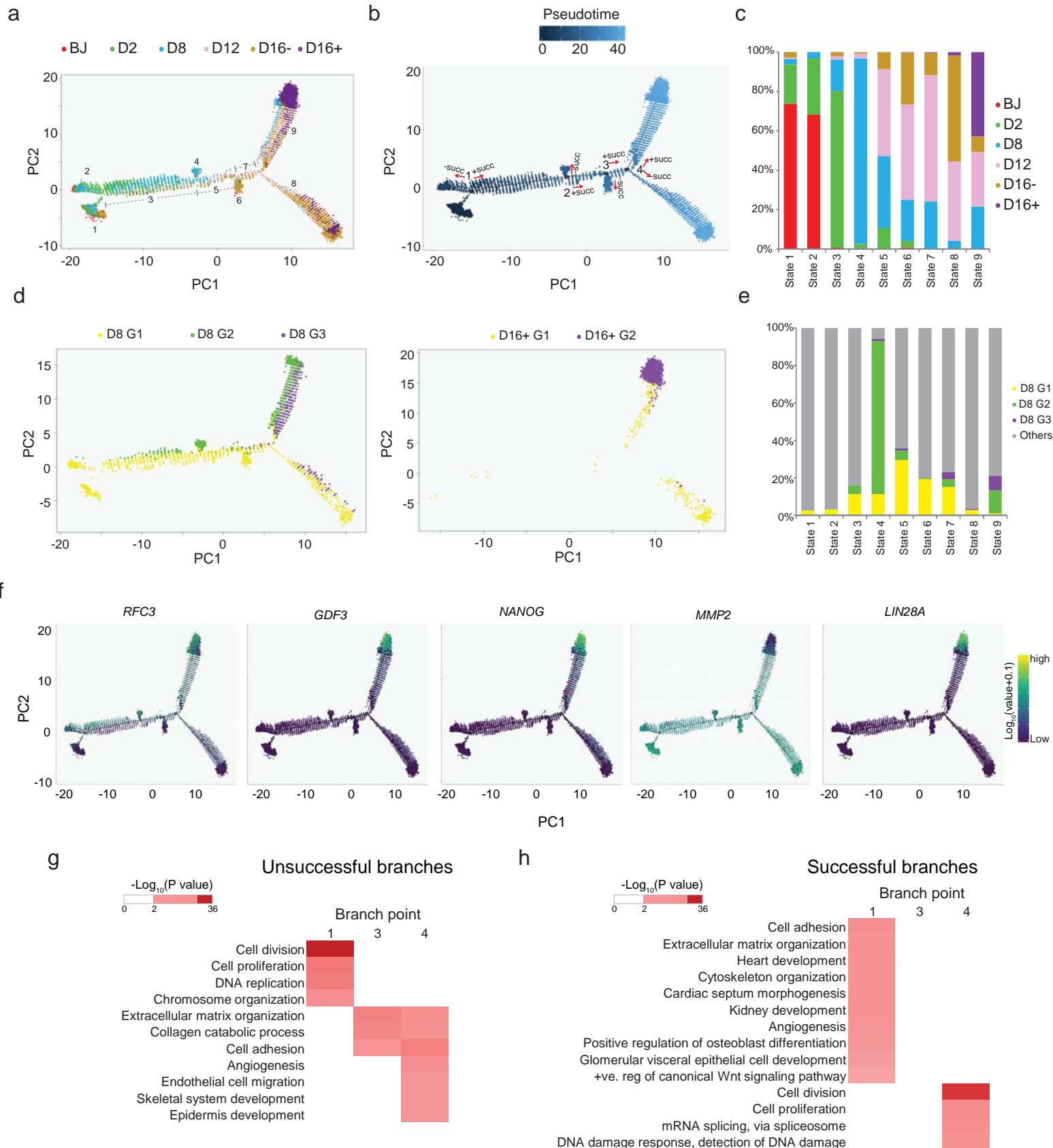
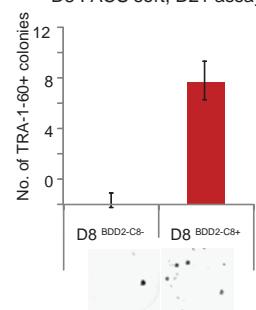


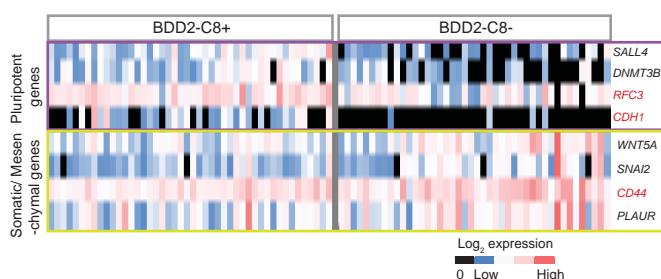
Figure 3

a

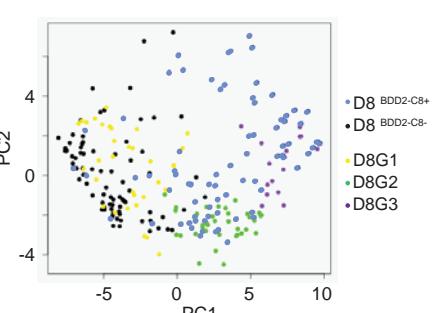
D8 FACS sort, D21 assay



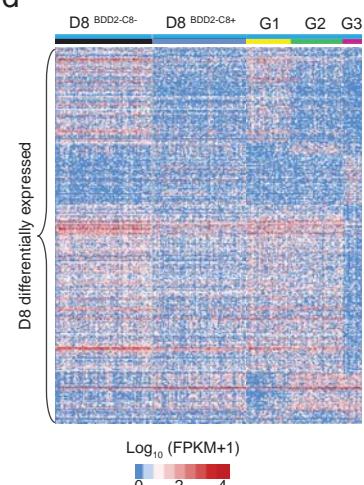
b



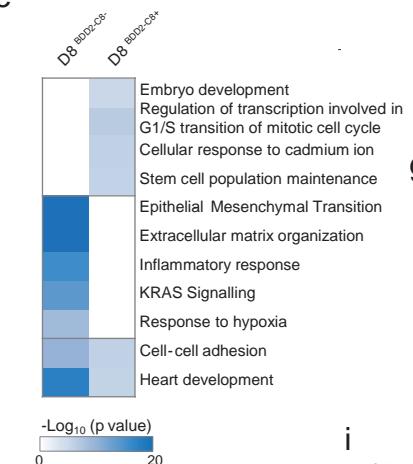
c



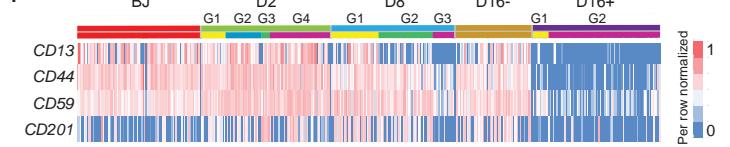
d



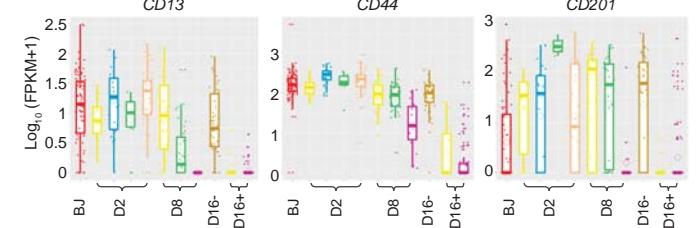
e



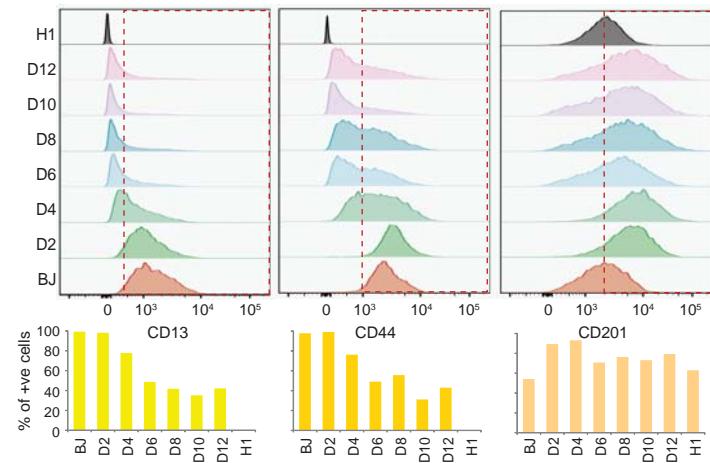
f



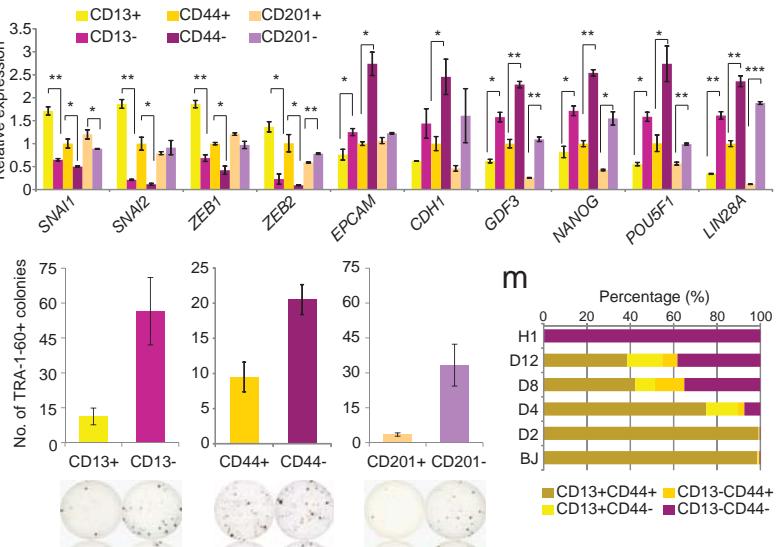
g



h



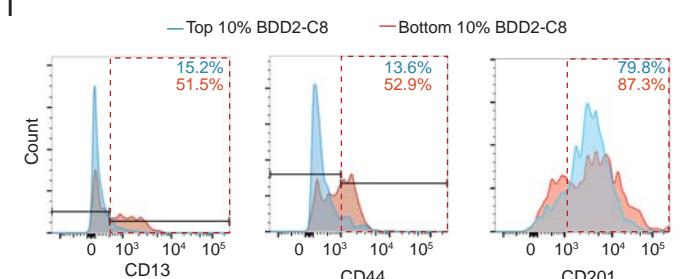
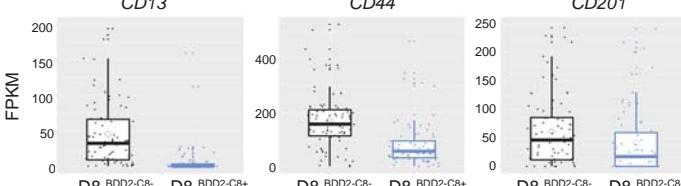
i



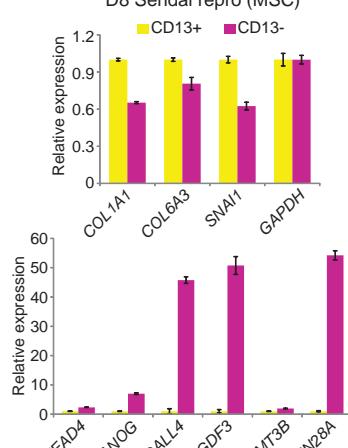
j



k



n



o

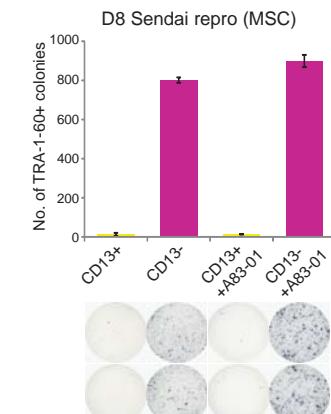


Figure 4

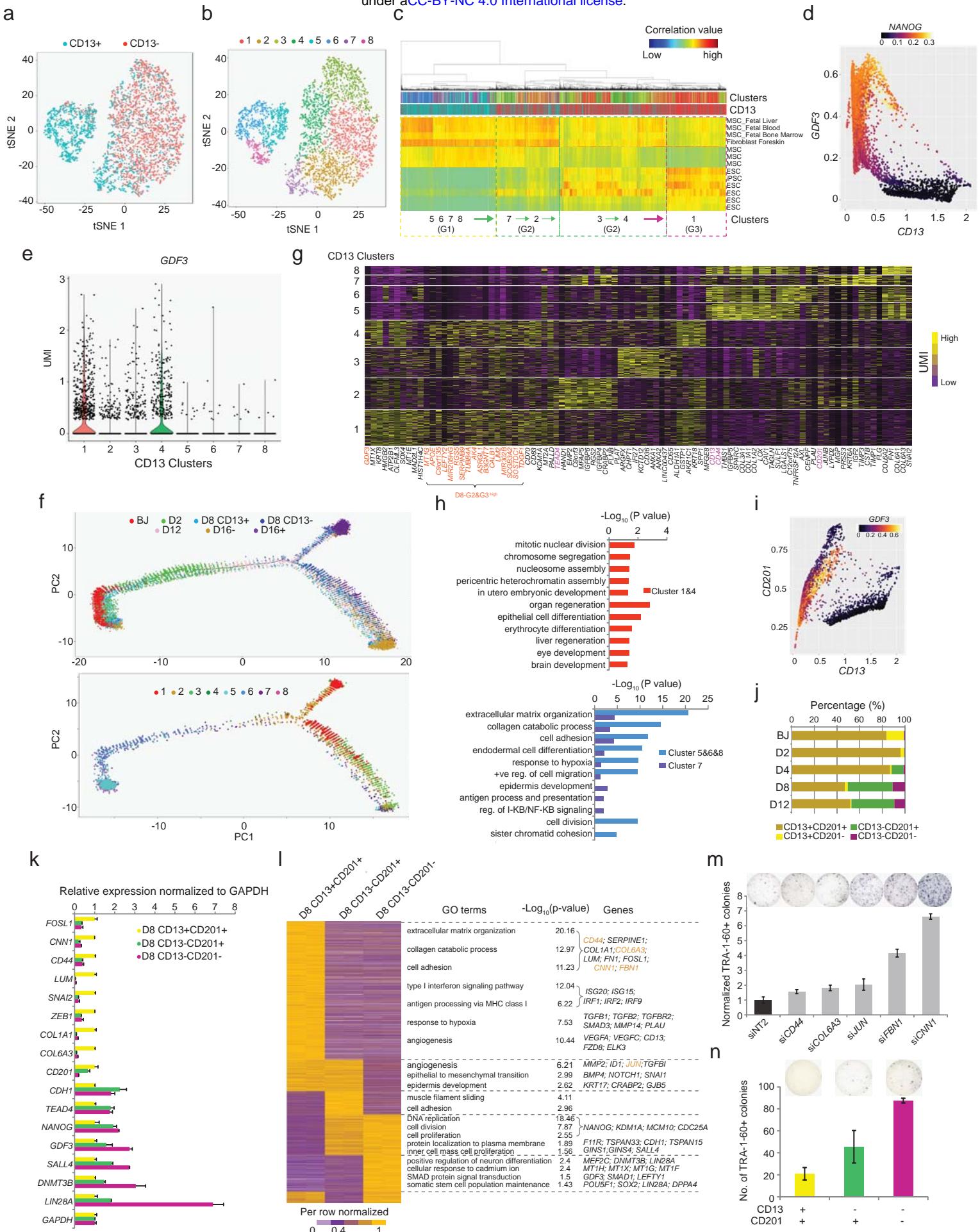


Figure 5

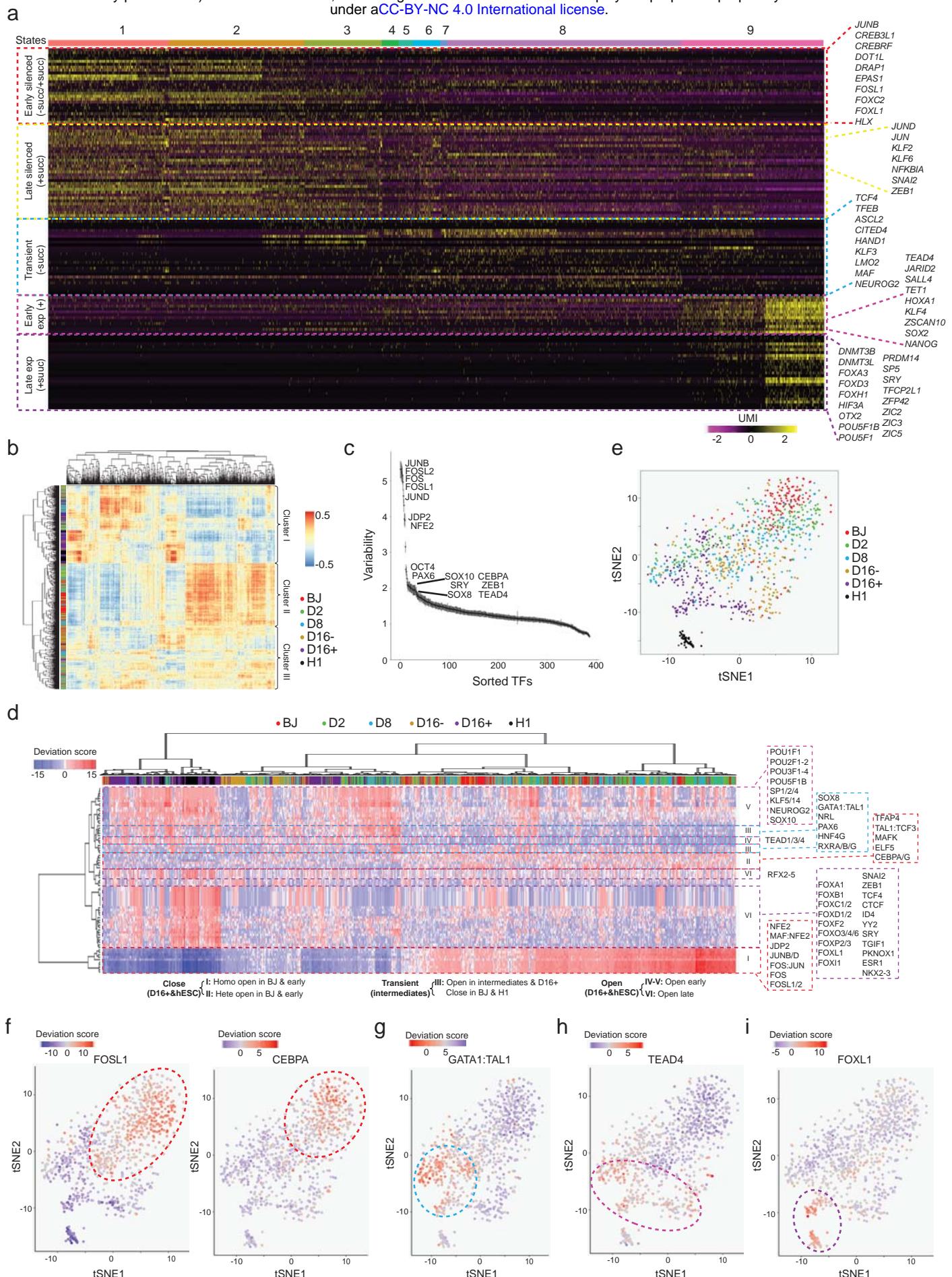


Figure 6

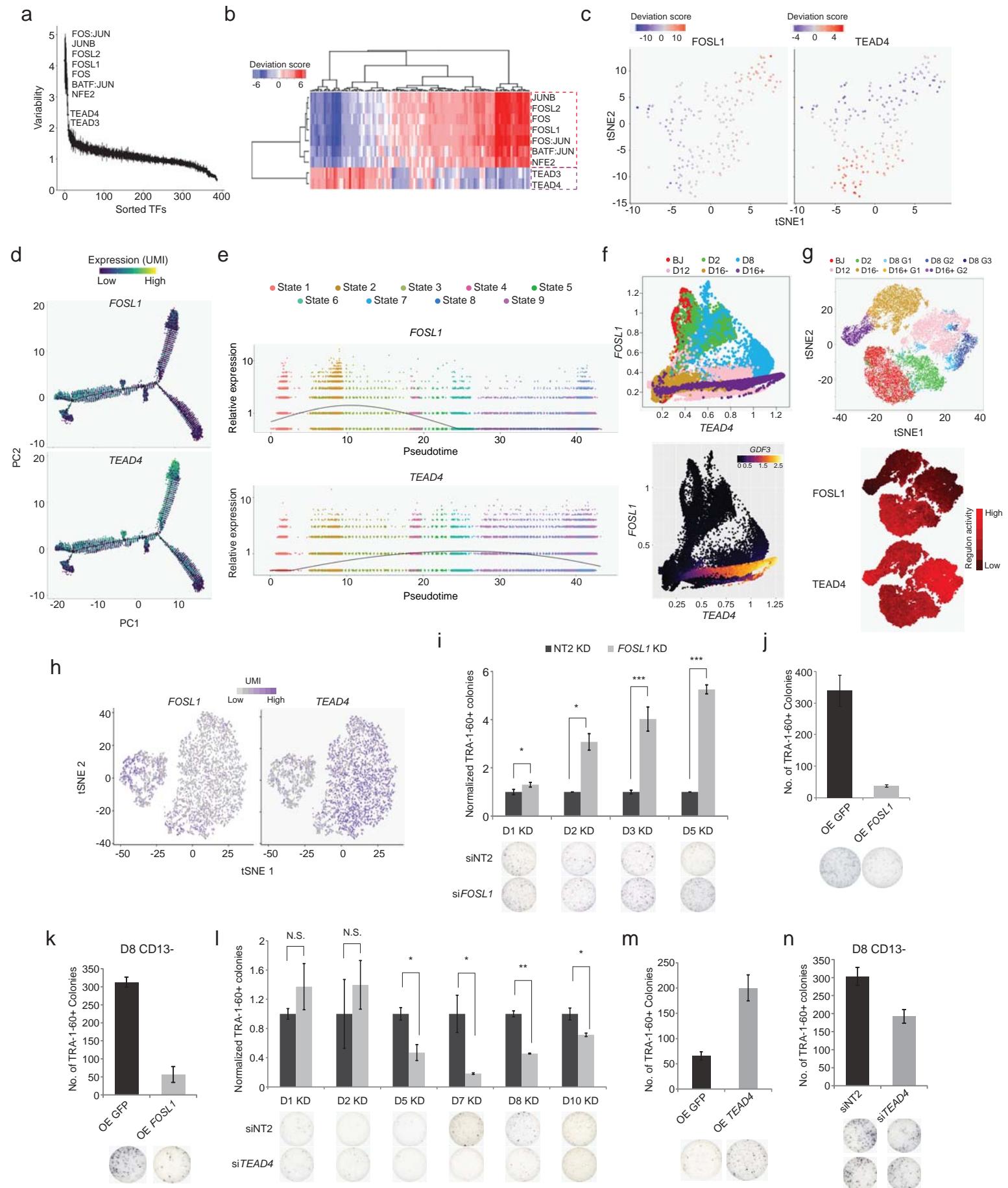


Figure 7

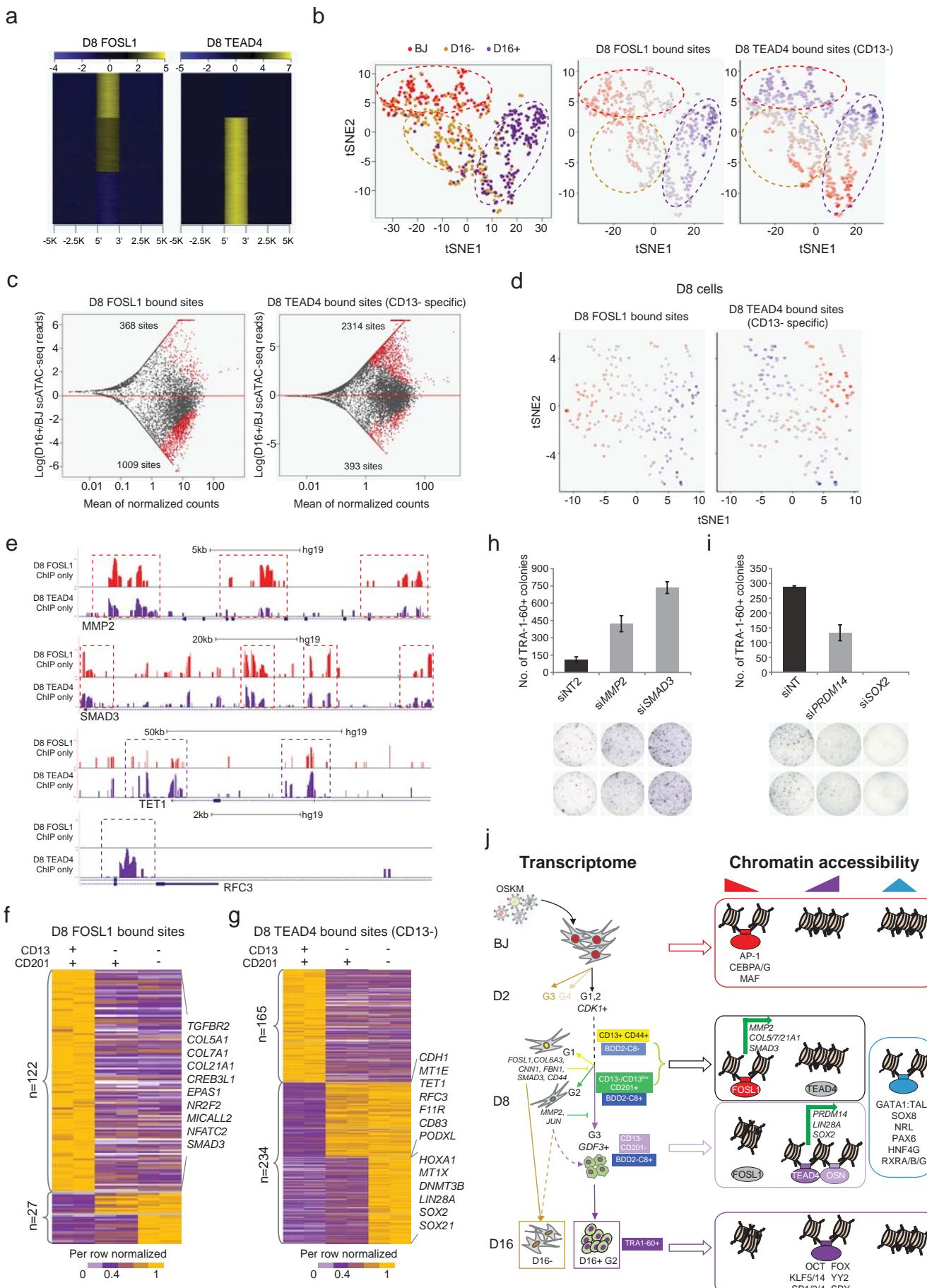


Figure 8