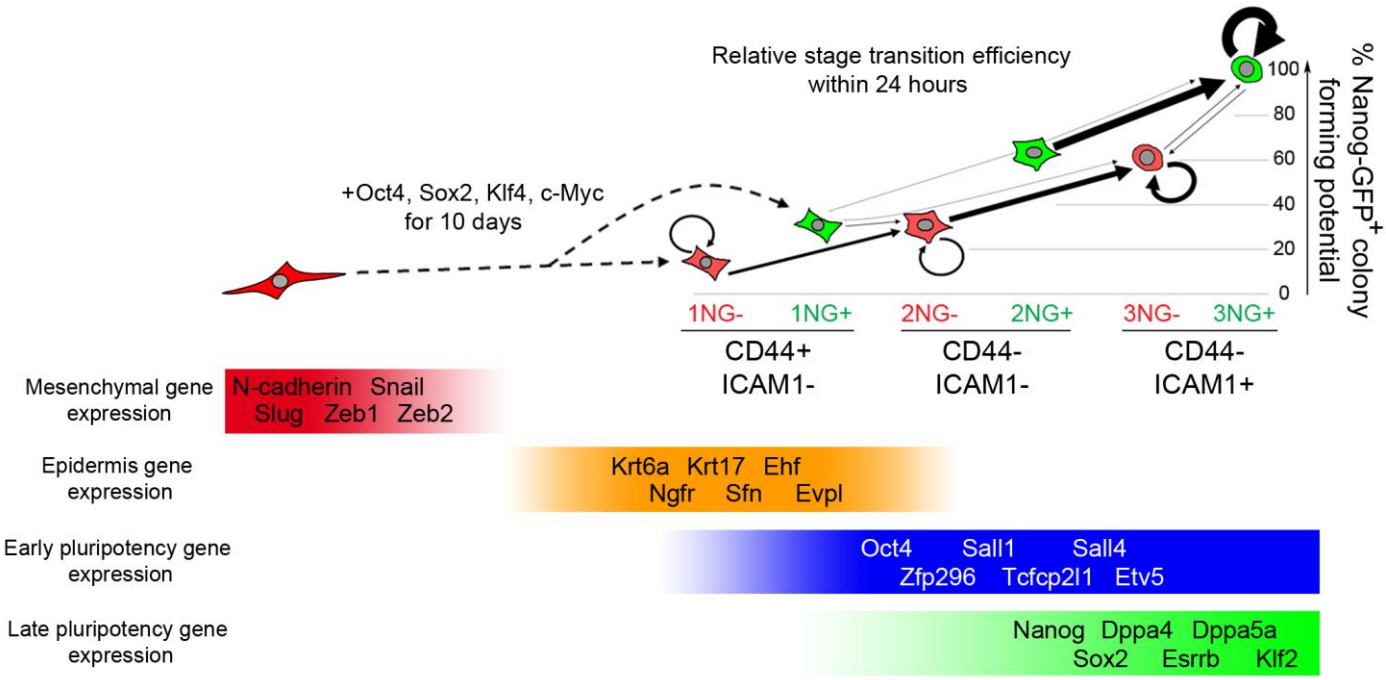
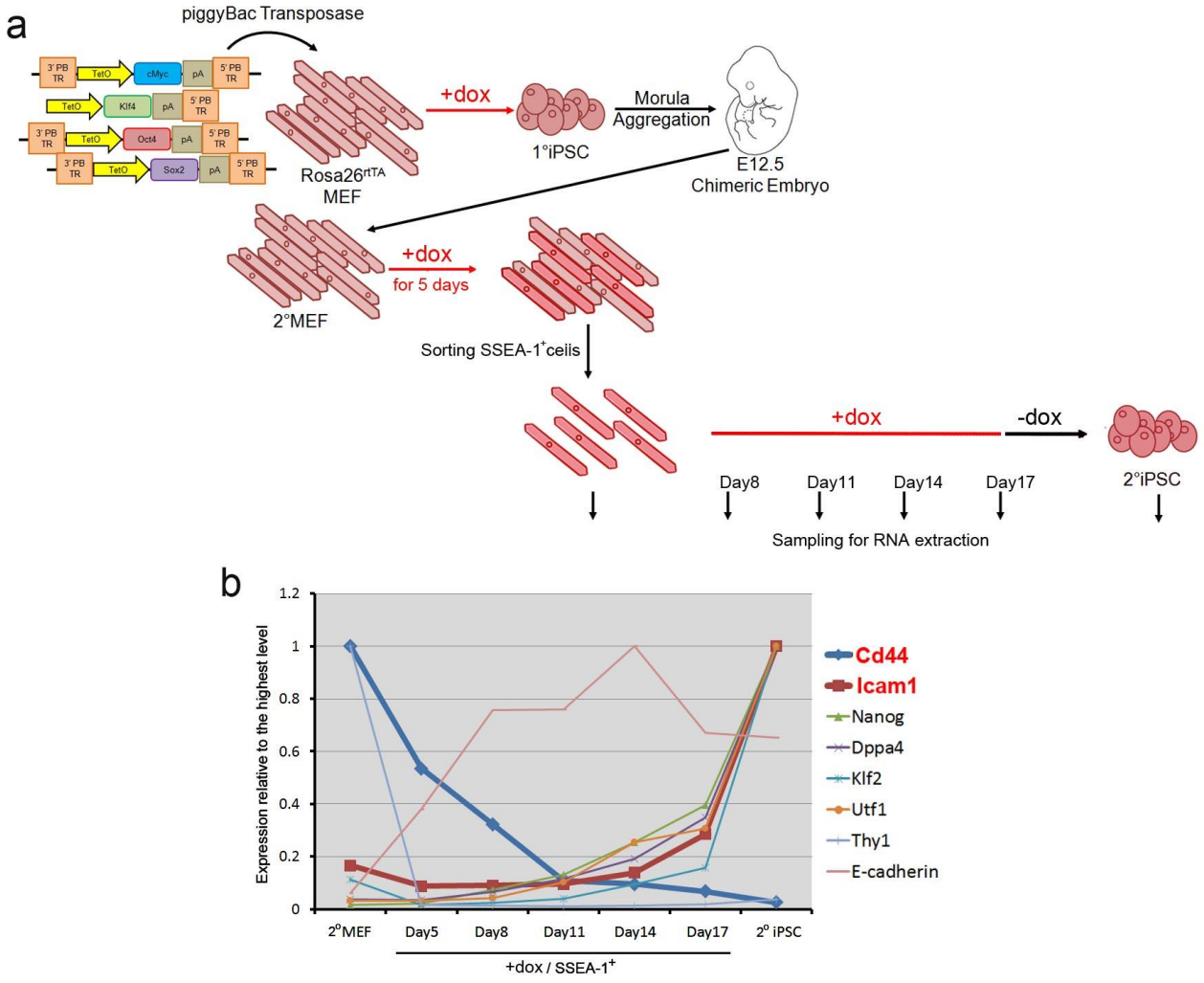
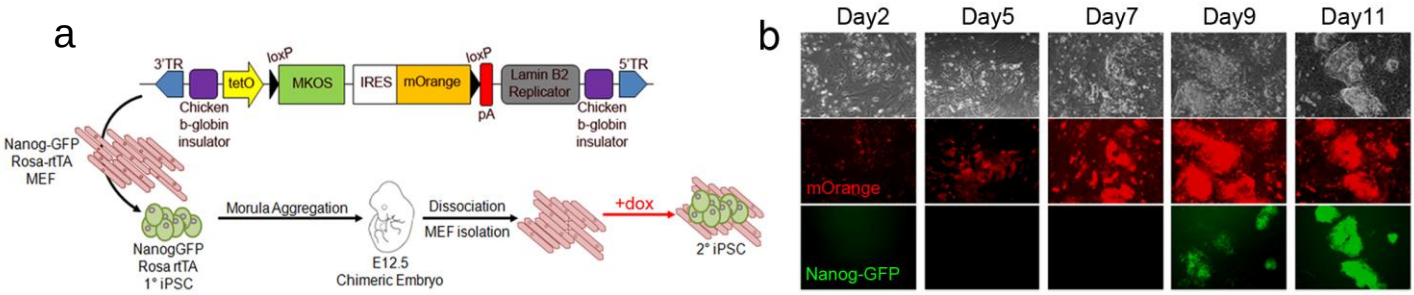


Supplementary Figure 1. A Route map to iPSCs defined by CD44 and ICAM1 expression change. Sequential changes of CD44 and ICAM1 expression allowed isolation of subpopulations at different stages of reprogramming, the progression of which was accompanied by a gradual increase of iPSC colony formation efficiency. Differences in the transition rate of each subpopulation from one stage to the next revealed preferential routes to iPSCs. Global gene expression profiling highlighted transient up-regulation of multiple epidermis genes. There were two groups of pluripotency genes; those displaying early initiation of expression which overlapped with epidermis gene expression (Early) and those which were up-regulated in parallel with down-regulation of epidermis genes (Late).

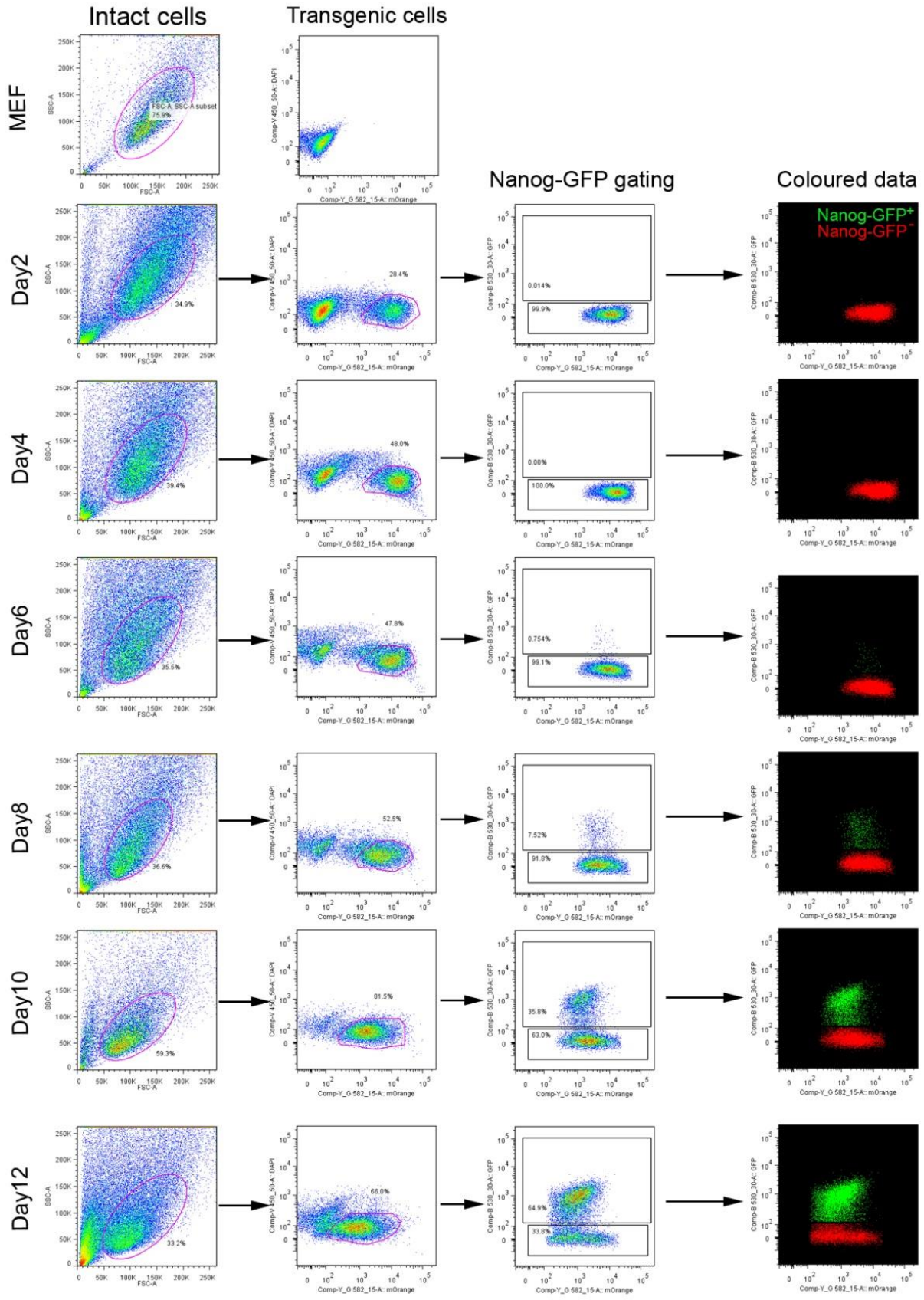


Supplementary Figure 2. Identification of novel reprogramming cell surface markers, Cd44 and Icam1. a. Secondary (2^o) reprogramming was carried out using the 6c cell line generated by 4 *piggyBac* (PB) transposons carrying reprogramming factors *c-Myc*, *Klf4*, *Oct4* and *Sox2*. Subsequently RNA was extracted from SSEA-1^+ cells at day5, 8, 11, 14, 17 as well as from 2^o iPSCs for microarray analysis to identify novel reprogramming cell surface markers. This experiment was performed in the absence of VitC and Alki. **b.** Microarray analysis of 6c 2^o reprogramming identified Cd44 and Icam1 as potential cell surface markers to dissect the reprogramming process. *Thy1* expression is already down-regulated in SSEA1^+ cells at day5 and *E-cadherin* expression plateaus at day8, suggesting previously identified markers are not suitable for investigation of the later stages of reprogramming.

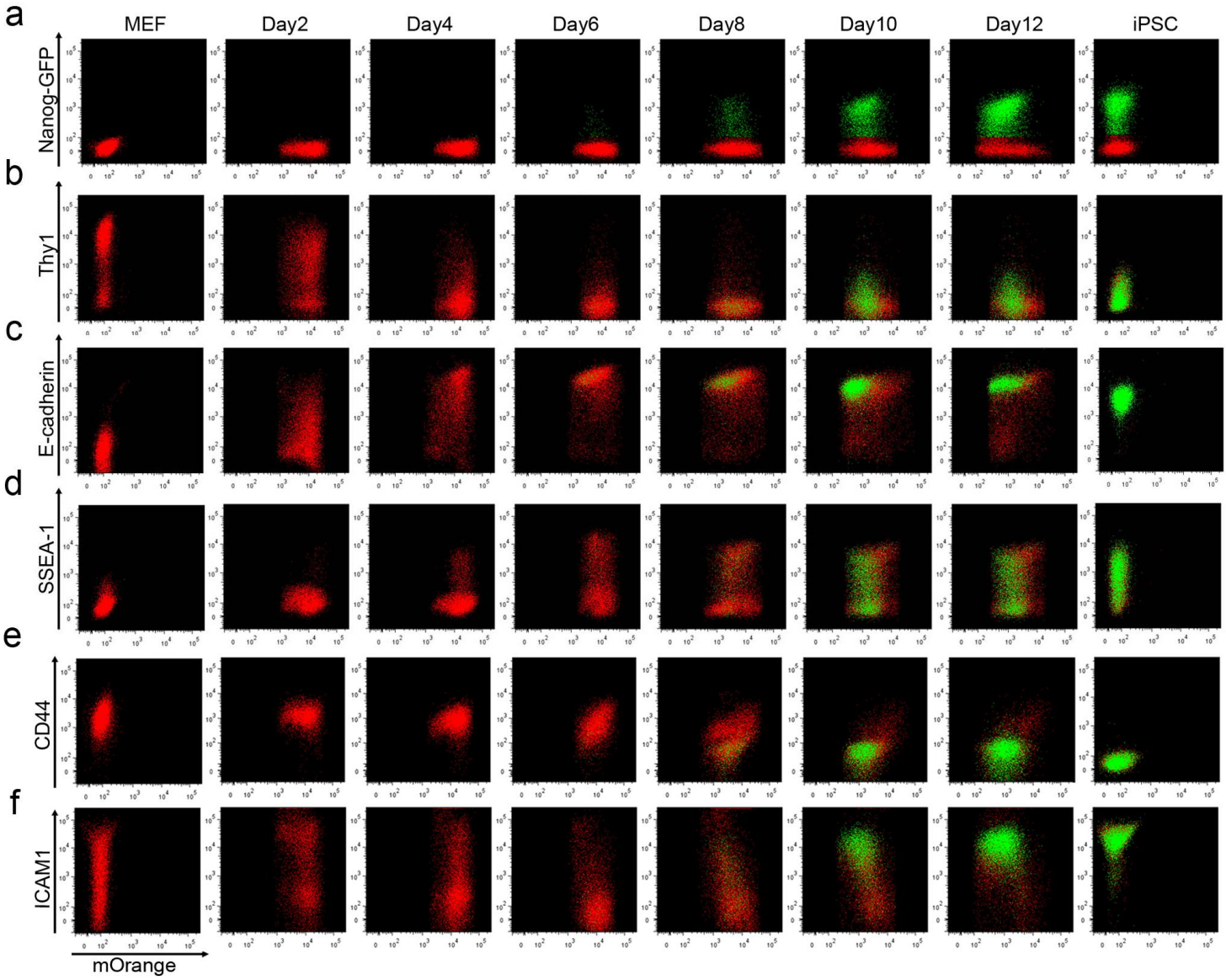
.



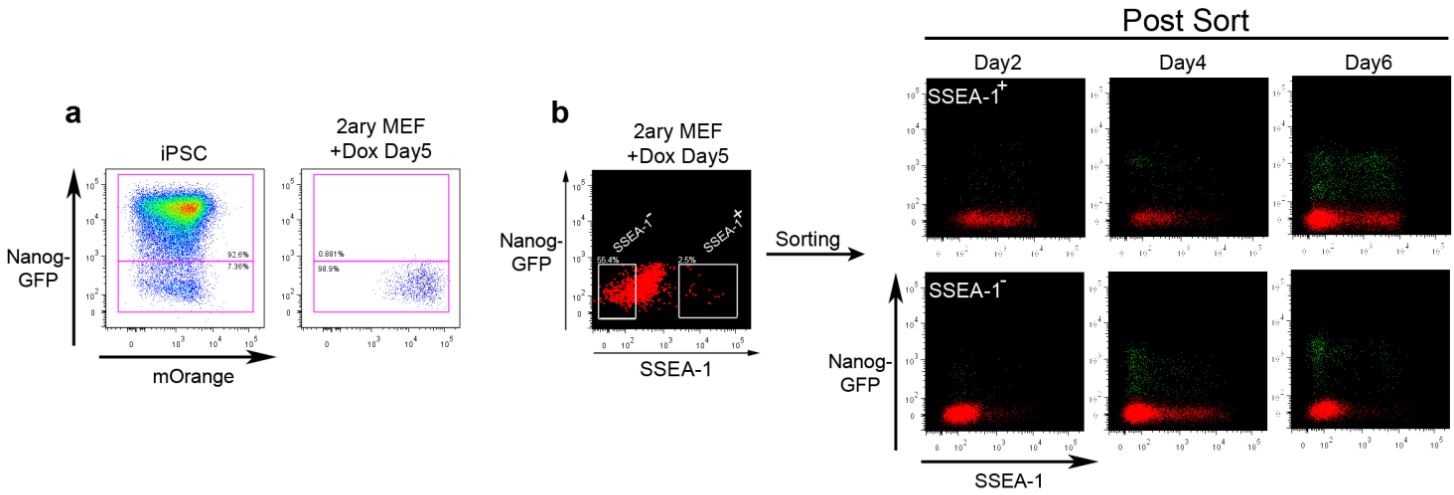
Supplementary Figure 3. A *piggyBac* (PB) secondary (2°) reprogramming system with 2A peptide-linked reprogramming cassette MKOS followed by ires mOrange. a. The reprogramming PB transposon with insulators and replicator introduced into Nanog-GFP MEFs. **b.** Upon administration of dox, induction of reprogramming factors was observed as mOrange expression. Nanog-GFP expression was observed at later time points.



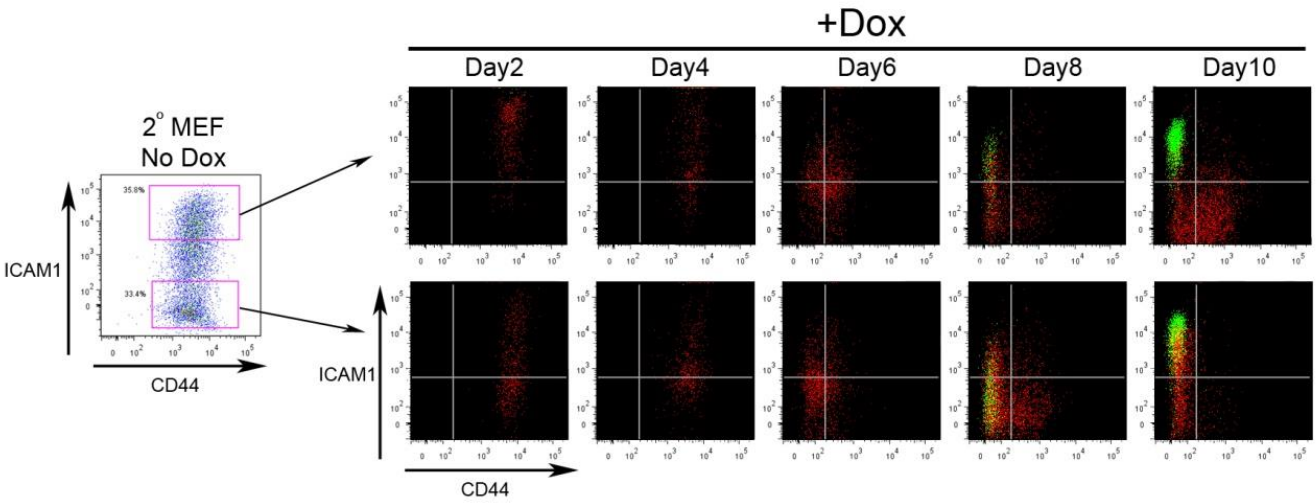
Supplementary Figure 4. Gating strategy for secondary reprogramming. Intact cells were gated using side and forward scatter (Intact cells). Transgenic cells were gated from wild-type cells using mOrange reporter (Transgenic cells). Nanog-GFP^{+} and Nanog-GFP^{-} cells were determined based on wild-type MEFs and coloured in green and red respectively.



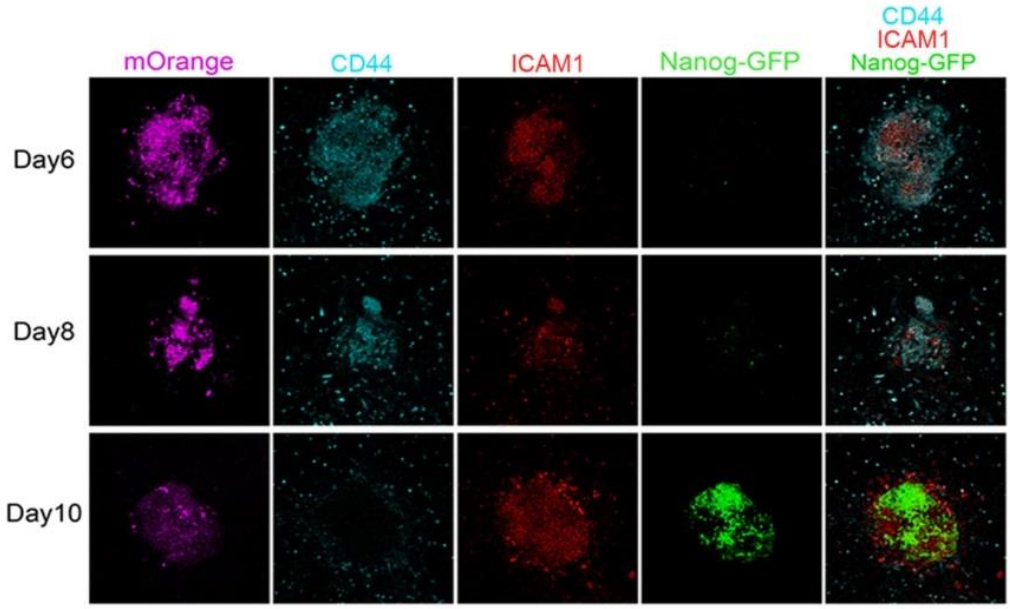
Supplementary Figure 5. Expression pattern of previously used and novel markers during 2° reprogramming. a. Nanog-GFP expression during secondary reprogramming. Note appearance of GFP^+ cells at day 6 from mOrange expressing cells. **b, c.** Thy1 negative population (b) and E-cadherin positive population (c) plateau by day 4 of 2^o reprogramming. **d.** SSEA-1 expression level in established iPSCs are heterogeneous and Nanog-GFP^+ cells appear from both SSEA-1^+ and SSEA-1^- cells. **e.** Downregulation of CD44 occurs later than that of Thy1, more closely correlating with the appearance of Nanog-GFP^+ cells. **f.** ICAM1 expression is heterogeneous in MEFs, but the majority of cells become ICAM1^- by around day 6. Re-upregulation of ICAM1 closely correlates with Nanog-GFP expression.



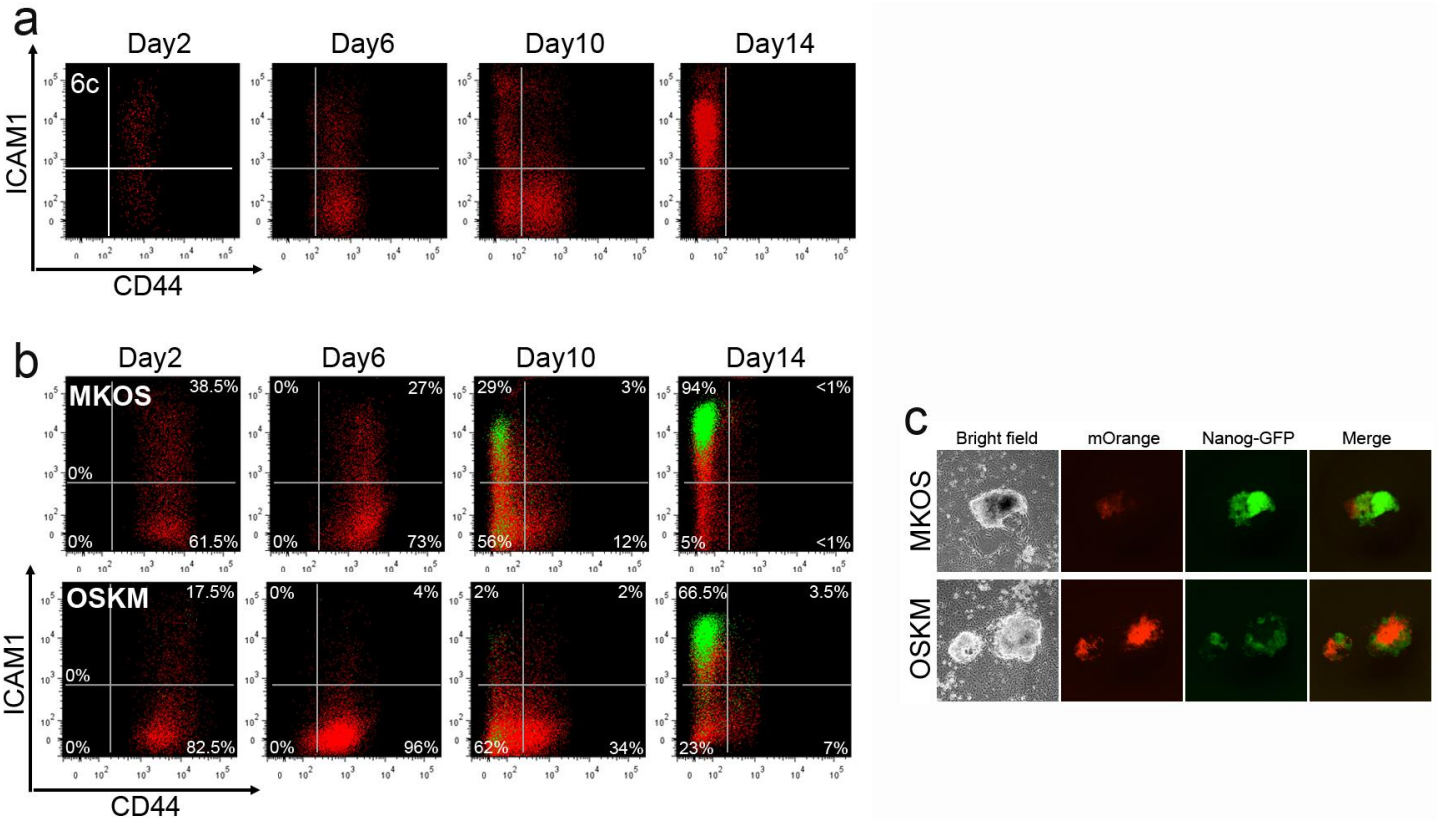
Supplementary Figure 6. SSEA-1 expression does not predict the appearance of Nanog-GFP^{+} cells. a. Gating strategy for Nanog-GFP^{-} cells at day 5 of reprogramming. **b.** Sorting strategy for Nanog-GFP^{-}, SSEA-1$^{+/-}$ cells at day 5 of reprogramming. Cells were isolated, replated in reprogramming conditions and reanalyzed every 48hours. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.



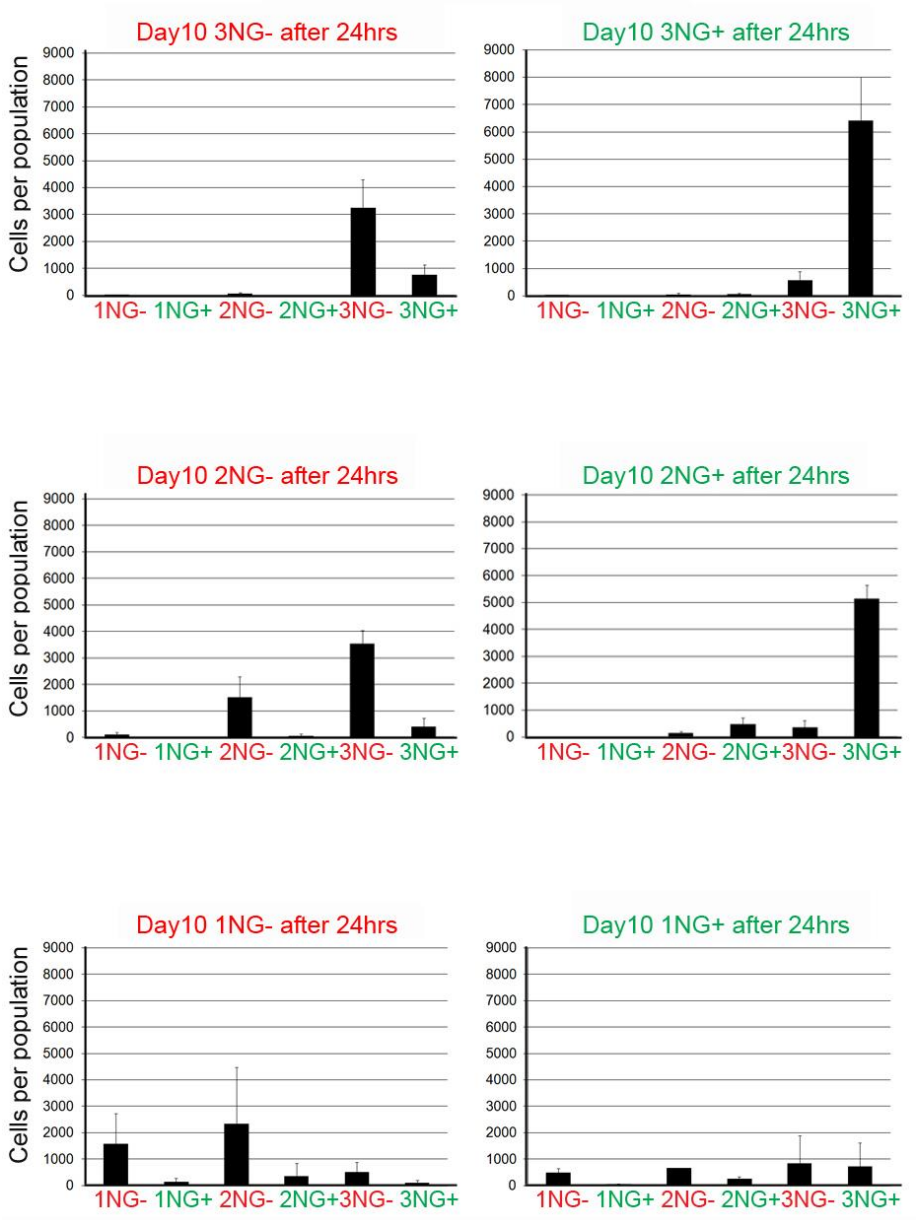
Supplementary Figure 7. Similar reprogramming kinetics with ICAM1$^{+/-}$ MEFs. ICAM1^{+} and ICAM1^{-} secondary MEF were sorted before initiating reprogramming. CD44, ICAM1, Nanog-GFP expression was monitored every two days during reprogramming. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.



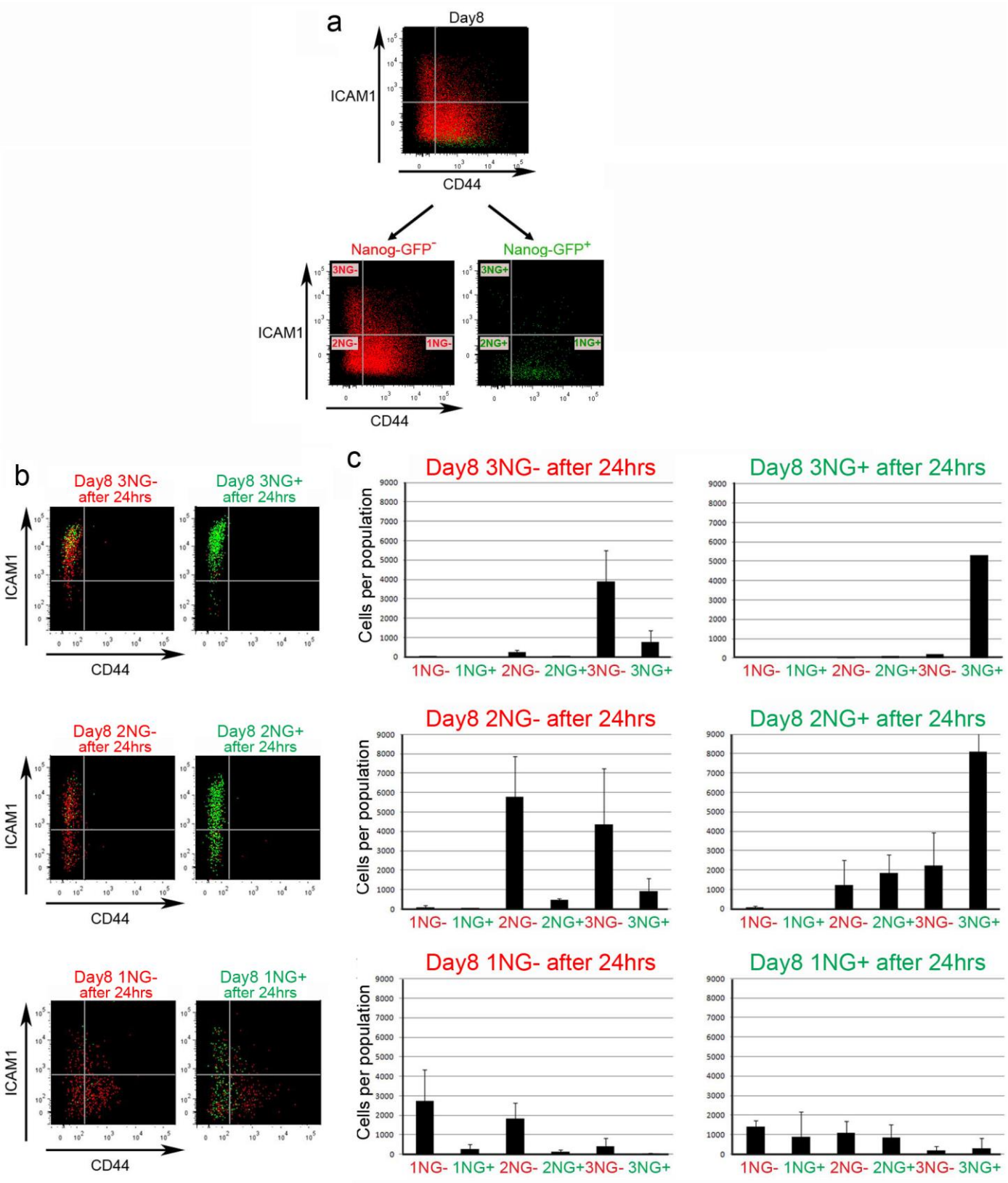
Supplementary Figure 8. Immunofluorescence for CD44 and ICAM1 at day 6, 8 and 10 after reprogramming initiation. Cells in a single colony have distinct CD44, ICAM1, Nanog-GFP expression, indicating clonal analysis is not sufficient to isolate cells in similar stages. Note expression of mOrange tends to be low in colonies with Nanog-GFP^{+} cells consistent with our flow cytometry data.



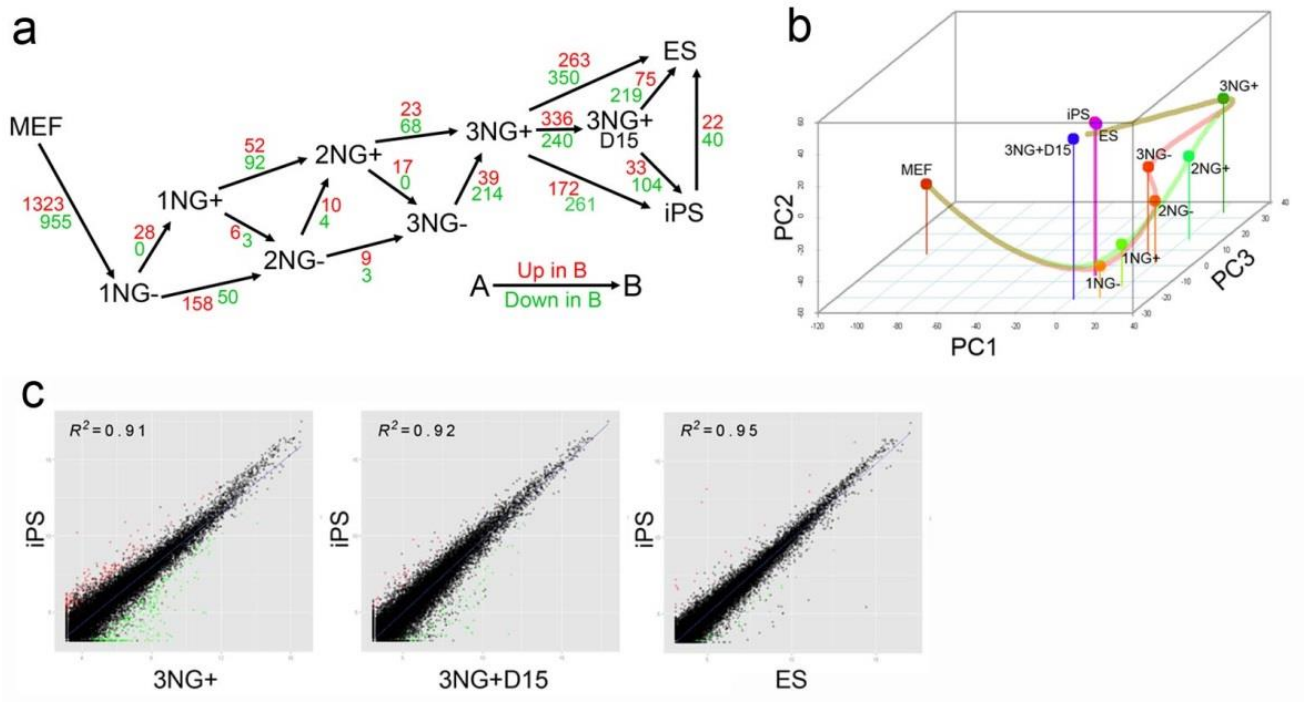
Supplementary Figure 9. The reprogramming pathway is conserved in different reprogramming systems. a, Non-polycistronic PB 2^{o} reprogramming with the 6c cell line. **b.** Primary PB reprogramming using both MKOS and OSKM polycistronic cassettes. **c.** Typical colonies arising from MKOS, OKMS primary PB reprogramming.



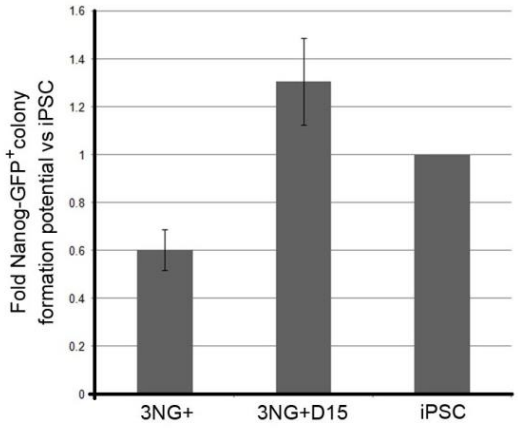
Supplementary Figure 10. Total number of cells in each gate after 24 hours for day 10 sorted populations. Each reprogramming population was sorted at day 10 and cells cultured in reprogramming conditions for 24 hours. Cells were then harvested and CD44/ICAM1/Nanog-GFP expression was re-analysed. Total cell numbers found in each gate were plotted. This data highlighted the rapid expansion of cells once they entered the 3NG+ gate. The error bars represent the standard deviation of three independent experiments.



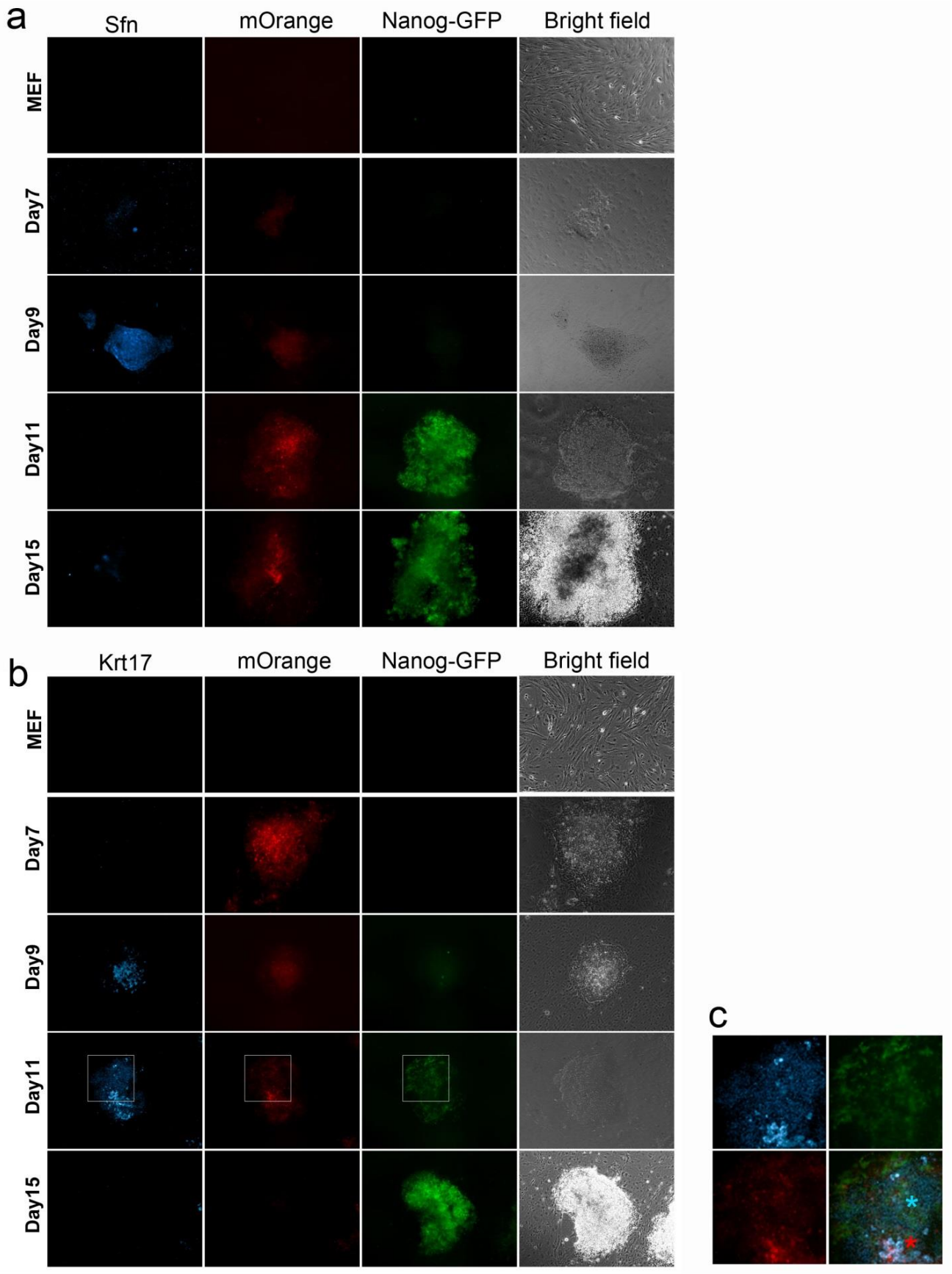
Supplementary Figure 11. Behavior of day8 sorted subpopulations are similar to that of day 10 subpopulations. a. Sorting strategy at day 8 of reprogramming. **b.** Each subpopulation sorted at day8 were replated in reprogramming conditions, and reanalyzed after 24 hours. **c.** Total cell numbers in each gate after 24 hour analysis for each sorted population. The error bars represent the standard deviation of three independent experiments.



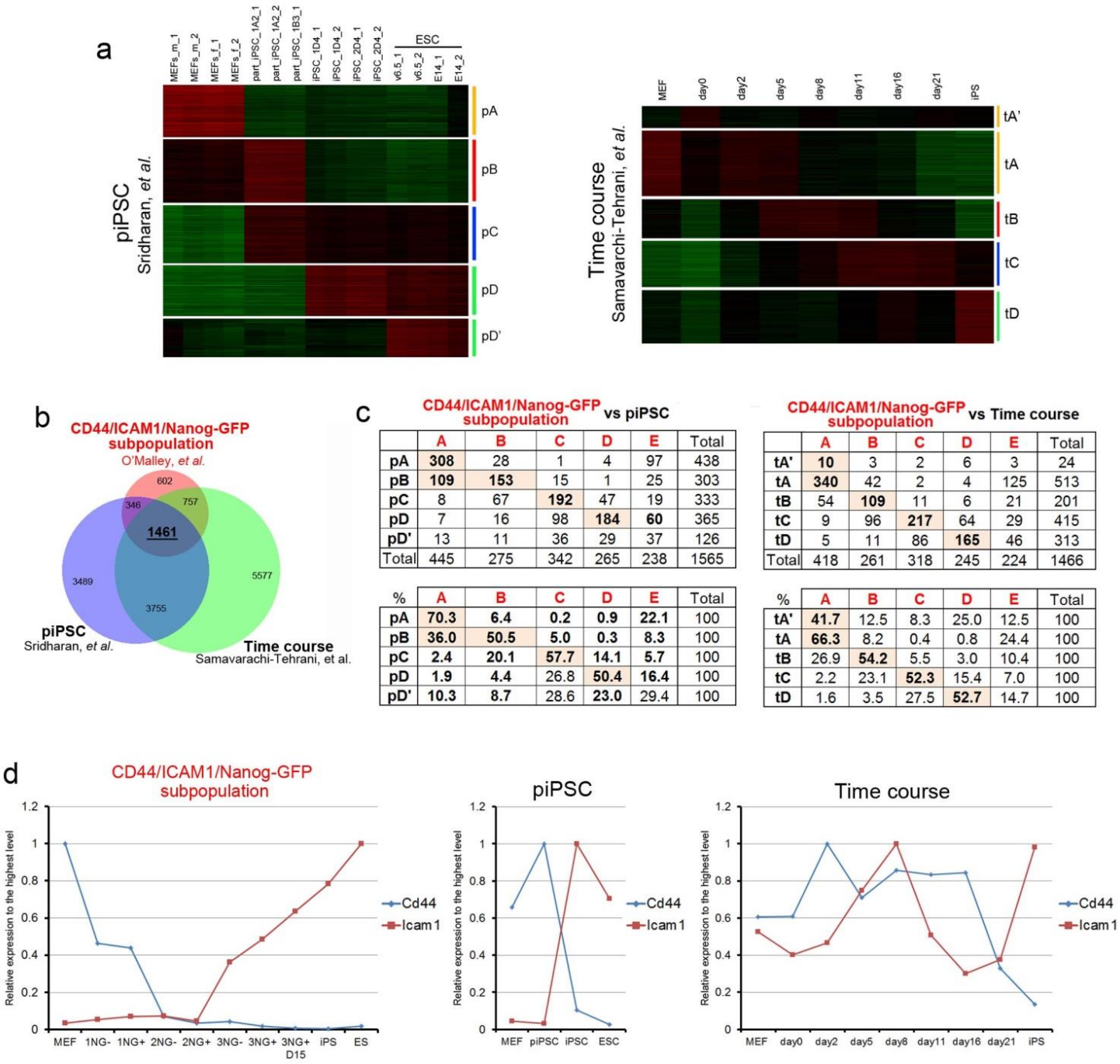
Supplementary Figure 12. Differentially expressed genes (DEGs) and Principal Component Analysis of reprogramming intermediates. a. Number of DEGs between samples identified via both edgeR and DESeq are indicated with arrows as shown. P-values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. **b.** Using these DEGs (total 3,171), Principal Components Analysis (PCA) was performed. The green and red lines connecting the samples are based on the result that 3NG- cells were frequently produced by 2NG- and 1NG- cells (Figure 3d), while 2NG+ cells eventually appeared from 1NG+ cells (Figure 3b). **c.** Comparison of gene expression profiles between iPSCs and 3NG+, 3NG+D15, ESCs. Green and red color represents up- or down-regulated genes identified by both edgeR and DESeq.



Supplementary Figure 13. Nanog-GFP Colony formation potential of 3NG+ cells in the absence of dox. 3NG+ cells sorted at day10 (3NG+) and day15 (3NG+D15) of reprogramming were plated at clonal density on feeders in the absence of dox. The number of Nanog-GFP$^+$ colonies was counted after 10 days. 3NG+ cells from day10 post transgene induction showed reduced ability (60%) to generate colonies compared to established iPSCs, while cells from day15 generated similar colony numbers to iPSCs. This suggests that about 40% of day10 3NG+ cells have not acquired exogenous reprogramming factor independent self-renewal capacity, but this trait can be acquired within an additional 5 days in the presence of dox. The error bars represent the standard deviation of three independent experiments.



Supplementary Figure 14. Transient up-regulation of Sfn and Krt17 during reprogramming. The highest protein level of both Sfn and Krt17 was observed around day9 post reprogramming (**a, b**). **c.** Higher magnification images of white squares in b. While Sfn protein was down-regulated before Nanog-GFP expression (a), Krt protein was detectable in Nanog-GFP expressing cells in the earlier stage, probably due to the protein stability (blue asterisk in c). Higher Krt17 expression was observed in mOrange^{high}, Nanog-GFP^{-} cells (red asterisk in c).



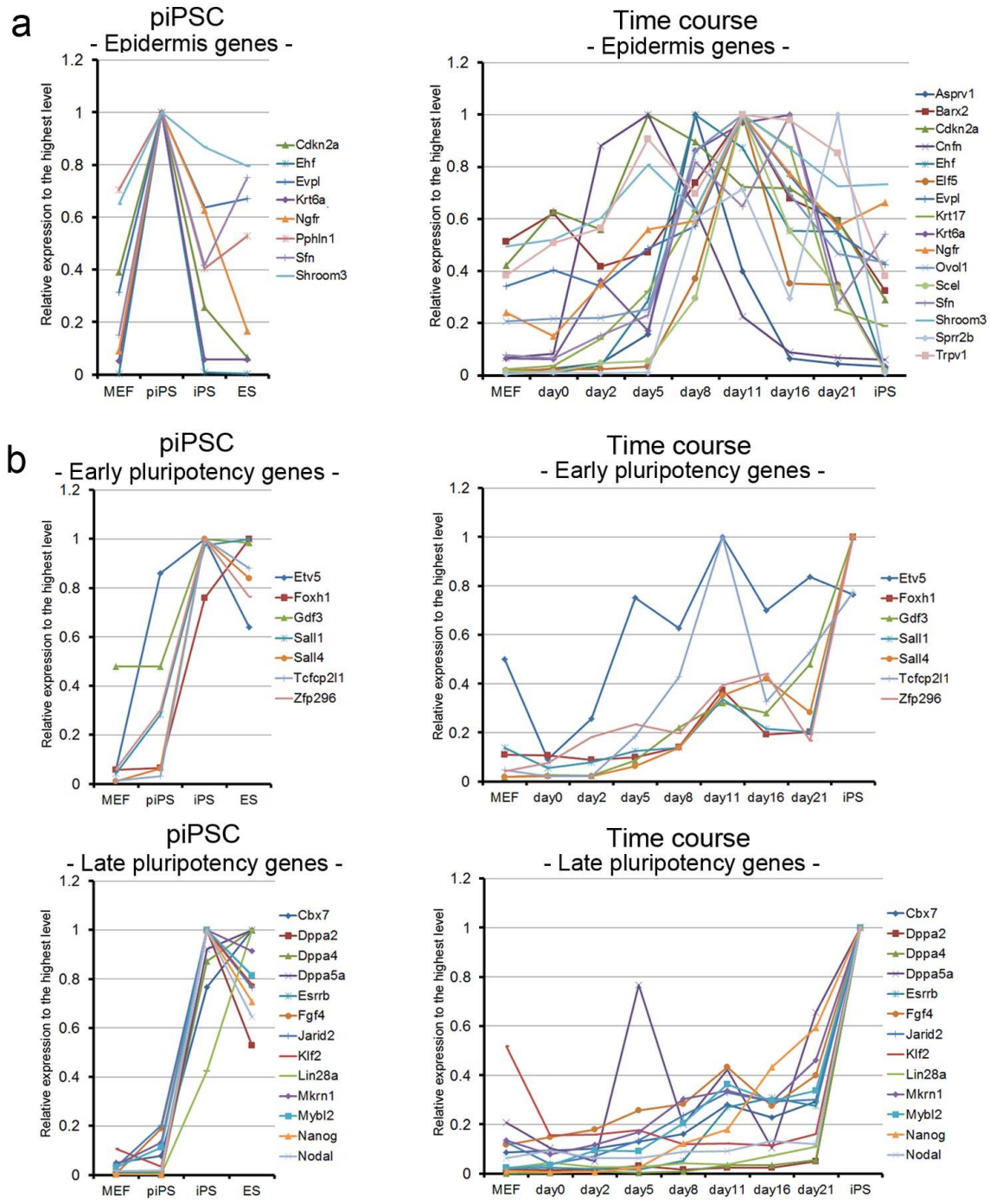
	A	B	C	D	E	Total
pA	**308**	28	1	4	97	438
pB	**109**	**153**	15	1	25	303
pC	8	67	**192**	47	19	333
pD	7	16	98	**184**	**60**	365
pD'	13	11	36	29	37	126
Total	445	275	342	265	238	1565

%	A	B	C	D	E	Total
pA	**70.3**	**6.4**	**0.2**	**0.9**	**22.1**	100
pB	**36.0**	**50.5**	**5.0**	**0.3**	**8.3**	100
pC	**2.4**	**20.1**	**57.7**	**14.1**	**5.7**	100
pD	**1.9**	**4.4**	26.8	**50.4**	**16.4**	100
pD'	**10.3**	**8.7**	28.6	**23.0**	29.4	100

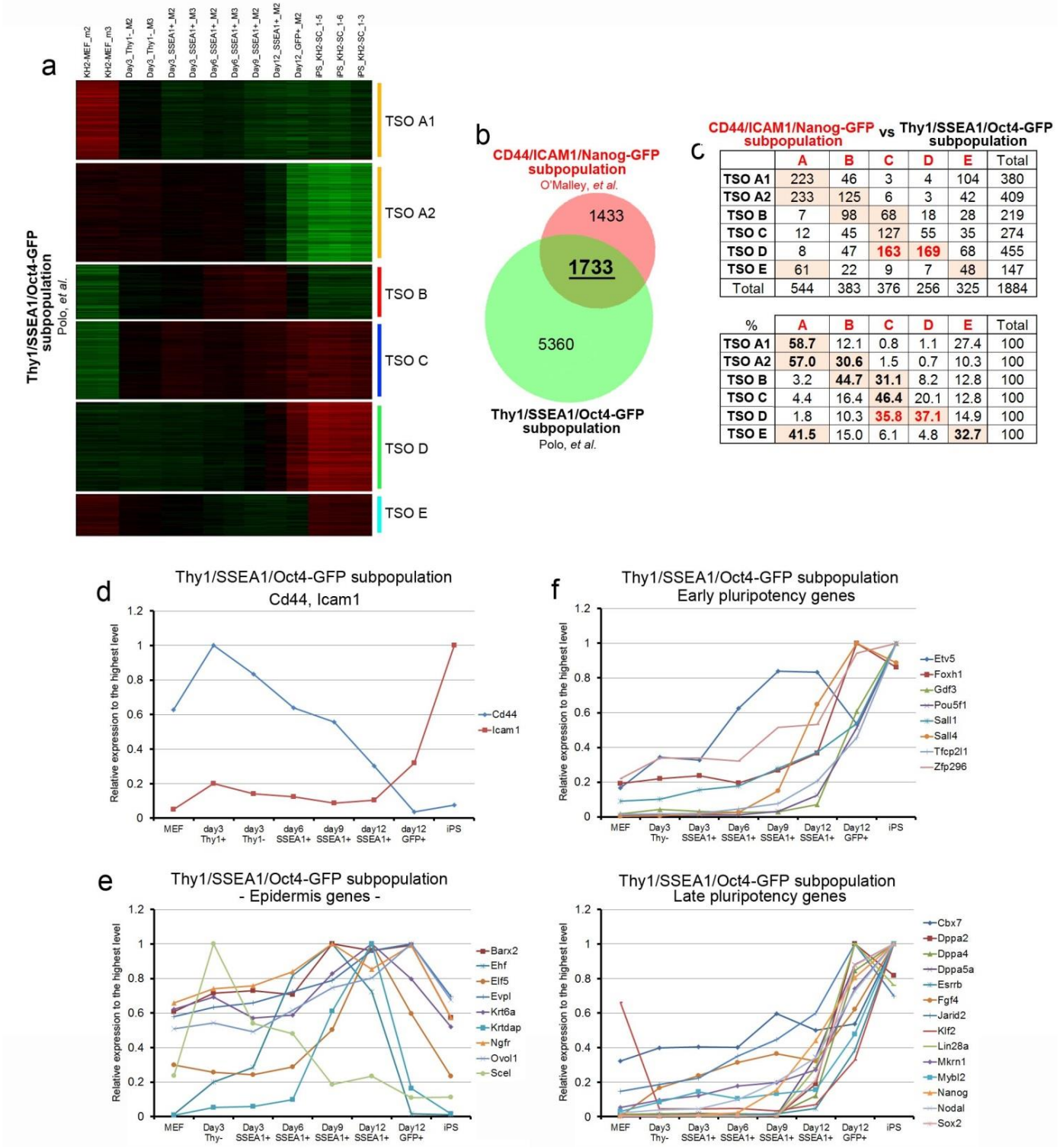
	A	B	C	D	E	Total
tA'	**10**	3	2	6	3	24
tA	**340**	42	2	4	125	513
tB	54	**109**	11	6	21	201
tC	9	96	**217**	64	29	415
tD	5	11	86	**165**	46	313
Total	418	261	318	245	224	1466

%	A	B	C	D	E	Total
tA'	**41.7**	12.5	8.3	25.0	12.5	100
tA	**66.3**	8.2	0.4	0.8	24.4	100
tB	26.9	**54.2**	5.5	3.0	10.4	100
tC	2.2	23.1	**52.3**	15.4	7.0	100
tD	1.6	3.5	27.5	**52.7**	14.7	100

Supplementary Figure 15. Comparative analysis using published reprogramming time course and partially reprogrammed cell data sets. a. Heat maps from time course[11] and partially reprogrammed iPSCs (piPSC)[27] data sets with differentially expressed genes (DEGs) with > 2.0 fold change and FDR ≤ 0.05, respectively. The DEGs were classified into 5 categories based on their expression pattern as identified from our dataset. A complete list of DEGs in each category is available in Supplementary Table 6 and 7. **b.** Venn diagram of all DEGs from this (Subpopulation), time course and piPSC data sets, highlighting 1461 DEGs common to all analyses. **c.** These 1461 genes were used to compare the overlap within each category against our subpopulation data set. Total number of genes belonging to each category is indicated in the upper tables, and their percentages against each Subpopulation A-E category are shown below. Overlap of more than 30% with Subpopulation A-E genes are highlighted in pink. Note that in the time course and piPSC microarrary datasets some genes are represented by multiple probe sets, resulting in several genes appearing in more than one category, with the total number of common genes standing at1466 and 1565, respectively. **d.** Cd44 and Icam1 expression profiles from Subpopulation, time course and piPSC data sets.



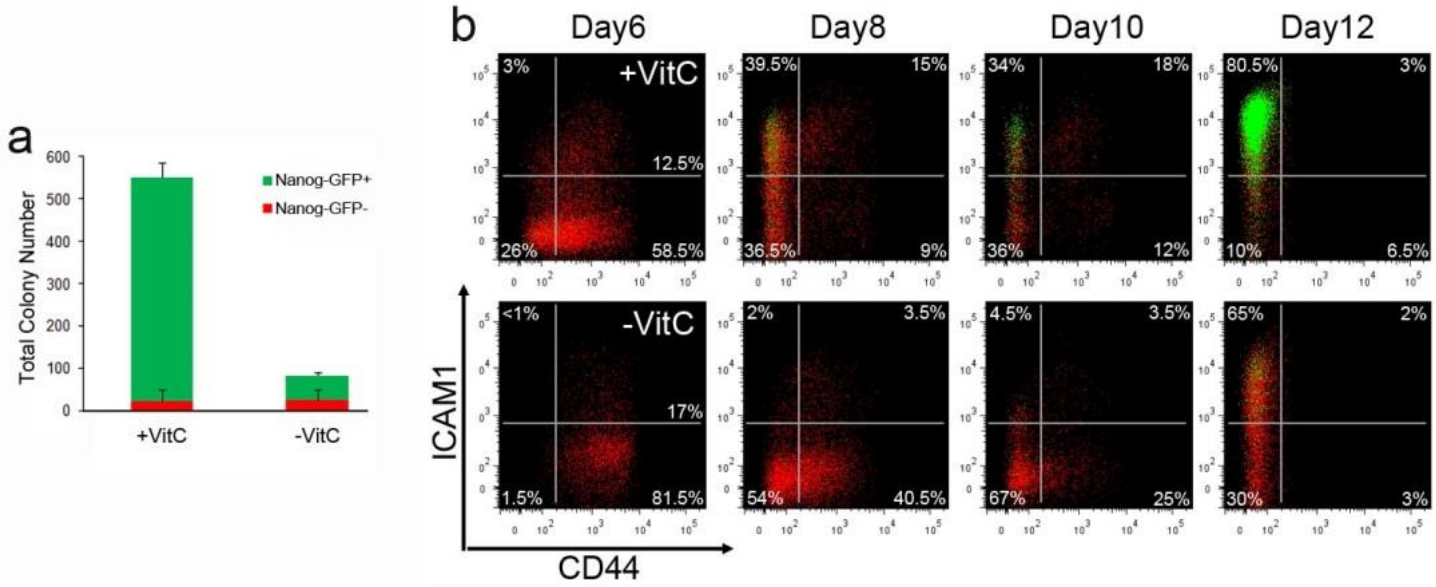
Supplementary Figure 16. Conserved transient epidermis gene up-regulation and limited resolution at the later stage of reprogramming without the sorting strategy. a. Expression pattern of epidermis-related genes, identified in Subpopulation Group B, common to time course and piPSC datasets. Data are shown as relative expression against the highest single value among the samples. Signal values are summarized in Supplementary Table 4. The data indicated transient up-regulation of epidermis-related genes is a common feature during reprogramming. **b.** Expression pattern of 19 pluripotency-related genes from time course and piPSC datasets. Genes are grouped according to their expression pattern from the Subpopulation dataset, Early (up-regulated in 1N populations) and Late (gradually up-regulated through later stages of reprogramming). It is notable that there is a large increase in the expression level of many pluripotency genes between the last time point (day21) and iPSCs in the Timecourse data, suggesting that proportionally there were few reprogrammed cells by day21. In general, such bulk population analysis may not be suitable to investigate how pluripotency genes are up-regulated. Consistent with the fact that piPSCs have a low potential to generate iPSCs, most pluripotency genes are not expressed in piPSCs.



CD44/ICAM1/Nanog-GFP subpopulation vs Thy1/SSEA1/Oct4-GFP subpopulation	A	B	C	D	E	Total
TSO A1	223	46	3	4	104	380
TSO A2	233	125	6	3	42	409
TSO B	7	98	68	18	28	219
TSO C	12	45	127	55	35	274
TSO D	8	47	163	169	68	455
TSO E	61	22	9	7	48	147
Total	544	383	376	256	325	1884

%	A	B	C	D	E	Total
TSO A1	58.7	12.1	0.8	1.1	27.4	100
TSO A2	57.0	30.6	1.5	0.7	10.3	100
TSO B	3.2	44.7	31.1	8.2	12.8	100
TSO C	4.4	16.4	46.4	20.1	12.8	100
TSO D	1.8	10.3	35.8	37.1	14.9	100
TSO E	41.5	15.0	6.1	4.8	32.7	100

Supplementary Figure 17. Comparative analysis of another marker system (Polo et al. Cell, 2012). a. Heat maps from Thy1/SSEA1/Oct4-GFP (TSO) subpopulation datasets with differentially expressed genes (DEGs). A complete list of DEGs in each category is available in Supplementary Table 8. **b.** Venn diagram highlighting 1733 DEGs common to our CD44/ICAM1/Nanog-GFP subpopulation dataset. **c.** These 1733 genes were used to compare the overlap within each category against our subpopulation dataset. Overlap of more than 30% with our A-E genes are highlighted in pink. Almost equal numbers of TSO D genes belong to our C and D groups (red), indicating CD44/ICAM1/Nanog-GFP sorting strategy gives higher resolution at the late stage of reprogramming. Note that in the TSO microarrary dataset some genes are represented by multiple probe sets, resulting in several genes appearing in more than one category, with the total number of common genes standing at 1884. **d, e, f.** Expression pattern of Cd44 and Icam1, epidermis-related genes, early and late pluripotency genes from TSO Subpopulation dataset, respectively.



Supplementary Figure 18. Secondary reprogramming in the presence of dox, Alki either with or without VitaminC (+ or -VitC). a. Both total and Nanog-GFP^{+} colony numbers decreased in the absence of VitC. **b.** Cells reprogrammed in the absence of VitC displayed delayed reprogramming kinetics.