doi:10.1038/nature12243

High-resolution analysis with novel cell-surface markers identifies routes to iPS cells

James O'Malley[1], Stavroula Skylaki[2], Kumiko A. Iwabuchi[1], Eleni Chantzoura[1], Tyson Ruetz[1], Anna Johnsson[3], Simon R. Tomlinson[1], Sten Linnarsson[3] & Keisuke Kaji[1]

The generation of induced pluripotent stem (iPS) cells presents a challenge to normal developmental processes. The low efficiency and heterogeneity of most methods have hindered understanding of the precise molecular mechanisms promoting, and roadblocks preventing, efficient reprogramming. Although several intermediate populations have been described[1–7], it has proved difficult to characterize the rare, asynchronous transition from these intermediate stages to iPS cells. The rapid expansion of minor reprogrammed cells in the heterogeneous population can also obscure investigation of relevant transition processes. Understanding the biological mechanisms essential for successful iPS cell generation requires both accurate capture of cells undergoing the reprogramming process and identification of the associated global gene expression changes. Here we demonstrate that in mouse embryonic fibroblasts, reprogramming follows an orderly sequence of stage transitions, marked by changes in the cell-surface markers CD44 and ICAM1, and a Nanog–enhanced green fluorescent protein (Nanog–eGFP) reporter. RNA-sequencing analysis of these populations demonstrates two waves of pluripotency gene upregulation, and unexpectedly, transient upregulation of several epidermis-related genes, demonstrating that reprogramming is not simply the reversal of the normal developmental processes. This novel high-resolution analysis enables the construction of a detailed reprogramming route map, and the improved understanding of the reprogramming process will lead to new reprogramming strategies.

Several reports have suggested that reprogramming progresses in an ordered manner[3,5,6,8–10]. To identify markers whose expression changed concurrent with pluripotency gene expression, we performed time course microarray analysis using a piggyBac transposon-based secondary reprogramming system[3,11] (Supplementary Fig. 2a). Of a number of candidate cell-surface markers, *Cd44* and *Icam1* (also known as *CD54*) demonstrated the most dynamic expression changes throughout secondary mouse embryonic fibroblast (MEF) reprogramming (Supplementary Fig. 2b). For further investigation, we generated an efficient secondary reprogramming system in which doxycycline-mediated induction of the reprogramming factors could be monitored by an mOrange reporter placed after the 2A-peptide-linked reprogramming cassette *c-Myc-Klf4-Oct4-Sox2* (MKOS)[12], and endogenous *Nanog* promoter activation could be followed by expression of enhanced green fluorescent protein (eGFP)[13] (Supplementary Fig 3). Reprogramming cultures were supplemented with vitamin C and an Alk inhibitor, both of which enhance reprogramming efficiency[10,14,15]. In this secondary reprogramming system, Nanog–eGFP$^+$ cells appeared as early as day 6, and >60% of mOrange$^+$ transgene-expressing cells were found to be Nanog–eGFP$^+$ by day 12 (Supplementary Figs 4 and 5a). Most mOrange$^+$ transgene-expressing cells lost expression of Thy1 (also known as CD90) and gained E-cadherin (also known as Cdh1) expression by day 4 (Supplementary Fig. 5b, c). Expression of stage-specific embryonic antigen 1 (SSEA-1, also known as Fut4) barely changed after day 8, with a gradual gain of Nanog–eGFP$^+$ cells in both SSEA-1$^+$ and SSEA-1$^-$ cell populations (Supplementary Fig. 5d). Consistent with heterogeneous expression of SSEA1 in iPS and embryonic stem (ES) cells, it was not possible to delineate the reprogramming process accurately using SSEA-1 (Supplementary Fig. 6). By contrast, the appearance of CD44$^-$ and ICAM1$^+$ cells at later time points closely correlated with Nanog–eGFP expression (Supplementary Fig. 5e, f). Double staining for CD44 and ICAM1 revealed that a distinct series of population changes occur during reprogramming (Fig. 1). Initially, MEFs displayed high CD44 and broad ICAM1 expression, with most becoming ICAM1$^-$ by day 6, along with the appearance of a minor CD44$^-$ ICAM1$^-$ cell population. By day 8, CD44$^-$ populations appeared enriched, and at day 12 almost all cells displayed an iPS/ES-cell-like CD44$^-$ ICAM1$^+$ profile, of which more than 60% expressed Nanog–eGFP. Consistent with the observation that Nanog expression is not necessarily a sign of completed reprogramming[16], Nanog–eGFP$^+$ cells were observed even before cells obtained this iPS/ES-cell-like phenotype (CD44$^-$ ICAM1$^+$). Both ICAM1$^+$- and ICAM1$^-$-sorted MEFs demonstrated similar fluorescence-activated cell sorting (FACS) profile changes during reprogramming (Supplementary Fig. 7). Immunofluorescence for CD44 and ICAM1 revealed that reprogramming is not synchronized even within individual colonies (Supplementary Fig. 8). Secondary reprogramming of the non-polycistronic iPS cell line 6c (refs 3, 11) and primary reprogramming using MKOS and *Oct4-P2A-Sox2-T2A-Klf4-E2A-cMyc* (OSKM)[17] piggyBac transposons resulted in similar ICAM1 and CD44 profile changes, indicating their suitability for use in other systems and contexts (Supplementary Fig. 9). These findings demonstrated the asynchronous but stepwise manner of reprogramming, and highlighted the potential usefulness of CD44 and ICAM1 to isolate intermediate reprogramming subpopulations.

Next, we aimed to confirm that the observed CD44/ICAM1 profile changes reflected the transition of individual cells from one stage to the next, and not merely the loss of one major population and expansion of another minor population. CD44$^+$ ICAM1$^-$ (gate 1), CD44$^-$ ICAM1$^-$ (gate 2) and CD44$^-$ ICAM1$^+$ (gate 3) cell populations, either Nanog–eGFP$^+$ (that is, 1NG$^+$, 2 NG$^+$ and 3NG$^+$) or Nanog–eGFP$^-$ (1NG$^-$, 2NG$^-$ and 3NG$^-$), were isolated by cell-sorting at day 10 of reprogramming and re-plated in reprogramming conditions (Fig. 2a). After 3 days, both NG$^+$ and NG$^-$ cells progressed in the order of gates 1 to 2 to 3 (Fig. 2b). This progression correlated well with increased Nanog–eGFP$^+$ colony-forming potential (c.f.p.), with 3NG$^+$ cells displaying similar clonogenicity to fully reprogrammed iPS cells (Fig. 2c). Of cells with the same CD44/ICAM1 profile, Nanog–eGFP expression correlated with a higher c.f.p. (for example, 1NG$^-$ versus 1NG$^+$).

To examine the progression of the reprogramming process more accurately, cells from each gate were sorted, and their expression of CD44/ICAM1/Nanog–eGFP was re-analysed after 24 h (Fig. 2d). On the basis of total cell numbers in each gate after 24 h (Supplementary Fig. 10), we generated a reprogramming route map representing differences in the

[1]MRC Centre for Regenerative Medicine, University of Edinburgh, Edinburgh BioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK. [2]Stem Cell Dynamics Research Unit, Helmholtz Center Munich, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [3]Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Scheeles väg 1, SE-171 77 Stockholm, Sweden.

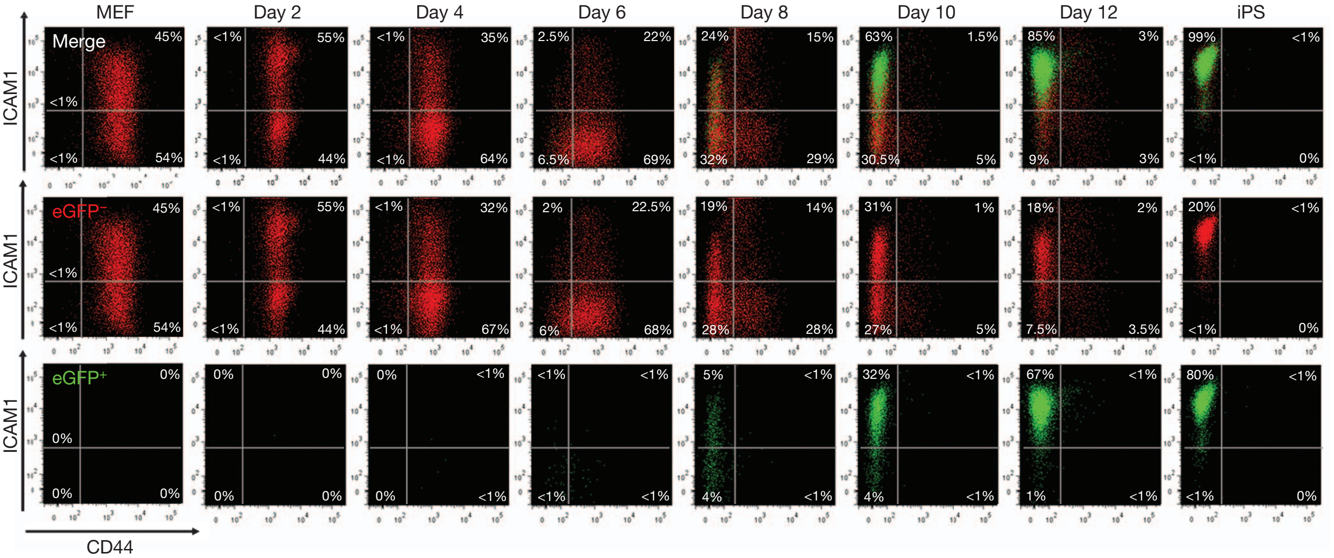


Figure 1 | FACS analysis during secondary reprogramming of MEFs with CD44/ICAM1 double staining. Loss of CD44 expression was rapidly followed by ICAM1 upregulation and Nanog–eGFP expression. By day 12, most cells displayed an ICAM$^+$/CD44$^-$ ES-cell-like profile. Red denotes Nanog–eGFP$^-$ cells; green denotes Nanog–eGFP$^+$ cells.

efficiency of these stage transitions and in Nanog–eGFP$^+$ c.f.p. (Fig. 2e). Similar results were obtained when each subpopulation was sorted at day 8 (Supplementary Fig. 11). This analysis revealed that reaching a Nanog–eGFP$^+$ state is a rate-limiting step—as few cells overcame this barrier in the 24 h assay—and those that do so reprogram more efficiently than their Nanog–eGFP$^-$ counterparts, consistent with the role of *Nanog* as an accelerator of reprogramming and the gateway to pluripotency[18,19].

To determine global gene expression changes during these stage transitions, we carried out RNA-sequencing analysis using a highly multiplexed sample bar-coding system[20–26] (see Methods and Supplementary Table 1). Hierarchical clustering using the complete list of differentially expressed genes (DEGs) revealed four major branches: (1) MEFs; (2) 1NG$^{-/+}$ and 2NG$^-$; (3) 2NG$^{-/+}$ and 3NG$^{-/+}$; and (4) 3NG$^+$ sorted at day 15 (3NG$^+$D15), iPS and ES cells (Fig. 3a). There was a prominent gene expression difference between 3NG$^+$ and 3NG$^+$D15 cells, with the latter being more similar to iPS and ES cells (Fig. 3a and Supplementary Fig. 12), possibly reflecting the observed difference in the c.f.p. in the absence of doxycycline (Supplementary

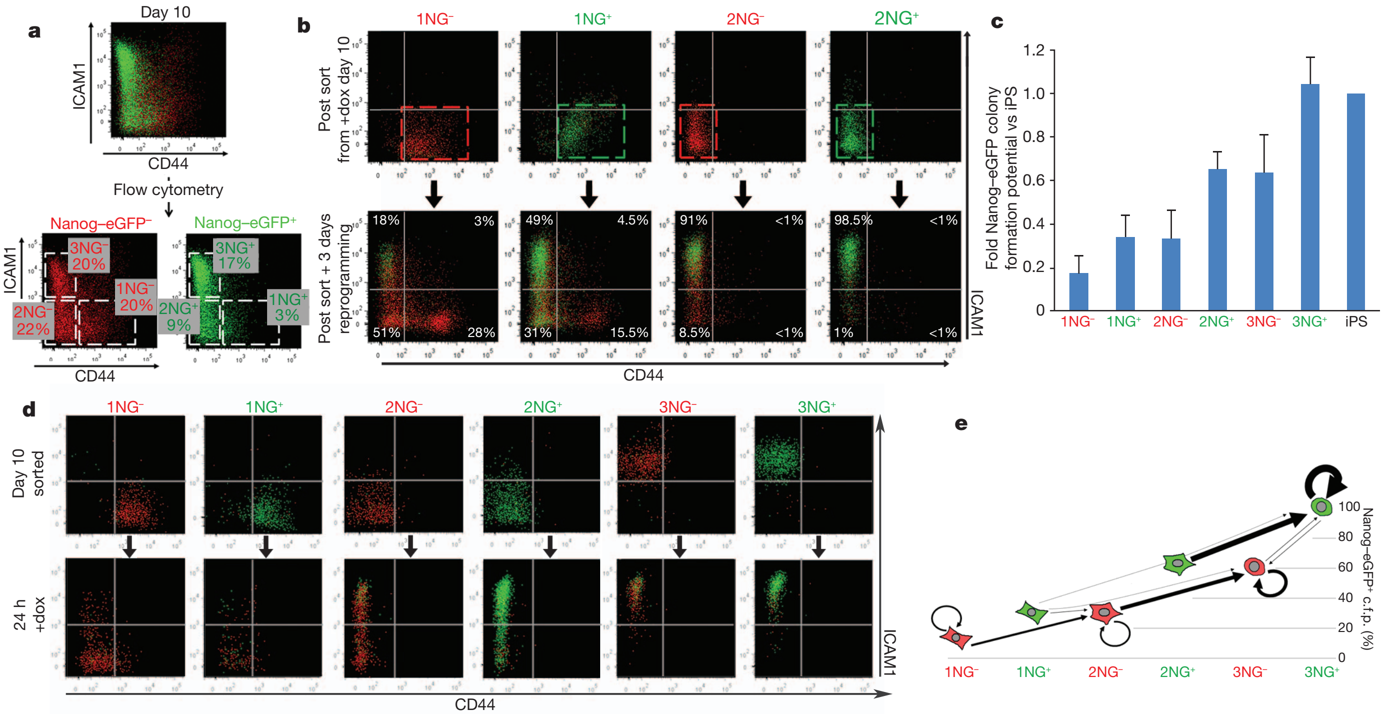
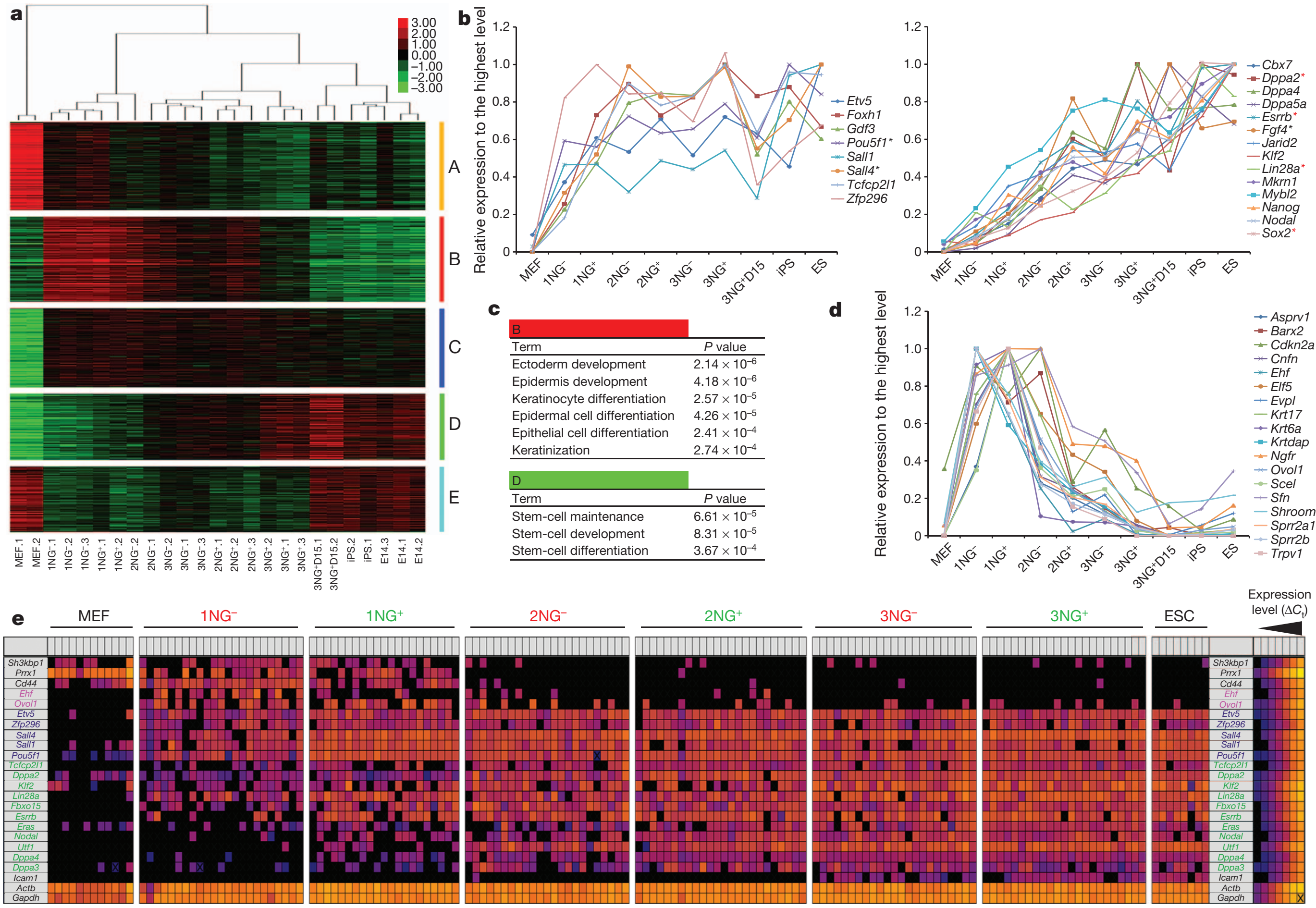


Figure 2 | CD44/ICAM1 subpopulations represent distinct stages of reprogramming. **a**, Nanog–eGFP$^+$ (NG$^+$) and Nanog–eGFP$^-$ (NG$^-$) cells were subdivided into CD44$^+$ ICAM1$^-$ (gate 1), CD44$^-$ ICAM1$^-$ (gate 2) and CD44$^-$ ICAM1$^+$ (gate 3) populations at day 10 of reprogramming. **b**, FACS analysis of sorted subpopulations after a 3-day culture in the presence of doxycycline (dox). **c**, Relative probability to generate Nanog–eGFP$^+$ iPS cell colonies from each subpopulation compared to fully reprogrammed iPS cells. Error bars represent s.d., $n = 3$. **d**, Expression of CD44, ICAM1 and Nanog–eGFP was re-analysed 24 h after sorting. **e**, Major transitions (>500 cells) of each population within 24 h. The *y* axis indicates relative c.f.p. after a further 10 days. Arrow size reflects relative cell numbers.



B	
Term	*P* value
Ectoderm development	2.14×10^{-6}
Epidermis development	4.18×10^{-6}
Keratinocyte differentiation	2.57×10^{-5}
Epidermal cell differentiation	4.26×10^{-5}
Epithelial cell differentiation	2.41×10^{-4}
Keratinization	2.74×10^{-4}

D	
Term	*P* value
Stem-cell maintenance	6.61×10^{-5}
Stem-cell development	8.31×10^{-5}
Stem-cell differentiation	3.67×10^{-4}

Figure 3 | Global gene expression changes during the stage transition. **a**, Hierarchical clustering of samples with DEGs and expression heat map. Groups A–E represent different expression patterns. **b**, Early (left) and late (right) upregulation of pluripotency-related genes. Black and red asterisks indicate early and late pluripotency genes, respectively, previously identified by single-cell quantitative PCR (qPCR)[5]. **c**, Epidermal and stem-cell gene enrichment in gene list B and D, respectively. **d**, Transient upregulation of 18 epidermis/keratinocyte-related genes during reprogramming. **e**, Single-cell gene expression analysis. Each square represents one reaction chamber from one cell. Colour corresponds to ΔC_t value, as shown in the legend.

Fig. 13). The DEGs between these two populations may be involved in the establishment of an exogenous-factor-independent self-renewal state. Principal component analysis clearly distinguished $2NG^+$ from $3NG^-$ cells, consistent with the higher probability of the former to reach the $3NG^+$ state within 24 h (Supplementary Figs 10 and 12b). DEGs could be classified into five distinct expression pattern groups (A–E) (Fig. 3a and Supplementary Tables 2 and 3). Group A contained readily downregulated fibroblast-related genes. Group D comprised factors gradually upregulated towards iPS cells, in which ES cell genes were highly enriched ($P \leq 0.000367$) (Fig. 3c). However group C, which contained genes upregulated at early stages and maintained throughout reprogramming, also included some pluripotency-related factors. To extend this finding, we examined the expression pattern of 22 pluripotency-related genes in our data set[27,28]. Interestingly, 8 pluripotency genes, including endogenous *Oct4* (also known as *Pou5f1*), were already upregulated at the $1NG^+/2NG^-$ stages to the level found in $3NG^+$ cells (Fig. 3b, left), whereas 14 pluripotency genes were more gradually upregulated in the later stage reprogramming populations (Fig. 3b, right, and Supplementary Table 4). This early and late pluripotency gene upregulation was confirmed at the single cell level[5] (Fig. 3e), highlighting the high resolution of the CD44/ICAM1 sorting system.

We also identified two additional gene expression patterns displaying transient upregulation (group B) or downregulation (group E) exclusively in the intermediate stages of reprogramming. This finding indicates that reprogramming from MEFs to iPS cells is not simply the loss of MEF genes and gain of ES cell genes. Gene Ontology analysis revealed that genes related to ectoderm/epidermis development and keratinocyte differentiation were highly enriched in group B ($P \leq 0.000274$) (Fig. 3c, d and Supplementary Tables 3–5). Although SFN and KRT17 were barely detectable by immunofluorescence in MEFs and iPS cells, transient upregulation was observed in the intermediate stages of reprogramming (Supplementary Fig. 14). Single-cell PCR confirmed the co-expression of epidermis genes (*Ehf* and *Ovol1*) with early pluripotency genes in the $1NG^{-/+}$ stage (Fig. 3e). Consistent with our data, analysis of three published microarray data sets incorporating partially reprogrammed iPS cells[1], a time course experiment[3] and a subpopulation analysis with Thy1, SSEA-1 and Oct4–eGFP (ref. 6) confirmed transient epidermal gene expression during reprogramming (Supplementary Figs 15–17 and Supplementary Tables 6–8). Partially reprogrammed cells from B cells also displayed similar epidermis gene expression[4], whereas two factor-reprogramming (Oct4 and Sox2) of MEFs did not[29]. Therefore, this intermediate state could be a consequence of the use of Klf4 that is important for efficient reprogramming, and demonstrates that the reprogramming process is not simply a reversion of normal differentiation (summarized in Supplementary Fig. 1). It would be intriguing to investigate whether similar transient gene expression changes can be seen in reprogramming of ectoderm or endoderm lineages. Downregulation of these epidermis genes coincided with upregulation of 'late' pluripotency genes. Future examination of this rapid switch in gene expression may provide a new insight into the molecular mechanism of reprogramming.

The integrative data analysis described above demonstrated that this CD44/ICAM1/Nanog–eGFP marker system could uniquely provide high-resolution information during late pluripotency gene upregulation, enabling the discrimination of 'reprogramming' from 'expansion of reprogrammed cells' (Fig. 3b and Supplementary Figs 16b and 17f). This system also refines investigation of the kinetics of reprogramming. It has recently been shown that vitamin C increases reprogramming efficiency by facilitating histone 3 Lys 9 (H3K9) demethylation[7], and that reprogramming factors fail to bind trimethylated H3K9-rich regions in the initial stages of reprogramming[30]. We carried out reprogramming in the absence of vitamin C and observed not only a decrease in the iPS cell colony number, but also a marked delay in the transition from one stage of reprogramming to the next (Supplementary Fig. 18). Similar analyses can be performed using our marker system to investigate the mechanism of action of other factors that alter reprogramming efficiency. Isolation and analysis of subpopulations affected by these factors could reveal the downstream genes specifically involved in, and required for, successful reprogramming. Further studies using this high-resolution analysis system have the potential to make a considerable contribution towards revealing the molecular mechanisms of reprogramming.

METHODS SUMMARY

The vector PB-TAP IRI 2LMKOSimO, a modified version of polycistronic reprogramming vector pCAG2LMKOSimO (ref. 12), containing insulator and replicator sequences and driven by the tetO$_2$ promoter, was constructed as described in the Methods. This vector was used to generate iPS cell line D6s4B5 from reverse tetracycline transactivator (rtTA)-expressing MEFs carrying a Nanog–eGFP reporter[13]. D6s4B5 iPS cells were used to generate chimaeric embryos from which MEFs were isolated at embryonic day 12.5. Transgenic MEFs were cultured in doxycycline (300 ng ml^{-1}), vitamin C (10 μg ml^{-1}) and Alk inhibitor (500 nM), and collected for flow cytometry analysis (BD Fortessa), carried out using antibodies for CD44 and ICAM1 every 2–3 days. Cells were sorted (BD FACS Aria II) at day 10 or 15, and replated on gelatin for analysis at 24 h, or at clonal density on irradiated MEFs for Nanog–eGFP$^+$ c.f.p. 10 days after cell sorting. All flow cytometry data were analysed using FlowJo (Tree Star). Immunofluorescence was carried out using confocal microscopy (Leica TSC SP2). RNA from sorted samples was extracted using Trizol (Invitrogen), and 10 ng total RNA was used for multiplexed RNA-sequencing[20,21]. Data were analysed using GeneProf[22], and DEGs were identified using edgeR and DESeq Bioconductor libraries[23–25]. Gene Ontology enrichment was calculated using DAVID[26].

Full Methods and any associated references are available in the online version of the paper.

Received 19 September 2012; accepted 3 May 2013.
Published online 2 June 2013.

1. Sridharan, R. *et al.* Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136,** 364–377 (2009).
2. Golipour, A. *et al.* A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. *Cell Stem Cell* **11,** 769–782 (2012).
3. Samavarchi-Tehrani, P. *et al.* Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell Stem Cell* **7,** 64–77 (2010).
4. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).
5. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150,** 1209–1222 (2012).
6. Polo, J. M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151,** 1617–1632 (2012).
7. Chen, J. *et al.* H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Genet.* **45,** 34–42 (2013).
8. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell* **2,** 230–240 (2008).
9. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2,** 151–159 (2008).
10. Li, R. *et al.* A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* **7,** 51–63 (2010).
11. Woltjen, K. *et al.* piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458,** 766–770 (2009).
12. Kaji, K. *et al.* Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **458,** 771–775 (2009).
13. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450,** 1230–1234 (2007).
14. Esteban, M. A. *et al.* Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. *Cell Stem Cell* **6,** 71–79 (2010).
15. Maherali, N. & Hochedlinger, K. Tgfβ signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc. *Curr. Biol.* **19,** 1718–1723 (2009).
16. Chan, E. M. *et al.* Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nature Biotechnol.* **27,** 1033–1037 (2009).
17. Carey, B. W. *et al.* Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proc. Natl Acad. Sci. USA* **106,** 157–162 (2009).
18. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).
19. Silva, J. *et al.* Nanog is the gateway to the pluripotent ground state. *Cell* **138,** 722–737 (2009).
20. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167 (2011).
21. Islam, S. *et al.* Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nature Protocols* **7,** 813–828 (2012).
22. Halbritter, F., Vaidya, H. J. & Tomlinson, S. R. GeneProf: analysis of high-throughput sequencing experiments. *Nature Methods* **9,** 7–8 (2012).
23. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
24. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
25. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
26. Huang, d. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
27. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–1061 (2008).
28. Xu, H., Lemischka, I. R. & Ma'ayan, A. SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. *BMC Syst. Biol.* **4,** 173 (2010).
29. Nemajerova, A., Kim, S. Y., Petrenko, O. & Moll, U. M. Two-factor reprogramming of somatic cells to pluripotent stem cells reveals partial functional redundancy of Sox2 and Klf4. *Cell Death Differ.* **19,** 1268–1276 (2012).
30. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151,** 994–1004 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Nagy and K. Woltjen for providing the 6c iPS cell line, I. Chambers for providing TNG mice, S. Monard and O. Rodrigues for assistance with flow cytometry, and T. Kunath, T. Burdon, S. Lowell and N. Festuccia for discussions and comments on the manuscript. We also thank L. Robertson for technical assistance, and Biomed unit staff for mouse husbandry. This work was supported by ERC grants ROADTOIPS (no. 261075) and BRAINCELL (no. 261063), and the Anne Rowling Regenerative Neurology Clinic. J.O.'M. and T.R. are funded by an MRC PhD Studentship and a Darwin Trust of Edinburgh Scholarship, respectively.

Author Contributions J.O.'M. designed and performed flow cytometry analysis and sorting experiments, prepared RNA for sequencing, carried out immunofluorescence imaging, and collected, analysed and interpreted data, and wrote the manuscript. S.S. analysed RNA-sequencing and published microarray data sets. K.A.I. carried out single-cell PCR analysis. E.C. performed primary reprogramming and FACS analysis. T.R. carried out immunofluorescence and confocal imaging. S.R.T. performed microarray analysis to identify cell-surface marker candidates. A.J. and S.L. performed multiplexed RNA-sequencing and collected data. K.K. conceived the study, identified the surface markers, generated the D6s4B5 iPS cell line, analysed RNA-sequencing data, supervised experiment design and data interpretation, and wrote the manuscript.

Author Information RNA-sequencing data are deposited in the ArrayExpress under accession number E-MTAB-1654. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.K. (keisuke.kaji@ed.ac.uk).

METHODS

Vector construction. The piggyBac transposon PB-TAP containing the $tetO_2$ promoter, an attR1R2 Gateway cloning cassette (Invitrogen) and rabbit β-globin poly A signal, was provided by A. Nagy. To minimize silencing of the reprogramming vector, a chicken β-globin insulator[31] was inserted into the PacI site between the piggyBac 3′-terminal repeat (3′-TR) and the $tetO_2$ promoter, and a human lamin B2 (LMB2) replicator[32] plus another chicken β-globin insulator were inserted into the EcoRV site between the rabbit β-globin poly A signal and the piggyBac 5′-TR, to generate PB-TAP IRI. The BamHI fragment containing *loxP*-flanked MKOS reprogramming cassette followed by ires-mOrange (2LMKOSimO) from pCAG2LMKOSimO (ref. 12) was inserted into a Gateway entry vector pENTR 2B (Life Technologies), to generate attP2LMKOSimO pENTR. Finally the attP2LMKOSimO cassette was Gateway-cloned into the PB-TAP IRI to yield reprogramming piggyBac transposon PB-TAP IRI attP2LMKOSimO. Similarly, reprogramming piggyBac transposon PB-TAP IRI 2LOSKMimO was generated after transferring the OSKM reprogramming cassette[17] into attP2LMKOSimO pENTR replacing the MKOS cassette. Plasmid sequences are available on request.

Generation of a primary iPS cell line D6s4B5. Embryos at 12.5 days post coitum (d.p.c.) were obtained from $Rosa^{rtTA/rtTA}$, $Nanog^{eGFP/+}$, $Col1a1^{+/+}$ mice, which were derived by crossing TNG mice[13] and B6;129-$Gt(ROSA)26Sor^{tm1(rtTA*M2)Jae}$ $Col1a1^{tm2(tetO\text{-}Pou5f1)Jae}$/J (Jackson Laboratory). The embryos were decapitated, eviscerated, dissociated with 0.25% trypsin and 0.1% EDTA, and plated in MEF medium (GMEM, 10% FBS, penicillin–streptomycin, 1× non-essential amino acids (Invitrogen), 1 mM sodium pyruvate and 0.05 mM 2-mercaptoethanol). The PB-TAP IRI attP2LMKOSimO (500 ng) and pCyL43 piggyBac transposase expression vector[33] (2 μg) were introduced into the MEFs by nucleofection (Amaxa) as before[12], and cells were cultured in ES cell medium (MEF medium supplemented with 1,000 U ml^{-1} leukaemia inhibiting factor (LIF)) in the presence of 1.0 μg ml^{-1} doxycycline (Sigma) for an initial 8 days, and thereafter 0.5 μg ml^{-1} doxycycline. Pluripotency of a clonal iPS cell line D6 was confirmed by teratoma formation, and a subclone D6s4B5 was used for secondary reprogramming. To compare CD44 and ICAM1 profiles of primary reprogramming with PB-TAP IRI attP2LMKOSimO and PB-TAP IRI 2LOSKMimO vectors, MEFs were nucleofected as above and cultured in the presence of 1.0 μg ml^{-1} doxycycline, 10 μg ml^{-1} vitamin C (Sigma) and 500 nM Alk inhibitor A 83-01 (TOCRIS Bioscience).

Secondary reprogramming. Each chimaeric embryo was collected at 12.5 d.p.c., dissociated and cultured in MEF medium. One-twentieth of the dissociated cells were exposed to doxycycline (300 ng ml^{-1}) for 2 days, and the proportion of transgenic MEFs was measured by FACS analysis of mOrange expression. For FACS time course and colony counting experiments, secondary transgenic MEFs were diluted to 5% and 30% by addition of 129 wild-type MEFs and plated in a gelatinized 6-well-plate at 1×10^5 cells per well (5,000 and 30,000 transgenic MEFs per well, respectively). For sorting experiments, MEFs were plated at 2×10^5 cells per gelatinized 100 mm plate (1×10^4 transgenic MEFs per plate). Cells were cultured in reprogramming medium, which is ES cell medium supplemented with 300 ng ml^{-1} doxycycline, 10 μg ml^{-1} vitamin C and 500 nM Alk inhibitor. Medium was changed every 2 days.

Flow cytometry and cell sorting. Cell-surface marker analysis was performed with the following eBioscience antibodies: ICAM-1-biotin (13-0541; 1/100), CD44-biotin (17-0441; 1/100), CD44- allophycocyanin (APC) (17-0441; 1/300), streptavidin-phycoerythrin (PE)-Cy7 (25-4317-82; 1/1500), SSEA-1-647 (51-8813; 1/50), E-cadherin-biotin (13-3249; 1/100), Thy1-APC (17-0902, 1/300) and CD2-biotin (13-0029; 1/100). For sorting experiments, dead cells were excluded using 4′,6-diamidino-2-phenylindole (DAPI) nucleic acid stain (Invitrogen) (0.5 ng ml^{-1}). Cells were incubated in 0.25% trypsin and 1 mM EDTA (Life Technologies) for 1–2 min at 37 °C, collected in GMEM media containing 10% FCS and counted. Staining was carried out in FACS buffer (2% FCS in PBS) at $\sim 1 \times 10^6$ cells ml^{-1} for 15–30 min at 4 °C, and followed by washing with FACS buffer, sorting and/or analysis with FACSAriaII and LSRFortessa (both BD Biosciences), respectively. Excitation laser lines and filters used for each fluorophore are summarized in Supplementary Table 9. Data were analysed using FlowJo software (Tree Star). Intact cells were identified based on forward and side light scatter, and subsequently analysed for fluorescence intensity. Additional gating was carried out as outlined in Supplementary Fig. 2. For colony formation assays, sorted cells were plated on γ-irradiated MEFs in 12-well plates at 3.5×10^3 cells per well. Nanog–$eGFP^+$ colonies were quantified 10 days after sorting. For 24 h or time-course analysis, sorted cells were plated in gelatinized 48-well plate at 1×10^4 cells per well. In both cases, cells were cultured in reprogramming medium after sorting.

Immunofluorescenceand confocal microscopy imaging. Images of cells stained with ICAM-1-biotin (1/100), CD44-APC (1/300) and streptavidin-PE-Cy7 (1/1,500) antibodies described above were captured with a confocal microscope (Leica TSC SP2) and Leica confocal software. Cells stained with anti-Krt17 (LifeSpan BioSciences) and anti-Sfn (Sigma) antibodies and anti-Rabbit IgG CF633 secondary antibody (Sigma) were imaged with a fluorescence microscopy (Olympus).

Multiplexed RNA sequencing and data analysis. RNA was isolated with TRI reagent (Sigma) following the manufacturer's instructions. RNA quality and concentration was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies). Using 10 ng RNA, reverse transcription with bar-coded primers, complementary DNA amplification, and sequencing with Illumina HiSeq 2000 were performed as previously described[20,21]. Quality control of the obtained reads and alignment to the mouse reference genome (NCBI37/mm9) were performed using the GeneProf web-based analysis suite with default parameters[22]. Gene expression read counts were exported and analysed in R to identify DEGs, using the edgeR and DESeq Bioconductor libraries[23–25]. For both methods, low expression transcripts (less than 13 reads in all samples) were filtered out, and P values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. Genes listed as DEGs by both methods in any two subpopulation comparison indicated in Supplementary Table 1 and Supplementary Fig. 12a (total 3,171) were used for further analysis. Hierarchical clustering and K-means clustering ($K = 5$) was performed using Cluster 3.0, and Java Treeview was used for visualization[34,35]. This multiplexed RNA-sequencing technology reads only the 5′ end of transcript, thus detecting only endogenous *Oct4* and *Sox2*. *Nanog* expression was detectable in Nanog–$eGFP^-$ populations owing to the reporter system. Principal components analysis was performed in R and plotted with the scatterplot3d library[36]. Gene Ontology enrichment was calculated using the DAVID functional annotation bioinformatics tool[26]. Gene Ontology term enrichment analysis was carried out with a modified Fisher exact P value. The three additional published studies[1,3,6] (GEO accession numbers GSE21757, GSE14012 and GSE42379) were analysed in a similar way. For the time course data, the analysis was performed as following: data were robust multi-array average (RMA)[37] normalized using the expression console from Affymetrix and, because no replicates were provided, fold changes between two samples were calculated in Excel. Genes with more than 1.5-fold changes were classified as DEGs. For the Plath and Polo data set, data were RMA-normalized using the 'affy' package[38] in R, and DEGs were identified using the 'limma' package[38] in R with fold change ≥ 1.5 and FDR ≤ 0.05, or fold change ≥ 1.5 where no replicates were available. Subsequently, K-means clustering of the identified DEGs was performed for all studies. Selected gene expression data are shown as the relative expression against the highest signal among the samples using an averaged signal value (reads per million) of duplicates/triplicates.

Single-cell gene expression analysis. Single-cell qPCR was performed as described previously[5] with slight modifications. In brief, 22 sets of TaqMan gene expression assays (Applied Biosystems; Supplementary Table 9) were pooled at a final concentration of 180 nM per primer set and 50 μM per probe. Individual cells were sorted directly into 10 μl RT-PreAmp Master Mix (5 μl of CellsDirect reaction mix (Invitrogen), 2.5 μl of pooled assays, 0.2 μl of SuperScript III (Invitrogen), 1.3 μl of water) using FACSAria II. Cell lysis and sequence-specific reverse transcription were performed at 50 °C for 15 min. Reverse transcriptase was inactivated by heating to 95 °C for 2 min. Subsequently, in the same tube, cDNA went through sequence-specific amplification by denaturing at 95 °C for 15 s, and annealing and amplification at 60 °C for 4 min for 22 cycles. Preamplified products were diluted fivefold with water and analysed in 48.48 dynamic arrays on a biomark system (Fluidigm) following the Fluidigm protocol. C_t values were calculated and visualized using BioMark real-time PCR analysis software (Fluidigm). Each assay was performed in replicate.

31. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev. Genet.* **7,** 703–713 (2006).
32. Fu, H. *et al.* Preventing gene silencing with human replicators. *Nature Biotechnol.* **24,** 572–576 (2006).
33. Wang, W. *et al.* Chromosomal transposition of *PiggyBac* in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **105,** 9290–9295 (2008).
34. Saldanha, A. J. Java Treeview–extensible visualization of microarray data. *Bioinformatics* **20,** 3246–3248 (2004).
35. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
36. Ligges, U. & Maechler, M. scatterplot3d — an R package for visualizing multivariate data. *J. Stat. Softw.* **8,** 1–20 (2003).
37. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
38. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of *Affymetrix GeneChip* data at the probe level. *Bioinformatics* **20,** 307–315 (2004).