nature cell biology

# Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells

Yuki Ohi[1,2,3,12], Han Qin[1,2,3], Chibo Hong[4], Laure Blouin[1,2,3], Jose M. Polo[5,6], Tingxia Guo[3,7], Zhongxia Qi[8], Sara L. Downey[4], Philip D. Manos[6,9], Derrick J. Rossi[6,9,10], Jingwei Yu[8], Matthias Hebrok[3,7], Konrad Hochedlinger[5,6], Joseph F. Costello[4], Jun S. Song[11,12,13] and Miguel Ramalho-Santos[1,2,3,13]

**Human induced pluripotent stem (iPS) cells are remarkably similar to embryonic stem (ES) cells, but recent reports indicate that there may be important differences between them. We carried out a systematic comparison of human iPS cells generated from hepatocytes (representative of endoderm), skin fibroblasts (mesoderm) and melanocytes (ectoderm). All low-passage iPS cells analysed retain a transcriptional memory of the original cells. The persistent expression of somatic genes can be partially explained by incomplete promoter DNA methylation. This epigenetic mechanism underlies a robust form of memory that can be found in iPS cells generated by multiple laboratories using different methods, including RNA transfection. Incompletely silenced genes tend to be isolated from other genes that are repressed during reprogramming, indicating that recruitment of the silencing machinery may be inefficient at isolated genes. Knockdown of the incompletely reprogrammed gene *C9orf64* (chromosome 9 open reading frame 64) reduces the efficiency of human iPS cell generation, indicating that somatic memory genes may be functionally relevant during reprogramming.**

Human iPS cells can be derived from differentiated cells by activation of key transcription factors and hold enormous promise in regenerative medicine[1]. Although iPS cells are remarkably similar to ES cells, there may be important differences between them. Human iPS cells have been suggested to be less efficient than ES cells in targeted differentiation to neural and blood lineages[2,3]. Transcriptional differences have also been described and proposed to represent a persistent memory of the original somatic cells in iPS cells[4–6]. However, it has recently been countered that the transcriptional differences observed may largely be due to laboratory-specific batch effects[7,8].

The present confusion surrounding this issue derives from the poor overlap between gene sets attributed to somatic cell memory in different studies, and from a lack of correlation between gene expression and epigenetic information. Transcriptional differences between human iPS cells and ES cells could not be explained by differences in histone modification patterns[4,7]. Recent studies have identified differences in DNA methylation between iPS and ES cells in both mouse and human cells[9–14]. Mouse iPS cells have been shown to retain a DNA methylation memory of the original somatic cell that may bias iPS cell differentiation towards lineages related to that cell[12,14]. However, the DNA methylation differences found between iPS cells and ES cells were largely not demonstrated to correlate with gene expression differences[9–14]. A further limitation stems from the fact that iPS cells generated in different laboratories by different methodologies are often used for comparison[4,5]. In addition, most human iPS cells analysed so far, including in two very recent studies of genome-wide DNA methylation[9,13], are derived

[1]Departments of Ob/Gyn and Pathology and Center for Reproductive Sciences, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143, USA. [2]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California San Francisco, 35 Medical Center Way, San Francisco, California 94143, USA. [3]Diabetes Center, University of California, San Francisco, California 94143, USA. [4]Department of Neurosurgery, Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, 1450 Third Street, San Francisco, California 94158, USA. [5]Department of Medicine, Howard Hughes Medical Institute, Cancer Center and Center for Regenerative Medicine, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6]Department of Stem Cell and Regenerative Biology, Harvard Stem Cell Institute, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Department of Medicine, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143, USA. [8]Department of Laboratory Medicine, University of California San Francisco, 185 Berry Street, San Francisco, California 94107, USA. [9]Stem Cell Program, Children's Hospital Boston, Boston, Massachusetts 02115, USA. [10]Immune Disease Institute, Program in Cellular and Molecular Medicine, Department of Pathology, Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts 02115, USA. [11]Institute for Human Genetics, Department of Epidemiology and Biostatistics, and Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, 513 Parnassus Avenue, San Francisco, California 94143, USA. [12]These authors contributed equally to this work.
[13]Correspondence should be addressed to J.S.S. or M.R-S. (e-mail: SongJ@humgen.ucsf.edu or mrsantos@diabetes.ucsf.edu)

Received 14 February 2011; accepted 16 March 2011; published online 17 April 2011; DOI: 10.1038/ncb2239

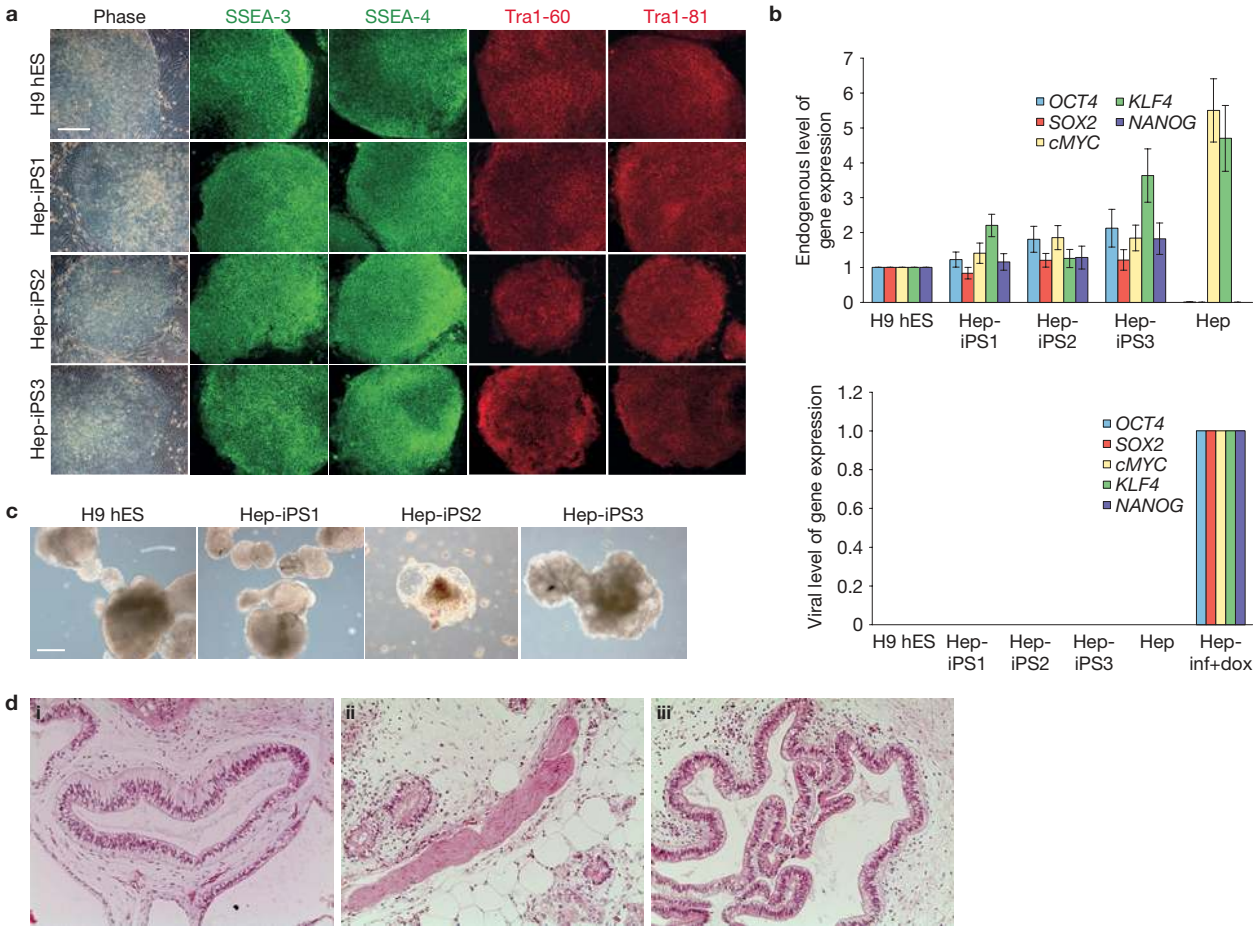


**Figure 1** Pluripotency validation for the derived Hep-iPS cells used for the microarray studies. (**a**) The three Hep-iPS clones used in this analysis showed strong, positive immunostaining for all analysed specific markers for human ES (hES) cells. SSEA, stage-specific embryonic antigen. Tra1-81, tumour rejection antigen 1-81. Scale bar, 300 μm. (**b**) All Hep-iPS clones showed high expression levels of endogenous pluripotency markers and negligible levels of transgene expression by quantitative rtPCR. Values were standardized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and ubiquitin B (Ubb), then normalized to H9 ES cells (endogenous) or 5-factor-infected hepatocytes + doxycycline (Hep-inf+dox) at 4 days (viral). Data are from triplicate reactions. Error bars represent standard deviations. (**c**) All Hep-iPS clones formed embryoid bodies *in vitro* when grown under non-attachment conditions. Shown here are d8 embryoid bodies and control ES-cell-derived embryoid bodies. Scale bar, 200 μm. (**d**) Pluripotency of the Hep-iPS cell clones was further confirmed by their ability to form teratomas *in vivo*, comprised of tissues derived from all three germ layers. (i) Neural tissue (ectoderm). (ii) Striated muscle and adipocytes (mesoderm). (iii) Gut-like epithelium (endoderm). Also see Supplementary Fig. S2 for pluripotency validation of Fib-iPS cells used for the microarray analysis. Mel-iPS cells have previously been described[17].

from fibroblasts, thus limiting the evaluation of a potential memory of the original somatic cell in iPS cells.

We report here a systematic comparison of human iPS cells generated from different somatic cell types. Importantly, all iPS cells analysed by transcriptional profiling were generated with the same methodology and analysed in parallel. Our data allow us to distinguish different types of somatic cell memory in human iPS cells, which can be partially explained by incomplete promoter DNA methylation. We find that the somatic memory gene *C9orf64* regulates the efficiency of iPS cell generation, and that incompletely silenced genes tend to be isolated from other genes destined to be silenced during reprogramming.

## RESULTS

### Generation of human iPS cells from somatic cells representative of all three embryonic germ layers

We used a doxycycline-inducible lentivirus transgene system[15,16] to generate iPS cells (Supplementary Fig. S1). To have a broad range of starting differentiated states, somatic cells representative of the three embryonic germ layers were reprogrammed to iPS cells: adult hepatocytes (Hep) for endoderm, newborn foreskin fibroblasts (Fib) for mesoderm and adult melanocytes (Mel) for ectoderm (Supplementary Fig. S1). The Mel-iPS cell lines have been previously described[17]. iPS cell pluripotency was extensively validated, including colony morphology, growth rate, marker expression, transgene independence, formation of embryoid bodies and development of teratomas[17] (Fig. 1 and Supplementary Fig. S2). Integration analysis indicates that all iPS cell lines used are independent clones (data not shown). We focused our analysis in this study on low-passage iPS cells (below passage 20), because they are expected to be more informative about the molecular mechanisms that underlie reprogramming.

### Transcriptional profiling of iPS cells and ES cells

The expression levels in Hep, Fib, Mel and the iPS cells derived from them were profiled in triplicate. In addition, three independent

well-established ES cell lines, H1, H7 and H9, and their 8-day (d8) embryoid bodies were also profiled individually. All samples were analysed using Affymetrix ST 1.0 microarrays (Supplementary Fig. S1). A hierarchical clustering of the data correctly classified the cell types as shown in Fig. 2a. The three iPS cell types clustered together with the ES cells, forming a single branch of pluripotent cell samples. Figure 2b further shows that all somatic cells underwent extensive reprogramming towards an ES cell-like transcriptional profile.

### iPS cells retain a transcriptional memory of the original somatic cell

We used the equal-variance $t$ statistic to find a global pattern of differential gene expression between iPS and ES cells. We plotted the gene expression differences between iPS cells and ES cells against the differences between the original somatic cells and ES cells and fitted locally weighted scatter plot smoothing (LOESS) regression curves to each plot (Fig. 3a and Supplementary Data S1; see Methods). We then carried out bootstrap simulations to model noise in gene expression under the assumption that iPS and ES cells are truly identical and that their differences arise from random fluctuations. The actual regression curves lie well outside the intervals of simulated curves, revealing that genes that were highly expressed in somatic cells tend to be repressed but remain higher in iPS cells when compared with ES cells, and conversely for genes expressed at low levels in somatic cells (Fig. 3b). This pattern was observed for all three types of iPS cell analysed (Fig. 3a,b and Supplementary Fig. S3a for Fib and Mel).

To find a confident set of differentially expressed genes, we used a robust statistical method, differential expression via distance synthesis (DEDS), which combines $t$-test, moderated $t$-test, fold change and significance analysis of microarrays into a summary statistic[18]. DEDS has been shown to outperform the individual statistics on spike-in data sets, and its synthesis approach also makes it robust against the limitations of individual tests[18]. At 5% false discovery rate (FDR), this analysis confirmed that a very significant proportion (~50–60%) of the genes differentially expressed between iPS cells and ES cells represent a memory of the differential expression that already existed between the original somatic cells and ES cells (Fig. 3c, upper Venn diagrams). That is, a statistically significant ($10^{-6} > P > 10^{-16}$, Fisher's exact test) number of genes that were higher in iPS cells relative to ES cells resulted from incomplete silencing during reprogramming. Similarly, a statistically significant ($10^{-9} > P > 10^{-32}$, Fisher's exact test) number of genes that were lower in iPS cells relative to ES cells were the result of incomplete reactivation during reprogramming. No statistically significant overlap was found between genes that change in opposite directions in iPS cells and somatic cells, relative to ES cells (Fig. 3c, lower Venn diagrams). Our analysis thus demonstrates that iPS cells retain a transcriptional memory of the original somatic cells.

We next examined whether transcriptional memory in iPS cells is cell-type-specific or associated with multiple differentiated states. In support of a cell-type-specific transcriptional memory, ~8 ± 2% of the genes differentially expressed between an iPS cell type and ES cells were already differentially expressed specifically in the original somatic cell (but not the other somatic cells), relative to ES cells (Supplementary Fig. S3b). However, most of the genes differentially expressed between each iPS cell type and ES cells (52 ± 5% of total) were found to also be differentially expressed in two or all three somatic cell types relative to ES cells, indicating that they may represent a memory of a general differentiated state. Finally, ~24 ± 2% of genes differentially expressed

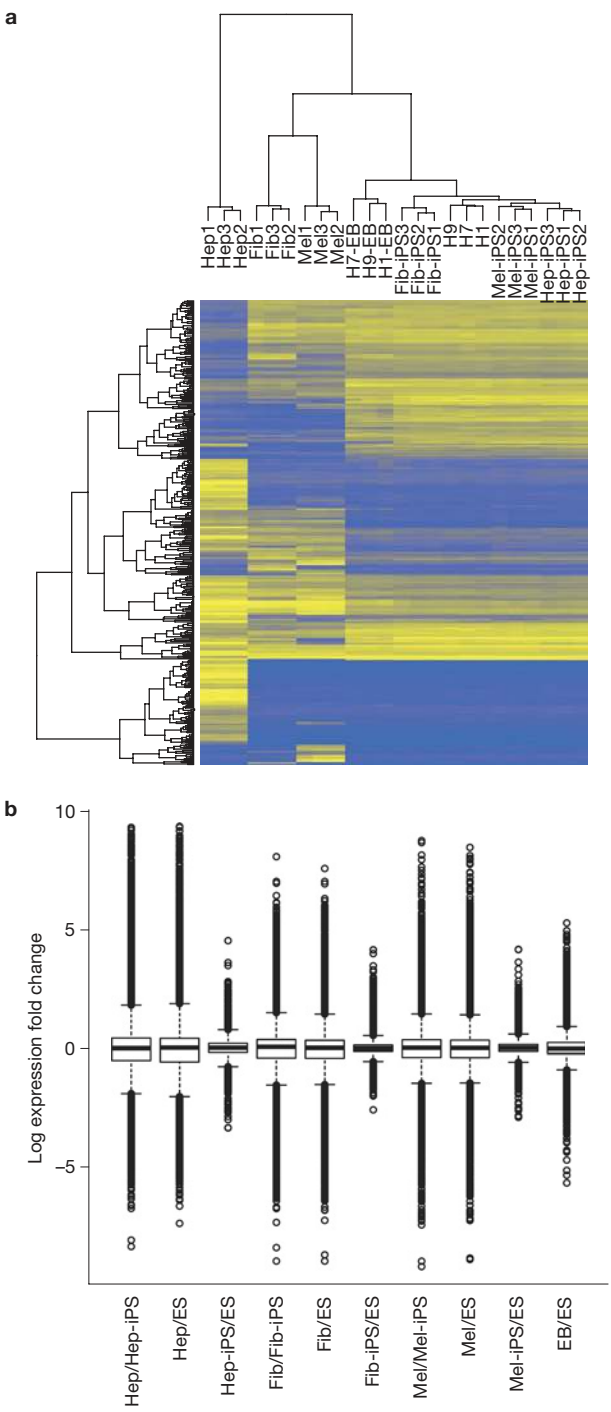


**Figure 2** Multiple cell types undergo extensive transcriptional reprogramming to the human iPS cell state. (**a**) Average-linkage hierarchical clustering of the RMA-normalized expression profiles shows that the replicate data cluster together tightly, confirming the reproducibility of the experiments, and that the somatic cells have been successfully reprogrammed. EB, embryoid bodies. (**b**) The box plot of log expression fold changes for all RefSeq genes further shows that the iPS cells have been reprogrammed to closely resemble the transcriptional profiles of ES cells. The black centre line represents the median. The upper and lower edges of the box represent the first and third quartiles, and they define the inter-quartile range. Outliers farther than 1.5 times the inter-quartile range from the box are shown as circles.

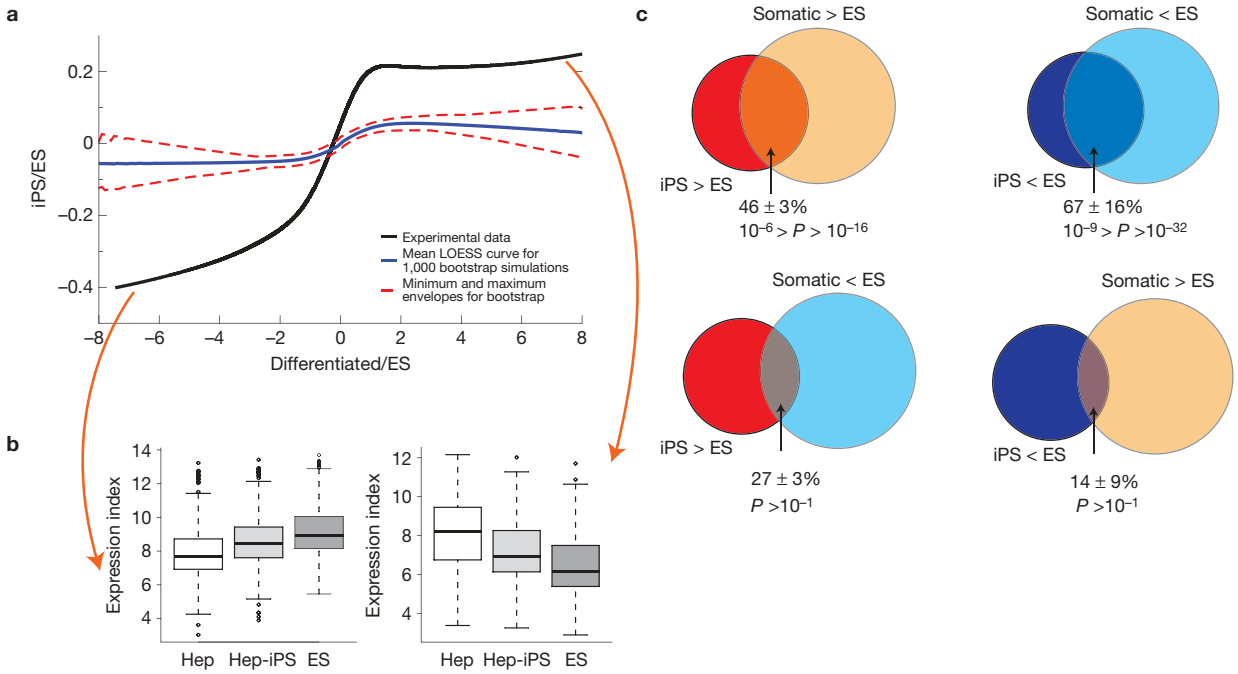


**Figure 3** iPS cells retain a transcriptional memory of the original somatic cell. (**a**) LOESS curves fitted to the scatter plots of *t*-test log *P* values for hepatocyte and hepatocyte-derived iPS cells: $-\log(P)$ and $\log(P)$ are plotted for fold changes greater than 1 and less than 1, respectively. The black line is a curve fitted to our data, and other curves are fitted to the 1,000 bootstrap simulation data sets obtained by assuming identically distributed iPS and ES cell expression levels. The black line shows clear deviation from the null hypothesis iPS = ES and thus reflects the trend that the transcriptional memory of the originating cell type is retained in low-passage iPS cells: genes that were higher (or lower) in the somatic cell than in ES cells tend to be significantly repressed (or induced) during reprogramming, but nevertheless remain higher (or lower) in iPS cells than in ES cells. (**b**) Box plots of expression levels for 191 genes that are higher in both iPS cells and somatic cells relative to ES cells (upper right corner genes in **a**) and 391 genes that are lower in both iPS cells and somatic cells relative to ES cells (lower left corner genes) at a *t*-test *P*-value cutoff of 0.01. The plots illustrate progressive convergence of somatic gene expression towards the ES cell state. (**c**) Top, Venn diagrams for progressively reprogrammed genes (somatic > iPS > ES or somatic < iPS < ES). Bottom, Venn diagrams for over-reprogrammed genes (somatic < ES < iPS or somatic > ES > iPS). The *P* values for the overlaps are from Fisher's exact test, and show significant overlaps only for progressively reprogrammed genes. The standard deviations indicate variation among the three cell types.

between each iPS cell type and ES cells were not differentially expressed between the original somatic cells and ES cells, and thus reflect aberrant transcriptional reprogramming (Supplementary Fig. S3b).

In addition, we do not find evidence of persistent expression of master transcriptional regulators of specific cell types. Microphthalmia-associated transcription factor (MITF) is a master regulator of melanocyte differentiation[19] and regulates a class of melanocyte-specific genes. Our data show that MITF and its target genes *TYR* (tyrosinase) and *TRPM1* (transient receptor potential cation channel, subfamily M, member 1) were successfully suppressed in Mel-iPS cells to levels similar to ES cells (DEDS $q$ value = 0.3 for MITF). Similarly, the hepatocyte nuclear factor (HNF) transcription factors and their target genes highly expressed in hepatocytes[20] were reprogrammed in Hep-iPS to the ES cell state (the minimum DEDS $q$ value for HNFs was 0.2). These findings indicate that key lineage-specifying transcription factors do not seem to play a major role in establishing a persistent somatic transcriptional memory in iPS cells.

Finally, we found no evidence that Hep-iPS cells are more efficient than Fib-iPS cells in targeted differentiation towards endoderm at both the messenger RNA and protein level (Supplementary Figs S4 and S5). We cannot exclude that differentiation biases towards the somatic cell type of origin may be observed using other targeted differentiation assays, as has been described in mouse iPS cells[12,14]. Taken together, our data indicate that low-passage human iPS cells retain a transcriptional memory of the somatic cells, with common as well as cell-specific components.

## DNA methylation can partially explain somatic gene expression in iPS cells

We next analysed available data on genome-wide DNA methylation in ES cells and fibroblasts[21]. The top incompletely silenced genes in iPS cells, such as *C9orf64*, *TRIM4* (tripartite motif-containing 4) and *COMT* (catechol *O*-methyltransferase), showed preponderant promoter DNA methylation only in H1 ES cells and not in IMR90 lung fibroblasts (Supplementary Fig. S6). To carry out an unbiased assessment of the contribution of differential DNA methylation to the observed differential expression between iPS and ES cells (Supplementary Fig. S7a), the CpG (cytosine–phosphate–guanine) islands of all genes higher in each iPS cell type relative to ES cells were examined for cytosines differentially methylated between IMR90 and H1 (ref. 21). Genes incompletely repressed in Fib-iPS cells showed a strong trend to be DNA methylated at their promoters in H1 ES cells, but not IMR90 fibroblasts (Fig. 4a): the Pearson correlation coefficient between the log expression fold-change Fib-iPS/ES and $mC_{ES>IMR90}$ was 0.80 ($R^2 = 0.64$ for 12 RefSeq genes with DEDS $q$ value < 0.05). Strikingly, a similar correlation was found for Hep-iPS (Pearson correlation = 0.37 for 56 RefSeq genes with DEDS $q$ value < 0.05) and for Mel-iPS (Pearson correlation = 0.74 for 14 RefSeq genes with

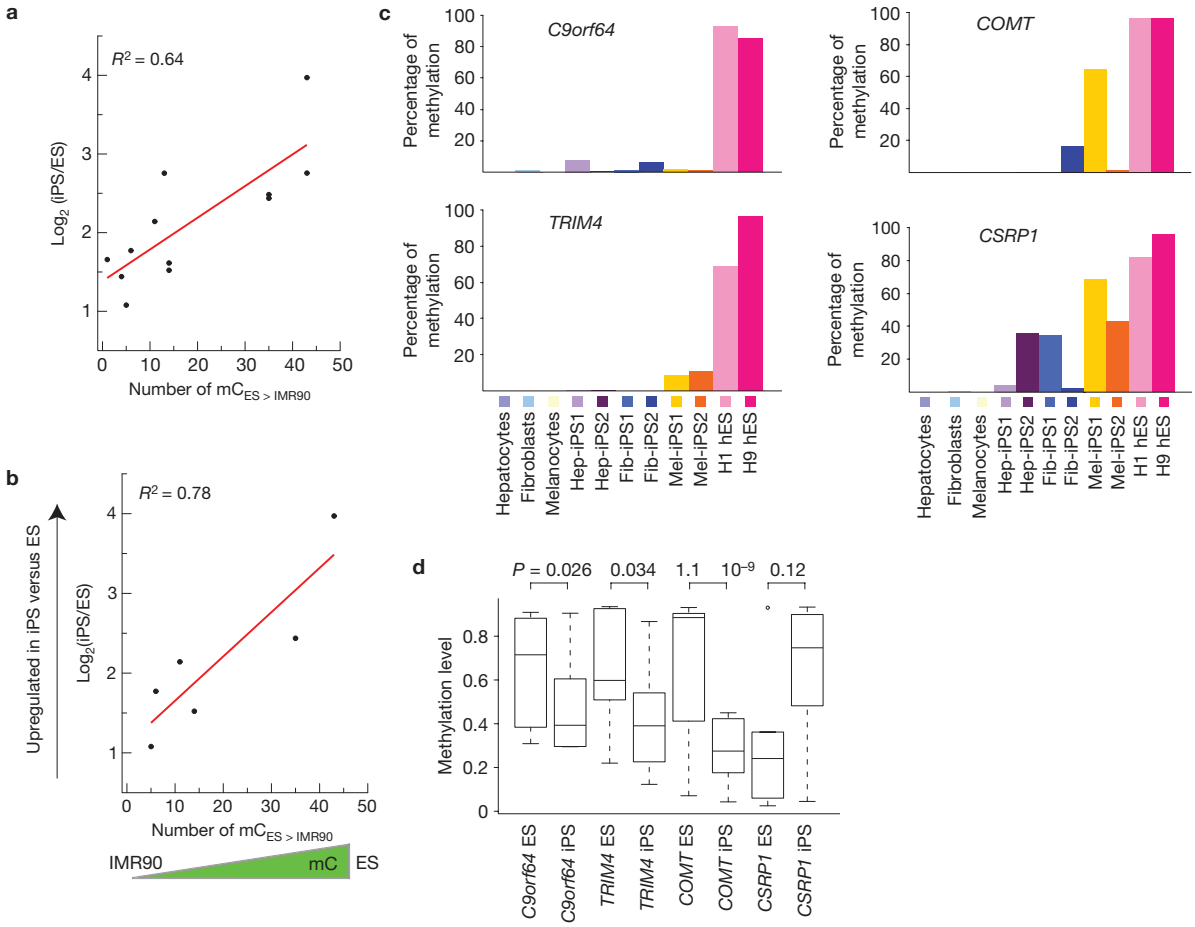


**Figure 4** DNA methylation can partially explain somatic gene expression in iPS cells. (**a**) The genes that maintain higher expression levels in Fib-iPS cells when compared with ES cells tend to be also methylated at higher levels in H1 ES cells when compared with the fibroblast cell line IMR90. The Pearson correlation coefficient between the log expression fold change and single-nucleotide resolution differences in CpG island methylation was 0.80 ($R^2 = 0.64$, $P$ value = 0.002). (**b**) The correlation was 0.88 ($R^2 = 0.78$, $P$ value = 0.02) for six genes with expression levels that remain higher in all three iPS cell types when compared with ES cells. $mC_{ES>IMR90}$ is the number of cytosines in CpG islands with higher levels of methylation in H1 than IMR90. (**c**) The overall level of DNA methylation of four of the top somatic genes whose expression persists in low-passage iPS cells. The level of DNA methylation was examined with bisulphite sequencing analysis in three types of somatic cell (hepatocytes, fibroblasts and melanocytes), two clones for each iPS cell type and H1 and H9 human ES (hES) cells. The detailed bisulphite sequencing data for all samples can be found in Supplementary Data S2. (**d**) Higher-passage iPS cells retain incomplete DNA methylation at somatic cell memory genes. CpG island methylation levels were examined for our validated somatic memory genes (**c**) in five ES cell lines and six iPS cell lines with passage number >30 (passage range 30–58, data from a recent study[9]). The box plot shows the difference in methylation levels between the higher-passage ES and iPS cells. One-sided Wilcoxon test $P$ values confirm that *C9orf64*, *TRIM4* and *COMT* are still resistant to promoter DNA methylation (that is, they are hypomethylated) in high-passage iPS cells relative to high-passage ES cells. No significant difference in the level of DNA methylation was found for the more variable of the four genes, *CSRP1*.

DEDS $q$ value < 0.05, Methods and Supplementary Fig. S7b). Figure 4b shows that the correlation remains high if we consider only those genes that were differentially expressed in all three iPS cells when compared with ES cells, indicating that the contribution of DNA methylation to expression variation is not dependent on cell type. A similar analysis using CpG shores, 2-kilobase (kb)-long flanking regions of CpG islands that have previously been associated with incomplete reprogramming[22], yielded only a weak explanation of $R^2 = 0.02$ for the observed variance in differential expression. Our data thus indicate that incomplete establishment of new promoter CpG DNA methylation may occur during reprogramming.

We next carried out bisulphite sequencing analysis of promoter CpG methylation for four of the top somatic genes whose expression persists in iPS cells, *C9orf64*, *TRIM4*, *COMT* and *CSRP1* (cysteine and glycine-rich protein 1; Fig. 4c). Consistent with the high expression levels of *C9orf64*, *TRIM4* and *COMT* in somatic and iPS cells (Supplementary Table S1), the promoters of these three genes were depleted of CpG methylation in these cell types, but heavily methylated in ES cells (Fig. 4c and Supplementary Data S2). Consistent with the pattern of gene expression, *CSRP1* exhibited greater variability, but also showed the trend of being most methylated in ES cells, intermediately methylated in all iPS cells and least methylated in the somatic cells. We validated differential methylation using four other independent human ES cell lines and four other independent iPS cell lines, including iPS cells generated with different methods such as RNA transfection (Supplementary Fig. S7c). In addition, *C9orf64*, *TRIM4* and *COMT* were also insufficiently methylated in six late-passage iPS cell lines when compared with five late-passage ES cell lines (all above passage 30),

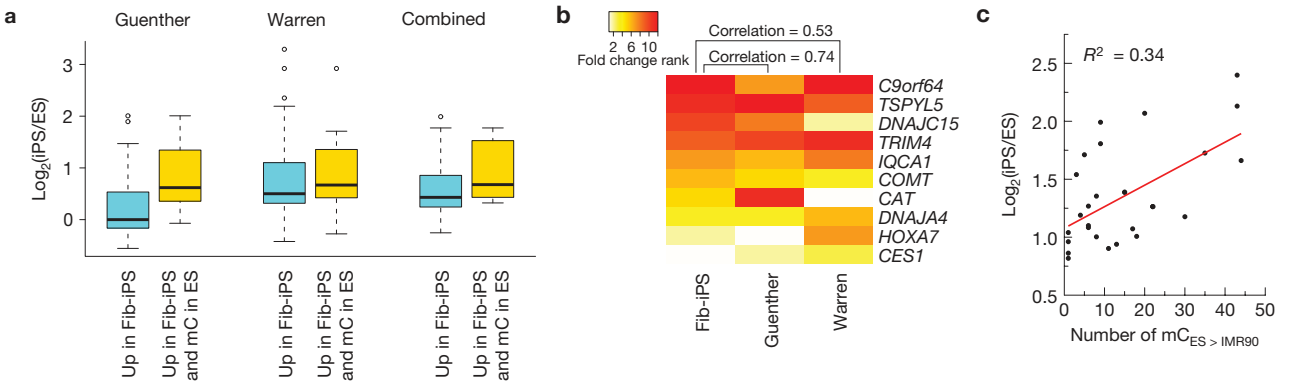


**Figure 5** Meta-analysis of DNA-methylation-associated transcriptional memory in independent data sets. (**a**) Thirty-seven genes expressed at higher levels in our Fib-iPS cells relative to ES cells tend to show higher expression in the iPS cells generated by Guenther *et al.*[7] and Warren *et al.*[29], but there is high variability when expression data alone are used (cyan box plots). However, when we use 10 differentially expressed genes from our data that were also differentially DNA methylated in ES cells, a greater proportion show persistent higher expression in the iPS cells of the two data sets (yellow box plots). (**b**) The heat map shows the iPS/ES fold-change ranking of the 10 genes that are higher in our Fib-iPS cells and also methylated in ES cells. (The higher the rank, the greater the fold change.) Shown are the Spearman rank correlation coefficients of fold changes between our data and those of Guenther *et al.*[7] and Warren *et al.*[29]. (**c**) Twenty-nine genes were expressed at significantly higher levels in iPS cells relative to ES cells in a pooled analysis of the Guenther *et al.*[7] and Warren *et al.*[29] data sets and were also differentially methylated in ES cells[21]. Differential expression was determined by applying meta-DEDS analysis to the pooled data set at a stringent cutoff of 0% FDR. The figure shows that the fold-change levels of those genes correlate significantly with DNA methylation levels ($P$ value $= 9.9 \times 10^{-4}$): the higher the fold change in iPS cells relative to ES cells, the higher the level of promoter DNA methylation in H1 ES cells relative to IMR90 fibroblasts.

which were analysed in a recent study[9] (Fig. 4d). These data indicate that the hypomethylated state of somatic cell genes can persist and correlate with expression in human iPS cells.

## Meta-analysis confirms recurrent transcriptional memory associated with DNA methylation

We next sought to determine whether the genes associated with somatic cell memory in our data showed similar expression trends in other published data sets. A pooled analysis of eight different studies[4,11,23–28] comparing human iPS cells and ES cells revealed that the most incompletely silenced genes in our data, *C9orf64* and *TRIM4*, are within the top differentially expressed genes in these other studies, with an expression ~4-fold higher in iPS than in ES cells (see Methods). We also compared our data with two recent studies that report large data sets comparing iPS cells with ES cells[7,29]. Guenther *et al.*[7] profiled 7 different ES cell lines and 14 fibroblast-derived iPS cell lines, 6 of which had been treated to excise the reprogramming factors from the genome. Warren *et al.*[29] used synthetic mRNAs to reprogram four different types of fibroblast and also profiled H1 and H9 ES cell lines. We first pooled together the two data sets using meta-DEDS (mDEDS; ref. 30), again synthesizing the aforementioned four statistical tests. At 5% FDR, 37 genes are higher in our Fib-iPS cells relative to ES cells, and 10 of them had higher DNA methylation levels in ES cells. Of these 37 genes, 68% are also higher in the pooled Guenther/Warren iPS cells when compared with ES cells (Fig. 5a, 'combined', Fisher test $P = 7.4 \times 10^{-12}$ for the overlap). Strikingly, 9 out of the 10 differentially methylated genes were significantly higher in those iPS cells (Fig. 5a, Fisher test $P = 8.3 \times 10^{-7}$ for the overlap). Not only was the overlap between the genes significant, but their expression levels relative to ES cells also correlated well with our data (Fig. 5b). To test the robustness of this meta-analysis, we also analysed the Guenther and Warren data sets separately: 5 out of our 10 genes (Fisher test $P = 5.8 \times 10^{-10}$ for the overlap) were also expressed at higher levels in the Guenther iPS cells. Seven out of our ten genes (Fisher test $P = 1.0 \times 10^{-5}$ for the overlap) were also expressed at higher levels in the Warren iPS cells. Finally, even at the more stringent cutoff of 0% FDR estimated by mDEDS, 6 out of 10 genes (*C9orf64*; testis-specific Y-encoded-like protein 5, *TSPYL5*; *TRIM4*; IQ motif containing with AAA domain 1, *IQCA1*; DnaJ (heat-shock protein 40) homologue, subfamily C, member 15, *DNAJC15*; catalase, *CAT*, Fisher test $P = 2.1 \times 10^{-12}$ for the overlap) are still found to be expressed at higher levels in the pooled iPS cells.

We directly assessed a correlation between transcription and DNA methylation in the pooled data sets (Fig. 5c). We carried out the expression/DNA methylation regression analysis described earlier (Fig. 4a,b) with the pooled Guenther and Warren data at 0% mDEDS FDR. Figure 5c shows that the log (iPS/ES) fold changes correlate significantly with promoter DNA methylation levels in H1 ES cells (Pearson correlation = 0.58, $t$ distribution $P$ value $= 9.9 \times 10^{-4}$), similar to what we had observed for our data (Fig. 4a,b). These results provide an independent validation of our findings that differences in levels of DNA methylation at certain somatic cell genes may underlie their expression in low-passage iPS cells, independent of laboratory-specific variability and reprogramming methods.

## The incompletely reprogrammed gene *C9orf64* regulates the efficiency of iPS cell generation

We tested whether the expression of incompletely reprogrammed genes in iPS cells is spurious or has any relevance for reprogramming. We carried out RNA-mediated interference (RNAi) for the top incompletely reprogrammed gene, *C9orf64*, in the context of iPS cell generation. We found that RNAi against *C9orf64* during generation of human iPS cells, using three independent shRNAs, significantly decreased the total number of colonies staining positive for Tra1-81, compared with infection with the four factors alone or together with a non-targeting shRNA control (Fig. 6a). The *C9orf64*-knockdown phenotype could be rescued by overexpression of an RNAi-immune complementary DNA (Supplementary Fig. S8). *C9orf64* inhibition did not substantially reduce total cell numbers during the first 10 days

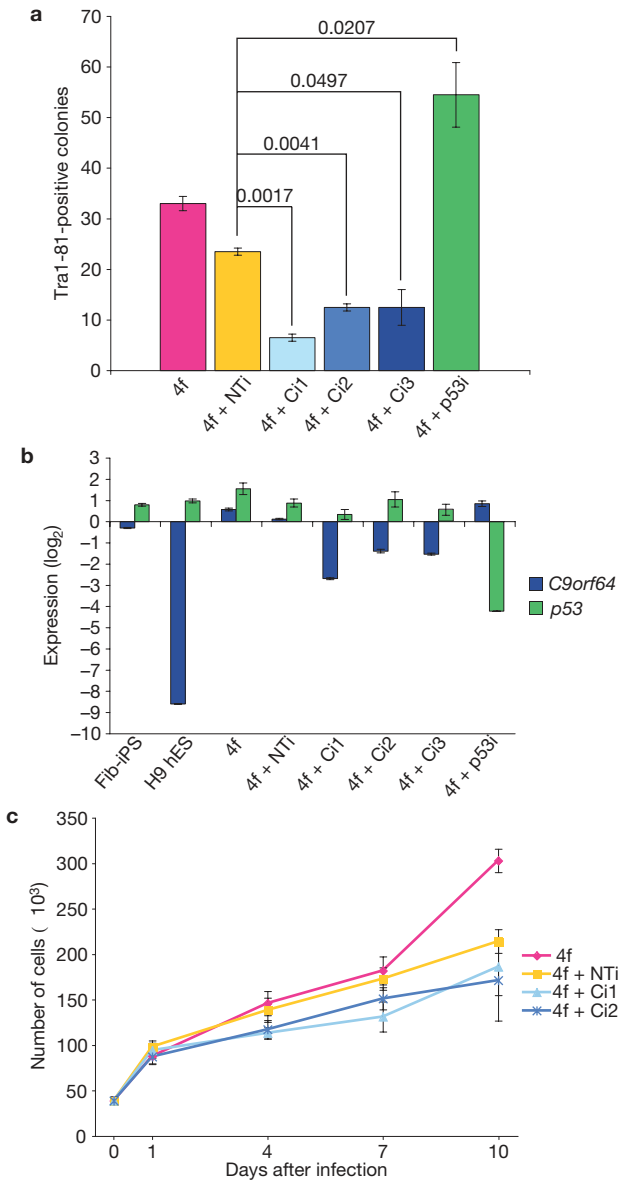


**Figure 6** The somatic cell memory gene *C9orf64* is required for efficient generation of iPS cells. (**a**) The number of Tra1-81-positive iPS cell colonies was counted on d20 after infection of BJ foreskin fibroblasts with 4f alone (4f), 4f+non-targeting shRNA (4f+NTi), 4f+*C9orf64* shRNA (three different short hairpins targeting *C9orf64* were independently tested, 4f+Ci1, 4f+Ci2 and 4f+Ci3) or 4f+ *p53* shRNA (4f+p53i). Infections were carried out in duplicate. Knockdown of *C9orf64* resulted in a significant reduction in the number of Tra1-81-positive iPS cell colonies when compared with 4f alone, 4f+NTi or 4f+p53i. (**b**) Reduction in the levels of *C9orf64* expression achieved by each of the three shRNA constructs was confirmed by quantitative rtPCR. The 4f, 4f+NTi and 4f+p53i conditions showed no significant reduction in the level of *C9orf64* expression. Fib-iPS and H9 human ES (hES) cells served as positive and negative controls for *C9orf64* expression, respectively. *p53* expression analysis further validated the specificity of the shRNAs. Values were standardized to GAPDH and Ubb, then normalized to uninfected BJ fibroblasts. Note $\log_2$ scale in *y* axis: for example, −2 equals a fourfold reduction, −3 equals an eightfold reduction, and so on. Data are from triplicate reactions. (**c**) Growth curves of fibroblasts infected with 4f, 4f+NTi, 4f+Ci1 and 4f+Ci2, counted on d0, d1, d4, d7 and d10 post-infection. Infections were carried out in triplicate. *C9orf64* RNAi did not substantially alter total cell numbers during the first 10 days of reprogramming. In all panels, data shown are representative of two independent experiments, and error bars represent standard deviations.

of reprogramming, before the appearance of colonies (Fig. 6c). These results indicate that *C9orf64* is required for efficient iPS cell generation, although its mode of action remains to be determined.

### Proximity in the genome affects efficiency of gene silencing in iPS cells

We next sought to gain insight into the mechanisms that underlie persistent expression of somatic genes in iPS cells. DNA methyltransferases (DNMTs) were detected at equivalent levels in iPS cells and ES cells (Fig. 7a), indicating that the differential methylation observed between iPS cells and ES cells cannot be attributed to insufficient DNMT levels. There is no correlation between the density of promoter CpGs and the extent to which somatic genes are silenced (data not shown). Interestingly, we found a non-random pattern in the genomic locations of incompletely silenced genes: they tend to be isolated from other genes that undergo silencing on reprogramming (Fig. 7b). These findings indicate that the recruitment of the silencing machinery, including DNMTs, may be inefficient or delayed at certain somatic genes that are 'left behind' owing to their isolation.

## DISCUSSION

Our data document how remarkably similar to human ES cells are iPS cells generated from different somatic cell types. Nevertheless, we find that iPS cells retain a residual transcriptional memory of the somatic cells, and provide data in support of inefficient promoter DNA methylation as the underlying mechanism. Many factors may contribute to variability in gene expression in human iPS cells, including genetic background, starting somatic cell, method used for reprogramming, culture conditions, passage number and batch effects in microarray studies. Some of these same factors may also affect ES cells and have complicated an analysis of the potential transcriptional differences between human iPS cells and ES cells[4–8]. The strength of our study resided in comparing human iPS cells generated from different somatic cell types using the same methodology and analysed in parallel. Our use of gene expression and DNA methylation, rather than gene expression alone, allowed us to find evidence for somatic cell memory in other studies.

It has been shown that promoter DNA demethylation, a pre-requisite for gene reactivation, can be inefficient during generation of iPS cells[12,31]. We report here that DNA methylation and silencing of somatic genes may also contribute to reprogramming (Fig. 8). A complex balance between DNA demethylation and methylation is therefore likely to be critical for reprogramming. Our data indicate that care should be taken when using small molecules that promote DNA demethylation in iPS cells, and that an evaluation of the DNA methylation status of somatic cell genes may be warranted in the validation of new human iPS cell lines.

It is important to point out that most of our findings pertain to low-passage ($<P20$) human iPS cells, and that many of the differences relative to ES cells are expected to be attenuated, although possibly not completely abolished (see Fig. 4d), with extensive passaging[4,14]. The expression profile of ES cells, on the other hand, has been suggested to be relatively stable with passaging[4]. It will nevertheless be important to determine whether variability between ES cell lines, or any gene expression changes that ES cells may develop with continued culture, are also mediated by differential DNA methylation.

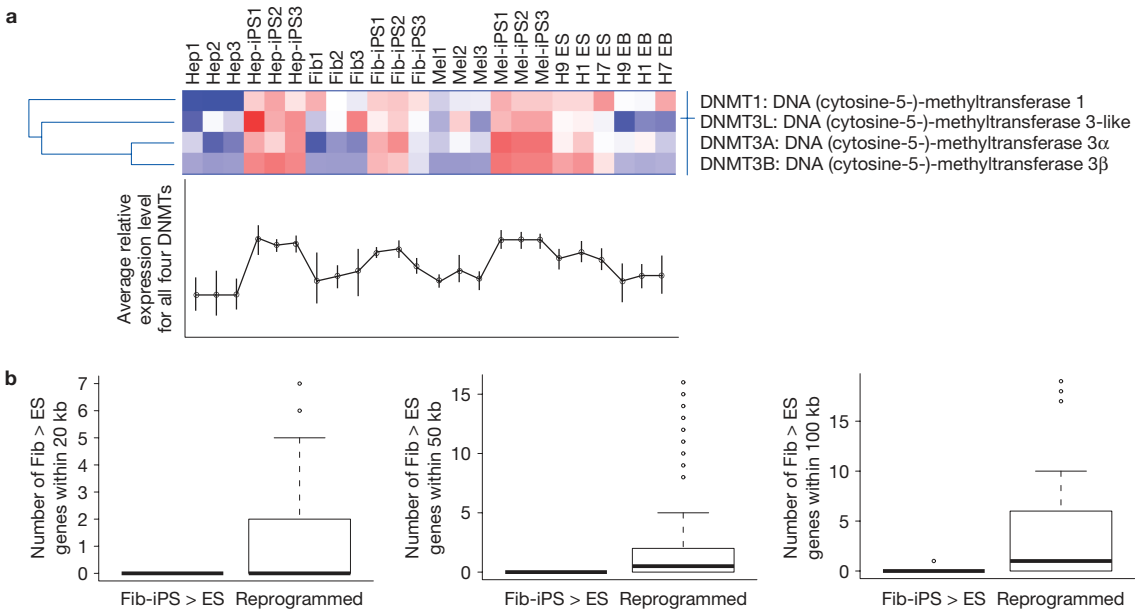


**Figure 7** Proximity in the genome affects efficiency of gene silencing in iPS cells. (**a**) The heat map shows the expression levels of four different DNMTs in all of the cell types analysed in our microarray study. The expression level of these DNMTs is relatively equivalent between iPS cells and ES cell controls, indicating that any differential DNA methylation in iPS cells is not due to insufficient DNMT expression. EB, embryoid bodies. The error bars represent the standard deviations of the relative expression level between each of the DNMTs for each cell type. (**b**) We considered 62 silenced genes ('Reprogrammed') and 5 genes whose expression persists in iPS cells ('Fib-iPS > ES'), with at least 10 cytosines methylated only in ES and also showing higher expression in Fib than ES cells. The 'Reprogrammed' genes tend to have nearby genes that also require silencing, whereas the 'Fib-iPS > ES' genes are more isolated. The one-sided Wilcoxon *P* values for the difference in the number of nearby genes between the two groups are 0.054, 0.022 and 0.028 for the 20-kb, 50-kb and 100-kb distance restrictions, respectively. The local density of genes, irrespective of expression status, was also slightly lower near genes whose expression persists in iPS cells.

The *C9orf64* RNAi data indicate that some somatic genes may continue to be expressed in low-passage iPS cells because they play an active role during reprogramming. *C9orf64* is a conserved gene of unknown function with no known protein domains. It is possible that it is required to stabilize an intermediate stage with characteristics of both the somatic and the reprogrammed state, although further studies will be required to address this.

Our data indicate that gene density can affect the efficiency with which genes are silenced. The proximity of multiple genes being repressed may synergize in recruiting the silencing machinery, whereas silencing may be inefficient or delayed in more isolated regions, where stochastic events thought to underlie the reprogramming process[32,33] may have a lower probability of occurring (Fig. 8). It will be of interest to determine how positional effects in the genome affect the efficiency of epigenetic and transcriptional reprogramming.

Interestingly, several of the somatic cell memory genes reported here have been associated with cancer. *TSPYL5* is silenced and DNA methylated in a subset of cancers[34–36]. *C9orf64* is deleted in some cases of acute myeloid leukaemia[37], and its promoter region is methylated in some breast cancer cell lines[38]. *CSRP1* has been proposed to be a tumour suppressor silenced by DNA methylation in hepatocellular carcinoma[39]. It is therefore possible that deletion or epigenetic silencing of genes associated with somatic cell memory may contribute to cancer progression. Indeed, our preliminary findings indicate that the incompletely silenced genes reported here show a significant trend for downregulation during progression of hepatocellular carcinoma (data not shown). Our results prompt an evaluation of the role of somatic cell memory genes in cancer models. □

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturecellbiology/

*Note: Supplementary Information is available on the Nature Cell Biology website*


### ACKNOWLEDGEMENTS

The authors wish to thank S. Fisher, O. Genbachev, A. Leavitt and B. Conklin for expert advice on culturing human ES cells, D. Subramanyam and R. Blelloch for the Adult Fibroblast-iPS 1 cell line, L. Ta, A. Williams and A. Holloway at the Gladstone Institutes, J. Bolen at the Mouse Pathology Core Facility for expert assistance, J. Utikal for technical advice, and J. Yang and A. Campain for sharing their meta-DEDS code. We thank members of the Santos laboratory, R. Blelloch, H. Willenbring, S. Fisher and M. Grskovic for helpful discussions and critical reading of the manuscript. Work in the Santos laboratory is supported by CIRM, JDRF, an NIH Director's New Innovator Award and the Leona M. and Harry B. Helmsley Charitable Trust. Y.O. was partially supported by the UCSF Diabetes Center and a T32 grant from the NICHD to the UCSF Center for Reproductive Sciences. Work in M.H.'s laboratory was supported by grants from the JDRF and the Leona M. and Harry B. Helmsley Charitable Trust. T.G. was supported by the JDRF and the Leona M. and Harry B. Helmsley Charitable Trust. S.L.D. was partially supported by CIRM. J.S.S. was partially supported by the PhRMA Foundation.


### AUTHOR CONTRIBUTIONS

Y.O., J.S.S. and M.R-S. conceived the project. J.M.P., K.H., P.D.M. and D.J.R. provided reagents. Z.Q. and J.Y. provided assistance with data analysis. C.H. and S.L.D. carried out the bisulphite sequencing analysis under supervision of J.F.C. T.G. carried out the targeted differentiation to endoderm analysis under supervision of M.H. J.S.S. carried out all of the bioinformatic analyses. Y.O., H.Q. and M.R-S. designed and Y.O. and H.Q. carried out all other experiments with technical assistance from L.B. Y.O, J.S.S. and M.R-S. wrote the manuscript with input from the other authors.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

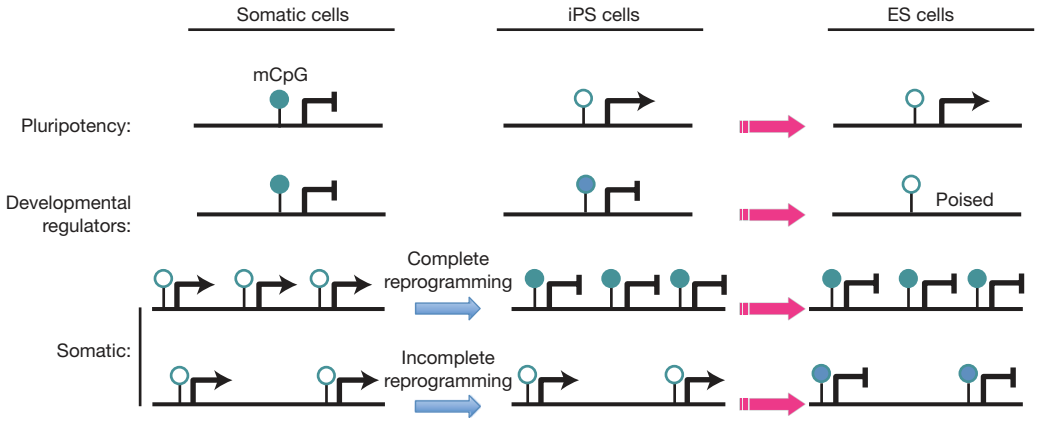


**Figure 8** Model for the role of DNA methylation in reprogramming to the human iPS cell state. It has previously been shown that DNA demethylation and reactivation of pluripotency genes are essential components of reprogramming[31] (top). In addition, incomplete demethylation of genes silenced in the somatic cell, including developmental regulators of other lineages, has been shown to persist in mouse iPS cells and may affect their differentiation[12,14] (middle). We report here that differential methylation of somatic cell genes underlies their differential expression in human iPS cells (bottom 'somatic' panel), and that somatic genes whose expression persists in low-passage iPS cells tend to be isolated from other genes that undergo silencing. Clustering of genes requiring simultaneous repression may facilitate recruitment of the silencing machinery, including DNMTs, and regional DNA methylation (top 'somatic' panel). Extensive passaging (pink arrows) may lead to further epigenetic silencing of somatic genes in human iPS cells. Arrows indicate active transcription, and hooks indicate repression.

Published online at http://www.nature.com/naturecellbiology
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Yamanaka, S. A fresh look at iPS cells. *Cell* **137,** 13–17 (2009).
2. Feng, Q. *et al.* Hemangioblastic derivatives from human induced pluripotent stem cells exhibit limited expansion and early senescence. *Stem Cells* **28,** 704–712 (2010).
3. Hu, B. Y. *et al.* Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency. *Proc. Natl Acad. Sci. USA* **107,** 4335–4340 (2010).
4. Chin, M. H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5,** 111–123 (2009).
5. Ghosh, Z. *et al.* Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PLoS One* **5,** e8975 (2010).
6. Marchetto, M. C. *et al.* Transcriptional signature and memory retention of human-induced pluripotent stem cells. *PLoS One* **4,** e7076 (2009).
7. Guenther, M. G. *et al.* Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* **7,** 249–257 (2010).
8. Newman, A. M. & Cooper, J.B. Lab-specific gene expression signatures in pluripotent stem cells. *Cell Stem Cell* **7,** 258–262 (2010).
9. Bock, C. *et al.* Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144,** 439–452 (2011).
10. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27,** 353–360 (2009).
11. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41,** 1350–1353 (2009).
12. Kim, K. *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* **467,** 285–290 (2010).
13. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471,** 68–73 (2011).
14. Polo, J. M. *et al.* Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.* **28,** 848–855 (2010).
15. Hockemeyer, D. *et al.* A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell* **3,** 346–353 (2008).
16. Maherali, N. *et al.* A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell* **3,** 340–345 (2008).
17. Utikal, J., Maherali, N., Kulalert, W. & Hochedlinger, K. Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *J. Cell Sci.* **122,** 3502–3510 (2009).
18. Yang, Y. H., Xiao, Y. & Segal, M. R. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* **21,** 1084–1093 (2005).
19. Hemesath, T. J. *et al.* Microphthalmia, a critical factor in melanocyte development, defines a discrete transcription factor family. *Genes Dev.* **8,** 2770–2780 (1994).
20. Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303,** 1378–1381 (2004).
21. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
22. Irizarry, R. A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18,** 780–790 (2008).
23. Avery, K., Avery, S., Shepherd, J., Heath, P. R. & Moore, H. Sphingosine-1-phosphate mediates transcriptional regulation of key targets associated with survival, proliferation, and pluripotency in human embryonic stem cells. *Stem. Cells Dev.* **17,** 1195–1205 (2008).
24. Baker, D. E. *et al.* Adaptation to culture of human embryonic stem cells and oncogenesis *in vivo*. *Nat. Biotechnol.* **25,** 207–215 (2007).
25. Li, S. S. *et al.* Target identification of microRNAs expressed highly in human embryonic stem cells. *J. Cell Biochem.* **106,** 1020–1030 (2009).
26. Lowry, W. E. *et al.* Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl Acad. Sci. USA* **105,** 2883–2888 (2008).
27. Soldner, F. *et al.* Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* **136,** 964–977 (2009).
28. Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324,** 797–801 (2009).
29. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7,** 618–630 (2010).
30. Campain, A. & Yang, Y. H. Comparison study of microarray meta-analysis methods. *BMC Bioinform.* **11,** 408–418 (2010).
31. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).
32. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).
33. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat. Biotechnol.* **25,** 1177–1181 (2007).
34. Chapman, E. J., Kelly, G. & Knowles, M. A. Genes involved in differentiation, stem cell renewal, and tumorigenesis are modulated in telomerase-immortalized human urothelial cells. *Mol. Cancer Res.* **6,** 1154–1168 (2008).
35. Jung, Y., Park, J., Bang, Y. J. & Kim, T. Y. Gene silencing of TSPYL5 mediated by aberrant promoter methylation in gastric cancers. *Lab. Invest.* **88,** 153–160 (2008).
36. Kim, T. Y., Zhong, S., Fields, C. R., Kim, J. H. & Robertson, K. D. Epigenomic profiling reveals novel and frequent targets of aberrant DNA methylation-mediated silencing in malignant glioma. *Cancer Res.* **66,** 7490–7501 (2006).
37. Sweetser, D. A. *et al.* Delineation of the minimal commonly deleted segment and identification of candidate tumour-suppressor genes in del(9q) acute myeloid leukemia. *Genes Chromosomes Cancer* **44,** 279–291 (2005).
38. Cai, L. Y. *et al.* Identification of PRTFDC1 silencing and aberrant promoter methylation of GPR150, ITGA8 and HOXD11 in ovarian cancers. *Life Sci.* **80,** 1458–1465 (2007).
39. Hirasawa, Y. *et al.* Methylation status of genes upregulated by demethylating agent 5-aza-2'-deoxycytidine in hepatocellular carcinoma. *Oncology* **71,** 77–85 (2006).

DOI: 10.1038/ncb2239

## METHODS

**Lentivirus production.** The doxycycline-inducible lentiviral vectors and a lentiviral vector constitutively expressing a reverse tetracycline transactivator (rtTA) used in our study have been previously described[16]. For virus production, 293T cells at 60–70% confluency were transfected in 10 cm plates with 4 μg of the lentiviral vectors together with 1 μg each of the packaging plasmids VSV-G, MDL-RRE and RSVr using Fugene 6 (Roche). Viral supernatants were collected after 72 h, filtered and concentrated with 1 ml of cold PEG-it Virus Precipitation Solution (System Biosciences) for every four volumes of virus. The virus supernatant and PEG-it mixture was incubated overnight at 4 °C. The mixture was centrifuged at 1,500 × g for 30 min at 4 °C, resuspended in 100 μl cold phosphate-buffered saline (PBS) and stored at −80 °C. Lentiviral infections were carried out in 1 ml of medium using 10 μl rtTA, 5 μl each of octamer-binding transcription factor 4 (*OCT4*), sex-determining region Y-box 2 (*SOX2*), Krueppel-like factor 4 (*KLF4*) and *NANOG*, and 2 μl *cMYC* per well of a six-well plate. Polybrene (8 μg ml$^{-1}$; Sigma) was used for each infection.

**Cell culture and human iPS cell generation.** Human primary newborn foreskin (BJ) fibroblasts were obtained from ATCC (reference #: CRL-2522) and cultured in DMEM with 10% FBS, 1× glutamine, 1× non-essential amino acids, 1× sodium pyruvate, 2×penicillin/streptomycin, and 0.06 mM β-mercaptoethanol (fibroblast medium). For lentiviral infections of fibroblasts, 50,000 cells were plated per well of a six-well plate and infected overnight. The day after infection, the virus was removed and replaced with fresh fibroblast medium. At 48 h after infection, the infected cells from a single well of a six-well plate were trypsinized and seeded onto irradiated mouse embryonic feeders in DMEM/F12 with 20% KSR (knockout serum replacement), 0.5× glutamine, 1×non-essential amino acids, 2×penicillin/streptomycin, 0.1 mM β-mercaptoethanol, 10 ng ml$^{-1}$ basic fibroblast growth factor (bFGF; human ES cell medium) containing 2% FBS and 1 μg ml$^{-1}$ doxycycline in 10 cm plate format. The melanocytes were obtained from Promocell (reference #: C-12402). The *NANOG* transgene was not used for deriving Mel-iPS cells (only the doxycycline-inducible 4 factors were used[17]).

Adult human primary hepatocytes were obtained from Lonza (reference #: CC-2703W6) and cultured in human hepatocyte growth medium (HCM, Lonza). Hepatocytes were received as non-proliferating monolayers of cells shipped in a six-well plate format. On arrival, the shipping medium was replaced with fresh HCM and the cells were allowed to recover in a 5% $CO_2$, 37°C incubator for approximately 2 h before infection. Virus infections were carried out in 1 ml HCM per well of a six-well plate on two subsequent days. The day after the last infection, cells were mechanically dissociated into single cells and seeded in HCM onto irradiated mouse embryonic feeders in a 10 cm plate format. The following day, cells were transferred to human ES cell medium containing 1 μg ml$^{-1}$ doxycycline and fed with this medium daily until the appearance of human ES cell-like colonies (up to 40 days). In all cases of human somatic cell reprogramming, Tra1-81 staining of live cells was carried out as previously published[26].

**Immunohistochemical analysis.** Human ES and iPS cells were fixed directly in culturing plates (for pluripotency marker analysis) or on glass coverslips (for the targeted differentiation analysis) with 4% paraformaldehyde, and permeabilized with 0.1% Triton X-100. Cells were then stained with primary antibodies against SSEA-3 (MAB4303, Millipore, 1:100), SSEA-4 (MAB4304, Millipore, 1:100), Tra1-60 (ab16288, Abcam, 1:100), Tra1-81 (MAB4381, Millipore, 1:100), FOXA2 (07633, Upstate, 1:200), SOX17 (AF1924, R&D Systems, 1:1,000) and HNF1b (AF3330, R&D Systems, 1:100). Respective secondary antibodies were conjugated to either Alexa Fluor 594 or Alexa Fluor 488 (Invitrogen) and used at 1:500. Cell counting was done with Cellprofiler 2.0.

**Quantitative PCR.** RNA was isolated from cells using the RNeasy Mini RNA Isolation kit (Qiagen). cDNA was produced with the High-Capacity cDNA Reverse Transcription kit (Applied BioSystems) using random primers. Quantitative real-time PCR (rtPCR) reactions were carried out in triplicate with the SYBR Green quantitative rtPCR Master Mix (Applied BioSystems) and run on an Applied BioSystems 7900HT Sequence Detection System. Primer sequences are listed in Supplementary Table S3.

**Stochastic differentiation to embryoid bodies.** Human ES and iPS cells were lifted from feeder cells using a 1:1 ratio of dispase/collagenase IV mix (1 mg ml$^{-1}$ each). The dispase/collagenase IV mixture (1 ml) was used per well of a six-well plate. Cells were then grown in suspension culture with Knockout DMEM containing 20% FBS, 0.5× glutamine, 1× non-essential amino acids, and 0.1 mM β-mercaptoethanol. Embryoid bodies were collected and analysed at d8 for markers of the three embryonic germ layers.

**Directed differentiation to endoderm.** iPS and ES cells were differentiated towards endoderm using a published protocol[40] (Supplementary Fig. S4a). Two clones each of Hep-iPS cells and Fib-iPS cells and two lines of ES cells (H1 and H9) were used in this analysis. Cells were collected on d3 (definitive endoderm stage) and d6 (primitive gut tube stage) after differentiation and processed for either quantitative rtPCR or immunohistochemical analysis.

**Teratoma induction.** Human ES and iPS cells were grown to 70–80% confluency in a six-well plate format and one entire plate-worth of cells was used to inject one immunocompromised SCID/Beige mouse subcutaneously into two sites near the hind flanks. Each six-well-plate-worth of cells was pelleted and resuspended in 140 μl of DMEM/F12 and immediately before injection, 60 μl of Matrigel (BD Biosciences) was mixed with the cells for a total volume of 200 μl. The cell/Matrigel mix (100 μl) was injected into each site. Tumours developed after 6–12 weeks and were processed for histological analysis.

**Expression data analysis.** The Affymetrix ST 1.0 expression data were normalized together using the robust multichip average (RMA) and the latest RefSeq probe mapping to the reference human genome[41,42]. To minimize redundancy, RefSeq probes corresponding to the same Gene Symbol were combined if they showed no within-array variation for all 24 samples. This filtering process yielded a final list of 26,532 RefSeq genes. The equal-variance $t$-test was used to assess the significance of differential expression between groups. The expression profiles of the three ES cell lines were pooled together into one group. Analysis of variance (ANOVA) was carried out to find 453 genes that are significantly different among the eight groups (Hep, Hep-iPS, Fib, Fib-iPS, Mel, Mel-iPS, ES, EB) at a $P$-value cutoff of $10^{-14}$. Figure 2a shows the average-linkage clustering of the samples using those genes.

**Bootstrap simulation.** Assuming the null hypothesis that the log expression levels for each gene are identically distributed in ES and iPS cells, we estimated a normal null distribution separately for each gene by using maximum likelihood on the pooled data set of three iPS and three ES replicates. Six independent samples were then drawn from the normal distribution for each gene and grouped into three ES versus three iPS; one complete parametric bootstrap simulation consisted of such re-sampling for all RefSeq genes on the microarray. A LOESS curve was fitted to $t$-test $P$ values for each bootstrap simulation. The entire process was repeated 1,000 times, and Fig. 3a shows the enveloping curves for the simulated LOESS regression.

**Independent confirmation of incompletely silenced genes.** We pooled together 24 iPS cell and 18 ES cell expression profiles from Gene Expression Omnibus (GSE18226, GSE14711, GSE9865, GSE16654, GSE6561, GSE7896, GSE9440, GSE15176). The data were normalized together using RMA and then corrected for potential batch effects using an empirical Bayes method[43].

**Meta-analysis of published iPS cell expression profiles.** The data from Guenther *et al.*[7] (GSE23402) and Warren *et al.*[29] (GSE23583) were normalized together using RMA, as described above. We used the Bioconductor package DEDS (ref. 18). We carried out 2,000 permutations and used 5% FDR as a cutoff for deciding differential expression. Meta-DEDS was used to pool together the two data sets, again applying 2,000 permutations and 5% or 0% FDR.

**CpG methylation analysis.** We consider a CpG island to be associated with a gene if it contains the transcription start site of the gene or if one of its edges lies within 2 kb from the transcription start site of the gene. Using the DEDS method[18], 64 RefSeq genes were found to be expressed at a higher level in Fib-iPS cells than ES cells at the $q$-value cutoff of 0.05. Among the 64 genes, 12 genes had differentially methylated cytosines between IMR90 and ES cells in their CpG islands located within 2 kb.

For the 12 genes, we define $f$ = log expression fold change between Fib-iPS and ES. (Note that $f > 0$ if the expression is higher in the iPS cell.)

Let $mC_{ES>IMR90}$ = number of C with higher methylation in ES when compared with IMR90.

Let $mC_{IMR90>ES}$ = number of C with higher methylation in IMR90 when compared with ES.

The Pearson correlation between $f$ and $mC_{ES>IMR90}$ in the corresponding CpG island is 0.80 and the $P$ value for the correlation is $1.9 \times 10^{-3}$. (The corresponding correlation and $P$ value are 0.37 and $5.1 \times 10^{-3}$ for Hep-iPS and 0.74 and $2.5 \times 10^{-3}$ for Mel-iPS.) Six genes were differentially expressed in all iPS cells when compared with ES cells at a DEDS $q$-value cutoff of 0.05 and had differentially methylated CpG islands between IMR90 and ES cells. The Pearson correlation between $f$ and $mC_{ES>IMR90}$ for those genes is 0.88 and $P$ value = 0.02.

A least-squares linear regression model was fitted to the log differential expression fold changes with $mC_{ES>IMR90}$ and $mC_{IMR90>ES}$ as two predictors. Only $mC_{ES>IMR90}$, and not $mC_{IMR90>ES}$, contributed significantly to the model. The statistical package R was used for the computations.

**Clonal bisulphite sequencing.** Total genomic DNA underwent bisulphite conversion following an established protocol[44] with modification of: 95 °C for

1 min, 50 °C for 59 min for a total of 16 cycles. Regions of interest were amplified with PCR primers (Supplementary Table S2) and were subsequently cloned using pCR2.1/TOPO (Invitrogen). Individual bacterial colonies were subjected to PCR using vector-specific primers and sequenced using an ABI 3700 automated DNA sequencer.

**RNAi in reprogramming.** Newborn foreskin fibroblasts were seeded at 30,000 cells per well of a six-well plate the day before infection. Cells were infected with 0.5 µl each of concentrated retroviruses (obtained from the Harvard Gene Therapy Initiative), leading to the overexpression of *OCT4*, *SOX2* and *KLF4*, and 0.05 µl in the case of *cMYC*, alone or in combination with 50 µl of non-concentrated lentivirus for a non-targeting shRNA (5′-ATCTCGCTTGGGCGAGAGTAAG-3′), *C9orf64* shRNA (three independent shRNAs—shRNA1: 5′-CATGTTTGCTGATTATAGA-3′; shRNA2: 5′-CTTTGATATTTAGAGAACA-3′; shRNA3: 5′-GAGGTTATAGGAAATTGAT-3′) or a *p53* shRNA (5′-GACTCCAGTGGTAATCTACT-3′). Cells were infected in 1 ml human ES cell medium (see the section, Cell culture and human iPS cell generation) and 8 µg ml$^{-1}$ polybrene. Cells remained in the presence of virus for 48 h and on the day after virus addition, 1 ml of fibroblast medium was added. At 48 h after infection, virus was removed and cells were cultured in ES cell medium. On d20–d28 after infection, Tra1-81 staining of live cells was carried out to identify fully reprogrammed iPS cell colonies.

**Accession numbers.** The microarray data are available from Gene Expression Omnibus under access number GSE23034.

40. Kroon, E. *et al.* Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells *in vivo*. *Nat. Biotechnol.* **26,** 443–452 (2008).
41. Dai, M. H. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** E175 (2005).
42. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).
43. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
44. Grunau, C., Clark, S. J. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.* **29,** E65 (2001).

nature cell biology



DOI: 10.1038/ncb2239

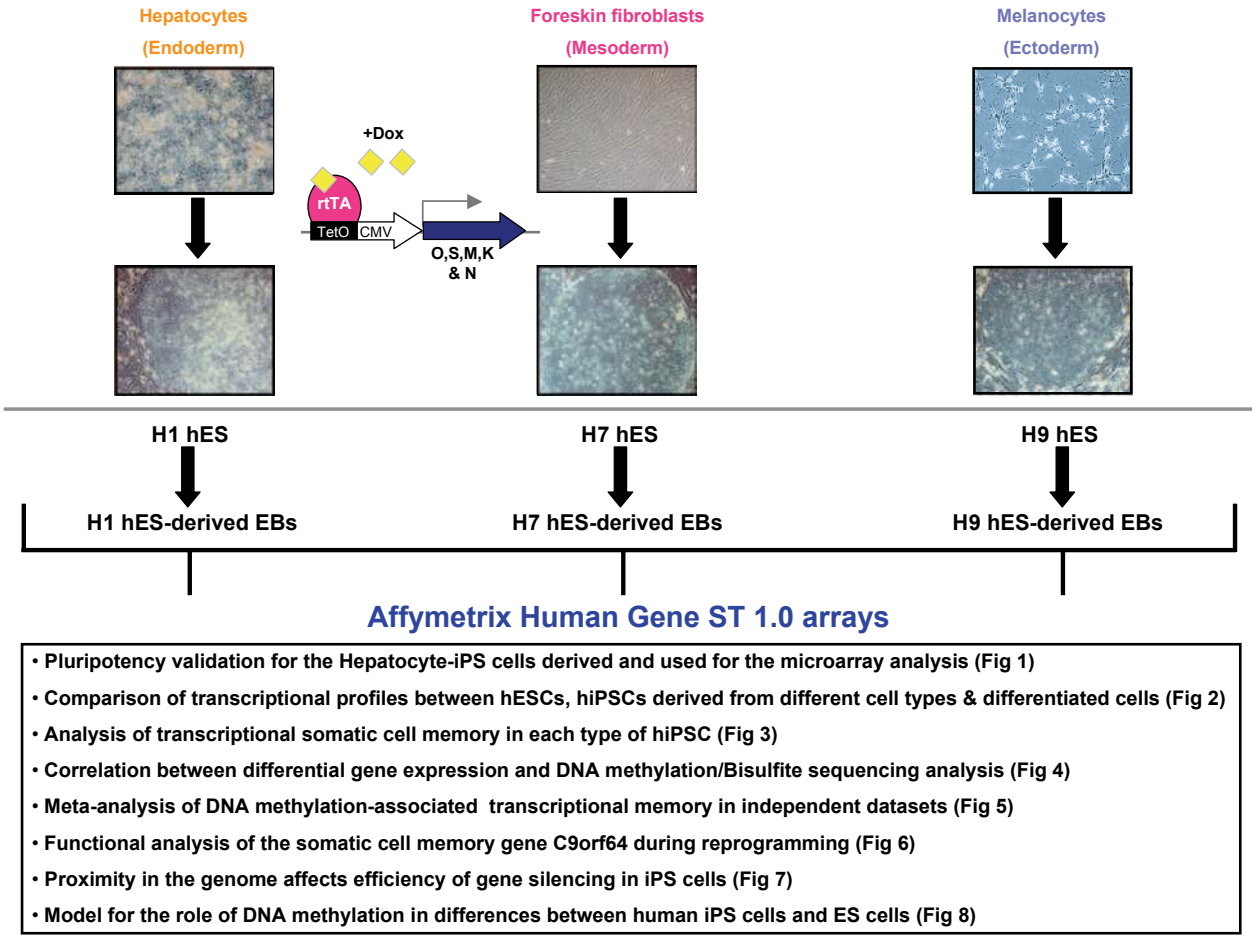


**Figure S1** Flowchart of analysis and data reported in the manuscript. Human somatic cells representative of the 3 embryonic germ layers were reprogrammed to pluripotency using the dox-inducible lentiviral transgene system for overexpression of the reprogramming factors OCT4 (O), SOX2 (S), cMYC (M), KLF4 (K) and NANOG (N). Hepatocyte-derived iPS cells, newborn foreskin fibroblast-derived iPS cells and melanocyte-derived iPS cells represented the endodermal, mesodermal and ectodermal lineages, respectively. All 3 types of iPS cells, their parental somatic cell counterparts, 3 lines of human ES cells (H1, H7, H9) and Embryoid Bodies derived from these ES cells were hybridized to Affymetrix Human Gene ST 1.0 arrays and transcriptionally profiled in parallel. Triplicates of independent clones were used for all cell types except for the somatic cells, since each somatic cell represents a single clone. For all 3 types of parental somatic cells, technical triplicates were used for the analysis. Shown in the box beneath the flowchart is a brief description of each of the figures presented in the main text of the manuscript.

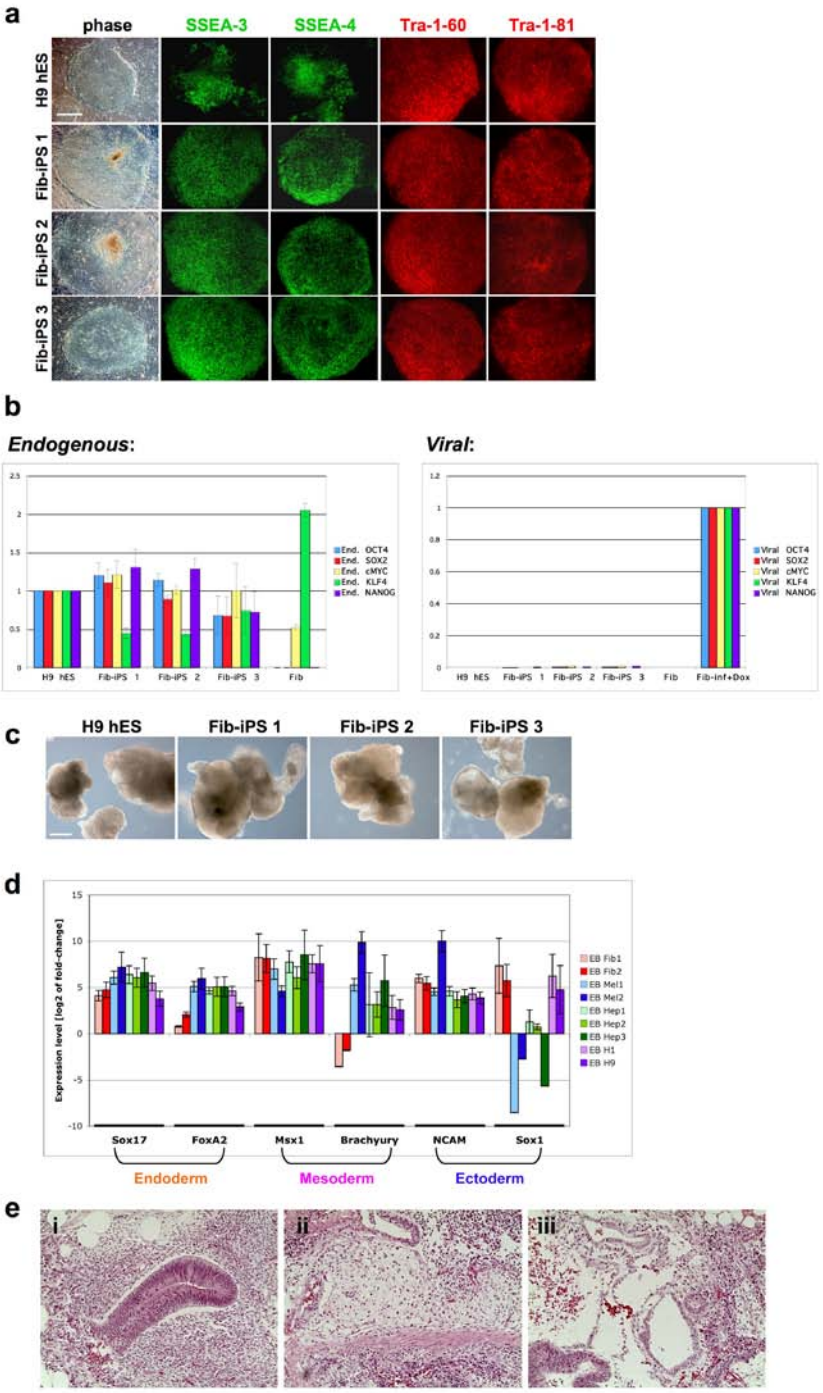


**Figure S2** Pluripotency validation for the Fibroblast-iPS cells derived and used for the microarray studies. (a) All 3 Fib-iPS clones used in this analysis showed strong, positive staining for the human ES cell specific-markers SSEA-3, SSEA-4, Tra1-60 and Tra1-81, comparable to that observed in control H9 ES cells. Scale bar represents 300 μm. (b) qRT-PCR was used to confirm both high expression levels of endogenous pluripotency genes in all 3 Fib-iPS cell clones, as well as negligible levels of transgene expression. Values were standardized to GAPDH and Ubb, then normalized to H9 ES cells (endogenous) or 5-factor infected BJ fibroblasts + dox for 4 days (viral). (c) All clones of Fib-iPS cells formed embryoid bodies in vitro when grown under non-attachment conditions. Shown here are d8 EBs alongside with control ES cell-derived EBs. Scale bar represents 200 μm. (d) The pluripotent nature of all iPS cells used in our analysis was confirmed by their ability to form EBs comprised of tissues derivative of the 3 germ layers in vitro (also see Fig. 1c and Supplementary Figure 2c). qRT-PCR analysis on d8 EBs derived from all types of iPS cells and ES cell controls confirmed the presence of the 3 embryonic germ layers. Values were standardized to GAPDH and Ubb. Expression fold changes shown in the graph are relative to H9 ES cells on d0. (e) Hematoxylin and eosin stain of teratomas generated from fibroblast-derived iPS cells injected subcutaneously into immunocompromised SCID mice. Structures derivative of all three germ layers could be identified. (i) Neural tissue (ectoderm), (ii) Striated muscle and mesenchyme (mesoderm), (iii) gut-like epithelium (endoderm). In **b and d**, data are from triplicate reactions and error bars represent standard deviations.

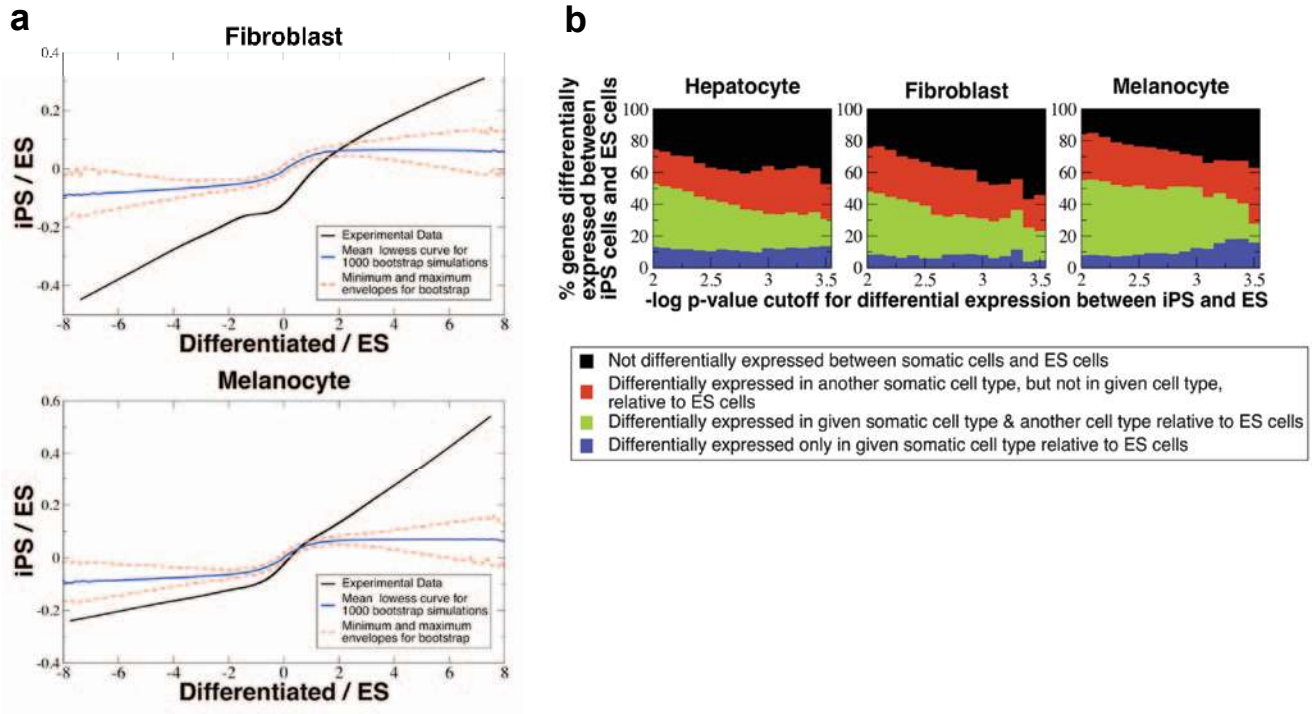


**Figure S3** iPS cells retain a transcriptional memory of the original somatic cell. (a) Similar to Fig. 3a, these figures show LOESS curves fitted to the scatter plots of t-test log p-values for fibroblast and melanocyte: -log(p) and log(p) are plotted for fold changes greater than 1 and less than 1, respectively. The black line is a curve fitted to our data, and other curves are fitted to the 1000 bootstrap simulation datasets obtained by assuming identically distributed iPS and ES cell expression levels. The black line shows clear deviation from the null hypothesis iPS=ES and thus reflects the trend that the transcriptional memory of the originating cell type is retained in iPS cells: genes that were higher (or lower) in the somatic cell than in ES cells tend to be significantly repressed (or induced) during reprogramming, but nevertheless remain higher (or lower) in iPS cells than in ES cells. (b) The partition of differentially expressed genes between iPS cells and ES cells according to their expression status in somatic cells. It is seen that ~50% of the genes were already differentially expressed between the corresponding somatic cell type and ES cells, while ~10% were differentially expressed only in the corresponding somatic cell type and not other cell types relative to ES cells. Cell type-specific somatic expression can thus explain ~10% of the observed incomplete reprogramming.

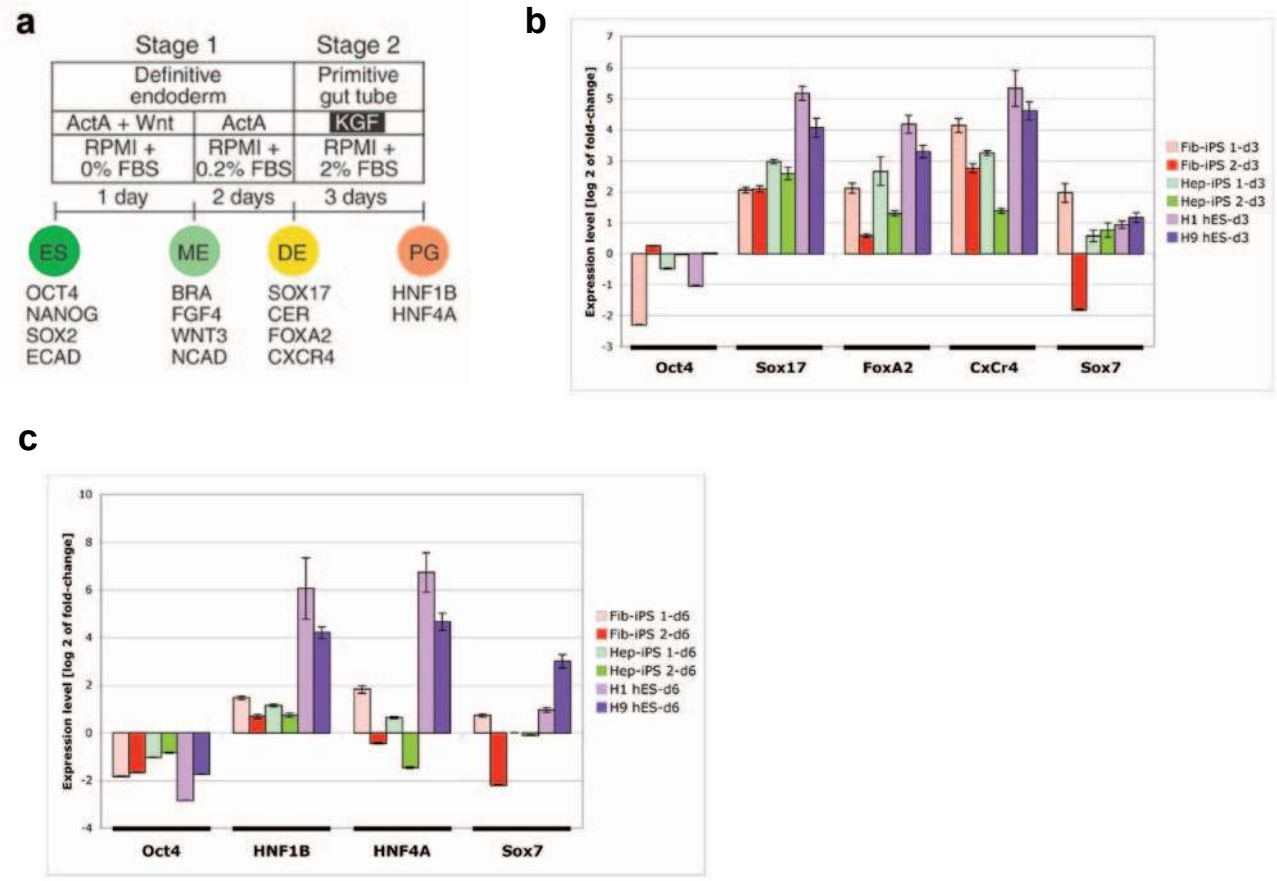


**Figure S4** Targeted differentiation of iPS and ES cells towards an endodermal fate. (a) A targeted differentiation approach based on a previously published protocol by Kroon et al., 2008 was used to differentiate 2 clones each of Fib-iPS and Hep-iPS cells and 2 lines of ES cells (H1 and H9) to the definitive endoderm stage (d3 of the protocol) and to the primitive gut tube stage (d6 of the protocol) using the growth conditions shown. (b) On d3 of the differentiation assay, qRT-PCR was used to analyze the expression levels of 3 definitive endoderm markers (SOX17, FOXA2, CXCR4) in each of the cell types analyzed. SOX7 (extraembryonic endoderm-specific marker) expression was used to monitor the formation of definitive endoderm since SOX17, FOXA2 and CXCR4 can be found in all endodermal tissues. Low SOX7 levels indicated that the tissue generated was definitive endoderm and low OCT4 expression levels indicated the loss of pluripotency in iPS and ES cells during differentiation. (c) On d6, the expression level of 2 primitive gut tube markers, HNF1B and HNF4A, was analyzed by qRT-PCR. The same controls as on d3 were used in this analysis. In both graphs, values were standardized to Ubb and TBP, and the expression fold changes for all genes in each cell type are referenced to a corresponding d0 sample. In **b and c**, data are from triplicate reactions and error bars represent standard deviations.

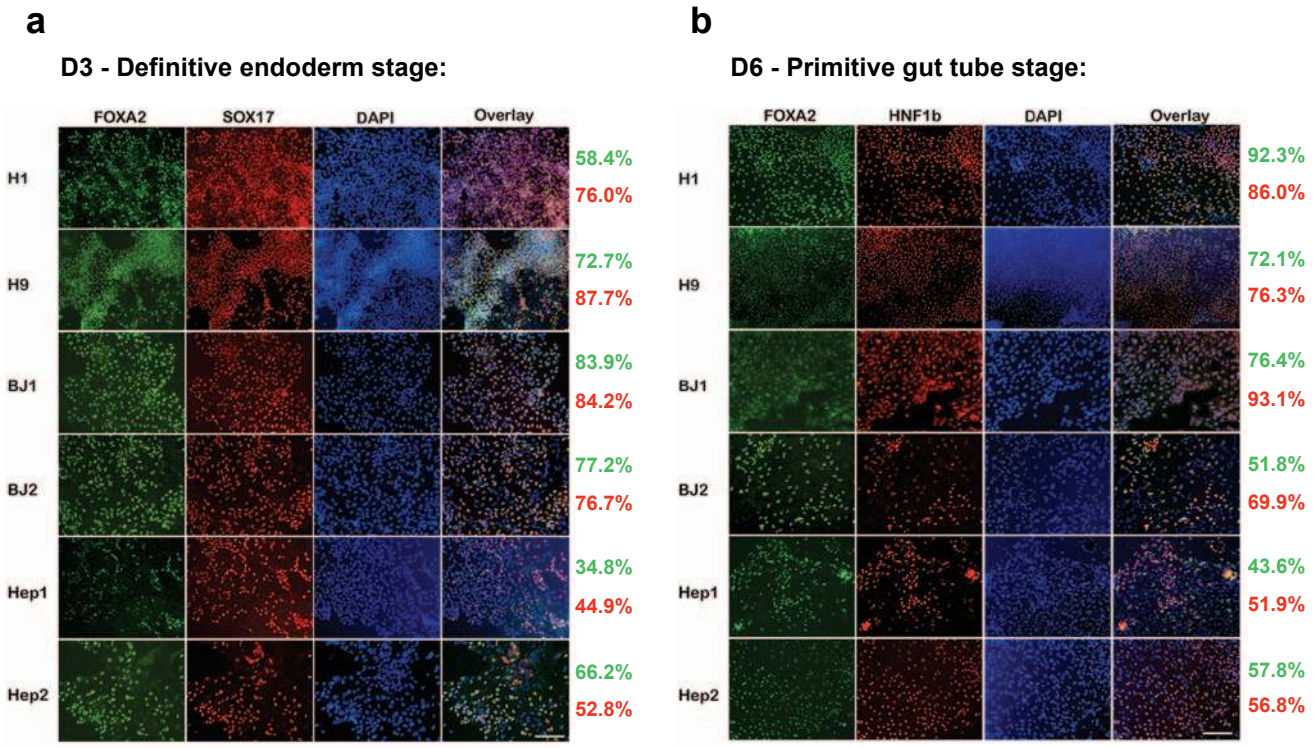


**Figure S5** Immunohistochemical analysis of iPS and ES cells differentiated towards endoderm. (a) On d3 of the targeted differentiation protocol, the same 2 clones of Fib-iPS (BJ 1 and 2) and Hep-iPS (Hep 1 and 2) cells, as well as 2 human ES cell lines (H1 and H9), analyzed by qRT-PCR in Supplementary Figure 4b were stained for the definitive endoderm markers FOXA2 and SOX17. The number of FOXA2- and SOX17-positive cells relative to the total number of DAPI-positive nuclei were quantified for each cell type analyzed. Values in green, percentage of FOXA2-positive cells; values in red, percentage of SOX17-postiive cells. (b) On d6 of the targeted differentiation protocol, the same cell lines were stained for FOXA2 and the primitive gut marker, HNF1b. The number of FOXA2- and HNF1b-positive cells relative to the total number of DAPI-positive nuclei were quantified for each cell type analyzed. Values in green, percentage of FOXA2-positive cells; values in red, percentage of HNF1b-postiive cells. Scale bars represent 200 μm.

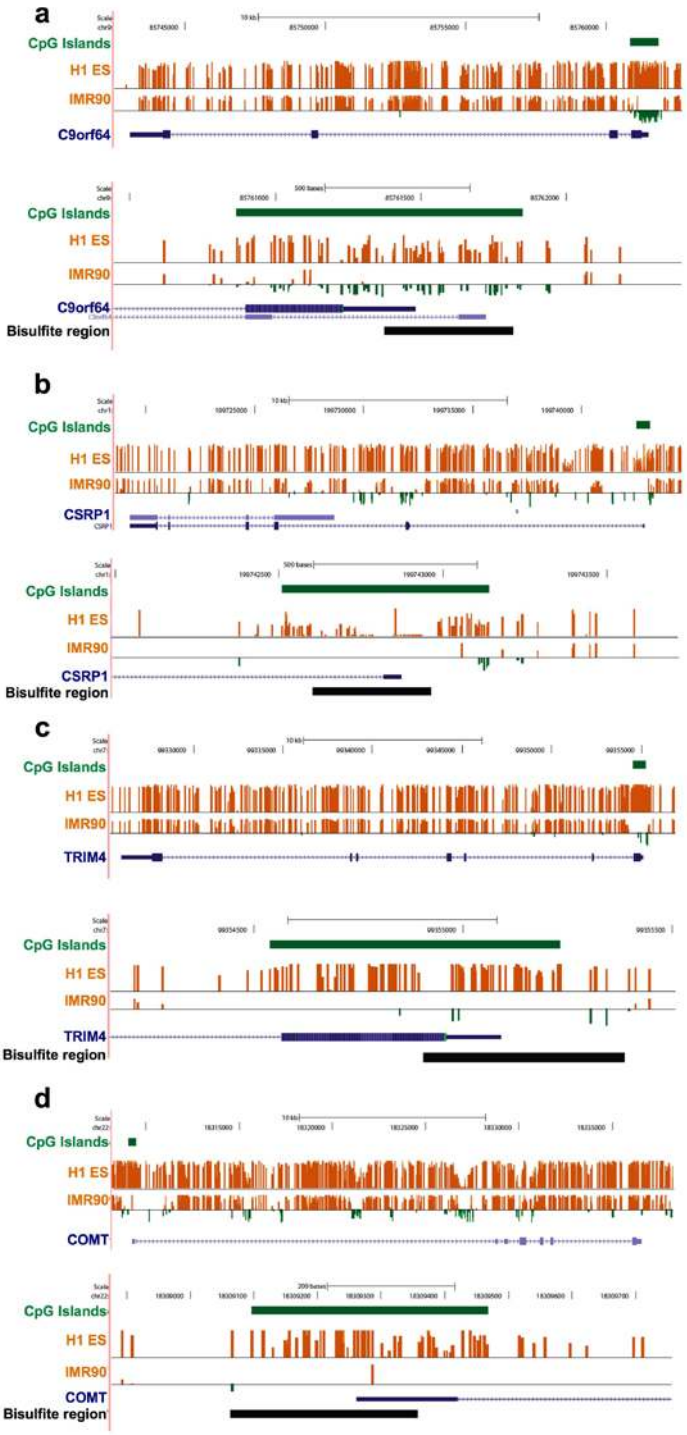


**Figure S6** Genome browser tracks of the 4 most incompletely silenced genes showing differential methylation between H1 ES cells and IMR90 cells. (a) Upper panel, a putative gene C9orf64 on chromosome 9, showing browser tracks for CpG islands, and DNA methylation levels for H1 ES and IMR90 cells assayed by MethylC-Seq (Lister et al. 2009). In the H1 ES and IMR90 cells methylation tracks, the Y-axis displays methylation scores of individual sites (CG, CHG and CHH). Methylation score is defined by the number of Cs and Ts at that position from MethylC-seq reads using the following formula: score = number of Cs / (C+T) * 1000 - 500. Data from both strands are combined. Scores range between -500 (unmethylated) and 500 (methylated), and the zero line is equivalent to 50% methylated. Negative scores are displayed as green bars and positive scores are displayed as orange bars. Lower panel, a close-up of the promoter near the region (black rectangle) that was assessed for methylation by bisulfite sequencing. (b-d) 3 additional genes which exhibit differential methylation between ES cells and IMR90 cells, and which lack epigenetic reprogramming in iPS cells (Supplementary Data 2) (b) CSRP1, (c) TRIM4, (d) COMT.

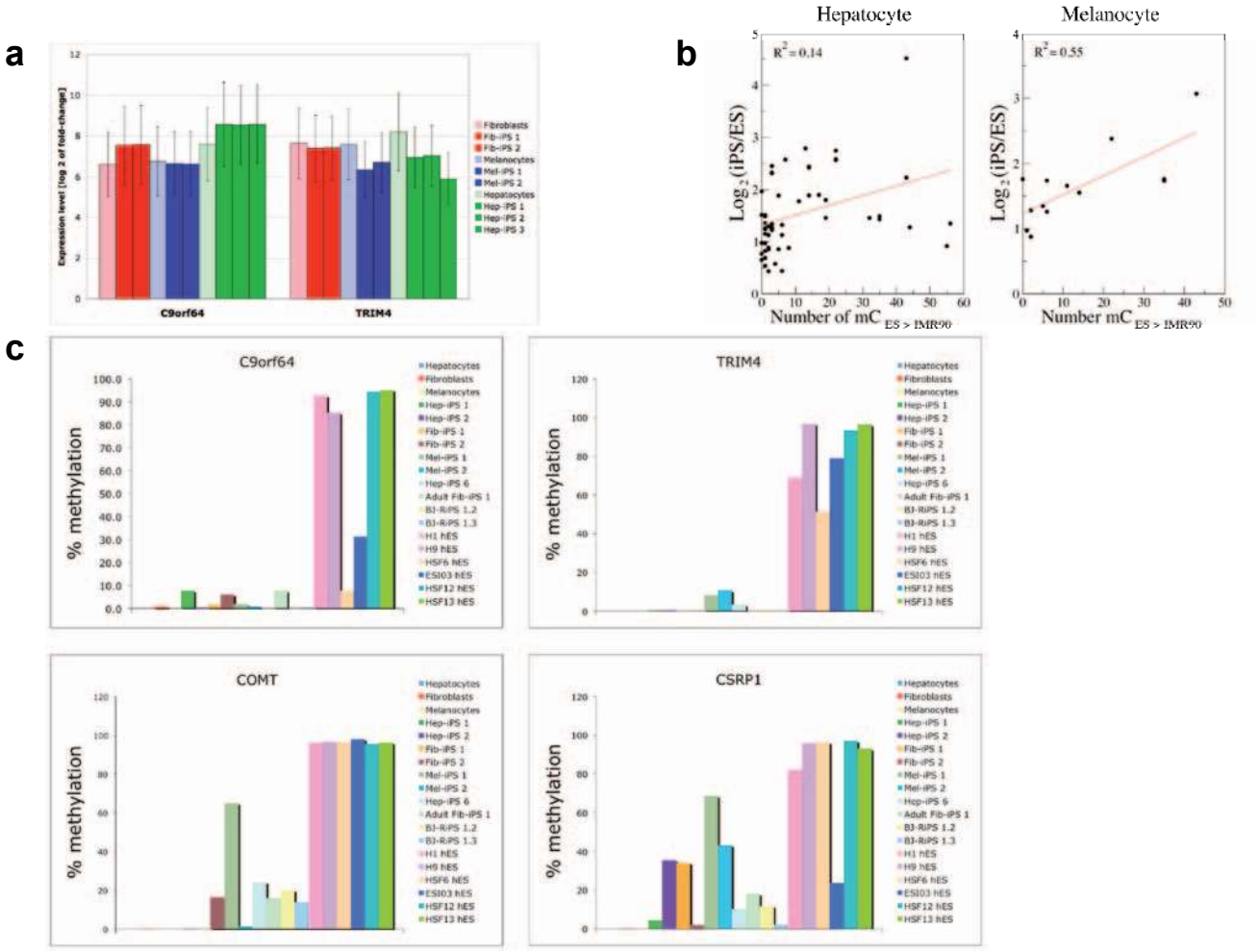


**Figure S7** DNA methylation and incompletely reprogrammed genes. (a) The high expression level of C9orf64 and TRIM4 in somatic cells and their iPS cell counterparts, relative to ES cells (in this case, H9 ES cells), was confirmed by qRT-PCR. Values were standardized to GAPDH and Ubb. Data are from triplicate reactions. Error bars represent standard deviations. (b) The genes which maintain higher expression levels in Hep-iPS cells and Mel-iPS cells compared to ES cells tend to be also methylated at higher levels in H1 ES cells compared to the fibroblast cell line IMR90. The Pearson correlation coefficient between the log expression fold change and single-nucleotide resolution differences in CpG island methylation was 0.37 and 0.74 for Hep-iPS cells and Mel-iPS cells, respectively. (c) We performed additional bisulfite sequence analysis in 4 human iPS cells (Hep-iPS 6, Adult Fib-iPS 1, BJ-RiPS 1.2 and BJ-RiPS 1.3), that were derived from different donor somatic cells and by reprogramming strategies different from those that were originally transcriptionally profiled by microarrays. Four additional ES cell lines (HSF6, ESI03, HSF12 and HSF13) were also analyzed, in parallel, as controls. The results are essentially the same as depicted in Figure 4c, while there is some variability in C9orf64 in ES cells, indicating that our findings are not clone- or methodology-dependent. Shown in the graphs is the % methylation observed at the promoters of C9orf64, TRIM4, CSRP1 and COMT in all the original cells that were transcriptionally-profiled by microarray, combined with the results from the newly analyzed iPS and ES cell lines.

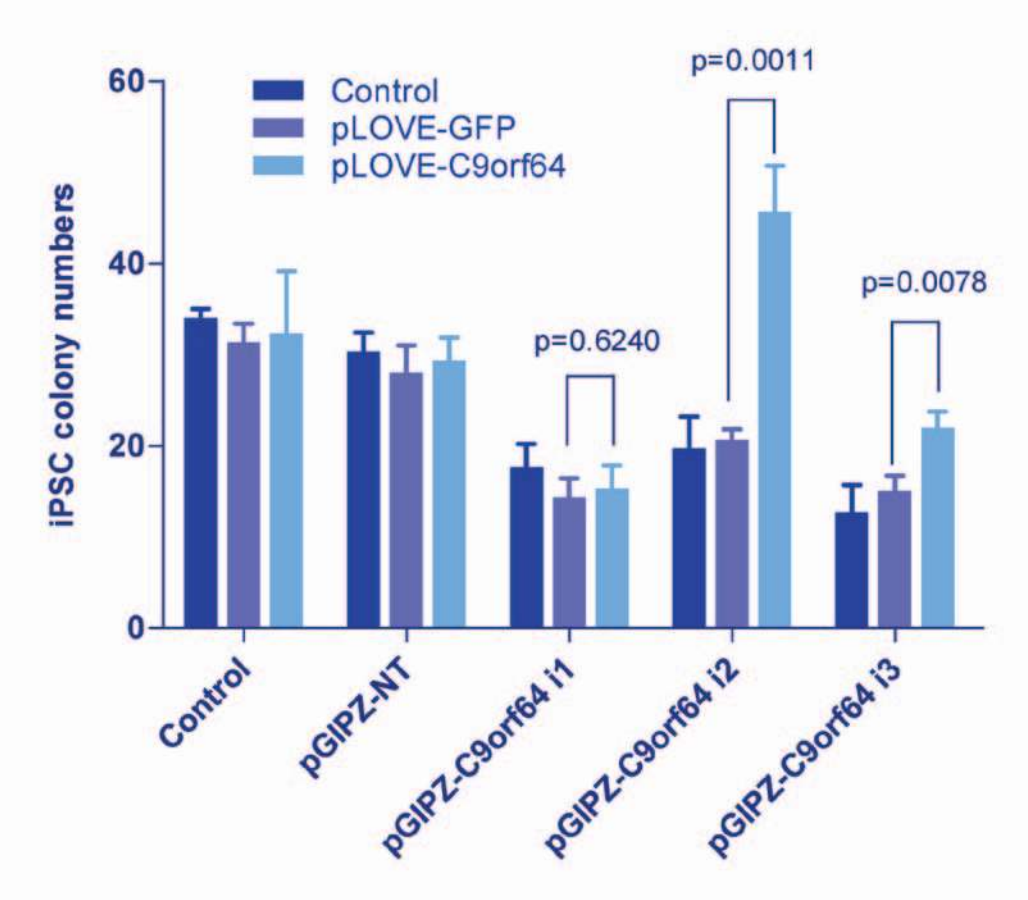


**Figure S8** C9orf64 overexpression rescues the deficiency in iPS colony number that results from C9orf64 inhibition. The number of iPS colonies was counted on d21 after infection of BJ foreskin fibroblasts with 4f alone (control), 4f+non-targeting shRNA (pGIPZ-NT) and 4f+C9orf64 shRNA (pGIPZ-C9orf64 i1/2/3). For each condition, C9orf64 was also overexpressed by pLOVE-C9orf64 lentivirus infection, and pLOVE-GFP was used as a negative control. Infections were performed in triplicate, and error bars represent standard deviations. C9orf64 overexpression resulted in a significant rescue of the reduction in number of iPS colonies by C9orf64 shRNA2 and shRNA3, which target the UTR region of the C9orf64 mRNA. No rescue was observed for C9orf64 shRNA1, which targets the ORF region of the C9orf64 mRNA.

**Supplementary Table 1**

| Gene | RefSeq | $Log_2$(Fib/ES) expression change | $CpG_{ES>IMR90}$ | $CpG_{IMR90>ES}$ |
|---|---|---|---|---|
| C9orf64 | NM_032307 | 3.971 | 43 | 0 |
| TRIM4 | NM_033017 | 2.437 | 35 | 0 |
| DNAJA4 | NM_001130182 | 1.521 | 14 | 0 |
| IQCA1 | NM_024726 | 2.141 | 11 | 0 |
| COMT | NM_000754 | 1.772 | 6 | 0 |
| CES1 | NM_001025194 | 1.079 | 5 | 0 |

**Supplementary Table 1. 6 genes differentially expressed between ESC and all 3 iPSC. DEDS 5% FDR cutoff was used to determine differential expression.**

**Supplementary Table 2**

| Gene name | Coordinates(hg18 version) | Bisulfite primer names | Sequence |
|---|---|---|---|
| C9orf64 | chr9:85,761,380-85,761,819 | Forward | TGTAGTTAAGGTAAAGGTTTTTTTT |
| | | Reverse | ACTCAATCCTCAACACCCAAATCTAC |
| CSRP1 | chr1:199,742,606-199,742,959 | Forward | GTGTTTAGGAAGTTTAGGAAGGTT |
| | | Reverse | CAATATACAAAACCCACTAATTAAC |
| TRIM4 | chr7:99,354,907-99,355,386 | Forward | ATAGTTTAGGTAGATGGGGTAGGTTAATTT |
| | | Reverse | CCTAAACCCTCAAACTTAAAAAAAA |
| COMT | chr22:18,309,072-18,309,357 | Forward | TTTGAGTAAGATTAGATTAAGAGGT |
| | | Reverse | ACAACCCTAACTACCCCAAAAACCC |

**Supplementary Table 2. Sequences for bisulfite primers used for methylation analysis.**

**Supplementary Table 3**

| Gene name | Forward primer sequence | Reverse primer sequence |
|---|---|---|
| GAPDH | CAATGACCCCTTCATTGACC | GACAAGCTTCCCGTTCTCAG |
| Ubb | TTGTTGGGTGAGCTTGTTTG | GTCTTGCCGGTAAGGGTTTT |
| TBP | TGTGCACAGGAGCCAAGAGT | ATTTTCTTGCTGCCAGTCTGG |
| C9orf64 | AGTGGGTACTGGTCCCTGTG | GTCGCGTAGTACGAGGCACT |
| Endogenous OCT4 | TGTACTCCTCGGTCCCTTTC | TCCAGGTTTTCTTTCCCTAGC |
| Endogenous SOX2 | GCTAGTCTCCAAGCGACGAA | GCAAGAAGCCTCTCCTTGAA |
| Endogenous cMYC | CGGAACTCTTGTGCGTAAGG | CTCAGCCAAGGTTGTGAGGT |
| Endogenous KLF4 | TATGACCCACACTGCCAGAA | TGGGAACTTGACCATGATTG |
| Endogenous NANOG | CAGTCTGGACACTGGCTGAA | CTCGCTGATTAGGCTCCAAC |
| Lentiviral OCT4 | CCCCTGTCTCTGTCACCACT | CCACATAGCGTAAAAGGAGCA |
| Lentiviral SOX2 | ACACTGCCCCTCTCACACAT | CATAGCGTAAAAGGAGCAACA |
| Lentiviral cMYC | AAGAGGACTTGTTGCGGAAA | TTGTAATCCAGAGGTTGATTATCG |
| Lentiviral KLF4 | GACCACCTCGCCTTACACAT | CATAGCGTAAAAGGAGCAACA |
| Lentiviral NANOG | ACATGCAACCTGAAGACGTG | CACATAGCGTAAAAGGAGCAA |
| TRIM4 | GAAGTGAAGAACGCCACACA | TCAACCAGGAAGTTGTGCAG |
| SOX17 | GGCGCAGCAGAATCCAGA | CCACGACTTGCCCAGCAT |
| FOXA2 | GGGAGCGGTGAAGATGGA | TCATGTTGCTCACGGAGGAGTA |
| MSX1 | CGAGAGGACCCCGTGGATGCAGAG | GGCGGCCATCTTCAGCTTCTCCAG |
| BRACHYURY | TGCTTCCCTGAGACCCAGTT | GATCACTTCTTTCCTTTGCATCAAG |
| NCAM | AGGAGACAGAAACGAAGCCA | GGTGTTGGAAATGCTCTGGT |
| SOX1 | ATGCACCGCTACGACATGG | CTCATGTAGCCCTGCGAGTTG |
| OCT4$^{T}$ | GCATAGTCGCTGCTTGATCG | TGGGCTCGAGAACCATGTG |
| CXCR4 | CACCGCATCTGGAGAACCA | GCCCATTTCCTCGGTGTAGTT |
| SOX7 | ACGCCGAGCTCAGCAAGAT | TCCACGTACGGCCTCTTCTG |
| HNF1B | TCACAGATACCAGCAGCATCAGT | GGGCATCACCAGGCTTGTA |
| HNF4A | CATGGCCAAGATTGACAACCT | TTCCCATATGTTCCTGCATCAG |

**Supplementary Table 3. Sequences for qRT-PCR primers used in our study.**

*OCT4$^{T}$ primer set was used only for the targeted differentiation analysis.