Transcriptional profiling of cells sorted by RNA abundance

Sandy Klemm[1,8], Stefan Semrau[2,8], Kay Wiebrands[3,4,8], Dylan Mooijman[3,4], Dina A Faddah[5,6], Rudolf Jaenisch[5,6] & Alexander van Oudenaarden[2–5,7]

We have developed a quantitative technique for sorting cells on the basis of endogenous RNA abundance, with a molecular resolution of 10–20 transcripts. We demonstrate efficient and unbiased RNA extraction from transcriptionally sorted cells and report a high-fidelity transcriptome measurement of mouse induced pluripotent stem cells (iPSCs) isolated from a heterogeneous reprogramming culture. This method is broadly applicable to profiling transcriptionally distinct cellular states without requiring antibodies or transgenic fluorescent proteins.

A common challenge in biology is to identify and isolate transcriptionally distinct subpopulations within a single tissue or cell type. Although a variety of techniques have been developed to discriminate among these alternative expression modes, the most widely used methods require either transgenic integration of fluorescent protein reporters or the availability of specific antibodies[1–3]. These approaches, however, are precluded for biological systems that are refractory to genetic manipulation (for example, primary human tissue) and for processes in which RNA— rather than protein—is the key discriminative marker (for example, noncoding RNA). Recently, flow cytometry has been used to sort cells using a spectrum of fluorescence labeling techniques in which oligonucleotide probes are hybridized to either DNA or RNA target sequences[4–7]. The principle limitation of these methods has been that RNA extracted from hybridized material is often highly degraded[8,9]. Although reverse transcription and quantitative PCR (RT-qPCR) has previously been reported for hybridized cells[10,11], unbiased transcriptome measurements require full-length RNA extraction. RNA degradation is partially mitigated by labeling RNA in live cells[7,12], but extended hybridization in *ex vivo* culture may obscure the molecular state of primary tissue. Given these limitations, we have developed a method for RNA labeling in cross-linked cells that permits full-length RNA isolation and unbiased transcriptional profiling.

The proposed RNA cell-sorting technique uses flow cytometry to measure the fluorescence of individual cells labeled with a single-molecule RNA FISH (smFISH) probe library[13,14]. As a proof-of-principle experiment, GFP transcripts were fluorescently labeled in cells that expressed the transgene under doxycycline control[15] (**Supplementary Fig. 1a–c**). To assess the sorting potential of the labeled RNA signal, we measured single-cell RNA fluorescence distributions by flow cytometry, which revealed a clear separation of high- and low-induction profiles (**Fig. 1a** and **Supplementary Fig. 1b**,**c**). Furthermore, we found the measured mRNA fluorescence to scale linearly with mRNA and protein abundance across a broad range of induction levels (**Fig. 1b**). We further confirmed the linearity of the labeled RNA fluorescence signal for a panel of endogenous genes by comparing the mean flow cytometry signal intensity with the average number of RNA molecules in iPSCs quantified by single-cell transcript counting[14] (**Fig. 1c**). We note that the absolute number of gene transcripts expressed in stem cells is dependent on both genetic background and medium conditions (**Supplementary Fig. 2**).

We then asked whether the observed RNA fluorescence signal provides an accurate measurement of single-cell transcript levels. For this purpose, we measured both GFP protein and labeled mRNA fluorescence in single cells and found a strong correlation (Pearson's correlation coefficient (ρ) = 0.77; **Supplementary Fig. 1d**), which is consistent with the direct dependence of protein production on mRNA abundance. Additionally, the correlation between mRNA and protein was confirmed across a broad range of GFP induction levels (**Supplementary Fig. 1e**). Next we tested the single-cell precision of the proposed RNA measurement by differentially labeling the 3′ and 5′ ends of a single transcriptional target, *Oct4*-IRES-GFP fusion mRNA (where *Oct4* is the gene *Pou5f1* and IRES is an internal ribosome entry site), and found that the labels were strongly correlated at the single-cell level (ρ = 0.90; **Supplementary Fig. 3a**). Furthermore, we separately labeled *Oct4* in mouse embryonic stem cells (mESCs) with either Alexa 594 or Cy5 fluorophores. We then mixed the differentially labeled cells and confirmed by flow cytometry that the two subpopulations were strongly anticorrelated (ρ = −0.81; **Supplementary Fig. 3b**). Having established measurements of both positive and negative correlation, we hypothesized that doxycycline-induced GFP expression would in principle be uncorrelated with every endogenous mRNA species in the genome. We confirmed this prediction across all measured GFP induction levels for the gene *Oct4* (ρ = 0.1; **Supplementary Fig. 3c**). Finally, we sorted cells above and below each quartile of the *Sox2* RNA fluorescence

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Hubrecht Institute – Royal Netherlands Academy of Arts and Sciences, Utrecht, The Netherlands. [4]University Medical Center Utrecht, Utrecht, The Netherlands. [5]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA. [7]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to S.K. (klemm@mit.edu) or A.v.O. (a.vanoudenaarden@hubrecht.eu).

RECEIVED 14 AUGUST 2013; ACCEPTED 20 FEBRUARY 2014; PUBLISHED ONLINE 30 MARCH 2014; DOI:10.1038/NMETH.2910

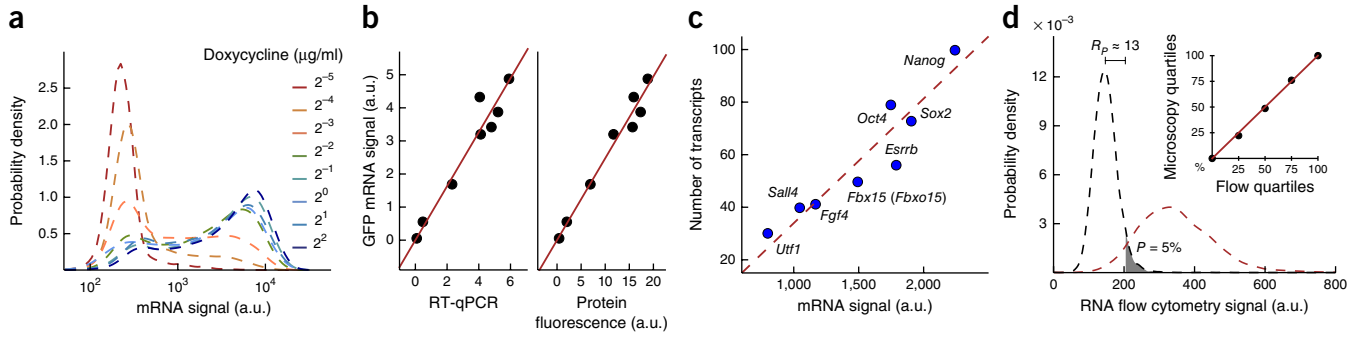

Figure 1 | Quantitative single-cell measurement of transcription. (**a**) Single-cell distribution of GFP mRNA fluorescence for indicated doxycycline induction levels. Dashed lines reflect kernel density estimates of the signal distribution for each induction level. (**b**) Linear and unbiased scaling of GFP mRNA fluorescence with RT-qPCR and GFP protein fluorescence. (**c**) Linear scaling of mean mRNA fluorescence with single-cell transcript quantification by classical (microscopy) smFISH for a panel of endogenous genes in iPSCs grown in 2i medium. (**d**) *Sox2* mRNA signal (red) and background (black) with shaded 95% quantile (gray) and molecular resolution R_P are reported for the $P = 5\%$ significance level. Inset, recapitulation of first (25%), second (50%) and third (75%) quartile sorting by classical smFISH transcript counting. Each experiment was performed once. a.u., arbitrary units.

distribution. Single-cell transcript counting in each fraction quantitatively recapitulated the flow cytometry quartile measurements (**Fig. 1d**). Taken together, these data suggest that flow cytometry can be used to quantitatively measure the abundance of mRNA in single cells hybridized with a complementary smFISH library.

To assess the resolution of the labeled RNA signal, we measured the expected number of transcripts required for a cell to be statistically resolved from the background (Online Methods). We measured a resolution of 57 transcripts for a library of 20-nt probes hybridized to the *Sox2* gene in mouse iPSCs (**Supplementary Fig. 4a**). Under more stringent hybridization conditions, a 30-nt library improved the resolution to 24 transcripts (**Supplementary Fig. 4b**). We confirmed this estimate by sorting cells and directly measuring the difference in the number of transcripts required for a pair of cells to be resolved with 95% confidence (**Supplementary Fig. 4c**,**d**). The sorting resolution was further improved by measuring RNA abundance and sorting in specific cell-cycle phases (**Supplementary Fig. 5a**,**b**). For G1 cells, we estimated the molecular resolution as 30 molecules for the 20-nt probe library and 13 molecules for the 30-nt library (**Fig. 1d** and **Supplementary Fig. 5c**,**d**).

In order to measure the transcriptome of sorted subpopulations, RNA is extracted from hybridized cells by cross-link reversal (Online Methods). Under standard hybridization conditions[14], the molecular integrity of the extracted RNA is attenuated in a time- and temperature-dependent manner, owing to enzymatic RNA degradation as well as hydrolysis-mediated RNA fragmentation (**Supplementary Fig. 6a**). We have addressed this by developing an RNA-preserving hybridization buffer (RPHB) that facilitates efficient isolation of full-length RNA following hybridization (**Supplementary Fig. 6b**). RPHB employs a nearly saturating salt concentration to eliminate enzymatic RNA degradation by precipitating proteins, and it includes a high concentration of the chelating agent EDTA, which inhibits RNA degradation by sequestration of metal ions involved in RNA hydrolysis. We tested RPHB by extracting RNA from live and RPHB-hybridized mESCs and comparing relative expression levels by RT-qPCR and microarrays (**Fig. 2a**,**b**), establishing that RNA extracted from RPHB-hybridized material is quantitatively unbiased. Furthermore, we found that the error distribution between technical microarray replicates was identical for live and hybridized samples (**Supplementary Fig. 6c**). We then examined the genome-wide expression fold change between NIH-3T3 fibroblasts and J1 mESCs. A comparison of fold-change measurements for both live and hybridized samples revealed a strong correlation ($\rho = 0.94$) over the full dynamic range of the microarray (**Fig. 2c**). Finally, we isolated mESCs and fibroblasts by RNA FACS from an artificial mixture of these cell types; subpopulation transcriptome measurements on the sorted fractions recapitulated the respective cell-type signatures (**Supplementary Figs. 6d** and **7**).

One of the motivating applications for RNA-based sorting has been to transcriptionally profile iPSCs during the process of cellular reprogramming[16]. Following disruption of the somatic state during reprogramming, individual cells stochastically reactivate the pluripotency machinery at widely different rates[17] and contribute to a diverse collection of coexisting subpopulations[18]. Whereas iPSCs are independent of ectopic reprogramming factors, incompletely reprogrammed cells require sustained reprogramming factor expression to be competitively maintained in culture. To interrogate these subpopulations, we isolated cells that had reactivated the endogenous *Sox2* locus—an established reprogramming marker[18,19]—from a background of partially reprogrammed cells. The *Sox2*$^+$ and *Sox2*$^-$ transcriptomes were then compared with iPSCs derived by reprogramming factor withdrawal.

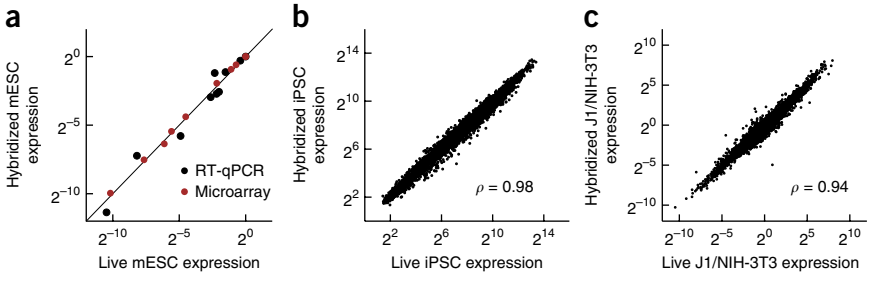

Figure 2 | RNA extraction and transcriptional sorting. (**a**) Gene expression (normalized to *Gapdh*) for *Nanog*, *Sox2*, *Klf2*, *Rnh* (*Rnh1*), *Znf7* (*Zfp7*), *Rex1* (*Zfp42*), *Stella* (*Dppa3*), *Tubb5* and *Ubc* as measured by RT-qPCR (black) and microarray (red) in hybridized mESCs and in live mESCs. (**b**) Genome-wide microarray expression measurements for live and hybridized cells. (**c**) Fold change in genome-wide expression between NIH-3T3 cells and J1 mESCs for hybridized and live samples. Each experiment was performed once.

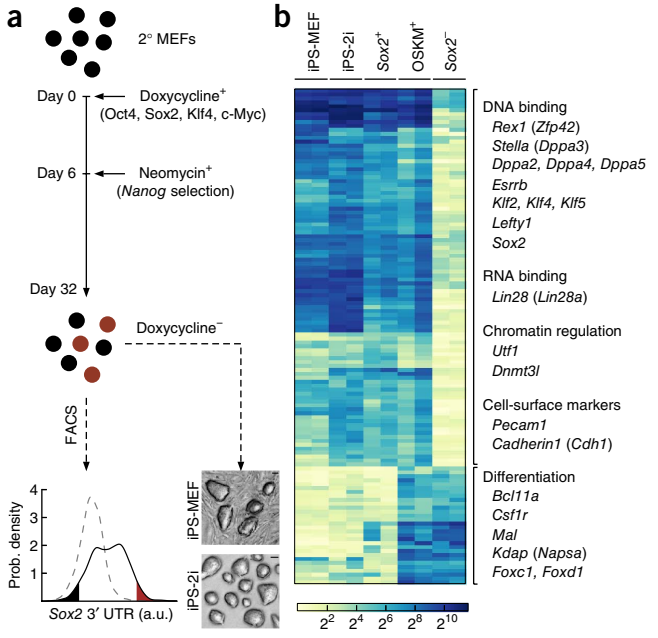

Figure 3 | Isolation and transcriptional profiling of iPSCs. (**a**) Secondary MEF reprogramming by OSKM expression and *Nanog* selection (Online Methods); isolation of iPSCs by OSKM withdrawal in MEF coculture (iPS-MEF) and 2i medium (iPS-2i); and RNA FACS for reprogrammed *Sox2*$^+$ (brown) and nonreprogrammed *Sox2*$^-$ (black) cells (scale bars, 25 μm). The plot shows bimodal *Sox2* 3′ UTR signal (black) relative to the nonspecific background signal (gray) in day 32 OSKM$^+$ cells. a.u., arbitrary units. (**b**) Hierarchical clustering of genes differentially expressed between *Sox2*$^+$ and *Sox2*$^-$ cells (top 5% shown for each of two biological replicates).

Secondary mouse embryonic fibroblasts (2° MEFs; Online Methods) were reprogrammed by exposure to doxycycline-induced expression of Oct4, Sox2, Klf4 and c-Myc (OSKM) for 32 consecutive days (**Fig. 3a**), at which point iPSCs were passaged in coculture with MEFs (iPS-MEF) and in feeder-free 2i medium (iPS-2i) as OSKM-independent colonies (**Fig. 3a**). The distribution of endogenous *Sox2* expression, measured by flow cytometry using an smFISH probe library designed against the noncoding 3′ UTR of *Sox2*, was bimodal (**Fig. 3a**) for day 32 OSKM$^+$ cells, reflecting an underlying diversity of reprogramming depth among individual cells. The upper and lower *Sox2* expression tails were sorted and transcriptionally profiled in comparison with OSKM-independent iPSCs, revealing an unambiguous pluripotency signature for the positive fraction (**Fig. 3b**). Although many somatic marks were repressed in both *Sox2*$^+$ and *Sox2*$^-$ cells, a broad spectrum of iPSC-specific genes were differentially upregulated in *Sox2*$^+$ cells, including those encoding transcription factors, RNA-binding proteins, chromatin regulators and cell-surface markers. Reciprocally, *Sox2*$^-$ cells expressed a class of differentiation-associated genes that were repressed in iPSCs. The striking similarity between *Sox2*$^+$ cells and iPSCs suggests that cells within the *Sox2*$^+$ subpopulation are reprogrammed and give rise to stable iPSCs under OSKM withdrawal. By leveraging a noncoding transcriptional element (*Sox2* 3′ UTR), these experiments illustrate the flexibility of RNA FACS and suggest the potential for a broader subpopulation analysis of cellular reprogramming.

Single-cell heterogeneity in gene expression is a common phenomenon for a variety of developmental and homeostatic processes. The principle focus of this work has been to develop a fluorescent measure of RNA in single cells, which facilitates high-resolution sorting as well as efficient and unbiased RNA isolation. This technique extends flow cytometry to a new class of applications based on direct quantification of RNA.

Accession codes. NCBI Gene Expression Omnibus: microarray data generated in this study are available under accession numbers GSE55671, GSE55672 and GSE55919.

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Lou for performing the microarray experiments and M. Bienko, N. Crosetto and N. Slavov for critical reading of the manuscript. This work was supported by the US National Institutes of Health (NIH) National Cancer Institute Physical Sciences Oncology Center at Massachusetts Institute of Technology (U54CA143874), an NIH Pioneer award (8 DP1 CA174420-05), a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award to A.v.O. and an NWO Rubicon award to S.S. R.J. was supported by NIH grants HD 045022 and R37CA084198. D.A.F. was supported by a Vertex Scholarship, a US National Science Foundation Graduate Research Fellowship and Jerome and Florence Brill Graduate Student Fellowship. Support for S.K. was provided by the Koch Institute for Integrative Cancer Research Graduate Fellowship.

AUTHOR CONTRIBUTIONS

A.v.O., S.K. and S.S. developed the idea of transcriptionally profiling RNA-sorted cells. S.S. and S.K. demonstrated the compatibility of smFISH with flow cytometry. S.K. performed all of the experiments; K.W. collaborated on the RNA integrity measurements in **Supplementary Figure 4** and the reverse cross-linking controls in **Figure 2**. S.K., S.S., K.W. and A.v.O. developed the reverse cross-linking protocol. S.K. and D.M. optimized the RNA flow sorting procedure. S.K. conceived and experimentally validated the RNA-preserving hybridization buffer (RPHB). D.A.F. and R.J. produced the 2° reprogrammable MEFs. S.K. analyzed the data, developed the analytic estimate of the molecular resolution, prepared the figures, and wrote the manuscript in collaboration with A.v.O., who guided the project. All authors read and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. *Science* **263**, 802–805 (1994).
2. Chambers, I. *et al. Nature* **450**, 1230–1234 (2007).
3. Dietrich, J.-E. & Hiiragi, T. *Development* **134**, 4219–4231 (2007).
4. Rufer, N., Dragowska, W., Thornbury, G., Roosnek, E. & Lansdorp, P. *Nat. Biotechnol.* **16**, 743–747 (1998).
5. Prigodich, A.E. *et al. Anal. Chem.* **84**, 2062–2066 (2012).
6. Robertson, K.L., Verhoeven, A.B., Thach, D.C. & Chang, E.L. *RNA* **16**, 1679–1685 (2010).
7. Rhee, W.J. & Bao, G. *BMC Biotechnol.* **9**, 30 (2009).
8. Masuda, N., Ohnishi, T., Kawamoto, S., Monden, M. & Okubo, K. *Nucleic Acids Res.* **27**, 4436–4443 (1999).
9. Specht, K. *et al. Am. J. Pathol.* **158**, 419–429 (2001).
10. Yamada, H. *et al. Cytometry A* **77**, 1032–1037 (2010).
11. Maruo, R. *et al. Mol. Biotechnol.* **49**, 42–47 (2011).
12. Larsson, H.M. *et al. PLoS ONE* **7**, e49874 (2012).
13. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. *Science* **280**, 585–590 (1998).
14. Raj, A., Bogaard, P.V.D., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. *Nat. Methods* **5**, 877–879 (2008).
15. Beard, C., Hochedlinger, K., Plath, K., Wutz, A. & Jaenisch, R. *Genesis* **44**, 23–28 (2006).
16. Takahashi, K. & Yamanaka, S. *Cell* **126**, 663–676 (2006).
17. Hanna, J. *et al. Nature* **462**, 595–601 (2009).
18. Buganim, Y. *et al. Cell* **150**, 1209–1222 (2012).
19. Golipour, A. *et al. Cell Stem Cell* **11**, 769–782 (2012).

ONLINE METHODS

Molecular resolution. The molecular resolution is defined as the expected number of fluorescently labeled mRNA molecules required for a cell to be statistically resolved from the background. The resolution depends on the measured signal, the distribution of mRNA, and the background fluorescence. The transcript distribution was measured by counting smFISH labeled mRNA in single cells as previously described[14]. The background fluorescence—which accounts for both cellular autofluorescence and nonspecific probe binding—was estimated by hybridizing a mock sample with a 1:20 mixture of fluorescently labeled and unlabeled probe libraries (yielding a 95% attenuation of the specific signal). The resolution R_P was calculated for the significance level P under the null distribution given by the mock signal (**Supplementary Note**).

FISH buffers and probes. Oligonucleotide libraries with 20-nt probes were designed and fluorescently labeled as previously described[14]. Probes for the 30-nt *Sox2* library were similarly designed such that the predicted melting temperature of individual probes deviates from the median by no more than 5° C. See **Supplementary Data** for all probe libraries used in this study. The following buffers were used in this study. RPHB: 300 mM NaCl, 30 mM sodium citrate, 2.1 M ammonium sulfate, 10 mM EDTA, 1 mg/ml *Escherichia coli* tRNA, 500 µg/ml BSA, 25% (40%, v/v) formamide for 20- (30-)nt probe library, pH 5.2; wash buffer: 25% (40%, v/v) formamide for 20- (30-)nt probe library, 2× SSC; flow buffer: 2× SSC; sorting buffer: 200 mM NaCl, 20 mM sodium citrate, 1.5 M ammonium sulfate, 5 mM EDTA, pH 5.2, 2× SSC; reverse cross-linking buffer: 100 mM NaCl, 10 mM pH 8.0 Tris, 1 mM EDTA, 0.5% SDS (v/v), 500 µg/ml proteinase K. RPHB incorporates components of both RNALater Solution (Ambion) as well as the previously reported smFISH hybridization buffer[14], which was used to measured the dependence of RNA degradation on hybridization conditions (**Supplementary Fig. 4a**).

RNA FISH. Cells were fixed in 4% paraformaldehyde for 5 min, centrifuged at 1,000*g* for 5 min, and washed in 70% ethanol (EtOH). Following overnight ethanol permeabilization (70% EtOH) at 4 °C, cells were resuspended in RPHB with labeled probes (0.5–1 ng/µl) and incubated at 30 °C for 12 h. Following hybridization, an equal volume of wash buffer was added and mixed thoroughly with each sample. Cells were then pelleted by centrifugation and resuspended in wash buffer for 30 min at 30 °C. After the previous wash step was repeated, cells were resuspended in flow buffer and maintained at 4 °C in preparation for sorting.

Flow cytometry, FACS and RNA extraction. Cells were sorted by FACS into 4 °C sorting buffer using a BD Biosciences Aria II flow cytometer. Analytic flow cytometry measurements were performed on a BD Biosciences LSR Fortessa platform. Unless otherwise noted, cell size was controlled for by costaining cells with Hoechst 33342 (Invitrogen) and selecting for diploid DNA content. FACS-sorted cells were centrifuged as before and resuspended in reverse cross-linking buffer at 50 °C for 1 h. Total RNA was isolated by phenol-chloroform (Trizol, Invitrogen) using the manufacturer's protocol.

Microarrays. Microarray assays were performed by the Massachusetts Institute of Technology (MIT) BioMicro Center. NIH-3T3 fibroblasts and mESCs were assayed using Eukaryotic Exon 1.0 ST arrays from Agilent, whereas Mouse 430A 2.0 Affymetrix chips were used for MEFs and iPSCs.

Cell lines and media. Embryonic cells and iPSCs were grown as indicated: (i) cocultured with irradiated MEF cells (Global Stem) or (ii) cultured in 2i conditions with both glycogen synthase kinase 3β (Stemgent) and mitogen-activated protein kinase kinase inhibitors (Stemgent). Embryonic stem cells and iPSCs were grown with leukemia inhibitory factor (10^3 units/ml, Millipore) and 15% (10% for E14 cells) heat-inactivated FBS (Hyclone) together with Knockout DMEM (Gibco), L-glutamine (Gibco), MEM non-essential amino acids (Gibco), penicillin-streptomycin (Gibco) and β-mercaptoethanol (Sigma). NIH-3T3 fibroblast and primary MEF cells were cultured without inhibitors in 10% serum. J1 (2i medium) and E14 mESCs were used for the post-hybridization RNA extraction controls, and J1 mESCs (2i medium) were used for the NIH-3T3–mESC sorting experiments. The GFP induction experiments were performed using KH2:GFP cells[15] with constitutive R26 M2rtTA expression and a tetracycline-inducible EGFP construct targeted to the *ColA1* locus. Reprogrammable secondary MEFs and KH2:GFP cells were provided by the lab of R.J. and were recently authenticated and confirmed to be mycoplasma negative.

Secondary somatic cell generation and reprogramming. Secondary mouse embryonic fibroblasts (2° MEFs) were isolated from chimeric embryos as previously described[20], providing doxycycline-inducible Oct4, Sox2, Klf4 and c-Myc expression; neomycin selection at the *Nanog* locus; and a fluorescent transcriptional reporter for Oct4. The parental pluripotent stem cells used to generate the secondary line were derived by replacing an *Oct4* allele with an *Oct4*-IRES-GFP sequence in *Nanog*-Neo iPSCs[20]. Secondary MEFs were plated at optimal density[20] and passaged after 48 h. Doxycycline (2 µg/ml, Stemgent) was added 24 h after replating, marking the start of reprogramming. Neomycin selection (Stemgent G418, 1 µg/ml) was applied beginning at day 6 and maintained throughout the reprogramming time course. Reprogrammed iPSCs were stabilized in coculture with MEFs (iPS-MEF) and in feeder-free conditions (iPS-2i).

20. Wernig, M. *et al. Nat. Biotechnol.* **26**, 916–924 (2008).

doi:10.1038/nmeth.2910