Article

Reprogramming roadmap reveals route to human induced trophoblast stem cells

https://doi.org/10.1038/s41586-020-2734-6

Received: 5 February 2019

Accepted: 24 June 2020

Published online: 16 September 2020

Check for updates

Xiaodong Liu[1,2,3,19], John F. Ouyang[4,19], Fernando J. Rossello[1,2,3,16,19], Jia Ping Tan[1,2,3], Kathryn C. Davidson[1,2,3], Daniela S. Valdes[1,2,3], Jan Schröder[1,2,3], Yu B. Y. Sun[1,2,3], Joseph Chen[1,2,3], Anja S. Knaupp[1,2,3], Guizhi Sun[1,2,3], Hun S. Chy[3,5], Ziyi Huang[3,5], Jahnvi Pflueger[6,7], Jaber Firas[1,2,3], Vincent Tano[1,2,3], Sam Buckberry[6,7], Jacob M. Paynter[1,2,3], Michael R. Larcombe[1,2,3], Daniel Poppe[6,7], Xin Yi Choo[1,2,3], Carmel M. O'Brien[3,5], William A. Pastor[8,9,17], Di Chen[8,9], Anna L. Leichter[10], Haroon Naeem[11], Pratibha Tripathi[1,2], Partha P. Das[1,2], Alexandra Grubman[1,2,3], David R. Powell[11], Andrew L. Laslett[3,5], Laurent David[12,13], Susan K. Nilsson[3,5], Amander T. Clark[8,9,14,15], Ryan Lister[6,7], Christian M. Nefzger[1,2,3,18], Luciano G. Martelotto[10], Owen J. L. Rackham[4]✉ & Jose M. Polo[1,2,3]✉

The reprogramming of human somatic cells to primed or naive induced pluripotent stem cells recapitulates the stages of early embryonic development[1–6]. The molecular mechanism that underpins these reprogramming processes remains largely unexplored, which impedes our understanding and limits rational improvements to reprogramming protocols. Here, to address these issues, we reconstruct molecular reprogramming trajectories of human dermal fibroblasts using single-cell transcriptomics. This revealed that reprogramming into primed and naive pluripotency follows diverging and distinct trajectories. Moreover, genome-wide analyses of accessible chromatin showed key changes in the regulatory elements of core pluripotency genes, and orchestrated global changes in chromatin accessibility over time. Integrated analysis of these datasets revealed a role for transcription factors associated with the trophectoderm lineage, and the existence of a subpopulation of cells that enter a trophectoderm-like state during reprogramming. Furthermore, this trophectoderm-like state could be captured, which enabled the derivation of induced trophoblast stem cells. Induced trophoblast stem cells are molecularly and functionally similar to trophoblast stem cells derived from human blastocysts or first-trimester placentas[7]. Our results provide a high-resolution roadmap for the transcription-factor-mediated reprogramming of human somatic cells, indicate a role for the trophectoderm-lineage-specific regulatory program during this process, and facilitate the direct reprogramming of somatic cells into induced trophoblast stem cells.

Human embryonic stem (ES) cells are derived from the epiblast of pre-implantation blastocysts; alternatively, human induced pluripotent stem (iPS) cells are generated from adult cells such as fibroblasts through nuclear reprogramming mediated by transcription factors. Both cell types are pluripotent, as they can give rise to all cell types within the embryo but not the extra-embryonic tissues (that is, placenta). Conventionally, human ES and iPS cells are cultured in a primed state that resembles the post-implantation epiblast. However, recent advances in culture conditions have enabled the generation of naive ES and iPS cells that resemble the human pre-implantation epiblast, an earlier stage in embryonic development[1–3]. In contrast to the reprogramming of mouse somatic cells (for which comprehensive roadmaps of the reprogramming process have previously been reported[8–12]), few recent studies have revealed the details of reprogramming towards human pluripotency[13–15]. Moreover, variations in donor genetic background, culture conditions, reprogramming

[1]Department of Anatomy and Developmental Biology, Monash University, Clayton, Victoria, Australia. [2]Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Clayton, Victoria, Australia. [3]Australian Regenerative Medicine Institute, Monash University, Clayton, Victoria, Australia. [4]Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore Medical School, Singapore, Singapore. [5]Biomedical Manufacturing, Commonwealth Scientific and Industrial Research Organisation, Clayton, Victoria, Australia. [6]Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Perth, Western Australia, Australia. [7]The Harry Perkins Institute of Medical Research, Perth, Western Australia, Australia. [8]Department of Molecular Cell and Developmental Biology, University of California Los Angeles, Los Angeles, CA, USA. [9]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California Los Angeles, Los Angeles, CA, USA. [10]Single Cell Innovation Laboratory, University of Melbourne Centre For Cancer Research, The University of Melbourne, Melbourne, Victoria, Australia. [11]Monash Bioinformatics Platform, Monash University, Melbourne, Victoria, Australia. [12]Université de Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR1064, ITUN, F-44000, Nantes, France. [13]Université de Nantes, CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS016, CNRS UMS3556, F-44000, Nantes, France. [14]Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA, USA. [15]Jonsson Comprehensive Cancer Center, University of California Los Angeles, Los Angeles, CA, USA. [16]Present address: University of Melbourne Centre For Cancer Research, The University of Melbourne, Melbourne, Victoria, Australia. [17]Present address: Department of Biochemistry, McGill University, Montreal, Quebec, Canada. [18]Present address: Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [19]These authors contributed equally: Xiaodong Liu, John F. Ouyang, Fernando J. Rossello. ✉e-mail: owen.rackham@duke-nus.edu.sg; jose.polo@monash.edu

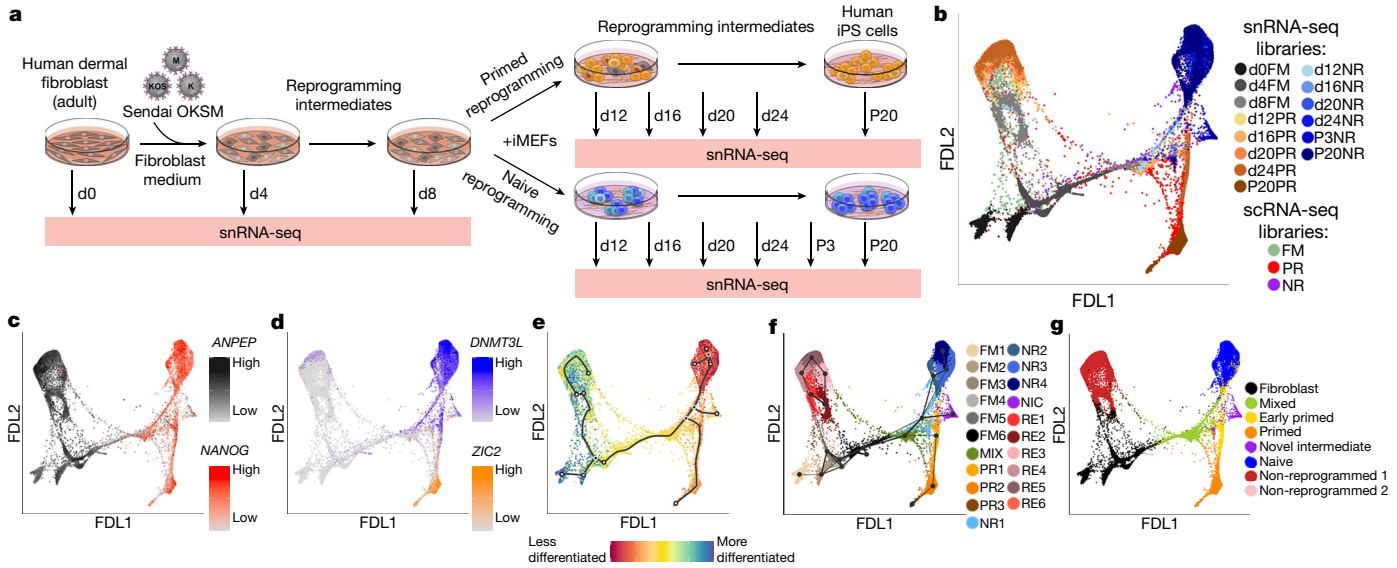


Fig. 1 | Charting a roadmap for reprogramming human cells. a, Experimental design. d, day; iMEFs, irradiated mouse embryonic fibroblasts; OKSM, OCT4, KLF4, SOX2 and MYC; P, passage. **b**, FDL of 43,791 cells, highlighting the snRNA-seq and scRNA-seq libraries. FM, fibroblast medium; NR, naive reprogramming; PR, primed reprogramming. **c**, Expression of marker genes associated with human fibroblasts (*ANPEP*) and shared pluripotency (*NANOG*). **d**, Naive pluripotency (*DNMT3L*) and primed pluripotency (*ZIC2*) on FDL. **e**, Cellular trajectory reconstruction using CytoTRACE and Monocle3. **f**, PAGA trajectory inference applied onto cell clusters in a FDL. Mix, shared clusters; NIC, novel intermediate clusters; RE, refractory cells. **g**, Predicted cell states using defined gene signatures on FDL. For more details on sample numbers and statistics, see 'Statistic and reproducibility' in Methods.

systems and isolation strategies for reprogramming intermediates can confound results[13–15].

Charting a roadmap for reprogramming human cells

To investigate the cellular transitions during the reprogramming of genetically matched adult human dermal fibroblasts into primed and naive iPS cells in a clinically relevant way, we used integration-free Sendai viruses to deliver the transcription factors *OCT4* (also known as *POU5F1*), *KLF4*, *SOX2* and *MYC*. We first cultured transduced cells in fibroblast medium, and then transitioned into medium for primed (KSR with FGF2) or naive (t2iLGoY) reprogramming (Methods). Primed and naive reprogramming intermediates and iPS cells were confirmed by morphological changes, the pluripotency marker TRA-1-60 and the naive-associated marker KLF17 (Extended Data Fig. 1a, b). To study each reprogramming pathway at single-cell resolution, we used two complementary strategies. First, we used a time-resolved strategy that collected intermediates at days 0, 4 and 8, then at days 12, 16, 20 and 24 (after the transition into primed or naive reprogramming medium), as well as from passage 3 (for naive reprogramming) and passage 20 (for primed and naive reprogramming), and subjected them to single-nucleus RNA sequencing (snRNA-seq) (Fig. 1a). Second, we used a medium-resolved strategy to assess the entire reprogramming experiment as a single process and control for any possible confounding effects, by pooling the complete trajectories into three libraries on the basis of medium composition (fibroblast medium, primed reprogramming and naive reprogramming) and subjecting them to single-cell RNA sequencing (scRNA-seq) (Extended Data Fig. 1c). We integrated the snRNA-seq and scRNA-seq datasets, which resulted in a dataset of 43,791 cells that robustly detected 11,549 genes (Extended Data Fig. 1d, Supplementary Tables 1, 2, Methods). We used force-directed layout (FDL)[16] to visualize the relationships between single cells, which has previously been used to characterize mouse reprogramming[12]. FDL shows that the cells separated into either primed or naive reprogramming trajectories (Fig. 1b, Extended Data Fig. 1e–i, Supplementary Video 1), and identified cells in different predicted stages of the cell cycle (Extended Data Fig. 1h). We further confirmed cell identity using the expression of known marker genes for fibroblasts (*ANPEP*), shared pluripotency (*NANOG*), primed pluripotency (*ZIC2*) and naive pluripotency (*DNMT3L*) (Fig. 1c, d, Extended Data Fig. 1j). We corroborated these findings by applying several complementary methods of dimensionality reduction, including principal component analysis (PCA), diffusion maps[17] and uniform manifold approximation and projection (UMAP), which produced equivalent results (Extended Data Fig. 1k–r). CytoTRACE[18]—which estimates cell potency—resolved the expected order, with naive cells appearing the least differentiated, followed by primed cells and then fibroblasts (Fig. 1e). Furthermore, a pseudotime trajectory analysis using the Monocle3[19] algorithm reinforced the observed major bifurcations that occur between naive and primed trajectories, fibroblasts and refractory cells (Fig. 1e). Altogether, these results show that the naive reprogramming trajectory is distinct from the primed trajectory rather than being an extension of it.

Alternative induced pluripotent conditions

To further characterize the cell populations that arise during reprogramming, we performed an unsupervised clustering analysis[20], which identified 21 cell clusters (Extended Data Fig. 2a). Notably, we observed a mixed cluster of naive reprogramming and primed reprogramming intermediates only near the bifurcation point of the trajectories. The clusters enabled us to apply partition-based graph abstraction (PAGA)[21] trajectory inference, which confirmed that the primed-reprogramming and naive-reprogramming trajectories bifurcate (Fig. 1f, Extended Data Fig. 2b–d). Furthermore, the mesenchymal–epithelial transition occurred early during reprogramming (Extended Data Fig. 2e). We performed a differential gene-expression analysis to identify cluster-specific marker genes, which were then combined to produce eight gene signatures (Extended Data Fig. 2f–h, Supplementary Table 3); two of these signatures robustly distinguished primed from naive iPS cells. Consistent with a previous study[14], we found that during primed reprogramming some cells activated the naive

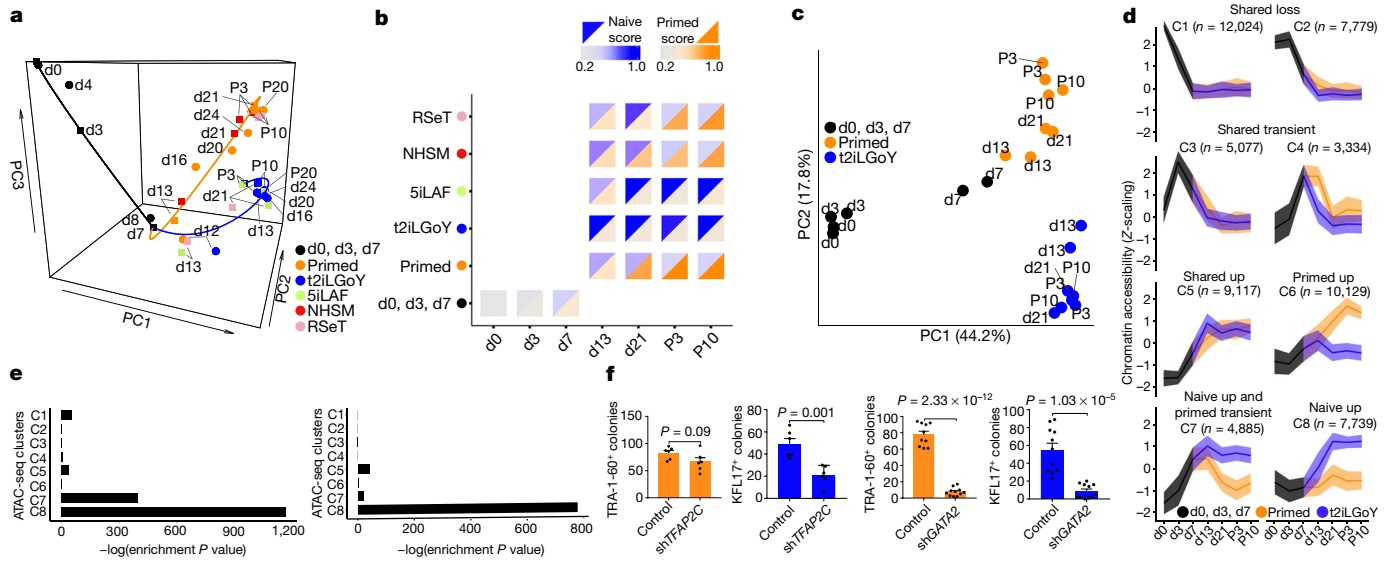


Fig. 2 | Distinct transcriptional regulatory programs drive primed and naive human reprogramming. a, PCA of the integrated bulk RNA-seq datasets (squares) from primed and several types of naive reprogramming intermediates with snRNA-seq (circles) datasets (Methods). $n \geq 2$. **b**, Naive and primed signatures scores of reprogramming intermediates under different conditions. **c**, PCA of ATAC-seq nucleosome-free signals, $n = 2$. **d**, Clustering analysis of ATAC-seq peaks during reprogramming. The number of peaks in each cluster is given. Solid lines and ribbons represent mean of standardized ATAC-seq signals across clusters ± s.d. **e**, Motif enrichment significance (−log(*P* value)) of *TFAP2C* (left) and *GATA2* (right) in ATAC-seq clusters (C1–C8). **f**, Reprogramming efficiency upon knockdown of *TFAP2C* using shRNA into primed (orange) ($n = 6$ each for control and sh*TFAP2C*) and naive (blue) ($n = 6$ each for control and sh*TFAP2C*) pluripotency, and reprogramming efficiency upon *GATA2* knockdown into primed ($n = 10$ for control, $n = 11$ for sh*GATA2*) and naive ($n = 11$ each for control and sh*GATA2*) pluripotency. For more details on sample numbers and statistics, see 'Statistic and reproducibility' in Methods.

signature but that these cells are still transcriptionally distinct from naive reprogramming intermediates (Fig. 1g, Extended Data Fig. 2g, i). Furthermore, our results demonstrated that reprogramming into naive pluripotency does not require a transition through a primed pluripotency state.

Analysis of the gene expression of pluripotency-associated cell-surface markers[22] across clusters informed a flow-cytometry isolation strategy to analyse purified populations of reprogramming intermediates using bulk-level assays (Extended Data Fig. 3a, Supplementary Fig. 1, Methods). Bulk RNA-seq data obtained from different time points during primed and naive reprogramming confirmed our isolation strategy (Extended Data Fig. 3b). The development of different culture conditions to propagate and maintain naive human ES and iPS cells has been a subject of active research[1–6], with different media producing iPS cells with a spectrum of naive characteristics[4]. To study the reprogramming pathways in different medium conditions, we therefore isolated reprogramming intermediates in additional naive media: 5iLAF[2], NHSM[1] and RSeT (Extended Data Fig. 3c–e, Methods). Harmonization of the RNA-seq data from the intermediates in different media with our snRNA-seq dataset revealed that cells cultured in NHSM follow the primed reprogramming trajectory, whereas the trajectory of cells cultured in 5iLAF overlaps with that of cells cultured in t2iLGoY. Day-13 and day-21 intermediates from RSeT medium transitioned along the naive t2iLGoY trajectory but ultimately switched branches, which established the fact that iPS cells cultured in RSeT (at passages 3 and 10) clustered near primed iPS cells (Fig. 2a, Extended Data Fig. 3f, Supplementary Table 4). These observations were confirmed by scoring these reprogramming intermediates using the primed and naive gene signatures (Fig. 2b, Extended Data Fig. 3g, Supplementary Table 5). We further examined cell heterogeneity during RSeT reprogramming by scRNA-seq, and identified both primed-like and naive-like intermediates (Supplementary Table 6). The primed-like cell population probably dominates over time, which would explain the observed switch in the reprogramming branch at bulk level (Extended Data Fig. 4a, b). Overall, these analyses revealed that reprogramming using various pluripotency conditions always follows the main naive or primed trajectories.

Chromatin dynamics during reprogramming

Cell-fate transitions during reprogramming are orchestrated by a dynamic reorganization of the epigenome[8,10,11,14]. To elucidate the chromatin accessibility landscape and the use of regulatory elements during reprogramming, we performed assay for transposase-accessible chromatin sequencing (ATAC-seq) on reprogramming intermediates isolated by flow cytometry (Supplementary Table 4). PCA of the ATAC-seq peaks (Fig. 2c, Extended Data Fig. 5a) and the integration of this analysis with the corresponding RNA-seq experiments (Extended Data Fig. 5b, c, Methods) revealed distinct changes in chromatin accessibility and a bifurcated trajectory, as observed in our transcriptional analyses. A closer inspection of population-identifying genes (*ANPEP*, *PRDM14*, *SOX11* and *DNMT3L*) revealed that the loss of accessibility of somatic regulatory elements is accompanied by a gain of open chromatin regions in regulatory elements and/or the promoters of genes associated with either primed or naive pluripotency (Extended Data Fig. 5d–f). To uncover the distinct dynamics of chromatin accessibility, we performed fuzzy clustering[23], which resulted in eight clusters (which we labelled C1–C8) (Supplementary Table 7) that we grouped by their behaviour over time (Fig. 2d). This analysis revealed: (1) a comparable distribution of peaks across genomic-region classes in all clusters (Extended Data Fig. 6a); (2) that regions of open chromatin in fibroblasts (C1 and C2) became progressively inaccessible ('shared loss') during reprogramming, concomitant with the downregulation of the associated genes (Fig. 2d, Extended Data Fig. 6b, c); (3) that transient clusters (C3 and C4) ('shared transient') exhibit an overrepresentation of genes associated with transcription, metabolism and the morphogenesis of various organs; (4) that regions with a gradual gain of accessibility for both primed and naive reprogramming (C5)

('shared up') are associated with embryonic development and stem cell maintenance; (5) that regions that specifically gained accessibility during primed reprogramming (C6) ('primed up') were associated with a range of embryonic developmental processes; (6) that two clusters (C7 and C8) ('naive up'; C7 is also 'primed transient') exhibit gain of naive-specific accessibility during reprogramming, and are associated with regulation of cell division, metabolism and cell polarity (Fig. 2d, Extended Data Fig. 6b, c, Supplementary Table 8).

Distinct programs drive reprogramming

To determine the specific transcription factors that drive these different programs, we identified the transcription factor binding-site motifs that are enriched in each cluster (Supplementary Table 9). Motif enrichment analysis of the shared loss regions uncovered transcription factors (such as *FOSL1*) that safeguard fibroblast cell identity, corroborating previous studies in mouse[10,11] (Extended Data Fig. 6d, e). C3 exhibited motifs for somatic transcription factors (for example, *FOSL1* and *JUNB*) and an enrichment for *OCT4*, *SOX2*, *NANOG* and *KLF4* binding motifs (Extended Data Fig. 6d, e). This redistribution of somatic transcription factors to transiently accessible regions that contain their binding motifs during reprogramming by *OCT4* and *SOX2* supports a similar effect that was previously described in mice[11], which potentially represents a pan-mammalian paradigm of somatic accessible-chromatin reorganization mediated by reprogramming factors. Two clusters (C7 and C8) show an unexpected significant (Supplementary Table 9) motif enrichment of TE-associated transcription factors (for example, *TFAP2C* and *GATA2*), and these transcription factors were specifically upregulated during reprogramming to the naive state or transiently upregulated in the primed state (as, for example, in C7) (Fig. 2e, Extended Data Fig. 6d–f). Furthermore, the shared-up C5 also exhibited enrichment for these same factors (Fig. 2e). To test whether these TE-associated transcription factors were passengers or drivers, we experimentally knocked them down during reprogramming using short hairpin (sh)RNA (Extended Data Fig. 6g, Supplementary Table 10). Although the absence of *TFAP2C* had a minor effect on the efficiency of primed reprogramming, naive reprogramming was greatly impaired (Fig. 2f). Knockdown of *GATA2* affected both primed and naive reprogramming, possibly as a result of *GATA2* expression being upregulated earlier in reprogramming (Extended Data Fig. 6f). Thus, these different transcriptional regulatory processes probably govern naive and primed branches of reprogramming.

Trophectoderm branch during reprogramming

We hypothesized that regulatory networks associated with the TE lineage synergistically govern the transition to naive pluripotency. Using our defined signatures, we calculated a primed and naive score of in vivo human embryo datasets from two previous studies[24,25] (Extended Data Fig. 7a, b, Methods). As expected, epiblast had the highest naive score (Supplementary Table 11), validating our approach. We next used the epiblast, primitive endoderm and TE signatures (Supplementary Table 12) from a previously published scRNA-seq human embryo dataset[25] to compute the epiblast, primitive endoderm and TE scores of our reprogramming intermediates. In addition to the expected upregulation and maintenance of the epiblast-associated transcriptional circuitry, TE-associated transcriptional programs were transiently activated during reprogramming into the naive states in t2iLGoY and 5iLAF media (Extended Data Fig. 7c–f). This was supported by a gene set enrichment analysis (Extended Data Fig. 7e). We found a subpopulation of cells to be highly enriched for the TE signatures in the single-cell trajectory of naive reprogramming (Fig. 3a, Extended Data Fig. 7g). This subpopulation forms a 'novel intermediate cluster' and its corresponding signature (the 'novel intermediate signature') shows high enrichment in the TE lineage of in vivo human blastocysts (Extended Data Fig. 7h).

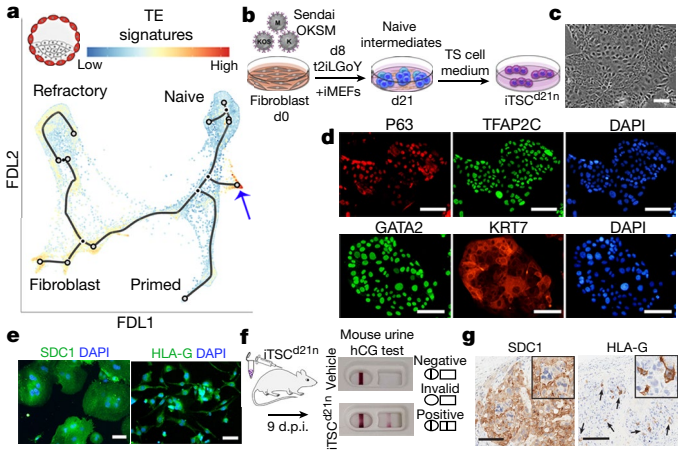


Fig. 3 | Derivation of iTS cells during reprogramming. a, In vivo TE signatures on FDL projection overlaid with single-cell trajectories constructed using Monocle3 (black lines). Blue arrow indicates the TE-enriched cell population. **b**, Experimental design for derivation of iTSC^{d21n}. **c**, Phase-contrast image of iTSC^{d21n}. Scale bar, 100 µm. **d**, Immunostaining of iTSC^{d21n} with P63, TFAP2C, GATA2 and KRT7. Scale bar, 100 µm. Representative images from $n = 4$. **e**, SDC1 and HLA-G immunostaining of syncytiotrophoblast (ST) (left) and extravillous trophoblast (EVT) (right) cells, respectively, differentiated from iTSC^{d21n}. Scale bar, 100 µm. Representative images from $n = 4$. **f**, Representation of iTSC^{d21n} engraftment assay by injection into NOD-SCID mice. The urine, blood serum and lesions were examined 9 days post-injection (d.p.i). Representative positive results for hCG pregnancy test from urine samples collected from iTSC^{d21n}-injected mice compared to the vehicle controls. $n = 3$. **g**, Immunohistochemical staining of SDC1 (left) and HLA-G (right) in the lesions collected from iTSC^{d21n} engrafts in NOD-SCID mice. No lesions were evident in vehicle controls. Arrows indicate HLA-G-positive trophoblast cells. Scale bar, 200 µm. Representative images from $n = 4$. For more details on sample numbers and statistics, see 'Statistic and reproducibility' in Methods.

Deriving induced trophoblast stem cells

We hypothesized that this TE-associated cell cluster could be stabilized to give rise to trophoblast stem (TS) cells. Thus, we transitioned naive reprogramming intermediates at day 21 into the recently reported human TS cell medium[7] (Fig. 3b); we observed the appearance of cells that morphologically resemble TS cells that we named 'induced TS cells' (hereafter, iTSC^{d21n}) (Fig. 3c). Further characterization showed that iTSC^{d21n} express key markers that define human TE and TS cells[7,26], such as P63, TFAP2C, GATA2 and KRT7 (Fig. 3d, Extended Data Fig. 8a). Moreover, these iTS cells express comparable levels of TS-cell marker genes, and are distinct from human fibroblasts and primed and naive iPS cells (Extended Data Fig. 8b). To functionally characterize iTSC^{d21n}, we examined the in vitro differentiation capacity of these cells to give rise to syncytiotrophoblast and extravillous trophoblast cells, the major trophoblast subtypes of the placenta[26]. This demonstrated that iTSC^{d21n} can be differentiated into syncytiotrophoblast cells characterized by SDC1-positive multinucleated cells, and extravillous trophoblast cells defined by upregulation of HLA-G, a key histocompatibility molecule that is expressed in placenta[7,26] (Fig. 3e, Extended Data Fig. 8c). The iTSC^{d21n} syncytiotrophoblast cells showed a significantly higher fusion index than that of iTSC^{d21n}, and secreted human chorionic gonadotropin (hCG) that could be detected using an over-the-counter hCG pregnancy test stick, and which we quantified by hCG enzyme-linked immunosorbent assay (ELISA) (Extended Data Fig. 8d–f). Next, we evaluated the in vivo differentiation potential of iTSC^{d21n} by

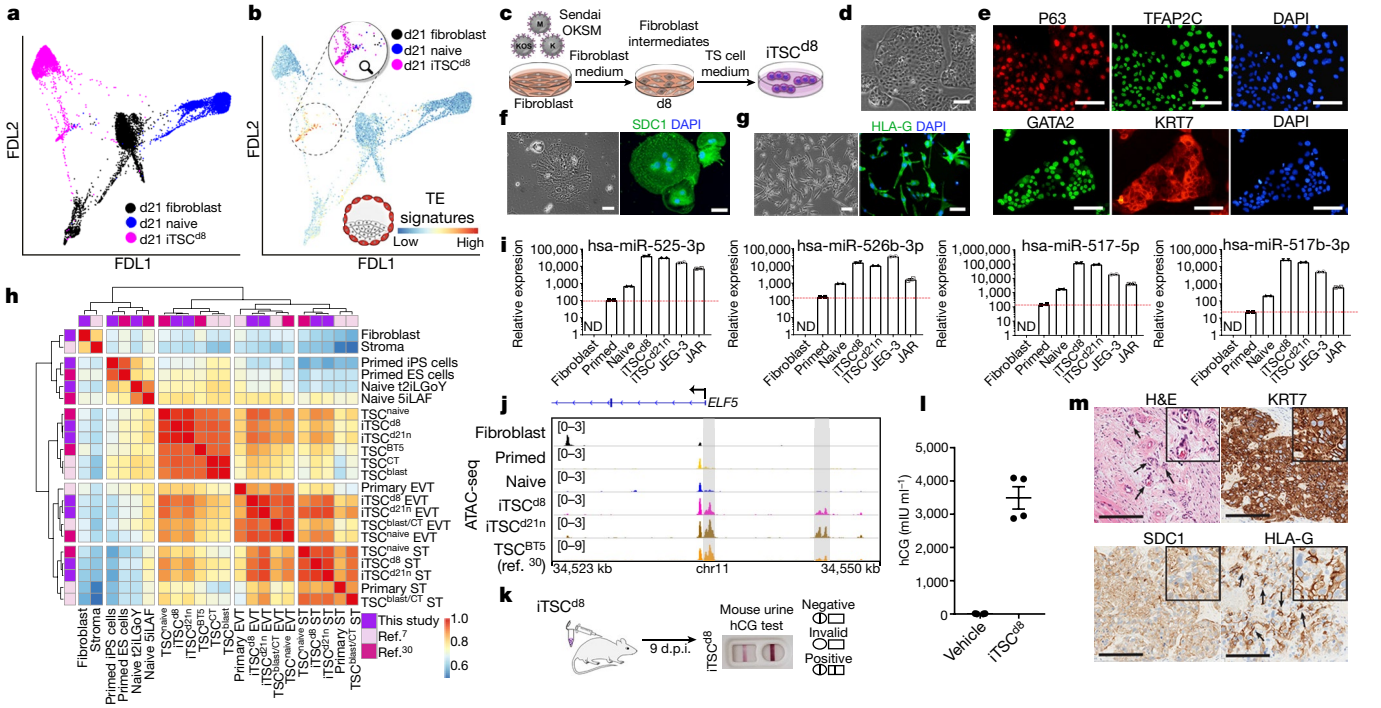


Fig. 4 | Direct derivation of iTS cells from human fibroblasts. a, FDL representation of scRNA-seq libraries of day-21 reprogramming intermediates (10,518 cells). **b**, TE signatures on FDL projections. The TE-like population is highlighted and coloured by library. **c**, Experimental design of direct derivation of iTSC^{d8} from fibroblasts. **d**, Phase-contrast image of iTSC^{d8}. Scale bar, 100 µm. **e**, Immunostaining of iTSC^{d8} for several TS cell makers. Scale bar, 100 µm. **f**, **g**, Phase-contrast and immunostaining of ST (**f**) and EVT (**g**) cells differentiated from iTSC^{d8}. Scale bar, 100 µm. $n = 4$ for **d**–**g**. **h**, Spearman correlation of transcriptomes from this study with published datasets[7,30]. Biological replicates ($n \geq 2$) are averaged before performing correlation. blast, blastocyst; blast/CT, blastocyst/cytotrophoblast; CT, cytotrophoblast. **i**, Expression of miRNAs of the chromosome 19 miRNA cluster, normalized to 1,000× expression of miR-103a, in the indicated cell lines. Mean ± s.e.m. ND, not detected. Red dotted line indicates the level in primed iPS cells. $n = 2$. **j**, ATAC-seq signal at *ELF5* region in the indicated cell types. Mean value of replicates ($n = 2$). TS cell peaks are marked in grey. **k**, Representative hCG test from urine samples collected from iTSC^{d8}-injected mice, $n = 3$. **l**, hCG protein level detected by hCG ELISA using mouse blood serum samples, $n = 4$. **m**, Haematoxylin and eosin (H&E) and immunohistochemical staining of KRT7, SDC1 and HLA-G in the lesions collected from iTSC^{d8} engrafts in NOD-SCID mice, $n = 4$. No lesions were evident in vehicle controls. Scale bar, 200 µm. For more details on sample numbers and statistics, see 'Statistic and reproducibility' in Methods.

subcutaneous injection into mice. Nine days post-injection, mouse urine was positive for hCG using the over-the-counter human pregnancy tests (Fig. 3f, Methods) and hCG was also detected in the blood serum (Extended Data Fig. 8g). We further confirmed engraftment and differentiation by histology analyses of the lesions formed, showing SDC1-positive syncytiotrophoblast-like cells and HLA-G-positive extravillous-trophoblast-like cells, comparable to previously reported primary tissue-derived TS cells[7] (Fig. 3g, Extended Data Fig. 8h,i). Importantly, these results demonstrate that iTSC^{d21n} are bipotent in vitro and in vivo. Finally, we used CD70^{low} to enrich TE-like cells from the novel intermediate cluster, and demonstrated that the identified TE-like cluster carries the greatest potential for the generation of iTSC^{d21n} (Extended Data Fig. 8j, k). Altogether, these results suggest that cell-fate specification is highly dynamic and plastic during the reprogramming of somatic cells in humans.

Reprogramming fibroblasts directly into iTS cells

To test whether iTS cells could be derived directly from human fibroblasts, we started reprogramming experiments and transitioned the day-8 intermediates into TS cell medium or naive medium, or kept them in fibroblast medium. We then performed scRNA-seq on these conditions at day 21 to assess the cellular heterogeneity (Extended Data Fig. 9a). A population of TE-like cells was observed, and closer examination revealed that this TE-like population contained cells from all three reprogramming conditions (Fig. 4a, b, Extended Data Fig. 9b–d, Supplementary Table 13). Furthermore, the day-21 fibroblast-medium intermediates also consist of cells with strong epiblast, primed and naive signatures (Extended Data Fig. 9e), and accordingly these intermediates were able to give rise to pluripotent and TS cell lines (Extended Data Fig. 9f–h). We noticed that the proportion of TE-like population was the highest in TS cell medium, as compared to fibroblast and naive media (Fig. 4b, Extended Data Fig. 9d). Therefore, we hypothesized that we could derive iTS cell lines more efficiently by directly transitioning day-8 intermediates into TS cell medium (hereafter, iTSC^{d8}), without the need to expose the cells to naive medium or prolonged culturing in fibroblast medium (Fig. 4c). As seen in Fig. 4d, iTSC^{d8} were successfully derived directly, and our transgene-free iTSC^{d8} (Extended Data Fig. 10a) have demonstrated the capacity to undergo >50 passages thus far without a reduction in growth rate. We then performed a comprehensive molecular and functional characterization of iTSC^{d8} on the basis of features defined for TS cells generated from primary sources[7,26–29]. This demonstrated that these iTSC^{d8} expressed key marker genes indicative of mononuclear trophoblasts[26] (Fig. 4e), and that they could differentiate into syncytiotrophoblasts and extravillous trophoblasts. The syncytiotrophoblasts expressed SDC1, displayed cell fusion and hCG secretion (Fig. 4f, Extended Data Fig. 10b–e). Extravillous trophoblasts were positive for the HLA-A, -B, -C pan marker (W632), but not the HLA-B marker, and—importantly—they expressed HLA-G (Fig. 4g, Extended Data Fig. 10f–h). We found that the expression of HLA-A, -B, -C pan marker (W632) was detected in iTS cells, similar to what was previously

reported in TS cells derived from blastocysts[7]. By contrast, trophoblast organoids are HLA-negative[28] which suggests that the culture conditions might support TS cells at different stages of gestation. Furthermore, our iTS cells and iTS-cell-derived syncytiotrophoblasts and extravillous trophoblasts share a common transcriptomic profile with the corresponding primary cell types of other published datasets[7,27–30] (Fig. 4h, Extended Data Fig. 10i–l, Supplementary Table 14). iTS cells also show higher levels of expression of microRNAs (miRNAs) from the chromosome 19 miRNA cluster, compared to fibroblast and iPS cells—a unique feature of primary trophoblast[26] (Fig. 4i). We observed specific open chromatin accessibility at the promoter and putative enhancer regions of the *ELF5* locus in iTS cells and in TSC^{BT5} (which are derived from human blastocysts[30]) (Fig. 4j) that has previously been found to be hypomethylated[7,26]. Finally, we showed that $iTSC^{d8}$ could engraft into mouse tissues, differentiate into the major trophoblast-lineage cell types of the placenta in vivo, and secrete hCG in urine and serum (Fig. 4k–m, Extended Data Fig. 10m). These results confirmed that $iTSC^{d8}$ derived directly from human fibroblasts are similar to primary TS cells.

Discussion

Here we present a detailed molecular roadmap of reprogramming into primed and naive human pluripotency at the single-cell level, for which we developed an interactive online tool (http://hrpi.ddnetbio.com/) to facilitate exploration of the dataset. This roadmap reveals that the two reprogramming trajectories diverge, and that a cell does not need to first acquire a primed pluripotent state to reprogram into a naive pluripotent state; this indicates that reprogramming to the naive state is not a reversion of the developmental pathway. On closer inspection, both the main naive and primed branches also exhibit alternative sub-branches. We hypothesize that these sub-branches could be true alternative pathways or metastable fates. For example, in the naive branch at least two sub-branches are apparent—one in which a TE-associated network is upregulated and one in which it is not. The fact that the knockdown of the transcription factors that are predicted to drive these networks impaired naive reprogramming (Fig. 2) suggests that both sub-branches are active, and that the reprogramming trajectories remain similar for different naive conditions (5iLAF and t2iLGoY). This indicates that each medium not only promotes a similar final pluripotency state (as has previously been shown[4]) but also drives the intermediate cells along similar trajectories. Together, these results present a 'push-or-pull' question of whether similar reprogramming trajectories are determined by being pulled towards a common final pluripotency state, or whether the specific culture medium pushes the cells along similar trajectories and—as a consequence—results in a similar final state.

The change in chromatin accessibility during primed and naive reprogramming also indicates a bifurcated trajectory. Early and transient chromatin accessibility clusters are enriched in OCT4, KLF4 and SOX2 motifs, which suggests binding of these transcription factors at initially closed regions and supports a pioneering effect of these factors, as previously reported[11,31]. Furthermore, the upregulation of TE-associated transcriptional networks during reprogramming into the epiblast-like state (naive) is unexpected (Figs. 2e, 3a), as one of the first cell-fate decisions that cells make during development is whether they will become trophoblast or epiblast. Our results reveal the coexistence of primed-like, naive-like and TE-like cells during reprogramming in the fibroblast medium without exposing them to any pluripotent or trophoblast media, suggesting that the combination of OCT4, KLF4, SOX2 and MYC can induce human fibroblasts to acquire pluripotent and trophoblast states. The direct reprogramming of fibroblasts into iTS cells is in contrast to the recently reported three-step approach, in which somatic cells must first be reprogrammed into iPS cells, then converted into the expanded-potential or naive stem cells before being differentiated into TS cells[30,32]. We envision that this direct approach will facilitate the generation of patient-specific iTS cells to study trophoblast dysfunction. Such studies are critically needed as this dysfunction leads to various complications during pregnancy, such as pre-eclampsia and intra-uterine growth restriction[7,26,28]. Furthermore, having stable, self-renewing bona fide isogenic human iPS cell and iTS cell lines will provide the opportunity to study human trophectoderm and trophoblast development, and to better understand their roles in coordinating events associated with cell-fate decisions during early human embryogenesis. With these cell lines, it would be possible to investigate the interaction between pluripotent and trophoblast stem cells in vitro and apply modern biochemical and molecular techniques at scale, rapidly increasing our ability to understand and intervene in developmental diseases. Finally, because both embryonic and extra-embryonic lineages can be derived, these results also hint at the possibility that there may be a totipotent state during reprogramming. Thus, if the conditions to stabilize these cells and stringently defined totipotency criteria are met[33], a totipotent cell type could eventually be derived by reprogramming.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2734-6.

1. Gafni, O. et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282–286 (2013).
2. Theunissen, T. W. et al. Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 524–526 (2014).
3. Takashima, Y. et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **162**, 452–453 (2015).
4. Liu, X. et al. Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming. *Nat. Methods* **14**, 1055–1062 (2017).
5. Kilens, S. et al. Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. *Nat. Commun.* **9**, 360 (2018).
6. Giulitti, S. et al. Direct generation of human naive induced pluripotent stem cells from somatic cells in microfluidics. *Nat. Cell Biol.* **21**, 275–286 (2019).
7. Okae, H. et al. Derivation of human trophoblast stem cells. *Cell Stem Cell* **22**, 50–63.e6 (2018).
8. Polo, J. M. et al. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
9. O'Malley, J. et al. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature* **499**, 88–91 (2013).
10. Chronis, C. et al. Cooperative binding of transcription factors orchestrates reprogramming. *Cell* **168**, 442–459.e20 (2017).
11. Knaupp, A. S. et al. Transient and permanent reconfiguration of chromatin and transcription factor occupancy drive reprogramming. *Cell Stem Cell* **21**, 834–845.e6 (2017).
12. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 1517 (2019).
13. Takahashi, K. et al. Induction of pluripotency in human somatic cells via a transient state resembling primitive streak-like mesendoderm. *Nat. Commun.* **5**, 3678 (2014).
14. Cacchiarelli, D. et al. Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. *Cell* **162**, 412–424 (2015).
15. Wang, Y. et al. Unique molecular events during reprogramming of human somatic cells to induced pluripotent stem cells (iPSCs) at naïve state. *eLife* **7**, e29518 (2018).
16. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9**, e98679 (2014).
17. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
18. Gulati, G. S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
19. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
20. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
21. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
22. O'Brien, C. M. et al. New monoclonal antibodies to defined cell surface proteins on human pluripotent stem cells. *Stem Cells* **35**, 626–640 (2017).

23. Kumar, L. & E Futschik, M. Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).
24. Yan, L. et al. Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
25. Petropoulos, S. et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
26. Lee, C. Q. E. et al. What is trophoblast? A combination of criteria define human first-trimester trophoblast. *Stem Cell Reports* **6**, 257–272 (2016).
27. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).
28. Turco, M. Y. et al. Trophoblast organoids as a model for maternal–fetal interactions during human placentation. *Nature* **564**, 263–267 (2018).
29. Haider, S. et al. Self-renewing trophoblast organoids recapitulate the developmental program of the early human placenta. *Stem Cell Reports* **11**, 537–551 (2018).
30. Dong, C. et al. Derivation of trophoblast stem cells from naïve human pluripotent stem cells. *eLife* **9**, e52504 (2020).
31. Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
32. Gao, X. et al. Establishment of porcine and human expanded potential stem cells. *Nat. Cell Biol.* **21**, 687–699 (2019).
33. Posfai, E., Schell, J. P., Janiszewski, A., Rovic, I. & Murray, A. Defining totipotency using criteria of increasing stringency. Preprint at https://www.biorxiv.org/content/10.1101/2020.03.02.972893v1 (2020).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture conditions

The experimental design, materials and reagents are described in the Life Sciences Reporting Summary. All cell lines used in this study were authenticated, and mycoplasma-tested as described in the Reporting Summary. Primary human adult dermal fibroblasts from three female donors were obtained from ThermoFisher (catalogue number C-013-5C and lot no. 1029000 for 38F, lot no. 1528526 for 55F and lot no. 1569390 for 32F); cells were recovered and plated in medium 106 (ThermoFisher) supplemented with low serum growth supplement (LSGS) (ThermoFisher) for expansion. The use of human embryonic stem cells (H9) was carried out in accordance with approvals from Monash University and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Human Research Ethics Offices. Conventional primed human iPS cells (established lines) and H9 ES cells (WiCell Research Institute, http://www.wicell.org) were maintained in a feeder-free system on vitronectin (VTN-N, Gibco)-coated tissue culture plastics in essential 8 medium (Gibco). Media were changed daily, and cells were passaged every 5 d using 0.5 mM EDTA (Invitrogen). Culture conditions used for human somatic cell reprogramming were prepared as previously described[4,34]. Fibroblast medium: DMEM (ThermoFisher), 10% fetal bovine serum (FBS) (Hyclone), 1% nonessential amino acids (ThermoFisher), 1 mM GlutaMAX (ThermoFisher), 1% penicillin–streptomycin (ThermoFisher), 55 μM 2-mercaptoethanol (ThermoFisher) and 1 mM sodium pyruvate (ThermoFisher). Primed medium: DMEM/F12 (ThermoFisher), 20% knockout serum replacement (KSR) (ThermoFisher), 1 mM GlutaMAX (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 1% non-essential amino acids (ThermoFisher), 50 ng/ml recombinant human FGF2 (Miltenyi Biotec), 1% penicillin–streptomycin (ThermoFisher). Naive medium (t2iLGoY)[35]: 50:50 mixture of DMEM/F-12 (ThermoFisher) and neurobasal medium (ThermoFisher), supplemented with 2 mM L-glutamine (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 0.5% N2 supplement (ThermoFisher), 1% B27 supplement (ThermoFisher), 1% penicillin–streptomycin (ThermoFisher), 10 ng/ml human leukaemia inhibitory factor (LIF) (made in-house), 250 μM L-ascorbic acid (Sigma), 10μg/ml recombinant human insulin (Sigma), 1 μM PD0325901 (Miltenyi Biotec), 1 μM CHIR99021 (Miltenyi Biotec), 2.5 μM Gö6983 (Tocris), 10 μM Y-27632 (Abcam). Naive human stem cell medium (NHSM): culture condition adapted from ref. [1] with suggested modifications from the web page of the J. Hanna laboratory in 2014 was used. DMEM/F12 (ThermoFisher) supplemented with 10 mg/ml AlbuMAX I (ThermoFisher), 1% penicillin–streptomycin (ThermoFisher), 1 mM GlutaMAX (ThermoFisher), 1% nonessential amino acids (ThermoFisher), 10% KSR (ThermoFisher),1% N2 supplement (ThermoFisher), 12.5 μg/ml recombinant human insulin (Sigma), 50 μg/ml L-ascorbic acid (Sigma), 20 ng/ml of recombinant human LIF (made in-house), 8 ng/ml FGF2 (Peprotech), 2 ng/ml recombinant TGFβ1 (Peprotech), 20 ng/ml human LR3-IGF1 (Prospec) and small molecule inhibitors: 1 μM PD0325901 (Miltenyi Biotec), 3 μM CHIR99021 (Miltenyi Biotec), 5 μM SP600125 (Tocris) 2 μM BIRB796 (Axon), 0.4 μM LDN193189 (Axon), 10 μM Y-27632 (supplemented daily to media from freshly thawed stock aliquot) (Abcam) and 1 μM Gö6983 (supplemented daily to media from freshly thawed stock aliquot) (Tocris). Naive 5iLAF medium[2,36]: 50:50 mixture of DMEM/F-12 (ThermoFisher) and neurobasal medium (ThermoFisher) supplemented with 1% N2 supplement (ThermoFisher), 2% B27 supplement (ThermoFisher), 1% nonessential amino acids (ThermoFisher), 1 mM GlutaMAX (ThermoFisher), 1% penicillin–streptomycin (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 50μg/ml bovine serum albumin (ThermoFisher), 1 μM PD0325901 (Miltenyi Biotec), 1 μM IM-12 (Millipore), 0.5 μM SB590885 (Tocris), 1 μM WH-4-023 (A Chemtek), 10 μM Y-27632 (Abcam), 20 ng/ml activin A (Peprotech), 8 ng/ml FGF2 (Miltenyi Biotec), 20 ng/ml human LIF (made in-house) and 0.5% KSR (ThermoFisher). Naive RSeT medium: 100 ml of RSeT 5× supplement, 1 ml of RSeT 500× supplement and 0.5 ml of RSeT 1,000× supplement into 398.5 ml of RSeT basal medium (Stem Cell Technologies) supplemented with 1% penicillin–streptomycin (ThermoFisher). Human TS cell medium[7]: DMEM/F-12, GlutaMAX (ThermoFisher) supplemented with 0.3% BSA (Sigma), 0.2% FBS (ThermoFisher), 1% ITS-X supplement (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 0.5% penicillin–streptomycin (ThermoFisher), 1.5 μg/ml L-ascorbic acid (Sigma), 5 μM Y27632 (Abcam), 2 μM CHIR99021 (Miltenyi Biotec), 0.5 μM A83-01 (Sigma), 1 μM SB431542, 50 ng/ml EGF (Peprotech) and 0.8 mM valproic acid (VPA) (Sigma).

Reprogramming experiments

The t2iLGoY medium was used for naive reprogramming, as it has previously been shown[4] that this medium can be used to reprogram fibroblasts into naive iPS cells that possess all the hallmarks of naive pluripotency and maintain a more stable karyotype when compared to other conditions. Human somatic cell reprogramming into primed and naive pluripotent states experiments and the subsequent culture of primed and naive iPS cells were performed as previously described[4,34]. In brief, the reprogramming of human fibroblasts was conducted using CytoTune-iPS 2.0 Sendai reprogramming kit, according to the manufacturer's instructions (ThermoFisher). Primary human adult dermal fibroblasts were seeded at a density of about $5–10 \times 10^4$ cells in fibroblast medium. As shown in Fig. 1a, cells were transduced with Sendai viruses in fibroblast medium at a multiplicity of infection (MOI) as follows: KLF4, OCT4 and SOX2, MOI = 5 or 10; MYC, MOI = 5 or 10; and KLF4 MOI = 6 or 12. Cells were reseeded onto a layer of iMEF feeders on day 7 and transitioned into different culture media (primed, t2iLGoY, NHMS, RSeT or 5iLAF) the next day. After 18–21 days, iPS cells could be passaged and expanded as previously described[34]. For shRNA knockdown experiments, a pair of U6 shRNA lentiviral vectors (VectorBuilder) for each gene was used. The shRNA sequences are provided in Supplementary Table 10. Lentiviral particles were generated using human embryonic kidney cells (293T) as previously described[11,37]. Human adult dermal fibroblasts were transduced with lentiviral vectors for one week and re-plated two days before Sendai transduction. Colony counts of *TFAP2C* and *GATA2* knockdown experiments are provided in Source Data for Fig. 2f. Knockdown experiments were validated by reverse-transcription qPCR (RT–qPCR), and primers used are listed in Supplementary Table 15. All cells were cultured in incubators at 37 °C, 5% O_2 and 5% CO_2. For the derivation of $iTSC^{d21n}$ during naive reprogramming, day-21 naive t2iLGoY reprogramming intermediates were transitioned into TS cell medium[7]. After 4–5 days, cells were passaged using TrypLE express (ThermoFisher) every 3–4 days at a 1:2–1:4 ratio. For the initial 4 passages, iTS cells were passaged onto iMEF feeders and cultured in a 37-°C, 5% O_2 and 5% CO_2 incubator. Starting from passage 5, $iTSC^{d21n}$ were passaged onto a tissue culture flask that was pre-coated with 5μg/ml collagen IV (Sigma) (for at least 1 h at 37 °C) and cultured in a 37-°C, 20% O_2 and 5% CO_2 incubator. For the direct derivation of $iTSC^{d8}$ from human fibroblasts, day-8 fibroblast reprogramming intermediates were transitioned into TS cell medium. After 10–13 days, $iTSC^{d8}$ were passaged onto iMEF feeders and cultured in a 37-°C, 5% O_2 and 5% CO_2 incubator as described for $iTSC^{d21n}$. Sendai detection in established iTS cell lines was performed as described in the Sendai reprogramming protocol (ThermoFisher). For the derivation of primed, naive iPS cells and iTS cells from day-21 fibroblast reprogramming intermediates, day-21 fibroblast reprogramming intermediates were transitioned into primed, naive or TS cell medium, and then cultured and expanded as described.

Differentiation of iTSC^{d21n} and iTSC^{d8} into ST and EVT in vitro

Differentiation of iTS cells into ST and EVT was performed as previously described[7]. For the differentiation of iTS cells into ST, iTS cells were seeded at a density of 1×10^5 cells per well onto a 6-well plate pre-coated with 2.5 µg/ml collagen IV (Sigma) and cultured in 2 ml of ST differentiation medium (DMEM/F-12, GlutaMAX (ThermoFisher) supplemented with 0.3% BSA (Sigma), 4% KSR (ThermoFisher), 1% ITS-X supplement (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 0.5% penicillin–streptomycin (ThermoFisher), 2.5 µM Y27632 (Abcam) and 2 µM forskolin (Selleckchem)). Media were replaced daily for the initial 4 days, and cells were analysed on day 6. The fusion index was used to quantify the efficiency of cell fusion, which is calculated by using the number of nuclei counted in the syncytia minus the number of syncytia, then divided by the total number of nuclei counted. For the differentiation of iTS cells into EVT, iTS cells were seeded at a density of 0.75×10^5 cells per well onto a 6-well plate pre-coated with 1 µg/ml collagen IV (Sigma) and cultured in 2 ml of EVT differentiation medium (DMEM/F-12, GlutaMAX (ThermoFisher) supplemented with 0.3% BSA (Sigma), 4% KSR (ThermoFisher), 1% ITS-X supplement (ThermoFisher), 0.1 mM 2-mercaptoethanol (ThermoFisher), 0.5% penicillin–streptomycin (ThermoFisher), 2.5 µM Y27632 (Abcam), 100 ng/ml NRG1 (Cell Signaling) and 7.5 µM A83-01 (Sigma)). Shortly after suspending the cells in the EVT differentiation medium, Matrigel (Corning) was overlaid to a 2% final concentration. On day 3 of differentiation, EVT differentiation medium without human NRG1 (Cell Signaling) and Matrigel (Corning) was added to a final concentration of 0.5%. On day 6 of differentiation, EVT differentiation media were replaced without NRG1 (Cell Signaling) or KSR (ThermoFisher), and Matrigel (Corning) was added to 0.5% final concentration. The cells were cultured for an additional 2 d before analyses were performed.

iTSC^{d21n} and iTSC^{d8} in vivo engraftment assay

Protocols and use of mice were undertaken with the approval of the Monash University Animal Welfare Committee following the 2004 Australian Code of Practice for the Care and Use of Animals for Scientific Purposes and the Victorian Prevention of Cruelty to Animals Act and Regulations legislation. iTS cells with 80% confluency were dissociated with TrypLE express (ThermoFisher) and counted. Ten million iTS cells were resuspended in 200 µl of a 1:2 mixture of Matrigel (Corning) and DMEM/F-12, GlutaMAX (ThermoFisher) supplemented with 0.3% BSA (Sigma) and 1% ITS-X supplement (ThermoFisher). The cellular mixture was then injected subcutaneously into dorsal flanks of male and female, 5–20-week-old NOD/SCID IL-2R-gamma knockout mice (100 µl into each flank). Mice were randomized between controls and iTS cells, but investigators were not blinded. Nine days after injection, urine, blood serum and lesions were collected from the mice for analysis. Mice urine and serum were used for the detection and measurement of hCG secretion as detailed in 'Pregnancy tests and hCG ELISA'. Collected lesions were fixed with 4% paraformaldehyde (PFA) (Sigma) overnight and subsequently embedded in paraffin. Lesions collected were less than 1 cm^3 in volume. Paraffin-embedded tissues were sectioned and stained with haematoxylin and eosin (H&E) or proceeded with immunohistochemistry staining of KRT7, HLA-G and SDC1 (Supplementary Table 16) at the Histology Platform at Monash University.

Pregnancy tests and hCG ELISA

iTS cells were seeded at a density of 0.5×10^5 cells per ml on a 12-well plate for ST differentiation as detailed in 'Differentiation of iTSC^{d21n} and iTSC^{d8} into ST and EVT in vitro'. The medium of the ST cells was replaced on day 4 and the conditioned medium was collected at day 6 and stored at −80 °C. As a control, iTS cells were also seeded at a density of 0.5×10^5 cells per ml on a 12-well plate and cultured in TSC medium. Two days later, the conditioned medium was collected and stored at −80 °C. The conditioned media were then tested using over-the-counter hCG pregnancy test sticks (Freedom) according to the manufacturer's recommendations. In addition, the hCG level within the medium was also measured using hCG ELISA kit (Abnova, ABNOKA4005) according to the manufacturer's instructions. Following the iTS cell engraftment assay, the collected mouse urine was tested using the over-the-counter hCG pregnancy test sticks and hCG level in blood serum was measured using hCG ELISA kit.

Flow cytometry analysis and fluorescent-activated cell sorting

All antibodies used in flow cytometry analysis and fluorescent-activated cell sorting (FACS) experiments are summarized in Supplementary Table 16. Cells were dissociated with TrypLE express (ThermoFisher), and DPBS (ThermoFisher) supplemented with 2% FBS (Hyclone) and 10 µM Y-27632 (Abcam) was used for antibody labelling steps and final resuspension of the samples. For spanning-tree progression analysis of density-normalized events (SPADE) analysis (Extended Data Fig. 3e), a three-step antibody labelling procedure was used: (1) rat anti-human IgM SSEA-3 (1:10, BD); mouse anti-human NLGN4X IgG2a (1:128, CSIRO CSTEM30[22]); (2) mouse anti-rat IgM PE (1:200, eBiosciences); BV605 goat anti-mouse IgG (1:100, BioLegend); and (3) BV421 mouse anti-human CD326 (EpCAM) (1:100, BioLegend); BUV395 mouse anti-human TRA-1-60 (1:100, BD); BV711 mouse anti-human CD24 (1:50, BD); mouse anti-human SSEA-4-PE-Vio770 (1:20, Miltenyi Biotec); mouse anti-human F11R IgG was conjugated to APC by the Walter and Eliza Hall Institute of Medical Research (WEHI) antibody facilities (1:200, CSIRO CSTEM27[22]); APC-Cy7 CD13 (1:500, BioLegend); Anti-TRA-1-85 (CD147)-VioBright FITC (1:20, Miltenyi Biotec). For FACS, antibody labelling was performed as: (1) mouse anti-human F11R IgG antibody (1:200, CSIRO CSTEM27); PE rat anti-human SSEA-3 IgM antibody (1:10, BD); (2) AF647 goat anti-mouse IgG antibody (1:2,000, ThermoFisher); mouse anti-rat IgM PE (1:200, eBiosciences); and (3) PE-Cy7 mouse anti-human CD13 (1:400, BD); BV421 mouse anti-human CD326 (EpCAM) (1:100, BioLegend); BUV395 mouse anti-human TRA-1-60 (1:100, BD). The antibody labelling steps were carried out in a volume of 500 µl per 1 million cells, and incubation time was 10 min on ice per step; after each antibody labelling step, cells were washed with 10 ml cold PBS and pelleted at 400*g* for 5 min. The cells were then resuspended in a final volume of 500 µl, and propidium iodide (PI) (Sigma) was added to a concentration of 2 µg/ml. Cell sorting was carried out with a 100-µm nozzle on an Influx instrument (BD Biosciences), and flow cytometry analysis was carried out using an LSRIIb or LSRIIa analyser (BD Biosciences). For Supplementary Fig. 1, reprogramming intermediates were isolated on day 3 into CD13$^+$F11R$^+$ and CD13$^+$F11R$^-$ subpopulations, and then reseeded into fibroblast medium (FM) condition for five days for flow cytometry reanalysis and for iPS cell formation confirmed by alkaline phosphatase (AP) staining according to the manufacturer's instructions (Vector laboratories). On day 7, CD13$^+$, CD13$^-$F11R$^+$TRA-1-60$^-$ and CD13$^-$F11R$^+$TRA-1-60$^+$ subpopulations were used for such analysis (reseeded in FM condition for one day and then transitioned into either primed or naive t2iL-GoY conditions). On day 13, CD13$^-$F11R$^+$TRA-1-60$^+$SSEA3$^+$EPCAM$^-$ and CD13$^-$F11R$^+$TRA-1-60$^+$SSEA3$^+$EPCAM$^+$ subpopulations were isolated for primed reprogramming, CD13$^-$F11R$^+$TRA-1-60$^+$SSEA3$^+$EPCAM$^+$ and CD13$^-$F11R$^+$TRA-1-60$^+$SSEA3$^-$EPCAM$^+$ subpopulations were isolated for naive reprogramming. For iTS cell purification, a two-step antibody labelling procedure was used: (1) mouse anti-human APA (1:100) and (2) BUV395 mouse anti-human TRA-1-60 (1:100, BD); APC rat anti-human and mouse CD49F (ITGA6) (1:20, Miltenyi Biotec); AF488 goat anti-mouse IgG1 antibody (1:2,000, ThermoFisher). iTS cell purification was performed on the reprogrammed cells at passage 9–10 by isolating TRA-1-60$^-$APA$^+$ITGA6$^+$ subpopulations and reseeding onto collagen-IV-coated 6-well plate for long-term passaging. For Extended Data Fig. 8k, enrichment of CD70^{high} and CD70^{low} populations was performed using a one-step antibody labelling procedure: anti-TRA-1-85 (CD147)-VioBright FITC (1:20, Miltenyi Biotec); PE-Cy7

mouse anti-human CD13 (1:400, BD); BV421 mouse anti-human CD326 (EpCAM) (1:100, BioLegend); BUV395 mouse anti-human TRA-1-60 (1:100, BD); APC mouse anti-human F11R (1:250, CSIRO CSTEM27); BUV737 mouse anti-human CD70 (1:100, BD). Details of these antibodies are provided in Supplementary Table 16. Labelled cells were resuspended in a final volume of 500 µl containing 2 µg/ml of PI (Sigma) for cell sorting. TRA185$^+$CD13$^-$F11R$^+$TRA-1-60$^+$EPCAM$^+$CD70^{high} and TRA185$^+$CD13$^-$F11R$^+$TRA-1-60$^+$EPCAM$^+$CD70^{low} subpopulations (denoted as CD70^{high} and CD70^{low} subpopulations, respectively) were isolated and reseeded onto a layer of iMEF feeders (24-well plate) at a density of 5×10^3 cells per well. On the next day after reseeding, the spent culture medium was replaced with the TS cell medium. Immunostaining for KRT7$^+$ colonies was then performed on day 9 after reseeding as described in 'Immunostaining'. We demonstrated that the CD70^{low} TE-like novel intermediates resulted in more KRT7$^+$ iTS cell colonies as compared to unenriched or CD70^{high} naive populations, indicating that the identified TE-like cluster carries the greatest potential for the generation of iTSC^{d21n} (Extended Data Fig. 8k). For HLA experiments, cells were labelled with HLA-A, B, C (W6/32) or HLA-Bw4 (1:1, Purcell lab), then AF647 goat anti-mouse IgG antibody (1:1,000, ThermoFisher), or cells were labelled with (1) HLA-G MEM-G/9 (1:500, Abcam); (2) AF488 goat anti-mouse IgG antibody (1:1,000, ThermoFisher); and (3) PE-Cy7 mouse anti-human HLA-A, B, C W6/32 (1:200, Biolegend).

Multidimensional analyses of flow cytometry data

To visualize the multidimensional flow cytometry data, we used SPADE[38]. SPADE trees were generated as previously described[39] using the Cytobank platform (http://www.cytobank.org). Samples were labelled with antibodies as described in 'Flow cytometry analysis and fluorescent-activated cell sorting' for flow cytometry analysis and all experiments were performed on the same day to warrant their use for comparison. The SPADE tree indicates a clear transition of cell populations at the early stages of reprogramming (from day 0 to day 7), with reprogramming in NHSM and RSeT conditions exhibiting a more primed-like transition (Extended Data Fig. 3e). In particular, the RSeT medium formed a separated branch on the SPADE tree, in contrast to reprogramming in 5iLAF and t2iLGoY media (Extended Data Fig. 3e).

RT–qPCR

RNA was extracted from cells using RNeasy micro kit (Qiagen) or RNeasy mini kit (Qiagen) and QIAcube (Qiagen) according to the manufacturer's instructions. Reverse transcription was then performed using SuperScript III cDNA Synthesis Kit (ThermoFisher) or QuantiTect reverse transcription kit (Qiagen, cat. no. 205311), real-time PCR reactions were set up in triplicates using QuantiFast SYBR Green PCR Kit (Qiagen) and then carried out on the 7500 Real-Time PCR system (ThermoFisher).

RT–qPCR for miRNAs

miRNA and total RNA was extracted from cells using miRNeasy Mini Kit (Qiagen, cat. no. 217004) according to the manufacturers' instructions. They were then converted to cDNA using TaqMan MicroRNA Reverse Transcription Kit (Life Technologies, cat. no. 4366596). qPCR reactions were performed using QuantiFast SYBR Green PCR Kit (Qiagen). Data obtained from miRNA qPCR were analysed as follows: In each sample, hsa-miR-103a was used for normalization to obtain ΔC_t value for each miRNA. $2^{-\Delta Ct}$ was then calculated for each miRNA to obtain the relative expression against hsa-miR-103a. The values obtained were multiplied by 1,000 and then the results were plotted in logarithmic scale[26] (Fig. 4i). All primers used are listed in the Supplementary Table 15.

Immunostaining

Cells were fixed in 4% PFA (Sigma), permeabilized with 0.5% Triton X-100 (Sigma) in DPBS (ThermoFisher) and blocked with 5% goat serum (ThermoFisher). All antibodies used in this study are described in Supplementary Table 16 (for example, primary antibodies used were rabbit anti-KLF17 polyclonal (1:500, Sigma) and mouse anti-TRA-1-60 IgM (1:300, BD)). Primary antibody incubation was conducted overnight at 4 °C on shakers followed by incubation with secondary antibodies (1:400) for 1 h. Secondary antibodies used in this study were: goat anti-mouse IgM AF488 (1:400, ThermoFisher) or goat anti-mouse IgM AF647 (1:400, Invitrogen) for TRA-1-60; and goat anti-rabbit IgG AF555 (1:400, ThermoFisher) or goat anti-rabbit IgG AF647 (1:400, ThermoFisher) for KLF17 (Supplementary Table 16). After labelling, cells were stained with 4′,6-diamidino-2-phenylindole, dihydrochloride (DAPI) (1:1,000, ThermoFisher) for 30 min. Images were taken by IX71 inverted fluorescent microscope (Olympus). For whole-well (24-well plates) scanning of TRA-1-60 positive colonies for primed condition, KLF17$^+$ colonies for naive condition, and KRT7$^+$ colonies for Extended Data Fig. 8k, a DMi8 microscope (Leica) was used, and the number of colonies in each well was quantified using ImageJ. For Extended Data Fig. 9g, NR2F2 was used as a trophoblast marker, as suggested by a recent study[40].

snRNA-seq of human reprogramming intermediates

For snRNA-seq experiments, day 0, day 4, day 8, day 12 primed, day 12 naive, day 16 primed, day 16 naive, day 20 primed, day 20 naive, day 24 primed, day 24 naive, iPS cell naive (passage 3), iPS cell primed (passage 20) and iPS cell naive (passage 20) were collected and cryopreserved. These collected samples were then subjected to FACS, for day 0, day 4, day 8, day 12 primed, day 12 naive, day 16 primed, day 16 naive, day 20 primed, day 20 naive, day 24 primed and day 24 naive samples were sorted for PI$^-$TRA-1-85$^+$ cells to remove dead cells and iMEF cells, and iPS cell primed (passage 3) and iPS cell naive (passage 3 and passage 20) samples were sorted for PI$^-$TRA-1-85$^+$CD13$^-$F11R$^+$TRA-1-60$^+$EPCAM$^+$ cells to get rid of dead cells and iMEF cells, as well as differentiated cells. snRNA-seq library preparation was then prepared separately on each time point, generating 14 libraries (Fig. 1a). Nuclei were prepared using the Frankenstein protocol for nuclei isolation from fresh and frozen tissue followed by 10x Genomics that can be found in https://www.protocols.io/view/frankenstein-protocol-for-nuclei-isolation-from-f-3fkgjkw. In brief, cells were thawed and pelleted at 500*g* for 5 min at 4 °C. Five hundred µl of chilled Nuclei EZ lysis buffer supplemented with 0.2 U/µl RNase inhibitor was added to the pellet of cells and resuspended gently with a 1,000-µl bore tip and rest on ice for 5 min to complete lysis. The homogenate was filtered once using a 70-µm Flowmi filter and centrifuged at 500*g* for 5 min at 4 °C. After removing the supernatant (leaving 50 µl behind) the nuclei pellet was washed with 1,000 µl of chilled nuclei wash and resuspension buffer (1× PBS, 1.0% BSA, 0.2 U/µl RNase inhibitor). The nuclei were again pelleted at 500*g* for 5 min at 4 °C, removing supernatant and leaving behind about 50 µl; then, nuclei were gently resuspended in 1,000 µl nuclei wash and resuspension buffer. Nuclei were pelleted, supernatant removed and resuspended in 300 µl of nuclei wash and resuspension buffer supplemented with DAPI (10 µg/ml). The nuclei suspension was filtered using a 40-µm Flowmi filter; nuclei integrity was visually inspected under a microscope; and we proceeded with cytometric analysis and sorting based on DNA content using 70-µm nozzle, gating for single nucleus and sorting directly into reverse transcription buffer without reverse transcription enzyme: 20 µl reverse transcription buffer, 3.1 µl TSO primer, 2 µl additive B and 30 µl H_2O. After sorting nuclei (1,000–7,000 nuclei depending on sample), volume was completed to 80 µl with H_2O, 8.3 µl reverse transcription enzyme C was added, mixed and proceeded with chip loading. All subsequent steps were carried out as described in the Chromium Single Cell 3′ Reagent Kits User Guide (v.3 Chemistry). Sequencing was done on a Illumina NovaSeq 6000 using a paired-end 2×150-bp sequencing strategy and aiming for 30,000 read-pairs per nucleus. Chromium barcodes were used for demultiplexing and FASTQ files were generated from the mkfastq pipeline using the Cellranger program

(v.3.0.2). Alignment to hg19 genome (GRCh37, CellRanger reference version 1.2.0, genome build GRCh37.p13, which contained the Sendai virus KLF4, MYC and SeV sequences as extra chromosomes) and unique molecular identifier (UMI) counting were then performed using Cellranger against Ensembl's GRCh37 genome annotation (version 82, including protein-coding, long intergenic non-coding RNA and antisense biotypes; modified as premRNA, UMIs assigned to transcript) containing the Sendai virus sequences as extra transcripts. The endogenous expression of Yamanaka factors was quantified by counting sequencing reads only against the 5′ and 3′ UTR regions of the transcripts of endogenous *OCT4*, *KLF4*, *SOX2* and *MYC*.

scRNA-seq of human reprogramming intermediates

For scRNA-seq experiments, day 0, day 3, day 7, day 13 primed, day 13 naive, day 21 primed, day 21 naive, iPS cells primed (passage 3) and iPS cells naive (passage 3) were collected and cryopreserved. These collected samples were then subjected to FACS, for day 0, day 3, day 7, day 13 primed, day 13 naive, day 21 primed and day 21 naive samples were sorted for $PI^{-}TRA\text{-}1\text{-}85^{+}$ cells to remove dead cells and iMEF cells, and iPS cells primed (passage 3) and iPS cells naive (passage 3) samples were sorted for $PI^{-}TRA\text{-}1\text{-}85^{+}CD13^{-}F11R^{+}TRA\text{-}1\text{-}60^{+}EPCAM^{+}$ cells to get rid of dead cells and iMEF cells, as well as differentiated cells. Three samples were prepared in Extended Data Fig. 1c for subsequent library preparation, sample one contained cells isolated from day 0, 3 and 7, samples two and three contained cells for primed (day 13, day 21 and iPS cells) and naive reprogramming (day 13, day 21 and iPS cells), respectively, and a small number of mixed day 0, 3 and 7 cells were added to sample two and three to capture the full reprogramming trajectories and also to account for potential batch effects. The collected cells were isolated, encapsulated and library constructed using Chromium controller (10x Genomics) as per the manufacturer's instructions (Chromium Single Cell 3′ Reagent Kit V2 User Guide, 10X Genomics document number CG00052 Revision 3). A total of 12 cDNA amplification cycles were used. A total of 16 cycles of library amplification were used. Sequencing was carried out using an Illumina NextSeq 500 using SBS V2 chemistry in a high-output mode according to the recommendations outlined by 10x Genomics Chromium Single Cell 3′ Reagent Kit V2 User Guide (10x Genomics document number CG00052 Revision 3), with the exception that the second read was extended to 115 bp instead of 98 bp. Libraries were diluted according to the manufacturer's instruction (NextSeq 500 System User Guide, Illumina document number 15046563 v02) and loaded at 1.8 pM. Chromium barcodes were used for demultiplexing and FASTQ files were generated from the mkfastq pipeline using the Cellranger program (v.2.1.0). Alignment and UMI counting were performed to the hg19 genome as per the snRNA-seq (except that mature mRNA was used rather than premRNA, UMIs assigned to exons). The same experimental procedure and the computational pipeline were also applied to generate the RSeT reprogramming scRNA-seq library shown in Extended Data Fig. 4a, b.

scRNA-seq of day-21 fibroblast, naive and iTSC^{d8} reprogramming intermediates

For Extended Data Fig. 9a, day-21 fibroblast, naive and iTSC^{d8} reprogramming intermediates were collected and sorted for $PI^{-}TRA\text{-}1\text{-}85^{+}$ cells to remove dead cells and iMEF cells. The collected cells were isolated, encapsulated and constructed using Chromium controller (10x Genomics) as per the manufacturer's instructions (Chromium Next GEM Single Cell 3′ Reagent Kit V3.3 User Guide). Sequencing was done on an Illumina NovaSeq 6000 using a paired-end (R1 28 bp and R2 87 bp) sequencing strategy and aiming for 20,000 read-pairs per cell. Chromium barcodes were used for demultiplexing and FASTQ files were generated from the mkfastq pipeline using the Cellranger program (v.3.1.0). Alignment and UMI counting were performed to the hg19 genome as per the scRNA-seq experiments.

snRNA-seq and scRNA-seq cell calling and quality control

To identify the cell-containing droplets, cell calling was performed on the raw_gene_bc_matrices generated by the Cellranger program as follows. All the cell barcodes are ranked in order of decreasing the number of total UMI counts. The $\log_{10}$-transformed total UMI counts (*y*-axis) were then plotted against the $\log_{10}$-transformed rank (*x*-axis). The first 'knee' point in this UMI-barcode rank plot represents a marked drop in the total UMI counts, shifting from cell-containing barcodes to the majority of non-cell-containing barcodes. To determine this knee point, a linear model was fitted on the UMI-barcode rank plot between the top n_{upper} and n_{lower} ranks. Barcodes that deviate negatively from the linear model by more than k_{cut} on the *y*-axis are then deemed to have passed the knee point and discarded. This cell calling procedure was performed on each library separately using $n_{upper} = 100$, $n_{lower} = 400$, $k_{cut} = 0.15$ for the snRNA-seq and $n_{upper} = 100$, $n_{lower} = 500$, $k_{cut} = 0.2$ for the scRNA-seq. This resulted in a total of 38,100 cells and 7,674 cells for the snRNA-seq and scRNA-seq, respectively. Quality control was first performed at the cell level. Cells with (i) extremely high total UMI counts (nUMI), (ii) low number of expressed genes (nGene), (iii) high percentage mitochondrial genes (pctMT) or (iv) low percentage housekeeping genes (gene list from ref. [41]) (*pctHK*) were discarded. Cutoffs of nUMI > 15,000, nGene < 1,200 and nUMI > 50,000, nGene < 1,800, pctMT > 12, pctHK < 10 were applied to discard cells for the snRNA-seq and scRNA-seq, respectively. No pctMT and pctHK cutoffs were applied in the case of snRNA-seq, as there were very few mitochondrial or housekeeping genes detected. Next, quality control was performed at the gene level. Genes with (i) low $\log_{10}$(average UMI) [$\log_{10}$(aveUMI)] or (ii) do not have at least 'minUMI' UMIs in at least 'minCell' cells were discarded. Cutoffs of $\log_{10}$(aveUMI) < −2.5, minUMI = 2, minCell = 10 and $\log_{10}$(aveUMI) < −2, minUMI = 2, minCell = 10 were applied to discard genes for the snRNA-seq and scRNA-seq, respectively. After quality control, 36,597 cells and 17,004 genes, and 7,194 cells and 12,246 genes, remained for the snRNA-seq and scRNA-seq, respectively.

snRNA-seq ambient RNA removal

From the UMI-barcode rank plot in the snRNA-seq libraries, we observed non-cell-containing barcodes with high total UMI counts (in the range of 500–750 UMIs as compared to 20–50 UMIs in the scRNA-seq libraries), indicating substantial ambient RNA contamination. To circumvent this, ambient RNA removal was then performed using the decontx algorithm[42] in the celda package (v.1.1.6). The decontx algorithm assumes that there are *K* cell populations and uses Bayesian variational inference to infer the ambient RNA contamination as a weighted combination of the *K* cell population distributions. Thus, the algorithm requires the raw UMI counts and population membership for each cell as input. To determine the cell population membership, we applied the Seurat (v.3.1.1) clustering pipeline[20] using the following functions with default settings unless otherwise stated: NormalizeData, FindVariableFeatures (with 2,000 features), ScaleData, RunPCA. The cell clusters were then obtained using the FindNeighbours (using the top 10 principal components (PCs)) and FindClusters (resolution = 0.5) functions. The Seurat clustering pipeline was applied to each snRNA-seq library separately and decontx was then performed on each library using the default settings. A random seed of 42 was used throughout the entire analysis.

snRNA-seq and scRNA-seq preprocessing and integration

To integrate both the snRNA-seq and scRNA-seq datasets, we used the Seurat v.3 integration technique (v.3.1.1)[43]. Seurat v.3 identifies 'anchors' or pairwise correspondences between cells in the two datasets, which is then used to harmonize the datasets. As part of the preprocessing step, the functions NormalizeData (with default settings), FindVariableFeatures (using 1,500 features) were applied to the snRNA-seq and scRNA-seq datasets separately. Furthermore, each cell was assigned cell-cycle scores (S score and G2M score) and a cell-cycle phase using

Seurat's CellCycleScoring function. The FindIntegrationAnchors function (using 1,500 features) was then executed to identify the anchors, followed by running the function IntegrateData on the genes that are detected in both datasets. This resulted in an integrated single-cell dataset comprising 43,791 cells and 11,549 genes (Supplementary Table 1). The list of feature genes is in Supplementary Table 2.

scRNA-seq and snRNA-seq dimension reduction and trajectory inference

To represent the single-cell data in a concise manner, we applied several dimension reduction techniques using the anchor feature genes identified in the data integration step. PCA was performed on the scaled gene expression using the RunPCA function in Seurat package (v.3.1.1). Following that, uniform manifold approximation and projection (UMAP) and *t*-distributed stochastic neighbour embedding (*t*-SNE) were implemented on the top 14 PCs (determined using an elbow plot) via the RunUMAP and RunTSNE functions, respectively. Diffusion maps were generated using the scanpy.pp.neighbours function (using the top 14 PCs) and scanpy.tl.diffmap function in the scanpy package (v.1.4.4.post1)[44]. FDL was generated using the scanpy.tl.draw_graph function in the scanpy package using the ForceAtlas 2 layout and initialized using the UMAP coordinates. To infer the trajectories present in our single-cell data, we applied three different approaches. First, we applied the 'Cellular Trajectory Reconstruction Analysis using gene Counts and Expression' (CytoTRACE, v0.1.0) algorithm[18], which orders the single cells on the basis of their differentiation potential. As our dataset comprises two different assays, we ran CytoTRACE in the integrated mode, which integrates the scRNA-seq and snRNA-seq data using the Scanorama method before calculating the differentiation potential. The raw counts were supplied as input and default settings were used. Second, we used Monocle3 (v.0.1.3)[19], which learns a trajectory graph from a dimension reduction. In particular, we did a modification in which we supplied the FDL dimension reduction calculated previously into Monocle3 and ran the cluster_cells (using $k = 30$ neighbours) and learn_graph functions in the Monocle3 package to obtain an FDL-based Monocle3 trajectory. Third, we used PAGA[21] which quantifies the connectivity between clusters of cells and generates an abstracted graph representing the trajectories observed during reprogramming. The PAGA algorithm was performed using the scanpy.tl.paga function in the scanpy package (v.1.4.4.post1) using the Seurat cell clusters as input. The generation of the cell clusters is described in 'scRNA-seq and snRNA-seq cell clustering'.

scRNA-seq and snRNA-seq cell clustering

The single cells were clustered using the FindNeighbours (using the top 14 PCs for consistency with the dimension reductions) and FindClusters function (resolution = 0.5) in the Seurat (v.3.1.1) package, which implements an unsupervised graph-based algorithm. This resulted in 21 clusters, which were then labelled using a combination of letters and a number (for example, cluster FM1) that were determined from the cell composition of the cluster (FM, fibroblast medium; mix, shared clusters; PR, primed reprogramming; NR, naive reprogramming; NIC, novel intermediate cluster; RE, refractory cells) and the ordering of the cell population along reprogramming trajectory.

snRNA-seq differential expression and identification of gene signatures

As the data integration introduces dependencies between data points, we chose to perform the differential expression analysis solely on the snRNA-seq. The snRNA-seq data were chosen over the scRNA-seq as the former has more cells and a larger number of detected genes. Before differential expression, we performed clustering on only the snRNA-seq using the procedure described in 'scRNA-seq and snRNA-seq cell clustering' (using the top 12 PCs instead of the top 14 PCs), generating 21 snRNA clusters (Extended Data Fig. 2d). Pairwise differential expression between the 21 snRNA clusters was performed using the Wilcoxon rank-sum test on the log-transformed gene expression. The Wilcoxon rank-sum test *P* values were then adjusted for multiple testing using the Benjamini–Hochberg procedure to yield the false discovery rate (FDR). Genes are deemed differentially expressed if the $\log_2$-transformed fold change is >1.5 and the FDR is <0.01.

To identify gene signatures, we first define cluster-specific marker genes for each of the 21 snRNA clusters. For each snRNA cluster, we define marker genes as genes that have an average $\log_2$-transformed fold change (averaged across all 20 pairwise differential expressions) of >1.5, and we also require the genes to be differentially expressed in at least 14 of the 20 pairwise differential expressions. Hierarchical clustering was then performed on the Jaccard similarity of the marker genes (Extended Data Fig. 2f) to identify overlapping gene sets that is, the gene signatures. Overall, we identified eight gene signatures (Supplementary Table 3) that we named fibroblast (snRNA-FM1, snRNA-FM2, snRNA-FM3, snRNA-FM4); mixed (snRNA-mix); early-primed (snRNA-PR1); primed: snRNA-PR2, snRNA-PR3, snRNA-PR4); novel intermediate signature (snRNA-nic); naive (snRNA-NR1, snRNA-NR2, snRNA-NR3, snRNA-NR4); non-reprogrammed (nonReprog)1 (snRNA-RE1, snRNA-RE3, snRNA-RE4, snRNA-RE5); nonReprog2 (snRNA-RE6). The marker genes for clusters snRNA-RE2 and snRNA-FM5 were not used, as there are very few genes. Furthermore, in the fibroblast, primed, naive and nonReprog1 gene signatures, which comprises marker genes from more than one cluster, we picked only genes that are called marker genes at least twice to be included in the gene signature. One mitochondrial gene was then removed, resulting in a total of 504 genes across all 8 gene signatures (Supplementary Table 3). We then determined the strength of each gene signature in every single cell by calculating the average expression of the genes of interest subtracted by the aggregated expression of a set of control genes[41]. The control genes were determined by binning all detected genes into 25 gene-expression bins and 100 genes were then randomly selected from the same bin for each gene in the gene signature. Every single cell was then assigned to one of the eight gene signatures on the basis of the highest gene signature strength. This was then used to track the cell identity changes during reprogramming (Extended Data Fig. 2i). The same gene signature calculations were also applied to determine the strength of TE, epiblast (EPI) and primitive endoderm (PE) gene signatures in each single cell (Fig. 3a, Extended Data Fig. 7g). Furthermore, gene signatures related to the S and G2M cell-cycle phases were calculated to predict the cell-cycle phase (Extended Data Fig. 1h). Single cells were assigned to the G1 phase if both S and G2M scores were less than zero. Otherwise, they were assigned either the S or G2M phase on the basis of the higher of the S and G2M scores.

scRNA-seq analysis of RSeT reprogramming

The RSeT reprogramming scRNA-seq dataset was analysed together with the FM, primed reprogramming and naive reprogramming scRNA-seq counterparts (Supplementary Table 6). The raw UMI counts of all four scRNA-seq libraries were combined and subjected to the same quality control cutoffs: $n_{upper} = 100$, $n_{lower} = 500$, $k_{cut} = 0.2$ for cell calling, nUMI > 50,000, nGene < 1,800, pctMT > 12, pctHK < 10 for cell quality control and $\log_{10}$(aveUMI) < −2, minUMI = 2, minCell = 10 for gene quality control. This resulted in 9,852 cells and 12,590 genes after quality control. Subsequently, the combined scRNA-seq datasets were analysed using a similar workflow as the previous scRNA-seq and snRNA-seq datasets. The dataset was preprocessed using Seurat v.3's NormalizeData (with default settings), FindVariableFeatures (using 1,500 features) functions. Next, PCA was performed, followed by other dimension algorithms (UMAP, *t*-SNE, diffusion maps and FDL) using the top 15 PCs. We found that the RSeT cells follow the naive trajectory, but we also observed a primed-like cluster of cells, expressing primed-associated markers such as *ZIC2* and *NLGN4X* (Extended Data Fig. 4a, b). It has previously been shown that primed cells have a growth

advantage over the naive population[4], and hence this could be the reason that they become the dominant population in the RSeT medium over time. These results suggest that RSeT is a more permissive condition that allows the derivation of a continuum of pluripotent states[4,6].

scRNA-seq analysis of day-21 reprogramming intermediates

The day-21 reprogramming intermediates scRNA-seq libraries were analysed using a similar workflow as the previous scRNA-seq and snRNA-seq datasets (Supplementary Table 13). In brief, quality control was performed at both cell and gene level with the following cutoffs: $n_{upper} = 100$, $n_{lower} = 500$, $k_{cut} = 0.2$ for cell calling, nUMI > 50,000, nGene < 1,800, pctMT > 12, pctHK < 10 for cell quality control and $\log_{10}$(aveUMI) < −2, minUMI = 2, minCell = 10 for gene quality control. This resulted in 10,518 cells and 12,611 genes after quality control. Subsequently, the dataset was preprocessed using Seurat v.3's NormalizeData (with default settings), FindVariableFeatures (using 1,500 features) functions. Next, PCA was performed, followed by other dimension algorithms (UMAP, *t*-SNE, diffusion maps and FDL) using the top 15 PCs. We also applied cell clustering (using the same top 15 PCs and resolution = 0.5), identifying 13 clusters. These clusters were then labelled using a combination of letters and a number (for example, cluster d21TR1) that were determined from the cell composition of the cluster (d21FM: fibroblast medium; d21NR: naive reprogramming; d21TR, TS cell reprogramming) and the ordering of the cell population along reprogramming trajectory. The strength of the eight gene signatures defined in this study is also calculated as per the previous scRNA-seq and snRNA-seq dataset.

RNA-seq of reprogramming intermediates

For the bulk RNA-seq of the FACS-purified reprogramming intermediates (Extended Data Fig. 3), RNA extraction was performed using the RNeasy micro kit (Qiagen, cat. no. 74004) from about $2–20 \times 10^4$ cells with QIAcube (Qiagen). The concentrations of RNA were measured by a Qubit RNA HS Assay Kit (ThermoFisher, cat. no. Q32855) on a Qubit 2.0 Fluorometer (ThermoFisher). About 25 ng of RNA was used for library construction with the SPIA kit (NuGen) and subsequently sequenced by a HiSeq 1500 or HiSeq 3000 sequencer (Illumina). Sequencing libraries were single-end with 50-nt length and a targeted number of reads of 20–30 million.

RNA-seq analysis of reprogramming intermediates

Bulk RNA-sequencing reads generated in this study, ref. [22] (d0 fibroblasts, $n = 2$ (donor identifiers 32F and 55F, biological replicates)) and ref. [4] (P3 t2iLGoY, P10 t2iLGoY, P3 RSeT, P10 RSeT, P3 NHSM, P10 NHSM, P3 5iLAF, P10 5iLAF; all conditions with $n = 2$ (32F and 55F)) were processed as follows: low-quality sequencing reads and were filtered and trimmed with Trimmomatic[45] (v.0.36, Phred score of 6 consecutive bases below 15, minimum read length of 36 nt) and mapped to a custom version of hg19 human genome (with modifications described in 'scRNA-seq sequencing and processing') with STAR (v.2.4.2a)[46]. Gene read counting was performed with featureCounts (v.1.5.2, unstranded)[47] against the custom version of Ensembl's GRCh37 annotation with modifications described in 'snRNA-seq of human reprogramming intermediates'. From the resulting counts table, we retained genes that have (i) at least 10 counts in one sample and (ii) at least 2 counts per million (CPM) in at least two samples so as to remove the lowly expressed genes. Library normalization was then performed using the rpkm function in the edgeR package (v.3.24.3) with the arguments normalized.lib.sizes = TRUE and prior.count = 1 to yield fragments per kilobase per million (FPKM). PCA was then performed on the log-transformed $\log_2$(FPKM + 1) on the top 500 most highly variable genes using the prcomp_irlba function in the irbla package (v.2.3.3). To show the reprogramming trajectory in the 3D PCA plots, cubic splines were fitted independently on each PC using the splinefun function in base R (v.3.5.1).

Projection of bulk RNA-seq samples onto single-cell data

To project the bulk RNA-seq samples of FACS-purified reprogramming intermediates onto the single-cell data, we treated each bulk RNA-seq sample as a single cell and performed the same Seurat v.3 integration technique that was previously used to integrate both the snRNA-seq and scRNA-seq. The same procedure was applied with the exception that the arguments k.filter = 20 and k.score = 10 were supplied to the FindIntegrationAnchors function to adjust for the fact that the bulk RNA-seq contains far fewer samples (50 samples) than the single-cell counterpart. We then aggregated the gene expression of the combined gene expression as follows. For the bulk RNA-seq, samples were aggregated on the basis of the medium condition and time point. For the single cells, the scRNA-seq cells and non-reprogrammed cells were removed and the remaining single nucleus was aggregated on the basis of the medium condition and time point.

Scoring of bulk RNA-seq samples using the primed or naive gene signatures and TE, EPI and PE signatures

For the bulk RNA-seq samples of reprogramming intermediates, we used a simple scoring system to determine the strength of different gene signatures (Supplementary Table 5). To compute the scores for each sample, the gene expression of the gene set of interest was first divided by the maximum gene expression across all samples to obtain a scaled gene expression ranging from 0 to 1. The scaled gene expression was then averaged across all the genes in the gene set to give the final score, which ranges from 0 to 1. This scoring system was applied to determine the strength of the primed and naive pluripotency using the genes in the primed and naive gene signatures determined from the single-cell data, respectively. We also used gene sets that are highly expressed in the EPI, PE and TE on the basis of a previous study[25]. In particular, we obtained the top 100 genes, ordered by differential expression FDR in that study, for each of the three lineages across embryonic day (E)5 to E7, giving rise to the ALL EPI, ALL PE and ALL TE gene sets. Furthermore, we also extracted the top 100 genes for each embryonic day, giving rise to day-specific EPI (E5-EPI, E6-EPI and E7-EPI), PE and TE gene sets. These gene sets can be found in Supplementary Table 11. To validate this scoring approach, gene set enrichment analysis on each medium and time point condition was performed as follows. Condition-specific differential expression was performed using the empirical Bayes quasi-likelihood *F*-tests in the edgeR package (v.3.24.3) between the condition of interest and the average expression of the remaining conditions. Gene set enrichment analysis was then performed on the log-transformed fold changes from these differential expression results using the fgsea package (v.1.8.0) with 10,000 permutations.

RNA-seq for characterization of iTS cells and iTS-cell-differentiated cells

For the bulk RNA-seq of the iTS cell and iTS-cell-differentiated cells, RNA-seq was performed with a multiplexing approach, using an 8-bp sample index[48] and a 10-bp UMI were added during initial poly(A) priming and pooled samples were amplified using a template-switching oligonucleotide. The Illumina P5 (5′ AAT GAT ACG GCG ACC ACC GA 3′) and P7 (5′ CAA GCA GAA GAC GGC ATA CGA GAT 3′) sequences were added by PCR and Nextera transposase, respectively. The library was designed so that the forward read (R1) uses a custom primer (5′ GCC TGT CCG CGG AAG CAG TGG TAT CAA CGC AGA GTA C 3′) to sequence directly into the index and then the 10-bp UMI. The reverse read (R2) uses the standard R2 primer to sequence the cDNA in the sense direction for transcript identification. Sequencing was performed on the NextSeq550 (Illumina), using the V2 High output kit (Illumina, no. TG-160-2005) in accordance with the Illumina Protocol 15046563 v.02, generating 2 reads per cluster composed of a 19-bp R1 and a 72-bp R2.

Analysis of the RNA-seq of iTS cells and iTS-cell-differentiated cells

The sequencing reads are demultiplexed using sabre (v.1.0) using the barcodes-sample table, and allowing up to one mismatch per barcode, and a minimum UMI length of 9 bp. The demultiplexed data has single reads per sample and UMIs are added to the read name. We use STAR (v.2.5.2b)[46] to align the reads to the GRCh37 Ensembl reference genome (v.87). Read deduplication based on UMIs was performed with je markdupes (v.1.2)[49] and transcript read counts calculated with featureCounts (v.1.5.2)[47]. From the resulting counts table, lowly expressed genes were filtered and library normalization was performed as per the bulk RNA-seq analysis of reprogramming intermediates. We then compared the similarity of the transcriptomes of our iTS cells, iTS-cell-derived EVTs and STs with published transcriptomic datasets: namely (i) blastocyst-derived TS cell gene expression from refs. [7,30]; (ii) trophoblast organoid gene expression from refs. [28,29]; and (iii) single-cell gene expression (only Smart-seq2) of the fetal–maternal interface from ref. [27]. The removeBatchEffect function in the limma package (v.3.38.3) was applied to our dataset and each of the three sets of external datasets separately to account for technical differences, followed by Spearman correlation between the two datasets.

ATAC-seq

ATAC-seq samples were prepared as previously described[50]. In brief, reprogramming intermediates and iPS cells were isolated by FACS (Supplementary Table 4) and about 65,000 cells were washed and lysed in ATAC-seq lysis buffer (10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630, 10 mM Tris pH 7.4). The transposition reaction was then carried out by using 22.5 µl of UltraPure Distilled Water (ThermoFisher, cat. no. 10977-015), 25 µl of Tagment DNA Buffer (Illumina, cat. no. 15027866) and 2.5 µl of Tagment DNA Enzyme 1 (Illumina, cat. no. 15027865) for each sample, and then incubated for 30 min at 37 °C, followed by immediate purification using a MinElute Reaction Cleanup Kit (Qiagen, cat. no. 28204) according to the manufacturer's instructions. Eleven µl of transposed DNA, 25 µl of the NEBNext High-Fidelity 2× PCR Master Mix (cat. no. M0541S) and 1.25 µM of the adaptor sequences, as previously published[50], were used in a 50-µl PCR reaction. PCR parameters were: 72 °C for 5 min, 98 °C for 30 s and 9 cycles of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 1 min. The prepared libraries were purified using a MinElute PCR purification kit (Qiagen, cat. no. 28004) followed by Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63880) according to the manufacturer's specifications, in which library fragments ranging from 200 to 700 bp were selected and sequenced on an Illumina HiSeq 1500 in 2×51 cycle paired-end mode.

ATAC-seq preprocessing and alignment

ATAC-seq reads (pair-end 51-nt reads) were adaptor-trimmed and filtered by base quality and length using Cutadapt v.1.8[51]using -a CTGTCTCTTATACACATCT, -A CTGTCTCTTATACACATCT, -q 20, and–minimum-length 18 options. Read pairs passing filters were mapped to the complete human genome (hg19 human genome (UCSC version, December 2011)) using Bowtie2 with -X 2000, –no-mixed and–no-discordant options[52]. Mapped sample reads were filtered for multi-mappers (mapping quality <10) and reads mapped to mitochondrial DNA using Jvarkit's[53] samjs. PCR duplicates were discarded using Picard's (http://broadinstitute.github.io/picard) MarkDuplicates tool. Sequencing reads aligned to known genomic blacklisted regions were also not considered for further analysis[54].

ATAC-seq peak calling and exploratory analysis

Peak calling was performed on each biological replicate with MACS2 callpeak subcommand[55] using–nomodel -f BAM–keep-dup all–gsize hs–shift -100–extsize 200–SPMR -B options. For downstream analysis we used an 'intersect and rescue' approach. This approach consisted of intersecting each time point and reprogramming medium biological replicates peak sets (bedtools intersect)[56], subcommand (-wa -wb -F/-f 0.3) and then filtering those peaks with a fold change over background of more than 5 fold change and at least 3 fold change in the other replicate. This created two intersection peaksets (major to 5 fold change in replicate 1 and major to 3 fold change in replicate 2 and vice versa), which were then combined and merged with bedtools merge (a minimum of 1 bp overlap). The union peak set of both replicates for each time point and reprogramming medium was then reduced by merging all peaks within 100 bp. Finally, a consensus peak set of all time points and reprogramming media was created using bedtools merge as described. Sequencing read counts for each biological replicate time point and medium were produced using featureCounts[47](-p -F SAF), FPKMs calculated (peaks with less than 5 FPKMs in at least 2 samples were discarded) then $\log_2$-transformed ($\log_2 + 1$) and quantile-normalized. Genome coverage plots were generated using wiggleplotr bioconductor package[57]. PCA was then performed on the $\log_2$-transformed FPKM on all features using the prcomp_irlba function in the irbla package (v.2.3.3). Human in vivo inner cell mass and ES cell samples from ref. [58], human blastocyst-derived TS cells (BT5) from ref. [30] were processed as described in 'ATAC-seq preprocessing and alignment'. We noted that regulatory elements of the fibroblast marker *ANPEP* became less accessible by day 7, accompanied by the downregulation of *ANPEP* gene expression. By contrast, this was followed by a gain of chromatin accessibility of regulatory elements and/or promoter regulatory elements of genes associated with shared pluripotency (*PRDM14*), primed pluripotency (*SOX11*) or naive pluripotency (*DNMT3L*) gain accessibility, which coincides with the upregulation of these pluripotency genes (Extended Data Fig. 5d, e). We also observed naive-specific open chromatin regions in proximity to or within the gene body of naive pluripotency factors such as *KLF17*, *ZNF729*, *NANOG* and *POU5F1* (*OCT4*) as previously reported in ATAC-seq datasets of in vivo human embryos[58,59] (Extended Data Fig. 5f). In particular, we found that the chromatin accessibility of two previously identified naive enhancers at the *OCT4* and *NANOG* loci[59], also detected in human inner cell mass[58], became gradually accessible up to day 7 while the cells were still in FM. Following this, these regions lost accessibility in the primed intermediates and iPS cells, while remaining open in naive cells (Extended Data Fig. 5f).

Integration of bulk ATAC-seq samples with bulk RNA-seq samples

To integrate the bulk ATAC-seq profiles with the bulk RNA-seq samples, we first selected ATAC-seq peaks that are within an activity distance of −100 to 10 bp around the transcription start site (TSS) of each gene and assigned these peaks to the corresponding gene. Next, we further integrated the two assays by performing upper quartile normalization, which makes the transcript counts and peak intensities distributions comparable and the removeBatchEffect command in the limma package (v.3.38.3) to the combined $\log_2$-transformed ($\log_2 + 1$) ATAC and RNA dataset, specifying that the terminal time points (namely, fibroblast-d0, primed-P10 and t2iLGoY-P10) are to be preserved using the design argument. PCA was then performed on this integrated dataset using the top 1,000 most highly variable genes. To characterize gene expression of genes associated with identified cluster peaks (see 'ATAC-seq peak calling and exploratory analysis'); annotated peaks with no genes associated to (intergenic) were discarded and in cases of peaks assigned to the same gene, the peak closest to the gene's TSS was selected. Bulk RNA-seq gene read counts were processed as described in 'RNA-seq analysis of reprogramming intermediates', $\log_2$-transformed FPKMs (log2 +1) and *z*-scores across all conditions calculated. Gene ontology (GO) analysis of genes associated to each cluster was then performed using the Metascape[60], web interface (https://metascape.org/) on GO biological processes with default settings. The top 20 enriched GO terms for each cluster are presented in Supplementary Table 8.

ATAC-seq fuzzy cluster analysis

Processing of the read counts for fuzzy clustering and *c*-means clustering was performed as previously described[11]. In summary, sequencing read counts of each biological replicate were aggregated, FPKMs calculated discarding peaks with less than 10 in any condition then $\log_2$-transformed ($\log_2 + 1$) and quantile-normalized. Only peaks with a coefficient of variation across time points and media higher than 20% were considered for clustering. This peak subset was *z*-scaled and *c*-means fuzzy clustering[23] was performed ($m = 1.243778$, 8 clusters) (Supplementary Table 7). A cluster membership threshold of 0.8 was used for downstream analysis.

ATAC-seq peak annotation and Motif analysis

Cluster peaks were annotated using Homer's annotatePeaks subcommand[61] and annotatr[62]. A motif enrichment analysis of cluster peaks was performed using Homer's findMotifsGenome (-size given) for known motifs (Supplementary Table 9).

Statistics and reproducibility

For the snRNA-seq and scRNA-seq experiments of the reprogramming roadmap, specific library information can be found in Fig. 1b and Supplementary Table 1. For time-resolved snRNA-seq experiments, a total of $n = 14$ biologically independent samples across 14 media and time points were included. Each sample was then subjected to snRNA-seq. The media and time points are d0-FM ($n = 1$), d4-FM ($n = 1$), d8-FM ($n = 1$), d12-PR ($n = 1$), d12-NR ($n = 1$), d16-PR ($n = 1$), d16-NR ($n = 1$), d20-PR ($n = 1$), d20-NR ($n = 1$), d24-PR ($n = 1$), d24-NR ($n = 1$), P3-NR ($n = 1$), P20-PR ($n = 1$) and P20-NR ($n = 1$). For the medium-resolved scRNA-seq experiments, a total of $n = 9$ biologically independent samples across 9 media and time points were included. The media and time points are d0-FM ($n = 1$), d3-FM ($n = 1$), d7-FM ($n = 1$), d13-PR ($n = 1$), d13-NR ($n = 1$), d21-PR ($n = 1$), d21-NR ($n = 1$), P3-PR ($n = 1$) and P3-NR ($n = 1$). These samples are then pooled into three scRNA-seq libraries, which are the FM library (d0-FM, d3-FM, d7-FM samples), primed reprogramming library (d0-FM, d3-FM, d7-FM, d13-PR, d21-PR, P3-PR samples) and the naive reprogramming library (d0-FM, d3-FM, d7-FM, d13-NR, d21-NR, P3-NR samples). The total number of cells used in the final analysis was 43,791 (Figs. 1b–g, 3a, Extended Data Figs. 7g,h, 8j). Detailed cell numbers for snRNA-seq and scRNA-seq in each figure are as follows: Fig. 1b, Extended Data Fig. 1e–g,k–r, 43,791 cells across 17 libraries (3,713 d0-FM cells, 3,511 d4-FM cells, 3,809 d8-FM cells, 2,472 d12-PR cells, 491 d12-NR cells, 4,506 d16-PR cells, 2,578 d16-NR cells, 2,680 d20-PR cells, 1,858 d20-NR cells, 2,148 d24-PR cells, 1,121 d24-NR cells, 2,169 P3-NR cells, 3,009 P20-PR cells, 2,532 P20-NR cells, 2,402 FM cells, 2,506 primed reprogramming cells and 2,286 naive reprogramming cells); Fig. 1f, Extended Data Fig. 2a–c, 43,791 cells across 21 clusters (2,691 FM1 cells, 1,326 FM2 cells, 955 FM3 cells, 1,098 FM4 cells, 862 FM5 cells, 1,424 FM6 cells, 1,474 mix cells, 1,756 PR1 cells, 3,069 PR2 cells, 646 PR3 cells, 1,042 NR1 cells, 879 NR2 cells, 4,270 NR3 cells, 6,049 NR4 cells, 505 NIC cells, 2,159 RE1 cells, 2,005 RE2 cells, 1,361 RE3 cells, 2,992 RE4 cells, 7,138 RE5 cells and 90 RE6 cells); Fig. 1g, 43,791 cells across 8 gene signatures (8,714 fibroblast cells, 2,575 mixed cells, 2,365 early-primed cells, 3,970 primed cells, 610 novel intermediate cells, 10,563 naive cells, 14,820 nonReprog1 cells and 174 nonReprog2 cells); Extended Data Fig. 1h, 43,791 cells across 3 cell-cycle phases (18,771 G1 cells, 12,090 S cells and 12,930 G2M cells); Extended Data Fig. 2d, 43,791 cells across 21 snRNA clusters (7,194 scRNA(unused) cells, 2,501 snRNA-FM1 cells, 1,197 snRNA-FM2 cells, 1,060 snRNA-FM3 cells, 1,392 snRNA-FM4 cells, 984 snRNA-FM5 cells, 1,164 snRNA-mix cells, 1,121 snRNA-PR1 cells, 638 snRNA-PR2 cells, 783 snRNA-PR3 cells, 1,592 snRNA-PR4 cells, 1,143 snRNA-NR1 cells, 3,020 snRNA-NR2 cells, 4,498 snRNA-NR3 cells, 1,039 snRNA-NR4 cells, 406 snRNA-NIC cells, 2,416 snRNA-RE1 cells, 1,160 snRNA-RE2 cells, 1,156 snRNA-RE3 cells, 6,530 snRNA-RE4 cells, 2,690 snRNA-RE5 cells and 107 snRNA-RE6 cells); Extended Data Fig. 2e, h, for gene expression trends, the normalized gene expression was averaged across all cells within the same cluster before log transformation; Extended Data Fig. 2f–h, pairwise differentially expressed genes between the 21 snRNA clusters were determined using two-sided Wilcoxon rank-sum test with *P* values adjusted for multiple testing using the Benjamini–Hochberg procedure, genes that (i) have an average $\log_2$-transformed fold change (averaged across all 20 pairwise differential expressions) of >1.5 and (ii) are differentially expressed ($\log_2$-transformed fold change >1.5 and FDR <0.01) in at least 14 of the 20 pairwise differential expressions are deemed cluster-specific marker genes for each of the 21 snRNA clusters. Hierarchical clustering was then performed on the Jaccard similarity of these marker genes to identify eight gene signatures (504 genes in total, 52 fibroblast genes, 67 mixed genes, 28 early-primed genes, 39 primed genes, 31 naive genes, 54 novel intermediate genes, 58 nonReprog1 genes and 175 nonReprog2 genes). For scRNA-seq of RSeT reprogramming, specific library information can be found in Extended Data Fig. 4a and Supplementary Table 6. In addition to the scRNA-seq experiments already mentioned, an additional $n = 3$ biological independent samples across 3 time points were included, namely d13 RSeT reprogramming (RR), d21-RR and P3-RR. These samples were then pooled into the RSeT reprogramming library containing the d0-FM, d3-FM, d7-FM, d13-RR, d21-RR, P3-RR samples. The total number of cells used in the final analysis (which included cells from the FM, primed reprogramming and naive reprogramming libraries already mentioned) was 9,852 (Extended Data Fig. 4). Detailed cell numbers for scRNA-seq in each figure are as follows: Extended Data Fig. 4a, 9,852 cells across 4 libraries (2,402 FM cells, 2,506 primed reprogramming cells, 2,286 naive reprogramming cells and 2,658 RSeT reprogramming cells). For scRNA-seq of day-21 reprogramming intermediates, specific library information can be found in Fig. 4a and Supplementary Table 13. A total of $n = 3$ biologically independent samples across 3 conditions were included. Each sample was then subjected to scRNA-seq. The conditions are d21 fibroblast medium (d21FM, $n = 1$), d21 naive reprogramming (d21NR, $n = 1$) and d21 TS cell reprogramming (d21TR, $n = 1$). The total number of cells used in the final analysis was 10,518 (Fig. 4a, b, Extended Data Fig. 9b–e). Detailed cell numbers for scRNA-seq of day-21 reprogramming intermediates in each figure are as follows: Fig.4a, 10,518 cells across 3 libraries (4,761 d21FM cells, 2,801 d21NR cells and 2,956 d21TR cells); Extended Data Fig. 9c, 10,518 cells across 13 clusters (89 d21FM1 cells, 531 d21FM2 cells, 329 d21FM3 cells, 268 d21FM4 cells, 480 d21FM5 cells, 315 d21FM6 cells, 2,797 d21FM7 cells, 147 d21NR1 cells, 899 d21NR2 cells, 1,771 d21NR3 cells, 301 d21TR1 cells, 629 d21TR2 cells and 1,962 d21TR3 cells); Extended Data Fig. 9b, d, the marked d21TR1 containing 301 cells comprises 6 d21FM cells, 16 d21NR cells and 279 d21TR cells. For bulk RNA-seq of reprogramming intermediates, specific library information can be found in Extended Data Fig. 3f and Supplementary Table 5. $n = 2$ biological replicates were obtained for each condition except for day 13 primed ($n = 3$), day 13 naive ($n = 3$) and passage 3 primed ($n = 4$) (Fig. 2a, Extended Data Figs. 3b, f, 5b, c). For the scoring of primed and naive signatures, gene expression trends and Spearman correlation comparisons, the FPKM values were averaged across replicates before $\log_2 + 1$ transformation (Fig. 2b, Extended Data Figs. 3g, 6f, 7d–f). Gene expression trends of genes associated with ATAC-seq peaks are shown as *z*-standardized values (Extended Data Fig. 6b, c). In Extended Data Fig. 7e, gene set enrichment analysis was then performed on the log-transformed fold changes from condition-specific differential expression results with 10,000 permutations. The *P* values from the gene set enrichment were then corrected for multiple testing via the Benjamini–Hochberg procedure to yield the FDR. The product of the normalized enrichment score (NES) and $-\log_{10}$(FDR) (NES × $-\log_{10}$(FDR)) is then plotted in the heat map in Extended Data Fig. 7e. For bulk RNA-seq of iTS-cell-related samples, specific library information can be found in Supplementary Table 14. $n = 2$ biological replicates were obtained for each condition except for $iTSC^{d21n}$ ($n = 3$), $iTSC^{d8}$ EVT ($n = 4$) and $iTSC^{d21n}$ EVT ($n = 4$) (Extended

Data Figs. 8b, 10b, i, j). For the Spearman correlation comparisons, the FPKM values were averaged across replicates before $\log_2$ transformation (Fig. 4h, Extended Data Fig. 10k, l). For ATAC-seq of reprogramming intermediates, specific library information can be found in Supplementary Table 5. $n = 2$ biological replicates were obtained for each condition. For PCA, each replicate peak counts FPKMs were calculated (peaks with less than 5 FPKMs in at least 2 samples were discarded), $\log_2$-transformed (log2 + 1) and quantile-normalized (Fig. 2c, Extended Data Fig. 5a). For fuzzy clustering, replicate counts were aggregated for each peak, FPKMs calculated (discarding peaks with less than 10 FPKM in any condition), $\log_2$-transformed ($\log_2 + 1$) and quantile-normalized. Peaks with a coefficient of variation <20% were discarded. This peak subset was *z*-scaled and *c*-means fuzzy clustering was performed ($m = 1.243778$, 8 clusters) (Supplementary Table 7). A cluster membership threshold of 0.8 was used for downstream analysis. The number of peaks per cluster is as follows: C1, 12,024; C2, 7,779; C3, 5,077; C 4, 3,334; C5, 9,117; C6, 10,129; C7, 4,885; C8, 7,739 (Fig. 2d). Cluster-specific peak trends are shown as the mean ± s.d. for each reprogramming medium (Fig. 2d). *P* values of motif enrichment analysis of cluster-specific peaks are calculated on the basis of a cumulative binomial distribution to then calculate the probability of detecting them in target sequences by chance (Fig. 2e). Chromatin accessibility trends for peak associated genes are shown as *z*-scaled across reprogramming stages calculated as described in 'ATAC-seq fuzzy cluster analysis' (Extended Data Fig. 6b, c). In Fig. 2f, for *TFAP2C* knockdown experiments, two reprogramming rounds were performed and for each round of reprogramming, $n = 3$ independent experimental replicates were transduced, reprogrammed and quantified separately for both scrambled controls and sh*TFAP2C* reprogramming into either primed or naive iPS cells. Primed: *P* value = 0.09, naive: *P* value = 0.001. Data are represented as mean ± s.e.m., the significance is determined statistically by two-tailed unpaired Student's *t*-test. For *GATA2* knockdown experiments, two reprogramming rounds ($n = 2$) were performed for primed reprogramming. For round 1: $n = 6$ independent experimental replicates were transduced, reprogrammed and quantified separately for both scrambled controls and sh*GATA2* reprogramming into either primed or naive iPS cells. For round 2: $n = 4$ independent experimental replicates for scrambled control primed reprogramming, $n = 5$ independent experimental replicates for scrambled control naive reprogramming, $n = 5$ independent experimental replicates for sh*GATA2* primed reprogramming and $n = 5$ independent experimental replicates for sh*GATA2* naive reprogramming. Primed: $P = 2.33 \times 10^{-12}$, naive: $P = 1.03 \times 10^{-5}$. Data are represented as mean ± s.e.m., the significance is determined statistically by two-tailed unpaired Student's *t*-test. For Fig. 3c–e, these experiments were repeated in $n = 4$ biological replicates (4 independent experiments from two donors) with similar results, and representative images are shown in the figure. For Fig. 3f, $n = 3$ biological replicates, 3 independent iTS cell lines were injected into three mice, and similar results were obtained, and representative results are shown in the figure. For Fig. 3g, 4 lesions were generated from iTS cell lines, collected and analysed, similar results were obtained and representative images are shown ($n = 4$ biological replicates). For Fig. 4d, e, the experiments were repeated independently ($n = 4$ biological replicates) with similar results and representative results are shown in the figure. For Fig. 4f, g, the experiments were repeated with 4 iTS cell lines obtained from the two donors (independently differentiated into STs and EVTs) with similar results, and representative images are shown in the figures ($n = 4$ biological replicates). For Fig. 4i, the experiments were repeated with 4 independent cell lines (obtained from the two donors) and each of the 4 experiments were performed in 2 technical replicates with similar results, and representative plots are shown in the figure ($n = 4$ biological replicates × 2 technical replicates). For Fig. 4k, $n = 3$ independent cell lines were injected to three mice, and similar results were obtained and a representative image is shown. For Fig. 4l, the serum of two independent experiments (2 iTS cell lines injected, 1 line per mouse) were measured in 2 technical replicates ($n = 2$ biological replicates × 2 technical replicates). Representative results are shown in the figure. For Fig. 4m, 4 lesions were generated, collected and analysed, similar results were obtained and representative images are shown ($n = 4$ biological replicates). For Extended Data Fig. 1a, more than 10 reprogramming experiments using two different donors were performed with similar results. Representative phase-contrast images are shown in the figure. For Extended Data Fig. 1b, representative images are shown from staining of $n = 2$ biological replicates. For Extended Data Fig. 3d, 4 experiments were independently performed (from two donors) with similar results and representative images are shown in the figures ($n = 4$ biological replicates). For Extended Data Fig. 3e, $n = 2$ biological replicates (from two donors) were used for analysis. For Extended Data Fig. 6g, the relative expression of *TFAP2C* and *GATA2* were measured in $n = 2$ independent experiments with technical replicates. Representative results are shown in the figure. For Extended Data Fig. 8a, the experiments were repeated independently with $n = 2$ biological replicates (from two donors) with similar results, and representative images are shown in the figures. For Extended Data Fig. 8c, these experiments were repeated in $n = 4$ biological replicates (4 independent experiments from two donors) with similar results, and representative images are shown in the figure. For Extended Data Fig. 8d, fusion index was used to quantify the efficiency of cell fusion, which is calculated by using the number of nuclei counted in the syncytia minus the number of syncytia, then divided by the total number of nuclei counted. The quantification was performed on $n = 5$ cell clusters counted randomly and independently across ST cells differentiated from two iTS cell lines (obtained from two different donors) with similar results, and representative results are shown in the figure. $P = 1.60 \times 10^{-7}$, data are represented as mean ± s.e.m., the significance is determined statistically by two-tailed unpaired Student's *t*-test. For Extended Data Fig. 8e, the conditioned media from $n = 6$ biological replicates (6 independent cell lines from 2 different donors were differentiated into STs) were tested for hCG pregnancy tests and similar results were obtained from these tests; representative results are shown in the figure. For Extended Data Fig. 8f, the conditioned media of two independent experiments (from two donors) were measured in 2 technical replicates ($n = 2$ biological replicates × 2 technical replicates). Representative results are shown in the figure. For Extended Data Fig. 8g, the serum of two independent experiments (2 iTS cell lines injected, 1 line per mouse) were measured in 2 technical replicates ($n = 2$ biological replicates × 2 technical replicates). Representative results are shown in the figure. For Extended Data Fig. 8h, 4 lesions were generated, collected and analysed ($n = 4$ biological replicates). For Extended Data Fig. 8i, 4 lesions were generated from iTS cell lines, collected and analysed, similar results were obtained and representative images are shown ($n = 4$ biological replicates). For Extended Data Fig. 8k, $n = 3$ independent experiments for unenriched and $CD70^{low}$ cells were performed and $n = 2$ for $CD70_{high}$ cells. For Extended Data Fig. 9g, the experiments were repeated independently with $n = 2$ biological replicates (from two donors) with similar results, and representative images are shown in the figures. For Extended Data Fig. 9h, the relative expression of *NANOG*, *ZIC2*, *KLF17*, *DPPA3*, *GATA2* and *KRT7* were measured in $n = 3$ independent experiments with technical replicates. For Extended Data Fig. 10a, the experiments were repeated with $n = 6$ biological replicates (3 independent cell lines derived from each of the two donors) with similar results and representative images are shown in the figure. For Extended Data Fig. 10c, fusion index was used to quantify the efficiency of cell fusion, which is calculated by using the number of nuclei counted in the syncytia minus the number of syncytia, then divided by the total number of nuclei counted. The quantification was performed on $n = 5$ cell clusters counted randomly and independently across ST cells differentiated from two iTS cell lines (obtained from two different donors) with similar results and representative results are shown in the figure. $P = 3.95 \times 10^{-7}$, data are represented as mean ± s.e.m., the

significance is determined statistically by two-tailed unpaired Student's *t*-test. For Extended Data Fig. 10d, the conditioned media from $n = 6$ biological replicates (6 independent cell lines from 2 different donors were differentiated into STs) were tested for hCG pregnancy tests and similar results were obtained from such tests, and representative results are shown in the figure. For Extended Data Fig. 10e, the conditioned media of two independent experiments (from two donors) were measured in 2 technical replicates ($n = 2$ biological replicates × 2 technical replicates). Representative results are shown in the figure. For Extended Data Fig. 10f–h, the experiments were repeated independently with $n = 4$ biological replicates with similar results and representative images are shown in the figure. For Extended Data Fig. 10m, 4 lesions were generated from iTS cell lines, collected and analysed ($n = 4$ biological replicates). For Supplementary Table 8, $n = 2$ biological replicates (from two donors) were used for data analysis presented in this supplementary table. GO enrichment *P* values are calculated on the basis of an accumulative hypergeometric distribution, and adjusted for multiple testing (*q*-values) using Benjamini-Hochberg adjustment. For Supplementary Table 9, $n = 2$ biological replicates (from two donors) were used. Motif enrichment *P* values are calculated on the basis of a cumulative binomial distribution. As described in ref. [61], the statistics assess the occurrence of motifs in target sequences versus a random background. From these motif occurrences, it then calculates the probability of detecting them in target sequences by chance. The software used for these calculations is described in 'ATAC-seq peak calling and exploratory analysis' and 'Integration of bulk ATAC-seq samples with bulk RNA-seq'.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

We developed an interactive online tool (http://hrpi.ddnetbio.com/) to facilitate exploration of the dataset, and for downloading all of the processed datasets. Raw and processed next-generation sequencing datasets have been deposited at the NCBI Gene Expression Omnibus (GEO) repository under accession numbers: GSE150311 (scRNA-seq experiments of intermediates during human primed and naive reprogramming); GSE150637 (scRNA-seq experiments of day 21 reprogramming intermediates cultured under fibroblast condition, naive pluripotent and trophoblast stem cell conditions); GSE147564 (snRNA-seq experiments of intermediates during human primed and naive reprogramming); GSE147641 (ATAC-seq experiments of intermediates during human primed and naive reprogramming); GSE150590 (ATAC-seq experiments of iTS cells); GSE149694 (bulk RNA-seq experiments of intermediates during human primed and naive reprogramming); and GSE150616 (bulk RNA-seq experiments of iTS cells and their derived placenta subtypes). Source data are provided with this paper.

Code availability

All data were analysed with standard programs and packages as detailed. Scripts can be found at https://github.com/SGDDNB/hrpi.

34. Liu, X., Nefzger, C. & Polo, J. Establishment and maintenance of human naive pluripotent stem cells by primed to naive conversion and reprogramming of fibroblasts. *Protoc. Exch.* https://doi.org/10.1038/protex.2017.099 (2017).
35. Guo, G. et al. Naive pluripotent stem cells derived directly from isolated cells of the human inner cell mass. *Stem Cell Reports* **6**, 437–446 (2016).
36. Pastor, W. A. et al. Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell* **18**, 323–329 (2016).
37. Larcombe, M. R. et al. Production of high-titer lentiviral particles for stable genetic modification of mammalian cells. *Methods Mol. Biol.* **1940**, 47–61 (2019).
38. Qiu, P. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
39. Nefzger, C. M. et al. A versatile strategy for isolating a highly enriched population of intestinal stem cells. *Stem Cell Reports* **6**, 321–329 (2016).
40. Meistermann, D. et al. Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm. Preprint at https://www.biorxiv.org/content/10.1101/604751v1 (2019).
41. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
42. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genom. Biol.* **21**, 57 (2020).
43. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
44. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
46. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
48. Grubman, A., Choo, X. Y., Chew, G., Ouyang, J. F. & Sun, G. Mouse and human microglial phenotypes in Alzheimer's disease are controlled by amyloid plaque phagocytosis through Hif1a. Preprint at https://www.biorxiv.org/content/10.1101/639054v1 (2019).
49. Girardot, C., Scholtalbers, J., Sauer, S., Su, S.-Y. & Furlong, E. E. M. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17**, 419 (2016).
50. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
53. Lindenbaum, P. JVarkit: java-based utilities for Bioinformatics, https://figshare.com/articles/JVarkit_java_based_utilities_for_Bioinformatics/1425030 (2015).
54. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
55. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Alasoo, K. et al. Transcriptional profiling of macrophages derived from monocytes and iPS cells identifies a conserved response to LPS and novel alternative transcription. *Sci. Rep.* **5**, 12524 (2015).
58. Wu, J. et al. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**, 256–260 (2018).
59. Pastor, W. A. et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. *Nat. Cell Biol.* **20**, 553–564 (2018).
60. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
61. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
62. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

Acknowledgements We thank staff at Monash Flowcore Facility for providing high-quality cell sorting services and technical input; S. Wang, T. Wilson and the University of Melbourne Centre for Cancer Research (UMCCR) core for assistance with next-generation library preparation and Illumina sequencing; J. Hatwell-Humble for assistance with the mouse work; and A. Purcell for providing the HLA antibodies. We acknowledge the use of the services and facilities of Micromon, Monash Micro Imaging and Monash Histology Platforms at Monash University. This work was supported by National Health and Medical Research Council (NHMRC) project grants APP1104560 to J. M. Polo and A. L. Laslett, APP1069830 to R.L., and a Monash University strategic grant awarded to C.M.N. X.L. was supported by the Monash International Postgraduate Research Scholarship, a Monash Graduate Scholarship and the Carmela and Carmelo Ridolfo Prize in Stem Cell Research. A.S.K. was supported by an NHMRC Early Career Fellowship APP1092280. J. M. Polo and R.L. were supported by Silvia and Charles Viertel Senior Medical Research Fellowships. J. M. Polo was also supported by an ARC Future Fellowship FT180100674. R.L was supported by a Howard Hughes Medical Institute International Research Scholarship. O.J.L.R. and J.F.O. were supported by a Singapore National Research Foundation Competitive Research Programme (NRF-CRP20-2017-0002). The Australian Regenerative Medicine Institute is supported by grants from the State Government of Victoria and the Australian Government.

Author contributions J. M. Polo conceptualised the study. O.J.L.R. and J. M. Polo supervised the study. X.L., J.F.O., F.J.R., O.J.L.R. and J. M. Polo designed the experiments and analysis. O.J.L.R devised the single-cell analysis pipeline and data integration. X.L. performed reprogramming experiments, collection and isolation of single cells, intermediates and functional validation of iTS cell experiments with support from C.M.N., J.P.T., K.C.D., D.S.V., Y.B.Y.S., J.C., J. M. Paynter, J.F., Z.H., P.T., P.P.D. and S.K.N.; X.L. and C.M.N. performed single-cell RNA-seq, FACS experiments, SPADE analysis and the molecular experiments of the cells with support from A.S.K. and J.C.; L.G.M. helped with snRNA-seq experiments with support from A. L. Leichter; M.R.L. helped with RT-PCR experiments. D.P. helped with sequencing of day-21 reprogramming intermediates scRNA-seq libraries. X.L. generated the lentiviral particles with the assistance of J.P.T., G.S.; J.P. helped with ATAC-seq experiments. H.S.C., C.M.O'B. and A. L. Laslett. provided reagents and technical assistance. H.N. and D.R.P helped with bulk RNA-seq analysis. J.F.O. performed the snRNA-seq, scRNA-seq and bulk RNA-seq analyses for the

human reprogramming intermediates and iTS cell experiments as well as the integration across the various datasets with support from F.J.R., J.S., J. M. Polo and O.J.L.R.; F.J.R. performed ATAC-seq analysis with support from V.T., X.Y.C, J.S., S.B., O.J.L.R., W.A.P., D.C., A.T.C., J. M. Polo and R.L.; J.F.O. and O.J.L.R. developed the interface for the interactive online tool. X.L., J.F.O., F.J.R., O.J.L.R. and J. M. Polo wrote the manuscript with input from K.C.D., A.G., A.T.C., L.D., C.M.N. and R.L. All authors approved of and contributed to, the final version of the manuscript.

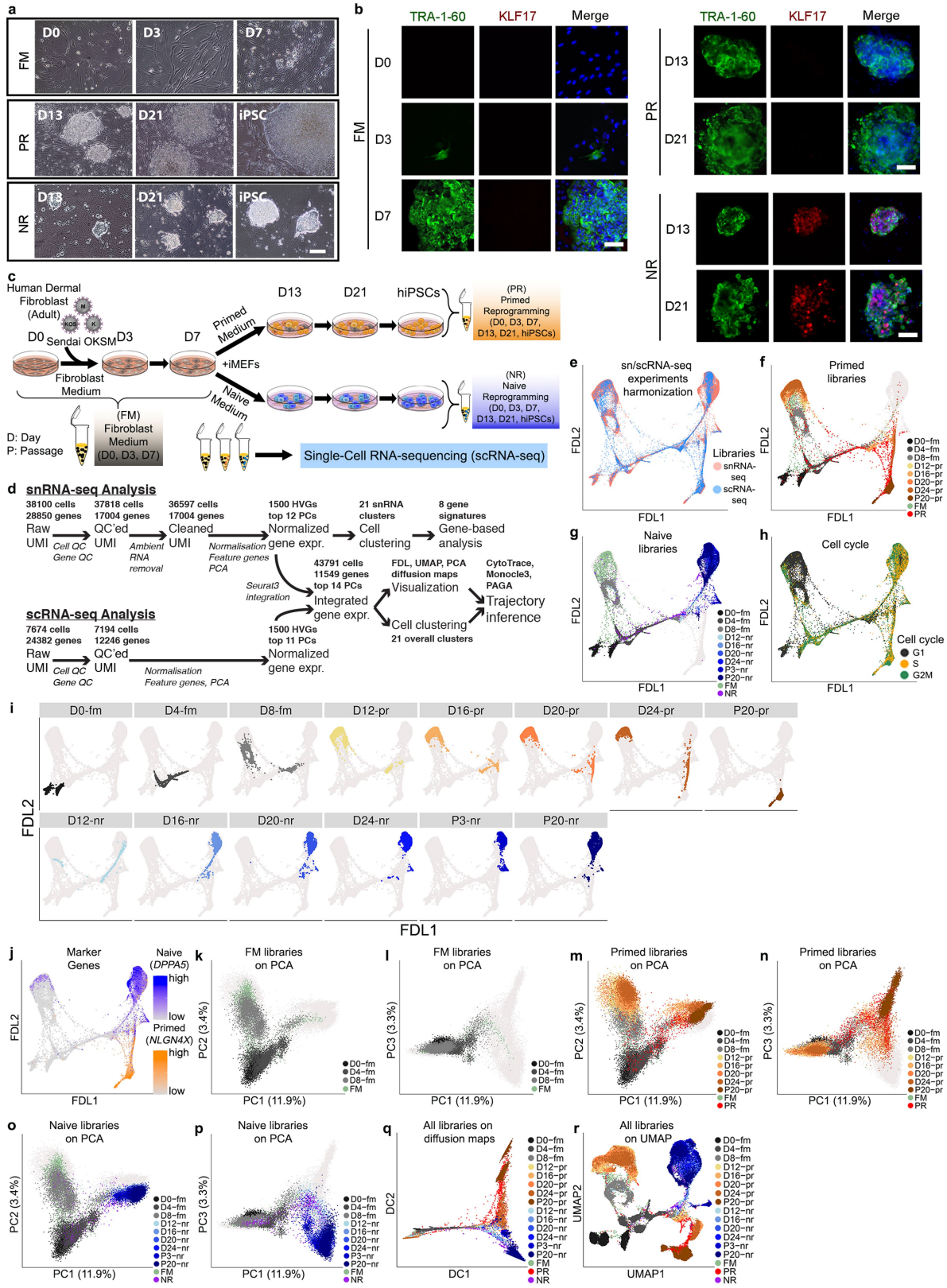
Competing interests O.J.L.R. and J. M. Polo. are co-inventors on a patent (WO/2017/106932) and are co-founders and shareholders of Mogrify Ltd., a cell therapy company. X.L., J.F.O., K.C.D., L.D., O.J.L.R. and J. M. Polo are co-inventors on a provisional patent application (application number: 2019904283) filed by Monash University, National University of Singapore and Université de Nantes related to work on derivation of iTS cells. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2734-6.

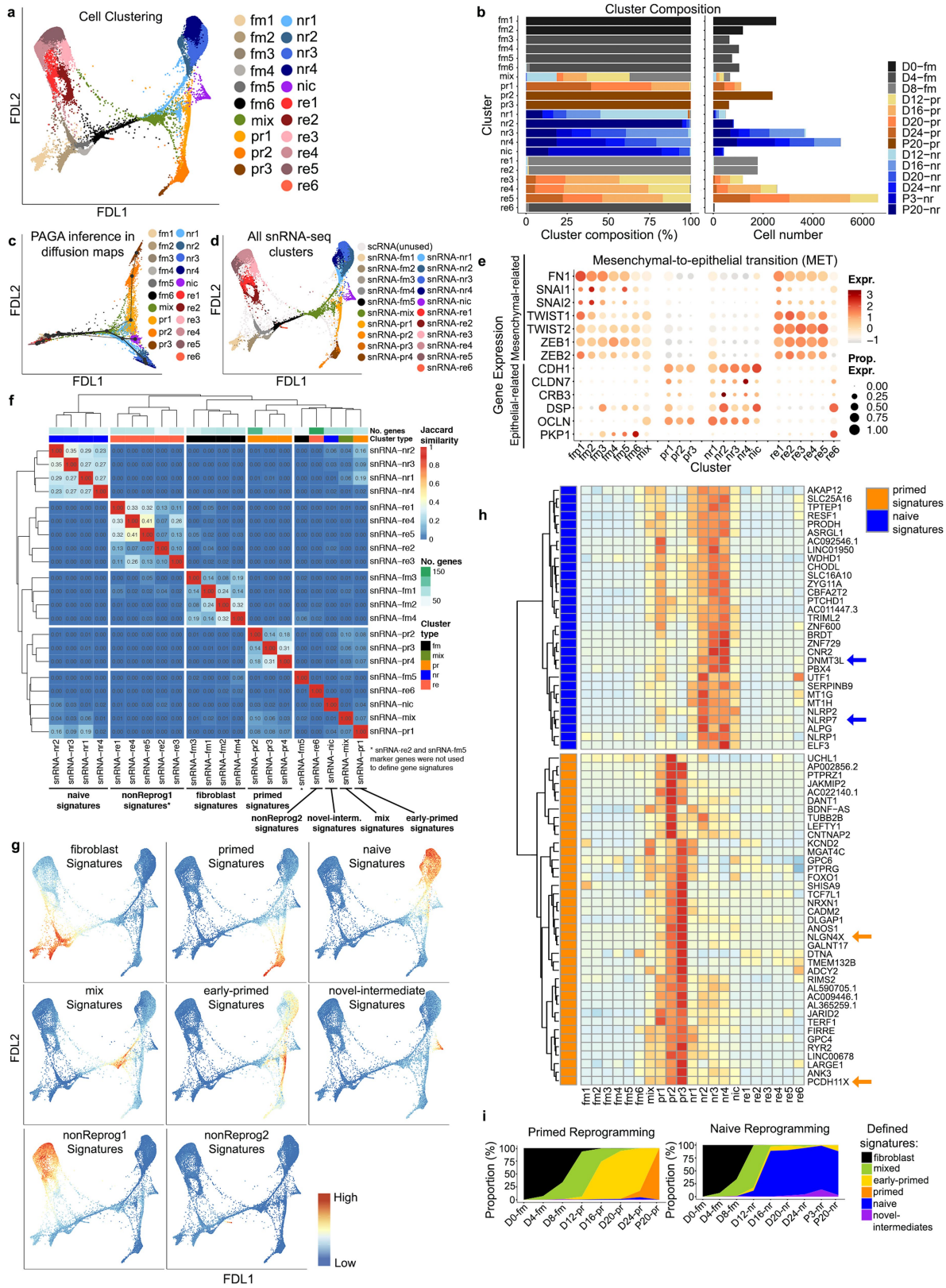
Correspondence and requests for materials should be addressed to O.J.L.R. or J.M.Polo.

Peer review information *Nature* thanks Ashley Moffett, Samantha A. Morris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



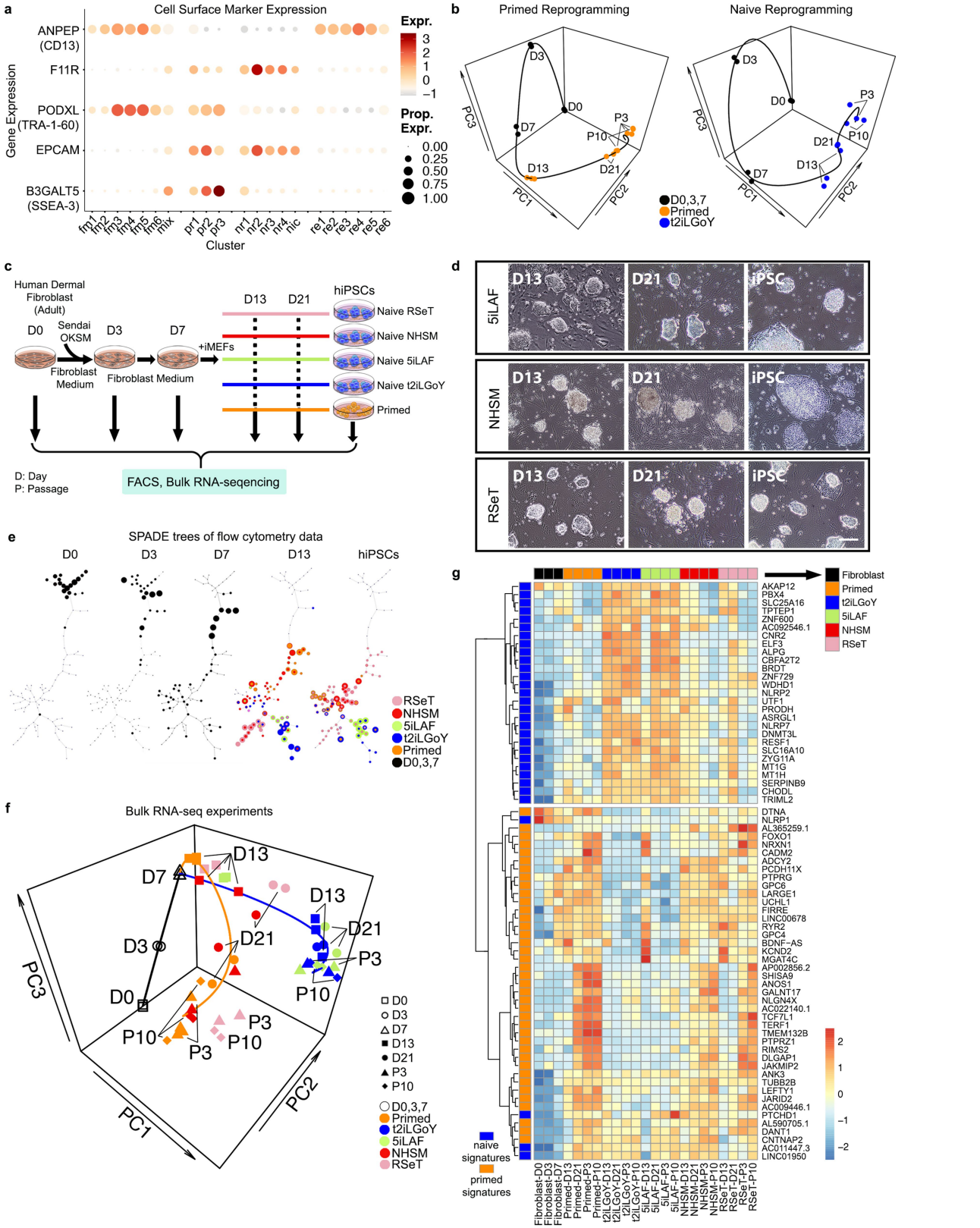
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Experimental designs, analysis pipelines for snRNA-seq and scRNA-seq. a, Morphological changes of cells undergoing reprogramming in fibroblast medium (FM); primed reprogramming (PR); naive reprogramming (NR). FM: D0, 3 and 7; PR: D13, D21 and iPS cells (iPSCs); and NR: D13, D21 and iPS cells, $n > 10$. Scale bar, 500 µm. **b**, Immunostaining at early stages (FM: D0, 3 and 7), during PR (D13 and D21) and NR (D13 and D21) with TRA-1-60 for primed colonies, KLF17 for naive colonies and DAPI for nuclei staining, $n = 2$. Scale bar, 50 µm. **c**, Experimental design for scRNA-seq libraries. iMEF, irradiated mouse embryonic fibroblasts. **d**, snRNA-seq and scRNA-seq data analysis strategy (Methods). **e**, Representation of integrated snRNA-seq and scRNA-seq experiments (43,791 cells) on FDL. **f**, **g**, Primed and naive libraries on FDL. **h**, FDL showing cells in predicted stages of the cell cycle. **i**, Reprogramming trajectories on FDL highlighting cells within each time point. **j**, Expression of genes associated with primed pluripotency (*NLGN4X*) and naive pluripotency (*DPPA5*) on FDL. **k**–**r**, PCA (**k**–**p**), diffusion maps (**q**) and UMAP (**r**) of snRNA-seq and scRNA-seq data. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.

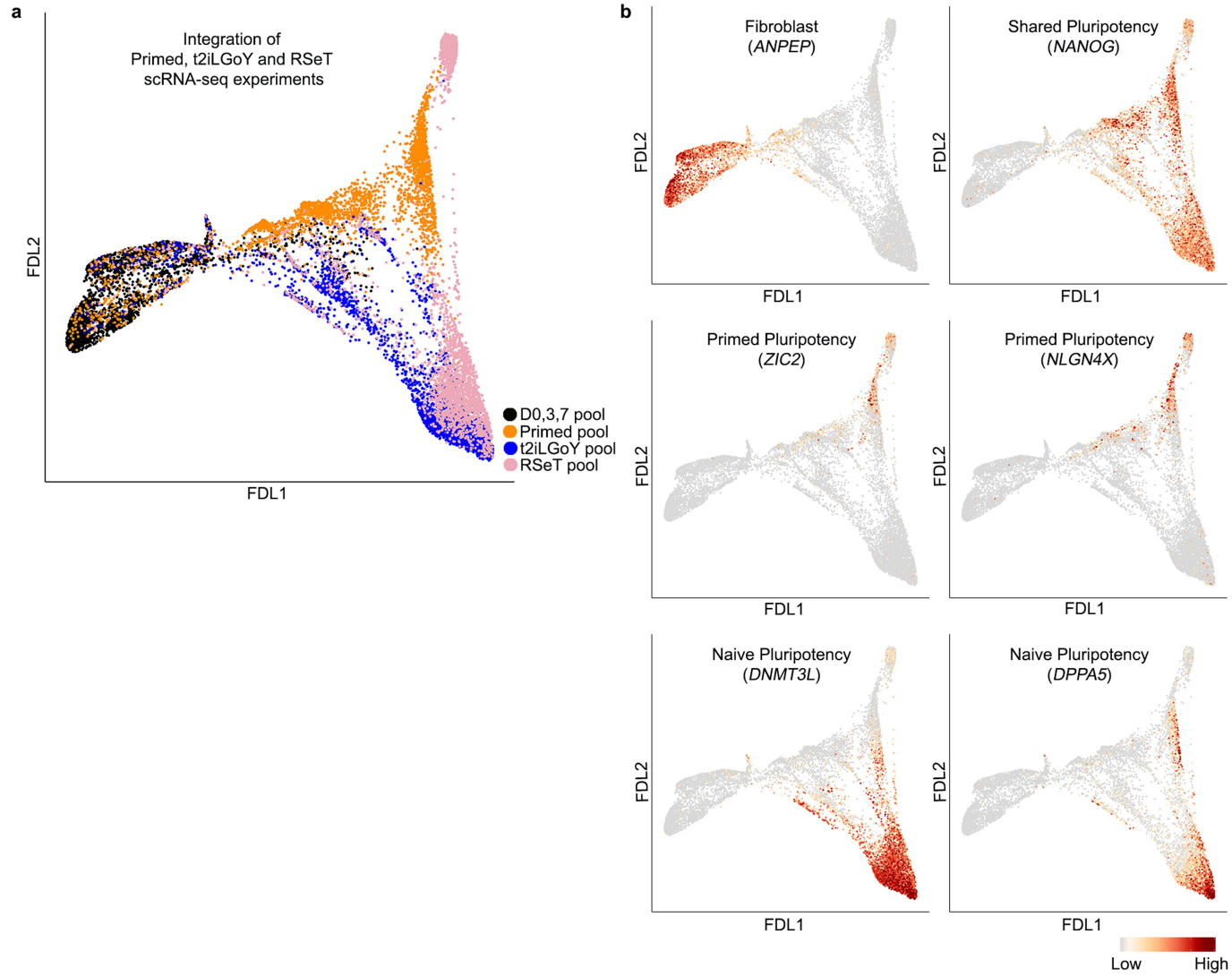


Extended Data Fig. 2 | See next page for caption.

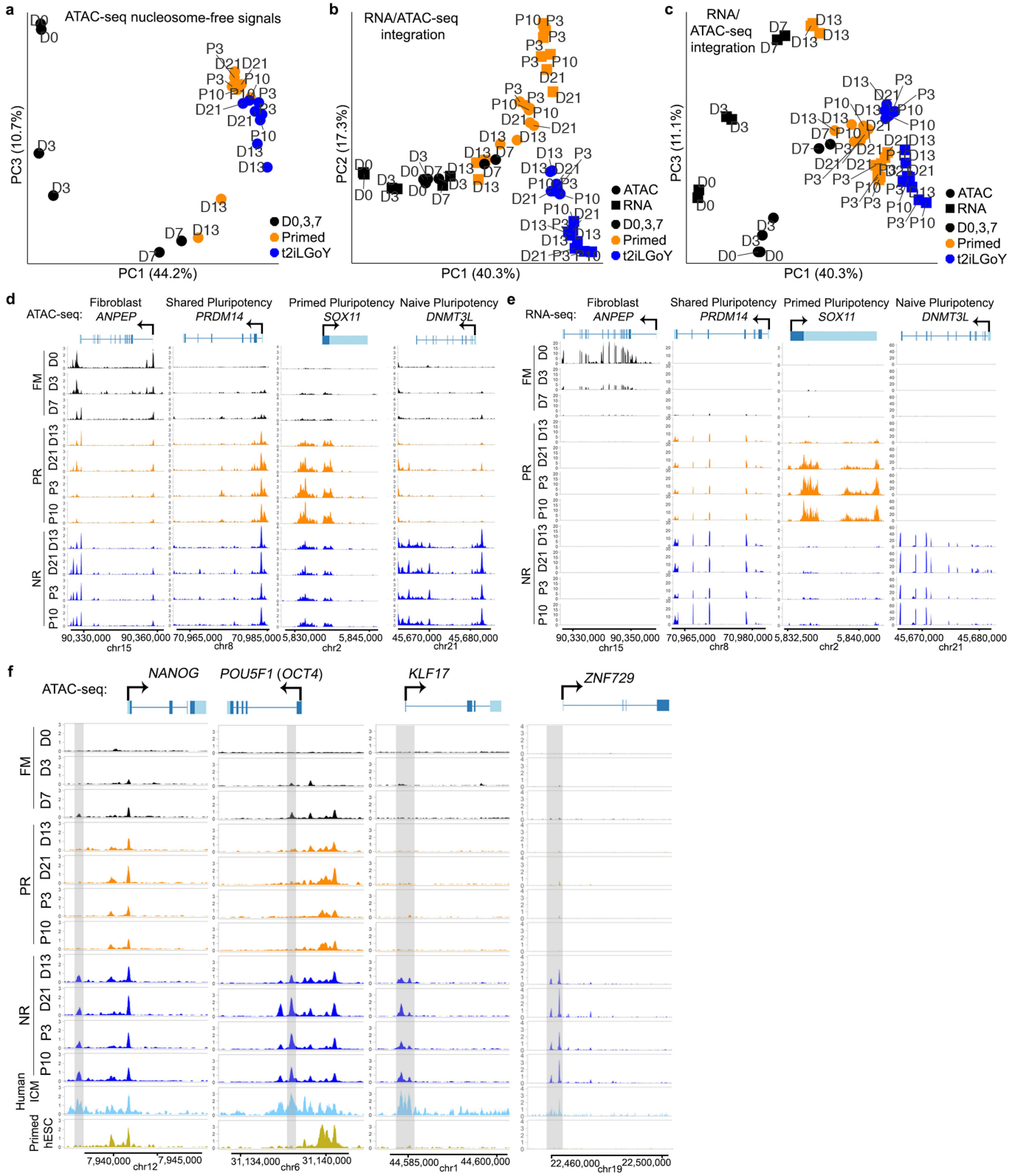
Extended Data Fig. 2 | Resolving the molecular hallmarks of primed and naive reprogramming trajectories. a, Unsupervised clustering projected onto the FDL shown in Fig. 1 (43,791 cells). fm1–fm6, fibroblast and early reprogramming intermediate cell clusters; mix, shared cell cluster; pr1–pr3, primed reprogramming cell clusters; nr1–nr4, naive reprogramming cell clusters; nic, novel intermediate cell cluster; re1–re6, refractory cell clusters. **b**, snRNA-seq time point and library contribution (composition and cell number) towards each cell cluster. **c**, PAGA trajectory inference on diffusion maps. **d**, snRNA-seq clusters, used to define gene signatures, on FDL. **e**, Dot plot showing the expression of mesenchymal and epithelial (MET)-associated genes across cell clusters. **f**, Jaccard similarity of snRNA-seq cluster-specific genes. Cluster-specific genes are then grouped to define the eight gene signatures, highlighted at the bottom. **g**, Defined gene signatures on FDL. **h**, Gene-expression heat map of the primed or naive pluripotency signatures across the cell clusters (coloured arrows indicate known marker genes). **i**, Area plots showing the transition and activation of the defined signatures during primed and naive reprogramming over time. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



Extended Data Fig. 3 | Isolation and characterization of intermediates during reprogramming into several naive human induced pluripotent states. a, Identification of cell-surface markers for the isolation of primed and naive reprogramming intermediates. **b**, PCA of bulk RNA-seq data of isolated intermediates during primed and naive reprogramming, $n \geq 2$. **c**, Experimental designs for the generation, isolation and profiling of intermediates during reprogramming into several naive human induced pluripotent states. **d**, Morphological changes during reprogramming under naive 5iLAF, NHSM and RSeT culture conditions (Methods), $n = 4$. Scale bar, 500 µm. **e**, Visualization of flow cytometry profiles (SPADE tree) of intermediates during reprogramming, $n = 2$. **f**, PCA of RNA-seq of primed and several types of naive reprogramming intermediates (Methods), $n \geq 2$. **g**, Heat map showing gene expression profiles of primed and naive pluripotency signatures genes (defined in snRNA-seq and scRNA-seq analysis) across reprogramming intermediates and iPS cells derived under all different culture conditions, $n \geq 2$. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.

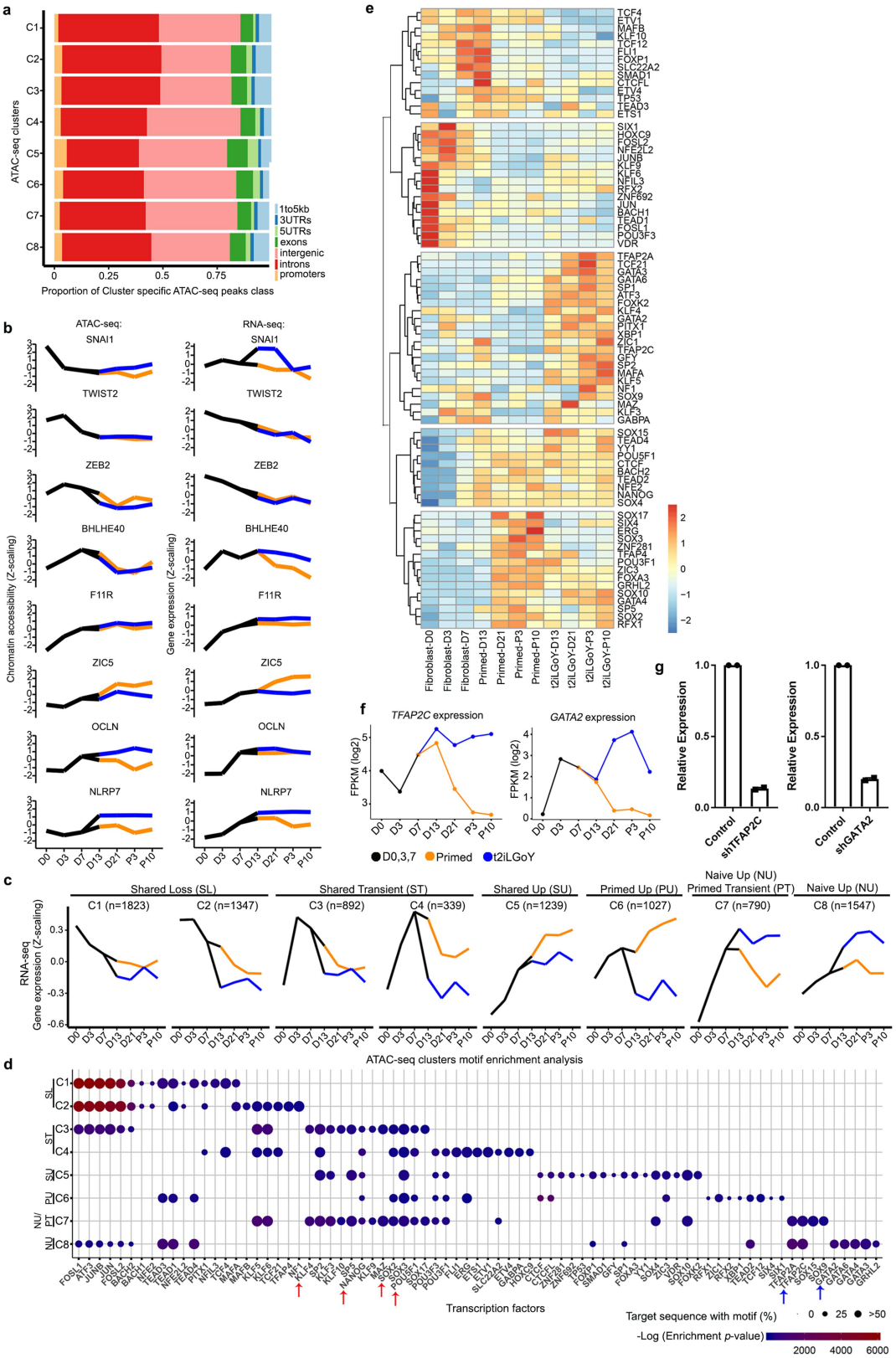


Extended Data Fig. 4 | Single-cell profiling of the reprogramming pathway into naive RSeT state. a, FDL of fibroblast, primed, naive t2iLGoY and RSeT scRNA-seq libraries (9,852 cells) (Methods). **b**, Expression profile of genes associated with human fibroblasts (*ANPEP*), shared pluripotency (*NANOG*), primed pluripotency (*ZIC2* and *NLGN4X*) and naive pluripotency (*DNMT3L* and *DPPA5*) on FDL. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



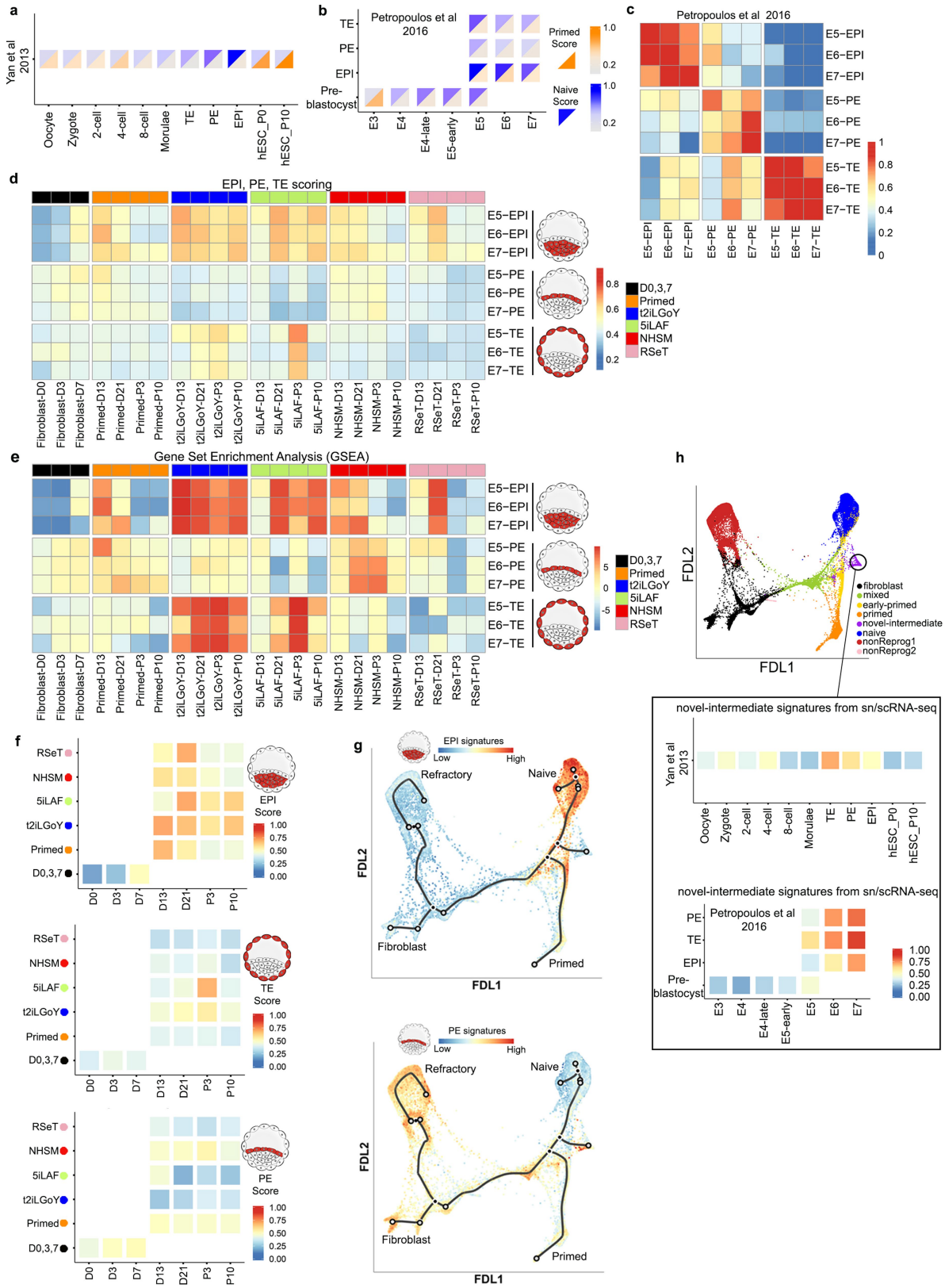
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Dynamics of chromatin state transitions during reprogramming into primed and naive human induced pluripotency. **a**, PCA plot of ATAC-seq nucleosome-free signals, PC1 versus PC3 related to Fig. 2c. ATAC-seq was performed using isolated reprogramming intermediates and iPS cells from FM (D0, D3 and D7), PR (D13, D21, P3 and P10), NR (D13, D21, P3 and P10), $n = 2$. FM, fibroblasts medium (black); PR, primed reprogramming (orange); NR, naive reprogramming (blue). **b**, **c**, PCA plot of the integration of RNA-seq and ATAC-seq experiments ($n \geq 2$). **d**, **e**, ATAC-seq and corresponding RNA-seq tracks of primed and naive reprogramming intermediates for fibroblast marker, *ANPEP*; shared pluripotency marker, *PRDM14*; primed-specific pluripotency marker *SOX11*; naive-specific pluripotency marker *DNMT3L*. Model of each gene is shown: coding sequences, light blue boxes, and exons, dark blue boxes; introns are shown as light blue connecting lines. **f**, Naive-reprogramming-specific ATAC-seq signals (in light grey) around core pluripotency factors *NANOG* and *POU5F1* (also known as *OCT4*), naive-reprogramming-specific *KLF17* and *ZNF729* in primed and naive reprogramming intermediates and iPS cells compared to human inner cell mass and primed ES cells (ESCs) ATAC-seq data[58]. For more details on sample numbers, see 'Statistics and reproducibility' in Methods.



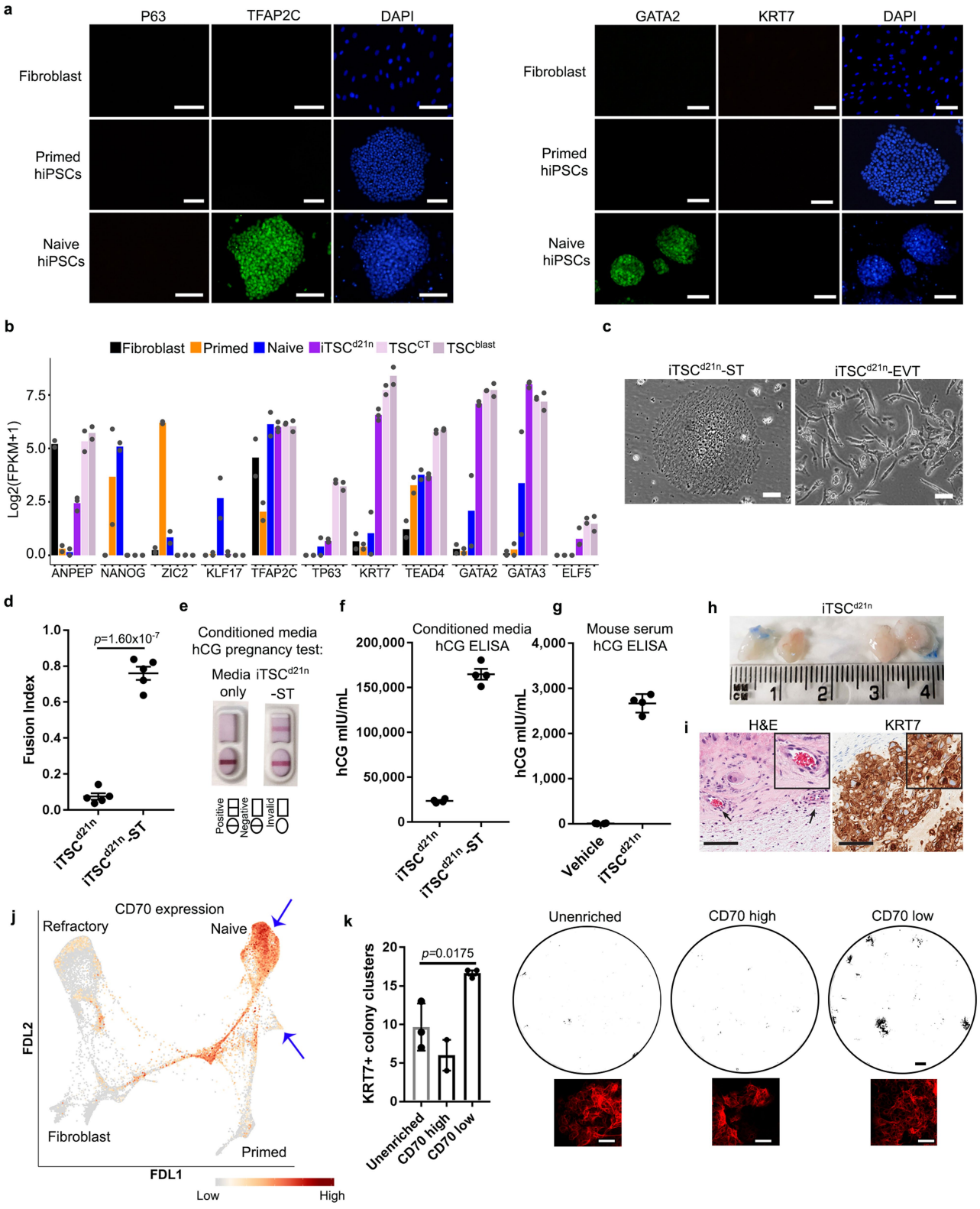
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Features of accessible chromatin landscape during reprogramming into primed and naive human induced pluripotency. **a**, Proportion of genomic regions in each of the ATAC-seq clusters. **b**, Averaged chromatin accessibility (*z*-scaled, $n = 2$) and gene expression (*z*-scaled, $n \geq 2$) of one representative gene from each of the ATAC-seq peak clusters. **c**, Standardized gene expression (averaged *z*-scaling) of genes associated with ATAC-seq cluster peaks (Methods). **d**, Transcription factor motif enrichment analysis of the ATAC-seq peak clusters. Motif enrichment (−log(*P* value)) heat map by colour and the size the percentage of sequences in the cluster featuring the motif. Red arrow points to *OCT4*, *SOX2*, *NANOG* and *KLF4* motifs in transient ATAC-seq cluster (C3), Blue arrow = enrichment of TE-associated transcription factors *TFAP2C* and *GATA2* (C7 and C8) are indicated by blue arrows. **e**, Gene-expression heat map transcription factors identified in the motif enrichment analysis in **d**. **f**, *TFAP2C* and *GATA2* gene expression during primed and naive reprogramming. **g**, Reverse transcription qPCR analysis of sh*TFAP2C* and sh*GATA2* compared to scrambled controls, $n = 2$. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



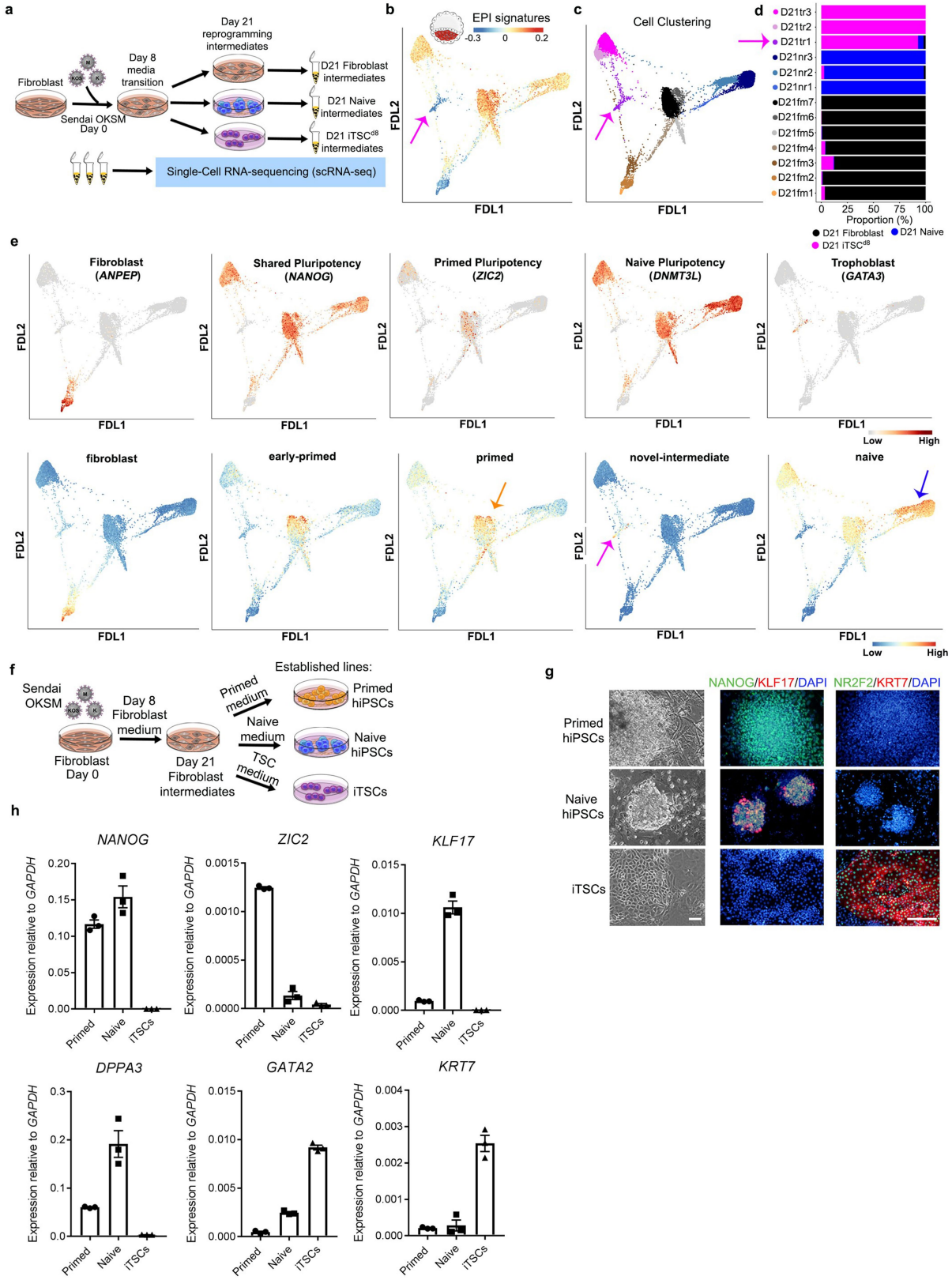
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Uncovering the transcriptional programmes of human fibroblast reprogramming into naive induced pluripotency. **a**, **b**, Primed and naive scores, using gene signatures defined in this study (Fig. 1g), on human preimplantation embryos at indicated embryonic stages based on scRNA-seq experiments from published studies[24,25]. **c**, EPI, PE and TE signatures score at indicated embryonic stages[25]. **d**, EPI, PE and TE gene signatures[25] from embryonic (E) day 5, 6 and 7 on intermediates and iPS cells reprogrammed under primed and different naive culture conditions (Methods). **e**, Gene set enrichment analysis (GSEA) (Methods) of the EPI, PE and TE gene signatures in reprogramming intermediates and iPS cells reprogrammed under primed and several naive culture conditions. **f**, EPI, PE and TE gene signatures scores in reprogramming intermediates and iPS cells reprogrammed under primed and several naive culture conditions. We used a combined gene signature across E5 to E7 for each lineage (Methods). **g**, EPI and PE signatures on FDL with single-cell trajectories constructed using Monocle3 (43,791 cells), related to Fig. 3a. **h**, Scoring of novel-intermediate signatures defined in this study (Extended Data Fig. 2f, g) on human preimplantation embryos of different lineages at indicated embryonic stages based on scRNA-seq experiments from published studies[24,25]. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



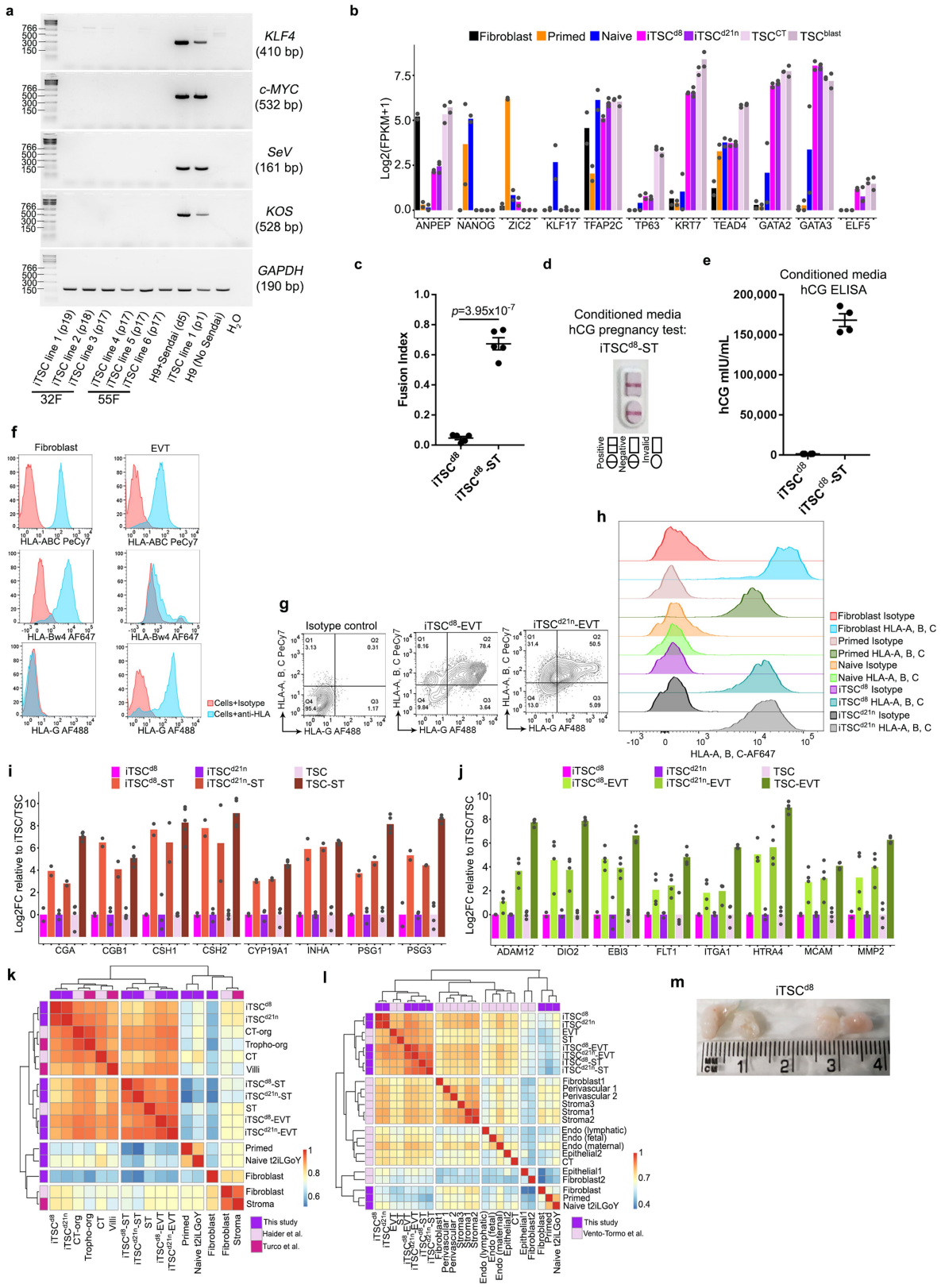
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Characterization of iTSC^{d21n}. a, Immunostaining of fibroblast, primed, naive t2iLGoY iPS cells with P63, TFAP2C, GATA2 and KRT7, $n = 2$. Scale bar, 100 µm. **b**, Gene expression of trophoblast genes in fibroblasts, primed, naive t2iLGoY iPS cells, iTSC^{d21n} and TS cells (TSCs) derived from a human blastocyst (TSC^{blast})[7] and first-trimester placental trophoblast (TSC^{CT})[7], mean of replicates, $n = 2$. **c**, Phase-contrast image of ST and EVT cells differentiated from iTSC^{d21n}, $n = 4$. Scale bar, 100 µm. **d**, Fusion index of iTSC^{d21n} ST and iTSC^{d21n}, $n = 5$, data are represented as mean ± s.e.m. *P* values by two-tailed unpaired Student's *t*-test. **e**, Representative results for over-the-counter hCG pregnancy test for medium of ST cells differentiated from iTSC^{d21n} and control medium, $n = 6$. **f**, hCG levels in iTSC^{d21n}- and iTSC^{d21n}-ST conditioned medium, detected by ELISA, $n = 4$. **g**, hCG level in mouse blood serum detected by ELISA, $n = 4$. **h**, Lesions collected from subcutaneously engrafted iTSC^{d21n} in NOD-SCID mice, $n = 4$. **i**, Haematoxylin and eosin, and immunohistochemical staining of KRT7 in the lesions from **h**. No evident lesions were observed in vehicle controls, $n = 4$. Scale bar, 200 µm. **j**, Distinct level of CD70 expression in naive and TE populations (indicated by blue arrows) on FDL projection of snRNA-seq and scRNA-seq datasets. **k**, Quantification of KRT7^{+} colony clusters after 9 d of transitioning into TS cell medium of unenriched, CD70^{high} and CD70^{low} populations, $n = 2$ or 3 independent experiments, data are mean ± s.e.m. *P* values by two-tailed unpaired Student's *t*-test. Representative images of whole-well scans (top panels) (scale bar, 1 mm) and KRT7 immunostaining (bottom panels) (scale bar, 100 µm). For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Cellular heterogeneity of fibroblast and iTSC^{d8} reprogramming intermediates revealed by scRNA-seq. a, Experimental designs and preparation of scRNA-seq libraries of day-21 fibroblast, naive and TSC^{d8} reprogramming intermediates. **b**, Strength of EPI signatures on FDL (10,518 cells). The cell population not enriched for EPI signatures but enriched for TE signatures is indicated by a purple arrow, related to Fig. 4b. **c**, Representation of 13 cell clusters from unsupervised clustering projected onto the FDL, fibroblast medium cell clusters: D21fm1–D21fm7; naive reprogramming cell clusters: D21nr1–D21nr3; trophoblast reprogramming cell clusters: D21tr1–D21tr3. **d**, Contribution of each scRNA-seq library (%) to the composition of cell clusters. D21tr1 cluster is indicated by a purple arrow. **e**, Expression of genes associated with human fibroblasts (*ANPEP*), shared pluripotency (*NANOG*), primed pluripotency (*ZIC2*), naive pluripotency (*DNMT3L*) and trophoblast (*GATA3*) on FDL projection of day-21 fibroblast, naive and TSC^{d8} reprogramming intermediate scRNA-seq libraries (top panels). Defined fibroblast, early-primed, primed, novel-intermediate and naive signatures (Extended Data Fig. 2f) on the FDL projection (bottom panels). **f**, Experimental designs to validate the potential of day-21 fibroblast reprogramming intermediates for the derivation of primed, naive iPS cells and iTS cells. **g**, Phase-contrast images of primed, naive iPS cells and iTS cells generated from day 21 fibroblast reprogramming intermediates, $n = 2$. Scale bar, 50 µm. Immunostaining of primed, naive iPS cells and iTS cells with NANOG, KLF17, NR2F2, KRT7 and DAPI for nuclei staining, $n = 2$. Scale bar, 200 µm. **h**, Reverse-transcription qPCR analysis of *NANOG*, *ZIC2*, *KLF17*, *DPPA3*, *GATA2* and *KRT7* expression in primed, naive iPS cells and iTS cells generated from day-21 fibroblast reprogramming intermediates, $n = 3$. Data are mean ± s.e.m. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Characterization of iTSC^{d8}. **a**, Sendai viral transgenes in iTS cell lines with positive and negative controls, $n = 6$. **b**, Gene expression of trophoblast genes in fibroblasts, primed iPS cells, naive t2iLGoY iPS cells, iTSC^{d8} and iTSC^{d21n} compared to TSCs derived from a human blastocyst (TSC^{blast}) and first-trimester placental trophoblast (TSC^{CT})[7], data are presented as mean ($n = 2$). **c**, Cell fusion index of iTSC^{d8} ST and iTSC^{d8}. $n = 5$, data are mean ± s.e.m. *P* values by two-tailed unpaired Student's *t*-test. **d**, Representative results for hCG pregnancy test obtained from medium of ST cells differentiated from iTSC^{d8}, $n = 6$. **e**, hCG levels of iTSC^{d8}- and iTSC^{d8}-ST-conditioned medium detected by ELISA, $n = 4$. **f**, Representative flow cytometry analysis of pan HLA-A, B, C class I marker (W6/32), HLA-Bw4 and HLA-G in fibroblasts and EVTs, $n = 4$. **g**, Representative flow cytometry analysis of pan HLA class I marker (W6/32) and HLA-G in iTSC^{d8} EVT and iTSC^{d21n} EVT. **h**, Representative flow cytometry analysis of pan HLA class I marker (W6/32) in fibroblasts, primed iPS cells, naive t2iLGoY iPS cells, iTSC^{d8} and iTSC^{d21n}, $n = 4$. **i**, **j**, Expression of ST genes in iTSC^{d8}- and iTSC^{d21n}-derived ST cells (**i**) and expression of EVT genes in iTSC^{d8} and iTSC^{d21n}-derived EVT cells (**j**). **k**, **l**, Spearman correlation of the transcriptomes of fibroblast, primed and naive t2iLGoY iPS cells, iTSC^{d8} and iTSC^{d21n}, iTSC^{d8} ST and iTSC^{d21n} ST, iTSC^{d8} EVT and iTSC^{d21n} EVT generated in this study with trophoblast organoids samples from refs. [28,29] (**k**) and single-cell fetal–maternal interface samples from ref. [27] (**l**), $n \geq 2$, replicates are averaged before performing correlation. **m**, Lesions collected from subcutaneously engrafted iTSC^{d8} in NOD-SCID mice, $n = 4$. For more details on sample numbers and statistics, see 'Statistics and reproducibility' in Methods.

nature research

Corresponding author(s): Owen J. L. Rackham
Jose M. Polo

Last updated by author(s): May 28, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Motic Image Plus 2.0, LightCycler 480 software (Roche, version 1.5.1.62), BD FACSDivia software version 8.0.1 for LSRIIa/b, BD FACS TM software, version 1.2.0.142 for influx sorters, DP controller version 2.1.1.183 and DP manager version 2.1.1.163 for fluorescence microscope imaging, Leica application suite X version 3.7.1.21655 for DMi8.
Data analysis	Data was analyzed using: GraphPad Prism (v7), ImageJ (v1.8.0_112), FlowJo (v10), Cytobank 7.2.0, Cellranger (v2.1.0, v3.0.2, v3.1.0), base R(3.5.1), data.table(v1.12.2), Matrix(v1.2-17), ggplot2(v3.2.1), pheatmap(v1.0.12), shiny(v1.2.0), celda(v1.1.6), Seurat(v3.1.1), limma(v3.38.3), uwot(v0.1.4), irlba(v2.3.3), reticulate(v1.12), monocle3(v0.1.3), SingleCellExperiment(v1.4.1), scanpy(v1.4.4.post1), edgeR(v3.24.3), CytoTRACE(v0.1.0), metascape (v3.0), cutadapt (v1.8), STAR (v 2.4.2a), Trimmomatic (v 0.36), bowtie2 (v2.3.2), samjs (version 0b19b0e4e79be99da6616394cc096a38724d7d48), picard MarkDuplicates (v2.18.0-SNAPSHO), MACS2 (v2.1.1.20160309), bedtools2 (v2.25.0), featureCounts (v1.5.2), mfuzz (v2.38.0), homer (v4.10.3), wiggleplotr (1.2.0), annotatr (v1.4.1), ComplexHeatmap (1.17.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All processed data sets are available at http://hrpi.ddnetbio.com. Raw and processed next generation sequencing datasets were deposited at the Gene Expression Omnibus (GEO) repository with the following accession numbers: GSE150311: scRNA-seq experiments of intermediates during human primed and naive reprogramming; GSE150637: scRNA-seq experiments of of day 21 reprogramming intermediates cultured under fibroblast condition, naive pluripotent and

trophoblast stem cell conditions; GSE147564: snRNA-seq experiments of intermediates during human primed and naive reprogramming; GSE147641: ATAC-seq experiments of intermediates during human primed and naive reprogramming; GSE150590: ATAC-seq experiments of induced trophoblast stem cells; GSE149694: bulk RNA-seq experiments of intermediates during human primed and naïve reprogramming; GSE150616: bulk RNA-seq experiments of induced trophoblast stem cells and their derived placenta subtypes. All source data related to this study are available in the source data table related to each figure. There is no restrictions on data availability.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We performed time-coure profiling of primed and naive reprogramming using snRNA-seq libraries (from 38F donor) and scRNA-seq libraries (from 32F donor) and similar results/findings were observed. Also, we generated time-course bulk RNA-seq and ATAC-seq from two different donors with similar results. Furthermore, we generated iTSCs using two different donor fibroblasts in multiple runs of reprogramming experiments (at least 4 independent replicates). These suggest that our sample size is sufficient.
Data exclusions	No data were excluded.
Replication	Each experiment was reproduced at least in 4 biological replicates (from two donors) if not otherwise stated. Please refer to figure legends and methods for details. All replications were successful.
Randomization	No randomization methods were utilized only for Animals used for experiments that were randomly allocated.
Blinding	The investigators were not blinded during data collection and analysis. We did not considered blinding required in these type of experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Details of all antibodies used in this study were provided in Supplementary Table 15 with catalog number, lot number and commercial sources supplied.

For flow cytometry:
Antibody, Company, Catalogue Number, Lot Number, Dilution:
PE-Cy7 mouse anti-human CD13 BD Biosciences Cat# 561599 Lot 8248674 1:400
APC-Cy7 CD13 Biolegend Cat# 301710 N/A 1:500
BUV395 mouse anti-human TRA-1-60 BD Biosciences Cat# 563878 Lot 9232453 1:100
Anti-TRA-1-85 (CD147)-VioBright FITC Miltenyi Biotec Cat#130-107-106 Lot 5190821362 1:20
PE-SSEA3 BD Biosciences Cat#560237 Lot 8248664 1:10
F11R-APC CSIRO CSTEM27APC, O'Brien et al., 2017 N/A 1:200
PE-Vio770 mouse anti-SSEA-4 Miltenyi Biotec Cat# 130-105-051 Lot 5171206186 1:20
BV 605 goat anti-mouse IgG Biolegend Cat# 405327 Lot B212043 1:100
PE mouse anti-Rat IgM eBiosciences Cat# 12-4342-82 Lot 4274948 1:200
AF647 goat anti-mouse IgG secondary ThermoFisher Cat#A21235 Lot 1596074 1:2000
Mouse IgG1 κ Isotype Control BD Biosciences Cat# 550878 Lot 8151786 1:100

Mouse IgG2a, κ Isotype Control BD Biosciences Cat# 554126 Lot 2153856 1:100
BV711 mouse anti-human CD24 BD Biosciences Cat# 563401 Lot 7125705 1:50
BV 421 mouse anti-human CD326 (EpCAM) Biolegend Cat# 324220 Lot B287899 1:100
Mouse anti-human F11R IgG2a CSIRO CSTEM27, O'Brien et al., 2017 N/A 1:200
Mouse anti-human NLGN4X IgG2a CSIRO CSTEM30, O'Brien et al., 2017 N/A 1:128
Mouse anti-human CD249 (APA) IgG1 BD Biosciences Cat#564532 Lot 8199856 1:100
APC anti-human/mouse CD49f (ITGA6) Miltenyi Biotec Cat#130-097-250 Lot 5190515229 1:20
BUV737 mouse anti-human CD70 BD Biosciences Cat#612856 Lot 9186789 1:100
PE-Cy7 mouse anti-human HLA-A,B,C Biolegend Cat#311430 Lot B290872 1:200
Mouse anti-human HLA-Bw4 IgG1 Purcell Lab N/A N/A 1:1
Mouse anti-HLA-G IgG1 [MEM-G/9] Abcam Cat#ab7758 Lot GR3215298-13 1:50

For Immunostaining:
Antibody, Company, Catalogue Number, Lot Number, Dilution:
Rabbit anti-KLF17 polyclonal Sigma-Aldrich Cat# HPA024629 Lot A105885 1:500
Mouse anti-TRA-1-60 IgM BD Biosciences Cat# 560071 Lot 5205659 1:300
Goat anti-rat IgM AF488 secondary ThermoFisher Cat#A-21212 Lot 1206007 1:400
Goat anti-rabbit IgG AF555 secondary ThermoFisher Cat#A-21428 Lot 1786491 1:400
Goat anti-mouse IgM AF488 secondary ThermoFisher Cat#A-21042 Lot 1896382 1:400
Goat anti-mouse IgG-AF488 secondary ThermoFisher Cat#A-11029 Lot 2066710 1:400
Goat anti-mouse IgG1-AF488 secondary ThermoFisher Cat#A-21121 Lot 1964382 1:400
Goat anti-mouse IgG2a-AF555 secondary ThermoFisher Cat#A-21137 Lot 1899521 1:400
Mouse anti-GATA2 IgG2a Sigma-Aldrich Cat#WH0002624M1 Lot G2151-2D11 1:100
Mouse anti-TFAP2C IgG1 Santa Cruz Biotechnology Cat#sc-12762 Lot K1318 1:200
Rabbit anti-p63 IgG Cell Signaling Technology Cat#13109 Lot 3 1:800
Mouse anti-SDC1 IgG1 abcam Cat#ab181789 Lot GR3178572-9 1:200
Rabbit anti-CK7 (KRT7) IgG abcam Cat#ab181598 Lot GR3214132-7 1:100
Mouse anti-HLA-G IgG1 [MEM-G/1] abcam Cat#ab7759 Lot GR3262011-5 1:50
Mouse anti-NR2F2 IgG2a abcam Cat#ab41859 Lot GR3291672-1 1:100

For Immunohistochemistry:
Antibody, Company, Catalogue Number, Lot Number, Dilution:
Mouse anti-SDC1 IgG1 Abcam Cat#ab181789 Lot GR3178572-9 1:200
Rabbit anti-CK7 (KRT7) IgG Abcam Cat#ab181598 Lot GR3214132-7 1:8000
Mouse anti-HLA-G IgG1 [MEM-G/1] Abcam Cat#ab7759 Lot GR3262011-5 1:50

Validation

Antibodies obtained from the commercial source were validated by the suppliers, detailed validation analysis relevant literatures are provided on the company website for the products used in this study. Some antibodies were validated in a previously published study as indicated in methods or relevant literature was cited.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Human fibroblasts sourced from ThermoFisher (Catalogue number, C-013-5C and lot#1029000 for 38F, lot#1528526 for 55F and lot#1569390 for 32F) for reprogramming experiments. H9 human embryonic stem cells were used with collaborator lab of Andrew Laslett. JEG-3 cells were obtained from ATCC (Cat no: HTB-36, Lot number: 70000179), JAR cells were also obtained from ATCC (Cat no. HTB-144, Lot number: 70023736).

Authentication

Human dermal fibroblasts were authenticated by ThermoFisher, as stated in the certificate of analysis, these primary fibroblast cell lines were derived from tissues of donor (age and sex identified). H9 human embryonic stem cells were authenticated by short tandem repeat (STR) analysis. JEG3 and JAR cell lines were authenticated by STR analysis.

Mycoplasma contamination

Fibroblasts lines were tested by ThermoFisher, H9 human embryonic stem cells were tested by Laslett lab. JEG3 and JAR cells were tested by ATCC, they are all mycoplasma negative. Furthermore, cell lines were regularly tested and were mycoplasma negative.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

Mouse (Mus musculus) NOD/SCID IL-2R Gamma KO, male and female, 5-20 weeks of age were used.

Wild animals

No wild animals were used.

Field-collected samples

No field-collected samples were used.

Ethics oversight	Protocols and use of animals were undertaken with the approval of the Monash University Animal Welfare Committee following the 2004 Australian Code of Practice for the Care and Use of Animals for Scientific Purposes and the Victorian Prevention of Cruelty to Animals Act and Regulations legislation.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were dissociated with TrypLE express (ThermoFisher), and DPBS (ThermoFisher) supplemented with 2% FBS (Hyclone) and 10µM Y-27632 (Abcam) was used for antibody labeling steps and final resuspension of the samples. The antibody labeling steps were carried out in a volume of 500 µl per 1 million cells, and incubation time was 10 mins on ice per step; after each antibody labeling step, cells were washed with 10 ml cold PBS and pelleted at 400× g for 5 mins. The cells were then resuspended in a final volume of 500 µl, and propidium iodide (PI) (Sigma) was added to a concentration of 2µg/ml. Cell sorting was carried out with a 100 µm nozzle on an Influx instrument (BD Biosciences), and flow cytometry analysis was carried out using an LSRIIb or LSRIIA analyser (BD Biosciences).
Instrument	LSRIIb, LSRIIa analyser or BD Influx cell sorters (BD).
Software	Collection: FACSDiva software suit (version 8.0.1) (BD) for analysers, FACS TM software suit (version 1.2.0.142) (BD) for influx sorters. Analysis: FlowJo (v10) (FlowJo, LLC) & Cytobank 7.2.0 (Cytobank, Inc.).
Cell population abundance	Abundance of distinct cell populations of interest was determined using appropriate negative controls and puritiy of sorted populations as determined by post sort reanalysis.
Gating strategy	Standard gating setting commonly utilized at flowcore facility of Monash University were used. Cell debris was excluded using a FSC vs SSC gate; aggregates were excluded via a FSC-H vs FSC-W approach; dead cells were defined as PI high/positve and gated out; furthermore iMEF feeder cells were gated out via the FITC channel (TRA-1-85 negative). Apart from using appropriate isotype, FMO and unstained controls, positive, negative control cell samples were used to set appropriate gates and determine real positive cell populations and confirmed by post sort reanalysis.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.