Human reprogramming roadmap unveils route to induced trophoblast stem cells

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[8,11,17], Di Chen[8,11], Anna L. Leichter[12], Haroon Naeem[13], Pratibha Tripathi[1,2], Partha P. Das[1,2], Alexandra Grubman[1,2,3], David R. Powell[13], Andrew L. Laslett[3,5], Laurent David[14,15], Susan K. Nilsson[3,5], Amander T. Clark[8,9,10,11], Ryan Lister[6,7], Christian M. Nefzger[1,2,3,18], Luciano G. Martelotto[12], Owen J. L. Rackham[4]* and Jose M. Polo[1,2,3]*

[1]Department of Anatomy and Developmental Biology, Monash University, Victoria, Australia

[2]Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Victoria, Australia

[3]Australian Regenerative Medicine Institute, Monash University, Victoria, Australia

[4]Program in Cardiovascular and Metabolic Disorders, Duke-National University of Singapore Medical School, Singapore

[5]Biomedical Manufacturing, Commonwealth Scientific and Industrial Research Organisation, Victoria, Australia

[6]Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Western Australia, Australia

[7]The Harry Perkins Institute of Medical Research, Western Australia, Australia

[8]Department of Molecular Cell and Developmental Biology, University of California Los Angeles, CA, USA

[9]Molecular Biology Institute, University of California Los Angeles, CA, USA

[10]Jonsson Comprehensive Cancer Center, University of California Los Angeles, CA, USA

[11]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California Los Angeles, CA, USA

[12]Single Cell Innovation Laboratory, University of Melbourne Centre For Cancer Research, The University of Melbourne, Victoria, Australia

[13]Monash Bioinformatics Platform, Monash University, Victoria, Australia

[14]Nantes Université, CHU Nantes, Inserm, CRTI, UMR 1064, ITUN, Nantes, France.

[15]Nantes Université, CHU Nantes, Inserm, CNRS, SFR Santé, FED 4203, Inserm UMS 016, CNRS UMS 3556, Nantes, France.

Present address: [16]University of Melbourne Centre For Cancer Research, The University of Melbourne, Victoria, Australia

Present address: [17]Department of Biochemistry, McGill University, Montreal, Canada

Present address: [18]Institute for Molecular Bioscience, University of Queensland, Queensland, Australia

[19]These authors contributed equally

*Correspondence:

owen.rackham@duke-nus.edu.sg (O.J.L.R.); jose.polo@monash.edu (J.M.Polo)

Summary Paragraph

Reprogramming human somatic cells to primed or naive induced pluripotent stem cells (iPSC) recapitulates the different stages of early human embryonic development[1–6]. The molecular mechanism underpinning the reprogramming of human somatic cells to primed or naive induced pluripotency remains largely unexplored, impeding our understanding and limiting rational improvements to reprogramming protocols. To address this, we reconstructed molecular reprogramming trajectories using single-cell transcriptomics. This revealed that reprogramming into primed and naive human pluripotency follows diverging and distinct trajectories. Moreover, genome-wide accessible chromatin analyses showed key changes in regulatory elements of core pluripotency genes, and orchestrated global changes in chromatin accessibility over time. Integrated analysis of these datasets unveiled an unexpected role of trophectoderm (TE) lineage-associated transcription factors and the existence of a subpopulation of cells that enter a TE-like state during reprogramming. Furthermore, this TE-like state could be captured, allowing the derivation of induced Trophoblast Stem Cells (iTSCs). iTSCs are molecularly and functionally similar to TSCs derived from human blastocysts or first-trimester placental trophoblasts[7]. Altogether, these results provide a high-resolution roadmap for transcription factor-mediated human

reprogramming, revealing an unanticipated role of the TE-lineage specific regulatory program during this process and facilitating the direct reprogramming of somatic cells into iTSCs.

Keywords

Naive human induced pluripotent stem cells, Primed human induced pluripotent stem cells, Reprogramming, Induced trophoblast stem cells, Syncytiotrophoblast, Extravillous trophoblast, Yamanaka factors, Non-integrating, Epiblast, Trophectoderm, Pre-implantation.

Main Text

Human embryonic stem cells (hESCs) are derived from the epiblast of preimplantation blastocysts. Alternatively, human induced pluripotent stem cells (hiPSCs) are generated from adult cells, such as fibroblasts, by transcription factor (TF)-mediated nuclear reprogramming. Both cell types are pluripotent since they can give rise to all cell types within the embryo, but not the extraembryonic tissues (i.e. placenta). Conventionally, hESCs/hiPSCs are cultured in the primed state resembling the post-implantation epiblast, however recently culture conditions have enabled the generation of naive hESCs/hiPSCs, resembling human preimplantation epiblast, an earlier stage in embryonic development[1–3]. Contrary to mouse reprogramming, where comprehensive roadmaps of the reprogramming process have been reported[8–12], few recent studies have revealed details of reprogramming towards human pluripotency[13–15]. Moreover, variations in donor genetic background, culture conditions, reprogramming systems and isolation strategies for reprogramming intermediates can confound results[13–15].

Charting a human reprogramming roadmap

To investigate the cellular transitions during the reprogramming of genetically matched adult human dermal fibroblasts into primed and naive hiPSCs in a clinically relevant way, we utilised integration-free Sendai viruses to deliver the TFs *OCT4/POU5F1*, *KLF4*, *SOX2,* and *c-MYC* (OKSM). Transduced cells were first cultured in fibroblast medium (fm) and then transitioned into media for either primed reprogramming (pr) or t2iLGoY naive reprogramming (nr) (see Methods). Primed and naive reprogramming intermediates and hiPSCs 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 employed two complementary

strategies: (1) 'time-resolved' to track changes happening with respect to time, by collecting intermediates at Day 0 (D0), D4, D8, D12-pr, D12-nr, D16-pr, D16-nr, D20-pr, D20-nr, D24-pr, D24-nr, Passage 3 (P3-nr), P20-pr, P20-nr and subjecting them to single-nucleus RNA sequencing (snRNA-seq) (Fig. 1a); (2) 'media-resolved' 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 based on the medium compositions (libraries FM, PR, and NR) and subjecting them to single-cell RNA sequencing (scRNA-seq) (Extended Data Fig. 1c). We integrated the sn and scRNA-seq datasets, resulting in a dataset of 43,791 cells, robustly detecting 11,549 genes (Extended Data Fig. 1d, Supplementary Table 1,2, see Methods). To visualize the relationships between single cells, we employed force-directed layout (FDL)[16], previously used to characterise mouse reprogramming[12]. FDL shows that 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). Cell identity was further confirmed by 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 further corroborated these findings by applying several complementary dimensionality reduction methods such as principal component analysis (PCA), diffusion maps[17] (DM), and 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 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 the naive reprogramming trajectory is distinct from the primed, rather than an extension of it.

Alternative induced pluripotent conditions

To further characterise the cell populations arising during reprogramming, we performed unsupervised clustering analysis[20], identifying 21 cell clusters (Extended Data Fig. 2a). Notably, we only observed Naive Reprogramming (NR) and Primed Reprogramming (PR) intermediates near the trajectories bifurcation point. The clusters allowed us to apply Partition-based graph abstraction (PAGA)[21] trajectory inference, which confirmed that PR and NR trajectories bifurcate (Extended Data Fig. 2b-d, Fig. 1f). Furthermore, the mesenchymal-epithelial transition (MET) 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 different gene signatures (Extended Data Fig. 2f-h, Supplementary Table 3), with two of these robustly distinguishing primed and naive human hiPSCs. Consistent with a previous study[14], we found that some cells during PR activated the naive signatures, but these cells are still transcriptionally distinct from naive reprogramming intermediates (Extended Data Fig. 2g, Fig. 1g). Furthermore, the 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, see Methods). Bulk RNA-seq 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 hESCs/hiPSCs has been a subject of active research[1–6], with different media producing hiPSCs with a spectrum of naive characteristics[4]. Thus, to study the reprogramming pathways in different media conditions we isolated reprogramming intermediates in other naive media including 5iLAF[2], NHSM[1], and RSeT (Extended Data Fig. 3c-e). Harmonisation of the RNA-seq of the different media-intermediates with the snRNA-seq dataset revealed that NHSM cells follow the previously identified primed reprogramming trajectory, whereas 5iLAF overlaps with that of t2iLGoY. Day 13 and 21 intermediates of the RSeT condition transitioned along the naive t2iLGoY trajectory but ultimately switched branches, establishing that RSeT hiPSCs (Passage 3 and 10) clustered near primed hiPSCs (Fig. 2a, Extended Data Fig. 3f, Supplementary Table 4). These observations were confirmed by scoring these reprogramming intermediates using the primed and naive signatures defined previously (Fig. 2b and Supplementary Table 5). We further examined cell heterogeneity during RSeT reprogramming by scRNA-seq, identifying both primed-like and naive-like intermediates (Supplementary Table 6). The primed-like cell population likely dominates over time, explaining 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 reorganisation of the epigenome[8,10,11,14]. To elucidate the chromatin accessibility landscape and the use of regulatory elements (RE) during reprogramming, we performed Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) on flow-cytometry-isolated reprogramming intermediates (Supplementary Table 4). PCA of the ATAC-seq peaks (Fig. 2c, Extended Data Fig. 5a) and its integration with RNA-seq experiments (Extended Data Fig. 5b,c, see 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*, *DNMT3L*) revealed that loss of accessibility of somatic regulatory elements is accompanied by a gain of open chromatin regions in RE and/or 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], resulting in eight clusters (C1-8) (Supplementary Table 7) and grouped them by their behaviour over time (Fig. 2d). This analysis revealed: (1) Comparable distribution of peaks across genomic region classes in all clusters (Extended Data Fig. 6a); (2) Regions of open chromatin in fibroblasts (C1 and C2) became progressively inaccessible [shared loss (SL)] during reprogramming, concomitant with downregulation of the associated genes (Fig. 2d, Extended Data Fig. 6b,c); (3) Transient clusters (C3 and C4) [shared transient (ST)] exhibit overrepresentation of genes associated with transcription, metabolism, and various organ morphogenesis; (4) Regions with a gradual gain of accessibility for both primed and naive reprogramming (C5) [shared up (SU)] are associated with embryonic development and stem cell maintenance; (5) Regions that specifically gained accessibility during primed reprogramming (C6) [primed up (PU)] were associated with a range of embryonic developmental processes; (6) Two clusters (C7, C8) [naive up (NU), (C7 is also primed transient (PT)] 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 specific TFs that drive these different programs, we identified TF binding-site motifs enriched in each cluster (Supplementary Table 9). Motif enrichment analysis of the SL regions uncovered TFs (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 TFs (e.g. *FOSL1*, *JUNB*) and an enrichment for *OCT4*, *SOX2*, *NANOG* and *KLF4* binding motifs

(Extended Data Fig. 6d,e). This redistribution of somatic TFs to transiently accessible regions harbouring their binding motifs during reprogramming by *OCT4*/*SOX2* supports a similar effect previously described in mice[11], potentially representing a pan-mammalian paradigm of somatic accessible chromatin reorganization mediated by reprogramming factors. Interestingly, two clusters (C7 and C8) show an unexpected significant motif enrichment of trophectoderm (TE) associated TFs (e.g. *TFAP2C*, *GATA2*), and these TFs were specifically upregulated during reprogramming to the naive state or transiently upregulated in the primed state (e.g. C7) (Extended Data Fig. 6d-f, Fig. 2e). Furthermore, the shared C5 cluster also exhibited enrichment for the same factors (Fig. 2e). To test whether these TE-associated TFs were passengers or drivers, we experimentally knocked them down during reprogramming using short hairpin (sh) RNAs (Extended Data Fig. 6g, Supplementary Table 10). While the absence of *TFAP2C* showed a minor effect on the efficiency of primed reprogramming, naive reprogramming was greatly impaired (Fig. 2f). Knockdown (KD) of *GATA2* affected both primed and naive reprogramming, possibly being a result of *GATA2* expression being upregulated earlier in reprogramming (Fig. 2f). Thus, these different transcriptional regulatory processes likely govern naive and primed branches of reprogramming.

Trophectoderm branch during reprogramming

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

Deriving induced trophoblast stem cells

We hypothesised that this TE-associated cell cluster could be stabilised to give rise to trophoblast stem cells (TSCs). Thus, we transitioned naive reprogramming intermediates at day 21 (d21n) into the recently reported human TSC medium[7] (Fig. 3b). Remarkably, we observed the appearance of cells that morphologically resemble TSCs, which we named induced TSC (iTSC^{d21n}) (Fig. 3c). Further characterization showed that iTSC^{d21n} express key markers that define human TE and TSCs[7,26] such as P63, TFAP2C, GATA2, and KRT7 (Fig. 3d, Extended Data Fig. 8a). Moreover, these iTSCs express comparable levels of TSC marker genes and are distinct from human fibroblasts and primed and naive hiPSCs (Extended Data Fig. 8b). To functionally characterize the iTSC^{d21n}, we examined their *in vitro* differentiation capacity to give rise to syncytiotrophoblast (ST) and extravillous trophoblast (EVT) cells, the major trophoblast subtypes of the placenta[26]. This demonstrated that iTSC^{d21n} can be differentiated into ST cells characterised by SDC1-positive multinucleated cells and EVT cells defined by upregulation of HLA-G, a key histocompatibility molecule expressed in placenta[7,26] (Fig. 3e, Extended Data Fig. 8c). The iTSC^{d21n}-ST cells showed significantly higher fusion index compared to iTSC^{d21n} and secreted human chorionic gonadotropin (hCG) that could be detected using an over-the-counter (OTC) human hCG pregnancy test stick and quantified by hCG ELISA (Extended Data Fig. 8d-f). Next, we evaluated the *in vivo* differentiation potential of iTSC^{d21n} by subcutaneous injection into mice. Nine days post-injection (P.I.), mouse urine was positive for hCG using the OTC human pregnancy tests (Fig. 3f, see 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 ST-like cells and HLA-G-positive EVT-like cells comparable to the reported primary tissue-derived TSCs[7] (Extended Data Fig. 8h,i, Fig. 3g). Importantly, these results demonstrate that iTSCs^{d21n} are bipotent *in vitro* and *in vivo*. Finally, we used CD70-low to enrich TE-like cells from the 'nic' cluster and demonstrated that the identified TE-like cluster carries the greatest potential for iTSC^{d21n} generation (Extended Data Fig. 8j,k). Altogether, this suggests that cell fate specification is highly dynamic and plastic during human somatic cell reprogramming.

Reprogramming fibroblasts directly into iTSCs

To test whether iTSCs could be derived directly from human fibroblasts, we started reprogramming experiments and transitioned the day 8 intermediates into (1) TSC or (2)

naive medium, or (3) 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 intermediates also consist of cells with strong epiblast, primed, and naive signatures (Extended Data Fig. 9e), and accordingly they were able to give rise to pluripotent and trophoblast stem cell lines (Extended Data Fig. 9f-h). We noticed that the proportion of TE-like population was the highest in TSC media compared to fibroblast and naive media (Fig. 4b, Extended Data Fig. 9d). Therefore, we hypothesized we could derive iTSC lines more efficiently by directly transitioning day 8 intermediates into TSC media (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, iTSCs^{d8} were successfully derived directly, and our transgene-free iTSCs^{d8} (Extended Data Fig. 10a) have demonstrated the capacity to undergo >50 passages thus far without a growth rate reduction. We then performed a comprehensive molecular and functional characterisation of iTSC^{d8} based on features defined for TSCs generated from primary sources[7,26–29]. This demonstrated that: (1) These iTSC^{d8} expressed key marker genes indicative of mononuclear trophoblasts[26] (Fig. 4e), and (2) they could differentiate into STs and EVTs. The STs expressed SDC1, displayed cell fusion and hCG secretion (Fig. 4f-g, Extended Data Fig. 10b-e). EVTs expressed HLA-A, B, C pan markers, but not HLA-B marker, and importantly they did express HLA-G (Extended Data Fig. 10f-h). We found that the expression of HLA-A, B, C was detected in iTSCs, similar to what was reported in TSCs derived from blastocysts[7]. In contrast, trophoblast organoids are HLA-negative[28] suggesting that the culture conditions might support TSCs at different stages of gestation. (3) Furthermore, our iTSCs and iTSC-derived STs/EVTs share a common transcriptomic profile with the corresponding primary cell types in other published datasets (Fig. 4h, Extended Data Fig. 10i-l, Supplementary Table 14). (4) iTSCs also shows higher levels of expression of microRNAs (miRNAs) from the chromosome 19 miRNA cluster (C19MC) compared to fibroblast and hiPSCs, a unique feature of primary trophoblast[26] (Fig. 4i). (5) We observed specific open chromatin accessibility at the promoter and putative enhancer regions of the *ELF5* locus in our iTSCs and TSC^{BT5} (data from[30]) (Fig. 4j), which has previously been found to be hypomethylated[7,26]. (6) 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). Thus, these results

confirmed that iTSC^{d8} derived directly from human fibroblasts are similar to the primary TSCs.

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 easy exploration of the dataset. This roadmap revealed that the two reprogramming trajectories diverge, and in order for a cell to reprogram into a naive pluripotent state it does not need to first acquire a primed pluripotent state, indicating 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 hypothesise 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 where a TE-associated network is upregulated and one where it is not. The fact that the knockdown of TFs predicted to be driving those 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), indicating that each medium promotes not only a similar final pluripotency state, as we have shown previously[4], but also drives the intermediate cells along similar trajectories. Together, these results present a 'push or pull' question: are similar reprogramming trajectories determined by being pulled towards a common final pluripotency state, or do the specific culture media pushes the cells along similar trajectories, and as a consequence result in similar final states?

The change in chromatin accessibility during primed and naive reprogramming also indicate a bifurcated trajectory. Early and transient chromatin accessibility clusters are enriched in OKS motifs, suggesting binding of these TFs at initially closed regions and supporting 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 (Fig. 2e, 3a), since one of the first cell fate decisions that cells make during development is whether they will become trophoblast or epiblast. Interestingly, our results revealed 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 OKSM can induce human fibroblasts to acquire

pluripotent and trophoblast states. The direct reprogramming of fibroblasts into iTSCs is in contrast to the recently reported three-step-approach where somatic cells must first be reprogrammed into hiPSCs, then converted into the expanded-potential or naive stem cells before being differentiated into TSCs[30,32]. We envision that this direct approach will facilitate the generation of patient-specific iTSCs to study trophoblast dysfunction. Such studies are critically needed as this dysfunction leads to various complications during pregnancy, such as preeclampsia and intrauterine growth restriction[7,26,28]. Furthermore, having stable, self-renewing, *bona fide* isogenic human iPSC and iTSC lines will provide a unique 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. As such, 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, since both embryonic and extraembryonic lineages can be derived, these results also hint at the intriguing 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.

References

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**, (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**, (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. *bioRxiv* (2020).

Main Fig. Legends

Fig. 1 | Charting a human reprogramming roadmap. a, Experimental design. **b,** FDL of 43,791 cells, highlighting the snRNA-seq and scRNA-seq libraries. **c,** Expression of marker genes associated with human fibroblasts (*ANPEP*), 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 on FDL. **g,** Predicted cell states using defined gene signatures on FDL. For more details on sample number and statistics, please see statistics and reproducibility section.

Fig. 2 | Distinct transcriptional regulatory programs drive primed and naive human reprogramming. a, PCA of the integrated bulk RNA-seq of primed and several types of naive reprogramming intermediates with snRNA-seq datasets (see Methods), $n \geq 2$. **b,** Naive and primed signatures scores of reprogramming intermediates under different conditions. **c,** PCA of ATAC-seq signals, $n=2$. **d,** Clustering analysis of ATAC-seq peaks during reprogramming. 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* and *GATA2* in ATAC-seq clusters (C1-C8). **f,** Reprogramming efficiency upon *TFAP2C* KD into primed ($n=6$ each for control and

sh*TFAP2C*) and naive (*n*=6 each for control and sh*TFAP2C*) pluripotency, and reprogramming efficiency upon *GATA2* KD 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 number and statistics, please see statistics and reproducibility section.

Fig. 3 | Derivation of iTSCs during reprogramming. a, *In vivo* TE signatures on FDL projection overlaid with single-cell trajectories constructed using Monocle3 (black lines). Blue arrow indicates TE-enriched cell population. **b,** Experimental design for derivation of iTSC^{d21n}. **c,** Phase-contrast image of iTSC^{d21n}. Scale bar, 100µm. **d,** Immunostaining of iTSC^{d21n} with P63, TFAP2C, GATA2, KRT7. Scale bar, 100µm. Representative images from *n*=4. **e,** SDC1 and HLA-G immunostaining of ST and EVT cells, respectively, differentiated from iTSC^{d21n}. Scale bar, 100µm. 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. 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 and HLA-G in the lesions harvested from iTSC^{d21n}-engrafts in NOD-SCID mice. No evident lesions were observed in vehicle controls. Scale bar, 200µm. Representative images from *n*=4. For more details on sample number and statistics, please see statistics and reproducibility section.

Fig. 4 | Direct derivation of iTSCs from human fibroblasts. a, FDL representation of scRNA-seq libraries of day 21 reprogramming intermediates (10,518 cells). **b,** TE signatures on FDL projections, TE-like population is highlighted and coloured by the library. **c,** Experimental design of direct derivation of iTSC^{d8} from fibroblasts. **d,** Phase-contrast image of iTSC^{d8}. Scale bar, 100µm. **e,** Immunostaining of iTSC^{d8} for several TSC makers. Scale bar, 100µm. **f,** Phase-contrast and immunostaining of ST and **g,** EVT cells differentiated from iTSC^{d8}. Scale bar, 100µm. *n*=4 for **d-g**. **h,** Spearman correlation of transcriptomes from this study with published datasets. Biological replicates (*n*≥2) are averaged prior to performing correlation. **i,** C19MC miRNAs expression normalised to miR-103a, mean ± s.e.m., not detected (ND), red dotted line indicates level in primed hiPSCs. *n*=2. **j,** ATAC-seq signal at *ELF5* region in indicated cell types (TSC^{BT5} derived from human blastocysts[30]), mean value of replicates (*n*=2), TSC 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,** Hematoxylin and eosin, and

immunohistochemical staining of KRT7, SDC1 and HLA-G in the lesions harvested from iTSC^{d8}-engrafts in NOD-SCID mice, n=4, no evident lesions were observed in vehicle controls. Scale bar, 200μm. For more details on sample number and statistics, please see statistics and reproducibility section.

METHODS

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, mycoplasma tested as described in the Reporting Summary. Primary human adult dermal fibroblasts (HDFa) from three different female donors were obtained from ThermoFisher (Catalogue number C-013-5C and lot#1029000 for 38F, lot#1528526 for 55F and lot#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 iPSCs (established lines) and H9 ESCs (WiCell Research Institute, Madison, WI, 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 days using 0.5 mM EDTA (Invitrogen). Culture conditions used for human somatic cell reprogramming were prepared as described previously[4,34]. **Fibroblast medium**: DMEM (ThermoFisher), 10% Fetal Bovine Serum (FBS, Hyclone), 1% Nonessential amino acids (ThermoFisher), 1mM GlutaMAX (ThermoFisher), 1% Penicillin-streptomycin (ThermoFisher), 55μM 2-mercaptoethanol (ThermoFisher) and 1mM sodium pyruvate (ThermoFisher). **Primed medium**: DMEM/F12 (ThermoFisher), 20% Knockout Serum Replacement (KSR, ThermoFisher), 1mM GlutaMAX (ThermoFisher), 0.1mM 2-mercaptoethanol (ThermoFisher), 1% Non-essential amino acids (ThermoFisher), 50ng/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 2mM L-Glutamine (ThermoFisher), 0.1mM 2-mercaptoethanol (ThermoFisher), 0.5% N2 supplement (ThermoFisher), 1% B27 supplement (ThermoFisher), 1% Penicillin-streptomycin (ThermoFisher), 10ng/ml human leukemia inhibitory factor (LIF, made in house), 250μM L-

ascorbic acid (Sigma), 10µg/ml recombinant human insulin (Sigma), 1µM PD0325901 (Miltenyi Biotec), 1µM CHIR99021 (Miltenyi Biotec), 2.5µM Gö6983 (Tocris), 10µM Y-27632 (Abcam). **Naive Human Stem cell Medium (NHSM)**: culture condition adapted from Gafni and colleagues[1] with suggested modifications from the Hanna laboratory's web page in 2014 was used. DMEM/F12 (ThermoFisher) supplemented with 10mg/ml AlbuMAX I (ThermoFisher), 1% Penicillin-streptomycin (ThermoFisher), 1mM GlutaMAX (ThermoFisher), 1% Nonessential amino acids (ThermoFisher), 10% KSR (ThermoFisher), 5ml N2 supplement (ThermoFisher), 12.5µg/ml recombinant human insulin (Sigma), 50µg/ml L-ascorbic acid (Sigma), 20ng/ml of recombinant human LIF (made in house), 8ng/ml FGF2 (Peprotech), 2ng/ml recombinant TGF-β1 (Peprotech), 20ng/ml human LR3-IGF1 (Prospec), and small molecule inhibitors: 1µM PD0325901 (Miltenyi Biotec), 3µM CHIR99021 (Miltenyi Biotec), 5µM SP600125 (Tocris), 2µM BIRB796 (Axon), 0.4µM LDN193189 (Axon), 10µM Y-27632 (supplemented daily to media from freshly thawed stock aliquot; Abcam) and 1µM 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), 1mM GlutaMAX (ThermoFisher), 1% Penicillin-streptomycin (ThermoFisher), 0.1mM 2-mercaptoethanol (ThermoFisher), 50µg/ml Bovine Serum Albumin (ThermoFisher), 1µM PD0325901 (Miltenyi Biotec), 1µM IM-12 (Millipore), 0.5µM SB590885 (Tocris), 1µM WH-4-023 (A Chemtek), 10µM Y-27632 (Abcam), 20ng/ml Activin A (Peprotech), 8ng/ml FGF2 (Miltenyi Biotec), 20ng/ml human LIF (made in house) and 0.5% KSR (ThermoFisher). **Naive RSeT medium**: 100ml of RSeT 5X supplement, 1ml of RSeT 500X supplement and 0.5ml of RSeT 1000X supplement into 398.5ml of RSeT Basal Medium; (Stem Cell Technologies) supplement with 1% Penicillin-streptomycin (ThermoFisher). **Human trophoblast stem cell (TSC) medium**[7]: DMEM/F-12, GlutaMAX (ThermoFisher) supplemented with 0.3% BSA (Sigma), 0.2% FBS (ThermoFisher), 1% ITS-X supplement (ThermoFisher), 0.1mM 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 naive t2iLGoY medium was used for naive reprogramming as we have previously shown that it can be used to reprogram fibroblasts into naive hiPSCs, with all the hallmarks of naive pluripotency and maintains a more stable karyotype when compared to other conditions[4]. Human somatic cell reprogramming into primed and naive pluripotent states experiments and subsequent culture of primed and naive hiPSCs were performed as previously described[4,34]. Briefly, reprogramming of human fibroblasts was conducted using CytoTune-iPS 2.0 Sendai reprogramming kit according to the manufacturer's instructions (ThermoFisher). Primary HDFa were seeded at a density of ~5-10×10^4 cells in fibroblast medium. As shown in Fig. 1a, cells were transduced with Sendai viruses in FM at the multiplicity of infection (MOI) as follows, KOS MOI=5 or 10, c-MYC MOI=5 or 10, 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, 5iLAF) the next day. After 18-21 days, hiPSCs could be passaged and expanded as described previously[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 described previously[11,37]. HDFa were transduced with lentiviral vectors for one week and replated two days before Sendai transduction. Colony counts of *TFAP2C*, *GATA2* knockdown experiments are provided in Source Data Fig. 2f. Knockdown experiments were validated by qRT-PCR, and primers used are listed in Supplementary Table 15. All cells were cultured at 37 °C, 5% O_2 and 5% CO_2 incubators. For the derivation of $iTSC^{d21n}$ during naive reprogramming, day 21 naive t2iLGoY reprogramming intermediates were transitioned into TSC 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, iTSCs 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 tissue culture flask that was pre-coated with 5μg/ml Collagen IV (Sigma) (for at least one hour 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 TSC medium. After 10-13 days, $iTSC^{d8}$ can be passaged onto iMEF feeders and cultured in a 37 °C, 5% O_2 and 5% CO_2 incubator as described for $iTSC^{d21n}$ above. Sendai detection in established iTSC cell lines was performed as described in the Sendai reprogramming protocol (ThermoFisher). For the derivation of primed, naive hiPSCs and iTSCs from d21 fibroblast reprogramming

intermediates, day 21 fibroblast reprogramming intermediates were transitioned into primed, naive or TSC media, and then cultured and expanded as described above.

Differentiation of iTSC^{d21n} and iTSC^{d8} into ST and EVT *in vitro*. Differentiation of iTSCs into ST and EVT was performed as previously described[7]. For the differentiation of iTSCs into ST, iTSCs were seeded at a density of 1 x 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), 0.1mM 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. 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 iTSCs into EVT, iTSCs were seeded at a density of 0.75 x 10^5 cells per well onto a 6-well plate pre-coated with 1 μg/ml Col 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), 0.1mM 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 hNRG1 (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 hNRG1 (Cell Signaling) or KSR (ThermoFisher), and Matrigel (Corning) was added to 0.5% final concentration. The cells were cultured for an additional 2 days before analyses were performed.

iTSC^{d21n} and iTSC^{d8} *in vivo* engraftment assay. 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. iTSCs with 80% confluency were dissociated with TrypLE express (ThermoFisher) and counted. 10^7 iTSCs 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 weeks of age NOD/SCID IL-2R Gamma KO mice (100 µl into each flank). Mice were randomised between controls and iTSCs, but not blinded. Nine days after injection, urine, blood serum, and lesions were collected from the mice for analysis. Mice urine and serum were utilized for the detection and measurement of hCG secretion as detailed below. Collected lesions were fixed with 4% Paraformaldehyde (PFA, Sigma) overnight and subsequently embedded in paraffin. Lesions collected were less than $1cm^3$ in volume. Paraffin-embedded tissues were sectioned and stained with hematoxylin-eosin (H&E) or proceeded with immunohistochemistry staining of KRT7, HLA-G, SDC1 (Supplementary Table 16) at the Histology Platform at Monash University.

Pregnancy tests and hCG ELISA. iTSCs were seeded at a density of 0.5×10^5 cells/ml on a 12-well plate for ST differentiation as detailed in the above section. 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 controls, iTSCs were also seeded at a density of 0.5×10^5 cells/ml on a 12-well plate and cultured in TSC medium. 2 days later, the conditioned medium was collected and stored at -80°C. The conditioned media were then tested using OTC hCG pregnancy test sticks (Freedom) according to the manufacturer's recommendations. In addition, the hCG level within the media was also measured using hCG ELISA kit (Abnova, ABNOKA4005) according to the manufacturer's instructions. Following the iTSC engraftment assay, the collected mouse urine was tested using the OTC hCG pregnancy test sticks as described above and hCG level in blood serum was measured using hCG ELISA kit as described above.

Flow cytometry analysis and fluorescent activated cell sorting (FACS). All antibodies used in flow cytometry analysis and FACS experiments were summarized in Supplementary Table 16. 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. For SPADE analysis (Extended Data Fig. 3e), a three-step antibody labeling 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). (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 labeling was performed as below: (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). (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 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). For Supplementary Fig. 1, reprogramming intermediates were isolated on day 3 into CD13+F11R+ and CD13+F11R- subpopulations, and then reseeded into FM condition for five days for flow cytometry reanalysis and for hiPSC 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 t2iLGoY 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 iTSCs purification, a two-step antibody labeling procedure was used: (1) mouse anti-human APA (1:100) (2) BUV395 mouse anti-human TRA-1-60 (1:100, BD); APC rat anti-human & mouse CD49F (ITGA6) (1:20, Miltenyi Biotec); AF488 goat anti-mouse IgG1 antibody (1:2,000, ThermoFisher). iTSCs purification was performed on the reprogrammed cells at passage 9-10 by isolating TRA160-APA+ITGA6+ subpopulations and reseeding onto Col IV-coated 6-well plate for long-term passaging. For Extended Data Fig. 8k, enrichment of CD70-high, 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,

673 BD); BV421 mouse anti-human CD326 (EpCAM) (1:100, BioLegend); BUV395 mouse anti-
674 human TRA-1-60 (1:100, BD); APC mouse anti-human F11R (1:250, CSIRO CSTEM27);
675 BUV737 mouse anti-human CD70 (1:100, BD). Details of these antibodies are provided in
676 Supplementary Table 16. Labeled cells were resuspended in a final volume of 500 µl
677 containing 2µg/ml of propidium iodide (PI) (Sigma) for cell sorting. TRA185+CD13-
678 F11R+TRA-1-60+EPCAM+CD70-high and TRA185+CD13-F11R+ TRA-1-
679 60+EPCAM+CD70-low subpopulations denoted as CD70-high and CD70-low
680 subpopulations respectively were isolated and reseeded onto a layer of iMEF feeders (24-well
681 plate) at a density of 5 x 10^3 cells per well. On the next day after reseeding, the spent culture
682 medium was replaced with the TSC medium. Immunostaining for KRT7 positive colonies
683 was then performed on day 9 after reseeding as described below. We demonstrated that the
684 CD70-low TE-like novel intermediates resulted in more KRT7+ iTSC colonies as compared
685 to unenriched or CD70-high naive populations, indicating that the identified TE-like cluster
686 carries the greatest potential for the generation of $iTSC^{d21n}$ (Extended Data Fig. 8k). For HLA
687 experiments, cells were labeled with HLA-A, B, C (W6/32) or HLA-Bw4 (1:1, Purcell lab),
688 then AF647 goat anti-mouse IgG antibody (1:1000, ThermoFisher). Or cells were labeled
689 with (1) HLA-G MEM-G/9 (1:500, Abcam); (2) AF488 goat anti-mouse IgG antibody
690 (1:1000, ThermoFisher); (3) PE-Cy7 mouse anti-human HLA-A, B, C W6/32 (1:200,
691 Biolegend).

692 **Multidimensional analyses of flow cytometry data.** To visualise the multidimensional flow
693 cytometry data, we employed spanning-tree progression analysis of density-normalized
694 events (SPADE)[38]. SPADE trees were generated as described previously[39] using the
695 Cytobank platform (http://www.cytobank.org). Samples were labeled with antibodies as
696 described above for flow cytometry analysis and all experiments were performed on the same
697 day to warrant their use for comparison. The SPADE tree indicates a clear transition of cell
698 populations at the early stages of reprogramming (from day 0 to day 7), with reprogramming
699 in NHSM and RSeT conditions exhibiting a more primed-like transition (Extended Data Fig.
700 3e). In particular, the RSeT media formed a separated branch on the SPADE tree, in contrast
701 to reprogramming in 5iLAF and t2iLGoY (Extended Data Fig. 3e).

702 **Quantitative RT-PCR.** RNA was extracted from cells using RNeasy micro kit (Qiagen) or
703 RNeasy mini kit (Qiagen) and QIAcube (Qiagen) according to the manufacturer's
704 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).

Quantitative RT-PCR 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 was analyzed as follows: In each sample, hsa-miR-103a was used for normalization to obtain ΔCt value for each miRNA. 2^ (-ΔCt) was then calculated for each miRNA to obtain the relative expression against hsa-miR-103a. The values obtained were multiplied by 1000 and then the results were plotted in logarithmic scale[26] (Fig. 4i). All primers used were listed in the Supplementary Table 15.

Immunostaining. Cells were fixed in 4% Paraformaldehyde (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 were described in Supplementary Table 16. For example, primary antibodies used: rabbit anti-KLF17 polyclonal (1:500, Sigma), 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 hour. 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, goat anti-rabbit IgG AF555 (1:400, ThermoFisher) or goat anti-rabbit IgG AF647 (1:400, ThermoFisher) for KLF17 (Supplementary Table 16). After labeling, cells were stained with 4',6-Diamidino-2-Phenylindole, Dihydrochloride (DAPI, 1:1000, 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 positive colonies for naive condition, and KRT7 positive colonies for Extended Data Fig. 8k, 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].

Single-nucleus RNA-sequencing (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, hiPSC naive (passage 3), hiPSC primed (passage 20) and hiPSC naive (passage 20) were collected and cryopreserved. These collected samples were then subjected to FACS, for D0, D4, D8, D12 primed, D12 naive, D16 primed, D16 naive, D20 primed, D20 naive, D24 primed and D24 naive samples were sorted for PI negative, TRA-1-85 positive cells to remove dead cells and iMEF cells, while hiPSC primed (passage 3) and hiPSC naive (passage 3 and passage 20) samples were sorted for PI negative, TRA-1-85 positive, CD13 negative, F11R positive, TRA-1-60 positive, EPCAM positive 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 timepoint, 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 protocols.io. Briefly, cells were thaw and pelleted at 500xg for 5 minutes at 4°C. 500 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 1000 µL bore tip and rest on ice for 5' to complete lysis. The homogenate was filtered once using a 70 µm Flowmi filter and centrifuged at 500xg for 5 minutes at 4°C. After removing the supernatant (leaving 50 µL behind) the nuclei pellet was washed with 1000 µL of chilled Nuclei Wash and Resuspension Buffer (1x PBS, 1.0% BSA, 0.2 U/µl RNase Inhibitor). The nuclei were again pelleted at 500g for 5 minutes at 4°C, remove supernatant leaving behind ~50 µL and gently resuspend nuclei in 1000 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). Nuclei suspension was filtered using a 40 µm Flowmi filter, nuclei integrity was visually inspected under a microscope, and 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 RT Enzyme: 20 µL RT Buffer, 3.1 µL TSO primer, 2 µL Additive B and 30 µL H_2O. After sorting nuclei (1000-7000 nuclei depending on sample), complete volume to 80 µL with H_2O, add 8.3 uL RT Enzyme C, mixed and proceeded with chip loading. All the steps from forward were carried out as described in the Chromium Single Cell 3' Reagent Kits User Guide (v3 Chemistry). Sequencing was done on a Illumina NovaSeq 6000 using a paired-end 2x150 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 (v3.0.2). 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 UMI counting were then performed using Cellranger against Ensembl's GRCh37 genome annotation (version 82, including protein-coding, lincRNA and antisense byotipes) containing the Sendai virus sequences as extra transcripts. The endogenous expression of Yamanaka factors was quantified by only counting sequencing reads against the 5' and 3' UTR regions of the endogenous OKSM transcripts.

Single cell RNA-sequencing (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, hiPSC primed (passage 3) and hiPSC naive (passage 3) were collected and cryopreserved. These collected samples were then subjected to FACS, for D0, D3, D7, D13 primed, D13 naive, D21 primed and D21 naive samples were sorted for PI negative, TRA-1-85 positive cells to remove dead cells and iMEF cells, while hiPSC primed (passage 3) and hiPSC naive (passage 3) samples were sorted for PI negative, TRA-1-85 positive, CD13 negative, F11R positive, TRA-1-60 positive, EPCAM positive 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 D0, 3 and 7, samples two and three contained cells for primed (D13, D21, hiPSCs) and naive reprogramming (D13, D21, hiPSCs) respectively, and a small number of mixed D0, 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 115b instead of 98b. Libraries were diluted according to the manufacturer's instruction "NextSeq 500 System User Guide" Illumina document number 15046563 v02 and loaded at 1.8pM. Chromium barcodes were used for demultiplexing and FASTQ files were generated from the mkfastq pipeline using the Cellranger program (v2.1.0). Alignment and UMI counting were performed to the hg19 genome as per the snRNA-seq. 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 harvested and sorted for PI negative, TRA-1-85 positive 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 28bp and R2 87bp) 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 (v3.1.0). Alignment and UMI counting were performed to the hg19 genome as per the scRNA-seq experiments.

snRNA-seq and scRNA-seq cell calling, 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 log10-transformed total UMI counts (Y-axis) were then plotted against the log10-transformed rank (X-axis). The first “knee” point in this UMI-barcode rank plot represents a drastic 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 Tirosh et al[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 are very little mitochondrial or housekeeping genes detected.

Next, quality control was performed at the gene level. Genes with (i) low log10 (average UMI) [*log10aveUMI*] or (ii) do not have at least *minUMI* UMIs in at least *minCell* cells were discarded. Cutoffs of *log10aveUMI* < -2.5, *minUMI* = 2, *minCell* = 10 and *log10aveUMI* < -2, *minUMI* = 2, *minCell* = 10 were applied to discard genes for the snRNA-seq and scRNA-seq respectively. After quality control, 36,597 cells / 17,004 genes and 7,194 cells / 12,246 genes remain 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 (v1.1.6). 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 (v3.1.1) 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 *FindNeighbors* (using top 10 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 employed the Seurat v3 integration technique (v3.1.1)[43]. Seurat v3 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. Principal component analysis (PCA) was performed on the scaled gene expression using the *RunPCA* function in Seurat package (v3.1.1). Following that, Uniform Manifold Approximation and Projection (UMAP) and t-Distributed Stochastic Neighbor 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.neighbors* function (using the top 14 PCs generated above) and *scanpy.tl.diffmap* function in the scanpy package (v1.4.4.post1)[44]. Force-directed layout 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 based on 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 prior to calculating the differentiation potential. The raw counts were supplied as input and default settings were used. Second, we employed Monocle3 (v0.1.3)[19] which learns a trajectory graph from a dimension reduction. In particular, we did a modification where 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 partition-based graph abstraction (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 (v1.4.4.post1) using the Seurat cell clusters as input. The generation of the cell clusters will be described in the next section.

scRNA-seq and snRNA-seq cell clustering. The single cells were clustered using the *FindNeighbors* (using the top 14 PCs for consistency with the dimension reductions) and *FindClusters* function (resolution = 0.5) in the Seurat (v3.1.1) package, which implements an

unsupervised graph-based algorithm. This resulted in 21 clusters which are then labeled using a combination of letters and a number (e.g. cluster fm1) which were determined from the cell composition of the cluster (fm: fibroblast medium, mix: shared clusters, pr: primed reprogramming, nr: naive reprogramming, nic: novel intermediates cluster, re: refractory cells) and the ordering of the cell population along reprogramming trajectory.

snRNA-seq differential expression, 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 was chosen over the scRNA-seq as the former has more cells and a larger number of detected genes. Prior to differential expression, we performed clustering on only the snRNA-seq using the procedure described earlier (using the top 12 PCs instead), 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 are then adjusted for multiple testing using the Benjamini–Hochberg procedure to yield the false discovery rate (FDR). Genes are deemed differentially expressed if the log2 fold change (LFC) 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 LFC (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 i.e. the gene signatures. Overall, we identified eight gene signatures (Supplementary Table 3), namely fibroblast (snRNA-fm1, snRNA-fm2, snRNA-fm3, snRNA-fm4); mixed (snRNA-mix); early-primed (snRNA-pr1); primed: snRNA-pr2, snRNA-pr3, snRNA-pr4); novel intermediates signature (snRNA-nic); naive (snRNA-nr1, snRNA-nr2, snRNA-nr3, snRNA-nr4); nonReprog1 (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 only pick 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 504 genes across all eight gene

signatures (Supplementary Table 3). We then determine 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 are then randomly selected from the same bin for each gene in the gene signature. Every single cell is then assigned to one of the 8 gene signatures based on the highest gene signature strength. This is 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, EPI and PE gene signatures in each single cell (Fig. 3a and 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 are assigned to the G1 phase if both S and G2M scores are less than zero. Otherwise, they are assigned either the S or G2M phase based on the higher of the S and G2M scores.

scRNA-seq analysis of RSeT reprogramming. The RSeT reprogramming (RR) scRNA-seq dataset was analyzed together with the FM, PR, and NR 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 QC and *log10aveUMI* < -2, *minUMI* = 2, *minCell* = 10 for gene QC. This resulted in 9,852 cells / 12,590 genes after quality control. Subsequently, the combined scRNA-seq datasets are analyzed using a similar workflow as the previous scRNA-seq and snRNA-seq dataset. The dataset was preprocessed using Seurat v3'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 force-directed layout) 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). We have previously 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 are analyzed using a similar workflow as the previous scRNA-seq and snRNA-seq dataset (Supplementary Table 13). Briefly, quality control (QC) 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 QC and *log10aveUMI* < -2, *minUMI* = 2, *minCell* = 10 for gene QC. This resulted in 10,518 cells / 12,611 genes after quality control. Subsequently, the dataset was preprocessed using Seurat v3's *NormalizeData* (with default settings), *FindVariableFeatures* (using 1500 features) functions. Next, PCA was performed, followed by other dimension algorithms (UMAP, t-SNE, diffusion maps and force-directed layout) 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 are then labeled using a combination of letters and a number (e.g. cluster D21tr1) which were determined from the cell composition of the cluster (D21fm: fibroblast medium, D21nr: naive reprogramming, D21tr: TSC reprogramming) and the ordering of the cell population along reprogramming trajectory. The strength of the 8 gene signatures defined in this study is also calculated as per the previous scRNA-seq and snRNA-seq dataset.

RNA-sequencing (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#74004) from ~2–20 × 10^4 cells with QIAcube (Qiagen). The concentrations of RNA were measured by a Qubit RNA HS Assay Kit (ThermoFisher, Cat#Q32855) on a Qubit 2.0 Fluorometer (ThermoFisher). ~25 ng of RNA was used for library construction with the SPIA kit (NuGen) and subsequently sequenced by 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, O'Brien et al[22]. [D0 fibroblasts, *n*= 2 (32F and 55F biological replicates)] and Liu et al[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 36nt) and mapped to a custom version of hg19 human genome (with modifications described above in

the scRNA-seq sequencing and processing section) with STAR (v 2.4.2a)[46]. Gene read counting was performed with featureCounts (v1.5.2, unstranded)[47] against the custom version of Ensembl's GRCh37 annotation with modifications described above in the snRNA-seq/scRNA-seq sequencing and processing section. 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 (v3.24.3) with the arguments normalized.lib.sizes = TRUE and prior.count = 1 to yield fragments per kilobase per million (FPKM). Principal component analysis (PCA) was then performed on the log-transformed log2 (FPKM+1) on the top 500 most highly variable genes using the *prcomp_irlba* function in the irbla package (v2.3.3). 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 (v3.5.1).

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 treat each bulk RNA-seq sample as a "single-cell" and performed the same Seurat v3 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 a lot fewer samples (50 samples) than the single-cell counterpart. We then aggregate the gene expression of the combined gene expression as follows. For the bulk RNA-seq, samples were aggregated based on the media condition and timepoint. For the single cells, the scRNA-seq cells and non-reprogrammed cells were removed and the remaining single nucleus was aggregated based on the media condition and timepoint.

Scoring of bulk RNA-seq samples using the primed/naive gene signatures and TE/EPI/PE signatures. For the bulk RNA-seq samples of reprogramming intermediates, we employ 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 utilized gene sets that are highly expressed in the epiblast (EPI), primitive endoderm (PE) and trophectoderm (TE) based on the 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 E5 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, 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 media/timepoint condition was performed as follows. Condition-specific differential expression was performed using the empirical Bayes quasi-likelihood F-tests in the edgeR package (v3.24.3) between the condition of interest and the average expression of the remaining conditions. Gene set enrichment analysis was then performed on the log fold changes from these differential expression results using the fgsea package (v1.8.0) with 10,000 permutations.

RNA-seq for characterization of iTSCs and iTSC-differentiated cells. For the bulk RNA-seq of the iTSC and iTSC-differentiated cells, RNA-seq was performed with a multiplexing approach, using an 8 bp sample index[48] and a 10 bp unique molecular identifier (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) utilizes 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, #TG-160-2005) in accordance with the Illumina Protocol 15046563 v02, generating 2 reads per cluster composed of a 19 bp R1 and a 72 bp R2.

Analysis of the RNA-seq of iTSCs and iTSC-differentiated cells. The sequencing reads are demultiplexed using sabre (v1.0) using the barcodes-sample table, and allowing up to one mismatch per barcode, and a minimum UMI length of 9bp. The demultiplexed data has single reads per sample and UMIs are added to the read name. We use STAR (v2.5.2b)[46] to align the reads to the GRCh37 Ensembl reference genome (v87). Read deduplication based on UMIs

was performed with je markdupes (v1.2)[49] and transcript read counts calculated with featureCounts (v1.5.2)[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 iTSC, iTSC-derived EVT/STs with published transcriptomic datasets, namely (i) blastocyst-derived TSCs gene expression from Okae et al[7]. and Dong et al[30].; (ii) trophoblast organoids gene expression from Haider et al[29]. and Turco et al[28]. and (iii) single-cell gene expression (only Smart-seq2) of the fetal-maternal interface from Vento-Tormo et al[27]. The *removeBatchEffect* function in the limma package (v3.38.3) 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.

Assay for transposase-accessible chromatin using sequencing (ATAC-seq). ATAC-seq samples were prepared as previously described[50]. Briefly, reprogramming intermediates and hiPSCs were isolated by FACS (Supplementary Table 4) and ~65k 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#10977-015), 25 µl of Tagment DNA Buffer (Illumina, Cat#15027866) and 2.5 µl of Tagment DNA Enzyme 1 (Illumina, Cat#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#28204) according to the manufacturer's instructions. 11 µl of transposed DNA, 25µl of the NEBNext High-Fidelity 2x PCR Master Mix (Cat#M0541S) and 1.25µM of the adaptor sequences as published previously[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#28004) followed by Agencourt AMPure XP beads (Beckman Coulter, Cat#A63880) according to the manufacturer's specifications, where library fragments ranging from 200 to 700 bp were selected and sequenced on an Illumina HiSeq 1500 in 2x51 cycle paired-end mode.

ATAC-seq preprocessing and alignment. ATAC sequencing reads (pair-end 51nt 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 use an "intersect and rescue" approach. This approach consisted of intersecting each time point and reprogramming media 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 (FC) and at least 3 FC in the other replicate. This created two intersection peaksets (major to 5 FC in replicate 1 and major to 3 FC in replicate 2 and *vice versa*), which were then combined and merged with bedtools merge (a minimum of 1 bp overlap). The union peakset of both replicates for each timepoint and reprogramming media 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 above. Sequencing read counts for each biological replicate time point and media 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 normalised. Genome coverage plots were generated using wiggleplotr bioconductor package[57]. Principal component analysis (PCA) was then performed on the log2-transformed FPKM on all features using the *prcomp_irlba* function in the irbla package (v2.3.3). Human *in vivo* ICM and hESCs samples from Wu et al[58], human blastocyst-derived TSCs (BT5) from Dong et al[30] were processed as described above. We noted that RE of the fibroblast marker *ANPEP* became less accessible by day 7, accompanied by the downregulation of *ANPEP* gene expression. In contrast, This is followed by a gain of chromatin accessibility of regulatory elements and/or promoters RE 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 or within the gene body of naive pluripotency factors such as *KLF17*, *ZNF729*, *NANOG* and *POU5F1* (*OCT4*)

as were 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 (ICM)[58], became gradually accessible up to day 7 whilst the cells were still in FM. Following this, these regions lost accessibility in the primed intermediates and hiPSCs, 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, 10 bp around the TSS of each gene and assigned these peaks to the corresponding gene. Next, we further integrate the two assays by performing upper quartile normalisation, which makes the transcript counts and peak intensities distributions comparable and the *removeBatchEffect* command in the limma package (v3.38.3) to the combined $\log_2$ transformed ($\log_2 + 1$) ATAC/RNA dataset, specifying that the terminal timepoints, namely Fibroblast-D0, Primed, t2iLGoY, to be preserved using the design argument. PCA was then performed on this integrated dataset using the top 1000 most highly variable genes. To characterise gene expression of genes associated with identified cluster peaks (see details below); 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 above, log2 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 normalised. Only peaks with a coefficient of variation across timepoints 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 sn/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/timepoints were included. Each sample is then subjected to snRNA-seq. The media/timepoints (D: day, P: passage, fm: fibroblast medium, pr: primed reprogramming medium, nr: naive reprogramming medium) 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), P20-nr (n=1). For the media-resolved scRNA-seq experiments, a total of n=9 biologically independent samples across 9 media/timepoints were included. The media/timepoints 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), P10-pr (n=1), P10-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), PR library (D0-fm, D3-fm, D7-fm, D13-pr, D21-pr, P10-pr samples), NR library (D0-fm, D3-fm, D7-fm, D13-nr, D21-nr, P10-nr samples). The total number of cells used in the final analysis was 43,791 (Fig. 1b-g, 3a and Extended Data Fig. 7g,h,8j). Detailed cell numbers for sn and scRNA-seq in each figure are as follows. Fig. 1b and 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 PR cells, 2,286 NR cells); Fig. 1f and 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, 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-interm. cells, 10,563 naive cells, 14,820 nonReprog1 cells, 174 nonReprog2

cells); Extended Data Fig. 1h: 43,791 cells across 3 cell cycle phases (18,771 G1 cells, 12,090 S cells, 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, 107 snRNA-re6 cells); Extended Data Fig. 2e,h: For gene expression trends, the normalised gene expression was averaged across all cells within the same cluster prior to log transformation; Extended Data Fig. 2f-h: Pairwise DEGs 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 LFC (averaged across all 20 pairwise differential expressions) of > 1.5 and (ii) are differentially expressed (LFC > 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-interm. genes, 58 nonReprog1 genes, 175 nonReprog2 genes). For scRNA-seq of RSeT reprogramming, specific library information can be found in Extended Data Fig. 4a and Supplementary Table 6. On top of the scRNA-seq experiments mentioned earlier, an additional n=3 biological independent samples across 3 timepoints were included, namely D13-rr (rr: RSeT reprogramming), D21-rr, P10-rr. These samples are then pooled into the RR library containing the D0-fm, D3-fm, D7-fm, D13-rr, D21-rr, P10-rr samples. The total number of cells used in the final analysis (which included cells from the FM, PR and NR libraries mentioned above) 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 PR cells, 2,286 NR cells, 2,658 RR 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 is then subjected to scRNA-seq. The conditions are D21 fibroblast medium (D21fm, n=1), D21 naive reprogramming (D21nr, n=1), D21 TSC reprogramming (D21tr, n=1). The total number of cells used in the final analysis was 10,518 (Fig. 4a,b and 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, 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, 1,962 D21tr3 cells); Extended Fig 9b and Extended Data Fig. 9d: The marked D21tr1 containing 301 cells comprises 6 D21fm cells, 16 D21nr cells, 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 and Extended Data Fig. 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 prior to log2 + 1 transformation (Fig. 2b and Extended Data Fig. 3g, 6f, 7d-f). Gene expression trends of genes associated with ATAC-seq peaks are shown as z-standardised values (Extended Data Fig. 6b,c). In Extended Data Fig. 7e, gene set enrichment analysis was then performed on the log 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 normalised enrichment score (NES) and -log10(FDR) [NES * -log10(FDR)] is then plotted in the heatmap in Extended Data Fig. 7e. For bulk RNA-seq of iTSC-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 Fig. 8b, 10b,i,j). For the Spearman correlation comparisons, the FPKM values were averaged across replicates prior to log2 transformation (Fig. 4h and 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), log2 transformed (log2 + 1) and quantile normalised (Fig. 2c and 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), log2 transformed (log2 + 1) and quantile normalised. 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, 12024; C2, 7779; C3, 5077; C 4, 3334; C5, 9117; C6, 10129; C7, 4885; C8, 7739 (Fig. 2d). Cluster specific peak trends are shown as the mean +/- SD for each reprogramming media (Fig. 2d). P-values of motif enrichment analysis of cluster specific peaks are calculated based on 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 above (Extended Data Fig. 6b,c). In Fig. 2f, for *TFAP2C* KD 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 hiPSCs. 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* KD 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 hiPSCs. 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 n=4 biological replicates (4 independent experiments from two donors) with similar results and representative images were shown in the figures. For Fig. 3f, n=3 biological replicates, 3 independent iTSC cell lines were injected into three mice, and similar results were obtained, and representative results were shown in the figure. For Fig. 3g, 4 lesions were generated from iTSC lines, harvested 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 were shown in the figures. For Fig. 4f-g, the experiments were repeated with 4 iTSC cell lines obtained from the two donors were independently differentiated into STs and EVTs) with similar results and representative images were 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 were shown in the figure (n=4 biological replicates x 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 iTSC lines injected, 1 line per mouse)) were measured in 2 technical replicates (n=2 biological replicates x 2 technical replicates). Representative results were shown in the figure. For Fig. 4m, 4 lesions were generated, harvested 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 were 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 were shown in the figures (n=4 biological replicates). For Extended Data Fig. 3e, n=2 biological replicates (from two donors) were used for analysis in this figure. For Extended Data Fig. 6g, the relative expression of *TFAP2C* and *GATA2* were measured in n=2 independent experiments with technical replicates. Representative results were 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 were shown in the figures. For Extended Data Fig. 8c, these experiments were repeated n=4 biological replicates (4 independent experiments from two donors) with similar results and representative images were shown in the figures. 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 iTSC lines (obtained from two different donors) with similar results and representative results were 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 such tests, and representative results were 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 x 2 technical replicates). Representative results were shown in the figure.

For Extended Data Fig. 8g, the serum of two independent experiments (2 iTSC lines injected, 1 line per mouse)) were measured in 2 technical replicates (n= 2 biological replicates x 2 technical replicates). Representative results were shown in the figure. For Extended Data Fig. 8h, 4 lesions were generated, harvested and analysed (n=4 biological replicates). For Extended Data Fig. 8i, 4 lesions were generated from iTSC lines, harvested 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 were 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 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 were 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 iTSC lines (obtained from two different donors) with similar results and representative results were 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 were 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 x 2 technical replicates). Representative results were 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 were shown in the figure. For Extended Data Fig. 10m, 4 lesions were generated from iTSC lines, harvested and analysed (n=4 biological replicates). For Supplementary Table 7, n=2 biological replicates (from two donors) were used for data analysis presented in this supplementary table. GO Enrichment p-values are calculated based on an accumulative hypergeometric distribution, and adjusted for multiple testing (q-values) using Benjamini-

Hochberg adjustment. For Supplementary Table 8, n=2 biological replicates (from two donors) were used in this supplementary table. Motif enrichment P-values are calculated based on a cumulative binomial distribution. As described in Heinz S., et al[61], the statistics assess the occurrence of motifs in target sequences vs 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 the Methods section.

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 easy exploration of the dataset and download of all processed datasets. Raw and processed next generation sequencing datasets were 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 induced trophoblast stem cells; GSE149694: bulk RNA-seq experiments of intermediates during human primed and naive reprogramming; GSE150616: bulk RNA-seq experiments of induced trophoblast stem cells and their derived placenta subtypes. Source Data for four Figures and ten Extended Data Figures are provided within the online content of this paper.

Code availability

All data were analysed with standard programs and packages as detailed above. 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. *Protocol Exchange* (2017) doi:10.1038/protex.2017.099.

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. doi:10.1101/604751.

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., Corbett, S. E., Koga, Y., Wang, Z. & Johnson, W. E. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *BioRxiv* (2019).

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 Hif1α. *bioRxiv* (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.journal* **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. *figshare* (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).

Extended Data Fig. Legends

Extended Data Fig. 1 | Experimental designs, analysis pipelines for single-nucleus and single-cell RNA-sequencing. a, Morphological changes of cells undergoing reprogramming in FM: Fibroblasts Medium; PR: Primed Reprogramming; NR: Naive Reprogramming. (FM: D0, 3, 7), PR (D13, D21, hiPSCs) and NR (D13, D21, hiPSCs), n >10, Scale bar, 500μm. **b,** Immunostaining at early stages (FM: D0, 3, 7), during PR (D13, D21) and NR (D13, D21) with TRA-1-60 for primed colonies, KLF17 for naive colonies and DAPI for nuclei staining, n=2. Scale bar, 50μm. **c,** Experimental design for single-cell RNA-seq (scRNA-seq) libraries. FM(scRNA-seq)/fm(snRNA-seq): Fibroblasts Medium; PR/pr: Primed Reprogramming; NR/nr: Naive Reprogramming; iMEF: irradiated Mouse Embryonic Fibroblasts. **d,** Single-nucleus (sn) and single-cell (sc) RNA-seq data analysis strategy (see Methods for details). **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 timepoint. **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 sn/scRNA-seq data. For more details on sample number and statistics, please see statistics and reproducibility section.

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 intermediates 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 timepoint/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,** Dotplot 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 heatmap 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 number and statistics, please see statistics and reproducibility section.

Extended Data Fig. 3 | Isolation and characterisation 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 ReST culture conditions (see Methods), $n=4$. Scale bar, 500µm. **e,** Visualisation 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 (see Methods), $n \geq 2$. **g,** Heatmap showing gene expression profiles of primed and naive pluripotency signatures genes (defined in sn/scRNA-seq analysis) across reprogramming intermediates and hiPSCs derived under all different culture conditions, $n \geq 2$. For more details on sample number and statistics, please see statistics and reproducibility section.

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, naive RSeT scRNA-seq libraries (9,852 cells, see Methods). **b,** Expression profile of genes associated with human fibroblasts (*ANPEP*), shared pluripotency (*NANOG*), primed pluripotency (*ZIC2*, *NLGN4X*) and naive pluripotency (*DNMT3L*, *DPPA5*) on FDL. For more details on sample number and statistics, please see statistics and reproducibility section.

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 vs PC3 related to Fig. 2c. ATAC-seq was performed using isolated reprogramming intermediates and hiPSCs from FM (D0, D3, D7), PR (D13, D21, P3, P10), NR (D13, D21, P3, 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* (*OCT4*), naive-reprograming-specific *KLF17* and *ZNF729* in primed and naive reprogramming intermediates and hiPSCs compared to human ICM and primed hESCs ATAC-seq data[58]. For more details on sample number, please see statistics and reproducibility section. For more details on sample number and statistics, please see statistics and reproducibility section.

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* ≥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 (see Methods). **d,** TF motif enrichment analysis of the ATAC-seq peak clusters. Motif enrichment (-log*P* value) heatmap by colour and the size the percentage (%) of sequences in the cluster featuring the motif. Red arrow points to *OCT4*/*SOX2*/*NANOG*/*KLF4* motifs in transient ATAC-seq cluster (C3), Blue arrow=and enrichment of TE-associated TFs *TFAP2C*/*GATA2* (C7 and C8) are indicated by blue arrows. **e,** Gene expression heatmap TFs identified in the motif enrichment analysis in **d**. **f,** *TFAP2C* and *GATA2* gene expression during primed and naive reprogramming. **g,** qRT-PCR analysis of sh*TFAP2C* and sh*GATA2* compared to scrambled controls, *n*=2. For more details on sample number and statistics, please see statistics and reproducibility section.

Extended Data Fig. 7 | Uncovering the transcriptional programs 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, TE gene signatures[25] from embryonic (E) day 5, 6, 7 on intermediates and hiPSCs reprogrammed under primed and different naive culture conditions (see Methods). **e,** Gene set enrichment analysis (GSEA, see methods) of the EPI, PE and TE gene signatures in reprogramming intermediates and hiPSCs reprogrammed under primed and several naive culture conditions. **f,** EPI, PE and TE gene signatures scores in reprogramming intermediates and hiPSCs reprogrammed under primed and several naive culture conditions. We used a combined gene signature across E5 to E7 for each lineage (see 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 number and statistics, please see statistics and reproducibility section.

Extended Data Fig. 8 | Characterisation of iTSC^{d21n}. a, Immunostaining of fibroblast, primed, naive t2iLGoY hiPSCs with P63, TFAP2C, GATA2, KRT7, *n*=2. Scale bar, 100µm. **b,** Gene expression of trophoblast genes in fibroblasts, primed, naive t2iLGoY hiPSCs, iTSC^{d21n} and 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, 100μm. **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 OTC hCG pregnancy test for media of ST cells differentiated from $iTSC^{d21n}$ and control media, *n*=6. **f,** hCG levels in $iTSC^{d21n}$ and $iTSC^{d21n}$-ST conditioned media, detected by ELISA, *n*=4. **g,** hCG level in mouse blood serum detected by ELISA, *n*=4. **h,** Lesions harvested from subcutaneously engrafted $iTSC^{d21n}$ in NOD-SCID mice, *n*=4. **i,** Hematoxylin and eosin, and immunohistochemical staining of KRT7 in the lesions from **h**, no evident lesions were observed in vehicle controls, *n*=4. Scale bar, 200μm. **j,** Distinct level of CD70 expression in naive and TE populations (indicated by blue arrows) on FDL projection of sn/scRNA-seq datasets. **k,** Quantification of KRT7+ colony clusters after 9 days of transitioning into TSC media of unenriched, CD70 high and CD70 low populations, *n*=2-3 independent experiments, data are represented as mean ± s.e.m., *p* values by two-tailed unpaired Student's *t*-test. Representative images of whole-well scans (top panels, scale bar, 1mm) and KRT7 immunostaining (bottom panels, scale bar, 100μm). For more details on sample number and statistics, please see statistics and reproducibility section.

Extended Data Fig. 9 | Cellular heterogeneity of fibroblast and $iTSC^{d8}$ reprogramming intermediates revealed by scRNA-seq. a, Experimental designs and preparation of single-cell RNA-seq (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, and **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 intermediates scRNA-seq libraries (upper 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 hiPSCs and iTSCs. **g,** Phase-contrast images of primed, naive hiPSCs and iTSCs

generated from day 21 fibroblast reprogramming intermediates, *n*=2. Scale bar, 50µm. Immunostaining of primed, naive hiPSCs and iTSCs with NANOG, KLF17, NR2F2, KRT7 and DAPI for nuclei staining, *n*=2. Scale bar, 200µm. **h,** qRT-PCR analysis of *NANOG*, *ZIC2*, *KLF17*, *DPPA3*, *GATA2*, *KRT7* expression in primed, naive hiPSCs and iTSCs generated from day 21 fibroblast reprogramming intermediates, *n*=3. Data are represented as mean ± s.e.m. For more details on sample number and statistics, please see statistics and reproducibility section.

Extended Data Fig. 10 | Characterisation of iTSC^{d8}. a, Sendai viral transgenes in iTSC lines with positive and negative controls, *n*=6. **b,** Gene expression of trophoblast genes in fibroblasts, primed hiPSCs, naive t2iLGoY hiPSCs, 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 represented as mean ± s.e.m., *p* values by two-tailed unpaired Student's *t*-test. **d,** Representative results for hCG pregnancy test obtained from media of ST cells differentiated from iTSC^{d8}, *n*=6. **e,** hCG levels of iTSC^{d8} and iTSC^{d8}-ST conditioned media 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 hiPSCs, naive t2iLGoY hiPSCs, iTSC^{d8} and iTSC^{d21n}, *n*=4. **i,** Expression of ST genes in iTSC^{d8} and iTSC^{d21n}-derived ST cells and **j,** expression of EVT genes in iTSC^{d8} and iTSC^{d21n}-derived EVT cells. **k,** Spearman correlation of the transcriptomes of fibroblast, primed and naive t2iLGoY hiPSCs, 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 Haider et al.[29] and Turco et al.[28] and **l,** Single-cell fetal-maternal interface samples from Vento-Tormo et al[27], $n \geq 2$, replicates are averaged prior to performing correlation. **m,** Lesions harvested from subcutaneously engrafted iTSC^{d8} in NOD-SCID mice, *n*=4. For more details on sample number and statistics, please see statistics and reproducibility section.

Acknowledgments

The authors thank the staff at Monash Flowcore Facility for providing high-quality cell sorting service and technical input. The authors acknowledge the use of the services and facilities of Micromon, Monash Micro Imaging, and Monash Histology Platforms at Monash University. Furthermore, the authors thank Dr. Sen Wang and Dr. Trevor Wilson at the ACRF Centre for Cancer Genomic Medicine at the MHTP Medical Genomics Facility, and the University of Melbourne Centre for Cancer Research (UMCCR) core for assistance with next-generation library preparation and Illumina sequencing. We also thank Jess Hatwell-Humble for assistance with the mouse work. We thank A. Purcell (Monash University) for providing the HLA antibodies. This work was supported by National Health and Medical Research Council (NHMRC) project grants APP1104560 to J.M.Polo and A.L.L., 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 cells analysis pipeline and data integration. X.L. performed reprogramming experiments, collection and isolation of single cells, intermediates and functional validation of iTSC experiments with support from C.M.N., J.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 single nucleus-RNA-seq experiments with support from A.L.; 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.T., G.S.; J.P. helped with ATAC-seq experiments.

H.S.C., C.M.B., and A.L.L. provided reagents and technical assistance. H.N. and D.R.P helped with bulk RNA-seq analysis. J.F.O. performed the sn/scRNA-seq and bulk RNA-seq analyses for the human reprogramming intermediates and iTSC 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

Although not directly related to this manuscript, O.J.L.R. and J.M.Polo. are co-inventors of the 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 iTSCs. The other authors declare no competing interests.

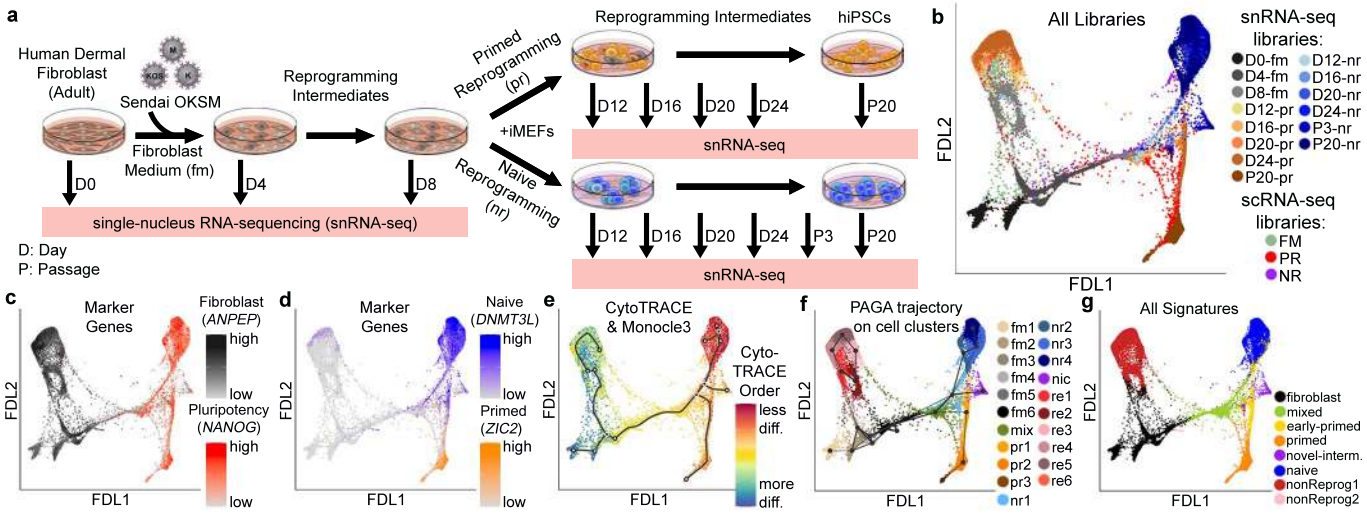
Additional information

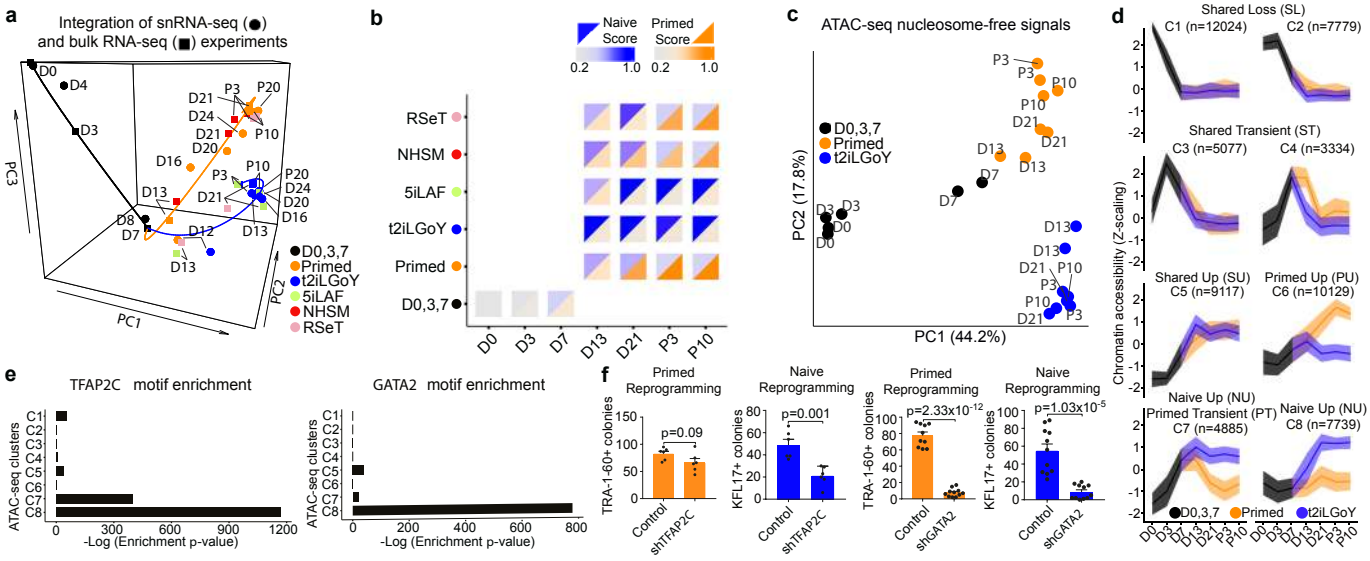
Supplementary information is available for this paper.

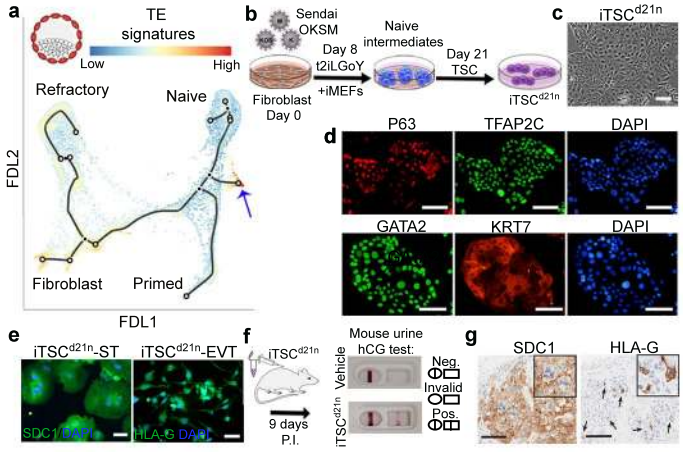
Correspondence and requests for materials should be addressed to O.J.L.R. or J.M.Polo.

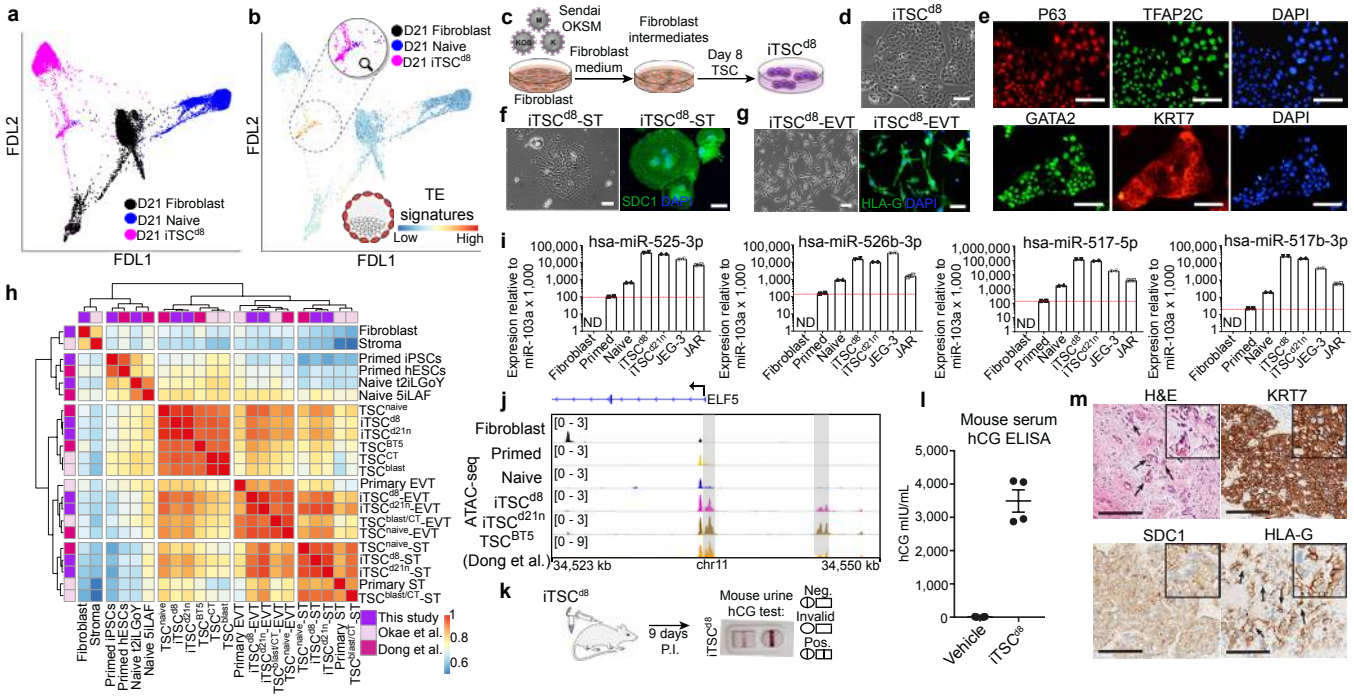
Peer review information *Nature* thanks

Reprints and permissions information is available at http://www.nature.com/reprints.


a
Human Dermal Fibroblast (Adult)
Sendai OKSM
Fibroblast Medium (fm)
Reprogramming Intermediates
D0
D4
D8
single-nucleus RNA-sequencing (snRNA-seq)
D: Day
P: Passage
Primed Reprogramming (pr)
+iMEFs
Naive Reprogramming (nr)
Reprogramming Intermediates
hiPSCs
D12
D16
D20
D24
P20
snRNA-seq
P3
b
All Libraries
FDL2
FDL1
snRNA-seq libraries:
D0-fm
D4-fm
D8-fm
D12-pr
D16-pr
D20-pr
D24-pr
P20-pr
D12-nr
D16-nr
D20-nr
D24-nr
P3-nr
P20-nr
scRNA-seq libraries:
FM
PR
NR
c
Marker Genes
Fibroblast (ANPEP)
high
low
Pluripotency (NANOG)
high
low
d
Marker Genes
Naive (DNMT3L)
high
low
Primed (ZIC2)
high
low
e
CytoTRACE & Monocle3
Cyto-TRACE Order
less diff.
more diff.
f
PAGA trajectory on cell clusters
fm1
fm2
fm3
fm4
fm5
fm6
mix
pr1
pr2
pr3
nr1
nr2
nr3
nr4
nic
re1
re2
re3
re4
re5
re6
g
All Signatures
fibroblast
mixed
early-primed
primed
novel-interm.
naive
nonReprog1
nonReprog2


a
Integration of snRNA-seq (●) and bulk RNA-seq (■) experiments
PC1
PC2
PC3
D0,3,7
Primed
t2iLGoY
5iLAF
NHSM
RSeT
b
Naive Score
Primed Score
0.2
1.0
RSeT
NHSM
5iLAF
t2iLGoY
Primed
D0,3,7
c
ATAC-seq nucleosome-free signals
PC1 (44.2%)
PC2 (17.8%)
d
Shared Loss (SL)
C1 (n=12024)
C2 (n=7779)
Shared Transient (ST)
C3 (n=5077)
C4 (n=3334)
Shared Up (SU)
C5 (n=9117)
Primed Up (PU)
C6 (n=10129)
Naive Up (NU)
Primed Transient (PT)
C7 (n=4885)
Naive Up (NU)
C8 (n=7739)
Chromatin accessibility (Z-scaling)
e
TFAP2C motif enrichment
GATA2 motif enrichment
ATAC-seq clusters
-Log (Enrichment p-value)
f
Primed Reprogramming
Naive Reprogramming
TRA-1-60+ colonies
KFL17+ colonies
p=0.09
p=0.001
p=2.33x10-12
p=1.03x10-5
Control
shTFAP2C
shGATA2


a
TE signatures
Low
High
Refractory
Naive
Fibroblast
Primed
FDL2
FDL1
b
Sendai
OKSM
Fibroblast
Day 0
Day 8
t2iLGoY
+iMEFs
Naive intermediates
Day 21
TSC
iTSCd21n
c
iTSCd21n
d
P63
TFAP2C
DAPI
GATA2
KRT7
DAPI
e
iTSCd21n-ST
iTSCd21n-EVT
SDC1/DAPI
HLA-G DAPI
f
iTSCd21n
9 days
P.I.
Mouse urine
hCG test:
Vehicle
iTSCd21n
Neg.
Invalid
Pos.
g
SDC1
HLA-G


a
FDL2
FDL1
D21 Fibroblast
D21 Naive
D21 iTSC^d8
b
FDL2
FDL1
D21 Fibroblast
D21 Naive
D21 iTSC^d8
TE signatures
Low
High
c
Sendai OKSM
Fibroblast
Fibroblast medium
Fibroblast intermediates
Day 8 TSC
iTSC^d8
d
iTSC^d8
e
P63
TFAP2C
DAPI
GATA2
KRT7
DAPI
f
iTSC^d8-ST
iTSC^d8-ST
SDC1 DAPI
g
iTSC^d8-EVT
iTSC^d8-EVT
HLA-G DAPI
h
Fibroblast
Stroma
Primed iPSCs
Primed hESCs
Naive t2iLGoY
Naive 5iLAF
TSC^naive
iTSC^d8
iTSC^d21n
TSC^BT5
TSC^CT
TSC^blast
Primary EVT
iTSC^d8-EVT
iTSC^d21n-EVT
TSC^blast/CT-EVT
TSC^naive-EVT
TSC^naive-ST
iTSC^d8-ST
iTSC^d21n-ST
Primary ST
TSC^blast/CT-ST
This study
Okae et al.
Dong et al.
1
0.8
0.6
i
hsa-miR-525-3p
hsa-miR-526b-3p
hsa-miR-517-5p
hsa-miR-517b-3p
Expresion relative to miR-103a x 1,000
ND
Fibroblast
Primed
Naive
iTSC^d8
iTSC^d21n
JEG-3
JAR
j
ELF5
ATAC-seq
Fibroblast
Primed
Naive
iTSC^d8
iTSC^d21n
TSC^BT5
(Dong et al.)
[0 - 3]
[0 - 9]
34,523 kb
chr11
34,550 kb
k
iTSC^d8
9 days P.I.
Mouse urine hCG test:
Neg.
Invalid
Pos.
l
Mouse serum hCG ELISA
hCG mIU/mL
Vehicle
iTSC^d8
m
H&E
KRT7
SDC1
HLA-G