ARTICLE

https://doi.org/10.1038/s41586-018-0744-4

Single-cell mapping of lineage and identity in direct reprogramming

Brent A. Biddy[1,2,3], Wenjun Kong[1,2,3], Kenji Kamimoto[1,2,3], Chuner Guo[1,2,3], Sarah E. Waye[1,2,3], Tao Sun[1,2,3,4] & Samantha A. Morris[1,2,3]*

Direct lineage reprogramming involves the conversion of cellular identity. Single-cell technologies are useful for deconstructing the considerable heterogeneity that emerges during lineage conversion. However, lineage relationships are typically lost during cell processing, complicating trajectory reconstruction. Here we present 'CellTagging', a combinatorial cell-indexing methodology that enables parallel capture of clonal history and cell identity, in which sequential rounds of cell labelling enable the construction of multi-level lineage trees. CellTagging and longitudinal tracking of fibroblast to induced endoderm progenitor reprogramming reveals two distinct trajectories: one leading to successfully reprogrammed cells, and one leading to a 'dead-end' state, paths determined in the earliest stages of lineage conversion. We find that expression of a putative methyltransferase, *Mettl7a1*, is associated with the successful reprogramming trajectory; adding Mettl7a1 to the reprogramming cocktail increases the yield of induced endoderm progenitors. Together, these results demonstrate the utility of our lineage-tracing method for revealing the dynamics of direct reprogramming.

Direct lineage reprogramming bypasses pluripotency to convert cell identity between somatic states, yielding clinically valuable cell types[1]. However, these conversion strategies are generally inefficient, producing incompletely converted and developmentally immature cells that fail to fully recapitulate target cell identity[2,3]. The considerable heterogeneity that arises during reprogramming has hindered the study of the molecular mechanisms underlying lineage conversion. Single-cell RNA-sequencing analysis (scRNA-seq) has enabled fully converted cells to be distinguished from partially reprogrammed intermediates[4,5], although these analytical approaches typically result in the loss of spatial, temporal and lineage information. Elegant computational approaches can infer missing observations[6,7], but reconstruction of true reprogramming trajectories using these tools remains challenging. Although sophisticated lineage tracing solutions to connect cell history with fate are emerging, these protocols are either not compatible with high-throughput scRNA-seq[8–11], or require genome editing strategies that are not readily deployed in some systems[12–15].

To enable simultaneous single-cell profiling of cell identity and clonal history, we have developed 'CellTagging', a straightforward, high-throughput cell tracking method. Sequential lentiviral delivery of CellTags (heritable random barcodes) enables the construction of multi-level lineage trees. Here, we apply CellTagging to transcription factor-induced direct lineage reprogramming of mouse embryonic fibroblasts (MEFs) to induced endoderm progenitors (iEPs), a self-renewing cell type that has both hepatic and intestinal potential[3,16]. Generation of iEPs represents a prototypical cell fate engineering methodology, reflecting the inefficiency and infidelity of many reprogramming protocols[2,3]. CellTagging and tracking more than 100,000 cells during conversion to iEPs reveals two distinct trajectories: a route towards successfully reprogrammed cells, and an alternate path to a putative 'dead-end' state, marked by re-expression of fibroblast genes. Although few cells are successfully reprogrammed, clonally related cells tend to follow the same trajectories, suggesting that their reprogramming outcome may be determined from the earliest stages of lineage conversion. These clonal dynamics and lineages can be explored on our companion website, CellTag Viz (http://www.celltag.org/). In later stages of conversion, our analyses reveal expression of a putative methyltransferase, *Mettl7a1*, along the successful reprogramming trajectory. Adding this factor to the reprogramming cocktail increases the yield of successfully converted iEPs. Together, these findings demonstrate the utility of CellTagging for lineage reconstruction, providing molecular insights into reprogramming that serve to improve the outcome of this generally inefficient process.

Combinatorial indexing of cells to track clonal history

CellTagging is a lentivirus-based approach to uniquely label individual cells with heritable barcode combinations. CellTags are highly expressed and readily captured within each single-cell transcriptome, enabling recording of clonal history over time, in parallel with cell identity (Fig. 1a). Recovery of CellTag expression, followed by filtering and error correction, ensures sensitive and specific identification of clonally related cells (Extended Data Fig. 1a–g). The efficacy of this barcoding approach is demonstrated by CellTagging a 'species mix' of genetically distinct human 293T cells and MEFs (Extended Data Fig. 1h–j). This is further supported by labelling two independent biological replicates with the same CellTag library: whereas individual CellTags appear in both pools of cells, no combinatorial signatures of 2 or more CellTags are shared between replicates, confirming that clones are derived from distinctly labelled cells ($n = 4{,}141$ cells expressing 3.0000 ± 0.0004 (mean ± s.e.m.) unique CellTags per cell, Fig. 1b, c). Finally, CellTagging does not perturb cell physiology or reprogramming efficiency (Extended Data Fig. 2). Together, these data validate the utility of CellTagging to deliver unique, heritable labels into cells, permitting clonal relationships to be tracked longitudinally, with a high degree of confidence.

We next applied CellTagging to the direct reprogramming of fibroblasts to iEPs, driven by retroviral overexpression of the transcription factors FOXA1 and HNF4α (encoded by *Foxa1* and *Hnf4a*, respectively) in four independent biological replicates. To enable lineage reconstruction, we devised a sequential CellTagging scheme in which

[1]Department of Developmental Biology, Washington University School of Medicine in St Louis, St Louis, MO, USA. [2]Department of Genetics, Washington University School of Medicine in St Louis, St Louis, MO, USA. [3]Center of Regenerative Medicine, Washington University School of Medicine in St Louis, St Louis, MO, USA. [4]Present address: Nanomedicine Research Center, Department of Neurosurgery, Cedars–Sinai Medical Center, Los Angeles, CA, USA. *e-mail: s.morris@wustl.edu

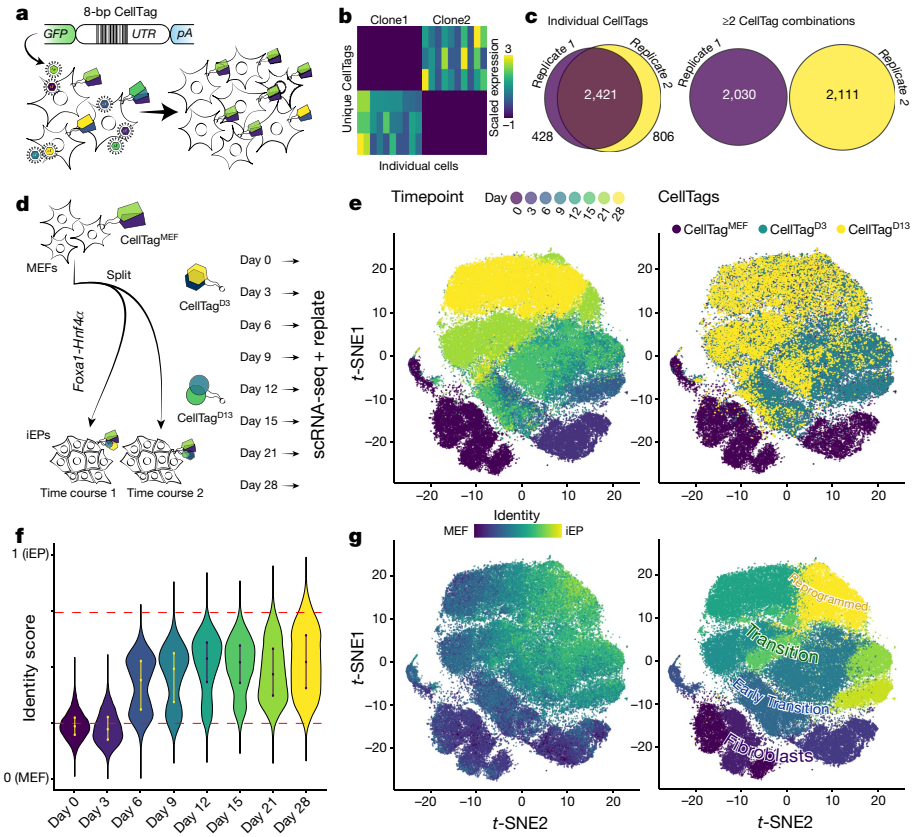


Fig. 1 | CellTagging: clonal tracking applied to reprogramming. a, The CellTagging workflow: a lentiviral construct contains an 8-bp random CellTag barcode in the 3′ untranslated region (UTR) of GFP, followed by an SV40 polyadenylation signal. Transduced cells express unique combinations of CellTags, resulting in distinct, heritable signatures, enabling tracking of clonally related cells. **b**, Representative CellTag expression in two clones, defined by unique combinations of three CellTags ($n = 10$ cells per clone). **c**, Left, overlap of individual CellTags in two independent biological replicates tagged with the same CellTag library. Right, CellTag signatures are not shared between the two replicates (replicate 1, $n = 8{,}535$ cells; replicate 2, $n = 11{,}997$ cells). **d**, Experimental approach: MEFs are tagged with the CellTag^{MEF} library, expanded for two days and then split for cell fate reprogramming in two independent biological replicates. Additional CellTagging was performed at 3 days (CellTag^{D3}) and 13 days (CellTag^{D13}) after initiation of reprogramming. Every 3–7 days, a sample of cells was collected for scRNA-seq, and the remaining cells were cultured. **e**, Visualization of scRNA-seq data. Projection of time points and CellTag expression onto a *t*-SNE plot (time courses 1 and 2, $n = 85{,}010$ cells). **f**, Scoring single-cell identity via quadratic programming. Cells scoring >0.75 (upper red line) are classified as iEPs; cells scoring <0.25 (lower red line) are classified as fibroblasts ($n = 85{,}010$ cells). **g**, Left, projection of identity scores onto the *t*-SNE plot. Right, designations of *t*-SNE clusters: fibroblast, early transition, transition and reprogrammed.

fibroblasts were transduced with an initial CellTag library, CellTag^{MEF}. Following a 48-h expansion period, these cells were split into independent biological replicates for reprogramming. Tagging with a second library (CellTag^{D3}) was performed at the end of the 3-day period of transcription factor delivery, followed by a third round (CellTag^{D13}) 13 days after the start of reprogramming, coinciding with the phenotypic emergence of iEPs. After sequencing, CellTags are assigned to rounds by demultiplexing on the basis of a short motif preceding the random CellTag region. Cells were collected every 3–7 days over the 28-day time course. A sample of cells from each time point was fixed in methanol for high-throughput droplet microfluidics-based scRNA-seq (Drop-seq[17] and 10x Genomics[18] platforms), and the remaining cells were replated to enable clonal growth and lineage reconstruction (Fig. 1d). In total, 104,887 single-cell transcriptomes were captured. Downstream analysis focused on data captured using the 10x Genomics platform (85,010 high-quality single-cell transcriptomes, merging time courses 1 and 2; Fig. 1e, Extended Data Fig. 3a–c, Supplementary Table 1). Canonical correlation analysis[19] demonstrates consistent replication across the sequencing technologies and biological replicates (Extended Data Fig. 3d, e).

Parallel capture of reprogramming and clonal dynamics

Using *t*-distributed stochastic neighbour embedding[6] (*t*-SNE), the 28-day reprogramming process resolves into 13 clusters of transcriptionally distinct cells (Extended Data Figs. 3f, g, 5a). CellTag expression is detected in 99% of cells, and CellTag^{MEF} expression is detected across all time points, CellTag^{D3} is detected from day 6, and CellTag^{D13} is detected from day 15 (Fig. 1e). Of 85,010 sequenced cells, 55,571 (65%) passed the threshold of at least two CellTags per cell that is required for tracking (Extended Data Fig. 4). To investigate dynamics of reprogramming, we first analysed gene expression for each cluster, revealing progressive silencing of fibroblast identity (Extended Data Fig. 5a, b, Supplementary Tables 2, 3). To track emergence of iEPs, we used quadratic programming[5] to score individual cell identities as a fraction of starting and target cell types, revealing that iEP identity is progressively gained from day 6 of reprogramming. Projection of identity scores onto the *t*-SNE plot localizes iEPs to cluster 2, coinciding with reprogramming days 21 and 28 (Fig. 1f, g). Further examination of this iEP-containing cluster identifies new markers, including apolipoprotein A1 (APOA1, encoded by *Apoa1*; Extended Data Fig. 5a, b, Supplementary Table 3). Immunostaining for APOA1 demonstrates protein-level co-expression with the canonical iEP marker E-cadherin (CDH1)[3,16] (Extended Data Fig. 5c–e). Although previous studies show that only around 1% of cells are successfully reprogrammed[3,16], we observe a high proportion of cells expressing *Apoa1*, beginning from day 6 ($62.5 \pm 5.5\%$; Extended Data Fig. 5b, d, e). Together, these observations suggest that many cells initiate reprogramming but few complete the transition to iEPs. Using expression of these markers,

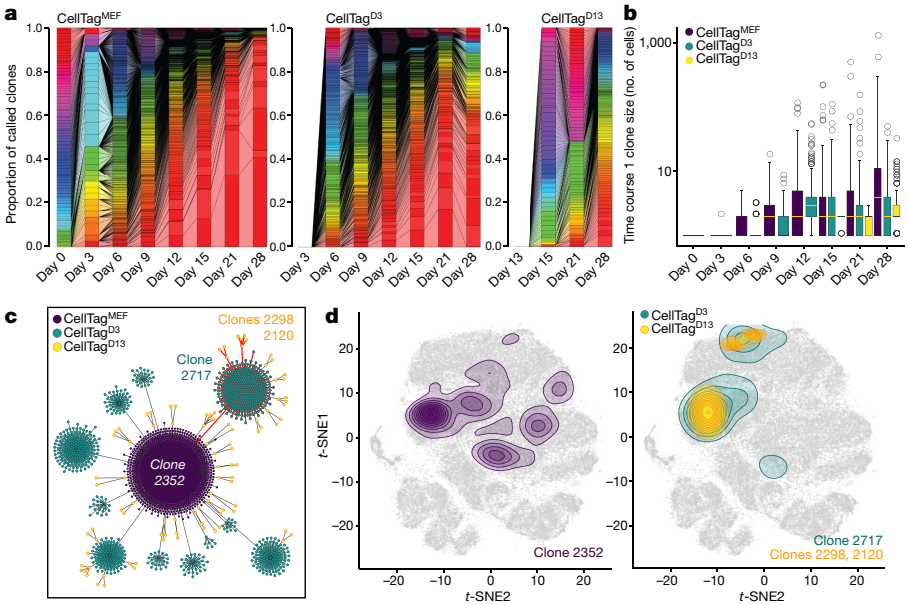


Fig. 2 | Tracking clonal dynamics of reprogramming and constructing lineage trees. a, Connected bar plots showing individual clones as a proportion of all clones during reprogramming, for each CellTagging round (time course 1, $n = 12{,}932$ cells, 1,031 clones). **b**, Mean number of cells per clone, per time point, for each round of CellTagging ($n = 1{,}031$ clones). **c**, Reconstruction and visualization of lineages using force-directed graph drawing. Each node represents an individual cell, and edges represent clonal relationships between cells: purple, CellTag^{MEF} clones; blue, CellTag^{D3} clones; yellow, CellTag^{D13} clones. **d**, Contour plots, representing cell density of each clone, projected onto the *t*-SNE for the lineage highlighted in red in **c** ($n = 2{,}199$ cells). All lineages and clone distributions can be explored with CellTag Viz (http://www.celltag.org/).

together with cell identity scores, we broadly partition the process into four phases: fibroblast, early transition, transition and reprogrammed (Fig. 1g, Extended Data Fig. 5b).

We next integrated clonal relationships into this single-cell landscape: from the 55,571 cells passing the threshold to support clone calling, we identified 27,020 cells possessing clonal relatives, on the basis of shared CellTag signatures. Defining a clone as three or more related cells, we identified 706 CellTag^{MEF} clones and 884 CellTag^{D3} clones. Because CellTag^{D13} clones had less time to expand, we also included related cell pairs for this later labelling, resulting in 561 clones (Supplementary Table 4). Consistent with the above validation experiments, examination of 10 major clones (defined as the ten largest clones based on number of cells) based on CellTag^{D3}-labelled replicates shows that the CellTag combinations used to identify clonally related cells were unique (Extended Data Fig. 6a). CellTags are reliably detected over a 10-week period; although their expression gradually diminishes over time, they are not silenced (Extended Data Figs. 4c, 6b–d). This demonstrates the advantage of our CellTag combinatorial indexing method for reliably labelling cells and tracking them over extended periods.

During reprogramming, we observed extensive clonal growth: CellTag^{MEF} clones reached an average size of 47 ± 22 cells per clone by day 28 (Fig. 2a, b, Extended Data Fig. 7a–d). Expanding at a similar rate, CellTag^{D3} clones were first detected from day 6, whereas smaller clones arose from CellTag^{D13}-labelling (Fig. 2a, b). In some instances, we observed rapid expansion of an individual clone during reprogramming (Extended Data Fig. 7d). This could not be reconciled with viral integration analysis (Supplementary Table 5), suggesting that the clonal growth we observed was associated with iEPs entering a self-renewing, progenitor-like state. As a consequence of this rapid expansion, iEPs were derived from only a small number of clones. We next sought to connect these clonal relationships over time, to trace the origins of successfully reprogramming cells. In this approach, we assume that the identity or state of each cell that we capture is representative of its collective clone. We find that gene expression is highly correlated among clonally related cells, suggesting that family members are likely to behave in a similar manner and share reprogramming outcomes (Extended Data Fig. 7e, f).

Lineage and reprogramming trajectory reconstruction

Sequential CellTagging enables the reconstruction of lineage trees and reprogramming trajectories. First, we apply force-directed graphing to construct hundreds of multi-level lineages (Extended Data Fig. 8a, b), which can be explored at http://www.celltag.org/. Figure 2c shows a representative lineage stemming from one CellTag^{MEF} clone, branching into CellTag^{D3} and CellTag^{D13} descendants. Next, to visualize the distribution of clonally related cells, we use contour plotting in combination with the *t*-SNE plot. This reveals considerable overlap of clones belonging to the same lineage, supporting our observation that clonally related cells are transcriptionally similar (Fig. 2d, Extended Data Fig. 8c, d). From these analyses, we observe enrichment or depletion of iEPs within many lineages. To quantify this, we re-clustered cells in the later stages of reprogramming, providing high-coverage clone information. Within this subset, 8% of cells are classified as fully reprogrammed iEPs (Fig. 3a; Extended Data Fig. 9a, b). We then performed randomized testing to identify major clones that were significantly enriched for or depleted of iEPs, yielding 20 iEP-enriched clones in which 20–50% of cells are fully reprogrammed. By contrast, we found 24 iEP-depleted clones in which less than 3% of cells are classified as iEPs (Fig. 3b).

iEP-enriched and iEP-depleted clones are clearly segregated on contour plots, suggesting the existence of discrete reprogramming trajectories; this is also supported by orthogonal pseudotemporal ordering analysis (Fig. 3c, d, Extended Data Fig. 9c, d). Quantification of these trajectories reveals a bifurcation at day 21, when successfully reprogramming clones transition through clusters 6 and 7, leading to the reprogrammed state at day 28. Conversely, these transition clusters are bypassed on the iEP-depleted trajectory, on which clones traverse cluster 4 on day 21, entering a putative reprogramming 'dead-end' by day 28 (Fig. 3e; Pearson's correlation coefficient, $r = -0.84$). To investigate the timing of the commitment to these trajectories, we quantified occupancy of CellTag^{D13}-labelled cells in reprogrammed and putative dead-end states (cluster1 and 3, respectively) at day 28. The distribution of clonally related cells between these states shows that they are restricted to one of the two states, indicating that reprogramming outcome is determined by day 13 (in $88 \pm 8\%$ of restricted clones; Extended Data Fig. 9e). These divergent routes appear to be rooted in distinct transcriptional states

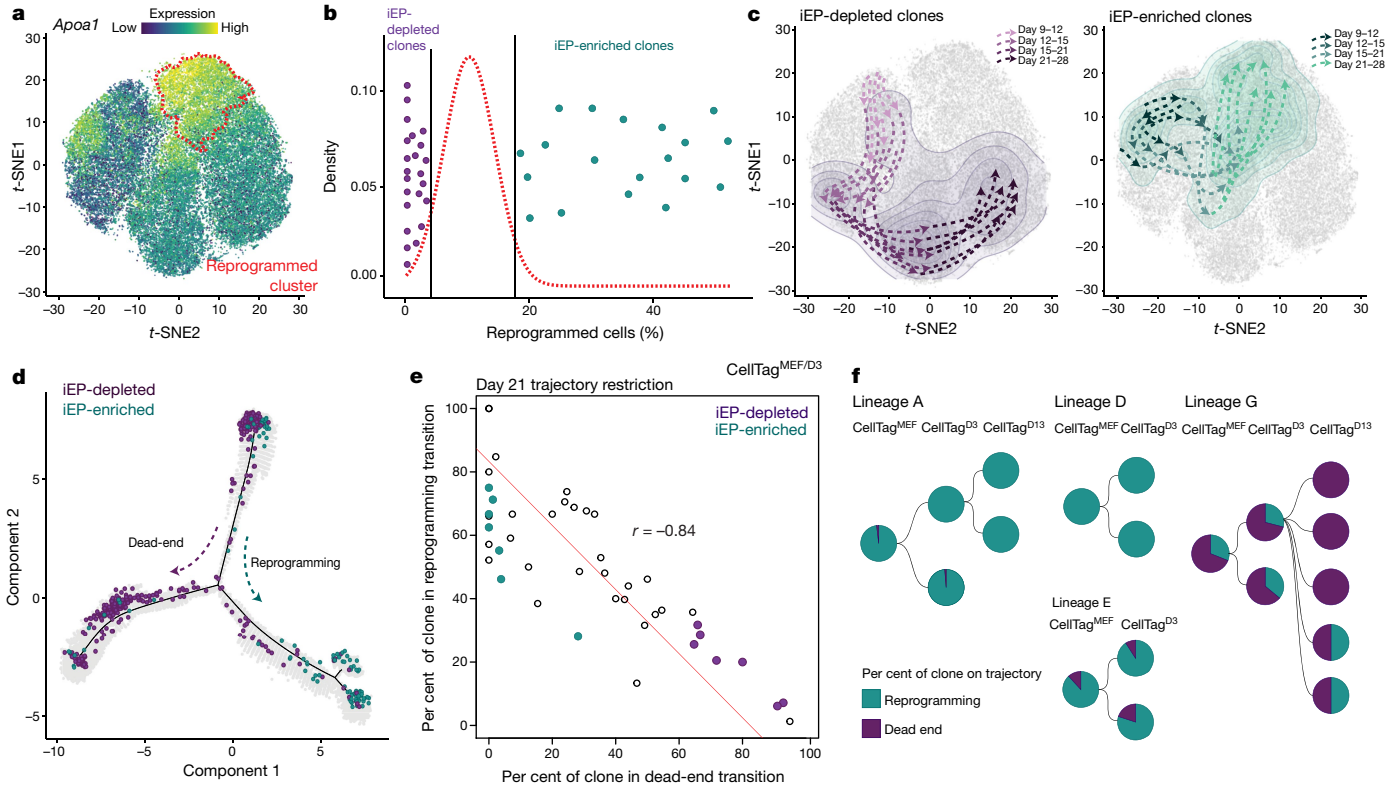


Fig. 3 | Mapping reprogramming trajectories and timing of cell fate commitment. a, *Apoa1* expression in a subset of cells from time courses 1 and 2 ($n = 48{,}515$ cells). Fully reprogrammed iEPs are outlined in red (cluster 1). **b**, Density plot of the mean proportion of reprogrammed cells for groups of randomly selected cells (defined by cluster 1 occupancy; $n = 59$ groups, 10,259 cells). Randomized testing of 59 CellTag$^{MEF/D3}$ clones (≥35 cells per clone, $n = 10{,}259$ cells) identifies iEP-enriched clones ($n = 20$ clones, 6,128 cells; $P < 0.05$) and iEP-depleted clones ($n = 24$ clones, 3,177 cells; $P < 0.05$). **c**, Clones spanning all time points were selected for further analysis. Trajectories showing connections between areas of highest clonal density across each day of reprogramming for iEP-depleted clones (left, $n = 7$ clones, 2,270 cells) and iEP-enriched clones (right, $n = 7$ clones, 1,037 cells). **d**, Pseudotemporal ordering of the time course 1 and 2 subset in **a**, with overlay of individual cells derived from iEP-enriched and iEP-depleted clones, defining reprogramming and dead-end trajectories ($n = 14$ clones, 3,307 cells). **e**, Proportions of clones occupying clusters 6 and 7 (reprogramming transition) or cluster 4 (dead-end transition) at reprogramming day 21 ($r = -0.84$, Pearson's correlation; $n = 44$ clones, 9,305 cells). **f**, Lineage trees of related clones, with the proportion of each clone contributing to reprogramming or dead-end trajectories shown ($n = 1{,}185$ cells).

as early as day 6 (Fig. 3c), suggesting that they are established early in the reprogramming process.

The existence of early-labelled clones that are biased in their reprogramming outcome, in addition to the shared transcriptional signatures that we observe between clonally related cells, suggests that cells do not reprogram in a stochastic manner. Here, sequential CellTagging and quantification of reprogramming outcome for each clone within a lineage allows us to probe the probability with which cells successfully generate iEPs. To study this, we identified lineages of CellTag^{D3}-labelled clones arising from common CellTag^{MEF}-labelled ancestors. For each clone within a lineage, we calculated the proportion of cells occupying reprogramming and dead-end trajectories. In a stochastic model of reprogramming, we would expect the post-reprogramming-induction, CellTag^{D3}-labelled clones from a common ancestor to follow different reprogramming trajectories. However, Fig. 3f shows that CellTag^{D3}-descendant clones reprogram with similar efficiencies to each other, and to their CellTag^{MEF}-labelled parent, particularly for those lineages reprogramming at high efficiency (Pearson's correlation coefficient, $r = 0.71$; Extended Data Fig. 9f). This suggests that reprogramming outcome may be determined at early stages. We considered the possibility that an 'elite' cell type that is predisposed to reprogram exists in the highly heterogeneous fibroblast starting population. To investigate this possibility, cells were first tagged and then split for reprogramming in two biological replicates. We identified 84 clones that appeared across both replicates; only 4 clones reprogrammed in both replicates (Supplementary Table 6), arguing against the existence of an elite reprogramming cell type in the fibroblast population.

Mettl7a1 delineates successful reprogramming

To investigate the molecular characteristics underpinning the distinct reprogramming paths, we compared cells between reprogramming and dead-end trajectories ($n = 2{,}074$ cells). Along the reprogramming trajectory, iEP identity scores gradually increase over time. By contrast, partial fibroblast identity is re-established with progression along the dead-end trajectory, supporting the suggestion that this represents a reprogramming impasse (Fig. 4a). Significant changes in gene expression between these two trajectories are apparent, including key elements of Wnt, Igf2 and HGF signalling pathways. The dead-end trajectory is enriched for imprinted gene expression (*Dlk1* and *Peg3*), in concert with reactivation of fibroblast gene expression and silencing of reprogramming transgenes. Many of these differences in gene expression are evident from day 6, including marked upregulation of *Apoa1* and concomitant downregulation of *Col1a2* on the reprogramming trajectory, supporting our observations that these outcomes are established from early stages. We did not detect significant differences in transgene expression between the two trajectories at these early stages, suggesting that transgene expression level is not a bifurcation driver (Fig. 4b, c, Extended Data Fig. 10a, b, Supplementary Table 7).

Focusing on later stages of reprogramming, we performed differential expression analysis of the trajectory bifurcation at day 21 (Supplementary Table 7). *Mettl7a1*, an as-yet-uncharacterized putative methyltransferase, was transiently and significantly upregulated along the successful reprogramming trajectory (Fig. 4b, c). METTL3, a related methyltransferase-like protein, catalyses N^6-methyladenosine (m^6A) modification of mRNA, and regulates stem-cell differentiation

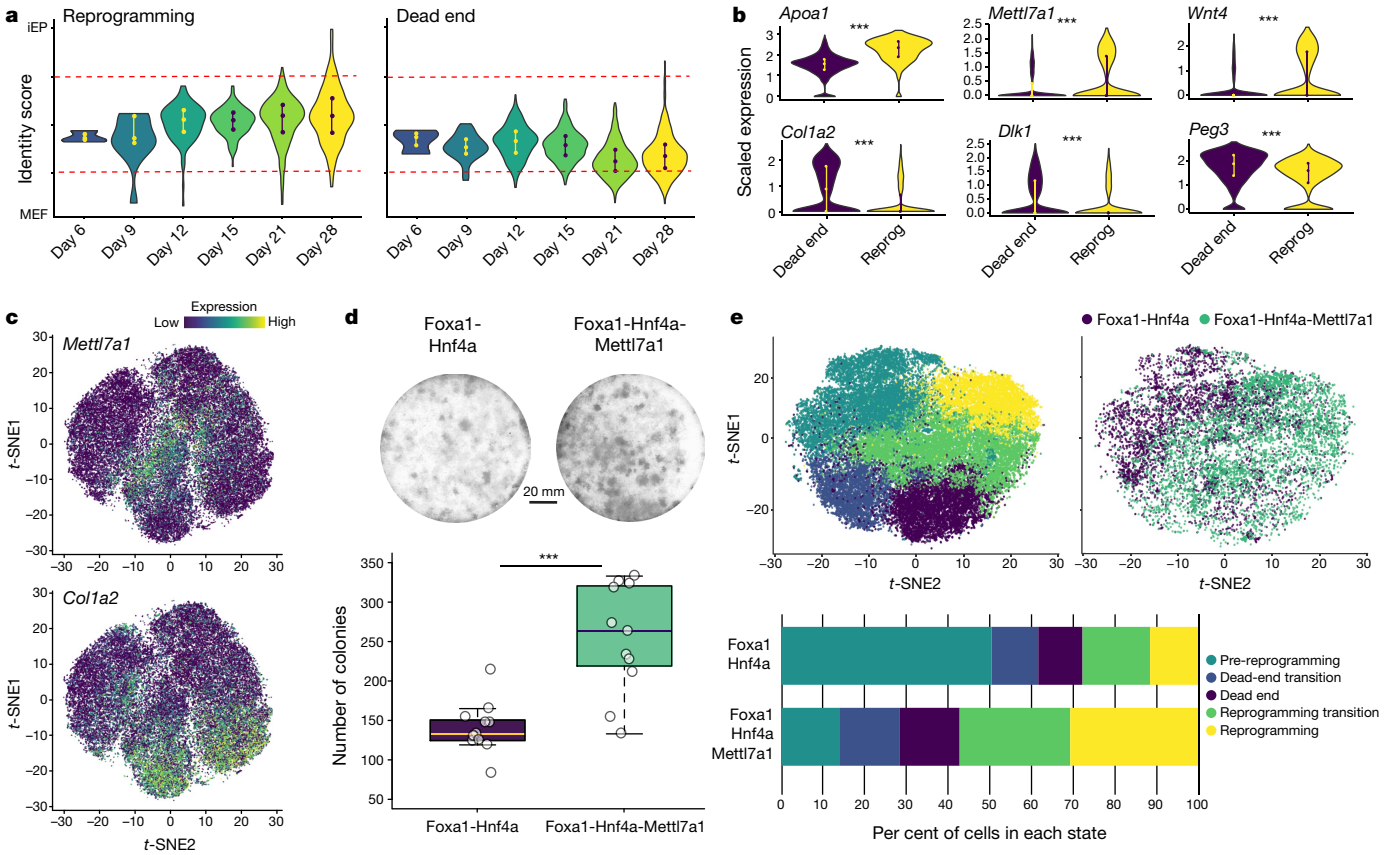


Fig. 4 | Molecular hallmarks of reprogramming trajectories. **a**, Identity scores of cells on the reprogramming (left, $n = 7$ clones, 1,037 cells) and dead-end trajectories (right, $n = 7$ clones, 1,037 cells, random downsampling from 2,270 cells) from reprogramming days 6 to 28. Cells scoring >0.75 (upper red line) are classified as iEPs, cells scoring <0.25 (lower red line) are classified as fibroblasts. **b**, Violin plots of significantly different ($P < 0.001$, permutation test, one-sided) gene expression between reprogramming and dead-end trajectories ($n = 14$ clones, 2,074 cells). **c**, Projection of *Mettl7a1* and *Col1a2* expression onto the *t*-SNE plot ($n = 48{,}515$ cells). **d**, Colony-formation assay (E-cadherin immunohistochemistry) for cells reprogrammed with Foxa1-Hnf4a or Foxa1-Hnf4a-Mettl7a1. Scale bar, 20 mm. Bottom, blinded and automated colony quantification. $n = 22$ technical replicates, 3 independent biological replicates; $P = 8 \times 10^{-5}$, one-sided *t*-test. **e**, Top, scRNA-seq analysis of 6,559 cells reprogrammed with Foxa1-Hnf4a and 6,559 cells (10,161 cells before random downsampling) reprogrammed with Foxa1-Hnf4a-Mettl7a1, 14 days after the start of reprogramming. Bottom, quantification of distribution of Foxa1-Hnf4a-Mettl7a1-reprogrammed cells across reprogramming stages, relative to that of Foxa1-Hnf4a-reprogrammed cells.

and reprogramming to pluripotency[20,21]. We therefore focused on *Mettl7a1* in the context of enhancing reprogramming efficiency. Addition of Mettl7a1 to the standard Foxa1-Hnf4a reprogramming cocktail resulted in a twofold increase in iEP colony formation (Fig. 4d). scRNA-seq of cells reprogrammed with Foxa1-Hnf4a or Foxa1-Hnf4a-Mettl7a1 reprogrammed cells shows that addition of Mettl7a1 to the reprogramming cocktail results in a threefold increase in the number of cells entering the fully reprogrammed state (Fig. 4e, Extended Data Fig. 10c–g). Inclusion of CellTags in these reprogramming experiments shows that under both control and Mettl7a1 conditions, the average number of cells per clone did not differ significantly between the two conditions (Extended Data Fig. 10h, i). Thus, Mettl7a1, rather than expanding existing iEPs, promotes a true increase in reprogramming efficiency.

Discussion

Here we have developed and validated a combinatorial indexing strategy, CellTagging, which enables simultaneous analysis of clonal history and cell identity at single-cell resolution. Our longitudinal dissection of Foxa1-Hnf4a-mediated direct lineage reprogramming to iEPs reveals two distinct conversion trajectories: one that leads to successful reprogramming, and one that leads to a dead-end state. We observe strong parallels between direct lineage reprogramming and induction of pluripotency: For instance, during induction of pluripotency, almost all cells initiate reprogramming, although transition to a fully pluripotent state is rare. This is characterized by two waves, or phases; in the second phase, a subset of cells are able to stably maintain the core pluripotency network[4,22]. In this context, the later bifurcation leading to the iEP state may parallel this second phase of reprogramming to pluripotency. Our identification of Mettl7a1 as a pro-reprogramming factor suggests that it may have an important role in the stabilization of iEP identity in later stages of lineage conversion.

Fibroblast-to-iEP conversion also shares a common feature with reprogramming to pluripotency with respect to inefficiency. On the basis of the low frequency of pluripotent cell generation, studies have suggested that the initiation and early phases of reprogramming are stochastic processes[4,23]. Our method of sequential CellTagging and lineage reconstruction enables reprogramming probabilities to be quantified. Tracking reprogramming outcome of clones derived from a shared ancestor strongly suggests that, in many cases, the trajectory of cell fate conversion is determined from the outset. If these early stages of reprogramming were stochastic, we would expect to see heterogeneity in reprogramming outcome between clones of the same lineage; however, we observe that clones of the same lineage follow similar reprogramming trajectories. Consistent with earlier studies[23], our CellTagging-and-split approach shows that clonally related cells—split into independent biological replicates—do not share reprogramming outcome, arguing against the existence of an elite cell type that is primed to reprogram. It is important to note here that, although we control the stoichiometry of the reprogramming factors, we do not control copy number or location of integration, which may produce a variable outcome between biological replicates.

The evidence presented here suggests the existence of a privileged cell state in which reprogramming potential is predetermined. This is supported by several recent studies from reprogramming to pluripotency that also suggest the existence of a privileged state, or that cells can be coaxed into such a state via transient factor expression[24–28]. Furthermore, DNA barcode-based clonal analyses support a deterministic model of reprogramming[29]. Finally, scRNA-seq in combination with computational trajectory reconstruction suggests that reprogramming outcome can be predicted as early as two days following initiation via factor expression[30]. The next challenge will be to uncover the molecular hallmarks of this permissive state, enabling further improvements in reprogramming cells towards any desired cell identity with high efficiency and fidelity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0744-4.

Received: 3 June 2017; Accepted: 3 October 2018;
Published online: 05 December 2018

1. Vierbuchen, T. & Wernig, M. Direct lineage conversions: unnatural but useful? *Nat. Biotechnol.* **29**, 892–907 (2011).
2. Cahan, P. et al. CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).
3. Morris, S. A. et al. Dissecting engineered cell types and enhancing cell fate conversion via *CellNet*. *Cell* **158**, 889–902 (2014).
4. Buganim, Y. et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
5. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
6. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
7. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
8. Rodriguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
9. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
10. Porter, S. N., Baker, L. C., Mittelman, D. & Porteus, M. H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines *in vitro* and *in vivo*. *Genome Biol.* **15**, R75 (2014).
11. Yao, Z. et al. A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* **20**, 120–134 (2017).
12. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
13. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
14. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
15. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
16. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390–393 (2011).
17. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
18. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. 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).
20. Chen, T. et al. m^{6}A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* **16**, 289–301 (2015).
21. Batista, P. J. et al. m^{6}A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707–719 (2014).
22. Polo, J. M. et al. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
23. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462**, 595–601 (2009).
24. Guo, S. et al. Nonstochastic reprogramming from a privileged somatic cell state. *Cell* **156**, 649–662 (2014).
25. Babos, K. N. et al. Balancing dynamic tradeoffs to drive cellular reprogramming. Preprint at https://www.biorxiv.org/content/early/2018/08/17/393934 (2018).
26. Rais, Y. et al. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* **502**, 65–70 (2013).
27. Di Stefano, B. et al. C/EBPα poises B cells for rapid reprogramming into induced pluripotent stem cells. *Nature* **506**, 235–239 (2014).
28. Di Stefano, B. et al. C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. *Nat. Cell Biol.* **18**, 371–381 (2016).
29. Yunusova, A. M., Fishman, V. S., Vasiliev, G. V. & Battulin, N. R. Deterministic versus stochastic model of reprogramming: new evidence from cellular barcoding technique. *Open Biol.* **7**, (2017).
30. Schiebinger, G. et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. Preprint at https://www.biorxiv.org/content/early/2017/09/27/191056 (2017).

Acknowledgements We thank members of the Morris laboratory, and T. Druley and R. Mitra for critical discussions; S. McCarroll, E. Macosko and M. Goldman for advice establishing Drop-seq; B. Treutlein for quadratic programming assistance; J. Dick for the gift of the pSMAL backbone; K. Kniepkamp for help with CellTag Viz; and The Genome Technology Access Center in the Department of Genetics. This work was funded by National Institutes of Health (NIH) grants R01-GM126112, R21-HG009750; P30-DK052574; Silicon Valley Community Foundation, Chan Zuckerberg Initiative Grants HCA-A-1704-01646 and HCA2-A-1708-02799; The Children's Discovery Institute of Washington University and St. Louis Children's Hospital MI-II-2016-544. S.A.M. is supported by a Vallee Scholar Award; B.A.B.: NIH-T32HG000045-18; C.G.: NIH-5T32GM007200-42; S.E.W.: NIH-5T32GM007067-44; K.K.: Japan Society for the Promotion of Science Postdoctoral Fellowship.

Reviewer information *Nature* thanks L. Perié, M. Porteus, L. Vallier and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.A.B. and S.A.M. conceived the research. S.A.M. led experimental work, assisted by B.A.B., W.K., C.G., S.E.W. and T.S. B.A.B. and W.K. led computational analysis, assisted by K.K. and supervised by S.A.M. All authors participated in interpretation of data and writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0744-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0744-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.A.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. Except where stated, the investigators were not blinded to allocation during experiments and outcome.

Mice and derivation of mouse embryonic fibroblasts. MEFs were derived from embryonic day (E)13.5 C57BL/6J embryos. (The Jackson laboratory: 000664). Heads and visceral organs were removed and the remaining tissue was minced with a razor blade and then dissociated in a mixture of 0.05% trypsin and 0.25% collagenase IV (Life Technologies) at 37 °C for 15 min. After passing the cell slurry through a 70-μM filter to remove debris, cells were washed and then plated on 0.1% gelatin-coated plates, in DMEM supplemented with 10% FBS (Gibco), 2 mM L-glutamine and 50 mM β-mercaptoethanol (Life Technologies). All animal procedures were based on animal care guidelines approved by the Institutional Animal Care and Use Committee.

Lenti- and retrovirus production. Lentiviral particles were produced by transfecting 293T-17 cells (ATCC: CRL-11268) with the pSMAL-CellTag construct (see below), along with packaging constructs pCMV-dR8.2 dvpr (Addgene plasmid 8455), and pCMV-VSVG (Addgene plasmid 8454). Constructs were titred by serial dilution on 293T cells. Hnf4a-T2A-Foxa1 and Mettl7a1 were cloned into the pGCDN-Sam retroviral construct and packaged with pCL-Eco (Novus Biologicals, NBP2-29540), titred on fibroblasts. We opted to generate a bicistronic Hnf4a-Foxa1 construct, based on the T2A sequence to increase the consistency of reprogramming via maintenance of exogenous transcription factor stoichiometry. Virus was collected 48 h and 72 h after transfection and applied to cells immediately following filtering through a low-protein binding 0.45-μm filter.

CellTagging methodology. To generate CellTags, we introduced an 8-bp variable region into the 3′UTR of GFP in the pSMAL lentiviral construct[31], using a gBlock gene fragment (Integrated DNA Technologies) and megaprimer insertion. This approach relies on the presence of 60-bp 'arms' in the gene fragment that are homologous to the desired plasmid insertion site. The fragments were then introduced into the plasmid using PCR, followed by DpnI (New England Biolabs) treatment to digest non-modified plasmid. All the recovered DNA from bacterial transformation (Stellar Competent Cells, Takara Biosciences) was grown overnight in liquid culture, followed by maxi-prep extraction of the plasmid DNA. This complex library of CellTag constructs was used to generate lentivirus (above) which was then used to transduce fibroblasts at a multiplicity of infection of ~3–4. For CellTag versions 2 and 3, a short 6-bp sequence was also included, just upstream of the variable CellTag region. For CellTag version 2, this sequence motif is GTGATG. For CellTag version 3, this sequence motif is TGTACG. For both Drop-seq and 10x Genomics-based experiments, the starting fibroblast population was transduced with CellTag version 1 (denoted as $CellTag^{MEF}$) for 24 h, followed by washing and culture for a further 48 h. At this point, cells were split, with one portion taken for Drop-seq/10x Genomics and two portions replated for reprogramming to iEPs in two biological replicates. For 10x Genomics-based experiments, cells were tagged again, immediately following 72 h of reprogramming, with CellTag version 2 (denoted as $CellTag^{D3}$). One further round of CellTagging followed at day 13 post-initiation of reprogramming with CellTag version 3 (denoted as $CellTag^{D13}$). Pooled CellTag libraries have been deposited at Addgene: https://www.addgene.org/pooled-library/morris-lab-celltag/, pSMAL-CellTag-V1 (pooled library #115643); pSMAL-CellTag-V2 (pooled library #115644); pSMAL-CellTag-V3 (pooled library #115645).

Generation and collection of iEPs. Early passage MEFs (<passage 6) were reprogrammed with modifications to the described protocols[16]. We modified this protocol, transducing cells every 12 hours for 3 days, with fresh Hnf4a-T2A-Foxa1 retrovirus in the presence of 4 μg/ml protamine sulfate (Sigma-Aldrich). These transduced cells were then cultured on 0.1% gelatin-treated plates for 1 week in hepato-medium (DMEM:F-12, supplemented with 10% FBS, 1 μg/ml insulin (Sigma-Aldrich), 100 nM dexamethasone (Sigma-Aldrich), 10 mM nicotinamide (Sigma-Aldrich), 2 mM L-glutamine, 50 mM β-mercaptoethanol (Life Technologies), and penicillin–streptomycin, containing 20 ng/ml epidermal growth factor (Sigma-Aldrich)). After 7 days of culture, the cells were transferred onto plates coated with 5 μg/cm^2 Type I rat collagen (Gibco, A1048301). For Drop-seq based experiments (two independent biological replicates), with a cell capture rate of 5%, 2×10^5 cells were initially seeded, and cells were collected every 7 days. At each collection, cells were gently dissociated in TrypLE Express (Gibco), and 1.5×10^5 cells were collected for Drop-seq, replating and culturing the remaining cells. For 10x Genomics-based experiments, with a cell encapsulation rate of up to 60%, 5×10^4 cells were initially seeded and collected every 3–7 days. At each cell collection, 3×10^4 dissociated cells were fixed in methanol, and the remaining cells were replated and cultured. Methanol fixation was performed as previously described[32]. In brief, cells were collected and washed in phosphate buffered saline (PBS), followed by resuspension in ice-cold 80% methanol in PBS, with gentle vortexing. These cells were stored at −80 °C for up to three months, and processed in the same batch on the 10x Genomics platform (below). iEP lines at the end of reprogramming tested negative for mycoplasma.

Immunostaining. iEP cells were grown in 4-Chamber Culture Slides (Falcon #354114) and fixed in 4% paraformaldehyde. Cells were permeabilized in 0.1% Triton-X100, followed by blocking in 10% fetal bovine serum in PBS (blocking buffer). Primary antibody, goat apolipoprotein A-I antibody (1:100, Novus Biologicals, NB600-609, lot: 30506) or mouse E-cadherin antibody (1:50, BD Biosciences, 610181, Clone: 36/E-cadherin, lot: 7187865) in blocking buffer was applied overnight before washing and applying secondary antibody: Alexa Fluor 555 rabbit anti-goat IgG (1:1000, Invitrogen A-21431) or Alexa Fluor 488 goat anti-mouse IgG (1:1000, Invitrogen A-32723), diluted in blocking buffer. Nuclear staining was performed with 300 nM DAPI in PBS. Slides were mounted with ProLong Gold antifade reagent (Invitrogen P36930). Images were captured using a Zeiss Axio Imager Z2 fluorescent microscope.

Mettl7a1 reprogramming and colony formation assay. Mouse *Mettl7a1* (NM 027334, Origene: MC205948) was sub-cloned into the retroviral vector, pGCDN-Sam[16], and retrovirus was produced as described above. For comparative reprogramming experiments, MEFs (1.2×10^5 cells per 6-cm plate, in 3 independent biological replicates) were serially transduced over 72 h (as above), followed by splitting and seeding at 4×10^4 cells per well of a 6-well plate to generate technical replicates. In control experiments, virus produced from an empty vector control expressing only GFP was added to the Foxa1-Hnf4a reprogramming cocktail. In Mettl7a1 experiments, virus produced from the Mettl7a1-IRES-GFP construct was added to virus containing Hnf4a and Foxa1. *Mettl7a1* overexpression was confirmed by preparing RNA from cells transduced with Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 using the RNeasy kit (Qiagen). Following cDNA synthesis (Maxima cDNA synthesis kit, Life Tech), quantitative reverse transcription with PCR (qRT–PCR) was performed to quantify *Mettl7a1* overexpression (TaqMan Probe: Mm03031185_sH, TaqMan qPCR Mastermix, Applied Biosystems). Cells were reprogrammed for two weeks, at which point the cells in some wells were dissociated and fixed in methanol for 10x Genomics-based single-cell analysis (details below). The remaining wells were processed for colony-formation assays: cells were fixed on the plate with 4% paraformaldehyde, permeabilized in 0.1% Triton-X100 then blocked with Mouse on Mouse Elite Peroxidase Kit (Vector PK-2200). Mouse E-cadherin antibody (1:100, BD Biosciences) was applied for 30 min before washing and processing with the VECTOR VIP Peroxidase Substrate Kit (Vector SK-4600). Colonies were visualized on a flatbed scanner, adding heavy cream to each well to increase image contrast. Colonies were counted, using the colony counter ImageJ plugin (https://imagej.nih.gov/ij/plugins/colony-counter.html). These analyses were blinded.

Drop-seq. Cells were dissociated using TrypLE Express (Gibco), washed in PBS containing 0.01% BSA and diluted to 100 cells/μl, then processed by Drop-seq within 15 min of their collection. Drop-seq was performed as previously described[17] (http://mccarrolllab.com/dropseq/). In brief, cells and beads were diluted to an estimated co-occupancy rate of 5% upon co-encapsulation: 1×10^5 cells/ml and 1.2×10^5 beads/ml. Two independent lots of beads (Macosko-2011-10, ChemGenes) were used: 091615 (time course 3) and 032516B (time course 4). Emulsions were collected and broken using 1 ml of Perfluorooctanol (Sigma) for 15 ml of emulsion, followed by washing in 6× saline-sodium citrate (SSC) buffer to recover beads. Reverse transcription was then performed using the Maxima H Minus Reverse Transcriptase kit (EP0752, Life Tech). After treatment with 2,000 U/ml of ExonucleaseI (New England Biolabs), aliquots of 2,000 beads (representing ~100 single-cell transcriptomes for a cell-bead co-encapsulation rate of 5%) were amplified by PCR for 13 cycles, using Kapa HiFi Hotstart Readymix (Kapa Biosystems). The PCR product resulting from this reaction was purified by addition of 0.6× AMPure XP beads (Beckman Coulter). Six hundred picograms of this purified cDNA product from an estimated 5,000 cells was tagmented using Nextera XT according to the manufacturer's protocol (Illumina). The resulting cDNA library was again purified using 0.6× AMPure XP beads, followed by 1× AMPure XP beads. cDNA concentrations were assessed by Tapestation (Agilent) analysis. Libraries were sequenced on an Illumina HiSeq 2500, with custom priming (Read1CustSeqB Drop-seq primer).

10x Genomics procedure. For single-cell library preparation on the 10x Genomics platform, we used: the Chromium Single Cell 3′ Library and Gel Bead Kit v2 (PN-120237), Chromium Single Cell 3′ Chip kit v2 (PN-120236) and Chromium i7 Multiplex Kit (PN-120262), according to the manufacturer's instructions in the Chromium Single Cell 3′ Reagents Kits V2 User Guide. Just before cell capture, methanol-fixed cells were placed on ice, spun at 3,000 r.p.m. for 5 min at 4 °C, followed by resuspension and rehydration in PBS, according to a previously described method[32]. Seventeen thousand cells were loaded per lane of the chip, aiming for capture of 10,000 single-cell transcriptomes. All samples were processed in parallel, on the same day. Resulting cDNA libraries were quantified on an Agilent Tapestation and sequenced on an Illumina HiSeq 3000.

Viral integration analysis. Genomic DNA was prepared from control MEFs and iEPs derived from clone 1 (time course 4), using the DNeasy Blood & Tissue kit (Qiagen). Sample quality was assessed by Qubit DNA Assay Kit and gel electro-

phoresis. Library construction was carried out using the Nextera XT Library prep kit (Illumina) following the manufacturer's recommendations. The lentivirus integration boundary sequence was enriched by amplification using primers specific for lentivirus long terminal repeat (LTR) and the Nextera XT adaptor sequence. Two separate PCR reactions were performed for each sample, one for 3′ LTR and another for 5′ LTR. The final PCR was performed to add Illumina sequencing adapters with unique barcodes for each sample. The libraries for each sample were pooled into a final library and assessed by Qubit DNA assay, Agilent Bioanalyzer and qRT–PCR. The library was sequenced on the NextSeq 500 system using the 150 Cycle High Output flow cell. Fastq data was extracted from the NextSeq system using bcl2fastq and the quality control of the data was performed using FastQC. Fastq reads were aligned to the mouse reference genome (GRCm38) using BWA MEM. De-duplication was performed using Samtools. Peak calling and comparison between two samples for putative lentivirus integration site was performed using MACS2.

Library preparation and sequencing of CellTag plasmid libraries for whitelist generation. Library construction was carried out using the Nextera XT Library prep kit (Illumina), following the manufacturer's recommendations. The CellTag region was enriched by amplification using primers specific for the pSMAL lentivirus GFP UTR and the Nextera XT adaptor sequence. A final PCR was performed to add Illumina sequencing adapters. The libraries for each CellTag version were pooled and assessed by Tapestation (Agilent). The library was sequenced on an Illumina MiSeq. Reads that contained the CellTag motif were identified (see 'CellTag demultiplexing'). A 90% percentile cut-off in terms of reads reported for each CellTag was used to select CellTags for inclusion on the whitelist of cell barcodes.

10x Genomics and Drop-seq alignment, digital gene expression matrix generation. The Cell Ranger v.2.1.0 pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest) was used to process data generated using the 10x Chromium platform. This pipeline was used in conjunction with a custom reference genome, created by concatenating the sequences corresponding to the Hnf4a-T2A-Foxa1 transgene and the GFP-CellTag transgene as new chromosomes to the mm10 genome. The unique UTRs in the Hnf4a-T2A-Foxa1 and GFP-CellTag transgene constructs allowed us to monitor transgene expression. To create Cell Ranger-compatible reference genomes, the references were rebuilt according to instructions from 10x Genomics (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references). To achieve this, we first created a custom gene transfer format (GTF) file, containing our transgenes, followed by indexing of the FASTA and GTF files, using Cell Ranger mkgtf and mkref functions. Following this step, the default Cell Ranger pipeline was implemented, with the filtered output data used for downstream analyses. For Drop-seq analysis, raw reads were processed, filtered, and aligned as previously described[17], including correction of barcode synthesis errors. This process and the required tools, are further outlined online in the Drop-seq Alignment Cookbook (http://mccarrolllab.com/dropseq/). To facilitate downstream analyses the reference genome used during alignment was modified to include the transgenic sequences above. Processed reads were aligned to a custom genome build, using STAR. Across all experiments, the mean number of confidently mapped reads per cell was 38,259 (Supplementary Table 1).

Following alignment, digital gene expression (DGE) matrices were generated for each time point, for all time courses. Drop-seq DGEs were aggregated using a custom R script. Merged 10x Genomics DGE files were generated using the aggregation function of the Cell Ranger pipeline. We then performed initial filtering of these DGE files as a quality control step. We first removed cells with a low number (<200) of unique detected genes. We then removed cells for which the total number of unique molecular identifiers (UMIs) (after log transformation) was not within three standard deviations of the mean. This was followed by the removal of outlying cells with an unusually high or low number of UMIs given their number of reads by fitting a loess curve (span = 0.5, degree = 2) to the number of UMIs with number of reads as predictor (after log transformation), removing cells with a residual more than three standard deviations away from the mean. This process was also used to remove cells for with unusually high or low number of genes given their number of UMIs. Finally, we removed cells in which the proportion of the UMI count attributable to mitochondrial genes was greater than 10% (for Drop-seq-based experiments) or 20% (for 10x Genomics-based experiments).

Data normalization and scoring of cell cycle phase. Following DGE filtering, cell cycle scores were generated for each cell and data were normalized. Cell cycle scores were generated using a pre-defined classifier to assign cell cycle phase for each cell. This classifier was built from training data by identifying pairs of genes where the difference in expression within each pair changed sign across phases. Cell cycle phase was assigned to each cell by examination of the sign of the difference in test data. After calculating the cell cycle scores, the data was normalized using the 'deconvolution' method. This method pools cells and combines the expression values of the cells in a pool. The pooled expression values are used to calculate size-factors for normalization. These pool-based normalization factors can then be deconvoluted into cell-specific normalization factors, which are then used to normalize the expression of each cell. This deconvolution normalization method is an attempt to address the abundance of zero counts that is prevalent to scRNA-seq. The cell cycle scores and data normalization was facilitated by the Scater package[33], available on Bioconductor.

CellTag demultiplexing. Reads containing the CellTag sequence were extracted from the processed and filtered BAM files produced by the 10x Genomics and Drop-seq pipelines. Reads that contained the CellTag motif were identified from the following sequences: CellTagV1 ($CellTag^{MEF}$): CCGGTNNNNNNNNGAATTC, CellTagV2 ($CellTag^{D3}$): GTGATGNNNNNNNNGAATTC, CellTagV3 ($CellTag^{D13}$): TGTACGNNNNNNNNGAATTC. Following extraction of reads from the BAM file, a custom gawk script was used to parse the output, capturing the read ID, sequence, cell barcode, UMI, CellTag sequence and aligned genes for each read. This parsed output was then used to construct a cell barcode × CellTag UMI matrix. CellTags were grouped by cell barcodes and then the number of unique UMIs for each cell barcode–CellTag pair was counted. The matrix was then filtered to remove any cell barcodes not found in the filtered Cell Ranger and Drop-seq output files. Finally, the CellTags were filtered to remove any that were represented by $\leq$1 UMI. The construction and filtering of the CellTag UMI matrix was accomplished using a custom R script. Using this matrix, an error-correction step was then performed to amend PCR and sequencing errors: CellTags one edit-distance apart were collapsed on a cell-by-cell basis, using Starcode[34], an algorithm to determine which sequence pairs lie within a given Levenshtein distance, merging matched pairs into clusters of similar sequences. This filtered CellTag UMI count matrix was then used for all downstream clone and lineage analysis.

CellTag filtering and clone calling. The CellTag matrix was initially filtered by removing CellTags that do not appear on the whitelists generated for each CellTag plasmid library (see 'Library preparation and sequencing of CellTag plasmid libraries for whitelist generation'). CellTags appearing in $>$5% of cells in the first time point were also removed as this would suggest dominance of the library by individual CellTags that would interfere with accurate clone-calling. The requirement for this filtering was rare. Cells expressing more than 20 CellTags (likely to correspond to cell multiplets), and less than 2 CellTags per cell were filtered out. To identify clonally related cells, Jaccard analysis using the R package Proxy was used to calculate the similarity of CellTag signatures between cells. A Jaccard score of $>$0.7 was used as a cut-off to identify cells highly likely to be related, on the basis of our experimental findings. We found this cut-off to be stringent enough for unrelated cells not to be connected, but in a small number of instances, we found related cells that were not connected, probably owing to CellTag errors that were not corrected, or CellTag dropout. These related cells were united as part of lineage construction, below. Clones were defined as groups of 3 or more related cells (for $CellTag^{MEF}$, $CellTag^{D3}$), or 2 or more related cells (for $CellTag^{D13}$) identified using a custom R script. Clones were visualized using the Corrplot package with hierarchical clustering, contour plotting using ggplot2, or using force-directed network graphs (see below). Clones were called on cells pre-filtered for numbers of genes, UMIs and mitochondrial RNA content.

Seurat, Monocle and quadratic programming analyses. After filtering and normalization, the R package Seurat[6] was used to cluster and visualize cells. As the data were already normalized, they were loaded into Seurat without normalization, scaling or centring. Along with the expression data, metadata for each cell was collected, including information such as clone identity, cell cycle phase, and time point (Supplementary Table 4). Seurat was used to remove unwanted variation, regressing out number of UMIs, proportion of mitochondrial UMIs and cell cycle scores. Next, highly variable genes were identified and used as input for dimensionality reduction via principal component analysis (PCA). The resulting PCs and the correlated genes were examined to determine the number of components to include in downstream analysis. These PCs were then used as input to cluster the cells, visualizing these clusters using *t*-SNE. Semi-supervised Monocle[7] analysis was used to order cells in pseudotime, based on expression of the fibroblast marker *Col1a2* and the iEP marker *Apoa1*. Quadratic programming[5] was used to score fibroblast and iEP identity. This approach was modified to use bulk expression data of MEF and iEP collected previously[16] and whole transcriptome profiles of the two cell types were used for identity score calculation. The R package QuadProg was used for quadratic programming to generate cell identity scores. Investigators were blinded to allocation in the orthogonal pseudotemporal ordering analysis.

Lineage visualization via construction of force-directed network graphs. Network graphs were constructed by integrating all data for all rounds of CellTagging. In the graphs, each node represents an individual cell, and edges represent clonal relationships between cells. First, using a custom R-based script, cells were assembled into sub-clusters, according to $CellTag^{MEF}$, $CellTag^{D3}$, and $CellTag^{D13}$ information. Then, these sub-clusters were connected to each other to build lineages of related cells, connected across the different rounds of CellTagging—that is, two different $CellTag^{D3}$ clones sharing the same $CellTag^{MEF}$

labels are part of the same lineage. Using this approach, we identified collisions in $4.5 \pm 1.1\%$ of clones—a collision is defined as one clone sharing two or more parents. In these cases, we inspected the CellTag signature for each clone and united any clones that had been split, reducing the collision rate to $0.9 \pm 0.6\%$. The resulting networks were visualized as force-directed network graphs using Cytoscape 3.6.0 and Allegro Layout. Allegro spring-electric was used as the layout protocol to render force-directed network graphs. Individual graphs for each lineage can be explored with our Shiny-based interactive platform, CellTag Viz (http://www.celltag.org/).

Trajectory discovery by randomized testing. To identify clones with an enriched or depleted rate of iEP generation, we used randomized testing to evaluate whether each clone (of at least 35 cells in size) possesses a similar percentage of fully reprogrammed cells relative to a randomly selected population of the same size. Here, the percentage of reprogrammed cells is defined as the proportion of cells within each group found in the reprogrammed cluster, as defined by Seurat. Two groups, cells of the clone and that of the overall population, are compared with the null percentage calculated using the cells in each clone. Let N represent the number of cells in each clone and M represent the remaining cell population size. We pool the two groups of cells (size $= N + M$) and resample N random cells, without replacement, from the pooled cells $(N + M)/N$ times such that every possible separation with ending groups of size N and M can be sampled and captured. During this process, the percentage is calculated based on the N randomly sampled cells. With the percentage calculated, P values can be evaluated based on the proportion of randomly sampled cells with a percentage greater than or equal to the null percentage. Using the P value of <0.05 (>0.95 for the other tail), we identified clones that were enriched or depleted for reprogrammed cells. These calculations were performed using a custom R-based script. Clones with at least 35 cells were selected to increase the statistical power of this analysis. For permutation testing to analyse differences in trajectory-specific gene expression, a custom Python-based script was used.

Reagent and protocol availability. Pooled CellTag libraries have been deposited and are available from Addgene: https://www.addgene.org/pooled-library/morris-lab-celltag/. A working protocol can be accessed via protocols.io https://doi.org/10.17504/protocols.io.vawe2fe.

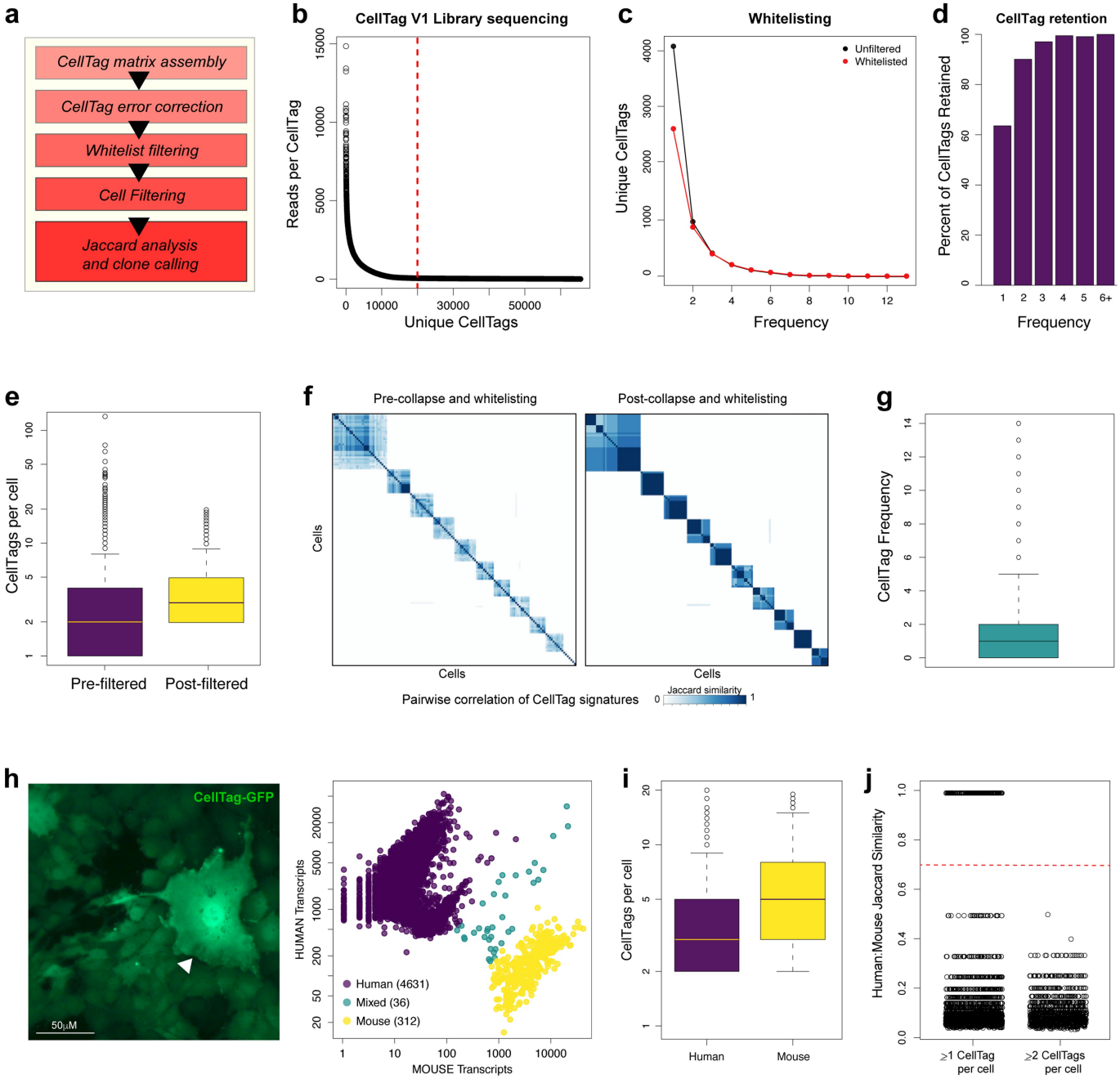
Code availability. Code for processing of CellTag data, clone-calling, and construction of lineage trees is available on GitHub (https://github.com/morris-lab).

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

Data availability

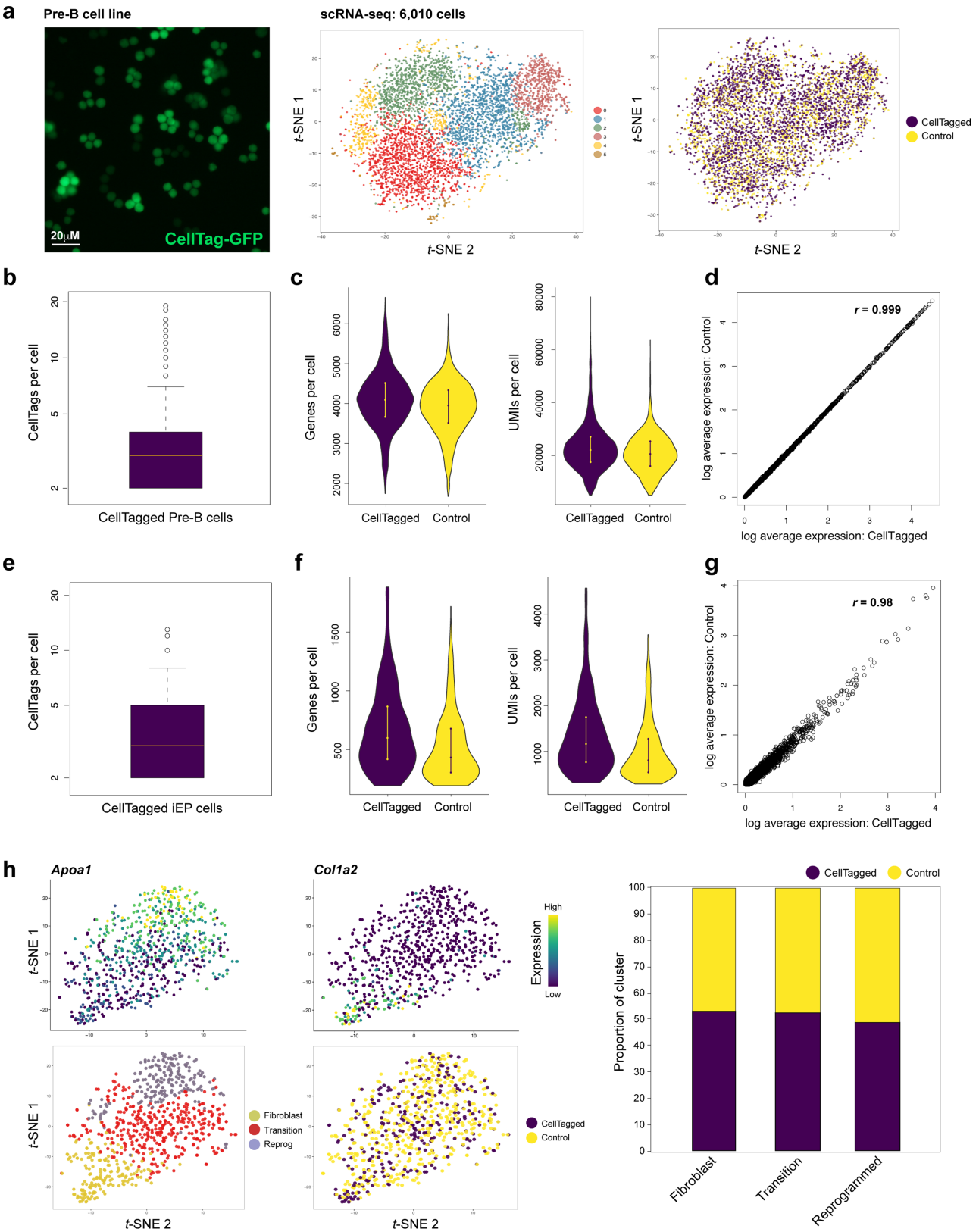
All source data, including sequencing reads and single-cell expression matrices, are available from the Gene Expression Omnibus (GEO) under accession code GSE99915.

31. van Galen, P. et al. The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. *Nature* **510**, 268–272 (2014).
32. Alles, J. et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol.* **15**, 44 (2017).
33. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **347**, btw777 (2017).
34. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).



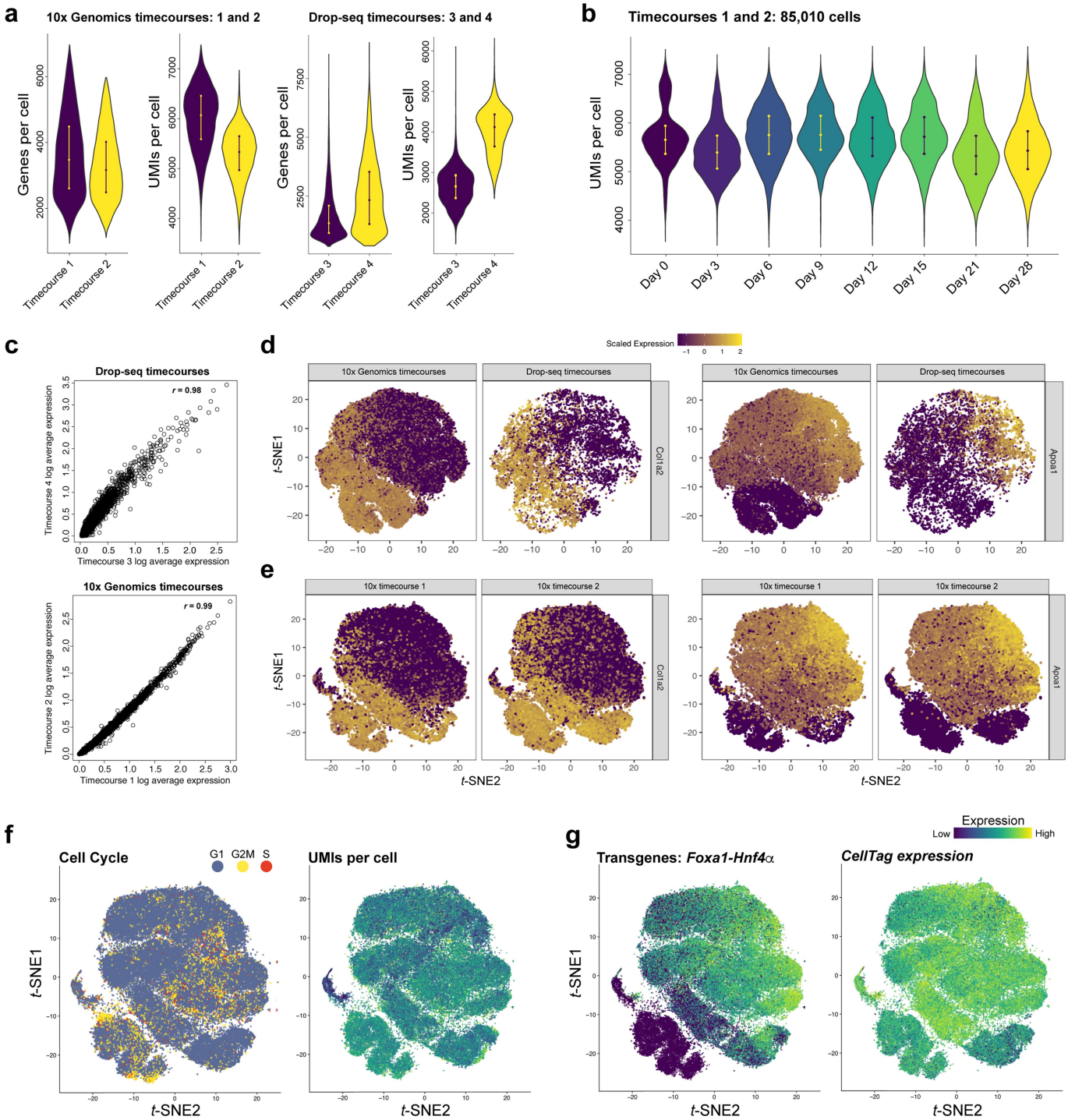
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | CellTag processing and species-mixing validations. a, Schematic of the CellTag processing and filtering pipeline: CellTag sequences are first extracted from aligned sequencing reads, followed by construction of a matrix of CellTag expression in each cell. To mitigate potential artefacts arising as a result of PCR and sequencing errors, we implemented an error-correction step, collapsing similar barcodes one edit-distance apart, on a cell-by-cell basis. An initial filtering step then removes any CellTags that do not appear on a whitelist of CellTags that are confirmed to exist in the complex lentiviral library. A second filtering step removes cells expressing less than two or more than 20 unique CellTags. Using this filtered dataset, Jaccard analysis is then applied (using the R package, Proxy) to identify related cells, based on CellTag signature similarity, allowing clones to be called. **b**, Generation of the CellTag whitelist. Following CellTag lentiviral plasmid sequencing, CellTags were extracted from the raw fastq files via identification of the adjacent motifs as described in Methods (see Methods, 'CellTag demultiplexing'). A 90th percentile cut-off in terms of reads reporting each CellTag was used to select CellTags for inclusion on the whitelist. Of a possible 65,536 unique combinations, we detected 19,973 sequences passing this 90th percentile of read counts. Data for CellTag version 1 ($CellTag^{MEF}$) is shown here. Whitelist creation was also performed for CellTag versions 2 ($CellTag^{D3}$) and 3 ($CellTag^{D13}$). **c**, **d**, CellTag frequency (**c**), that is, how many times each CellTag is detected in a population of transduced cells, before (black) and after (red) removal of CellTags that do not feature on the whitelist. This whitelisting predominantly results in the removal of CellTags that appear only once; singletons that are likely to arise owing to sequencing and PCR errors. This is reflected in the histogram in **d**, showing that only 60% of singleton CellTags detected are retained, whereas over 90% of CellTags appearing in two or more cells are retained. **e**, Mean CellTags per cell pre- and post-CellTag pipeline filtering. Cells in this figure correspond to the cells shown in Fig. 1b, c (replicate 1: $n = 8{,}535$ cells; replicate 2: $n = 11{,}997$ cells). **f**, Pairwise correlation scores (Jaccard similarity) and hierarchical clustering of 10 major clones arising from this tag and trace experiment. Hierarchical clustering is based on each cell's Jaccard correlation relationships with other cells, where each defined 'block' of cells represents a clone. Left, scoring and clustering of pairwise correlations, before whitelisting and filtering. Right, after whitelisting and filtering, pairwise correlations are stronger and more cells are detected within each clone ($n = 869$ cells). **g**, CellTag frequency metric: each detected CellTag appears in less than two cells ($n = 9{,}072$ cells in total) at the start of the experiment, on average. The library is therefore not dominated by any abundant CellTags, which would potentially generate false-positive results. **h**, A species mixing experiment, consisting of a mixture of human 293T cells and MEFs (left), labelled with ~3–5 CellTags per cell and expressing GFP as a result. A fibroblast (white arrow) is visible within a colony of 293T cells. Scale bar, 50μ M. Seventy-two hours after transduction, cells were collected and processed for Drop-seq. Right, following sequencing and alignment, cells were assigned to their corresponding species, revealing a low rate of doublet formation ($n = 4{,}631$ human cells, 312 mouse cells, 36 mixed). **i**, Mean CellTags per cell for human and mouse cells in the species-mixing experiment. CellTag transcripts were detected in 70% of cells ($n = 3{,}493/4{,}979$ cells). Of the tagged population, each cell expressed 5 CellTags on average: 3.800 ± 0.002 in human cells, and 5.90 ± 0.02 in mouse cells (mean ± s.e.m.). **j**, For each cell, CellTag signatures were extracted and Jaccard similarity analysis was performed to assess the frequency of CellTag signature overlap between the two species. To establish a false-positive baseline, we initially compared CellTag overlap between mouse and human populations, as these cells are not related. From the analysis of 4,943 cells, we identified 200 instances of mouse–human cell pairings out of a possible 1.5×10^7 pairs sharing the same individual CellTags. This demonstrates that reliance on only one CellTag per cell does not uniquely label cells with high confidence. Excluding cells represented by only one CellTag removes this noise, resulting in no detection of cross-species CellTag signatures (Jaccard similarity index <0.7). This highlights the importance of combinatorial labelling, and the efficacy of our approach to uniquely label unrelated cells.

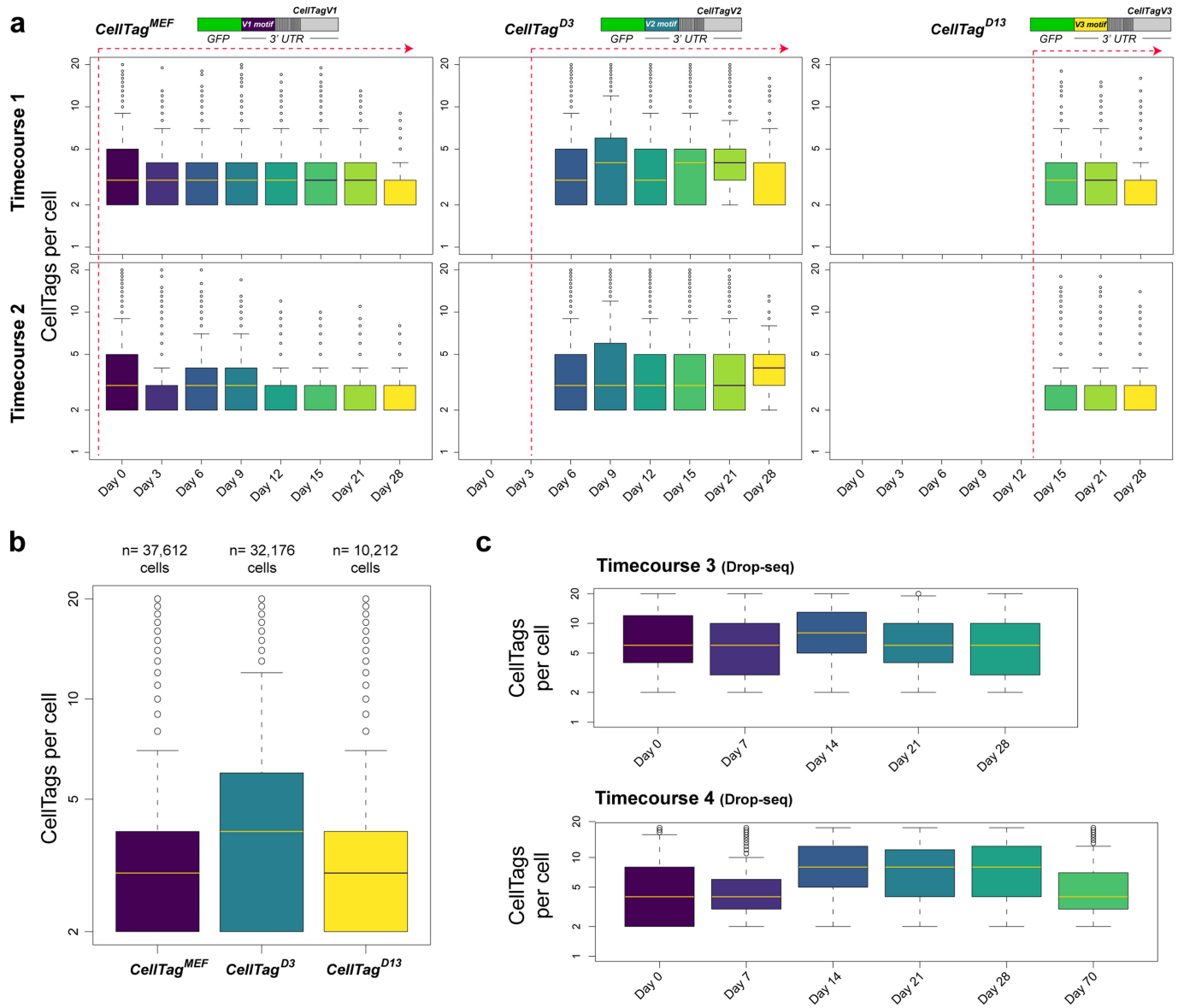


Extended Data Fig. 2 | See next page for caption.

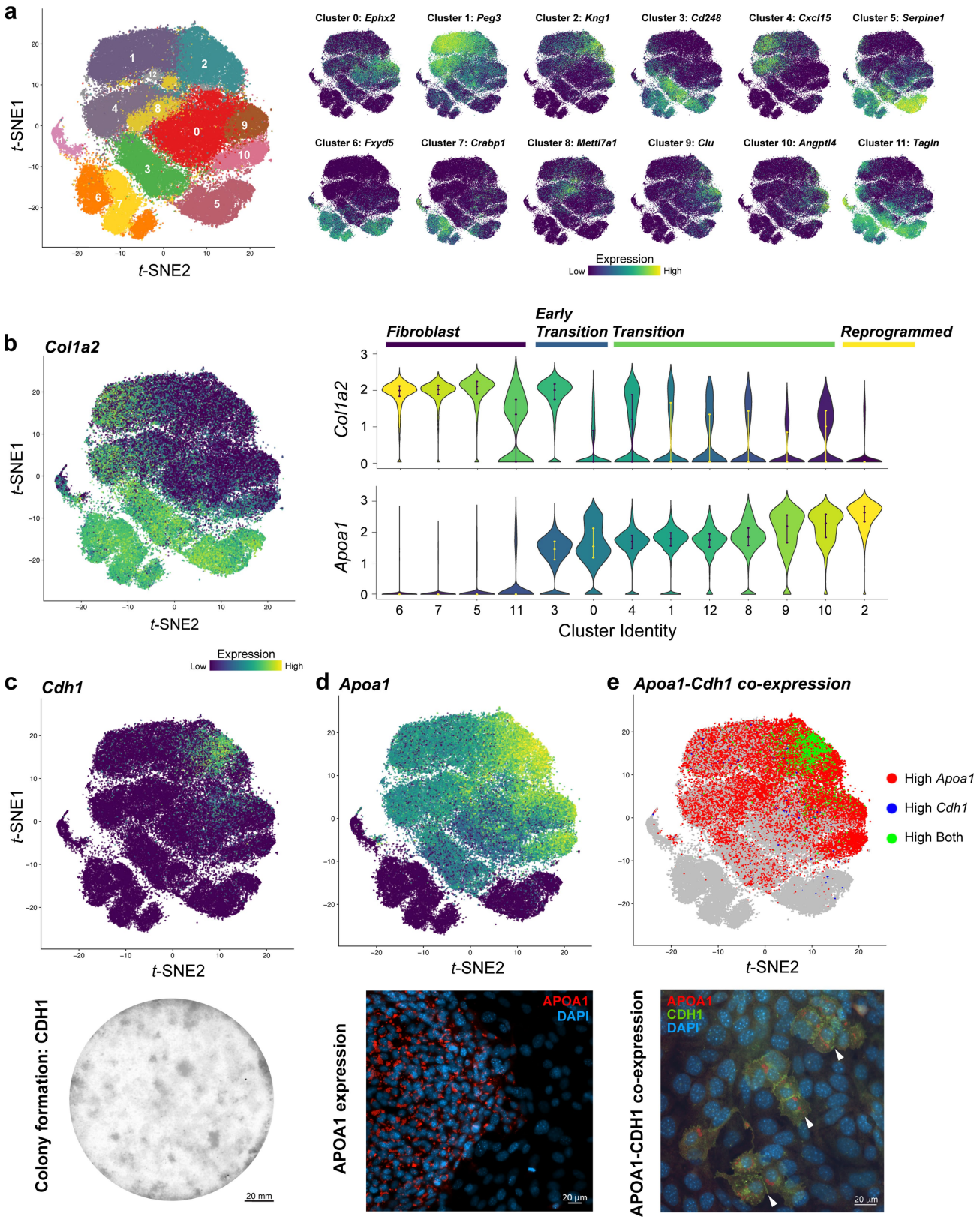
Extended Data Fig. 2 | CellTagging does not perturb cell physiology or reprogramming efficiency. To assess the potential effect of CellTagging on cell physiology we performed scRNA-seq on CellTag-labelled cells and unlabelled control cells 72 h after tagging. **a**, Left, fluorescent image of CellTag-labelled, GFP-expressing, pre-B cell line, HAFTL-1. Right, 10x Genomics-based scRNA-seq of CellTag-labelled ($n = 3{,}943$ cells) and non-tagged control cells ($n = 2{,}067$ cells). Cells were clustered using Seurat, resulting in a *t*-SNE plot with 6 clusters of transcriptionally distinct cells. CellTag-labelled and control cells were evenly distributed across these populations. **b**, The CellTag-labelled B-cell population expresses a mean of 3.50 ± 0.02 CellTags per cell. **c**, We detect no observable differences in numbers of genes or UMIs per cell in either population. **d**, Average gene expression values between CellTag-labelled and control cells are highly correlated ($r = 0.999$, Pearson's correlation), demonstrating that our labelling approach does not induce significant changes in gene expression. These experiments were performed independently twice with similar results. **e**, To assess the potential effect of CellTagging on reprogramming outcome, we induced lineage conversion (MEF to iEP) of CellTagged cells in parallel with unbarcoded control cells, followed by three weeks of culture and processing on the Drop-seq platform ($n = 773$ cells passing quality control). A mean of 3.30 ± 0.09 CellTags per cell are expressed in a labelled reprogrammed cell population. **f**, There are no observable differences in numbers of genes or UMIs per cell in either the labelled or unlabelled populations. **g**, Average gene expression values between CellTagged and control cells are highly correlated ($r = 0.98$, Pearson's correlation), again demonstrating that our labelling approach does not induce significant changes in gene expression. **h**, Seurat clustering of cells, in which cells in fibroblast (*Col1a2*-high), transition, and fully reprogrammed (*Apoa1*-high) states can be identified. Right, barcoded and control cells are distributed fairly evenly across these reprogramming stages. Some variation is expected between these independent biological replicates. These experiments were performed independently twice with similar results.



Extended Data Fig. 3 | scRNA-seq metrics and quality control of cell clustering. a, Numbers of genes and UMIs per cell for 10x Genomics-based (time course 1, $n = 30{,}733$ cells and time course 2: $n = 54{,}277$ cells) and Drop-seq-based (time course 3, $n = 5{,}932$ cells and time course 4: $n = 5{,}414$ cells) reprogramming time courses. In these cross-platform comparisons, we apply more stringent filtering of Drop-seq data to include only those cells with 1,000 or more UMIs. For Drop-seq experiments, with a cell capture rate of 5%, 2×10^5 MEFs were initially seeded for reprogramming. For 10x Genomics experiments, with a cell encapsulation rate of up to 60%, 5×10^4 MEFs were initially seeded for reprogramming. **b**, Mean numbers of UMIs per cell at each captured time point during reprogramming ($5{,}570.0 \pm 2.2$), in two independent biological replicates (10x Genomics, time courses 1 and 2): cells were captured at days 3, 6, 9, 12, 15, 21 and 28, along with the initial MEF population (day 0). **c**, Average gene expression values of 10x Genomics and Drop-seq replicates are highly correlated at day 0, demonstrating technical consistency ($r = 0.99$, and $r = 0.98$, respectively, Pearson's correlation). **d**, Alignment of independent 10x Genomics replicates (time courses 1 and 2) with Drop-seq replicates (time courses 3 and 4) using canonical correlation analysis[19]. Left, expression of MEF marker *Col1a2*. Right, iEP marker *Apoa1*. Overlay of data from these two sources demonstrates a high level of technical and biological consistency between the two technologies. **e**, Alignment of 10x Genomics replicates (time course 1 and 2) using canonical correlation analysis. Expression of *Col1a2* (left), *Apoa1* (right). Integration of these two replicates demonstrates a high level of technical and biological consistency. **f**, Projections of cell cycle phase and UMIs per cell onto *t*-SNE alignment of time courses 1 and 2 shows that clustering is independent of these factors. **g**, Reprogramming factor expression (using detection of bicistronic Hnf4a-T2A-Foxa1 transgene expression) and CellTag expression across time courses 1 and 2.

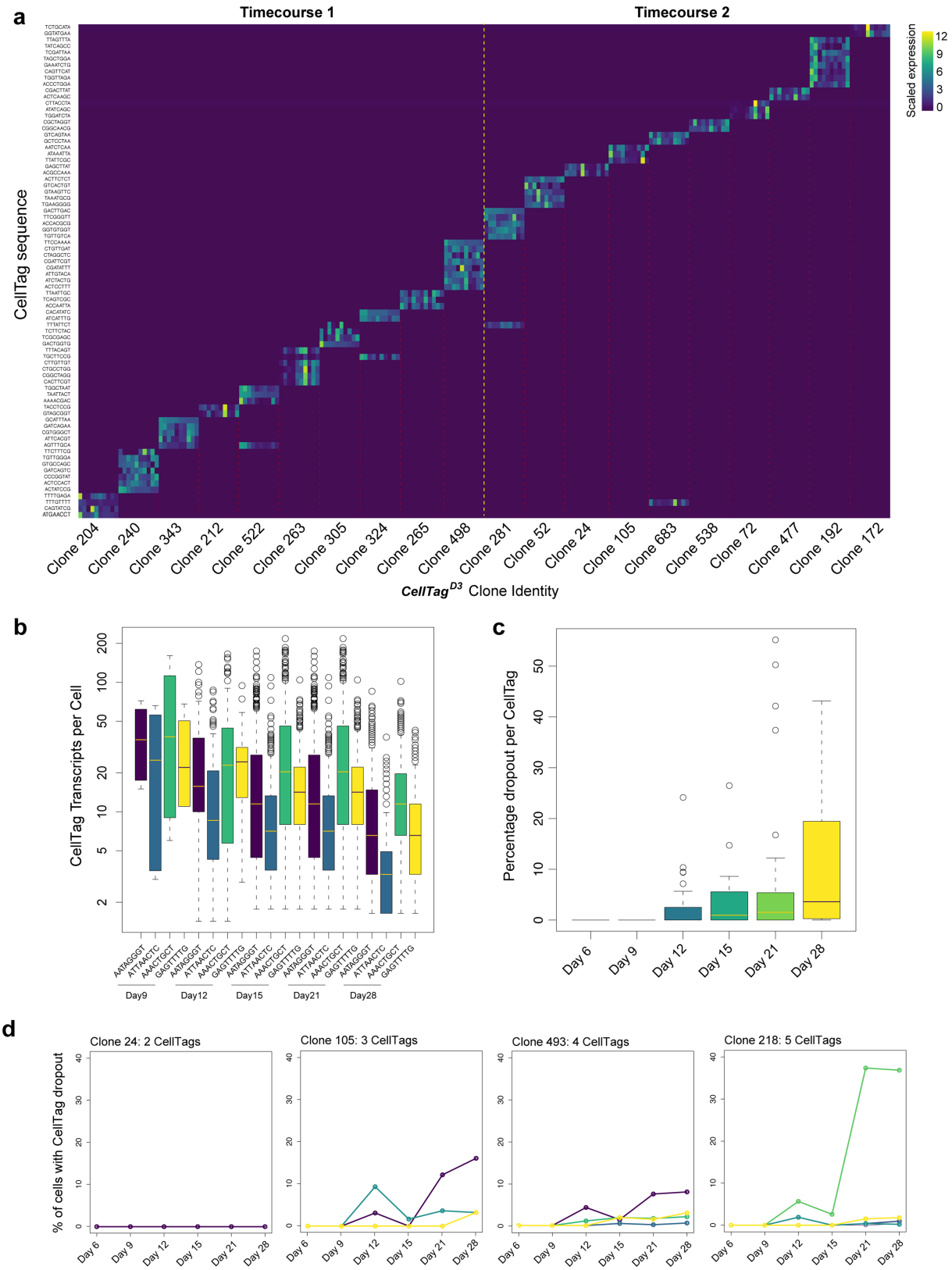


Extended Data Fig. 4 | CellTag expression metrics. a, Mean counts of CellTags expressed per cell, following whitelisting and filtering for time course 1 ($n = 19{,}581$ cells passing filtering) and 2 ($n = 38{,}943$ cells passing filtering), broken down by time point and CellTag version. Red dashed lines denote time of CellTag transduction. **b**, Mean number of CellTags expressed per cell, post-whitelisting and filtering, for each round of barcoding across time courses 1 and 2. $CellTag^{MEF}$: 3.40 ± 0.01 CellTags per cell, $n = 37{,}612$ cells; $CellTag^{D3}$: 4.50 ± 0.02 CellTags per cell, $n = 32{,}176$ cells; $CellTag^{D13}$: 3.20 ± 0.02 CellTags per cell, $n = 10{,}212$ cells. Sixty-five per cent of sequenced cells pass the ≥ 2 CellTag expression threshold to support tracking. **c**, Mean CellTags per cell following whitelisting and filtering for both Drop-seq time courses, broken down by time point. All cells with 200 or more genes were included in this analysis (time course 1: $n = 10{,}038$ cells; time course 2: $n = 9{,}839$ cells). CellTags were introduced only in MEFs, before reprogramming in these experiments. In Drop-seq time courses, we detected a mean of 7.80 ± 0.07 CellTags per cell, across 61% of cells (12,086/19,877 cells) passing the tracking threshold.



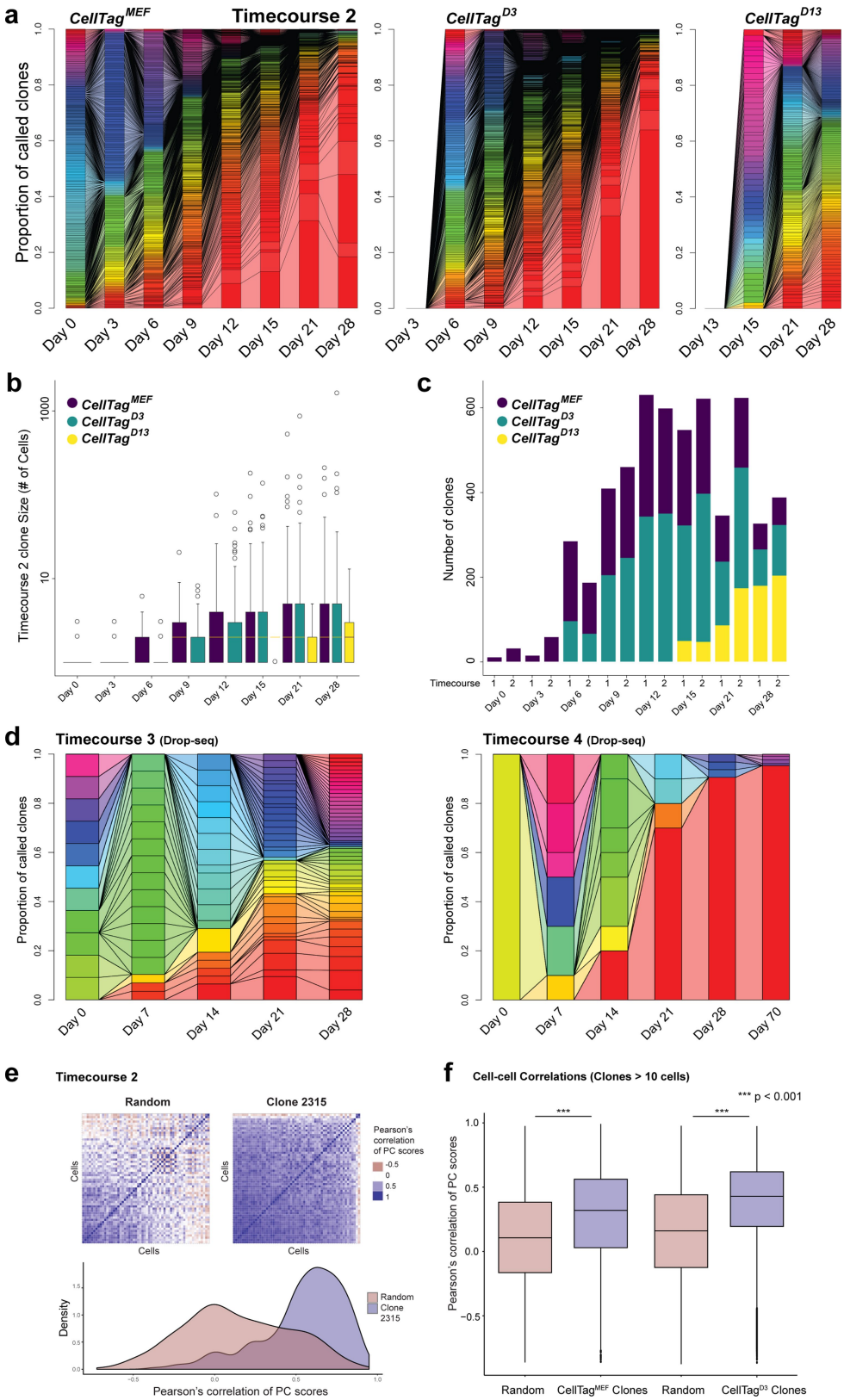
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Assignment of cluster identities based on mRNA and protein expression. a, Top enriched gene expression associated with each cluster, projected onto the reprogramming *t*-SNE plot ($n = 85{,}010$ cells). **b**, Left, expression of *Col1a2*, projected onto the *t*-SNE plot. Top right, violin plot of *Col1a2* expression levels in each cluster. Bottom right, violin plot of *Apoa1* expression levels in each cluster, ordered by gain of expression over the course of reprogramming. Clusters are classified as one of four reprogramming stages: fibroblast, clusters 5, 6, 7, 11; early transition, clusters 0, 3; transition, clusters, 1, 4, 8, 9,10, 12; and reprogrammed, cluster 2. *Apoa1* is not expressed in the fibroblast clusters. **c**, Top, expression of the iEP marker[3,16] *Cdh1* (E-cadherin), projected onto the *t*-SNE plot, highlighting the location of fully reprogrammed cells. Bottom, staining of CDH1 protein in iEP colonies emerging following three weeks of reprogramming (control shown is from Fig. 4d). Scale bar, 20 mm. **d**, Top, expression of the novel iEP marker, apolipoprotein A1, *Apoa1*, projected onto the *t*-SNE plot. Bottom, immunofluorescence of APOA1 protein in an iEP colony, following three weeks of reprogramming. APOA1 (red) is localized to vesicles. This is a representative image selected from five independent biological replicates. Scale bar, 20 μm. **e**, Top, co-expression of *Apoa1* and *Cdh1* at the transcript level within the same individual cells in the fully reprogrammed cluster confirms *Apoa1* as a marker of iEP emergence. Bottom, immunofluorescence of APOA1 and CDH1 protein in iEPs. White arrows mark emerging iEP colonies co-expressing both proteins. APOA1 expression (red) is found localized to vesicles of CDH1-positive cells (green), where the most intense CDH1 staining is observed at cell–cell junctions. This is a representative image selected from three independent biological replicates. Scale bar, 20 μm.



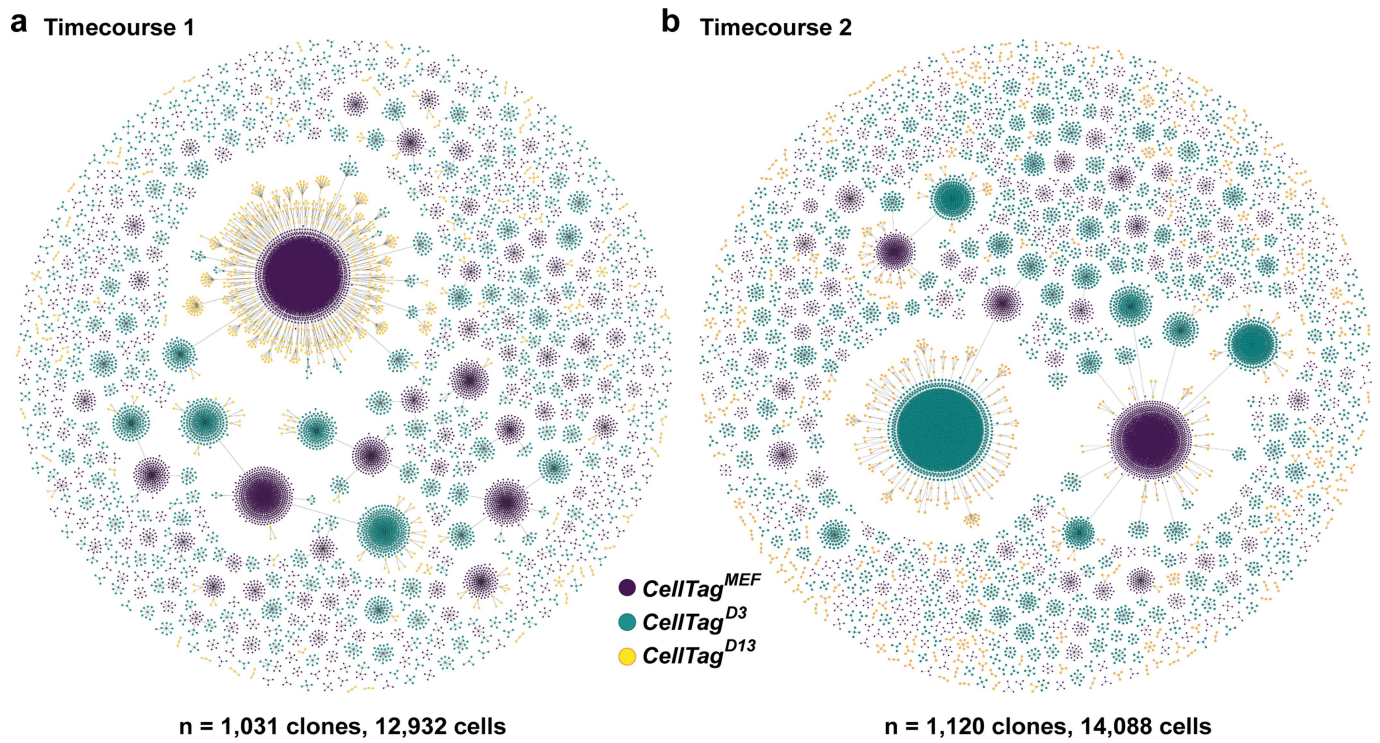
Extended Data Fig. 6 | See next page for caption.

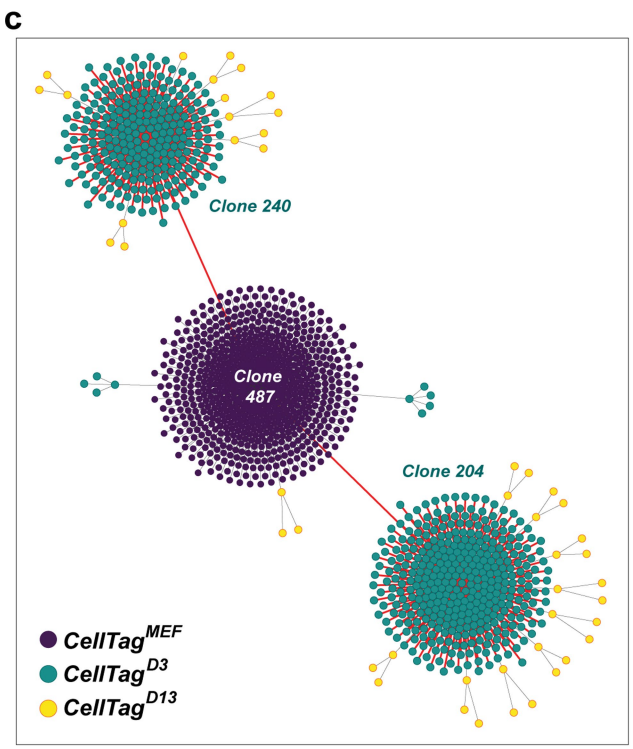
Extended Data Fig. 6 | Combinatorial CellTag labelling to identify clonally related cells. a, Heat map showing scaled expression of individual CellTags in 20 major clones from cells labelled with $CellTag^{D3}$ ($n = 10$ representative cells per clone, time courses 1 and 2). The dashed yellow line marks separation between the two time courses. Dashed red lines mark separation between independent clones. Although some CellTags are shared between these independent biological replicates, the combined CellTag signatures are unique. **b**, Expression levels of individual CellTags per cell over three weeks in a representative clone labelled by four unique CellTags. Expression diminishes over time, but is not completely silenced. **c**, To assess CellTag silencing, we selected 10 major clones ($n = 6{,}728$ cells), defining the intact CellTag signature for each clone at reprogramming day 6. We then assessed loss, or 'dropout' of CellTags from each signature over the time course to day 28. By week 4, expression of an individual CellTag is lost in 1 out of 10 cells—that is, expected CellTag expression was not detected in $11 \pm 2\%$ of cells. Conversely, CellTag expression is retained in almost 90% of cells by day 28. Later rounds of CellTag labelling ($CellTag^{D13}$) are less prone to this effect, with CellTags dropping out in only $3.0 \pm 1.5\%$ of cells. **d**, We mapped CellTag expression across four representative clones, in which expression of each CellTag is plotted over time. The *y* axis denotes the percentage of cells within each clone in which expression of specific CellTags has dropped out. Typically, only one CellTag exhibits dropout, and expression of the other CellTags is maintained. We do not observe complete silencing, that is, loss of expected CellTag expression in 100% of cells. This demonstrates the advantage of our CellTag combinatorial indexing method to reliably label cells and track them over an extended period of time. For example, reliance on the expression of a single, longer barcode would not be effective following integration into a region that later becomes silenced.

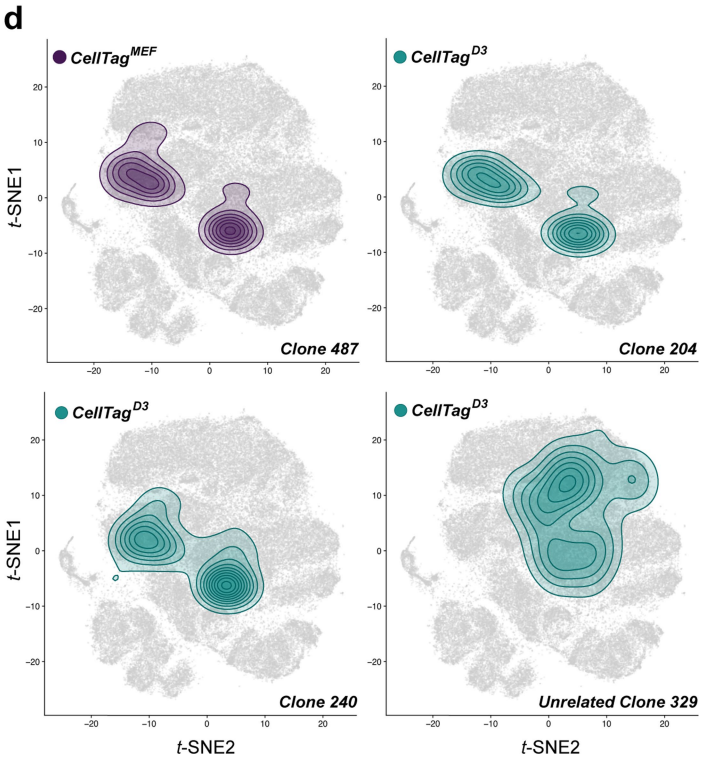


Extended Data Fig. 7 | See next page for caption.

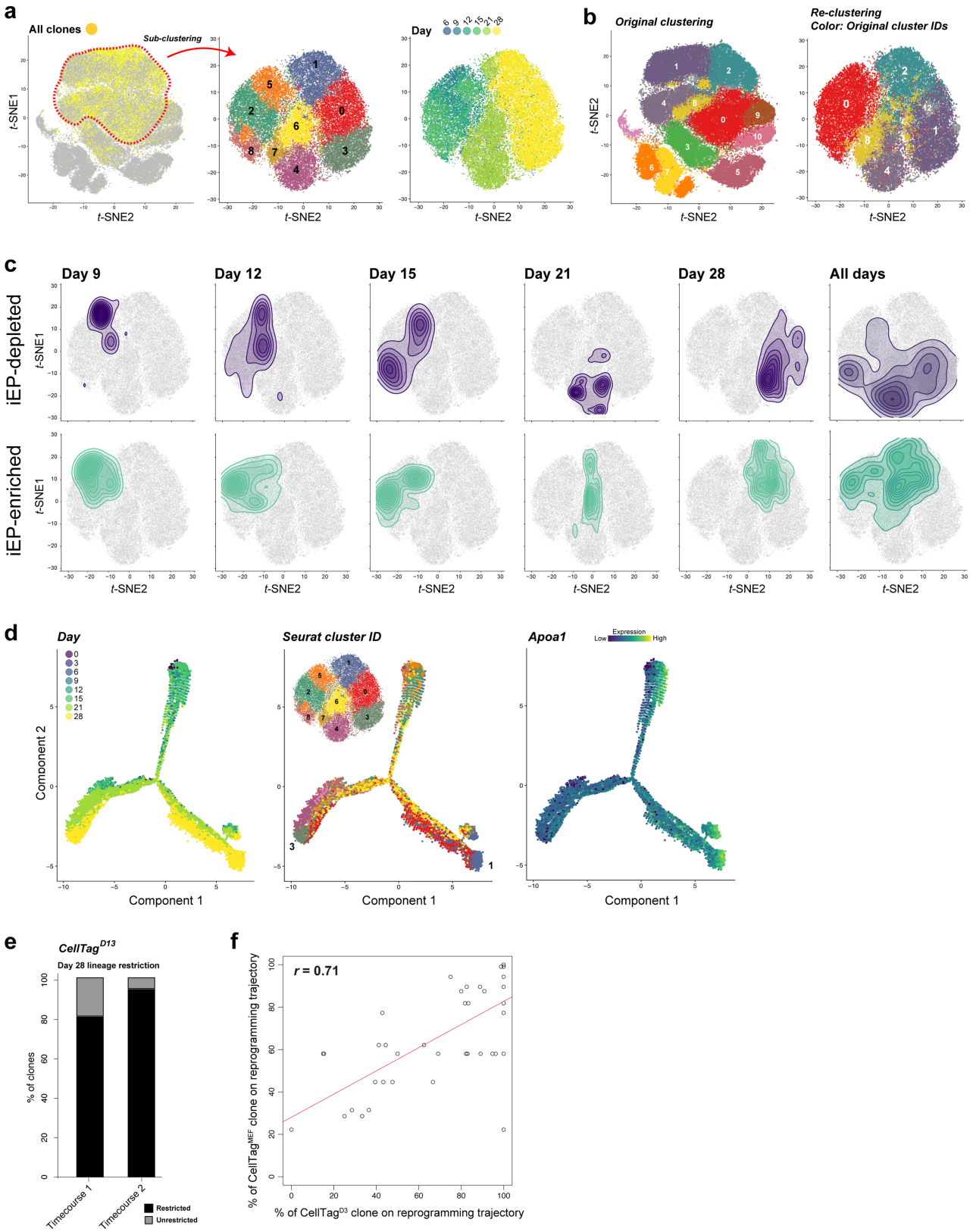
Extended Data Fig. 7 | Visualizing growth of clones and gene expression correlation within clones. a, Connected bar plots showing individual clones as a proportion of all clones at each reprogramming time point for time course 2, for each round of CellTagging ($n = 14{,}088$ cells across 1,120 clones). Connected bars denote clonal expansion and growth over time. **b**, Average number of cells per clone, per time point, for each round of CellTag labelling (time course 2, $n = 1{,}120$ clones). **c**, Number of clones detected at each time point, for each round of CellTagging over reprogramming time courses 1 ($n = 1{,}031$ clones) and 2 ($n = 1{,}120$ clones). The number of clones detected gradually increases over time as the probability of capture increases with clonal growth. The number of clones then begins to decrease as the growth of some individual clones out-competes other clones, which are lost from the population over time. **d**, Connected bar plots showing individual clones as a proportion of all clones called at each reprogramming time point for Drop-seq replicate 1 ($n = 103$ clones) and Drop-seq replicate 2 ($n = 37$ clones). In replicate 2, a single clone progressively dominates the culture over 10 weeks of growth. In our viral integration analyses (Supplementary Table 5), we detect three viral integration sites in the cells of this clone. We did not detect any differential expression of genes proximal to these integration sites. Similarly, analysis of gene expression enrichment in 12 dominant clones across two biological replicates does not reveal any common signature of these clones to explain their rapid expansion (data not shown). This suggests that the clonal growth we observe is a normal part of the iEP reprogramming process, in which the cells enter a progenitor-like state. Even so, these analyses do not exclude the acquisition of genetic and epigenetic changes endowing these expanding clones with increased fitness. **e**, Correlation of principal component (PC) scores in clonally related cells (clone 2315, $n = 58$ cells) relative to a random sampling of cells. Correlation between PC scores was used as a proxy for transcriptional similarity between cells. Clonally related cells were much more closely correlated, relative to randomly selected cells. **f**, Quantification of correlation analysis for all time course 2 clones consisting of 10 cells or more, for $CellTag^{MEF}$ ($n = 78$ clones, 3,963 cells) and $CellTag^{D3}$-labelled clones ($n = 109$ clones, 6,265 cells). Mean correlation scores for clonally related cells are significantly higher than random cell groupings ($P < 0.001$, *t*-test, one-sided). We tagged cells both before and after the 72-h reprogramming window, expecting substantial heterogeneity to be introduced by serial viral transduction. On the contrary, there is only a slight but insignificant increase in PC score correlation between $CellTag^{MEF}$ and $CellTag^{D3}$-labelled, clonally related cells.





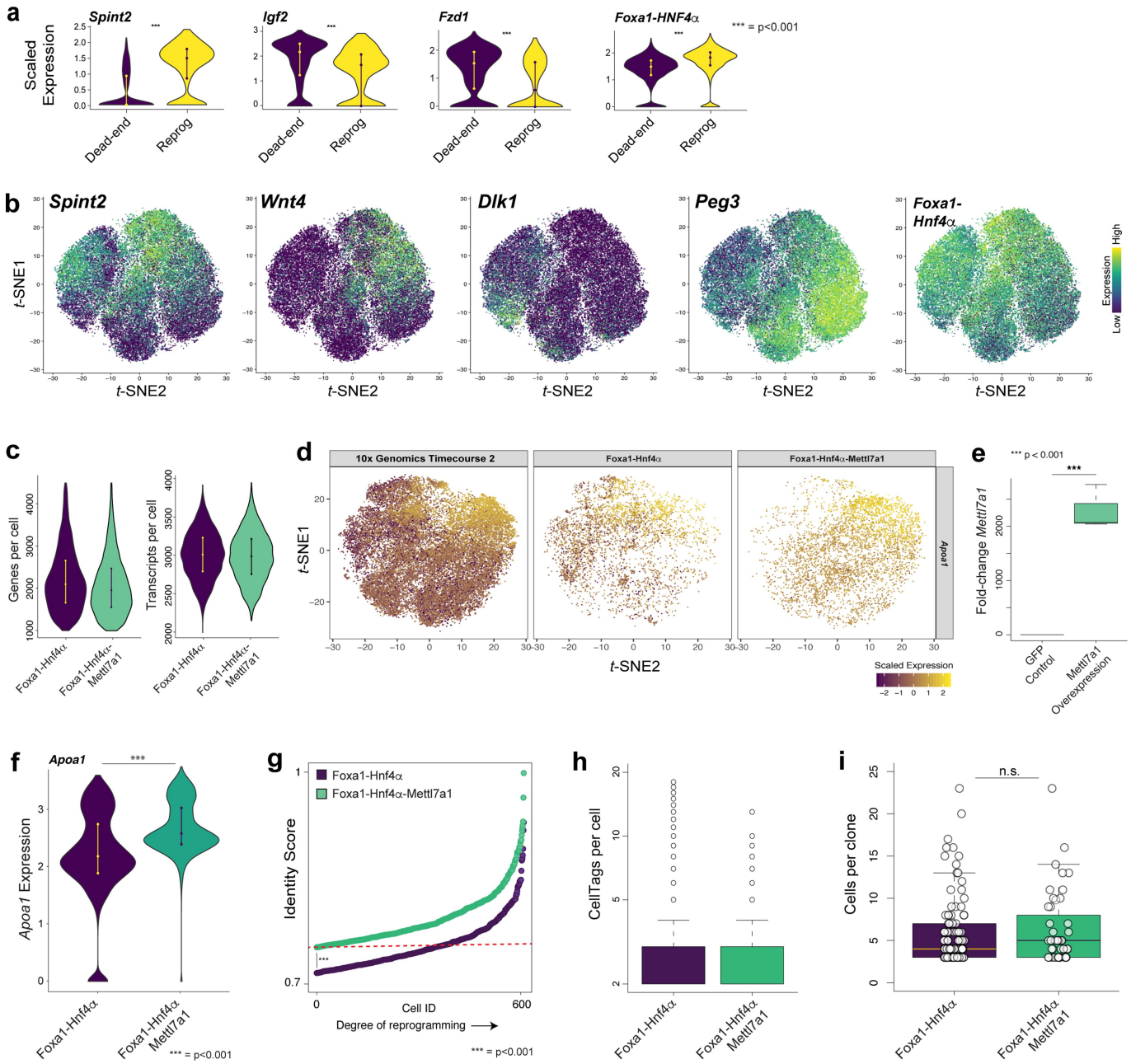


Extended Data Fig. 8 | Reconstruction and visualization of lineages via force-directed graph drawing. a, **b**, Force-directed graph of all clonally related cells and lineages reconstructed from time course 1 (1,031 clones, 12,932 cells) (**a**) and time course 2 (1,120 clones, 14,088 cells) (**b**). All lineages and clone distributions can be interactively explored via our companion website, CellTag Viz (http://www.celltag.org/). **c**, In this tree, we follow CellTag^{MEF} clone 487 from time course 1 and its descendants. Each node represents an individual cell, and edges represent clonal relationships between cells. Purple, CellTag^{MEF} clones; blue, CellTag^{D3} clones; yellow, CellTag^{D13} clones. In the lineage highlighted in red, we follow the CellTag^{MEF} clone ($n = 678$ cells), branching into two CellTag^{D3} lineages (clone 204 ($n = 363$ cells) and clone 240 ($n = 260$ cells)). **d**, Contour plots, representing cell density of each clone, projected onto the *t*-SNE plot, for the lineage shown in **c**. Top left, cells belonging to clone 487 (CellTag^{MEF}). Clones 204 and 240 (CellTag^{D3}) descend from this first clone, exhibiting a high degree of overlap within 2D space, on the *t*-SNE plot. An unrelated CellTag^{D3} clone, 329 ($n = 38$ cells), does not overlap with this lineage, demonstrating the high degree of similarity between cells belonging to the same lineage.



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

Extended Data Fig. 9 | Mapping reprogramming trajectories and timing of cell fate decisions. a, Projection of all clones (yellow, $n = 2{,}151$ clones, 27,020 cells) across reprogramming time courses 1 and 2 ($n = 85{,}010$ cells). A subset of clusters with the highest density of detected clones, outlined in red (clusters 0, 1, 2, 4, 8, and 12), were extracted from this larger dataset and re-clustered to generate a higher-resolution *t*-SNE plot, focusing on reprogramming days 6 to 28 ($n = 48{,}515$ cells). **b**, Left, original cluster identities of all cells ($n = 85{,}010$ cells). Right, subset of 48,515 cells, coloured by original cluster identity. **c**, Contour plots of iEP-depleted clone distribution (top panels, ($n = 7$ clones, 1,037 cells)) and iEP-enriched clone distribution (bottom panels, ($n = 7$ clones, 2,270 cells)) broken down by reprogramming day, and across days 9–28 (far right). These specific clones were selected from the larger iEP-depleted and iEP-enriched groups, as they included cells distributed across all time points, enabling their trajectories to be defined. In these distributions, clusters 8, 4 and 3 are iEP-depleted, thus representing the dead-end trajectory. Conversely, clusters 2, 6 and 1 are iEP-enriched, representing the reprogramming trajectory. These trajectories divide cluster 0 into two halves, but re-clustering does not increase resolution (data not shown). Deeper sequencing of a larger number of cells may provide further insights into this cluster in future studies. **d**, Monocle2 psuedotemporal ordering of cells in the subset of cells ($n = 48{,}515$ cells), coloured by day of reprogramming (left panel), Seurat cluster ID (middle panel) and *Apoa1* expression (right panel). Monocle2 uses dimension reduction to represent each single cell in 2D space and effectively 'connects the dots' to construct a reprogramming trajectory. In this analysis, we performed semi-supervised ordering using *Col1a2* (marking fibroblast identity) expression as a start point and *Apoa1* expression (marking iEP identity) as an endpoint. The branched trajectory generated by Monocle2 is in general agreement with our clonal analyses. **e**, Restriction of $CellTag^{D13}$ clones (time course 1, $n = 79$ clones, 240 cells; time course 2, $n = 30$ clones, 148 cells) to either the reprogrammed cluster (cluster 1), or the dead-end cluster (cluster 3) at day 28. Of the clones from these two biological replicates, $88 \pm 8\%$ exhibit adherence to one of these trajectories by day 13 of reprogramming. **f**, We identified lineages in which multiple $CellTag^{D3}$-labelled clones share a common $CellTag^{D0}$-labelled ancestor. The proportion of each clone on the reprograming trajectory (defined as occupancy of clusters 2, 6 and 1 on the t-SNE plot of the subset of clusters), and proportion of each clone on the dead-end trajectory (defined as occupancy of clusters 8, 4 and 3) was calculated. We then plotted the proportion of each $CellTag^{MEF}$-labelled clone on the reprogramming trajectory against that of its $CellTag^{D3}$-labelled descendants ($r = 0.71$, Pearson's correlation, $n = 13$ lineages, 57 clones, 6,035 cells).



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

Extended Data Fig. 10 | *Mettl7a1* expression is upregulated on the reprogramming trajectory, and promotes iEP generation. a, Violin plots of significantly different gene expression between reprogramming and dead-end trajectories ($n = 2{,}074$ cells). **b**, Projection of gene expression onto the *t*-SNE plot ($n = 48{,}515$ cells). *Wnt4* and *Spint2* expression is significantly upregulated along the reprogramming trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). *Dlk1* and *Peg3* expression is significantly upregulated along the dead-end trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). Expression of the Foxa1-Hnf4a transgene is significantly downregulated along the dead-end trajectory ($P < 0.001$, permutation test, one-sided, $n = 1{,}037$ cells). **c**, Mean numbers of genes and transcripts per cell following 10x Genomics-based scRNA-seq analysis: Foxa1-Hnf4a reprogrammed cells ($n = 6{,}559$ cells) and Foxa1-Hnf4a-Mettl7a1 reprogrammed cells ($n = 10{,}161$ cells), collected 14 days after initiation of reprogramming. For subsequent analyses, the Foxa1-Hnf4a-Mettl7a1 experimental group was randomly downsampled for direct comparison to the Foxa1-Hnf4a experimental group ($n = 6{,}559$ cells for both groups). **d**, The Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 scRNA-seq datasets were merged with cells from time course 2, using canonical correlation analysis[19], to help place these two experimental groups on the previously defined trajectories. Expression levels of *Apoa1* are projected onto this *t*-SNE plot. **e**, Confirmation of *Mettl7a1* expression by qRT–PCR, following transduction of cells with Foxa1-Hnf4a-GFP versus Foxa1-Hnf4a-Mettl7a1 retroviruses ($^{**}P = 5.3 \times 10^{-3}$, *t*-test, one-sided). **f**, Violin plot of mean *Apoa1* expression in cells reprogrammed with Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1. Addition of Mettl7a1 to the reprogramming cocktail results in a significant increase in *Apoa1* expression, supporting observations that this factor increases the yield of fully reprogrammed cells ($P < 0.001$, permutation test, one-sided). **g**, Plot of identity scores of Foxa1-Hnf4a (purple) and Foxa1-Hnf4a-Mettl7a1 (green) reprogrammed cells. Cells are ordered according to an increase in iEP identity. Red dashed line indicates a cut-off of 0.75; above this score cells are considered as iEPs. Threefold-more Foxa1-Hnf4a-Mettl7a1 cells classify as iEPs, relative to Foxa1-Hnf4a cells, represented as a significant increase in iEP score ($P < 0.001$, permutation test, one-sided). **h**, Box plot of mean CellTag expression between Foxa1-Hnf4a (3 ± 0.05 CellTags per cell) and Foxa1-Hnf4a-Mettl7a1 (2.5 ± 0.04 CellTags per cell) experimental groups. The box plots show the median, first and third quantile, and error bar with outliers. **i**, Box plot of cells per clone for Foxa1-Hnf4a and Foxa1-Hnf4a-Mettl7a1 experimental groups, following data processing via our CellTag demultiplexing and clone calling pipeline. Clone size does not significantly differ between these two groups: Foxa1-Hnf4a, 6.0 ± 0.4 cells per clone ($n = 99$ clones, 595 cells); Foxa1-Hnf4a-Mettl7a1: 6.30 ± 0.65 cells per clone ($n = 43$ clones, 277 cells), demonstrating that the addition of Mettl7a1 enhances iEP yield by increasing the number of unique reprogramming events. For comparison, average clone size at ~day 14 for time course replicates 1 and 2 is ~8 cells per clone.

nature research

Corresponding author(s): Samantha A Morris

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	Data was analyzed using the R-based packages, Seurat, Monocle 2, Proxy. Allegro Spring-Electric was used as the layout protocol to render force-directed network graphs. Custom scripts were used for some data processing steps and all code is available on GitHub: https://github.com/morris-lab. ImageJ, Cell Ranger v2.1.0 pipeline, Cystoscope 3.6.0, Allegro Layout were also used.

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

Data

Policy information about availability of data

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

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

All raw data is available on GEO under accession code GSE99915. Processed data and metadata is available at http://celltag.org/. In supplementary data table 1, a list of figures that have associated raw data is available.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	n/a
Data exclusions	Data were excluded from figure 5, via random downsampling to equalize sample size.
Replication	To verify reproducibility of experimental findings, reprogramming timecourses were performed as four independent biological replicates. For investigation of reprogramming efficiency, 11 independent biological replicates were performed. All attempts at replication were successful.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Pooled CellTag libraries are deposited and available from Addgene: pSMAL-CellTag-V1 (https://www.addgene.org/115643); pSMAL-CellTag-V2 (https://www.addgene.org/115644); pSMAL-CellTag-V3 (https://www.addgene.org/115645).

Antibodies

Antibodies used	Primary antibody, goat anti-Apolipoprotein A-I antibody (1:100, Novus Biologicals, NB600-609, lot: 30506) and mouse anti-E-Cadherin (1:50, BD Biosciences, 610181, Clone: 36/E-Cadherin, lot: 7187865)

Validation	Validation is provided on the manufacturers websites, and negative controls were also performed on non-expressing cell lines.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Mouse embryonic fibroblasts were derived from E13.5 mouse embryos.
Authentication	Cell were derived directly from mouse embryos.
Mycoplasma contamination	Cell lines tested as mycoplasma negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.