nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-020-0480-9

Sci-fate characterizes the dynamics of gene expression in single cells

Junyue Cao[1]✉, Wei Zhou[1,2], Frank Steemers[3], Cole Trapnell[1,4] and Jay Shendure[1,4,5,6]✉

Gene expression programs change over time, differentiation and development, and in response to stimuli. However, nearly all techniques for profiling gene expression in single cells do not directly capture transcriptional dynamics. In the present study, we present a method for combined single-cell combinatorial indexing and messenger RNA labeling (sci-fate), which uses combinatorial cell indexing and 4-thiouridine labeling of newly synthesized mRNA to concurrently profile the whole and newly synthesized transcriptome in each of many single cells. We used sci-fate to study the cortisol response in >6,000 single cultured cells. From these data, we quantified the dynamics of the cell cycle and glucocorticoid receptor activation, and explored their intersection. Finally, we developed software to infer and analyze cell-state transitions. We anticipate that sci-fate will be broadly applicable to quantitatively characterize transcriptional dynamics in diverse systems.

During development of organisms, as well as during myriad physiological and pathophysiological processes, individual cells traverse a manifold of molecularly and functionally distinct states. However, although experimental methods for profiling various aspects of single-cell biology have recently proliferated, almost all such methods deliver only a static snapshot of each cell. To address this in part, 'pseudotime' methods computationally place individual cells along a continuous trajectory based on their transcriptomes[1–6]. However, pseudotime infers rather than directly measures transcriptional dynamics, is dependent on sufficient representation along the trajectory and may fail to capture the detailed dynamics of individual cells (for example, directionality, multiple superimposed potentials)[7]. In contrast, time-lapse microscopy can experimentally measure transcriptional dynamics, but is limited to visualization of a few marker genes in a few cells, and as such may be insufficient to decipher the complexity of many biological systems.

In the present study, we describe a technique, sci-fate, for measuring the dynamics of gene expression in large numbers of single cells and at the level of the whole transcriptome. In brief, we integrated protocols for labeling newly synthesized mRNA with 4-thiouridine (4sU)[8,9] with single-cell combinatorial indexing RNA-sequencing (sci-RNA-seq[10]). As a proof of concept, we applied sci-fate to a model system of cortisol response, characterizing expression dynamics in >6,000 single cells. From these data, we quantify the dynamics of the transcription factor (TF) modules that underpin the cell cycle, glucocorticoid receptor (GR) activation and other processes, and develop a framework for inferring the distribution of cell-state transitions. The methods described in the present study may be broadly applicable to quantitatively characterize transcriptional dynamics in diverse systems.

Results

Overview of sci-fate. Briefly, sci-fate relies on the following steps (Fig. 1a): (1) cells are incubated with 4sU, a thymidine analog, to label newly synthesized RNA[11–17]. (2) Cells are harvested, fixed with 4% paraformaldehyde (PFA), and then subjected to a thiol (SH)-linked alkylation reaction, which covalently attaches a carboxyamidomethyl group to 4sU by nucleophilic substitution[8]. (3) Cells are distributed by dilution to four 96-well plates. The first sci-RNA-seq molecular index is introduced via in situ reverse transcription (RT) with a poly(T) primer bearing both a well-specific barcode and a degenerate unique molecular identifier (UMI). During first-strand complementary DNA synthesis, modified 4sU is a template for guanine rather than adenine incorporation. (4) Cells from all wells are pooled and then redistributed by FACS to multiple 96-well plates. (5) Double-stranded complementary DNA is synthesized. After Tn5 transposition, cDNA is PCR-amplified via primers recognizing the Tn5 adaptor on the 5′-end and the RT primer on the 3′-end. These primers also bear a well-specific barcode that introduces the second sci-RNA-seq molecular index. (6) PCR amplicons are subjected to massively parallel DNA-sequencing. As with other sci-methods[10,18–25], most cells pass through a unique combination of wells, such that their contents are marked by a unique combination of barcodes that can be used to group reads derived from the same cell. (7) The subset of each cell's transcriptome corresponding to newly synthesized transcripts is distinguished by T→C conversions in reads mapping to mRNAs (see Methods).

For quality control, we first tested sci-fate with a mixture of HEK293T (human) and NIH/3T3 (mouse) cells under four conditions: with versus without 4sU labeling (200 nM, 6 h), and with versus without the SH-linked alkylation reaction. The resulting transcriptomes were overwhelmingly species coherent (>99% purity for both human and mouse cells, 2.7% collisions; see Supplementary Fig. 1a,b) with similar mRNA recovery rates (overall median 21,342 UMIs per cell; see Supplementary Fig. 1c). However, only with 4sU labeling and SH-linked alkylation did we observe a substantial proportion of reads bearing T→C conversions, that is, newly synthesized transcripts (46% and 31% for treated human and mouse cells, respectively, versus 0.8% for untreated cells; see Supplementary Fig. 1d). The aggregated transcriptomes of cells derived from sci-fate and conventional sci-RNA-seq were highly correlated (Spearman's correlation $r = 0.99$; see Supplementary Fig. 1e,f), suggesting that short-term labeling and conversion do not substantially bias transcript counts.

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA. [3]Illumina, San Diego, CA, USA. [4]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [5]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. [6]Howard Hughes Medical Institute, Seattle, WA, USA. ✉e-mail: cao1025@uw.edu; shendure@uw.edu

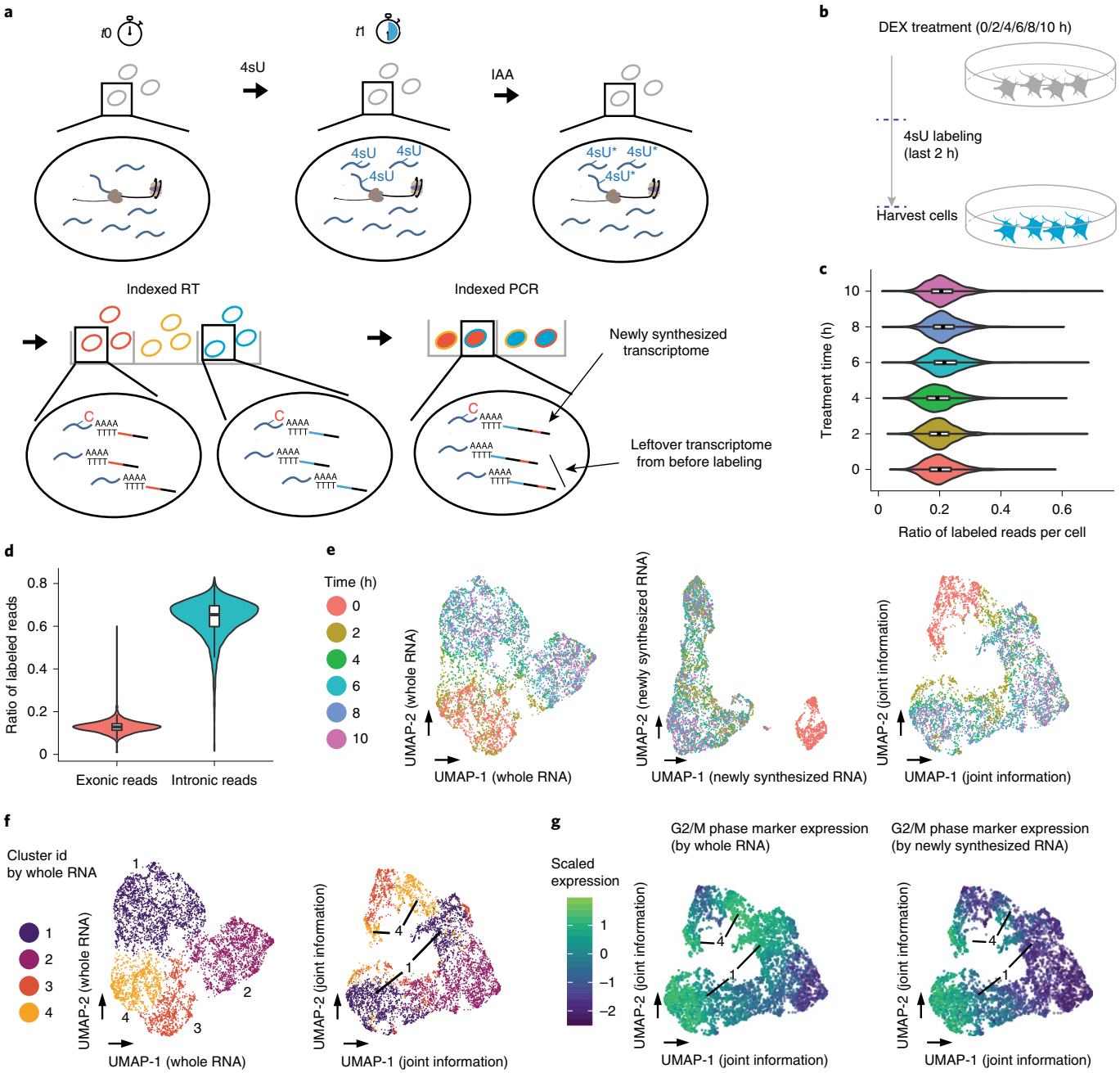

Fig. 1 | Sci-fate enables joint profiling of whole and newly synthesized transcriptomes. a, The sci-fate workflow. Key steps are outlined in the text. IAA, iodoacetamide. Asterisk, chemically modified 4sU. **b**, Experimental scheme. A549 cells were treated with dexamethasone for varying amounts of time ranging from 0 h to 10 h. Cells from all treatment conditions were labeled with 4sU 2 h before harvest for sci-fate. **c**, Violin plot showing the fraction of 4sU-labeled reads per cell for each of the six treatment conditions. Cell number, $n = 1{,}054$ (0 h), 1,049 (2 h), 949 (4 h), 1,262 (6 h), 1,041 (8 h) and 1,325 (10 h). For all violin plots in this figure: thick lines in the middle are the medians; upper and lower box edges are the first and third quartiles, respectively; whiskers are 1.5 times the interquartile range; and circles are outliers. **d**, Violin plot showing the fraction of 4sU-labeled reads per cell ($n = 6{,}680$), split out by the subsets that map to exons versus introns. **e**, UMAP visualization of A549 cells ($n = 6{,}680$) based on their whole transcriptomes (left), newly synthesized transcriptomes (middle) or joint analysis, that is, combining the top PCs from each (right). **f**, Same as left and right of **e**, respectively, but colored by cluster ID from UMAP based on whole transcriptomes. **g**, Same as right of **e**, but colored by normalized expression of G2/M-marker genes by their overall expression levels (left) or their levels of newly synthesized transcripts (right). UMI counts for these genes are scaled by library size, log(transformed), aggregated and then mapped to *Z*-scores.

Profiling of transcriptome dynamics in cortisol response. To investigate the transcriptional dynamics of cortisol response[26], we applied sci-fate to an in vitro model wherein dexamethasone (DEX), a synthetic mimic of cortisol, activates glucocorticoid receptor (GR), which binds to thousands of locations across the genome and rapidly alters gene expression[27–30]. Specifically, we treated lung adenocarcinoma-derived A549 cells for 0, 2, 4, 6, 8 or 10 h with 100 nM DEX. In each condition, cells were incubated with 4sU (200 nM) for the 2 h immediately preceding harvest. We then performed a 384 × 192 sci-fate experiment (Fig. 1b). Each of the 6 conditions was

represented by 64 wells during the first round of indexing, such that all samples could be processed in a single sci-RNA-seq experiment to minimize batch effects.

After filtering out low-quality cells, potential doublets and a small subgroup of differentiated cells (see Methods), we obtained single-cell profiles for 6,680 cells (median 26,176 UMIs corresponding to mRNAs detected per cell). A median of 20% mRNA UMIs were labeled per cell (Fig. 1c; see Supplementary Fig. 2a–c). The proportion of newly synthesized mRNAs was markedly higher in reads mapping to intronic (65%) versus exonic (13%) regions ($P < 2.2 \times 10^{-16}$, two-sided Wilcoxon's signed-rank test; Fig. 1d, and see Supplementary Fig. 2d,e), consistent with the expectation that the intronic reads are more likely to have been recently synthesized. We also compared intronic reads and newly synthesized mRNA for RNA-velocity analysis[31] and observed a subjectively consistent picture, suggesting that they capture similar information (see Supplementary Fig. 2f).

In exploring these data, we first asked whether the newly synthesized versus whole-transcriptome data convey identical or distinct information with respect to cell state. Performing dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP)[32] on whole transcriptomes failed to separate DEX-untreated (0 h) versus DEX-treated (2+ h) cells (Fig. 1e, left, and see Supplementary Fig. 2g). In contrast, applying UMAP to the newly synthesized subset of the single-cell transcriptomes readily separated DEX-untreated versus -treated cells (Fig. 1e, center). These patterns are probably a consequence of the fact that, in DEX-treated cells, the newly synthesized transcriptome more faithfully reflects the GR response itself. Illustrating this, the classic markers for GR response, *FGD4* (ref. [27]) and *FKBP5* (ref. [33]), exhibited the highest-fold induction in comparisons of the newly synthesized transcriptome at 0 h versus 2 h, but the magnitude of their induction was dampened in comparisons of the whole transcriptome between the same time points (see Supplementary Fig. 2h,i and Supplementary Table 1).

To jointly make use of the information conveyed by the whole and newly synthesized transcriptomes, we combined their top principal components (PCs) for UMAP analysis. This approach separates cells that had experienced no (0 h), recent (2 h) or extended (4+ h) DEX treatment (Fig. 1e, right). With this joint approach, cells corresponding to two clusters defined by analysis of whole transcriptomes (clusters 1 and 4; Fig. 1f, left) each split into two groups (Fig. 1f, right). Examining the levels of newly synthesized mRNAs corresponding to cell-cycle markers[34], one pair of these new groups corresponds to G2/M-phase cells (high levels of both overall and newly synthesized G2/M markers), and the other to early G0/G1-phase cells (high levels of overall but low levels of newly synthesized G2/M markers) (Fig. 1g, and see Supplementary Fig. 2j,k). Of note, cells from the 2-h time point exhibited a distribution of cell-cycle states according to this joint information (Fig. 1e,g). Overall, these analyses illustrate how joint analysis of the newly synthesized and whole components of single-cell transcriptomes can recover cell-state information that is not easily obtained from the whole transcriptomes alone.

TF module activity decomposes cellular processes. Multiple, dynamic gene, regulatory processes are concurrently under way in this in vitro system—minimally, the GR response and the cell cycle. We speculated that these might be disentangled, and their intersection probed, by first identifying the TF modules driving new mRNA synthesis in relation to each process.

TF modules, comprising candidate links between TFs and their regulated genes, were identified as follows: for each gene, across the 6,680 cells, we computed correlations between the levels of newly synthesized mRNA for that gene and the overall expression level of each of the 859 expressed TFs, using LASSO (least absolute shrinkage and selection operator) regression. Out of 1,086 links involving TFs characterized by ENCODE[35], 807 were validated by TF-binding sites near the genes' promoters[35], a 4.3-fold enrichment relative to background expectation (odds ratio for validation = 2.89 for LASSO-identified links versus 0.67 for background, $P < 2.2 \times 10^{-16}$, two-sided Fisher's exact test). These covariance links were further filtered by chromatin immunoprecipitation–sequencing (ChIP–seq) binding[36], and supplemented with additional covariance links validated by motif[37] enrichment analysis (Fig. 2a; see Methods). Altogether, we identified 986 links between 29 TFs and 532 genes (see Supplementary Fig. 3a,b and Supplementary Table 2). As a control, we permutated the cell order of the cell × TF expression matrix (T_i) (see Methods), and then repeated the analysis. No links were identified after permutation. Some of the identified TF–gene regulatory relationships are supported by a manually curated database of TF networks (TRRUST[38]), for example E2F1 (top enriched TF of *E2F1*-linked genes = E2F1, adjusted $P = 8 \times 10^{-7}$)[39], NFE2L2 (top enriched TF of *NFE2L2*-linked genes = NFE2L2, adjusted $P = 0.003$)[39] and SREBF2 (top enriched TF of *SREBF2*-linked genes = SREBF2, adjusted $P = 0.0006$)[39].

The 29 TFs with one or more gene links included well-established GR response effectors such as *CEBPB*[40], *FOXO1* (ref. [41]) and *JUNB*[42] (see Supplementary Fig. 3c,d). This group also included several TFs not previously implicated in the GR response, including *YOD1* and *GTF2IRD1*, both of which exhibited greater expression and activity in DEX-treated cells (see Supplementary Fig. 3e,f). The main TFs driving cell-cycle progression were also identified, for example *E2F1*, *E2F2*, *E2F7*, *BRCA1* and *MYBL2* (ref. [43]). Notably, the expression levels of TFs such as *E2F1* were more highly correlated with the levels of newly synthesized than overall target gene mRNAs (see Supplementary Fig. 3g). We also observed regulatory links corresponding to TFs involved in cell differentiation such as *GATA3* (ref. [44]), mostly expressed in a subset of quiescent cells, as well as TFs involved in oxidative stress response such as *NRF1* (ref. [45]) and *NFE2L2* (*NRF2*)[46].

We calculated a measure of each of these 29 TFs' activities in each cell, based on the normalized aggregation of the levels of newly synthesized mRNA for all its target genes. We then computed the absolute correlation coefficient between each pair of TFs with respect to their activity across the 6,680 cells. Hierarchical clustering of these pairwise correlations identified several major TF modules, that is sets of TFs that appear to be regulating the same process (Fig. 2b). A first large TF module corresponds to all cell cycle-related TFs in the set, for example *E2F1* and *FOXM1* (ref. [43]). A second large TF module corresponds to GR response-related TFs, for example *FOXO1*, *CEBPB*, *JUNB* and *RARB*[40–42]. The other modules include one corresponding to GR-activated G1/G2/M-phase cells (*KLF6*, *TEAD1* and *YOD1*; see Supplementary Fig. 3h), and another corresponding to probably differentiating, GR-activated, G1-phase cells (*GATA3* and *AR*; see Supplementary Fig. 3h)[44,47]. Additional TFs or TF modules appear to capture other processes that are heterogeneous in this population of cells, including *NRF1* and *NFE2L2* for stress response/apoptosis (top enriched pathway of *NFE2L2*-linked genes: ferroptosis, adjusted $P = 1 \times 10^{-5}$)[39,45,46,48], *KLF5* for DNA-damage repair (top pathway: ATM signaling, adjusted $P = 0.018$)[39,49] and *SREBF2* for cholesterol homeostasis (top pathway: 'SREBF and miR33 in cholesterol and lipid homeostasis', adjusted $P = 9 \times 10^{-6}$)[39,50].

To assign cell-cycle states to individual cells, we first ordered cells by their cell cycle-linked TF module activity. This resulted in a smooth, almost circular trajectory, in which the levels of newly synthesized mRNA corresponding to known cell-cycle markers was dynamic (Fig. 2c)[34]. We observed a gap between late G2/M phase and early G1 phase, consistent with the dramatic cell-state change during cell division. By unsupervised clustering of the activities of individual TFs within the cell cycle-linked TF module, we identified nine cell-cycle states spanning the early, middle and late cell-cycle

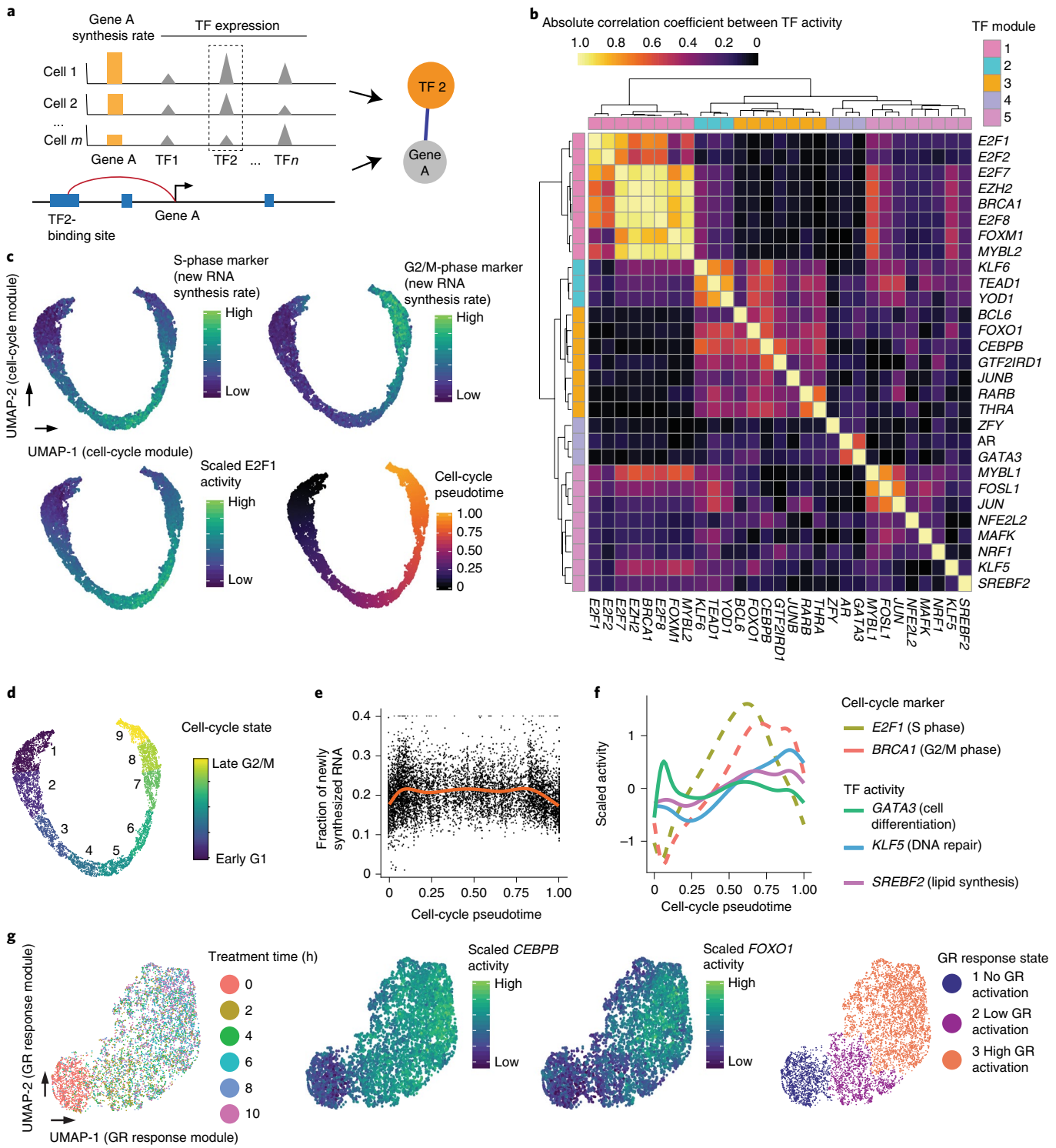

Fig. 2 | Characterizing TF modules driving concurrent, dynamic gene, regulatory processes in populations of single cells. a, Schematic of approach used to identify links between TFs and their regulated genes. **b**, Heatmap showing the absolute Pearson's correlation coefficient between the activities of pairs of TFs (cell number, $n = 6{,}680$). **c**, UMAP visualization of A549 cells ($n = 6{,}680$) based on the activity of cell cycle-related TF module, colored by levels of newly synthesized mRNA corresponding to S-phase markers (top left), G2/M-phase markers (top right) and *E2F1* activity (bottom left). The bottom right panel is colored by pseudotime based on the point position on the principal curve estimated using the princurve package[64]. **d**, Same as **c**, but colored according to nine cell-cycle states defined by unsupervised clustering analysis. In broad terms, cell-cycle states 1–3 correspond to G1 phase, 4–6 to S phase and 7–9 to G2/M phase. **e**, Scatter plot showing the changes in the fraction of newly synthesized mRNA in each cell ($n = 6{,}680$) along cell-cycle progression. The red line is the smoothed curve estimated using the geom_smooth function[65]. **f**, Similar to **e**, but showing smoothed activity of selected TF modules as a function of cell cycle pseudotime. **g**, UMAP visualization of A549 cells ($n = 6{,}680$) based on the activity of GR response-related TF module, colored by DEX treatment time (left), *CEBPB* or *FOXO1* activity (middle panels), or cluster ID from unsupervised clustering (right). Throughout the figure, to calculate the TF module activity, newly synthesized UMI counts for genes linked to module-assigned TFs are scaled by library size, log(transformed), aggregated and then mapped to *Z*-scores.

phases (Fig. 2d). Early G1- and late G2/M-phase cells exhibited decreased synthesis of new RNA relative to other parts of the cell cycle, possibly due to chromosomal condensation during mitosis (Fig. 2e)[51–53]. Other (that is, non-cell-cycle) TF modules exhibited different dynamics in relation to cell-cycle progression (Fig. 2f). For example, *GATA3* activity peaks in the early G1 phase, potentially reflecting a cell-differentiation pathway distinct from cell-cycle reentry[44]. In contrast, the modules of *KLF5* and *SREBF2*, associated with DNA repair and lipid homeostasis, respectively, exhibited greater activity from S phase to G2 phase, possibly related to roles in DNA replication and cell division, respectively[54].

Similarly, the cells can also be ordered into a smooth trajectory based on GR response-linked TF module activity. As expected, this trajectory correlates well with DEX treatment time, as well as the activity of GR response-related TFs (Fig. 2g). By unsupervised clustering of the activities of individual TFs within the GR response-linked TF module, we identified GR response states corresponding to no, low and high levels of activation (Fig. 2g).

We next sought to explore the intersection of the nine cell-cycle states (Fig. 2d) and the three GR response states (Fig. 2g). Each of 27 possible state combinations was represented by some cells, with the smallest group corresponding to 1.1% of the overall dataset ($n=74$ cells, intersection of 'early G2/M' cell-cycle state and 'no GR activation' state; see Supplementary Fig. 4a,b). Although we observe several TF modules that appear specific to certain intersections of the cell cycle and GR response (*KLF6/TEAD1/YOD1* and *GATA3/AR*; Fig. 2b), several observations support the conclusion that the dynamics of the cell cycle and GR response operates largely independently. First, we observe minimal correlation between the activities of the primary TF modules for the cell-cycle and the GR response across the 6,680 cells (Pearson's correlation $r=0.004$; Fig. 2b). Second, the relative proportions of each of the 27 possible state combinations are readily predicted by proportions of cell-cycle and GR response states, that is with no interaction term (see Supplementary Fig. 4b).

Inferring single-cell transcriptome dynamics with sci-fate. We next sought to develop a strategy to use sci-fate data to infer the past transcriptional state of each cell, that is at the onset of 4sU labeling, which might in turn allow us to relate cells derived from different time points. The inference of past transcriptional state requires knowledge of two parameters—first, the detection rate of newly synthesized transcripts (that is, the proportion of newly synthesized transcripts containing one or more detected T→C mutations) and, second, the degradation rate of each mRNA species. In this section, we discuss how each of these parameters can be estimated directly from the sci-fate data generated for this experiment. A more detailed consideration is provided in Methods.

Under the assumption that mRNA degradation rates are not affected by DEX treatment (this assumption is validated in the following paragraph), it is relatively straightforward to estimate sci-fate's detection rate for newly synthesized transcripts. Each sci-fate transcriptome in this dataset consists of two components—the newly synthesized transcriptome, the detection rate of which we hope to estimate, and the 'leftover' transcriptome, that is transcripts that were present at the onset of 4sU labeling, minus any degradation over the course of the 2 h. Comparing the 0-h (untreated) and 2-h DEX treatment groups, we expect that their leftover bulk transcriptomes (at the onset of 2-h 4sU labeling) should be identical, as should sci-fate's detection rate for newly synthesized transcripts. As such, an equation can be constructed relating the transcriptomes of these treatment groups to each other (see Methods). For each of the 186 genes exhibiting the largest differences in new transcription between the two conditions, we solved this equation to estimate sci-fate's detection rate. As these estimates were largely consistent across genes and robust to sequencing depth (see Supplementary Fig. 5a–e), we used their median value (82%) as sci-fate's estimated detection rate for all subsequent analyses.

We next sought to estimate the degradation rate of each mRNA species. As noted above, the bulk transcriptome at each time point in our experiment can be broken down into the newly synthesized transcriptome and the leftover transcriptome. Furthermore, the leftover transcriptome should equal the bulk transcriptome from the time point 2 h earlier, provided that we correct for mRNA degradation over that interval. From these assumptions, an equation can be constructed and solved to estimate the mRNA half-life of each gene, which we did independently for each 2-h interval of the experiment (see Methods and Supplementary Table 3). As a first quality check, we simply compared these estimated mRNA degradation rates between time points, and found them to be both consistent and robust to sequencing depth (see Supplementary Fig. 5f,g; median Pearson's $r=0.92$). As a second quality check, we compared them with orthogonally generated estimates of mRNA half-lives from the literature[9]. Despite the fact that different technologies were used on different cell lines (A549 versus K562), the estimates of mRNA half-lives were reasonably consistent (see Supplementary Fig. 5h; Pearson's $r=0.76$). Of note, the absolute differences in estimated mRNA half-lives between sci-fate and previous techniques could be due to the use of different cell lines or systematic differences between the techniques.

With these parameters in hand, we next estimated the past transcriptional state of each cell in our dataset (see Methods and Supplementary Fig. 6a,b), and sought to use these estimated states to link individual cells to each other across time points (Fig. 3a). Specifically, for each cell B (for example, a cell from the 2-h time point), we used a recently developed alignment method[34] to identify a cell A profiled at an earlier time point (for example, a cell from the 0-h time point), wherein A's current state was closest to B's estimated past state. In this framework, A can be regarded as the parent state of B. Applying this strategy to each of the five intervals comprising our experiment, we constructed a set of linkages spanning the entire dataset and time course (Fig. 3b).

A key contrast with conventional pseudotime is that, with sci-fate, each cell is now characterized not only by its present state, but also by specific linkages to a series of distinct cells matching its predicted past and/or future states (Fig. 3c). To evaluate whether these mini-trajectories contain structure, we applied UMAP and unsupervised clustering, which resulted in three distinct trajectory clusters (Fig. 3d). To annotate these, we checked the proportions of each of the aforementioned three GR response states and nine cell-cycle states in each of them, as a function of time. As expected, all three trajectories exhibited a rapid transition from no GR activation to low/high GR activation (Fig. 3e). However, each trajectory appears to correspond to a different starting point with respect to the cell cycle (Fig. 3f). Trajectory 1 corresponds to cells that transition from G2/M to G1 phase over the course of the 10-h experiment. Trajectory 2 corresponds to cells that transition from late S phase to G2/M phase over the course of the experiment. Trajectory 3 corresponds to cells that transition from G1 to either S phase or G1 arrest over the course of the experiment. The inference of G1 arrest subsequent to DEX treatment is consistent with the dynamics of cell-state proportions in this experiment as well as with previous research[55,56]. As a control, we clustered the cell-state transition trajectories by simply aligning neighboring time points without knowledge of newly synthesized mRNA; this failed to recover the expected cell-cycle dynamics (see Supplementary Fig. 6c).

Inferred cell transitions recapitulate expected dynamics. We next sought to evaluate whether the distribution of cell-state transitions inferred by sci-fate are consistent with the expected dynamics. We assigned each cell into one of the 27 states (3 GR response × 9 cell-cycle states) and computed a cell-state transition network (Fig. 4a),

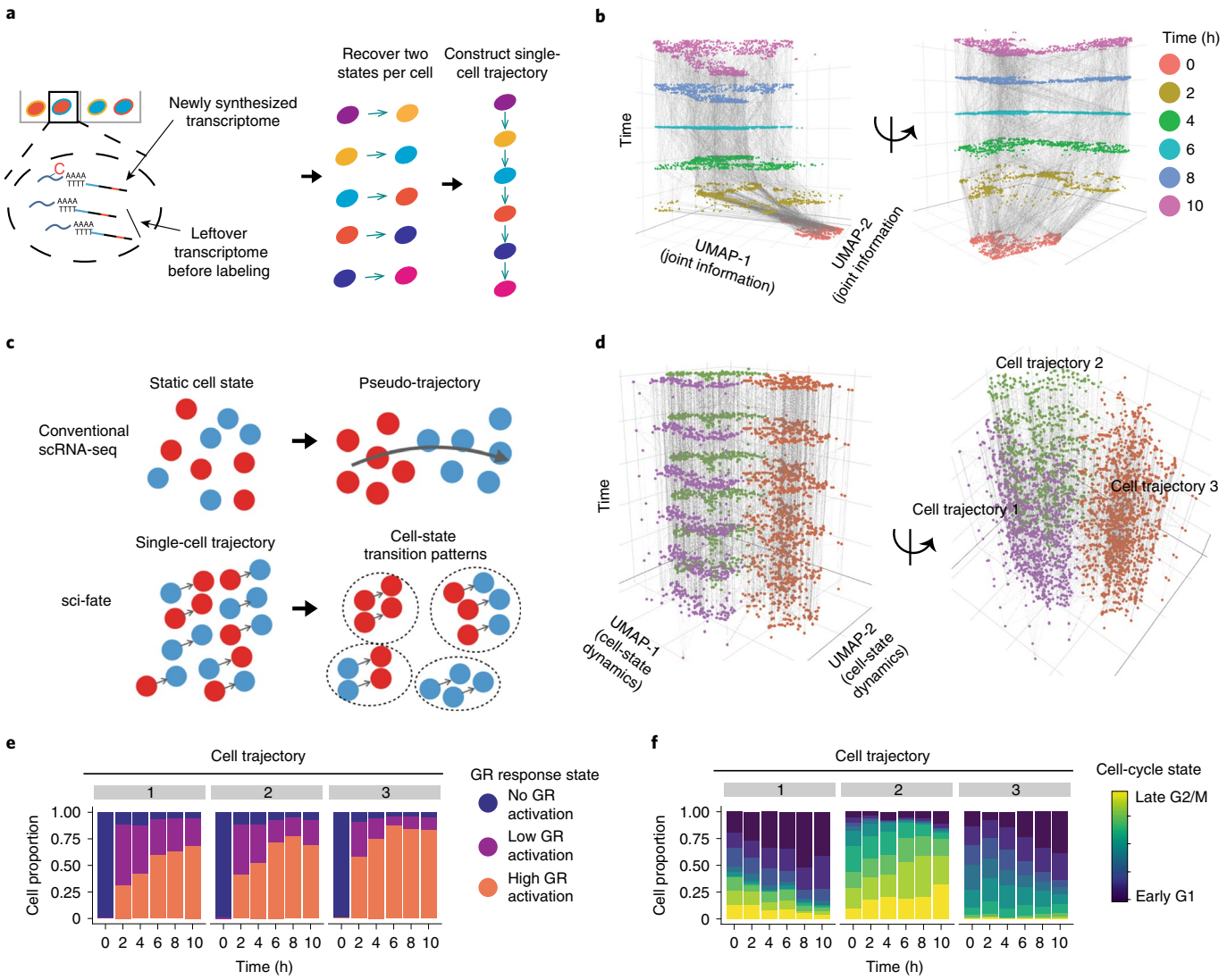

Fig. 3 | Inferring single-cell transcriptional dynamics with sci-fate. a, Schematic of approach for linking cells based on estimated past transcriptional states to reconstruct single-cell transition trajectories. **b**, Three-dimensional plot of all cells (cell number, $n = 6{,}680$). The *x* and *y* coordinates correspond to the joint information UMAP space shown in the far-right panel of Fig. 1e. The *z* coordinate as well as colors correspond to DEX treatment time. Linked parent and child cells are connected with gray lines. **c**, Schematic comparing conventional scRNA-seq and sci-fate for cell trajectory analysis. **d**, Similar to **b**, except the *x* and *y* coordinates correspond to the UMAP space based on the single-cell transition trajectories across the six time points (cell number, $n = 6{,}680$). **e**,**f**, Barplots showing the contributions of the three GR response states (**e**) and the nine different cell-cycle states (**f**) to each of three cell-trajectory clusters.

with the assumption that the cell-state transitions in this experiment follow a Markov process, with transition probabilities that do not change over time. This assumption is validated in part by the observation that the distributions of predicted cell-state transitions, estimated from the last three time intervals (4–6 h, 6–8 h, 8–10 h), are highly correlated with each other (see Supplementary Fig. 7a) despite varied cell-state proportions at 4 h versus the later time points (see Supplementary Fig. 7b). Consistent with DEX treatment, transitions are highly biased from the G1 to S, S to G2/M and G2/M to G1 phase of the cell cycle (Fig. 4a). As a control analysis, cell-state transition networks were similarly derived, but based either on randomly permutated cell-state transition links or on links derived from mature mRNAs only; these both failed to recapitulate the expected pattern of cell-cycle transitions (see Supplementary Fig. 7c).

The 27 states shown in Fig. 4a each correspond to subsets of cells, the transcriptomes of which are similar, making use of the joint information provided by distinguishing between old (>2 h) and new (<2 h) transcripts. it corresponds to expected phenomena, for example irreversible progression through the GR response, as well as irreversible progression through the cell cycle. As examples, S-phase cells without GR activation (0-h treatment) mostly transit into S-phase cells with GR activation (2-h treatment), whereas G2/M-phase cells with no GR activation (0-h treatment) mostly transit to G2/M- or G1-phase cells with GR activation (2-h treatment) (Fig. 4b). For comparison, overlaying the same UMAP coordinates with RNA-velocity vectors[31] recovered similar patterns, but only when treatment time information was incorporated into the RNA-velocity analysis (see Supplementary Fig. 7d).

Can we use this framework for a better understanding of the characteristics of transcriptional states that govern their dynamics? As a first approach, we calculated pairwise Pearson's distance between the aggregated transcriptomes of each of the 27 states. As expected, the greater the distance between any pair of states, the lower the proportional representation of that transition in the

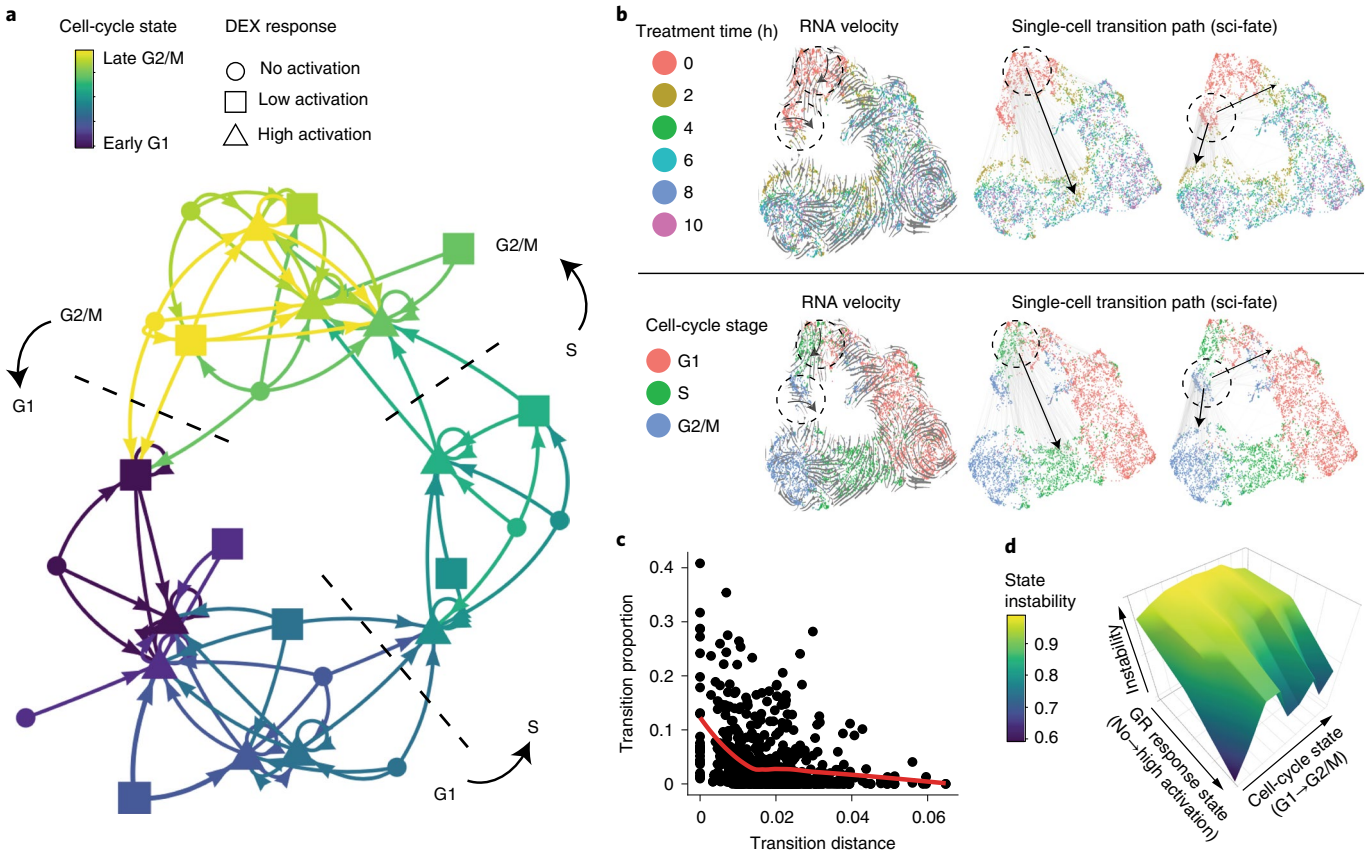

Fig. 4 | Constructing a state transition network for GR response and cell cycle. a, Cell-state transition network. The nodes are 27 cell states characterized by combinations of cell-cycle and GR activation states. The links represent frequent cell-state transition trajectories (transition proportion >10%) between cell states. This threshold for defining a link corresponds to approximately 2 s.d. from the mean transition proportion calculated after permutating cell-transition links ($n = 729$). **b**, The *x* and *y* coordinates correspond to the joint information UMAP space shown in the far-right panel of Fig. 1e, colored by DEX treatment time (top) or inferred cell-cycle state (bottom). Gray lines represent inferred cell-state transition links between parent and child cells (middle: cell-state transition links starting from cells at the S phase and no GR activation stage (link number, $n = 433$); right: cell-state transition links starting from cells at the G2/M phase and no GR activation stage (link number, $n = 365$)). Black arrows show main cell-state transition directions. **c**, Scatter plot showing the relationship between transition distance (Pearson's distance) and transition proportion ($n = 729$), together with the red LOESS (locally estimated scatter-plot smoothing) smoothed line by ggplot2[65]. **d**, Three-dimensional plot showing the cell-state stability landscape. The *x* axis represents GR response states (from no to low to high activation state). The *y* axis represents the cell-cycle states ordered from G1 to G2/M. The *z* axis represents cell-state instability, defined as the proportion of cells inferred to be moving out of a given state between time points.

network (Spearman's correlation coefficient = −0.38; Fig. 4c). As a second approach, we computed 'instability' as the proportion of cells inferred to be moving out of a given state between time points (Fig. 4d). As expected, states corresponding to no GR activation were the least stable by this metric. Furthermore, among high GR activation states, states corresponding to early G1 were the most stable. These representations of the data are consistent with the transition network, wherein the states corresponding to high GR activation and early G1 are a frequent 'destination' of all nearby states (purple triangles in Fig. 4a).

Discussion

Sci-fate captures information analogous to RNA velocity[31], which distinguishes 'older' and 'newer' transcripts based on their splicing status. On the one hand, RNA velocity is more straightforward than sci-fate, because it makes use of information that is indirectly captured by many single-cell-profiling technologies, whereas sci-fate requires 4sU-labeling steps that cannot necessarily be used in all contexts. On the other hand, sci-fate lends itself to experimental control in a way that RNA velocity does not, as the timing and length of 4sU labeling can be specified whereas, with RNA velocity, it is a product of endogenous splicing dynamics. Furthermore, as we show, an experimental design that couples the labeling of newly synthesized mRNA to a time series enables the quantitative analysis of cells with complex transcriptional histories and futures.

While our manuscript was under review, two methods directed at the same goal, scSLAM-seq and NASC-seq, were reported[57,58]. Although there are similarities including the labeling strategy, we note major differences with respect to performance, accuracy and scalability: (1) as sci-fate uses combinatorial indexing, we successfully measured newly synthesized mRNA in >6,000 cells in one experiment, compared with <200 cells for scSLAM-seq or NASC-seq. Given that sci-fate is easily adaptable to three-level combinatorial indexing[25], it should already be possible to profile newly synthesized mRNA as >1 million cells per experiment. (2) Sci-fate costs <US$0.20 per cell for library preparation with two-level indexing, and <US$0.01 per cell with three-level indexing[25]. By comparison, both scSLAM-seq and NASC-seq utilize smart-seq which costs ~US$11 per cell for library preparation[59]. On a related point, sci-fate required an order of magnitude fewer raw reads per cell (~200,000 sci-fate versus ~2 million with scSLAM-seq), but achieved a greater number of genes detected per cell. (3) A key feature of sci-fate is that

we performed in situ 4sU chemical conversion in bulk fixed cells, resulting in a high reaction efficiency and low mRNA loss. In contrast, scSLAM-seq and NASC-seq require extraction of mRNA from each cell, followed by bead-based purification and chemical conversion. As a result, sci-fate exhibits higher efficiency to detect low abundance transcripts (median 6,500 genes detected per cell with sci-fate versus ~4,000 with scSLAM-seq, despite 1/10th of the raw sequencing depth). Furthermore, sci-fate exhibits a higher detection rate of newly synthesized mRNA (82% in sci-fate versus <50% in scSLAM-seq). (4) The signal-to-noise ratio (labeled versus unlabeled cells) of sci-fate is 20- to 58-fold, compared with only ~10-fold for scSLAM-seq or NASC-seq. This is partly due to the fact that the sci-fate library preparation is strand specific, whereas smart-seq is not. (5) Sci-fate enables direct counting of newly synthesized versus pre-existing mRNA via 3′-tagged UMIs[60], which are used by neither scSLAM-seq nor NASC-seq. Additional advantages of sci-fate include compatibility with fixed cells and the ability to concurrently process multiple independent biological samples within a single experiment. Finally, it is notable that in situ 4sU chemical conversion requires cell permeabilization and, at least in our experience, PFA fixation, neither of which is straightforward to introduce on droplet-based single-cell RNA-sequencing (scRNA-seq) platforms such as 10x Genomics.

We note that, although sci-fate enables quantification of mRNA synthesis in single cells, we remain in need of methods for measuring mRNA degradation rates in single cells. Related to this, our simplifying assumption that gene-specific degradation rates are constant across our DEX time course might not be a good choice in other systems. Specifically, in systems where the gene-specific degradation rates are expected or observed to vary substantially over time, these should be estimated for each time interval separately.

Sci-fate can be broadly applied to most in vitro systems to quantitatively characterize cell-state dynamics within short time windows (for example, one to several hours). For even shorter time frames, a concern is that the signal-to-noise ratio will drop as the rate of labeling falls toward the background rate of 0.8%. For longer time frames, a time-series approach can be adopted as in the main experiment described in the present study.

A major limitation of sci-fate is that 4sU-labeling experiments are generally performed within in vitro cell culture models. However, recent studies have shown that 4sU can be used together with transgenic *UPRT*-expressing mice to stably label cell type-specific nascent RNA transcription in vivo[61–63], suggesting that sci-fate, with further optimizations to enhance 4sU incorporation and detection rate, can potentially be used to profile single-cell transcriptional dynamics in vivo and at scale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0480-9.

Received: 9 June 2019; Accepted: 6 March 2020;
Published online: 13 April 2020

References

1. 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).
2. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
3. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
4. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
5. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
6. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* **19**, 477 (2018).
7. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).
8. Herzog, V. A. et al. Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
9. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* **15**, 221–225 (2018).
10. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
11. Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. & Boothroyd, J. C. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* **23**, 232–237 (2005).
12. Dolken, L. et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959–1972 (2008).
13. Miller, C. et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **7**, 458–458 (2014).
14. Duffy, E. E. et al. Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol. Cell* **59**, 858–866 (2015).
15. Schwalb, B. et al. TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
16. Rabani, M. et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).
17. Miller, M. R., Robinson, K. J., Cleary, M. D. & Doe, C. Q. TU-tagging: cell type-specific RNA isolation from intact complex tissues. *Nat. Methods* **6**, 439–441 (2009).
18. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
19. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
20. Ramani, V. et al. Massively multiplex single-cell Hi-C. Preprint at *bioRxiv* https://doi.org/10.1101/065052 (2016)..
21. Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
22. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
23. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
24. Yin, Y. et al. High-throughput single-cell sequencing with linear amplification. *Mol. Cell.* **76**, 676–690.e10 (2019).
25. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
26. Buckingham, J. C. Glucocorticoids: exemplars of multi-tasking. *Br. J. Pharmacol.* **147**, S258 (2006).
27. Reddy, T. E. et al. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* **19**, 2163–2171 (2009).
28. John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
29. Reddy, T. E., Gertz, J., Crawford, G. E., Garabedian, M. J. & Myers, R. M. The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. *Mol. Cell. Biol.* **32**, 3756–3767 (2012).
30. Vockley, C. M. et al. Direct GR binding sites potentiate clusters of TF binding across the human genome. *Cell* **166**, 1269–1281.e19 (2016).
31. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494 (2018).
32. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Software* **3**, 861 (2018).
33. Binder, E. B. The role of FKBP5, a co-chaperone of the glucocorticoid receptor in the pathogenesis and therapy of affective and anxiety disorders. *Psychoneuroendocrinology* **34**, S186–S195 (2009).
34. 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).
35. ENCODE Project Consortium et al. A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
36. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* **306**, 636–640 (2004).
37. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
38. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

39. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
40. Boruk, M., Savory, J. G. A. & Haché, R. J. G. AF-2-dependent potentiation of CCAAT enhancer binding proteinβ -mediated transcriptional activation by glucocorticoid receptor. *Mol. Endocrinol.* **12**, 1749–1763 (1998).
41. Qin, W. et al. Identification of functional glucocorticoid response elements in the mouse FoxO1 promoter. *Biochem. Biophys. Res. Commun.* **450**, 979–983 (2014).
42. Sheela Rani, C. S., Elango, N., Wang, S.-S., Kobayashi, K. & Strong, R. Identification of an activator protein-1-like sequence as the glucocorticoid response element in the rat tyrosine hydroxylase gene. *Mol. Pharmacol.* **75**, 589 (2009).
43. Fischer, M. & Müller, G. A. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* **52**, 638–662 (2017).
44. Chou, J., Provot, S. & Werb, Z. GATA3 in development and cancer differentiation: cells GATA have it! *J. Cell. Physiol.* **222**, 42–49 (2010).
45. Biswas, M. & Chan, J. Y. Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. *Toxicol. Appl. Pharmacol.* **244**, 16 (2010).
46. Ryoo, I.-G. & Kwak, M.-K. Regulatory crosstalk between the oxidative stress-related transcription factor Nfe2l2/Nrf2 and mitochondria. *Toxicol. Appl. Pharmacol.* **359**, 24–33 (2018).
47. Heer, R., Robson, C. N., Shenton, B. K. & Leung, H. Y. The role of androgen in determining differentiation and regulation of androgen receptor expression in the human prostatic epithelium transient amplifying population. *J. Cell. Physiol.* **212**, 572–578 (2007).
48. Meixner, A., Karreth, F., Kenner, L., Penninger, J. M. & Wagner, E. F. Jun and JunD-dependent functions in cell proliferation and stress response. *Cell Death Differ.* **17**, 1409–1419 (2010).
49. Li, M. et al. Krüppel-like factor 5 promotes epithelial proliferation and DNA damage repair in the intestine of irradiated mice. *Int. J. Biol. Sci.* **11**, 1458–1468 (2015).
50. Eberlé, D., Hegarty, B., Bossard, P., Ferré, P. & Foufelle, F. SREBP transcription factors: master regulators of lipid homeostasis. *Biochimie* **86**, 839–848 (2004).
51. Shermoen, A. W. & O'Farrell, P. H. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell* **67**, 303–310 (1991).
52. Palozola, K. C. et al. Mitotic transcription and waves of gene reactivation during mitotic exit. *Science* **358**, 119–122 (2017).
53. Parsons, G. G. & Spencer, C. A. Mitotic repression of RNA polymerase II transcription is accompanied by release of transcription elongation complexes. *Mol. Cell. Biol.* **17**, 5791–5802 (1997).
54. Sanchez-Alvarez, M., Zhang, Q., Finger, F., Wakelam, M. J. O. & Bakal, C. Cell cycle progression is an essential regulatory component of phospholipid metabolism and membrane homeostasis. *Open Biol.* **5**, 150093 (2015).
55. Harmon, J. M., Norman, M. R., Fowlkes, B. J. & Thompson, E. B. Dexamethasone induces irreversible G1 arrest and death of a human lymphoid cell line. *J. Cell. Physiol.* **98**, 267–278 (1979).
56. Greenberg, A. K. et al. Glucocorticoids inhibit lung cancer cell growth through both the extracellular signal-related kinase pathway and cell cycle regulators. *Am. J. Respir. Cell Mol. Biol.* **27**, 320–328 (2002).
57. Erhard, F. et al. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* **571**, 419–423 (2019).
58. Hendriks, G.-J. et al. NASC-seq monitors RNA synthesis in single cells. *Nat. Commun.* **10**, 3138 (2019).
59. Baran-Gale, J., Chandra, T. & Kirschner, K. Experimental design for single-cell RNA sequencing. *Briefings in functional genomics* **17**, 233–239 (2018).
60. Chen, W. et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **19**, 70 (2018).
61. Matsushima, W. et al. SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development* **145**, dev164640 (2018).
62. Sharma, U. et al. Small RNAs are trafficked from the epididymis to developing mammalian sperm. *Dev. Cell* **46**, 481–494 (2018).
63. Gay, L. et al. Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. *Genes Dev.* **27**, 98–115 (2013).
64. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502 (1989).
65. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

Methods

Mammalian cell culture. All mammalian cells were cultured at 37 °C with 5% CO_2, and were maintained in high-glucose Dulbecco's modified Eagle's medium (Gibco catalog no. 11965) for HEK293T and NIH/3T3 cells or Dulbecco's modified Eagle's medium/F12 medium for A549 cells, both supplemented with 10% fetal bovine serum and 1× penicillin–streptomycin (Gibco, catalog no. 15140122; 100 U ml^{-1} of penicillin, 100 µg ml^{-1} of streptomycin). Cells were trypsinized with 0.25% trypsin–ethylenediaminetetraacetic acid (EDTA) (Gibco, catalog no. 25200-056) and split 1:10 three times a week.

Sample processing for sci-fate. For HEK293T and NIH/3T3 cells, cells were incubated with 200 µM 4sU for 6 h before cell harvest. A549 cells were treated with 100 nM DEX for 0, 2, 4, 6, 8 or 10 h. Cells in all treatment conditions were incubated with 200 µM 4sU for the last 2 h before cell harvest. Of note, an excessively long labeling time (the extreme of which results in all transcripts being labeled) may result in information loss. Through a test experiment on HEK293T cells, we found that the transcriptome degradation rate is around 10% per h for each cell. We then selected 2 h as the labeling time window, such that about 80% of detected transcriptomes per cell would be previously synthesized and usable to infer the previous cell state. In theory, a shorter labeling time enables more accurate cell past state inference but also requires sampling of more time points to cover a continuous process with a given time interval (that is, 10 h in our DEX-treatment experiment). A short labeling time would also potentially be affected by greater noise. In our data, the background labeling rate ('labeled' reads ratio in non-labeled cells) is 0.8%. Given that 2 h of labeling results in detection of ~20% of transcripts as newly synthesized, to keep at least a 10:1 signal-to-noise ratio, the minimum labeling period should be at least 50 min.

All cell lines (A549, HEK293T and NIH/3T3 cells) were trypsinized, spun down at 300*g* for 5 min (4 °C) and washed once in 1× ice-cold phosphate-buffered saline (PBS). All cells were fixed with 4 ml ice-cold 4% PFA (EMS) for 15 min on ice. After fixation, cells were pelleted at 500*g* for 3 min (4 °C) and washed once with 1 ml PBSR (1× PBS, pH 7.4, 0.2 mg ml^{-1} bovine serum albumin (New England Biolabs), 1% SuperRnaseIn (Thermo) and 10 mM dithiothreitol (DTT)). After washing, cells were resuspended in PBSR at 2–10 million cells ml^{-1}, flash frozen and stored in liquid nitrogen. PFA-fixed cells were thawed in a 37 °C water bath, spun down at 500*g* for 5 min, and incubated with 500 µl PBSR including 0.2% Triton X-100 for 3 min on ice. Cells were pelleted and resuspended in 500 µl nuclease-free water including 1% SuperRnaseIn. Then, 3 ml of 0.1 N HCl was added into the cells for a 5-min incubation on ice[23]; 3.5 ml Tris-HCl, pH 8.0, and 35 µl of 10% Triton X-100 were added into cells to neutralize the HCl. Cells were pelleted and washed with 1 ml PBSR. Cells were resuspended in 100 µl PBSR (without DTT), and 100 µl PBSR (without DTT) with fixed cells was incubated with a mixture including 40 µl iodoacetamide (100 mM), 40 µl sodium phosphate buffer (500 mM, pH 8.0), 200 µl dimethylsulfoxide and 20 µl H_2O, at 50 °C for 15 min. The reaction was quenched by 8 µl DTT (1 M) and 8.5 ml PBS[66]. Of note, the cell loss rate is high (>95%) in the chemical conversion and centrifugation steps. Cells were pelleted and resuspended in 100 µl PBSI (1× PBS, pH 7.4, 0.2 mg ml^{-1} bovine serum albumin, 1% SuperRnaseIn). For all later washes, cells were pelleted by centrifugation at 500*g* for 5 min (4 °C).

The following steps are similar for the sci-RNA-seq protocol with PFA-fixed nuclei[10,19]. Briefly, cells were distributed into four 96-well plates. For each well, 500–5,000 cells (2 µl) were mixed with 1 µl of 25 µM anchored oligo(dT) primer (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10-bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where 'N' is any base and 'V' is either A, C or G; Integrated DNA Technologies) and 0.25 µl of 10 mM dNTP mix (Thermo), denatured at 55 °C for 5 min and immediately placed on ice. Then, 1.75 µl of first-strand reaction mix, containing 1 µl of 5× Superscript IV First-Strand buffer (Invitrogen), 0.25 µl of 100 mM DTT (Invitrogen), 0.25 µl SuperScript IV reverse transcriptase (200 U µl^{-1}, Invitrogen) and 0.25 µl RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was added to each well. RT was carried out by incubating plates at the following temperature gradient: 4 °C for 2 min, 10 °C for 2 min, 20 °C for 2 min, 30 °C for 2 min, 40 °C for 2 min, 50 °C for 2 min and 55 °C for 10 min. All cells were then pooled, stained with 4′,6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at 50 cells per well into 5 µl EB buffer. Cells were gated based on the DAPI stain such that singlets were discriminated from doublets and sorted into each well. Then, 0.66 µl mRNA Second Strand Synthesis buffer (New England Biolabs) and 0.34 µl mRNA Second Strand Synthesis enzyme (New England Biolabs) were added to each well, and second strand synthesis was carried out at 16 °C for 180 min. Each well was then mixed with 5 µl Nextera TD buffer (Illumina) and 0.025 µl i7-only TDE1 enzyme (provided by Illumina), and incubated at 55 °C for 5 min to carry out tagmentation. The reaction was stopped by adding 12 µl DNA-binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 36 µl AMPure XP beads (Beckman Coulter), eluted in 16 µl EB buffer (Qiagen) and transferred to a fresh multi-well plate.

For PCR reactions, each well was mixed with 2 µl of 10 µM P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; Integrated DNA Technologies), 2 µl of 10 µM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′; IDT) and 20 µl NEBNext High-Fidelity 2× PCR Master Mix (New England Biolabs). Amplification was carried out using the following program: 72 °C for 5 min, 98 °C for 30 s, and 18–22 cycles of 98 °C for 10 s, 66 °C for 30 s, 72 °C for 1 min) and a final 72 °C for 5 min. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined using Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% tris-borate–EDTA–polyacrylamide gel electrophoresis (PAGE) gel. Libraries were sequenced on the NextSeq 500 platform (Illumina) using a v.2 150-cycle kit (read 1: 18 cycles, read 2: 130 cycles, index 1: 10 cycles, index 2: 10 cycles).

Read alignment and downstream processing. Read alignment and gene count matrix generation for the scRNA-seq was performed using the pipeline that we developed for sci-RNA-seq[10] with minor modifications. Reads were first mapped to a reference genome with STAR v.2.5.2b[67], with gene annotations from GENCODE v.19 for humans and GENCODE v.M11 for mice. For experiments with HEK293T and NIH/3T3 cells, we used an index combining chromosomes from both humans (hg19) and mice (mm10). For the A549 experiment, we used human genome build hg19.

The single-cell sam files were first converted into alignment tsv files using the sam2tsv function in jvarkit[68]. Next, for each single-cell alignment file, mutations matching the background SNPs were filtered out. For background SNP reference of A549 cells, we downloaded the paired-end, bulk RNA-seq data for A549 cells from ENCODE[36] (sampled name: ENCFF542FVG, ENCFF538ZTA, ENCFF214JEZ, ENCFF629LOL, ENCFF149CJD, ENCFF006WNO, ENCFF828WTU, ENCFF380VGD). Each paired-end fastq files was first adaptor clipped using trim_galore-0.4.1 (ref. [69]) with default settings, aligned to human hg19 genome build with STAR v.2.5.2b[67]. Unmapped and multiple mapped reads were removed using samtools v.1.3[70]. Duplicated reads were filtered out by MarkDuplicates function in picard 1.105[71]. De-duplicated reads from all samples were combined and sorted with samtools v.1.3[70]. Background SNPs were called using the mpileup function in samtools v.1.3[70] and the mpileup2snp function in VarScan 2.3.9[72]. For the HEK293T and NIH/3T3 test experiment, a background SNP reference was generated in a similar pipeline, with the aggregated single-cell sam data from control condition (no 4sU labeling and no iodoacetamide treatment condition).

For each single-cell alignment file, all mutations with quality score ≤13 were removed. Mutations at both ends of each read were mostly due to sequencing errors, and thus were also filtered out. Mutations mapping to the background SNP reference were filtered out. For each read, we checked where there were T→C mutations for sense strand or A→G mutations for antisense strand, and labeled these mutated reads as newly synthesized.

Each cell was characterized by two digital gene expression matrices from the full sequencing data and newly synthesized RNA data as described in the read alignment and downstream processing steps. Genes with expression in fewer than five cells were filtered out. Cells with <2,000 UMIs or >80,000 UMIs were discarded. Cells with doublet score >0.2 by doublet analysis pipeline Scrublet v.0.2[73] were removed.

The dimensionality of the data was first reduced using principal components analysis (after selecting the top 2,000 genes with the highest variance) on a digital gene expression matrix on either full gene expression data or the newly synthesized gene expression data by Monocle 3/alpha[2,74]. The top ten PCs were selected for dimensionality reduction analysis with UMAP v.0.3.2, a recently proposed algorithm based on Riemannian geometry and algebraic topology to perform dimension reduction and data visualization[32]. For joint analysis, we combined the top ten PCs calculated on the whole transcriptome and the top ten PCs on the newly synthesized transcriptome for each single cell before dimension reduction with UMAP. Cell clusters were done via the densityPeak algorithm implemented in Monocle 3/alpha[2,74]. We first performed UMAP analysis on joint information of all processed cells, and identified an outlier cluster (724 of 7,404 cells). These cells were marked by high-level expression of *GATA3*, a marker of differentiated cells[44], and were filtered out before downstream analysis.

Linking TFs to their regulated genes. We sought to identify links between TFs and their regulated genes based on expression covariance. Cells with >10,000 UMIs detected, and genes (including TFs) with newly synthesized reads detected in >10% of all cells were selected. On average, these TFs are detected as being expressed in ~58% of cells. Of note, a small number of TFs with high expression overall, but low expression within newly synthesized reads, were filtered out at this step (14 TFs with expression in >50% of cells; 75 of TFs with expression in >20% of cells, filtered consequent to this). The full gene expression and newly synthesized gene count per cell were normalized by cell-specific library size factors computed on the full gene expression matrix by estimateSizeFactors in Monocle 3/alpha[2,74], log(transformed), centered, then scaled by scale() function in R. For each gene detected, a LASSO regression model was constructed with package glmnet v.2.0[75] to predict the normalized expression levels, based on the normalized expression of 853 TFs annotated in the 'motifAnnotations_hgnc' data from package RcisTarget v.1.2.1[37], by fitting the following model:

$$G_i = \beta_0 + \beta_t T_i$$

where G_i is the adjusted gene expression value for gene *i*. It is calculated with the newly synthesized mRNA count (g_i) for each cell, normalized by the cell-specific

size factor (SG_i) estimate using estimateSizeFactors in Monocle 3/alpha[2,74] on the full expression matrix of each cell, and log(transformed):

$$G_i = \ln\left(\frac{g_i}{SG_i} + 0.1\right)$$

To simplify downstream comparison between genes, we standardize the response G_i before fitting the model for each gene i with the scale() function in R.

Similar to G_i, T_i is the adjusted TF expression value for each cell. It is calculated using the full TF expression count (t_i) for each cell, normalized by the SG_i estimate using estimateSizeFactors in Monocle 3/alpha[2,74] on the full expression matrix of each cell, and log(transformed):

$$T_i = \ln\left(\frac{t_i}{SG_i} + 0.1\right)$$

Before fitting, T_i is standardized with the scale() function in R. β_0 and β_i are regression coefficients in LASSO regression.

Although negative correlations between a TF's expression and a gene's new synthesis rate could reflect the activity of a transcriptional repressor, we felt that the more probable explanation for negative links reported by glmnet was mutually exclusive patterns of cell-state-specific expression and TF activity. Thus, during prediction, we excluded TFs with negatively correlated expression with a potential target gene's synthesis rate, and also low regression coefficient (≤0.03) links. We identified a total of 6,103 links between TFs and regulated genes. A modified strategy without filtering negatively correlated TF-gene synthesis rate pairs identified 47 additional repressive TF-gene links between 9 TFs and 46 genes (see Supplementary Table 2).

Our approach aims to identify TFs that may regulate each gene, by finding the subset that can be used to predict its expression in a regression model. However, a TF with expression correlated with a gene's expression does not definitively mean that it is directly regulating that gene. To identify putatively direct targets within this set, we intersected the links with TFs profiled in ENCODE ChIP–seq experiments[36]. Of the 6,103 links between TFs and genes by LASSO regression, 1,086 links have TFs characterized in ENCODE, 807 of which were validated by target TF-binding sites near gene promoters from ENCODE[35], a 4.3-fold enrichment relative to background expectation (odds ratio for validation = 2.89 for links identified in LASSO regression versus 0.67 for background; $P < 2.2 \times 10^{-16}$, two-sided, Fisher's exact test). Only gene sets with significant enrichment of the correct TF ChIP–seq binding sites were retained (two-sided, Fisher's exact test, false discovery rate of 5%), and further pruned to remove indirect target genes without TF-binding data support. Ultimately, 591 links were retained using this approach.

To expand the set of validated TF–gene links, we further applied the package SCENIC[37], a pipeline to construct gene-regulatory networks based on the enrichment of target TF motifs in the 10-kb window around genes' promoters. Each co-expression module identified by LASSO regression was analyzed using *cis*-regulatory motif analysis with RcisTarget v.1.2.1[37]. Only modules with substantial motif enrichment (normalized enrichment score reported by RcisTarget > 3) of the correct TF regulator were retained, and pruned to remove indirect target genes without motif support. We filtered the TF–gene links using three correlation coefficient thresholds (0.3, 0.4 and 0.5), and combined all links validated by RcisTarget[37]. In total, 509 links were validated using this motif-based approach.

Combining both approaches, we identified a total of 986 TF–gene regulatory links using the covariance between TF expression and gene synthesis rate, validated by DNA-binding data or motif analysis. To evaluate the possibility that the links were artifacts of regularized regression, we permutated the order of single-cell TF expression (T_i) and performed the same analyses. No links were identified after this permutation.

Applying a similar strategy to all mRNAs (rather than only newly synthesized mRNAs) revealed 2,108 TF–gene links, with 532 identified by both approaches. The 448 TF–gene links uniquely identified by analysis of newly synthesized mRNA exhibited higher correlations between TF expression and newly synthesized mRNA than all mRNAs (mean Spearman's correlation of 0.19 versus 0.16, respectively; $P = 5.3 \times 10^{-8}$, two-sided, Wilcoxon's rank-sum test). The TF–gene links identified exclusively by analysis of all mRNAs corresponded to lower mRNA synthesis rates for linked genes (mean UMI count, normalized by size factor for newly synthesized mRNA: 1.20 for genes with links by newly synthesized mRNAs versus 0.97 for genes with links identified solely through analysis of all mRNAs; $P < 2.2 \times 10^{-16}$, two-sided, Wilcoxon's rank-sum test).

Ordering cells based on the activity of functional TF modules. To calculate TF 'activity' in each cell, newly synthesized UMI counts for genes linked to each of the 27 TFs were scaled by library size, log(transformed), aggregated and then mapped to *Z*-scores. As TFs with highly correlated or anti-correlated activity suggest that they may function in linked biological processes, we calculated the absolute Pearson's correlation coefficient between each pair of TF activity, and based on this we clustered TFs using a ward.d2 clustering method in package pheatmap v.1.0.12[76]. Five functional TF modules were identified and annotated based on their functions.

To characterize the dynamics of cells in relation to potentially independent cellular processes, cells were ordered by the activity of cell cycle-related TFs (TF module 1) or GR activity-related TFs (TF module 3) with UMAP[32] (metric = 'cosine', n_neighbors = 30, min_dist = 0.01). The cell-cycle progression trajectory was validated by cell-cycle gene markers in Seurat v.2.3.4[34]. Three cell-cycle phases were identified using the densityPeak algorithm implemented in Monocle 3/alpha[2,74], on the UMAP coordinates ordered by cell-cycle TF modules. As each main cell-cycle phase still showed variable TF activity and cell-cycle marker expression, we segmented each phase to early/middle/late states by *k*-means clustering ($k = 3$), and recovered a total of nine cell-cycle states. Three GR response states were identified using the densityPeak algorithm implemented in Monocle 3/alpha[2,74].

Past transcriptome state recovery from sci-fate. To infer the past transcriptome state (that is, the cell state before 4sU labeling commenced), we assume that mRNA half-lives are consistent across different DEX-treatment durations. This assumption is further validated by a self-consistency check later. Under this assumption, the partly degraded bulk transcriptome before the 2-h 4sU labeling should be the same between no DEX and 2-h DEX-treated cells. Thus, for any given gene, differences in whole transcriptomes (bulk) between these time points should be equal to differences in the newly synthesized transcriptomes (bulk), corrected by technique's detection rate:

$$A_{0h}/S_{0h} - (N_{0h}/S_{0h})/\alpha = A_{2h}/S_{2h} - (N_{2h}/S_{2h})/\alpha$$

where A_{0h} is the aggregated UMI count for all cells in the no DEX treatment group; S_{0h} is the library size (total UMI count of cells) at no DEX treatment; N_{0h} is the aggregated newly synthesized UMI count for all cells in the no DEX treatment group; A_{2h} is the aggregated UMI count for all cells in the 2-h DEX treatment group; S_{2h} is the library size (total UMI count of cells) in the 2-h DEX treatment group; N_{2h} is the aggregated newly synthesized UMI count for all cells in the 2-h DEX treatment group; and α is the detection rate for each gene in sci-fate. As cells from different time points were profiled in the same experiment and the UMI counts detected per cell were similar across conditions (see Supplementary Fig. 2a,b), we assume the same overall RNA amount in the 0-h and 2-h samples, and normalize the aggregated gene count reads by total counts of each time point. For experiments where this assumption may not stand, spike-in standards could be used to control for differences in the overall amount of mRNA between conditions. In theory, one detection rate can be calculated for each gene. However, for genes with minor differences of the newly synthesized rate between two conditions, the estimated α is dominated by noise. We thus selected genes showing higher differences in the normalized, newly synthesized rate between two conditions: we first tested a series of thresholds for gene filtering and calculated the α for each gene. We then plotted the relationship between threshold and the ratio of genes with out-of-range α values (<0 or >1). We selected the threshold that was at the knee point of the plot, resulting in 186 genes being selected (see Supplementary Fig. 5a). The differences in newly synthesized mRNA of these genes correlate highly with the differences in mRNA expression level (Pearson's $r = 0.93$; see Supplementary Fig. 5b), suggesting that the new RNA detection rate is stable across genes (see Supplementary Fig. 5c). In the present study we use the median detection rate across 186 selected genes to estimate α.

We next computed the mRNA degradation rate across each 2-h interval. As the A549 cell population can be regarded as stable without external perturbation, for 2-h DEX-treated cells, its past state (before 2-h 4sU labeling) should be the same as the 0-h DEX-treated cells. Expanding on this logic, the past state (before 4sU labeling) for T = 0-/2-/4-/6-/8-/10-h DEX-treated cells should be similar to the profiled T = 0-/0-/2-/4-/6-/8-h cells, respectively:

$$A_{t1}/S_{t1} - (N_{t1}/S_{t1})/\alpha = A_{t0}/S_{t0} \times \beta$$

A_{t1} is the aggregated UMI count for all cells in $t1$; S_{t1} is the library size (the total UMI count of cells) at $t1$; N_{t1} is the aggregated newly synthesized UMI count for all cells at $t1$; α is the estimated detection rate of sci-fate; A_{t0} is the aggregated UMI count for all cells in $t0$; S_{t0} is the library size (the total UMI count of cells) at $t0$; and β is 1 − gene-specific degradation rate between $t0$ and $t1$, and is related to the mRNA half-life γ by:

$$\beta = (1/2)^{(t1-t0)/\gamma}$$

The β was calculated for each of the 14,587 genes across each 2-h interval of DEX treatment. As with the self-consistency check, the gene degradation rates are highly correlated across different DEX-treatment times (see Supplementary Fig. 5g). We therefore used the average degradation rate for each gene for downstream analysis.

With the overall sci-fate detection rate as well as per-gene degradation rates estimated, the past transcriptome state of each cell can be estimated using:

$$a_{t1} - n_{t1}/\alpha = a_{t0} \times \beta$$

a_{t1} is the single-cell UMI count in $t1$; n_{t1} is the single-cell newly synthesized UMI count at $t1$; and α is the detection rate for each gene in sci-fate. In the present

study, we use the median detection rate across 186 selected genes as its estimates; β is 1 − gene-specific degradation rate between $t0$ and $t1$. The a_{t0} is the estimated single-cell transcriptome in a past time point $t0$, with all negative values (15.6% of values on average) converted to 0.

The detection rate (α) of the newly synthesized transcriptome is experiment specific and depends mainly on the 4sU-labeling concentration. Generally, a lower 4sU concentration will lead to a lower 4sU incorporation rate into newly synthesized mRNA, which will reduce the detection rate of newly synthesized transcripts. In addition, the length of sequencing reads may affect the detection rate, as short sequencing reads will reduce the possibility of detecting incorporated 4sU. In the case of our experiments and as noted earlier, we used relatively long reads (on average, 75 bp of transcript-derived sequence obtained per read) to potentially increase detection rate. We also designed the experiments such that all treatment conditions share the same 4sU treatment concentration and incubation time in the same cell type. Furthermore, all cells were sequenced in the same sequencing run, such that all conditions were expected to share a similar α. For sci-fate experiments with different labeling conditions or sequencing settings, the α would need to be re-estimated for each part of the experiment. Of note, our simplifying assumption that gene-specific degradation rates are constant across our DEX time course might not be a good choice in other systems. Specifically, in systems where the gene-specific degradation rates are expected or observed to substantially vary over time, these should be estimated for each time interval separately.

Linkage analysis to build single-cell state trajectory. The goal of what we call here 'linkage analysis' is to associate each cell with parent and child cells at different time points, that is single-cell state trajectories. Our approach is based on the fact that the past transcriptome states (before 2-h 4sU labeling) of cells at $t1$ should share the same cell population distribution with the profiled transcriptome states of cells at $t0$ (2 h earlier than $t1$), assuming that there is no cell apoptosis. We thus applied a published manifold alignment strategy to identify common cell states between two datasets, based on common sources of variation[34]. As a result, whole transcriptomes from $t0$ cells and recovered past transcriptomes from $t1$ cells are aligned in the same UMAP space. This analysis is based on an assumption that, for intermediate time points, we are oversampling the space of physiologically distinct states in this time course. Violation of this and other assumptions can be detected by outliers during alignment of the two datasets. For each cell A from $t1$, we selected its nearest neighbor in $t0$ as its parent state in the alignment UMAP space. Similarly, for each cell from $t0$, we selected its nearest neighbor in $t1$ as its child cell state. Of note, it is not necessary for the link to be bidirectional: the parent state of one cell may be linked to a different child cell. After the parent and child states were identified for each cell (except cells at the start and end time points), we then extend each cell trajectory by searching for the linked parent cell of each cell's parent, and similarly the linked child cell of each cell's child. Thus, each single cell can be characterized by a single-cell state transition path across all six time points spanning 10 h. As multiple cells (>50) are profiled for each of the 27 defined cell states, stochastic cell-state transition processes can also potentially be captured.

Dimensionality reduction and clustering analysis. For dimensionality reduction on single-cell transcriptomes, the top five PCs for full transcriptomes and the top five PCs for newly synthesized transcriptomes were selected for each state, and combined in temporal order along a single-cell state trajectory for UMAP analysis. The main cell trajectory types were identified using a density peak-clustering algorithm[77].

With cell-state proportion at the beginning time point (0-h treatment) and cell-state transition probabilities estimated from the data, we first predicted the cell-state distribution after 2 h, assuming that the cell-state transitions in DEX treatment are cell-autonomous, time-independent, Markovian processes. Similarly, the cell-state distribution at later time points can be predicted from the cell-state distribution 2 h earlier.

For RNA-velocity analysis of these same data, single-cell spliced/unspliced expression matrices were generated using the command line interface of velocyto v.0.17[31] with the default run_smartseq2 mode on single-cell bam files. Cell transition direction inference was performed with an optimized, scalable, RNA-velocity analysis toolkit scVelo v.0.1.17 and scanpy v.1.4.1 with default settings[31,78]. To integrate treatment time information into the RNA-velocity analysis of the 0-h and 2-h time points, we prohibited transitions of cells within 0 h, and instead selected the future state as the cell at the 2-h time point with the highest transition probability.

Cell-state instability and cell-state distance calculations. We defined cell-state instability as the proportion of cells in a given state 'moving' to any other state at the next time point. To calculate cell-state distances, we first sampled an equal number ($n = 50$) of cells from each state, and separately aggregated the full transcriptome and newly synthesized transcriptome of sampled cells of that state (that is, in this 'joint transcriptome', each gene is represented by two columns, one for the whole transcriptome and one for the newly synthesized transcriptome). The cell-state distance is calculated as (1 – Pearson's correlation coefficient between the joint transcriptomes of two different states).

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

Data availability

The data generated by this study can be downloaded in raw and processed forms from the National Center for Biotechnology Information's Gene Expression Omnibus (GSE131351).

Code availability

Scripts for processing sci-fate sequencing were written in Python and R with code available at https://github.com/JunyueC/sci-fate_analysis.

References

66. Muhar, M. et al. SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science* **360**, 800–805 (2018).
67. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
68. Lindenbaum, P. JVarkit: java-based utilities for Bioinformatics. *figshare* (2015).
69. Krueger, F.. Trim Galore. *GitHub* https://github.com/FelixKrueger/TrimGalore (2019).
70. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Broad Institute. Picard Tools. *GitHub* http://broadinstitute.github.io/picard/ (2019).
72. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
73. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Preprint at *bioRxiv* https://doi.org/10.1101/357368 (2018).
74. Cole Trapnell Lab. Monocle release. *GitHub* https://github.com/cole-trapnell-lab/monocle-release (2019).
75. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
76. Kolde, R. pheatmap. *GitHub* https://github.com/raivokolde/pheatmap (2018).
77. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
78. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Acknowledgements

We thank the members of the Shendure labs for helpful discussions, particularly X. Huang, R. Blecher, B. Martin, F. Chardon and R. Qiu; J. Rose, D. Maly, L. VandenBosch, and T. Reh's lab for sharing the NIH/3T3 cell line; and J. McFaline-Figueroa for sharing the A549 cell line. This work was funded by the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), grants from the National Institutes of Health (grant nos. DP1HG007811 and R01HG006283 to J.S.; grant no. DP2 HD088158 to C.T.), the W. M. Keck Foundation (to C.T. and J.S.), the Dale. F. Frey Award for Breakthrough Scientists (to C.T.), the Alfred P. Sloan Foundation Research Fellowship (to C.T.) and the Brotman Baty Institute for Precision Medicine. J.S. is an investigator of the Howard Hughes Medical Institute.

Author contributions

J.S. and J.C. designed the research. J.C. developed the technique and performed the experiments with the assistance of F.S. J.C. performed the computation analysis with suggestions from W.Z. and C.T. J.S. and J.C. wrote the paper.

Competing interests

F.J.S. declares competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents and data disclosed in this article.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0480-9.

Correspondence and requests for materials should be addressed to J.C. or J.S.

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

nature research

Corresponding author(s): Jay Shendure

Last updated by author(s): Feb 20, 2020

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used except for Illumina RTA basecalling.
Data analysis	Common, freely available DNA sequencing data analysis software was used to analyze data, as described in Methods: bcl2fastq/v2.16, python/v2.7.13, trim_galore/v0.4.1, samtools/v1.3, picard/1.105, STAR/v 2.5.2b, R/3.5.0, VarScan/2.3.9 , scanpy/v1.4.1, scrublet/v0.2, EnrichR/v1.0, UMAP/v0.3.2, reticulate/v1.10, Monocle3/alpha, Seurat/2.3.4, pheatmap/1.0.12, RcisTarget/v1.2.1, glmnet/v.2.0, velocyto/v.0.1.17, and scVelo/v.0.1.17. Scripts for processing sci-fate sequencing were written in python and R with code available at https://github.com/JunyueC/sci-fate_analysis.

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

Data

Policy information about availability of data

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

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

The data generated by this study can be downloaded in raw and processed forms from the NCBI Gene Expression Omnibus (GSE131351).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No explicit calculations were performed to determine sample size.
Data exclusions	No data were excluded from the study.
Replication	The technique was tested and validated in two independent experiments using different cell lines. For A549 cell experiment, each of the six treatment conditions was represented by 64 replicate wells during the first round of processing in sci-fate. All attempts at replication were successful.
Randomization	The order of samples are randomized during drug treatment, and during sample processing in sci-fate.
Blinding	Investigators were blinded to group allocation during data collection (sequencing) and analysis.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T, NIH/3T3 and A549 cells were from ATCC
Authentication	None of the cell lines were authenticated.
Mycoplasma contamination	Cell lines were not tested for Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.