**Learning lineage dynamics by coherent sparse optimization**

Shou-Wen Wang\*,1, Michael J Herriges2,3, Kilian Hurley4,5, Darrell N. Kotton2,3 and Allon M. Klein\*,1

1 Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA

2 Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA 02118, USA

3 The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA

4 Department of Medicine, Royal College of Surgeons in Ireland, Education and Research Centre, Beaumont Hospital, Dublin, Ireland

5 Tissue Engineering Research Group, Royal College of Surgeons in Ireland, Dublin, Ireland

\*Email: [shouwen\_wang@hms.harvard.edu](mailto:shouwen_wang@hms.harvard.edu) (S.W.W.); [allon\_klein@hms.harvard.edu](mailto:allon_klein@hms.harvard.edu) (A.M.K.)

**Abstract**

Single-cell genome-wide profiling offers an approach to map transitional cell states during cell differentiation, disease onset, and drug response. Lineage-tracing, in which cells are labeled with hereditary markers, offers an approach to establishing dynamic relationships between cell states. Here, we develop a computational approach to infer cell dynamics and fate decision boundaries from single-cell genomics integrated with lineage tracing. To do so, we extend the statistical problem of compressed sensing to enforce coherent, sparse clonal relationships in time series data (CoSpar). The approach is compatible with one or more clonal time points. In simulations and published datasets, CoSpar is robust to severe down-sampling of clonal data, and it identifies fate biases associated with underlying progenitor heterogeneity. In differentiation of human pluripotent stem cells, we predict and confirm a role for Leukemia Inhibitory Factor in expanding lung endoderm progenitors. CoSpar is available at <https://cospar.readthedocs.io/>.

**Introduction**

In tissue development, regeneration, and disease, cells differentiate into distinct, reproducible phenotypes. A ubiquitous challenge in studying these processes is to order the sequence of events in differentiation1–3, and to identify events that drive cells towards one phenotype or another. This challenge is common to understanding mechanisms in embryo development, stem cell self-renewal, cancer cell drug resistance, and tissue metaplasia1–3.

At least two observational strategies help to order cellular events. Single-cell genome-wide profiling – such as by single-cell RNA sequencing (scRNA-seq) – offers a universal and scalable approach to establishing dynamic states by densely sampling cells at different stages3–10. Alone, however, scRNA-seq does not establish which early differences between cells drive or correlate with future fate2,11–13. Lineage tracing offers a complementary family of methods that can clarify long-term dynamic relationships across multiple cell cycles. To carry out lineage tracing, individual cells are labeled naturally or experimentally at an early time point1–3. The state of their clonal progeny is analyzed at one or more later time points (Fig. 1**a**).

Recently, a number of efforts from us and others have integrated lineage-tracing with single-cell genome-wide profiling (hereafter LT-scSeq), using unique, heritable, and expressed DNA barcodes2,13–21. These technologies identify cells that share a common ancestor and define their genomic state in an unbiased manner. LT-scSeq experiments have been used successfully to identify when fate decisions occur13,14, novel markers for stem cells22, and pathways whose manipulation alters cellular outcomes14,16. The simplest of these methods label cells at one time point13 (Fig. 1**b**); more complex methods allow the accumulation of barcodes over successive cell divisions and can thus reveal the sub-structure of clones2,13–21 (Fig. 1**c**).

Emerging LT-scSeq methods have been successful at revealing novel regulators of cell fate14,16 and the potency of early progenitors13,14. But they also present challenges that may limit their utility in practice. We identified at least five technical and biological challenges that affect design and interpretation, detailed in (Fig. 1**f**). These include stochastic differentiation and variable expansion of clones28,29 (Fig. 1**f-i**); cell loss during analysis (Fig. 1**f-ii**); barcode homoplasy wherein cells acquire the same barcode despite not having a lineage relationship30,31 (Fig. 1**f-iii**); access to clones only at a single time point23,24 (Fig. 1**f-iv**); and clonal dispersion due to a lag time between labeling cells and the first sampling (Fig. 1**f-v**).

Here we develop an approach to analyze LT-scSeq experiments in a manner that is robust to these problems. To do so, we begin with a generalized model of clonal dynamics in which cells can divide, differentiate, or be lost from the sampled tissue in a stochastic manner, with rates that are state-dependent (Supplementary Fig. 1**a**). Defining the features of this model can be considered the goal of LT-scSeq. We make it our narrower goal to learn the fraction of progeny of cells, initially in one state, which are found to occupy a second state after some time interval (Supplementary Fig. 1**b,c**). When calculated across all pairs of initial and final states, these fractions encode differentiation bias and fate hierarchies (Fig. 1**d**). They can reveal genes whose early expression is predictive of future fate choice.

**Results**

**Dynamic inference from clonal data with state information.**

One formalization of dynamic inference is to identify a transition map, a matrix , which describes the probability of a cell, initially in some state at time , giving rise to progeny in a state at time (Fig. 1**d**)32. We define specifically as the fraction of progeny from state *i* that occupy state *j* (Supplementary Fig. **1c**). This transition matrix does not capture all we can learn about cell dynamics: it already combines the effects of cell division, loss, and differentiation (Supplementary Fig. **1d**). As will be seen, even learning will prove useful for several applications (Fig. 1**d**).

The transition map has properties that make it amenable to inference. We expect it to be a sparse matrix, since most cells can access just a few states during an experiment (Fig. 1**e**, left panel). And we expect it to be locally coherent, meaning that neighboring cell states share similar fate outcomes (Fig. 1**e**, right panel). These constraints could improve robustness in map inference (see illustration in Supplementary Fig. 1**e**). Box 1 formalizes these constraints and establishes the technical foundation for inferring a transition map by coherent sparse (CoSpar) optimization (see schema in Fig. 2**a**; Supplementary Fig. 2). As inputs, CoSpar requires a clone-by-cell matrix that encodes the clonal information at time , and a data matrix for observed cell states (e.g. from scRNA-seq).

CoSpar is formulated assuming that we have information on the same clones at more than one time point. When we observe clones at only one time point , we jointly optimize the transition map and the initial clonal data (Fig. 2**b**; Methods). In this joint optimization, one must initialize the transition map, and we have shown that the final result is robust to initialization (Supplementary Fig. 3**e**; Supplementary Fig. 4**c,d**). This approach can be used for clones with nested structure (Supplementary Fig. 4**f-h**). Finally, coherence and sparsity provide reasonable constraints when no clonal information is available, offering an approach to infer transition maps from state heterogeneity alone. We have extended CoSpar to this case.

|  |
| --- |
| **Box 1: Coherent Sparse Optimization**  In a model of stochastic differentiation, cells in a clone are distributed across states with a time-dependent density profile . A transition map directly links clonal density profiles between time points:    From multiple clonal observations, our goal is to learn *T.* To do so, we denote as a clone-by-cell matrix and introduce *S* as a matrix of cell-cell similarity over all observed cell states, including those lacking clonal information. The density profiles of all observed clones are estimated as .  With enough clonal information could in principle be learnt by matrix inversion. However, the number of clones will always be far less than the number of states. To constrain the map, we require that: 1) is a sparse matrix (Fig. 1**e**, left panel); 2) is locally coherent (Fig. 1**e**, right panel); and 3) is a non-negative matrix. With these requirements, the inference becomes an optimization problem:    quantifies the sparsity of the matrix through its L1 norm, and quantifies the local coherence of ( being the Graph Laplacian of the cell state similarity graph, and being the local divergence). The remaining constraints enforce the observed clonal dynamics, non-negativity of , and map normalization, respectively. At , the minimization takes the form of *Lasso*33, an algorithm for compressed sensing. Our formulation extends compressed sensing from vectors to matrices, and to enforce local coherence. The local coherence extension is reminiscent of the *fused Lasso* problem30. An iterative, heuristic approach solves the CoSpar optimization efficiently (Fig. 2**a**; Supplementary Fig. 2). See Methods and Supplementary Notes 1-3 for further details. |

**CoSpar infers dynamics robustly in simulations.**

We used computer simulations to establish the basic performance of CoSpar. We modeled cells progressing through a sequence of gene expression states either towards a single fate (Fig. 3**a**) or bifurcating into two fates (Fig. 3**e**), with clones sampled in a manner representative of LT-scSeq experiments13,14. We then used the simulated data to infer the transition rates and fate biases of the simulated process. With 1000 clones, mean transition rates inferred by CoSpar were within 3 standard deviations () of the actual transition rate distribution 98% of the time (TPR>98%, Fig. 3**d**). In simulations of bifurcating dynamics, CoSpar predicted the distribution of progeny fates with 85% Pearson correlation to ground truth (Fig. 3**j**). Inferences remained similarly accurate with as few as 30 clones (Fig. 3**d**). These results established that CoSpar can quantitatively infer transition maps in idealized conditions.

We examined how two errors typical of LT-scSeq -- barcode homoplasy, and clonal dispersion -- would alter dynamic inference. Enforcing sparsity and local coherence should make inference insensitive to both sources of errors. To simulate homoplasy, we allowed different clones to acquire the same barcode (Fig. 3**a**). CoSpar only detectably lost accuracy when all lineage barcodes mixed more than ten clones on average (Fig. 3**b-d**) – a degree of homoplasy far higher than expected in most experiments. The method should thus be robust to most LT-scSeq applications, and may even be suitable for parsing lineage-tracing experiments with very few barcodes23,34–36. To study clonal dispersion, we simulated sampling clones at increasing times post-barcoding. CoSpar proved largely insensitive to dispersion resulting from lag time: even after five cell cycles it recovered fate biases and mean transition rates with no loss of accuracy (Fig. 3**g-i**).

CoSpar performance can be contrasted with that of approaches used in previous work, which average the transitions between cells observed in each clone at different time points13. While this approach works for perfect data, it is severely affected by both lag time and barcode homoplasy (Fig. 3**d,g,i**).

**CoSpar predicts early fate bias in hematopoiesis.**

We applied CoSpar to published datasets from three independent experiments. The first experiment tracked hematopoietic progenitor cells (HPCs) differentiating in culture, with clones sampled on days 2, 4 and 6 post-barcoding (Fig. 4**a,b**)13. During this time, cells progressed from a heterogeneous pool of HPC states into ten identifiable differentiated cell types. We used all clonal data to generate a ground truth for the early fate bias towards either the monocyte or neutrophil fate, using the simple method from Weinreb et.al.13 (Fig. 4**c**). Below, we evaluated the accuracy of fate prediction using its Pearson correlation with the ground truth.

As a baseline for comparison, we applied CoSpar to predict HPC fate bias using state information alone (Fig. 4**e**). Even without access to any clonal data, CoSpar could resolve early fate bias at a performance close to the upper bound defined by cross-validation of the ground-truth data (CoSpar correlation R=0.69; ground-truth R=0.72) (Fig. 4**e,g**; Supplementary Fig. 6**a**). This performance reflects improvements from enforcing coherence and sparsity (R=0.51-0.54 prior to CoSpar; Fig. 4**d**; Supplementary Fig. 3**f**). However, the prediction based on state information alone is limited because it is sensitive to the choice of distance metric used in analysis (Fig. 4**g**; Supplementary Fig. 3**e**).

Clonal information eliminated the sensitivity to distance metric. To show this, we applied CoSpar to data restricted in time, or restricted in its quality and depth. Using even a single time point of clonal data (day-6), CoSpar recovered the early fate bias (Fig. 4**f**; R=0.68), and it did so robustly over a range of parameters and choices of distance metrics (Fig. 4**g**;Supplementary Fig. 3**e**). Further, it recovered the differentiation hierarchy seen in the correlation of clonal barcodes across all cell types (Supplementary Fig. 3**c,d**). When using a sub-sampled dataset from the top 15% most dispersed clones as ranked by day-4 intra-clone distance (Fig. 4**b**), CoSpar performed similarly well, and outperformed the original method used to analyze this data13 (Fig. 4**h,i**;Supplementary Fig. 3**a,b**)**.** Thus, CoSpar successfully facilitates analysis of clones at a single time point, or using a fraction of the original data collected in this example.

These benchmarks suggest that CoSpar should be able to predict fate biases not previously recognized. We investigated fate biases in the Gata1+ states that give rise to five mature fates: megakaryocyte (Mk), erythrocyte (Er), mast cell (Ma), basophil (Ba), and eosinophil (Eos) (Fig. 4**a,k**). In culture, Mk and Er arise from a common progenitor (MEP), while Ba, Eos and Ma are produced by a different progenitor (BEMP)43. Existing studies of these progenitors are hampered by the lack of good markers. While molecular signatures of FACS-sorted MEP have been explored recently44–46, less is known about the transcriptome identity of BEMPs. This dataset provides an opportunity to predict the molecular identity of these early progenitors. The original method used to analyze this data finds very few genes distinguishing BEMPs and MEPs (Supplementary Fig. 3**g-i**). Yet using all clonal information, CoSpar predicts an early fate decision boundary between MEP and BEMPs (Fig. 4**j,k**). Differences between the putative BEMPs and MEPs are evident in scRNA-seq data alone, but clonal data provide confidence that the differences are associated with functional fate bias. The fate bifurcation correlates with the expression of known marker genes (*Slc14a1* for Mk47, *Thy1* for Ba33; Fig. 4**l**), and a transcription factor (TF) *Cebpa* that regulates Eos and Ba differentiation30. We identified 377 known and novel putative fate-associated genes (Fig. 4**m**; Supplementary Table 1).

**CoSpar reveals early fate bias in reprogramming.**

The second experiment we analyzed tracked cells during the reprogramming of fibroblast cells over 28 days into endoderm progenitors (Fig. 5**a**)14. In this experiment, approximately 30% of cells successfully reprogrammed; the remainder failed. Clonal analysis with cumulative barcoding was used to identify these cells early and predicted features that regulate their fate (Fig. 5**b,c**). We used clones strongly enriched in one of the two fates, identified by the original study, to generate the ground truth for early fate bias, and we then used it to benchmark CoSpar.

To evaluate CoSpar, we revisited this experiment after discarding over 90% of clones, and we specifically retained clones that show the least bias in reprogramming outcomes. Despite deliberately using down-sampled low-quality data, CoSpar recapitulated fate bias: the predicted progenitors of reprogrammed and failed cells share 73 out of 100 marker genes with the ground truth population (Fig. 5**f**), including genes previously showing strong positive and negative association with reprogramming success (*Apoa1*, *Spint2*, *Col1a2*, *Peg3*), as well as *Mettl7a1*, which was found to improve reprogramming14. These genes could be associated with fate bias using as few as ten clones, even when deliberately selecting clones with minimal fate bias (Fig. 5**d,e**;Supplementary Fig. 4**b**). By contrast, the analytical approach used in the original study14 failed to identify fate-predictive gene expression after such severe reduction in data quality (Fig. 5**e,f**; Supplementary Fig. 4**b**). Further, CoSpar performed robustly when using only clonal data from the final time point of the experiment (Fig. 5**g,h**; Supplementary Fig. 4**c-e**).

As in hematopoiesis, it is instructive to see the information encoded in clonal relationships. When applying CoSpar without clonal data, we found that CoSpar could predict the same early fate biases (Fig. 5**g**, Middle panel), but is again sensitive to the distance metric used (Fig. 5**h**). A different distance metric performs best here from the hematopoiesis dataset, suggesting that there is no simple ‘best-practice’ approach to dynamic inference in the absence of clonal data.

Finally, we applied CoSpar to predict fate bias at the earliest available time point after reprogramming is initiated (day 3), where no clonal information is available and fate bias remains unexplored14. Using clonal information, CoSpar predicts strong fate biases (Fig. 5**i**), arguing that future reprogramming success is established very early on. This prediction is supported by the differential expression of transgene *FoxA1-HNF4a* (a TF cocktail to induce reprogramming), the reprogramming marker gene *Apoa1,* and failed trajectory marker *Col1a2* and *Dlk1*14 (Fig. 5**j)**. We also found other known and putative drivers or marker genes that correlate with fate bias on day 3 (Fig. 5**k**; Supplementary Table 2).

**CoSpar predicts early fate bias during lung directed differentiation.**

In the third experiment, human pluripotent stem cells were differentiated into distal lung alveolar epithelial cells (induced alveolar epithelial type 2 cells, or iAEC2s)23,34. Here, clonal and transcriptomic information were profiled jointly on day 27 after initial barcoding on day 17, and a separate time-course experiment produced scRNA-seq data for 6 time points, including days 17 and 21(Fig. 6**a**). In this study, Hurley et al. reported the existence of clones derived from multipotent cells on day 17 but did not investigate their fate biases. A re-examination of the clonal data, however, suggests strong fate biases as early as day 17. Out of the 272 clones, 25% were enriched in either the iAEC2 or non-iAEC2 clusters (FDR=0.01), and clonal compositions differed significantly from that of randomized clones (Fig. 6**b**). Accordingly, clonal representation of iAEC2s anti-correlates with other fates (Supplementary Fig. 5**b,c**). We investigated signatures that could predict effectors of fate bias among day 17 progenitors.

Applying CoSpar, we assigned a putative fate bias to each of the cells seen on day 17. CoSpar predicts some cells to be strongly biased in cell fate (Fig. 6**c**), and also the existence of unbiased multipotent states; these strongly overlap with highly proliferating cell states on day 17 and are consistent with large clones hosting multiple endodermal lineages on day 27 (Supplementary Fig. 5**d**). As a control, we expected weaker fate biases earlier in differentiation, which is confirmed by applying CoSpar to cells two days earlier (day 15, Supplementary Fig. 5**e-g**). Among genes differentially expressed between the two biased populations on day 17, we identified several established TFs that regulate lung differentiation: *CEBPD, NKX2-1, SOX9, SOX11* (Fig. 6**d,e**;Supplementary Table 3)23,35–37.

To gain confidence in the analysis, we sought to test a prediction made using CoSpar relating fate bias to cell state at day 17. We found that on day 17 of the iAEC2 differentiation protocol Leukemia Inhibitory Factor Receptor (*LIFR*) has reduced expression (Fig. 6**e**) in cells biased towards non-iAEC2 fates and high expression in cells with a low fate bias that express proliferative markers (Fig. 6**e**). Previous work has shown that its ligand, LIF, is expressed in the adult mouse distal lung38, and in fetal rat lungs LIF is expressed in developing lung epithelial and mesenchymal cells. While LIF supplementation or repression in an explanted fetal mouse lung model altered lung growth and branching morphogenesis39, whether LIF affects the proliferation or differentiation of developing lung epithelial progenitors remains unknown. To test if LIF influences developing lung endoderm, we used the same developmental directed differentiation protocol employed by Hurley et al. to generate lung epithelial progenitors and iAEC2s23,34. We differentiated human BU3 NGST induced pluripotent stem cells (iPSCs), which feature a GFP reporter knocked into the locus downstream of *NKX2-1*, a transcription factor expressed in lung epithelial cells23,34, and purified NKX2-1GFP+ cells on day 15. We then added 0, 5, or 50ng/ml of recombinant human LIF at days 17-19 of differentiation, the time point we had computationally predicted that LIFR expression levels correlate with high proliferative marker expression and low fate bias. At day 29, we observed a -fold (mean±SD, n=3 donors) increase in the yield of lung endoderm cells (identified by NKX2- expression) in response to prior, transient LIF exposure (Fig. 6**g,h**). This result provides evidence that predictions from CoSpar can lead to actionable insights into development and differentiation, and it opens the specific possibility for achieving increased yields through guided optimization of differentiation protocols.

**Discussion**

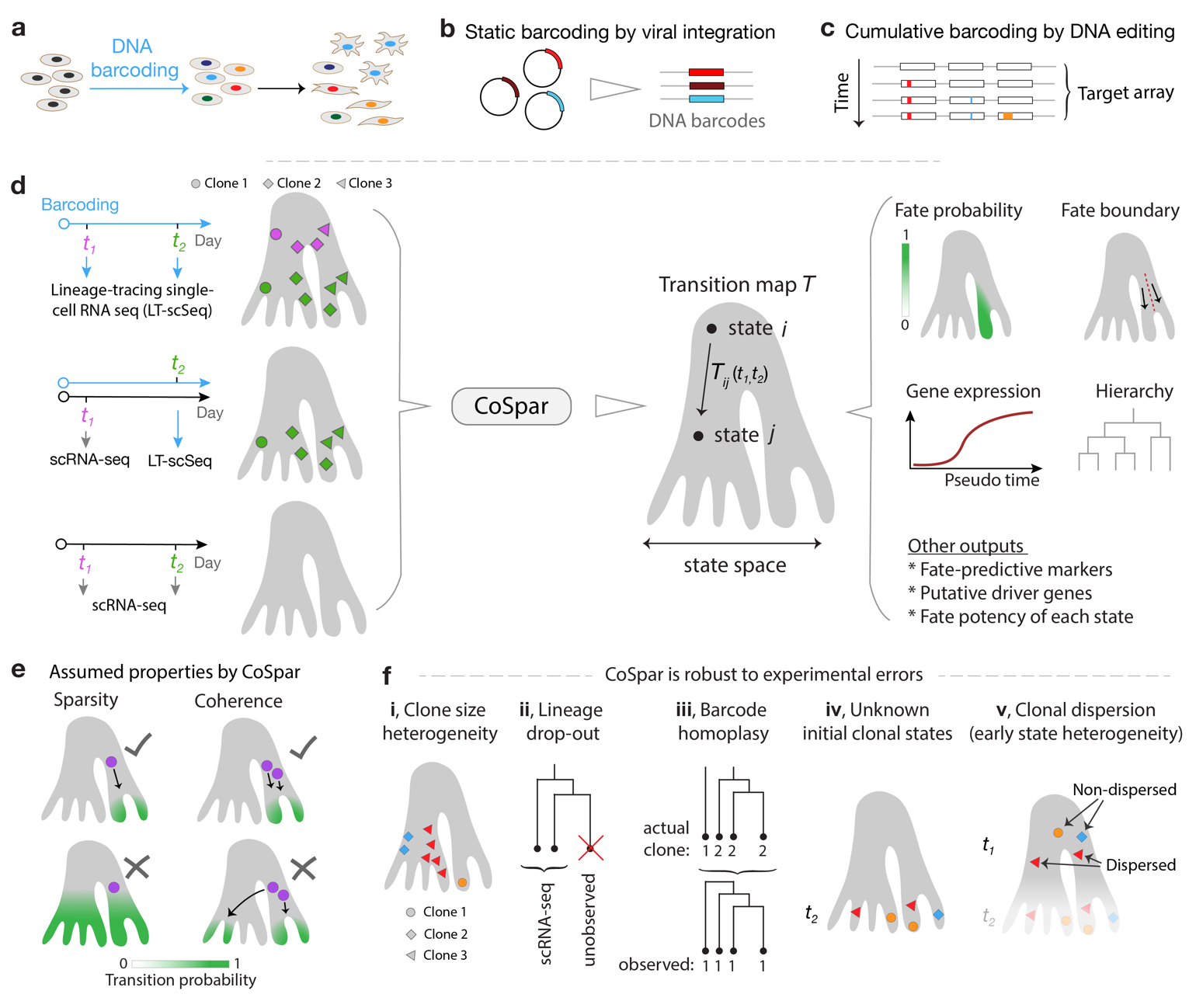
Here we have developed a mathematical framework for systematically inferring dynamic transitions by integrating state and clonal information. Extending compressed sensing, our method takes advantage of reasonable assumptions on the nature of biological dynamics: that cells in similar states behave comparably, and that cells limit their possible dynamics to give sparse transitions. Using reference datasets, we demonstrated that coherent sparse optimization relates molecular heterogeneity of cells to their future fate outcomes, in a manner that is robust to typical sources of experimental error (Fig. 1**f**), using as little as 5-10% data originally collected in prior experiments. The computational methods used in each original study to analyze clonal data were not similarly robust. CoSpar also successfully predicted early fate biases in these datasets using only clonal information from the last time point. When clonal data was removed entirely, results were sensitive to the choice of distance metric, and no single approach optimally inferred fate bias across all data sets.

The robustness of CoSpar could greatly simplify the design of LT-scSeq experiments, by enabling experiments with fewer cells, fewer clones, or fewer time points. In all three datasets considered here, CoSpar reveals clear early fate boundaries that were not previously reported, yet in agreement with the heterogeneity of key transcription factors and fate determinants. We predicted novel transcription factors and markers in each case, and they could facilitate enriching and manipulating the desired fate outcomes. As an example, we experimentally tested CoSpar’s prediction that LIFR expression levels in developing lung endoderm clones was associated with future outcomes by functionally validating that LIF supplementation in a 3-day window could augment the yield of lung epithelial descendants ten days later.

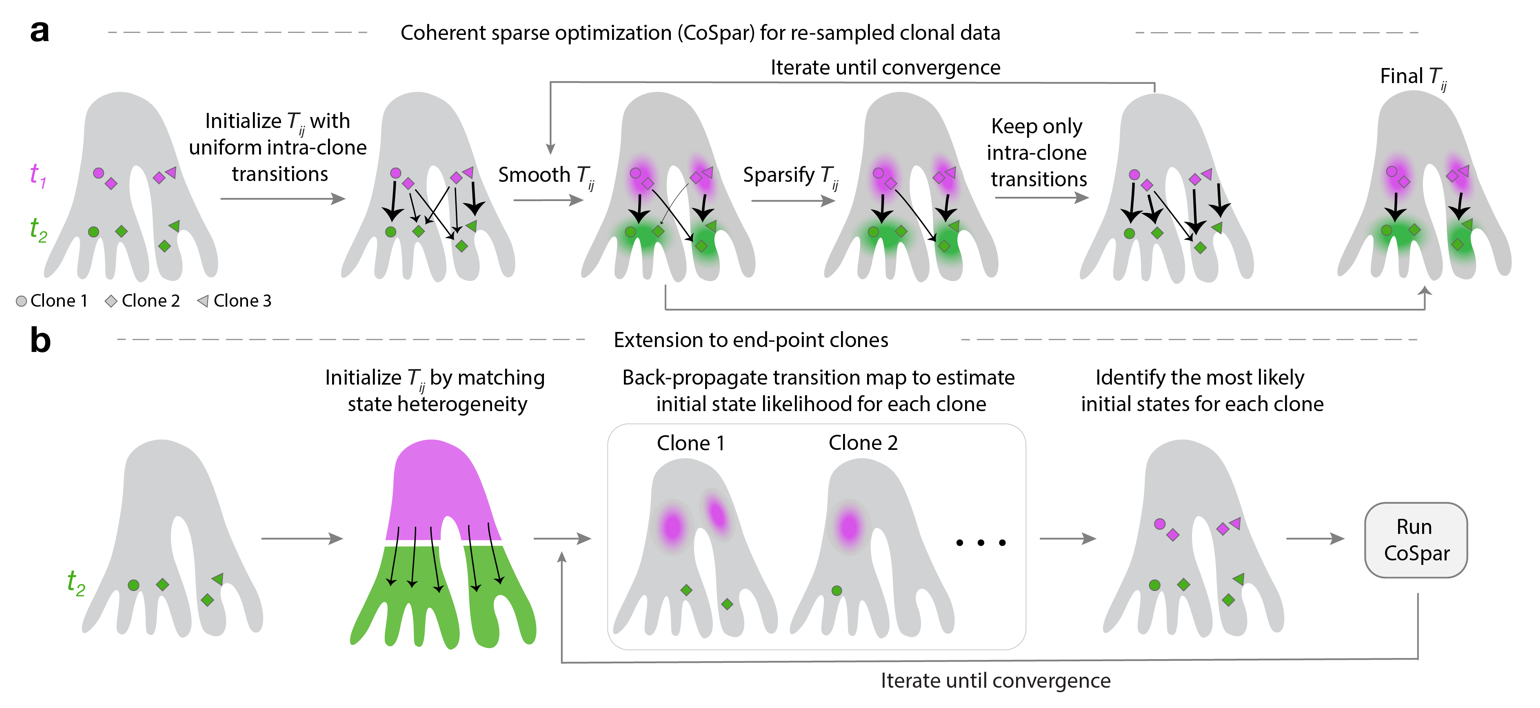
The examples we have analyzed specifically relate to LT-scSeq implemented using LARRY13,23 and CellTagging14, but CoSpar is not limited to these technologies. The state measurement can be transcriptomic (via scRNA-seq or RNA fluorescence in situ hybridization (FISH)40), as shown above, as well as proteomic and epigenomic; and lineage tracing can be achieved with static DNA barcodes13,23, endogenous mutations41, or exogenous DNA constructs that accumulate mutations over time, like CRISPR-based editing2,17,18,42,43. CoSpar can thus facilitate interpretation of the rapidly evolving field of LT-scSeq, and thus accelerate exploration of development and disease.

CoSpar also has limitations, which directly follow from its central assumption. By enforcing coherent fate choices between similar cells (Fig. 2**a**), CoSpar becomes sensitive to choices in measuring cell-cell similarity, and to the degree of smoothing used in implementing the algorithm (Supplementary Fig. 2**c**). Thus, CoSpar will fail to identify fate biases when heterogeneity relevant to cell fate is not measured, or when it is filtered out during data analysis, or due to over- or under-smoothing. In addition, when inferring progenitor bias from clones observed at a single late time point, CoSpar performs by matching heterogeneity in initial states to the structure of clones observed later. CoSpar can fail in this case when heterogeneity in the later population cannot be related to heterogeneity in the initial population. Despite these caveats, CoSpar provided sensible predictions in the cases examined here.

We expect coherent sparse optimization to be useful for applications beyond dynamic inference. Several problems require learning locally coherent maps from few and noisy measurements. Such problems occur, for example, when: integrating two sets of measurements in the same system44,45 (batch correction and multi-omics); decoding spatial transcriptomes from composite FISH measurements46; and inferring responses of a system to individual perturbations from composite perturbation readouts47–49. Outside of biology, the association of measurements in one modality with sparse measurements in another can occur in marketing and social networks50. Forcing coherence and sparsity constraints could greatly improve map inference in general, reducing the cost of data acquisition and enabling new discoveries.

****

**Fig. 1. Integrative analysis of lineage tracing and transcriptome data. a**, Lineage-tracing single cell genomics (LT-scSeq) experiments simultaneously measure cell phenotypes and clonal lineage (indicated by colors). **b-c,** LT-scSeq assays encode lineage information with static DNA barcodes or cumulative barcoding. **d,** CoSpar unifies analysis of different experimental designs to infer transition maps, which encode the probability of a progenitor cell in state *i* giving rise to a progeny observed in a later state *j*. Such maps reveal fate boundaries, lineage hierarchy, putative markers, and putative fate-determinants. **e,** Two key assumptions constrain dynamic inference by CoSpar. **f,** Five experimental challenges occur in clonal analysis. Single labeled cells can give rise to clones with a wide dispersion in size; LT-scSeq loses cells during analysis leading to loss of clonal structure; barcode homoplasy occurs when cells from different clones present the same barcode due to experimental limitations; progenitor states are not observed when clones are only observed upon tissue dissociation; clonal dispersion occurs when early clonal states are heterogeneous due to the lag time between barcoding and profiling.



**Fig. 2. The CoSpar algorithm.** **a,** When clones are resampled at two time points, a transition

map is inferred by iteratively enforcing observed clonal transitions, coherence (by smoothing)

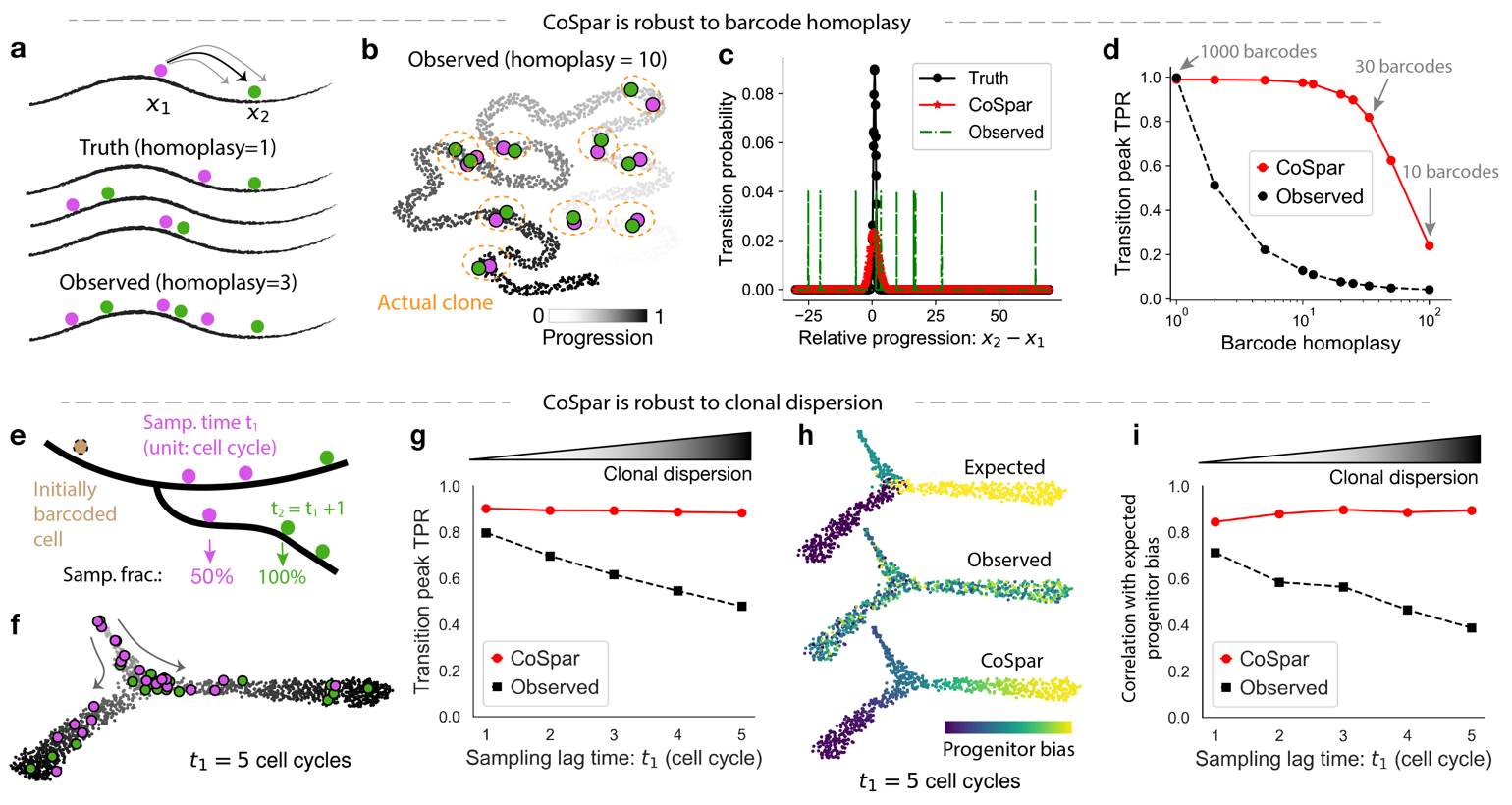
and sparsity until convergence is achieved. (See details and derivation in Methods and

Supplemental Note 3). **b,** When clones are observed only once, we infer their progenitor fate

bias and identity by first initializing a transition map without clonal information, then iteratively

(1) back-propagating the map to predict clonal progenitor identity and (2) learning the transition

map as in **a** until the map and progenitor identities jointly converge.

****

**Fig. 3. Proof-of-concept with simulated data**. **a-d**, Benchmarking transition map inference

with barcode homoplasy errors. **a,** Schematics of a simplified simulated LT-scSeq experiment

to evaluate the accuracy of CoSpar and its robustness to barcode homoplasy errors.

Homoplasy is simulated by assigning multiple clones with the same barcode. **b,** UMAP

embedding of simulated data. Cells labeled with one barcode are shown, with moderate

homoplasy (10 clones / barcode). **c,** Distribution of true and inferred transition map matrix

elements. Observed transitions are broadly distributed due to homoplasy errors, which

associate progenitor cells and their progeny across different clones. CoSpar suppresses such

transitions by enforcing sparsity and coherence. **d,** CoSpar is robust to severe barcode

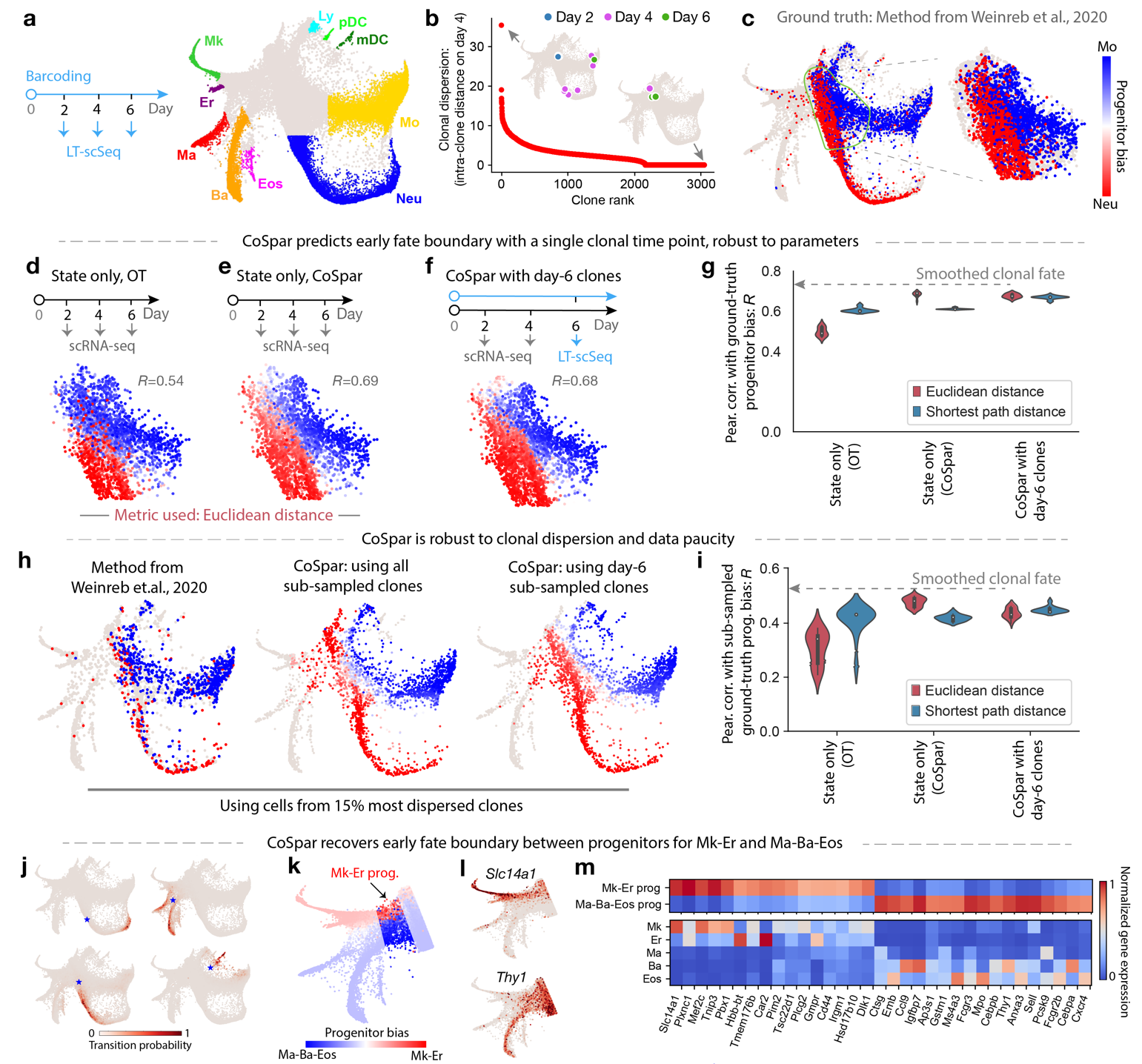
homoplasy, as seen from the fraction of predicted transitions within 3 standard deviations of

the true peak (TPR). **e-i,** Benchmarking transition map inference with clonal dispersion.

**e,** Schematics of a second simulated LT-scSeq experiment including variable lag times between clonal labeling and observation. **f,** UMAP embedding of simulated data, with one example clone shown. The clone is first observed 5 cell divisions after initial labeling.

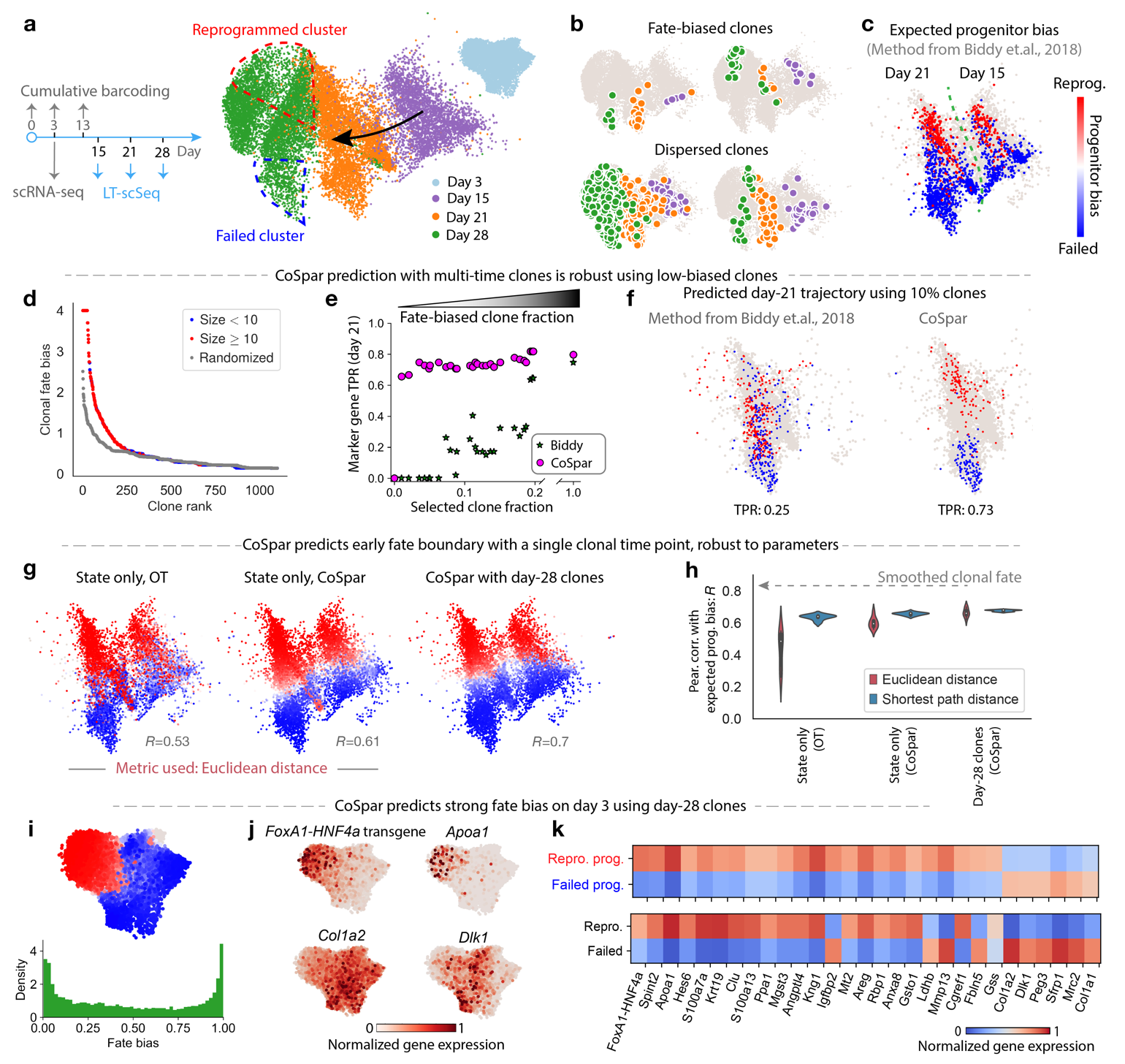
**g,** Quantitative evaluation of dynamic inference as a function of the sampling lag time. Growing lag time leads to higher clonal dispersion. Legend and transition peak TPR are defined as in **d**. **h,** Progenitor bias evaluated from the true and inferred transition maps with a simulated sampling lag time of five cell cycles. All clones are highly dispersed, providing no observed bias among early and late states; imposing sparsity enables recovering the true bias.

**i**, Quantification of the correlation between true and inferred progenitor bias (shown in **h**), over different sampling lag times.

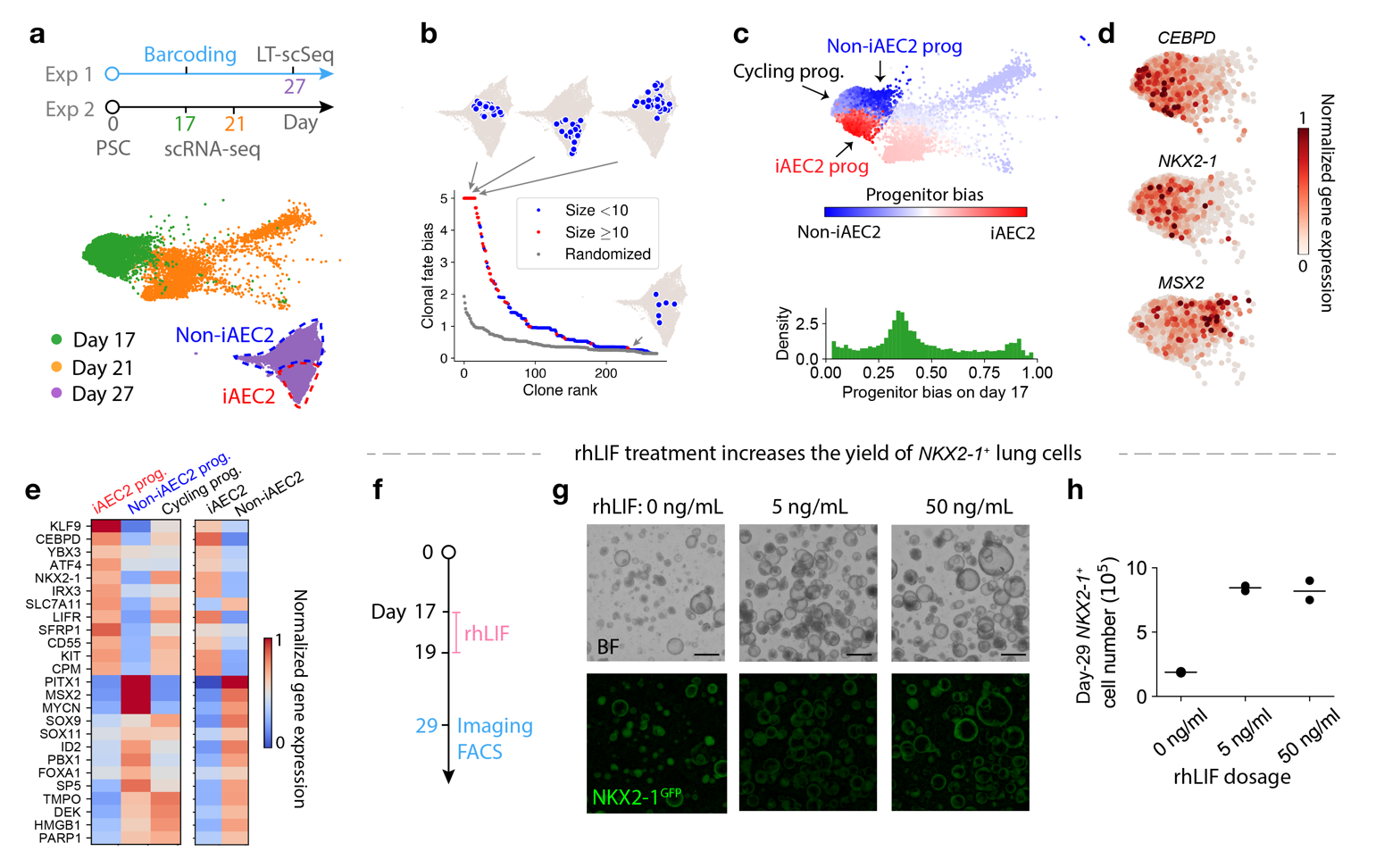


**Fig. 4. Benchmarking CoSpar and prediction of progenitor bias in hematopoiesis**. **a,** Experimental design and SPRING visualization of the hematopoiesis dataset from Ref13. Early hematopoietic progenitors differentiate into megakaryocyte (Mk), erythrocyte (Er), mast cell (Ma), basophil (Ba), eosinophil (Eos), neutrophil (Neu), monocyte (Mo), lymphoid precursor (Ly), migratory (ccr7+) DC (mDC), plasmacytoid DC (pDC). **b,** Clones ranked by intra-clone dispersion (i.e., mean intra-clone graph distance) over the observed cell states after 4 days of differentiation. Two illustrative clones are shown. **c**, Bias towards Mo or Neu fate evaluated from all clonal data using the original method in Weinreb et al13. Bias among early progenitors (right panel) serves as ground truth for benchmarking. **d,e**, Baseline inference of progenitor bias using optimal transport (OT) or CoSpar, using only state information but no clonal data.

**f,** CoSpar inference of progenitor bias using clonal data from a single-clonal time point. **g,** Violin plot showing the distribution of fate prediction outcomes, quantified by the Pearson correlation of the inferred fate bias with the ground truth. The distribution reflects differences in parameters for the OT method (which is used to initialized CoSpar) and choice of distance metric used, showing that clonal data reduces sensitivity to parameter choices in data analysis. Dashed line shows the upper limit expected from cross-validation of benchmarking. **h,** Fate bias inferred using only the 15% most dispersed clones (ranked in panel **b**). **i**, Violin plots showing the distribution in inference performance with the down-sampled data (quantified as in **f**) across parameter values*.* **j-m**, Predicting the transcriptional identity of Gata1+ Mk-Er and Ma-Ba-Eos progenitors using CoSpar. **j,** Representative values of the inferred transition map for 2-day transitions from 4 example cell states (indicated by \*). **k**, Heat map of predicted progenitor bias towards Mk-Er and Ma-Ba-Eos fates, overlaid on the state embedding. **l,** Expression of selected genes correlating strongly with predicted fate bias. **m**, Expression heat map for selected genes differentially expressed between the Mk-Er and Ma-Ba-Eos progenitors. Full list of fate-associated genes is provided in Supplementary Table 1.

****

**Fig. 5. Progenitor bias in fibroblast reprogramming**. **a**, Experimental design and UMAP visualization of cell reprogramming from fibroblast cells to induced endoderm progenitors (iEP) by ectopic expression of a transgene *FoxA1-HNF4a* on day 014. Schema shows time points for scRNA-seq only (grey arrows) and LT-scSeq (blue arrows). **b**, The UMAP visualization overlaid with examples of individual clones. Cells are colored by time point as in **a**. **c**, UMAP visualization of transcriptomes on days 15 and 21 of reprogramming, colored by progenitor bias towards successful or failed reprogramming fates, using cells in clones selectively filtered for strong fate bias as in the original study14. **d-f**, Benchmarking CoSpar using clones with weak fate bias. **d**, Clones ranked by consistency in the fate outcomes of their constituent cells [fate bias defined as -log(p-value), Fisher Exact test]. **e,** Accuracy in predicting the fate outcome of cells observed on day 21 using data from progressively fewer fate-biased clones. Predictions use the original method (Biddy et al.14) or CoSpar. Accuracy is assessed in the true positive rate (TPR) of identifying genes associated with fate outcomes previously reported in Ref14. **f**, UMAP visualization showing the cell states on day 21 predicted to undergo successful or failed reprogramming, when using the 10% clones with lowest fate bias. **g-h**, CoSpar predicts early progenitor bias with a single clonal time point, robust to parameters. **g**, Progenitor bias on days 15 and 21 predicted using only state information; or with end-point (day 28) clonal information only. **h**, Violin plots as in Fig. 4**g** quantifying prediction accuracy over a range of parameters, showing consistent improvement by imposing coherence, sparsity, and enforcing clonal relationships. **i-k**, Predicting early fate determination within 3 days of transgene expression. **i,** Predicted progenitor bias of cells on day 3. **j,** Expression on day-3 states of selected genes predicted to correlate with successful or failed reprogramming. **k,** Expression of additional genes differentially expressed on day 3 between cells predicted to succeed or to fail reprogramming. See the full list at Supplementary Table 2.

****

**Fig. 6. Progenitor bias during hPSCs differentiation into endodermal lineages**. **a**, Experimental design and UMAP visualization for differentiating human pluripotent stem cells (hPSC) into induced alveolar epithelium (iAEC2) lung cells and other endodermal cell types.

**b**, Clones ranked by fate bias towards iAEC2 fate (bias defined as in Fig. 5**d**), with representative biased (top) and dispersed (bottom) clones shown. **c,** Predicted progenitor bias of cells towards iAEC2 fate on day 17 of differentiation, overlaid on the state embedding and shown as a histogram. Cycling progenitors are identified as cells enriched in *TOP2A* or *MKI67* (Supplementary Fig. 5**d**). **d**,**e** Expression on day-17 states of selected genes predicted to correlate with iAEC2 and non-iAEC2 fates. In **e**, expression is shown alongside the corresponding expression in mature cells on day 27. **f-h,** rhLIF treatment increases the yield of *NKX2-* lung endoderm differentiated from human iPSCs. **f,** BU3 NGST cells were differentiated as previously described23 with the exception that cells were treated with 0, 5, and 50 ng/ml of rhLIF from D17-19 . **g,** Representative images of the cells on D29.Scale bar = 500µm**. h,** Quantification by flow cytometry of the total number of NKX2- cells resulting after completion of the protocol (n=3 shown for each condition).

**References**

1. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).

2. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0223-2.

3. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166–179 (2018).

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. 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).

6. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).

7. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).

8. Qiu, X. *et al.* Mapping Vector Field of Single Cells. 696724 (2019) doi:10.1101/696724.

9. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0591-3.

10. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

11. Tritschler, S. *et al.* Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, (2019).

12. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2467–E2476 (2018).

13. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).

14. Biddy, B. A. *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).

15. Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).

16. Rodriguez-Fraticelli, A. E. *et al.* Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* (2020) doi:10.1038/s41586-020-2503-6.

17. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

18. 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).

19. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).

20. Bowling, S. *et al.* An engineered CRISPR/Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* 797597 (2019) doi:10.1101/797597.

21. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

22. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).

23. Hurley, K. *et al.* Reconstructed Single-Cell Fate Trajectories Define Lineage Plasticity Windows during Differentiation of Human PSC-Derived Distal Lung Progenitors. *Cell Stem Cell* (2020) doi:10.1016/j.stem.2019.12.009.

24. 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).

25. Hormoz, S. *et al.* Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst* **3**, 419-433.e8 (2016).

26. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).

27. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 91–108 (2005).

28. Yu, V. W. C. *et al.* Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell* **167**, 1310-1322.e17 (2016).

29. Weissman, T. A. & Pan, Y. A. Brainbow: new resources and emerging biological applications for multicolor genetic labeling and analysis. *Genetics* **199**, 293–306 (2015).

30. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).

31. Ferreira, R., Ohneda, K., Yamamoto, M. & Philipsen, S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.* **25**, 1215–1227 (2005).

32. Lu, Y.-C. *et al.* The Molecular Signature of Megakaryocyte-Erythroid Progenitors Reveals a Role for the Cell Cycle in Fate Specification. *Cell Rep.* **25**, 2083-2093.e4 (2018).

33. Arinobu, Y. *et al.* Developmental checkpoints of the basophil/mast cell lineages in adult murine hematopoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 18105–18110 (2005).

34. Jacob, A. *et al.* Differentiation of Human Pluripotent Stem Cells into Functional Lung Alveolar Epithelial Cells. *Cell Stem Cell* **21**, 472-488.e10 (2017).

35. Rockich, B. E. *et al.* Sox9 plays multiple roles in the lung epithelium during branching morphogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4456-64 (2013).

36. Perl, A.-K. T., Kist, R., Shan, Z., Scherer, G. & Whitsett, J. A. Normal lung development and function after Sox9 inactivation in the respiratory epithelium. *Genesis* **41**, 23–32 (2005).

37. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).

38. Quinton, L. J. *et al.* Leukemia inhibitory factor signaling is required for lung protection during pneumonia. *J. Immunol.* **188**, 6300–6308 (2012).

39. Nogueira-Silva, C. *et al.* Leukemia inhibitory factor in rat fetal lung development: expression and functional studies. *PLoS One* **7**, e30517 (2012).

40. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).

41. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).

42. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

43. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

44. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* (2019) doi:10.1038/s41586-019-1773-3.

45. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).

46. Cleary, B. *et al.* Compressed sensing for imaging transcriptomics. *bioArxiv* 743039 (2020) doi:10.1101/743039.

47. Nitzan, M., Casadiego, J. & Timme, M. Revealing physical interaction networks from statistics of collective dynamics. *Sci Adv* **3**, e1600396 (2017).

48. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).

49. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

50. Aggarwal, C. C. *Recommender Systems: The Textbook*. (Springer, Cham, 2016).