Learning dynamics by computational integration of single cell genomic and lineage information

Shou-Wen Wang*,[1] and Allon M. Klein*,[1]

1 Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA
*Email: shouwen_wang@hms.harvard.edu (S.W.W.); allon_klein@hms.harvard.edu (A.M.K.)

Abstract

A goal of single cell genome-wide profiling is to reconstruct dynamic transitions during cell differentiation, disease onset, and drug response. Single cell assays have recently been integrated with lineage tracing, a set of methods that identify cells of common ancestry to establish *bona fide* dynamic relationships between cell states. These integrated methods have revealed unappreciated cell dynamics, but their analysis faces recurrent challenges arising from noisy, dispersed lineage data. Here, we develop coherent, sparse optimization (CoSpar) as a robust computational approach to infer cell dynamics from single-cell genomics integrated with lineage tracing. CoSpar is robust to severe down-sampling and dispersion of lineage data, which enables simpler, lower-cost experimental designs and requires less calibration. In datasets representing hematopoiesis, reprogramming, and directed differentiation, CoSpar identifies fate biases not previously detected, predicting transcription factors and receptors implicated in fate choice. Documentation and detailed examples for common experimental designs are 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 events occurring during differentiation[1–3], 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 metaplasia[1–3].

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 observing dynamic states by densely sampling cells at different stages[3–10]. However, scRNA-seq alone does not identify which early differences between cells drive or correlate with fate[2,11–13]. Conversely, 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 at an early time point[1–3]. 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 barcodes[2,13–21]. These technologies identify cells that share a common ancestor and define their genomic state in an unbiased manner. LT-scSeq experiments have been used to successfully identify when fate decisions occur[13,14], novel markers for stem cells[16], and pathways which control cell fate choice[14,16]. The simplest of these methods labels cells at one time point[13] (Fig. 1**b**); more complex methods allow the accumulation of barcodes over successive cell divisions to reveal the substructure of clones[2,13–21] (Fig. 1**c**).

Emerging LT-scSeq methods have been successful at revealing novel regulators of cell fate[14,16] and the fate potential of early progenitors[13,14], but they also present challenges that may limit their utility in practice. We identified at least five technical and biological challenges that affect experimental design and interpretation (Fig. 1**f**). These include stochastic differentiation and variable expansion of clones[22] (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 relationship[2] (Fig. 1**f-iii**), access to clones only at a single time point[23,24] (Fig. 1**f-iv**), and clonal dispersion due to a lag time between labeling cells and the first sampling (Fig. 1**f-v**). Addressing these problems should greatly simplify the design and interpretation of LT-scSeq assays and put them in the hands of a wider research community. To our knowledge, there is not yet an analysis method that systematically overcomes these problems.

Here, we develop a robust and generalizable computational approach to analyze LT-scSeq experiments. We begin with a model of clonal dynamics in which cells divide, differentiate, or are lost from the sampled tissue in a stochastic manner, with rates that are state-dependent (Supplementary Fig. 1**a**). We use this model to learn from the data the fraction of progeny of cells, initially in one state, which are found to occupy a second state after some time interval (Fig. 1**d,** Supplementary Fig. 1**b,c**). Our approach captures differentiation bias and fate hierarchies, and can reveal genes whose early expression is predictive of future fate choice.

Results

Dynamic inference from clonal data with state information.

A formalization of dynamic inference is to identify a transition map, a matrix $T_{ij}(t_1, t_2)$[7,25]. We define $T_{ij}(t_1, t_2)$ specifically as the fraction of progeny of a cell, initially in some state i at time t_1, that occupy state j at time t_2(Fig. 1**d**, Supplementary Fig. **1c**). This transition matrix represents a coarse-grained view of the cell dynamics : it already combines the effects of cell division, loss, and differentiation (Supplementary Fig. **1d**). As will be seen, even learning $T_{ij}(t_1, t_2)$ will prove useful for several applications (Fig. 1**d**).

We make two reasonable assumptions about the nature of biological dynamics to constrain inference of the transition map. We assume the map to be a sparse matrix, since most cells can access just a few states during an experiment (Fig. 1**e**, left panel). And we assume the map to be locally coherent, meaning that cells in similar states should share similar fate outcomes (Fig. 1**e**, right panel). These constraints together force transition maps to be parsimonious and smooth, which makes them robust to practical sources of noise in LT-scSeq experiments (Supplementary Fig. 1**e**). Box 1 formalizes the two constraints and lays out 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 $I(t)$ that encodes the clonal information at time t, 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. More often, one might observe clones at only one time point t_2. For these cases CoSpar jointly optimizes the transition map T and the initial clonal data $I(t_1)$ (Fig. 2**b**; Methods). In this joint optimization, one must initialize the transition map; 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 to the common problem of predicting dynamics from state heterogeneity alone without lineage data[7]. We extended CoSpar to this case. Thus, CoSpar is flexible to different experimental designs, as summarized in Fig. 1**d**.

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 $P(t)$. A transition map T directly links clonal density profiles $P(t_{1,2})$ between time points:

$$P_i(t_2) = \sum_j P_j(t_1) T_{ji}(t_1, t_2).$$

From multiple clonal observations, our goal is to learn T. To do so, we denote $I(t)$ 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 $P(t) \approx I(t)S(t)$.

With enough clonal information, $T(t_1, t_2)$ 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) T is a sparse matrix (Fig. 1**e**, left panel); 2) T is locally coherent (Fig. 1**e**, right panel); and 3) T is a non-negative matrix. With these requirements, the inference can be formulated as the following optimization problem:

$$\min_T \underbrace{\|T\|_1}_{\text{Sparsity}} + \underbrace{\alpha\|LT\|_2}_{\text{Coherence}}, \quad s.t. \underbrace{\|\mathbf{P}(t_2) - \mathbf{P}(t_1)T(t_1,t_2)\|_2 \leq \epsilon}_{\text{Clonal dynamics}} \; ; \; T \geq 0 \; ; \; \text{Normalization}$$

$||T||_1$ quantifies the sparsity of the matrix T through its L1 norm, and $||LT||_2$ quantifies the local coherence of T (L being the Graph Laplacian of the cell state similarity graph, and LT being the local divergence). The remaining constraints enforce the observed clonal dynamics, non-negativity of T, and map normalization, respectively. At $\alpha = 0$, the minimization takes the form of *Lasso*[26], 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* problem[27]. 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.

Computer simulations validate that CoSpar recovers dynamics with quantitative accuracy, and they establish that CoSpar inference is robust to two errors typical of LT-scSeq -- barcode homoplasy and clonal dispersion. 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 experiments[13,14]. With 1000 clones – typical of real experiments – mean transition rates inferred by CoSpar were within 3 standard deviations of the actual transition rate 98% of the time (TPR>98%, Fig. 3**d**) and the distribution of progeny fates showed 85% Pearson correlation to ground truth (Fig. 3**j**). Inferences remained similarly accurate with as few as 30 clones (Fig. 3**d**). CoSpar was robust to barcode homoplasy, and only detectably lost accuracy when all lineage barcodes mixed more than ten clones on average (Fig. 3**a-d**). This degree of homoplasy is far higher than expected in most experiments. Further, CoSpar was robust to clonal dispersion, simulated by sampling clones at increasing times post-barcoding (Fig. 3**f-i**). Conversely, approaches used in previous work, which average the transitions between cells observed in each clone at different time points[13], are 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 method from Weinreb et.al.[13] (Fig. 4**c**).

As a baseline for comparison, we applied CoSpar to predict HPC fate bias using state information alone (Fig. 4**e**). For this and further comparisons, we report the accuracy of fate prediction using Pearson correlation of predicted fate bias with that observed using all clonal data ('ground truth'). 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 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 method from Weinreb et al., which was used to analyze this data originally[13] (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), and Ba, Eos and Ma are produced by a different progenitor (BEMP)[30,31]. Existing studies of these progenitors are hampered by the lack of good markers. While molecular signatures of FACS-sorted MEP have been explored recently[32], less is known about the transcriptomic 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**). Applying CoSpar, we predict an early fate decision boundary between MEP and BEMPs (Fig. 4**j,k**), which correlates with the early expression of genes later associated with the resulting cell types (*Slc14a1* for Mk[32], *Thy1* for Ba[33]; Fig. 4**l**), and with the transcription factor (TF) *Cebpa* that regulates Eos and Ba differentiation[30]. We identified 377 known and novel putative fate-associated genes (Fig. 4**m**; Supplementary Table 1). Differences between the putative BEMPs and MEPs are evident in scRNA-seq data, and clonal data integrated by CoSpar supports that the differences are associated with functional fate bias. This analysis highlights that CoSpar can identify fate-predictive genes from limited LT-scSeq data.

CoSpar reveals early fate bias in reprogramming.

The second experiment we analyzed tracked cells during the reprogramming of fibroblast cells over 28 days into endodermal 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 reprogramming[14]**.** 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 study[14] 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 unexplored[14]. 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 identified multiple genes predicted to correlate with fate bias on day 3 and whose significance in reprogramming has not been previously established (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].

Discussion

Here we have developed a computational framework for systematically inferring dynamic transitions by integrating state and clonal information. It extends the problem of 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 published 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 sensitive to clonal dispersion and to down-sampling of data. 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.

The examples we have analyzed specifically relate to LT-scSeq implemented using LARRY[13,23] and CellTagging[14], but CoSpar is not limited to these technologies. The state measurement can be transcriptomic (via scRNA-seq or RNA fluorescence in situ hybridization (FISH)[38]), as shown above, as well as proteomic and epigenomic; and lineage tracing can be achieved with static DNA barcodes[13,23], endogenous mutations[39], or exogenous DNA constructs that accumulate mutations over time, like CRISPR-based editing[2,17,18,40,41]. 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 necessarily leans more strongly on state information, and it might fail 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.

Coherent sparse optimization could prove 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 system[42,43] (batch correction and multi-omics), decoding spatial transcriptomes from composite FISH measurements[44], and inferring responses of a system to individual perturbations from composite perturbation readouts[45–47]. Outside of biology, the association of measurements in one modality with sparse measurements in another can occur in marketing and social networks[48]. Forcing coherence and sparsity constraints could greatly improve map inference in general, reducing the cost of data acquisition and enabling new discoveries.

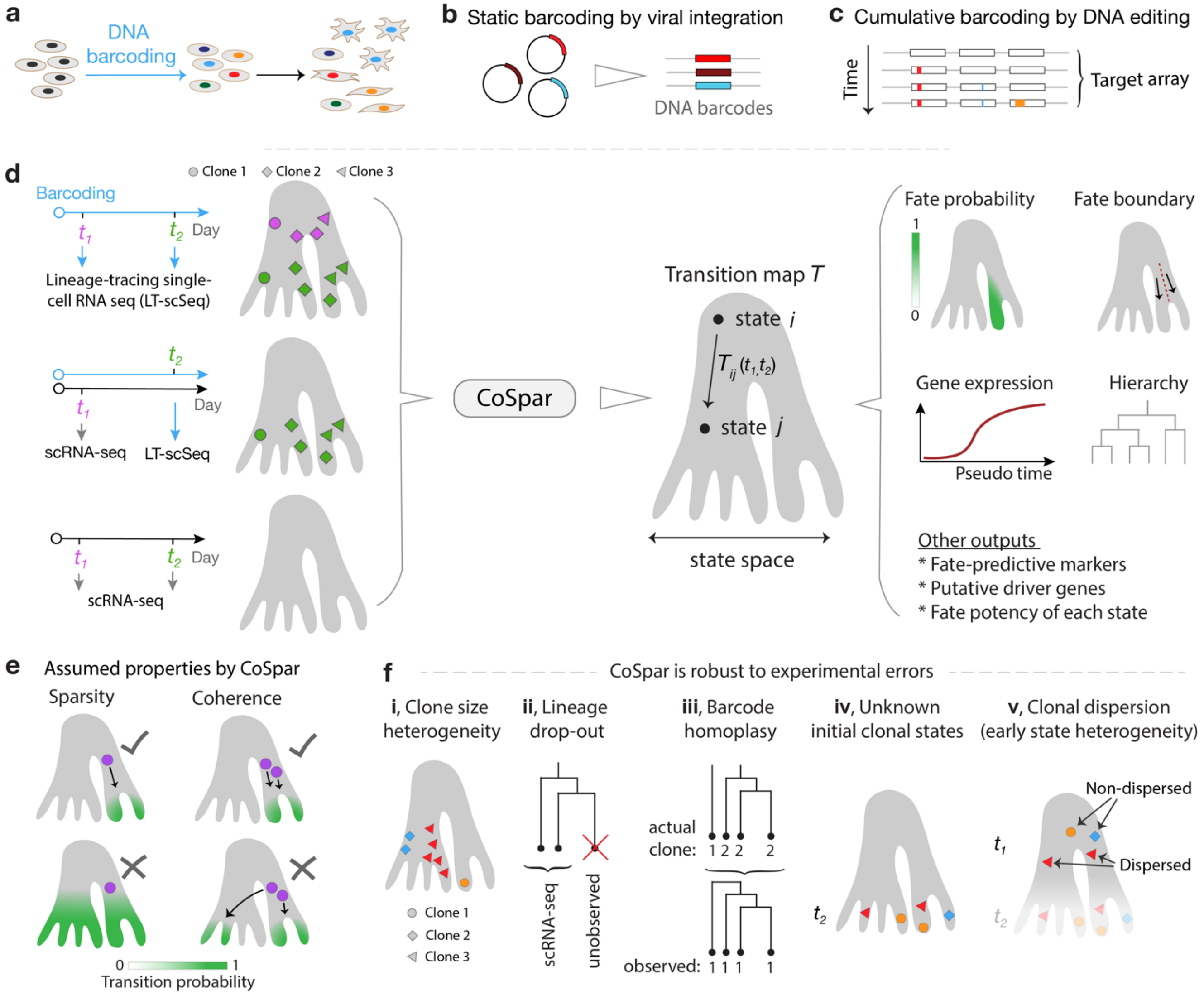


Fig. 1. Integrative analysis of lineage tracing and transcriptome data. a, Lineage-tracing single cell genomics (LT-scSeq) experiments simultaneously measure cell phenotypes and clonal lineage (indicated by colors). **b-c,** LT-scSeq assays encode lineage information with static DNA barcodes or cumulative barcoding. **d,** CoSpar unifies analysis of different experimental designs to infer transition maps (see text) to reveal fate boundaries, lineage hierarchy, putative markers, and putative fate-determinants. Here and below, the shaded gray regions schematically show a manifold of observed single cell genomic states. **e,** Two key assumptions constrain dynamic inference by CoSpar. **f,** Stereotypical challenges 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.

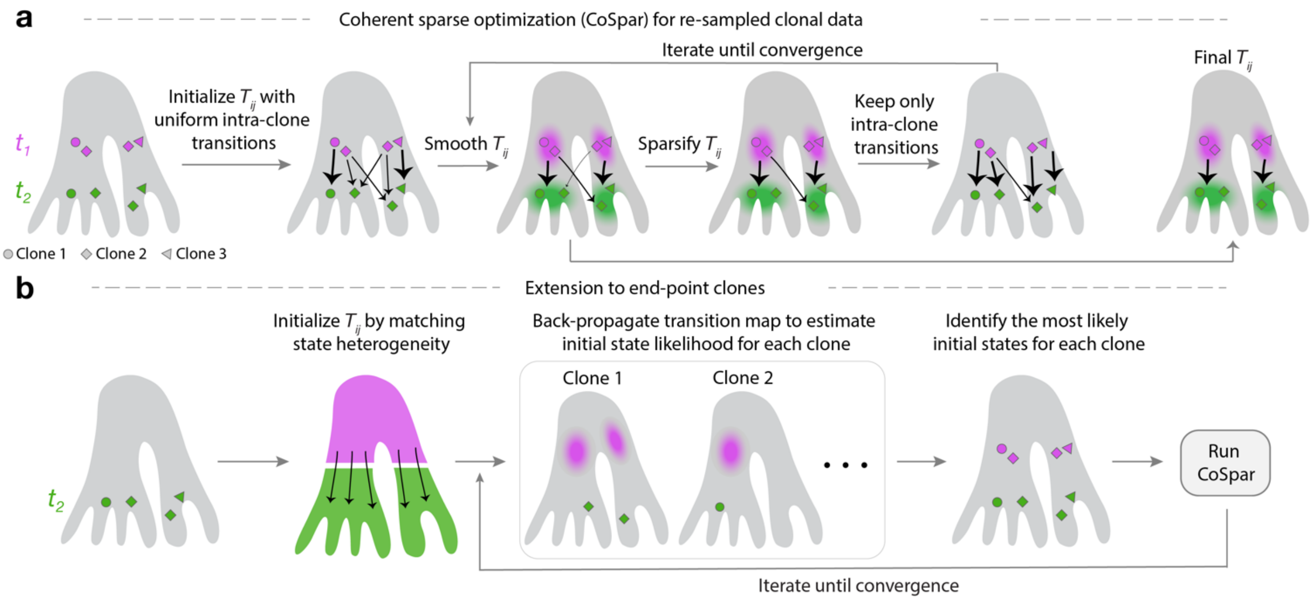

a
Coherent sparse optimization (CoSpar) for re-sampled clonal data
Iterate until convergence
Final T_{ij}
Initialize T_{ij} with uniform intra-clone transitions
Smooth T_{ij}
Sparsify T_{ij}
Keep only intra-clone transitions
t_1
t_2
● Clone 1 ◇ Clone 2 ◁ Clone 3
b
Extension to end-point clones
Initialize T_{ij} by matching state heterogeneity
Back-propagate transition map to estimate initial state likelihood for each clone
Identify the most likely initial states for each clone
Clone 1
Clone 2
t_2
Run CoSpar
Iterate until convergence

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.

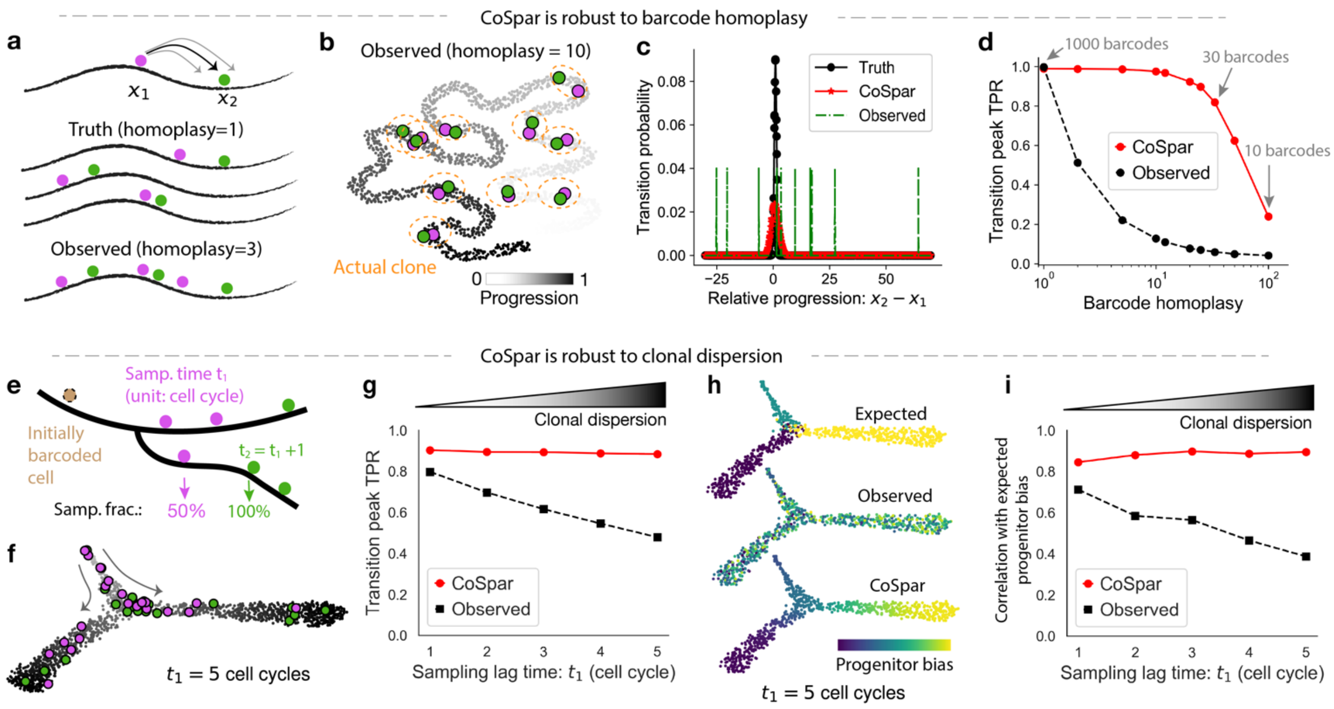


Fig. 3. Proof-of-concept with simulated data. **a-d**, Benchmarking transition map inference with barcode homoplasy errors. **a,** Schematics of a simplified simulated LT-scSeq experiment to evaluate the accuracy of CoSpar and its robustness to barcode homoplasy errors. Homoplasy is simulated by assigning 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.

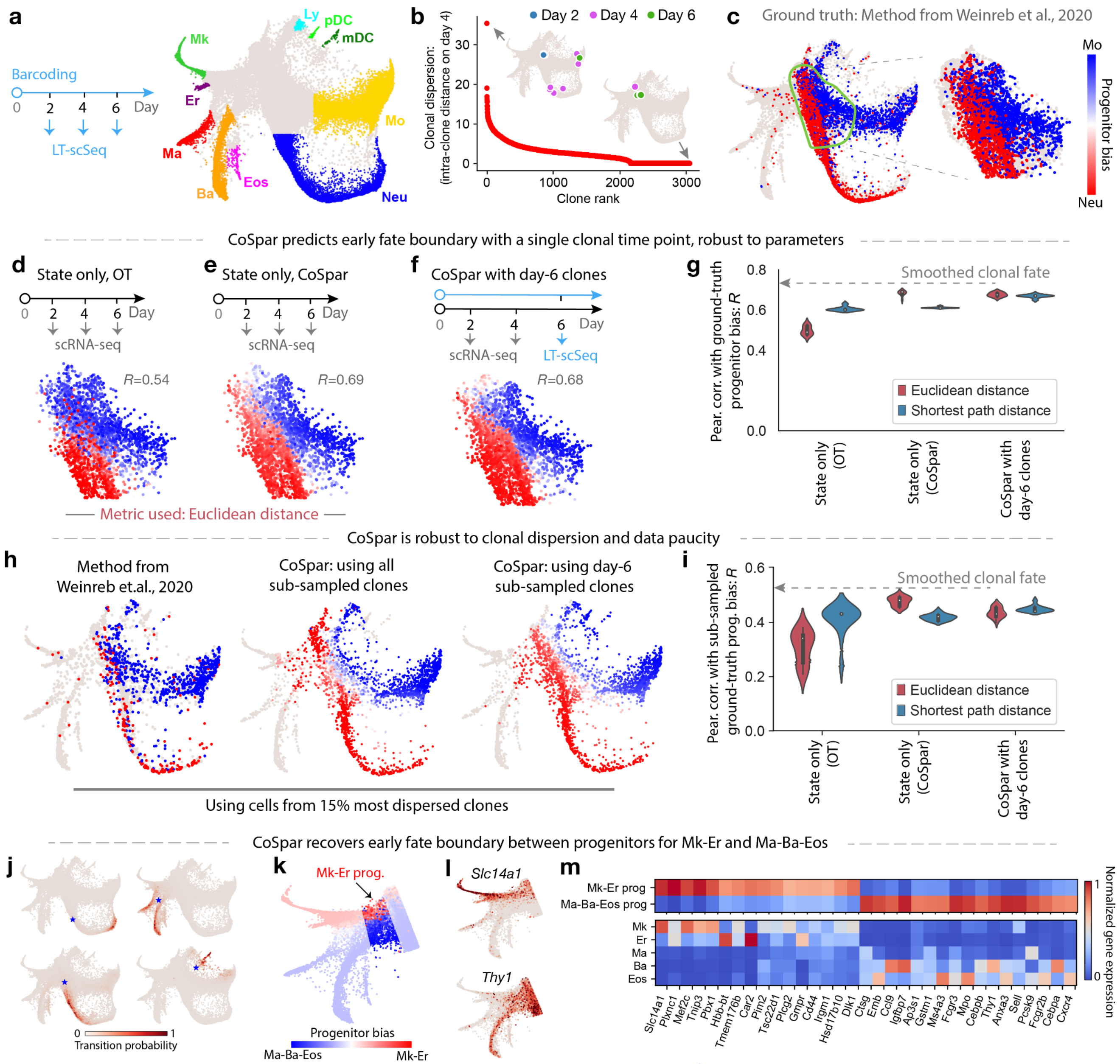


Fig. 4. Benchmarking CoSpar and prediction of progenitor bias in hematopoiesis. a, Experimental design and SPRING visualization of the hematopoiesis dataset from Weinreb et al.[13]. Early hematopoietic progenitors differentiate into megakaryocyte (Mk), erythrocyte (Er), mast cell (Ma), basophil (Ba), eosinophil (Eos), neutrophil (Neu), monocyte (Mo), lymphoid precursor (Ly), migratory (ccr7+) dendritic cells (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 al[13]. 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 initialize 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.

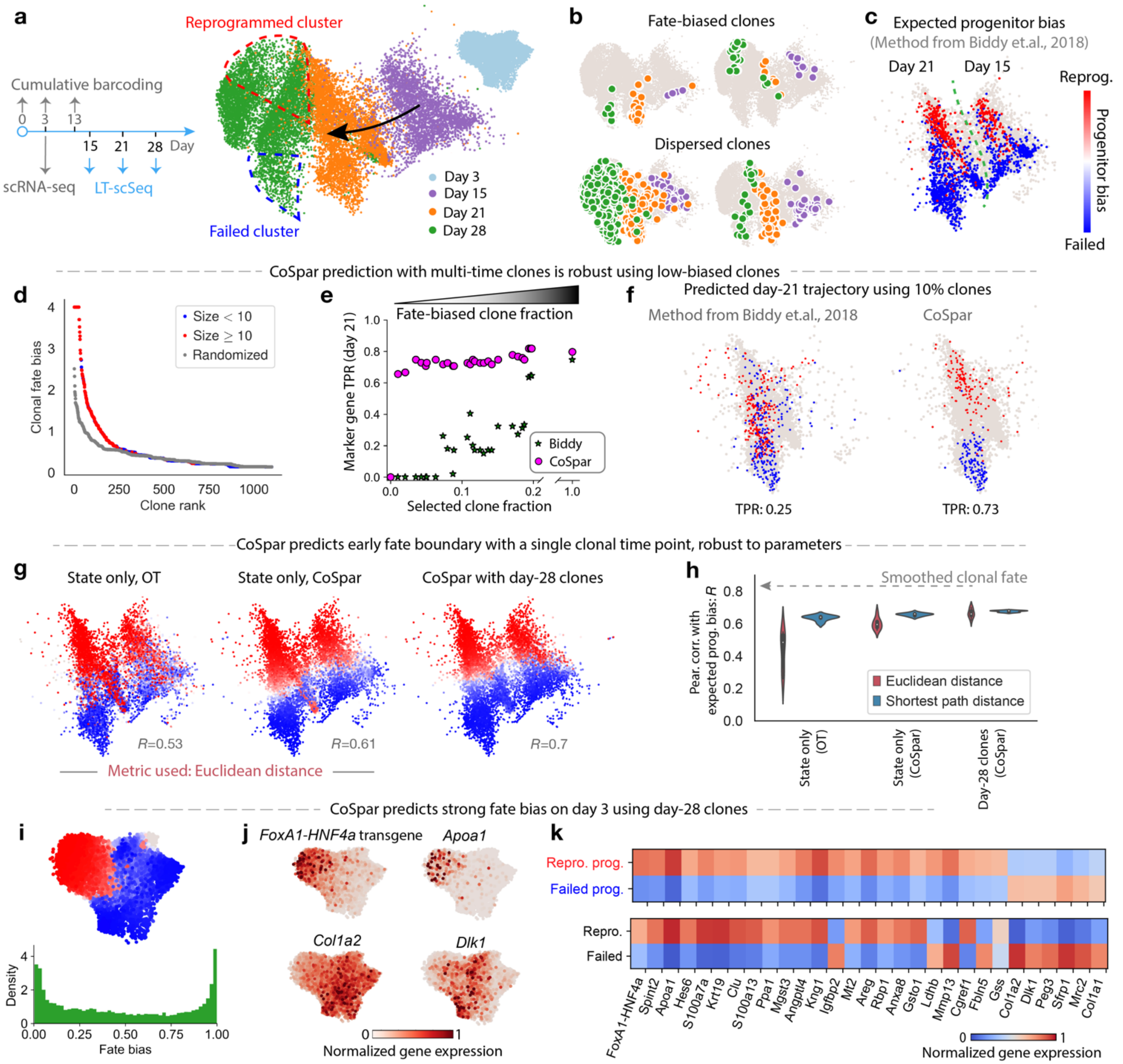


Fig. 5. Progenitor bias in fibroblast reprogramming. a, Experimental design and UMAP visualization of cell reprogramming from fibroblast cells to induced endoderm progenitors (iEP) by ectopic expression of a transgene *FoxA1-HNF4a* on day 0[14]. Schema shows time points for scRNA-seq only (grey arrows) 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 study[14]. **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 Ref[14]. **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.

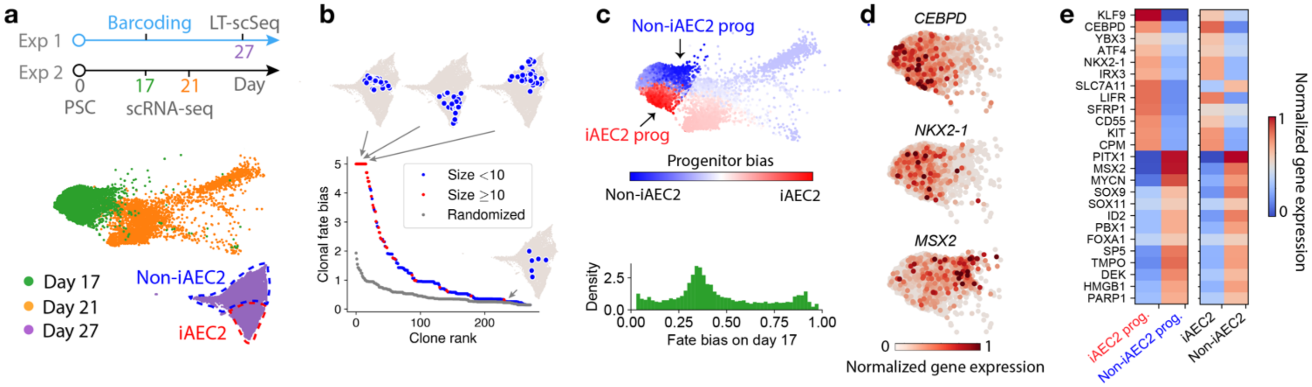


Fig. 6. Progenitor bias during hPSCs differentiation into endodermal lineages. **a**, Experimental design and UMAP visualization for differentiating human pluripotent stem cells (hPSC) into induced alveolar epithelium (iAEC2) lung cells and other endodermal cell types. **b**, Clones ranked by fate bias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

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

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

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

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

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

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

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

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

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

47. 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).

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

Methods

Definitions: states, transition maps, and clones. To formalize the problem of learning biological dynamics, we first define basic terminology. The observed **state** of a cell can include information on its transcriptome, epigenome, proteome, metabolic state, phospho-proteome, structural organization, or a combination of all of these. It may also include information on the environment of the cell, such as the transcriptome of neighboring cells, extracellular matrix composition, etc. These are quantified by a set of n features, $X \in \mathbb{R}^n$. Although X is continuous, it will be mathematically convenient to treat the accessible set of states as discrete. This is reasonable because experiments only sample a finite number of cells, so resolution into X is limited in practice. For convenience, we enumerate cell state as X_i, or more concisely as state i.

In a dynamical cellular system, cells are observed to occupy a distribution of states at consecutive times, with $P_i(t)$ giving the fraction of cells in state i at time t. We consider the **finite-time transition map** $T_{i'i}(t_1, t_2)$ as relating between experimental timepoints through the relationship[1]:

$$P_i(t_2) = \sum_{i'} P_{i'}(t_1) T_{i'i}(t_1, t_2) \quad (1)$$

The goal of our analysis is to learn $T_{i'i}(t_1, t_2)$, which in turn encodes information on the fate potential of cells in each state i, and the rate by which cells transition between states. In typical population-sampling experiments such as scRNA-seq, the transition map is shaped by the dynamics of cells, and by the rates of cell division and loss from the tissue (see Supplementary Note 1; Supplementary Fig. 1**d**). Errors in lineage tracing affect how well we can recover the transition map (see Supplementary Note 2).

Previous work has sought to infer $T_{i'i}(t_1, t_2)$, from $P_{i'}(t_1)$, $P_i(t_2)$ only[1]. Here we greatly constrain the inference problem using the dynamics of clones. By **clone** we mean a set of cell states (≥ 0 cells) that arise from a common ancestor cell. Experimentally, we use “clone” to mean a set of (≥ 1) cell states that share the same barcode, a genetically heritable element. Clones may be labeled by a static barcode, or by accruing barcodes through mutation or further integration events that label sub-clones. Barcode accrual allows a cell to associate with multiple detected clonal barcodes.

Data structures. Denoting the number of cells at time t as N_t, and the number of clones as M, we define:

$I(t) \in \{0,1\}^{M \times N_t}$: clone-by-cell matrix for the observed clonal data at time t, with discrete entries 0 or 1 indicating whether a cell belongs to a clone or not. We use $I_{mi}(t)$ to indicate its value for m-th clone at state i. For convenience, we sometimes use I_t to represent the matrix.

$\mathcal{I}_t^m$: the set of cell states at time t that belong to m-th clone.

$S(t) \in [0,1]^{N_t \times N_t}$: state-similarity matrix among cell states at time t.

$T \in \mathbb{R}^{N_{t_1} \times N_{t_2}}$: matrix of transition probability from cell states at t_1 to states at t_2.

$\pi \in \mathbb{R}^{N_{t_1} \times N_{t_2}}$: transition matrix that only allows intra-clone transitions (inter-clone transition amplitudes are set to 0).

$P_{\mathcal{C}_{t_2}} \in [0,1]^{N_{t_1}}$: fate map, i.e., a vector of probability for each initial cell state to transition to cluster $\mathcal{C}_{t_2}$at time t_2.

Dynamic inference with CoSpar. CoSpar seeks to minimize an objective function with a close connection to compressed sensing, as discussed in the main text. A heuristic, efficient algorithm implements the optimization through an iterative procedure (see main text for the objective function, and Supplementary Note 3 for its mathematical connection with compressed sensing). Referring to Fig. 2**a**, in each iteration, we 1) threshold the map to promote sparsity; 2) enforce clonal constraints by setting inter-clone transitions to be zero and performing clone-wise normalization; 3) locally average the transition map to promote coherence. These steps are described by the following pseudo-code. Full implementation and user guide are available at https://cospar.readthedocs.io.

```
function CoSpar ($I_{t_1}, I_{t_2}$)
    Initialization: $T_{ij}^{(0)} = 1\ \forall i, j$.
    For $l \leftarrow 1, 2, \dots, n_{cs}$ do
        $n \leftarrow n_{sm}(l)$
        Build similarity matrix: $S \leftarrow S^{(n)}$
        $\pi \leftarrow \mathcal{P}\left(\theta\left(T^{(l-1)}, \nu_{cs}\right)\right)$.
        Smoothing: $T^{(l)} \leftarrow [S(t_1)]^{+} \pi S(t_2)$.
        If $\mathrm{mean}_i \left[\mathrm{Corr}\left(T_{i\cdot}^{(l)}, T_{i\cdot}^{(l-1)}\right)\right] > 1 - \epsilon_{cs}$: Break
    return $T^{(l)}$, $\pi$
```

Here, + is a symbol for matrix transposition. Operators $\theta, \mathcal{P}$ and $S^{(n)}$ are defined below:

Definition of operators $\theta, \mathcal{P}$*.* Operator θ implements row-wise thresholding to promote sparsity:

$$[\theta(T, \nu)]_{ij} = \begin{cases} T_{ij} & \text{if } T_{ij} \geq \nu \max_j\ T_{ij} \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

where $\nu \in [0,1]$ is a parameter that tunes sparsity.

Operator $\mathcal{P}$ carries out clonal projection and normalization:

$$[\mathcal{P}(T)]_{ij} = \sum_m \frac{\tilde{\pi}_{ij}^m}{\sum_{i'j'} \tilde{\pi}_{i'j'}^m}, \tag{3}$$

where $\tilde{\pi}_{ij}^{m} = T_{ij}$ if the transition $i \to j$ occurs within clone m, and otherwise $\tilde{\pi}_{ij} = 0$. The normalization penalizes large clones, which tend to be more heterogeneous and less informative.

CoSpar has two outputs: the smoothed transition map T and the map π that only allows intra-clone transitions.

Similarity matrices $S^{(n)}$. We currently know of no natural choice for establishing the similarity of two states X_i, X_j. We found that a Graph diffusion process[2,3] recovered ground-truth results well in the simulations and experimental down-sampling analyses. CoSpar constructs a weighted kNN graph of observed cell states from a PCA embedding using the method proposed by UMAP[4], leading to a graph connectivity w_{ij} from state i to j that properly takes care of the heterogeneity of local cell density, with $w_{ii} = 0$. To make sure that transitions between two states are reversible, we symmetrize the connectivity: $\bar{w}_{ij} = (w_{ij} + w_{ji})/2$. Then, the random walk matrix is

$$\mathcal{M}_{ij} = \beta\delta_{ij} + \frac{(1-\beta)\,\bar{w}_{ij}}{\sum_k \bar{w}_{ik}},$$

where β controls the probability to stay at the original state after a unit step. We then introduce a family of similarity matrices:

$$S^{(n)} = [\mathcal{M}^n]^+. \tag{4}$$

The default method implemented in *scanpy.pp.neighbors* was used to construct the kNN graph at a specified neighbor number k_{cs}, with $\beta = 0.1$ and $k_{cs} = 20$.

Annealing steps $[n_1, n_2, \ldots]$. CoSpar iterates through different depths n of $S^{(n)}$, inspired by simulated annealing for finding the optimal solution in a rugged energy landscape[5]. Specifically, we use the sequence $\vec{n}_{sm} = [n_1, n_2, \ldots]$ to indicate the depths at each iteration.

Parameter choices. The following parameters of CoSpar are adjustable: 1) parameters used for building the random walk matrices $\mathcal{M}(t_{1,2})$, including β and k_{cs}; 2) the sequence $\vec{n}_{sm} = [n_1, n_2, \ldots]$ for generating annealing similarity matrix $S^{(n)}$; 3) the threshold ν_{cs} for promoting sparsity; and 4) parameters n_{cs} and ϵ_{cs} used to control iteration and convergence. We found 3 iterations are sufficient to obtain a convergent map (Supplementary Fig. 2**b,d**). Throughout this paper, we used a fixed iteration run $n_{cs} = 3$, and ignored ϵ_{cs} for computational efficiency. We also set $k_{cs} = 20$ and $\beta = 0.1$. We found CoSpar is more robust to ν_{cs} than to $\vec{n}_{sm}$ (Supplementary Fig. 2**a,c**)**.** Other parameters are given for each respective dataset below.

Extending CoSpar to single-time clones. When clonal data are available only at a single time point, dynamic inference is implemented as shown schematically in Fig. **2b**. Here only measurements on $I(t_2)$ are available. We jointly optimize the initial clonal data $I(t_1)$ and the transition map T. An iterative algorithm is used as defined here:

```
Function JointOptimization ($I_{t_2}$)
    $T^{(0)} \leftarrow T_{init}$
    For $l \leftarrow 1, 2, \ldots, n_{JO}$ do
        Infer $\hat{I}_{t_1}(T^{(l-1)}, I_{t_2})$
        $T^{(l)} \leftarrow$ CoSpar $(\hat{I}_{t_1}, I_{t_2})$
        If $\mathrm{mean}_i \left[\mathrm{Corr}\left(T_{i\cdot}^{(l)}, T_{i\cdot}^{(l-1)}\right)\right] > 1 - \epsilon_{JO}$: Break
    Return $T^{(l)}, \hat{I}_{t_1}$
```

Initialize the map, T_{init}. CoSpar uses optimal transport (OT) to construct the initialized map $T(t_1, t_2) = T_{init}$. Given an initial state distribution at t_1 and a later density at t_2, OT finds a map T_{int} that minimizes the transport cost to move the initial distribution to the later one. The approach is related to that developed in Waddington-OT (WOT)[1], but with a minor modification. To construct the OT cost matrix[1], approximated by a cell-cell distance matrix, CoSpar offers two approaches: 1) Euclidean distance in the selected PCA space, as implemented in WOT; 2) shortest path distance on a kNN graph of the state manifold. We found that shortest-path distance generally performs better than Euclidean distance (Supplementary Fig. 3**e**; Supplementary Fig. 4**c,d**). CoSpar accepts two parameters for this initialization: a k_{OT} for constructing the kNN graph, and a regularization parameter ϵ_{OT}.

Alternative initialization T_{init}. OT provides a reasonable initialization when the cell-cell distance matrix contains sufficient information to match the state heterogeneity at selected time points. When this assumption fails (e.g. owing to large differentiation effects over the observed time window, or batch effects), we initialize T using an alternative approach, in which we generate an artificial clonal matrix based on highly variable genes at both time points: $(\hat{I}_{t_1,}\hat{I}_{t_2}) \leftarrow \mathrm{HighVar}$, and then use it to calculate the initial transition map, $T_{init} \leftarrow \mathrm{CoSpar}\ (\hat{I}_{t_1}, \hat{I}_{t_2})$. See Supplementary Note 4 for further details.

Inferring the clonal matrix $\hat{I}_{t_1}(T, I_{t_2})$. Given a transition map T, CoSpar updates the clonal matrix $\hat{I}(t_1)$ based on the principle of maximum likelihood:

$$\hat{I}_{t_1} = \underset{I_{t_1}}{\mathrm{argmax}}\, P(I_{t_1} | T, I_{t_2}), \qquad (5)$$

under two constraints:
1) all initial states are clonally labeled, i.e. $\sum_{i,m} I_{mi}(t_1) = N_{t_1}$;
2) the fraction of cells with a given clonal barcode structure is constant over time. Note that this constraint represents a simplification as all clones initially derive from single cells and only develop to be heterogeneous in size over time. We provide an alternative enforcing each clone to have the same size at t_1, which is true for static barcoding at t_1. We found that the former constraint gives robust results over all tested datasets.

These two constraints are integrated as follows. With $\vec{\zeta} \in \{0,1\}^M$ indicating a clonal barcode combination, and $\mathcal{J}_t^{\zeta}$ indicating the set of cell states at time t with barcode combination $\vec{\zeta}$, the total number of cells with the barcode structure $\vec{\zeta}$ at time t is $N_t^{\zeta} \equiv |\mathcal{J}_t^{\zeta}|$. We enforce the constraint:

$$N_{t_1}^{\zeta} = N_{t_1} \frac{N_{t_2}^{\zeta}}{N_{t_2}^{*}},$$

where $N_{t_2}^{*}$ is the number of clonally labeled cells at t_2. As $N_{t_1}^{\zeta}$ is generally non-integer, we sample the cell number probabilistically from $\{\lfloor N_{t_1}^{\zeta} \rfloor, \lceil N_{t_1}^{\zeta} \rceil\}$, with a mean of $N_{t_1}^{\zeta}$, where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ take the floor and ceil of a number, respectively.

We provide a heuristic implementation for this optimization. First, rank all observed barcode structures $\vec{\zeta}$ from small to large values of $N_{t_1}^{\zeta}$. Then, sequentially infer the initial structure of each clone $\vec{\zeta}$:

1) compute from T the fate probability $P_{\mathcal{J}_{t_2}^{\zeta}}(i)$ that each state i in t_1 transitions to $\mathcal{J}_{t_2}^{\zeta}$, as defined below by Eq. (6);

2) select among not-yet-clonally-labeled cell states at t_1 the top $N_{t_1}^{\zeta}$ most likely initial cell states as the hypothetical initial states for this clone, and update the clonal matrix $\hat{I}(t_1)$ accordingly.

Parameter choices. The joint optimization accepts additional parameters 1) for initializing T (k_{OT} and ϵ_{OT} for the OT method, and gene selection parameter HighVar_gene_pctl for the HighVar method); and 2) for controlling iteration and convergence, i.e., n_{JO} and ϵ_{cs}. We found that one iteration is sufficient to obtain a convergent map for all tested datasets in this paper (Supplementary Fig. 2**e**). We set $k_{OT} = 5$, $\epsilon_{OT} = 0.02$, $n_{JO} = 1$ and ignored ϵ_{JO} throughout this paper. The remaining parameters are provided for each dataset below.

Toolkits for transition map analysis.

Fate map. From a transition map T, we can compute the probability for early states to enter a given set of states $\mathcal{C}_{t_2}$ (a fate cluster). This is a key output of CoSpar, and will be used to generate other important outputs including progenitor probabilities, fate boundary, and fate coupling, etc. We first row-normalize the transition map: $\tilde{T}_{ij} = T_{ij} / \sum_k T_{ik}$. The fate probability for an initial cell state i is given by

$$P_{\mathcal{C}_{t_2}}(i) = \sum_{j\in\mathcal{C}_{t_2}} \tilde{T}_{ij}. \tag{6}$$

The fate probability satisfies $P_{\mathcal{C}_{t_2}} \in [0,1]$.

Progenitor map. We compute the probability that a set of later states $\mathcal{C}_{t_2}$ originate from a given initial state by normalizing the fate probabilities $P_{\mathcal{C}_{t_2}}(i)$ towards the fate cluster $\mathcal{C}_{t_2}$:

$$\tilde{P}_{\mathcal{C}_{t_2}}(i) = \frac{P_{\mathcal{C}_{t_2}}(i)}{\sum_i P_{\mathcal{C}_{t_2}}(i)} \tag{7}$$

The progenitor probability satisfies $\tilde{P}_{\mathcal{C}_{t_2}} \in [0,1]$.

Progenitor bias. We compute the bias by which an early state contributes differently to two fate clusters. Given two progenitor maps $\tilde{P}_{\mathcal{A}}$ and $\tilde{P}_{\mathcal{B}}$ towards cluster $\mathcal{A}$ and $\mathcal{B}$, we compute the bias as

$$Q_i = \frac{\tilde{P}_{\mathcal{A}}(i)}{\tilde{P}_{\mathcal{A}}(i) + \tilde{P}_{\mathcal{B}}(i)} \tag{8}$$

The progenitor bias is within the range $[0,1]$. We set state i to have a neutral bias Q_i=0.5, if it a small contribution to both fates: $\tilde{P}_{\mathcal{A}}(i) + \tilde{P}_{\mathcal{B}}(i) \le \nu_0\, \tilde{P}^*$, where $\tilde{P}^*$ is the maximum progenitor probability across both fates, i.e., $\tilde{P}^* = \max\limits_{i,\mathcal{C}\in(\mathcal{A},\mathcal{B})} \tilde{P}_{\mathcal{C}}(i)$. We set $\nu_0 = 0.05$ in this paper.

Predictive genes. We perform differential gene expression (DGE) analysis between cells with different progenitor biases. The biased population towards fate $\mathcal{A}$ or $\mathcal{B}$ are given by

$$\mathcal{A}^* = \{\arg_i Q_i > \nu_{bias,\mathcal{A}}\}, \qquad \mathcal{B}^* = \{\arg_i Q_i < \nu_{bias,\mathcal{B}}\}, \tag{9}$$

where $\nu_{bias,\mathcal{A}}$ and $\nu_{bias,\mathcal{B}}$ are the corresponding thresholds. We perform DGE analysis between these two populations using the Wilcoxon rank-sum test with Benjamini-Hochberg correction. We rank the enriched genes (FDR<0.05) according to the expression fold change between population $\mathcal{A}^*$ and $\mathcal{B}^*$.

*Fate coupling (Supplementary Fig. **3d,f**).* We define fate coupling as the correlation of fate maps towards two fates. Specifically, we first compute the fate map $P_{\mathcal{C}}$ towards selected fate clusters. $P_{\mathcal{C}}$ is a $N_{t_1} \times n$ matrix where n is the number of selected fates, represented by cell sets $\mathcal{C}^{(1)}{}_{t_2}, \ldots, \mathcal{C}^{(n)}{}_{t_2}$. The raw coupling is given by

$$Y = P_{\mathcal{C}}^{+} P_{\mathcal{C}}. \tag{10}$$

Here, $Y_{ll'}$ sums over "joint probability" between fate cluster l and l' across all initial states. We normalize the coupling as $\tilde{Y}_{ll'} = Y_{ll'}/\sqrt{Y_{l'l'}Y_{ll}}$, which brings the self-coupling $\tilde{Y}_{ll}$ to 1, and $\tilde{Y}_{ll'} \in [0, 1]$.

*Clonal fate bias (Fig. 5****d****; Fig. 6****b****).* We evaluate the fate bias of a clone towards/against a given cluster as in[6] by quantifying the statistical significance of a clone's occupancy of a given transcriptional state (e.g. a cluster), when compared to that expected from a random sampling of cells. The P-value (or P_{value}) is computed with Fisher Exact test, accounting for the clone size. We then transform it into clonal fate bias $-\log_{10} P_{\text{value}}$, and rank each clone accordingly. We also provide the same rank plot for randomly sampled clones.

Analyzing simulated datasets.

*Linear differentiation (Fig. 3****a-d****, Supplementary Fig. 2****a-c****).* A cell trajectory was parameterized as a one-dimensional interval of length L. The dynamics were simulated with a homogenous transition map corresponding to a biased random walk $T_{x_1,x_2} = \mathcal{N}(x_2 - x_1; 1, \sigma)$, where $\mathcal{N}(\cdot\,; 1, \sigma)$ is the Gaussian distribution with mean 1 and standard deviation σ. Specifically, clones were simulated from this map by sampling $x_1 \sim \text{Uniform}(0, \text{L})$, and then $x_2 = x_1 + 1 + \xi$ with $\xi \sim$ Gaussian(0,σ). Each pair (x_1, x_2) defines a clone. A total of *N* clones were simulated. To simulate barcode homoplasy, clones were randomly mixed to give *M*<*N* clonal barcodes of uniform size. All observations of cell states were embedded in a 50-dimensional space $Z = (z_1, \dots, z_{50})$ by setting $z_1 = x$, and adding independent Gaussian noise $z_k = 0.2\xi$ to each of the remaining 49 dimensions. We used $\sigma = 0.5,\ L = 100,\ N = 1000$. The number of detected clonal barcodes *M* was variable as shown in the figure panels. CoSpar was applied with $\nu_{cs} = 0.2$, $\vec{n}_{sm}$=[5,5,5].

*Bifurcation and cell sampling (Fig. 3****e-i****).* A cell trajectory was parameterized as a one-dimensional interval of length *L*/2 bifurcating into two one-dimensional intervals of further length *L*/2 corresponding to fates A and B. To simulate a clonal resampling experiment, for each clone an initial barcoded cell was seeded at $x_0 \sim \text{Uniform}(0, \text{L})$ at $t = 0$. Cells were simulated to divide once at each unit time step, and all cells progressed along the trajectory according to a random walk, with $T_{x_1,x_2}(t_1, t_2) = \mathcal{N}(x_2 - x_1; t_2 - t_1, \sigma\sqrt{t_2 - t_1})$. As each cell transitions past the bifurcation point (L/2) it chose between fates A, B with probability 1/2. At $t = t_1$, we sampled cell states in each clone with a success rate 0.5 per cell. Successfully sampled cells were removed, and the remaining unobserved cells continued to divide and progress as described. The state of all remaining cells was profiled at $t_2 = t_1 + 1$. The observed cell states were embedded in a 50-dimensional observation space *Z* by first embedding in two-dimensions,

$$(z_1, z_2) = \begin{cases} (x, 0), & \text{if } x < L/2 \\ \left(\frac{x}{2}, \frac{x}{2}\right), & \text{if } x \geq L/2, \text{fate} = \text{A} \\ \left(\frac{x}{2}, -\frac{x}{2}\right), & \text{if } x \geq L/2, \text{fate} = \text{B} \end{cases}$$

and then adding independent Gaussian noise $z_k = 0.2\xi$ to each of the remaining 48 dimensions. We set $\sigma = 1, t_1 = 5,\ L = 10$. *M*=100 clones were simulated. CoSpar was applied with $v_{cs} = 0.2$, $\vec{n}_{sm}$=[10,10,10].

Evaluating CoSpar with simulated data. We defined the TPR (Fig. 3**d,g**) as the fraction of rows of the inferred transition map, T_{x_1,x_2}, for which the maximum transition rate is within 3σ of the expected peak position, i.e. $\text{TPR } = E[H(3\sigma - |\text{argmax}_{\Delta\text{x}}\, T_{\text{x}_1,\text{x}_1+\Delta\text{x}} - 1|)]$ where $E(\cdot)$ is the mean over all rows of T, and $H(z)$={1 for *z*>0; 0 otherwise}. The progenitor bias for the bifurcation model (Fig. 3**h,i**) was calculated according to Eq. (8). Each of the TPR and progenitor bias comparisons (Fig. 3**d,g,i**) shows averages after application of CoSpar to 5 independent simulations.

Benchmarking and applying CoSpar to hematopoiesis.

Pre-processing. Data[7] is available at Gene Expression Omnibus (GEO), accession number GSE140802. Data was preprocessed as originally described[7]: 1) UMI counts were normalized in each cell to the average across all cells; 2) highly variable genes were selected using the SPRING gene filtering function (filter_genes using parameters *min_vscore_pctl* =85 ,*min_counts*=3, *min_cells*=3)[8]; and 3) genes correlated with cell cycle were excluded from the highly-variable gene list (genes with correlation $C > 0.1$ to the signature genes defined by *Ube2c, Hmgb2, Hmgn2, Tuba1b, Ccnb1, Tubb5, Top2a*, and *Tubb4b*). The 2-dimensional embedding and state annotation of cells were as in[7], also available at the GEO website (GSE140802). We selected the top 40 Principal Components (PCs). Unless otherwise stated, we constructed kNN graph with $k = 20$ for downstream analysis.

Applying CoSpar. Code detailing implementation of CoSpar to the data is provided at https://cospar.readthedocs.io/. In brief, we evaluated the progenitor fate bias, identified putative driver genes, and computed the fate coupling as described above. The default parameters are $v_{cs} = 0.1, \vec{n}_{sm} = [20,15,10]$, and we initialize the transition map using the OT method for joint optimization.

*Intra-clone dispersion (Fig. 4**b**).* We quantified the intra-clone dispersion of a clone m as the maximum cell-cell distance $d(m,t)$ within a clone at time t $(t = 2,4,6)$, where the distance was measured by the shortest-path distance in the kNN graph at $k = 5$. Fig. 4**b** shows the dispersion normalized by the mean dispersion on day 2.

Transition map using the method from Weinreb et al[Citation error] *(Fig. 4**c,h**; Supplementary Fig. 3**a,b**, **g-i**).* We selected clones that have a unique fate at a later time point, where each mature fate cluster was defined as in Weinreb et al (see annotations at Fig. 4**a**). Multi-fate clones were discarded. Given this clone matrix $I^w(t)$, with $t = 2{,}4{,}6$, we computed the transition map as $T_{in}^w(t_1, t_2) = [I_{t_1}^w]^+ I_{t_2}^w$, where any initial cell state has the same probability to transition to any later cell state observed in the same clone. The ground truth progenitor bias in Fig. 4**c** shows the progenitor bias Q_i on day 2 and day 4 computed from $T_{in}^w(2,4), T_{in}^w(2,6)$, $T_{in}^w(4,6)$ using Eq. (8).

*Fate map reconstruction error (Supplementary Fig. 3**a,b**).* To allow comparison between methods, we used $\pi(4,6)$ from CoSpar with $\nu_{cs} = 0.2$ or $T_{in}^w(4,6)$ from the Weinreb method, constructed from sub-sampled clones on day 4-6, to compute the fate map $P_{\mathcal{C}}(i, t = 4)$ towards cells annotated with a given fate (cell set $\mathcal{C}$) according to Eq. (6). We evaluated the inferred maps by comparing them to a ground-truth fate map $P_{\mathcal{C}}^{true}(i, t = 2)$ from the Weinreb method with all clones from day 2-4. We evaluated the prediction using the Wasserstein distance[9] between the two distribution $P_{\mathcal{C}}$ and $P_{\mathcal{C}}^{true}$, restricted to the progenitor state space $\bar{\mathcal{C}}$ (i.e., excluding states belonging to fate $\mathcal{C}$). Note that $P_{\mathcal{C}}(i, t = 4)$ maps the fate probability of cells sampled on day 4, while $P_{\mathcal{C}}^{true}(i, t = 2)$ is for cells sampled on day 2. To compare the fate maps for these non-overlapping cell subsets, we computed the OT map T^{OT} from day-2 states to day-4 states with $k_{OT} = 5$ and $\epsilon_{OT} = 0.02$, using shortest-path distance. The Wasserstein distance is given by $d_{wass} = \sum_{i,j\in\bar{\mathcal{C}}} P_{\mathcal{C}}(i) T_{ij}^{OT} P_{\mathcal{C}}^{true}(j)$. We computed the Wasserstein distance for 3 major fates: Neutrophils, Monocytes, and Basophils, and reported the average.

*Waddington-OT (Supplementary Fig. 3**f**; Supplementary Fig. 4**e**).* Results shown were obtained using the WOT package (https://github.com/broadinstitute/wot)[1], using default parameters: $\epsilon_{OT} = 0.05$, $\lambda_1 = 1, \lambda_2 = 50$.

Benchmarking and applying CoSpar to fibroblast reprogramming.

Pre-processing. Data was downloaded from GEO, accession number GSE99915. We followed the same processing as described above for hematopoiesis, and removed cell-cycle-correlated genes with correlation score $|C| > 0.03$. We used UMAP (scanpy.tl.umap with *min_dist*=0.3) to generate the embedding.

In this dataset, cells were barcoded at three time points (day 0, 3, and 13). Following Biddy et.al.[6], we concatenated day-0 and day-3 barcodes to form a unique clonal ID for downstream analysis. However, keeping 3 barcodes per cell, thus allowing nested clonal structure, works equally well (Supplementary Fig. 4**f-h**). We also inherited their annotation for the reprogrammed cluster (obtained by email communication with the authors), and used their selected clones to define the ground truth for reprogramming and failed trajectories. The failed cluster (Fig. 5**a**) was defined as a leiden cluster (scanpy.tl.leiden with *resolution*=1.5) in the cells sampled at day 28, which highly expresses *Col1a2* (Supplementary Fig. 4**a**), a gene expressed in fibroblasts that failed reprogramming[6]. The reprogrammed and failed cluster were used to define the progenitor bias in this dataset.

Applying CoSpar. The default parameters are $\nu_{cs} = 0.2, \vec{n}_{sm} = [15{,}10{,}5]$, and we initialize the transition map using the OT method for joint optimization. See jupyter notebook implementation at https://cospar.readthedocs.io/.

*Selecting dispersed clones (Fig. 5**d,e**).* We first calculated for each clone the fraction γ_o of cells within the reprogrammed cluster. Dispersed clones are defined as occupying both the reprogrammed cluster and other states on day 28, thus having intermediate values of γ_o. We selected dispersed clones satisfying $R_- \le \gamma_o < R_+$, where $R_- = x$ and $R_+ = 0.4 - 2x$, and x parameterizes the window. This parameterization was chosen so that we could evenly exclude clones at both sides of the window when adjusting x. The fraction of clones within this window was used as an indicator for each sub-sampled dataset in Fig. 5**e**.

*Transitions using the method from Biddy et al (Fig. 5**e,f**).* Following Biddy et.al.[6], we first identified clones that are enriched or depleted in the reprogrammed cluster according to Fisher's Exact test. Among statistically significant clones ($P_{value} \le 0.05$), we selected cell states belonging to reprogramming clones ($\gamma_o > 0.4$) as putative reprogramming population $\mathcal{D}_r$, and classified cell states of low-reprogramming clones ($\gamma_o < 0.4$) as putative failed population $\mathcal{D}_f$.

To boost the performance for downstream analysis, we made the following modification to the original method in Biddy et.al.[6]. For a putative population ($\mathcal{D}_r$ or $\mathcal{D}_f$), we enriched for high-fidelity states by iteratively excluding clones with γ_o closest to 0.4 until the total number of cells in $\mathcal{D}_r$ or $\mathcal{D}_f$ was at or below 3,000.

*Calculating marker gene TPR (Fig. 5**e,f**, Supplementary Fig. 4**b**).* For a putative reprogramming ($\mathcal{D}_r$) and failed ($\mathcal{D}_f$) population predicted by either CoSpar or the Biddy method, we assessed their accuracy by the overlap of their top differentially expressed genes with those from the reference population (defined by the fate-biased clones selected by Biddy et.al.[6]).

To predict population $\mathcal{D}_r$ and $\mathcal{D}_f$ with CoSpar, we inferred T with $\nu_{cs} = 0.4$ and threshold the fate map $P_\mathcal{C}$ built from the intra-clone transition map $\pi = \mathcal{P}\big(\theta(T, 0)\big)$ as follows:

$$\mathcal{D}_x = \left\{\arg_i P_{\mathcal{C}_x}(i) > \nu_t \max P_{\mathcal{C}_x}\right\},\ \ x \in \{r, f\}$$

where, to enrich for high-fidelity states, $\nu_t = \max\,(0.5, \omega)$ and ω was chosen such that $|\mathcal{D}_x|$ is the largest value below 500.

For both CoSpar and the Biddy prediction, when $|\mathcal{D}_x| \le 200$, we increased the total cell number up to 200 by adding the nearest neighbors of selected cell states using the kNN graph defined by the full dataset. This step supports the statistical power of the differential gene expression (DGE) analysis.

Finally, we performed DGE analysis between $\mathcal{D}_r$ and $\mathcal{D}_f$, identified enriched genes for each population, and compared them with the reference. Specifically, we first calculated the P-value for each gene using the Wilcoxon rank-sum test, with Benjamini-Hochberg correction. We ranked them according to the expression fold change between $\mathcal{D}_r$ and $\mathcal{D}_f$, kept the top 50 genes enriched in $\mathcal{D}_r$ and another top 50 in $\mathcal{D}_f$, and excluded statistically insignificant ones (adjusted P-value ≥ 0.05). Denoting the resulting gene set for predicted population $\mathcal{D}_x$ as $\mathcal{E}_x$, and that from the corresponding reference population as $\mathcal{E}_x^{true}$, the marker gene TPR for this putative population is given by

$$TPR_x = \frac{|\mathcal{E}_x \cap \mathcal{E}_x^{true}|}{\max\{|\mathcal{E}_x|, |\mathcal{E}_x^{true}|\}}, \qquad x \in \{r, f\}$$

The final marker gene TPR for a given method (CoSpar or the Biddy method) was $(TPR_r + TPR_f)/2$.

Application of CoSpar to in vitro differentiation of lung endoderm.

Pre-processing. Data was downloaded from GEO, accession numbers GSE137805 and GSE137811. We selected highly variable genes using filter_genes function (*min_vscore_pctl*=80 ,*min_counts*=3, *min_cells*=3), and normalized the UMI counts per cell to 10000. We used the top 40 PCs to construct kNN graph with $k = 20$ for downstream analysis. We inherited the original embedding on day 17 and 21 by Hurley et.al.[10] (available at https://kleintools.hms.harvard.edu/tools/springViewer_1_6_dev.html?cgi-bin/client_datasets/nacho_springplot/allMerged), and used UMAP (scanpy.tl.umap with *min_dist*=0.3) to generate the embedding for day-15 and day-27 cells. The iAEC2 cluster is defined as the day-27 leiden cluster (scanpy.tl.leiden with *resolution*=0.5) that highly express *SFTPB* and *SFTPC* (Supplementary Fig. 5**a**), marker genes for iAEC2 cells[10].

Applying CoSpar. To apply joint optimization (Fig. 6**c**; Supplementary Fig. 5**f,g**), we initialized the transition map using the HighVar method with *HighVar_gene_pctl*=80, and ran CoSpar with $\nu_{cs} = 0.2$, $\vec{n}_{sm} = [20,15,10]$. See jupyter notebook implementation at https://cospar.readthedocs.io/.

Data availability

All data analyzed in this article are publicly available through online sources. The annotated data, results, and Python implementation are available at https://cospar.readthedocs.io/. The raw data for the hematopoiesis dataset can be accessed at Gene Expression Omnibus (GEO) database with accession number GSE140802, the reprogramming dataset via GSE99915, and the lung dataset with GSE137805 and GSE137811.

Code availability

The results reported in this paper and our Python implementation are available at https://cospar.readthedocs.io/.

Acknowledgements

SWW acknowledges support by Damon Runyon Computational Biology Fellowship. AMK acknowledges support by NIH Grant 1R01HL14102-01. We thank Tal Scully for helping with figures. We thank Darrell Kotton, Killian Hurley, and Michael J Herriges for feedback and support with AEC2 analysis, and for comments on the manuscript.

Author contributions

SWW and AMK conceived the project. SWW devised the computational method, wrote the package, and carried out CoSpar analyses. SWW and AMK wrote the manuscript. AMK supervised the project.

Competing interests

AMK is a founder of 1CellBio, Inc.

Supplementary references

1. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).
2. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
3. Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A. & Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **30**, 83–98 (2013).
4. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

5. van Laarhoven, P. J. M. & Aarts, E. H. L. Simulated annealing. in *Simulated Annealing: Theory and Applications* (eds. van Laarhoven, P. J. M. & Aarts, E. H. L.) 7–15 (Springer Netherlands, 1987).

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

7. 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).

8. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).

9. Peyré, G. & Cuturi, M. Computational Optimal Transport. *arXiv [stat.ML]* (2018).

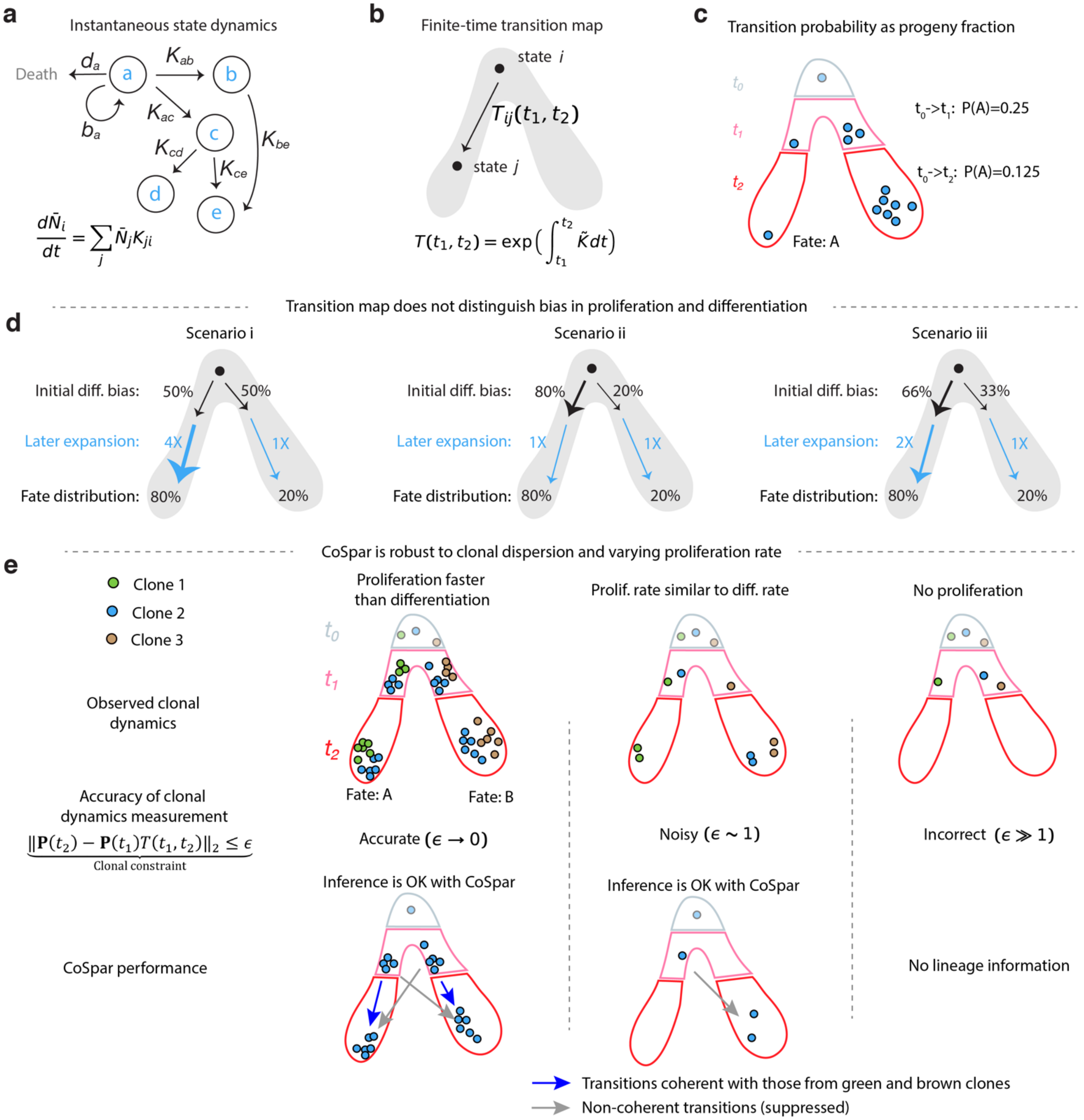
10. Hurley, K. *et al.* Reconstructed Single-Cell Fate Trajectories Define Lineage Plasticity Windows during Differentiation of Human PSC-Derived Distal Lung Progenitors. *Cell Stem Cell* (2020) doi:10.1016/j.stem.2019.12.009.

Supplementary Information for

Learning dynamics by computational integration of single cell genomic and lineage information

Shou-Wen Wang* and Allon M. Klein*

*Corresponding authors: shouwen_wang@hms.harvard.edu (S.W.W.), allon_klein@hms.harvard.edu (A.M.K.)

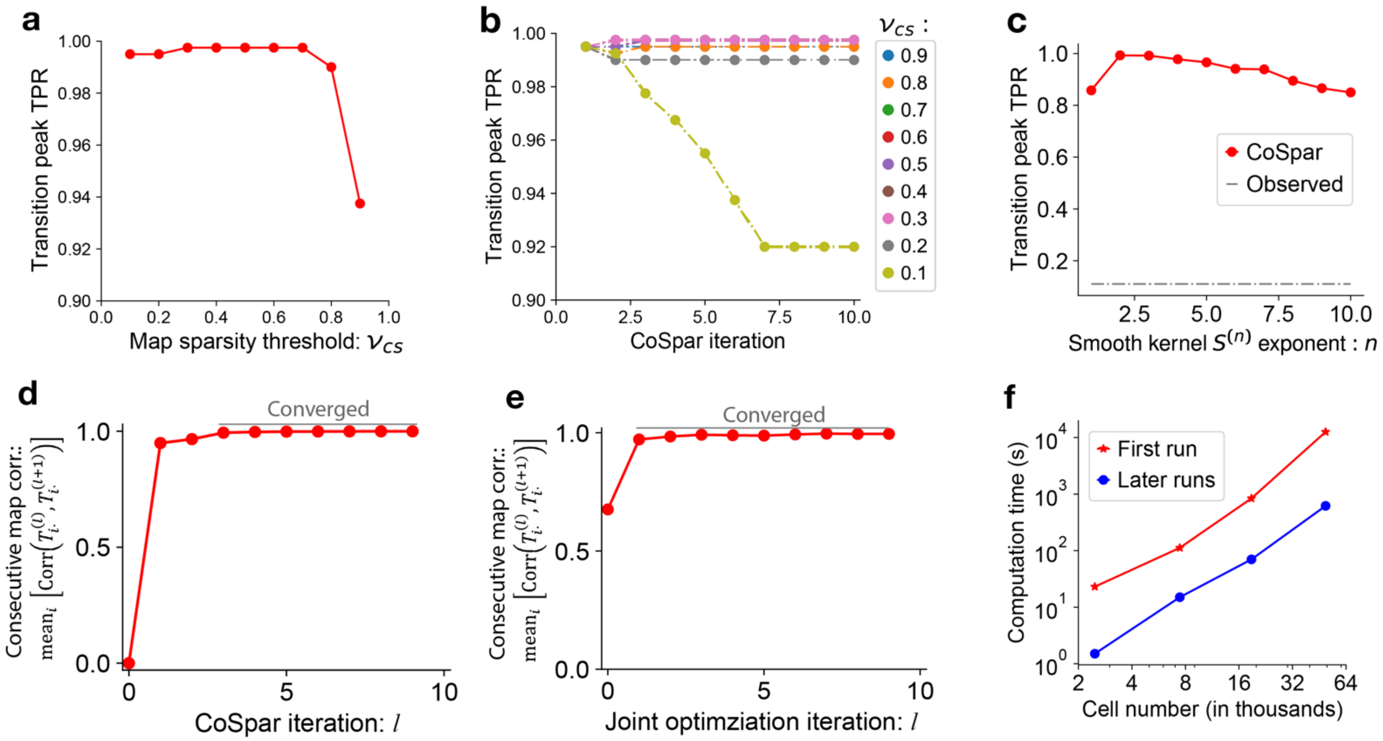

a
Instantaneous state dynamics
Death
d_a
K_{ab}
a
b
b_a
K_{ac}
c
K_{be}
K_{cd}
K_{ce}
d
e
$\frac{d\bar{N}_i}{dt} = \sum_j \bar{N}_j K_{ji}$
b
Finite-time transition map
state i
$T_{ij}(t_1, t_2)$
state j
$T(t_1,t_2) = \exp\left(\int_{t_1}^{t_2} \tilde{K} dt\right)$
c
Transition probability as progeny fraction
t_0
t_1
t_2
t_0->t_1: P(A)=0.25
t_0->t_2: P(A)=0.125
Fate: A
Transition map does not distinguish bias in proliferation and differentiation
d
Scenario i
Initial diff. bias: 50% 50%
Later expansion: 4X 1X
Fate distribution: 80% 20%
Scenario ii
Initial diff. bias: 80% 20%
Later expansion: 1X 1X
Fate distribution: 80% 20%
Scenario iii
Initial diff. bias: 66% 33%
Later expansion: 2X 1X
Fate distribution: 80% 20%
CoSpar is robust to clonal dispersion and varying proliferation rate
e
● Clone 1
● Clone 2
● Clone 3
Proliferation faster than differentiation
Prolif. rate similar to diff. rate
No proliferation
t_0
t_1
t_2
Observed clonal dynamics
Fate: A
Fate: B
Accuracy of clonal dynamics measurement
$\underbrace{\|\mathbf{P}(t_2) - \mathbf{P}(t_1)T(t_1,t_2)\|_2 \leq \epsilon}_{\text{Clonal constraint}}$
Accurate ($\epsilon \to 0$)
Noisy ($\epsilon \sim 1$)
Incorrect ($\epsilon \gg 1$)
Inference is OK with CoSpar
Inference is OK with CoSpar
CoSpar performance
No lineage information
Transitions coherent with those from green and brown clones
Non-coherent transitions (suppressed)

Supplementary Fig. 1. Models, assumptions and limitations of Coherent Sparse Optimization.
a, Simple example of the class of stochastic models that CoSpar seeks to learn. In such models, each node represents an observed cell state. In practice, thousands of measured states are included; here only five are shown. At each state cells self-renew, die, or differentiate with state-specific rates. The mean fraction of cells in each state evolves according to coupled first-order equations as shown. See Supplementary Note 1 for details.
b, The empirically-observed finite-time transition map can be interpreted through its relation to the transition rate matrix K (see panel **a**). See Supplementary Note 1 for details.

c, Schematics illustrating the operational, experimentally-accessible definition of a transition probability, as the average fraction of progeny derived from an initial cell *i* at t_0 that differentiates into a target state *j* at later times. As defined, transition probabilities are sensitive to biases in fate choice, and to differential rates of cell division and cell loss.

d, Schematics exemplifying that transition maps cannot distinguish fate bias from differences in net rates of cell expansion (division – loss). Three different underlying dynamics lead to the same transition maps.

e, Schematics clarifying the robustness of CoSpar to clonal dispersion (demonstrated in Fig. 3). i), When cells undergo extensive proliferation prior to fate bifurcation and clonal sampling, each clone densely samples several differentiation trajectories. By imposing sparsity and coherence, CoSpar re-enforces a minimal number of transitions that explain dynamics across all clones. ii), At lower rates of proliferation, fewer cells from each clone are sampled, and it may lead to observing clonally-related cells at different time-points on different trajectories, as shown (blue clone sampled towards fate A at t_1, and towards fate B at t_2). By enforcing coherence between clones rooted in neighboring states, CoSpar may still recover a correct transition map. In this case, there is a trade-off in the CoSpar cost function between minimizing the clone transition map error and maximizing coherence. iii), Lacking proliferation, one cannot establish clonal relationships that constrain dynamic inference.

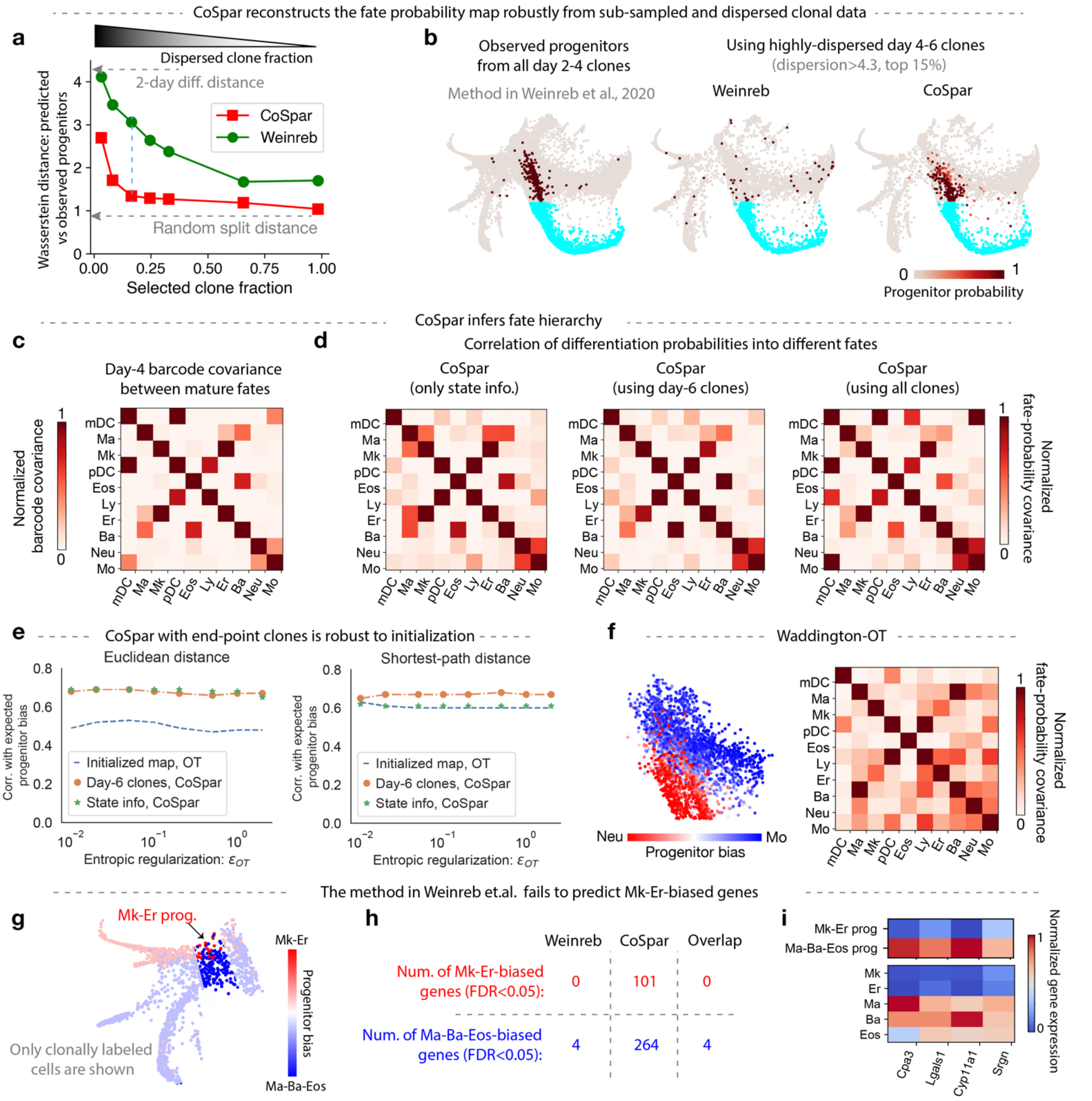


Supplementary Fig. 2. Evaluating CoSpar performance across parameter sweeps.

a-c, Performance of CoSpar using simulations as in Fig. 3**a-d** with a range of algorithm parameters (see Methods for parameter definitions): (**a**) sparsity threshold $\nu_{cs} \in [0,1]$; (**b**) number of iterations, showing convergence; (**c**) smoothing kernel exponent.

d,e, Demonstration of algorithm convergence, seen in the correlation between maps from consecutive iterations against the number of iterations, for the two algorithms (CoSpar, and Joint CoSpar, see Methods). The maps analyzed here correspond to those from the down-sampled hematopoietic dynamics (Fig. 4**h**).

f, Computational time to convergence, as a function of total cell number. In the first run, CoSpar will generate (and save) a similarity matrix, which is very costly (red curve). CoSpar can use similarity matrices generated previously to speed up computation (blue curve).



Supplementary Fig. 3. Benchmarking CoSpar in hematopoiesis.
a, CoSpar reconstructs transition maps from sub-sampled and dispersed clonal data. Here, we evaluate the prediction error as the Wasserstein distance between fraction of cell progeny predicted to occupy a given fate, compared to that obtained from the 'ground truth' transition map constructed using all clonal data rooted in day 2 clones (see main text). In **a**, the prediction error is assessed for a decreasing fraction of day 4-6 clones, obtained by progressively excluding less dispersed clones that contribute the strongest signal (see Fig. 4**b**). Green curve is obtained by applying the method from the original paper. A lower bound on the error (random split distance) is the Wasserstein distance between random 50% partitions of the ground-truth data. The largest observed errors are

comparable to the Wasserstein distance between populations separated by two days of progressive differentiation (upper grey arrow).

b, The ground truth and predicted fate maps for neutrophils cluster using the 15% most dispersed clones. These plots illustrate one value on the plot in **a**.

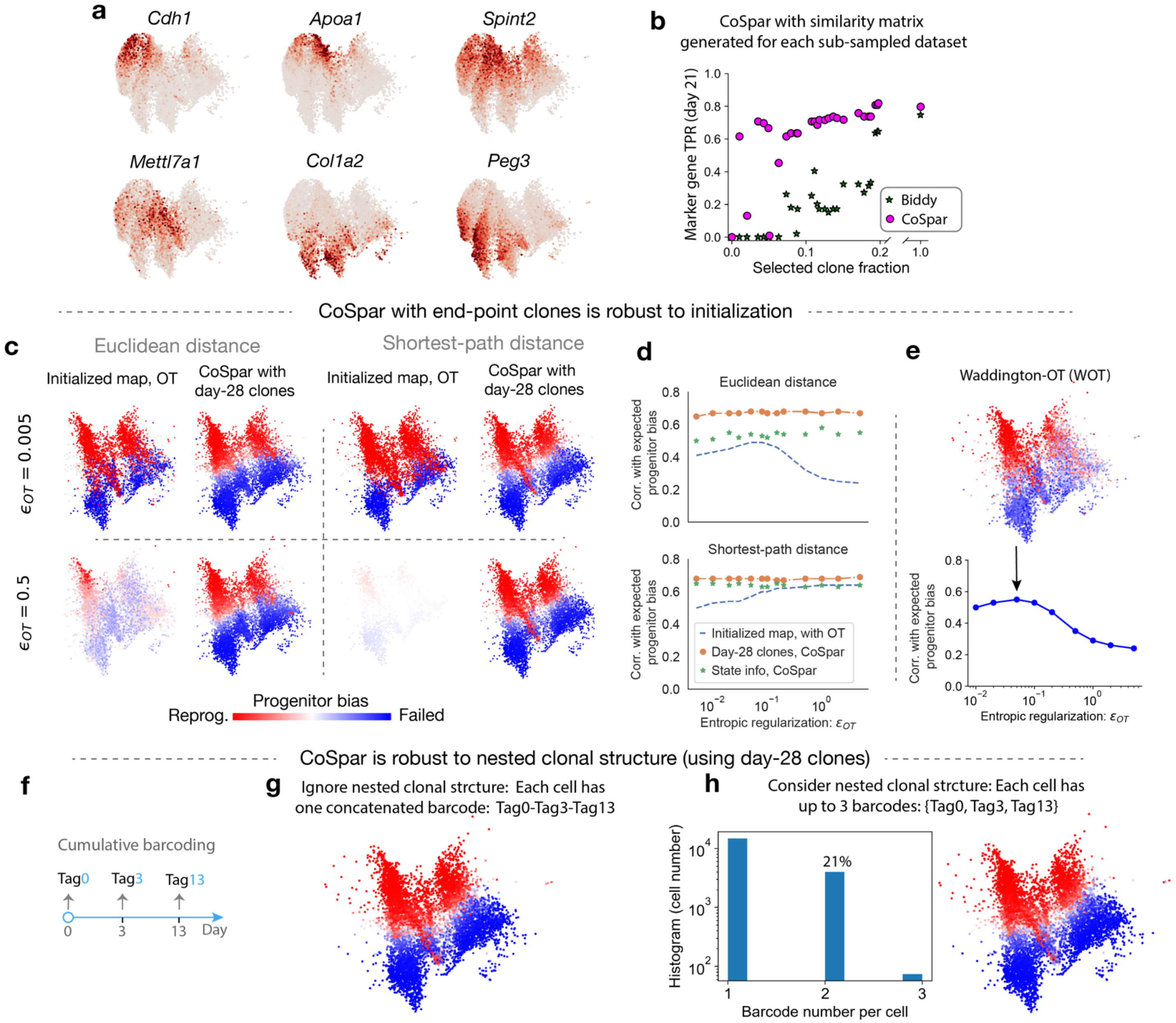
c, The normalized covariance of clonal barcode abundances between different cell types, calculated using all data on day 4 of differentiation[1].

d, The correlation of predicted transition probabilities of progenitors, inferred with CoSpar using different data indicated (See Methods).

e, Joint CoSpar optimization is robust to initialization and choice of distance metric. This panel accompanies Fig. 4**g**. Plots show the correlation of progenitor biases calculated from the transition maps for different initialization choices of the transition map. Optimal transport (OT) is used to initialize the transition map from state information alone prior to CoSpar. Plots scan the OT entropic regularization strength ϵ_{OT}.

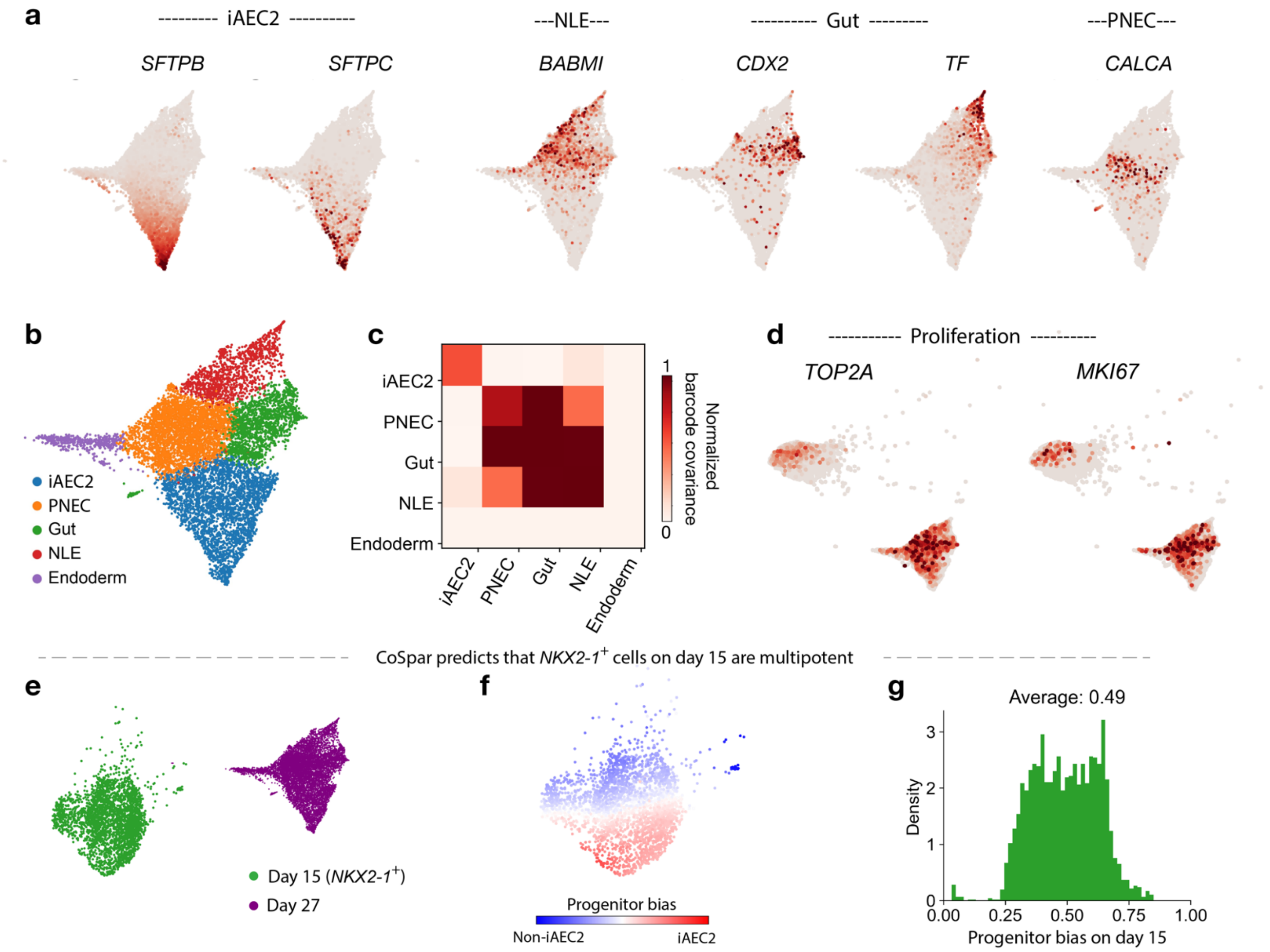
f, Application of Waddington-OT (WOT) to hematopoiesis dataset. WOT was applied to the same data in Ref[2], where clonal data was used to tune the local cell proliferation rates. When WOT is applied without access to any clonal information, performance is degraded as seen by comparing the plots here to the ground truth. Plots are to be compared with those in panels **c**,**d** and Fig. 4**c**. WOT is applied with default parameters (ϵ_{OT} =0.05).

g-i, Predicting early fate boundaries in the Gata1+ lineages using the original method from Ref[2]. **g**, Predicted progenitor bias among the Gata1+ cells on the state embedding. **h**, Comparison of the number of differentially expressed genes (FDR<0.05) identified from different methods of clonal analysis. **i**, Gene expression heat map for all differentially expressed genes identified with the Weinreb method[2].



Supplementary Fig. 4. Benchmarking CoSpar in fibroblast reprogramming.
a, Expression of selected marker genes on UMAP visualizations from day 15, 21 and 28.
b, Reproduction of results in Fig. 5**e** using a similarity matrix obtained from each sub-sampled dataset. Results are seen to be robust to sub-sampling strategies.
c-e, Transition maps inferred by CoSpar with access only to end-point clonal information are robust to the choice of initialization. These panels accompany Fig. 5**h**. **c**, Visualization of the progenitor bias derived from the initialized transition map and the corresponding CoSpar prediction, for different entropic regularizations and distance metrics as indicated. **d**, Parameter sweep quantifying the stability of the predicted progenitor bias. **e**, Progenitor bias prediction from Waddington-OT[3], which relies only on state information. Upper panel: the predicted progenitor bias on the state manifold at ϵ_{OT}=0.05. Lower panel: progenitor bias correlation with ground truth across different ϵ_{OT} values.
f-h, CoSpar analysis with clonal barcodes integrated at sequential time points. The analysis was done with clonal data on day 28. **f**, The cumulative barcoding scheme in the reprogramming experiment.

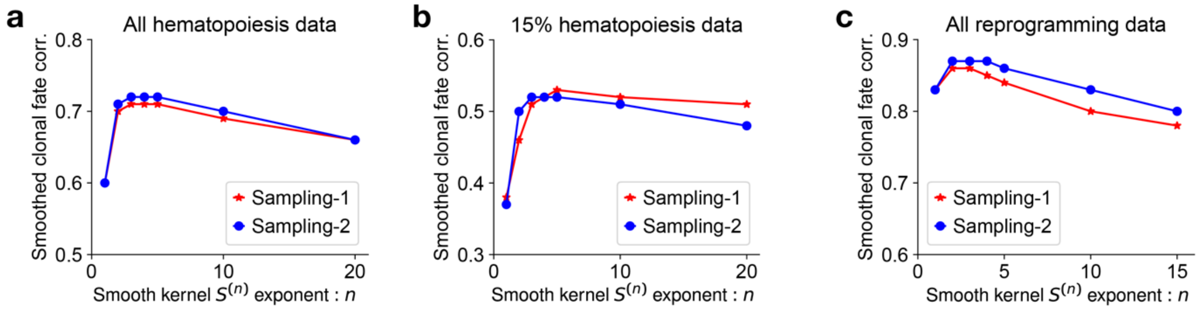
Cells were barcoded on day 0, 3, and 13. **g**, A progenitor bias prediction generated by concatenating all tags from all three time points into a single clonal barcode for each cell, thus ignoring the nested clonal structure in the data. **h**, Equivalent results of CoSpar analysis with nested clonal structure, carried out by treating Tag0, Tag3 and Tag13 as independent barcodes for a cell, such that each cell may have up to three barcodes. Left panel shows the histogram of barcode number per cell.



Supplementary Fig. 5. Marker gene expression and clonal structure during differentiation into alveolar cells and other endodermal cells.

a, Expression of genes associated (in Ref[4]) with iAEC2 cells, non-lung endoderm (NLE), gut endoderm, and pulmonary neuroendocrine cells (PNEC).

b, Leiden clustering of day-27 cell states. Cluster are named based on their corresponding gene expression.
c, Normalized barcode covariance on day 27 among all clusters, showing evidence of clonal partitioning of iAEC2 cells.
d, Expression of two representative genes marking proliferating cells (*TOP2A* and *MKI67*) on day 17 and 27 state manifold, showing that cells predicted by CoSpar to show low commitment on day 17 appear proliferating (Fig. 6**c**).
e-g, CoSpar predicts that lineage restriction occurs after day 15, except for a rare fraction of cells committed to non-iAEC2 fates. **e**, UMAP visualization of cell states on day 15 and 27. **f**, CoSpar-predicted progenitor bias among cells on day 15. **g**, Histogram of the progenitor bias on day 15 (shown in panel **f**). Unlike on day 17 (Fig. 6**c**), here progenitor bias is concentrated at 50%.



Supplementary Fig. 6. Establishing upper bounds for fate prediction after data loss. In this paper, performance of CoSpar was compared to previously published methods by discarding clonal data and then examining the fidelity of fate predictions in the face of data loss. Supporting the results reported in Figs. 4**g,i** and 5**h**, we obtain an upper bound for fate prediction, by randomly sampling 50% cells from the full ground-truth dataset in each case to predict the progenitor bias of remaining cells, with different smoothing exponents *n*. Prediction was carried out by first inferring the progenitor bias Q_i^{tr} from the training data (denoted by *tr*) to predict the bias Q_i^{tst} of the test data, by imputation via graph diffusion: $Q_i^{tst} = \sum_j {S_{ij}}^{(n)} Q_j^{tr}$. Results show that, in all the three cases considered, a smoothing exponent *n*=3 provided the best correlation between the imputed and actual values of Q_i^{tst}. These correlation values are indicated by the upper dashed grey lines in Figs. 4**g,i** and 5**h**.

CONTENTS

Supplementary Note 1: Connecting transition maps to models of differentiation 10
Supplementary Note 2: The effect of noisy measurement on transition map inference 11
Supplementary Note 3: Coherent sparse optimization 12
Supplementary Note 4: Transition map initialization with HighVar 13

References 14

Supplementary Note 1: Connecting transition maps to models of differentiation

This note grounds the finite-time transition map in a stochastic model of cell differentiation. In doing so it also clarifies what cannot be learnt from the transition map.

We begin by considering a Markov model of differentiation represented by an arbitrary graph of finite size, where each node represents a cell state. In this model, each cell probabilistically undergoes proliferation, death, and differentiation with rates that are specific to the cell state. A clone is a realization of such a stochastic branching process, seeded as a single barcoded cell in some cell state. Starting from a cell state i, k_{ij} is the differentiation rate to a different state j; b_i is the probability of a cell dividing into two cells; and d_i is the cell loss rate for cells in state i. We assume that these rates are first-order (independent of the number of cells in a state). These rates can vary with time to reflect changes in the tissue environment. Supplementary Fig. 1**a** shows a simplified example of such a model.

This model is useful in its simplicity, but it is clearly not general: being a Markov process, it assumes that we have a complete measurement of the variables that could affect state dynamics, such as the transcriptome, epigenome, and extracellular environment. This is unlikely to be true. Incomplete state measurement leads to a non-Markovian dynamics[5]. Nonetheless, our model may be a useful approximation as it generates predictions of biomarkers and fate regulators, and their correlation with fate bias.

Our goal in this paper is to learn the structure of such a graphical model (e.g. Supplementary Fig. 1**a**) and its rate constants, from LT-scSeq data. To learn a model from data, we focus most simply on the mean dynamics of cell number at each state. To do so, one could consider a complete stochastic description using the chemical master equation[6], which gives the distribution evolution over the extended state space $N \times X = \{(N_i, X_i) \ \ \forall\, i; \text{ and } N_i = 1, 2, ...\}$, where N_i is the number of cells at state i and X_i is the corresponding state. However, because we assume a first-order model, there exists a closed-form equation for the dynamics of average cell number $\bar{N}_i(t)$ at state i and time t,

$$\frac{d}{dt}\bar{N}_i(t) = \sum_j \bar{N}_j(t) K_{ji}, \tag{1}$$

where $K_{ij} \equiv (1-\delta_{ij})k_{ij} + \delta_{ij}(b_i - d_i - \sum_{k\neq i} k_{ik})$, with $\delta_{ij} = \{1 \text{ if } i = j; \text{ otherwise } 0\}$, is the instantaneous transition rate from state i to j that includes all cellular processes: division, cell death, and differentiation. This mean dynamics only captures the net effect of cell number change $(b_i - d_i)$, and does not distinguish whether it is from cell proliferation or loss.

To make contact with experiment, we represent the number of cells at each state as a fraction of the total cell number to obtain the cell density:

$$P_i(t) \equiv \frac{\bar{N}_i(t)}{\bar{N}(t)}, \tag{2}$$

where $\bar{N}(t) \equiv \sum_j \bar{N}_j(t)$ is the total cell number at time t. The dynamics of the cell density $P_i(t)$ is

$$\frac{d}{dt}P_i(t) = \sum_j P_j \tilde{K}_{ji}(t), \tag{3}$$

where $\tilde{K}_{ji}(t) \equiv K_{ji} - \delta_{ji}\bar{\alpha}(t)$, and $\bar{\alpha}(t) \equiv \sum_k P_k(t)(b_k - d_k)$ is the average growth rate of the population at time t. Diagonal elements in $\tilde{K}$ reflect whether net growth in each state is larger (positive) or smaller (negative) than the population average.

We now can ground the transition map T in terms of the model. Integrating Eq. (3) from time t_1 to t_2 leads to the relation

$$P_i(t_2) = \sum_j P_j(t_1) T_{ji}(t_1, t_2), \tag{4}$$

where the intrinsic finite-time transition map

$$T = \exp\Big(\int_{t_1}^{t_2} \tilde{K} dt\Big) \tag{5}$$

is obtained from matrix exponentiation of the corrected instantaneous transition rate matrix $\tilde{K}$.

The transition probability T_{ij} is the fraction of progenies from initial state i that ends at later state j (Supplementary Fig. 1**b**). To see this, we can sum over all states in Eq. (4), and noting that $\sum_i P_i(t) = 1$, we have $1 = \sum_j P_j(t_1) \sum_i T_{ji}$. This equation is valid for any distribution $P_j(t_1)$ and therefore the transition map satisfies the conservation property

$$\sum_j T_{ij} = 1. \tag{6}$$

Owing to its normalization (Eq. 6), the transition map that is experimentally accessible captures the most interesting property we want: the probability of a cell to differentiate into different cell types. A certain initial state i can transition to multiple states over time window t, i.e., T has multiple non-zero entries associated with the i-th row.

Nonetheless, it is important to note that T_{ij} is shaped both by differences in transition rates between states, and by the collective effect of proliferation and cell death along the trajectories between state i and j. Mathematically, although proliferation and cell death only affect the diagonal terms in the instantaneous transition matrix $\tilde{K}$, the matrix exponentiation in Eq. (5) will propagate this effect to the off-diagonal terms in the finite-time transition matrix T. For this reason, empirical transition maps alone obscure differences between biases in proliferation and choice towards competing fates, as illustrated in Supplementary Fig. 1**d**.

Supplementary Note 2: The effect of noisy measurement on transition map inference

In Eq. (5), the transition map is seen to emerge from stochastic state transitions accumulating over time. In practice, an inferred map is also shaped by sources of noise associated with measurement and subsequent dimensionality reduction of the data. In this note, we examine the errors propagated from different technical sources into the observed transition map T. As might be expected, we show that technical sources of error lead to a 'blurred' transition map, delocalized over the cell state graph. The smoothing kernels connecting the true and observed transition map can be understood as a matrix product of error kernels associated with each individual source of uncertainty.

a. Measurement errors. We will consider the errors associated with correctly assigning transition rates from a state i at time t_1 to state j at time t_2. Such a transition contributes to mass at matrix element $T_{ij}(t_1, t_2)$ of the transition map. At time t_2, errors in measurement re-assign cells from state j to another state k, with a probability ϵ_{jk} normalized such that $\sum_k \epsilon_{jk} = 1$. With such an error, the observed transition map now becomes $T_{ij}^{(\text{obs.})} = \sum_k T_{ik}\epsilon_{kj}$. A similar error may occur at t_1. Because technical errors may differ between time points, we will denote $\epsilon^{(i)}$ as the error in measuring the state of a cell at time t_i. Accounting for errors in two time points, the observed transition map now becomes:

$$T_{ij}^{(\text{obs.})} = \sum_{k,l} \epsilon_{ki}^{(1)} T_{kl} \epsilon_{lj}^{(2)}.$$

b. Clonal dispersion. In LT-scSeq experiments, the cells sampled at t_1 are clonally related to those that give rise to cells sampled at t_2. But being distinct, they may occupy different states. As above, we consider the error in estimating transition rates from state i at t_1 to state j at t_2. At t_1, a clonally-related state, k, is observed instead of state i, with a probability that we shall denote σ_{ik}. This probability satisfies normalization $\sum_k \sigma_{ik} = 1$. Accounting for this clonal dispersion, the observed transition map relates to the true transition map through the relation:

$$T_{ij}^{(\text{obs.})} = \sum_k \sigma_{ki} T_{kj}.$$

Note that because cells divide, more than one cell may be observed in a clone at time t_1. In this case, the error kernel σ_{ki} no longer has a unique definition because choices in constructing the transition map may assign more or less weight to particular cells within each clone. By enforcing local coherence, CoSpar strongly weights σ_{ki} towards states k that are close to i, thus reducing errors in the transition map as compared to using a 'naive' clonal analysis method such as we have previously reported[2], which weights all cells in a clone at t_1 equally.

Compounding clonal dispersion and measurement error, we recognize the the observed transition map has the form:

$$T^{(\text{obs.})}(t_1, t_2) = S_1^T T(t_1, t_2) S_2,$$

where $S_1 = \epsilon^{(1)}\sigma$ and $S_2 = \epsilon^{(2)}$.

Supplementary Note 3: Coherent sparse optimization

Our goal in dynamic inference is to learn the finite-time transition map, as defined in Eq. (4), for the set of observed cell states in a given experiment. After imposing sparsity and coherence constraints (see main text), we obtain the cost function,

$$\min_T ||T||_1 + \alpha||LT||_2, \text{ s.t. } \sum_m ||P(t_2; m) - P(t_1; m)T(t_1, t_2)||_2 \le \epsilon;\ T \ge 0;\ \text{Normalization.} \tag{7}$$

Here, $P(t_{1,2}; m)$ is a row-vector representing the distributions of cell states within the m-th clone. $L_{ij} = 1 - \bar{w}_{ij}/\sum_j \bar{w}_{ij}$ is the normalized graph laplacian, with w_{ij} the graph connectivity of the nearest neighbor kNN graph of cell states. Defining $\mathbf{P}(t)$ as a clone-by-cell matrix resulting from concatenation of individual clonal distribution: $\{P(t; m), m = 0, 1, 2...\}$, we note that $\sum_m ||P(t_2; m) - P(t_1; m)T(t_1, t_2)||_2 = ||\mathbf{P}(t_2) - \mathbf{P}(t_1)T(t_1, t_2)||_2$, which gives the form of the cost function given in the main text. For joint optimization, the cost function is additionally minimized over $\mathbf{P}(t_1)$, i.e. $\min_{\mathbf{P}(t_1)}[\cdots]$.

Before continuing, we note the relationship of this optimization problem to past literature. Absent the coherence constraint ($\alpha = 0$), this optimization problem reduces to sparse optimization by lasso regression. To our knowledge, only one study has explored the extension of lasso to enforce coherence with relation to a data embedding, called 'fused lasso' optimization[7]. Fused lasso is however different in three important ways from Eq. (7). First, it suppresses the first-order derivative of the inference target to promote coherence. Second, fused lasso was developed for 1-d or 2-d datasets, assuming a natural ordering for the observed cell states. Third, like lasso, the inference object of fused lasso is a vector. In contrast, the coherent sparse optimization in Eq. (7) is generalized to arbitrary graphs; it suppresses the second-order derivative (the curvature) to enforce coherence; and it is generalized to matrix inference.

Our goal is now to ground the optimization problem in LT-scSeq data, and to propose an algorithm that approximates solution of Eq. (7). To make connection with raw clonal data, we approximate the density profile matrices $\mathbf{P}(t)$ as,

$$\mathbf{P}(t) = I(t)S(t), \tag{8}$$

where $I(t)$ is a clone-by-cell matrix observed at time t, and $S(t)$ is a cell-cell similarity matrix at time t. Note that Eq. (8) integrates the state information (encoded in $S(t)$) and clonal information (encoded in $I(t)$) into $\mathbf{P}$. This local smoothing operation indirectly imposes coherent transitions in this system.

We now discuss implementation of the optimization problem. Eq. (7) might be formulated as a quadratic programming problem, and be solved accordingly as in fussed lasso[7]. However, this strategy is very expensive computationally[7]. There could be ways to solve the optimization efficiently and exactly, and we leave it as an open problem. Instead, we provide an efficient yet heuristic way to solve the optimization. Specifically, we break down individual elements of the objective function, and propose a simple alternative for each of them.

1. *Sparsification.* Instead of including the sparsity term $||T||_1$ into the objective function, we directly apply a pre-defined thresholding to the transition map at each iteration: $T \leftarrow \theta(T, \nu)$, where

$$[\theta(T, \nu)]_{ij} = \begin{cases} T_{ij}, & \text{if } T_{ij} \ge \nu \max_j T_{ij} \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

2. *Transitions within clones.* To enforce Eq. (4) for each observed clone, we consider a clonal transition map π^m for the m-th clone, which allows only intra-clone transitions and conserves the total transition flux within a clone. We do so by projecting the transition map T and performing clone-wise normalization: $\pi^m \leftarrow \mathcal{P}_m(T)$:

$$[\mathcal{P}_m(T)]_{ij} = \frac{\tilde{\pi}^m_{ij}}{\sum_{i'j'} \tilde{\pi}^m_{i'j'}}, \tag{10}$$

where $\tilde{\pi}^m_{ij} = T_{ij}$ if the transition $i \to j$ occurs within clone m, and otherwise $\tilde{\pi}^m_{ij} = 0$. The composite map capturing all intra-clone transitions is then,

$$\mathcal{P}(T) = \sum_m \mathcal{P}_m(T) \tag{11}$$

A map constructed in this way, $\pi = \mathcal{P}(T)$, will satisfy the following equation approximately:

$$I(t_2) \approx I(t_1)\pi(t_1; t_2), \tag{12}$$

which is the clonal constraint for directly observed cell states[8]. The map $\pi(t_1; t_2)$ can be used to specify T, but being constrained to clones it is no longer coherent.

3. *Coherence.* To enforce coherence, we begin by noting that Eqs. (4), (8) and (12) together lead to the relationship $T(t_1; t_2) = S^{-1}_{t_1} \pi(t_1; t_2) S_{t_2}$. As similarity matrices S are generally non-invertable, we introduce a pseudo-inverse,

$$T(t_1; t_2) \approx S^{+}_{t_1} \pi(t_1; t_2) S_{t_2}. \tag{13}$$

Eq. (13) smoothes the transition map learnt within-clones, π, over nearby states to get a transition map T across all states. T is now a locally continuous map and satisfies the coherence constraint: similar initial cell states have similar fate outcomes.

This approach to calculating T leads to minimization of the term $\alpha||LT||_2$ in Eq. (7), although the parameter α establishing the relative weight of coherence is no longer explicitly identifiable in the procedure. It is instead reflected in the extent of smoothing.

These three steps, carried out sequentially and iteratively, define the CoSpar algorithm given in methods. Note that normalization is performed clone-wise in Eq. (11). The non-negativity constraint, $T \geq 0$, is implicitly satisfied in the above steps. In our strategy, Eq. (13) is the most time-consuming step as it involves multiplication of large matrices. CoSpar is nonetheless efficient as it carries out matrix multiplication *only* at Eq. (13), and we find that it converges within a few iterations (Supplementary Fig. **2d**).

Supplementary Note 4: Transition map initialization with HighVar

The HighVar method provides an approach to initialize the joint optimization of T and $I(t_1)$ (see Methods). The approach is loosely motivated by the expectation that cells similar in gene expression between time points may share clonal origin. This expectation can be violated; we use it only to initialize numerical optimization.

HighVar consists of three steps: 1) Select highly variable genes that are expressed at both t_1 and t_2; 2) For each highly variable gene (indexed by m), threshold its expression to form a binary expression matrix $\hat{x}_{im} \in \{0, 1\}$ for all states observed at t_1 and t_2 to generate pseudo clonal data $\hat{I}(t_1)$ and $\hat{I}(t_2)$ from the binary expression matrix; 3) Run CoSpar with $\hat{I}(t_1)$ and $\hat{I}(t_2)$. The pseudo-clonal data $\hat{I}(t_1)$ and $\hat{I}(t_2)$ are discarded, and the resulting map T is used to initialize CoSpar with the true clonal data.

For the first step, we use the SPRING gene filtering function filter_genes with an adjustable gene variability percentile parameter HighVar_gene_pctl to select highly variable genes[9]. For the second step we discretize the gene expression of each highly-variable gene, sequentially, with a gene-specific threshold η_m:

$$\hat{I}_{im} = H\Big(x_i(m) - \eta_m\Big) \times Z_{im},$$

where $H(\cdot)$ is the Heaviside step function ($H(x) = 1$ if $x > 0$; otherwise 0), $Z_{im} = [1 - H(\sum_{m^*=0}^{m-1} \hat{I}_{im^*})]$ sums over previously considered genes to ensure that the same cell is not assigned to more than one pseudo-clone. The gene-specific threshold η_m is chosen such that every pseudo clone has the same number of cells at each time point N_t/M, where N_t is the number of observed cells at time t and M is the total number of highly variable genes (i.e., pseudo clones). In case N_t/M is not an integer, we use its ceil, i.e., $\lceil N_t/M \rceil$, and stop the clonal matrix update when all cells are clonally labeled.

[1] C. Weinreb and A. M. Klein, Proceedings of the National Academy of Sciences **117**, 17041 (2020).

[2] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein, Science **367** (2020).

[3] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, *et al.*, Cell **176**, 928 (2019).

[4] K. Hurley, J. Ding, C. Villacorta-Martin, M. J. Herriges, A. Jacob, M. Vedaie, K. D. Alysandratos, Y. L. Sun, C. Lin, R. B. Werder, *et al.*, Cell Stem Cell **26**, 593 (2020).

[5] S.-W. Wang, K. Kawaguchi, S.-i. Sasa, and L.-H. Tang, Phys. Rev. Lett. **117**, 070601 (2016).

[6] D. T. Gillespie, The journal of physical chemistry **81**, 2340 (1977).

[7] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**, 91 (2005).

[8] One can appreciate that this equation is approximately satisfied because $I(t_1)\pi(t_1;t_2)$ gives a matrix with non-zero values on at clonally observed states at t_2. Therefore $I(t_1)\pi(t_1;t_2)$ has the same sparse structure as $I(t_2)$ but will differ in the exact non-zero values because $I(t_2)$ is strictly binary.

[9] C. Weinreb, S. Wolock, and A. M. Klein, Bioinformatics **34**, 1246 (2018).