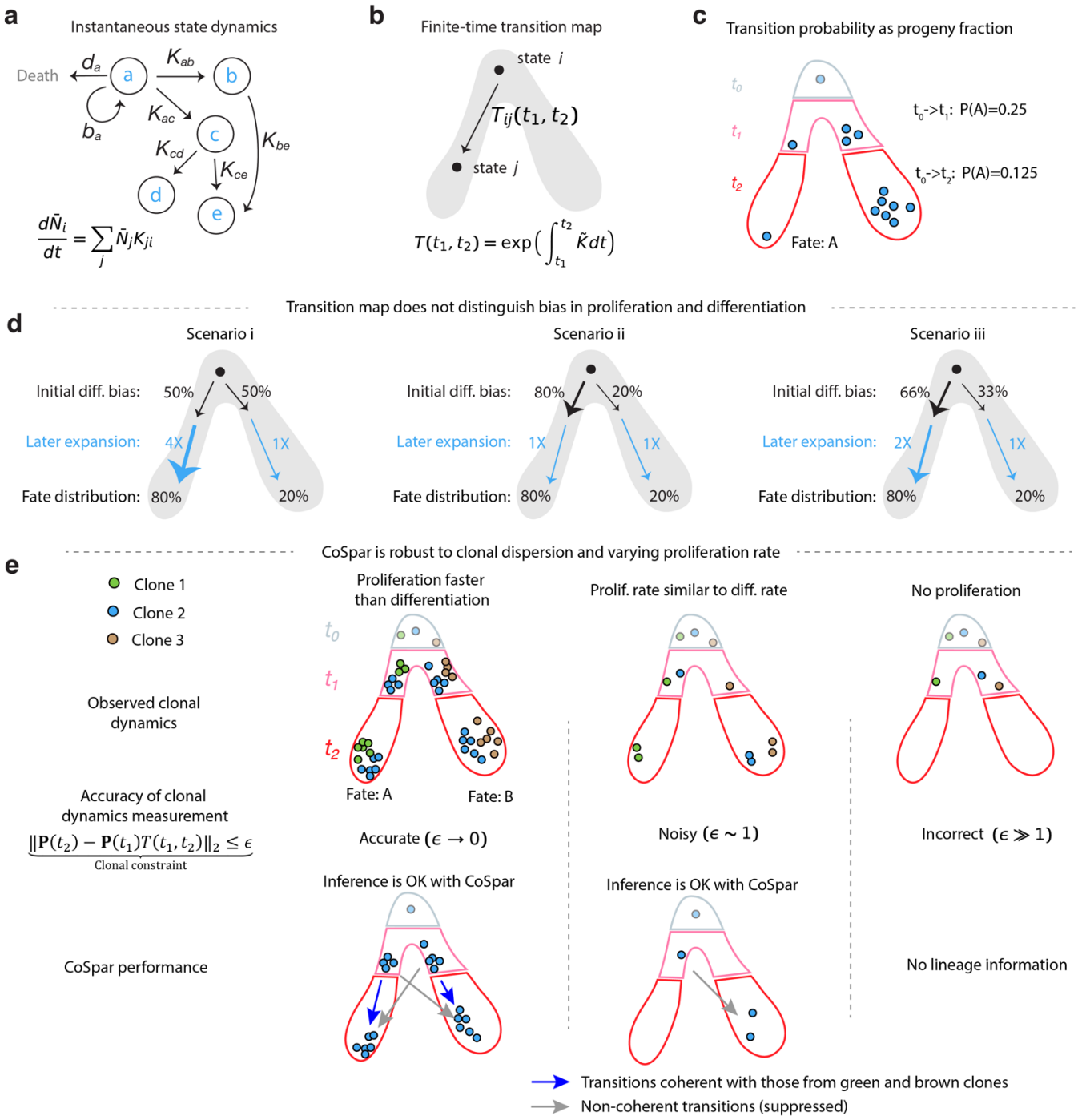
Supplementary Information for

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

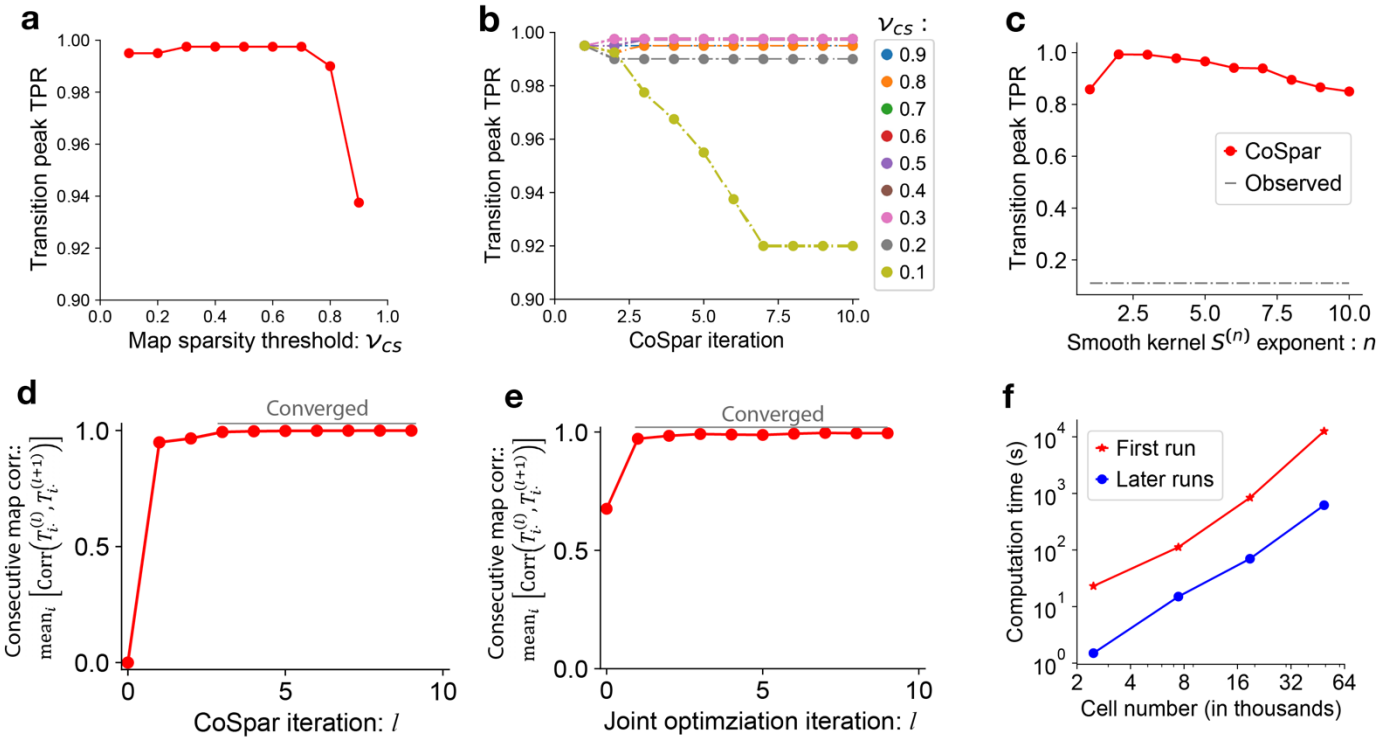
Shou-Wen Wang* and Allon M. Klein*

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

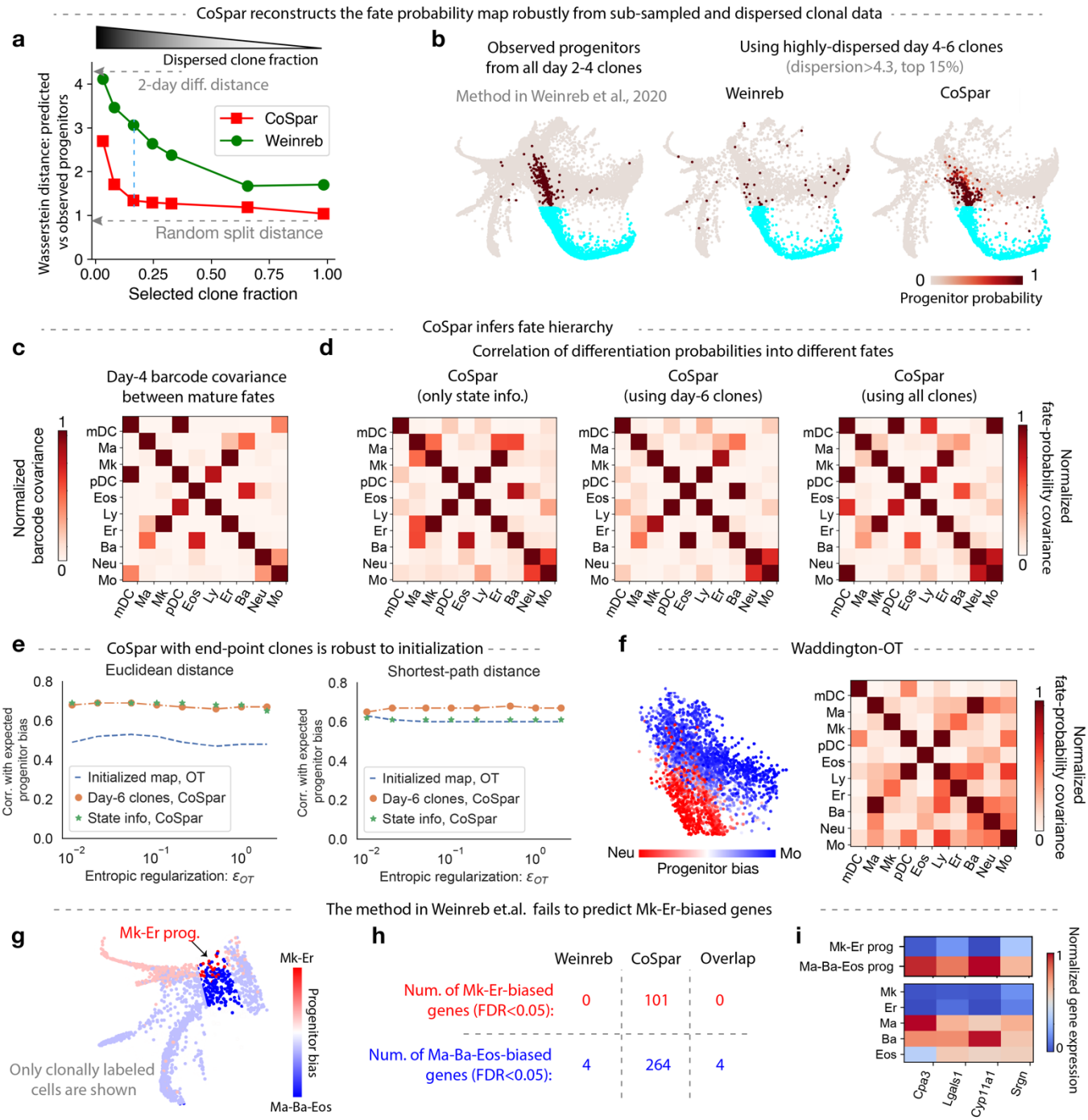


Supplementary Fig. 1. Models, assumptions and limitations of Coherent Sparse Optimization. **a**, Simple example of the class of stochastic models that CoSpar seeks to learn. In such models, each node represents an observed cell state. In practice, thousands of measured states are included; here only 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.

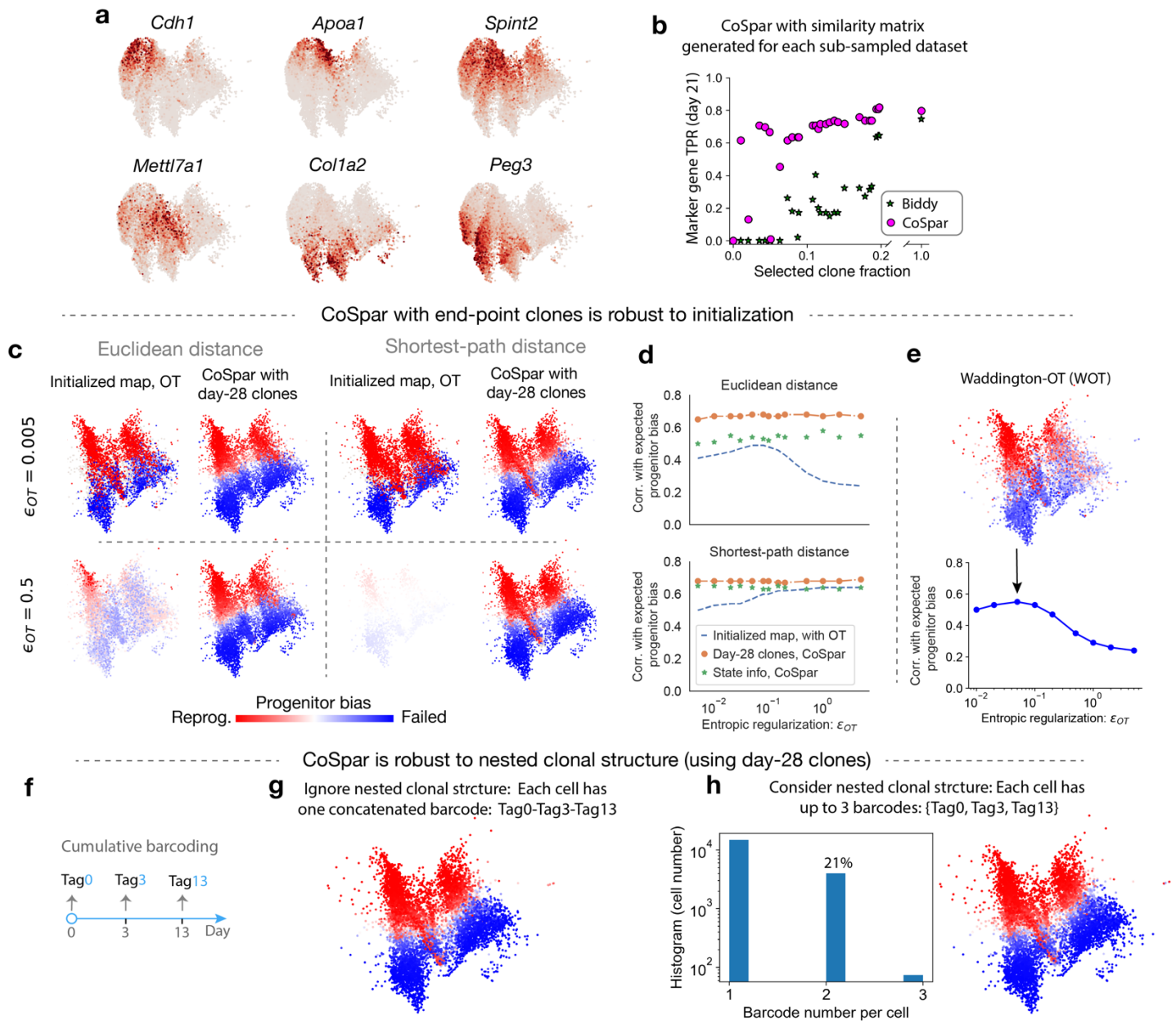


Supplementary Fig. 2. Evaluating CoSpar performance across parameter sweeps.
a-c, Performance of CoSpar using simulations as in Fig. 3**a-d** with a range of algorithm parameters (see Methods for parameter definitions): (**a**) sparsity threshold $\nu_{cs} \in [0,1]$; (**b**) number of iterations, showing 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).



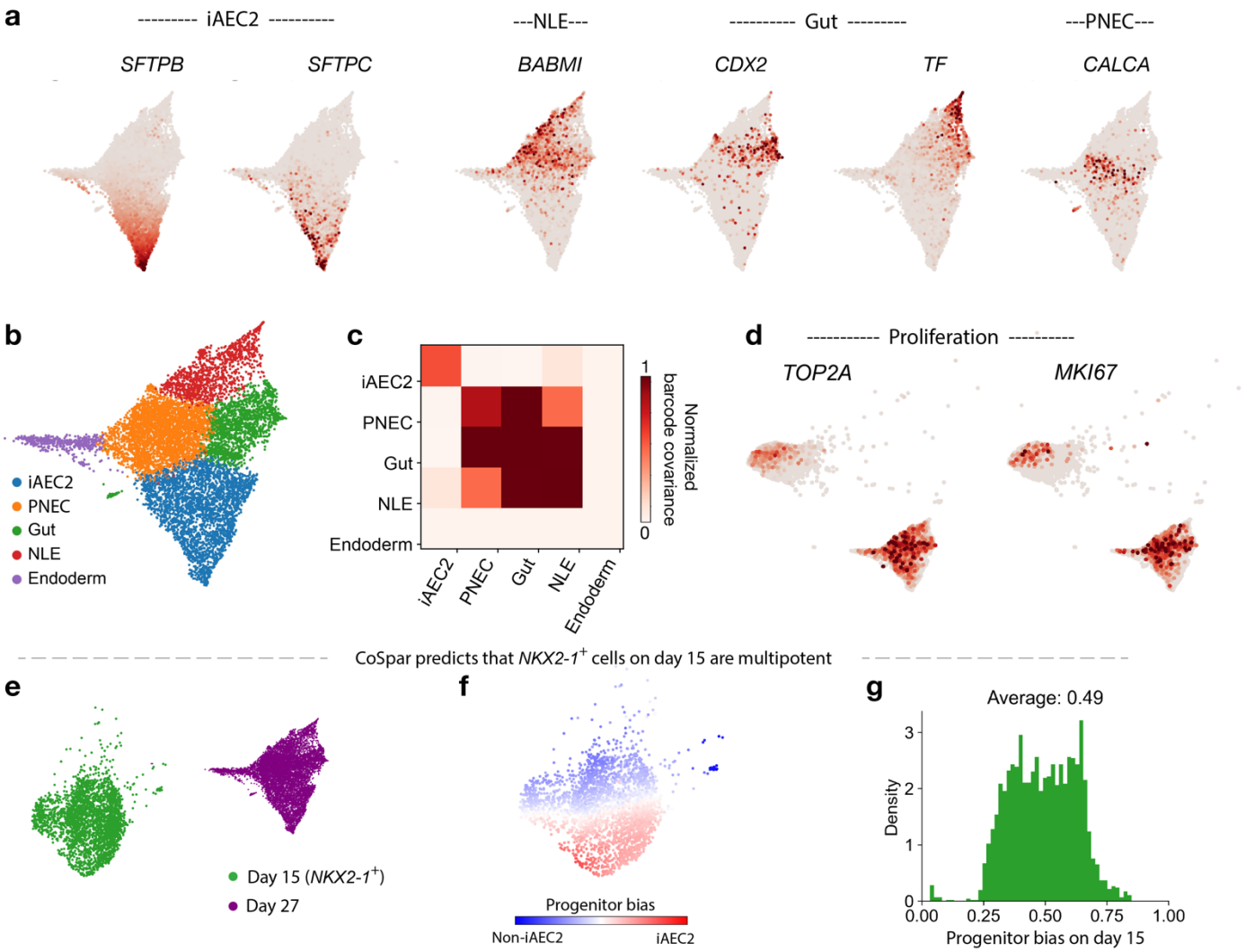
Supplementary Fig. 3. Benchmarking CoSpar in hematopoiesis.
a, CoSpar reconstructs transition maps from sub-sampled and dispersed clonal data. Here, we evaluate the prediction error as the Wasserstein distance between fraction of cell progeny predicted to occupy a given fate, compared to that obtained 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].



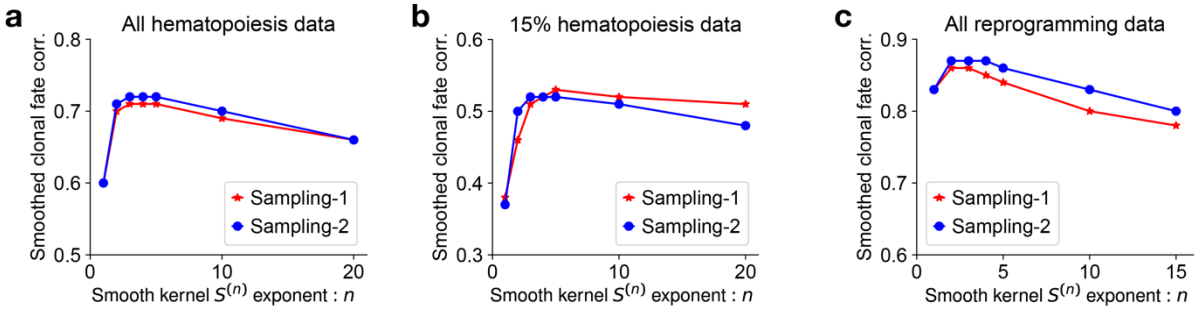
Supplementary Fig. 4. Benchmarking CoSpar in fibroblast reprogramming.
a, Expression of selected marker genes on UMAP visualizations from day 15, 21 and 28.
b, Reproduction of results in Fig. 5**e** using a similarity matrix obtained from each sub-sampled dataset. Results are seen to be robust to sub-sampling 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

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 $T^{(\text{obs.})}_{ij} = \sum_k T_{ik}\epsilon_{kj}$. 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:

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

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:

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

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,

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

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

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

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. 2**d**).

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).