ARTICLES
https://doi.org/10.1038/s41587-022-01209-1

nature biotechnology

CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information

Shou-Wen Wang[1]✉, Michael J. Herriges[2,3], Kilian Hurley[4,5], Darrell N. Kotton[2,3] and Allon M. Klein[1]✉

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. In this study, we developed coherent, sparse optimization (CoSpar) as a robust computational approach to infer cell dynamics from single-cell transcriptomics integrated with lineage tracing. Built on assumptions of coherence and sparsity of transition maps, CoSpar is robust to severe downsampling and dispersion of lineage data, which enables simpler experimental designs and requires less calibration. In datasets representing hematopoiesis, reprogramming and directed differentiation, CoSpar identifies early 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/.

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 toward 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 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. 1a).

Recently, several efforts from us and others have integrated lineage tracing with single-cell RNA sequencing (hereafter LT-scSeq) using unique, heritable and expressed DNA barcodes[2,12,14–19]. 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[12,15], find novel markers for stem cells[18] and reveal pathways that control cell fate choice[15,18]. The simplest of these methods labels cells at one time point[12] (Fig. 1b); more complex methods allow the accumulation of barcodes over successive cell divisions to reveal the substructure of clones[2,12,14–20] (Fig. 1c).

Emerging LT-scSeq methods have been successful at revealing regulators of cell fate[15,18] and the fate potential of early progenitors[12,15], but they also present challenges that might limit their utility in practice. At least five technical and biological challenges affect experimental design and interpretation (Fig. 1f): stochastic differentiation and variable expansion of clones[21] (Fig. 1f(i)); cell loss during analysis (Fig. 1f(ii)); barcode homoplasy wherein cells acquire the same barcode despite not having a lineage relationship[2] (Fig. 1f(iii)); access to clones at only a single time point[22,23] (Fig. 1f(iv)); and errors in determining the state of clonal progenitors due to a lag time between labeling cells and the first sampling ('clonal dispersion') (Fig. 1f(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.

In this study, we advanced on recent efforts[24,25] to develop robust, computationally efficient and generalizable approaches 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. 1a). 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. 1d and Supplementary Fig. 1b,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 $T_{ij}(t_1, t_2)$[9]. 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 occupies state j at time t_2 (Fig. 1d and Supplementary Fig. 1c). This transition map averages the effects of cell division, loss and differentiation (Supplementary Fig. 1d), but it nonetheless proves useful for several applications[9] (Fig. 1d).

We make two assumptions about the nature of biological dynamics to constrain inference of the transition map. We assume the

[1]Department of Systems Biology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [2]Center for Regenerative Medicine of Boston University and Boston Medical Center, Boston, MA, USA. [3]The Pulmonary Center and Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [4]Department of Medicine, Royal College of Surgeons in Ireland, Education and Research Centre, Beaumont Hospital, Dublin, Ireland. [5]Tissue Engineering Research Group, Royal College of Surgeons in Ireland, Dublin, Ireland. ✉e-mail: shouwen_wang@hms.harvard.edu; allon_klein@hms.harvard.edu

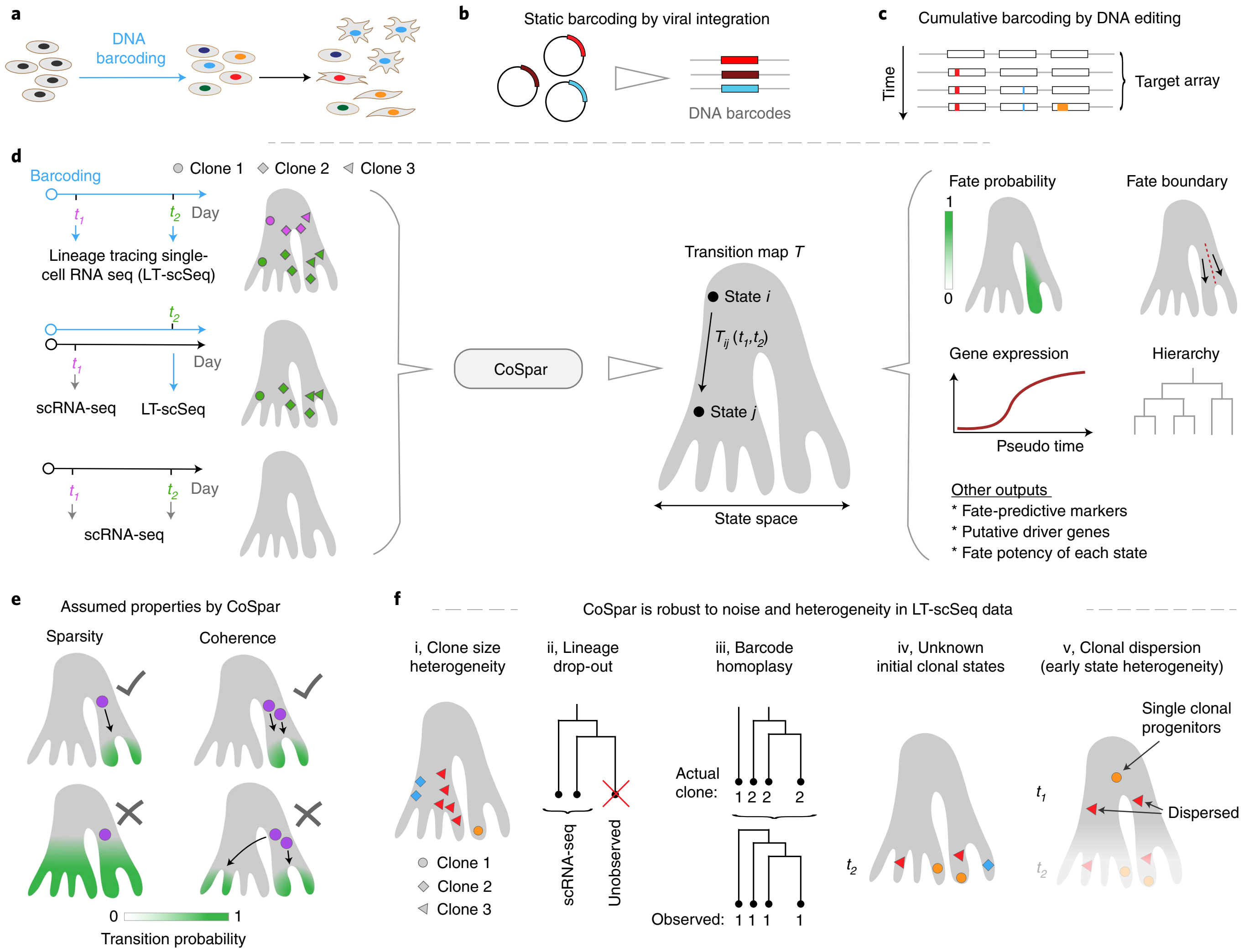

Fig. 1 | Integrative analysis of lineage tracing and transcriptome data. a, LT-scSeq experiments simultaneously measure cell phenotypes and clonal lineages (indicated by colors). **b**, **c**, LT-scSeq assays encode lineage information with static DNA barcodes or cumulative barcoding. **d**, CoSpar unifies 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 transcriptomic states. **e**, Two key assumptions constrain dynamic inference by CoSpar. **f**, Stereotypical challenges in clonal analysis: (i) single labeled cells can give rise to clones with a wide dispersion in size; (ii) LT-scSeq loses cells during analysis due to inefficient cell capture or loss of barcode information, leading to loss of clonal structure; (iii) barcode homoplasy occurs when cells from different clones present the same barcode due to experimental limitations; (iv) clonal progenitors are unobserved when clones are seen only upon tissue dissociation; (v) early-time clonal dispersion introduces errors in identifying the ancestor state of each clone (Supplementary Fig. 2).

map to be a sparse matrix, because most cells can access just a few states during an experiment (Fig. 1e, left). Additionally, we assume the map to be locally coherent, meaning that cells in similar states should share similar fate outcomes (Fig. 1e, right). These constraints together force transition maps to be parsimonious and smooth, which we reasoned would make them robust to practical sources of noise in LT-scSeq experiments (Supplementary Fig. 1e).

We formalize transition map inference starting from a model of stochastic division and differentiation dynamics in Supplementary Notes 1–4, leading to the optimization problem formalized in Fig. 2a. The resulting algorithm is rooted in past work as follows: without assuming coherence ($\alpha = 0$ in Fig. 2a), the minimization takes the form of Lasso[26], an algorithm for compressed sensing. Building on this, fused Lasso[27] imposes uniformity between vectors jointly subject to Lasso regression. CoSpar now extends the idea of joint optimization from vectors to graphs, and it enforces coherence (minimizing a second derivative of T) rather than uniformity (minimizing a first derivative). An iterative, heuristic approach approximates the CoSpar optimization efficiently (Fig. 2b), without explicitly defining α. As inputs, CoSpar requires a barcode-by-cell matrix $I(t)$ that encodes the clonal information at time t and a data matrix for observed cell states (for example, from scRNA-seq). Clonal data might have nested structure reflecting subclonal labeling (Supplementary Note 4). CoSpar usage is schematized in Supplementary Figs. 3 and 4 and detailed in https://cospar.readthedocs.io/.

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 a later time point t_2. For these cases, inference is not fully constrained, as one must learn both the transition map T and the initial clonal data $I(t_1)$ (Fig. 2c and Methods). We approximate a solution additionally constrained by a minimum global

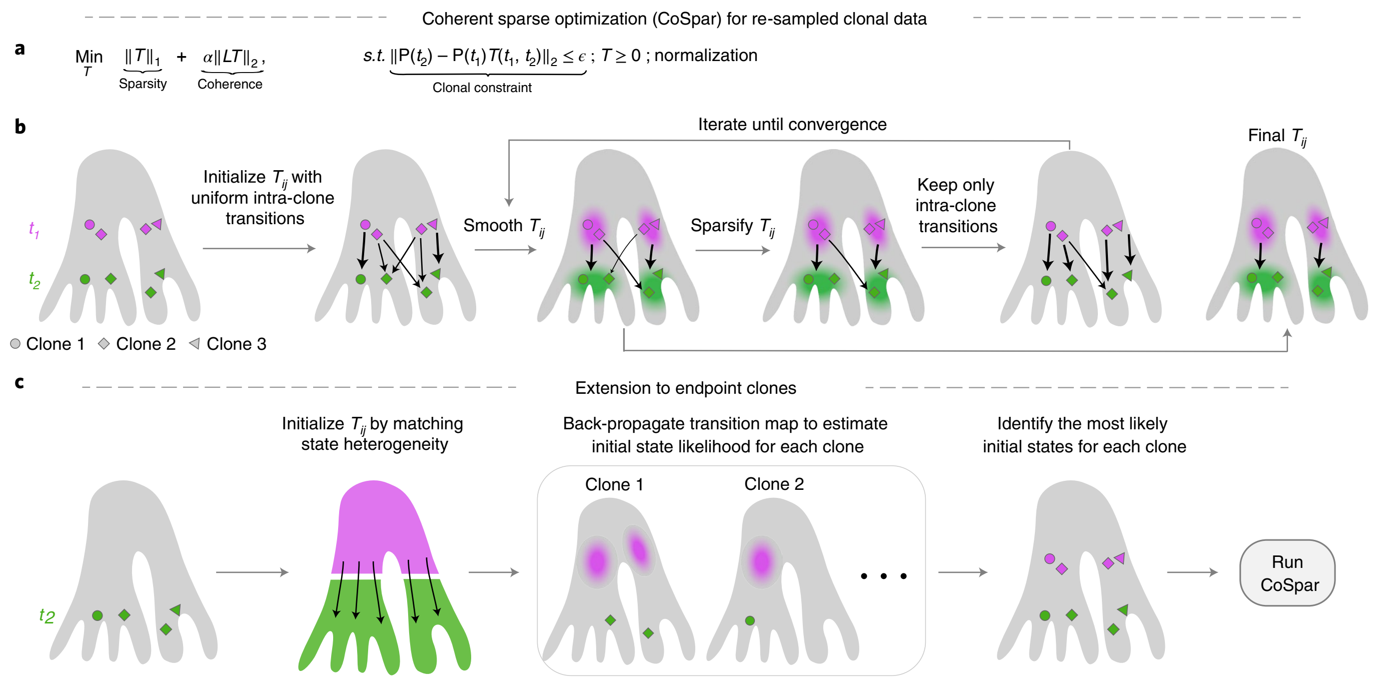
Coherent sparse optimization (CoSpar) for re-sampled clonal data
a
$\min_T \underbrace{\|T\|_1}_{\text{Sparsity}} + \underbrace{\alpha\|LT\|_2}_{\text{Coherence}}$, s.t. $\underbrace{\|P(t_2) - P(t_1)T(t_1, t_2)\|_2 \leq \epsilon}_{\text{Clonal constraint}}$; $T \geq 0$; normalization
b
Iterate until 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
c
Extension to endpoint 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
t2
Run CoSpar

Fig. 2 | The CoSpar algorithm. a, **b**, Transition map inference using resampled clones. **a**, Transition map inference formulated as an optimization problem. $||T||_1$ (L1 norm) quantifies the sparsity of the transition map matrix T; $||LT||_2$ (L2 norm) quantifies the local coherence of T with L being the graph Laplacian of the cell state similarity graph. A constraint enforces the observed clonal dynamics, with $\mathbf{P}(t) \in \Re^{M \times N_t}$ being the kernel-smoothed density profiles of M observed clones over N_t observed transcriptomes at time t (see Eqs. 7 and 8 in Supplementary Note 3 for complete definitions). **b**, Heuristic implementation of the CoSpar optimization. A transition map is inferred by iteratively enforcing observed clonal transitions, coherence and sparsity until convergence is achieved. This approximate solution does not explicitly specify the parameters α, ϵ (see details and derivation in Methods and Supplemental Note 3). **c**, When clones are observed only once, we infer their progenitor fate bias and identity by first initializing a transition map without clonal information and then (1) back-propagating the map to predict clonal progenitor identity and (2) learning the transition map.

transport cost[9,25] (Supplementary Note 5). We later show that this approach is robust to initialization for the datasets analyzed (Figs. 4 and 5 and Supplementary Figs 7e and 10d). Finally, coherence and sparsity provide constraints to the simpler problem of predicting dynamics from state heterogeneity alone without lineage data[9]. We extended CoSpar to this case. Thus, CoSpar is flexible to different experimental designs, as summarized in Fig. 1d.

In this and subsequent sections, we show that CoSpar recapitulates dynamics and is robust to the challenges typical of LT-scSeq (Fig. 1f) and to run parameters. We first show robustness to barcode homoplasy and early-time clonal dispersion using computer simulations. We modeled cells progressing through a sequence of gene expression states either toward a single fate (Fig. 3a) or bifurcating into two fates (Fig. 3e), with clones sampled in a manner representative of LT-scSeq experiments[12,15]. With 1,000 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 (true positive rate (TPR) > 98%; Fig. 3d), and the distribution of progeny fates showed 85% Pearson correlation to ground truth (Fig. 3i). Inferences remained similarly accurate with as few as 30 barcodes (Fig. 3d) and across a wide range of parameter values of the CoSpar algorithm (Supplementary Fig. 5d–f). CoSpar was robust to barcode homoplasy and only detectably lost accuracy when all lineage barcodes mixed more than ten clones on average (Fig. 3a–d). This degree of homoplasy is far higher than expected in most experiments. Furthermore, CoSpar was robust to early time clonal dispersion, simulated by sampling clones at increasing times after barcoding (Fig. 3f–i). Conversely, approaches used in previous work, which average the transitions between cells observed in each clone at different time points[12], are severely affected by both lag time and barcode homoplasy (Fig. 3d,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 after barcoding (Fig. 4a,b)[12]. 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 toward either the monocyte or neutrophil fate, using the method from Weinreb et al.[12] (Fig. 4c).

As a baseline for comparison, we applied CoSpar to predict HPC fate bias using state information alone (Fig. 4e). 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. 4e,g and Supplementary Fig. 6a). This performance reflects improvements from enforcing coherence and sparsity ($R = 0.51$–0.54 before CoSpar; Fig. 4d and Supplementary Fig. 7f), robust across a wide range of algorithm parameters (Supplementary Fig. 5a,b). However, the prediction based on state information alone is limited because it is sensitive to the choice of distance metric used in analysis (Fig. 4g and Supplementary Fig. 7e).

Clonal information eliminated the sensitivity to distance metric. To show this, we applied CoSpar to data restricted in time or in depth or depleted of lineage-restricted clones. Using even a single time point of clonal data (day 6), CoSpar recovered early fate bias (Fig. 4f; $R = 0.68$), and it did so robustly over a range of parameters and choices of distance metrics (Fig. 4g and Supplementary Fig. 7e). Furthermore, it recovered the differentiation hierarchy seen in the

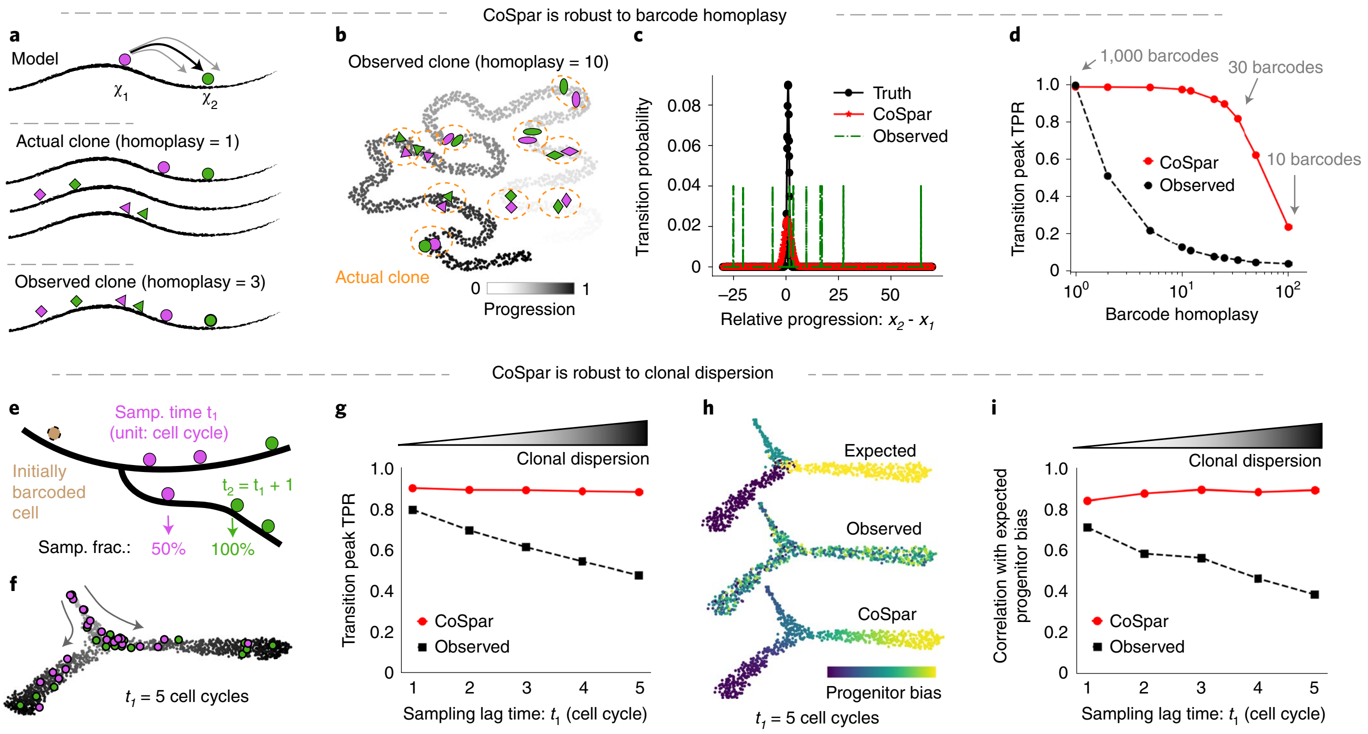

Fig. 3 | Proof of concept with simulated data. a–d, Benchmarking transition map inference with barcode homoplasy errors. **a**, Schematics of a simplified simulated LT-scSeq experiment to evaluate the accuracy of CoSpar and its robustness to barcode homoplasy errors. Homoplasy is simulated by assigning multiple clones with the same barcode. **b**, UMAP embedding of simulated data. Cells labeled with one barcode are shown, with moderate homoplasy (ten clones per 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 five 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.

correlation of clonal barcodes across all cell types (Supplementary Fig. 7c,d). When using a subsampled dataset from the top 15% most dispersed clones as ranked by day 4 intra-clone distance (Fig. 4b), CoSpar performed similarly well and outperformed the method from Weinreb et al., which was used to analyze these data originally[12] (Fig. 4h,i and Supplementary Fig. 7a,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. Benchmarking against two recent methods (LineageOT and SuperOT)[24,25] revealed that CoSpar predictions were more accurate and remained so when training on sparse data or using dispersed clones (Supplementary Figs. 8 and 9).

These benchmarks suggest that CoSpar should be able to resolve early fate biases in transcriptomic space and predict new fate regulators of known fate biases. 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. 4a,k). In culture, Mk and Er arise from a common progenitor (MEP), and Ba, Eos and Ma are produced by a different progenitor (BEMP)[28,29]. Although molecular signatures of fluorescence-activated cell sorted MEP have been explored recently[30], less is known about the transcriptomic identity of BEMPs. Applying CoSpar, we predict an early fate decision boundary between MEP and BEMPs (Fig. 4j,k), which correlates with the early expression of genes later associated with the resulting cell types (*Slc14a1* for Mk[30] and *Thy1* for Ba[31]; Fig. 4l) and with the transcription factor *Cebpa* that regulates Eos and Ba differentiation[29]. We identified 377 known and novel putative fate-associated genes (Fig. 4m and Supplementary Table 1). By contrast, the original method used to analyze these data finds very few genes distinguishing BEMPs and MEPs (Supplementary Fig. 7g–i). Differences between the putative BEMPs and MEPs are not fully identifiable without clonal information: using only state information, CoSpar and WaddingtonOT[9] recover only 25–60% of the fate-associated genes (Supplementary Fig. 7j,k). 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 that we analyzed tracked cells during the reprogramming of fibroblast cells over 28 d into endodermal progenitors (Fig. 5a)[15]. In this experiment, cells initially expand, with many cells lost due to senescence or cell death. After 28 d, 30% of resulting cells were successfully reprogrammed. Clonal analysis with cumulative barcoding was used to identify these cells early and predicted features that regulate their fate (Fig. 5b,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.

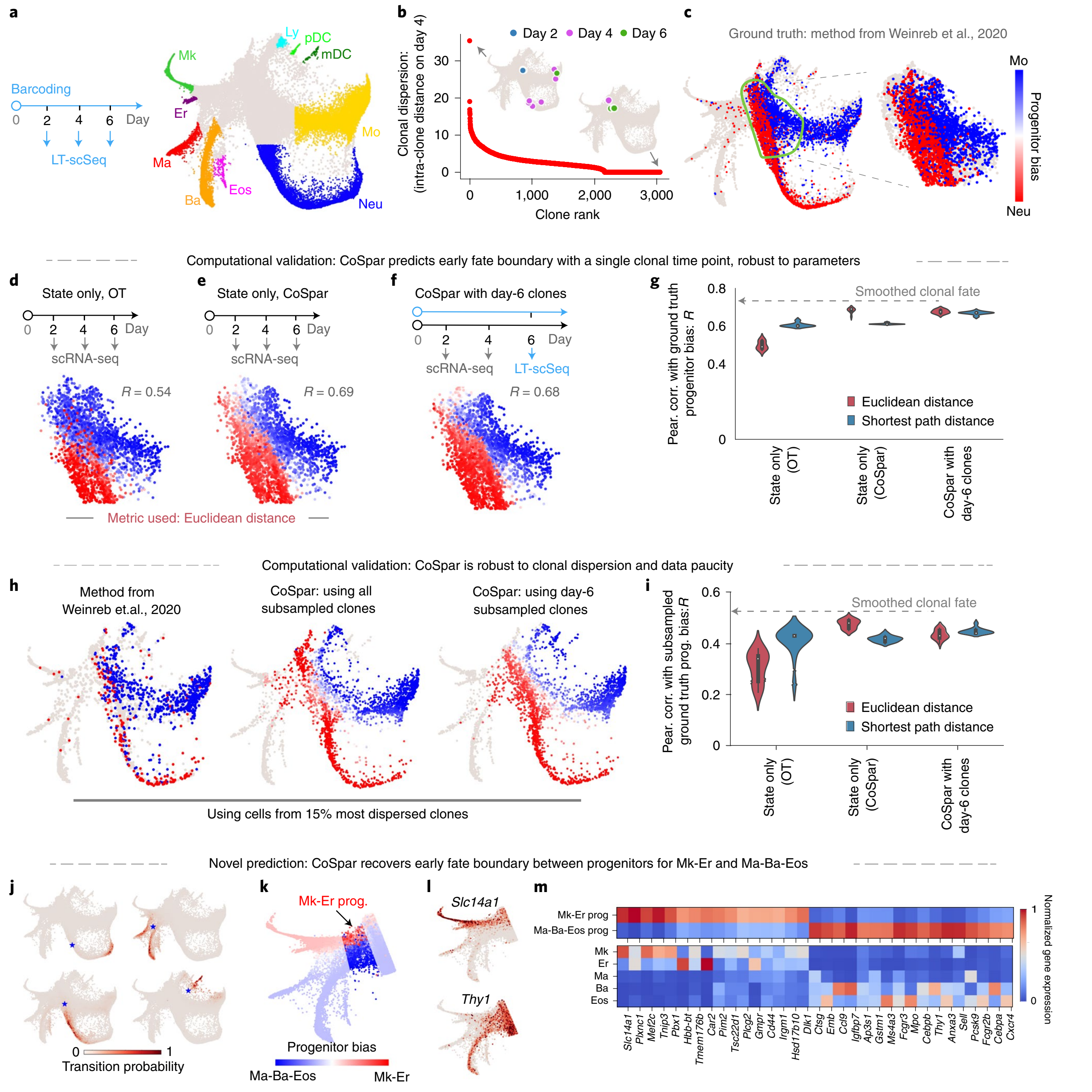

Fig. 4 | Benchmarking CoSpar and prediction of progenitor bias in hematopoiesis. Panels **a**–**i** benchmark CoSpar; panels **j**–**m** generate novel predictions. **a**, Experimental design and SPRING visualization of the hematopoiesis dataset from Weinreb et al.[12]. Early hematopoietic progenitors differentiate into megakaryocyte (Mk), erythrocyte (Er), mast cell (Ma), basophil (Ba), eosinophil (Eos), neutrophil (Neu), monocyte (Mo), lymphoid precursor (Ly), migratory (*Ccr7*$^+$) dendritic cell (mDC) and plasmacytoid dendritic cell (pDC). **b**, Clones ranked by intra-clone dispersion (that is, mean intra-clone graph distance between the observed cell states) after 4 d of differentiation. Two illustrative clones are shown. **c**, Bias toward Mo or Neu fate evaluated from all clonal data using the original method in Weinreb et al.[12]. Bias among early progenitors (right) serves as ground truth for benchmarking. **d**, **e**, Inference of progenitor bias using state information without clonal data, by OT or CoSpar algorithms. **f**, CoSpar inference of progenitor bias using clonal data from a single 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 is over parameter values for the OT method used to initialize CoSpar and choice of distance metric used, showing that clonal data reduce sensitivity to parameter choices in data analysis ($n = 8$ parameter sets; Supplementary Fig. 7e). The 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 **b**). **i**, Violin plots showing the distribution in inference performance with the downsampled data (quantified as in **f**) across parameter values ($n = 8$ parameter sets). **j**–**m**, Predicting the transcriptomic identity of Gata1$^+$ Mk-Er and Ma-Ba-Eos progenitors using CoSpar. **j**, Representative values of the inferred transition map for 2-d transitions from four example cell states (indicated by *). **k**, Heat map of predicted progenitor bias toward 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. The full list of fate-associated genes is provided in Supplementary Table 1. In **g** and **i**, white points indicate median; black bars span first to third quartiles.

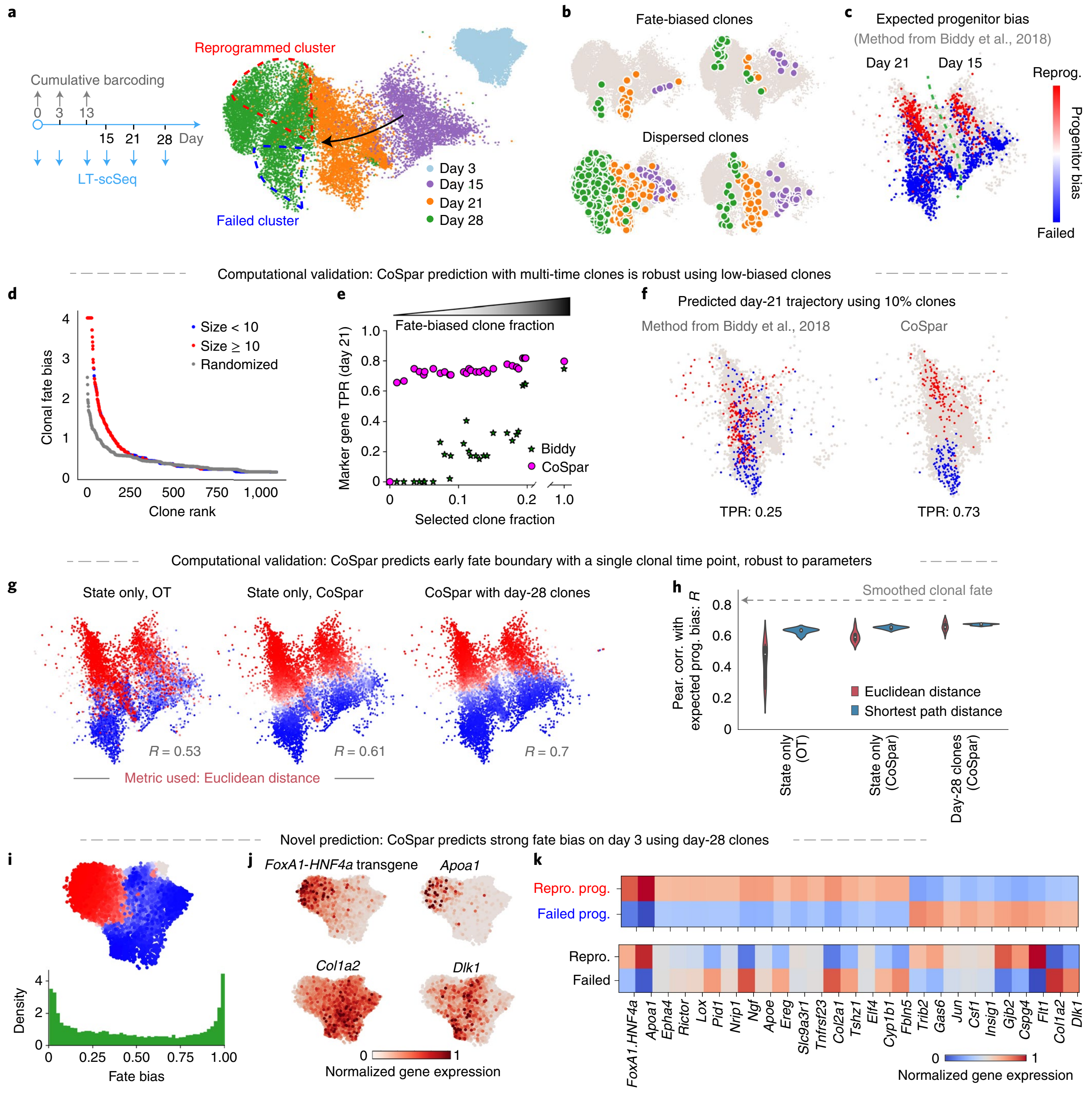

Fig. 5 | Progenitor bias in fibroblast reprogramming. Panels **a**–**h** benchmark CoSpar; panels **i**–**k** generate novel predictions. **a**, Experimental design and UMAP visualization of cell reprogramming from fibroblast cells to induced endoderm progenitors (iEPs) by ectopic expression of a transgene *FoxA1-HNF4a* on day 0 (ref. [15]). Schema shows time points for scRNA-seq only (gray 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 toward successful or failed reprogramming fates, using cells in clones selectively filtered for strong fate bias as in the original study[15]. **d**–**f**, Benchmarking CoSpar inference for 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), two-sided Fisher's 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.[15]) or CoSpar. Accuracy is assessed in the TPR of identifying genes associated with fate outcomes previously reported in ref. [15]. **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 state information or with endpoint (day 28) clonal information. **h**, Violin plots as in Fig. 4g quantifying prediction accuracy over a range of parameters (*n* = 13 parameter sets), showing consistent improvement by imposing coherence and sparsity and enforcing clonal relationships. **i**–**k**, Predicting early fate determination within 3 d 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 in Supplementary Table 2.

To evaluate CoSpar, we revisited this experiment after discarding 90% of clones, specifically retaining clones that show the least bias in reprogramming outcomes. Using these severely downsampled data consisting of dispersed clones, CoSpar still recapitulated 73 of 100 genes previously identified to discriminate reprogrammed and failed cells (Fig. 5f), including genes previously showing strong positive and negative association with reprogramming success (*Apoa1*, *Spint2*, *Col1a2* and *Peg3*) as well as *Mettl7a1*, which was found to improve reprogramming[15]. These genes could be associated with fate bias using as few as ten clones, even when deliberately selecting clones with minimal fate bias (Fig. 5d,e and Supplementary Fig. 10b). By contrast, the analytical approach used in the original study[15] failed to identify fate-predictive gene expression after such severe data reduction (Fig. 5e,f and Supplementary Fig. 10b). Furthermore, CoSpar performed robustly when using only clonal data from the final time point of the experiment (Fig. 5g,h and Supplementary Fig. 10c–e), and the result is robust whether we explicitly consider the nested clonal structure or not (Supplementary Fig. 10f–h). CoSpar was also robust to varying degrees of clone size heterogeneity and to cell dropout (Fig. 1f), as seen by subsampling clones or cells per clone (Supplementary Fig. 10j–l).

As in hematopoiesis, it is instructive to see what information on fate choice can be learned even without clonal relationships. When applying CoSpar without clonal data, we found that CoSpar could predict the same early fate biases (Fig. 5g, middle) but is again sensitive to the distance metric used (Fig. 5h). 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. Indeed, we are able to select the best metric here only because we are guided by the clonal information.

Finally, we applied CoSpar to predict fate bias at the earliest available time point after reprogramming is initiated (day 3). It is appreciated that fate choice occurs early during reprogramming, because subclones of cells labeled at day 3 acquire similar fates at later time points[15]. But which cells on day 3 reprogram successfully remains unknown as only seven clones labeled on day 3 were shared with other time points. CoSpar identified ~20% cells at day 3 with more than 80% putative bias toward reprogramming (Fig. 5i). These cells show a distinct transcriptional signature, with early expression of the transgene *FoxA1-HNF4a* used to induce reprogramming as well as *Apoa1* and early downregulation of *Col1a2* and *Dlk1* (ref. [15]) (Fig. 5j). We also identified multiple genes predicted to correlate with fate bias on day 3, whose significance in reprogramming has not been previously established, and that might offer targets for future investigation of early reprogramming success (Fig. 5k and Supplementary Table 2).

CoSpar predicts early fate bias during lung-directed differentiation. In the third experiment, human induced pluripotent stem cells (iPSCs) were differentiated into distal lung alveolar epithelial cells (induced alveolar epithelial type 2 cells (iAEC2s))[22,32]. 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 six time points, including days 17 and 21 (Fig. 6a). 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[22]. A re-examination of the clonal data, however, suggests strong fate biases as early as day 17. Of the 272 clones, 25% were enriched in either the iAEC2 or non-iAEC2 clusters (false discovery rate (FDR) = 0.01), and clonal compositions differed significantly from that of randomized clones (Fig. 6b). Accordingly, clonal representation of iAEC2s anti-correlates with other fates (Supplementary Fig. 11b,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. 6c) and others to be unbiased and multipotent; the latter 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. 11d). As a control, we expected weaker fate biases earlier in differentiation, which is confirmed by applying CoSpar to cells 2 d earlier (day 15; Supplementary Fig. 11e–g). Of genes differentially expressed between the two biased populations on day 17, we identified several established transcription factors that regulate lung differentiation: *CEBPD*, *NKX2-1*, *SOX9* and *SOX11* (Fig. 6d,e and Supplementary Table 3)[22,33–35].

We tested a prediction made using CoSpar relating fate bias to cell state at day 17. On this day of the differentiation protocol, leukemia inhibitory factor receptor (*LIFR*) showed reduced expression in cells biased toward non-iAEC2 fates and high expression in cells with a low bias to any fate and which also express proliferative markers (Fig. 6e). Previous work has shown that its ligand, LIF, is expressed in the adult mouse distal lung[36], and, in fetal rat lungs, LIF is expressed in developing lung epithelial and mesenchymal cells. Although LIF supplementation or repression in an explanted fetal mouse lung model altered lung growth and branching morphogenesis[37], whether LIF affects the proliferation or differentiation of developing lung epithelial progenitors remains unknown. To test if LIF influences developing lung endoderm, we used the same directed differentiation protocol to generate lung epithelial progenitors (expressing *NKX2-1*) and subsequently iAEC2s using BU3 NGST, an iPSC line. This iPSC line carries a green fluorescent protein (GFP) reporter targeted to the endogenous lung epithelial selective *NKX2-1* locus and a tdTomato reporter targeted to the *SFTPC* locus to identify iAEC2s[22,32]. After purifying NKX2-1$^{\mathrm{GFP+}}$ cells on day 15, we added 0, 5, or 50 ng ml^{-1} of recombinant human LIF (rhLIF) at days 17–19 of differentiation, the time point at which LIFR expression levels correlate with proliferation markers and fate predictions. At day 29, in response to transient LIF exposure, we observed a 3.1-fold (3.1 ± 1.4, $n = 5$) increase in the total cell number, suggesting a strong effect of LIF on epithelial progenitor proliferation. We additionally observed a decrease in both the fraction and number of cells reaching iAEC2 fate, identified by co-expression of *NKX2-1* and *SFTPC* (fraction dropping to 21 ± 14% of the untreated condition and cell count dropping to 53 ± 26%; Fig. 6g–I), indicating that LIF biases cell fate during differentiation. This result provides confidence that CoSpar predictions can identify targetable pathways in development and differentiation.

Discussion

We have developed a 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 similarly 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, using as little as 5–10% of data originally collected in previous experiments. The computational methods used in each original study to analyze clonal data were sensitive to such data reduction. CoSpar also successfully predicted early fate biases in these datasets using clonal information from only a single time point. When clonal data were removed entirely, results were sensitive to the choice of distance metric, and no single approach optimally inferred fate bias across all datasets. Both constraints of coherence and sparsity improved fate predictions, especially when the clonal data are highly dispersed (Supplementary Fig. 5a–c).

The robustness of CoSpar could greatly simplify the design of LT-scSeq experiments by enabling experiments with fewer cells

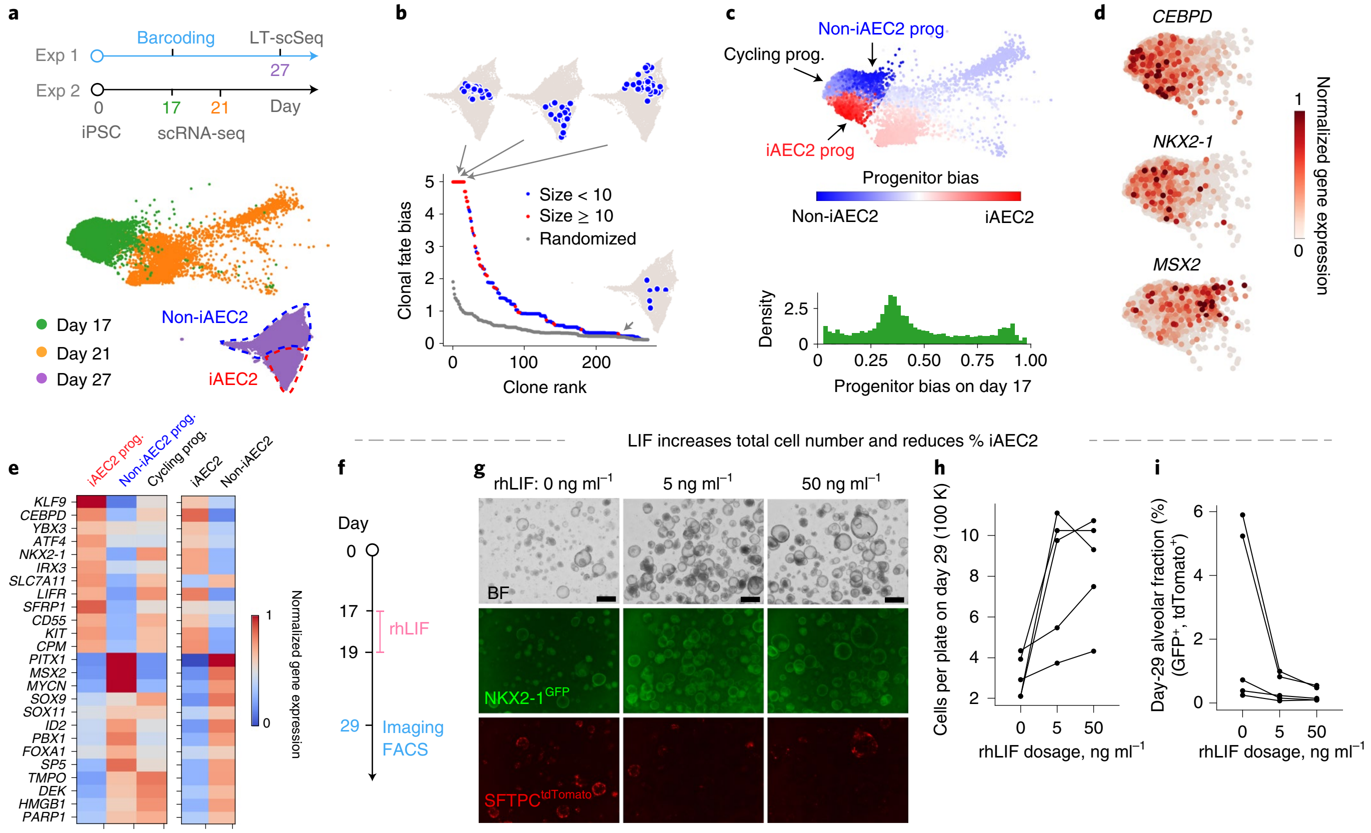

Fig. 6 | Progenitor bias during human iPSC differentiation into endodermal lineages. a, Experimental design and UMAP visualization for differentiating human iPSCs into iAEC2 lung cells and other endodermal cell types. **b**, Clones ranked by fate bias toward iAEC2 fate (bias defined as in Fig. 5d), with representative biased (top) and dispersed (bottom) clones shown. **c**, Predicted progenitor bias of cells toward iAEC2 fate on day 17 of differentiation, overlaid on the state embedding and shown as a histogram. Cycling progenitors are identified as cells enriched in *TOP2A* or *MKI67* (Supplementary Fig. 11d). **d**, **e**, Expression on day-17 states of selected genes predicted to correlate with iAEC2 and non-iAEC2 fates. In **e**, expression is shown alongside the corresponding expression in mature cells on day 27. **f**–**h**, rhLIF treatment increases total cell number and reduces the percentage of iAEC2 cells differentiated from human iPSCs. **f**, BU3 NGST iPSCs were differentiated as previously described[22], with the exception that cells were treated with 0, 5, and 50 ng ml^{-1} of rhLIF from days 17–19. **g**, Representative images of the cells on day 29. Scale bar, 500 µm. **h**, Quantification by flow cytometry of the total number of cells resulting after completion of the protocol ($n = 5$ biologically independent samples shown for each condition). **i**, Fraction of alveolar cells (iAEC2s; defined as GFP and tdTomato double positive) on day 29 at different rhLIF dosages (see Methods for details). FACS, fluorescence-activated cell sorting.

(Supplementary Fig. 10l), fewer clones or fewer time points. In all three datasets considered here, CoSpar reveals early fate boundaries that were not previously reported and yet are in agreement with the heterogeneity of key transcription factors and fate determinants. We predicted novel transcription factors, receptors and markers in each case, and they could facilitate manipulating fate outcomes, as exemplified here in the case of iPSC differentiation.

The examples that we analyzed specifically implement LT-scSeq using LARRY[12,22] and CellTagging[15], but CoSpar is not limited to these technologies. State measurements can be transcriptomic (via scRNA-seq or RNA fluorescence in situ hybridization (FISH)[38]) as well as proteomic and epigenomic; and lineage tracing can be achieved with static DNA barcodes[12,22], endogenous mutations[39] or exogenous DNA constructs that accumulate mutations over time, like CRISPR-based editing[2,14,19,20,40]. CoSpar can thus facilitate interpretation of the rapidly evolving field of LT-scSeq and, thus, accelerates exploration of development and disease.

CoSpar also has limitations, which directly follow from its central assumption. First, CoSpar is still limited to learning only the average fate biases of observed states; it does not separate biases in division versus differentiation rates, and it does not distinguish stochastic versus deterministic clonal biases due to hidden variables[12]. By enforcing coherent fate choices between similar cells (Fig. 2a,b), CoSpar becomes sensitive to choices in measuring cell–cell similarity and to the degree of smoothing used in implementing the algorithm (Supplementary Fig. 5a,f). 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 is lost due to oversampling or undersmoothing. In addition, when inferring progenitor bias from clones observed at a single late time point, CoSpar necessarily leans 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[41,42] (batch correction and multi-omics), decoding spatial transcriptomes from composite FISH measurements[43] and inferring responses of a system to individual perturbations from composite perturbation readouts[44–46]. Outside of biology, the association of measurements in one modality with sparse measurements in another can occur in marketing and social networks[47]. Forcing coherence and sparsity constraints could greatly improve map inference in general, reducing the cost of data acquisition and enabling discoveries.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01209-1.

Received: 6 May 2021; Accepted: 4 January 2022;
Published online: 21 February 2022

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.* **21**, 410–427 (2020).
3. Kester, L. & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179 (2018).
4. Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
5. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
6. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
7. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
8. Qiu, X. et al. Mapping vector field of single cells. Preprint at *bioRxiv* https://doi.org/10.1101/696724 (2019).
9. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943 (2019).
10. 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).
11. Tritschler, S. et al. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).
12. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
13. 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. USA* **115**, E2467–E2476 (2018).
14. 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).
15. Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
16. Bowling, S. et al. An engineered CRISPR/Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* **181**, 1410–1422 (2019).
17. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
18. Rodriguez-Fraticelli, A. E. et al. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* **583**, 585–589 (2020).
19. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
20. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
21. 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).
22. 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* **26**, 593–608 (2020).
23. 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).
24. Prasad, N., Yang, K. & Uhler, C. Optimal transport using GANs for lineage tracing. Preprint at https://arxiv.org/abs/2007.12098 (2020).
25. Forrow, A. & Schiebinger, G. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nat. Commun.* **12**, 4940 (2021).
26. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. 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. Ser. B Stat. Methodol.* **67**, 91–108 (2005).
28. Ferreira, R., Ohneda, K., Yamamoto, M. & Philipsen, S. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.* **25**, 1215–1227 (2005).
29. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).
30. 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 (2018).
31. Arinobu, Y. et al. Developmental checkpoints of the basophil/mast cell lineages in adult murine hematopoiesis. *Proc. Natl Acad. Sci. USA* **102**, 18105–18110 (2005).
32. Jacob, A. et al. Differentiation of human pluripotent stem cells into functional lung alveolar epithelial cells. *Cell Stem Cell* **21**, 472–488 (2017).
33. 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).
34. Rockich, B. E. et al. *Sox9* plays multiple roles in the lung epithelium during branching morphogenesis. *Proc. Natl Acad. Sci. USA* **110**, E4456–E4464 (2013).
35. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
36. Quinton, L. J. et al. Leukemia inhibitory factor signaling is required for lung protection during pneumonia. *J. Immunol.* **188**, 6300–6308 (2012).
37. Nogueira-Silva, C. et al. Leukemia inhibitory factor in rat fetal lung development: expression and functional studies. *PLoS ONE* **7**, e30517 (2012).
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 (2019).
40. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
41. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).
42. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
43. Cleary, B. et al. Compressed sensing for highly efficient imaging transcriptomics. *Nat. Biotechnol.* **39**, 936–942 (2021).
44. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
45. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896 (2016).
46. Nitzan, M., Casadiego, J. & Timme, M. Revealing physical interaction networks from statistics of collective dynamics. *Sci. Adv.* **3**, e1600396 (2017).
47. Aggarwal, C. C. *Recommender Systems: The Textbook* (Springer, 2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 might also include information on the environment of the cell, such as the transcriptome of neighboring cells and extracellular matrix composition. 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 sample only 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 accordance with common practice in scRNA-seq analysis[13], we use the experimentally observed set of cell states to define the set $\{X_i\}$. Therefore, the number of accessible states will be the same as the number of observed cells.

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 time points through the relationship[9]:

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

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 (Supplementary Note 1 and Supplementary Fig. 1d). Errors in lineage tracing affect how well we can recover the transition map (Supplementary Note 2).

Seminal work sought to infer $T_{i'i}(t_1,t_2)$, from $P_{i'}(t_1)$, $P_i(t_2)$ only[9]. One can greatly constrain the inference problem using the dynamics of clones[24,25]. 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. CoSpar works with data generated from both static and cumulative barcoding. For cumulative barcoding, each unique mutation or integration is considered a barcode, such that each cell can express more than one barcode. For more details, refer to Supplementary Note 4.

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

$I(t) \in \{0, 1\}^{M\times N_t}$: barcode-by-cell association matrix for the observed clonal data at time t, with discrete entries 0 or 1 indicating whether a cell contains the corresponding barcode 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.

$S^{(n)}(t) \in [0, 1]^{N_t\times N_t}$: state similarity matrix among N_t cell states at time t. n indicates the depth of graph diffusion used to create the similarity matrix.

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

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

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

$n_{df}(l)$: the depth n of graph diffusion (df) to create $S^{(n)}$ at l-th iteration of CoSpar.

n_{cs}: maximum number of iterations to carry out the CoSpar algorithm, unless the stopping criterion is reached.

Dynamic inference with CoSpar. CoSpar seeks to minimize an objective function with a close connection to compressed sensing (Fig. 2a). A heuristic, efficient algorithm implements the optimization through an iterative procedure. Referring to Fig. 2b, in each iteration, we (1) threshold the map to promote sparsity; (2) enforce clonal constraints by setting inter-clone transitions to be 0 and performing clone-wise normalization; and (3) locally average the transition map in high-dimensional state space to promote coherence. These steps are described by the following pseudo-code. The mathematical connection between our implementation and the objective function at Fig. 2a is detailed in Supplementary Note 3. Users need to provide the barcode-by-cell association matrix I_t and the count matrix from LT-scSeq measurements (for building similarity matrix S). 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 $l \leftarrow 1, 2, \ldots, n_{cs}$ do
    $n \leftarrow n_{df}(l)$
    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 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 θ, $\mathcal{P}$ and $S^{(n)}$ are defined below:

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

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

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

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

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

where $\tilde{\pi}_{ij}^m = T_{ij}$ if the transition $i \rightarrow j$ occurs within clone m and, otherwise, $\tilde{\pi}_{ij} = 0$. The normalization penalizes large clones, which tend to be more heterogeneous and less informative. For the case of cumulative barcoding, these choices of normalization and summation over all barcodes serve to place greater weight on smaller subclones than on larger parent clones (see Supplementary Note 4 for details).

CoSpar has two outputs: the smoothed transition map T and the map π that allows only 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[48,49] recovered ground truth results well in the simulations and experimental downsampling analyses. CoSpar constructs a weighted k-nearest neighbor (kNN) graph of observed cell states from a principal component analysis (PCA) embedding using the method proposed by uniform manifold approximation and projection (UMAP)[50], 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:

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

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 graph diffusion $S^{(n)}$, inspired by simulated annealing for finding the optimal solution in a rugged energy landscape[51]. Specifically, we use the sequence $\vec{n}_{df} = [n_1, n_2, \ldots]$ 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 $\vec{n}_{df} = [n_1, n_2, \ldots]$ 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 that three iterations are sufficient to obtain a convergent map (Supplementary Fig. 5f,g). 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 that CoSpar is more robust to ν_{cs} than to $\vec{n}_{df}$ (Supplementary Fig. 5d,e). We recommend setting $\nu_{cs} = 0.1$–0.2 (that is, trimming away transitions weaker than 10–20% of the maximum inferred from clonal data) for most applications. Other parameters are given for each respective dataset below. The subscript cs in these parameters means CoSpar.

Extending CoSpar to single time clones. When clonal data are available at only a single time point (that is, only $I(t_2)$ is available), dynamic inference is implemented as shown schematically in Fig. 2c. For the raionale, see Supplementary Note 5.

```
Function Joint Inference ($I_{t_2}$)
  $T^{(0)} \leftarrow T_{init}$
  Infer $\hat{I}_{t_1}(T^{(0)}, I_{t_2})$
  $\boldsymbol{T} \leftarrow$ CoSpar ($\hat{I}_{t_1}, I_{t_2}$)
  Return $T, \hat{I}_{t_1}$
```

These steps are defined below:

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)[9] but with simplification. WOT generalizes OT to allow non-uniform growth rates for each cell state. To obtain an approximate initial map, we avoid this generalization as it introduces additional tunable parameters. To construct the OT cost matrix[9], approximated by a cell–cell distance matrix, CoSpar offers two approaches: (1) Euclidean distance in the selected PCA space and (2) shortest path distance on a kNN graph of the state manifold. CoSpar accepts two parameters for this initialization: a k_{OT} for constructing the kNN graph and a regularization parameter ϵ_{OT}.

Alternative initialization (HighVar). 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 (for example, 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, $(\hat{I}_{t_1}, \hat{I}_{t_2}) \leftarrow \text{HighVar}$, and then use it to calculate the initial transition map, $T_{init} \leftarrow \text{CoSpar}(\hat{I}_{t_1}, \hat{I}_{t_2})$. For further details, see Supplementary Note 6. We found that this method can integrate datasets across different experiments that have strong batch effects, as demonstrated using the dataset on iPSC differentiation into iAEC2 cells reported in Fig. 6 of the paper. In this dataset, days 17 and 21 are from one experiment without lineage tracing, and day 27 is from a separate experiment; we found strong batch effects here, and the initialized map based on OT resembled a random initialization.

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:

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

under two constraints:

1. All initial states are clonally labeled—that is $\sum_{i,m} \hat{I}_{mi}(t_1) = N_{t_1}$.
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 $\mathcal{I}_t^{\zeta}$ 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 $N_t^{\zeta} \equiv |\mathcal{I}_t^{\zeta}|$. 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 a non-integer, we sample the cell number probabilistically from $\left\{ \left\lfloor N_{t_1}^{\zeta} \right\rfloor, \left\lceil N_{t_1}^{\zeta} \right\rceil \right\}$, with a mean of $N_{t_1}^{\zeta}$, where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ take the floor and ceiling 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 $P_{\mathcal{I}_{t_2}^{\zeta}}(i)$ that each state i in t_1 transitions to $\mathcal{I}_{t_2}^{\zeta}$, as defined below by Eq. (1);
2. select among not yet clonally labeled cell states at t_1 the top $N_{t_1}^{\xi}$ 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 inference accepts additional parameters for initializing T (k_{OT} and ϵ_{OT} for the OT method and gene selection parameter HighVar_gene_pctl for the HighVar method). We set $k_{OT} = 5$, $\epsilon_{OT} = 0.02$. The remaining parameters are provided for each dataset below.

Toolkit 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. 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

$$\boldsymbol{P}_{\mathcal{C}_{t_2}}(\boldsymbol{i}) = \sum_{j \in \mathcal{C}_{t_2}} \tilde{\boldsymbol{T}}_{ij}. \qquad (1)$$

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 $P_{\mathcal{C}_{t_2}}(i)$ toward the fate cluster $\mathcal{C}_{t_2}$:

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

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 $\tilde{P}_{\mathcal{B}}$ toward cluster $\mathcal{A}$ and $\mathcal{B}$, we compute the bias as

$$\boldsymbol{Q_i} = \frac{\tilde{\boldsymbol{P}}_{\mathcal{A}}(\boldsymbol{i})}{\tilde{\boldsymbol{P}}_{\mathcal{A}}(\boldsymbol{i}) + \tilde{\boldsymbol{P}}_{\mathcal{B}}(\boldsymbol{i})}. \qquad (2)$$

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

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

$$\mathcal{A}^* = \left\{ \text{arg}_i Q_i > \nu_{bias,\mathcal{A}} \right\}, \mathcal{B}^* = \left\{ \text{arg}_i Q_i < \nu_{bias,\mathcal{B}} \right\},$$

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. 7d,f). We define fate coupling as the correlation of fate maps toward two fates. Specifically, we first compute the fate map $P_{\mathcal{C}}$ toward 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 $\mathcal{C}_{t_2}^{(1)}, \ldots, \mathcal{C}_{t_2}^{(n)}$. The raw coupling is given by

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

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 (Figs. 5d and 6b). We evaluate the fate bias of a clone toward/against a given cluster as in ref. [15] by quantifying the statistical significance of a clone's occupancy of a set of transcriptomic states (for example, a cluster), when compared to that expected from a random sampling of cells. The P value (or P_{value}) is computed with Fisher's 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. 3a–d and Supplementary Fig. 5d–f).* 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. Here, $T_{x_1,x_2} = \mathcal{N}(x_2 - x_1; 1, \sigma)$, 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~Uniform(0, L) and then $x_2 = x_1 + 1 + \xi$ with ξ~ 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 $Z = (z_1, \ldots, z_{50})$ by setting $z_1 = x$ and adding independent Gaussian noise $z_k = 0.2\xi$ to each of the remaining 49 dimensions. We used $\sigma = 0.5$, $L = 100$ and $N = 1{,}000$. The number of detected clonal barcodes M was variable, as shown in the figure panels. CoSpar was applied with $\nu_{cs} = 0.2$ and $\vec{n}_{df} = [5, 5, 5]$.

Bifurcation and cell sampling (Fig. 3e–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~Uniform(0, 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 and B with probability 1/2. At $t = t_1$, we sampled cell states in each clone with a success rate of 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 $\sigma = 1$, $t_1 = 5$ and $L = 10$. $M = 100$ clones were simulated. CoSpar was applied with $\nu_{cs} = 0.2$ and $\vec{n}_{df} = [10{,}10{,}10]$.

Evaluating CoSpar with simulated data. We defined the TPR (Fig. 3d,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—that is, TPR $=$ $E[H(3\sigma - |\text{argmax}_{\Delta x} - T_{x_1,x_1+\Delta x} - 1|)]$ where $E(\cdot)$ is the mean over all rows of T, and $H(z) = \{1 \text{ for } z > 0; 0 \text{ otherwise}\}$. The progenitor bias for the bifurcation model (Fig. 3h,i) was calculated according to Eq. (2). Each of the TPR and progenitor bias comparisons (Fig. 3d,g,i) shows averages after application of CoSpar to five independent simulations.

Benchmarking and applying CoSpar to hematopoiesis. *Pre-processing.* Data[12] are available at the Gene Expression Omnibus (GEO) under accession number GSE140802. Data were pre-processed as originally described[12]: (1) unique molecular identifier (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 and min_cells = 3)[52]; 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 two-dimensional embedding and state annotation of cells were as in ref. [12], also available at the GEO website (GSE140802). We selected the top 40 principal components (PCs). Unless otherwise stated, we constructed a 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 $\nu_{cs} = 0.1$, $\vec{n}_{df} = [20, 15, 10]$, and we initialize the transition map using the OT method for joint inference.

Intra-clone dispersion (Fig. 4b). 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 $k = 5$. Figure 4b shows the dispersion normalized by the mean dispersion on day 2.

Transition map using the method from Weinreb et al.[12] (Fig. 4c,h and Supplementary Fig. 7a,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 in Fig. 4a). Multi-fate clones were discarded. Given this clone matrix $I^w(t)$, with $t = 2{,}4{,}6$, we computed the transition map as $T^w_{in}(t_1, t_2) = [I^w_{t_1}]^+ I^w_{t_2}$, 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. 4c shows the progenitor bias Q_i on day 2 and day 4 computed from $T^w_{in}\,(2, 4)$, $T^w_{in}\,(2, 6)$ and $T^w_{in}\,(4, 6)$ using Eq. (2).

Fate map reconstruction error (Supplementary Fig. 7a,b). To allow comparison between methods, we used $\pi(4, 6)$ from CoSpar with $\nu_{cs} = 0.2$ or $T^w_{in}(4, 6)$ from the Weinreb et al. method, constructed from subsampled clones on days 4–6, to compute the fate map $P_{\mathcal{C}}\,(i, t = 4)$ toward cells annotated with a given fate (cell set $\mathcal{C}$) according to Eq. (1). We evaluated the inferred maps by comparing them to a ground truth fate map $P^{true}_{\mathcal{C}}\,(i, t = 2)$ from the Weinreb et al. method with all clones from days 2–4. We evaluated the prediction using the Wasserstein distance[53] between the two distribution $P_{\mathcal{C}}$ and $P^{true}_{\mathcal{C}}$, restricted to the progenitor state space $\bar{\mathcal{C}}$ (that is, 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, whereas $P^{true}_{\mathcal{C}}\,(i, t = 2)$ 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 the shortest path distance. The Wasserstein distance is given by $d_{wass} = \sum_{i,j \in \bar{\mathcal{C}}} P_{\mathcal{C}}\,(i)\, T^{OT}_{ij} P^{true}_{\mathcal{C}}(j)$. We computed the Wasserstein distance for three major fates—neutrophils, monocytes and basophils—and reported the average.

WOT (Supplementary Figs. 7f and 10e). Results shown were obtained using the WOT package (https://github.com/broadinstitute/wot)[9], using default parameters: $\epsilon_{OT} = 0.05$, $\lambda_1 = 1$ and $\lambda_2 = 50$. We used a uniform growth rate for each cell.

Neuron network prediction (Supplementary Fig. 8). This section deals with comparison of CoSpar's performance to that of an alternative algorithm, SuperOT. SuperOT trains a neural network to predict the most likely fate of a progenitor cell at some later time point. To do so, it trains on clones represented by two time points and can be tested using clones not included in the training set. To compare CoSpar performance to SuperOT, we generated comparable predictions from CoSpar. To this end, we first applied CoSpar to infer a Transition Map $T(t_1, t_2)$ and then coarse-grained the transition map into a Fate Map $P_{\mathcal{C}_{t_2}}$ (see above). We then defined the predicted fate choice of each cell as its most likely fate according in the Fate Map: $\text{fate}_i = \underset{\mathcal{C}_{t_2}}{\text{argmax}}\, P_{\mathcal{C}_{t_2}}\,(i)$. Then, we used this labeled dataset to train the neuron network MPLClassifier to predict fate choices of observed clones in a test dataset. We ran sklearn.neural_network.MLPClassifier with the following modified parameters to train the model: random_state = 1, max_iter = 300 and alpha = 0.1. For both CoSpar and SuperOT, the resulting fate predictions for the majority fate per clone were correlated with that observed in the test dataset.

Benchmarking and applying CoSpar to fibroblast reprogramming. *Pre-processing.* Data were downloaded from the 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 (days 0, 3 and 13). Following Biddy et al.[15], we concatenated day-0 and day-3 barcodes to form a unique clonal ID for downstream analysis. However, keeping three barcodes per cell, thus allowing nested clonal structure, works equally well (Supplementary Fig. 10f–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. 5a) 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. 10a), a gene expressed in fibroblasts that failed reprogramming[15]. The reprogrammed and failed clusters were used to define the progenitor bias in this dataset.

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

Selecting dispersed clones (Fig. 5d,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 cell states on day 28, thus having intermediate values of γ_o. We selected dispersed clones satisfying $R_- \leq \gamma_o < R_+$, where $R_- = x$ and $R_+ = 0.4 - 2x$, and x parameterizes the selection 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 subsampled dataset in Fig. 5e.

Transitions using the method from Biddy et al. (Fig. 5e,f). Following Biddy et al.[15], we first identified clones that are enriched or depleted in the reprogrammed cluster according to Fisher's exact test. Among statistically significant clones ($P_{value} \leq 0.05$), 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.[15]. 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. 5e,f and Supplementary Fig. 10b). For a putative reprogramming ($\mathcal{D}_r$) and failed ($\mathcal{D}_f$) population predicted by either CoSpar or the Biddy et al. method, we assessed their accuracy by the overlap of their top DEGs with those from the reference population (defined by the fate-biased clones selected by Biddy et al.[15]).

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 $\pi = \mathcal{P}\,(\theta\,(T, 0))$ as follows:

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

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 et al. prediction, when $|\mathcal{D}_x| \leq 200$, 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 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 $P \geq 0.05$). 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

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

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

Application of CoSpar to in vitro differentiation of lung endoderm. *Pre-processing.* Data were downloaded from the GEO, accession numbers GSE137805 and GSE137811. We selected highly variable genes using the filter_genes function (min_vscore_pctl = 80, min_counts = 3 and min_cells = 3) and normalized the UMI counts per cell to 10,000. We used the top 40 PCs to construct a *k*NN graph with *k* = 20 for downstream analysis. We inherited the original embedding on days 17 and 21 by Hurley et al.[22] (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. 11a), marker genes for iAEC2 cells[22].

Applying CoSpar. To apply joint inference (Fig. 6c and Supplementary Fig. 11f,g), we initialized the transition map using the HighVar method with HighVar_gene_pctl = 80 and ran CoSpar with $\nu_{cs} = 0.2$, $\vec{n}_{df} = [20, 15, 10]$. See Jupyter Notebook implementation at https://cospar.readthedocs.io/.

Directed differentiation of iPSCs into lung epithelium. We performed directed differentiation of BU3 NGST human iPSCs into NKX2.1$^+$ lung epithelial cells as previously described[22,32]. The BU3 NGST line carries GFP and tdTomato reporters targeted to the endogenous lung epithelial selective *NKX2-1* and *SFTPC* loci[32]. On day 15 of differentiation, NKX2-1^{GFP+} cells were sorted and resuspended in undiluted growth factor-reduced Matrigel (Corning) at a dilution of 500 cells per microliter. Cells were fed every other day with previously described CK + DCI media[22,32] supplemented with 10 µm of Y-27632 (Rock Inhibitor) (CK + DCI + RI). On days 17–19, this media was supplemented with 0, 5 or 50 ng ml^{-1} of rhLIF (R&D Systems). On day 29, Z-stack images of live organoids were taken and processed on a Keyence BZ-X710 fluorescence microscope. Z-stacks were used to generate full-focus projections using BZ-X Analyzer software (version 1.3.1.1), followed by background subtraction and intensity correction. Cells were then collected and digested into a single-cell suspension as previously described[22] and analyzed by flow cytometry to assess the yield of cells delineated by the fluorescent reporters and DRAQ7 (live/dead stain).

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

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 the Gene Expression Omnibus database with accession number GSE140802, the reprogramming dataset with accession number GSE99915 and the lung dataset with accession numbers GSE137805 and GSE137811.

Code availability

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

References

48. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
49. 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).
50. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
51. van Laarhoven, P. J. M. & Aarts, E. H. L. in *Simulated Annealing: Theory and Applications* (eds van Laarhoven, P. J. M. & Aarts, E. H. L.) 7–15 (Springer Netherlands, 1987).
52. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).
53. Peyré, G. & Cuturi, M. Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.* **11**, 355–607 (2019).

Acknowledgements

S.-W.W. is a Damon Runyon Quantitative Biology Fellow supported by the Damon Runyon Cancer Research Foundation (02-20). A.M.K. acknowledges support by National Institutes of Health (NIH) grants R01HL14102-01 and R01-CA218579. K.H. would like to acknowledge funding from the Health Research Board Emerging Clinical Scientist Award ECSA-2020-011. D.N.K. is supported by NIH grants R01HL095993, U01TR001810 and N01 75N92020C00005. We thank T. Scully for helping with figures.

Author contributions

S.-W.W. and A.M.K. conceived the project. S.-W.W. devised the computational method, wrote the package and carried out CoSpar analyses. K.H. and D.N.K. designed and supervised, and M.J.H. carried out and analyzed, iPSC differentiation experiments. S.-W.W. and A.M.K. wrote the manuscript. A.M.K. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01209-1.

Correspondence and requests for materials should be addressed to Shou-Wen Wang or Allon M. Klein.

Peer review information *Nature Biotechnology* thanks Samantha A. Morris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature research

Corresponding author(s): Shou-Wen Wang
Allon M. Klein

Last updated by author(s): Dec 19, 2021

Reporting Summary

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

Statistics

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

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	CoSpar (v0.1.8); Waddington-OT (v1.0.8; https://broadinstitute.github.io/wot/) ; LineageOT (v0.1.0; https://lineageot.readthedocs.io)

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

Data

Policy information about availability of data

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

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

The 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.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Cell culture studies were performed using n=5. This number was chosen with the expectation that it would provide 100% power to detect statistically significant differences in cell populations based on preliminary runs while also allowing for analysis of multiple independent differentiations.
Data exclusions	No data was excluded
Replication	5 replicates are provided as discussed in sample size.
Randomization	Randomization was not applicable as we used only one cell type and three treatment conditions in equal numbers.
Blinding	Blinding was not used as cells needed to be maintained in specific conditions throughout experiments and by the time of analysis different culture conditions were easily identified, regardless of blinding.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
□	☒ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	BU3 NGST cell line
Authentication	Yes
Mycoplasma contamination	This cell line has tested negative for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study