INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Check for updates

Cite this article: Wu X, Wollman R, MacLean AL. 2023 Single-cell Ca^{2+} parameter inference reveals how transcriptional states inform dynamic cell responses. *J. R. Soc. Interface* **20**: 20230172.
https://doi.org/10.1098/rsif.2023.0172

Received: 24 March 2023
Accepted: 10 May 2023

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
systems biology

Keywords:
Bayesian inference, dynamical systems, calcium signalling, single cell, spatial transcriptomics

Author for correspondence:
Adam L. MacLean
e-mail: macleana@usc.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6662626.

THE ROYAL SOCIETY
PUBLISHING

Single-cell Ca^{2+} parameter inference reveals how transcriptional states inform dynamic cell responses

Xiaojun Wu[1], Roy Wollman[2,3] and Adam L. MacLean[1]

[1]Department of Quantitative and Computational Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA
[2]Department of Integrative Biology and Physiology, and [3]Department of Chemistry and Biochemistry, University of California Los Angeles, Los Angeles, CA, USA

XW, 0000-0002-7613-8415; ALM, 0000-0003-0689-7907

Single-cell genomic technologies offer vast new resources with which to study cells, but their potential to inform parameter inference of cell dynamics has yet to be fully realized. Here we develop methods for Bayesian parameter inference with data that jointly measure gene expression and Ca^{2+} dynamics in single cells. We propose to share information between cells via transfer learning: for a sequence of cells, the posterior distribution of one cell is used to inform the prior distribution of the next. In application to intracellular Ca^{2+} signalling dynamics, we fit the parameters of a dynamical model for thousands of cells with variable single-cell responses. We show that transfer learning accelerates inference with sequences of cells regardless of how the cells are ordered. However, only by ordering cells based on their transcriptional similarity can we distinguish Ca^{2+} dynamic profiles and associated marker genes from the posterior distributions. Inference results reveal complex and competing sources of cell heterogeneity: parameter covariation can diverge between the intracellular and intercellular contexts. Overall, we discuss the extent to which single-cell parameter inference informed by transcriptional similarity can quantify relationships between gene expression states and signalling dynamics in single cells.

1. Introduction

Models in systems biology span systems from the scale of protein/DNA interactions to cellular, organ and whole organism phenotypes. Their assumptions and validity are assessed through their ability to describe biological observations, often accomplished by simulating models and fitting them to data [1–4]. Under the framework of Bayesian parameter inference and model selection, the available data are used along with prior knowledge to infer a posterior parameter distribution for the model [5]. The posterior distribution characterizes the most likely parameter values to give rise to the data as well as the uncertainty that we have regarding those parameters. Thus, parameter inference provides a map from the dynamic phenotypes that we observe in experiments to the parameters of a mathematical model.

Single-cell genomics technologies have revealed a wealth of information about the states of single cells that was not previously accessible [6]. This ought to assist with the characterization of dynamic phenotypes. However, it is much less clear how to draw maps between dynamic phenotypes of the cell and single-cell states as quantified via genomic measurements. The challenge in part lies in the combinatorial complexity: even if a small fraction of genes contain information regarding the phenotype of interest, say a few hundred, this is more than enough to characterize any feasible number of states of an arbitrarily complex dynamical process.

This leads us to a central question: can the integration of single-cell gene expression data into a framework for parameter inference improve our understanding of the cellular phenotypes of interest? Here, various sources of transcriptional noise must be taken into account [7–9], which we propose to address by taking a global view and comparing cells first by their similarity across many genes, and, after inference, by their similarity in posterior parameter distributions. Our previously work provides the ideal data for this approach: we jointly measured dynamics and gene expression in the same single cells [10]. Here, we apply our new parameter inference framework to study Ca^{2+} signalling dynamics and signal transduction in response to adenosine triphosphate (ATP) in human mammary epithelial (MCF10A) cells.

Ca^{2+} signalling regulates a host of cellular responses in epithelial cells, from death and division to migration and molecular secretion, as well as collective behaviours, such as organogenesis and wound healing [11–13]. In response to ATP binding to purinergic receptors, a signalling cascade is initiated whereby phospholipase C (PLC) is activated and in turn hydrolyses phosphatidylinositol 4,5-bisphosphate (PIP2), producing inositol 1,4,5-trisphosphate (IP3) and diacylglycerol (DAG). The endoplasmic reticulum (ER) responds to IP3 by the activation of Ca^{2+} channels: the subsequent release of calcium from the ER into the cytosol produces a spiked calcium response. To complete the cycle and return cytosolic calcium levels to steady state, the sarco/ER Ca^{2+}-ATPase (SERCA) channel pumps the Ca^{2+} from the cytosol back into the ER [14,15]. Ca^{2+} signalling is highly conserved, regulating cell phenotypic responses across mammals, fish and flies [12,16,17] as well as in prokaryotes [18]. Since the Ca^{2+} response to ATP occurs quickly in epithelial cells: on a timescale that is almost certainly faster than gene transcription, we work under the assumption that the transcriptional state of the cell does not change in the duration of the experiment.

Our ability to measure gene expression in thousands of single cells has not only led to new discoveries but has also fundamentally changed how we identify and characterize cell states [19]. Technologies used to quantify gene expression in single cells include sequencing and fluorescent imaging. The latter permits the measurement of hundreds of genes in spatially resolved populations of single cells. Small molecule fluorescence *in situ* hybridization (smFISH) can be multiplexed to achieve this high resolution by protocols, such as MERFISH [20] and seqFISH [21]. Moreover, by coupling multiplexed smFISH with fluorescent imaging of Ca^{2+} dynamics using a GFP reporter in MCF10A cells, we are able to jointly capture the dynamic cell responses and the single-cell gene expression in the same single cells [10]. These data offer new potential to study the relationships between transcriptional states of cells and the dynamic phenotypes these may produce.

Models of gene regulatory networks and cellular signalling pathways described by ordinary differential equations (ODEs) capture the interactions between gene transcripts, proteins or other molecular species and their impact on cellular dynamics. Well-established dynamical systems theory offers a range of tools with which to analyse transient and equilibrium behaviour of ODE models [22]; it remains an open question whether or not it is appropriate to make equilibrium assumptions of living cells [23]. Constraining dynamic models of cellular/molecular processes with single-cell data via inference offers much potential to gain new insight into dynamics, albeit coming with many challenges, given, among other things, the complex sources of noise in these data and the lack of explicit temporal information in (snapshot) datasets gathered at one time point [24]. Parameter inference has provided insight into clonal relationships of single cells [25,26] and stem cell differentiation/cell state transitions [27,28]. Inference methods have also been applied to single-cell data for the discovery of new properties of single-cell oscillations [29,30] and cell–cell variability [31–33], as well as to study cell–cell communication [34]. New methods to infer the parameters of models of stochastic gene expression provide means to study single-cell dynamics in greater depth [35,36].

Here, we model Ca^{2+} dynamics via ODEs based on previous work [37,38]. We develop a parameter inference framework to fit Ca^{2+} response dynamics in many single cells. We perform inference of multiple cells sequentially, through the construction of ‘cell chains’. A cell chain is an ordering of cells, which can be random or directed by some measure, e.g. by similarity of gene expression or of Ca^{2+} dynamic response. Given a cell chain, we propose to infer the parameters of the Ca^{2+} ODE model in a single cell via a transfer of information from its cell predecessor in the chain. We achieve this by setting the prior of the current cell in the chain informed by the posterior of its predecessor. We will use this framework to assess the extent to which transcriptional cell states inform dynamic cell responses.

In the next section, we present the model and the methods implemented for parameter inference using Hamiltonian Monte Carlo in Stan [39]. We go on to study the results of inference: we discover that priors informed by cell predecessors accelerate parameter inference, but that cell chains with randomly sampled predecessors perform as well as those with transcriptional similarity-informed predecessors. Analysis of hundreds of fitted single cells reveals that cell-intrinsic versus cell-extrinsic posterior parameter relationships can differ widely, indicative of fundamentally different sources of underlying variability. By perturbing the posterior distributions, we assess model parameter sensitivities in light of Ca^{2+} dynamics. We also find that variability in single-cell gene expression is associated with variability in posterior parameter distributions, both for individual gene–parameter pairs and globally, via principal component analysis. We go on to cluster cells by their posterior parameter distributions, and discover that only for gene expression-based cell chains are there clear relationships between gene expression states and dynamic cell phenotypes.

2. Material and methods

2.1. A model of Ca^{2+} dynamics in response to ATP

We model Ca^{2+} signalling pathway responses in MCF10A human epithelial cells using nonlinear ODEs, as previously developed [37,38]. The model consists of four state variables: phospholipase C (PLC), inositol 1,4,5-trisphosphate (IP3), the fraction of IP3-activated receptor (h) and cytoplasmic Ca^{2+}. The four variables are associated with a system of four nonlinear ODEs describing the rates of change of the Ca^{2+} pathway species following ATP stimulation, to characterize dynamic responses in MCF10A cells. The equations are given by

$$\frac{\mathrm{d[PLC]}}{\mathrm{d}t} = \mathrm{ATP} \cdot \mathrm{e}^{-K_{\mathrm{ATP}}t} - K_{\mathrm{off,ATP}}[\mathrm{PLC}], \qquad (2.1)$$

$$\frac{\mathrm{d[IP3]}}{\mathrm{d}t} = V_{\mathrm{PLC}}\frac{[\mathrm{PLC}]^2}{K_{\mathrm{IP3}}^2 + [\mathrm{PLC}]^2} - K_{\mathrm{off,IP3}}[\mathrm{IP3}], \tag{2.2}$$

$$\frac{\mathrm{d}h}{\mathrm{d}t} = a([\mathrm{Ca}^{2+}] + d_{\mathrm{inh}})\left(\frac{d_{\mathrm{inh}}}{[\mathrm{Ca}^{2+}] + d_{\mathrm{inh}}} - h\right), \tag{2.3}$$

$$\frac{\mathrm{d}[\mathrm{Ca}^{2+}]}{\mathrm{d}t} = \beta\left(\varepsilon(\eta_1 m_\infty^3 h^3 + \eta_2)(c_0 - (1+\varepsilon)[\mathrm{Ca}^{2+}]) - \eta_3\frac{[\mathrm{Ca}^{2+}]^2}{k_3^2 + [\mathrm{Ca}^{2+}]^2}\right),$$

$$\beta = \left(1 + \frac{K_e[B_e]}{(K_e + [\mathrm{Ca}^{2+}])^2}\right)^{-1}$$

$$\text{and } m_\infty = \left(\frac{[\mathrm{IP3}]}{d_1 + [\mathrm{IP3}]}\right)\left(\frac{[\mathrm{Ca}^{2+}]}{d_5 + [\mathrm{Ca}^{2+}]}\right). \tag{2.4}$$

The equations describe a chain of responses following ATP binding to purinergic receptors: the activations of PLC, IP3, the $\mathrm{IP_3R}$ channel on the surface of the ER and finally the release of Ca^{2+} from the ER into the cytoplasm [38]. Ca^{2+} may also enter the ER through the $\mathrm{IP_3R}$ channel and the SERCA pump [38]. Our model differs from Yao *et al.* [38] in that we combine the product of two parameters in the previous model, $K_{\mathrm{on,ATP}}$ and ATP, into a single parameter, ATP. This reduction of the model parameter space removed the redundancy that would otherwise exist in the distributions of $K_{\mathrm{on,ATP}}$ and ATP. A description of each of the parameters in the model is given in table 1, where reference values for each of the model parameters are found in Lemon *et al.* [37] and Yao *et al.* [38].

2.2. Data collection and preprocessing

The data consist of a joint assay measuring Ca^{2+} dynamics and gene expression via multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) [20]. Ca^{2+} dynamics in a total of 5128 human MCF10A cells are measured via imaging for 1000 s (ATP stimulation at 200 s) using a GCaMP5 biosensor. Immediately following this step, 336 genes are measured by MERFISH [10]. The Ca^{2+} trajectories are smoothed using a moving average filter with a 20 s window size (electronic supplementary material, figure S1). After smoothing, data points occurring before ATP stimulation are removed. Data points for each Ca^{2+} trajectory after $t = 300$ are downsampled by a factor of 10; the trajectories are at or close to steady state by this time. After removing the data for the first 200 s and downsampling for the last 700 s, each processed trajectory consists of Ca^{2+} response data on 171 time points ($t = 200, 201, \ldots, 298, 299, 300, 310, 320, \ldots, 1000$). All numerical experiments in this work will evaluate Ca^{2+} response on those same 171 time points. Single-cell gene expression data are collected using MERFISH after the Ca^{2+} imaging as previously described [10,20].

2.3. Generating cell chains via cell–cell similarity

Cell–cell similarity is quantified via single-cell transcriptional states, i.e. by comparing x_m^i and x_m^j, the expression of m genes in cells i and j. We obtain a symmetric cell–cell similarity matrix, W, from the log-transformed MERFISH expression data via optimization in SoptSC [41]: entries $W_{i,j}$ denote the similarity between cells i and j. To create a chain of cells linked through their similarity in gene expression space, we:

1. Construct a graph $G = (V, E)$; each node is a cell and an edge is placed between two cells if they have a similarity score above zero.
2. For a choice of initial (root) cell, traverse G and record the order of cells traversed.

Ideally, each cell would be visited exactly once; however, this amounts to finding a Hamiltonian path in G, an NP-complete problem. Therefore, as a heuristic solution we use a depth-first search (DFS), which can be completed in linear time. From the current node, we randomly select an unvisited neighbour node and set this as the next current node, recording it once visited (pre-order DFS). If the current node has no unvisited neighbours, it backtracks until a node with unvisited neighbours is found. When there is no unvisited node left, every node in the graph has been visited exactly once. Given cases where the similarity matrix is sparse (as we have here), the DFS generates a tree that is very close to a straight path.

2.4. Bayesian parameter inference with posterior-informed prior distributions

We seek to infer dynamic model parameters in single cells, informed by cell–cell similarity via the position of a cell in a cell chain. We use the Markov chain Monte Carlo (MCMC) implementation: Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS) in Stan [39,42]. HMC improves upon the efficiency of other MCMC algorithms by treating the sampling process as a physical system and employing conservation laws [43]. From an initial distribution, the algorithm proceeds through intermediate phase of sampling (warm-up) until (one hopes) convergence to the stationary distribution. During warm-up, NUTS adjusts the HMC hyperparameters automatically [42].

The prior distribution over parameters is a multivariate normal distribution, with dimensions θ_j, $j = 1, \ldots, m$, where m is the number of parameters. This choice of prior makes it straightforward to pass information from the inferred posterior distribution of one cell to the next cell in line to be sampled, which will be described in §2.5. Let f be a numerical solution of the ODE model, and y_0 be the initial condition. Then, in each single cell, the Ca^{2+} response to ATP is generated by the following process:

$$\begin{aligned} \theta_j &\sim \mathcal{N}(\mu_{\theta_j}, \sigma_{\theta_j}^2) \\ \hat{y}(t) &= f(y_0, t; \theta) \\ \sigma &\sim \mathrm{Cauchy}(0, 0.05) \\ y(t) &\sim \mathcal{N}(\hat{y}(t), \sigma^2), \end{aligned}$$

where we truncate the prior so that each θ_i is bounded by 0 from below. The Cauchy distribution is chosen to generate the noise for observed Ca^{2+} response as it contains greater probability mass in its tails, thus encouraging NUTS to explore extreme values of the parameter space more frequently.

For the first cell in a chain, we use a relatively uninformative prior, the 'Lemon' prior (table 1), derived from parameter value estimates in previous work [37,38,40]. For the ith cell in a chain ($i > 1$), the prior distribution is constructed from the posterior distribution of the $(i-1)$th cell (§2.5). For each cell, NUTS is run for four independent chains with the same initialization. To simulate $\hat{y}(t)$ during sampling, we use the implementation of fourth- and fifth-order Runge–Kutta in Stan [39]. For each output trajectory y, its error is the Euclidean distance between y and data y^* for all 171 data points

$$\epsilon(y, y^*) := \left(\sum_{k=0}^{170}(y(t_k) - y^*(t_k))^2\right)^{1/2}.$$

The error of a posterior sample for a cell is the mean error of trajectories simulated from all draws in the sample

$$\epsilon_{\mathrm{sample}} := \frac{1}{N}\sum_{i=1}^{N}\epsilon(y_i, y^*),$$

Table 1. Definition and description of the ODE model parameters. Prior distributions are derived from [37,38,40].

name	description	prior distribution	unit
ATP	concentration of ATP that activates PLC	$\mathcal{N}(5, 4)$	s^{-1}
K_{ATP}	ATP decay rate	$\mathcal{N}(0.0083, 0.0025)$	s^{-1}
$K_{\text{off,ATP}}$	PLC degradation rate	$\mathcal{N}(1.25, 1)$	s^{-1}
V_{PLC}	maximum velocity for IP3 generation	$\mathcal{N}(1, 1)$	$\mu M \cdot s^{-1}$
K_{IP3}	equilibrium constant for IP3 generation through PLC	$\mathcal{N}(0.5, 0.01)$	μM
$K_{\text{off,IP3}}$	IP3 degradation rate	$\mathcal{N}(1.25, 1)$	s^{-1}
a	time constant of IP3 channel	$\mathcal{N}(1, 1)$	s^{-1}
d_{inh}	dissociation constant for IP3 channel calcium inhibiting subunit	$\mathcal{N}(0.4, 0.01)$	μM
K_e	dissociation constant for calcium buffer	$\mathcal{N}(10, 4)$	μM
B_e	concentration of calcium buffer	$\mathcal{N}(150, 25)$	μM
d_1	dissociation constant for IP3 channel IP3 activating subunit	$\mathcal{N}(0.13, 0.01)$	μM
d_5	dissociation constant for IP3 channel calcium activating subunit	$\mathcal{N}(0.0823, 0.01)$	μM
ϵ	ER to cytosolic volume	$\mathcal{N}(0.185, 0.01)$	—
η_1	IP3 channel permeability constant	$\mathcal{N}(575, 625)$	s^{-1}
η_2	ER leak permeability constant	$\mathcal{N}(5.2, 1)$	s^{-1}
η_3	Ca^{2+} pump permeability constant	$\mathcal{N}(45, 25)$	s^{-1}
c_0	concentration of free Ca^{2+} in the ER	$\mathcal{N}(4, 1)$	μM
k_3	SERCA pump dissociation constant	$\mathcal{N}(0.4, 0.01)$	μM

where N is the sample size and y_i is the output trajectory from the ith draw in the sample.

Convergence of NUTS chains is evaluated using the $\widehat{R}$ statistic: the ratio of between-chain variance to within-chain variance [39,44]. A typical heuristic used is $\widehat{R}$ between 0.9 and 1.1, indicating that for this set of chains the stationary distributions reached for a given parameter are well mixed. There are two caveats on our use of $\widehat{R}$ in practice:

1. For our model, we observe that well-fit (i.e. not overfit) Ca^{2+} trajectories did not require $\widehat{R} \in (0.9, 1.1)$ for all parameters. Thus, we assess $\widehat{R}$ only for the log posterior, using a more tolerant upper bound of 4.0.
2. There are cases where one chain diverges but three of the four are well mixed. In such cases, we choose to retain the three well-mixed chains as a sufficiently successful run. Thus if $\widehat{R}$ is above the threshold, before discarding the run, we compute $\widehat{R}$ for all three-wise combinations of chains, and retain the run if there exist three well-mixed chains.

2.5. Constructing and constraining prior distributions

We construct the prior distribution of the ith cell from the posterior of the $(i-1)$th cell. The prior mean for each parameter θ_j for the ith cell is set to $\mu_j^{(i-1)}$, the posterior mean of θ_j from the $(i-1)$th cell. The variance of the prior for θ_j is derived from $\sigma_j^{(i-1)}$, the posterior variance of θ_j from the $(i-1)$th cell. To (a) sufficiently explore the parameter space, and (b) prevent instabilities (rapid growth or decline) in marginal parameter posterior values along the cell chain, we scale each $\sigma_j^{(i-1)}$ by a factor of 1.5 and clip the scaled value to be between 0.001 and 5. The scaled and clipped value is then set as the prior variance for θ_j for the ith cell.

2.6. Dimensionality reduction and sensitivity analyses

To compare posterior samples from different cells, we use principal component analysis (PCA). Posterior samples are projected onto a subspace by first choosing a cell (the focal cell) and normalizing the posterior samples from other cells against the focal cell, either by min–max or z-score normalization. Min–max normalization transforms a vector x to $(x - x_{\min})/(x_{\max} - x_{\min})$, where $x_{\min}$ is the minimum and $x_{\max}$ the maximum of x. z-score normalization transforms x to $(x - \mu_x)/\sigma_x$, where μ_x is the mean and σ_x is the standard deviation of x. Normalizing to the focal cell amounts to setting $x_{\min}$, $x_{\max}$, μ_x, σ_x to be the values corresponding to the focal cell for all cells normalized. We perform PCA (implemented by scikit-learn 0.24 [45]) on the normalized focal cell posterior samples and project them into the subspace spanned by the first two principal components. The normalized samples from all other cells are projected onto the PC1–PC2 subspace of the focal cell.

We develop methods for within-posterior sensitivity analysis to assess how perturbations of model parameters within the bounds of the posterior distribution affect Ca^{2+} responses. Given $\tilde{\theta}$, the posterior distribution of a cell, each parameter θ_j is perturbed to two extreme values: $\tilde{\theta}_j^{(0.01)}$, the 0.01-quantile of $\tilde{\theta}_{\cdot,j}$, and $\tilde{\theta}_j^{(0.99)}$, the 0.99-quantile of $\tilde{\theta}_{\cdot,j}$. Nine 'evenly spaced' samples are drawn from the posterior range of $\tilde{\theta}$ for the parameter of interest, $\tilde{\theta}_{\cdot,j}$: the kth draw corresponds to a sample $\tilde{\theta}_{i,\cdot}$ such that $\tilde{\theta}_{i,j} = \tilde{\theta}_j^{(0.1k)}$, the 0.1$k$-quantile of $\tilde{\theta}_{\cdot,j}$. For each draw $\tilde{\theta}_{i,\cdot}$, we replace $\tilde{\theta}_{i,j}$ by either $\tilde{\theta}_j^{(0.01)}$ or $\tilde{\theta}_j^{(0.99)}$ and then simulate a Ca^{2+} response. The mean Euclidean distances between trajectories simulated from the evenly spaced samples and the perturbed samples are used to quantify the sensitivity of each parameter perturbation.

2.7. Correlation analysis and cell clustering of MERFISH data

Correlations between single-cell gene expression values and posterior parameters from the Ca^{2+} pathway model are determined for variable genes. We calculate the z-scores of posterior means for each parameter of a cell sampled from a population, and remove that cell if any of its parameters has a posterior mean z-score smaller than −3.0 or greater than 3.0. PCA is performed on log-normalized gene expression of remaining cells using scikit-learn 0.24 [45], which yields a loadings matrix A such that $A_{i,j}$ represents the 'contribution' of gene i to component j. We designate gene i as variable if $A_{i,j}$ is ranked top 10 or bottom 10 in the jth column of A for any $j \le 10$. For each variable gene, we calculate the Pearson correlation between its log-normalized expression value and the posterior means of individual model parameters. Gene–parameter pairs are ranked by their absolute Pearson correlations and the top 30 are selected for analysis. Gene–parameter pair relationships are quantified by linear regression using a Huber loss, which is more robust to outliers than mean squared error.

To cluster cells using their single-cell gene expression, raw count matrices are normalized, log-transformed, and scaled to zero mean and unit variance before clustering using the Leiden algorithm at 0.5 resolution [46], implemented in Scanpy 1.8 [47]. Marker genes for each cluster are determined by a t-test.

2.8. Clustering of cell posterior parameter distributions

Cells are clustered according to their posterior distributions. For each parameter, the posterior means for each cell are computed and scaled to [0, 1]. The distance between two cells is defined as the m-dimensional Euclidean distance between their posterior means (where m is the number of parameters). Given distances calculated between all pairs of cells, agglomerative clustering with Ward linkage is performed using SciPy 1.7 [48]. Marker genes for each cluster identified are determined using a t-test.

3. Results

3.1. Single-cell priors informed by cell predecessors enable computationally efficient parameter inference

To study the dynamic Ca^{2+} responses of cells to ATP stimulation, we fit the ODE model (equations (2.1)–(2.4)) to data in single cells using Bayesian parameter inference (figure 1*a*). Only those MCF10A cells classified as 'responders' to ATP were included—cells with very low overall responses (less than 1.8 Ca^{2+} peak height) were filtered out. To assess whether cell chains improve inference, we performed parameter inference of the ODE model in single cells fit either individually, each from the same prior (we used the 'Lemon' prior (table 1)), or fit via the construction of a cell chain. In a cell chain there is a transfer of information, whereby the posterior parameter distribution of one cell informs the prior distribution of the next cell in the chain. The first cell in the chain was fit using the Lemon prior. We are primarily interested in cell chains constructed using transcriptional similarity: we constructed cell chains based on a single-cell gene expression similarity metric and compared them with alternatives (see Material and methods; electronic supplementary material, table S1). We studied the effects of different choices of g, where $\pi_{i+1} = g(p_i)$, p_i is the posterior distribution for cell i, and π_{i+1} is the prior distribution for the following cell. We found that transformations via scaling and clipping were necessary to sufficiently explore the parameter space for each cell while maintaining stable marginal posterior distributions along a cell chain (electronic supplementary material, section S1.1, figure S2). We tested various numerical methods to solve the ODE system (stiff and non-stiff), and found that we could simulate Ca^{2+} responses sufficiently well using a non-stiff solver (electronic supplementary material, figure S3), so for inference runs with hundreds of single cells we proceeded to use a non-stiff solver.

Parameter inference of the ODE model via a cell chain (denoted *Similar-r1*) was more efficient and gave more accurate results than individually fit cells (figure 1*b*,*c*), with shorter overall computational run times and higher posterior model probabilities (electronic supplementary material, table S2). The model fit quality was also higher for the cell chain versus individually fit cells as assessed by the $\widehat{R}$ statistic (electronic supplementary material, table S3). To test whether these improved model fits are in part due to longer fitting times rather than the construction of the cell chain directly, we fit the same cell consecutively ten times: the fits improved over the 10 repeat epochs, but the only substantial improvements were seen for the first couple of epochs, after which improvements were minimal and the overall fit quality was comparable to the same cell fitted in the chain (electronic supplementary material, section S1.2, figure S4*a*–*f*), albeit with some evidence for overfitting in individual parameters (electronic supplementary material, figure S4*g*,*h*). Thus, the quality of fits obtained from fitting in a cell chain are not inherently due to more time spent running inference but are due to the transfer of information between different cells along the chain.

The advantage of transfer learning in cell chains can be demonstrated by the higher predictive power of sampled posterior distributions. We predicted Ca^{2+} responses of test cells for which the parameters have not been inferred (i.e. cells not in *Similar-r1*), using the initial conditions of the test cells but parameters from elsewhere. Each test cell was simulated using parameters sampled from: the posterior of a fitted cell with similar gene expression to the test cell; the posterior of a random fitted cell; and reference values from literature ('Lemon' values; see table 1). To compare predicted Ca^{2+} responses, we used the Euclidean distance between a predicted Ca^{2+} trajectory and data to quantify the prediction error. We found that posteriors of similar cells and posteriors of random cells had equally good predictive power on test cells: in both cases better than using Lemon values for prediction (figure 1*d*). These results illustrate how constructing priors for cells using posterior information from other cells offers greater ability to capture the dynamics of a new cell not previously modelled.

To assess whether cell chains ordered using gene expression information improve inference performance over cell chains ordered randomly, we compared inference runs of at least 500 cells in a chain, with priors informed by cell predecessor, where the chain construction was either random or gene expression similarity based. The performance of cell chains ordered randomly—evaluated by computational efficiency (sampling times) and accuracy of fits (model posterior probabilities)—was not significantly different to that of the similarity-based chains (electronic supplementary material, table S4). Therefore, although the use of a cell chain (priors informed by cell predecessors) improved inference relative to individually fit cells, the choice of cell predecessors (similarity-based versus randomly assigned) did not affect computational efficiency or the accuracy of fits.

We also studied the effects of HMC parameters on sampling. We found that sampling times were faster without loss of fit quality when we reduced the maximum tree depth (a parameter controlling the size of the search space) from 15 to 10, since rarely was a tree depth greater than 10 used in

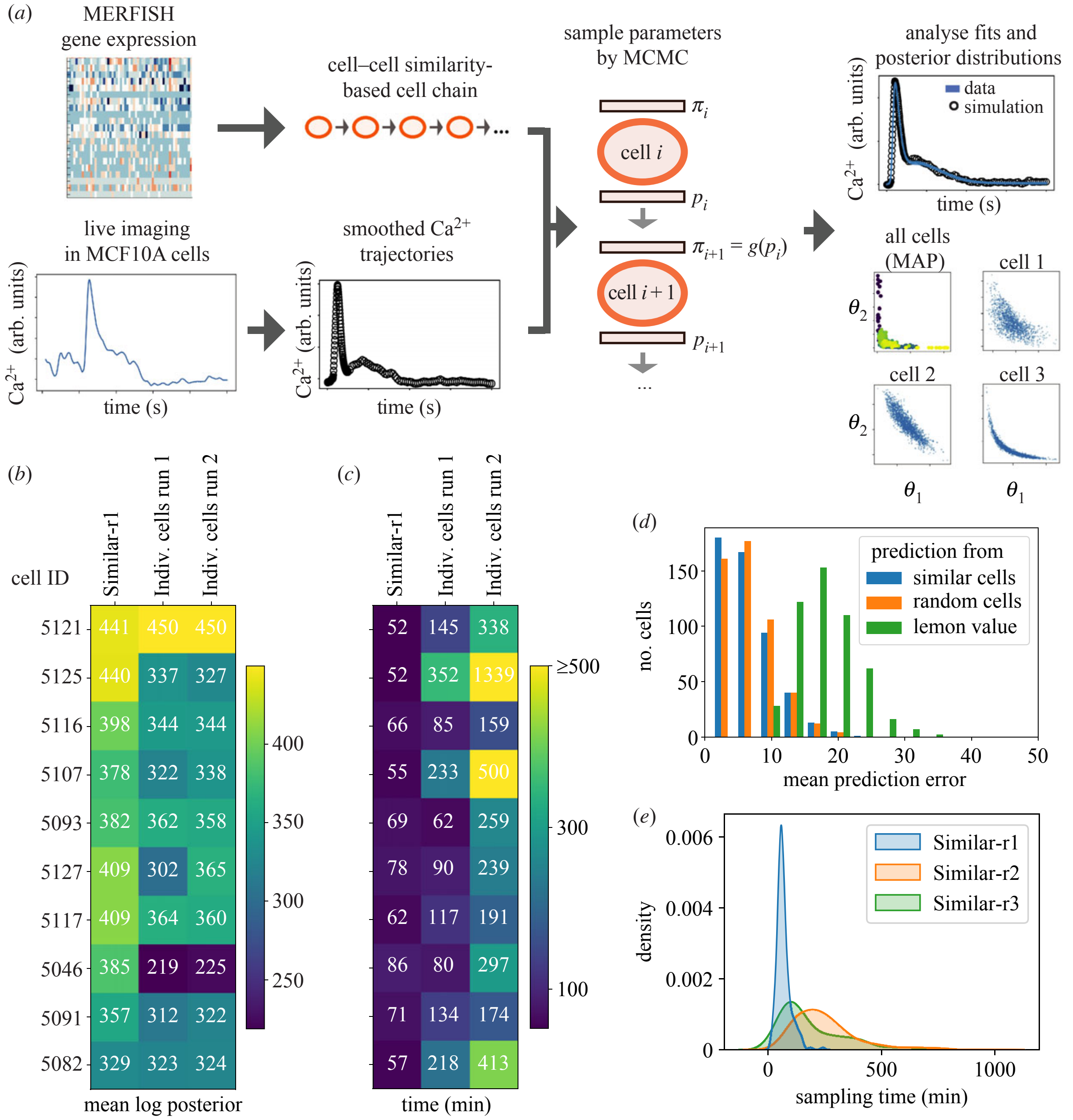

Figure 1. Cell chains improve performance of single-cell parameter inference. (*a*) Workflow for single-cell parameter inference along a cell chain. π_i and p_i, respectively, denote the prior and posterior distributions of cell *i*. MAP: maximum *a posteriori* value. (*b–c*) Comparison of inference for gene expression similarity-based cell chain (*Similar-r1*; first column) versus cells fit individually, i.e. without a cell chain (second two columns). Output metrics are the mean log posterior, where higher values denote better fits; and the sampling times. Each row represents a cell. For *Similar-r1*, one cell was sampled from a Lemon prior before Cell 5121 to inform the prior of that cell (not shown). Run 1 for the individually fitted cells has 500 warm-up steps and run 2 has 1000. (*d*) Comparison of predictions of the dynamics of an unfitted cell using: samples from the posterior distribution of a fitted cell with similar gene expression, samples from the posterior distribution of a random fitted cell, and reference values from literature ('Lemon' values; see table 1). (*e*) Comparison of the effects of HMC parameters. Parameters (*num. warmup steps*; *max. tree depth*) are for r1: (500, 10), r2: (1000, 15), r3: (500, 15).

practice; so this reduction did not negatively impact the model fits (electronic supplementary material, table S5). We also found that a warm-up period of 500 steps was sufficient for convergence of MCMC chains for most cells. Setting the maximum tree depth to 10 and the number of warm-up steps to 500 led to much faster sampling times for large populations of cells (figure 1*e*).

3.2. Analysis of single-cell posteriors reveals divergent intracellular and intercellular sources of variability

The posterior distributions of hundreds of cells show striking differences between marginal parameters: some are consistent across cells in a chain while others vary widely. To quantitatively assess this, we ran two similarity-based cell chains with identical cell ordering for the final 100 cells but with different initial cells. We found that while some marginal posterior parameters were similar for all cells (e.g. $K_{\mathrm{off,ATP}}$, figure 2*a*), others diverged for the same set of cells along a chain (e.g. d_5, figure 2*b*). Relative changes in marginal posteriors were seen to be tightly correlated. We computed the fold change in mean marginal posterior parameter values between consecutive cells along the chain (figure 2*a*,*b*, second row): the majority of consecutive cell pairs were tightly correlated both in direction and magnitude, even when the absolute values diverged. We obtained similar

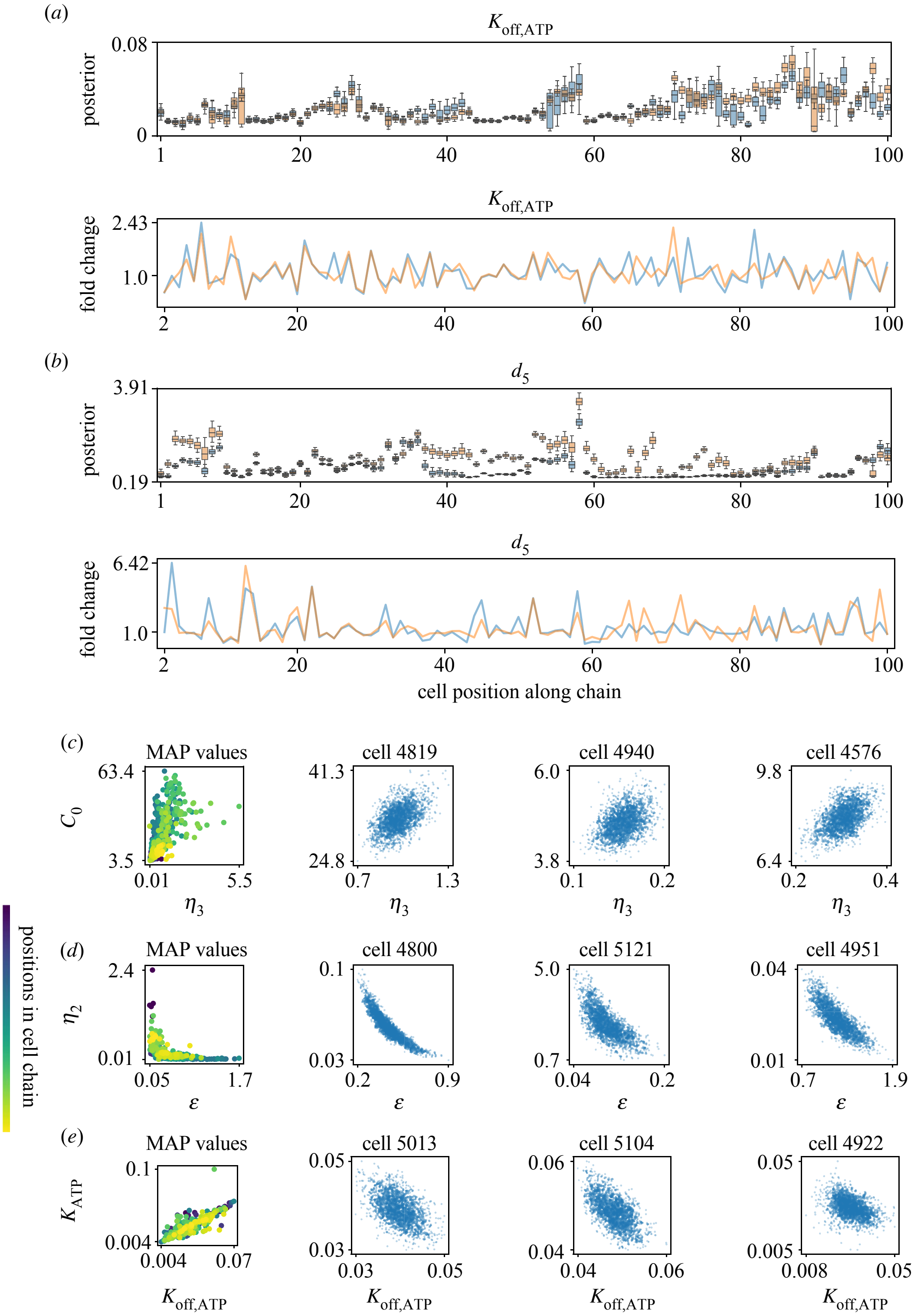

Figure 2. Parameter dependencies revealed by the analysis of marginal posterior distributions along cell chains. (*a*) Marginal parameter posterior distributions for the PLC degradation rate ($K_{off,ATP}$) in two parallel runs of the same similarity-based cell chain. Box plots of the first–third quartiles of the distribution with whiskers denoting its full range (upper). Marginal posterior mean fold changes of consecutive cells in the chain (lower). (*b*) As for (*a*), with the IP3 channel dissociation parameter (d_5). (*c*) Left: intercellular variability. Scatter plot of the maximum *a posteriori* (MAP) values of 500 cells for parameters η_3 and c_0. Colour indicates position along the chain. Right: intracellular variability. Scatter plots of 500 samples from the posterior of one cell from the chain; three representative cells are shown. (*d*) As for (*c*), with parameters ϵ and η_2. (*e*) As for (*c*), with parameters $K_{off,ATP}$ and K_{ATP}.

results for random cell chains run in parallel with different initial cells (electronic supplementary material, figure S5). Analysis of the posterior values of parameters relative to their 'Lemon' values (see table 1) revealed in some cases large distances between them (e.g. d_5 varies from 0.08 (Lemon) to posterior values greater than 3). There are several possible causes for these prior–posterior discrepancies, including differences in the biological system and differences in experimental inputs, e.g. the stimulus used or the amount of stimulus that cells receive.

Further analysis of the marginal posterior distributions revealed two uninformative ('sloppy' [49]) parameters. The posterior distributions of B_e and η_1 drifted, i.e. varied along the chain independent of the particular cell (electronic supplementary material, figure S6*a*,*b*). Given these insensitivities, we studied model variants where either one or both of these parameters were set to a constant. Comparing chains of 500 cells each, the reduced models performed as well as the original in terms of sampling efficiency and convergence (electronic supplementary material, figure S6*c–e* and table S6). Posterior predictive checks of the reduced models showed no significant differences in simulated Ca^{2+} trajectories. Thus, for further investigation into the parameters underlying single-cell Ca^{2+} dynamics, we analysed the model with both B_e and η_1 set to a constant. This cell chain is referred to as *Reduced-3*.

We discovered striking differences between intracellular and intercellular variability through analysis of the joint posterior distributions of parameters in chain *Reduced-3*. Several parameter pairs were highly correlated, as can be expected given their roles in the Ca^{2+} pathway, e.g. as activators or inhibitors of the same species. However, comparison of parameter correlations within (intra) and between (inter) cells yielded stark differences. Some parameter pairs showed consistent directions of correlation intercellularly (along the chain) and within single cells. The Ca^{2+} pump permeability (η_3) and the concentration of free Ca^{2+} (c_0) were positively correlated both inter- and intra-cellularly (figure 2*c*). Similarly, the ER-to-cytosolic volume (ϵ) and the ER permeability (η_2) were negatively correlated in both cases (figure 2*d*). However, the ATP decay rate (K_{ATP}) and the PLC degradation rate ($K_{\mathrm{off,ATP}}$) were positively correlated along the chain (posterior means) but—for many cells—negatively correlated within the cell (figure 2*e*). The distribution of MAP values is well-mixed, i.e. there is no evidence of biases arising due to a cell's position in the chain: the variation observed in the posterior distributions represents biological differences in the population. These differences may be in part explained by the differences in scale: intercellular parameter ranges are necessarily as large as (and sometimes many times larger than) intracellular ranges. On these different scales, parameters can be positively correlated over the large scale but negatively correlated locally, or vice versa. These divergent sources of variability at the inter- and intra-cellular levels highlight the complexity of the dynamics arising from a relatively simple model of Ca^{2+} pathway activation.

3.3. Quantifying the sensitivity of Ca^{2+} responses in a population of heterogeneous single cells

We conducted analysis of the sensitivity of Ca^{2+} responses to the model parameters. Typically, one defines a parameter sensitivity as the derivative of state variables with respect to that parameter [50,51]. Here, we are most interested in how the dynamics are affected by parameter perturbations over the range of their marginal posterior distributions. Thus, we evaluate the model response to a given parameter perturbation across its marginal posterior distribution in a population of cells as follows. First, sample from a cell's posterior distribution, and alter each sample such that the parameter of interest is set to an extreme value according to its marginal posterior distribution (0.01-quantile or 0.99-quantile). We then simulate trajectories from these altered samples (figure 3*a*), and use the distance between unperturbed and perturbed model trajectories to define the sensitivity of model output to that parameter, taking the mean of nine simulated trajectories.

We find that there is a lot of variation in the Ca^{2+} responses: sensitive to some model parameters and insensitive to others (figure 3*b*). Notably, the sensitivities of the least sensitive parameters had mean values of close to 1.0: similar to the distances obtained from the best-fit posterior values (electronic supplementary material, table S5), i.e. the Ca^{2+} response is insensitive to these parameters across the whole posterior range. The insensitive parameters were not simply those which had the highest posterior variance: there was little correlation between the inferred sensitivity and the posterior variance (electronic supplementary material, table S7), compare e.g. parameters d_1 and d_5.

Analysis of the Ca^{2+} responses to parameter perturbations provides means to predict how much Ca^{2+} responses are affected by changes in extracellular and intracellular dynamics (figure 3*c*,*d*). For example, low concentrations of ATP result in very low Ca^{2+} responses; increasing the concentration of ATP can more than double the peak response (figure 3*c*). The importance of IP3 in Ca^{2+} signal transduction is in agreement with the results of Yao *et al.* [38]; here we go further in that we can quantify the particular properties of the Ca^{2+} response affected by each parameter. In the case of $K_{\mathrm{off,IP3}}$, the main effect is also in the peak height of the Ca^{2+} response (figure 3*d*).

3.4. Variability in gene expression is associated with variability in Ca^{2+} dynamics

We studied variation between pairs of genes and parameters sampled from a cell population to assess whether relationships between them might exist. We found that several gene–parameter pairs were correlated. In general, the proportion of variance explained between a gene–parameter pair was low; this is to be expected given the many sources of variability in both the single-cell gene expression and the Ca^{2+} responses.

Analysis of the most highly correlated gene–parameter pairs (see Material and methods and electronic supplementary material, table S8) identified a number of genes that were correlated with multiple parameters, e.g. PPP1CC, as well as parameters that were correlated with multiple genes, e.g. η_3. Pairwise relationships were analysed via linear regression. The top four correlated gene–parameter pairs from a similarity-based cell chain are shown in figure 4*a–d*: cells are well mixed according to their positions along the chain, i.e. correlations are not due to local effects. The pairwise correlations overall are low, which we expect given single-gene inputs. Performing multiple regression could improve predictive power; however, our goal here is to study whether any evidence supports the existence of individual gene–parameter relationships. We performed the same analysis on a randomly ordered cell chain, where the same gene–parameter relationships were recapitulated, albeit with lower absolute correlation values (figure 4*e–h* and electronic supplementary material, table S9). There is no discernable influence of a cell's position in a chain on the gene–parameter relationship, confirming that these correlations among a cell population reflect the variability in the population rather than any sampling artefacts.

We compared the top genes ranked by gene–parameter correlations for four populations: from two randomly sampled and two similarity-informed cell chains. Gene–parameter pairs were sorted by their absolute Pearson correlation coefficients, and the genes ranked by their positions among sorted pairs. In total, we identified 75 correlated gene–parameter pairs for

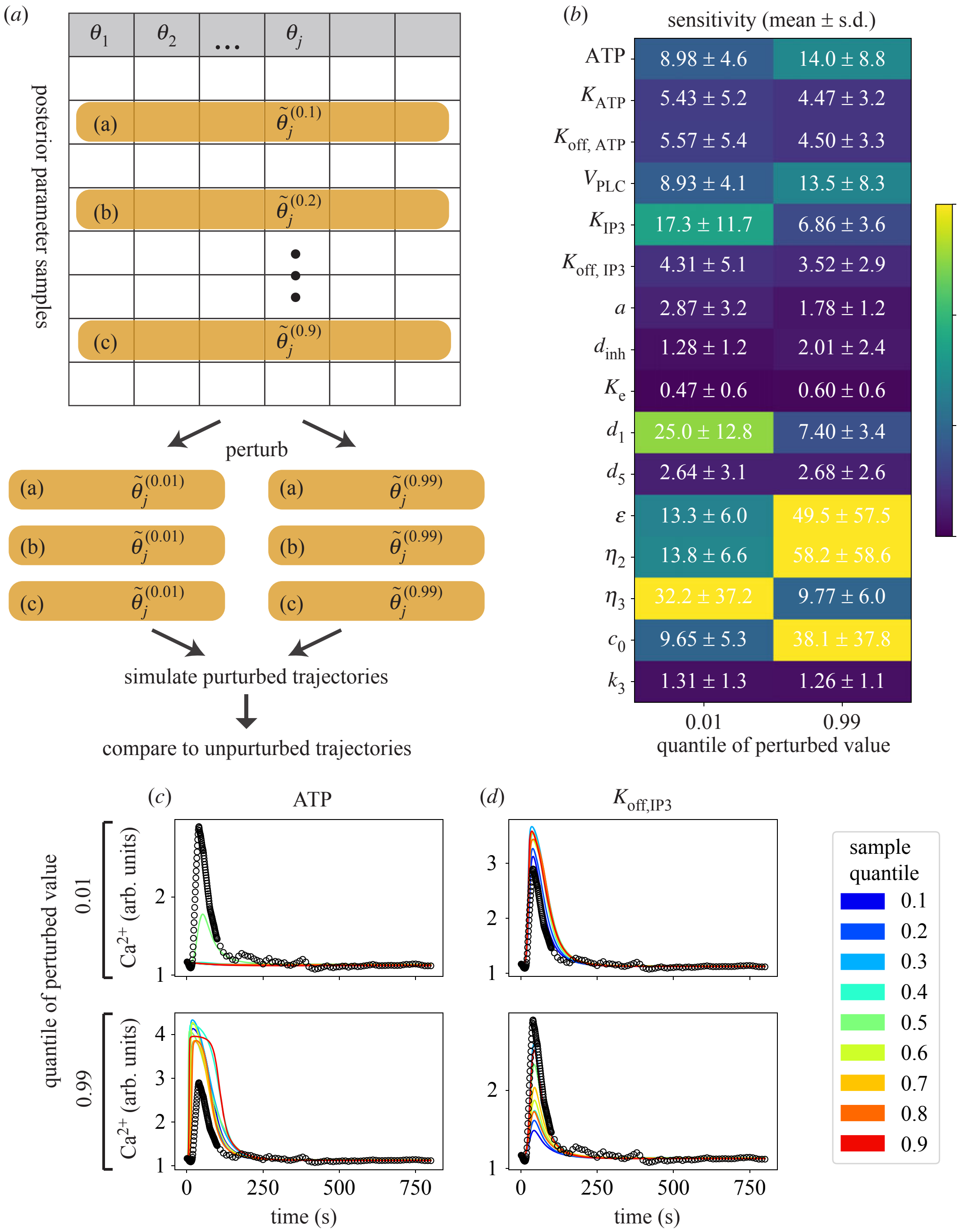

Figure 3. Sensitivity of Ca^{2+} responses to parameter perturbations. (*a*) Schematic diagram of approach to studying model sensitivities with respect to parameters sampled across the posterior range. The parameter to be perturbed (θ_j) is set to extreme values (0.01- or 0.99-quantile of the marginal distribution). (*b*) Parameter sensitivities for a population of fitted cells from a gene expression similarity-based chain (*Reduced-3* model). Sensitivities are reported in terms of model distances between baseline and perturbed trajectories. (*c*) Simulated model trajectories in response to perturbations in ATP concentration. (*d*) Simulated model trajectories in response to perturbations in $K_{\mathrm{off,IP3}}$.

the *Reduced-3* chain, applying a Bonferroni correction for multiple testing (electronic supplementary material, figure S7). Out of the top 30 genes, 25 appeared in the top 30 in at least three-quarters of the cell chains studied (figure 4*i*). Of these 25 genes, 20 also appeared as top-10 marker genes from unsupervised clustering (into three clusters) of the gene expression data directly (figure 4*i*). The high degree of overlap between these gene sets demonstrates that a subset of genes expressed in MCF10A cells explain not only their overall transcriptional variability but also their variability in Ca^{2+} model dynamics. These results are also suggestive of how information content pertaining to the heterogeneous Ca^{2+} cellular responses is encapsulated in the parameter posterior distributions.

Next, we turn our attention from the level of individual genes/parameters to that of the whole: what is the relationship between the posterior parameter distribution of a cell and its global transcriptional state? We used PCA for dimensionality reduction of the posterior distributions to address this question. We selected a cell (denoted the 'focal cell') from a similarity-based cell chain (*Reduced-3*) and decomposed its posterior distribution using PCA. We projected the posterior distributions of other cells onto the first two components of the focal cell (figure 4*j*,*k* and electronic supplementary material, figure S8*a*,*b*) to evaluate the overall similarity between the posterior distributions of cells relative to the focal cell. On PCA projection plots, posterior samples

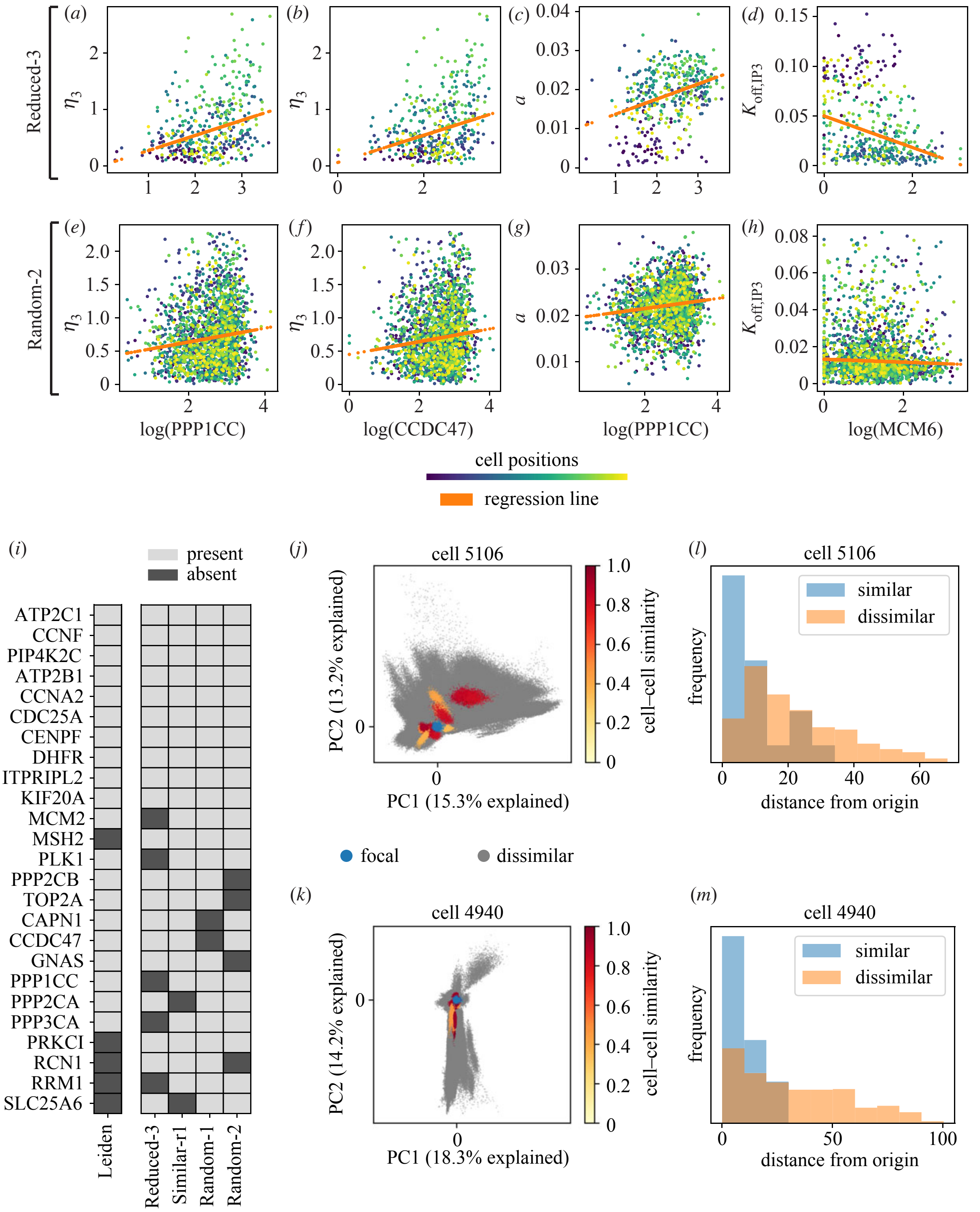

Figure 4. Variability in Ca^{2+} model dynamics is associated with variability in gene expression. (*a–d*) Top ranked gene–parameter pairs by Pearson correlation from similarity-based chain (*Reduced-3*). Log gene expression is plotted against marginal posterior means. Each dot represents a cell; colour denotes cell position along the chain. (*e–h*) The same gene–parameter pairs shown in (*a*), for cells sampled from a randomly ordered cell chain (*Random-2*). (*i*) Comparison of genes identified as present/absent in the top 30 genes from gene–parameter correlation analysis across different chains, and the top 30 marker genes from Leiden clustering directly on the gene expression space. (*j*) Projection of posterior samples (500 per cell) onto the first two principal components of the focal cell (Cell 5106), shown in blue. Posterior samples from dissimilar cells shown in grey; posterior samples from similar cells shown in yellow-red, where the shade indicates degree of transcriptional similarity. (*k*) As for (*j*), with PCA performed on a different focal cell (Cell 4940). (*l*) Histogram corresponding to (*j*): mean distances from the origin (focal cell 5106) to samples from surrounding cells. (*m*) Histogram corresponding to (*k*): mean distances from the origin (focal cell 4940) to samples from surrounding cells.

are coloured based on gene expression: samples are derived from cells that are either transcriptionally similar to the focal cell, or share no transcriptional similarity. Comparison of similar and dissimilar cells from the same population showed that cells that were transcriptionally similar were located closer to the focal cell than dissimilar cells (figure 4*l*,*m* and electronic supplementary material, figure S8*c*,*d*). By contrast, similar analysis of a random cell chain showed that transcriptionally similar cells were not located closer to the focal cell than dissimilar cells (electronic supplementary material, figure S9). Notably, proximity of posterior samples derived from transcriptionally similar cells was not driven by a cell's position along the chain (no block structure observed; electronic supplementary material, figure S10). Similarities between posterior distributions of transcriptionally similar cells were thus not driven by local cell–cell similarity, but rather underlie a global effect and denote a relationship between the transcriptional states of cells and the Ca^{2+} pathway dynamics that they produce.

3.5. Similarity-based posterior cell clustering reveals distinct transcriptional states underlying Ca^{2+} dynamics

To characterize the extent to which we can predict Ca^{2+} responses from knowledge of the model dynamics, we clustered 500 cells from a similarity-based cell chain (*Reduced-3*) based on the single-cell posterior distributions using hierarchical clustering (see Material and methods). Three clusters were obtained (figure 5*a*). Each cluster showed distinct Ca^{2+} dynamics: 'low-responders' exhibited lower overall Ca^{2+} peaks in response to ATP (figure 5*b*); 'early-responders' exhibited earlier overall Ca^{2+} peaks in response to ATP; and 'late-high-responders' exhibited robust Ca^{2+} responses with peaks that were later and higher than cells from other clusters (electronic supplementary material, figure S11). The distinct dynamic profiles can be explained by the model parameters that give rise to them: low-responders were characterized by high concentration of free Ca^{2+} in the ER (c_0) and low activation rates of IP$_3$R (electronic supplementary material, figures S11–S13). Early-responders were characterized by parameters leading to faster and earlier IP3 and PLC dynamics. Late-high-responders were characterized by small values of d_1 (electronic supplementary material, figure S13).

Comparison of posterior parameter clustering with the clustering done by Yao *et al.* [38] highlights similarities and distinctions. In both cases, one of the three clusters was characterized by larger responses to ATP and correspondingly higher values of d_{inh} (electronic supplementary material, figure S13). In Yao *et al.*, both d_1 and d_5 were smaller in cells with stronger Ca^{2+} responses; we found that d_1 was smaller in the late-high-responder cluster, but not in the early responders. In our results, d_5 was higher for the early-responders, in contrast with Yao *et al.* (electronic supplementary material, figure S13). We note that we set a stringent threshold for minimum peak Ca^{2+} response, i.e. we excluded non-responding cells, unlike Yao *et al.*, thus in a direct comparison most of the cells in our population would belong to the 'strong positive' cluster in [38].

To assess the Ca^{2+} dynamic clusters we obtained in light of single-cell gene expression, we performed two analyses for comparison. We clustered the same 500 cells based solely on their gene expression using community detection (Leiden algorithm in Scanpy [46]); and we clustered cells from a randomly ordered cell chain using the same approach as above for hierarchical clustering of posterior parameters. The cell clusters obtained based solely on gene expression can be distinguished based on the Ca^{2+} profiles observed: 'Ca-low', 'Ca-mid' and 'Ca-high' responses (figure 5*c*); overlapping partially with the similarity-based clusters obtained (figure 5*b*). By contrast, no distinct Ca^{2+} dynamic responses could be observed for the posterior clustering based on the random cell chain (figure 5*d* and electronic supplementary material, figure S14).

We analysed differential gene expression from each set of clusters obtained, from the similarity-based chain (figure 5*e*), the gene expression-based clustering (figure 5*f*) and the randomly ordered chain (figure 5*g*). Distinct markers for each cluster were observed for the similarity-based clustering and the gene expression-based clustering, but were not discernible for cell clustering on the random chain. Clustering of cell posteriors from the randomly ordered chain was thus unable to distinguish Ca^{2+} dynamic profiles nor gene expression differences. On the other hand, clustering posteriors from a similarity-based chain identified distinct gene expression profiles, and these overlapped with the marker gene profiles obtained by clustering on the gene expression directly. That is, parameter inference of single-cell Ca^{2+} dynamics from using a gene expression similarity-based chain enables the identification of cell clusters with distinct transcriptional profiles and distinct responses to ATP stimulation.

Analysis of the genes that are associated with each Ca^{2+} profile showed that low-responder cells were characterized by upregulation of CCDC47 and PP1 family genes (PPP1CC and PPP2CA). Early-responder cells were characterized by upregulation of CAPN1 and CHP1, among others. The late-high responder cells were characterized by increased expression CALM3 among others, although the marker genes for this cluster were less evident than the others. By comparison of marker genes, we saw considerable overlap between the early-responders (similarity-based clustering) and the Ca-mid responders (gene expression clustering). We also saw overlap between the low-responder and the Ca-low marker gene signatures. These results highlight that the posterior distributions of cells fit from similarity-based cell chains captured information about the underlying transcriptional states of the cells, linking cellular response parameters directly to transcriptional states. For example, the low-responder cells (similarity-based clustering) were distinguished by parameter d_5, characterizing the dynamics of IP3 dissociation, and were marked by high PPP1CC and CCDC47 expression.

Finally, we considered whether alternative means for cell chain construction could provide similar information. We constructed a cell chain using cells from *Reduced-3*, denoted '*Ca-similarity*' for which consecutive cells displayed similar Ca^{2+} responses (see electronic supplementary material, section S1.3). Clustering cells from *Ca-similarity* based on their Ca^{2+} responses (via *k*-means) showed that cells with different Ca^{2+} responses had distinct gene expression profiles (electronic supplementary material, figure S15). However, hierarchical clustering on the parameter posterior distributions from these cells after performing inference on *Ca-similarity* did not separate cells into clusters with distinct Ca^{2+} responses or distinct gene expression profiles (electronic supplementary material, figure S16*a,b*). This result makes sense when analysed via the similarity matrices obtained for *Ca-similarity* versus a gene expression similarity-based chain (electronic supplementary material, figure S16*c,d*). For the former, almost all pairs of neighbouring cells did not share similarity in gene expression despite their similarity in Ca^{2+} response. Whereas for *Reduced-3*, the gene expression similarity-based chain, all pairs of consecutive cells were similar in gene expression (electronic supplementary material, figure S16*d*).

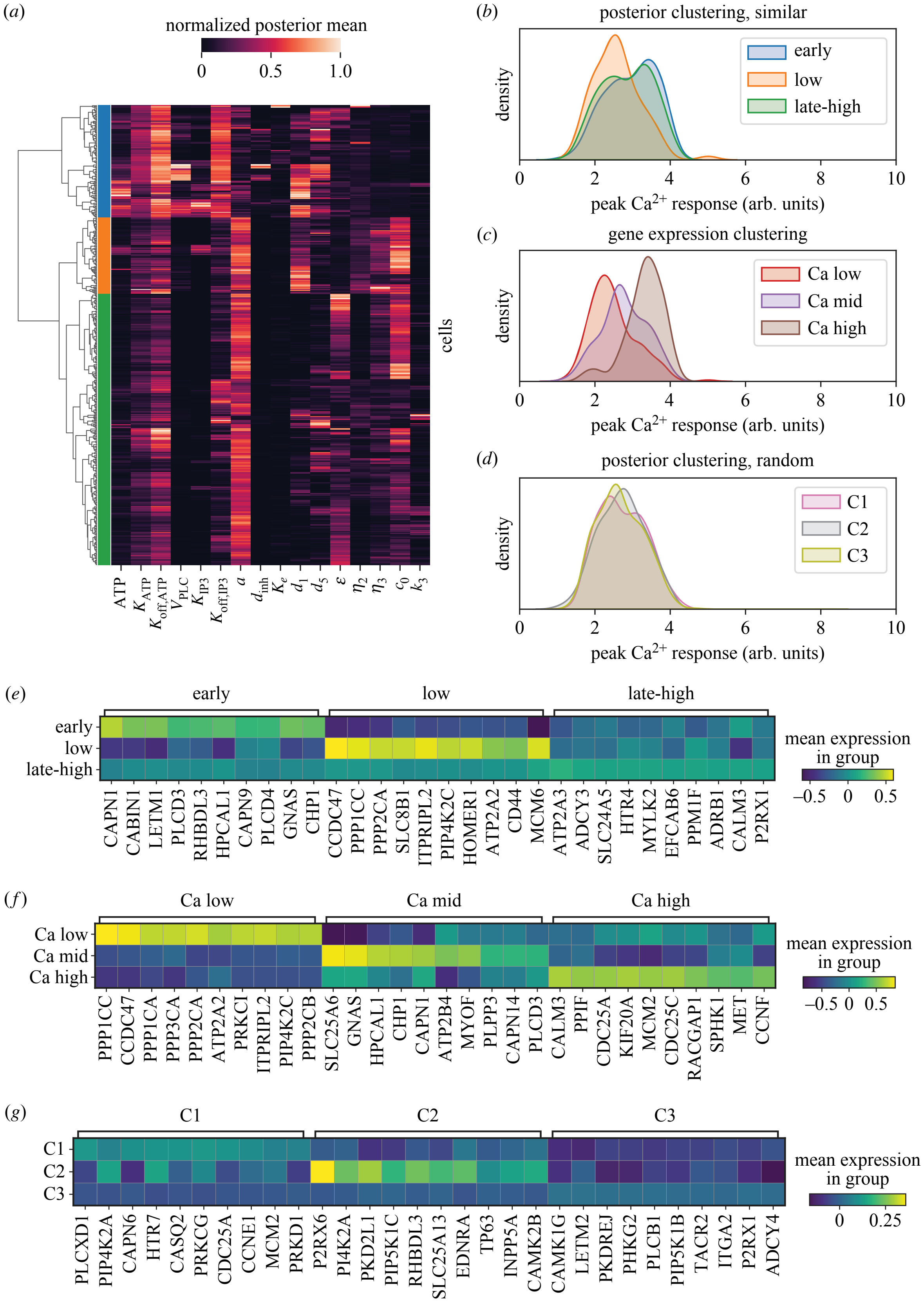

Figure 5. Clustering of cell posterior distributions reveals marker genes for Ca^{2+} states. (*a*) Agglomerative clustering on posterior means from a similarity-based chain (*Reduced-3*) using Ward linkage ($k = 3$ clusters). (*b*) Kernel density estimate of Ca^{2+} peak height from posterior clustering of a similarity-based chain. (*c*) Kernel density estimate of Ca^{2+} peak height from gene expression clustering (Leiden) for the same cell population as shown in (*a*–*b*). (*d*) Kernel density estimate of Ca^{2+} peak height from posterior clustering of a randomly ordered cell chain (*Random-2*). (*e*) Top-10 marker genes per cluster from similarity-based posterior clustering. (*f*) Top-10 marker genes per cluster from gene expression clustering of cells. (*g*) Top-10 marker genes per cluster from posterior clustering of randomly ordered cell chain.

4. Discussion

We have presented methods for inferring the parameters of a signalling pathway model, given data describing dynamics in single cells coupled with subsequent gene expression profiling. We hypothesized that via transfer learning we could use posterior information from a cell to inform the prior distribution of its neighbour along a ‘cell chain’ of transcriptionally similar cells. Implemented using Hamiltonian Monte Carlo sampling [42], we discovered that the cell chain construction for prior distributions did indeed lead to faster sampling of parameters. However, these improvements did not rely on the use of gene expression to construct priors: the performance of inference on cells in a chain ordered randomly was equally good. However, cell chains constructed using gene expression similarity contained more information (as defined by their posterior parameter distributions) about Ca^{2+} signalling responses. Clustering the posterior parameters identified important relationships between gene expression and the dynamic Ca^{2+} phenotypes, thus providing mappings from state to dynamic cell fate.

The model studied here is described by ODEs to characterize the Ca^{2+} signalling pathway, adapted from [37,40], consisting of 12 variables and (originally) a 40-dimensional parameter space. This was reduced to 19 parameters in Yao *et al.* [38] and 16 parameters in our work. Analysis of even a single 16-dimensional posterior distribution requires dimensionality reduction techniques, let alone the analysis of the posterior distributions obtained for populations of hundreds of single cells. Parameter sensitivity analysis highlighted the effects of specific parameter perturbations on the Ca^{2+} dynamic responses. Indeed, we advocate for the use of sensitivity analysis more generally as means to distinguish and pinpoint the effects of different parameter combinations for models of complex biochemical signalling pathways.

By unsupervised clustering of the posterior distributions, we found that distinct patterns of Ca^{2+} in response to ATP could be mapped to specific variation in the single-cell gene expression. In previous work using similar approaches for clustering [38], posterior parameter clusters predominantly revealed response patterns consisting of responders and non-responders; here we excluded those cells that did not exhibit a robust response to ATP. We were able to characterize subtler the Ca^{2+} response dynamics (described by 'early', 'low', and 'late-high' responders) and predict which transcriptional states give rise to each. This approach is limited since relatively little gene expression variance is explained by an individual model parameter: it may be possible to address this in future work by surveying a larger range of cell behaviours, e.g. by including a wider range of cellular responses or by considering higher-level covariance in the posterior parameter space. It also remains to be tested whether the given model of Ca^{2+} dynamics is appropriate to describe the signalling responses in cell types other than MCF10A cells.

Our ability to fit to the single cells tested came potentially at the expense of an unwieldy model size. With four variables and a 16-dimensional parameter space, the dimension of the model far exceeds that of the data: time series of Ca^{2+} responses in single cells. Without data with which to constrain the three additional model species, we needed to constrain the model in another way. We used an approach of 'scaling and clipping' for construction of the priors, i.e. setting ad hoc limits to control posterior variance. More effective (and less ad hoc) techniques could improve inference overall and may become necessary in the case of larger models. These include (in order of sophistication): tailoring the scaling/clipping choices to be parameter-specific; tailoring the choice of prior variance based on additional sources of data; or performing model reduction/identifiability analysis to further constrain the prior space before inference. Constructing priors from cells with similar gene expression also helped to curb the curse of dimensionality: sampling cells sequentially places a constraint on the model. Nonetheless, in the future more directed approaches to tackle model identifiability ought to be considered.

Connecting dynamic cell phenotypes to transcriptional states remains a grand challenge in systems biology. The limitations of deriving knowledge from gene expression data alone [24] have led to the proposal of new methods seeking to bridge the gap between states and fates [52]. Here, making use of technology that jointly measures Ca^{2+} dynamics and gene expression in single cells, we have shown that parameter inference informed by transcriptional similarity represents another way by which we can connect gene expression states to dynamic cellular phenotypes. The statistical framework employed improved our ability to perform parameter inference for single-cell models. We expect that improvements to statistical inference frameworks could be gained by similar approaches applied to other models of nonlinear cellular response dynamics. Current and future technologies combining higher-resolution dynamic cell measurements with high-throughput genomics will provide new data to inform these parameter inference methods and, we expect, lead to new discoveries of means of transcriptional control of biological dynamics.

Data accessibility. All code developed to simulate models and run parameter inference is released under an MIT license at: https://github.com/maclean-lab/singlecell-parinf. Parameter inference was developed in Python 3.6 and Stan 2.19. Posterior analyses were developed in Python 3.8. Ca^{2+} data and MERFISH data (processed files) are also available from the GitHub repository: https://github.com/maclean-lab/singlecell-parinf.

The data are provided in electronic supplementary material [53].

Authors' contributions. X.W.: conceptualization, investigation, methodology, software, writing—original draft, writing—review and editing; R.W.: conceptualization, investigation, writing—review and editing; A.L.M.: conceptualization, investigation, methodology, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. R.W. is a co-founder and equity holder of a BioCartography inc.

Funding. This work was supported by an Andrew J. Viterbi Fellowship in Computational Biology and Bioinformatics (to X.W.). A.L.M. was supported by the National Institutes of Health (grant no. R35GM143019) and the National Science Foundation (grant no. DMS2045327).

Acknowledgements. A.L.M. acknowledges support from the National Institutes of Health (R35GM143019) and the National Science Foundation (DMS2045327).

References

1. Wilkinson DJ. 2020 *Stochastic modelling for systems biology*. Chapman & Hall/CRC mathematical and computational biology, 3rd edn, first issued in paperback edition. Boca Raton, FL: CRC Press.
2. Hoops S *et al.* 2006 COPASI—a complex pathway simulator. *Bioinformatics* **22**, 3067–3074. (doi:10.1093/bioinformatics/btl485)
3. Liepe J, Filippi S, Komorowski M, Stumpf MPH. 2013 Maximizing the information content of experiments in systems biology. *PLoS Comput. Biol.* **9**, e1002888. (doi:10.1371/journal.pcbi.1002888)
4. Warne DJ, Baker RE, Simpson MJ. 2019 Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-

the-art. *J. R. Soc. Interface* **16**, 20180943. (doi:10.1098/rsif.2018.0943)

5. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202. (doi:10.1098/rsif.2008.0172)
6. Wagner A, Regev A, Yosef N. 2016 Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160. (doi:10.1038/nbt.3711)
7. Munsky B, Neuert G, van Oudenaarden A. 2012 Using gene expression noise to understand gene regulation. *Science* **336**, 183–187. (doi:10.1126/science.1216379)
8. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. 2011 Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474. (doi:10.1126/science.1198817)
9. Raj A, van Oudenaarden A. 2008 Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226. (doi:10.1016/j.cell.2008.09.050)
10. Foreman R, Wollman R. 2020 Mammalian gene expression variability is explained by underlying cell state. *Mol. Syst. Biol.* **16**, e9146. (doi:10.15252/msb.20199146)
11. Brodskiy PA, Zartman JJ. 2018 Calcium as a signal integrator in developing epithelial tissues. *Phys. Biol.* **15**, 051001. (doi:10.1088/1478-3975/aabb18)
12. Leiper LJ, Walczysko P, Kucerova R, Jingxing O, Shanley LJ, Lawson D, Forrester JV, McCaig CD, Zhao M, Collinson JM. 2006 The roles of calcium signaling and ERK1/2 phosphorylation in a Pax6+/−mouse model of epithelial wound-healing delay. *BMC Biol.* **4**, 27. (doi:10.1186/1741-7007-4-27)
13. Malèth J, Hegyi P. 2014 Calcium signaling in pancreatic ductal epithelial cells: an old friend and a nasty enemy. *Cell Calcium* **55**, 337–345. (doi:10.1016/j.ceca.2014.02.004)
14. Dupont G, Combettes L, Bird GS, Putney JW. 2011 Calcium oscillations. *Cold Spring Harbor Perspect. Biol.* **3**, a004226–a004226. (doi:10.1101/cshperspect.a004226)
15. Periasamy M, Kalyanasundaram A. 2007 SERCA pump isoforms: their role in calcium transport and disease. *Muscle Nerve* **35**, 430–442. (doi:10.1002/mus.20745)
16. Ma LH, Webb SE, Chan CM, Zhang J, Miller AL. 2009 Establishment of a transitory dorsal-biased window of localized Ca^{2+} signaling in the superficial epithelium following the mid-blastula transition in zebrafish embryos. *Dev. Biol.* **327**, 143–157. (doi:10.1016/j.ydbio.2008.12.015)
17. Balaji R, Bielmeier C, Harz H, Bates J, Stadler C, Hildebrand A, Classen A-K. 2017 Calcium spikes, waves and oscillations in a large, patterned epithelial tissue. *Sci. Rep.* **7**, 42786. (doi:10.1038/srep42786)
18. Domínguez DC, Guragain M, Patrauchan M. 2015 Calcium binding proteins and calcium signaling in prokaryotes. *Cell Calcium* **57**, 151–165. (doi:10.1016/j.ceca.2014.12.006)
19. MacLean AL, Hong T, Nie Q. 2018 Exploring intermediate cell states through the lens of single cells. *Curr. Opin. Syst. Biol.* **9**, 32–41. (doi:10.1016/j.coisb.2018.02.009)
20. Moffitt JR, Zhuang X. 2016 RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). In *Methods in enzymology* (eds Filonov GS, Jaffrey SR), vol. 572, pp. 1–49. Amsterdam, The Netherlands: Elsevier.
21. Eng C-HL *et al.* 2019 Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239. (doi:10.1038/s41586-019-1049-y)
22. Strogatz SH. 2015 *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*, 2nd edn. Boulder, CO: Westview Press, a member of the Perseus Books Group.
23. Chang AY, Marshall WF. 2019 Dynamics of living cells in a cytomorphological state space. *Proc. Natl Acad. Sci. USA* **116**, 21 556–21 562. (doi:10.1073/pnas.1902849116)
24. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. 2018 Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476. (doi:10.1073/pnas.1714723115)
25. Kuzmanovska I, Milias-Argeitis A, Mikelson J, Zechner C, Khammash M. 2017 Parameter inference for stochastic single-cell dynamics from lineage tree data. *BMC Syst. Biol.* **11**, 52. (doi:10.1186/s12918-017-0425-1)
26. Ruske LJ, Kursawe J, Tsakiridis A, Wilson V, Fletcher AG, Blythe RA, Schumacher LJ. 2020 Coupled differentiation and division of embryonic stem cells inferred from clonal snapshots. *Phys. Biol.* **17**, 065009. (doi:10.1088/1478-3975/aba041)
27. Stumpf PS *et al.* 2017 Stem cell differentiation as a non-Markov stochastic process. *Cell Syst.* **5**, 268–282.e7. (doi:10.1016/j.cels.2017.08.009)
28. Sha Y, Wang S, Zhou P, Nie Q. 2020 Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data. *Nucleic Acids Res.* **48**, 9505–9520. (doi:10.1093/nar/gkaa725)
29. Manning CS *et al.* 2019 Quantitative single-cell live imaging links HES5 dynamics with cell-state and fate in murine neurogenesis. *Nat. Commun.* **10**, 2835. (doi:10.1038/s41467-019-10734-8)
30. Burton J, Manning CS, Rattray M, Papalopulu N, Kursawe J. 2021 Inferring kinetic parameters of oscillatory gene regulation from single cell time-series data. *J. R. Soc. Interface* **18**, 20210393. (doi:10.1098/rsif.2021.0393)
31. Dharmarajan L, Kaltenbach H-M, Rudolf F, Stelling J. 2019 A simple and flexible computational framework for inferring sources of heterogeneity from single-cell dynamics. *Cell Syst.* **8**, 15–26.e11. (doi:10.1016/j.cels.2018.12.007)
32. Persson S, Welkenhuysen N, Shashkova S, Wiqvist S, Reith P, Schmidt GW, Picchini U, Cvijovic M. 2022 Scalable and flexible inference framework for stochastic dynamic single-cell models. *PLoS Comput. Biol.* **18**, e1010082. (doi:10.1371/journal.pcbi.1010082)
33. Loos C, Moeller K, Fröhlich F, Hucho T, Hasenauer J. 2018 A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability. *Cell Syst.* **6**, 593–603.e13. (doi:10.1016/j.cels.2018.04.008)
34. Cang Z, Nie Q. 2020 Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084. (doi:10.1038/s41467-020-15968-5)
35. Öcal K, Gutmann MU, Sanguinetti G, Grima R. 2022 Inference and uncertainty quantification of stochastic gene expression via synthetic models. *J. R. Soc. Interface* **19**, 20220153. (doi:10.1098/rsif.2022.0153)
36. Jiang Q, Fu X, Yan S, Li R, Du W, Cao Z, Qian F, Grima R. 2021 Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun.* **12**, 2618. (doi:10.1038/s41467-021-22919-1)
37. Lemon G, Gibson WG, Bennett MR. 2003 Metabotropic receptor activation, desensitization and sequestration—I: modelling calcium and inositol 1,4,5-trisphosphate dynamics following receptor activation. *J. Theor. Biol.* **223**, 93–111. (doi:10.1016/S0022-5193(03)00079-1)
38. Yao J, Pilko A, Wollman R. 2016 Distinct cellular states determine calcium signaling response. *Mol. Syst. Biol.* **12**, 894. (doi:10.15252/msb.20167137)
39. Carpenter B *et al.* 2017 *Stan*: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32. (doi:10.18637/jss.v076.i01)
40. J Rinzel Y-XL. 1994 Equations for InsP3 Receptor-mediated $[Ca^{2+}]i$ oscillations derived from a detailed kinetic model: a Hodgkin-Huxley like formalism. *J. Theor. Biol.* **166**, 461–473. (doi:10.1006/jtbi.1994.1041)
41. Wang S, Karikomi M, MacLean AL, Nie Q. 2019 Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* **47**, e66–e66. (doi:10.1093/nar/gkz204)
42. Hoffman MD, Gelman A. 2014 The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623.
43. Betancourt M. 2018 A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*. See http://arxiv.org/abs/1701.02434.
44. Kumar R, Carroll C, Hartikainen A, Martin O. 2019 ArviZ a unified library for exploratory analysis of Bayesian models in Python. *J. Open Sourc. Softw.* **4**, 1143. (doi:10.21105/joss.01143)
45. Pedregosa F *et al.* 2011 Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
46. Traag VA, Waltman L, van Eck NJ. 2019 From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233. (doi:10.1038/s41598-019-41695-z)
47. Wolf FA, Angerer P, Theis FJ. 2018 SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15. (doi:10.1186/s13059-017-1382-0)
48. Virtanen P *et al.* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python.

Nat. Methods **17**, 261–272. (doi:10.1038/s41592-019-0686-2)

49. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. 2007 Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, e189. (doi:10.1371/journal.pcbi.0030189)

50. Varma A, Morbidelli M, Wu H. 1999 *Parametric sensitivity in chemical systems*, 1st edn. Cambridge, UK: Cambridge University Press.

51. Komorowski M, Costa MJ, Rand DA, Stumpf MPH. 2011 Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl Acad. Sci. USA* **108**, 8645–8650. (doi:10.1073/pnas.1015814108)

52. Qiu X *et al.* 2022 Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690–711. (doi:10.1016/j.cell.2021.12.045)

53. Wu X, Wollman R, MacLean AL. 2023 Single-cell Ca^{2+} parameter inference reveals how transcriptional states inform dynamic cell responses. Figshare. (doi:10.6084/m9.figshare.c.6662626)