Supplementary Figures

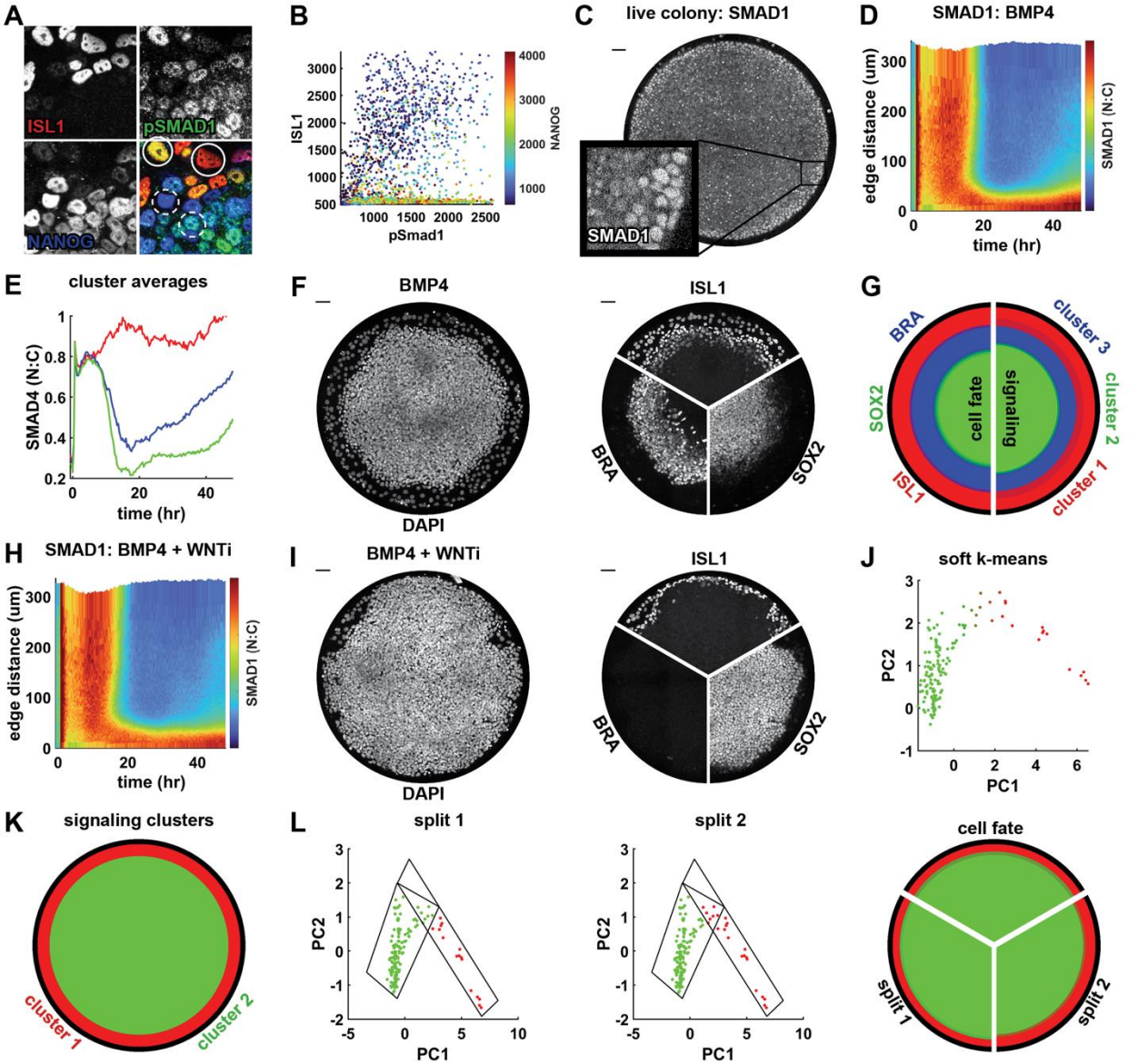

Figure 1 supplement: (A-B) detail of staining and quantification for pSMAD1, ISL1, and NANOG after BMP treatment in the presence of Wnt inhibitor (IWP2) showing low correlation between fate and final BMP signaling levels in a micropatterned colony. Solid circles indicate two amnion-like cells with high and low signaling levels respectively, and dashed circles similarly indicate two pluripotent cells at different signaling levels. (**C**) A representative micropatterned colony of RUES2 cells expressing RFP::SMAD1 at t = 30 hours after treatment with BMP4, showing nuclear localization of SMAD1 specifically at the colony edge. (**D**) Kymograph of mean SMAD1 signaling in N=5 micropatterned colonies treated with BMP4. (**E**) Mean signaling within the clusters of signaling histories in fig 1H. (**F**) Separated channel images showing the DAPI, ISL1, SOX2, and BRA stains corresponding to the colonies shown in Fig. 1J. (**G**) Comparison of the average profile of cell fate markers (left) and clusters of signaling histories (right) for BMP4-treated colonies. **(H)** kymograph showing average spatiotemporal dynamics of SMAD1 in micropatterned colonies treated with BMP4 and WNTi (IWP2). (**I**) Separated channel images showing the DAPI, ISL1, SOX2, and BRA stains corresponding to the colonies shown in Fig. 1N.**(J)** Scatterplot of the first two PCs of radially averaged signaling histories, colored for cluster assignment as in Fig 1G. **(K)** Predicted fate map based on the clustering in (J), determined as in Fig 1I. (**L**) Example of two different ways to assign the signaling histories corresponding to the 'elbow' of the PCA plot in colonies treated with BMP4 + WNTi, along with the resulting radial profiles for each assignment, compared to the profile of ISL1 and SOX2 expression in those colonies. Scale bars 50um.

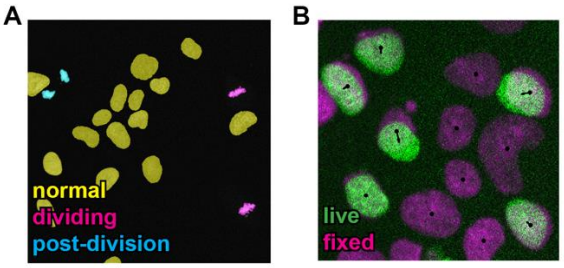

Figure 2 supplement: (A) Example image of nuclei with overlaid classification of cells as non-dividing (yellow), dividing (magenta), and immediately post-division (cyan). **(B)** Linking live to fixed cells. Image of live nuclei is shown in green, and fixed nuclei in magenta. Black arrows show links from the centroids of fixed nuclei to centroids of live nuclei.

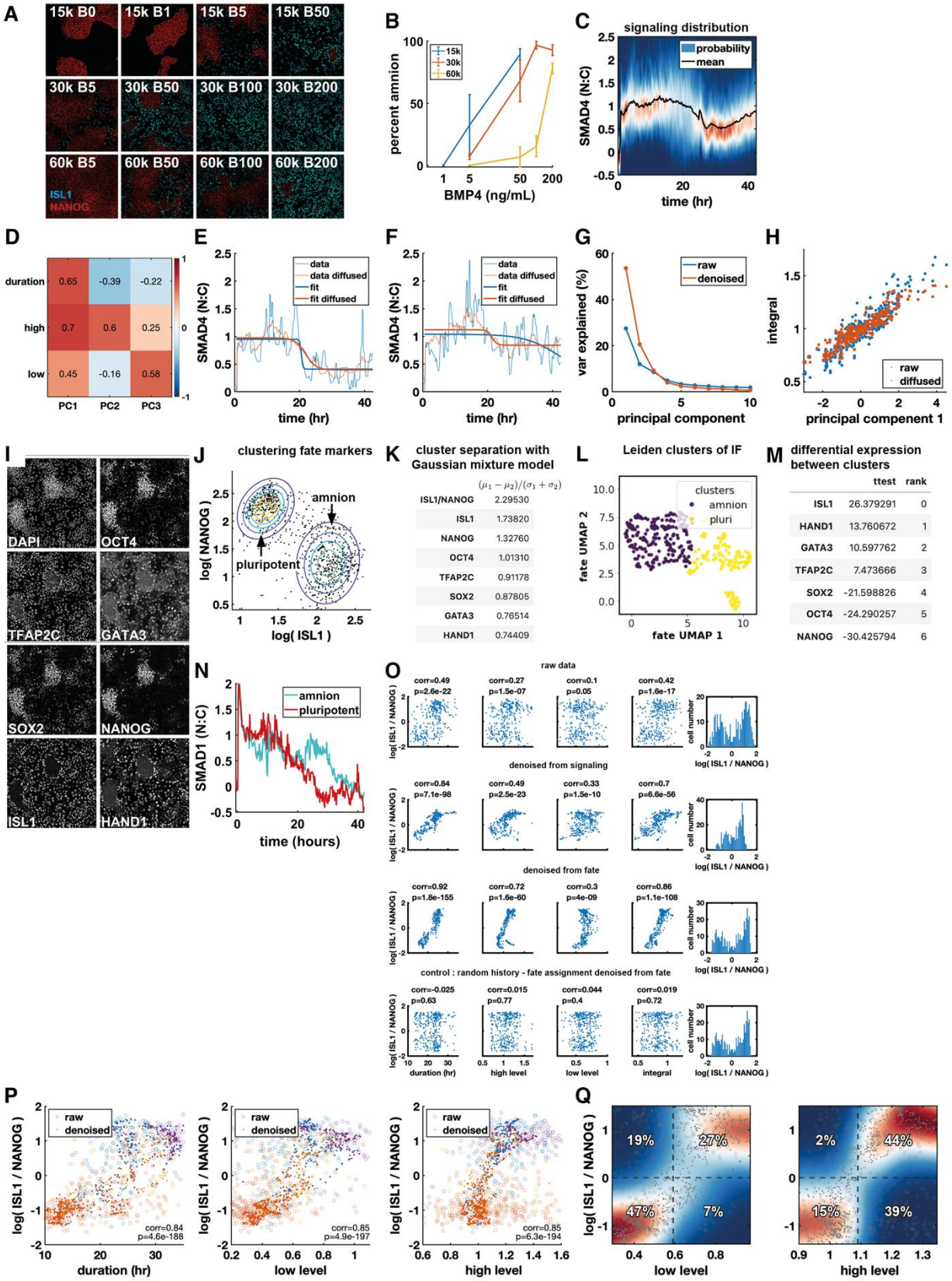

Figure 3 supplement: (**A**) immunofluorescence staining for NANOG and ISL1 for different conditions, first number is density, .e.g 15k = 15,000 cells / cm^2, second number is BMP4 dose, e.g. B1 = 1ng/ml BMP4. **(B)** Quantification of (A). **(C)** Heatmap plot of signaling distribution over time corresponding to Fig. 3AB. **(D)** Correlation of features and principal components before denoising. **(E-F)** Example signaling histories before and after denoising via data

diffusion with MAGIC, with sigmoid fits to the raw and denoised data. (**G**) Scatterplot of signal integral against principal component 1, with and without denoising. **(H)** Variance explained in signaling distribution from (C) by the first 10 PCs for raw and denoised signaling histories. (**I**) Representative single-channel IF images showing expression of all 7 stained genes in the same field of view. **(J)** Contour plot of a two-component Gaussian mixture model fit to fate marker expression, overlaid on a scatterplot of ISL1 vs. NANOG. (**K**) Table of values of a measure of cluster separation. The marginal distribution of the 7D Gaussian mixture model (GMM) is taken along each axis indicated and the separation of clusters along that direction is taken as the ratio of the difference in the means of the two GMM components to the sum of their standard deviations. A higher value indicates better separation. (**L**) UMAP plot showing the separation of cells into two clusters with Leiden clustering. (**M**) Table of differential expression of each marker between the Leiden clusters, showing highest absolute value for ISL1 and NANOG. (**N**) mean RFP::SMAD1 signaling in amnion and pluripotent cells. (**O**) Scatterplots of log(ISL1 / NANOG) and signaling features under various denoising schemes for data in Fig. 3A-M. **(P)** Scatter plots of signaling features vs. log(ISL1/NANOG) colored for condition with and without denoising for data in Fig. 3K-P. **(Q)** Heatmap of kernel density estimate after denoising of conditional distributions of log(ISL1 / NANOG) with respect to low level and high level of signaling, overlaid with a scatterplots of data points before (circles) and after denoising (dots). Dashed lines show separation of cells into amnion-like and pluripotent based on log(ISL1 / NANOG) or on signaling features. The percentage of cells in each quadrant is indicated, with correct assignments in the top right and bottom left quadrant of each heatmap.

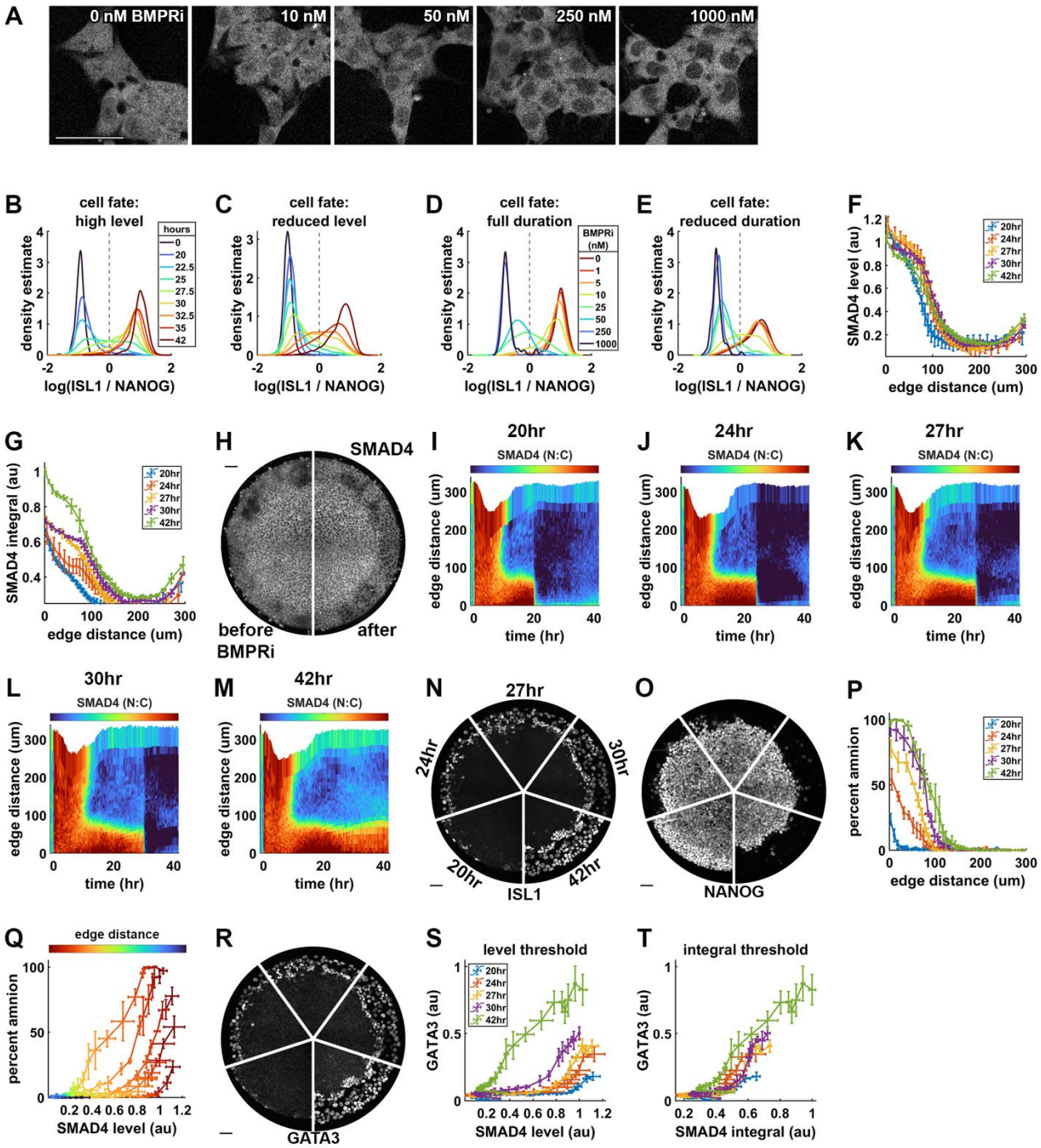

Figure 4 supplement: **(A)** GFP::SMAD4 images corresponding to data in Fig. 4H, showing sparsely seeded cells treated with 50ng/ml BMP4 and different doses of LDN193189 (BMPRi). (**B**) Kernel density estimate (KDE) of the log(ISL1/NANOG) distribution in each condition shown in 4D. (**C**) Kernel density estimate (KDE) of the log(ISL1/NANOG) distribution in each condition shown in 4E. Legend in (B). (**D**) KDE of the log(ISL1 / NANOG) distribution after 42h of differentiation for conditions in 4H. (**E**) KDE of the log(ISL1 / NANOG) distribution after 42h of differentiation for conditions in 4I. Legend in (D). **(F)** Average level of SMAD4 signaling before BMP inhibition as a function of distance from the colony edge for different durations of BMP signaling. **(G)** Integral of SMAD4 signaling at 42h as a function of distance from the colony edge for different durations of BMP signaling. (**H**) GFP::SMAD4 for a BMP treated micropatterned colony treated with 200ng/ml BMP4 shown before (29h, left) and after BMP signaling inhibition (31h, right). **(I-M)** Kymographs of average SMAD4 signaling in N=3 micropatterned colonies each for five signaling durations. **(N-O)** Representative IF images showing the spatial extent of ISL1 and NANOG expression in micropatterned colonies exposed to different durations of BMP signaling. **(P)** Quantification of percentage of

differentiated cells (ISL1+NANOG-) as a function of distance from the colony edge for different durations of BMP signaling. **(Q)** Percent amnion differentiation vs. level of BMP signaling as in Fig. 4P, colored for distance from the colony edge. **(R)** Representative IF images showing the spatial extent of GATA3 expression in micropatterned colonies exposed to different durations of BMP signaling. **(S)** GATA3 expression in radial bins as a function of SMAD4 level before removal of BMP4 for each signal duration. Error bars are standard deviation over the same radial bin in N = 3 replicate colonies. **(T)** GATA3 expression in radial bins as a function of total SMAD4 integral. Error bars are as in E. Scale bars 50um.

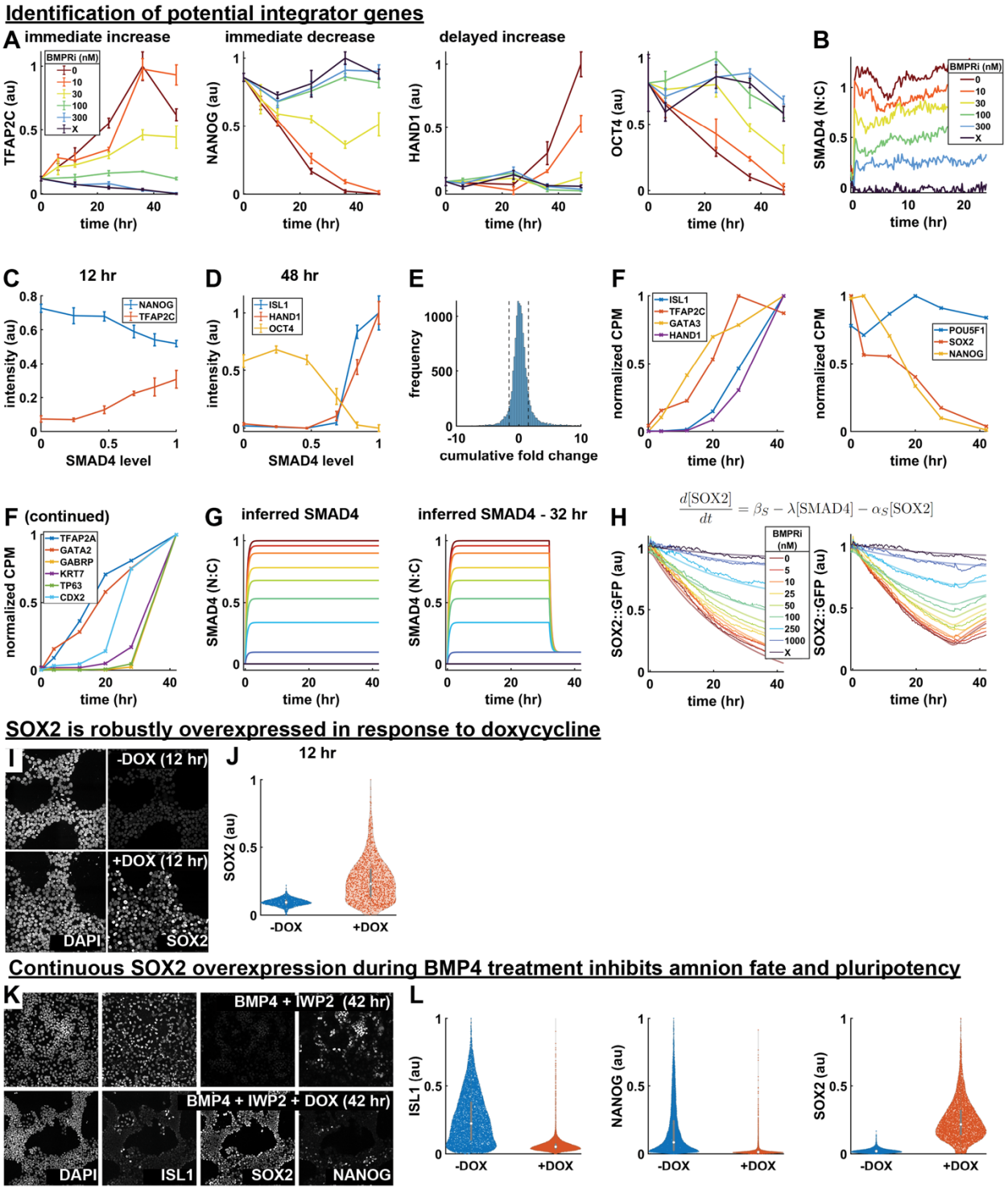

Figure 5 supplement: (**A**) Normalized expression of TFAP2C, NANOG, HAND1, and OCT4 over time for different signaling levels, measured with time-series IF. Error bars are standard deviation across N = 6 replicate images. (**B**) Average SMAD4 dynamics measured in the treatment conditions for which time-series IF was performed. (**C**) NANOG and TFAP2C expression at 12 hours, plotted against SMAD4 signaling level. (**D**) ISL1, HAND1, and OCT4 expression at 48 hours, plotted against SMAD4 signaling level, showing switch-like reliance. (**E**) Histogram of the cumulative log2 fold change over all genes in the time-series bulk RNA seq data. Genes with a fold change between the two dotted lines were not included in hierarchical clustering or subsequent analysis. (**F**) Expression over time for example genes measured with bulk RNA seq. Expression of amnion (left) and pluripotency (middle) genes that were also measured with time series IF, and additional amnion genes (right). (**G**) Idealized SMAD4 dynamics used as input to the ODE model, with level inferred from data in fig 4D. (**H**) GFP::SOX2 dynamics over the course of 42 hours of differentiation

with indicated treatments applied for 42 (left) or 32 (right) hours, overlaid with fits of the simple ODE model described by the equation above it for SOX2, and the equation for ISL1 as in Fig. 4K. (**I**) Representative IF images showing expression of SOX2 and NANOG after 12 hours in pluripotency conditions with or without doxycycline. (**J**) Violin plot of SOX2 expression after 12 hours with or without doxycycline in pluripotency conditions. (**L**) Representative IF images showing ISL1, SOX2, and NANOG expression after 42 hours of treatment with BMP4 + WNTi with or without doxycycline. (**M**) Violin plots of ISL1, NANOG, and SOX2 expression with or without doxycycline in differentiation conditions.

Algorithm for automated single-cell tracking

Seth Teague, Idse Heemskerk

March 29, 2023

Algorithm development

To construct tracks of single cells in time-lapse live-cell microscopy data, we took a "tracking by detection" approach (Magnusson et al. 2015), dividing the problem into two steps: (1) segmentation (detection) of all cells in each frame of the time-lapse, and (2) building tracks by linking segmented cells frame-to-frame. A custom single-cell tracking algorithm, based on the approach to particle tracking proposed in (Jaqaman et al. 2008), and similar to the implementation in the popular Fiji plugin Trackmate (Tinevez et al. 2017), was written in MATLAB and integrated into the image processing pipeline.

The approach taken in (Jaqaman et al. 2008) is to find an approximately optimal solution globally by breaking the tracking problem into two steps. Following this approach, we first link cells one-to-one or one-to-none in consecutive frames, assigning zero or one links from each cell in one frame to cells in the subsequent frame. This is followed by a "gap-closing, merging, splitting" (GMS) step, which addresses common segmentation and linking errors. Gap-closing connects the end of a track in frame t_1 to the beginning of a track in frame $t_2 > t_1 + 1$, and is intended to account for nuclei leaving and re-entering the frame or that fail to be segmented in one or more frames. Merging connects the end of a track to the middle of another track, and accounts for two nuclei in frame t being segmented as a single nucleus in frame $t + 1$. Conversely, splitting connects the beginning of a track in frame t to the middle of a track in a previous frame, and accounts for two nuclei in frame t being segmented as a single nucleus in frame $t - 1$ or to a cell dividing in frame $t - 1$. These two steps are each cast as a linear assignment problem (LAP), in which a cost is assigned to each possible assignment and the globally optimal solution of the LAP minimizes the sum of possible costs. Fast algorithms have been developed to find the optimal solution for a given cost matrix, so the essential problem is to determine an effective way to assign costs to possible assignments, generally based on the proximity and morphological similarity of nuclei to be linked.

In addition to the general difficulty of robustly tracking through a time-lapse with segmentation errors, an additional challenge is tracking through cell division. To account for this, we modified the approach in Jaqaman to account for both segmentation errors and cell division. To facilitate the identification of dividing cells, we used the object-classification pipeline in Ilastik (Berg et al. 2019) to label all nuclei as dividing (M-phase, with chromosomes aligned along the metaphase plate immediately prior to cell division) or non-dividing. In the original algorithm, at the frame-frame linking stage each cell in frame t is linked to at most one cell in frame $t + 1$, and splits are only assigned later. We maintain this general framework, looking for only one daughter cell for each cell marked as dividing during frame-frame linking, and aiming to identify the second daughter cell at the merging, splitting, gap closing step. Additionally, the cost function for linking or splitting from

dividing nuclei are modified to facilitate identification of progeny cells, as described below.

In addition to modifying the frame-frame linking and GMS steps to better handle cell divisions, we add a step for merge resolution. This is motivated by the observation that merging events occur purely due to segmentation errors and so our final tracks should not incorporate the merging of two cells into a single object. To resolve merges, we first look for a split from the merged track, indicating that two nuclei moved close together and then apart again, and determine which of the input tracks to the merge more closely matches each of the output tracks from the split. If there is no subsequent split, we assume that either a merge was followed by separation of the two nuclei that failed to be detected as a split, or that the merge was assigned in error. In either case, we determine which input track to the merge more closely matches the track after the merge, and discard the other link.

Frame-frame linking

To link cells in consecutive frames, we define the pairwise linking cost between each cell in frame t and each cell in frame $t+1$, as well as the cost for 'disappearance' of cells from frame t and 'appearance' of cells in frame $t+1$; that is, the cost for a cell in one frame to fail to be linked to any cell in the other. The cost for linking cell i in frame t to cell j in frame $t+1$ is based on the cells' xy positions, as well as the areas and intensities of the nuclei, and whether cell i is marked as dividing. The base cost for linking two cells is the squared euclidean distance between them, given by

$$D_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2$$

To obtain the final cost, we multiply this distance by weights based on the similarity of the two nuclei in area A and intensity I. If we define

$$d_A = \frac{2|A_i - A_j|}{A_i + A_j}, \quad d_I = \frac{2|I_i - I_j|}{I_i + I_j},$$

the the final cost is given by

$$c_{ij} = D_{ij}\,(1 + d_I)\,(1 + d_A)\,.$$

If cell i is labeled as dividing, an additional multiplicative weight is calculated based on the stereotypical rapid movement of daughter nuclei in opposite directions orthogonal to the orientation of the metaphase plate. This weight favors linking to prospective daughter cells found in a direction orthogonal to the metaphase plate. During image processing, the major and minor axes and orientation of an ellipse approximating the nucleus are calculated for each cell, and we use the orientation of cell i's major axis as the orientation of the metaphase plate. We define a normalized vector $\hat{v}$ orthogonal to that orientation. We additionally define the vector pointing from cell i to cell j,

$$\vec{u} = \begin{bmatrix} x_j - x_i \\ y_j - y_i \end{bmatrix},$$

and normalize it to $\hat{u} = \vec{u}/\|\vec{u}\|$. The additional weight for linking cell i to cell j is then

$$w = \frac{3}{2} - |\langle \hat{u}, \hat{v} \rangle|^3,$$

and the resulting overall cost is

$$c_{ij} = w \cdot D_{ij}\,(1 + d_I)\,(1 + d_A)\,.$$

The inner product $\langle \hat{u}, \hat{v} \rangle$ depends on the angle between $\hat{u}$ and $\hat{v}$ and varies from -1 (antiparallel) to 0 (orthogonal) to 1 (parallel). Our weight then varies from 3/2 (orthogonal) to 1/2 (either parallel or antiparallel as daughter cells travel in both directions). Note that the range of values taken by w is unaffected by cubing the inner product, but results in a wider range of angles close to $\pi/2$ producing close to the maximum weight.

We may then construct the cost matrix A with rows corresponding to prospective links from the n_t cells in frame t and columns corresponding to prospective links to the n_{t+1} cells in frame $t+1$, so that $A(i,j) = c_{ij}$, i.e.,

$$A = \begin{bmatrix} c_{11} & \cdots & c_{1n_{t+1}} \\ \vdots & \ddots & \vdots \\ c_{n_t 1} & \cdots & c_{n_t n_{t+1}} \end{bmatrix}.$$

For computational efficiency, we additionally take as an input a maximum linking distance that defines the maximum distance a cell is expected to move between consecutive frames. We treat links between cells at a distance greater than this cutoff as impossible by setting the linking cost to Inf (arbitrarily large). In practice, we used a maximum linking distance of about 15 µm. We additionally define the alternative costs for appearance and disappearance for each cell to be 105% of the maximum finite linking cost. Cost matrices for link rejection are constructed as follows: B_1 is an $n_t \times n_t$ diagonal matrix, with the cost for no link to be made to cell i in frame t at entry $B_1(i,i)$. All off-diagonal entries are set to Inf. Likewise, B_2 is an $n_{t+1} \times n_{t+1}$ diagonal matrix storing the costs to reject links to cells in frame $t+1$ and off-diagonal costs set to Inf. The resulting overall cost matrix is constructed as a block matrix as:

$$C = \begin{bmatrix} A & B_1 \\ B_2 & A^{\mathrm{T}} \end{bmatrix}.$$

Assignments are made by choosing one cost in each row such that no two costs come from the same column and the sum of the costs is minimized. This optimization is performed with the Jonker-Volgenant algorithm for LAPs (Jonker and Volgenant 1987) implemented in MATLAB (Cao 2023). Note that the inclusion of the transpose of A in the lower corner ensures that the number of assignments is the same along the rows and columns so that C is a square matrix, and column indices of the assignments in the first n_t rows will match the row indices of the last n_t columns. Likewise, row indices of the assignments to the first n_{t+1} columns will match the column indices of the last n_t rows.

Gap closing, merging, splitting

The GMS step aims to tie up loose ends (and beginnings) from the frame-frame linking step. Track ends are cells without a link to a cell in a subsequent frame and track beginnings are those without a link from a cell in a previous frame (note that these are not mutually exclusive: if a cell has no link in the previous or in the subsequent frame it is both the beginning and end of its own one-cell track). Unlike in frame-frame linking, this step is not local in time and optimizes over possible assignments in the entire time series at once. Each track end is matched to either a track beginning (gap closing), a mid-point of another track (merging), or is given no assignment (track termination). Conversely, each track start is matched to a track end (gap closing), a mid-point of another track (splitting), or is not linked (track initiation). The structure of the cost matrix constructed to handle these possible assignments is more complex, and is constructed as a block

matrix as

$$C = \begin{bmatrix} A_1 & A_2 & A_3 & \text{Inf} \\ B_1 & \text{Inf} & \text{Inf} & B_4 \\ C_1 & \text{Inf} & A_1^{\text{T}} & B_1^{\text{T}} \\ \text{Inf} & D2 & A_2^{\text{T}} & \text{Inf} \end{bmatrix}$$

Here A_1 contains costs for gap closing, A_2 for merging, A_3 for track termination, B_1 for splitting, and C_1 for track initiation. B_4 is a diagonal matrix with costs to reject splits and D_2 likewise has costs to reject merges. As in the frame-frame linking step, the cost matrix C is constructed to be a square matrix with the same possible assignments found along columns as along rows to satisfy the topological structure of the LAP. For instance, it can be seen that the first row of block matrices determines assignments from track ends, as does the third column of block matrices.

In constructing cost matrices, we again impose thresholds for computational efficiency, so links are only considered between cells within a maximum distance $\delta xy_{\max}$ (in practice, about 22.5 um) and a maximum number of time steps apart $\delta t_{\max}$ (in practice, five frames).

The matrix A_1 with costs for gap-closing is similar to the matrix of pairwise linking costs in the frame-frame linking step. Each entry of A_1 stores the cost to link a track end at nucleus i in frame t_1 to a track beginning at nucleus j in frame t_2 with $t_2 > t_1$. If the nuclei are within the threshold distances of one another, $t_2 \leq t_1 + \delta t_{\max}$ and $\|[x_j - x_i, y_j - y_i]^T\| \leq \delta xy_{\max}$, then the cost is given by

$$c_{ij} = \left[(x_i - x_j)^2 + (y_i - y_j)^2 + (t_2 - t_1)^2\right] \left(1 + \frac{|A_i - A_j|}{A_i}\right).$$

As in the frame-frame linking step, if cell i was labeled as dividing, this cost is multiplied by an additional weight w based on the angle between the normal vector to cell i's major axis and the vector between cell i and cell j. Similar to the frame-frame linking step, we build a diagonal cost matrix A_3 to reject links from each track end (cost for track termination) and C_1 to reject links to each track start (cost for track initiation). Like in the frame-frame linking step, these costs are taken to be slightly larger than the maximum finite gap-closing cost.

The matrix A_2 holds costs to merge track ends to midpoints of other tracks, where a track midpoint is any cell that is neither a track end nor a track start, i.e., that has a link both before and after it. Given cell i in frame t_1 that is a track end, we find all track midpoints within the time and distance cutoffs of the track end. For a given midpoint cell j in frame $t_2 \leq t_1 + \delta t_{\max}$, the cost to merge cell i into cell j is given by

$$m_{ij} = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right] \left(1 + \frac{|A_i + A_{j\text{prev}} - A_j|}{A_j}\right),$$

where $A_{j\text{prev}}$ is the area of the nucleus preceding cell j in its track, so that the cost of accepting a merge is lowest when the area of the merged nucleus is the sum of the areas of the two input nuclei. The alternative cost matrix to reject merging holds the cost of rejecting merges for each midpoint for which at least one merge is considered, and is given by

$$b_j = D_{\text{avg}} \left(1 + \frac{|A_{j\text{prev}} - A_j|}{A_j}\right),$$

where D_{avg} is the averaged squared frame-frame displacement for tracks constructed in the frame-frame linking step. Then, the cost for rejecting a merge is lower than the cost of accepting the merge if $|A_{j\text{prev}} - A_j| < |A_{j\text{prev}} + A_i - A_j|$, and if $D_{\text{avg}} < (x_i - x_j)^2 + (y_i - y_j)^2$.

The matrix B_1 holds costs to split track starts from midpoints of other tracks. Similar to the construction of the cost matrix for merging, we take a track start cell i in frame t_1, and find all track midpoints within the time and distance cutoffs. For a given midpoint cell j in frame t_2 with $t_1 > t_2 \geq t_1 - \delta t_{\max}$ that is not marked as dividing, the cost to split cell i from cell j is given by

$$s_{ij} = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right] \left(1 + \frac{|A_i + A_{j\text{next}} - A_j|}{A_j}\right),$$

so the cost is lower if the area before the split is closer to the sum of the areas of the two cells after the split. If cell j is marked as dividing, however, we assume that the first link is to one daughter cell and attempt to find and link to the other daughter cell. The position of the first daughter cell is used to find the expected position of the other, based on the observation that immediately after division, sibling cells move symmetrically away from the location of the parent nucleus prior to division. To find the expected position of the remaining sibling nucleus, the displacement of the first sibling from the parent is found, and the expected position is taken to be at the same displacement but in the opposite direction. The linking cost then uses the distance of each prospective cell from this expected position instead of the distance from the parent cell itself. The resulting linking cost is

$$s_{ij} = \left[(x_i - x_{\text{exp}})^2 + (y_i - y_{\text{exp}})^2\right] \left(1 + \frac{2|A_i - A_{j\text{daughter}}|}{A_i + A_{j\text{daughter}}}\right),$$

Where $A_{j\text{daughter}}$ is the area of the first daughter nucleus in the same frame as the track start. Much like for merging, the alternative cost to reject splits is

$$d_j = D_{\text{avg}} \left(1 + \frac{|A_{j\text{next}} - A_j|}{A_j}\right).$$

After all finite costs have been computed and the aggregate cost matrix is constructed as a block matrix, we numerically optimize to find the best solution to the LAP.

Merge resolution

We resolve merges with the aim of separating the individual tracks that were inputs into the merge at later time points. If a split occurs from the same track soon after a merge (within the maximum time cutoff for gap-closing), and the nucleus from which the split occurred was not labeled as dividing, we assign each input to the merge to one output from the split and discard the links in between; otherwise, one of the links in to the merge is discarded, depending on which input cell bears greater morphological similarity to the cell after the merge. These one or two assignments to resolve each merge are made as a (very small) LAP with costs as in the frame-frame linking step.

Linking live to fixed cells

At the end of live-cell imaging, each sample was fixed and immunofluorescence stained, and we include an additional step to link live cells at the end of the time-series to fixed cells. To ensure a consistent frame of reference, the image of nuclei in the last live frame is aligned to the DAPI stain of the fixed cells with phase correlation-based image registration and the positions of fixed nuclei are adjusted accordingly. Linking of individual cells is done in the same way as the frame-frame linking step during tracking, but with the linking cost based only on distance between nuclei and similarity in area. We do not consider nuclear intensity, which does not necessarily correlate between live data in which nuclei are labeled with fluorescent fusion proteins and fixed data where they are stained with DAPI. Sparse labeling introduces a potential complication to this step: only

10-20% of cells have nuclear markers in the live data, but every cell is stained for DAPI, including those that were not labeled live, so each live nucleus has potentially many more fixed nuclei nearby as candidates to which to link. However, we find that because there is little cell movement in the short time between the end of the live time lapse and the time that cells were fixed, our rigid image registration is robust to only a subset of nuclei being visible in the live data, and the alignment results in corresponding nuclei being very close in the aligned live and fixed data. To prevent erroneous linking to another nearby nucleus in the case that the true fixed nucleus corresponding to a given live cell failed to be properly segmented, we use a reduced maximum linking distance of 10 µm, or about one cell diameter, at this step.

Additional implementation details

The algorithm was implemented in MATLAB, and incorporated into a larger image-processing pipeline.

The the optimal set of assignments for each LAP in the tracking pipeline is computed numerically with a MATLAB implementation (Cao 2023) of the Jonker-Volgenant algorithm (Jonker and Volgenant 1987).

To account for shifts in the entire field of view between consecutive time points, we implemented a "dejittering" algorithm. We iterated over all frames in the time lapse and at each time loaded a maximal intensity projection of the z stack of images of nuclei at time t_i and t_{i+1} and used phase correlation to determine a global shift between the two images. At each time point, we applied the cumulative shift up to that time to the segmented cell positions in that time, effectively aligning the entire time-lapse to the field of view of the first frame. These updated cell positions are then used in the construction of cost functions for linking during tracking.

Parallelization is used to speed up the algorithm: at the frame-frame linking step, pairs of frames are linked in parallel, and in the gap closing, merging, splitting step, costs are computed in parallel as each block of the cost matrix is constructed.

References

Magnusson, Klas E. G. et al. (Apr. 2015). "Global Linking of Cell Tracks Using the Viterbi Algorithm". en. In: *IEEE Transactions on Medical Imaging* 34.4, pp. 911–929. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2014.2370951. URL: http://ieeexplore.ieee.org/document/6957576/ (visited on 03/29/2023).

Jaqaman, Khuloud et al. (Aug. 2008). "Robust single-particle tracking in live-cell time-lapse sequences". en. In: *Nature Methods* 5.8, pp. 695–702. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1237. URL: http://www.nature.com/articles/nmeth.1237 (visited on 05/02/2022).

Tinevez, Jean-Yves et al. (Feb. 2017). "TrackMate: An open and extensible platform for single-particle tracking". en. In: *Methods* 115, pp. 80–90. ISSN: 10462023. DOI: 10.1016/j.ymeth.2016.09.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S1046202316303346 (visited on 03/29/2023).

Berg, Stuart et al. (Dec. 2019). "ilastik: interactive machine learning for (bio)image analysis". en. In: *Nature Methods* 16.12, pp. 1226–1232. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0582-9. URL: http://www.nature.com/articles/s41592-019-0582-9 (visited on 05/02/2022).

Jonker, R and A Volgenant (1987). "A shortest augmenting path algorithm for dense and sparse linear assignment problems". en. In: *Computing* 38, pp. 325–340.

Cao, Li (2023). *LAPJV - Jonker-Volgenant Algorithm for Linear Assignment Problem V3.0*. URL: https://www.mathworks.com/matlabcentral/fileexchange/26836-lapjv-jonker-volgenant-algorithm-for-linear-assignment-problem-v3-0 (visited on 03/29/2023).

Mathematical model of BMP-SMAD4 integration

Seth Teague, Idse Heemskerk

March 30, 2023

Model development

We aimed to develop a mathematical model to explain how a simple gene regulatory network (GRN) could integrate BMP-SMAD4 signaling in time. In the simplest model, the expression of an integrator gene directly reflects the time integral of SMAD4 signaling. This is analogous to looking for genes with a rate of change that is a linear function of BMP-SMAD4 signaling, i.e., those that can be modeled with an ordinary differential equation (ODE) approximately as

$$\frac{d[\text{GENE}]}{dt} = \lambda[\text{SMAD4}](t),$$

where [GENE] gives the concentration of protein and [SMAD4](t) is the (time-varying) level of BMP signaling. Integration of the above results in production of the gene product that is directly proportional to the signaling integral, i.e.,

$$[\text{GENE}](t) = \lambda \int_0^t [\text{SMAD4}](\tau)\, d\tau + [\text{GENE}]_0,$$

where the initial concentration of protein is given by $[\text{GENE}]_0 = [\text{GENE}](0)$. More generally, we can allow an additional constant term β for constitutive production, so that protein production remains a linear function of signaling. Additionally, as protein products are not indefinitely stable, we add a decay term that is proportional to the current concentration of protein. The ODE model then becomes

$$\frac{d[\text{GENE}]}{dt} = \beta + \lambda[\text{SMAD4}](t) - \alpha[\text{GENE}].$$

Our screen for genes for which the rate of change is linear with SMAD4 signaling level found SOX2 to be a promising candidate, as measured at the protein level with immunofluorescence, and at the transcript level with RNA sequencing. If SOX2 is our integrator, its dynamics should roughly follow integrated SMAD4 signaling, and it should repress late-response amnion genes so that they are expressed only if SOX2 goes below a threshold level. We first aim to determine whether SOX2 dynamics can be plausibly modeled with the dynamics described above. SOX2 is negatively regulated by BMP signaling, so we rewrite the ODE as

$$\frac{d[\text{SOX2}]}{dt} = \beta - \lambda[\text{SMAD4}](t) - \alpha[\text{SOX2}], \tag{1}$$

where λ is taken to be positive. In the absence of BMP signaling, [SMAD4] $= 0$ and SOX2 expression tends to a steady-state value of β/α balancing constitutive production and decay. A model of the form

$$\frac{dy}{dt} = f(t) - \alpha y,$$

where f is a function of t but not y has the solution

$$y(t) = e^{-\alpha t} \int_0^t f(t) e^{\alpha \tau}\, d\tau + y_0 e^{-\alpha t}.$$

So the concentration of SOX2 is modeled by

$$[\text{SOX2}](t) = \beta/\alpha + ([\text{SOX2}]_0 - \beta/\alpha) e^{-\alpha t} - \lambda e^{-\alpha t} \int_0^t [\text{SMAD4}](\tau) e^{\alpha \tau}\, d\tau.$$

We assume that SOX2 is at or very near the steady state β/α in pluripotency maintenance conditions prior to stimulation of BMP signaling, i.e. $[\text{SOX2}]_0 = \beta/\alpha$. We further normalize this pretreatment expression level to one, so the expression for SOX2 as a function of time reduces to

$$[\text{SOX2}](t) = 1 - \lambda e^{-\alpha t} \int_0^t [\text{SMAD4}](\tau) e^{\alpha \tau}\, d\tau. \tag{2}$$

We see that the level of SOX2 protein reflects an exponentially-weighted integral of SMAD4 signaling, which closely approximates a direct integral of SMAD4 signaling for small α (where $e^{-\alpha t} \approx e^{\alpha t} \approx 1$). As a further simplification, if we consider conditions in which SMAD4 signaling is maintained at a steady level we can find the integral of [SMAD4] analytically and see that the SOX2 level exponentially decays towards the new steady state $1 - \lambda[\text{SMAD4}]/\alpha$ as described by

$$[\text{SOX2}](t) = 1 - \frac{\lambda[\text{SMAD4}]}{\alpha} \left(1 - e^{-\alpha t}\right).$$

Model fitting

To fit the model to experimental data, we measured GFP::SOX2 dynamics for 42 hours of BMP4-driven differentiation in conditions in which we could control the level and duration of signaling. Briefly, as described in the main text, we treated sparsely seeded hPSCs with a high dose of BMP4 to ensure uniformly high response, and controlled the signaling level via titration of a BMP receptor inhibitor (BMPRi). To control the duration, we shut down signaling by removing BMP4 and adding a high dose of BMPRi. We measured GFP::SOX2 dynamics at a range of signaling levels with durations of 42 and 32 hours (SI Fig 5H). As an indicator of differentiation to amnion, we measured ISL1 expression at the end of 42 hours with immunofluorescence. We approximated input SMAD4 dynamics as flat with levels measured in cells expressing GFP::SMAD4 in the same treatment conditions as the GFP::SOX2 cells (Fig 4; SI Fig 5G). We confirmed the linear relationship between SMAD4 signaling level and rate of SOX2 decay with a linear fit to the first 16 hours of GFP::SOX2 dynamics in each condition (Fig 5H). To determine the values of the parameters α, β, and λ in the model, we collected values of SOX2, SMAD4, and the slope of SOX2, averaged over short time windows to reduce the effect of measurement noise, and fit a plane defined by

$$\frac{d[\text{SOX2}]}{dt} = \beta - \lambda[\text{SMAD4}] - \alpha[\text{SOX2}],$$

to our measured values of $d[\text{SOX2}]/dt$, [SOX2], and [SMAD4]. Numerically integrating the model with the fitted parameters and with measured input SMAD4 dynamics, we see generally good agreement with measured SOX2, but with some discrepancies between model and data appearing later in the course of differentiation in conditions with the highest levels of signaling (SI Fig 5H, red and orange curves). In particular, the model predicts faster initial decay followed by a plateau

at later times, and for conditions in which BMP signals are removed, the model predicts sharp recovery of SOX2 levels. Signaling is completely shut down after addition of a high dose of BMPRi, so SOX2 should exponentially approach the initial level in all of these conditions. Because of this, the initial slope of recovery is expected to be highest for those with the lowest SOX2 level at the time of BMPRi addition. In contrast, SOX2 fails to robustly recover upon addition of BMPRi in high-signaling conditions (SI Fig 5H, right). Notably, there is a more pronounced recovery in SOX2 levels after BMPRi addition in conditions with lower signaling. We hypothesized that failure of SOX2 to recover after signaling shutdown in conditions with higher initial signaling reflected commitment to exit pluripotency. Because this effect seems to be pronounced only at later times in high-signaling conditions, we took it to be caused by repression of SOX2 by late-response amnion genes which only turn on in those conditions, and used ISL1 as a representative example of that class of genes. We modeled SOX2 as repressing expression of ISL1, as expected if it is our integrator gene. We additionally modeled repression of SOX2 by ISL1 so that SOX2 expression is further downregulated once ISL1 begins to be expressed. This fits the paradigm of mutually inhibitory regulatory programs specifying distinct cell fates that are widespread in development (Levine and Davidson 2005, Delás and Briscoe 2020). To implement this mutual repression mathematically, we take each gene to act on the other with Hill function dependence:

$$\frac{d[\text{SOX2}]}{dt} = \frac{\beta - \lambda_S[\text{SMAD4}]}{1 + ([\text{ISL1}]/K_{IS})^{n_S}} - \alpha_S[\text{SOX2}], \tag{3}$$

$$\frac{d[\text{ISL1}]}{dt} = \frac{\lambda_I[\text{SMAD4}]}{1 + ([\text{SOX2}]/K_{SI})^{n_I}} - \alpha_I[\text{ISL1}]. \tag{4}$$

In the above equations, the parameters K describe the threshold for 50% inhibition of one gene by the other and n describes the steepness of the Hill function. In our time series expression data for ISL1 (Fig 5A, SI Fig 5F), we see that it remains close to zero until 20-24 hours, when expression rapidly switches on, suggesting that there is a sharp threshold for regulation of ISL1. We therefore modeled repression of ISL1 by SOX2 with a switch-like Hill function by setting $n_I = 4$. On the other hand, repression of SOX2 by the amnion transcriptional program appears more graded, and we take $n_S = 2$. We used simulated annealing to fit equations (3) and (4) to measured SOX2 and ISL1 expression data, using the same SMAD4 input dynamics described above. Briefly, the Values of each parameter must be initialized: we used values found with the previous fit of SOX2 alone, i.e., $\lambda_I = \lambda_S = \lambda$, $\alpha_I = \alpha_S = \alpha$, and $\beta_S = \beta$. We further initialized the inhibition threshold coefficients K_{SI} and K_{IS} at 0.5. We then numerically evaluated the system of ODEs (3) and (4) with those parameters and the SMAD4 inputs described above, and calculated the mean squared error E between the target and calculated expression levels. Then for a set number of iterations, we do the following: perturb the parameters by applying Gaussian noise to each with a variance of 10^{-5} and calculate the new error E_{new} after running the model with the new parameters. If the new error is lower, accept these values as the new parameter values; otherwise, we may still accept the new parameter values with probability $\exp(-\Delta E/k_B T)$, where $\Delta E = E_{\text{new}} - E$, T is the 'effective temperature' for the annealing, and k_B is a tunable constant. The value of T linearly decreases to zero over the course of the iterations so that accepting a set of parameters resulting in a higher cost becomes increasingly unlikely as the algorithm progresses, allowing exploration of the parameter space at early iterations to avoid becoming trapped at a local minimum in the parameter landscape, and settling in to a specific minimum at the end.

We see that the resulting simulated SOX2 dynamics align more closely with the measured dynamics, and resolve the discrepancies mentioned above (Fig 5I). Furthermore, we see that the

relationship between the SMAD4 integral and ISL1 expression seen in the data is conserved for both signaling durations in our model (Fig 5J).

To generate the results shown in Fig 5IJ, we used the following parameter values:

Parameter	Value	Meaning
(α_S, α_I)	$(0.0363, 0.09)$	protein dilution + degradation rates
β	0.0329	constitutive SOX2 production rate
(λ_S, λ_I)	$(0.0397, 0.1240)$	coefficients for regulation by SMAD4
(n_S, n_I)	$(2, 4)$	Hill function coefficients
(K_{SI}, K_{IS})	$(0.266, 0.3995)$	inhibition thresholds

References

Levine, Michael and Eric H. Davidson (Apr. 2005). "Gene regulatory networks for development". en. In: *Proceedings of the National Academy of Sciences* 102.14, pp. 4936–4942. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0408031102. URL: https://pnas.org/doi/full/10.1073/pnas.0408031102 (visited on 03/30/2023).

Delás, M. Joaquina and James Briscoe (2020). "Repressive interactions in gene regulatory networks: When you have no other choice". en. In: *Current Topics in Developmental Biology*. Vol. 139. Elsevier, pp. 239–266. ISBN: 9780128131800. DOI: 10.1016/bs.ctdb.2020.03.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0070215320300508 (visited on 03/30/2023).