Supplementary Material
Probabilistic Modeling of Reprogramming to Induced Pluripotent Stem Cells

1 A two-type stochastic logistic process model for reprogramming dynamics

1.1 Notations

We designed a stochastic model to predict the dynamics of $\{Y(t) = (X_1(t), X_2(t))\}$, where $X_1(t)$ denotes the number of somatic cells (in our case, granulocyte-macrophage progenitors, GMPs) at time t, and $X_2(t)$ denotes the number of induced pluripotent stem cells (iPSCs) at time t. We define an invariant mapping $g : (X_1(t), X_2(t)) \to (S(t), prop(t))$ through $S(t) = X_1(t)+X_2(t)$ and $prop(t) = \frac{X_2(t)}{S(t)}$ so that our model prediction corresponds to a readout from the experimental data, specifically the percentage of Oct4-GFP+ wells. Denote the percentage of Oct4-GFP+ wells at time t by $prop(t)$. Our model considers a carrying capacity constraint M such that $S(t) \leq M \ \forall t$, where $t \in \mathbb{R}^+ \cup \{0\}$. Though outside the scope of this work, this assumption can be relaxed (Crawford et al., 2014). We designed the underlying dynamical process as a two-species stochastic version of the Verhulst logistic growth model by extending the model in one dimension proposed by Tan and Piantadosi (1991). Within this stochastic logistic process, the parameters λ_1 and λ_2 denote the initial proliferation rates per day per cell for GMPs and iPSCs, respectively (i.e., the "cell-intrinsic" rates for population sizes sufficiently small such that they are not yet impacted by the carrying capacity). The parameters ϕ_1 and ϕ_2 denote the apoptosis rates per day for GMPs and iPSCs, respectively, and γ represents the reprogramming rate from GMPs to iPSCs per day. The parameters can be arbitrary functions of time t. We assume that only the proliferation rates λ are affected by the presence of the carrying capacity. Note that in derivations in later sections, we treat all these cell-intrinsic rate parameters such as the proliferation rate, apoptosis rate and dedifferentiation rate as time-dependent variables to make our derivations more general, since the system with time-independent variables represents a special case of the system with time-dependent variables. Moreover, in **R**esult II of the main text, we suggest that a random reprogramming rate might explain the variability observed in the experimental data.

1.2 A one-type stochastic logistic process

We first review the stochastic logistic process in the one-type case (Tan and Piantadosi, 1991) as a building block for our two-dimensional extension. This process is also essential for estimating the proliferation and apoptosis rates for GMPs in both growth conditions (see Section III.1). The process $\{X(t), t \geq 0\}$ is a Markov process. Given j individuals at time t, the infinitesimal transition probabilities at time $t + \Delta t$ are given by

$$\begin{aligned} &P(X(t+\Delta t) = n + j | X(t) = n) \\ &= \begin{cases} \lambda(t)n(1-\frac{n}{M})\Delta t + o(\Delta t) & \text{if } j = 1, \\ \phi(t)n\Delta t + o(\Delta t) & \text{if } j = -1, \\ 1 - \lambda(t)n(1-\frac{n}{M})\Delta t - \phi(t)n\Delta t + o(\Delta t) & \text{if } j = 0, \\ o(\Delta t) & \text{else,} \end{cases} \end{aligned} \tag{1}$$

where $\lambda(t), \phi(t) > 0 \ \forall t \geq 0$, $j \leq M$, and $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$ for all $j = 0, 1, \cdots, M$. Then the master equation governing the probability mass function of $X(t)$ for the one-type stochastic logistic process is (Taylor and Karlin, 2014):

$$\begin{aligned} &\frac{\partial P(X(t) = n, t)}{\partial t} \\ &= P(X(t) = n-1, t) \cdot \lambda(t) \cdot (n-1) \cdot \left(1 - \frac{n-1}{M}\right) \\ &\quad + P(X(t) = n+1, t) \cdot \phi(t) \cdot (n+1) \\ &\quad - P(X(t) = n, t) \cdot \left\{ \ \lambda(t) \cdot n \cdot \left(1 - \frac{n}{M}\right) + \phi(t) \cdot n \ \right\}. \end{aligned} \tag{2}$$

Define $Q(u, z; t_0, t_1) = \sum_{v=0}^{M} z^v P_{uv}(t_0, t_1)$ to be the probability generating function (PGF) of $X(t_1)$ given $X(t_0) = u$, where $t_1 > t_0$ and $P_{uv}(t_0, t_1) = Pr\{X(t_1) = v | X(t_0) = u\}$ as a short-hand notation. The equation for PGF derived from the master equation (Tan and Piantadosi, 1991) is given by

$$\frac{\partial}{\partial t_1} Q(u, z; t_0, t_1) = (z-1) \cdot \left[\left\{ z \cdot \left(1 - \frac{1}{M}\right) \lambda(t) - \phi(t) \right\} \cdot \frac{\partial}{\partial z} Q(u, z; t_0, t_1) - z \left\{ \frac{z}{M} \lambda(t) - \phi(t) \right\} \cdot \frac{\partial^2}{\partial z^2} Q(u, z; t_0, t_1) \right] \tag{3}$$

Using Eq. (3), the differential equations for the first and second moments are given by

$$\begin{aligned} \frac{\partial}{\partial t} \kappa^{(1)} &= (\lambda(t) - \phi(t)) \kappa^{(1)} + \frac{\lambda(t)}{M} \kappa^{(2)}, \\ \frac{\partial}{\partial t} \kappa^{(2)} &= (\lambda(t) + \phi(t)) \kappa^{(1)} + \left\{ 2(\lambda(t) - \phi(t)) - \frac{\lambda(t)}{M} \right\} \kappa^{(2)} - \frac{2\lambda(t)}{M} \kappa^{(3)}, \end{aligned} \tag{4}$$

where $\kappa^{(k)} = \mathbb{E}[X^k]$. This system of two coupled differential equations is not solvable, but the first and second moments can be analytically approximated using the moment closure approximation (Murrell et al., 2004; Nåsell, 2003) by setting the higher centered moments as zero. The system of differential equations can be solved using Euler's method (Smith, 1965).

Although throughout the paper, we assume a carrying capacity in the sense that $X(t) \leq M \ \forall t$, where $t \in \mathbb{R}^+ \cup \{0\}$, this limitation can be relaxed by replacing the carrying capacity penalization $1 - 1/M$ by other functional forms, for instance e^{-M}, where $1 - 1/M$ can be interpreted as the first order Taylor series expansion of e^{-M}. It can be further generalized to e^{-cM} for different carrying capacity constraints.

In addition, we also showed the equations for mean and variance dynamics for the one-type stochastic linear birth-death process (without carrying capacity constraints) conditional on non-extinction at each time point for the results in **Figure 5** in the main text. We denote $\mu(t) = E[X(t)|X(t) > 0]$ and $\nu(t) = Var[X(t)|X(t) > 0]$. Then, assuming time-independent birth and death rates, we have

$$\begin{aligned} \mu(t) =& \lambda e^{(\lambda-\phi)t-\phi}/(\lambda-\phi) \\ \nu(t) =& \lambda\phi e^{2(\lambda-\phi)t} - (\phi^2 + \lambda - \phi\lambda)e^{(\lambda-\phi)t} + \phi(1-\phi)/(\lambda-\phi)^2. \end{aligned} \tag{5}$$

Then we can used the empirical mean and variance from the cell count data to estimate the parameters λ and ϕ.

1.3 A two-type stochastic logistic process

We extended the one-dimensional probabilistic logistic model to the two-type case using the notations introduced in Section 1.1. Although extensions to a larger number of types, i.e. m-type ($m \in \mathbb{Z}_{>0}$) generalized birth-death processes is possible, we decided not to further pursue such a generalization in this paper in order to not deviate too far from the subject matter. This process is defined by the

following infinitesimal transition probabilities:

$$
\begin{aligned}
&P(\boldsymbol{X}(t+\Delta t)=(n+j,m+k)|\boldsymbol{X}(t)=(n,m))\\
&=\begin{cases}
\lambda_1(t)n\left(1-\dfrac{n+m}{M}\right)\Delta t+o(\Delta t) & \text{if } j=1,k=0,\\
\phi_1(t)n\Delta t+o(\Delta t) & \text{if } j=-1,k=0,\\
\lambda_2(t)m\left(1-\dfrac{n+m}{M}\right)\Delta t+o(\Delta t) & \text{if } j=0,k=1,\\
\phi_2(t)m+o(\Delta t) & \text{if } j=0,k=-1,\\
\gamma(t)n+o(\Delta t) & \text{if } j=-1,k=1,\\
1-(\lambda_1(t)n+\lambda_2(t)m)\left(1-\dfrac{n+m}{M}\right)\Delta t-(\phi_1(t)n+\phi_2(t)m)\Delta t-\gamma(t)n\Delta t+o(\Delta t) & \text{if } j=0,k=0,\\
o(\Delta t) & \text{else,}
\end{cases}
\end{aligned}
\tag{6}
$$

where $\lambda_1(t),\lambda_2(t),\phi_1(t),\phi_2(t)\geq 0$, $n+m\leq M$ and $\lim\limits_{\Delta t\to 0} o(\Delta t)/\Delta t=0$ for all $m,n=0,1,\cdots,M$. Then the master equation governing the joint probability mass function of $\mathbf{X}(t)$ for the two-type stochastic logistic process is (Taylor and Karlin, 2014):

$$
\begin{aligned}
&\frac{\partial P(\boldsymbol{X}(t)=(n,m),t)}{\partial t}\\
&=P(\boldsymbol{X}(t)=(n-1,m),t)\cdot\lambda_1(t)\cdot(n-1)\cdot\left(1-\frac{n-1+m}{M}\right)\\
&\quad+P(\boldsymbol{X}(t)=(n,m-1),t)\cdot\lambda_2(t)\cdot(m-1)\cdot\left(1-\frac{n+m-1}{M}\right)\\
&\quad+P(\boldsymbol{X}(t)=(n+1,m),t)\cdot\phi_1(t)\cdot(n+1)+P(\boldsymbol{X}(t)=(n,m+1),t)\cdot\phi_2(t)\cdot(m+1)\\
&\quad+P(\boldsymbol{X}(t)=(n-1,m+1),t)\cdot\gamma(t)\cdot(n-1)\\
&\quad-P(\boldsymbol{X}(t)=(n,m),t)\cdot\left\{\begin{array}{c}\lambda_1(t)\cdot n\cdot\left(1-\dfrac{n+m}{M}\right)+\lambda_2(t)\cdot m\cdot\left(1-\dfrac{n+m}{M}\right)\\ +\phi_1(t)\cdot n+\phi_2(t)\cdot m+\gamma(t)\cdot n\ .\end{array}\right\}
\end{aligned}
\tag{7}
$$

Define $Q(u_1,u_2,z_1,z_2;t_0,t_1)=\sum_{v_1,v_2=0}^{v_1+v_2\leq M} z_1^{v_1}z_2^{v_2}P_{(u_1,u_2),(v_1,v_2)}(t_0,t_1)$ to be the probability generating function (PGF) of $\mathbf{X}(t_1)=(v_1,v_2)$ given $\mathbf{X}(t_0)=(u_1,u_2)$, where $t_1>t_0$ and $P_{(u_1,u_2),(v_1,v_2)}(t_0,t_1)=Pr\{\mathbf{X}(t_1)=(v_1,v_2)|\mathbf{X}(t_0)=(u_1,u_2)\}$ as a short-hand notation.

We then determined the probability generating function using the Kolmogorov forward equation ap-

proach (Taylor and Karlin, 2014):

$$
\begin{aligned}
&\frac{\partial}{\partial t}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
=&\Big[(z_1-1)\Big\{z_1\lambda_1(t)\Big(1-\frac{1}{M}\Big)-\phi_1(t)\Big\}-(z_1-z_2)\gamma(t)\Big]\cdot\frac{\partial}{\partial z_1}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&+\Big[(z_2-1)\Big\{z_2\lambda_2(t)\Big(1-\frac{1}{M}\Big)-\phi_2(t)\Big\}\Big]\cdot\frac{\partial}{\partial z_2}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&-\Big[z_1(z_1-1)z_2\frac{\lambda_1(t)}{M}+z_1z_2(z_2-1)\frac{\lambda_2}{M}\Big]\cdot\frac{\partial^2}{\partial z_1\partial z_2}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&-\Big[z_1^2(z_1-1)\frac{\lambda_1}{M}\Big]\cdot\frac{\partial^2}{\partial z_1^2}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&-\Big[z_2^2(z_2-1)\frac{\lambda_2^2}{M}\Big]\cdot\frac{\partial^2}{\partial z_2^2}Q(u_1,u_2,z_1,z_2;t_0,t_1).
\end{aligned}
\tag{8}
$$

Using Eq. (8), we can derive a system of coupled differential equations for the first and second moments

of $\mathbf{X}(t)$:

$$
\begin{aligned}
\frac{\partial m_1^{(1)}(t)}{\partial t} &= (\lambda_1 - \phi_1 - \gamma) m_1^{(1)}(t) - \frac{\lambda_1}{M}(m_{11}^{(2)}(t) + m_{12}^{(2)}(t)) \\
\frac{\partial m_2^{(1)}(t)}{\partial t} &= \gamma m_1^{(1)}(t) + (\lambda_2 - \phi_2) m_2^{(1)}(t) - \frac{\lambda_2}{M}(m_{12}^{(2)}(t) + m_{22}^{(2)}(t)) \\
\frac{\partial m_{11}^{(2)}(t)}{\partial t} &= (\lambda_1 + \mu_1 + \gamma) m_1^{(1)}(t) + \left(2\lambda_1 - 2\mu_1 - 2\gamma - \frac{\lambda_1}{M}\right) m_{11}^{(2)}(t) - \frac{\lambda_1}{M} m_{12}^{(2)}(t) - \frac{2\lambda_1}{M}(m_{111}^{(3)}(t) + m_{112}^{(3)}(t)) \\
\frac{\partial m_{12}^{(2)}(t)}{\partial t} &= -\gamma(m_1^{(1)}(t) - m_{11}^{(2)}(t)) + (\lambda_1 - \mu_1 + \lambda_2 - \mu_2 - \gamma) m_{12}^{(2)}(t) - \frac{\lambda_1 + \lambda_2}{M}(m_{112}^{(3)}(t) + m_{122}^{(3)}(t)) \\
\frac{\partial m_{22}^{(2)}(t)}{\partial t} &= (\lambda_2 + \mu_2) m_2^{(1)}(t) + \gamma m_1^{(1)}(t) + \left(2\lambda_2 - 2\mu_2 - \frac{\lambda_2}{M}\right) m_{22}^{(2)}(t) + \left(2\gamma - \frac{\lambda_2}{M}\right) m_{12}^{(2)}(t) \\
&\quad - \frac{2\lambda_2}{M}(m_{222}^{(3)}(t) + m_{122}^{(3)}(t)) \\
\frac{\partial m_{111}^{(3)}(t)}{\partial t} &= (\lambda_1 - \mu_1 - \gamma) m_1^{(1)}(t) + \left(3\lambda_1 + 3\mu_1 + 3\gamma - \frac{\lambda_1}{M}\right) m_{11}^{(2)}(t) - \frac{\lambda_1}{M} m_{12}^{(2)}(t) \\
&\quad + 3\left(\lambda_1 - \mu_1 - \gamma - \frac{\lambda_1}{M}\right) m_{111}^{(3)}(t) - \frac{3\lambda_1}{M} m_{112}^{(3)}(t) - \frac{3\lambda_1}{M} m_{1111}^{(4)}(t) - \frac{3\lambda_1}{M} m_{1112}^{(4)}(t) \\
\frac{\partial m_{112}^{(3)}(t)}{\partial t} &= \gamma(m_1^{(1)}(t) - 2m_{11}^{(2)}(t)) + (\lambda_1 + \mu_1 + \gamma) m_{12}^{(2)}(t) + \gamma m_{111}^{(3)}(t) \\
&\quad + \left(2\lambda_1 - 2\mu_1 + \lambda_2 - \mu_2 - 2\gamma - \frac{\lambda_1}{M}\right) m_{112}^{(3)}(t) - \frac{\lambda_1}{M} m_{122}^{(3)}(t) - \frac{2\lambda_1 + 2\lambda_2}{M}(m_{1112}^{(4)}(t) + m_{1122}^{(4)}(t)) \\
\frac{\partial m_{122}^{(3)}(t)}{\partial t} &= -\gamma(m_1^{(1)}(t) - m_{11}^{(2)}(t)) + (\lambda_2 + \mu_2 - 2\gamma) m_{12}^{(2)}(t) + \left(2\lambda_2 - 2\mu_2 + \lambda_1 - \mu_1 - \gamma - \frac{\lambda_2}{M}\right) m_{122}^{(3)}(t) \\
&\quad + \left(2\gamma - \frac{\lambda_2}{M}\right) m_{112}^{(3)}(t) - \frac{2\lambda_1 + 2\lambda_2}{M}(m_{1222}^{(4)}(t) + m_{1122}^{(4)}(t)) \\
\frac{\partial m_{222}^{(3)}(t)}{\partial t} &= (\lambda_2 - \mu_2) m_2^{(1)}(t) + \left(3\lambda_2 + 3\mu_2 - \frac{\lambda_2}{M}\right) m_{22}^{(2)}(t) - \frac{\lambda_2}{M} m_{12}^{(2)}(t) \\
&\quad + 3\left(\lambda_2 - \mu_2 - \frac{\lambda_2}{M}\right) m_{222}^{(3)}(t) - \frac{3\lambda_2}{M} m_{122}^{(3)}(t) - \frac{3\lambda_2}{M} m_{2222}^{(4)}(t) - \frac{3\lambda_2}{M} m_{1222}^{(4)}(t) \, .
\end{aligned}
\tag{9}
$$

where $m_i^{(1)}(t) = \mathbb{E}[X_i(t)]$, $m_{ij}^{(2)}(t) = \mathbb{E}[X_i(t)X_j(t)]$, $m_{ijk}^{(3)}(t) = \mathbb{E}[X_i(t)X_j(t)X_k(t)]$ and $m_{ijk\ell}^{(4)}(t) = \mathbb{E}[X_i(t)X_j(t)X_k(t)X_\ell(t)]$. Again, we used the moment closure analytical approximation approach (Murrell et al., 2004; Nåsell, 2003) by setting all the fourth central moments to zero, including $\mathbb{E}[(X_1(t) - \mathbb{E}[X_1(t)])^4]$, $\mathbb{E}[(X_1(t) - \mathbb{E}[X_1(t)])^3(X_2(t) - \mathbb{E}[X_2(t)])]$, $\mathbb{E}[(X_1(t) - \mathbb{E}[X_1(t)])^2(X_2(t) - \mathbb{E}[X_2(t)])^2]$, $\mathbb{E}[(X_1(t) - \mathbb{E}[X_1(t)])(X_2(t) - \mathbb{E}[X_2(t)])^3]$ and $\mathbb{E}[(X_2(t) - \mathbb{E}[X_2(t)])^4]$; as with the one-type logistic process, the two-type logistic process cannot be solved exactly (Tan and Piantadosi, 1991). Then, thanks to an anonymous reviewer, we used the following approximation to calculate the proportion of iPSCs at time

t following the standard multivariate Taylor expansion (Apostol, 1974):

$$\mathbb{E}\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right] \approx \frac{\mathbb{E}\left[X_2(t)\right]}{\mathbb{E}\left[X_1(t)+X_2(t)\right]} + \frac{\mathbb{E}\left[X_1(t)X_2(t)\right]\left(\mathbb{E}[X_2(t)]-\mathbb{E}[X_1(t)]\right)-\mathbb{E}[X_1(t)]\mathbb{E}[X_2(t)^2]+\mathbb{E}[X_1(t)^2]\mathbb{E}[X_2(t)]}{\left\{\mathbb{E}\left[X_1(t)+X_2(t)\right]\right\}^3} \tag{10}$$

The accuracy of the analytical approximation was validated using exact simulations based on (Gillespie, 1977) (**Figure S1**). Together with the approximation (Equation (10)), we also display the predictions of a much simpler formula

$$\mathbb{E}\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right] \approx \frac{\mathbb{E}\left[X_2(t)\right]}{\mathbb{E}\left[X_1(t)+X_2(t)\right]} \tag{11}$$

with dashed lines in **Figure S1**. The two approximations provide extremely similar results and interested readers can calculate either of the two approximations from the output of the R code provided in Section 4.

Unfortunately, we did not find that this approach was useful for approximating the variance of the proportion of iPSCs $Var\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right]$. To approximate $Var\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right]$, one needs to approximate $\mathbb{E}\left[\left(\frac{X_2(t)}{X_1(t)+X_2(t)}\right)^2\right]$, which is a nonlinear function of the distribution of $X_1(t)/(X_1(t)+X_2(t))$. More theoretical investigations are necessary to obtain an approximation. In the interest of the theme of our paper, we decided not to pursue this goal further and use variances based on computer simulations instead.

2 Parameter estimation for the two-type stochastic logistic process

We then sought to parameterize the newly designed two-type stochastic logistic process model using experimentally-derived data.

2.1 Estimating proliferation and apoptosis rates for GMPs and iPSCs

We first used equation Eq. (4) to calculate the expected number and variance of GMPs for different proliferation rates. This equation was initialized with the expected number and variance of GMPs at day 1 and setting the apoptosis rate as 0. Using this approach, the proliferation rate λ_1 was identified

as the value minimizing the mean squared error between the model prediction and cell counts at day 2. We then determined the expected number and variance of live cells using data on the percentage of living and dead cells. Then Eq. (4) was used again, by setting the proliferation rate to the estimate of λ_1, to calculate the expected number and variance of GMPs for different apoptosis rates. Finally, the apoptosis rate ϕ_1 was identified as the value minimizing the mean squared error between the model prediction and dead cell counts at day 2. The data as well as estimated proliferation and apoptosis rates are shown in **Table 1**. We also tabulated the model predicted mean and variance of cell counts at day 2. The variances of model prediction and experimental data are on a similar scale, indicating that constant proliferation and apoptosis rates are sufficient for modeling the dynamics of cellular reprogramming process.

For iPSCs, only cell doubling times were determined as approximately 10.22 hours; this value was used as the net growth rate $(\lambda_2 - \phi_2)$ of iPSCs. We did not perform experiments to estimate the apoptosis rate ϕ_2 of iPSCs. In the main text, we assume that the apoptosis rate of iPSCs is equal to the apoptosis rate of GMPs in each culture condition. In Section 3, we conduct sensitivity analyses for the apoptosis rate of iPSCs to demonstrate the robustness of our results.

2.2 Estimating the reprogramming rate

We then used exact simulations to calculate the expected percentage of iPSCs with $\frac{\mathbb{E}[X_2(t)]}{\mathbb{E}[X_1(t)+X_2(t)}$ for different reprogramming rates γ. These simulations were initialized by setting $(X_1(t=0)=1, X_2(t=0)=0)$ and $\lambda_1, \phi_1, \lambda_2, \phi_2$ equal to the estimates identified in Section 2.1. The estimate for γ was determined as the value that minimizes the mean squared error between model prediction and the empirical mean of the percentage of Oct4-GFP+ wells. Since the wells were diluted to 10% due to confluence at unrecorded times after day 10 for the OSKM condition, we randomly selected three time points after day 10 as the splitting points in the simulation. We then performed sensitivity analyses (Section 3) to demonstrate that this splitting does not significantly change the trajectories of the proportion of iPSCs in the simulations.

Remark 2.1. *Thanks to an anonymous reviewer, we provide an argument on the existence of different orders of moment dynamics of the two-type stochastic logistic process when any rate parameters are non-degenerate but drawn from an integrable probability distribution. First let us condition on all non-degenerate rate parameters, i.e. analyzing an ordinary stochastic logistic process with constant*

rate parameters. By definition, the random vector $\boldsymbol{X}(t)$ must be strictly bounded by $([0, M], [0, M]$. In addition, the domain of $\boldsymbol{X}(t)$ is $\mathbb{Z}_{\geq 0}$ (nonnegative integers). Therefore by discreteness the probability mass function $P(\boldsymbol{X}(t)|\theta)$ (where θ represents all rate parameters) governed by the master equation Eq. (7) is also strictly bounded within $[0, 1]$. Therefore, all moments of $\boldsymbol{X}(t)$, up to a finite order, must be bounded as well when conditioning on the non-degenerate rate parameters, and we can always find an upper bound which does not depend on the parameters. Since we assume integrability of the probability density function of the non-degenerate parameters, the marginal moments also have to be bounded.

2.3 Identifying distributions for the reprogramming rate in the OSKM condition

We then aimed to explore distributions of the unsynchronized reprogramming rate γ that were able to explain the additional variability not generated by the randomness of the logistic process per se. To this end, considering the fact that $\gamma \geq 0$, we explored lognormal distributions with different standard deviations (0, 0.25, 0.3, 0.35, $\cdots$, 1.05) (**Figure S2**). For a constant (synchronized) γ across all wells (**Figure S2 (a)**), we found a large deviation between the model prediction and experimental data, although the model still led to a good fit in terms of the mean percentage of iPSCs. We found that our logistic process model incorporating a random γ was able to fit the data, and also led to a variance which was on a similar scale as that of the data as γ increases. We selected the distribution of γ with the smallest mean squared error of the Fano factors between the model prediction and the experimental data (standard deviation = 0.75).

3 Sensitivity analyses

We then sought to explore whether a slight perturbation in the parameter values would alter the trajectories of the mean and variability of the percentage of iPSCs in either cell culture condition. The reason why we would like to perform such sensitivity analyses is that we do not have data to estimate the proliferation and apoptosis rates for iPSCs apart from the rough estimates available from the literature. All results are based on the analytical approximation, Equation (9), since we have already demonstrated that the analytical approximation is very close to the computer simulation results (**Figure S1**).

We first varied the apoptosis rate ϕ_2 for iPS cells in the range of $[0.1, 0.8]$; we found that, when ϕ_2

increases, the difference of the estimated value of γ between the OSKM and OSKM + AGi conditions decreases. However, the OSKM + AGi condition still leads to a higher γ than the OSKM condition when $\phi_2 \leq 0.5$, indicating a much lower net growth rate of iPS cells than somatic cell GMPs. Furthermore, when altering the apoptosis rate, the fit of the model prediction to the data decreases as indicated by the increased mean squared error between model prediction and empirical data (**Figure S3A** and **Figure S3B**).

We next performed sensitivity analyses for different values of the carrying capacity, $M = 100, 1,000$, $10,000$, $100,000, 1,000,000$, and $10,000,000$ (**Figure S3 C-D**). We observed that, when the carrying capacity varies, the model simulation of the percentage of iPS cells does not change dramatically, indicating robustness to estimation errors of M.

We then performed sensitivity analyses to demonstrate that the dilution strategy of cell populations does not significantly alter the trajectories of the percentage of iPS cells (**Figure S3 E-G**). We obtained consistent results for different splitting strategies: (1) no splitting of cells in the simulations; (2) splitting to 10% (i.e., keep 10%, discard 90%) on days 12, 20, and 31; and (3) splitting to 10% on days 15, 23, and 31. The correlations between either two strategies are greater than 0.99, indicating that different splitting strategies do not lead to dramatically different outcomes. Therefore we decided not to be concerned with such an "unnatural" modeling strategy in the main text.

Next, we performed sensitivity analyses to test whether a heterogeneous proliferation rate or apoptosis rate also leads to a good fit between the model prediction and experimental data in terms of both mean and variance. As shown in **Table 1** and Section 2.1, the variances of model prediction and experimental data are on a similar scale, indicating that a constant proliferation and apoptosis rate across wells is sufficient for modeling the dynamics of the cellular reprogramming process. Without this evidence, however, it would be impossible to identify which factor (proliferation/apoptosis rate or reprogramming rate) is heterogeneous in order to explain the excess variability in the observed experimental data. For example, in **Figure S4** and **Figure S5**, the proliferation/apoptosis rate for GMPs is a random variable from a log-normal distribution with different variances (0.2, 0.6, 0.8, 1.0, 1.2, 1.4, and 1.6) instead of a constant. We observed that the mean of the percentage of iPSCs from computer simulations led to a good fit compared with experimental data, and the variances also reached a similar scale as in the experimental data. Therefore we can only conclude that an unsynchronized reprogram-

ming rate contributes to the increased extent of variation of %iPSCs in the OSKM condition under the assumption of synchronized proliferation/apoptosis rate, supported by the data and calculations from **Table 1** and Section 2.1.

Finally we performed sensitivity analyses on the effects of increasing apoptosis and proliferation rates of iPSCs on the intrinsic variability of probability of iPSCs over time by assuming homogeneous reprogramming rates. As shown in **Figure S6**, we can conclude that if our guess on the apoptosis rate is incorrect, we still have the same conclusion as in the main text - the process with homogeneous reprograming rate underestimates the empirical variance.

4 R function to solve Equation (9)

In the following R code, we provide the R function to solve Equation (9). Notice that the returned vector of this function corresponds to the vector of different moments of the stochastic logistic process used in this paper, in which $X1$ corresponds to $\mathbb{E}[X_1(t)]$, $X2$ corresponds to $\mathbb{E}[X_2(t)]$, $X11$ corresponds to $\mathbb{E}[X_1(t)^2]$, $X12$ corresponds to $\mathbb{E}[X_1(t)X_2(t)]$, $X22$ corresponds to $\mathbb{E}[X_2(t)^2]$, $X111$ corresponds to $\mathbb{E}[X_1(t)^3]$, $X112$ corresponds to $\mathbb{E}[X_1(t)^2X_2(t)]$, $X122$ corresponds to $\mathbb{E}[X_1(t)X_2(t)^2]$, and $X222$ corresponds to $\mathbb{E}[X_2(t)^3]$.

```
MomentEquationsQuadraClosure <- function(t, X, params) {
    lambda1 = params[1]
    alpha1 = 1
    mu1 = params[2]
    beta1 = 0
    gamma1 = params[3]
    delta1 = 0

    lambda2 = params[4]
    alpha2 = 1
    mu2 = params[5]
    beta2 = 0
    gamma2 = 0
    delta2 = 0

    M = params[6]

    dX1 = (lambda1 - mu1 - gamma1) * X[1] + gamma2 * X[2] +
          (- lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1) / M * X[3] +
          (- lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1 - gamma2 * delta2) / M * X[4] -
          gamma2 * delta2 / M * X[5]
    dX2 = (lambda2 - mu2 - gamma2) * X[2] + gamma1 * X[1] +
          (- lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2) / M * X[5] +
          (- lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2 - gamma1 * delta1) / M * X[4] -
          gamma1 * delta1 / M * X[3]

    dX11 = (lambda1 + mu1 + gamma1) * X[1] + gamma2 * X[2] + (2 * lambda1 - 2 * mu1 - 2 * gamma1) * X[3] -
           (lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1) / M * X[3] + 2 * gamma2 * X[4] -
           (lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1 + gamma2 * delta2) / M * X[4] -
           gamma2 * delta2 / M * X[5] - 2 * (lambda1 * alpha1 - mu1 * beta1 - gamma1 * delta1) / M * X[6] +
           2 * (- lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1 - gamma2 * delta2) / M * X[7] -
```

```
       2 * gamma2 * delta2 / M * X[8]
dX12 = - gamma1 * X[1] - gamma2 * X[2] + (gamma1 + gamma1 * delta1 / M) * X[3] +
       (lambda1 - mu1 - gamma1 + lambda2 - mu2 - gamma2) * X[4] +
       (gamma1 * delta1 + gamma2 * delta2) / M * X[4] +
       (gamma2 + gamma2 * delta2 / M) * X[5] - gamma1 * delta1 / M * X[6] -
       (lambda1 * alpha1 - mu1 * beta1 + lambda2 * alpha2 - mu2 * beta2 - gamma2 * delta2) / M * X[7] -
       (lambda1 * alpha1 - mu1 * beta1 + lambda2 * alpha2 - mu2 * beta2 - gamma1 * delta1) / M * X[8] -
       gamma2 * delta2 / M * X[9]
dX22 = (lambda2 + mu2 + gamma2) * X[2] + gamma1 * X[1] + (2 * lambda2 - 2 * mu2 - 2 * gamma2) * X[5] -
       (lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2) / M * X[5] + 2 * gamma1 * X[4] -
       (lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2 + gamma1 * delta1) / M * X[4] -
       gamma1 * delta1 / M * X[3] - 2 * (lambda2 * alpha2 - mu2 * beta2 - gamma2 * delta2) / M * X[9] +
       2 * (- lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2 - gamma1 * delta1) / M * X[8] -
       2 * gamma1 * delta1 / M * X[7]

dX1111 = 4 * X[6] * X[1] - 6 * X[3] * X[1]^2 + 3 * X[1]^4
dX1112 = X[6] * X[2] + 3 * X[7] * X[1] - 3 * X[3] * X[1] * X[2] - 3 * X[4] * X[1]^2 + 3 * X[1]^3 * X[2]
dX1122 = 2 * X[7] * X[2] - X[3] * X[2]^2 + 2 * X[8] * X[1] - X[1]^2 * X[5] - 4 * X[4] * X[1] * X[2] +
         3 * X[1]^2 * X[2]^2
dX1222 = X[9] * X[1] + 3 * X[8] * X[2] - 3 * X[5] * X[1] * X[2] - 3 * X[4] * X[2]^2 + 3 * X[2]^3 * X[1]
dX2222 = 4 * X[9] * X[2] - 6 * X[5] * X[2]^2 + 3 * X[2]^4

dX111 = (lambda1 - mu1 - gamma1) * X[1] + gamma2 * X[2] + (3 * lambda1 + 3 * mu1 + 3 * gamma1) * X[3] -
        (lambda1 * alpha1 - mu1 * beta1 - gamma1 * delta1) / M * X[3] + 3 * gamma2 * X[4] -
        (lambda1 * alpha1 - mu1 * beta1 - gamma1 * delta1 + gamma2 * delta2) / M * X[4] -
        gamma2 * delta2 / M * X[5] + 3 * (lambda1 - mu1 - gamma1) * X[6] -
        (lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1) / M * X[6] +
        3 * (gamma2 - lambda1 * alpha1 / M - mu1 * beta1 / M - gamma2 * delta2 / M) * X[7] -
        3 * gamma2 * delta2 / M * X[8] - 3 * gamma2 * delta2 / M * dX1122 -
        3 * (lambda1 * alpha1 - mu1 * beta1 - gamma1 * delta1) / M * dX1111 +
        3 * (- lambda1 * alpha1 + mu1 * beta1 + gamma1 * delta1 - gamma2 * delta2) / M * dX1112


dX112 = gamma1 * X[1] - gamma2 * X[2] + (- 2 * gamma1 - gamma1 * delta1 / M) * X[3] +
        (lambda1 + mu1 + gamma1 - 2 * gamma2 - gamma1 * delta1 / M + gamma2 * delta2 / M) * X[4] +
        (gamma2 + gamma2 * delta2 / M) * X[5] + (gamma1 + 2 * gamma1 * delta1 / M) * X[6] +
        (2 * lambda1 - 2 * mu1 + lambda2 - mu2 - 2 * gamma1 - gamma2) * X[7] -
        (lambda1 * alpha1 + mu1 * beta1 - gamma1 * delta1 - 2 * gamma2 * delta2) / M * X[7] +
        (- lambda1 * alpha1 / M - mu1 * beta1 / M - gamma1 * delta1 / M + gamma2 * delta2 / M) * X[8] -
```

```
            gamma2 * delta2 / M * X[9] - gamma1 * delta1 / M * dX1111 + 2 * gamma2 * X[8] +
            (- 2 * lambda1 * alpha1 + 2 * mu1 * beta1 - lambda2 * alpha2 + mu2 * beta2) / M * dX1112 +
            (gamma1 * delta1 + gamma2 * delta2) / M * dX1112 +
            (- 2 * lambda1 * alpha1 + 2 * mu1 * beta1 - lambda2 * alpha2 + mu2 * beta2) / M * dX1122 +
            (2 * gamma1 * delta1 - gamma2 * delta2) / M * dX1122 - 2 * gamma2 * delta2 / M * dX1222

    dX122 = gamma2 * X[2] - gamma1 * X[1] + (- 2 * gamma2 - gamma2 * delta2 / M) * X[5] +
            (lambda2 + mu2 + gamma2 - 2 * gamma1 - gamma2 * delta2 / M + gamma1 * delta1 / M) * X[4] +
            (gamma1 + gamma1 * delta1 / M) * X[3] + (gamma2 + 2 * gamma2 * delta2 / M) * X[9] +
            (2 * lambda2 - 2 * mu2 + lambda1 - mu1 - 2 * gamma2 - gamma1) * X[8] -
            (lambda2 * alpha2 + mu2 * beta2 - gamma2 * delta2 - 2 * gamma1 * delta1) / M * X[8] +
            (- lambda2 * alpha2 / M - mu2 * beta2 / M - gamma2 * delta2 / M + gamma1 * delta1 / M) * X[7] -
            gamma1 * delta1 / M * X[6] - gamma2 * delta2 / M * dX2222 + 2 * gamma1 * X[7] +
            (- 2 * lambda2 * alpha2 + 2 * mu2 * beta2 - lambda1 * alpha1 + mu1 * beta1) / M * dX1222 +
            (gamma2 * delta2 + gamma1 * delta1) / M * dX1222 +
            (- 2 * lambda2 * alpha2 + 2 * mu2 * beta2 - lambda1 * alpha1 + mu1 * beta1) / M * dX1122 +
            2 * gamma2 * delta2 - gamma1 * delta1) / M * dX1122 - 2 * gamma1 * delta1 / M * dX1112

    dX222 = (lambda2 - mu2 - gamma2) * X[2] + gamma1 * X[1] + (3 * lambda2 + 3 * mu2 + 3 * gamma2) * X[5] -
            (lambda2 * alpha2 - mu2 * beta2 - gamma2 * delta2) / M * X[5] + 3 * gamma1 * X[4] -
            (lambda2 * alpha2 - mu2 * beta2 - gamma2 * delta2 + gamma1 * delta1) / M * X[4] -
            gamma1 * delta1 / M * X[3] + 3 * (lambda2 - mu2 - gamma2) * X[9] -
            (lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2) / M * X[9] +
            3 * (gamma1 - lambda2 * alpha2 / M - mu2 * beta2 / M - gamma1 * delta1 / M) * X[8] -
            3 * gamma1 * delta1 / M * X[7] - 3 * gamma1 * delta1 / M * dX1122 -
            3 * (lambda2 * alpha2 - mu2 * beta2 - gamma2 * delta2) / M * dX2222 +
            3 * (- lambda2 * alpha2 + mu2 * beta2 + gamma2 * delta2 - gamma1 * delta1) / M * dX1222

    list(c(dX1, dX2, dX11, dX12, dX22, dX111, dX112, dX122, dX222))
}
```

Then with the above R function, one can numerically solve the first to third order moment dynamics of the stochastic logistic processes explored in this paper with the deSolve package (Soetaert et al., 2010) in R. The "parms" variable is for the users to fill in.

```
library(deSolve)
T <- 100
step <- 0.001
Xinit <- c(1, 0, 1, 0, 0, 1, 0, 0, 0)
parms <- ...
```

```
approx.sol <- ode(times = seq(0, T, step), y = Xinit, method = 'euler',
 func = MomentEquationsQuadraClosure, parms)
approx.sol.mat <- as.data.frame(matrix(approx.sol, nrow = T / step, ncol = 10))
colnames(approx.sol.mat) <- c('X1', 'X2', 'X11', 'X12', 'X22', 'X111', 'X112', 'X122', 'X222')
```

References

Tom M Apostol. Mathematical Analysis. 1974.

Forrest W Crawford, Vladimir N Minin, and Marc A Suchard. Estimation for general birth-death processes. *Journal of the American Statistical Association*, 109(506):730–747, 2014.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

David J Murrell, Ulf Dieckmann, and Richard Law. On moment closures for population dynamics in continuous space. *Journal of Theoretical Biology*, 229(3):421–432, 2004.

Ingemar Nåsell. Moment closure and the stochastic logistic model. *Theoretical Population Biology*, 63 (2):159–168, 2003.

Gordon D Smith. Numerical Solution of Partial Differential Equations. 1965.

Karline Soetaert, Thomas Petzoldt, R Woodrow Setzer, et al. Solving differential equations in R: package desolve. *Journal of Statistical Software*, 33(9):1–25, 2010.

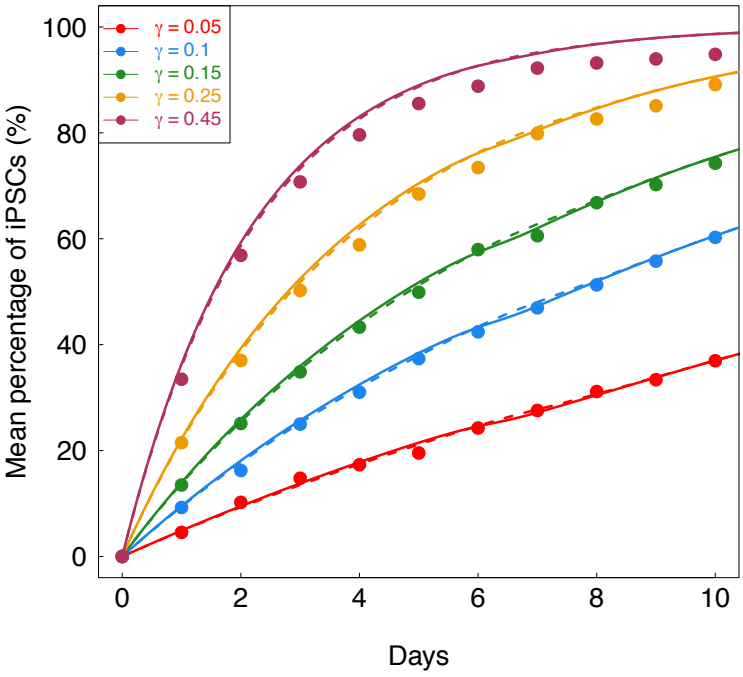
WY Tan and S Piantadosi. On stochastic growth processes with application to stochastic logistic growth. *Statistica Sinica*, 1:527–540, 1991.

Howard M Taylor and Samuel Karlin. *An Introduction to Stochastic Modeling*. Academic Press, 2014.

Supplementary Figures

Figure S1. Related to Figure 2. Consistency between analytical approximation and Gillespie simulations. a. Lines represent the analytical approximation; points denote the results from Gillespie simulation. Colors indicate different reprogramming rate γ, where $\lambda_1 = 1.84, \phi_1 = 0.09, \lambda_2 = 1.72, \phi_2 = 0.09$ in Panel A and $\lambda_1 = 1.72, \phi_1 = 0.06, \lambda_2 = 1.68, \phi_2 = 0.06$ in the Panel B. "——" represents the analytical approximation based on Equation (9) and Equation (10) by setting the fourth central moments to zero and ignoring all higher moments; "- - - -" represents the analytical approximation based on Equation (9) and Equation (11); "•" denotes the results from Gillespie computer simulations (Gillespie, 1977).

A. OSKM + AGi proliferation & apoptosis rates



B. OSKM proliferation & apoptosis rates

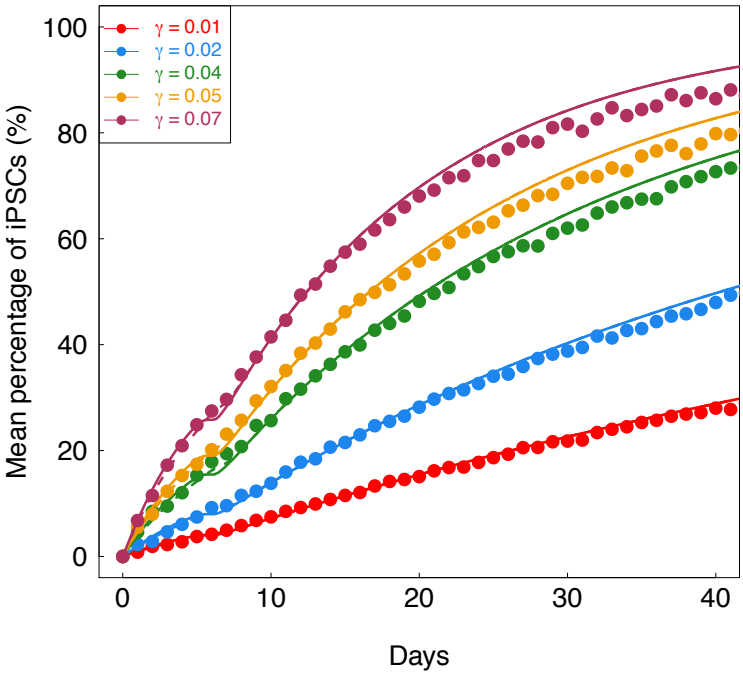
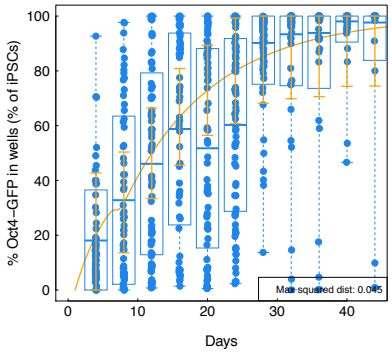
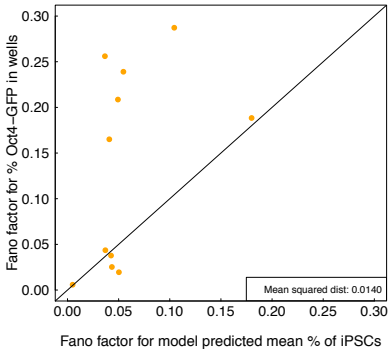


Figure S2. Related to Figure 2. Effects of a log-normally distributed reprogramming rate γ. Panels A and B show the comparison of the mean % iPSCs and Fano factors over time between experimental data and model prediction when $\gamma = 0.08$ is a degenerate constant, Panels C and D when $\gamma \sim LogNormal(log(0.10), 0.25)$, Panels E and F when $\gamma \sim LogNormal(log(0.08), 0.5)$, Panels G and H when $\gamma \sim LogNormal(log(0.08), 0.75)$, Panels I and J when $\gamma \sim LogNormal(log(0.10), 0.8)$. The maximum squared distance for the mean and mean squared distances for Fano factors are marked. In addition, the model-based prediction of Panel A is based on the analytical approximation; all other plots are based on computer simulations with 1,000 replicates.

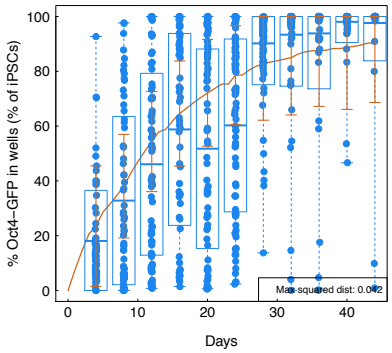
A. OSKM: Homogeneous reprogramming rate



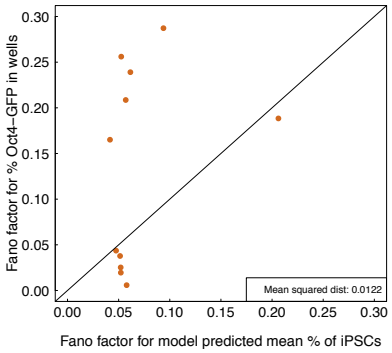
B. OSKM: Homogeneous reprogramming rate



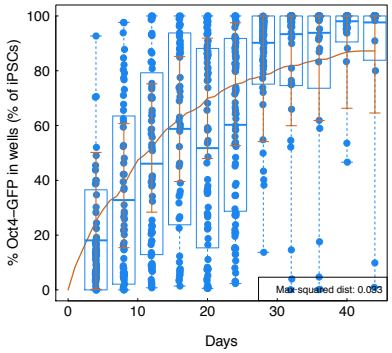
C. OSKM: Reprogramming rate = 0.25



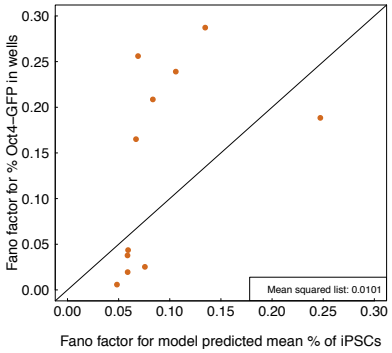
D. OSKM: Reprogramming rate = 0.25



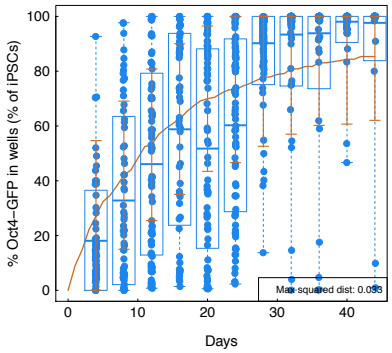
E. OSKM: Reprogramming rate = 0.50



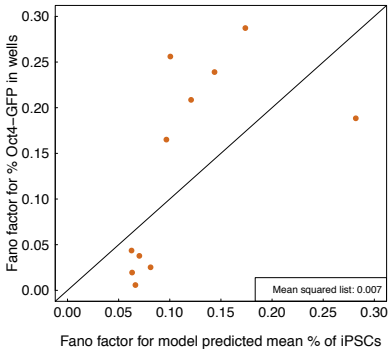
F. OSKM: Reprogramming rate = 0.50



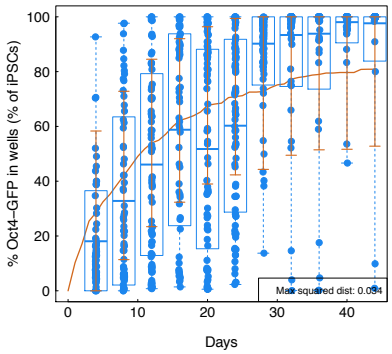
G. OSKM: Reprogramming rate = 0.75



H. OSKM: Reprogramming rate = 0.75



I. OSKM: Reprogramming rate = 1.00



J. OSKM: Reprogramming rate = 1.00

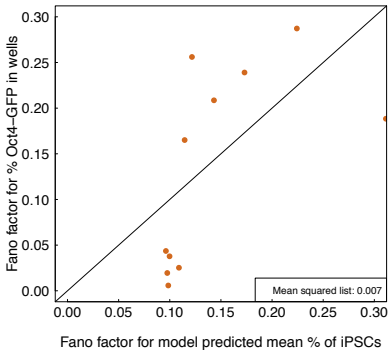
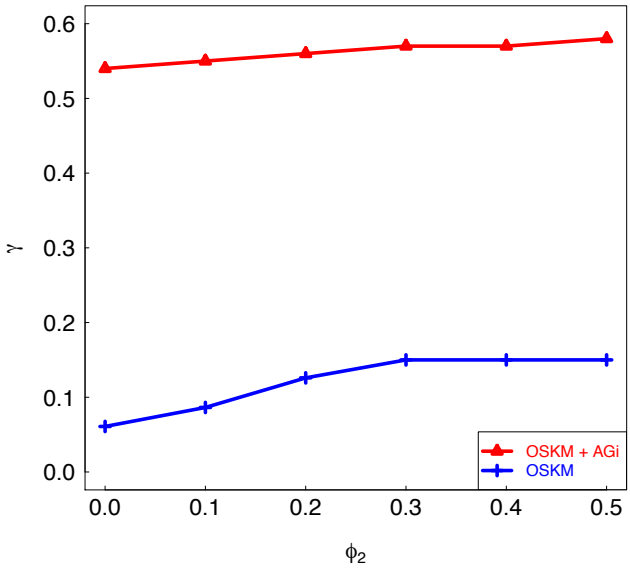
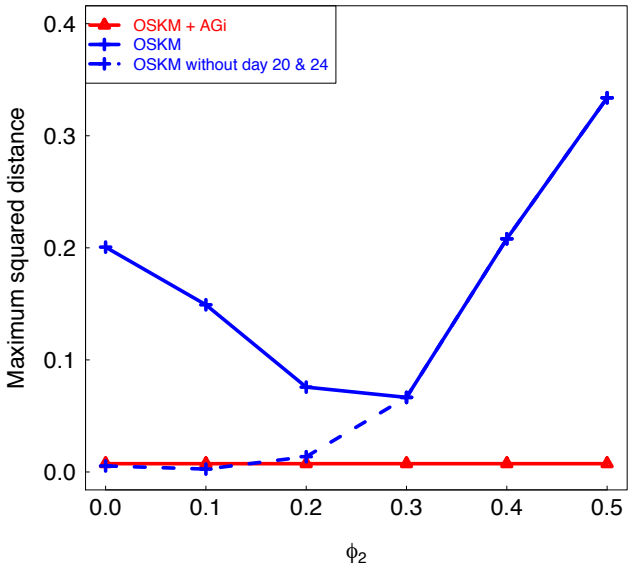


Figure S3. Related to Figure 2. Sensitivity analyses by altering the apoptosis rate for iPS cells, ϕ_2 (A, B); by altering the carrying capacity M (C, D) and by altering the cell splitting strategy (E - F). In this figure, we used the same values for λ_1, λ_2 and ϕ_1 as in Figure 2. Panel A shows the best fitting γ when changing ϕ_2. Panel B shows the mean squared error when changing ϕ_2. Blue: OSKM ; Red: OSKM + AGi. In C, D: Lines: mean trajectory of the percentage of iPS cells using analytical approximations. C. OSKM; D. OSKM + AGi. We then employed different splitting strategies to demonstrate the consistency between different possible splitting strategy in the OSKM condition. Panel E compares the proportion of iPS cells between no splitting and splitting at days 15, 23, and 31. Panel F compares the proportion of iPSCs between no splitting and splitting at days 12, 20, and 31. Panel G compares the proportion of iPSCs between splitting at days 12, 20, and 31 and splitting at days 15, 23, and 31.

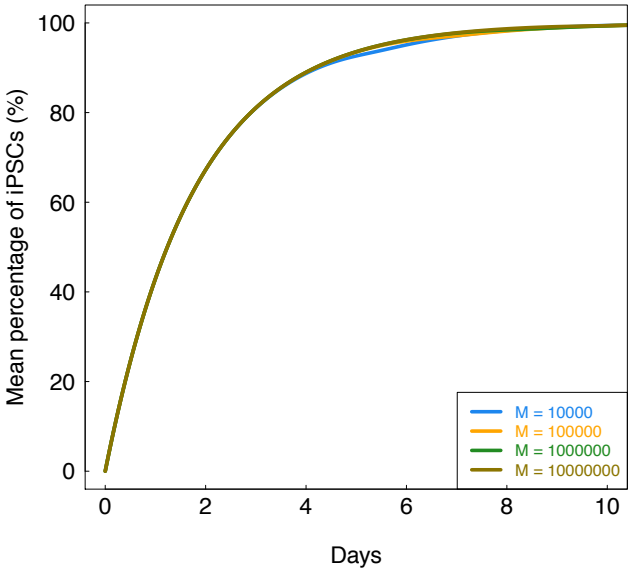
A. Altering apoptosis rate of iPSCs best fitted reprogramming rate



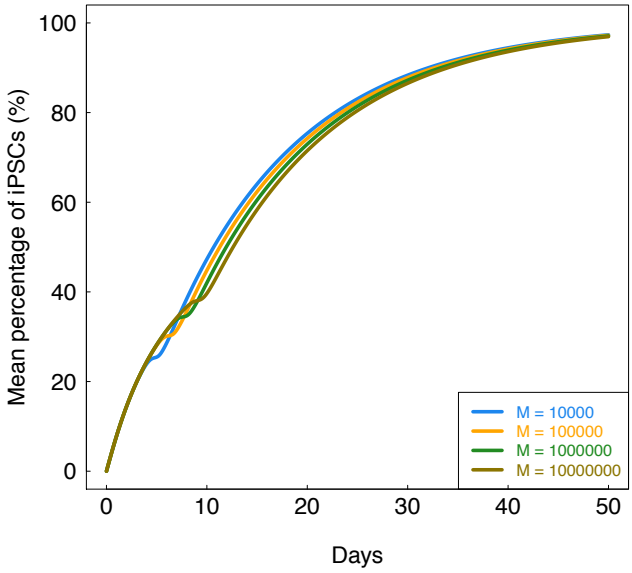
B. Altering apoptosis rate of iPSCs best fitted maximum squared distance



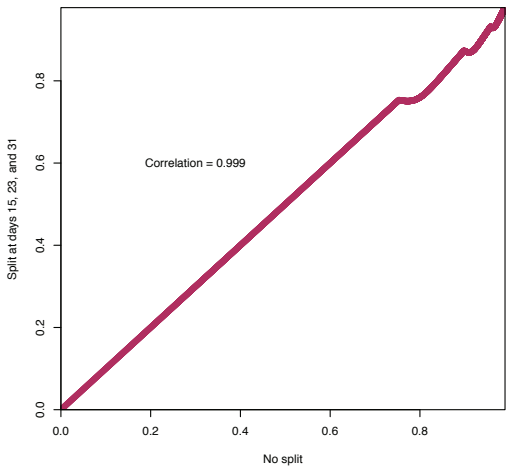
C. OSKM + AGi



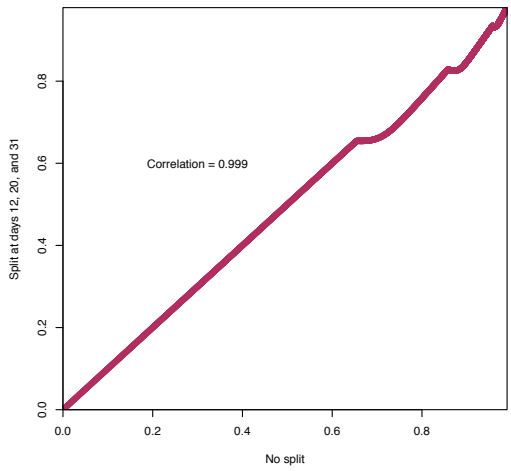
D. OSKM



E.



F.



G.

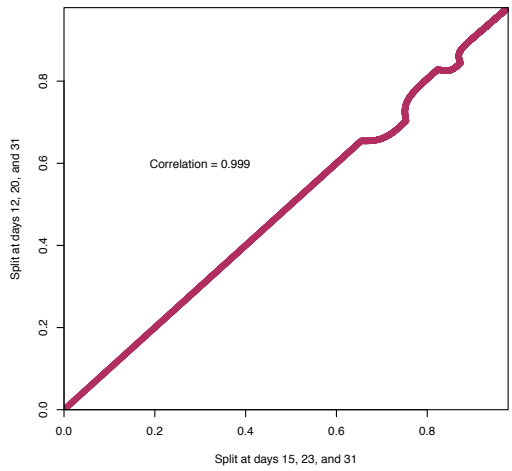
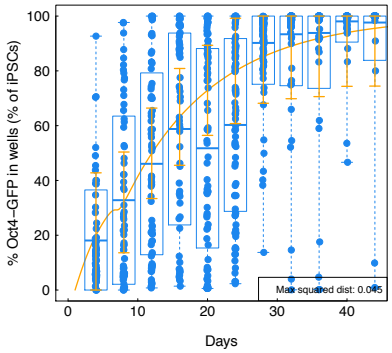
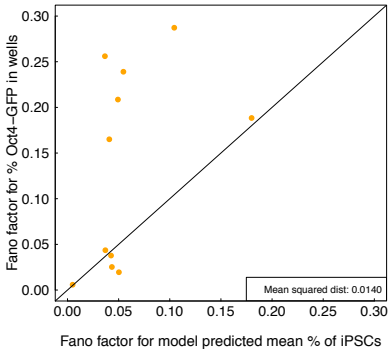


Figure S4. Related to Figure 2. Sensitivity analyses of random proliferation rates for GMPs. We used the value of λ_1 as obtained from **Table 1**. Panels A and B show the comparison between experimental data and model prediction when λ_1 has a log-normal distribution with sd 0, Panels C and D with sd 0.25, Panels E and F with sd 0.50, Panels G and H with sd 0.75, Panels I and J with sd 1.00.

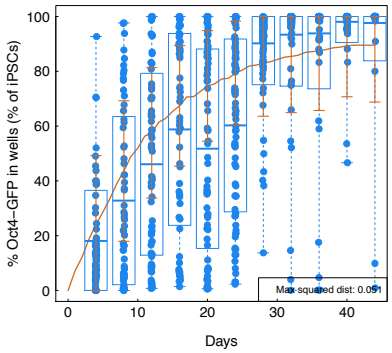
A. OSKM: Homogeneous proliferation rate



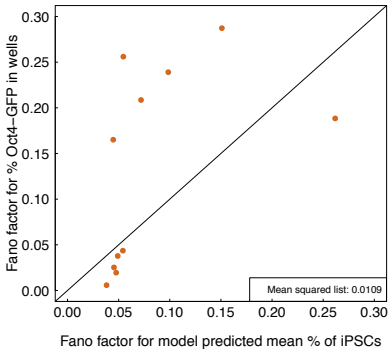
B. OSKM: Homogeneous proliferation rate



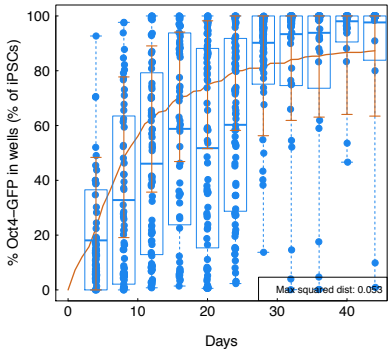
C. OSKM: SD = 0.25



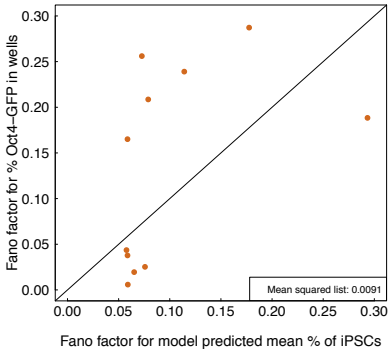
D. OSKM: SD = 0.25



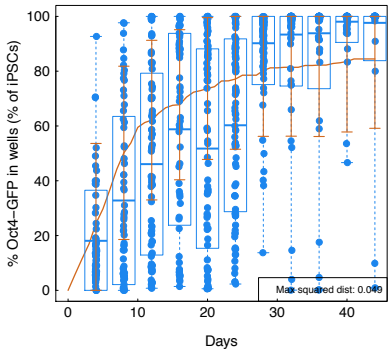
E. OSKM: SD = 0.50



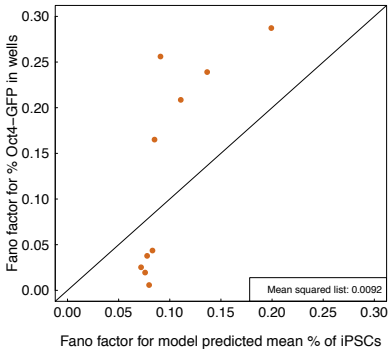
F. OSKM: SD = 0.50



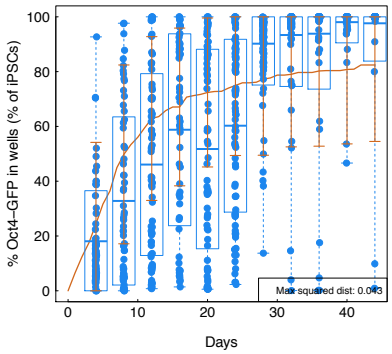
G. OSKM: SD = 0.75



H. OSKM: SD = 0.75



I. OSKM: SD = 1.00



J. OSKM: SD = 1.00

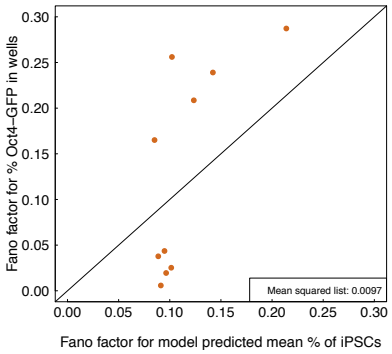
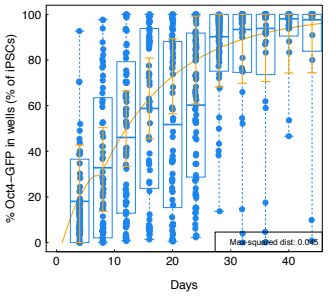
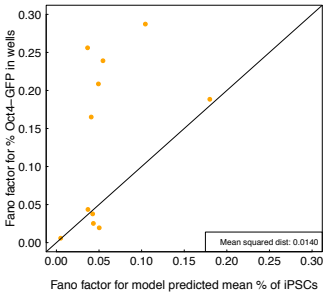


Figure S5. Related to Figure 2. Sensitivity analyses of random apoptosis rates for GMPs. We used the same value of ϕ_1 as obtained from **Table 1**. Panels A and B show the comparison between experimental data and model prediction when ϕ_1 has a log-normal distribution with sd 0, Panels C and D with sd 0.25, Panels E and F with sd 0.50, Panels G and H with sd 0.75, Panels I and J with sd 1.00, Panels K and L with sd 1.05.

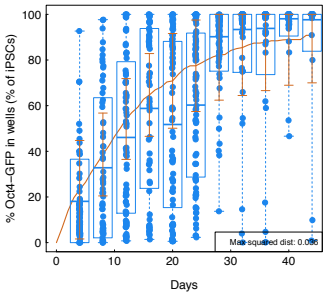
A. OSKM: Homogeneous apoptosis rate



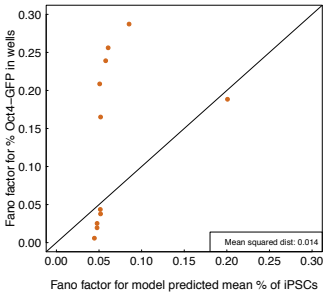
B. OSKM: Homogeneous apoptosis rate



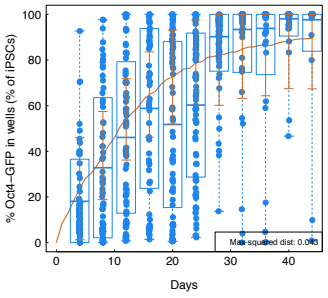
C. OSKM: SD = 0.25



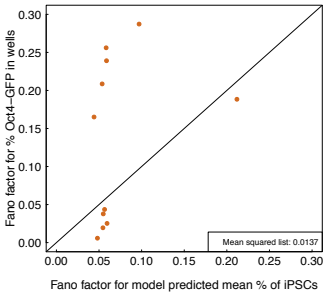
D. OSKM: SD = 0.25



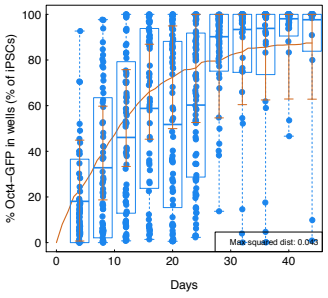
E. OSKM: SD = 0.50



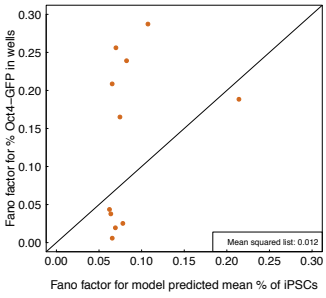
F. OSKM: SD = 0.50



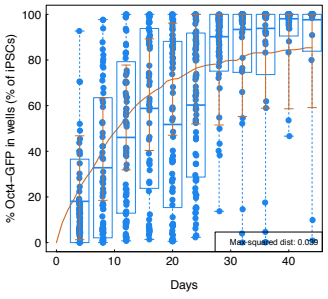
G. OSKM: SD = 0.75



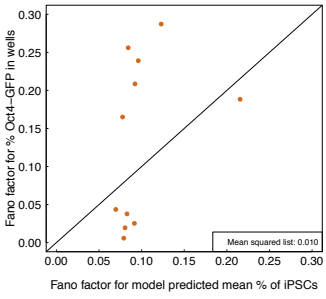
H. OSKM: SD = 0.75



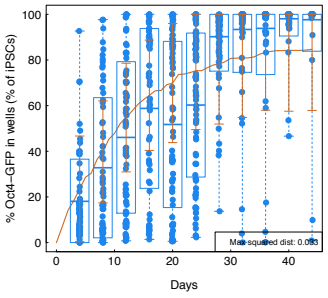
I. OSKM: SD = 1.00



J. OSKM: SD = 1.00



K. OSKM: SD = 1.05



L. OSKM: SD = 1.05

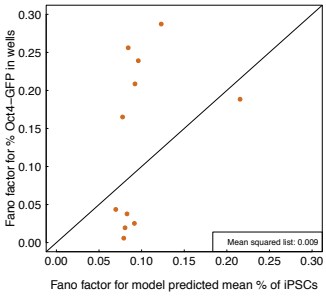


Figure S6: Related to Figure 2. Sensitivity analyses of the effects of increasing proliferation and apoptosis rates of iPSCs on the intrinsic variation for the probability of iPSCs over time. Here we compared the (A) mean difference and (B) mean squared difference between predicted variance and empirical variance by increasing the apoptosis rates of iPSCs while making sure the net growth rate of iPSCs the same. Then evaluating the model fit based on partial time points in the OSKM + AGi condition. The panels show the model-predicted percentage of iPSCs versus the empirical mean calculated from experimental data by estimating the reprogramming rate parameter using a subset of 2 (Panel C) and 3 (D) time points from the data, together with the estimated reprogramming rates γ using 2 to 6 points, respectively (Panel E). Use of more than three time points results in an estimated γ equal to the estimated γ using data from all time points. Evaluating the model fit based on partial time points in the OSKM condition. The panels show the model-predicted percentage of iPSC versus the empirical mean calculated from experimental data by estimating the reprogramming rate parameter using a subset of 2 (Panel F), 3-4 (Panel G), and 5-6 (Panel H) time points from the data, together with the estimated reprogramming rates γ using 2 to 10 points, respectively (Panel I). Use of more than six time points results in an estimated γ equal to the estimated γ using data from all time points.

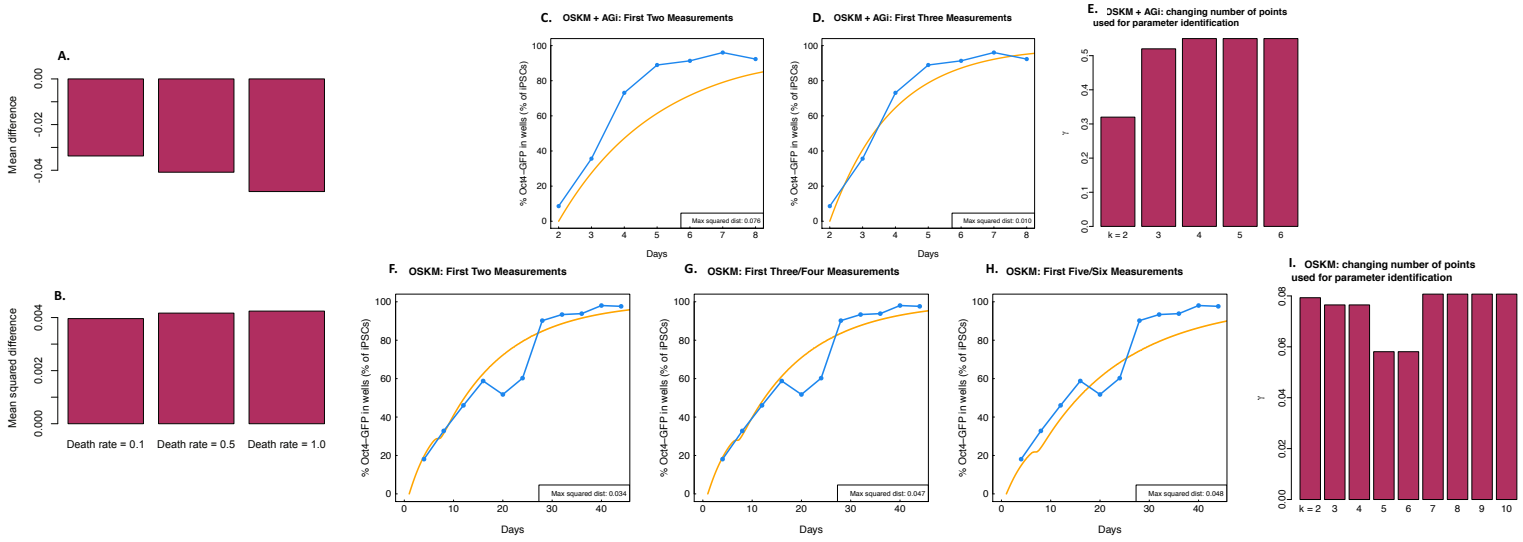
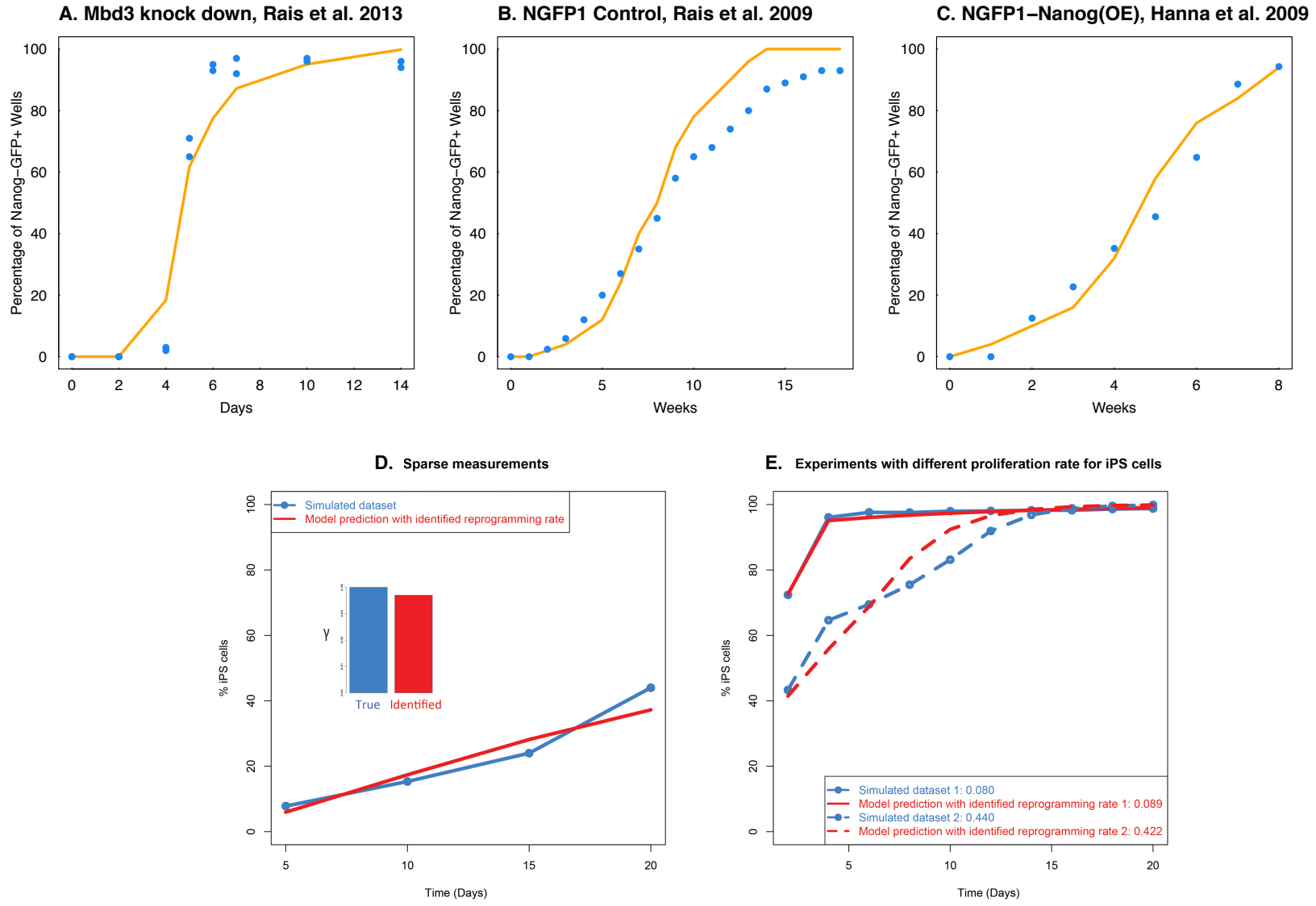

A.
Mean difference
0.00
-0.02
-0.04
B.
Mean squared difference
0.004
0.002
0.000
Death rate = 0.1
Death rate = 0.5
Death rate = 1.0
C. OSKM + AGi: First Two Measurements
% Oct4-GFP in wells (% of iPSCs)
Max squared dist: 0.076
Days
D. OSKM + AGi: First Three Measurements
% Oct4-GFP in wells (% of iPSCs)
Max squared dist: 0.010
Days
E. OSKM + AGi: changing number of points used for parameter identification
γ
k = 2
3
4
5
6
F. OSKM: First Two Measurements
% Oct4-GFP in wells (% of iPSCs)
Max squared dist: 0.034
Days
G. OSKM: First Three/Four Measurements
% Oct4-GFP in wells (% of iPSCs)
Max squared dist: 0.047
Days
H. OSKM: First Five/Six Measurements
% Oct4-GFP in wells (% of iPSCs)
Max squared dist: 0.048
Days
I. OSKM: changing number of points used for parameter identification
γ
k = 2
3
4
5
6
7
8
9
10

Figure S7. Related to Figure 4. Fitting the data from Figure 4 using the log-normal distribution with standard deviation 0.75. The figure shows the model-predicted percentage of replicates having surpassed a certain threshold of percent iPSC at each time point (red line) and the corresponding quantity measured from data (blue dots): A. NGFP1 Mbd3 knock down experiments; B. NGFP1 control experiment; C. NGFP1-NanogOE experiment. D. Evaluating the model prediction with sparse measurements. The reprogramming rate is $\gamma = 0.04$, the proliferation and apoptosis rates for somatic cells are 1.50 and 0.02, and proliferation and apoptosis rates for iPS cells are 0.09 and 0.02 per day. E. Evaluating the model prediction with different proliferation rates of the iPS cells and reprogramming rates. Simulation 1: reprogramming rate $\gamma = 0.08$, proliferation and apoptosis rates for somatic cells are 1.50 and 0.02, and proliferation and apoptosis rates for iPS cells are 3.50 and 0.02; Simulation 2: Reprogramming rate $\gamma = 0.44$, proliferation and apoptosis rates for somatic cells are 1.50 and 0.02, and proliferation and apoptosis rates for iPS cells are 0.09 and 0.02.


A. Mbd3 knock down, Rais et al. 2013
B. NGFP1 Control, Rais et al. 2009
C. NGFP1−Nanog(OE), Hanna et al. 2009
Percentage of Nanog−GFP+ Wells
Days
Weeks
Weeks
D. Sparse measurements
Simulated dataset
Model prediction with identified reprogramming rate
γ
True
Identified
% iPS cells
Time (Days)
E. Experiments with different proliferation rate for iPS cells
Simulated dataset 1: 0.080
Model prediction with identified reprogramming rate 1: 0.089
Simulated dataset 2: 0.440
Model prediction with identified reprogramming rate 2: 0.422
% iPS cells
Time (Days)

Supplementary Tables

Resource	Data Type	Biomarker
Bar-Nur et al. 2014	% Oct4-GFP+ cells/well (GMP)	Oct4-GFP reporter
Vidal et al. 2014	% Oct4-GFP+ cells/colony (MEF)	Oct4-GFP reporter
Rais et al. 2013	% Nanog-GFP+ wells (MEF)	Nanog-GFP reporter and mCherry marker
Hanna et al. 2009	% Nanog-GFP+ wells (B cells)	Nanog-GFP reporter
Smith et al. 2010	Cell counts for each colony (MEF)	Stained for Nanog, E-Cadherin, and Alkaline Phosphatase

Table S1: Experimental data analyzed. Related to Figure 1

	OSKM			OSKM + AGi		
	Cell counts on day 1	Cell counts on day 2	% Live cells	Cell counts on day 1	Cell counts on day 2	% Live cells
Replicate 1	13,000	63,900	96.2	10,400	52,200	96.0
Replicate 2	11,700	66,600	92.6	10,100	58,200	88.3
Replicate 3	13,300	75,900	86.1	13,400	59,100	96.9
Mean	12,666.67	68,800	91.63	11,300	56,500	93.73
Standard Deviation	850.49	6,295.24	5.12	1,824.83	3751.00	4.73
Proliferation rate for GMPs λ_1	1.78			1.67		
Apoptosis rate for GMPs λ_1	0.09			0.06		
SD model prediction	4645.70			4286.96		
Number of cells at certain time	244,140,625 (at day 12) 2.17×10^{32} (at day 44)			390,625 (at day 8)		

Table S2: The number of cells on day 1 and day 2, together with the percentage of live and dead cells, and the estimations for the proliferation and apoptosis rates for GMPs in OSKM and OSKM + AGi conditions assuming a birth-death process without carrying capacity M. Related to Table 1

	Replicate 1	Replicate 2	Replicate 3	Median	Standard Deviation
Simulation 1	580,170	62,546	70,369	70,369	296,617.84
Simulation 2	29,256	72,066	36,068	36,068	45,796.67
Simulation 3	351,400	22,839	96,986	96,986	157,075.00
Simulation 4	20,165	52,488	32,475	32,475	16,313.76

Table S3: The model-predicted standard deviation of the number of cells at day 2 with random λ_1 generated from a lognormal distribution with sd 0.6 from 4 different simulations. Related to Figure 2

	Replicate 1	Replicate 2	Replicate 3	Median	Standard Deviation
Simulation 1	650,281	44,553	82,455	82,455	339,305.50
Simulation 2	80,286	31,650	331,408	80,286	160,874.00
Simulation 3	25,359	75,898	101,225	75,898	38,624.90
Simulation 4	52,225	39,874	88,052	52,225	25,024.12

Table S4: The model-predicted standard deviation of the number of cells at day 2 with random μ_1 generated from a log-normal distribution with sd 1.0 from 4 different simulations. Related to Figure 2