HHS Public Access

Author manuscript

*Phys Biol*. Author manuscript; available in PMC 2019 April 25.

Published in final edited form as:

*Phys Biol*. ; 15(1): 016001. doi:10.1088/1478-3975/aa90e0.

# Cellular reprogramming dynamics follow a simple one-dimensional reaction coordinate

**Sai Teja Pusuluri**[1,2], **Alex H. Lang**[3,4,2], **Pankaj Mehta**[3,4,5], **Horacio E. Castillo**[1,6]

[1]Department of Physics and Astronomy and Nanoscale and Quantum Phenomena Institute, Ohio University, Athens, OH, 45701, USA

[2]These authors contributed equally to this work.

[3]Physics Department, Boston University, Boston, Massachusetts 02215, USA

[4]Center for Regenerative Medicine, Boston University, Boston, MA, 02215

[5]Co-corresponding author: pankajm@bu.edu

[6]Co-corresponding author: castillh@ohio.edu

## Abstract

Cellular reprogramming, the conversion of one cell type to another, induces global changes in gene expression involving thousands of genes, and understanding how cells globally alter their gene expression profile during reprogramming is an open problem. Here we reanalyze time-course data on cellular reprogramming from differentiated cell types to induced pluripotent stem cells (iPSCs) and show that gene expression dynamics during reprogramming follow a simple one-dimensional reaction coordinate. This reaction coordinate is independent of both the time it takes to reach the iPSC state as well as the details of the experimental protocol used. Using Monte-Carlo simulations, we show that such a reaction coordinate emerges from epigenetic landscape models where cellular reprogramming is viewed as a "barrier-crossing" process between cell fates. Overall, our analysis and model suggest that gene expression dynamics during reprogramming follow a canonical trajectory consistent with the idea of an "optimal path" in gene expression space for reprogramming.

## INTRODUCTION

Biology is in the midst of the revolution spearheaded by the pioneering work of Takahashi and Yamanaka[1] on cellular reprogramming showing that it is possible to reprogram mouse embryonic fibroblasts (MEFs) to cells resembling embryonic stem cells (ESCs), commonly called induced pluripotent stem cells (iPSCs), by manipulating the expression of just four transcription factors (TFs). The idea of manipulating small sets of TFs to alter cell fates has proven extremely versatile and it is now possible to create iPSCs from a variety of cell types[2], as well as to perform direct conversions between two differentiated cell types such as MEFs and neurons[3]. Most reprogramming experiments have a similar design[4] (Fig. 1A). The starting cell type (e.g. MEF) is engineered with a construct containing the desired reprogramming genes. These genes are induced at the start of the experiment. After several days, the cell culturing conditions are switched to a medium favorable to the desired cell

type (e.g. stem cell media). At a later time, typically a few weeks, the exogenous genes are turned off. If all goes well, some percentage (often ≈ 0.01 – 1%, but sometimes much larger, approaching even 100%[7]) of cells successfully reprogram to the desired cell type.

Significant progress has been made towards understanding the mechanisms underlying cellular reprogramming[5,6] (which from now on we will use to include both reprogramming to iPSC as well as direct conversion), yet many questions remain. Cellular reprogramming requires global changes in gene expression involving hundreds of transcription factors and thousands of genes, but how cells dynamically alter their gene expression profile during reprogramming is still not well understood. Reprogramming rates seem to depend on the exact protocol used and can be changed by several orders of magnitude through careful genetic manipulations[7,8]. Experiments have also measured whole genome time courses during reprogramming but the high-dimensional nature of the measured trajectories makes them difficult to interpret[9]. Other experiments have examined gene-level events during reprogramming. Buganim et al.[10] analyzed reprogramming dynamics at the single-cell level and concluded that reprogramming initially is probabilistic but ends with a hierarchichal (i.e. ordered), deterministic stage. In contrast, Polo et al.[9] analyzed reprogramming dynamics with both population level and single-cell level measurements and concluded that reprogramming follows an early deterministic phase with many gene changes, followed by an intermediate phase with fewer changes, and ending with a deterministic phase with many gene changes. Recently, Chung et al.[11] measured single cell reprogramming dynamics and proposed that the intermediate phase of reprogramming is a “loosely ordered probabilistic phase” in which the timing between events is probabilistic, but the order of events is relatively deterministic. This highlights the need for a better understanding of gene expression dynamics during reprogramming.

Reprogramming involves global changes in gene expression and hence is intrinsically high dimensional. For this reason, it is common to use dimensional reduction techniques such as Principal Component Analysis (PCA) to project the dynamics onto a low-dimensional sub-space. However, dimensional reductions techniques such as PCA have several key limitations. The principal component vectors have no clear biological interpretation, making it difficult to extract biological meaning from the resulting low-dimensional dynamics. PCA also depends on the type and quality of the data included in the dataset, making it cumbersome to compare dynamical data across experiments and systems.

To overcome these challenges, we introduce a new technique for visualizing high-dimensional reprogramming dynamics, inspired by “epigenetic landscape” models for cellular identity. In Waddington’s original landscape idea[12], each cell type corresponds to the minimum of one of the basins of attraction in an abstract cell identity landscape. This idea has been refined by a variety of researchers, and has yielded a number of insights into the genetic basis of cellular identity[13–26]. Two of us recently proposed a landscape model[27] that takes global gene expression profiles (microarrays or RNA-Seq) and uses techniques inspired by spin physics and the Hopfield model to explicitly construct a cell identity landscape. Without any additional parameters, this model provided an explanation for the existence of partially-reprogrammed cell types and can identify TFs that have been used to successfully reprogram to multiple cell types. In this paper, we extend this previous work to

analyze reprogramming dynamics. Using a new linear-algebra based analysis method inspired by our landscape model, we show that the experimentally observed gene expression dynamics during reprogramming follow a simple, one-dimensional reaction coordinate. We also show that this reaction coordinate emerges in numerical simulations of our landscape model, suggesting that reprogramming can be understood as a "barrier crossing" process between landscape minima.

# I. RESULTS

## A. Mathematical Model and Data Analysis Method

In the epigenetic landscape, cell types correspond to minima of stable basins of attraction, and reprogramming between basins proceeds through stochastic fluctuations resulting from gene expression noise (Fig. 1B). Here, we briefly summarize the relevant features of the landscape model (see Lang et al.[27] for additional details). Before defining the epigenetic landscape, one needs to define the state space. We define the epigenetic state space from the genome-wide expression profiles of natural cell types (i.e. stable cell states either found in vivo or in established growth media) using a curated dataset of microarrays for $p = 63$ mouse cell types and approximately $N \sim 1400$ TFs (see Materials and Methods). These data are summarized in a cell type matrix, $\xi_i^\mu$, whose entries contain the expression level of TF $i$ in cell type $\mu$ (e.g. MEF, ESC). This construction can easily be extended to include genes beyond TFs.

The global gene expression level of TFs in an arbitrary gene expression state (i.e. a perturbed natural cell type or a completely general gene expression) can be summarized using an $N$-dimensional expression state vector $\vec{S}$, whose entries $S_i$ encode the expression level of TFs $i = 1 \ldots N$. Expression levels are treated as continuous variables when analyzing experimental data, and as binary variables which can be either on or off ($S_i = \pm 1$) when performing numerical simulations (see Materials and Methods).

Now that the epigenetic state space has been defined, it is possible to define the epigenetic landscape. In order to describe the epigenetic landscape in geometric terms, we will first introduce a measure of similarity between an arbitrary expression vector $\vec{S}$ and the expression vector, $\vec{\xi^\mu}$, for cell type $\mu$. One common similarity measure in gene expression space is the overlap or dot product,

$$m^\mu = \vec{S} \cdot \vec{\xi^\mu} = \frac{1}{N} \sum_{i=1}^{N} S_i \xi_i^\mu \tag{1}$$

which measures the correlations between an arbitrary state $\vec{S}$, and cell type $\mu$, given by $\vec{\xi^\mu}$. The overlap between cell type $\mu$ and state $\vec{S}$ is 1, −1, or 0 for the cases when $\vec{S}$ and $\vec{\xi^\mu}$ are perfectly correlated, anti-correlated, or uncorrelated, respectively. Since the number of cell types, $p = 63$, is much smaller than the number of TFs, $N \sim 1400$, a given set of overlaps $\{m^1, \cdots, m^p\}$ is consistent with many possible N-dimensional gene expression states $S_i$. A

special case of particular interest is the overlap (or correlation) between two cell types, $\mu$ and $\nu$, given by

$$A^{\mu\nu} = \vec{\xi^{\mu}} \cdot \vec{\xi^{\nu}} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} \xi_i^{\nu} . \quad (2)$$

In practice, the dot product is a poor measure of similarity because cell types are highly correlated with each other. For example, blood cell types share a common core set of gene expression and thus B cells and T cells have a 87% overlap in their gene expression profiles. As a consequence of this, if a gene expression state approaches a certain cell type in the landscape, the overlap may appear to show that it is roughly equally similar to that cell type and to other cell types. In other words, the overlap is too blunt an instrument to follow the evolution of gene expression states in the landscape.

For this reason, it is useful to introduce an alternative measure of similarity, the projections $a^{\mu}$, in which natural cell types have zero similarity with each other. For example, even though the overlap between the gene expressions for B cells and T cells is 87%, the projection of each one of those gene expressions onto the other is exactly zero. The projection of an arbitrary expression profile $S_i$ onto cell type $\mu$ is given by:

$$a^{\mu} = \sum_{\nu=1}^{p} (A^{-1})^{\mu\nu} m^{\nu} \quad (3)$$

where $A^{-1}$ is the inverse of the cell type correlation matrix $A$. According to this definition, the projection would reduce to the overlap if $A$ was the identity matrix, which would correspond to the cell types having exactly zero overlap with each other (i.e. if they were orthogonal). However, as indicated before, overlaps between cell types are actually quite large, and as a consequence, $m^{\mu}$ and $a^{\mu}$ are quite different. We can think of the N-dimensional expression state vector $\vec{S}$ as the sum of two contributions, $\vec{S} = \vec{S^{\perp}} + \vec{S^{\parallel}}$, such that the dot product of $\vec{S^{\perp}}$ with each one of the $p$ cell type states $\vec{\xi^{\mu}}$ is exactly zero, and $S_i^{\parallel} = a_i$ ("the projected state") is a linear combination of the cell type states, given by

$$S_i^{\parallel} = a_i = \sum_{\nu=1}^{p} a^{\mu} \xi_i^{\mu} . \quad (4)$$

This construction has the geometric interpretation depicted in Fig. 1C: $a_i$ is obtained by projecting (i.e. casting a shadow) of $S_i$ onto the hyperplane defined by the $p$ cell types in the matrix $\xi$ (represented as the gray plane) As indicated before, the benefit of this construction is that it automatically accounts for the correlations between cell types: the projection of a B cell with itself is one, while a B cell's projection on T cells is zero, and vice versa.

Projections are essential when constructing landscape models for cellular identity. In Lang et. al[27], it was shown that it is possible to define a Lyapunov function (commonly called an

energy), $H_{basin}$ in expression space. This energy function defines an energy landscape, in which the natural cell types are the global minima. Mathematically, the energy is defined by[27]:

$$H_{\text{basin}} = -\frac{N}{2}\sum_{\mu} m^{\mu}a^{\mu} \tag{5}$$

$$= -\frac{N}{2}\overrightarrow{S}\cdot\overrightarrow{S^{\parallel}} \tag{6}$$

$$= -\frac{1}{2}\sum_{i,j} S_i J_{ij} S_j \tag{7}$$

where Eq. (5) represents the landscape in terms of the projections $a^{\mu}$ and overlaps $m^{\mu}$ with the natural cell types, Eq. (6) represents it in terms of the projected state $\overrightarrow{S^{\parallel}}$, and Eq. (7) represents the landscape in terms of an effective interaction matrix, $J_{ij}$, between TFs, defined as

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\sum_{\nu=1}^{p}\xi_i^{\mu}(A^{-1})^{\mu\nu}\xi_j^{\nu} \tag{8}$$

We emphasize that this Lyapunov function represents an abstract “cellular identity energy surface” characterizing the stability of cell states and cannot be directly related to metabolism or ATP consumption. Additionally, the coefficients $J_{ij}$ do not represent real physical interactions between TFs. Instead, these are effective, symmetric interactions that are chosen so that they generate individual basins of attraction for each natural cell type. We have checked that the landscape construction is robust to the introduction of reasonable amounts of asymmetry in the interactions[27]. This can be seen by noting that in a given cell type (say $\mu = 1$), $S_i = \xi_i^1$ and $a^1 = m^1 = 1$, while the projection on all other cell types is zero, $a^{\nu} = 0$ ($\nu = 2 \ldots p$). Inserting these results into Eq. (5) or into Eq. (6) shows that each cell type is a global minimum with energy $H_{min} = -\frac{N}{2}$. (We notice here that if we had chosen an alternative energy landscape model, based entirely on the overlaps, namely the Hopfield model: $H_{\text{Hopfield}} = -\frac{N}{2}\sum_{\mu} m^{\mu}m^{\mu}$, then the cell types *would not be* the global minima of the landscape.)

The landscape represented by $H_{basin}$ does not incorporate differences in natural cell type stability, especially with regard to stability differences due to external factors such as cell-type specific favorable growth media. These differences are represented by an additional term in the landscape,

$$H = H_{\text{basin}} + H_{\text{culture}},$$
$$H_{\text{culture}} = -\sum_{\mu=1}^{p} b^{\mu} a^{\mu}, \tag{9}$$

where $b^{\mu} > 0$ indicates the presence of a culture medium that favors cell type $\mu$. For example, during reprogramming, when cells are initially grown in MEF culture, only $b^{MEF} \neq 0$, while later in reprogramming, after the media has been changed to ESC culture, only $b^{ESC} \neq 0$. Finally, to incorporate the fact that some transcription factors are overexpressed in the experiments the variables $S_i$ corresponding to overexpressed TFs are locked in the "on" ($S_i = +1$) state.

Our energy function uniquely defines the landscape. However, there are multiple ways to implement dynamics on this landscape. We want to represent the fact that the dynamics is governed by the landscape itself and by stochastic fluctuations due to gene expression noise. A standard way to describe this kind of dynamics is to represent the evolution of the state of the system by random, asynchronous Monte-Carlo updates (Glauber dynamics),[28]. At each simulation time $t$, one (non-locked) TF $i$ is randomly chosen, and the probabilities for its new value $S_i(t+1)$ are

$$P[S_i(t+1)] = \frac{e^{\beta h_i(t) S_i(t+1)}}{e^{\beta h_i(t)} + e^{-\beta h_i(t)}}, \tag{10}$$

with the local field on TF $i$ given by $h_i = -\frac{\partial H}{\partial S_i} = \sum_j J_{ij} S_j$. We have introduced the effective noise parameter $\beta = 1/T$ (i.e. inverse temperature) that controls the level of stochasticity. When $\beta \to \infty$ the update becomes deterministic: the new expression value $S_i(t+1)$ for the chosen transcription factor $i$ takes, with probability 1, the value $S_i(t+1) = +1$ if $h_i > 0$ and the value $S_i(t+1) = -1$ if $h_i < 0$, thus always minimizing the energy. When $\beta \to 0$, the landscape plays no role, and all expression states are equally likely. As long as there are fewer cell types than there are genes there exists a $\beta$ above which all of the cell types are attractors of the dynamics. Additionally, as indicated above, in some of the simulations a subset of the transcription factors is locked at a certain value. Concretely, in many of the simulations we discuss, the OSKM TFs are fixed to have the value +1 ("on").

Currently available static genomic data do not yet provide a clear way to relate biological time to the time variable in the Glauber dynamics. Although in this work we only show results from this particular choice of dynamics, in Ref.[27] it was shown that changing to an alternative dynamics does not significantly modify the results.

## B. MEF Reprogramming Dynamics

We begin by reanalyzing the experimentally available time course data on reprogramming in mice. Fig. 1D shows the first two principal components (PC) for 10 different reprogramming trajectories from MEF to iPSC from multiple labs. In the analysis, we have included partially reprogrammed cells (PRC), which are novel cell states only found during

incomplete reprogramming experiments. The plot shows dynamics projected onto the first two PCs, but in reality this system is high-dimensional and it takes 21 PCs to explain 80% of the variation in the data (see SI Fig. 1 for details). The PCA plot illustrates several important findings. First, reprogramming trajectories seem to group into two distinct clusters, and within each cluster, the starting points (day 0) and ending points (final iPSC) are near each other. Therefore, even for different experimental protocols, reprogramming seems to follow only a few paths. Second, these paths are distinct from partially reprogrammed cells (PRC). While several reprogramming data points seem to be near PRCs, this is an artifact of keeping only two PCs in our visualization. In fact, the PRCs only have a Spearman correlation of 90% with the closest reprogramming data point and approximately 80% correlation with the two closest trajectories. Third, the final state of failed trajectories (trajectories that did not successfully reprogram to iPSCs) is closer to their starting point rather than to iPSCs, suggesting that failed trajectories do not leave the basin of attraction of the initial cell type. While PCA allows easy visualization of the data, the Principal Components have no clear biological meaning, making it difficult to interpret the lower dimensional PCA dynamics.

For a chemical reaction, the trajectory that a system takes on its potential energy landscape when going from reactants to products is usually called the *reaction pathway*, and the *reaction coordinate* is defined as an abstract one dimensional coordinate which measures progress along the reaction pathway. However, in common use the two terms “reaction pathway” and “reaction coordinate” are often conflated. In the case of a chemical reaction, the energy is a function of the positions of all the atoms involved. In the case of the epigenetic landscape, the energy is a function of the expression state. As in the case of a chemical reaction, the reprogramming process can be thought of in terms of a trajectory in the (high-dimensional) epigenetic landscape, going from the initial cell type to the final cell type. We will refer to this trajectory either as the “reaction pathway” or (in a slight abuse of terminology) as the “reaction coordinate”.

In Figures 1E-1G, we have replotted the same time-courses data as in the PCA plots using projections on the starting and ending cell type. As in the PCA plot, the various symbols represent the actual data, while the lines connecting data show the time order of experimental points. In these plots, the starting (ending) states for each trajectory are defined as the initial (final) time point for the corresponding experiment. When calculating projections, the start (end) states replace MEF (ESC) in our cell type matrix $\xi$. This allows us to plot each experiment against its own start and end points. This additional step is necessary because different experiments define MEFs and iPSs differently.

The result of this analysis is shown in Fig. 1E. In contrast to the PCA plot, which contained two clusters (Fig. 1D), the reprogramming trajectories in the projected basis all follow a similar path: a straight line joining the starting cell type with the ending cell type in projection space.

In addition, we can look at how reprogramming trajectories project on other cell types besides the starting and ending cell types. To do so, we introduce a new quantity,

$$a_{\perp} = \sqrt{\sum_{\substack{1 \le \nu \le p \\ \nu \neq (\text{start,end})}} (a^{\nu})^2}, \tag{11}$$

that measures the magnitude of the projections perpendicular to the plane spanned by the starting and ending cell type. This is shown in Fig. 1F. Notice that faster trajectories have a smaller perpendicular projection on the remaining cell types than slower trajectories. Furthermore, the difference in speed between experiments arises largely from the fact that slower trajectories also appear to get stuck at particular points along the reaction coordinate for as long as two weeks. The most important aspect of these results, however, is that the typical magnitude $| a^{\nu}_{\text{typ}} | = a_{\perp} / \sqrt{p-2}$ of each one of the individual projections $a^{\nu}$ for $\nu \neq$ (start, end) is extremely small: since $a_{\perp}$ 0.5, we find that $| a^{\nu}_{\text{typ}} | \lesssim 0.064 \ll 1$. In other words, to a very good approximation we can think of the trajectory as being restricted to the two-dimensional plane spanned by $a_{\text{start}}$ and $a_{\text{end}}$.

We can summarize the results shown in Figs. 1E and 1F as follows: (i) the data for different reprogramming trajectories from different experiments all collapse together and (ii) this unique trajectory is very simple, just a straight line from the initial state to the final state. The data collapse among different experimental trajectories is more remarkable when considering the extreme heterogeneity in reprogramming rates across the plotted experiments. The Polo et al. experiment[9] represents a typical time course with reprogramming taking approximately two weeks, while Rais et al.[8] is the fastest trajectory (8 days) and ST (Samavarchi-Tehrani et al.)[29] is the slowest trajectory in our dataset (30 days). The second observation is also extremely surprising since reprogramming between cell types requires whole-scale reorganization of gene expression and therefore it would be reasonable to expect the trajectory to involve nontrivial changes in a number of dimensions similar to the dimension of the genome, i.e. thousands (if only TFs are considered) or even higher. However, in the projection-based model, the reaction pathway essentially only involves the projections onto the initial and final cell types, and within that two-dimensional space it automatically appears as a straight line without the need for any further transformations.

To compare these experimental trajectories to our mathematical model, it is useful to visualize these data in yet another way. In Fig. 1G, we have replotted the same data taking the z-axis as the energy per TF, which can be calculated directly from gene expression profiles using our landscape construction ($H_{\text{basin}}/N$). In making these plots we have ignored the contributions of the culture terms in Eq 9 to the energy in our model (see SI Figures and Material Methods). Notice that the faster trajectories follow a lower energy path while the slowest trajectory (ST) follows a high energy path and appears to spend time stuck in two different barriers between days 8 and 21. These observation suggest that the experimentally observed reprogramming dynamics are consistent with the idea of a “barrier crossing” process between the starting and ending cell types in a rough landscape (see Fig. 1B).

Further evidence for this barrier-crossing picture comes from numerical simulation using our landscape model (see Materials and Methods). The insets in Figs. 1E-1G show failed and successful (final state has $a_{\text{end}} > 0.8$) reprogramming trajectories from Monte-Carlo

simulations. There is a striking similarity between the model trajectories and experiment. Like in the experiments, successful reprogramming trajectories in our model follow a simple one-dimensional reaction coordinate in the projection space and reprogramming requires crossing a significant energy barrier. Supplementary Figures 2-4 contain more examples of successful and unsuccessful simulation trajectories. Supplementary Figure 12 shows simulation results for the projection onto the initial cell type and onto other cell types, different from the initial and final ones. Those projections onto other cell types are much smaller than the projections onto the initial and final cell types, and can be neglected to a very good approximation, as discussed above in the case of the experiments.

Finally, we note that the reaction coordinate can also be visualized using more traditional measures of distances such as the overlap (dot product) of the gene expression profile with the starting and ending states (see SI Fig. 5A). However, when using overlaps, each experiment has its own starting and ending point, making it hard to compare across experiments. Furthermore, overlaps are unable to discern the “barrier crossing” picture that emerges automatically from using projections (see SI Fig. 5B).

### C. B Cell Reprogramming Dynamics

The previous section considered reprogramming from MEF to iPSC. Here, we extend this analysis to consider two additional reprogramming experiments from B cells to IPSs[30,31]. In the first experiment, the standard Yamanaka reprogramming protocol (OSKM)[1] was used to reprogram B cell to iPSC. Unlike in MEFs, in B cells the OSKM protocol resulted in extremely low reprogramming yields. To increase the reprogramming yield, the protocol was then modified so that OSKM expression was preceded by pulsed expression of CEBP$\alpha$ (abbreviated C+OSKM). This modified protocol significantly increased the reprogramming yield. Figure 2A shows that for both experiments, reprogramming trajectories once again follow a simple reaction coordinate in projection space. Figure 2B extends these plots to the energy vs. reaction coordinate plane. Notice, that in both experiments, the energy of the trajectories first increase and then decrease. The higher yield trajectory (C+OSKM) makes steady progress over the energy barrier, while the low yield trajectory (OSKM) appears to meander through inefficient directions. Thus the reprogramming dynamics of B cells are similar to the reprogramming dynamics of MEF: in all cases reprogramming follows a simple one-dimensional reaction coordinate and can be understood as a barrier crossing process between minima.

The insets in these figures show results from numerical simulations using the landscape model. The simulations reveal a simple reaction coordinate. However, unlike in experiment, the simulated trajectories for the two protocols exhibit nearly identical dynamics. This likely reflects the limitations of the coarse-graining approximation used to construct the landscape model. In the model, TFs are treated as binary variables and all TFs are treated on equal footing – no distinction is made between more promiscuous TFs like CEBP$\alpha$ and more specific downstream factors. Despite these limitations, the phenomenological model still captures the qualitative phenomena seen in the experiments.

The similarity of the reprogramming trajectories from MEFs and B cells suggests a universal reaction coordinate for reprogramming: a straight line connecting the starting and ending

cell type in projection space. In particular, the cells directly convert from the initial cell type to the final cell type without producing any other cell type as an intermediate state. This can be seen best in Figures 2C and 2D where we have plotted reprogramming dynamics from both MEFs and B Cells on the same plots. These experimental data are consistent with numerical simulations using our landscape model which show that reprogramming trajectories always follow a straight line in projection space for both choices of starting cell type.

We also performed direct conversion simulation MEF to cardiomyocyte[46] the simulation results are shown in the supplementary Fig, 13-15. The direct conversion trajectories also support the idea of a universal reaction coordinate and the absence of any intermediate states corresponding to any well-defined cell type. We could not analyze experimental time course data for this direct conversion experiment due to lack of such data for this particular experiment.

### D. Insight into Dynamics from Our Mathematical Model

Given the strong agreement between experiment and the landscape model, it is interesting to ask if the model can provide further insights into reprogramming dynamics beyond those that can be directly gleaned from analyzing experimental time courses. As discussed in the introduction, there is an ongoing debate in the reprogramming literature about the order and organization of gene-level events during reprogramming[9–11]. To address this, we performed detailed simulations that allowed us to probe gene-level events during reprogramming from MEF to iPSC (see Materials and Methods). Experimentally, reprogramming times (as measured by reporters for pluripotency markers) are well described as a Poisson process, implying the existence of a single rate limiting step[7]. Our simulation results support the idea of a single rate limiting step to the turning on of pluripotency markers (see Fig. 3A). In our simulations, the time to turn-on pluripotency markers is calculated by measuring the time it takes a trajectory to have a significant projection on an iPSC state ($a^{end} = 0.3$). Additionally, our simulations show that the later phase of reprogramming (defined as the period of time when trajectories go from having a projection $a^{end} = 0.3$ to $a^{end} = 0.8$) follows a narrowly peaked distribution. Once reprogramming has started, it is very fast: the median time for the later phase is approximately 40 times shorter than the median time for the early phase. Therefore, the rate-limiting step is the early stage of reprogramming. Consistent with experiment[7,8], we find that almost all trajectories eventually reprogram. The most clear evidence for this statement comes from two observations based on the results shown in Fig. 3A: (i) by the end of the simulation, 97.90% of the trajectories have completed the early (slow) stage of reprogramming, and (ii) within the time range of the simulation, the early stage of reprogramming is very well described by a Poisson process, and there is no indication of that behavior changing at longer times. If the first stage of reprogramming is indeed a Poisson process, as it is strongly suggested by our results, then given a long enough time, all cells will reprogram. In agreement with Yamanaka[32] and with Hanna et al.[7], these results are inconsistent with an “elite” model of reprogramming in which only a special subset of cells are amenable to reprogramming.

To ask about the order of gene level events, we probed the gene level dynamics of 10 genes known to be mutually exclusive for either MEFs or ESCs (*Snai1*, *Snai2*, *Prrx1*, *Twist2*, *Twist1* and *Zfp42*, *Nanog*, *Utf1*, *Lin28a*, *Sall4*, respectively) for 224 successful reprogramming trajectories out of a total of 3000 attempts. Recall, that in our model, each gene is represented by a binary variable and can either be 'on' or 'off'. Since the dynamics of our landscape model are stochastic, these genes turn on and off at different values of the reaction coordinate in each of these 224 trajectories. To understand if there is any structure in the gene level dynamics, we counted the percentage of trajectories for which a gene was on at a given reaction coordinate using a moving average over a narrow range of values of the reaction coordinate. The results are shown in Fig. 3B (see SI Figure 10 for an example of non-averaged data). The MEF (ESC) genes gradually turn off (on) over time as expected. Furthermore, the order in which genes turn on and off is relatively stable, at least when averaged over trajectories. In contrast, individual simulation trajectories show much more variability in the order in which genes turn on. However, if we consider individual pairs of TFs, we find that their ordering tends to be consistent with what one would expect from Fig. 3B. For example, *Nanog* turns on before *Sall4* in 58% of trajectories, and *Snai1* turns off before *Twist1* in 71% of trajectories, but for *Twist1* and *Twist2*, there is no clear trend for one or the other to turn off first.

All the qualitative features of our simulations are consistent with the idea that reprogramming trajectories correspond to successful "barrier crossing" between two minima in a landscape. An important qualitative prediction of all barrier crossing is that reprogramming trajectories should be dominated by a small number of optimal paths, with some amount of fluctuations around those paths[33,34]. In particular, the facts that the early phase of reprogramming is well described by a Poisson process, the later phase is described by a narrow distribution of times, and that the median time for the early phase is much longer than the median time for the later phase are all features that would be expected of a simple barrier-crossing process. Furthermore, these simulations show that in a high-dimensional barrier crossing, genes can turn on in a temporally ordered manner (at least when averages over many reprogramming attempts) even though the process is driven entirely by stochasticity.

## II. DISCUSSION

A common metaphor used to describe cellular identity is Waddington's landscape, or the idea of a rugged "epigenetic landscape" in which cell types correspond to minima of basins of attraction. In this picture, cellular reprogramming is envisioned as a process in which one cell type is externally driven out of its basin of attraction, across a barrier, and eventually ends up in the basin of attraction of the desired cell type. Previously, two of us used ideas from spin physics to introduce a model of cellular identity that can be built from genome expression data. In this paper, we reanalyzed experimental data on reprogramming dynamics in terms of this model and found good agreement between the experiments and simulations of the model.

The model provides several interesting insights into reprogramming dynamics. We find that reprogramming dynamics proceed along a simple one-dimensional reaction coordinate and

must cross a significant energy barrier. Somewhat surprisingly, this reaction coordinate is independent of reprogramming dynamics. In terms of projections, we can simply describe the reaction coordinate as a straight line from ($a^{start} = 1$, $a^{end} = 0$) to ($a^{start} = 0$, $a^{end} = 1$). What makes this simple picture especially interesting is that we demonstrated its validity for two different types of reprogramming experiments (MEF or B Cell to iPSC). Based on simulations with our model, we believe that any cellular interconversion (reprogramming or direct conversion), will proceed along a similar, universal, reaction coordinate when described in terms of $a^{start}$, $a^{end}$ (see SI Figures 13-15). We expect this basic picture should also be valid in other organisms such as humans.

An important corollary coming from the shape of this universal reprogramming trajectory is the following: since at all points in the reprogramming trajectory the state of the cell has substantial projections onto the initial cell type, the final cell type, or both, and much smaller projections onto all other cell types (SI Fig. 12), we can conclude that the system does not go through any well-defined intermediate cell type at any point in the process. In particular, for the case of direct conversion (SI Figures 13-15), this leads to the prediction that it *is not* a two-step process of conversion from the initial cell type to iPSC followed by conversion from iPSC to the final cell type.

Our model also gives insight into the ongoing debate about the phases of reprogramming dynamics. A priori, reprogramming dynamics may be either probabilistic or deterministic with respect to both the timing and order of gene level events. Our simulations show the the initial phase of reprogramming follows a Poisson distribution – initiating reprogramming is a rare event. However, once initiated, reprogramming proceeds quickly and efficiently. This is reflected in our simulations by the observation that the dynamics of the reprogramming process at later stages are well described by a narrowly peaked distribution. Furthermore, we find that when averaged over many successful reprogramming trajectories, the order of gene level events are relatively reproducible. Our simulations strongly support Chung et al.[11] description of reprogramming as a “loosely ordered probabilistic process”.

Why have different dynamics experiments led to such drastically different conclusions? So far, each experiment has used different techniques, each of which have their own limitations. GFP reporters (for example[7]) provide precise timing data but are limited to small numbers of genes. Whole genome expression data (for example[9]) provides data on all genes, but both microarrays and RNA-Seq require populations of cells. Finally, single cell gene expression data (for example[10]) provides accurate details of gene expression, but only for a subset of genes (currently 48 with standard Fluidigm chips[10]). Therefore, depending on which technique is utilized, each experimentalist rightfully sees a different picture of reprogramming dynamics. However, viewing reprogramming as a loosely ordered probabilistic process unifies all of these different experimental pictures.

Besides examining the gene level reprogramming dynamics, our model provides a clearer picture of the global mechanism behind reprogramming. One of the most surprising aspects of reprogramming is that the over expression of just a few TFs (out of thousands) can lead to such drastic changes in the global gene expression profile. Our simulations suggest the underlying reason for this is the important role played by culturing conditions. In our model,

inducing the OSKM TFs in MEFs only changes the energy by 0.5% , which at the noise levels considered here, do not lead to any successful reprogramming event. However, by including the effect of cell culture in our simulations, we achieve 7.43% reprogramming rates. This suggests that culturing conditions likely play an important role in dictating reprogramming efficiencies. For example, it is claimed[35] that it is possible to use the OSKM factors, normally used to reprogram to iPSC, to instead reprogram to neuronal progenitors just by changing culture conditions. This highlights an important issue of experimental design for direct conversions to a given cell type. Before one searches for TFs to manipulate, it is essential to understand the correct culturing conditions for the desired cell type. Without the correct medium, direct conversion may prove exceedingly difficult. In our simulations, we have found that the culture term for a given cell type decreases the size of the basin of attraction of all the other cell types. We even find some reprogramming events when we bias the system just by introducing the culture term, without forcing expression of the OSKM TFs (this likely reflects the limitations of the model). However, when we compare simulations of MEF to ESC reprogramming at a certain noise level and for a certain duration, the ones where expression of the OSKM TFs is forced and the ESC culture term is present have a success rate 5 times higher than the ones where the ESC culture term is present but OSKM expression is not forced.

Even though our current models suggest that the growth medium is extremely important for reprogramming and direct conversion dynamics, the biological mechanism that gives rise to this effect remains elusive. The underlying reason for this is that our epigenetic landscape model utilizes a coarse-grained description based on “effective interactions” between TFs and cannot distinguish between direct regulation of epigenetic states by the growth medium, and indirect regulation due to signaling pathways or other regulatory layers. This highlights the need for more detailed, mechanistic models of cell fate and reprogramming dynamics.

The experimental analysis and simulations presented here suggest that reprogramming can be viewed as a “barrier crossing” process in a rugged landscape (see Figure 4). In all barrier crossings, the dynamics are dominated by a few “optimal paths”, suggesting that reprogramming dynamics are likely to be low-dimensional and fairly reproducible at the gene level. A consequence of this picture is the existence of a simple reaction coordinate that describes the progress along the optimal path. If the landscape picture is correct, the existence of a reaction coordinate is likely to be a generic feature of all reprogramming and direct conversion protocols. Directed differentiation is a closely related experimental technique that instead of using TFs to convert between cell types focuses on recapitulating embryonic development through sequences of signaling molecules[36]. It will be interesting to see if projections are also a useful reaction coordinate for directed differentiation experiments.

The results presented here are also likely to be applicable to other systems. Recently, it has been suggested the evolutionary dynamics of viruses such as HIV can also be understood using a Hopfield-inspired landscape model[37]. In evolutionary landscapes, crossing fitness valleys in rugged landscapes can be understood in terms of barrier crossings. For this reason, it is likely that the techniques developed here in the context of cellular reprogramming can be adapted to visualize evolutionary data on fitness landscape dynamics. More generally,

landscapes have proven to be an important tool for furthering our understanding of a variety of other biological problems, including protein folding[38–40]. The intuitions developed in the context of these other problems are also likely to be applicable to cellular reprogramming and, in the future, it will be interesting to explore these connections further.

## III. MATERIALS AND METHODS

### A. Data Analysis

Here we present details of the data analysis. All the experimental data used in this paper are available in the online Supplementary Files.

- SI_Metadata-Cell_Type_basis.txt. List of publicly available microarrays used to define cell types.
- SI_Metadata-Data_Analysis.txt. Data samples analyzed in this paper.
- SI_Cell_Type_Basis-Data_Analysis.txt. Zscore data that defines cell types for the data analysis.
- SI_Cell_Type_Basis-Simulations.txt. Binarized data that defines cell types for the simulations.
- SI_Data_Analysis.txt. Zscore data for the data samples.
- SI_trajectory_1.txt. Example simulation trajectory of reprogramming MEF to ESC.
- SI_trajectory_2.txt. Example simulation trajectory of reprogramming MEF to ESC.

The following abbreviations are used in the Supplementary Files. All GSE and GSM are data identifiers from NCBI GEO except that GSE labels E-MEXP are data identifiers from ArrayExpress. Sample_Name refers to the sample label in SI_Data_Analysis.txt. Plot_Label refers to the abbreviations used in Figures 1 and 2. Paper_Reference is an abbreviated citation of the source data which came from the following papers[8,9,29–31,41–44].

Microarrays were taken from public datasets and come from a variety of different microarray platforms. In order to compare the different platforms, the following analysis was done. The raw microarray data was converted to an expression level as follows. Microarray probe-to-gene map was created with Bioconductor 3.0. All raw microarray files were initially processed by robust mean averaging (RMA) and genes with multiple microarray probes were averaged. Since we were interested in cellular identity, only transcription factors, transcription co-factors, or chromatin remodeling genes were kept (for short hand, referred to as transcription factors, TF, throughout the text)[45].

While the above analysis was done for both experimental data and simulations, from this point on the analysis differed between the two cases. For the experimental data analysis, we only used TFs that were common to all of the different microarray platforms, leaving $N = 994$ TFs. In order to make robust comparisons across platforms the RMA output was converted to a rank order. Next, we wanted to convert this rank order to the z-score of a log-

normal distribution. We converted the rank to a percentile (for N genes, by dividing by N +1), and then this percentile into a normal z-score. For later mathematical convenience, we used a biased estimator (i.e. we normalized by N and not N-1) since then the Euclidean norm of each microarray gene expression was N. Therefore, for the data analysis each sample is described by a Gaussian distribution with a Euclidean norm of $N = 994$.

The $N$ transcription factors (TF) are labeled by Latin indices $i$ and the $p$ cell types are labeled by Greek indices $\mu$. When analyzing experiments, we keep the $N = 994$ TFs common to all of the experimental datasets. Each sample is a Gaussian distribution with mean equal to 0 and Euclidean norm equal to $N$. This implies a standard deviation of $\frac{N}{N-1} \approx 1$. When performing simulations, we use the complete set of $N = 1436$ TFs and each TF is either on (+1) or off (−1).

For the simulations, we followed similar steps to produce continuous TF expression levels for the cell type basis vector. However, in order to reduce the computational cost, we binarized the gene expression so that each TF is either on (+1) or off (−1). We then dropped all TFs that were always on or always off in every cell type, leaving $N = 1436$ TFs for the simulations.

## B. Supplementary Code

We have also included code to aid in understanding the manuscript and as a potential starting point for future work.

- SI_load_paper_data.py. This script loads in the data used in this paper.
- SI_model.py. This script creates the model based on the provided data.
- SI_make_plots.py. This script takes the data and model and creates plots of the time courses as viewed from the reaction coordinates.
- SI_dynamics_code.py. This script implements the dynamics used in the simulations. See below for more details.

## C. Simulations

We performed Monte Carlo (MC) simulations of a system containing $N = 1436$ TFs using the update rule given by Eq. (10), with noise parameter $\beta = 1.62$ (i.e. $T \approx 0.617$). When a culture term was introduced, it was to bias the system towards the ESC cell type, with $b^{\mu} = 0.03$ for $\mu =$ ESC and $b^{\mu'} = 0$ for all other cell types. The dynamics are qualitatively similar for a wide-range of values for $b^{\mu}$ and the quantitative differences will be explored in future work.

Most of the results reported in this paper correspond to simulations where the total number of steps was $t = 10^5$. For the case of simulations of MEF to ESC reprogramming, the OSKM transcription factors were locked “on” for the whole simulation, and the culture term was present from step $t = 5000$ until the end. In this case, 3000 trajectories were simulated, out of which 224 successfully reprogrammed, i.e. the reprogramming rate was 7.43%. For the simulations of B-cell to ESC reprogramming, the protocol was similar, and in this case 205

trajectories reprogrammed successfully out of a total of 3000, corresponding to a reprogramming rate of 6.83%.

In order to obtain additional details about the probability distributions of times associated with the reprogramming, which we show in Figure 3A and SI Figure 11, we performed an additional set of simulations of MEF to ESC reprogramming, with the only change being that the total number of steps was 30 times larger, i.e. $t = 3 \times 10^6$ instead of $t = 10^5$. In this set of much longer simulations, 2937 trajectories out of 3000 successfully reprogrammed from MEF to ESC, which corresponds to a reprogramming rate of 97.90%.

In the supplementary files, we have also included a Python script SI_dynamics_code.py that implements the dynamics. We would like to point out that this script is provided for illustrative purposes only, it would probably require a long time to reproduce all the computer simulations presented in this paper by using the script as provided. The reprogramming simulations discussed in the paper were actually performed using a FORTRAN code that we developed, and which runs much faster than the Python script.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

We thank members of the Boston University Center for Regenerative Medicine (CReM) for extremely useful discussions. STP acknowledges the Condensed Matter and Surface Science (CMSS) program at Ohio University for support through a studentship. AHL was supported by a National Science Foundation Graduate Research Fellowship (NSF GRFP) under Grant No. DGE-1247312. PM was supported by a Simon's Investigator in the Mathematical Modeling in Living Systems. This work was supported in part by Ohio University.

## References

1. Takahashi K and Yamanaka S, Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors, Cell 126, 663 (2006), URL http://www.sciencedirect.eom/science/article/B6WSN-4KM3YVR-1/2/aa5af00da760b7e107387b03941a512d. [PubMed: 16904174]
2. González F, Boué S, and Belmonte JCI, Methods for making induced pluripotent stem cells: reprogramming àla carte, Nat Rev Genet 12, 231 (2011), URL 10.1038/nrg2937. [PubMed: 21339765]
3. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Südhof TC, and Wernig M, Direct conversion of fibroblasts to functional neurons by defined factors, Nature 463, 1035 (2010), URL 10.1038/nature08797. [PubMed: 20107439]
4. Takahashi K, Okita K, Nakagawa M, and Yamanaka S, Induction of pluripotent stem cells from fibroblast cultures, Nat. Protocols 2, 3081 (2007), URL 10.1038/nprot.2007.418. [PubMed: 18079707]
5. Yamanaka S, Induced Pluripotent Stem Cells: Past, Present, and Future, Cell stem cell 10, 678 (2012), URL http://linkinghub.elsevier.com/retrieve/pii/S1934590912002378. [PubMed: 22704507]
6. Xu J, Du Y, and Deng H, Direct Lineage Reprogramming: Strategies, Mechanisms, and Applications, Cell Stem Cell 16, 119 (2015), URL http://www.sciencedirect.com/science/article/pii/S1934590915000144. [PubMed: 25658369]
7. Hanna J, Saha K, Pando B, van Zon J, Lengner CJ, Creyghton MP, van Oudenaarden A, and Jaenisch R, Direct cell reprogramming is a stochastic process amenable to acceleration, Nature 462, 595 (2009), URL 10.1038/nature08592. [PubMed: 19898493]

8. Rais Y, Zviran A, Geula S, Gafni O, Chomsky E, Viukov S, Mansour AA, Caspi I, Krupalnik V, Zerbib M, et al., Deterministic direct reprogramming of somatic cells to pluripotency, Nature 502, 65 (2013), URL 10.1038/nature12587. [PubMed: 24048479]
9. Polo JM, Anderssen E, Walsh RM, Schwarz BA, Nefzger CM, Lim SM, Borkent M, Apostolou E, Alaei S, Cloutier J, et al., A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells, Cell 151, 1617 (2012), URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412014249. [PubMed: 23260147]
10. Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, and Jaenisch R, Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and, a Late Hierarchic Phase, Cell 150, 1209 (2012), URL http://www.sciencedirect.com/science/article/pii/S0092867412010215. [PubMed: 22980981]
11. Chung K-M, Kolling IV FW, Gajdosik MD, Burger S, Russell AC, and Nelson CE, Single Cell Analysis Reveals the Stochastic Phase of Reprogramming to Pluripotency Is an Ordered Probabilistic Process, PLoS ONE 9, e95304 EP (2014), URL http://dx.doi.org/10.1371%2Fjournal.pone.0095304. [PubMed: 24743916]
12. Waddington CH, The Strategy of the Genes A Discussion of Some Aspects of Theoretical Biology (Allen and Unwin, London, 1957).
13. Kauffman SA, The Origins of Order: Self-Organization and Selection in Evolution (Oxford University Press, New York City, 1993).
14. Enver T, Pera M, Peterson C, and Andrews PW, Stem Cell States, Fates, and the Rules of Attraction, Cell Stem Cell 4, 387 (2009), URL http://linkinghub.elsevier.com/retrieve/pii/S1934590909001659. [PubMed: 19427289]
15. Huang S, Reprogramming cell fates: reconciling rarity with robustness, BioEssays 31, 546 (2009), URL 10.1002/bies.200800189. [PubMed: 19319911]
16. Wang J, Xu L, Wang E, and Huang S, The Potential Landscape of Genetic Circuits Imposes the Arrow of Time in Stem Cell Differentiation, Biophysical Journal 99, 29 (2010), URL http://www.sciencedirect.com/science/article/pii/S0006349510004248. [PubMed: 20655830]
17. Zhou JX and Huang S, Understanding gene circuits at cell-fate branch points for rational cell reprogramming, Trends in Genetics 27, 55 (2011), URL http://linkinghub.elsevier.com/retrieve/pii/S0168952510002222. [PubMed: 21146896]
18. Zhou JX, Brusch L, and Huang S, Predicting Pancreas Cell Fate Decisions and Reprogramming with a Hierarchical Multi-Attractor Model, PLoS ONE 6, e14752 (2011), URL http://dx.doi.org/10.1371%2Fjournal.pone.0014752. [PubMed: 21423725]
19. Wang J, Zhang K, Xu L, and Wang E, Quantifying the Waddington landscape and biological paths for development and differentiation, Proceedings of the National Academy of Sciences 108, 8257 (2011), URL http://www.pnas.org/content/108/20/8257.abstract.
20. Huang S, The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?, BioEssays 34, 149 (2012), URL 10.1002/bies.201100031. [PubMed: 22102361]
21. Li C and Wang J, Quantifying Cell Fate Decisions for Differentiation and Reprogramming of a Human Stem Cell Network: Landscape and Biological Paths, PLoS Comput Biol 9, e1003165 EP (2013), URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1003165. [PubMed: 23935477]
22. Banerji CRS, Miranda-Saavedra D, Severini S, Widschwendter M, Enver T, Zhou JX, and Teschendorff AE, Cellular network entropy as the energy potential in Waddington's differentiation landscape, Sci. Rep 3 (2013), URL 10.1038/srep03039.
23. Heinaniemi M, Nykter M, Kramer R, Wienecke-Baldacchino A, Sinkkonen L, Zhou JX, Kreisberg R, Kauffman SA, Huang S, and Shmulevich I, Gene-pair expression signatures reveal lineage control, Nat Meth 10, 577 (2013), URL 10.1038/nmeth.2445.
24. Xu L, Zhang K, and Wang J, Exploring the Mechanisms of Differentiation, Dedifferentiation, Reprogramming and Transdifferentiation, PLoS ONE 9, e105216 EP (2014), URL http://dx.doi.org/10.1371%2Fjournal.pone.0105216. [PubMed: 25133589]
25. Li C and Wang J, Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle, Proceedings of the National Academy of Sciences 111, 14130 (2014), URL http://www.pnas.org/content/111/39/14130.abstract.

26. Zhang B and Wolynes PG, Stem cell differentiation as a many-body problem, Proceedings of the National Academy of Sciences 111, 10185 (2014), URL http://www.pnas.org/content/111/28/10185.abstract.
27. Lang AH, Li H, Collins JJ, and Mehta P, Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes, PLoS Comput Biol 10, e1003734 EP (2014), URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1003734. [PubMed: 25122086]
28. Amit DJ, Modelling Brain Function: The World of Attractor Neural Networks (Cambridge University Press, New York, NY, USA, 1992).
29. Samavarchi-Tehrani P, Golipour A, David L, Sung H.-k., Beyer TA, Datti A, Woltjen K, Nagy A, and Wrana JL, Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming, Cell Stem Cell 7, 64 (2010), URL http://www.sciencedirect.com/science/article/pii/S1934590910001700. [PubMed: 20621051]
30. Di Stefano B, Sardina JL, van Oevelen C, Collombet S, Kallin EM, Vicent GP, Lu J, Thieffry D, Beato M, and Graf T, C/EBPalpha poises B cells for rapid reprogramming into induced pluripotent stem cells, Nature 506, 235 (2014), URL 10.1038/nature12885. [PubMed: 24336202]
31. Di Stefano B, Collombet S, and Graf T, Time-resolved gene expression profiling during reprogramming of C/EBP-pulsed B cells into iPS cells, Scientific Data 1, EP (2014), URL 10.1038/sdata.2014.8.
32. Yamanaka S, Elite and stochastic models for induced pluripotent stem cell generation, Nature 460, 49 (2009), URL 10.1038/nature08180. [PubMed: 19571877]
33. Roma DM, O'Flanagan RA, Ruckenstein AE, Sengupta AM, and Mukhopadhyay R, Optimal path to epigenetic switching, Physical Review E 71, 011902 (2005), URL http://link.aps.org/doi/10.1103/PhysRevE.71.011902.
34. Mehta P, Exponential sensitivity of noise-driven switching in genetic networks, Physical Biology 5 (2008), URL http://stacks.iop.org/1478-3975/5/i=2/a=026005.
35. Kim J, Efe JA, Zhu S, Talantova M, Yuan X, Wang S, Lipton SA, Zhang K, and Ding S, Direct reprogramming of mouse fibroblasts to neural progenitors, Proceedings of the National Academy of Sciences 108, 7838 (2011), URL http://www.pnas.org/content/108/19/7838.abstract.
36. Wilson AA, Ying L, Liesa M, Segeritz C-P, Mills JA, Shen SS, Jean J, Lonza GC, Liberti DC, Lang AH, et al., Emergence of a Stage-Dependent Human Liver Disease Signature with Directed Differentiation of Alpha-1 Antitrypsin-Deficient iPS Cells, Stem Cell Reports pp. 873–85 (2015), URL http://www.sciencedirect.com/science/article/pii/S2213671115000776. [PubMed: 25843048]
37. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, and Chakraborty AK, Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design, Immunity 38, 606 (2013), URL http://www.sciencedirect.com/science/article/pii/S1074761313001076. [PubMed: 23521886]
38. Bryngelson JD, Onuchic JN, Socci ND, and Wolynes PG, Funnels, pathways, and the energy landscape of protein folding: A synthesis, Proteins: Structure, Function, and Bioinformatics 21, 167 (1995), URL 10.1002/prot.340210302.
39. Onuchic J, Luthey-Schulten Z, and Wolynes PG, Theory of Protein Folding: The Energy Landscape Perspective, Annual Review of Physical Chemistry 48, 545 (1997), URL 10.1146/annurev.physchem.48.1.545.
40. Onuchic J and Wolynes PG, Theory of protein folding, Current Opinion in Structural Biology 14, 70 (2004), URL http://www.sciencedirect.com/science/article/pii/S0959440X04000107. [PubMed: 15102452]
41. Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, and Plath K, Role of the Murine Reprogramming Factors in the Induction of Pluripotency, Cell 136, 364 (2009), URL http://linkinghub.elsevier.com/retrieve/pii/S0092867409000075. [PubMed: 19167336]
42. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, and Meissner A, Dissecting direct reprogramming through integrative genomic analysis, Nature 454, 49 (2008), URL 10.1038/nature07056. [PubMed: 18509334]
43. Koga M, Matsuda M, Kawamura T, Sogo T, Shigeno A, Nishida E, and Ebisuya M, Foxd1 is a mediator and indicator of the cell reprogramming process, Nat Commun 5 (2014), URL 10.1038/ncomms4197.

44. Fluri DA, Tonge PD, Song H, Baptista RP, Shakiba N, Shukla S, Clarke G, Nagy A, and Zandstra PW, Derivation, expansion and differentiation of induced pluripotent stem cells in continuous suspension cultures, Nat Meth 9, 509 (2012), URL 10.1038/nmeth.1939.

45. Zhang H-M, Liu T, Liu C-J, Song S, Zhang X, Liu W, Jia H, Xue Y, and Guo A-Y, AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors, Nucleic Acids Research (2014), URL http://nar.oxfordjournals.org/content/early/2014/09/27/nar.gku887.abstract.

46. Ieda M, D Fu J, Delgado-Olguin P, Vedantham V, Hayashi Y, Bruneau BG, and Srivastava D, Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors, Cell (2010), URL http://linkinghub.elsevier.com/retrieve/pii/S0092867410007713.

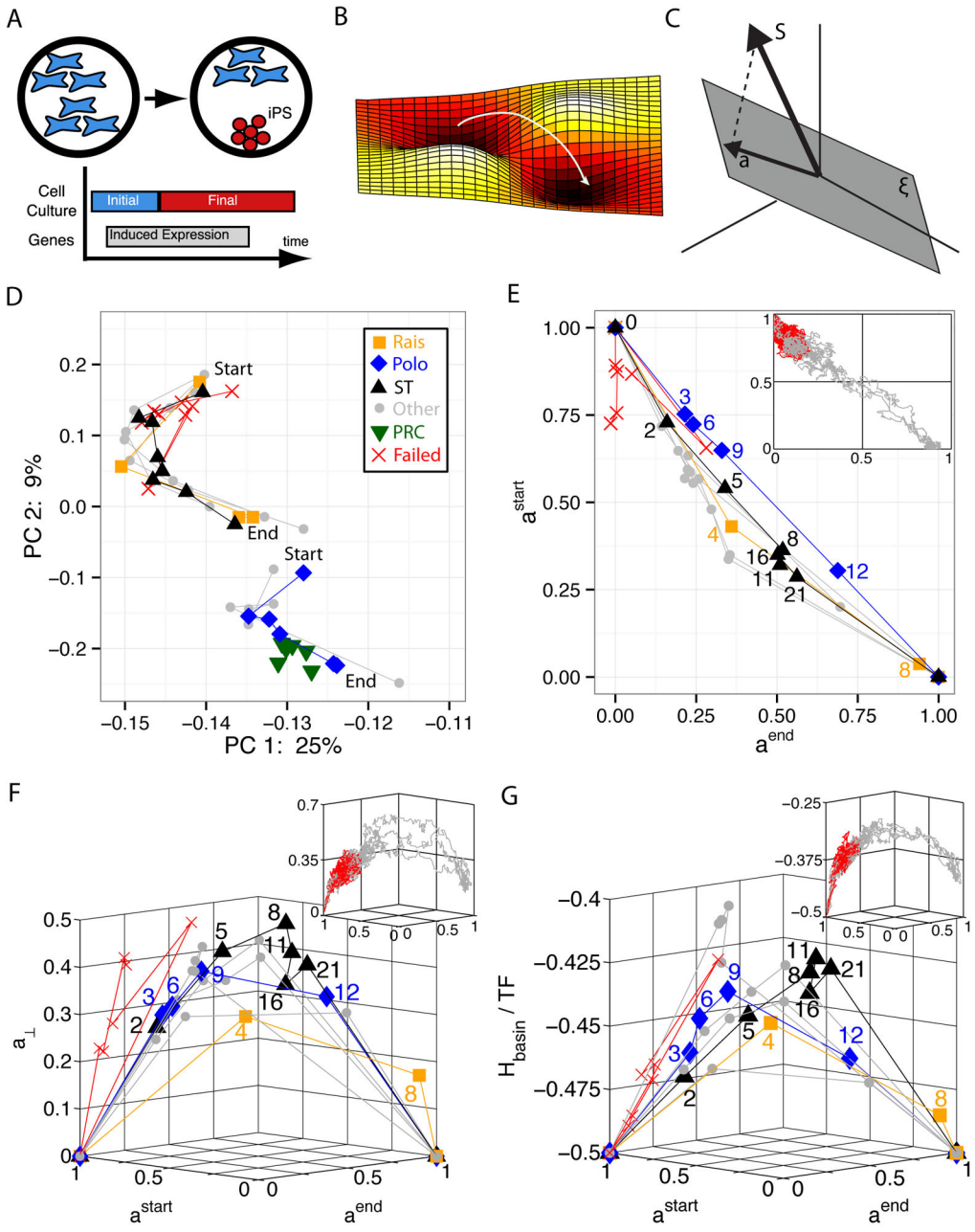


**FIG. 1: Cellular Reprogramming Reaction Coordinate.**

A. Transient expression of reprogramming genes plus switching culturing conditions probabilistically leads to the desired cell type. B. Reprogramming is commonly described as the crossing of a barrier in a high-dimensional landscape. C. Our proposed cellular identity landscape is based on the projection, *a*, of an arbitrary gene expression, *S,* onto the subspace (gray plane) spanned by the natural cell types, $\vec{\xi^{\mu}}$. D. Principal component analysis (PCA) of reprogramming from mouse embryonic fibroblasts (MEF) to induced pluripotent stem cells (iPSC) with start marking day 0 and end marking iPSC. Rais[8], Polo[9], and ST (Samavarchi-Tehrani)[29] are three successful trajectories in which the explicit time in days is labeled on

plots E, F, and G. Other represents additional successful trajectories, PRC are partially reprogrammed cells, and failed trajectories do not reprogram. E. Projection onto $a^{start}$ (MEF) and $a^{end}$ (iPSC) only. All successful trajectories follow a simple reaction coordinate in projection space, a straight line from ($a^{start} = 1$, $a^{end} = 0$) to ($a^{start} = 0$, $a^{end} = 1$). Insets in E, F, and G are simulation data with failed trajectories in red and successful trajectories in gray. See SI Fig. 2for larger version of simulations. F. Measure of projection on all other cell types, $a_{\perp}$ vs. the reaction coordinate. See SI Fig. 3 for larger version of simulations. G. Energy landscape of basins of attraction, $H_{\text{basin}}$, per transcription factor (TF) vs. reaction coordinate. See SI Fig. 4 for larger version of simulations.

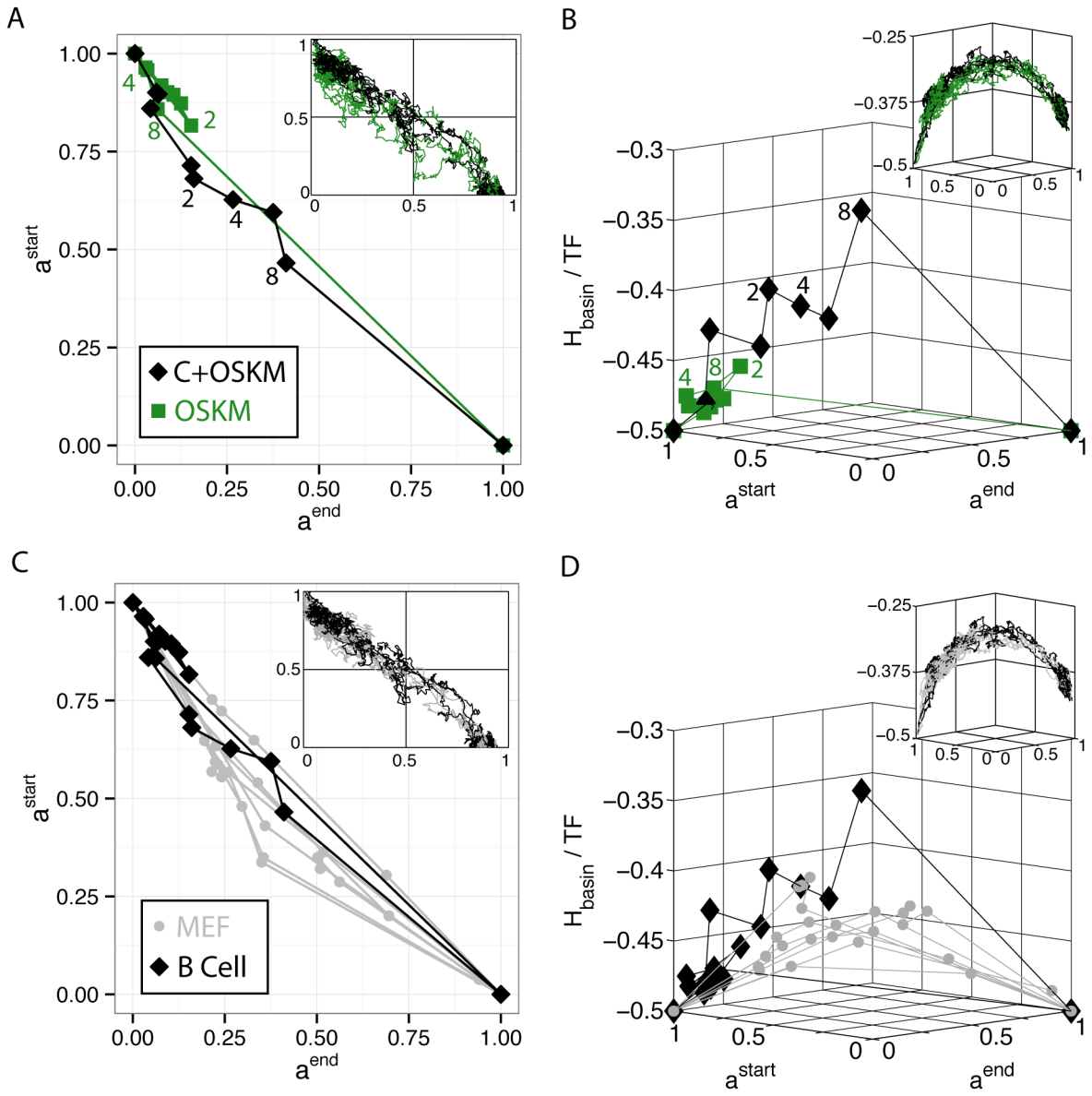


**FIG. 2: Universal Reaction Coordinate.**

A. Cellular reprogramming from $a^{start}$ (B Cells) to $a^{end}$ (iPSC) by Di Stefano et al.[30]. OSKM is the standard Yamanaka protocol, while C+OSKM is a pulse of C/EBPα followed by OSKM which led to higher reprogramming yield. All insets are simulation data of same data shown in main figure. See SI Fig. 6 for larger version of simulations. B. Energy landscape of basins of attraction, $H_{\text{basin}}$, per transcription factor (TF) vs. reaction coordinate. See SI Fig. 7 for larger version of simulations. C. Data collapse of trajectories to $a^{start}$ vs. $a^{end}$ for both MEF to iPSC (gray) and B Cell to iPSC (black). See SI Fig. 8 for larger version of simulations. D. Data collapse of trajectories when viewed as energy vs. reaction coordinate. See SI Fig. 9 for larger version of simulations

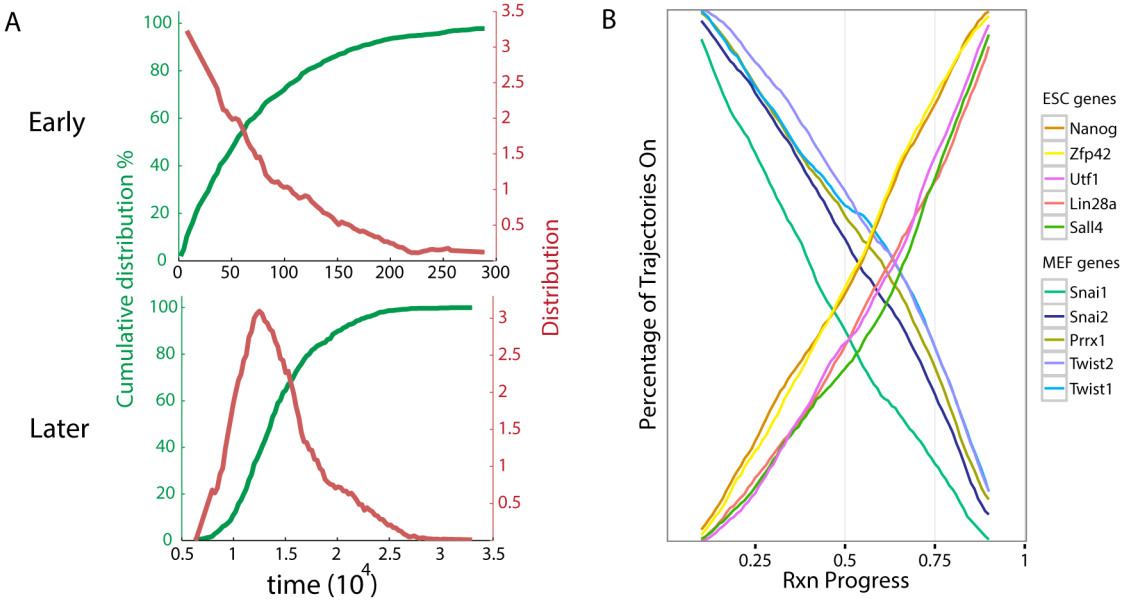


**FIG. 3: Nature of Reprogramming Dynamics in the Landscape Model.**

A. Cumulative distributions of timing show that the early ($a^{end}$ = 0 to $a^{end}$ = 0.3) and later ($a^{end}$ = 0.3 to $a^{end}$ = 0.8) stages of reprogramming are respectively a Poisson and a narrowly peaked distribution. See SI Figure 11 for early ($a^{end}$ = 0 to $a^{end}$ = 0.3), middle ($a^{end}$ = 0.3 to $a^{end}$ = 0.7) and late ($a^{end}$ = 0.7 to $a^{end}$ = 0.8) phases of reprogramming as Poisson, narrowly peaked and narrowly peaked distributions respectively. In order to study the complete timing distribution, the data shown here and in SI Figure 11 were obtained in a simulation of duration $t = 3 \times 10^6$ MC steps, which is 30 times longer than the simulations reported on in all other figures. B. Percentage of trajectories in which a gene is on vs. reaction coordinate. Data shown is a moving average of MEF (ESC) genes turning off (on) over time. See SI Figure 10 for example of non-averaged data.

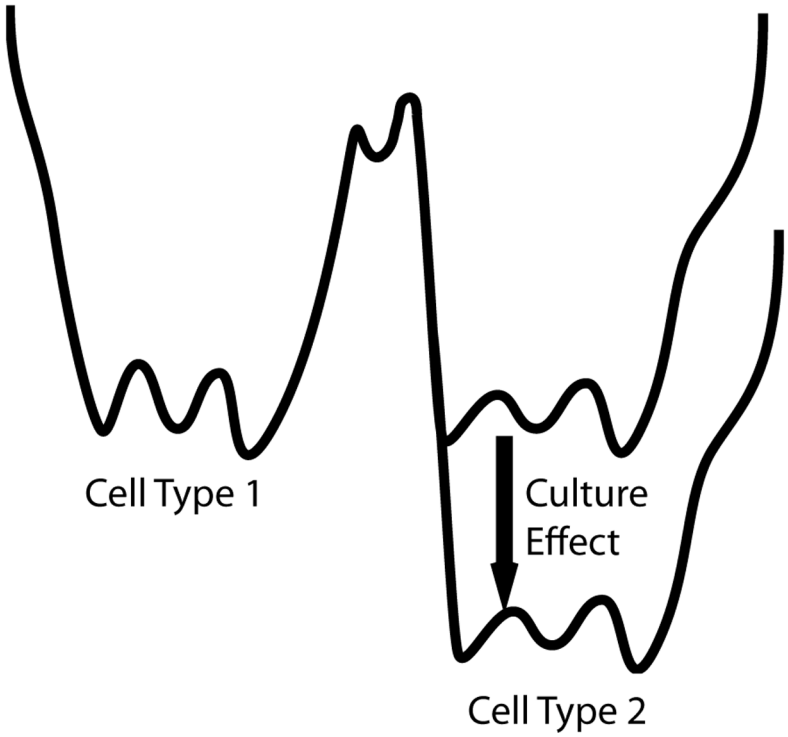


**FIG. 4: Culture Schematic.**

The correct culture conditions plays an essential role in reprogramming by stabilizing the final cell type.