Algorithm for cellular reprogramming

Scott Ronquist[a], Geoff Patterson[b], Lindsey A. Muir[c], Stephen Lindsly[a], Haiming Chen[a], Markus Brown[d], Max S. Wicha[e], Anthony Bloch[f], Roger Brockett[g], and Indika Rajapakse[a,f,1]

[a]Department of Computational Medicine and Bioinformatics, Medical School, University of Michigan, Ann Arbor, MI 48109; [b]Department of Curriculum Design, IXL Learning, Raleigh, NC 27560; [c]Department of Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109; [d]Department of Biological Sciences, University of Maryland, College Park, MD 20742; [e]Department of Hematology/Oncology, University of Michigan, Ann Arbor, MI 48109; [f]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109; and [g]John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA 02138

Edited by Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle, WA, and approved September 26, 2017 (received for review July 14, 2017)

The day we understand the time evolution of subcellular events at a level of detail comparable to physical systems governed by Newton's laws of motion seems far away. Even so, quantitative approaches to cellular dynamics add to our understanding of cell biology. With data-guided frameworks we can develop better predictions about, and methods for, control over specific biological processes and system-wide cell behavior. Here we describe an approach for optimizing the use of transcription factors (TFs) in cellular reprogramming, based on a device commonly used in optimal control. We construct an approximate model for the natural evolution of a cell-cycle–synchronized population of human fibroblasts, based on data obtained by sampling the expression of 22,083 genes at several time points during the cell cycle. To arrive at a model of moderate complexity, we cluster gene expression based on division of the genome into topologically associating domains (TADs) and then model the dynamics of TAD expression levels. Based on this dynamical model and additional data, such as known TF binding sites and activity, we develop a methodology for identifying the top TF candidates for a specific cellular reprogramming task. Our data-guided methodology identifies a number of TFs previously validated for reprogramming and/or natural differentiation and predicts some potentially useful combinations of TFs. Our findings highlight the immense potential of dynamical models, mathematics, and data-guided methodologies for improving strategies for control over biological processes.

cellular reprogramming | control theory | time series data | genome architecture | networks

In 1989, pioneering work by Weintraub et al. (1) successfully reprogrammed human fibroblasts into muscle cells via overexpression of transcription factor (TF) MYOD1, becoming the first study to demonstrate that the natural course of cell development could be altered. In 2007, Yamanaka and coworkers (2) changed the paradigm further by successfully reprogramming human fibroblasts into an embryonic stem-cell–like state [induced pluripotent stem cells (iPSCs)], using four TFs: POU5F1, SOX2, KLF4, and MYC. This work showed that a differentiated cell state could be reverted to a more pluripotent state. These discoveries have changed the trajectory of regenerative medicine, opening the possibility of generating needed cell types on demand for repairing damaged or diseased tissues. Ultimately, patient-derived fibroblasts could be used in autologous transplantations to minimize immune incompatibility.

These remarkable findings also demonstrate that the genome is a system capable of being controlled via an external input of TFs. In this context, determining how to push the cell from one state to another is, at least conceptually, a classical problem of control theory (3). The difficulty arises in the fact that the dynamics—and even proper representations of the cell state and inputs—are not well defined in the context of cellular reprogramming. Nevertheless, it seems natural to treat reprogramming as a problem in control theory, with the final state being the desired reprogrammed cell. In this paper, we provide such a framework based on empirical data and demonstrate the potential of this framework to provide insights into cellular reprogramming (4).

Our goal is to mathematically identify TFs that can directly reprogram human fibroblasts into a desired target cell type. As part of our methodology, we create a model for the natural dynamics of proliferating human fibroblasts, using time series data collected throughout the cell cycle. We couple data from bioinformatics with methods of mathematical control theory—a framework that we dub data-guided control (DGC). We use this model to determine a principled way to identify the best TFs for efficient reprogramming.

Previously, selection of TFs for reprogramming has been based largely on trial and error, typically relying on TF differential expression between cell types for initial predictions. Recent work has sought to predict TFs for reprogramming the cell state (5–8). Rackham et al. (7) devised a predictive method based on differential expression, as well as gene and protein network data. Our approach is fundamentally different in that we take a dynamical systems point of view, opening avenues for investigating efficiency (probability of conversion), timing (when to introduce TFs), and optimality (minimizing the number of TFs and amount of input).

Our method identifies TFs previously found to reprogram human fibroblasts into embryonic stem-cell–like cells, muscle cells, and many additional target cell types. Furthermore, our analysis predicts the points in the cell cycle at which the introduction of TFs might most efficiently affect a desired change of cell

Significance

Reprogramming the human genome toward any desirable state is within reach; application of select transcription factors drives cell types toward different lineages in many settings. We introduce the concept of data-guided control in building a universal algorithm for directly reprogramming any human cell type into any other type. Our algorithm is based on time series genome transcription and architecture data and known regulatory activities of transcription factors, with natural dimension reduction using genome architectural features. Our algorithm predicts known reprogramming factors, top candidates for new settings, and ideal timing for application of transcription factors. This framework can be used to develop strategies for tissue regeneration, cancer cell reprogramming, and control of dynamical systems beyond cell biology.

Author contributions: R.B. and I.R. designed research; S.R., G.P., A.B., R.B., and I.R. performed research; S.R., G.P., S.L., H.C., M.B., M.S.W., A.B., R.B., and I.R. analyzed data; and S.R., G.P., L.A.M., S.L., A.B., R.B., and I.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: indikar@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1712350114/-/DCSupplemental.

state. In addition, we demonstrate the efficacy of using topologically associating domains (TADs) for genome dimension reduction. Implicit in this approach is the notion of distance between cell types, which is measured in terms of the amount of transcriptional change required to transform one cell type into another. In this way, we are able to provide a comprehensive quantitative view of human cell types based on the respective distances between them.

Our framework separates into three parts:

i) Define the state. Use structure and function observations of the initial and target cell types' genomes to define a comprehensive state representation.

ii) Model the dynamics. Apply model identification methods to approximate the natural dynamics of the genome from time series data.

iii) Define and evaluate the inputs. Infer from bioinformatics (TF binding location and function) where TFs can influence the genome and then quantify controllability properties with respect to the target cell type.

The actual dynamics of the genome are undoubtedly very complicated, but as is often done in mathematical modeling studies, we use measurements to identify a linear approximation. This will take the form of a difference equation that is widely studied in the control systems literature, (9):

$$x_{k+1} = A_k x_k + B u_k. \qquad \textbf{[1]}$$

In this case, the three items listed above correspond respectively to the value of the state x_k at time k, the time-dependent state transition matrix A_k, and the input matrix B (along with the input function u_k).

Methods

Genome-State Representation and Dimension Reduction: $\mathbf{x_k}$. The state representation x in Eq. **1** is the foundation for any control system and is critical for controllability analysis. To fully represent the state of a cell, a high number of measurements would need to be taken, including gene expression, protein level, chromatin conformation, and epigenetic measurements. As a simplification, we assume that the gene expression profile is a sufficient representation of the cell state.

Gene expression for a given cell is dependent on a number of factors, including (but not limited to) cell type, cell-cycle stage, circadian-rhythm stage, and growth conditions. To best capture the natural fibroblast dynamics from population-level data, time series RNA-seq was performed on cells that were cell-cycle and circadian-rhythm synchronized in normal growth medium conditions (*SI Appendix*). Before data collection, all cells were temporarily held in the first stage of the cell cycle, G_0/G_1, via serum starvation. Upon release into the cell cycle, the population was observed every $\Delta t = 8$ h for 56 h, yielding eight time points (at 0 h, 8 h, 16 h, …, 56 h). Let $g_{i,k}$ be the measured activity of gene $i = 1, \dots, N$ at measurement time $k = 1, \dots, 8$, where N is the total number of human genes observed (22,083). Analysis of cell-cycle marker genes indicated that the synchronized fibroblasts took between 32 h and 40 h to complete one cell cycle after growth medium introduction (*SI Appendix*, Fig. S1). Because of this, we define $K = 5$ to be the total number of time points used for this model.

An obstacle to using g to represent x in a dynamical systems approach is the computational feasibility of studying a system with over 20,000 variables, necessitating a dimension reduction. A comprehensive genome-state representation should include aspects of both structure and function and simultaneously have low enough dimension to be computationally reasonable. Along these lines, we propose a biologically inspired dimension reduction based on TADs.

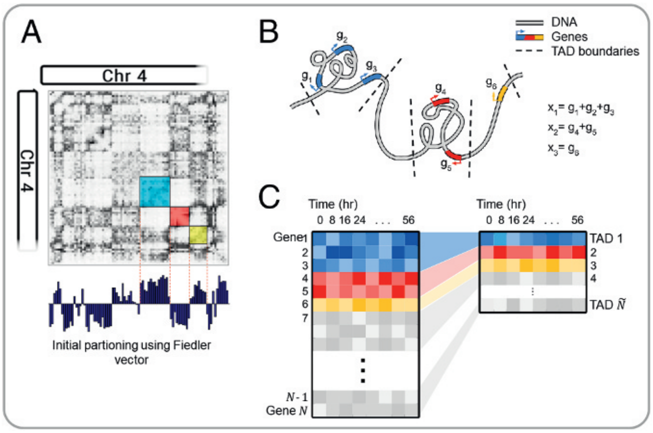


Fig. 1. Overview of TAD dimension reduction. (*A*) Partitioning the Hi-C matrix based on the Fiedler vector. (*B*) Cartoon depiction of TAD genomic structure. (*C*) TAD dimension reduction summary.

The advent of genome-wide chromosome conformation capture (Hi-C) allowed for the studying of higher-order chromatin structure and the subsequent discovery of TADs (10). TADs are inherent structural units of chromosomes: contiguous segments of the 1D genome for which empirical physical interactions can be observed (11). Moreover, genes within a TAD tend to exhibit similar activity, and TAD boundaries have been found to be largely cell-type invariant (11, 12). TADs group structurally and functionally similar genes, serving as a natural dimension reduction that preserves important genomic properties. Fig. 1 depicts an overview of this concept. We computed TAD boundaries from Hi-C data via an algorithm that uses Fielder vector partitioning, described in Chen et al. (13) (*SI Appendix*).

Let $tad(i) := j$ if gene i is contained within TAD j. We define each state variable $x_{j,k}$ to be the expression level of TAD $j = 1, \dots, \tilde{N}$ at time k, where $\tilde{N} = 2,245$ is the total number of TADs that contain genes. Specifically, $x_{j,k}$ is defined as the sum of the expression levels of all genes within the TAD, measured in reads per kilobase of transcript per million (RPKM); i.e.,

$$x_{j,k} := \sum_{\substack{i \text{ s.t.} \\ tad(i)=j}} g_{i,k}. \qquad \textbf{[2]}$$

The vector of all TAD activities at measurement k is denoted with a single subscript $x_k \in \mathbb{R}^{\tilde{N} \times 1}$, $k = 1, \dots, K$.

State Transition Matrix: $\mathbf{A_k}$. Given the data we have, perhaps the most direct way to model the evolution of TAD activity level would be to assume a model of the form $x_{k+1} = x_k + u_k$, where x_k and x_{k+1} come from data, and u_k is the exogenous input. However, in a time-varying situation, it will typically not be the best way to use the data to create a model because it fails to capture the idea that an input applied at one point in the cell cycle can be expected to have a different effect if applied at a different point in the cell cycle. Taken over a full cycle, the average value of the expression level of the 2,245 TADs is known within experimental error. Assuming that there is a function f which maps x_k to x_{k+1}, we can subtract the steady-state average, $\bar{x}$, and focus on measuring the deviation from average as the cycle evolves. With this in mind, we consider the first-order approximation $f(x) = \bar{x} + A(x - \bar{x})$, where A is allowed to depend on where one is in the cell cycle. This is a time-varying linear model for the variation from $\bar{x}$. If the model is to match data and capture variability over the cell cycle, we will need to have a different A for each time step. Using the principle that A should differ as little from the identity as possible, we let A_k be the identity plus a rank one matrix chosen to match the data for each time step k; we impose the condition that without inputs we have $x_{k+1} - \bar{x} = A_k(x_k - \bar{x})$.

Define a time-dependent state transition matrix A_k as

$$A_k := I_{\tilde{N}} + \frac{(x_{k+1}-x_k)x_k^T}{x_k^T x_k} \in \mathbb{R}^{\tilde{N}\times\tilde{N}}, \quad k = 1,2,3,4,5, \qquad [3]$$

where $I_{\tilde{N}}$ is the $\tilde{N} \times \tilde{N}$ identity matrix. Let the measured values of the state of the unforced evolution be $x_1, x_2, \cdots, x_5$; let the controls be labeled $u_1, u_2, \cdots, u_5$; let the values of the state with the controls acting be $z_2, z_3, \cdots, z_6$. Letting z denote the deviation from the cell-cycle average, we have

$$z_{k+1} = \left(I + \frac{1}{x_k^T x_k} \left((x_{k+1} - x_k) x_k^T \right) \right) z_k + Bu_k,$$

where A_k is as above. Solving this difference equation, we have

$$z_k = \prod_{i=1}^{k-1} A_i x_1 + \sum_{i=1}^{k-1} \prod_{j=i}^{k-1} A_{j-1} B u_i$$

with the understanding that $A_0 = I$. This explicit expression shows that the effect of the $u(i)$ cannot be inferred from the sum of the u_i because different u_i are weighted in different ways, dependent on the stage in the cell cycle at which it is applied. This is a significant point relating to the model and plays a significant role in determining the optimal times for inserting TFs.

Input Matrix and Input Signal: $\mathbf{B}, \mathbf{u}_k$. With the natural TAD-level dynamics established in the context of our control Eq. **1**, we turn our attention to quantifying methods for control.

A TF can regulate a gene positively or negatively by binding to a specific DNA sequence near a gene and encouraging or discouraging transcription. The degree to which a TF activates or represses gene expression depends on the specific TF–gene interaction, which is influenced by a variety of factors that are difficult to quantify. Let $w_{i,m}$ be the theoretical regulation weight of TF m on gene i, where $w_{i,m} > 0$ ($w_{i,m} < 0$) if TF m activates (represses) gene i, and $m = 1, \ldots, M$, where M is the total number of well-characterized TFs. Weights that are bigger in absolute value, $|w_{i,m}| \gg 0$, indicate stronger transcriptional influence, and weights equal to zero, $w_{i,m} = 0$, indicate no influence.

Extensive TF perturbation experiments would be needed to determine $w_{i,m}$ for each TF m on each gene i. Instead, we propose a simplified method to approximate $w_{i,m}$ from existing, publicly available data for TF binding sites (TFBSs), gene accessibility, and average activator/repressor activity. To determine the number of possible binding sites a TF m recognizes near gene i, the reference genome was scanned for the locations of potential TFBSs following methods outlined by Neph et al. (14) (*SI Appendix*). Position frequency matrices (PFMs), which give information on TF–DNA binding probability, were obtained for 547 TFs from a number of publicly available sources ($\therefore\ M = 547$). Let $c_{i,m}$ be the number of TF m TFBSs found within ±5 kb of the transcriptional start site (TSS) of gene i (*SI Appendix*, Fig. S2).

Although many TFs can do both in the right circumstances, most TFs have a tendency toward either activator or repressor activity (15). That is, if TF m is known to activate (repress) most genes, we can say with some confidence that TF m is an activator (repressor), so $w_{i,m} \geq 0$ ($w_{i,m} \leq 0$) for all i. To determine a TF's function, we performed a literature search for all 547 TFs and labeled 299 as activators and 124 as repressors (*SI Appendix*). The remaining TFs were labeled unknown for lack of conclusive evidence and were evaluated as both an activator and a repressor in separate calculations. Here, we define a_m as the activity of TF m, with 1 and -1 denoting activator and repressor, respectively.

TFBSs are cell-type invariant since they are based strictly on the linear genome. However, it is known that for a given cell type, certain areas of the genome may be opened or closed, depending on epigenetic aspects. To capture cell-type–specific regulatory information, we obtained publicly available gene accessibility data (DNase-seq) on human fibroblasts (GSM1014531). DNase-seq extracts cell-type–specific chromatin accessibility information genome-wide by testing the genome's sensitivity to the endonuclease DNase I and sequencing the nondigested genome fragments. These data are used for our initial cell type to determine which genes are available to be controlled by TFs (16). Here, we define s_i to be the DNase I sensitivity information (accessibility; open/close) of gene i in the initial state, with 1 and 0 denoting accessible and inaccessible, respectively (*SI Appendix*).

We approximate $w_{i,m}$ as

$$w_{i,m} := a_m s_i c_{i,m}, \qquad [4]$$

so that the magnitude of influence is equal to the number of observed consensus motifs $c_{i,m}$, except when the gene is inaccessible ($s_i = 0$) in which case $w_{i,m} = 0$.

Since we are working off a TAD-dimensional model, our input matrix B must match this dimension. Let b_m be a 2,245-dimensional vector, where the jth component is

$$b_{j,m} := \sum_{\substack{i \text{ s.t.} \\ tad(i)=j}} w_{i,m}, \qquad [5]$$

and define a matrix $B = \begin{bmatrix} b_1\ b_2\ \cdots\ b_M \end{bmatrix}$.

The amount of control input is captured in u_k, which is an $\mathbb{R}^{M\times 1}$ vector representing the quantity of the external TFs we are inputting to the system (cell) at time k. This can be controlled by the researcher experimentally through manipulation of the TF concentration (17). In this light, we restrict our analysis to $u_k \geq 0$ for all k, as TFs cannot be subtracted from the cell. $u_{m,k}$ is defined as the amount of TF m to be added at time point k.

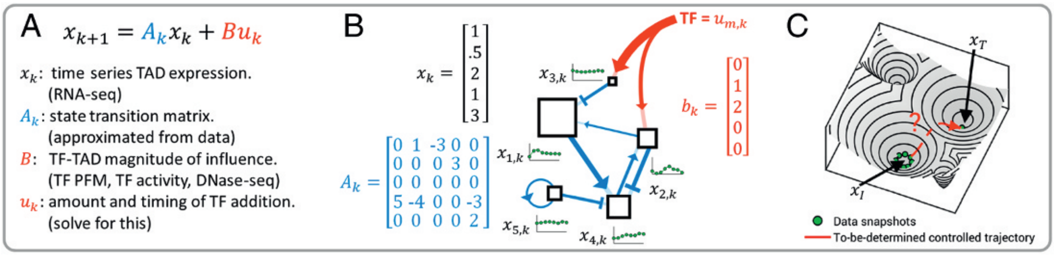

A
$x_{k+1} = A_k x_k + Bu_k$
x_k: time series TAD expression. (RNA-seq)
A_k: state transition matrix. (approximated from data)
B: TF-TAD magnitude of influence. (TF PFM, TF activity, DNase-seq)
u_k: amount and timing of TF addition. (solve for this)
B
$x_k = \begin{bmatrix} 1 \\ .5 \\ 2 \\ 1 \\ 3 \end{bmatrix}$
$A_k = \begin{bmatrix} 0 & 1 & -3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 5 & -4 & 0 & 0 & -3 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$
TF = $u_{m,k}$
$x_{3,k}$
$x_{1,k}$
$x_{2,k}$
$x_{5,k}$
$x_{4,k}$
$b_k = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 0 \end{bmatrix}$
C
x_T
x_I
?
Data snapshots
To-be-determined controlled trajectory

Fig. 2. DGC overview. (*A*) Summary of control equation variables. (*B*) Each TAD is a node in a dynamic network. The blue connections represent the edges of the network and are determined from time series fibroblast RNA-seq data. The green plots represent the expression of each TAD changing over time. The red arrows indicate additional regulation imposed by exogenous TFs. (*C*) A conceptual illustration of the problem: Can we determine TFs to push the cell state from one basin to another?

With all variables of our control Eq. **1** defined, we can now attempt to predict which TFs will most efficiently achieve cellular reprogramming from some x_I (initial state; fibroblast in our setting) to x_T (target state; any human cell type for which compatible RNA-seq data are available) through manipulation of u_k. An overview of our DGC framework is given in Fig. 2.

Selection of TFs. Our general procedure for scoring TFs is explained as follows. Eq. **1** has an explicit solution that is given below. The first few terms are

$$\begin{aligned} z_2 &= A_1x_1 + Bu_1 \\ z_3 &= A_2A_1x_1 + A_2Bu_1 + Bu_2 \\ z_4 &= A_3A_2A_1x_1 + A_3A_2Bu_1 + A_3Bu_2 + Bu_3 \\ &\vdots \end{aligned}$$

This shows how z_4 depends on u_1, u_2, and u_3.

If x_T is a target condition, then the Euclidean distance $\|\cdot\|$ can be used to measure how close a state is to the target state. We define

$$d = \|x_T - z_6(u)\|, \qquad \mathbf{[6]}$$

where the notation $z_6(u)$ is used to emphasize the dependence of z_6 on u. Considering all possible input signals, one can compute the optimal control that finds the minimum distance for a given initial and target cell type. Let u_* denote the optimal u used to minimize d and d_* denote this minimum distance value.

When appropriate, we write $z_6(u)$ to emphasize the fact that the final state depends on the input. The Euclidean distance $\|\cdot\|$ can be used to measure how close a given state is to the target. If there were no restrictions on the u terms, the control that minimizes the distance between z_6 and the target could be computed without difficulty. However, there are reasons for restricting the number of different TFs used in any one trial. Transfection of cells with too many TFs can lower the efficiency of transfection and even lead to cell death. Moreover, many confirmed direct reprogramming experiments use ≤4 TFs to achieve reprogramming. For these reasons, we modify the optimization problem by adding the constraint that there are no more than a fixed number of TFs (components of u) used in a given trial.

Let $\hat{p}$ be a set of integers that identifies the subset of the components of u (read: TFs) that are allowed to be nonzero. For example, $\hat{p} = \{1, 4, 7\}$ refers to TFs 1, 4, and 7. Let p be the number of elements in $\hat{p}$. Given a set of TFs, $\hat{p}$, we determine the quantity and timing of TF input, u_{*k}, that minimizes the difference between x_6 and the target cell state, x_T. Mathematically, this can be written as

$$\begin{array}{ll} \underset{u}{\text{minimize}} & \|x_T - z_6(u)\| \\ \text{subject to} & \begin{cases} u_{m,k} \geq 0, & k = 1, ..., 5 \\ u_{m,k} = 0, & \text{if } m \notin \hat{p} \\ u_{m,k+1} \geq u_{m,k} & \end{cases}. \end{array} \qquad \mathbf{[7]}$$

We use MATLAB's *lsqnonneg* function to solve Eq. **7**, which gives u_{*k} and d_*.

Let $d_0 := \|x_T - x_0\|$ be the distance between the final state and target state with no control input. Define a score $\mu := d_0 - d_*$, which can be interpreted as the improvement provided by a particular choice of u. This can be calculated for each $\hat{p}$ and sorted (high to low) to determine which TF or TF combination is the best candidate for direct reprogramming between x_0 and x_T.

We consider different scenarios for the type of input regime in the results. The first one assumes the input signal is constant $u_1 = u_k = \bar{u}$, intended to mimic empirical regimes where TFs are given at a single time point. Later, we also consider inputting TFs at different times $\hat{k}$, which can be viewed mathematically as requiring $u_{m,k} = 0$ for all $k < \hat{k}$, and $u_{m,k}$ is a constant value for all $k \geq \hat{k}$. This is intended to mimic inputting a TF at time $\hat{k}$, which will continue to express at a constant level until time point $k = 6$.

Remark. Subsets of TFs were chosen for each calculation based on the following criteria: ≥10-fold expression increase in target state compared with initial state and ≥10 RPKM in target state. These criteria are used to select differentially expressed TFs and TFs that are sufficiently active in the target state.

Results

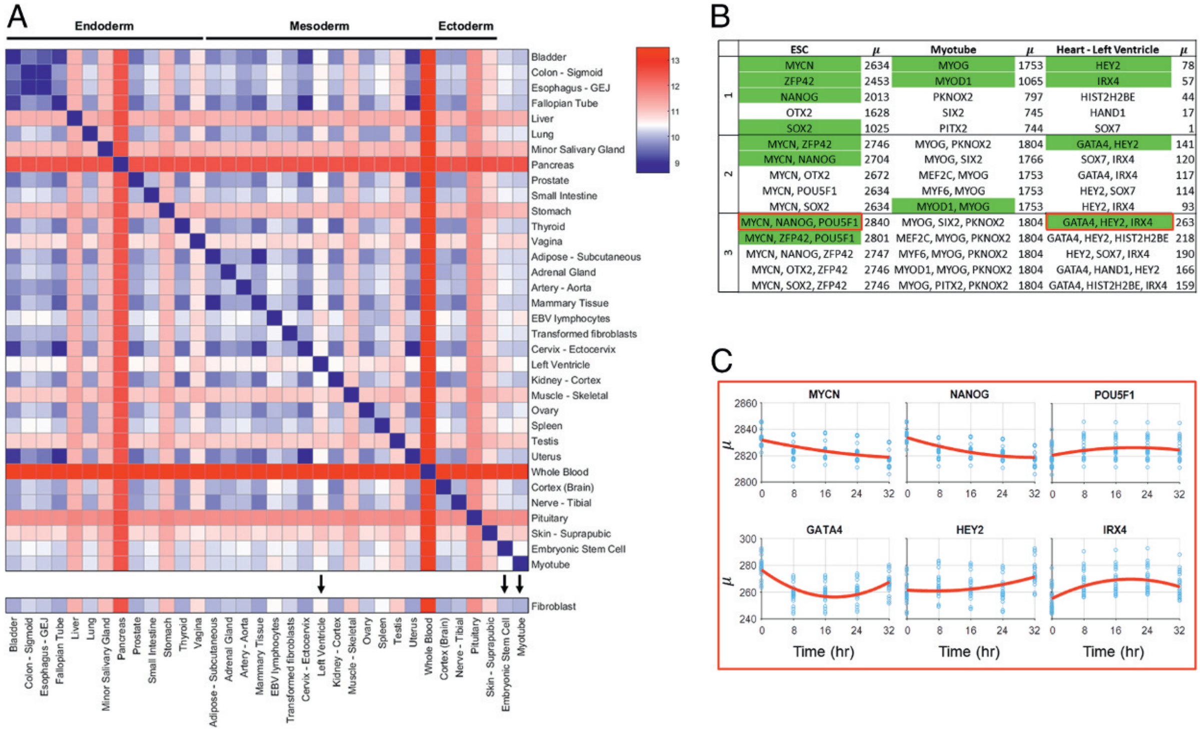
Quantitative Measure Between Cell Types. To best use our algorithm to predict TFs for reprogramming, compatible data on target cell types must be collected. For this, we explore a number of publicly available databases where RNA-seq has been collected, along with RNA-seq data collected in our laboratory. The ENCODE Consortium has provided data on myotubes and embryonic stem cells (ESCs) (*SI Appendix*) (18). The GTEx portal provides RNA-seq data on a large variety of different human tissue types (19). Although each GTEx experiment is performed on tissue samples, thus containing multiple different cell types, we use these data as more general cell-state targets.

To give a numerical structure to cell-type differences, conceptually similar to Waddington's epigenetic landscape, we calculate d_0 between all cell types collected. Fig. 3*A* shows d_0 values for 32 tissue samples collected from the GTEx portal, along with ESC, myotube, and our fibroblast data (additional cell-type d_0 values shown in *SI Appendix*). GTEx RNA-seq data are scaled to keep total RPKM difference between time series fibroblast and GTEx fibroblast RNA-seq minimal (*SI Appendix*).

TF Scores. To assess our method's predictive power, a subset of target cell types is presented here that has validated either TF reprogramming methods or TFs highly associated with the target cell type. Additional predicted TFs for reprogramming are included in *SI Appendix*. We note that although experimentally validated TFs provide the best current standard for comparison, we believe experimental validation with our predicted TFs may provide more efficient and comprehensive reprogramming results. For all reprogramming regimes presented in this section, fibroblast is used as the initial cell type due to the availability of synchronized time series data, and all TFs are introduced at $k = 1$ (11).

For conversion of fibroblast to myotubes, the top predicted single-input TFs are MYOG and MYOD1, both of which are known to be crucial for myogenesis. While MYOD1 is the classic master regulator reprogramming TF for myotube conversion, activation of downstream factor MYOG is necessary for full conversion (20). For fibroblast to ESC conversion, a number of TFs known to be necessary for pluripotency are predicted, including MYCN, ZFP42, NANOG, and SOX2 (2). With the knowledge that no single TF has been shown to fully reprogram a fibroblast to an embryonic state, combinations of TFs are more informative for this analysis. The top-scoring combination of three TFs is MYCN, NANOG, and POU5F1—three well-known markers for pluripotency (2). Interestingly, POU5F1 scores poorly when input individually, but is within the top set of three TFs when used in combination with MYCN and NANOG. Left ventricle reprogramming includes TFs that are known to be necessary for natural differentiation in the top score for all one to three combinations. These include GATA4 (a known TF in fibroblast to cardiomyocyte reprogramming), HEY2, and IRX4 (21–23).

Time-Dependent TF Addition. Fibroblast to ESC conversion was of particular interest in our analysis as this is a well-studied



	ESC	μ	Myotube	μ	Heart - Left Ventricle	μ
1	MYCN	2634	MYOG	1753	HEY2	78
	ZFP42	2453	MYOD1	1065	IRX4	57
	NANOG	2013	PKNOX2	797	HIST2H2BE	44
	OTX2	1628	SIX2	745	HAND1	17
	SOX2	1025	PITX2	744	SOX7	1
2	MYCN, ZFP42	2746	MYOG, PKNOX2	1804	GATA4, HEY2	141
	MYCN, NANOG	2704	MYOG, SIX2	1766	SOX7, IRX4	120
	MYCN, OTX2	2672	MEF2C, MYOG	1753	GATA4, IRX4	117
	MYCN, POU5F1	2634	MYF6, MYOG	1753	HEY2, SOX7	114
	MYCN, SOX2	2634	MYOD1, MYOG	1753	HEY2, IRX4	93
3	MYCN, NANOG, POU5F1	2840	MYOG, SIX2, PKNOX2	1804	GATA4, HEY2, IRX4	263
	MYCN, ZFP42, POU5F1	2801	MEF2C, MYOG, PKNOX2	1804	GATA4, HEY2, HIST2H2BE	218
	MYCN, NANOG, ZFP42	2747	MYF6, MYOG, PKNOX2	1804	HEY2, SOX7, IRX4	190
	MYCN, OTX2, ZFP42	2746	MYOD1, MYOG, PKNOX2	1804	GATA4, HAND1, HEY2	166
	MYCN, SOX2, ZFP42	2746	MYOG, PITX2, PKNOX2	1804	GATA4, HIST2H2BE, IRX4	159

Fig. 3. Quantitative measure between cell types and TF scores. (*A*) d_0 values between GTEx tissue types and ESC, myotube, and fibroblast. Tissue types and cell types with black arrows have predicted TFs for reprogramming from fibroblasts shown in *B*. (*B*) Table of predicted TFs for a subset of cell and tissue types. Top five TFs for combinations of one to three are shown. Green labeled TFs are highly associated with the differentiation process of the target cell type and/or validated for reprogramming. These TFs are discussed in the main text. (*C*) Time-dependent scores for selected combinations of three TFs for fibroblast to ESC and fibroblast to "heart - left ventricle." *x* axis refers to time of TF addition, and *y* axis refers to μ.

regime with a number of validated TFs (with a variety of reported efficiencies), and this conversion is promising for its regenerative medicine application. High-scoring TFs yield many that are known markers for pluripotency, but the top combination of three, MYCN, NANOG, and POU5F1, has not been used specifically together, to our knowledge. Here, we analyzed how the TF combination would score if input at different points throughout the cell cycle.

Time-dependent analysis of the top-scoring ESC TFs reveals that scores vary widely, depending on the time of input. MYCN and NANOG show a strong preference for input at the beginning of the cell cycle, while POU5F1 shows a slight preference for input toward the end of the cell cycle, with the highest score achieved when MYCN and NANOG are input at 0 h and POU5F1 is input at 32 h. Analysis on how the time of input control affects μ is shown in Fig. 3*C*. Time-dependent analysis was also conducted for the top combination of three TFs for fibroblast to left ventricle. This analysis predicted that the best reprogramming results would occur if GATA4 is given immediately (0 h), with IRX4 and HEY2 given later (24 and 32 h, respectively).

Discussion

The results from this algorithm show promise in their prediction of known reprogramming TFs and demonstrate the importance of including time series data for gene network dynamics. Time of input control has shown to have an impact on the end cell state, in line with what has been shown in natural differentiation (24).

While we believe that this is the best model currently available for predicting TFs for reprogramming, we are aware of its limitations and assumptions. TAD-based dimension reduction is based on the observation that genes within them correlate in expression over time, although we lack definitive proof of regulation by shared transcriptional machinery (11). This assumption was deemed necessary for dimension reduction in the context of deriving transition matrix A_k. With finer time steps in RNA-seq data, the assumption may not be necessary for TF prediction, at the cost of increased computation time. Additionally, a 5-kb window flanking the TSS of each gene was used to ensure that all potential regulators are found, at the cost of potential inclusion of false positive motifs.

Although this program can score TFs relative to other TFs in a given reprogramming regime, it is difficult to predict a μ threshold that would guarantee conversion. Additionally, rigorous experimental testing will be required to validate these findings and determine how our u vector translates to TF concentration. This is a product of the large number of assumptions that we have made to develop the initial framework for a reprogramming algorithm. With finer resolution in the time series gene expression, more subtle aspects of the genomic network may be observed, allowing for better prediction.

Our proposed DGC framework successfully identified known TFs for fibroblast to ESC and fibroblast to muscle cell reprogramming regimes. We use a biologically inspired dimension reduction via TADs, a natural partitioning of the genome. This comprehensive state representation was the foundation of our framework, and the success of our methods motivates further investigation of the importance of TADs as functional units to control the genome.

A dynamical systems view of the genome allows for analysis of timing, efficiency, and optimality in the context of reprogramming. Our framework is the first step toward this view. The successful implementation of time-varying reprogramming regimes would open unique avenues for direct reprogramming. This template can be used to develop regimes for changing any cell into any other cell, for applications that include reprogramming cancer cells and controlling the immune system. Our DGC framework is well equipped for designing personalized cellular reprogramming regimes. Finally, this framework can serve as a general technique for investigating the controllability of networks strictly from data.

Materials and Methods

Hi-C and RNA-seq data were collected from cell-cycle– and circadian-rhythm–synchronized proliferating human fibroblasts of normal karyotype. Data were collected every 8 h, spanning 56 h. Publicly available data were used for target cell types. Detailed materials and methods are provided in Chen et al. (11), Dataset S1, and in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Robert Oakes, Emily Crossette, and Sijia Liu for their critical reading of the manuscript and helpful discussions. We especially thank James Gimlett and Srikanta Kumar at Defense Advanced Research Projects Agency (DARPA) for support and encouragement. This work is supported, in part, by the DARPA Biochronicity, Deep-Purple, and FunCC programs.

1. Weintraub H, et al. (1989) Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc Natl Acad Sci USA* 86:5434–5438.
2. Takahashi K, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861–872.
3. Brockett R (1970) *Finite Dimensional Linear Systems* (Wiley, New York).
4. Rajapakse I, Groudine M, Mesbahi M (2011) Dynamics and control of state-dependent networks for probing genomic organization. *Proc Natl Acad Sci USA* 108:17257–17262.
5. Cahan P, et al. (2014) Cellnet: Network biology applied to stem cell engineering. *Cell* 158:903–915.
6. D'Alessio AC, et al. (2015) A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep* 5:763–775.
7. Rackham O, et al. (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat Genet* 48:331–335.
8. Michael DG, et al. (2016) Model-based transcriptome engineering promotes a fermentative transcriptional state in yeast. *Proc Natl Acad Sci USA* 113:E7428–E7437.
9. Aström KJ, Murray RM (2010) *Feedback Systems: An Introduction for Scientists and Engineers* (Princeton Univ Press, Princeton).
10. Dixon JR, et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485:376–380.
11. Chen H, et al. (2015) Functional organization of the human 4D nucleome. *Proc Natl Acad Sci USA* 112:8002–8007.
12. Dixon JR, Gorkin DU, Ren B (2016) Chromatin domains: The unit of chromosome organization. *Mol Cell* 62:668–680.
13. Chen J, Hero AO, III, Rajapakse I (2016) Spectral identification of topological domains. *Bioinformatics* 32:2151–2158.
14. Neph S, et al. (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150:1274–1286.
15. Ernst J, et al. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat Biotechnol* 34:1180–1190.
16. Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489:75–82.
17. Brewster RC, et al. (2014) The transcription factor titration effect dictates level of gene expression. *Cell* 156:1312–1323.
18. ENCODE Project Consortium, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
19. Lonsdale J, et al. (2013) The genotype-tissue expression (GTEx) project. *Nat Genet* 45:580–585.
20. Weintraub H (1993) The MyoD family and myogenesis: Redundancy, networks, and thresholds. *Cell* 75:1241–1244.
21. Ieda M, et al. (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* 142:375–386.
22. Fischer A, Schumacher N, Maier M, Sendtner M, Gessler M (2004) The notch target genes hey1 and hey2 are required for embryonic vascular development. *Genes Dev* 18:901–911.
23. Nelson DO, Jin DX, Downs KM, Kamp TJ, Lyons GE (2014) Irx4 identifies a chamber-specific cell population that contributes to ventricular myocardium development. *Dev Dyn* 243:381–392.
24. Loh KM, et al. (2016) Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types. *Cell* 166:451–467.