Algorithm for cellular reprogramming

Scott Ronquist\textsuperscript{a}, Geoff Patterson\textsuperscript{b}, Lindsey A. Muir\textsuperscript{c}, Stephen Lindsly\textsuperscript{a}, Haiming Chen\textsuperscript{a}, Markus Brown\textsuperscript{d}, Max S.acha\textsuperscript{e}, Anthony Blochl\textsuperscript{f}, Roger Brockett\textsuperscript{g}, and Indika Rajapakse\textsuperscript{h,i}

\textsuperscript{a}Department of Computational Medicine and Bioinformatics, Medical School, University of Michigan, Ann Arbor, MI 48109; \textsuperscript{b}Department of Curriculum Design, IXI Learning, Raleigh, NC 27560; \textsuperscript{c}Department of Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, MI 48109; \textsuperscript{d}Department of Biological Sciences, University of Maryland, College Park, MD 20742; \textsuperscript{e}Department of Hematology/Oncology, University of Michigan, Ann Arbor, MI 48109; \textsuperscript{f}Department of Genetics, University of Michigan, Ann Arbor, MI 48109; \textsuperscript{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.

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.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This is an open access article distributed under the PNAS license.

1To whom correspondence should be addressed. Email: indika@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):

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

[Eqn 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: \( x_k \). 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, chromosome 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 \( i = 1, \ldots, N \) at measurement time \( k = 1, \ldots, 8 \), 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.

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 \( \text{tad}(i) := j \) if gene \( i \) is contained within TAD \( j \). We define each state variable \( x_{k,j} \) to be the expression level of TAD \( j = 1, \ldots, N \) at time \( k \), where \( N = 2,245 \) is the total number of TADs that contain genes. Specifically, \( x_{k,j} \) 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.,

\[ x_{j,k} := \sum_{i \in \text{tad}(i)=j} g_{i,k}. \]  

[Eqn 2]

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

State Transition Matrix: \( A_k \). 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 \( f(x) = x + A(x - \bar{x}), \) 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 \( A \) be the
where \( A \) influence. 

The red arrows indicate additional regulation imposed by exogenous TFs. ([4]) A conceptual illustration of the problem: Can we determine TFs to push the network and are determined from time series fibroblast RNA-seq data. The green plots represent the expression of each TAD changing over time.

**Input Matrix and Input Signal:** \( B, u_k \). With the natural TAD-level dynamics established in the context of our control Eq. I, 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, \( |w_{i,m}| > 0 \), 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 (\( \sim M = 547 \)). 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/closed) of gene \( i \) in the initial state, with 1 and 0 denoting accessible and inaccessible, respectively ([SI Appendix]).

We approximate \( w_{i,m} \) as

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

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 \( b_{i,j} \) be a 2,245-dimensional vector, where the \( j \)th component is

\[
b_{j,m} := \sum_{i,k \in \text{TAD}(i,j)} w_{i,m},
\]

and define a matrix \( B = [b_1 b_2 \cdots b_M] \).

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 \( k \).

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 \( x_t \) (initial state; fibroblast in our setting) to \( x_f \) (target state; any human cell type for which compatible RNA-seq data are available) through manipulation of \( u_t \). 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

\[
\begin{align*}
z_1 &= A_1 x_1 + B u_t \\
z_2 &= A_2 A_1 x_2 + A_2 B u_t + B u_2 \\
z_3 &= A_3 A_2 A_1 x_3 + A_3 A_2 B u_2 + A_3 B u_2 + B u_3 \\
&\vdots
\end{align*}
\]

This shows how \( z_t \) depends on \( u_1, u_2, \) and \( u_3 \).

If \( x_f \) 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

\[
d = \| x_T - z_0(u) \|, \tag{6}
\]

where the notation \( z_0(u) \) is used to emphasize the dependence of \( z_0 \) 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 \( u_o \) denote the optimal \( u \) used to minimize \( d \) and \( d_o \) denote this minimum distance value.

When appropriate, we write \( z_0(u) \) 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 \( z_0 \) 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 \( \leq 4 \) 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 \( p \) 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 \( \hat{p} \), \( \hat{p} \), we determine the quantity and timing of TF input, \( u_{\hat{p},k} \), that minimizes the difference between \( z_0 \) and the target cell state, \( x_T \). Mathematically, this can be written as

\[
\begin{align*}
\text{minimize} & \quad \| x_T - z_0(u) \| \\
\text{subject to} & \quad \begin{cases} 
| u_{m,k} | & \geq 0, & k = 1, \ldots, 5, \\
| u_{m,k} | & = 0, & m \notin \hat{p}, \\
| u_{m,k+1} + u_{m,k} | & \geq 0, & k \geq 0 \end{cases}
\end{align*} \tag{7}
\]

We use MATLAB’s \texttt{lsonomneg} function to solve Eq. 7, which gives \( u_{\hat{p},k} \) and \( d_o \).

Let \( d_0 := \| x_T - z_0 \| \) be the distance between the final state and target state with no control input. Define a score \( \mu := d_0 - d_o \), 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 \( u_1 = u_2 = u \), 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: \( \geq 10\)-fold expression increase in target state compared with initial state and \( \geq 10 \) 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 \( d_{\hat{p}} \) between all cell types collected. Fig. 3A shows \( d_0 \) values for 32 tissue samples collected from the GTeX portal, along with ESC, myotube, and our fibroblast data (additional cell-type \( d_{\hat{p}} \) 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 comparisons, 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 MYOD and MYOD1, both of which are known to be crucial for myogenesis. While MYOD1 is the classic master regulator reprogramming TF for myofiber 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

---

Ronquist et al.
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 $\mu$ is shown in Fig. 3C. 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 $\mu$ threshold that would guarantee conversion. Additionally, rigorous experimental testing will be required to validate these findings and determine how our $x$ 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.

Ronquist et al.  
PNAS | November 7, 2017 | vol. 114 | no. 45 | 11837
Supporting Information (SI)

Data. The fibroblast (FIB) data (Hi-C and RNA-seq) used for this application was originally collected and published in a paper by Chen et al. (1). We refer the reader to this paper for a full description of technical protocols. Embryonic Stem Cell (ESC) and myotube (MT) data was downloaded from NCBI-GEPO (GSE23316 ENCODE Caltech RNA-seq and GSE52529) (2). 53 different tissue RNA-seq samples were downloaded from GTEx portal (3). 51 different immune cell type RNA-seq samples were obtained from the BLUEPRINT Epigenome project (4).

Hi-C and Construction of TADs. We computed TAD boundaries from genome-wide chromosome conformation capture (Hi-C) data using an algorithm described in Chen et al. (5). The algorithm was applied to averaged time series Hi-C data from proliferating human fibroblast (FIB) at 100 kilo-base pair (kb) resolution, which identified 2,562 TADs across all autosomal chromosomes (i.e. excluding Chromosomes X and Y). Of the 2,562 TADs, 317 contained no genes and were excluded from our analysis, leaving 2,245 TADs. These TADs ranged in size from a few hundred kb to several Mb, and contained on average 9-10 genes (standard deviation of 18 genes); one gene at minimum, and 249 genes maximum.

Construction of B Matrix. TF binding site position frequency matrix (PFM) information was obtained from Neph et al. and MotifDB, which is a collection of publicly available PFM databases, such as JASPAR, Jolma et al. cispl_1.02, stamlab, hPDI, and UniPROBE (6, 7). TRANSFAC PFM information was included as well. Motif scanning of the human reference genome (hg19) was performed using FIMO of the MEME suite, in line with methods established by Neph et al. (6). DNase-seq information for human fibroblast was derived from ENCODE for fibroblast (GSM1014531). If a narrow peak is found within the ±5kb of a gene TSS, the region is classified as open. TF function information was determined through an extensive literature search.

Scaling of RNA-seq. Due to differences in data collection procedures, the RNA-seq RPKM values obtained from the GTEx portal were of lower value, on average, compared to our fibroblast dataset, thus favoring repressor TFs for μ scoring. In order to account for this in our model, we scaled all GTEx RNA-seq data by a factor that solves the equation

$$\min_{\alpha} \| g_{FIB,UM} - \alpha g_{FIB,GTEX} \|$$  \[1\]

where \( g_{FIB,UM} \) is the gene-level RNA-seq vector average of our fibroblast data, \( g_{FIB,GTEX} \) is the gene-level RNA-seq vector of “Cells - Transformed fibroblasts” from the GTEx portal, and \( \alpha \) is a scalar that solves this equation. For this data, \( \alpha = 2.6113 \) and all GTEx data used as a target state was scaled by this factor.

Removal of MicroRNA. MicroRNA were removed from this analysis due to their high variance in RPKM values and unpredictable function.

TF Scores - BLUEPRINT Project Data. A number of immune cell types extracted from the BLUEPRINT Project revealed promising predicted TF results when fibroblast is used as the starting point (4). \( d_0 \) values between cell types are shown in Fig. S4.

For fibroblast to macrophage direct reprogramming, a number of factors scoring highly in our algorithm are known to play a role in macrophage reprogramming or the differentiation. SP11 (along with CEBPA) has been verified experimentally to reprogram fibroblasts into macrophage-like cells (20). IKZF1 has been shown to be crucial for macrophage polarization via the IRF4/IRF5 pathway (21). MYB has been shown to be crucial for the upstream cell type HSC (22).

For fibroblast to HSC direct reprogramming, the top scoring individual factor is highly associated with the target phenotype and has been shown to support HSC growth and regeneration (23). ERG (in combination with GATA2, LMO2, RUNX1c, and SCL) is a confirmed reprogramming factor for fibroblast to HSC in mice (24).

For fibroblast to erythroblast reprogramming, ERG is a promising factor as it is required for the maintenance of the upstream cell type HSC (24). NFIA is shown to promote the erythroid lineage from HSC differentiation (25).
Alternative Computation of $u$. Below is an example of how $u$ can be computed without the constraint that $u_k,m \geq 0$. Assume $u_k := \bar{u}$ is constant for all $t$. Then

$$x_{k+1} = A_k x_k + B u_k$$  \[2\]

can be written as

$$x_4 = A_3 A_2 A_1 x_1 + C \bar{u},$$  \[3\]

where

$$C := A_3 A_2 B + A_3 B + B.$$

We seek the control $\bar{u}$ that minimizes the distance between $x(3)$ and the target $x_T$:

$$\min_{\bar{u}} \|x_T - A_3 A_2 A_1 x_1 - C \bar{u}\|.$$  \[4\]

We can see that an exact solution exists if

$$x_T - A_3 A_2 A_1 x_1 \in \text{span}(C),$$  \[5\]

and is given by

$$A_3 A_2 A_1 x_1 + C \bar{u} = x_T$$  \[6\]

$$C \bar{u} = x_T - A_3 A_2 A_1 x_1,$$  \[7\]

$$\bar{u} = C^T (x_T - A_3 A_2 A_1 x_1).$$  \[8\]

where $C^T$ denotes the Moore-Penrose pseudoinverse of the matrix $C$, computed using the singular value decomposition of $C$. Even when Eq. 5 is not satisfied, it is well established that the control Eq. 8 solves Eq. 4.

Define

$$d = \left\| (I_N - CC^T)(x_T - A_3 A_2 A_1 x_1) \right\|,$$  \[9\]

$$\mu := d_0 - d_*,$$  \[10\]

$\mu$ can be used to compare between potential TFs for a defined initial state ($x_I$), target state ($x_T$), and TF number ($p$). The larger the value of $\mu$, the higher the relative score for its corresponding TF set.

We note that accurate TF predictions for some desired target cell types may not depend on minimizing distance alone, but also the amount of "energy" required for the system to reach $d_*$. We denote energy here with $\mu_2$ and define it as:

$$e(u) = \sum_{k=0}^{K-1} u_k^T \cdot u_k = \mu_2.$$  \[11\]

$\mu_2$ is analogous to the amount of a TF that needs to be added to the system to achieve $d_*$. In the case where two different TFs achieve the same $\mu$ score, $\mu_2$ can be computed to decide the better candidate (i.e. lower $\mu_2$ implies a better TF).

Data Sources. A summary of the data used for this algorithm is shown below, with citations and accession numbers or website link, where applicable.

- **Gene Expression**
  - Fibroblast: Chen et al. (1)
  - GTEx: https://www.gtexportal.org/home/ (3)
  - ESC: GSE23316 (2)
  - Myotube: GSE52529 (2)
  - BluePrint Epigenome: http://www.blueprint-epigenome.eu/ (4)

- **DNase-seq**
  - Fibroblast: GSM1014531 (26)

- **Hi-C TAD boundaries**
  - Fibroblast: Chen et al. (1)

- **TF PWM**
  - Neph et al. (6)
  - MotifDB + FIMO (7, 27)

DGC Framework Benchmarking. In order to set a standard for success, we show here how experimentally validated TFs score well using our algorithm without imposing any TF threshold prior to analysis. Experimental validation of novel predictions will set a better standard for success, and while our lab and collaborators are working towards this goal, this is not a trivial undertaking.

As an initial test, we show here where the Yamanaka factors (KLF4, MYC, SOX2, and POU5F1) rank for fibroblast to embryonic stem cell reprogramming, in comparison to randomly selected combinations of four TFs using our methods ($n = 669$). Results show KLF4, MYC, SOX2, and POU5F1 ranking 12/669 (1.8%). Histogram of scores are depicted in Figure S5.

We also show where MYOD1 ranks for fibroblast to myotube, in comparison to randomly selected TFs ($n = 669$, full set of TFs included in our analysis, plus the case where a TF is ranked as both an activator and a repressor). Results show MYOD1 ranking 57/669 (8.6%). Histogram of scores are depicted in Figure S6.

For further statistical analysis, we attempt the benchmarking method performed in Michael et al. (28). This paper attempts to solve a conceptually similar problem to the TF prediction problem, where they attempt to predict the control imposed given an initial state, a network, and many target states with known “ground truths” (the TF manipulated to achieve this state is known). Though conceptually similar, our problems are different in the following ways:

- Often, there are multiple TF combinations that can result in successful reprogramming to a target state
- The number of TFs used in many validated direct reprogramming experiments is either computationally too time-consuming, is not included in our set of 547, or is used in combination with other molecules that we cannot model currently (e.g. small molecules and inhibitors)
• The data collected from each experiment for a given target state is performed in many different labs using different protocols. We know that this is a limitation in our paper as well, but point this out here for comparison to Michael et al. data.

• There is a much smaller number of “goal-state expression profiles” in our context, where the TFs for direct reprogramming are known, and are included in our model

Despite these crucial differences, we believe this is a very thorough and convincing benchmarking method, and we have attempted to try this on our dataset.

We first selected 10 validated cell reprogramming experiments where we believe we have a good approximation for the target state, and a sufficient number of the TFs included in our list of 547 TFs. These experimentally verified test cases are derived from PMID: 25658369, 18035408, 18029452, 18849973, 2748593. Figure S7 summarizes the data used for this analysis.

We can evaluate where our algorithm ranked a known reprogramming combination for a target (as a percentile of all combinations scored), without imposing any thresholding, n = 669 for all random TF combinations. Plotting this data similar to Figure 2B and 2C in Michael et al. yields Figure S8.

SI Figures.
Fig. S1. Analysis of cell cycle marker gene expression. Gene expression RNA-seq data for 39 genes that have been shown in the literature to be cell cycle regulated (29). Cell cycle phases shown include (A) G1/S, (B) S, (C) G2, (D) G2/M. Raw data of gene expression over time (left), with smoothed/interpolated expression over time with standard deviation (right). The expression curves for each gene have been standardized by subtracting their mean and dividing by the standard deviation over the eight time points. x-axis denotes sample time point \( k \), referring to 0, 8, 16, \ldots, 56 h after growth medium introduction. y-axis is normalized expression.
Fig. S2. Visualization of input matrix $B$. (A) Visualization of the $22,083 \times 547$ $c_{i,m}$ matrix: identified TF-to-gene interactions based on TFBSs. The color at entry $(i, m)$ represents how many TF $m$ TFBSs were observed within $\pm 5$kb of gene $i$'s TSS. The color axis has been truncated to $[0, 10]$ but note that more than 10 TFBSs were observed for many (gene,TF) pairs. Certain columns (TFs) are dramatically highlighted compared to others, some of which have been labeled by name along the horizontal axis. Some gene names are labeled along the vertical axis, none of which particularly stand out. Both genes and TFs are sorted alphabetically. (B) A histogram for the non-zero values of $c_{i,m}$. The log-scale on the vertical axis emphasizes that most of the gene TSS regions contain much less than 25 TFBSs for a given TF. The SP1 TFBS, for example, is observed 249 times in a 10kb TSS centered on a gene.
Fig. S3. Quantitative measure between cell types and TF scores. (A) $d_0$ values between all GTEx tissue types. (B) TF scores for an extended list of target cell types. $\gamma -$ fibroblast.
Fig. S4. Quantitative measure between cell types and TF scores for BLUEPRINT Project database. (A) $d_0$ values between BLUEPRINT Project cell types. (B) TF scores for an extended list of target cell types. $x_I$ = fibroblast.
Fig. S5. Histogram of TF scores for fibroblast to ESC reprogramming, for combinations of 4 TFs, without imposing any overexpression thresholding. We show here where the known reprogramming TF combination “Yamanaka Factors” (KLF4, MYC, SOX2, and POU5F1) rank in comparison to randomly selected combinations using our methods (n = 669). Results show KLF4, MYC, SOX2, and POU5F1 ranking 12/669 (1.8%).
Fig. S6. Histogram of TF scores for fibroblast to myotube reprogramming, for combinations of 1 TFs, without imposing any overexpression thresholding. We show here where MYOD1 ranks in comparison to randomly selected TFs using our methods (n = 669). Results show MYOD1 ranking 52/669 (7.8%).
<table>
<thead>
<tr>
<th>Initial cell population</th>
<th>Target cell type</th>
<th>Target cell type expression profile used</th>
<th>&quot;Ground truth&quot; Transcription factors and other inputs</th>
<th>Inputs in our dataset (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibroblast</td>
<td>ESC</td>
<td>ESC (GEO: GSE23316)</td>
<td>SOX2, POU5F1, KLF4, MYC</td>
<td>SOX2, POU5F1, KLF4, MYC (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>ESC</td>
<td>ESC (GEO: GSE23316)</td>
<td>SOX2, POU5F1, NANOG</td>
<td>SOX2, POU5F1, NANOG (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>ESC</td>
<td>ESC (GEO: GSE23316)</td>
<td>SOX2, POU5F1</td>
<td>SOX2, POU5F1 (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Myotube</td>
<td>Myotube (GEO: GSE52529)</td>
<td>MYOD1</td>
<td>MYOD1 (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Skeletal Muscle</td>
<td>Skeletal Muscle (GTEX)</td>
<td>MYOD1</td>
<td>MYOD1 (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Hepatocyte</td>
<td>Liver (GTEX)</td>
<td>HNF1A, HNF4A, HNF4, CEBPA, ATF4, PAX8, POU5F1, HNF1A, CEBPA, ATF4, PAX8, POU5F1, p53-siRNA MYC (75%)</td>
<td></td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Hepatocyte</td>
<td>Liver (GTEX)</td>
<td>HNF1A, HNF4A, FOXA3, SV40 large T antigen</td>
<td>HNF1A, HNF4A (50%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Neuron</td>
<td>Neuron (PMID: 2518741)</td>
<td>SOX10</td>
<td>SOX10 (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Neuron</td>
<td>Neuron (PMID: 2518741)</td>
<td>SOX2</td>
<td>SOX2 (100%)</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>Neuron</td>
<td>Neuron (PMID: 2518741)</td>
<td>POU3F2, ASCL1, MYT1L, Lhx3, MNX1, ISL1, NEUROG2</td>
<td>POU3F2, Lhx3, MNX1, ISL1, NEUROG2 (71%)</td>
</tr>
</tbody>
</table>

**Fig. S7.** Table overview of target states and target TF combinations for statistical benchmarking.
**Fig. S8.** Cumulative percentage of correct predictions (TFs that have been experimentally validated; y-axis) selected above the percentage rank indicated on the x-axis. The 10 goal states used are shown in Table S7. Text boxes show where a given goal state ranked, with the number in parenthesis giving the exact rank in proportion to all TFs ranked. Our algorithm ranks known reprogramming TFs (green) above what would be expected by random chance (blue), without relying on TF overexpression.