Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes

Alex H. Lang
Dept. of Physics, Boston University, Boston, MA 02215

Hu Li
Dept. of Biomedical Engineering, Boston University, Boston, MA, 02215 and
Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, 02215

James J. Collins
Dept. of Biomedical Engineering, Boston University, Boston, MA, 02215
Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, 02215
Howard Hughes Medical Institute, Boston, MA, 02215 and
Center for BioDynamics, Boston University, Boston, MA 02215

Pankaj Mehta
Dept. of Physics, Boston University, Boston, MA 02215 and
Center for Regenerative Medicine, Boston University, Boston, MA, 02215
(Dated: November 16, 2012)

A common metaphor for describing development is a rugged epigenetic landscape where cell fates are represented as attracting valleys resulting from a complex regulatory network. Here, we introduce a framework for explicitly constructing epigenetic landscapes that combines genomic data with techniques from physics. Each cell fate is a dynamic attractor, yet cells can change fate in response to external signals. Our model suggests that partially reprogrammed cells are a natural consequence of high-dimensional landscapes and predicts that partially reprogrammed cells should be hybrids that co-express genes from multiple cell fates. We verify this prediction by reanalyzing existing data sets. Our model reproduces known reprogramming protocols and identifies candidate transcription factors for reprogramming to novel cell fates, suggesting epigenetic landscapes are a powerful paradigm for understanding cellular identity.

Understanding the molecular basis of cellular identity and differentiation is a major goal of modern biology. This is especially true in light of the work of Takahashi and Yamanaka demonstrating that the overexpression of just four transcription factors (TFs) is sufficient to convert somatic fibroblasts into cells resembling embryonic stem cells (ESCs), dubbed induced pluripotent stem cells (iPSCs) [1]. The idea of using a small set of TFs to reprogram cell fate has proven to be extremely versatile and reprogramming protocols now exist for generating neurons [2], cardiomyocytes [3], liver cells [4, 5], and human iPSCs [6]. Despite these revolutionary experimental advances, cell fate is still poorly understood mechanistically and theoretically. Recent experiments suggest cell fates can be viewed as high-dimensional attractor states of the gene regulatory networks underlying cellular identity [7]. In particular, cell fates are characterized by a robust gene expression and epigenetic state resulting from the complex interplay of transcriptional regulation, chromatin regulators, non-coding and micro RNAs, and signal transduction pathways.

These experiments have renewed interest in the idea of an 'epigenetic landscape' that underlies cellular identity [8–11]. In the landscape metaphor, cell fates are viewed as attracting basins in a rugged landscape and differentiation proceeds through signal-dependent 'low-energy' valleys that connect cell fates (see Figure 1). Cells can also change fates due to probabilistic barrier crossings between valleys as in cellular reprogramming[12, 13].

Traditionally, epigenetic landscapes have been modeled as low-dimensional systems using either a small subset of genes [14] or a few cell fates [15]. These methods cannot easily be scaled to higher dimensions. Instead, inspired in part by the success of statistical energy landscapes in protein folding [16], we avoid these limitations by combining large-scale genomic data with techniques from spin glass physics and neural networks [17–20]. Using only data of mouse microarray gene expression states as input, we construct a parameter-free epigenetic landscape model for cellular identity with 95 stable cell fates and 1152 TFs. Each cell fate is a robust attractor, yet cells can deterministically switch fates in response to external signals. Our model provides a unified framework to discuss differentiation and reprogramming. It also naturally explains the existence of partially reprogrammed cell fates (stable cell fates found in reprogramming experiments but not *in vivo*) as 'spurious' attractors resulting from the high dimensionality of the landscape. Our model predicts, and we verify, that partially reprogrammed cells are hybrids that co-express TFs of multiple naturally occurring cell fates. Finally, our model reproduces known reprogramming protocols to iPSCs,

heart, and liver, and can be used for designing reprogramming protocols to novel cell fates. Taken together, these results suggest that epigenetic landscapes represent a powerful framework for understanding the molecular circuitry and dynamics that gives rise to cell fate.

Any landscape model must reproduce several experimental observations. Most importantly, all cell fates must be robust attractors, yet allow cells to change fate through rare stochastic transitions as in cellular reprogramming experiments [21, 22]. A common result of reprogramming is not the desired cell fate, but partially reprogrammed cells [23, 24]. These results suggest that the landscape is rugged and may contain additional spurious attractors corresponding to cell fates that do not naturally occur *in vivo*. In addition, environmental and external signals can control cell fates. Some environments stabilize particular cell fates (Fig. 1B). A dramatic example of this is a protocol for reprogramming to neural progenitor cells (NPCs) that is identical to Yamanaka's protocol for reprogramming to ESC except for the culturing media [25]. Other external signals deterministically switch cell fates, as occurs in normal development (Fig. 1C) [26]. Together, these imply the landscape is a dynamic entity that depends on environmental signals, and we will show that our landscape successfully incorporates all of these experimental observations.

CONSTRUCTING EPIGENETIC LANDSCAPES

The epigenetic state space

Cellular identity and differentiation are largely controlled by epigenetics, especially histone modifications (HMs)[27] (Fig. 2A). Consequently, the ideal data set for constructing epigenetic landscapes are the genome-wide HM states for genes in all cell fates. However, unlike microarrays, global HM data are limited to a few cell fates [28, 29]. To circumnavigate this problem, we used available HM data and compared them to microarray gene expression levels for available cell fates; this created a conditional probability distribution of having a HM given a TF expression level (Fig. 2B). We found a sharp threshold which distinguished genes with the activating modification of histone 3 tri-methylation at lysine 4 (K4) from genes with the inactivating modification of histone 3 tri-methylation at lysine 27 (K27) and poised/bivalent genes (both K4 and K27). This threshold implies that continuous TF expression levels can be used to infer the discrete HM states. We then used 393 relevant whole genome microarrays (details in SI) to create a binary TF state for $N = 1152$ TFs (labeled by Latin indices i, j) in $p = 95$ cell fates (labeled by Greek indices μ, ν). A TF was designated active, $(+1)$, if its expression level in the corresponding microarray was above the threshold and inactive, (-1), otherwise (Fig. 2C). These binary (i.e. on/off) TF data are the only biological input into our model. We restricted our considerations to TFs due to their importance in cellular reprogramming and differentiation. However, our model can be easily generalized to include other important genes.

Motivation from attractor neural networks

The Takahashi and Yamanaka reprogramming experiments [1] are reminiscent of content-addressable memory and attractor neural networks. A content-addressable memory allows one to retrieve a full memory based on sufficient partial information. To paraphrase the original Hopfield paper[17], if we want to recall a paper citation, for example,"John J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, 1982," a content-addressable memory allows us to recall the full citation with the partial recall "Hopfield Neural networks 1982" or other sufficient details. In the Yamanaka reprogramming protocol, overexpressing only four TFs is enough for a fibroblast to "recall" the global TF expression of an ESC. A content-addressable memory is naturally represented as a basin of attraction in a dynamical system, with partial recall corresponding to entering the basin of attraction and full recall corresponding to reaching the minimum of the basin. Hopfield attractor neural networks [17, 18, 20] are a general method to take an input set of vectors ("memories") and explicitly construct a unique, global, landscape such that each input vector is a global minimum and has a basin of attraction.

Mathematical model

We now give a brief mathematical description of how we construct epigenetic landscapes. An arbitrary state of a cell is represented as a vector S_i of length $N = 1152$, with the ith component of the vector $+1$ if TF i is active and -1 if it is inactive. A cell fate $\mu = 1 \dots 95$ is characterized by its binary TF expression vector, ξ_i^μ. Our data set determine the ξ_i^μ and these are the only biological input into the landscape.

A direct application of the original Hopfield neural network [17] fails for the real biological data set we constructed. The original Hopfield method works for vectors whose components are random, independent variables with equal probability of on $(+1)$ and off (-1) states. This leads to Gaussian noise between memories [18]. However, cell fates are highly correlated (Fig. S1), as shown by the cell type correlation matrix $A^{\mu\nu} = 1/N \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$ which characterizes the correlation between cell fate μ and ν. These correlations lead to noise that globally destabilizes all basins of attraction. The Hopfield construction can be generalized to

the projection-method which can incorporate correlated "memories" [19].

Mathematically, the high correlation between cell fates implies that the overlap (vector dot-product), $m^\mu = 1/N \sum_{i=1}^{N} \xi_i^\mu S_i$, is a poor measure of how close a cell with expression state S_i is to a cell fate μ. Instead, the "projection" accurately characterizes correlated vectors and is given by $a^\mu = \sum_{\nu=1}^{p} (A^{-1})^{\mu\nu} m^\nu$, of S_i on a cell fate μ. The projection measures the orthogonal projection of a state S_i onto the subspace spanned by naturally occurring cell fates, $\{\xi_i^\mu\}$ (see Fig. 2D and SI).

We now explicitly construct the landscape using the potential energy (Lyapunov function), H, given by

$$H = -\frac{1}{2}\sum_{i,j} S_i J_{ij} S_j - \frac{N}{2}\sum_{\mu} \chi^\mu a^\mu - \frac{1}{2}\sum_{i,j} S_i K_{ij}(\vec{c}) S_j \tag{1}$$

The coupling constants $J_{ij} = 1/N \sum_{\mu=1}^{p} \sum_{\nu=1}^{p} \xi_i^\mu (A^{-1})^{\mu\nu} \xi_j^\nu$ represent effective, correlation-based, interactions between TFs that may be indirect and give rise to stable basins of attraction around each cell fate. The χ^μ are an environment-dependent local field that can stabilize or destabilize a specific cell fate μ. The $K_{ij}(\vec{c})$ are differentiation couplings that switch between cell types and can depend on the environment through the external input $\vec{c}$. Future work will explore the exact dependence of switching on the external environment. See Box 1 for technical description, and see SI for more details of the Hamiltonian and a geometric interpretation.

The dynamics of the network proceed by random, asynchronous updates [20] according to the probability

$$P[S_i(t+1)] = \frac{e^{\beta h_i(t) S_i(t)}}{e^{\beta h_i(t)} + e^{-\beta h_i(t)}} \tag{2}$$

with the local field on TF i given by $h_i = -\frac{\partial H}{\partial S_i}$, and β an effective noise parameter that controls the level of stochasticity resulting from biochemical noise (see SI). This update time cannot be directly related to biological time. Instead, inspired by experiments showing that reprogramming rates scale with cell division rates [12], and the observation that cellular divisions produce HM errors [30], we introduced an additional source of stochasticity into the dynamics by periodically flipping a fixed percentage of TF states to mimic cell division (see SI).

RESULTS

Cell fates are dynamic attractors that are responsive to signals

We tested our model using two *in silico* experiments. To verify that naturally occurring cell fates are dynamic attractors, we randomly perturbed the TF state of cells, S_i, from the ESC state and then tracked the TF state over time. Fig. 2E shows the projection of the TF state on the ESC state as a function of time. For a large number of starting conditions, after an initial transient, the system relaxes back to the ESC state (red bracket), explicitly demonstrating the existence of a large basin of attraction [7]. Originally, the interactions, J_{ij}, between TFs are symmetric, while biologically realistic interactions should be non-symmetric, implying a non-Lyapunov pseudo-potential [31]. The symmetry condition can be relaxed, for example by randomly deleting 20% of interactions (Fig. 2E Diluted). The cell fates remain robust attractors even for these pseudo-potentials. Our model can also deterministically switch between cell fates in response to differentiation signals. For example, the common myeloid progenitor (CMP) is a blood cell fate that *in vivo* can differentiate into either granulo-monocytic progenitors (GMP) or megakaryocyte-erythroid progenitors (MEP). In Fig. 2F, we show *in silico* experiments where we start the system in the CMP state and show the trajectories after applying either the GMP (signal 1, blue) or MEP (signal 2, red) differentiation signal, resulting in branching to two distinct cell fates.

Partially reprogrammed cells as "spuirous" attractors

Partially reprogrammed cells have a natural interpretation in our model as spurious attractors arising from the high dimensionality ($N = 1152$) of the state space [20]. These spurious attractors are guaranteed by topology of high-dimensional vector spaces and can be interpreted as potential cell fates that do not occur *in vivo*. Naively, one would expect partially reprogrammed cells to have minimal projection on natural cell fates since in high-dimensional vector spaces, any two random vectors are orthogonal (Fig. S2). The literature on neural networks [19] specifically predicts that in our model, spurious attractors should be hybrid cells that co-express genes from multiple cell fates, not necessarily including the beginning or ending cell fate. Reanalyzing all existing genome-wide data sets on partially reprogrammed cells (Table 1) validates our prediction. The purity of the partially reprogrammed cell colonies is important because a heterogenous sample could mimic hybrids. However, nearly all studied partially reprogrammed cells grew as homogenous colonies. Therefore, the limited experimental data does not support the idea of partially reprogrammed cells being explained by cell culture heterogeneity. In addition, reexamination of a claimed iPSC-to-NSC conversion [32] shows that the resulting cell fate is more accurately characterized as a partially reprogrammed cell. This illustrates how the techniques developed here can be used to improve the classification of reprogrammed cells.

TABLE I. **Partially reprogrammed cells as spurious attractors.** Partially reprogrammed cell lines (first column) and their significant projections (i.e 2 std above noise) onto "natural" cell fates based on microarray data. The last column indicates whether partially reprogrammed cell lines were homogeneous. The iPSC-NSC was originally classified as an NSCs [32] but is more accurately characterized as a partially reprogrammed hybrid of NPCs and NSCs. Abbreviations: iPSC, induced pluripotent stem cell; NSC, neural stem cell; NPC, neural progenitor cell; ESC, embryonic stem cell; MEF, mouse embryonic fibroblast; MPP, multi-potent progenitor.

Cell line	Start	Goal	Highest projecting states (projection)	Homogeneous?
1A2 [23]	MEF	ESC	MEF (0.187), ESC (0.163), muscle-smooth (0.126)	Yes
1B3 [23]	MEF	ESC	MEF (0.161), ESC (0.149)	Yes
BIV1+ [24]	B Cell	ESC	melanocyte (0.1548), MPP (0.136)	No info
BIV1- [24]	B Cell	ESC	ESC (0.434), pro-mesoderm (0.162), MPP (0.153), neural crest SC (0.127)	No info
MCV6 [24]	MEF	ESC	ESC (0.264), MPP (0.145), pre-erythroid (0.133), smooth muscle (0.132)	Yes
MCV8 [24]	MEF	ESC	ESC (0.180), MPP (0.160), melanocyte (0.143) large intestine (0.135)	No
iPSC-NSC [32]	iPSC	NSC	NPC (0.428), NSC (0.281)	Yes

Identifying transcription factors for cellular reprogramming

Our landscape model also provides a quantitative method to identify candidate TFs for reprogramming. First, recent experiments provide evidence that reprogramming TFs should be based only on final, not initial, cell fates[22]. Second, intuitively, reprogramming candidates should be both highly expressed and highly "predictive" of the desired cell fate. Since TF expression levels are well-fit by a log-normal distribution (Fig. S3), the log-normal z-score naturally defines high and low TF expression levels. Within our landscape, the "predictivity" of a TF, for a given cell fate, is measured by its contribution to the potential energy of that cell fate. This "projection-contribution" is mathematically represented as $\eta_i^\mu = 1/N \sum_{\nu=1}^{p} (A^{-1})^{\mu\nu} \xi_i^\nu$ for TF i in cell fate μ. To obtain a single quantitative rank, we can multiply the z-score by the projection-contribution.

We validate our candidate reprogramming TFs by comparing to existing protocols. Fig. 3 shows every TF's projection-contribution versus z-score in ESC, heart (cardiomyocytes), and liver (hepatocytes). Note that neurons were not included due to lack of compatible microarray data. ESC are examined in Fig. 3A and Fig. 3B. We have explicitly labeled TFs from existing reprogramming protocols [21, 22]. As expected, these TFs are both predictive and highly expressed (upper right hand corner of graphs). To check the biological validity of our predictions (SI for details), we analyzed the GO Annotation of our top 100 candidates for ESC reprogramming (Table S1). Within these top TFs, 9 have successfully been used in reprogramming, 8 are known pluripotency TFs (involved in maintaining stem cell fate), while 32 have no known function as of yet and are intriguing reprogramming candidates.

ESCs are unusual in that the highly expressed TFs are also among the most predictive (e.g. Pou5f1/Oct4, Nanog). The true benefit of using both z-score and projection-contribution is best demonstrated by examining reprogramming protocols to heart (Fig. 3C) and liver (Fig. 3D) (Table S2). For example, in heart (cardiomyocytes) reprogramming[3], Gata4 is ranked 68th in z-score TF expression, but our combined ranking of z-score and projection-contribution ranks Gata4 as 6th. Similar results are seen in liver (hepatocyte) protocols [4, 5] (see SI).

In most cell fates, the highly expressed TFs are not necessarily predictive, suggesting landscapes may be useful for rationally-designing reprogramming protocols to novel cell fates. Using our landscape model, we have identified the top 100 candidates for overexpression and as well as the top 100 candidates for knockouts for all cell fates where we have reliable data (see SI). Besides being candidates for reprogramming, the predictive TFs can be used as markers for each cell fate.

DISCUSSION

Our work suggests that epigenetic landscapes are a powerful paradigm for understanding cellular identity and reprogramming. Remarkably, despite having no free parameters or biological knowledge beyond binary TF states of cell fates, our landscape models can explain partially reprogrammed cell fates and identify TFs used in cellular reprogramming protocols.

The epigenetic landscapes of cellular identity constructed in this paper are related to but distinct from the standard Waddington landscape for development [8]. In both landscapes, cell fates are viewed as valleys of a potential. However, Waddington's landscape is concerned with how developmental signals give rise to different cell fates. Hence, the axis of the original Waddington landscape can be interpreted as a combination of time, signals, as well the state of molecular components such as transcription factors. In more modern language, Waddington's landscape is concerned with the geometry of bifurcations in space of possible developmental signals (see [33] for an example). In contrast, in the landscapes considered in this paper the state space is composed solely of the epigenetic state of transcription factors and developmental signals, $\vec{c}$, couple to the landscape through the signal-dependent parameters $K_{ij}(\vec{c})$.

Currently, our model has several limitations centered around dynamics. First, for ESC reprogramming, our model does not highlight the importance of the non-specific transcription factor Myc (see SI for a detailed discussion). Second, as mentioned previously, our use of asynchronous dynamics cannot be directly related to biological time. Lastly, our landscape does not accurately reflect the dynamics of reprogramming. Simulations of reprogramming with known protocols, such as the Yamanaka protocol, lead to rates of reprogramming that are comparable to the rates from a reprogramming simulation with a randomly selected protocol. This is likely due to the fact that cell fates are extremely stable and hence reprogramming is extremely rare and hence hard to stochastically sample.

Our model has several possible extensions. Following previous work on neural network [34], our landscape can be generalized to continuous TF expression levels. Additionally, our landscapes can provide a multitude of predictions given the correct data. Our framework can easily be generalized to include microRNAs, other genes, or the human epigenetic landscape given appropriate data sets. This opens up possibilities of improving upon the high reprogramming rates achieved by overexpressing microRNAs [35] or synthetic mRNAs [36]. Another attractive element of the framework presented here is that it allows for a quantitative analysis of whole genome-wide expression states (see Table 1). This is likely to yield a more accurate classification of reprogrammed cells and allow for the classification as well as the identification of diseased cell fates. Finally, our epigenetic landscape may also prove useful for designing more efficient directed differentiation protocols [37]. Overall epigenetic landscapes provide a unifying framework for cell identity, reprogramming, and directed differentiation.

MATERIALS AND METHODS

All microarray data utilize the Affymetrix GeneChip Mouse Genome 430 2.0 platform and were downloaded using ArrayExpress (www.ebi.ac.uk/arrayexpress). The details of all 393 microarrays can be found under accession GSE (to be determined). Microarray probe-to-gene map was created with Bioconductor 2.10. All raw microarray files were processed in one batch by robust mean averaging (RMA) in MATLAB. Since we were interested in cellular identity, only transcription factors, transcription factor co-factors, or chromatin remodeling genes were kept (for short hand, referred to as transcription factors (TF) throughout the text) [38].

Ideally, we would like the global histone modification (HM) state of all cell fates since these are the primary input into our model. However, global HM data are limited [28, 29]. Consequently, we used the global HM data for these three cell fates and compared them to microarray TF expression levels. This allowed us to create a conditional probability distribution of each HM for a given TF expression level (Fig. 2B). We found a sharp cutoff (≈ 5.5) which distinguished TFs with the activating modification of histone 3 tri-methylation at lysine 4 (K4) from TFs with the inactivating modification of histone 3 tri-methylation at lysine 27 (K27), poised/bivalent TFs (both K4 and K27), and no HM (most likely DNA methylation).

This conditional probability distribution allowed us to create a binary expression state for each cell fate from our microarray data. TFs with expression above 5.5 were designated on, $(+1)$, while TFs below 5.5 were designated off, (-1). The conclusions presented in the paper are robust to the threshold choice (not shown, but similar conclusions reached for a cutoff of 5 or 6). See SI for full details.

ACKNOWLEDGEMENTS

We thank members of the Mehta Group, Collins lab, and Laertis Ikonomou, Darrell Kotton, and other members of Boston University Center for Regenerative Medicine (CReM) for stimulating discussions. In addition, we thank Laertis Ikonomou, Darrell Kotton, and Kristian Moss Bendtsen for a detailed reading of the manuscript. AHL was supported by a Boston University Dean's Fellowship and a National Science Foundation

Graduate Research Fellowship (NSF GRFP) under Grant No. DGE-0741448. PM was supported by a Sloan Fellowship. AHL, JJC, and PM designed the research. HL helped contribute relevant microarray databases. AHL and PM performed the research, analyzed the data, and wrote the paper. The data reported in this paper are tabulated in the Supporting Online Material and archived at ArrayExpress Archive (www.ebi.ac.uk/arrayexpress) under GSE (to be determined).

[1] Takahashi, K & Yamanaka, S. (2006) *Cell* **126**, 663–676.
[2] Vierbuchen, T, et al. (2010) *Nature* **463**, 1035–1041.
[3] Ieda, M, et al. (2010) *Cell* **142**, 375–386.
[4] Sekiya, S & Suzuki, A. (2011) *Nature* **475**, 390–393.
[5] Huang, P, et al. (2011) *Nature* **475**, 386–389.
[6] Yu, J, et al. (2007) *Science* **318**, 1917–1920.
[7] Huang, S, Eichler, G, Bar-Yam, Y, & Ingber, D. E. (2005) *Phys Rev Lett* **94**.
[8] Waddington, C. H. (1957) *The Strategy of the Genes.* (Allen and Unwin, London).
[9] Kauffman, S. A. (1993) *The Origins of Order: Self-Organization and Selection in Evolution.* (Oxford University Press, USA).
[10] Enver, T, Pera, M, Peterson, C, & Andrews, P. W. (2009) *Cell Stem Cell* **4**, 387–397.
[11] Zhou, J. X & Huang, S. (2011) *Trends in Genetics* **27**, 55–62.
[12] Hanna, J, et al. (2009) *Nature* **462**, 595–601.
[13] Flöttmann, M, Scharp, T, & Klipp, E. (2012) *Frontiers in Physiology* **3**.
[14] MacArthur, B. D, Ma'ayan, A, & Lemischka, I. R. (2009) *Nat Rev Mol Cell Biol* **10**, 672–681.
[15] Huang, S, Guo, Y.-P, May, G, & Enver, T. (2007) *Dev Biol* **305**, 695–713.
[16] Bryngelson, J, Onuchic, J, Socci, N, & Wolynes, P. (1995) *Proteins: Structure, Function, and Bioinformatics* **21**, 167–195.
[17] Hopfield, J. J. (1982) *Proc Natl Acad Sci U S A* **79**, 2554–2558.
[18] Amit, D. J, Gutfreund, H, & Sompolinsky, H. (1985) *Phys Rev A* **32**, 1007–1018.
[19] Kanter, I & Sompolinsky, H. (1987) *Phys Rev A* **35**, 380–392.
[20] Amit, D. (1992) *Modeling brain function: The world of attractor neural networks.* (Cambridge Univ Pr).
[21] González, F, Boué, S, & Belmonte, J. C. I. (2011) *Nat Rev Genet* **12**, 231–242.
[22] Buganim, Y, et al. (2012) *Cell* **150**, 1209–1222.
[23] Sridharan, R, et al. (2009) *Cell* **136**, 364–377.
[24] Mikkelsen, T. S, et al. (2008) *Nature* **454**, 49–55.
[25] Kim, J, et al. (2011) *Proc Natl Acad Sci U S A* **108**, 7838–7843.
[26] Davidson, E. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution.* (Academic Press).
[27] Jenuwein, T & Allis, C. D. (2001) *Science* **293**, 1074–1080.
[28] Mikkelsen, T. S, et al. (2007) *Nature* **448**, 553–560.
[29] Meissner, A, et al. (2008) *Nature* **454**, 766–770.
[30] Ben-David, U, Mayshar, Y, & Benvenisty, N. (2011) *Cell Stem Cell* **9**, 97–102.
[31] Wang, J, Xu, L, Wang, E, & Huang, S. (2010) *Biophysical Journal* **99**, 29–39.
[32] Kim, J. B, et al. (2009) *Cell* **136**, 411–419.
[33] Corson, F & Siggia, E. D. (2012) *Proc Natl Acad Sci U S A* **109**, 5568–5575.
[34] Hopfield, J. J. (1984) *Proc Natl Acad Sci U S A* **81**, 3088–3092.
[35] Yoo, A. S, et al. (2011) *Nature* **476**, 228–231.
[36] Warren, L, et al. (2010) *Cell Stem Cell* **7**, 618–630.
[37] Longmire, T. A, et al. (2012) *Cell Stem Cell* **10**, 398–411.
[38] Zhang, H.-M, et al. (2012) *Nucleic Acids Research* **40**, D144–D149.

Box 1. Overview of our quantitative cellular identity landscape

- N transcription factors (TF) labeled by Latin indices i and p cell types labeled by Greek indices μ. Each TF is either on $(+1)$ or off (-1).
- A general network state is represented by a vector S_i of length N. A cell type μ is represented by the vector ξ_i^μ.
- We require all ξ_i^μ to be attractors in the landscape. This is ensured by constructing a correlation-based interaction network

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\sum_{\nu=1}^{p}\xi_i^\mu (A^{-1})^{\mu\nu}\xi_j^\nu \tag{3}$$

  with $A^{\mu\nu}$ the correlation matrix between cell types:

$$A^{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu \xi_i^\nu. \tag{4}$$

- The commonly used overlap order parameter is the "magnetization"

$$m^\mu = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu S_i. \tag{5}$$

- However, cell types are highly correlated with each other. Thus, a more accurate order parameter is the "projection", a^μ, on each cell type

$$a^\mu = \sum_{\nu=1}^{p}(A^{-1})^{\mu\nu}m^\nu. \tag{6}$$

- The specific "projection-contribution" of a given TF i to the projection on cell type μ is given by

$$\eta_i^\mu = \frac{1}{N}\sum_{\nu=1}^{p}(A^{-1})^{\mu\nu}\xi_i^\nu \tag{7}$$

- Cell types may be stabilized by external conditions (such as growing in a favorable culture media). For example, cell type μ is favored by $\chi^\mu a^\mu$, where χ characterizes the degree of the stabilization.
- External signals, $\vec{c}$, can induce switching between cell types and is represented by the matrix $K_{ij}(\vec{c})$. See SI for more detail.
- The landscape is defined by function H. Term 1 is an energy, or Lyapunov function, that generates the basins of attraction. Term 2 stabilizes specific cell types (ie culture term). Term 3 is non-Lyapunov and represents switching due to external signaling.

$$H = -\frac{1}{2}\sum_{ij}S_i J_{ij}S_j - \frac{N}{2}\sum_{\mu}\chi^\mu a^\mu - \frac{1}{2}\sum_{ij}\sum_{\mu\nu}S_i K_{ij}(\vec{c})S_j. \tag{8}$$

- TFs can stochastically switch states. A TF is biased towards a state by its interactions with the network through its local field $h_i = -\frac{\partial H}{\partial S_i}$. The dynamics are stochastic and controlled by a global noise parameter β. At each time step, one TF is updated with the probability of state $S_i(t+1)$ at time $t+1$ related to the state $S_i(t)$ and local field $h_i(t)$ at time t by

$$P[S_i(t+1)] = \frac{Exp[\beta h_i(t)S_i(t)]}{Exp[\beta h_i(t)] + Exp[-\beta h_i(t)]}. \tag{9}$$

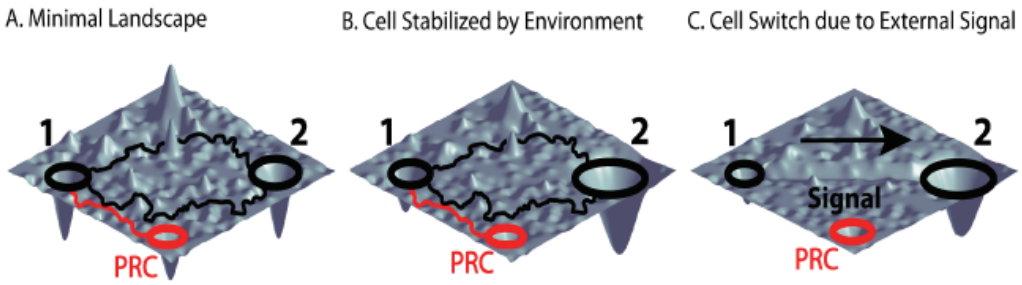

FIG. 1. **Phenotypic Landscape.** These landscapes envision cell fates as attractors in a signal-dependent landscape. (A) The minimal cellular identity landscape. Each cell fate is a basin of attraction (black circles). Reprogramming between different cell fates (1 and 2) can occur probabilistically via different trajectories (black paths). Partially reprogrammed cells (PRC) exist as smaller, spurious, basins of attraction (red circle) that can be experimentally observed by reprogramming experiments (example trajectory in red). (B) Same cellular identity landscape in the presence of a stabilizing environment (ex. favorable culturing medium) for cell fate 2. The environment increases the radius and depth of the cell fate 2 basin of attraction. (C) Landscape in the presence of an external signal that gives rise to differentiation from cell fate 1 to cell fate 2 (ex. growth factors associated with differentiation). Notice the low energy path between the cell fates that drives switching from cell fate 1 to cell fate 2.

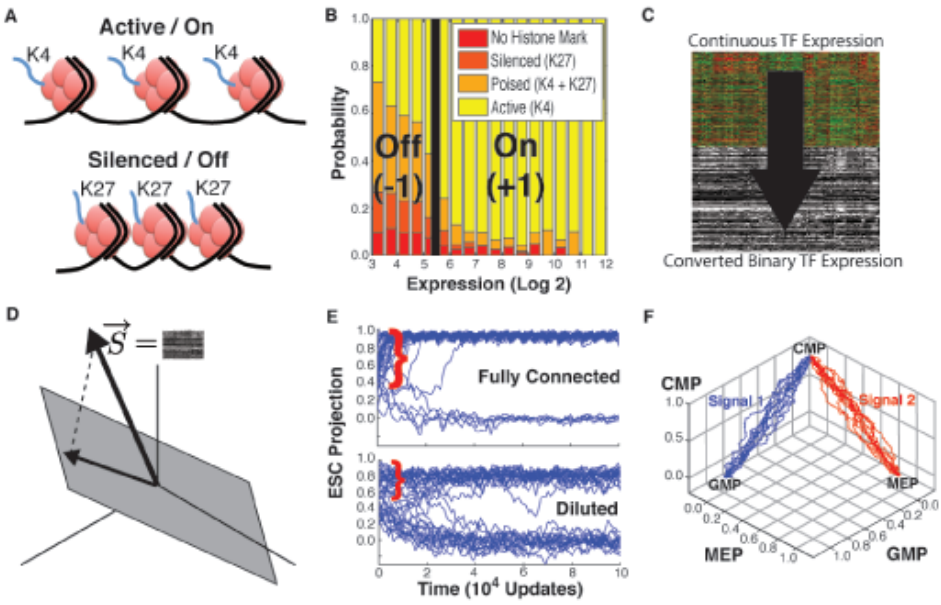

FIG. 2. **Overview of model.** (A) Histone 3 tri-methylation at lysine 4 (K4) is associated with active genes, while histone 3 tri-methylation at lysine 27 (K27) is associated with repressed genes. (B) Conditional probability distribution of histone modification (HM) given transcription factor (TF) expression levels derived by comparing microarray data with HM data from [28, 29]. Notice the sharp threshold (black line) between expression levels of active and inactive TFs. (C) Using (B), continuous TF expression levels is converted into binary states. (D) An arbitrary state is represented by a vector $\vec{S}$ of ± 1, with each dimension in the vector space representing the state of a TF. The natural cell fates form a subspace (gray plane). The landscape model is based on the orthogonal projection of the TF state onto this subspace. (E) The dynamics of the landscape model for different initial conditions for a fully connected interaction matrix J_{ij} and a diluted interaction matrix where 20% of interactions have been randomly deleted. Plot shows the projection of S on embryonic stem cells (ESC) as function of time. Notice the large basins of attraction (red bracket). (F) Simulations showing how a common myeloid progenitor (CMP) can differentiate into either granulo-monocytic progenitors (GMP) or megakaryocyte-erythroid progenitors (MEP) in response to two distinct external signals.

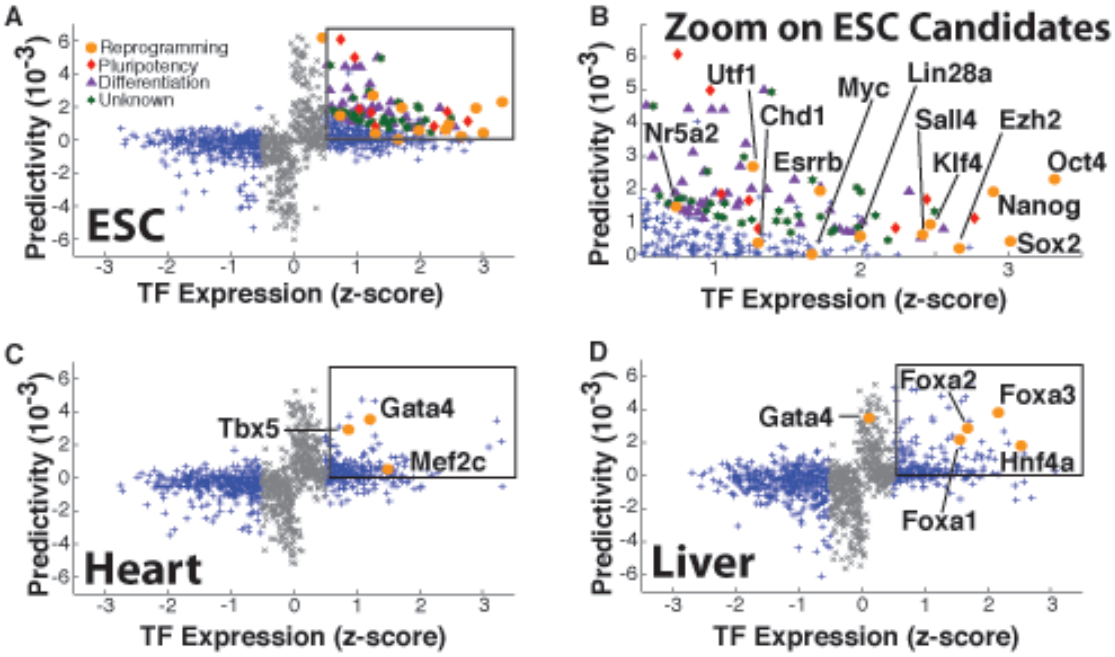

FIG. 3. **Identifying reprogramming candidates.** For a given cell fate, we plot every transcription factor's (TF) predictivity (aka energy projection-contribution, η_i^μ) vs TF expression level. Gray TFs are within the the uncertainty introduced by making data discrete (binary). All reprogramming TFs are in a pre-existing protocol. Boxed TFs are potential reprogramming candidates since they have both high expression level and high projection-contribution. (A) Embryonic stem cell graph (ESC). (B) Zoom in of box in (A). TFs used in known reprogramming protocols are labeled [21, 22]. (C) Heart. Labeled TFs can reprogram fibroblasts to cardiomyocytes [3]. (D) Liver. Labeled TFs can reprogram fibroblasts to liver cells. One protocol used Hnf4a plus any of Foxa1, Foxa2, or Foxa3 [4] while another used Gata4, Foxa3, Hnf1a, and deletion of p19Arf [5]. Hnf1a and p19Arf were not differentially expressed in our microarrays and are not shown.

Supplementary Information: Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming gene

Alex H. Lang, Hu Li, James J. Collins, and Pankaj Mehta

Contents

I. DATA SET CONSTRUCTION

All microarray data utilize the Affymetrix GeneChip Mouse Genome 430 2.0 platform and were downloaded using ArrayExpress (www.ebi.ac.uk/arrayexpress). The details of all 393 microarrays can be found under accession GSE (to be determined). Microarrays are all control or natural conditions. Microarray probe-to-gene map was created with Bioconductor 2.10. All raw microarray files were processed in one batch by robust mean averaging (RMA) in MATLAB, and genes with multiple microarray probes were averaged. At this point, the data set consisted of 393 microarrays with 20877 genes. Since we were interested in cellular identity, only transcription factors, transcription factor co-factors, or chromatin remodeling genes were kept (for short hand, referred to as transcription factors (TF) throughout the text)[1], leaving 1612 TFs.

Ideally, we would like the global histone modification (HM) state of all cell fates since these are the primary input into our model. However, global HM data are limited to embryonic stem cells (ESC), mouse embryonic fibroblasts (MEF), and neural progenitor cells (NPC)[2,3]. Consequently, we used the global HM data for these three cell fates and compared them to microarray TF expression levels. This allowed us to create a conditional probability distribution of each HM for a given TF expression level (Fig. 2B). We found a sharp cutoff (≈ 5.5) which distinguished TFs with the activating modification of histone 3 tri-methylation at lysine 4 (K4) from TFs with the inactivating modification of histone 3 tri-methylation at lysine 27 (K27), poised/bivalent TFs (both K4 and K27), and no HM (most likely DNA methylation).

This conditional probability distribution allowed us to create a binary expression state for each cell fate from our microarray data. TFs with expression above 5.5 were designated on, $(+1)$, while TFs below 5.5 were designated off, (-1). The conclusions presented in the paper are robust to the threshold choice (not shown, but similar conclusions reached for a cutoff of 5 or 6). After the binarization of TF expression, all TFs that were not differentially expressed across cell fates (i.e. TFs that are always on / always off in every cell fate) were dropped, leaving 1152 TFs. The binarized TF expression for the 95 cell fates was found by first binarizing all 393 microarrays and then taking the majority vote for each cell state (with ties broken by averaging the continuous data). The final result was the binary expression state for 95 cell fates and 7 partially reprogrammed cell fates with 1152 TFs.

Several self-consistency checks were performed on the data. First, the correlation matrix $A^{\mu\nu}$ (explained in main text and below) was calculated for the original continuous data and for the binarized data (Fig. S1). Both correlation matrices are consistent with each other showing binarization does not change the global correlations. Note that in the correlation matrix, cell fates have been grouped by tissue type, leading to a block diagonal form. Second, the TF expression is well fit by a log-normal distribution (Fig. S3), and the mean of this distribution (5.35) is consistent with the threshold (5.5). Third, the expression state of all cell fates was constructed from multiple microarray experiments. These different experiments were compared with each other and were within 2 standard deviations for all cell fates (see Fig. S2 for definition of std). This demonstrates that microarrays from multiple laboratories can be directly compared.

A complete list of microarrays used in this study is included as an Excel file. In addition, the processed data (continuous and binary) are also included as an Excel file.

II. OVERVIEW OF MODEL

A. Epigenetic landscapes from the projection method

Here we expand upon the explanation of the quantitative model from the main text. The state of a cell is represented as a vector S_i of length $N = 1152$, with each dimension of the vector (labeled by Latin indices i, j) corresponding to the binary TF expression, either $+1$ if TF i is active and -1 if it is inactive. There are $p = 95$ cell fates (labeled by Greek indices μ, ν) and cell fate μ is a N by 1 vector characterized by its binary expression state, ξ_i^μ. The TF expression for each cell fate, ξ_i^μ, is given by our microarray dataset. This is the only input of biological data in our landscape.

In general, cell fates are highly correlated (Fig. S1), and it is useful to define a correlation matrix $A^{\mu\nu} = 1/N \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu$ which characterizes the correlation between cell fate μ and ν. The high correlation between cell fates implies that the overlap (vector dot-product), $m^\mu = 1/N \sum_{i=1}^{N} \xi_i^\mu S_i$, is a poor measure of how close a cell with expression state S_i is to a cell fate μ. A more accurate characterization is provided by the "projection", $a^\mu = \sum_{\nu=1}^{p} (A^{-1})^{\mu\nu} m^\nu$, of S_i on a cell fate μ. The projection measures the orthogonal projection of a state S_i onto the subspace spanned by naturally occurring cell fates, $\{\xi_i^\mu\}$ (see Fig. 2D). The "projection-contribution" is mathematically represented as $\eta_i^\mu = 1/N \sum_{\nu=1}^{p} (A^{-1})^{\mu\nu} \xi_i^\nu$ for TF i in cell fate μ. This can be interpreted as the energy contribution of TF i to cell fate μ. This is also a measure of the "predictivity" of a TF for a given cell fate.

Using the state space, we can explicitly construct the landscape, H, as:

$$H = -\frac{1}{2} \sum_{i,j} S_i J_{ij} S_j - \frac{1}{2} \sum_i X_i S_i - \frac{1}{2} \sum_{i,j} S_i K_{ij}(\vec{c}) S_j \tag{1}$$

with the dynamics set by random, asynchronous update according to the probability

$$P[S_i(t+1)] = \frac{e^{\beta h_i(t) S_i(t)}}{e^{\beta h_i(t)} + e^{-\beta h_i(t)}} \tag{2}$$

with the local field on each TF given by $h_i = -\frac{\partial H}{\partial S_i}$.

The $H_{basin} = -1/2 \sum_{i,j} S_i J_{ij} S_j$ produces stable basins of attraction (Fig. 1A). This is done by inferring a correlation-based, TF interaction matrix[4]

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \sum_{\nu=1}^{p} \xi_i^\mu (A^{-1})^{\mu\nu} \xi_j^\nu \tag{3}$$

with effective interactions that may be indirect.

This stabilizing term of the landscape can be rewritten in terms of the overlap and projection as follows:

$$H_{basin} = -\frac{1}{2} \sum_{i,j} S_i J_{ij} S_j = -\frac{1}{2N} \sum_{i,j} \sum_{\mu\nu} S_i \xi_i^\mu (A^{-1})^{\mu\nu} \xi_j^\nu S_j = -\frac{N}{2} \sum_\mu m^\mu a^\mu \tag{4}$$

A simple geometric picture illustrates that $H_{basin} = -1/2 \sum_{i,j} S_i J_{ij} S_j$ makes each cell type a global minimum of the landscape. An arbitrary vector can be rewritten in terms of its projection in the cell fate subspace and its orthogonal component δS_i[4],

$$S_i = \sum_\mu a^\mu \xi_i^\mu + \delta S_i \tag{5}$$

Then, the distance of an arbitrary vector S to the cell fate subspace is given by Δ,

$$\Delta = (\sum_i (\delta S_i)^2)^{1/2} \tag{6}$$

which can be rewritten as

$$\frac{\Delta^2}{N} = 1 - \sum_{\mu} a^{\mu} m^{\mu} \tag{7}$$

This allows us to rewrite the stabilizing term of the landscape as

$$H_{basin} = -\frac{N}{2} + \frac{1}{2}\Delta^2 \tag{8}$$

This provides a very clear interpretation of the landscape as the global distance of an arbitrary vector S to the natural cell fate subspace[4].

To demonstrate cell fate stability, an ESC state for the fully connected interaction matrix was randomly perturbed (Fig. 2E, Fully Connected) and evolved in time. For a variety of initial conditions, it relaxed back to the ESC, demonstrating a large basin of attraction (red bracket). While our original construction is a true potential or Lyapunov function, this condition can be relaxed. Random dilution ($J_{ij} = 0$) creates a pseudo-potential or non-Lyapunov function, which has a smaller, noisier basin of attraction, see (Fig. 2E, Diluted) where 20% of J_{ij} were removed. Note, we show the trajectories of the diluted interaction matrix to demonstrate the robustness of the model, but further experimental data is needed to introduce biologically realistic (i.e. not random) dilution.

The $H_{culture} = -1/2 \sum_i X_i S_i$ stabilizes specific cell fates (Fig. 1B). This can be explicitly shown as:

$$H_{culture} = -\frac{1}{2}\sum_i X_i S_i = -\frac{N}{2}\sum_i \chi^{\mu}\eta_i^{\mu} S_i = -\frac{N}{2}\sum_i \chi^{\mu} a^{\mu} \tag{9}$$

where cell fate μ is favored χ^{μ} which characterizes the degree of environmental stabilization[5].

The $H_{switch} = -1/2 \sum_{i,j} S_i K_{ij}(\vec{c}) S_j$ represents switching due to external signals $\vec{c}$ (Fig. 1C). The original interpretation as an interaction in TF space can be written in terms of cell fates as

$$H_{switch} = -\frac{1}{2}\sum_{i,j} S_i K_{ij}(\vec{c}) S_j = -\frac{1}{2N}\sum_{i,j}\sum_{\mu,\nu} S_i \xi_i^{\mu} G^{\mu\nu}(\vec{c}) \eta_j^{\nu} S_j = -\frac{N}{2}\sum_{\mu,\nu} m^{\mu} G^{\mu\nu}(\vec{c}) a^{\nu} \tag{10}$$

where $G^{\mu\nu}$ characterizes the switching from the switching from cell fate ν to cell fate μ and in general is a complicated function of external signals $\vec{c}$. For now the $\vec{c}$ is included to show where external signals couple to switching, but more work is needed to determine the explicit interactions between the environment and cell fate switching.

We demonstrate this term with a biological example. The common myeloid progenitor (CMP) is a blood cell fate that *in vivo* can differentiate into either granulo-monocytic progenitors (GMP) or megakaryocyte-erythroid progenitors (MEP). We demonstrate (Fig. 2F) that signal 1 (red),

$$G^{GMP,CMP} = 1 \tag{11}$$

(with rest of $G^{\mu\nu} = 0$, i.e. signal 2 absent) can convert CMP to GMP, while a separate application of signal 2 (blue)

$$G^{MEP,CMP} = 1 \tag{12}$$

(with rest of $G^{\mu\nu} = 0$, i.e. signal 1 absent) can differentiate MEP to GMP. These switching terms are explicitly non-Lyapunov because they break symmetry; for example, signal 1 switches CMP to GMP, but not GMP to CMP.

The standard dynamics is random, asynchronous update[5]. The interaction network biases a TF to its state by its local field $h_i = -\frac{\partial H}{\partial S_i}$. The dynamics are stochastic and controlled by a global noise parameter β, which is the inverse temperature, $\beta = 1/T$. At each time step, one TF is probabilistically updated based only on its local field, h_i, and the global noise, β. The update time has no physical meaning, while the biologically relevant time scale is cell division[6]. Since cellular division produces HM errors[7], a physical time scale is introduced by periodically flipping a fixed percentage of TF states. The system is still stable under this additional noise, as shown in Fig. 2E which incorporates both asynchronous dynamics with $\beta = 1/0.45 = 2.2$ and 2% errors every 5000 time steps. The noise due to the temperature β is equivalent to partial annealing or smooth fluctuations about the landscape while the bursts of noise due to cell division is equivalent to quantum annealing or jumps through phase space which can tunnel through barriers in the landscape.

B. Model in matrix notation

For completeness, we again present the model above but now in explicit matrix notation. S is a N by 1 vector and ξ is a p by N matrix that is based on biological data. The correlation matrix is p by p

$$A = \frac{1}{N}\xi\xi^T \tag{13}$$

The overlap and projection are both p by 1 vectors

$$m = \frac{1}{N}\xi S \tag{14}$$

$$a = A^{-1}m = \frac{1}{N}A^{-1}\xi S = (\xi\xi^T)^{-1}\xi S \tag{15}$$

and the projection-contribution is a p by N matrix

$$\eta = \frac{1}{N}A^{-1}\xi = (\xi\xi^T)^{-1}\xi \tag{16}$$

The TF interaction matrix that creates stable basins of attraction a N by N matrix

$$J = \frac{1}{N}\xi^T A^{-1}\xi = \xi^T(\xi\xi^T)^{-1}\xi \tag{17}$$

One can verify that this is a projection matrix since $J^2 = J$. The culture stabilization term in TF space is a N by 1 vector given by X. In cell fate space it is represented as $X = N\eta^T\chi^T$.

The TF interaction matrix that induces switches between cell fates is a N by N matrix

$$K = \frac{1}{N}\xi^T G\eta = \xi^T G(\xi\xi^T)^{-1}\xi \tag{18}$$

where the cell fate switching matrix G is a p by p matrix.

The landscape can be written in TF space as

$$H = -\frac{1}{2}S^T JS - \frac{1}{2}X^T S - \frac{1}{2}S^T KS \tag{19}$$

Or the landscape can be given in terms of cell fates as

$$H = -\frac{N}{2}m^T a - \frac{N}{2}\chi^T a - \frac{N}{2}m^T Ga \tag{20}$$

C. Partially reprogrammed cells as spurious attractors

One of the most generic properties of all attractor neural network constructions is that in addition to the desired attractors, ξ_i^μ, the non-linearity of the dynamical process and the high dimensionality of the underlying space induces additional attractors, which are termed spurious attractors[5]. In general, these spurious attractors have higher energy than the stored patterns, ξ_i^μ, and hence smaller basins of attractions. For the traditional Hopfield model, these spurious attractors take the form of odd-hybrids (i.e. hybrids of $3, 5, 7, \ldots$ of the ξ_i^μ)[5]. However, for the projection method any hybrid of ξ_i^μ is a spurious attractor of the system[4]. This can be easily understood by noting that in the projection method, the entire subspace spanned by the ξ_i^μ are spurious attractors (see geometric picture above).

As discussed in the main text, the prediction of spurious attractors in the projection method inspired us to reexamine data on existing partially reprogrammed cells. Surprisingly, we found that partially reprogrammed cells could be thought of as hybrids of existing cell fates. While in this paper we use discrete states, Hopfield[8] has shown that the landscape construction can be generalized to continuous, sigmoidal states. This means that a continuous Hopfield model effectively interpolates continuously between a set of discrete states. The agreement between theory and experiment provides hints that the fundamental units of cellular identity are discrete states that are read out in a continuous manner, suggesting the tantalizing possibility that the discreteness of histone modifications (HM) is an important component of the networks that underly cellular identity.

D. Predicting candidate TFs for reprogramming

Our candidate TFs for overexpression should have high projection-contribution ($\eta > 0$) and high expression ($zscore > 0.5$). Note that conversely, negative projection-contribution ($\eta < 0$) and low TF expression ($zscore < -0.5$) indicates candidates for knockouts in reprogramming. TF within 0.5 std of the binary threshold are excluded since their projection-contribution can change significantly with changes in the threshold (data not shown). Within TFs that satisfy the above criteria, we can rank reprogramming candidates by a single number by multiplying projection-contribution by z-score TF expression.

See the SI file TF Reprogramming Candidates for our top 100 candidates for overexpression and our top 100 candidates for knockout for a variety of cell fates. We only made predictions for cell fates that are homogeneous. For example, most of our neural data is based on dissection of the brain, and hence is a heterogenous mixture of various neurons and other cell fates. To make predictions for various types of neurons, homogenous microarrays of each neuron type are needed.

In the SI file Processed Data, all data used for the predictions can be found. Another type of prediction can be made from the data. For example, we have microarrays for multiple B cell precursors, pre B cells, pre pro B cells, and pro B cells. However, if one is interested in reprogramming to any of the B cell precursors, since our predictions are based on linear algebra, the ranking for all three precursors can be averaged to produce rankings for a general B cell precursor.

E. Limitations of model

One limitation of our model is that it misses the importance of the oncogene Myc to reprogramming. Many protocols use Myc[9], but it can be replaced (with no deleterious effect) by short hairpin RNAs (shRNAs)[10], or dropped completely from protocols at the expense of speed and less efficient reprogramming[11]. Thus, Myc appears to increase the rate (not outcome) of reprogramming while our current framework has limited information about dynamics.

Another limitation is the use of Affymetrix GeneChip Mouse Genome 430 2.0 platform. This microarray is useful since there exists much public data on a variety of cell fates. However, upon closer examination of one cell fate, lung, we discovered the inaccuracy of this microarray. Nkx2-1 is known to be a marker of early lung development[12], yet the Affymetrix GeneChip Mouse Genome 430 2.0 shows no differential expression of Nkx2-1. Therefore, our current data may miss key TFs due to a poorly matched probe. For future research, more accurate predictions can be made using more modern microarrays such as Affymetrix Mouse Gene 1.0 ST or RNA-Seq data.

---

[1] Zhang, H.-M, et al. (2012) *Nucleic Acids Research* **40**, D144–D149.
[2] Mikkelsen, T. S, et al. (2007) *Nature* **448**, 553–560.
[3] Meissner, A, et al. (2008) *Nature* **454**, 766–770.
[4] Kanter, I & Sompolinsky, H. (1987) *Phys Rev A* **35**, 380–392.
[5] Amit, D. (1992) *Modeling brain function: The world of attractor neural networks.* (Cambridge Univ Pr).
[6] Hanna, J, et al. (2009) *Nature* **462**, 595–601.
[7] Ben-David, U, Mayshar, Y, & Benvenisty, N. (2011) *Cell Stem Cell* **9**, 97–102.
[8] Hopfield, J. J. (1984) *Proc Natl Acad Sci U S A* **81**, 3088–3092.
[9] González, F, Boué, S, & Belmonte, J. C. I. (2011) *Nat Rev Genet* **12**, 231–242.
[10] Onder, T. T, et al. (2012) *Nature* **483**, 598–602.
[11] Wernig, M, Meissner, A, Cassady, J. P, & Jaenisch, R. (2008) *Cell Stem Cell* **2**, 10–12.
[12] Longmire, T. A, et al. (2012) *Cell Stem Cell* **10**, 398–411.
[13] Young, R. A. (2011) *Cell* **144**, 940–954.
[14] Yu, J, et al. (2007) *Science* **318**, 1917–1920.
[15] Takahashi, K & Yamanaka, S. (2006) *Cell* **126**, 663–676.

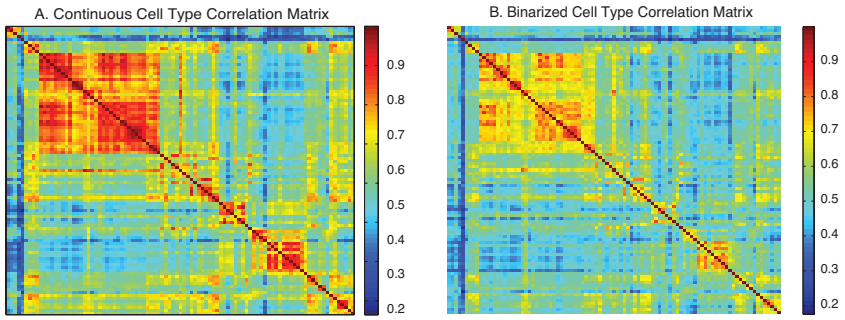

FIG. 1: **Cell fate correlation matrices.** (A) Correlation matrix between cell fates for continuous data. (B) Correlation matrix for binarized data.

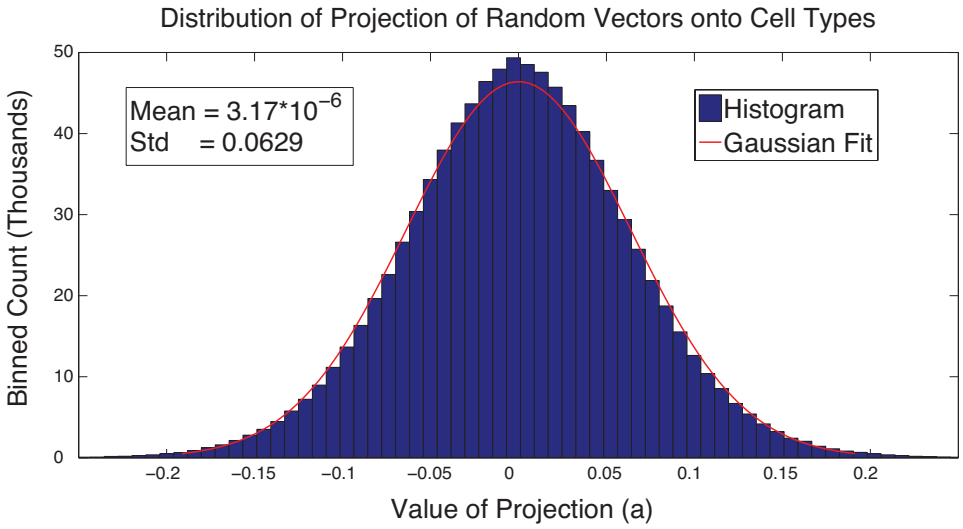

FIG. 2: **Projection of a random vector on a given cell fate.** Ten thousand binarized random vectors were created in MATLAB and projected onto the cellular sub-space. The histogram shows the distribution of the projections. The red line is a Gaussian fit to the histogram. The mean is practically zero while the standard deviation is 0.0629.

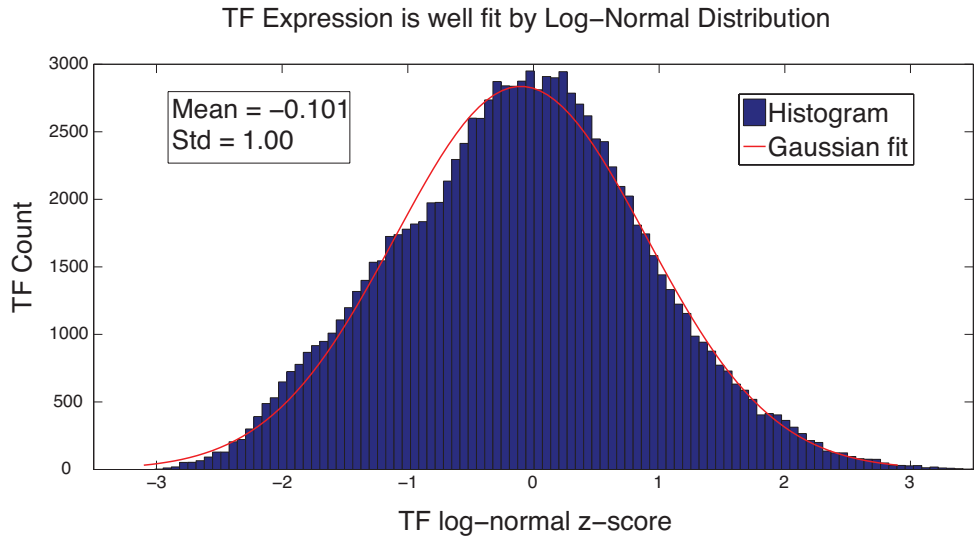

FIG. 3: **TF Expression is Log-Normal.** The figure shows the log-normal z-score with zero corresponding to the the previously determined binary cutoff point (5.5 in raw expression value) instead of the log-normal mean (5.35 in raw expression value).

TABLE I: Classifying ESC Reprogramming Candidates. Justification for transcription factor (TF) classification in Figure 3A and 3B. Table has top 100 embryonic stem cell (ESC) reprogramming candidates (as ranked by z-score times predictivity, η_i^μ). Classification of each TF is either justified by paper citation or GO Process term. Reprogramming TFs are in a pre-existing reprogramming protocol, pluripotency TFs help maintain the ESC state, differentiation TFs are expressed in ESC but help induce cell fate change *in vivo*, and unknown TFs do not yet have a known function and hence are intriguing reprogramming candidates. Note that reprogramming TFs Myc, Chd1, and Ezh2 not in our top 100 candidates.

TF	Z-Score	η_i^μ (10^{-3})	Z$*\eta_i^\mu$ (10^{-3})	Classification	Citation or GO Term
Pou5f1	3.31	2.32	7.68	Reprogramming	[9]
Zfp819	1.38	4.96	6.86	Unknown	biological process
Gbx2	1.33	5.01	6.68	Differentiation	midbrain-hindbrain boundary development
Nanog	2.90	1.94	5.62	Reprogramming	[9]
Prdm14	0.97	5.01	4.84	Pluripotency	[13]
Olig1	1.06	4.43	4.69	Differentiation	neuron fate commitment
Foxd3	0.74	6.10	4.54	Pluripotency	[14]
Zic3	2.32	1.95	4.52	Differentiation	cell differentiation
Msc	0.98	4.43	4.33	Differentiation	branchiomeric skeletal muscle development
Nr0b1	2.44	1.71	4.18	Pluripotency	negative regulation of cell differentiation
Gli1	1.21	3.40	4.11	Differentiation	lung development
Tcfl5	1.96	2.07	4.07	Unknown	regulation of transcription, DNA-dependent
Zfp936	1.99	1.94	3.86	Unknown	biological process
Zfp105	1.66	2.30	3.82	Unknown	regulation of transcription, DNA-dependent
Neurod1	0.82	4.45	3.64	Differentiation	positive regulation of cell differentiation
Rex2	1.20	2.99	3.58	Unknown	biological process
Pir	1.72	2.08	3.57	Differentiation	monocyte differentiation
Stat4	0.88	4.06	3.56	Signaling	signal transduction
Bcl3	1.54	2.30	3.55	Differentiation	marginal zone B cell differentiation
Utf1	1.26	2.70	3.40	Reprogramming	[9]
Esrrb	1.71	1.97	3.38	Reprogramming	[9]
Gm13152	2.50	1.34	3.35	Unknown	biological process
Otx2	1.17	2.71	3.17	Differentiation	cell fate specification
Zfp42	2.77	1.14	3.16	Pluripotency	[15]
Trip6	1.79	1.60	2.87	Other	regulation of transcription, DNA-dependent; cell adhesion

TF	Z Score	η (10^{-3})	Z$*\eta$ (10^{-3})	Classification	Citation or GO Term
Sycp3	1.35	2.01	2.72	Differentiation	spermatogenesis
Zfp568	0.85	3.13	2.66	Differentiation	convergent extension involved in neural plate elongation
Trps1	0.58	4.53	2.62	Unknown	negative regulation of transcription, DNA-dependent
Sall1	1.51	1.70	2.57	Differentiation	neural tube development
2610305D13Rik	2.09	1.22	2.56	Unknown	biological process
Tfap2c	1.51	1.70	2.56	Differentiation	cell differentiation
Zfp423	1.25	2.04	2.54	Differentiation	cell differentiation
Gli2	1.78	1.37	2.45	Differentiation	cell differentiation
Gm13212	0.95	2.56	2.42	Unknown	biological process
Htatip2	0.53	4.55	2.41	Differentiation	cell differentiation
Arid5b	0.93	2.55	2.36	Differentiation	adipose tissue development
Klf4	2.47	0.95	2.35	Reprogramming	[9]
Dnmt3b	1.97	1.19	2.33	Epigenetics	DNA methylation
Aire	1.30	1.78	2.32	Immune	humoral immune response
Zfp57	1.56	1.46	2.28	Epigenetics	DNA methylation involved in embryo development
Pbrm1	2.00	1.04	2.09	Differentiation	heart development
Tgif1	2.56	0.81	2.07	Differentiation	positive regulation of neuron differentiation
Sox15	1.23	1.68	2.06	Pluripotency	[15]
Zic5	1.09	1.89	2.05	Differentiation	cell differentiation
Zfp473	1.72	1.18	2.03	Unknown	regulation of transcription, DNA-dependent
5730507C01Rik	1.44	1.37	1.96	Unknown	biological process
Foxp1	1.04	1.87	1.95	Pluripotency	embryo development
Zfp229	1.57	1.22	1.92	Unknown	biological process
Six4	1.00	1.90	1.91	Differentiation	thymus development
Dmrt1	1.16	1.64	1.90	Differentiation	Sertoli cell differentiation

TF	Z Score	η (10^{-3})	Z$*\eta$ (10^{-3})	Classification	Citation or GO Term
Zscan10	2.23	0.84	1.88	Pluripotency	stem cell differentiation
Hhex	1.06	1.68	1.78	Differentiation	B cell differentiation
Cebpb	0.78	2.24	1.74	Differentiation	neuron differentiation
Eomes	0.57	3.01	1.72	Differentiation	cell differentiation
Foxh1	1.15	1.50	1.72	Differentiation	axial mesoderm development
Zbtb8a	1.97	0.87	1.71	Unknown	regulation of transcription, DNA-dependent
Hsf2bp	1.06	1.59	1.69	Unknown	biological process
Sall4	2.41	0.64	1.55	Reprogramming [9]	
Zfp955b	1.49	1.04	1.55	Unknown	biological process
Rara	0.94	1.64	1.54	Differentiation	chondroblast differentiation
Zfp934	0.97	1.59	1.54	Unknown	biological process
Rest	1.58	0.97	1.54	Differentiation	cardiac muscle cell myoblast differentiation
Tcf15	1.81	0.84	1.52	Differentiation	muscle organ development
Zfp647	1.42	1.06	1.50	Unknown	regulation of transcription, DNA-dependent
Id3	1.85	0.81	1.50	Differentiation	epithelial cell differentiation
Zfp217	1.81	0.81	1.46	Unknown	regulation of transcription, DNA-dependent
Pitx2	0.90	1.60	1.45	Differentiation	cardiac muscle cell differentiation
Egr1	1.90	0.74	1.40	Differentiation	T cell differentiation
Id4	0.93	1.51	1.40	Differentiation	cerebral cortex neuron differentiation
Smarca1	0.73	1.91	1.39	Differentiation	neuron differentiation
Zfp799	0.84	1.66	1.39	Unknown	biological process
Tead2	1.94	0.72	1.39	Differentiation	notochord development
Usf1	0.80	1.72	1.38	DNA Repair	response to UV
Zfp760	1.25	1.10	1.37	Unknown	biological process
Plagl1	0.78	1.75	1.36	Unknown	regulation of gene expression

TF	Z Score	η (10^{-3})	Z$*\eta$ (10^{-3})	Classification	Citation or GO Term
Sox2	3.01	0.44	1.33	Reprogramming	[9]
Hmgb2	2.40	0.54	1.30	Differentiation	positive regulation of erythrocyte differentiation
Rbpms	1.79	0.72	1.29	Unknown	regulation of transcription, DNA-dependent
Rhox6	0.94	1.38	1.29	Unknown	biological process
Grhl1	0.75	1.71	1.28	Unknown	regulation of transcription, DNA-dependent
Prdm5	0.83	1.51	1.25	Epigenetic	chromatin modification
Gm5595	1.05	1.17	1.23	Unknown	biological process
Relb	0.83	1.48	1.23	Differentiation	T-helper 1 cell differentiation
Zbtb12	1.02	1.20	1.22	Unknown	biological process
Sp5	0.60	2.03	1.22	Differentiation	bone morphogenesis
Lin28a	1.99	0.60	1.19	Reprogramming	[9]
Zfp553	1.18	0.99	1.17	Unknown	regulation of transcription, DNA-dependent
E2f1	0.88	1.32	1.17	Differentiation	forebrain development
Peg3	0.76	1.49	1.14	Apoptosis	apoptotic process
Zfp449	0.62	1.82	1.13	Unknown	regulation of transcription, DNA-dependent
Nr5a2	0.73	1.49	1.09	Reprogramming	[9]
Grhl3	0.70	1.56	1.08	Differentiation	ectoderm development
Nfya	0.68	1.59	1.08	Unknown	positive regulation of transcription DNA-dependent
Sox11	0.78	1.37	1.07	Differentiation	embryonic skeletal system morphogenesis
Cenpi	1.52	0.71	1.07	Unknown	biological process
Smad1	1.29	0.83	1.07	Pluripotency	[13]
Etv4	0.82	1.30	1.06	Differentiation	stem cell differentiation
Cenpn	2.18	0.48	1.04	Unknown	biological process
Zfp296	1.98	0.52	1.03	Unknown	biological process
Maff	0.92	1.12	1.03	Differentiation	regulation of epidermal cell differentiation

TABLE II: Comparison of rankings. Here we compare our ranking based on mulitplying z-score by projection-contribution versus the z-score only ranking. Note that in liver, Gata4 is within 0.5 std of the binary cutoff and therefore its projection-contribution is unreliable.

TF	Protocol	Our Rank	Z-score Rank
Gata4	Heart	6	68
Tbx5	Heart	13	129
Mef2c	Heart	50	38
Foxa3	Liver	2	13
Foxa2	Liver	9	36
Hnf4a	Liver	10	6
Foxa1	Liver	19	48
Gata4	Liver	No rank	415