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-GEO (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 Chromsomes 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.* cispb_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 fibroblasts 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

$$\underset{\alpha}{\text{minimize}} \quad \|\mathbf{g}_{FIB,UM} - \alpha \mathbf{g}_{FIB,GTEX}\| \qquad [1]$$

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 α 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 - Additional GTEx Data. For fibroblast to Adipose-Subcutaneous, the highest scoring factor is EBF1, a known maintainer of brown adipocyte identity, and a known promoter of adipogenesis in fibroblasts (8). The 2nd highest scoring marker, PPARG, has been shown to be involved in adipose differentiation, and can be used individually to achieve reprogramming from fibroblasts (9). Curiously, ATF3 is implicated here as being useful for adipocyte differentiation although its function has been shown to repress PPARG and stymie cell proliferation (10). Upon further research, using time dependent addition, ATF3 addition scores best when added towards the end of reprogramming process.

Two Brain tissue samples, cerebellum and hippocampus, predict TFs necessary for natural differentiation. Interestingly, our algorithm selects different TFs for each conversion, with factors linked specifically to each tissue. For cerebellum, NEUROD1, has been shown to be required for granule cell differentiation, while ZIC1 and ZIC4 are both known to promote cellebular-specific neuronal function (11, 12). The top scoring combination of 3 TFs are all similarly known to be important in neurogenesis (NEUROD1, ZBTB18, UNCX) (13, 14). Hippocampus TF scoring includes FOXG1 as the top predicted factor, a factor specifically needed in hippocampus development. OLIG2, FOXG1, and GPD1 are the top scoring set for hippocampus reprogramming, all of which have been shown to been necessary for hippocampus function.

Colon TF scoring finds known differentiation factor in natural colon secretory linage development, ATOH1, as the highest scoring individual factor (15). The top scoring combination of 2 TFs includes ATOH1 along with CDX2, another known factor necessary for full differentiation of colon cells, specifically small intestine maturation (16). Liver cell reprogramming similarly finds known factors for differentiation in the top score of all 3 combinations: HNF4A, CUX2, PROX1 (17–19). All TFs play a role in correct development of hepatic progenitor cell-types and hepatic stem cells, the cell types just above in lineage differentiation.

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. SPI1 (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

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

can be written as

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

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:

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

We can see that an exact solution exists if

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

and is given by

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

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

$$\bar{u} = C^{\dagger} \left(x_T - A_3 A_2 A_1 x_1\right), \quad [8]$$

where $C^{\dagger}$ 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

$$d = \left\| (I_N - CC^{\dagger})(x_T - A_3 A_2 A_1 x_1) \right\| \quad [9]$$

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

μ 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 μ, 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 μ_2 and define it as:

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

μ_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 μ score, μ_2 can be computed to decide the better candidate (i.e. lower μ_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.

1. Chen H, et al. (2015) Functional organization of the human 4d nucleome. *Proceedings of the National Academy of Sciences* 112(26):8002–8007.
2. Consortium EP, , et al. (2012) An integrated encyclopedia of dna elements in the human genome. *Nature* 489(7414):57–74.
3. Lonsdale J, et al. (2013) The genotype-tissue expression (gtex) project. *Nature genetics* 45(6):580–585.
4. Adams D, et al. (2012) Blueprint to decode the epigenetic signature written in blood. *Nature biotechnology* 30(3):224–226.
5. Chen J, Hero A, Rajapakse I (accepted 2016) Spectral identification of topological domains. *Bioinformatics*.
6. Neph S, et al. (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150:1274–1286.
7. Shannon P (2014) Motifdb: An annotated collection of protein-dna binding sequence motifs. *R package version* 1(0).
8. Jimenez MA, Åkerblad P, Sigvardsson M, Rosen ED (2007) Critical role for ebf1 and ebf2 in the adipogenic transcriptional cascade. *Molecular and cellular biology* 27(2):743–757.
9. Gregoire FM, Smas CM, Sul HS (1998) Understanding adipocyte differentiation. *Physiological reviews* 78(3):783–809.
10. Jang MK, Kim CH, Seong JK, Jung MH (2012) Atf3 inhibits adipocyte differentiation of 3t3-l1 cells. *Biochemical and biophysical research communications* 421(1):38–43.
11. Miyata T, Maeda T, Lee JE (1999) Neurod is required for differentiation of the granule cells in the cerebellum and hippocampus. *Genes & development* 13(13):1647–1652.
12. Frank CL, et al. (2015) Regulation of chromatin accessibility and zic binding at enhancers in the developing cerebellum. *Nature neuroscience* 18(5):647–656.
13. Cohen J, et al. (2016) Further evidence that de novo missense and truncating variants in zbtb18 cause intellectual disability with variable features. *Clinical Genetics*.
14. Sammeta N, Hardin DL, McClintock TS (2010) Uncx regulates proliferation of neural progenitor cells and neuronal survival in the olfactory epithelium. *Molecular and Cellular Neuroscience* 45(4):398–407.
15. VanDussen KL, Samuelson LC (2010) Mouse atonal homolog 1 directs intestinal progenitors to secretory cell rather than absorptive cell fate. *Developmental biology* 346(2):215–223.
16. Crissey MAS, et al. (2011) Cdx2 levels modulate intestinal epithelium maturity and paneth cell development. *Gastroenterology* 140(2):517–528.
17. DeLaForest A, et al. (2011) Hnf4a is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* 138(19):4143–4153.
18. Seth A, et al. (2014) Prox1 ablation in hepatic progenitors causes defective hepatocyte specification and increases biliary cell commitment. *Development* 141(3):538–547.
19. Vanden Heuvel GB, et al. (2005) Hepatomegaly in transgenic mice expressing the homeobox gene cux-1. *Molecular carcinogenesis* 43(1):18–30.
20. Feng R, et al. (2008) Pu. 1 and c/ebpα/β convert fibroblasts into macrophage-like cells. *Proceedings of the National Academy of Sciences* 105(16):6057–6062.
21. Bruns H, et al. (2016) the ikzf1-irf4 axis regulates macrophage polarization and macrophage-mediated anti-tumor immunity.
22. Perdiguero EG, Geissmann F (2016) The development and maintenance of resident macrophages. *Nature immunology* 17(1):2–8.
23. Naudin C, et al. (2017) Pumilio/foxp1 signaling drives expansion of hematopoietic stem/progenitor and leukemia cells. *Blood* 129(18):2493–2506.
24. Batta K, Florkowska M, Kouskoff V, Lacaud G (2014) Direct reprogramming of murine fibroblasts to hematopoietic progenitor cells. *Cell reports* 9(5):1871–1884.
25. Starnes L, et al. (2010) A transcriptome-wide approach reveals the key contribution of nfi-a in promoting erythroid differentiation of human [cd34. sup.+] progenitors and cml cells. *Leukemia* 24(6):1220–1224.
26. Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82.
27. Grant C, Bailey T, Noble W (2011) Fimo: scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.
28. Michael DG, et al. (2016) Model-based transcriptome engineering promotes a fermentative transcriptional state in yeast. *Proceedings of the National Academy of Sciences* 113(47):E7428–E7437.
29. Whitfield M, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* 13.6:1977–2000.

SI Figures.

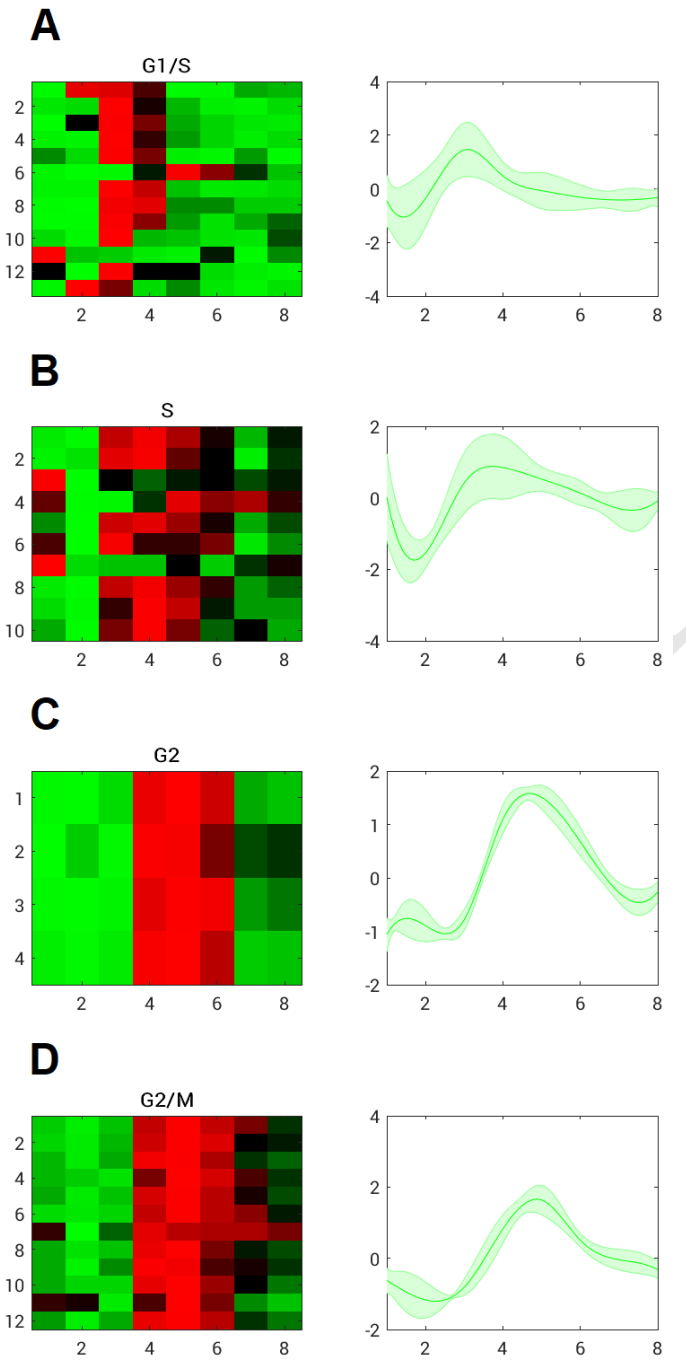


Fig. S1. Analysis of cell cycle marker gene expression. Gene expression RNA-seq data for 39 genes that have been shown in the literature to be cell cycle regulated (29). Cell cycle phases shown include (*A*) G_1/S, (*B*) S, (*C*) G_2, (*D*) G_2/M. Raw data of gene expression over time (left), with smoothed/interpolated 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, . . . , 56 h after growth medium introduction. y-axis is normalized expression.

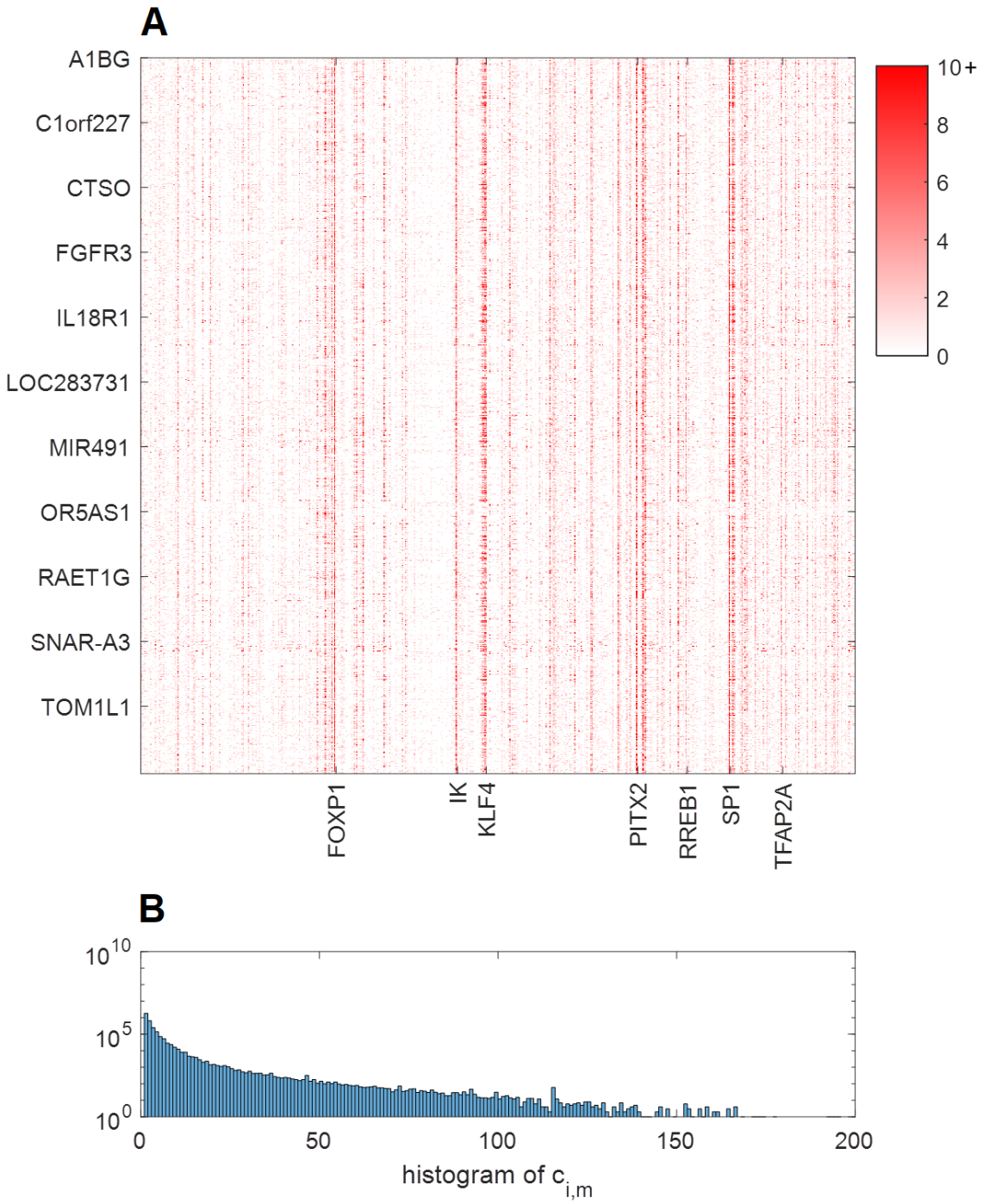
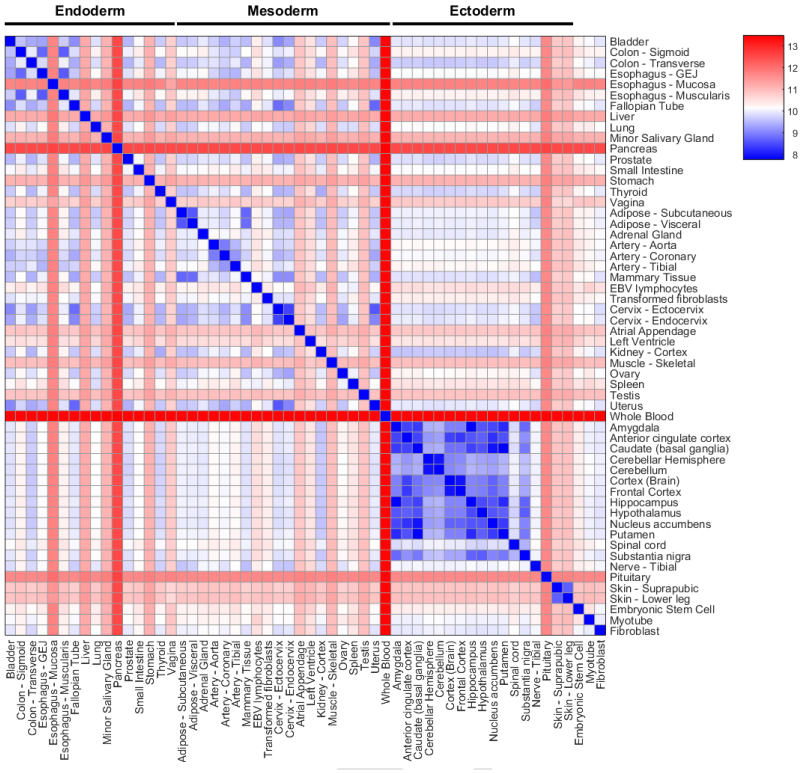


Fig. S2. Visualization of input matrix $\mathbf{B}$. (*A*) Visualization of the 22,083×547 $c_{i,m}$ matrix: identified TF-to-gene interactions based on TFBSs. The color at entry (i, m) represents how many TF m TFBSs were observed within $\pm$5kb of gene i's TSS. The color axis has been truncated to $[0, 10]$ 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.

A

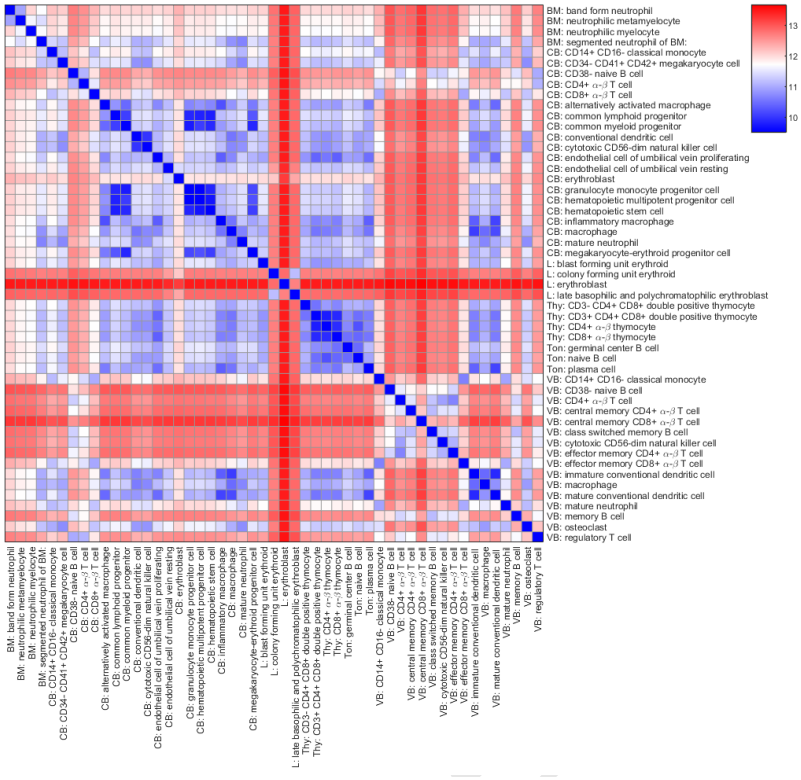


B

	ESC	μ	Myotube	μ	Adipose - Subcutaneous	μ	Brain - Cerebellum	μ	Brain - Hippocampus	μ	Colon - Transverse	μ	Heart - Left Ventricle	μ	Liver	μ
Combination of 1	MYCN	2634	MYOG	1753	EBF1	2839	NEUROD1	1772	FOXG1	372	ATOH1	1814	HEY2	78	HNF4A	81
	ZFP42	2453	MYOD1	1065	PPARG	2666	ZIC1	1725	KCNIP1	121	HNF4A	1693	IRX4	57	GATA4	62
	NANOG	2013	PKNOX2	797	TAL1	2182	EGR4	1622	NR2E1	103	ZBTB7C	1589	HIST2H2BE	44	ATF5	62
	OTX2	1628	SIX2	745	GPD1	1939	ZIC4	1517	HEY1	62	MYB	1483	HAND1	17	GPD1	47
	SOX2	1025	PITX2	744	FOS	1683	GPD1	1123	GPD1	61	SPIB	1381	SOX7	1	FOXA2	36
	FOXD3	0	MEF2C	675	ERG	1582	PAX6	977	SCRT1	41	HNF4G	1314	GATA4	0	NR1I2	34
	NR6A1	0	MYF6	136	SOX17	1165	PKNOX2	956	KLF15	27	NR1I2	1313	MEOX1	0	HNF1A	29
	POU5F1	0	PKNOX2	0	MEF2C	1018	ETV1	788	BHLHE22	0	EHF	1279	SOX18	0	GLYCTK	27
	ZIC2	0	SIX2	0	MAFB	945	HSF4	683	HLF	0	FOS	814	KLF15	0	MAFB	23
					MEOX1	733	NFIA	636	OLIG1	0	HOXB9	786	HIST2H2BE	0	FOXA1	15
Combination of 2	MYCN,ZFP42	2746	MYOG,PKNOX2	1804	EBF1,ATF3	2910	NEUROD1,ZBTB18	2230	FOXG1,GPD1	428	ATOH1,CDX2	1856	GATA4,HEY2	141	HNF4A,PROX1	95
	MYCN,NANOG	2704	MYOG,SIX2	1766	EBF1,FOS	2883	NEUROD1,UNCX	2083	OLIG2,FOXG1	389	ATOH1,ZBTB7C	1840	SOX7,IRX4	120	HNF4A,CUX2	93
	MYCN,OTX2	2672	MEF2C,MYOG	1753	EBF1,SOX7	2871	NEUROD1,EN2	2069	FOXG1,NR2E1	385	ATOH1,IRF8	1836	GATA4,IRX4	117	HNF4A,GPD1	82
	MYCN,POU5F1	2634	MYF6,MYOG	1753	EBF1,GPD1	2857	NEUROD1,SOX15	1983	FOXG1,KCNIP1	379	ATOH1,HOXB3	1832	HEY2,SOX7	114	HNF4A,MAFB	81
	MYCN,SOX2	2634	MYOD1,MYOG	1753	EBF1,IRF8	2854	NEUROD1,VSX1	1943	HLF,FOXG1	374	ATOH1,ISX	1830	HEY2,IRX4	93	HNF4A,GLYCTK	81
	MYCN,FOXD3	2634	MYOG,PITX2	1753	EBF1,PPARG	2852	ZIC1,UNCX	1927	SOX2,FOXG1	373	ATOH1,NKX3-2	1828	GATA4,HIST2H2BE	93	ATF5,HNF4A	81
	MYCN,NR6A1	2634	MYOG,PKNOX2	1753	EBF1,BHLHE40	2849	NEUROD1,KCNIP1	1904	OLIG1,FOXG1	372	ATOH1,SPIB	1826	HEY2,HIST2H2BE	84	ELF3,HNF4A	81
	MYCN,ZIC2	2634	MYOG,SIX2	1753	EBF1,ERG	2847	ZIC1,ZBTB18	1903	POU3F3,FOXG1	372	ATOH1,HNF4A	1822	SOX7,HIST2H2BE	83	FOXA1,HNF4A	81
	ZFP42,POU5F1	2523	MYOD1,SIX2	1069	EBF1,HEY1	2844	NEUROD1,PKNOX2	1900	SOX10,FOXG1	372	ATOH1,HOXB9	1821	MEOX1,HEY2	80	FOXA2,HNF4A	81
	NANOG,ZFP42	2454	MYOD1,PKNOX2	1067	EBF1,MAFB	2844	ZIC1,EN2	1894	SOX21,FOXG1	372	ATOH1,HAND1	1821	SOX18,HEY2	78	GATA4,HNF4A	81
Combination of 3	MYCN,NANOG,POU5F1	2840	MYOG,SIX2,PKNOX2	1804	EBF1,FOS,ATF3	2947	NEUROD1,ZBTB18,UNCX	2328	OLIG2,FOXG1,GPD1	445	ATOH1,SPIB,CDX2	1947	GATA4,HEY2,IRX4	263	HNF4A,CUX2,PROX1	106
	MYCN,ZFP42,POU5F1	2801	MEF2C,MYOG,PKNOX2	1804	EBF1,PPARG,ATF3	2944	NEUROD1,ZBTB18,EN2	2304	FOXG1,GPD1,KCNIP1	430	ATOH1,SPIB,IRF8	1903	GATA4,HEY2,HIST2H2BE	218	HNF4A,PROX1,MAFB	96
	MYCN,NANOG,ZFP42	2747	MYF6,MYOG,PKNOX2	1804	EBF1,ATF3,SOX7	2931	NEUROD1,ZBTB18,KCNIP1	2279	FOXG1,GPD1,NR2E1	430	ATOH1,HNF4A,CDX2	1893	HEY2,SOX7,IRX4	190	GATA4,HNF4A,PROX1	95
	MYCN,OTX2,ZFP42	2746	MYOD1,MYOG,PKNOX2	1804	EBF1,ATF3,ERG	2925	NEUROD1,ZIC1,ZBTB18	2272	HLF,FOXG1,GPD1	428	ATOH1,CDX2,ZBTB7C	1884	GATA4,HAND1,HEY2	166	ATF5,HNF4A,PROX1	95
	MYCN,SOX2,ZFP42	2746	MYOG,PITX2,PKNOX2	1804	EBF1,ATF3,GPD1	2924	NEUROD1,ZBTB18,SOX15	2263	OLIG1,FOXG1,GPD1	428	ATOH1,SPIB,HOXB3	1873	GATA4,HIST2H2BE,IRX4	159	HNF4A,PROX1,GLYCTK	95
	MYCN,ZFP42,FOXD3	2746	MYOG,PKNOX2,PKNOX2	1804	EBF1,FOS,SOX7	2921	EGR4,NEUROD1,ZBTB18	2262	POU3F3,FOXG1,GPD1	428	ATOH1,CDX2,HOXB9	1868	HEY2,SOX7,HIST2H2BE	147	HNF4A,PROX1,GPD1	95
	MYCN,ZFP42,NR6A1	2746	MYOG,PKNOX2,SIX2	1804	EBF1,ATF3,MAFB	2919	NEUROD1,ZBTB18,VSX1	2250	SOX10,FOXG1,GPD1	428	ATOH1,IRF8,ZBTB7C	1862	GATA4,HEY2,SOX7	143	ELF3,HNF4A,PROX1	95
	MYCN,ZFP42,ZIC2	2746	MEF2C,MYOG,SIX2	1766	EBF1,ATF3,IRF8	2918	NEUROD1,ZBTB18,ZIC4	2234	SOX2,FOXG1,GPD1	428	ATOH1,SPIB,NKX3-2	1861	GATA4,SOX7,IRX4	141	FOXA1,HNF4A,PROX1	95
	MYCN,OTX2,POU5F1	2716	MYF6,MYOG,SIX2	1766	EBF1,ATF3,BHLHE40	2917	NEUROD1,ZBTB18,HSF4	2234	SOX21,FOXG1,GPD1	428	ATOH1,SPIB,ISX	1860	GATA4,MEOX1,HEY2	141	FOXA2,HNF4A,PROX1	95
	MYCN,NANOG,OTX2	2704	MYOD1,MYOG,SIX2	1766	EBF1,FOS,GPD1	2916	NEUROD1,ZBTB18,GPD1	2231	SOX8,FOXG1,GPD1	428	ATOH1,NR1I2,CDX2	1860	GATA4,SOX18,HEY2	141	HLF,HNF4A,PROX1	95

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. x_I = fibroblast.

A



B

	Cord Blood - HSC	μ	Cord Blood - HMPP	μ	Cord Blood - Macrophage	μ	Cord Blood - B cell	μ	Cord_Blood - Neutrophil	μ	Cord Blood - Erythroblast	μ	Cord Blood - Natural Killer cell	μ	Cord Blood - Megakaryocyte cell	μ
Combination of 1	FOXP1	6304	FOXP1	5550	EHF	1724	FLI1	5850	GZF1	2770	ERG	3915	MYBL1	1657	MYB	3363
	IKZF1	5704	HMBOX1	4960	IKZF1	1705	REL	5768	MYB	2693	NFIA	3850	REL	833	FLI1	2691
	SPI1	5568	IKZF1	4707	SPI1	1568	RUNX3	5694	CREB5	2618	TFEC	3775	IKZF3	794	SPI1	2618
	TAL1	5403	ERG	4427	MYB	1556	SPI1	5393	FOXO3	2303	MYB	3611	RUNX3	737	IKZF3	2523
	ERG	5343	FLI1	4352	BHLHE41	1485	PAX5	5209	ERG	2199	ATF3	3608	HMBOX1	726	PAX5	2520
	HMBOX1	5316	MYB	4291	ITGB2	1393	IKZF1	5095	REL	2198	LMO2	3493	ERG	713	IKZF1	2519
	MYB	5296	SOX4	4213	REL	1323	IKZF3	4977	SPI1	2159	NFE2	3357	FLI1	703	ITGB2	2470
	FLI1	5221	ZNF350	4148	BATF	1256	HMBOX1	4880	IKZF1	2136	IKZF1	3347	IKZF1	633	BATF	2457
	ITGB2	4918	LMO2	3900	TFEC	1206	ITGB2	4755	RUNX3	2126	TAL1	3346	BATF	614	SPIB	2380
	LMO2	4875	ITGB2	3811	LMO2	1116	TCF7	4655	BATF	1981	ITGB2	3176	IRF4	573	NFE2	2301
Combination of 2	FOXP1,HMBOX1	6554	FOXP1,HMBOX1	5877	IKZF1,LMO2	2052	FLI1,HMBOX1	6253	MYB,TAL1	3596	ERG,ZSCAN16	4139	IKZF2,MYBL1	2490	MYB,TAL1	4160
	ERG,FOXP1	6341	FOXP1,CHD2	5636	EHF,LMO2	1958	REL,HMBOX1	6148	MYB,LMO2	3346	ERG,TFEC	4056	MYBL1,LMO2	2488	MYB,LMO2	3964
	SPI1,FOXP1	6334	FOXP1,NFIA	5621	SPI1,LMO2	1787	FLI1,FOXP1	6071	CREB5,GZF1	3070	ERG,E2F2	4053	MYBL1,IKZF1	2443	MYB,IKZF1	3903
	FOXP1,CHD2	6325	FOXP1,SOX4	5604	IKZF1,TFEC	1772	HMBOX1,RUNX3	6049	MYB,CHD2	3030	ERG,NFIA	4047	GFI1,MYBL1	2174	MYB,GFI1B	3680
	FLI1,FOXP1	6324	FOXP1,TFEC	5590	IKZF1,ITGB2	1759	REL,FOXP1	5992	MYB,GZF1	2995	ERG,HMBOX1	4035	MYBL1,ITGB2	2169	MYB,ITGB2	3674
	FOXP1,ZBTB16	6321	FOXP1,ZNF350	5590	EHF,IKZF1	1756	RUNX3,CHD2	5987	MYB,ITGB2	2975	NFIA,TFEC	4030	RORA,MYBL1	2096	MYB,TFEC	3600
	FOXP1,IKZF1	6318	FOXP1,ITGB2	5580	EHF,TFEC	1756	FLI1,REL	5978	MYB,IKZF1	2936	NFIA,ZSCAN16	4026	MYBL1,RUNX3	2063	MYB,MEIS1	3566
	MYB,FOXP1	6316	FOXP1,LMO2	5568	EHF,ITGB2	1743	FLI1,RUNX3	5943	MYB,GFI1B	2923	ERG,E2F8	3993	MYBL1,ZBTB16	2036	MYB,ZBTB16	3469
	ITGB2,FOXP1	6313	FOXP1,IKZF1	5559	EHF,HMBOX1	1732	FLI1,TCF7	5927	MYB,RUNX3	2910	ERG,ITGB2	3945	BCL6B,MYBL1	1966	MYB,HMBOX1	3382
	FOXP1,TAL1	6312	FOXP1,ZBTB16	5554	BHLHE41,EHF	1728	FLI1,CHD2	5927	CREB5,MYB	2833	ATF3,ERG	3941	NFIB,MYBL1	1925	MYB,IKZF3	3375
Combination of 3	FOXP1,HMBOX1,ZBTB16	6713	FOXP1,HMBOX1,SOX4	6197	EHF,IKZF1,LMO2	2243	FLI1,HMBOX1,FOXP1	7071	MYB,GZF1,TAL1	4151	ATF3,ERG,E2F2	4504	MYBL1,ZBTB16,LMO2	2736	MYB,GFI1B,LMO2	4282
	FOXP1,HMBOX1,HOXA6	6600	FOXP1,HMBOX1,ZBTB16	6005	BHLHE41,IKZF1,LMO2	2142	REL,HMBOX1,FOXP1	6882	CREB5,MYB,TAL1	4051	ERG,NFIA,ZSCAN16	4457	IKZF2,MYBL1,LMO2	2717	MYB,GFI1B,TAL1	4246
	ERG,FOXP1,HMBOX1	6576	FOXP1,HMBOX1,NFIA	5984	EHF,MYB,LMO2	2083	HMBOX1,RUNX3,FOXP1	6614	MYB,CHD2,TAL1	3825	ERG,TFEC,E2F2	4435	IKZF2,MYBL1,HMBOX1	2675	MYB,IKZF1,TAL1	4240
	SPI1,FOXP1,CHD2	6576	FOXP1,HMBOX1,ZNF350	5969	MYB,IKZF1,LMO2	2083	FLI1,TCF7,FOXP1	6563	FOXO3,MYB,TAL1	3783	ERG,TFEC,ZSCAN16	4406	MYBL1,IKZF1,ZBTB16	2603	MYB,ITGB2,TAL1	4219
	MEIS1,FOXP1,HMBOX1	6575	FOXP1,HMBOX1,CHD2	5953	IKZF1,LMO2,ZBTB16	2078	SPI1,HMBOX1,FOXP1	6457	MYB,GZF1,LMO2	3747	ERG,NFIA,E2F2	4364	MYBL1,IKZF1,LMO2	2564	MYB,LMO2,TAL1	4204
	FOXP1,HMBOX1,CHD2	6572	MEIS1,FOXP1,HMBOX1	5948	CEBPA,IKZF1,LMO2	2072	FLI1,HMBOX1,IKZF1	6398	CREB5,MYB,LMO2	3720	NFIA,TFEC,ZSCAN16	4354	IKZF2,MYBL1,IKZF1	2562	MYB,TAL1,TFEC	4201
	FLI1,FOXP1,HMBOX1	6566	FOXP1,HMBOX1,BBX	5937	BHLHE41,EHF,LMO2	2070	FLI1,REL,FOXP1	6391	MYB,ZBTB16,TAL1	3649	ATF3,NFIA,E2F2	4324	BCL6B,MYBL1,LMO2	2560	MYB,IKZF3,TAL1	4192
	SPI1,FOXP1,HMBOX1	6565	FOXP1,HMBOX1,TFEC	5915	IKZF1,ITGB2,LMO2	2068	FLI1,IKZF3,FOXP1	6374	MYB,LMO2,TAL1	3635	NFIA,TFEC,E2F2	4286	GFI1,MYBL1,LMO2	2559	MYB,IKZF1,LMO2	4179
	ITGB2,FOXP1,HMBOX1	6564	FOXP1,HMBOX1,ITGB2	5906	SPI1,IKZF1,LMO2	2064	FLI1,RUNX3,CHD2	6344	MYB,ITGB2,TAL1	3633	ERG,MYB,ZSCAN16	4272	IKZF2,MYBL1,ITGB2	2558	MYB,HMBOX1,TAL1	4168
	FOXP1,HMBOX1,NFIA	6560	FOXP1,HMBOX1,LMO2	5904	IKZF1,HMBOX1,LMO2	2052	RUNX3,CHD2,FOXP1	6332	MYB,GFI1B,TAL1	3630	ATF3,MYB,E2F2	4263	REL,MYBL1,LMO2	2548	MYB,TAL1,ZBTB16	4164

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.

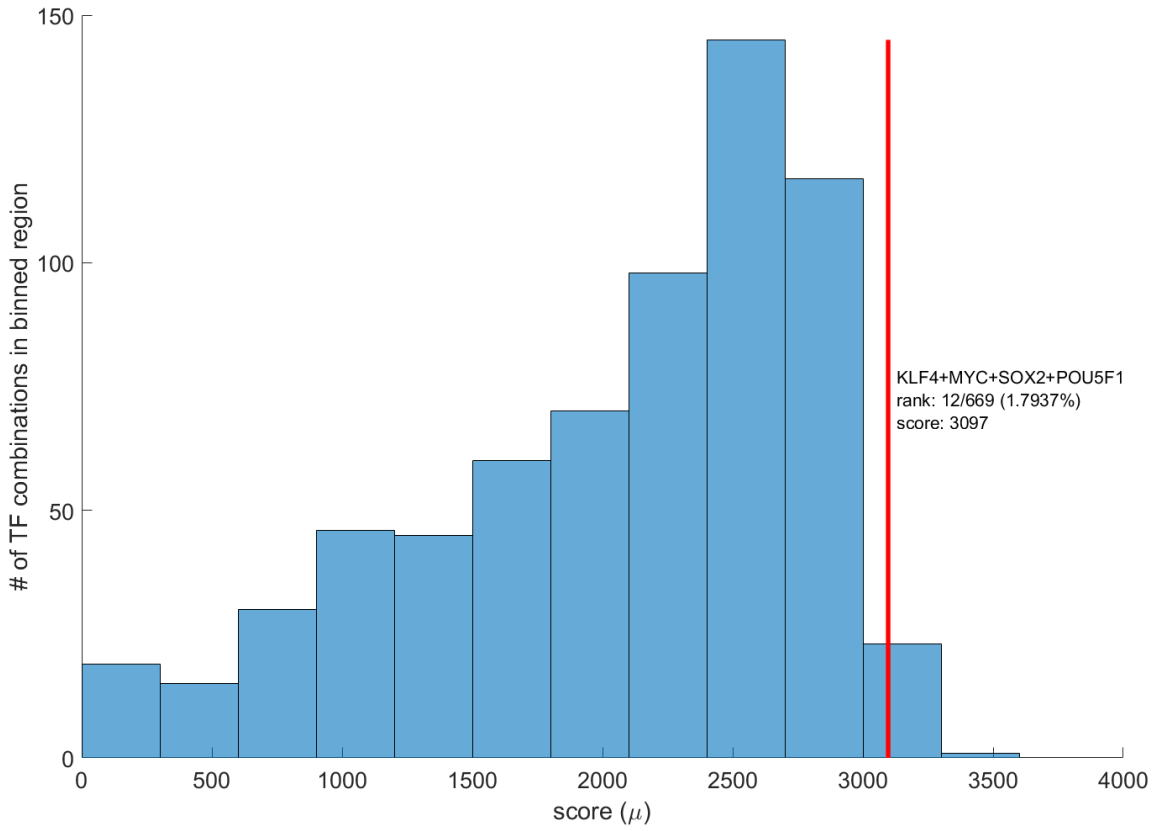


Fig. S5. Histogram of TF scores for fibroblast to ESC reprogramming, for combinations of 4 TFs, without imposing any overexpression thresholding. We show here where the known reprogramming TF combination "Yamanaka Factors" (KLF4, MYC, SOX2, and POU5F1) rank in comparison to randomly selected combinations using our methods (n = 669). Results show KLF4, MYC, SOX2, and POU5F1 ranking 12/669 (1.8%).

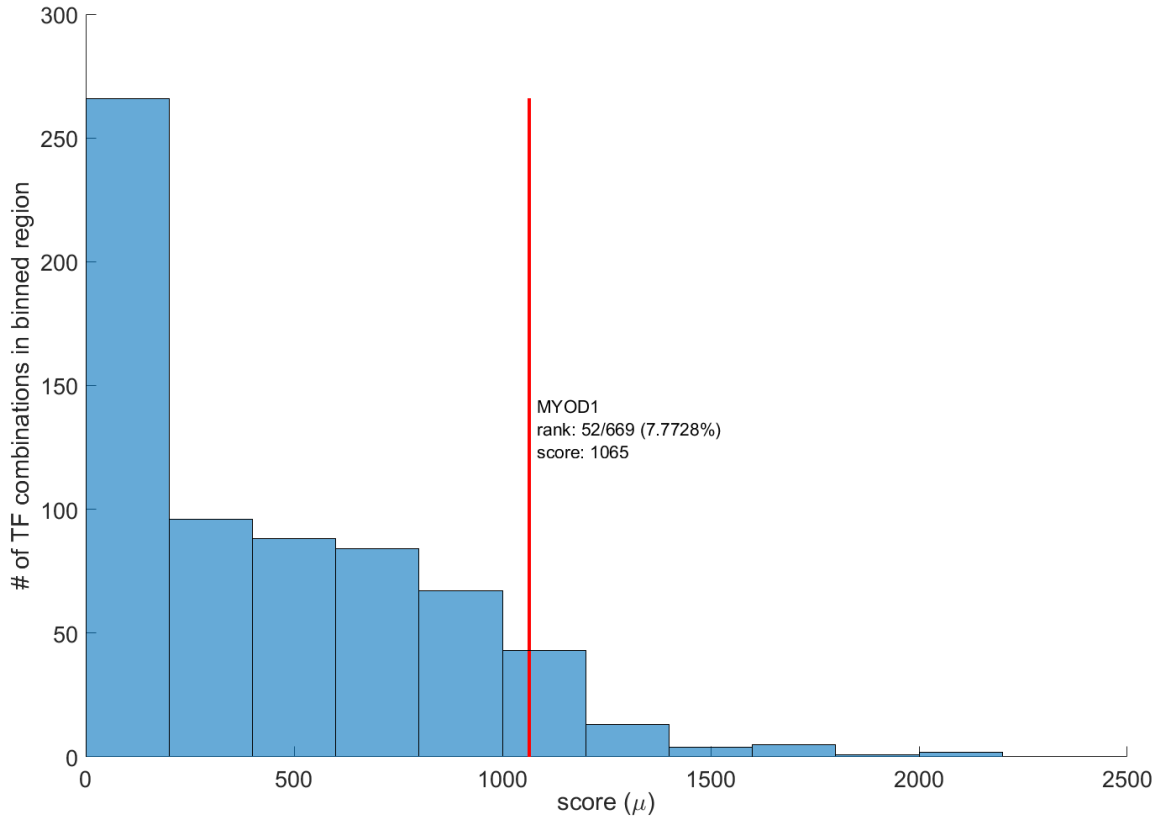


Fig. S6. Histogram of TF scores for fibroblast to myotube reprogramming, for combinations of 1 TFs, without imposing any overexpression thresholding. We show here where MYOD1 ranks in comparison to randomly selected TFs using our methods (n = 669). Results show MYOD1 ranking 52/669 (7.8%).

Initial cell population	Target cell type	Target cell type expression profile used	"Ground truth" Transcription factors and other inputs	Inputs in our dataset (%)
Fibroblast	ESC	ESC (GEO: GSE23316)	SOX2, POU5F1, KLF4, MYC	SOX2, POU5F1, KLF4, MYC (100%)
Fibroblast	ESC	ESC (GEO: GSE23316)	SOX2, POU5F1, NANOG	SOX2, POU5F1, NANOG (100%)
Fibroblast	ESC	ESC (GEO: GSE23316)	SOX2, POU5F1	SOX2, POU5F1 (100%)
Fibroblast	Myotube	Myotube (GEO: GSE52529)	MYOD1	MYOD1 (100%)
Fibroblast	Skeletal Muscle	Skeletal Muscle (GTEx)	MYOD1	MYOD1 (100%)
Fibroblast	Hepatocyte	Liver (GTEx)	HNF1A, HNF4A, HNF6, CEBPA, ATF5, PROX1, p53-siRNA MYC	HNF1A, HNF4A, CEBPA, ATF5, PROX1, MYC (75%)
Fibroblast	Hepatocyte	Liver (GTEx)	HNF1A, HNF4A, FOXA3, SV40 large T antigen	HNF1A, HNF4A (50%)
Fibroblast	Neuron	Neuron (PMID: 25186741)	SOX10	SOX10 (100%)
Fibroblast	Neuron	Neuron (PMID: 25186741)	SOX2	SOX2 (100%)
Fibroblast	Neuron	Neuron (PMID: 25186741)	POU3F2, ASCL1, MYT1L, LHX3, MNX1, ISL1, NEUROG2	POU3F2, LHX3, MNX1, ISL1, NEUROG2 (71%)

Fig. S7. Table overview of target states and target TF combinations for statistical benchmarking.

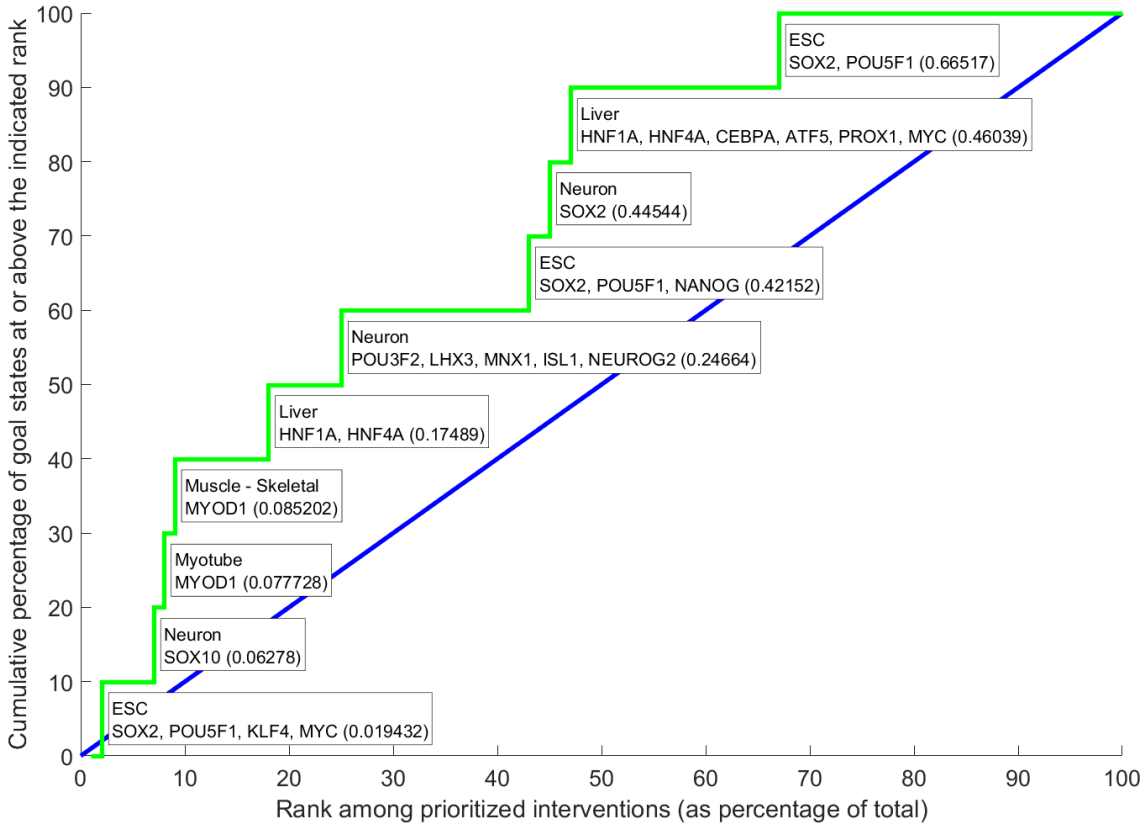


Fig. S8. Cumulative percentage of correct predictions (TFs that have been experimentally validated; y-axis) selected above the percentage rank indicated on the x-axis. The 10 goal states used are shown in Table S7. Text boxes show where a given goal state ranked, with the number in parenthesis giving 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.