ISCI, Volume *6*

Supplemental Information

Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming

Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, Steve Smale, Alfred Hero, Lindsey A. Muir, and Indika Rajapakse

SUPPLEMENTAL FIGURES

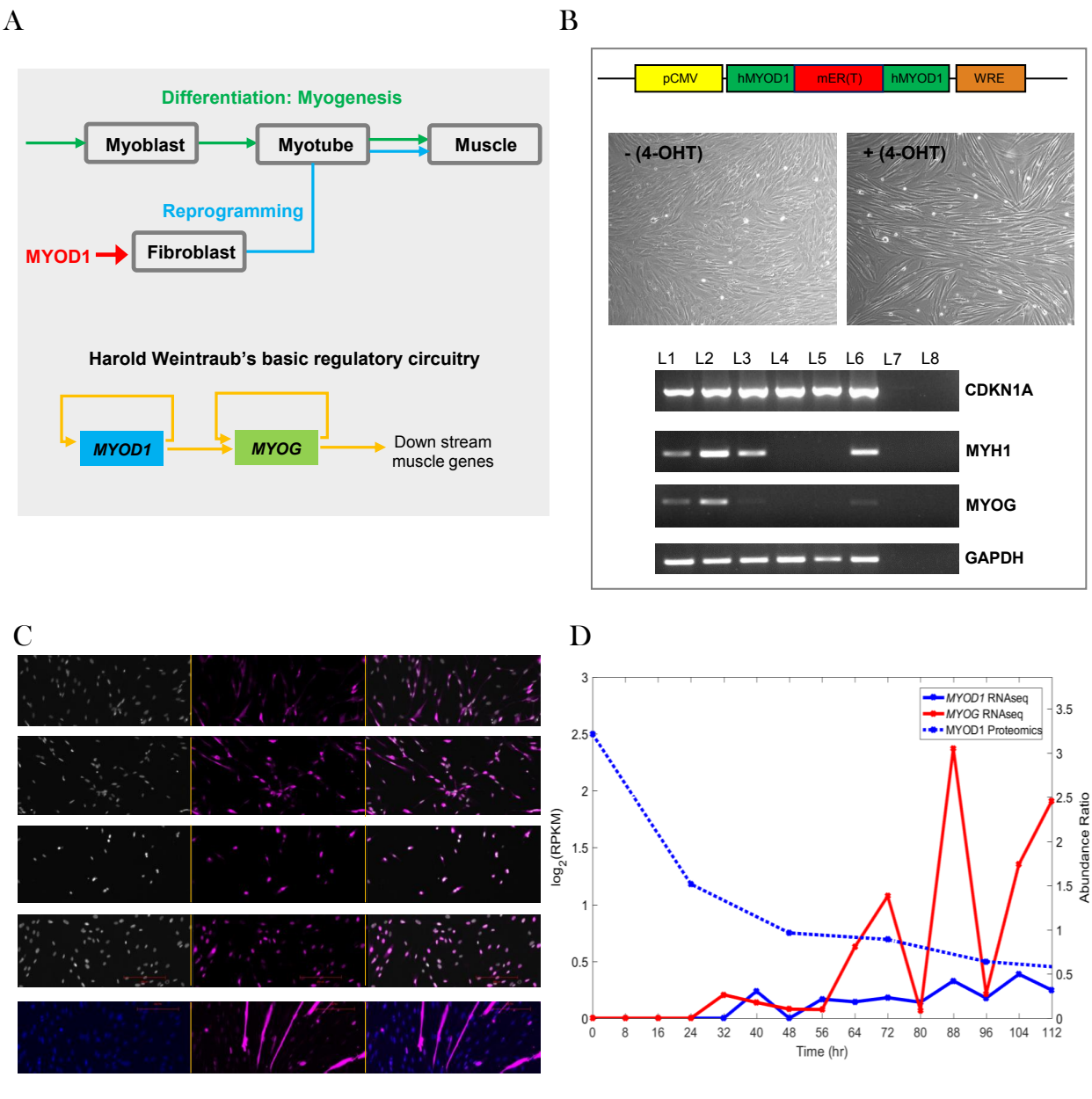
A
Differentiation: Myogenesis
Myoblast
Myotube
Muscle
Reprogramming
MYOD1
Fibroblast
Harold Weintraub's basic regulatory circuitry
MYOD1
MYOG
Down stream muscle genes
B
pCMV
hMYOD1
mER(T)
hMYOD1
WRE
- (4-OHT)
+ (4-OHT)
L1 L2 L3 L4 L5 L6 L7 L8
CDKN1A
MYH1
MYOG
GAPDH
C
D
MYOD1 RNAseq
MYOG RNAseq
MYOD1 Proteomics
log2(RPKM)
Abundance Ratio
Time (hr)

Figure S1

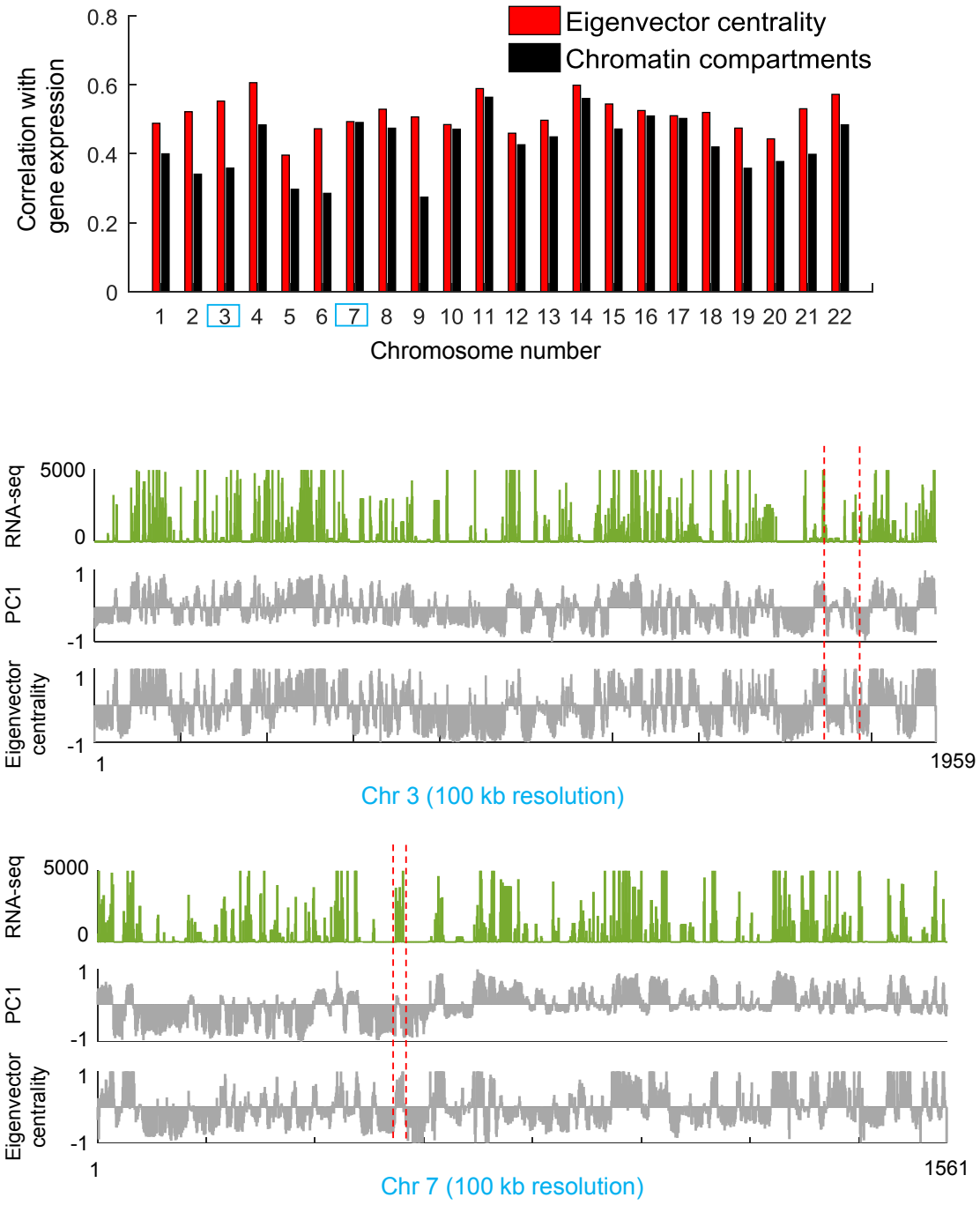
Eigenvector centrality
Chromatin compartments
Correlation with gene expression
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome number
RNA-seq
5000
0
PC1
1
-1
Eigenvector centrality
1
-1
1
1959
Chr 3 (100 kb resolution)
RNA-seq
5000
0
PC1
1
-1
Eigenvector centrality
1
-1
1
1561
Chr 7 (100 kb resolution)

Figure S2

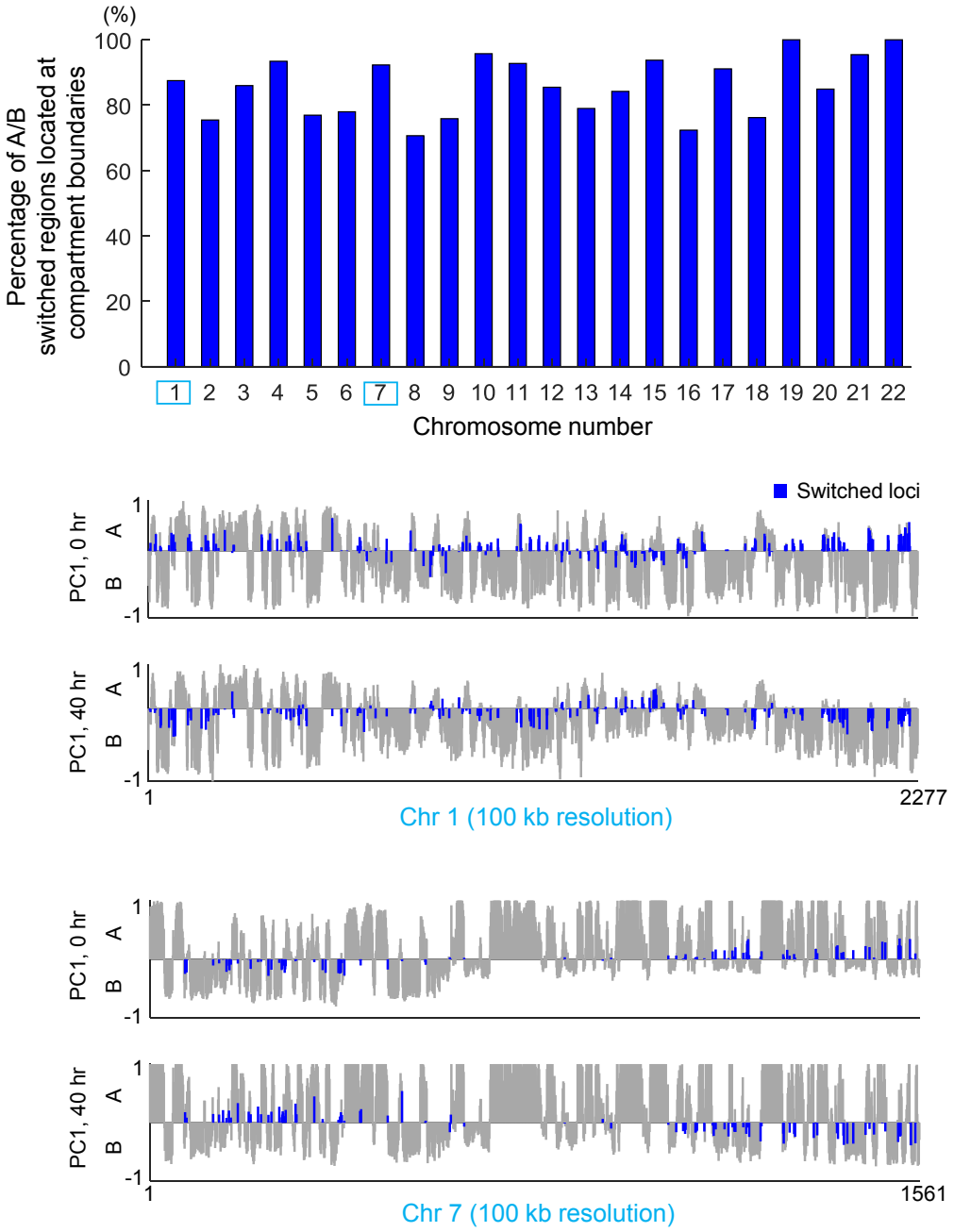
(%)
Percentage of A/B switched regions located at compartment boundaries
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome number
Switched loci
PC1, 0 hr
PC1, 40 hr
A
B
1
-1
1
2277
Chr 1 (100 kb resolution)
1
1561
Chr 7 (100 kb resolution)

Figure S3

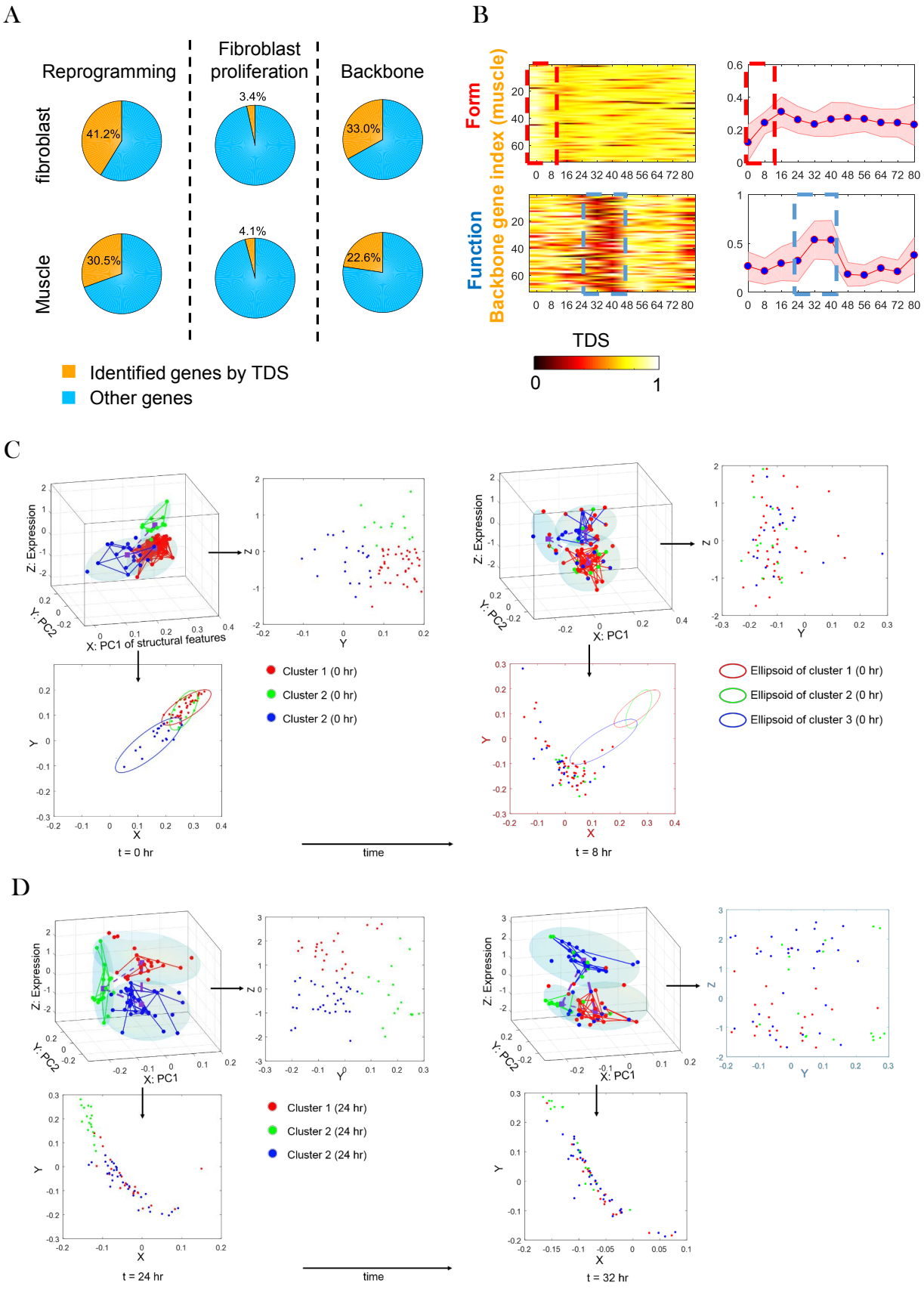
A
Reprogramming
Fibroblast proliferation
Backbone
fibroblast
Muscle
41.2%
3.4%
33.0%
30.5%
4.1%
22.6%
Identified genes by TDS
Other genes
B
Form
Function
Backbone gene index (muscle)
TDS
0
1
C
Z: Expression
X: PC1 of structural features
Y: PC2
Cluster 1 (0 hr)
Cluster 2 (0 hr)
Cluster 2 (0 hr)
Ellipsoid of cluster 1 (0 hr)
Ellipsoid of cluster 2 (0 hr)
Ellipsoid of cluster 3 (0 hr)
t = 0 hr
time
t = 8 hr
D
Cluster 1 (24 hr)
Cluster 2 (24 hr)
Cluster 2 (24 hr)
t = 24 hr
t = 32 hr

Figure S4

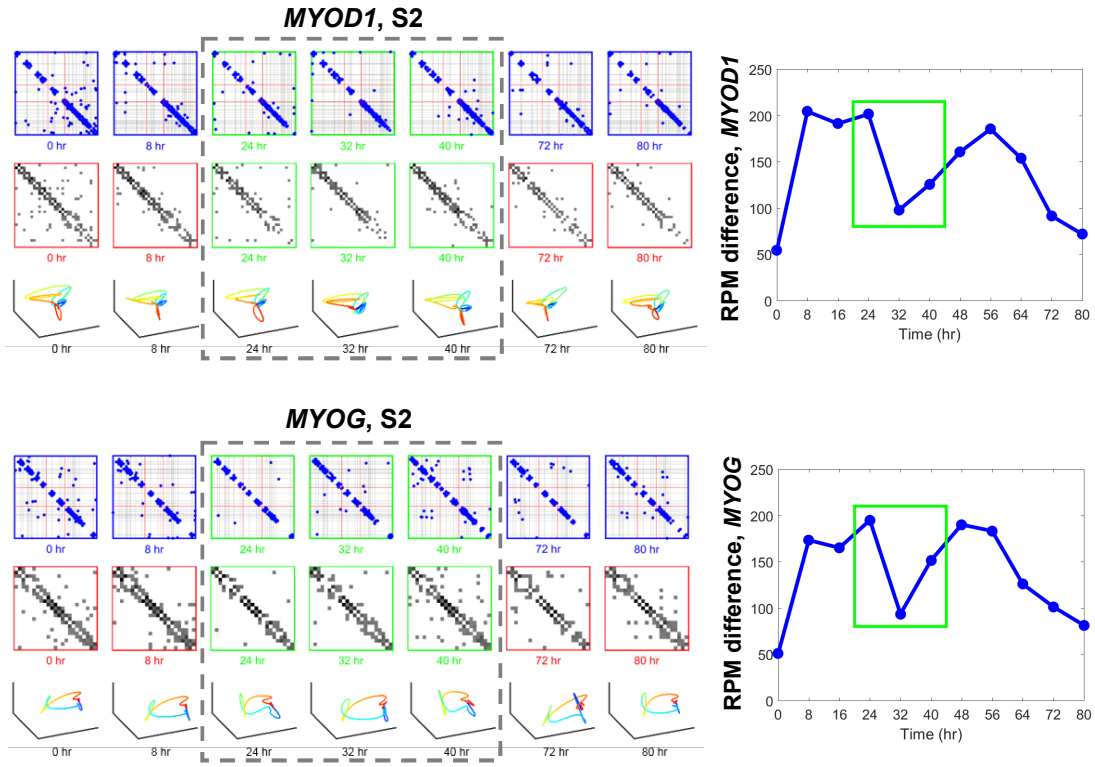
MYOD1, S2
0 hr
8 hr
24 hr
32 hr
40 hr
72 hr
80 hr
RPM difference, MYOD1
Time (hr)
MYOG, S2
RPM difference, MYOG
Time (hr)

Figure S5

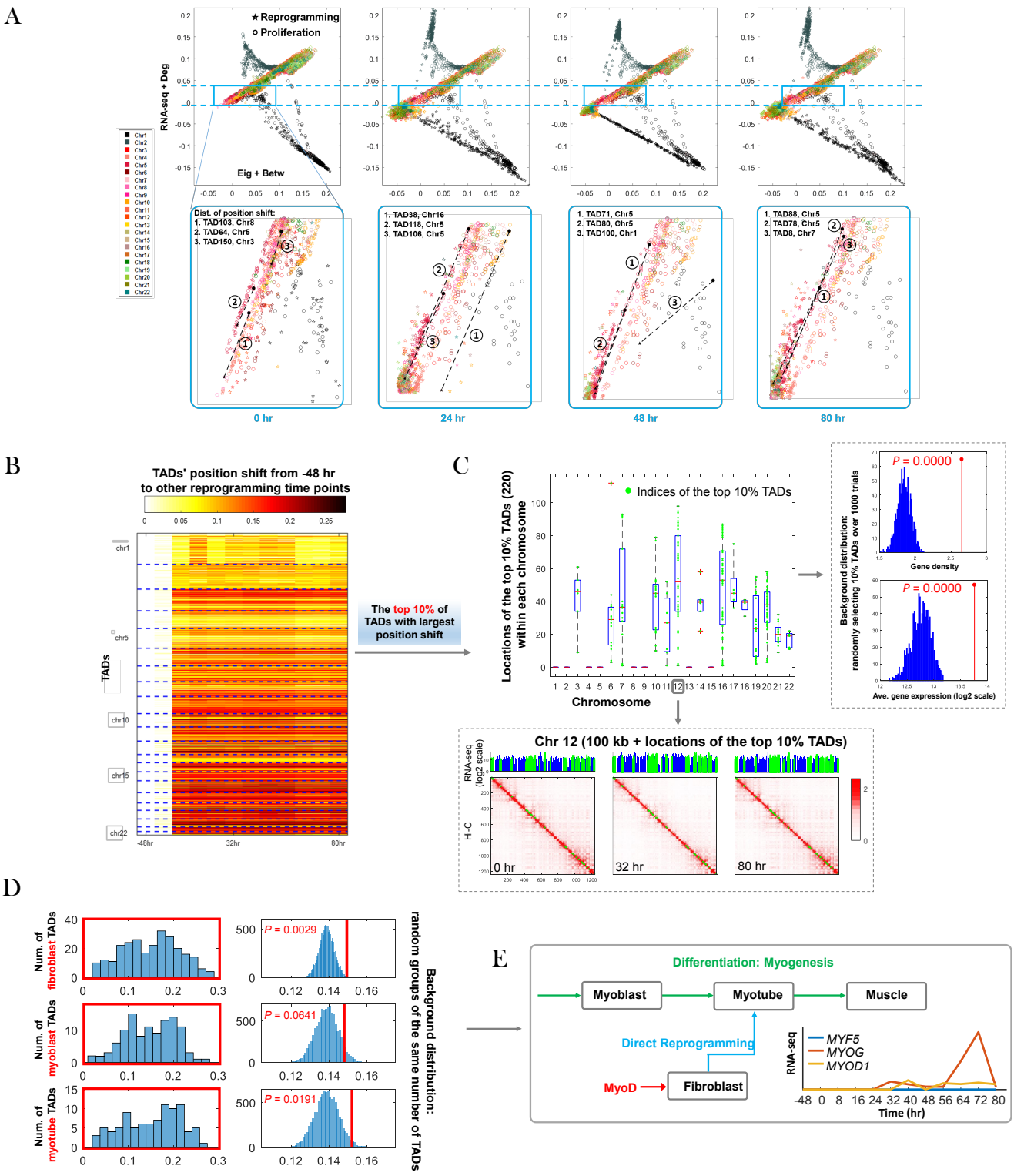
A
★ Reprogramming
○ Proliferation
RNA-seq + Deg
Eig + Betw
Dist. of position shift:
1. TAD103, Chr8
2. TAD64, Chr5
3. TAD150, Chr3
1. TAD38, Chr16
2. TAD118, Chr5
3. TAD106, Chr5
1. TAD71, Chr5
2. TAD80, Chr5
3. TAD100, Chr1
1. TAD88, Chr5
2. TAD78, Chr5
3. TAD8, Chr7
0 hr
24 hr
48 hr
80 hr
B
TADs' position shift from -48 hr
to other reprogramming time points
TADs
-48hr
32hr
80hr
The top 10% of TADs with largest position shift
C
Locations of the top 10% TADs (220) within each chromosome
Indices of the top 10% TADs
Chromosome
Background distribution: randomly selecting 10% TADs over 1000 trials
P = 0.0000
Gene density
P = 0.0000
Ave. gene expression (log2 scale)
Chr 12 (100 kb + locations of the top 10% TADs)
RNA-seq (log2 scale)
Hi-C
0 hr
32 hr
80 hr
D
Num. of fibroblast TADs
Num. of myoblast TADs
Num. of myotube TADs
P = 0.0029
P = 0.0641
P = 0.0191
Background distribution: random groups of the same number of TADs
E
Differentiation: Myogenesis
Myoblast
Myotube
Muscle
Direct Reprogramming
MyoD
Fibroblast
RNA-seq
MYF5
MYOG
MYOD1
Time (hr)

Figure S6

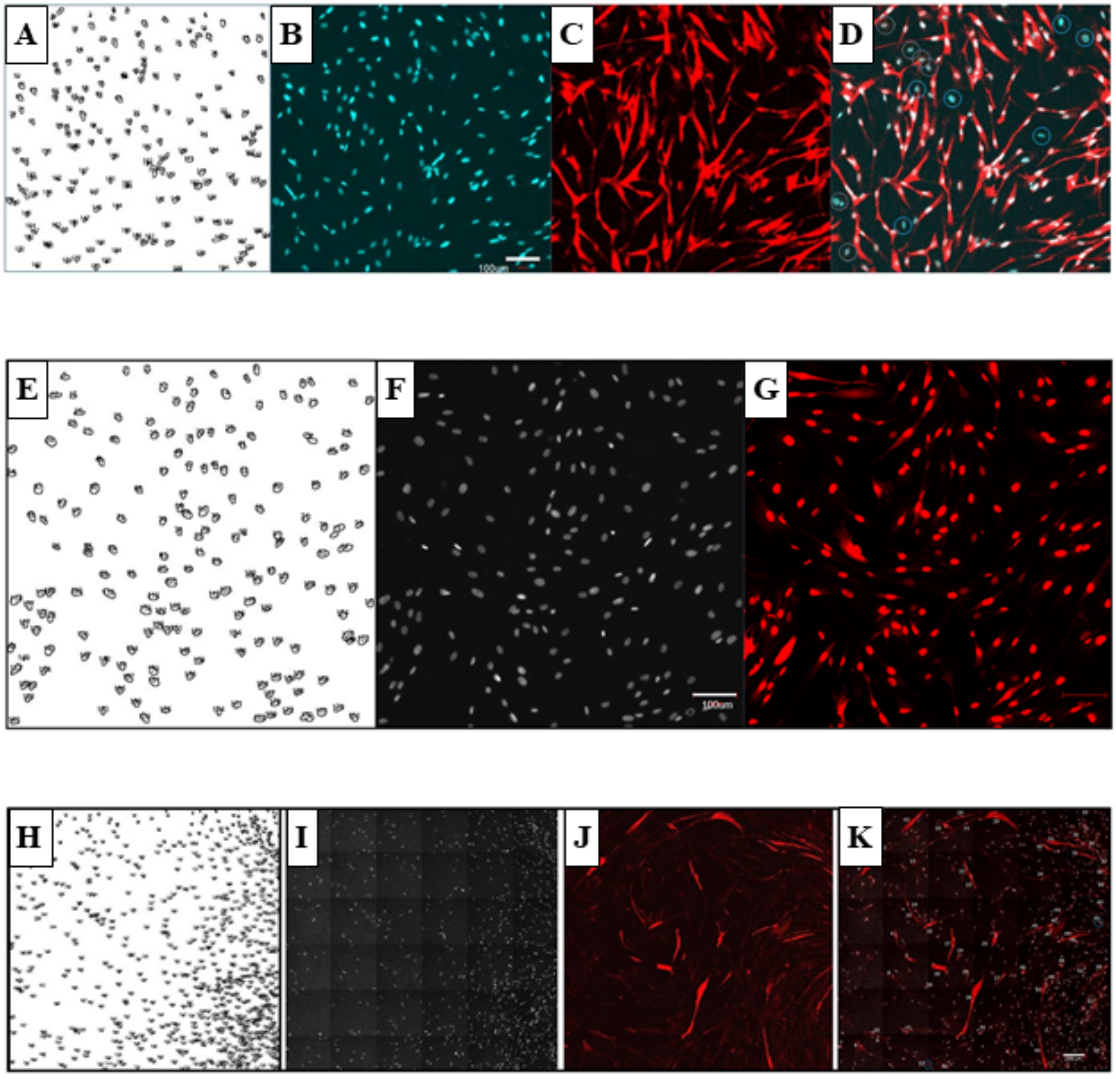

Figure S7

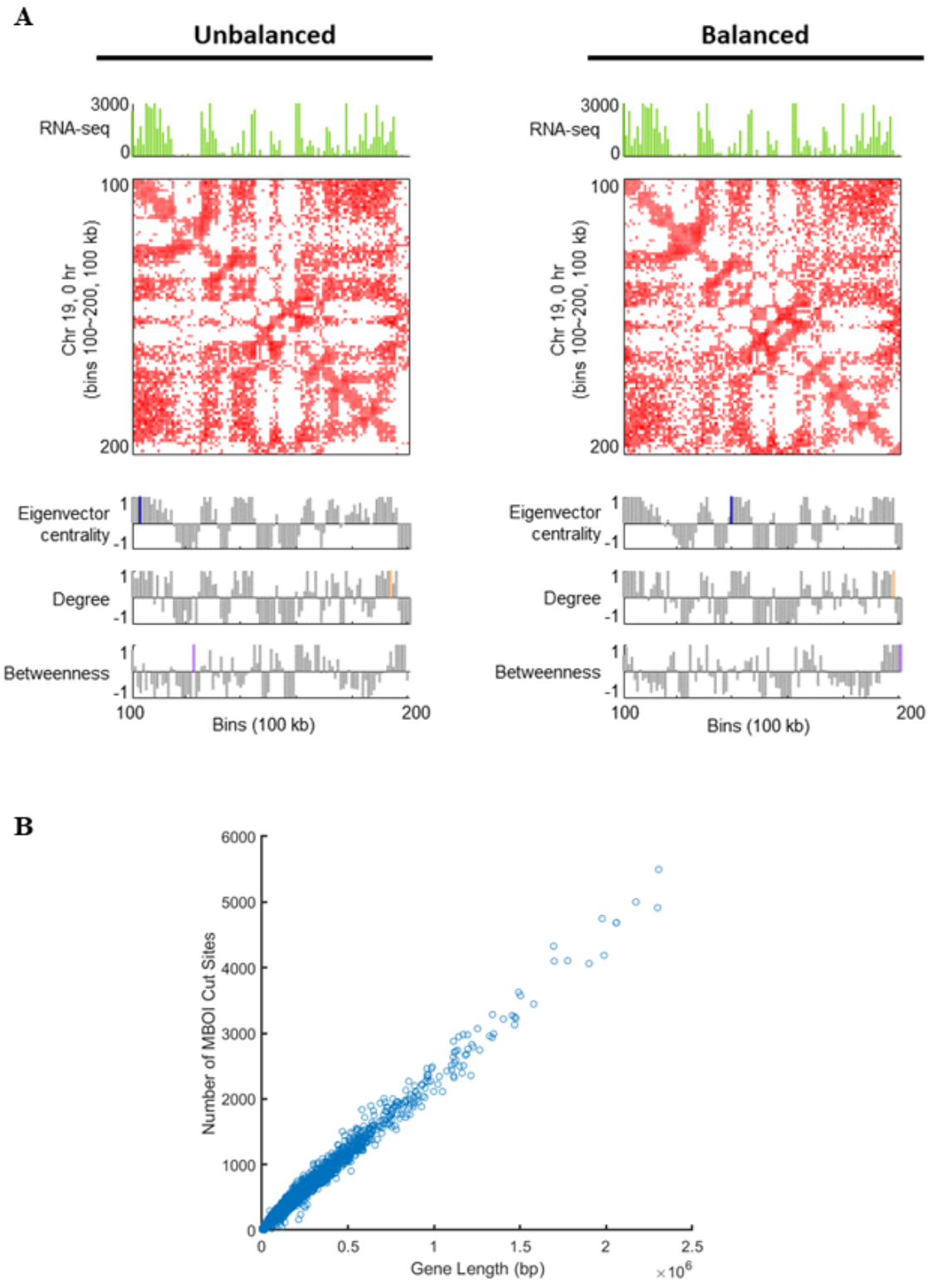
A
Unbalanced
Balanced
RNA-seq
3000
0
100
200
Chr 19, 0 hr
(bins 100~200, 100 kb)
Eigenvector centrality
Degree
Betweenness
1
-1
100
200
Bins (100 kb)
B
Number of MBOI Cut Sites
6000
5000
4000
3000
2000
1000
0
0.5
1
1.5
2
2.5
Gene Length (bp)
×10^6

Figure S8

SUPPLEMENTAL FIGURE TITLES AND LEGENDS

Figure S1: Myogenic conversion of human fibroblasts; Related to Figure 1.

(A) *Top*: Potential transition pathways for MYOD1-mediated fibroblast to muscle cell reprogramming. *Bottom*: Basic gene regulatory circuitry for myogenesis.

(B) *Top*: The cassette for the myogenic reprogramming lentiviral construct, expressing a fusion protein with the mouse mER(T) domain (red box) inserted within the human MYOD1 (green boxes) between amino acids 174 and 175. *Middle*: Light microscope images of cells without (left) or with (right) 4-OHT treatment at differentiation day 3. *Bottom*: RT-PCR validation of gene expression at day 3. Lanes L1/2/3/6, samples transduced with L-MYOD1; L4, not transduced; L5, transduced with an empty lentiviral vector; L7, RT-negative control; L8, no template negative control. Two key MYOD1 downstream genes, *MYOG* & *MYH1* are activated by the expression of L-MYOD1. *GAPDH* is used as an internal control, and *CDKN1A* (P21) is universally expressed.

(C) Left panels, DAPI; middle panels, representative immunostaining for MYOD1 (top four rows) and MYH1 (bottom row); right panels, overlay of left and middle panels.

(D) Time-series RNA-seq (solid line) and proteomic (dashed line) quantification of RNA and protein abundance, respectively, for MYOD1 (blue) and MYOG (red).

Figure S2: Eigenvector centrality refines active and inactive chromatin domains; Related to Figure 2.

Eigenvector centrality yields a higher correlation with gene expression than conventionally defined chromatin partitioning, determined by the first principal component (PC1) of the spatial correlation matrix of Hi-C data (Lieberman-Aiden et al., 2009). Chromosomes 3 and 7 are shown as examples.

Figure S3: Chromatin compartment change appears at boundary regions; Related to Figure 2.

Over 70% of A/B switched bins are at A/B boundary loci. Chromosomes 1 and 7 are shown as examples of chromatin compartment switching from 0 to 40 hrs.

Figure S4: Backbone genes in fibroblast and muscle gene module; Related to Figure 3.

(A) Pie charts showing the portion of backbone genes within each gene module. *Left*: Portion of genes recognized by form-function TDS during cellular reprogramming. *Middle*: Portion of the aforementioned genes that are also active during fibroblast proliferation. *Right*: Backbone genes given by the set of genes extracted from reprogramming but excluding those from proliferation.

(B) Heatmap of form and function TDS for muscle-related backbone genes.

(C) 3D configuration of muscle-related backbone genes in form-function space from 0 to 8 hrs, highlighting significant form change. The edge represents Hi-C contact between genes. Three clusters of genes at 0 hr are marked by red, green, and blue, respectively. The 3D ellipsoid determined by MVE provides the clustering envelope at the current time, where its centroid is marked by a purple square.

(D) 3D configuration of muscle-related backbone genes in form-function space from 24 to 32 hrs, highlighting significant function change.

Figure S5: Genomic dynamics of *MYOD1* and *MYOG*; Related to Figure 4A.

Top left or *Bottom left*: First row depicts Hi-C contact maps of *MYOD1* (or *MYOG*) at base pair scale, where blue points are contacts, red lines depict gene boundaries, and dashed black lines depict MboI cut-sites. Middle rows show Hi-C matrices binned by MboI cut sites and normalized by RPM. Bottom row shows 3D gene models, given by cubic Bézier curves that fits 3D representation of MboI binned contact matrices using Laplacian eigenmaps (Methods). *Top right* or *Bottom right*: Summation of entry-wise differences of Hi-C matrices for *MYOD1* (or *MYOG*) between time points.

Figure S6: A direct pathway from fibroblasts to myotubes; Related to Figure 4A.

(A) 2D representations of TAD-scale form-function features at time 0, 24, 48 and 80 hrs. The star marker represents the coordinate of a TAD at the reprogramming time instant. The circle marker represents the TAD at the stage of fibroblast proliferation (−48 hr). A specified region of data configuration (top plots) is magnified in bottom plots, where three topologically associating domains (TADs) with the 1st, 10th and 20th largest position shift (from proliferation to reprogramming) are marked.

(B) Heatmap of TADs' position shift from −48 hr to reprogramming time points.

(C) TADs with top 10% largest position shift. *Top left*: Locations of the identified TADs over chromosomes. *Bottom left*: Example of identified TADs (green color) at Chromosome 12 (100 kb-binned Hi-C) together with gene expression at time 0, 32 and 80 hrs. *Right*: P values of gene density and average gene expression.

(D) Position shift of TADs that involve fibroblast, myoblast, myotube, and skeletal muscle related genes, respectively. *Left*: Histograms of TADs' position shift for each gene module of interest. *Right*: P value of average position shift for each gene module.

(E) Direct pathway from fibroblasts to myotubes evidenced by gene expression of three myogenic regulatory factors: *MYF5*, *MYOD1* and *MYOG*.

Figure S7: Reprogramming Efficiency; Related to Figure 1A.

(A-D) Cytoplasmic MYOD after lentiviral transduction; (E-G) Translocation efficiency; (H-K) Percentage of Cells Expressing Myosin Heavy Chain (MYH1), 3 days after the end of 4OHT treatment. Scale bar: 100 µm.

(A) 185 nuclei/cell count.

(B) Original nuclei.

(C) MYOD1 cytoplasmic distribution.

(D) 173 cells expressing cytoplasmic MYOD1, and 12 cells without expression for a 94% transduction efficiency.

(E) 183 nuclei counted.

(F) Original Nuclei.

(G) Nuclear MYOD1 signal in all nuclei, but varied intensity, with 16 of the cells showing both cytoplasmic and nuclear staining.

(H) 739 nuclei/cells counted.

(I) Original nuclei.

(J) MYH1 positive cells.

(K) Overlay of nuclei and count of 58 MYH1 positive cells (7.8%).

Figure S8: Balanced vs unbalanced Hi-C analysis; Related to Figure 1B and Figure 2A.

(A) Similarity between analysis performed on balanced vs unbalanced matrices.

(B) Correlation between gene length and the number of restriction enzyme cut sites.

SUPPLEMENTAL TABLES AND TITLES

Table S1
Title: Identified genes at A/B switched loci. Related to Figure 2 and S3.

Table S2
Title: Gene clusters with significant function and form change during time. Related to Figure 3.

Table S3
Title: Gene modules of interest. Related to Figure 3, 6 and S6.

Table S4
Title: Core myogenic genes that steer cellular reprogramming. Related to Figure S4.

Table S5
Title: List of miRNAs that significantly change expression level over the reprogramming time course. Related to Figure 5.

Table S6
Title: JTK output for E-box circadian genes. Related to Figure 6B2.

Table S7
Title: Hi-C resolutions used for analysis in the indicated sections and figures. Related to all main document figures.

Table S8
Title: Number of sequenced and mapped reads for each Hi-C and RNA-seq sample. Related to all main document figures.

TRANSPARENT METHODS

Generation of a human MYOD1-expressing construct

We generated a lenti-construct (lenti-hMYOD1-mER(T)) expressing the human myogenic differentiation factor 1 protein (hMYOD1) fused with a tamoxifen-specific binding domain (mER(T)) derived from mouse estrogen receptor 1 (Kimura et al., 2008). The open reading frame (ORF) for the fusion protein was synthesized at IDT (Integrated DNA technologies) as one gBLOCK, and cloned into the NheI/EcoRI sites of a lenti-vector (obtained from the University of Michigan Vector Core). The expression of the fusion protein is driven by a CMV promoter. The lenti-viral particles were produced at the University of Michigan Vector Core facility for transduction of human BJ fibroblasts with normal karyotype (Cat# CRL2522, ATCC).

Cell culture, lentiviral transduction, and induction of MYOD1 reprogramming

BJ cells were propagated in growth medium (GM) composed of DMEM (Cat# 11960069, Thermo Fisher Scientific), 10% fetal bovine serum (Cat# 10437028, Thermo Fisher Scientific), 1x non-essential amino acids (Cat#11140050, Thermo Fisher Scientific), and 1x Glutamax (Cat# 35050061, Thermo Fisher Scientific). The day before viral transductions, fibroblasts at the 7th passage were plated in 6-well plates or T75 flasks in 13 mL of GM. We plated 1×10^5 cells per well in 6-well plates for RNA extraction, and 2×10^6 cells per flask T75 flasks for Hi-C and proteomics sampling. The cells were incubated in an incubator at 37° C with 5% of CO2.

Lentiviral transduction was performed the next day after plating the cells. We used a MOI (multiplicity of infection) of 15 to transduce the cells in 8 mL GM plus 4 µg/mL of polybrene (Cat# 107689, Sigma-Aldrich). The transduction incubation was carried out in an incubator at 37° C with 5% CO2 for 12 hours. After the incubation, the transduction medium was removed, and the cells were washed with PBS (Cat# 10010049, Thermo Fisher Scientific), then fed with 13 mL of fresh GM to continue incubation for 24 hours.

To induce myogenic reprogramming, we treated the cells transduced with lenti-hMYOD1-mER(T) with (Z)-4-Hydroxytamoxifen (4-OHT) (Cat# H7904, Sigma-Aldrich) to a final concentration of 1 µM in GM for two days. Treatment with 4-OHT induces nuclear translocation of the cytoplasmic hMYOD1-mER(T) protein and initiation of myogenic reprogramming (Kimura et al., 2008). To induce differentiation after 4-OHT treatment, we washed the cells twice with PBS, and changed to differentiation medium consisting of DMEM supplemented with 2% horse serum (Kimura et al., 2008).

Reprogramming Efficiency

At 48 hours post transduction, we detected MYOD1 expression in the cytoplasm in approximately 94% of the cells using an anti-MYOD1 antibody for immunocytochemistry analysis (Figures 2A-D). After a 1 µM daily addition of 4-OHT for two consecutive days, we observed translocation of MYOD1 from the cytoplasm into the nucleus in 100% of the cells expressing MYOD1. MYOD1 positive percentage: 93.6% to 96.8% (Figures 2E-G). In these experiments, we did not evaluate fibroblast markers at single cell resolution (e.g., by im-

munocytochemistry). By 3 days post-4-OHT treatment, we confirmed expression of myosin heavy chain 1 (MYH1), detected in approximately 8% of the MYOD1 expressing cells (Figures 2H-K). Certainly heterogeneity is a caveat of all population-level Hi-C or RNA-seq data, and there is clearly heterogeneity in our reprogramming cell population. Selection is one way to reduce heterogeneity, but we aimed to minimize time between transduction and reprogramming, maintain a low and consistent passage number, and also limit external perturbation as much as possible. Despite these caveats, our goal here was to acquire signatures of reprogramming across the population of cells, and in our data we discerned gene expression patterns consistent with reprogramming based on discrimination from the known fibroblast signature.

Crosslinking of cells for Hi-C

At each time point across the time course, cells in T75 flasks were washed with 10 mL PBS, then incubated with 15 mL of 1% formaldehyde prepared in PBS at room temperature for 10 min. To quench the crosslinking reaction, 2.5 M glycine was added to the flask to a final concentration of 0.2 M, and incubated for 5 min at room temperature on a rocking platform, then on ice for at least 15 min to stop crosslinking completely. The cells were removed from plates by scraping and transferred into 15 mL tubes. The crosslinked cells were collect by centrifugation at 800 x g for 10 min at 4° C. Collected cells were washed in 1 mL ice-cold PBS briefly, and centrifuged at 800 x g for 10 min at 4° C. After centrifugation, the supernatant was removed completely, and the cells were snap-frozen in liquid nitrogen and stored at −80° C for Hi-C library construction.

RNA-seq and small RNA-seq

We used the miRNeasy Mini Kit (Cat# 217004, Qiagen) for total RNA isolation according to the manufacturer's manual. The RNA samples extracted from each sampling time point were treated with RNase-Free DNAase I (Cat# 79254, Qiagen) to clean up any DNA contamination.

All RNA-seq and small RNA-seq data were generated at the University of Michigan Sequencing Core facility. RNA quality control (QC) was performed at the Core. The QC results from the TapeStation analysis (Agilent, Technologies) showed that the samples' RNA integrity number (RIN) was > 9.8. The RNA-seq libraries were prepared according to the TruSeq RNA Library Prep Kit v2 chemistry (Cat# RS-122-2001, Illumina). The small RNA-seq libraries were prepared with the NEBNext® Small RNA Library Prep Set for Illumina (Cat# E7330S, New England Biolabs, NEB).

We sequenced the mRNA species for each samples to produce the RNA-seq dataset, and the small RNA species to obtain the miRNA-seq dataset. Sequence reads were generated on the Illumina HiSeq 2500 platform with the V4 single end 50-base cycle. We used an in house pipeline for sequence read QC (FastQC), genome mapping and alignment (Tophat & Bowtie2), and expression quantification (Cufflinks). We used edgeR (Robinson et al., 2010) for differential expression analysis.

Generation of Hi-C libraries for sequencing

We adapted the in situ Hi-C protocols from Rao et al (Rao et al., 2014) with slight modifications. Briefly, we used 1% formaldehyde for chromatin cross-linking. We used approximately 2.5×10^6 cells for each Hi-C library construction. The chromatin was digested with restriction enzyme (RE) MboI (Cat# R0147M, NEB) overnight at 37° C with rotation. RE fragment ends were filled in and marked with biotin-14-dATP (Cat# 19524016, Thermo Fisher Scientific), and ligated with T4 DNA ligase (NEB, M0202). After the chromatin decross-linking and DNA isolation, DNA samples were sheared on a Covaris S2 sonicator to produce fragments ranging in size of 200-400 bp. The biotinylated DNA fragments were directly pulled down with the MyOne Streptavidin C1 T1 beads (Cat# 65001, Thermo Fisher Scientific). The ends of pulled down DNA fragments repaired, and ligated to indexed Illumina adaptors. The DNA fragments were dissociated from the bead by heating at 98° C for 10 minutes, separated on the magnet, and transferred to a clean tube.

Final amplification of the library was carried out in multiple polymerase chain reactions (PCR) using Illumina PCR primers. The reactions were performed in 25 µL scale consisting of 25 ng of DNA, 2 µL of 2.5mM dNTPs, 0.35 µL of 10 µM each primer, 2.5 µL of 10X PfuUltra buffer, PfuUltra II Fusion DNA polymerase (Cat# 600670, Agilent). The PCR cycle conditions were set to 98° C for 30 seconds as the denaturing step, followed by 14 cycles of 98° C 10 seconds, 65° C for 30 seconds, 72° C for 30 seconds, then with an extension step at 72° C for 7 minutes.

After PCR amplification, the products from the same library were pooled and fragments ranging in size of 300-500 bp were selected with AMPure XP beads. The size selected libraries were sequenced to produce paired-end Hi-C reads on the Illumina HiSeq 2500 platform with the V4 of 125 cycles.

Generation of Hi-C matrices

We standardized an in house pipeline to process Hi-C sequence data. With this pipeline, FastQC (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`) was used for quality control of the raw sequence reads. Paired-end reads with excellent quality were mapped to the reference human genome (HG19) using Bowtie2 (Langmead and Salzberg, 2012) , with default parameter settings and the "–very-sensitive-local" preset option, which produced a SAM formatted file for each member of the read pair (R1 and R2). HOMER was run with the recommended settings. Uninformative paired-end reads were filtered using the "makeTagDirectory" program with the "–tbp 1 -removePEbg -restrictionSite GATC -both -removeSelfLigation -removeSpikes 10000 5" settings. Unnormalized raw Hi-C matrices were generated with "analyzeHiC" with the "-raw" and "-res 1000000" or "-res 100000" settings to produce the raw contact matrix at 1 Mb resolution or 100 kb resolution, respectively.

Hi-C Normalization

These Hi-C data were not balanced/iteratively corrected. Balancing our Hi-C matrices does not change the overall structure of these matrices significantly, and results obtained from balanced matrices are similar to results obtained on non-balanced matrices. To show this, we have recreated manuscript Figure 2A for both balanced and unbalanced matrices (Figures S8A). Centrality measurements that are crucial to our analysis throughout the

paper (eigenvector, degree, and betweenness) are very similar when computed on balanced matrices. This was performed at 100 kb resolution using the Knight-Ruiz algorithm for balancing before Toeplitz normalization (Knight and Ruiz, 2013). Furthermore, since we use a 4-cutter restriction enzyme, MBOI, the number of cuts sites per gene is strongly correlated with gene length. We have calculated the number of MBOI cut sites vs the length of each gene to show this correlation (Figures S8B). These measures are highly correlated (R^2 = 0.988), leading us to believe that the number of cut-sites per gene is not skewing our analysis.

Reverse transcriptional polymerase chain reaction (RT-PCR) analysis

The cDNA templates for RT-PCR were synthesized from 1 µg RNA using the SuperScript® III First-Strand Synthesis System (Cat# 18080051, Thermo Fisher Scientific). Targets amplicons of corresponding genes were amplified in 20 µL reactions using the following settings: initial denaturation was performed at 95° C for 5 min, followed by 30 cycles at 95° C for 15 seconds, 56° C for 30 seconds, and 72° C for 20 seconds. The PCR reactions were then incubated for a final extension step at 72° C for 5 min. The products were analyzed on 1.5% agarose gel. The gel image was taken on an imaging station (Universal Hood II, Bio Rad).

Immunocytochemistry analysis

Cells were grown in appropriate media on washed and autoclaved 12mm round 1.5 glass coverslips placed in 12 well culture plates. At harvest, coverslips were rinsed briefly in phosphate-buffered saline pH 7.4 (PBS), treated with 4% paraformaldehyde in PBS for 10 min at room temperature, then washed three times in PBS at 5 minutes per wash. Cells were dehydrated in a series of ice-cold ethanol concentration steps, 50%, 70%, 90% and 100% at 5 minutes per step, and stored at 4° C until staining. Rehydration reversed the concentration series, with two washes in cold PBS at the end. Cells were permeabilized for 10 min in a PBS 0.25% Triton X-100 solution at RT, and then washed in PBS three times for 5 min per wash. Blocking of non-specific antibody binding was performed with 1% BSA PBST (PBS + 0.1% Tween 20) for 30 minutes, followed by immunostaining using primary antibody (DSHB anti-MHC MF20 diluted 1:20, and/or Thermofisher anti-MyoD diluted 1:250) in 1% BSA in PBST in a humidified chamber for 1 hr at room temperature (RT). The primary solution was removed, cells were washed three times in PBS at 5 min per wash, and the fluorescent secondary, Alexa Fluor 594 goat anti-mouse IgG in 1% BSA PBST was applied for 1 hr at RT in the dark. The secondary antibody solution was then removed and the cells were washed three times with PBS for 5 min each in the dark. Cells were mounted on slides with Prolong Gold anti-fade reagent with DAPI, and imaged.

QUANTIFICATION AND STATISTICAL ANALYSIS

Scale-adaptive gene expression

Hi-C matrices are commonly created at fixed resolution, or "bins" (e.g., 100kb, 1Mb). However, RNA-seq data (FPKM) are generated at the gene level and genes have variable length. For consistent analysis of form and function, we transform the RNA-seq data from gene level

to bin level, namely,

$$R_{\text{bin}_i} = \sum_{j \in \{\text{genes at bin } i\}} \frac{L_{j,\text{bin}_i}}{L_j} \frac{R_j L_j}{1000} = \sum_{j \in \{\text{genes at bin } i\}} \frac{R_j L_{j,\text{bin}_i}}{1000},$$

where L_j is the length of gene j in base pairs (bp), $\frac{L_j}{1000}$ is the length of gene j in kilobases (kb), L_{j,bin_i} is the length of the portion of gene j belonging to bin i, R_j signifies the FPKM value of gene j, and R_{bin_i} denotes the total RNA-seq RPM value at bin i.

Scale-adaptive Hi-C matrix

It is expected that loci that are close together in linear bp distance are more likely to be ligated together than distant pairs. This makes a Hi-C matrix highly diagonally dominant and conceals the contact pattern embedded in the matrix. In order to alleviate this effect, we normalize the counts by their contact probability as a function of the linear distance, namely, each entry of the matrix is normalized by its expected contact value (expected-observed method). This is equivalent to normalization of the Hi-C matrix by a Toeplitz structure whose diagonal constants are the mean values calculated along diagonals of the observed matrix; see details in (Chen et al., 2015, SI).

Similar to scale-adaptive gene expression, we are also able to construct gene-resolution Hi-C contact maps by calculating the contact frequency between two genes, which is normalized by the lengths of the genes (Chen et al., 2015). Moreover, to construct TAD-scale contact matrices, we first normalize both intra- and inter-chromosome Hi-C matrices at 100 kb resolution, and then compute the density of genome contacts between TADs. TAD boundaries here are defined based on (Dixon et al., 2012). Given TADs i and j, the resulting contact map $\mathbf{T}$ is given by

$$[\mathbf{T}]_{ij} = \frac{\sum_{m \in \text{TAD}_i} \sum_{n \in \text{TAD}_j} [\tilde{\mathbf{H}}]_{mn}}{L_i L_j},$$

where $\tilde{\mathbf{H}}$ is the normalized Hi-C matrix (100kb-binned Hi-C in our analysis), and L_i is the size of TAD_i. Since the TAD-scale contact matrix is dense, we apply thresholding to make the matrix more sparse by retaining only interactions that exceed the 50th-percentile of Hi-C contacts at the TAD scale.

Network representation of 4DN: graph Laplacian and Fiedler number

Let $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ denote a weighted undirected graph at time t, where $\mathcal{V}$ is a node set with cardinality $|\mathcal{V}| = n$, and $\mathcal{E}_t \subset \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ is an edge set at time t. The Hi-C matrix $\mathbf{H}_t$ can then be interpreted as an adjacency matrix corresponding to $\mathcal{G}_t$, where $(i, j) \in \mathcal{E}_t$ if there exists interactions between node i and j with edge weight $[\mathbf{H}_t]_{ij} > 0$ and $[\mathbf{H}_t]_{ij} = 0$ otherwise. Here nodes represent fixed-size bins, genes or TADs. It is often the case that a graph/network is represented through the graph Laplacian matrix, $\mathbf{L}_t = \mathbf{D}_t - \mathbf{H}_t$, where $\mathbf{D}_t = \text{diag}(\mathbf{H}_t \mathbf{1})$ is the degree matrix of $\mathcal{G}_t$, $\mathbf{1}$ denotes the vector of all ones, and $\text{diag}(\mathbf{x})$ signifies the diagonal matrix with diagonal vector $\mathbf{x}$. Given $\mathbf{L}_t$, the Fiedler number

and the Fiedler vector are defined by the second smallest eigenvalue and its corresponding eigenvector. It is known from spectral graph theory (Chung, 1997) that $\mathcal{G}_t$ is connected (namely, there exists a path between every pair of distinct nodes) if and only if the Fiedler number is nonzero. The entrywise signs of the Fiedler vector encodes information on network partitioning. For a network with Fiedler number equal to zero, we can extract its largest connected component (LCC), namely, the largest subgraph with nonzero Fiedler number.

Structural feature extraction via network centrality measures

A network/graph centrality measure is a quantity that evaluates the influence of each node to the network, and thus provides essential topological characteristics of nodes (Newman, 2010). In what follows, we introduce the key centrality measures used in our analysis and elaborate on the rationale behind them.

• Degree. A nodal degree is defined as the sum of edge weights (namely, Hi-C contacts) associated with each node,

$$\text{degree}(i,t) = \sum_{j=1}^{n} [\mathbf{H}_t]_{ij}, \tag{1}$$

where $\text{degree}(i,t)$ denotes the degree of node i at time t. We remark that $\text{degree}(i,t)$ exhibits the spatial proximity between node i to other nodes.

• Eigenvector centrality. The eigenvector centrality is defined as the principal eigenvector of the adjacency matrix, corresponding to its largest eigenvalue, namely

$$\text{eig}(i,t) = [\mathbf{v}_t]_i = \frac{1}{\lambda_1(\mathbf{H}_t)} \sum_{j=1}^{n} [\mathbf{H}_t]_{ij} [\mathbf{v}_t]_j, \tag{2}$$

where $\lambda_1(\mathbf{H}_t)$ is the maximum eigenvalue of $\mathbf{H}_t$ in magnitude, and $\mathbf{v}_t$ is the associated eigenvector, namely $\lambda_1(\mathbf{H}_t)\mathbf{v}_t = \mathbf{H}_t\mathbf{v}_t$. It is clear from (2) that the eigenvector centrality relies on the principle that a node has more influence if it is connected to many nodes which in turn are also considered to be influential. Different from degree centrality, the eigenvector centrality takes the full network topology into account.

• Betweenness. Betweenness is the fraction of shortest paths that pass through a node relative to the total number of shortest paths in the connected network. The betweenness of node i at time t is defined as

$$\text{betweenness}(i,t) = \sum_{k \in \mathcal{V}, k \neq i} \sum_{\substack{j \in \mathcal{V} \\ j \neq i, j > k}} \frac{\sigma_{kj}(i,t)}{\sigma_{kj}(t)}, \tag{3}$$

where $\sigma_{kj}(t)$ is the total number of shortest paths from node k to j at time t, and $\sigma_{kj}(i,t)$ is the number of such shortest paths passing through node i. Betweenness characterizes potential hub nodes in the network, and thus a node with high betweenness has the potential to disconnect the network if it is removed.

Other centrality measures can also be used, such as clustering coefficient, closeness and hop walk statistics, which differ in what type of influence is to be emphasized (Newman, 2010).

Integration of form and function

The extracted centrality feature vectors can then be combined with function vector (i.e., gene expression) to create a form-function feature matrix $\mathbf{X}_t \in \mathbb{R}^{n\times m}$, where n is the size of the Hi-C matrix, m is the number of extracted features, and t is the time step.

Data representation on low-dimensional non-linear manifolds

Information redundancy exists in the data matrix $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T]^T \in \mathbb{R}^{nk\times m}$, where k is the length of time horizon ($k = 12$ in our dataset). For example, the degree centrality and the eigenvector centrality could be correlated, and the replicates of RNA-seq data are strongly correlated. Therefore, data points given by rows of $\mathbf{X}$ are lying on a manifold with a smaller intrinsic dimensionality m' (often $m' \ll m$) that is embedded in the m-dimensional feature space. The goal of dimensionality reduction is to transform dataset $\mathbf{X}$ into $\mathbf{Y}$ with lower dimensionality m', while retaining the geometry of the data as much as possible (Van Der Maaten et al., 2009).

Laplacian eigenmap is a non-linear dimensionality reduction technique to find a low-dimensional data representation by preserving local properties of the underlying manifold. We remark that the linear dimensionality reduction technique, principal component analysis (PCA), is also applicable but it cannot adequately handle the nonlinearity embedded in the dataset. The method of Laplacian eigenmaps contain the following steps

- Normalize dataset $\mathbf{X} = [\mathbf{X}_1^T, \ldots, \mathbf{X}_k^T]^T$ to make different features comparable

$$\mathbf{X}_t(:,i) = \mathbf{X}_t(:,i)/\sigma_i,\ \ \sigma_i = \max_t\{\|\mathbf{X}_t(:,i)\|_2\}$$

$$\mathbf{X}_t(:,i) = \mathbf{X}_t(:,i) - \mu_i\mathbf{1},\ \ \mu_i = \frac{1}{kn}\sum_{t=1}^{k}\sum_{j=1}^{n}\mathbf{X}_t(j,i),$$

  where $\mathbf{X}_t(:,i)$ denotes the ith column of $\mathbf{X}_t$, the first transformation ensures that different features are all treated on the same scale, and the second transformation is to zero out the mean of the data.

- Construct a neighborhood graph in which every node is linked with its p nearest neighbors. The edge weight is computed using the heat kernel function, leading to a sparse adjacency matrix $\mathbf{W}$ with entries

$$[\mathbf{W}]_{ij} = e^{-\frac{\|\mathbf{X}(i,:)-\mathbf{X}(j,:)\|_2^2}{\sigma}},\ \text{if there is an edge between } i \text{ and } j,$$

  where σ is the heat kernel parameter, and we choose $\sigma = 200$ in our analysis (Van Der Maaten et al., 2009).

- Compute the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} = \text{diag}(\mathbf{W1})$. We then solve the generalized eigenvalue problem

$$\mathbf{Ly} = \lambda \mathbf{Dy} \tag{4}$$

for m' smallest nonzero eigenvalues. The resulting eigenvectors $\{\mathbf{y}_i\}_{i=1}^{m'}$ form the low-dimensional data representation $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{m'}]$.

After dimensionality reduction, we can also evaluate the significance of each feature that contributes to the low-dimensional data representation $\mathbf{Y}$. Let us consider a linear approximation $\mathbf{Y} \approx \mathbf{XQ} = [\mathbf{XQ}(:,1), \ldots \mathbf{XQ}(:,m')]$, and $\mathbf{Q} \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. It is clear that there exists a one-to-one correspondence between the columns of $\mathbf{Y}$ and the columns of $\mathbf{Q}$,

$$\mathbf{Y}(:,j) = \sum_i \mathbf{X}(:,i)Q(i,j).$$

Here $Q(i,j)$ signifies the contribution of the ith feature in $\mathbf{X}$ to the jth component of the obtained low-dimensional column-space $\mathbf{Y}$. The feature score (FS) for the ith feature corresponding to the jth dimension of the subspace is

$$\text{FS}(i,j) = \frac{|Q(i,j)|}{\sum_i |Q(i,j)|}. \tag{5}$$

Fitting the data: minimum volume ellipsoid

The minimum volume ellipsoid (MVE) estimator is the first high-breakdown robust estimator of multivariate location and scatter (Van Aelst and Rousseeuw, 2009). Geometrically, the MVE estimator finds the minimum volume ellipsoid covering, or enclosing a given set of data points. Let $\mathcal{X} = \{\mathbf{x}_i \,|\, \mathbf{x}_i \in \mathbb{R}^m,\ i \in \{1, 2 \ldots, n\}\}$ denote the dataset of interest, where n is the number of data points, and m is the number of features (or the dimension of the intrinsic low-dimensional manifolds). The ellipsoid that fits into $\mathcal{X}$ can be parametrized as

$$\mathcal{W}_{\mathbf{Q},\mathbf{b}} = \{\mathbf{x} \in \mathbb{R}^m \mid \|\mathbf{Qx} - \mathbf{b}\|_2 \leq 1\}, \tag{6}$$

where $\mathbf{Q} \in \mathbb{R}^{m \times m}$ and $\mathbf{b} \in \mathbb{R}^m$ are unknown parameters. The center and the shape of the ellipsoid $\mathcal{E}_{\mathbf{Q},\mathbf{b}}$ is given by $\mathbf{c} \coloneqq \mathbf{Q}^{-1}\mathbf{b}$, and $\mathbf{\Lambda} \coloneqq \mathbf{Q}^2$ since the ellipsoid (6) can be reformulated as $\mathcal{W}_{\mathbf{Q},\mathbf{b}} = \{\mathbf{x} \in \mathbb{R}^m \mid (\mathbf{x} - \mathbf{c})^T\mathbf{\Lambda}(\mathbf{x} - \mathbf{c})\Sigma \leq 1\}$. Finding the minimum volume ellipsoid can be cast as a convex problem

$$\begin{array}{ll} \underset{\mathbf{Q},\mathbf{b}}{\text{minimize}} & \det(\mathbf{Q}^{-1}) \\ \text{subject to} & \|\mathbf{Qx}_i - \mathbf{b}\|_2 \leq 1,\ i \in \mathcal{N}_\alpha \\ & \mathbf{Q} \text{ is positive definite,} \end{array}$$

where $\mathcal{N}_\alpha$ denotes the set of data within a α confidence region, determined by Mahalanobis distances of data below $\alpha = 97.5\%$ quantile of the chi-square distribution with l degrees of

freedom (Van Aelst and Rousseeuw, 2009). The MVE estimates the shape of the uncertainty ellipsoid for $\mathcal{X}$, which is different from its sample covariance. The latter is the maximum likelihood estimate under the assumption of Gaussian distribution.

Temporal difference score (TDS)

TDS is introduced to evaluate the temporal difference of form-function characteristics. Let $\mathbf{X}_t \in \mathbb{R}^{n\times m}$ denote data matrix associated with n nodes of a network and m features. TDS of node i at time t is defined as

$$\mathrm{TDS}(i,t) = \frac{\sum_{t'\in\mathcal{N}_t} \mathrm{dist}(\mathbf{X}_t(i,:), \mathbf{X}_{t'}(i,:))}{|\mathcal{N}_t|}, \tag{7}$$

where $\mathcal{N}_t$ defines the time window around t, namely, $\mathcal{N}_t = \{t-1, t\}$, and $\mathrm{dist}(\cdot)$ is a generic distance function between the ith row of $\mathbf{X}_t$ and $\mathbf{X}_{t'}$. In our analysis, $\mathbf{X}_t$ can represent either network centrality features from Hi-C data or gene expression.

A/B compartment switching analysis

A/B compartments were identified through methods conceptually similar to those described in (Lieberman-Aiden et al., 2009). Intra-chromosomal Hi-C matrices $\mathbf{H}$ were binned at the 100-kb level, with unmappable regions and/or regions with no identified contacts removed. Matrices were Toeplitz normalized based on linear genome distance to derive $\tilde{\mathbf{H}}$ (See Scale-adaptive Hi-C matrix). The entrywise sign of the principal component of the spatial correlation matrix associated with $\tilde{\mathbf{H}}$ (PC1) is used to identify A/B compartments. To determine A/B switching with concordant gene expression, we determined 100-kb bins that switched A/B compartments and whose entry-wise sign change was in the 50th percentile of total change. This was done to reduce noise in A/B compartment switch identification. All genes that overlap with defined A/B switch regions were analyzed for differential expression. Genes that had a mean FPKM value greater than 0.1, and had log2 fold change expression greater than 1 or less than -1 were kept.

Divergence of datasets and statistical significance

To depict the transition into the myogenic lineage, we studied human fibroblast proliferation (Chen et al., 2015) and MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage, over a 56-hr time course. First, we found an intrinsic low-dimensional (3D) manifold of centrality-based form-function features under the setting of both proliferation and reprogramming. This was given by the principal subspace of form-function data at the first two time points (corresponding to the fibroblast-like stage). Second, we obtained the 3D data representation of form-function features after projection onto the common subspace for proliferation and reprogramming, and tracked the centroids of the fitted ellipsoids (given by MVE estimates) over time. The trajectory of the centroids was then smoothed using the cubic spline. Last, we provided a statistical significance for the deviation in trajectory of proliferation and reprogramming at the 32 hr bifurcation, where the P value is defined from the multivariate Hotelling's T-Square test associated with the null hypothesis that the

centroids of proliferation and reprogramming are identical at a given time point.

Bifurcation identification at single gene level
Hi-C contacts within a ±5 kb window around a gene location are extracted. A $\{d+1, d+1, t\}$ tensor $\mathbf{A}_{i,j,t}$ is contructed based on the number of MboI cut-sites (GATC) found, d, within the region of interest, for each time point sampled, t. Each element i, j, t of $\mathbf{A}$ represents the number of contacts found between cut sites $\{i-1, i\}$ and $\{j-1, j\}$ at time t, divided by the total number of contacts found for each time point (RPM). The element-wise difference between time points is calculated, and the summation of difference (absolute value) between t and $t+1$ is recorded.

Identification of genes of interest
Genes of interest (GOIs) are mainly extracted through Gene Ontology (GO), with a few GOI subsets curated through other means. GO-extracted lists include myotube, myoblast, skeletal muscle, fibroblast, and circadian. "Muscle" genes are the union of myoblast, myotube, and skeletal muscle genes. Additional circadian related subsets were extracted from JTK analysis and literature reviews (core circadian), and additional cell cycle subsets were extracted from literature reviews (Table S3).

Statistical significance of TDS of genes
Given a set of genes, the significance test is made by comparing the average TDS of those genes with a random background distribution. The background distribution is generated by the average TDS of randomly selected gene sets (same size) over 1000 trials. The probability of the right-tailed event is used as P value.

Identification of MYOD/MYOG mediated oscillatory gene expression
Kallisto was used in RNA-seq quantification to obtain TPM (transcripts per million) expression results (Bray et al., 2016). BioCycle was used to identify oscillating transcripts after the 32 hr bifurcation point with a P value of 0.1 (Agostinelli et al., 2016). Transcripts found to be non-oscillatory before the bifurcation point were identified with a reported P value greater than 0.4. Phase, predicted through a neural network in BioCycle, was used to identify synchronous oscillating transcripts. Synchronous is defined as oscillating transcripts that are in-phase or antiphase within +/- 2 hours. MYOD1 and MYOG gene targets were found by identifying transcription factor binding sites for the respective motifs 10kb upstream or 1kb downstream of transcription start sites (TSS) using MotifMap with a Bayesian Branch Length Score > 1.0 and an FDR < 0.25 (Daily et al., 2011; Xie et al., 2009).

Super enhancer-promoter region dynamics
SE-P regions for skeletal muscles were downloaded from (Hnisz et al., 2013) (BI_Skeletal_Muscle). The Hi-C contacts between the SE and the associated gene TSS (±1kb) were extracted over time. SE-P contacts were normalized by dividing by the total number of contacts per sample, then multiplying by 100,000,000 (arbitrary scalar to best show trends). To determine the top upregulated genes, the linear regression slope of log2(FPKM) over time was calculated

and sorted for each gene, high to low. To determine significance, we first normalized the contacts by dividing by the total number of contacts for each SE-P region over time (so that all SE-P regions are on the same relative scale). We then performed a t-test between 16-24 hr and -48,0-8 hr normalized contacts.

DATA AND SOFTWARE AVAILABILITY

The dataset and codes will be reported when the paper is accepted.

REFERENCES

Agostinelli, F., Ceglia, N., Shahbaba, B., Sassone-Corsi, P. and Baldi, P. (2016), 'What time is it? deep learning approaches for circadian rhythms', *Bioinformatics* **32**, i8–i17.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016), 'Near-optimal probabilistic rna-seq quantification', *Nature biotechnology* **34**(5), 525–527.

Chen, H., Chen, J., Muir, L. A., Ronquist, S., Meixner, W., Ljungman, M., Ried, T., Smale, S. and Rajapakse, I. (2015), 'Functional organization of the human 4d nucleome', *Proceedings of the National Academy of Sciences* **112**(26), 8002–8007.

Chung, F. R. (1997), *Spectral graph theory*, Vol. 92, American Mathematical Soc.

Daily, K., Patel, V. R., Rigor, P., Xie, X. and Baldi, P. (2011), 'Motifmap: integrative genome-wide maps of regulatory motif sites for model species', *BMC bioinformatics* **12**(1), 495.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012), 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature* **485**(7398), 376–380.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A. and Young, R. A. (2013), 'Super-enhancers in the control of cell identity and disease', *Cell* **155**(4), 934–947.

Kimura, E., Han, J. J., Li, S., Fall, B., Ra, J., Haraguchi, M., Tapscott, S. J. and Chamberlain, J. S. (2008), 'Cell-lineage regulated myogenesis for dystrophin replacement: a novel therapeutic approach for treatment of muscular dystrophy', *Human molecular genetics* **17**(16), 2507–2517.

Knight, P. A. and Ruiz, D. (2013), 'A fast algorithm for matrix balancing', *IMA Journal of Numerical Analysis* **33**(3), 1029–1047.

Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with bowtie 2', *Nature methods* **9**(4), 357–359.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. and Dekker, J. (2009), 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science* **326**(5950), 289–293.

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. et al. (2014), 'A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell* **159**(7), 1665–1680.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010), 'edger: a bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics* **26**(1), 139–140.

Van Aelst, S. and Rousseeuw, P. (2009), 'Minimum volume ellipsoid', *Wiley Interdisciplinary Reviews: Computational Statistics* **1**(1), 71–82.

Van Der Maaten, L., Postma, E. and Van den Herik, J. (2009), 'Dimensionality reduction: a comparative', *J Mach Learn Res* **10**, 66–71.

Xie, X., Rigor, P. and Baldi, P. (2009), 'Motifmap: a human genome-wide map of candidate regulatory motif sites', *Bioinformatics* **25**(2), 167–174.