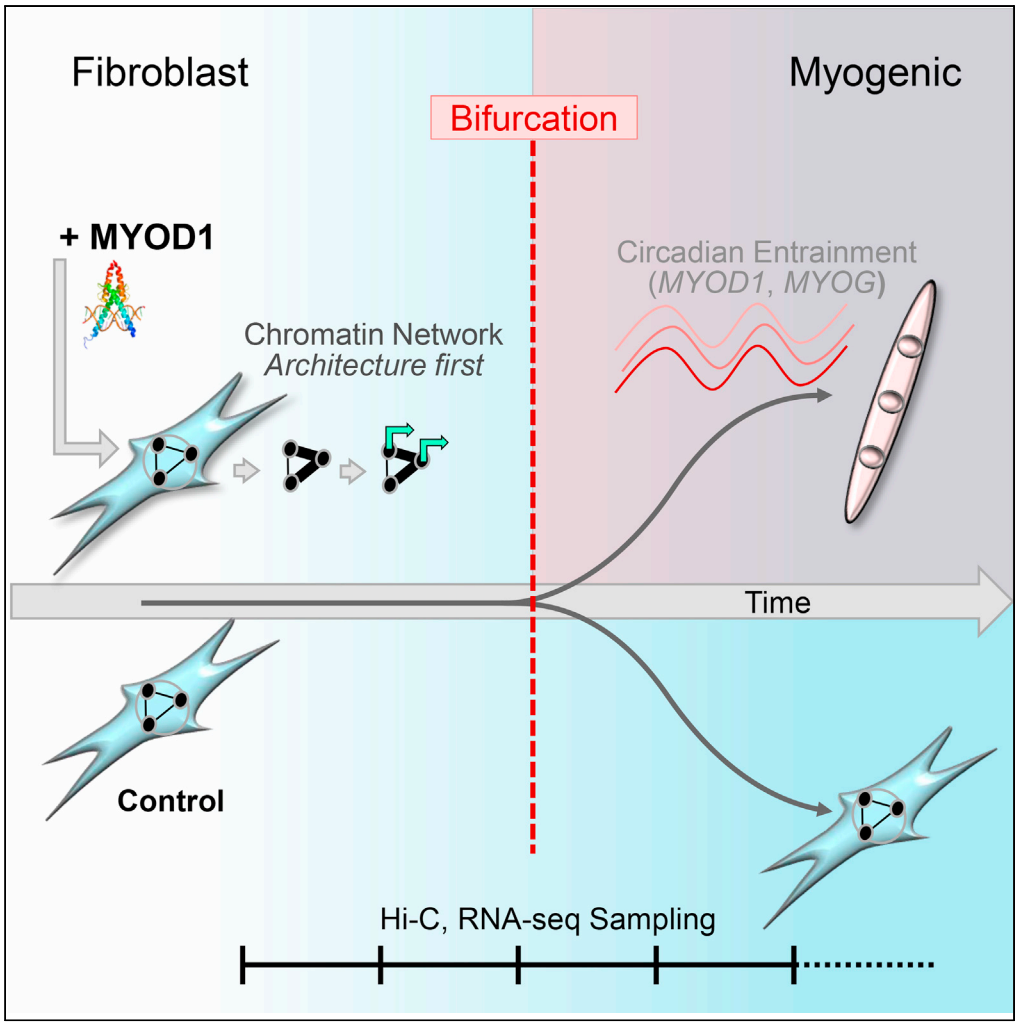
iScience

CellPress

Article

Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming



Sijia Liu, Haiming Chen, Scott Ronquist, ..., Alfred Hero, Lindsey A. Muir, Indika Rajapakse

indikar@umich.edu

HIGHLIGHTS

4D Nucleome analysis of direct human fibroblast to muscle reprogramming

A space-time bifurcation marks transit to a new cell identity

Chromatin reorganization precedes significant transcriptional changes

Myogenic master regulators have a role in entraining biological rhythms

Liu et al., iScience 6, 232–246
August 31, 2018 © 2018
https://doi.org/10.1016/j.isci.2018.08.002

iScience

CellPress

Article

Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming

Sijia Liu,[1,2,8] Haiming Chen,[1,8] Scott Ronquist,[1,8] Laura Seaman,[1] Nicholas Ceglia,[3] Walter Meixner,[1] Pin-Yu Chen,[4] Gerald Higgins,[1] Pierre Baldi,[3,8] Steve Smale,[5,6] Alfred Hero,[2] Lindsey A. Muir,[1] and Indika Rajapakse[1,7,9,*]

SUMMARY

Genome architecture has emerged as a critical element of transcriptional regulation, although its role in the control of cell identity is not well understood. Here we use transcription factor (TF)-mediated reprogramming to examine the interplay between genome architecture and transcriptional programs that transition cells into the myogenic identity. We recently developed new methods for evaluating the topological features of genome architecture based on network centrality. Through integrated analysis of these features of genome architecture and transcriptome dynamics during myogenic reprogramming of human fibroblasts we find that significant architectural reorganization precedes activation of a myogenic transcriptional program. This interplay sets the stage for a critical transition observed at several genomic scales reflecting definitive adoption of the myogenic phenotype. Subsequently, TFs within the myogenic transcriptional program participate in entrainment of biological rhythms. These findings reveal a role for topological features of genome architecture in the initiation of transcriptional programs during TF-mediated human cellular reprogramming.

INTRODUCTION

During cellular reprogramming, the mechanisms by which a small number of transcription factors (TF) (Takahashi et al., 2007), or a single TF as in Weintraub's work (Weintraub et al., 1989; Weintraub, 1993), impose new transcriptional programs that supersede established cell identities are not well understood. Unbiased technologies such as genome-wide chromosome conformation capture (Hi-C) and RNA sequencing (RNA-seq) are yielding ever higher resolution data that are essential to refining our notions of cell identity formation and maintenance. Yet these platforms are not well integrated analytically to understand the interplay between architecture and transcription. Furthermore, the dynamical nature of both architecture and transcription, during cellular reprogramming and natural biological rhythms, is challenging to capture experimentally and informatically and therefore is not well resolved. Thus the multi-platform genome-wide temporal capture of cells during reprogramming with integrated analytic approaches will be valuable for gaining insight into the mechanisms of reprogramming (Rajapakse and Groudine, 2011) and in line with the 4D Nucleome (4DN) movement (Chen et al., 2015; Dixon et al., 2015; Fortin and Hansen, 2015; Krijger et al., 2016).

Although genome architecture is a key element in transcriptional programs, its role in TF-mediated reprogramming is poorly understood, partially due to limited temporal data and analytic methods. During differentiation, architecturally defined regions can change their overall gene expression to facilitate a transcriptional program that supports a new cell state (Chen et al., 2015; Dixon et al., 2012, 2015; Lieberman-Aiden et al., 2009). These regions can be defined based on Hi-C contact maps into 2 major compartments: open, transcriptionally active chromatin, classically termed compartment A, or closed, transcriptionally inactive chromatin, termed compartment B (Chen et al., 2015; Lieberman-Aiden et al., 2009). Of critical importance in analyzing genome architecture data are sophisticated approaches to extract the most prominent, biologically relevant features. Our recent technique based on spectral graph theory extracts critical architectural information from Hi-C data, showing utility in defining chromatin domains at many scales (Chen et al., 2016).

In this work, we examined the dynamical interactions between the genome architectural features and transcription in human fibroblasts undergoing MYOD1-mediated reprogramming into the myogenic lineage. Sampling across a time course during reprogramming, we captured architecture by Hi-C, transcription by

[1]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA

[3]Department of Computer Science, University of California-Irvine, Irvine, CA 92697, USA

[4]AI Foundations, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

[5]Department of Mathematics, City University of Hong Kong, Hong Kong 999077, China

[6]Department of Mathematics, University of California, Berkeley, CA 94720, USA

[7]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA

[8]These authors contributed equally

[9]Lead Contact

*Correspondence: indikar@umich.edu

https://doi.org/10.1016/j.isci.2018.08.002

RNA-seq, and protein content by proteomics. To better understand the features of genome architecture and expression in a dynamical setting, we adopt a network point of view. Nodes of the network correspond to genomic loci that can be partitioned at different scales, for example, into larger scale 1-Mb regions or smaller scale gene-level regions. The edges of the network indicate contact between two genomic loci, with edge weights given by Hi-C entries. From the network perspective, A/B compartments are identified as distinct connected nodes of a network.

To further reveal chromatin spatial organization, we use network centrality measures. Using network centrality enables identification of nodes that play influential topological roles in the network (Newman, 2010). A number of centrality measures exist, each specialized to a particular type of nodal influence. For example, degree centrality characterizes the local connectedness of a node as measured by the number of edges connecting to this node, whereas betweenness centrality is a global connectedness measure that quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Eigenvector centrality is a neighborhood connectedness property in which a node has high centrality if many of its neighbors also have high centrality. In other words, a node is important if it is connected to other important nodes. For reference, Google's PageRank algorithm uses a variant of eigenvector centrality (Lohmann et al., 2010).

By examining different centrality measures we have discovered important features in Hi-C data largely overlooked in previous studies. We also found that cells undergoing reprogramming have significant architectural reorganization before changes in transcription, navigate through a critical transition point into the myogenic lineage, and subsequently show potent activation of the myogenic program that ties into regulation of biological rhythms.

RESULTS

Myogenic Reprogramming of Human Fibroblasts

We converted primary human fibroblasts into the myogenic lineage using the TF and master regulator MYOD1, following Weintraub's method for myogenic reprogramming (Weintraub, 1993) (Figure S1A). Fibroblasts were transduced with a lentiviral construct that expressed human MYOD1 fused with a tamoxifen-inducible ER(T) domain (L-MYOD1) (Kimura et al., 2008). With 4-hydroxytamoxifen (4-OHT) treatment, transduced cells showed nuclear translocation of L-MYOD1, morphological changes consistent with expression of key myogenic genes downstream of *MYOD1* (*MYOG* and *MYH1*) (Figure S1B), and myogenic differentiation (Figures S1C and S1D). These data demonstrate the conversion of fibroblasts into the myogenic lineage by L-MYOD1 (see Figure S7).

We used this system to delineate the dynamics of architecture and transcription underlying direct cellular reprogramming. Analyses were carried out on transduced, 4-OHT-treated cells, sampling at 8-hr intervals for RNA-seq (3 replicates per time point, small RNA-seq, and single replicate per time point, Hi-C; see Figure S8) and at 24-hr intervals for proteomics (Figure 1A).

We evaluated up to 16 time points (−48, 0, …, 112 hr) for genome architecture (form) through Hi-C and for transcription (function) through RNA-seq. The resulting time series data were studied at different scales (Figure 1B). The scale was based on units of length along the linear genome (1 Mb, 100 kb) or by structurally/functionally defined units of the genome, such as topologically associating domains (TADs) or individual genes.

Network Representation of Genomic Time Series Data

With the aid of network representation, we captured multiple topological properties of genome architecture using the concept of network centrality (Methods). For this analysis, we interpreted 100-kb-resolution Hi-C and RNA-seq data as measurements of dynamical networks, where Hi-C contact maps depict network topologies and RNA-seq data characterize the function of nodes (Figure 2A).

We found that network centrality measures such as degree centrality, eigenvector centrality, and betweenness centrality quantified different architectural features that reflect the importance of specific genomic loci within the network. Beyond the simplest measure, degree centrality, we found that eigenvector centrality identified architecturally defined regions of active/inactive gene expression (A/B compartments). In addition, across chromosomes, eigenvector centrality yielded a higher correlation with transcriptional

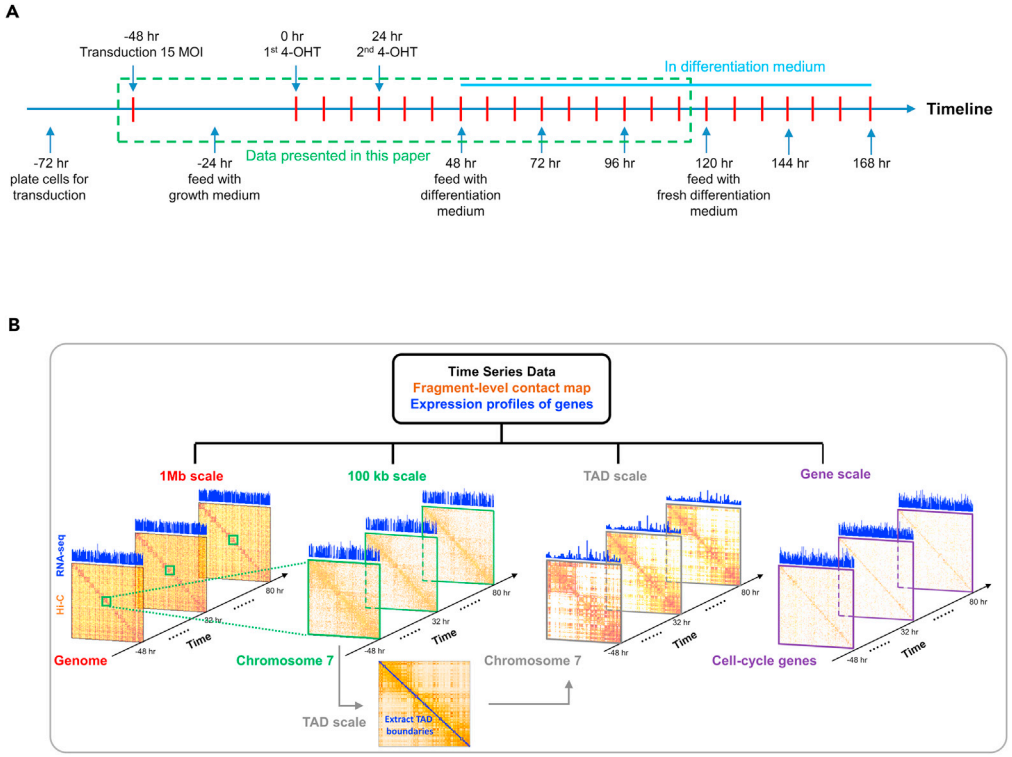


Figure 1. Myogenic Reprogramming of Human Fibroblasts

(A) Time course of MYOD1-mediated reprogramming. The time window outlined in green corresponds to time points at which both genome architecture and transcription were captured by Hi-C (single replicates) and RNA-seq (in triplicate).
(B) Scale-adaptive Hi-C matrices and gene expression. The considered scales include 1 Mb, 100 kb, TAD, and gene level.

activity than conventionally defined A/B compartments (Figures 2B and S2), which are derived from the first principal component of a spatial correlation Hi-C matrix (Lieberman-Aiden et al., 2009).

Betweenness centrality recognized regions that switched A/B compartment assignment between time steps. The values of betweenness at A/B switched regions were significantly higher than other centrality measures (Figure 2C). We then observed that A/B switched regions tended to be at boundaries between other A/B compartments and determined that this observation held for 70% of switched regions (Figure S3). Altogether, these results suggested that betweenness centrality detected boundary regions between open and closed chromatin that had a high propensity for altered architecture between time steps. We speculate that these regions, or "bridge nodes" in the network, serve as architectural buffers between largely active and inactive transcriptional chromatin that could limit access of transcriptional machinery to undesired regions.

We then sought to determine which genes showed differential expression within A/B switched regions. In A/B switched regions between 0 and 40 hr, we identified 175 genes (Table S1) that had at least 2-fold difference in expression (Methods). From this set, 47% of genes that change from compartment A to B had concordant gene expression (decrease), whereas 67% of genes that change from compartment B to A had concordant gene expression (increase).

Architectural Changes Precede Activation of the Myogenic Program

Given the cell state trajectory, it was unclear whether MYOD1-mediated reprogramming induced rewiring of genome architecture before the role of MYOD1 in mediating muscle gene transcription, or vice versa (Kosak and Groudine, 2004; Rajapakse and Groudine, 2011). To answer this question, we focused on form and function dynamics of 22,083 genes genome-wide, where the form is depicted by inter-gene Hi-C contact maps (Methods) and the function corresponds to RNA-seq Fragments Per Kilobase of

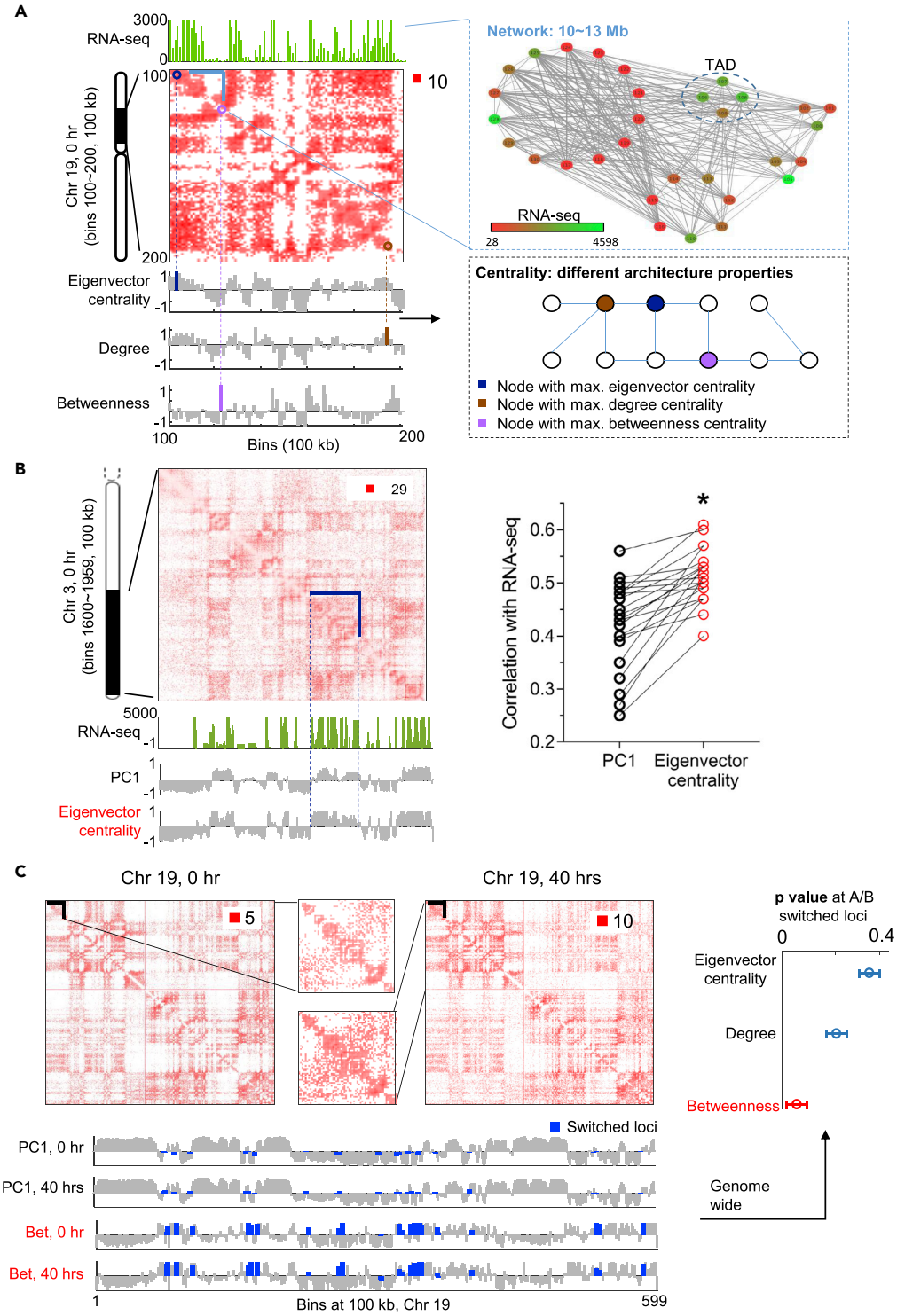


Figure 2. Network Representation of Genomic Time Series Data

(A) Mapping genomic form (Hi-C) and function (RNA-seq) to network architecture and node dynamics. *Top left*: Hi-C contact map (Toeplitz normalized, Methods) and RNA-seq at 100 kb resolution for chromosome 19. *Top right*: Network representation in which edge width indicates the Hi-C contact number and node color implies the magnitude of RNA-seq FPKM value. *Bottom left*: Network features given by eigenvector centrality, degree centrality, and betweenness centrality

Figure 2. *Continued*

scores. The bars marked by different colors correspond to maximum centrality values. *Bottom right*: An illustrative network under different centrality measures.

(B) Eigenvector centrality indicates chromatin compartments, termed A and B. *Top left*: Hi-C contact map of chromosome 3 at 100 kb resolution. *Bottom left*: RNA-seq, the first principal component (PC1) of the Hi-C correlation matrix, and eigenvector centrality (in terms of its *Z* score). *Right*: Correlation between RNA-seq, PC1, and eigenvector centrality extracted from Hi-C data for all chromosomes. Eigenvector centrality is a better indicator for chromatin compartments, marked by asterisk.

(C) Betweenness centrality indicates A/B switched loci. *Top left*: Hi-C contact map of chromosome 19 (100 kb resolution) at time points 0 and 40 hr. *Bottom left*: A/B partition and betweenness centrality (in terms of its *Z* score) at 0 and 40 hr. The blue color represents A/B switched bins from 0 to 40 hr. The switched loci tend to have large betweenness centrality scores. *Right*: Significance of betweenness centrality at A/B switched loci. The p value is determined by comparing the average betweenness value at A/B switched bins with a random background distribution of other centrality values under the same number of bins. p Values are computed for all chromosomes and shown through an error bar plot in which the circle represents the p value averaged over all chromosomes and the horizontal error bar is determined by the SD of p values for all chromosomes.

transcript per Million (FPKM) values (Figure 3A). The form-function evolution is then evaluated by determining the difference in network centrality features (extracted from inter-gene contact maps) and gene expression between successive time points. We refer to this measure as the temporal difference score (TDS; Methods). Based on TDS at successive time points (Figure 3B), we found that a significant form change at 8 hr preceded a significant function change at 16 hr.

For deeper understanding of form-function evolution during the reprogramming process, we applied K-means clustering (with 2 clusters) on both form and function data, separately. This was done to identify subsets of genes that yielded the most significant temporal change (Figures 3C and 3D). From this analysis, we found that genes contained within each cluster of high TDS, which are responsible for function and form change, at most have 20% overlap (Table S2). This suggests that the mechanism of form evolution could be different from that of function evolution and that these two mechanisms are steered by different sets of genes. Furthermore, we investigated 4 gene modules extracted from Gene Ontology (GO): fibroblast, myotube, cell cycle, and circadian genes (Table S3). We then contrasted our reprogramming data with data on human fibroblast proliferation. Data on proliferating human fibroblasts were previously obtained using similar methods over a time course (Chen et al., 2015) after cell cycle and circadian rhythm synchronization, with collection of RNA-seq and Hi-C every 8 hr. We found that the pattern of form-function evolution during reprogramming is quite different from fibroblast proliferation (Figure 3E). Consistent with findings represented in Figure 3B, the effects of nuclear reorganization were detectable before transcription changes, that is, form preceded function. Given these results, we propose that chromatin architectural changes facilitate the orchestrated activation of transcriptional networks associated with the adoption of a new cell identity.

We then sought to identify which genes may be responsible for these form-function dynamics. Within GO fibroblast and muscle gene modules, a significant proportion (>30%) of genes had form change at 8 hr and function change at 32–40 hr (Figure S4A). For comparison, less than 5% of these genes for each module showed similar form-function changes in fibroblast proliferation data. From these sets of genes, we extracted 77 fibroblast genes and 72 muscle genes that had significant change during reprogramming but low activity in proliferation. This yielded core or "backbone" genes, which had distinct form-function evolution during reprogramming (Figures S4B–S4D). The statistical significance of temporal change of the identified genes was $p < 0.05$ when compared with proliferation data (Methods).

Fibroblasts Navigate a Critical Transition En Route to the Myogenic Lineage

Genome dynamics during a direct transition between cell identities are poorly understood. We hypothesized that genome-wide data could be used to pinpoint a definitive transition. From our data we sought to identify the time at which cells transitioned into the myogenic state and which features of architecture and transcription define it. We therefore compared our reprogramming data with previously generated data on proliferating fibroblasts (Chen et al., 2015), where the time window of divergence between the datasets, or bifurcation, would indicate transition into a new cell identity.

To facilitate comparison, from each dataset we extracted a low-dimensional genome-wide form-function representation. This was done by first integrating centrality-based network features with transcription and then extracting the low-dimensional representation of the data using the dimension reduction

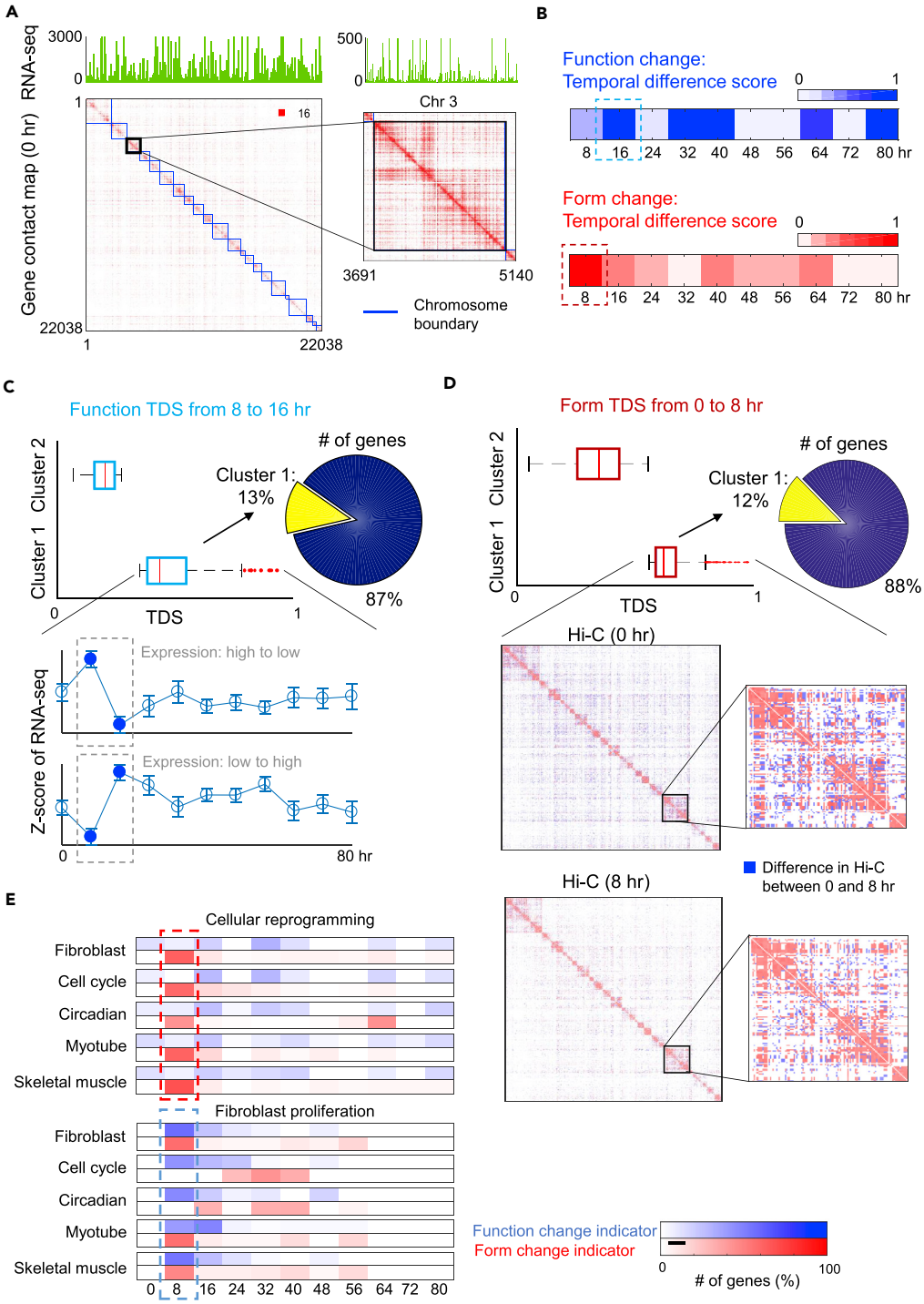


Figure 3. Changes in Genome Architecture Precede Activation of the Myogenic Program

(A) Genomic architecture (form) and gene expression (function) given by a Hi-C contact map and RNA-seq. Hi-C and RNA-seq are constructed at gene-level resolution.

(B) Function and form change at successive time points evaluated by temporal difference score (TDS; Methods) of RNA-seq and network centrality features of Hi-C data, respectively. The significant form change (at 8 hr) occurs before the function change (at 16 hr).

Figure 3. *Continued*

(C) Illustration of function TDS from 8 to 16 hr. Genes are divided into 2 clusters by applying K-means to their TDS values. Cluster 1 contains genes with the largest temporal change in RNA-seq. The gene expression can either decrease or increase from 8 to 16 hr.

(D) Illustration of form TDS from 0 to 8 hr. Two gene clusters are obtained by applying K-means to their TDS values. Hi-C contact maps associated with a subset of genes in cluster 1 are shown from 0 to 8 hr, where the blue color indicates the Hi-C difference between the 2 time points.

(E) Form-function change indicators for gene modules of interest during cellular reprogramming (top) and fibroblast proliferation (bottom), respectively. Here each row represents one gene module of interest, each column represents a time step, and the amount of change, as a percentage of total change over time for each module, is depicted by color. Percentage is determined by finding the number of genes with significant form-function change for each module and time step and dividing this number by the total number of significant gene changes for each module over time (row).

technique of Laplacian eigenmaps at 1 Mb resolution (Methods). Within each dataset, the form-function representation, fitted by a minimum volume ellipsoid (MVE) (Methods), showed distinct configurations at different time points (Figure 4A). Comparing between datasets, we observed a striking divergence, or bifurcation, at 32 hr ($p = 0.0048$), suggesting an abrupt shift in the genomic system during a transition from a fibroblast state to a myogenic state.

In examining local genome dynamics, distinct transitions were also observed for myogenic genes *MYOD1* and *MYOG*. Endogenous *MYOD1* and *MYOG* expressions were first detected around 32 hr. In addition, the transition was identified in intra-gene Hi-C contact maps of *MYOD1* and *MYOG* (Figure S5, Methods). Here the difference between gene-level Hi-C matrices at successive time points revealed a pattern strikingly similar to what was found in genome-wide dynamics. Taken together, our results are consistent with 3 phases for reprogramming from our data: fibroblast, bifurcation, and myogenic.

Phase Portraits Show Chromosome Architectural Changes Outpacing Transcriptional Changes

To further quantitate form-function dynamics on the chromosome level, we evaluated 2D phase portraits at 100 kb resolution. On a 2D plane, we designate one axis as a measure of form in terms of network connectivity (Methods) and the other as a measure of function in terms of average RNA-seq FPKM value. The portrait of 4DN is then described by a form-function domain, made up of 8 time points [0,56] (hr) for each chromosome (Figure 4B).

The portraits of 4DN for reprogramming and fibroblast proliferation showed similar positional patterns for chromosomes across time points. The centroid of the fitted form-function ellipsoid (MVE estimate; Methods) for each chromosome was shifted for reprogramming versus proliferation as illustrated for chromosomes 12, 5, and 13 (Figure 4C). Comparing the 32-hr critical transition point for chromosomes between datasets illustrates that the horizontal shift in form is greater than the vertical shift in function. Across time points, we found that most chromosomes undergo more form change (86.4%) than function change (13.8%) (Figure 4D). Furthermore, the area of the chromatin ellipse characterizes the variance (uncertainty) of 4DN (Figure 4D). Ellipsoids associated with reprogramming have larger volumes than those for fibroblast proliferation. Taken together, our results demonstrate a more complex dynamical behavior for reprogramming that is measurable through form-function dynamics, with notable involvement of genome architecture.

Fibroblasts Bypass a Myoblast-Like State during Myogenic Reprogramming

Intermediate stages of reprogramming are of consequence in the design of cell-based therapeutic strategies, as risks and benefits of cells that, for example, retain proliferative potential must be weighed. We therefore sought to further examine the pathway into the myogenic lineage, to determine whether the data support transit through a myoblast-like state or directly to a more differentiated myotube-like state. For this analysis, we considered TADs as functional units of the genome and identified those with significant form-function changes as playing important roles during reprogramming. Previous work showed that the boundaries of TADs remain stable between cell types (Dixon et al., 2012, 2015); however, the dynamical TAD-level interactions and functional changes during cellular reprogramming are not well understood. We interpreted the genome as a network of TADs (Figure S6), where network vertices corresponded to TADs, and edge weights were given by the interaction frequency between 2 TADs from Hi-C (retaining only interactions that exceeded the 50th percentile of inter-TAD contacts; see Methods). The function

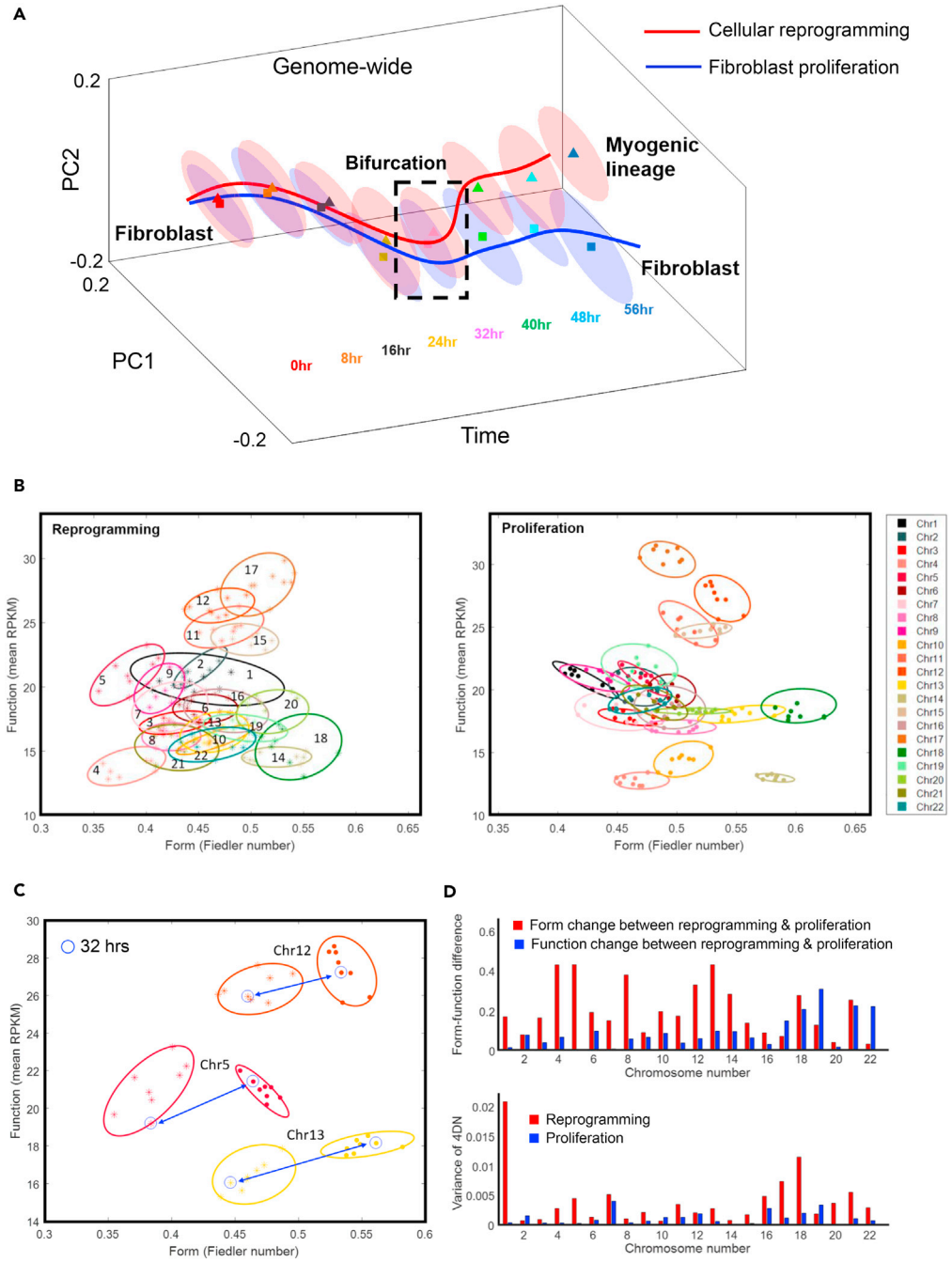


Figure 4. Fibroblasts Navigate a Critical Transition En Route to the Myogenic Lineage

(A) Cell state trajectory of MYOD1-mediated reprogramming and fibroblast proliferation (Chen et al., 2015). Ellipsoids represent low-dimensional data representations obtained by applying Laplacian eigenmaps (Methods) to network form-function features. The branching trajectory shows a critical transition, or bifurcation, at 32 hr ($p < 0.01$).

(B) Portrait of 4DN in the context of reprogramming and proliferation, respectively. It is described by a form-function domain (2D), constructed from 8 time points, for each chromosome. The fitted ellipsoid is obtained from the MVE estimate (Methods).

(C) Shift of form-function domains of chromosomes at 32 hr. Chromosomes 5, 12, and 13 show the most significant changes of all chromosomes.

(D) Form-function differences between cellular reprogramming and fibroblast proliferation, indicated by centroids and volumes of form-function ellipsoids for each chromosome. *Top*: Comparison between form change (horizontal shift) and function change (vertical shift) for each chromosome. *Bottom*: Variance of 4DN, given by volumes of chromosome ellipsoids under different cell dynamics.

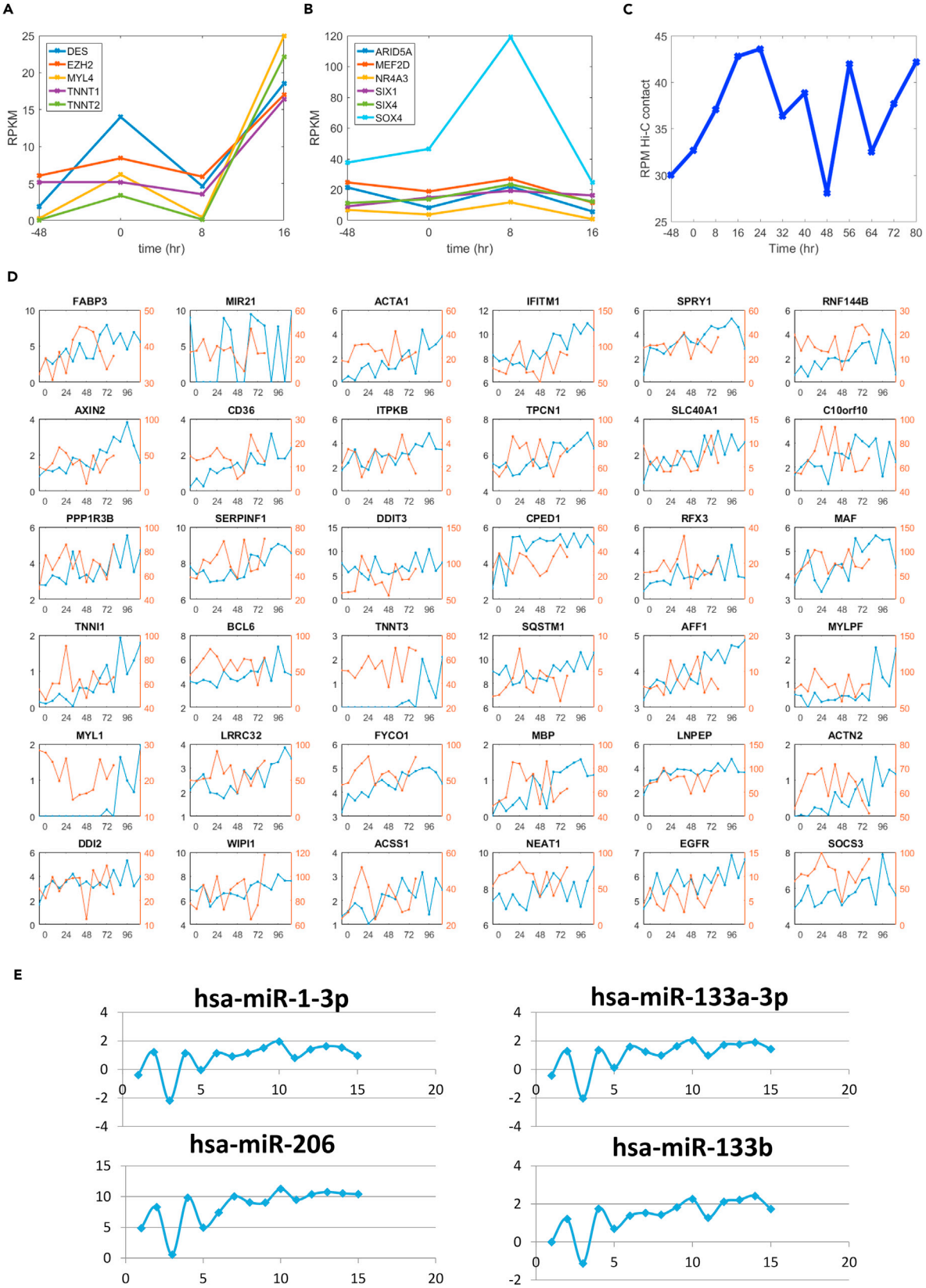

A
DES
EZH2
MYL4
TNNT1
TNNT2
RPKM
time (hr)
B
ARID5A
MEF2D
NR4A3
SIX1
SIX4
SOX4
RPKM
time (hr)
C
RPM Hi-C contact
Time (hr)
D
FABP3
MIR21
ACTA1
IFITM1
SPRY1
RNF144B
AXIN2
CD36
ITPKB
TPCN1
SLC40A1
C10orf10
PPP1R3B
SERPINF1
DDIT3
CPED1
RFX3
MAF
TNNI1
BCL6
TNNT3
SQSTM1
AFF1
MYLPF
MYL1
LRRC32
FYCO1
MBP
LNPEP
ACTN2
DDI2
WIPI1
ACSS1
NEAT1
EGFR
SOCS3
E
hsa-miR-1-3p
hsa-miR-133a-3p
hsa-miR-206
hsa-miR-133b

Figure 5. Increased Genomic Contacts among Myogenic Regulatory Elements Set the Stage for Reprogramming

(A) Early-phase expression dynamics of genes related to muscle cell terminal differentiation and chromatin remodeling. Genes encoding proteins involved in adult muscle function, including components of the contractile apparatus (DES, MYL4, TNNT1, TNN2), and EZH2, a repressor that is involved in myogenesis.

(B) Chromatin remodeling factors and master transcription factors act cooperatively with MYOD1 to drive proliferating human fibroblasts into muscle cells. These factors include ARID5A, part of the BAF47 muscle remodeling complex that acts in cooperation with MYOD1; MEF2D, which drives differentiation of myotubes to skeletal and cardiac muscle; NR4A3 (aka NOR1) involved in differentiation of myotubes into smooth muscle; and SIX1, SIX4, and SOX4, which control the differentiation of myotubes into muscle cells.

(C) Form and function of super enhancers and associated genes over time. Average Hi-C (read per million; RPM) contact between potential super enhancer and associated gene TSS regions over time, as defined by Hnisz et al. (2013).

(D) Top upregulated SE-P genes, $\log_2$(FPKM) (blue), and SE-P Hi-C normalized contact (red; see Methods) over time.

(E) Four muscle-specific miRNAs have significantly increased expression levels in the later time points relative to the baseline control. X axis, sampling time points; y axis, log-scale differences at other time points compared with baseline (−48 hr).

associated with a TAD was characterized by the sum of RNA-seq values of the set of genes contained within the TAD-defined region.

We applied network centrality analysis (Methods) to extract the 2D representation of dynamical form-function features at the TAD scale, using previously defined TAD boundaries (Figure S6A) (Dixon et al., 2012). The TAD-TAD network was constructed based on Hi-C matrices at 100 kb resolution, which facilitated the evaluation of whole genome characteristics. The resulting configuration of chromosomes was robust over time, but TADs within a chromosome showed form-function shifts. This can also be observed by contrasting the fibroblast stage (before 4-OHT; −48 hr) with the subsequent reprogramming time points (0, …, 80 hr) (Figure S6B).

We extracted the top 10% (220) of TADs whose positions change the most; these TADs are associated with the largest deviations from the fibroblast stage due to reprogramming (Figure S6C). We found that the identified TADs had high gene density and that genes within them are highly expressed ($p < 0.001$; see Methods). This implies that a core set of genes might exist within these TADs that induce significant form-function changes.

With this motivation, we focused on TADs containing genes related to fibroblasts and myogenesis to determine whether cells transitioned through a myoblast-like state. Gene sets were extracted from GO (Table S3), and for myogenesis included myoblast, myotube, and skeletal muscle. We found that TADs containing fibroblast or myotube genes had significant position shifts over time with $p = 0.0029$ or 0.0191, respectively (Figure S6D, and Methods). By contrast, the position shifts of TADs that contained myoblast genes were not statistically significant.

A direct pathway of reprogramming is further supported by the expression analysis of 3 myogenic regulatory factors: *MYF5*, *MYOD1*, and *MYOG* (Weintraub et al., 1991; Bentzinger et al., 2012). It is known from the hierarchy of TFs regulating progression through natural myogenic differentiation (Bentzinger et al., 2012) that *MYF5* is expressed in myoblasts, whereas *MYOD1* and *MYOG* are upregulated in myotubes. In our data, *MYF5* was not activated during reprogramming, whereas *MYOD1* and *MYOG* were expressed after the 32-hr critical transition point (Figures S1F and S6E).

Early-Stage Chromatin Remodeling and microRNA Dynamics

We additionally sought to understand the regulatory dynamics during reprogramming, including early-stage gene expression dynamics related to chromatin remodeling, super enhancer dynamics, and microRNA (miRNA) expression. Examination of early-stage RNA-seq data [-48, 16] (hr) revealed endogenous mechanisms relevant to *MYOD1* transcriptional activation including muscle stage-specific markers and chromatin remodeling factors (see Figure 5A). At 16 hr, the combined upregulation of *DES*, *MYL4*, *TNNT1*, and *TNNT2* suggests myogenic differentiation (Gard and Lazarides, 1980; Schiaffino et al., 2015). *EZH2* has been associated with both "safeguarding" the transcriptional identity of skeletal muscle stem cells and terminal differentiation of myoblasts into mature muscle (Juan et al., 2011). *ARID5A*, a regulator of the myotube BAF47 chromatin remodeling complex, is significantly upregulated at 8 hr ($p = 7.2 \times 10^{-5}$) and may act to enhance MYOD1 binding to target promoters (Joliot et al., 2014). *NR4A3*, *MEF2D*, *SIX4*, *SIX1*, and *SOX4* expression are also increased at 8 hr, all of which have important regulatory functions during differentiation in the myogenic lineage (see Figure 5B) (Ferrán et al., 2016; Bentzinger et al., 2012; Jang et al., 2015).

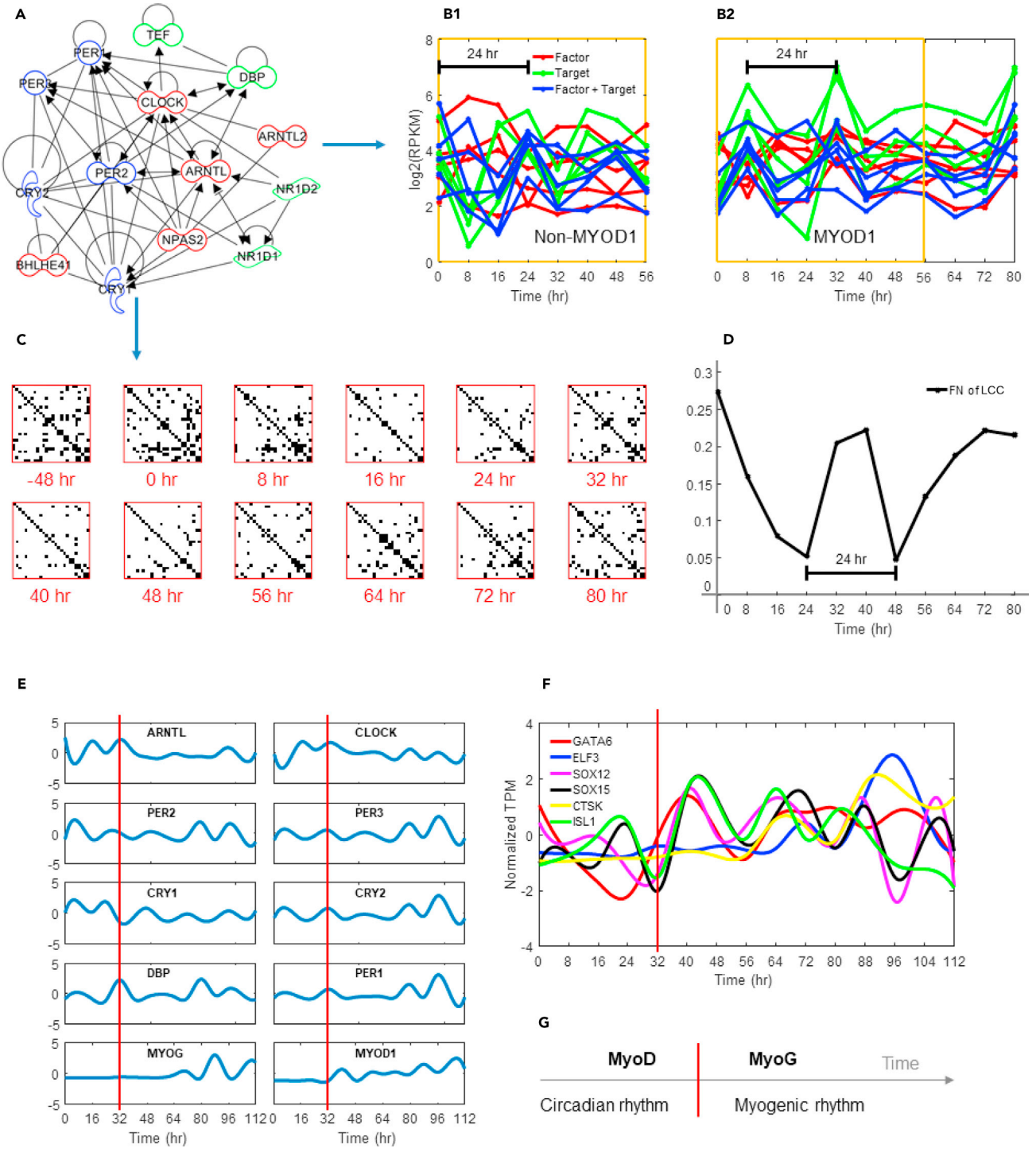


Figure 6. Myogenic Genes Participate in Entrainment of Biological Rhythms

(A) Gene network interactions between circadian E-box genes, derived from Ingenuity Pathway Analysis.

(B) Core circadian gene expression over time. (B1) Dexamethasone synchronization. (B2) L-MYOD1 synchronization. Target and factor correspond to genes with E-box targets and TFs that bind to E-box genes, respectively.

(C) Hi-C contacts between 26 core circadian genes over time (see Table S3). Rows and columns correspond to core circadian genes; contacts are binary (i.e., any contact between genes at a given time are shown).

(D) Network connectivity of the largest connected component (LCC; Methods) of the studied Hi-C contact maps at different time points.

Figure 6. *Continued*

(E) Normalized gene expression (FPKM, cubic spline) highlighting oscillation dampening after the bifurcation time 32 hr (red line) and the switch to differentiation medium for select core circadian genes; MYOD1 and MYOG also shown.

(F) Normalized transcripts per million (TPM) of transcription factors that are targeted by *MYOG* or *MYOD1* (*ELF3*) and that only showed oscillation after the critical transition at 32 hr (red line).

(G) Conceptual diagram of biological rhythm entrainment during MYOD1-mediated reprogramming, where the red line signifies the bifurcation event.

We also investigated how muscle-related super enhancer-promoter (SE-P) interactions change over time throughout MYOD1-mediated reprogramming. To capture these dynamics, we extracted the Hi-C contact between skeletal muscle super enhancer regions and the associated genes' transcription start sites (±1 kb), as determined by Hnisz et al. (2013) (618 SE-P regions; Methods). We observed that for these skeletal muscle SE-P Hi-C regions, the strongest amount of contact occurred relatively early in the reprogramming process, peaking 16–24 hr post-L-MYOD1 addition to the nucleus ($p = 4.17 \times 10^{-9}$, Figure 5C; Methods). Exact SE-P contact versus function trends were variable, but a number of important myogenesis genes, such as *TNNI1*, *MYLPF*, *ACTN2*, and *TNNT3*, show strong upregulation in function over time, with an increase in SE-P contact post-MYOD1 activation. Contact versus function trends for the top 36 upregulated genes are shown in Figure 5D (Methods).

We measured the abundance of 2,588 miRNA species with reads from small RNA-seq. Using the edgeR software (Robinson et al., 2010) for data analysis, we identified 266 miRNA species that were significantly up- or downregulated in expression levels over the time course relative to the baseline control (false discovery rate [FDR] <0.05) (Table S5). Among these significant miRNAs, miR-1-3p, miR-133a-3p, and miR-206 have been previously identified as myogenic factor-regulated, muscle-specific species (McCarthy, 2011; Rao et al., 2006). We observed that the 3 miRNAs, plus miR-133b (FDR = 0.09), significantly increased in expression levels after 4-OHT treatment (Figure 5E). Their expression patterns were highly similar to that observed in mouse C2C12 cell differentiation (Rao et al., 2006). The observation of muscle-specific miRNAs, particularly miR-206, which had 1,000-fold greater expression at later time points than baseline (Figure 5E), further supports MYOD1-mediated reprogramming of fibroblasts to myotubes. Notably, the cardiac-specific species miR-1-5p, miR-208a, and miR-208b (McCarthy, 2011) were not detected in our samples.

Linking Myogenic Genes with Entrainment of Biological Rhythms

A number of studies have explored the link between MYOD1 and circadian genes ARNTL and CLOCK, revealing that ARNTL and CLOCK bind to the core enhancer of the *MYOD1* promoter and subsequently induce rhythmic expression of *MYOD1* (Andrews et al., 2010; Zhang et al., 2012). Here we discovered that upon *MYOD1* activation, circadian genes exhibited robust synchronization in gene expression, suggesting MYOD1 feedback onto the circadian gene network. Further inspection showed that core circadian genes (Table S3) that contain E-boxes displayed the most profound synchronization initially, starting with an uptick in gene expression just after MYOD1 activation (Figures 6A–6D). Analysis using JTK_CYCLE (Hughes et al., 2010) confirmed our observation; all E-box circadian genes were found to have a synchronized period of 24 hr, with a maximum lag of 4 hr between genes, with the exception of *CRY1* (Table S6).

Consistent with a critical transition point, the subset of transcripts with oscillatory behavior was different before and after the 32 hr time point. Endogenous *MYOD1* and *MYOG* expression began close to 32 hr, and both transcripts displayed oscillatory expression. In addition, circadian transcript oscillations dampened at 40 hr, coinciding with the switch to low-serum differentiation medium (Figure 6E). To determine which newly oscillating transcripts were potential targets of MYOD1 and MYOG, we further investigated which transcripts have MYOD1 or MYOG binding motifs in their promoters using MotifMap (Daily et al., 2011), and which were synchronized in expression with MYOD1 and MYOG. Among the oscillating transcripts that fit these criteria, we found 6 TFs that were oscillatory only after the 32-hr critical transition point, have upstream MYOG binding sites, and were synchronized in expression with MYOG. Of these 6 TFs, only *ELF3* was found to have binding motifs for *MYOD1*, as well as synchronized expression with *MYOD1* (Figure 6F). Several of the 6 oscillatory TFs targeted by MYOG or MYOD1 are associated with muscle developmental and differentiation processes, including SOX15 (Meeson et al., 2007), GATA6 (Xie et al., 2015), ISL1 (Pacheco-Leyva et al., 2016), and ELF3 (Böck et al., 2014).

Robust synchronization in the expression of circadian genes that are downstream targets of MYOD1 suggests MYOD1 feedback onto circadian gene circuits. After the 32-hr critical transition point, MYOG was associated with synchronized expression of a subset of important myogenic TFs. These findings support

regulatory roles for MYOD1 and MYOG in entraining circadian and cell type-specific biological rhythms (Figure 6G).

DISCUSSION

In this study, we analyzed MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage from a dynamical network perspective. Distinct from previous studies, we generated an enriched time series dataset including Hi-C, RNA-seq, miRNA, and proteomics data. This provides a comprehensive genome-wide form-function description over time and allows us to detect early-stage cell fate commitment changes during cellular reprogramming. We found both global and local phenomena supporting a critical transition point between cell identities during reprogramming. Capturing these dynamics may help us identify genes that are key players in other reprogramming settings and develop a more universal understanding of the process and requirements for reprogramming between any two cell types.

Our data further suggest a direct pathway of reprogramming from fibroblasts to myotubes that bypasses a myoblast intermediate and is associated with the expression of *MYOD1* and *MYOG*, but not *MYF5*. Related results have been described in studies on control of the cell cycle during muscle development, in which *MYOD1* and *MYF5* are involved in the determination of myogenic cell fate, with a switch from *MYF5* to *MYOG* during muscle cell differentiation (Singh and Dilworth, 2013; Zeng et al., 2016). Moreover, it has been theorized (Del Vecchio et al., 2017) that a reprogrammed biosystem with positive perturbation (i.e., overexpression of one or more specific TFs like MYOD1) would bypass the intermediate state and move directly toward the terminally differentiated state. This claim is consistent with our finding, where the intermediate and terminally differentiated states correspond to myoblast and myotube stages, respectively. Understanding the intermediates of direct reprogramming will be important in the design of potential therapeutics, as their properties must be fully evaluated to understand the risk and efficiency of reprogramming, and to optimize the scalability of cell number, taking into account the proliferative capacity at different stages.

A number of studies have explored the link between *MYOD1* and circadian genes *ARNTL* and *CLOCK*, revealing that ARNTL and CLOCK bind to the core enhancer of the *MYOD1* promoter and subsequently induce rhythmic expression of *MYOD1* (Andrews et al., 2010; Zhang et al., 2012). We found that upon activation of L-MYOD1, the population of cells exhibits robust synchronization in circadian E-box gene expression. Among these E-box targets are the *PER* and *CRY* gene family, whose protein products are known to repress CLOCK-ARNTL function, thus repressing their own transcription. In addition, E-box target NR1D1, which is synchronized upon addition of L-MYOD1, competes with ROR proteins to repress *ARNTL* transcription directly. This adds another gene network connection under MYOD1 influence, indirectly acting to repress *ARNTL*, leading us to posit that MYOD1 can affect CLOCK-ARNTL function through E-Box elements, in addition to CLOCK-ARNTL's established activation effect on MYOD1. Furthermore, these oscillations dampen just after the 32-hr critical transition point, after which MYOG entrains the oscillations of a distinct subset of myogenic TFs. Therefore, MYOD1-mediated reprogramming and circadian synchronization are mutually coupled, consistent with other systems that modulate cell fate (Umemura et al., 2014; Dierickx et al., 2018).

Our proposed biological and computational technologies shed light on the hypothesis that nuclear reorganization occurs at the time of cell specification and both precedes and facilitates activation of the transcriptional program associated with differentiation (or reprogramming), i.e., form precedes function (Rajapakse and Groudine, 2011). The alternative hypothesis is that function precedes form, that is, nuclear reorganization occurs as a consequence of differential transcription and is a consequence of, rather than a regulator of, differentiation programs (Kosak and Groudine, 2004). Our findings support that nuclear reorganization occurs before gene transcription during cellular reprogramming, i.e., form precedes function, and that dynamical nuclear reorganization plays a key role in defining cell identity. However, our data do not establish a causal relationship, and for this, additional experiments will be necessary. For example, Hi-C and RNA-seq can be supplemented using MYOD1 chromatin immunoprecipitation sequencing to identify the regions of greatest adjacency differences between cell types that correlate with transcription and/or MYOD1 binding.

As demonstrated by our study, network centrality-based analysis allows us to study Hi-C architecture from multiple views and facilitates quantitative integration with gene expression. Accordingly, the detailed connections between network architecture and network function in the context of the genome can be used to

probe genomic reorganization during normal and abnormal cell differentiation. It will also be helpful to determine whether nuclear architectural remodeling can be both temporally and molecularly separated from transcriptional regulation.

Understanding the dynamical process of cellular reprogramming is critical in regenerative medicine to improve our ability to guide cells toward repair and regeneration of tissue in injury and disease. Furthermore, identifying an architectural function for TFs that is distinct from transcription would define a new molecular function with an as yet unknown role in development and disease.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, eight figures, and eight tables and can be found with this article online at https://doi.org/10.1016/j.isci.2018.08.002.

ACKNOWLEDGMENTS

We thank the University of Michigan Sequencing Core, and especially Jeanne Geskes, for assistance. We thank John Hogenesch for helpful discussions on circadian rhythms. We thank Daniel Burns and Stephen Lindsly for critical reading of the manuscript and helpful discussions. We extend special thanks to James Gimlett and Srikanta Kumar at Defense Advanced Research Projects Agency (DARPA) for support and encouragement. This work is supported in part by the DARPA Biochronicity, Deep-Purple, and FunCC Programs, and the Smale Institute. We also acknowledge the seminal work of Mark Groudine and late Hal Weintraub, whose ideas continue to guide our thinking.

AUTHOR CONTRIBUTIONS

All authors including N.C.,P. B., and S.S. participated in the discussion of the results. S.L., H.C., S.R., L.A.M., and I.R. prepared the manuscript with input from all authors. N.C. and P. B. performed computational analyses and interpreted the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 19, 2018
Revised: June 23, 2018
Accepted: July 31, 2018
Published: August 31, 2018

REFERENCES

Andrews, J.L., Zhang, X., McCarthy, J.J., McDearmon, E.L., Hornberger, T.A., Russell, B., Campbell, K.S., Arbogast, S., Reid, M.B., Walker, J.R., et al. (2010). CLOCK and BMAL1 regulate MyoD and are necessary for maintenance of skeletal muscle phenotype and function. Proc. Natl. Acad. Sci. USA *107*, 19090–19095.

Bentzinger, C.F., Wang, Y.X., and Rudnicki, M.A. (2012). Building muscle: molecular regulation of myogenesis. Cold Spring Harb. Perspect. Biol. *4*, 1–16.

Böck, M., Hinley, J., Schmitt, C., Wahlicht, T., Kramer, S., and Southgate, J. (2014). Identification of ELF3 as an early transcriptional regulator of human urothelium. Dev. Biol. *386*, 321–330.

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. Proc. Natl. Acad. Sci. USA *112*, 8002–8007.

Chen, J., Hero, A.O., and Rajapakse, I. (2016). Spectral identification of topological domains. Bioinformatics *32*, 2151–2158.

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*, 495.

Del Vecchio, D., Abdallah, H., Qian, Y., and Collins, J.J. (2017). A blueprint for a synthetic genetic feedback controller to reprogram cell fate. Cell Syst. *4*, 109–120.e11.

Dierickx, P., Van Laake, L.W., and Geijsen, N. (2018). Circadian clocks: from stem cells to tissue homeostasis and regeneration. EMBO Rep. *19*, 18–28.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

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*, 376–380.

Ferrán, B., Martí-Pàmies, I., Alonso, J., Rodríguez-Calvo, R., Aguiló, S., Vidal, F., Rodríguez, C., and Martínez-González, J. (2016). The nuclear receptor NOR-1 regulates the small muscle

protein, X-linked (SMPX) and myotube differentiation. Sci. Rep. *6*, 1–11.

Fortin, J.-P., and Hansen, K.D. (2015). Reconstructing a/b compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. *16*, 180.

Gard, D.L., and Lazarides, E. (1980). The synthesis and distribution of desmin and vimentin during myogenesis in vitro. Cell *19*, 263–275.

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*, 934–947.

Hughes, M.E., Hogenesch, J.B., and Kornacker, K. (2010). Jtk_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J. Biol. Rhythms *25*, 372–380.

Jang, S., Kim, J., Kim, C., An, J., Johnson, A., Song, P., Rhee, S., and Choi, K. (2015). KAT5-mediated SOX4 acetylation orchestrates chromatin remodeling during myoblast differentiation. Cell Death Dis. *6*, 1–11.

Joliot, V., Ait-Mohamed, O., Battisti, V., Pontis, J., Philipot, O., Robin, P., Ito, H., and Ait-Si-Ali, S. (2014). The SWI/SNF subunit/tumor suppressor BAF47/INI1 is essential in cell cycle arrest upon skeletal muscle terminal differentiation. PloS One *9*, 1–11.

Juan, A.H., Derfoul, A., Feng, X., Ryall, J.G., Dell'Orso, S., Pasut, A., Zare, H., Simone, J.M., Rudnicki, M.A., and Sartorelli, V. (2011). Polycomb ezh2 controls self-renewal and safeguards the transcriptional identity of skeletal muscle stem cells. Genes Dev. *25*, 789–794.

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. Hum. Mol. Genet. *17*, 2507–2517.

Kosak, S.T., and Groudine, M. (2004). Form follows function: the genomic organization of cellular differentiation. Genes Dev. *18*, 1371–1384.

Krijger, P.H.L., Di Stefano, B., de Wit, E., Limone, F., van Oevelen, C., de Laat, W., and Graf, T. (2016). Cell-of-origin-specific 3d genome structure acquired during somatic cell reprogramming. Cell Stem Cell *18*, 597–610.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lohmann, G., Margulies, D.S., Horstmann, A., Pleger, B., Lepsien, J., Goldhahn, D., Schloegl, H., Stumvoll, M., Villringer, A., and Turner, R. (2010). Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. PLoS One *5*, 1–8.

McCarthy, J.J. (2011). The MyomiR network in skeletal muscle plasticity. Exerc. Sport Sci. Rev. *39*, 150.

Meeson, A.P., Shi, X., Alexander, M.S., Williams, R., Allen, R.E., Jiang, N., Adham, I.M., Goetsch, S.C., Hammer, R.E., and Garry, D.J. (2007). Sox15 and fhl3 transcriptionally coactivate Foxk1 and regulate myogenic progenitor cells. EMBO J. *26*, 1902–1912.

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

Pacheco-Leyva, I., Matias, A.C., Oliveira, D.V., Santos, J.M., Nascimento, R., Guerreiro, E., Michell, A.C., van De Vrugt, A.M., Machado-Oliveira, G., Ferreira, G., et al. (2016). CITED2 cooperates with isl1 and promotes cardiac differentiation of mouse embryonic stem cells. Stem Cell Rep. *7*, 1037–1049.

Rajapakse, I., and Groudine, M. (2011). On emerging nuclear order. J. Cell Biol. *192*, 711–721.

Rao, P.K., Kumar, R.M., Farkhondeh, M., Baskerville, S., and Lodish, H.F. (2006). Myogenic factors that regulate expression of muscle-specific microRNAs. Proc. Natl. Acad. Sci. USA *103*, 8721–8726.

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*, 139–140.

Schiaffino, S., Rossi, A.C., Smerdu, V., Leinwand, L.A., and Reggiani, C. (2015). Developmental myosins: expression patterns and functional significance. Skelet. Muscle *5*, 22.

Singh, K., and Dilworth, F.J. (2013). Differential modulation of cell cycle progression distinguishes members of the myogenic regulatory factor family of transcription factors. FEBS J. *280*, 3991–4003.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Umemura, Y., Koike, N., Matsumoto, T., Yoo, S.-H., Chen, Z., Yasuhara, N., Takahashi, J.S., and Yagita, K. (2014). Transcriptional program of Kpna2/importin-α2 regulates cellular differentiation-coupled circadian clock development in mammalian cells. Proc. Natl. Acad. Sci. USA *111*, 5039–5048.

Weintraub, H. (1993). The MyoD family and myogenesis: redundancy, networks, and thresholds. Cell *75*, 1241–1244.

Weintraub, H., Davis, R., Tapscott, S., Thayer, M., Krause, M., Benezra, R., Blackwell, T.K., Turner, D., Rupp, R., Hollenberg, S., et al. (1991). The myoD gene family: nodal point during specification of the muscle cell lineage. Science *251*, 761–766.

Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B., and Miller, A.D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. Proc. Natl. Acad. Sci. USA *86*, 5434–5438.

Xie, Y., Jin, Y., Merenick, B.L., Ding, M., Fetalvero, K.M., Wagner, R.J., Mai, A., Gleim, S., Tucker, D., Birnbaum, M.J., et al. (2015). Phosphorylation of gata-6 is required for vascular smooth muscle cell differentiation after mTORC1 inhibition. Sci. Signal. *8*, 1–27.

Zeng, W., Jiang, S., Kong, X., El-Ali, N., Ball, A.R., Christopher, I., Ma, H., Hashimoto, N., Yokomori, K., and Mortazavi, A. (2016). Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. Nucleic Acids Res. *44*, e158.

Zhang, X., Patel, S.P., McCarthy, J.J., Rabchevsky, A.G., Goldhamer, D.J., and Esser, K.A. (2012). A non-canonical e-box within the MyoD core enhancer is necessary for circadian expression in skeletal muscle. Nucleic Acids Res. *40*, 3419–3430.

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

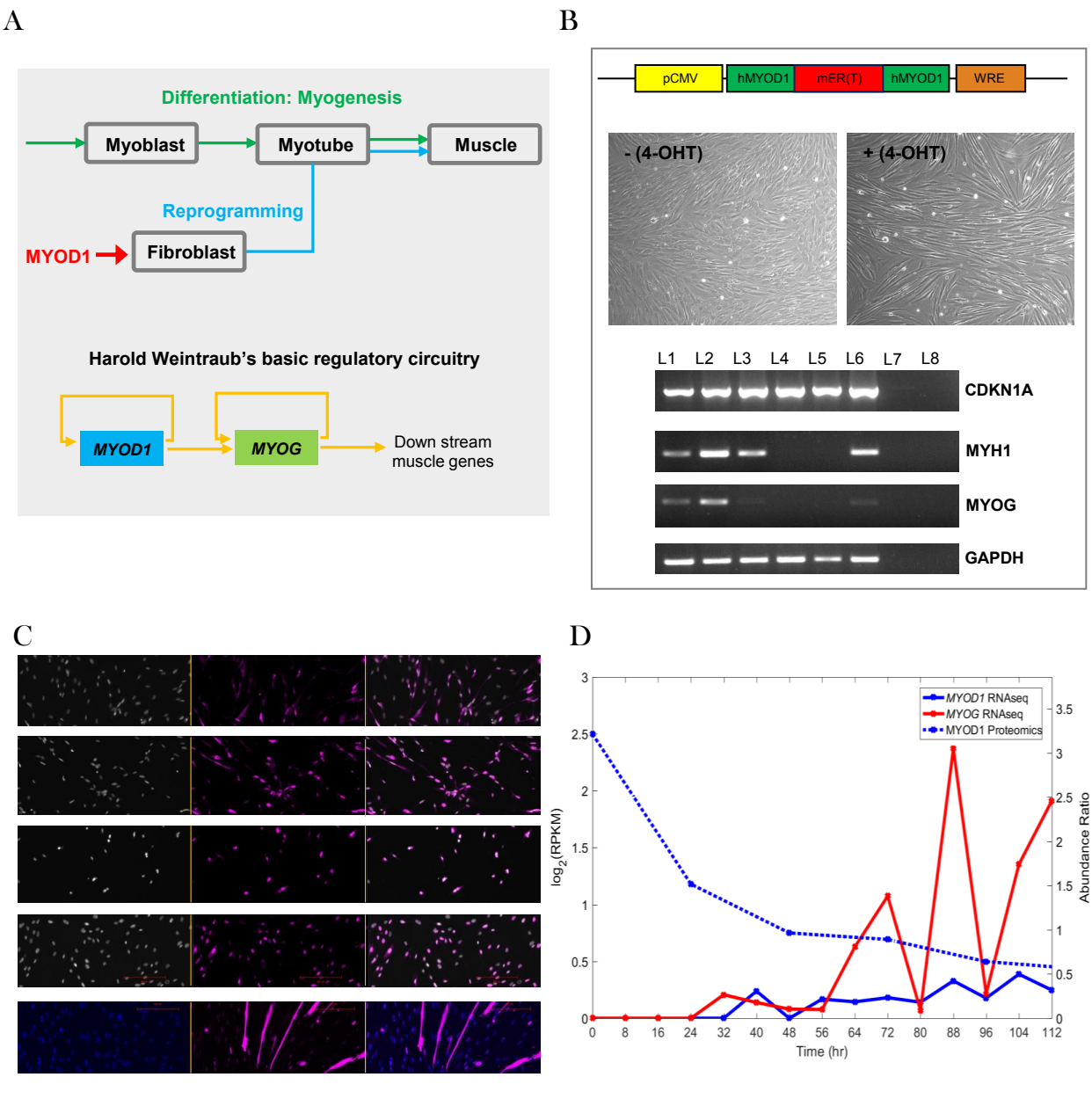

A
Differentiation: Myogenesis
Myoblast
Myotube
Muscle
Reprogramming
MYOD1
Fibroblast
Harold Weintraub's basic regulatory circuitry
MYOD1
MYOG
Down stream muscle genes
B
pCMV
hMYOD1
mER(T)
hMYOD1
WRE
- (4-OHT)
+ (4-OHT)
L1 L2 L3 L4 L5 L6 L7 L8
CDKN1A
MYH1
MYOG
GAPDH
C
D
MYOD1 RNAseq
MYOG RNAseq
MYOD1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

Figure S1

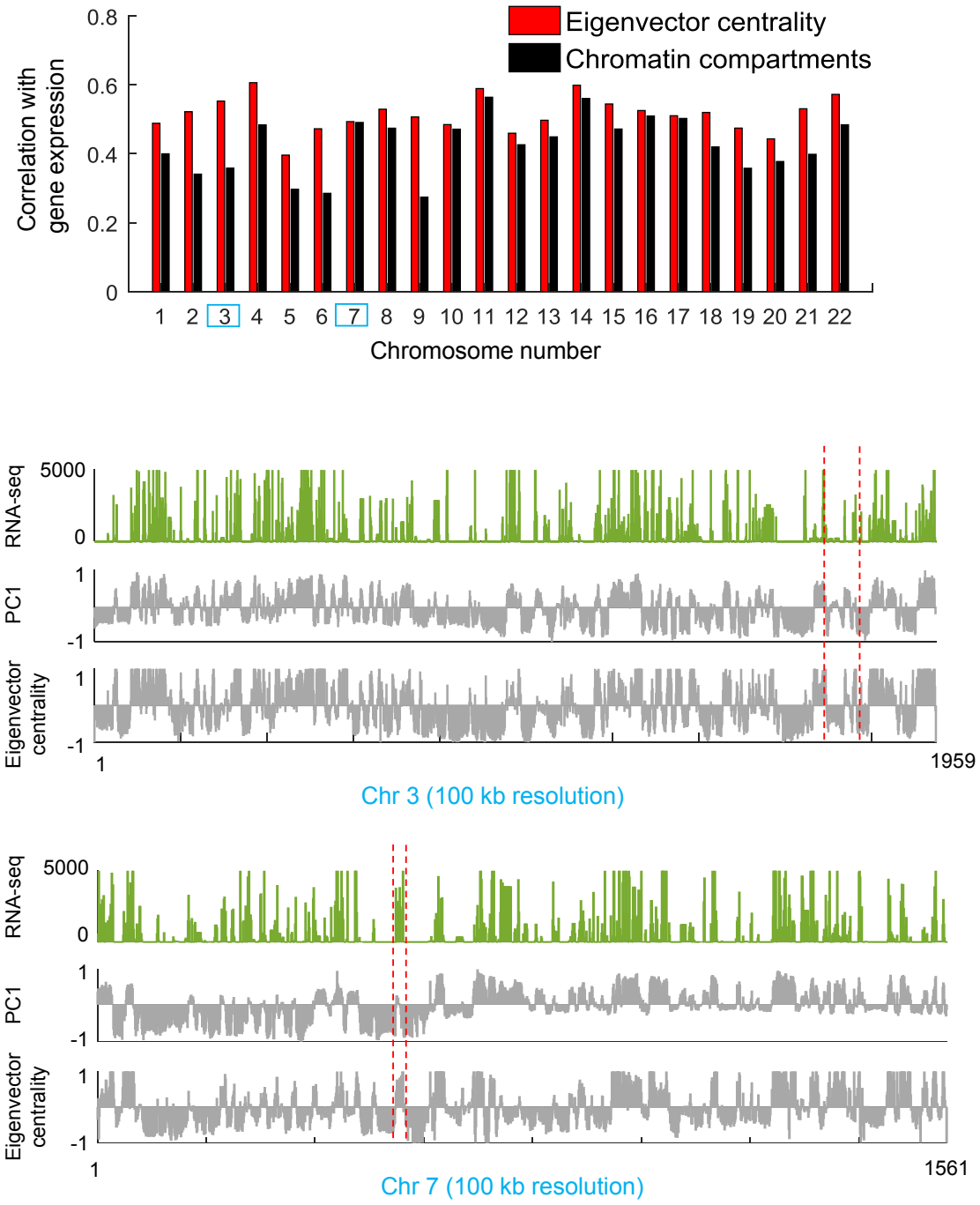

Eigenvector centrality
Chromatin compartments
Correlation with gene expression
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome number
RNA-seq
5000
0
PC1
1
-1
Eigenvector centrality
1
-1
1
1959
Chr 3 (100 kb resolution)
RNA-seq
5000
0
PC1
1
-1
Eigenvector centrality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

Figure S2

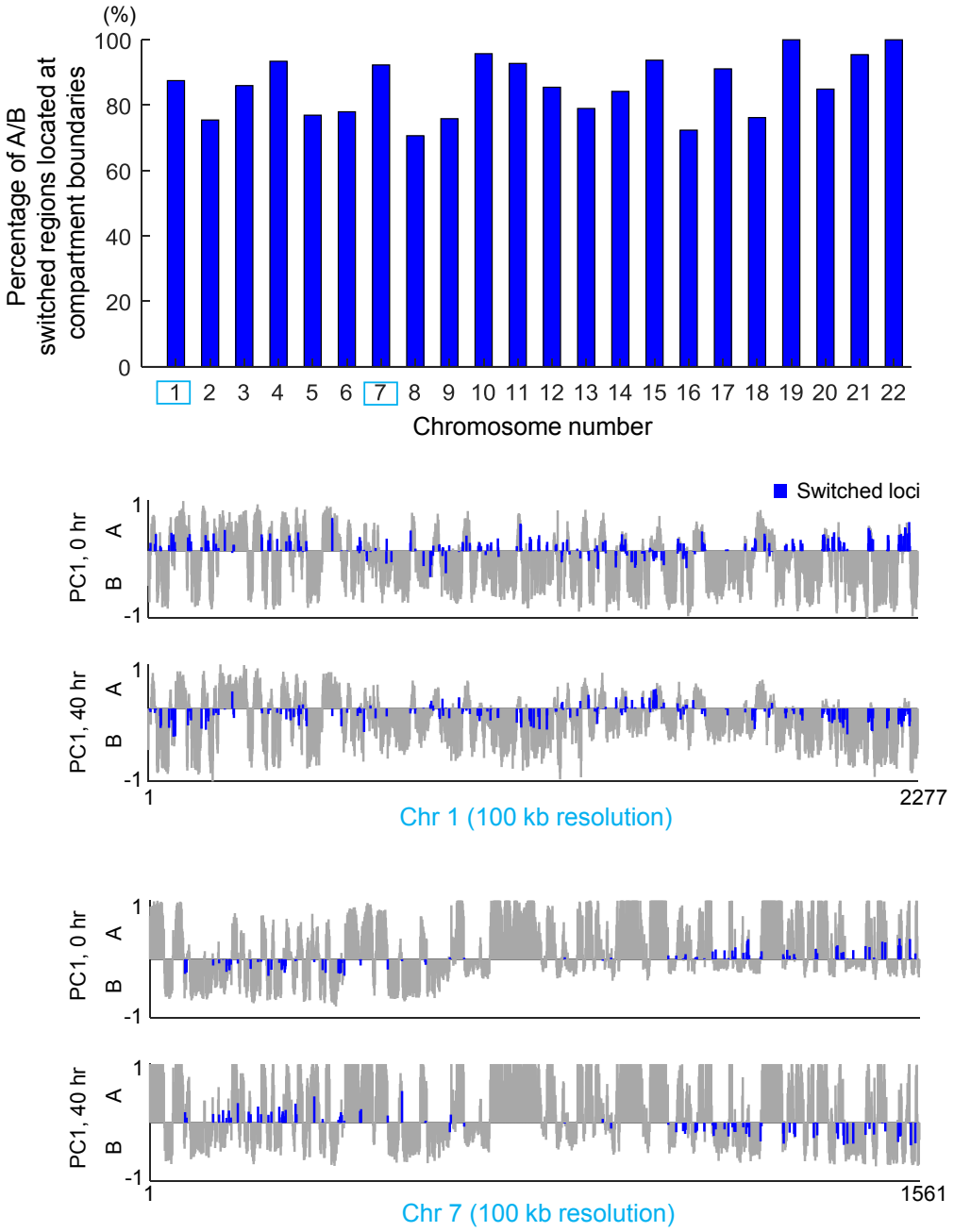

(%)
Percentage of A/B switched regions located at compartment boundaries
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome number
Switched loci
PC1, 0 hr
PC1, 40 hr
A
B
1
-1
1
2277
Chr 1 (100 kb resolution)
1
1561
Chr 7 (100 kb resolution)

Figure S3

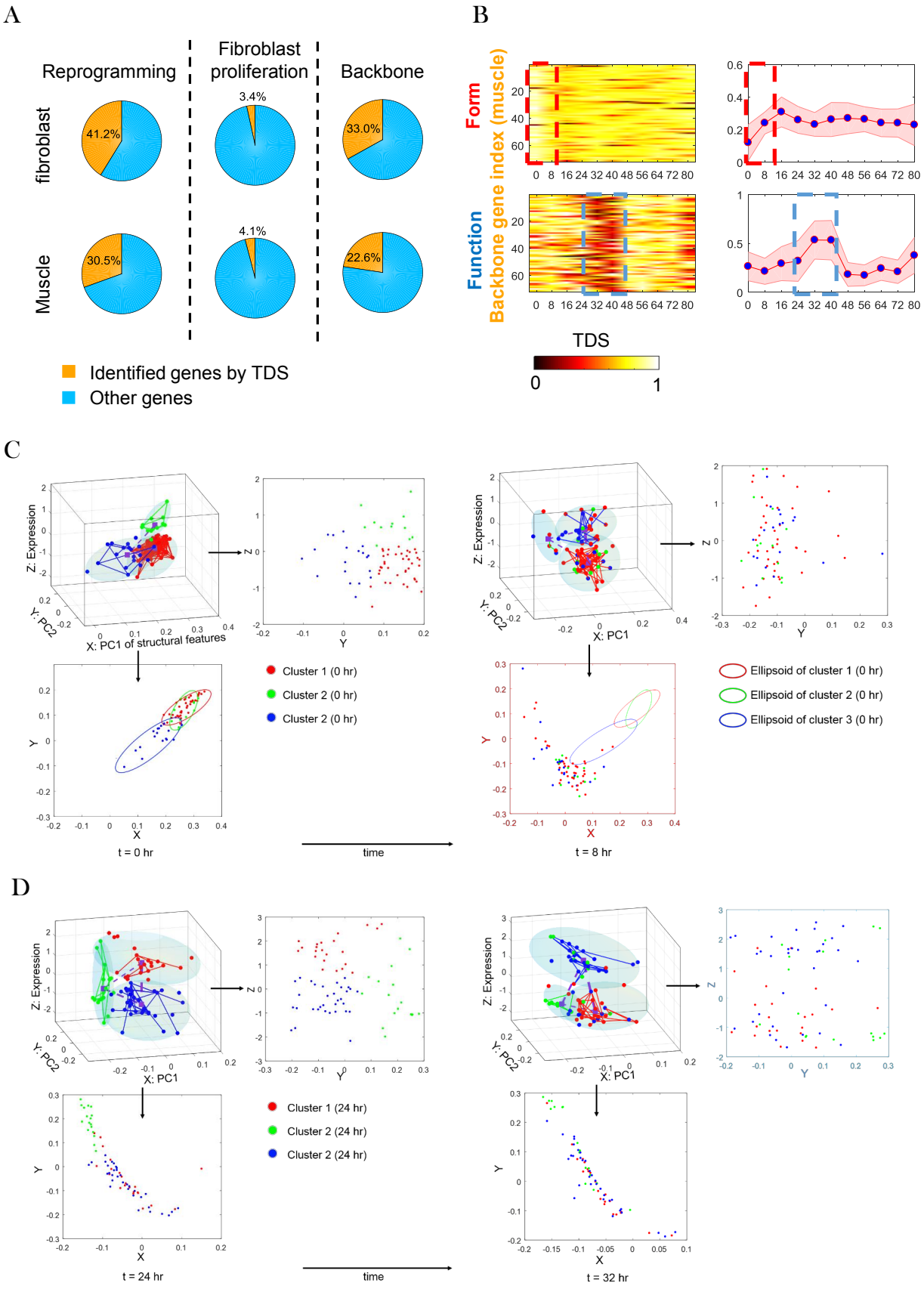

A
Reprogramming
Fibroblast proliferation
Backbone
fibroblast
Muscle
41.2%
3.4%
33.0%
30.5%
4.1%
22.6%
Identified genes by TDS
Other genes
B
Form
Function
Backbone gene index (muscle)
TDS
0
1
C
Z: Expression
X: PC1 of structural features
Y: PC2
Cluster 1 (0 hr)
Cluster 2 (0 hr)
Cluster 2 (0 hr)
Ellipsoid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

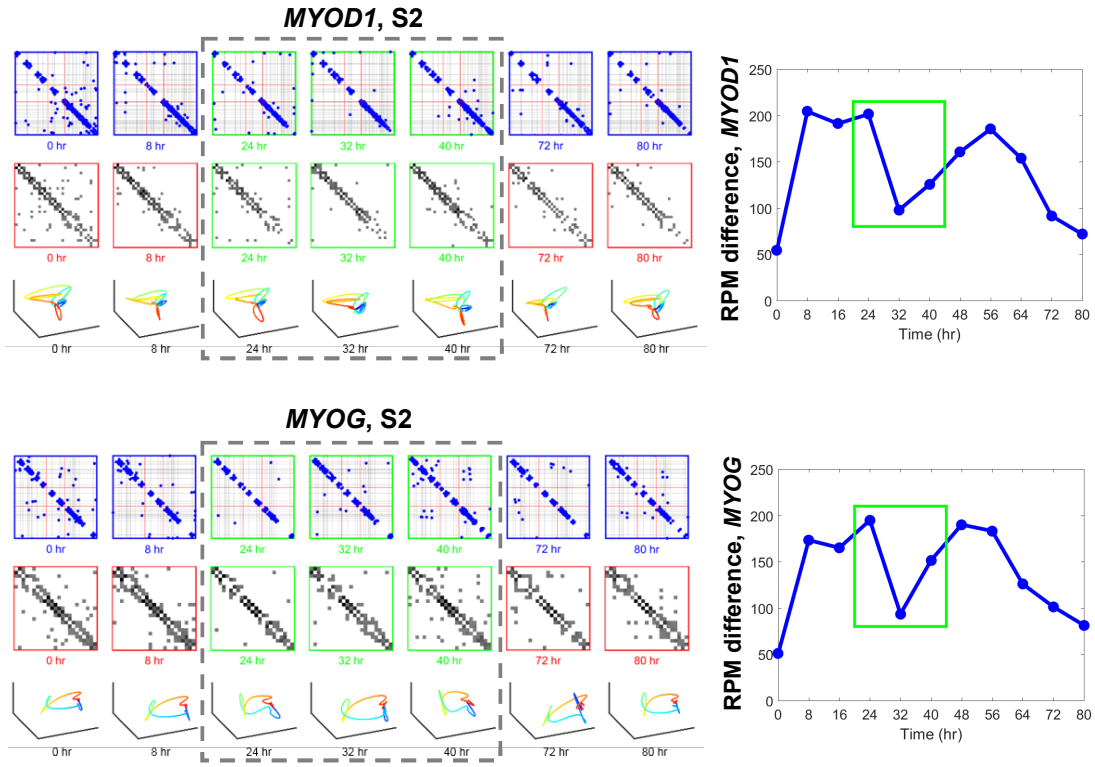

MYOD1, S2
0 hr
8 hr
24 hr
32 hr
40 hr
72 hr
80 hr
RPM difference, MYOD1
Time (hr)
MYOG, S2
RPM difference, MYOG
Time (hr)

Figure S5

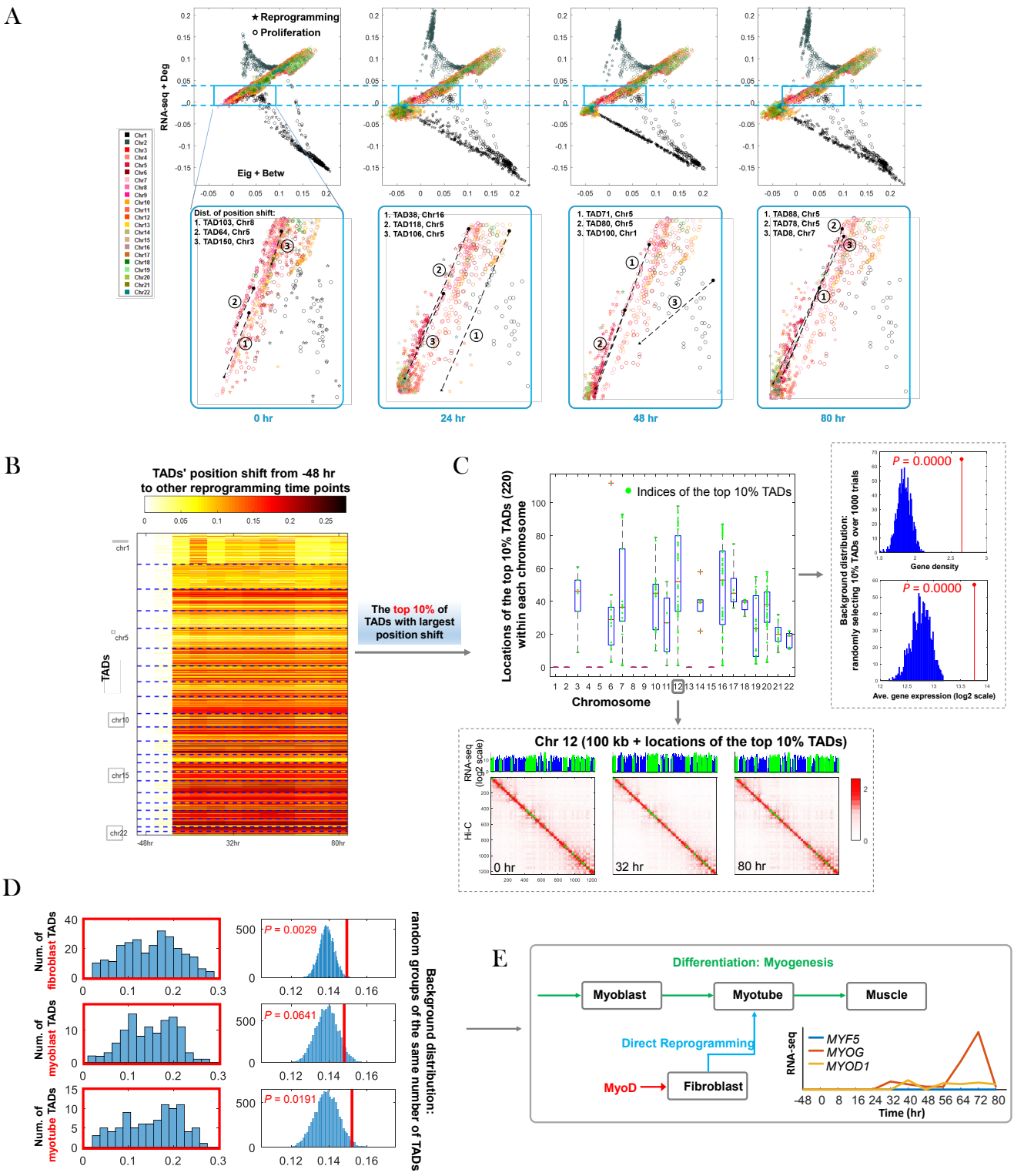

A
★ Reprogramming
○ Proliferation
RNA-seq + Deg
Eig + Betw
Dist. of position shift:
1. TAD103, Chr8
2. TAD64, Chr5
3. TAD150, Chr3
1. TAD38, Chr16
2. TAD118, Chr5
3. TAD106, Chr5
1. TAD71, Chr5
2. TAD80, Chr5
3. TAD100, Chr1
1. TAD88, Chr5
2. TAD78, Chr5
3. TAD8, Chr7
0 hr
24 hr
48 hr
80 hr
B
TADs' position 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

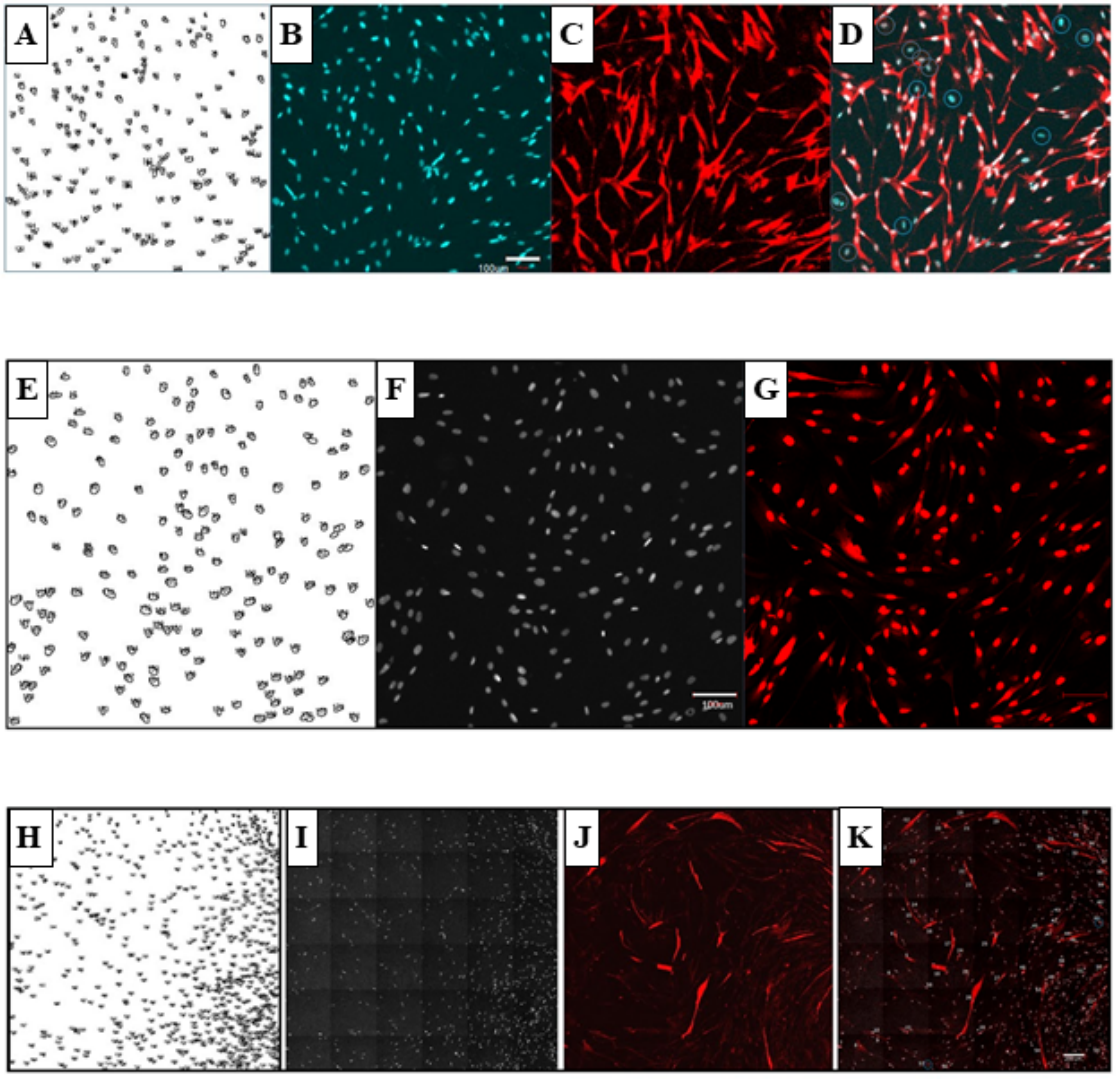


Figure S7

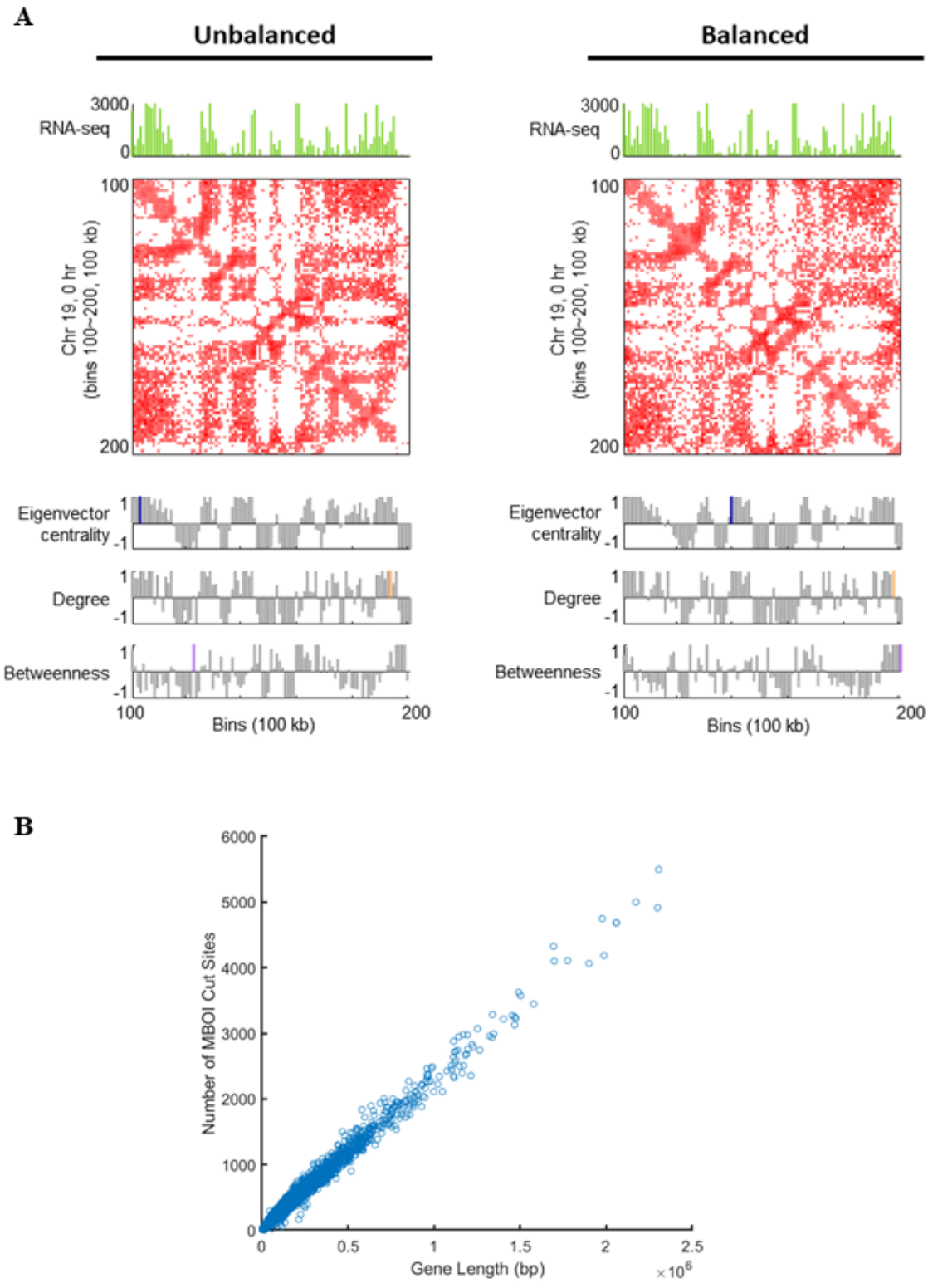

A
Unbalanced
Balanced
RNA-seq
3000
0
100
200
Chr 19, 0 hr
(bins 100~200, 100 kb)
Eigenvector centrality
Degree
Betweenness
1
-1
100
200
Bins (100 kb)
B
Number of MBOI Cut Sites
6000
5000
4000
3000
2000
1000
0
0.5
1
1.5
2
2.5
Gene Length (bp)
×10^6

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.