Genome Architecture Leads a Bifurcation in Cell Identity

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

[1]*Dept. of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[2]*Dept. of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109, USA*
[3]*Dept. of Computer Science, University of California-Irvine, Irvine, CA 92697, USA*
[4]*Dept. of Pediatrics & Communicable Diseases, University of Michigan, Ann Arbor, MI 48109, USA*
[5]*AI Foundations, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA*
[6]*Dept. of Biological Chemistry, University of California-Irvine, Irvine, CA 92697, USA*
[7]*Dept. of Mathematics, City University of Hong Kong, 999077 Hong Kong, China*
[8]*Dept. of Mathematics, University of California, Berkeley, CA 94720, USA*
[9]*Dept. of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA*
[*]*Contributed equally to this work*
[**]*Lead Contact, Correspondence: indikar@umich.edu*

SUMMARY

Genome architecture is important in transcriptional regulation, but its dynamics and role during reprogramming are not well understood. Over a time course, we captured genome-wide architecture and transcription during MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage. We found that chromatin reorganization occurred prior to significant transcriptional changes marking activation of the myogenic program. A global bifurcation event delineated the transition into a myogenic cell identity 32 hours after exogenous MYOD1 activation, an event also reflected in the local dynamics of endogenous MYOD1 and MYOG. These data support a model in which master regulators induce lineage-specific nuclear architecture prior to fulfilling a transcriptional role. Interestingly, early in reprogramming, circadian genes that are MYOD1 targets synchronized their expression patterns. After the bifurcation, myogenic transcription factors that are MYOG targets synchronized their expression, suggesting a cell-type specific rhythm. These data support roles for MYOD1 and MYOG in entraining biological rhythms.

Keywords: cellular reprogramming, 4D nucleome, bifurcation, network dynamics, network centrality, biological rhythms

INTRODUCTION

A comprehensive understanding of cell identity, how it is maintained and how it can be manipulated, remains elusive. Global analysis of the dynamical interplay between genome architecture (form) and transcription (function) brings us closer to this understanding (Rajapakse and Groudine, 2011). This dynamical interaction creates a genomic signature that we can refer to as the four-dimensional organization of the nucleus, or 4D Nucleome (4DN) (Chen et al., 2015; Dixon et al., 2015; Fortin and Hansen, 2015; Krijger et al., 2016). Genome technologies such as genome-wide chromosome conformation capture (Hi-C) are yielding ever higher resolution data that give a more complete picture of the 4DN, allowing us to refine cell types, lineage differentiation, and pathological contributions of cells in different diseases. High time resolution on a global scale gives key insight into biological processes. Of interest in regenerative medicine is understanding the dynamical process of cellular reprogramming.

Pioneering work by Weintraub et al. showed reprogramming of fibroblasts into muscle cells was possible through overexpression of a single transcription factor (TF), MYOD1, thus demonstrating that a different cell identity could supersede an established one (Weintraub, 1993; Weintraub et al., 1989). In 2007, when Yamanaka and colleagues reprogrammed human fibroblasts into an embryonic stem cell-like state with four TFs, POU5F1 (OCT4), SOX2, KLF4, and MYC, they showed that a pluripotent state could also supersede an established cell identity (Takahashi et al., 2007). These remarkable findings demonstrate the possibilities of controlling the genome and the cell identity through TFs. However, how TFs dynamically orchestrate genome architecture and transcription as a cell changes identities during reprogramming is not understood.

One exciting finding in recent reports was that Hi-C contact maps can be used to divide the genome into two major compartments, termed A and B (Chen et al., 2015; Lieberman-Aiden et al., 2009). Compartment A is associated with open chromatin (transcriptionally active), and compartment B with closed chromatin (transcriptionally inactive). The pattern of A/B compartmentalization is cell-type specific and reflects unique gene expression sig-

natures. Studies show that A/B compartment switching occurs during differentiation and reprogramming, where genomic regions previously assigned to one compartment change to a different compartment to facilitate the gene expression associated with a new cell state (Chen et al., 2015; Dixon et al., 2015; Fortin and Hansen, 2015; Krijger et al., 2016). These studies support conjecture that A/B compartments have a contributory but not a deterministic role in establishing cell-type specific patterns of gene expression (Dixon et al., 2015).

Previously we introduced a new technique from spectral graph theory to partition the genome into A/B compartments and identify topologically associating domains (TADs) (Chen, Hero and Rajapakse, 2016). This motivated us to study the 4DN from a network point of view, where nodes of the network correspond to genomic loci that can be partitioned at different scales: gene level, TAD level, and chromosome level. The edges of the network indicate contact between two loci, with contact weights given by Hi-C entries. Previous studies have extracted a single topological feature from the Hi-C matrix (e.g. A/B compartments), and then combined those results with gene expression (Chen et al., 2015; Dekker et al., 2013; Dixon et al., 2015; Fortin and Hansen, 2015; Krijger et al., 2016; Lieberman-Aiden et al., 2009). From the network perspective, A/B compartments are identified as distinct connected components of a network. As will be demonstrated here, other properties of the network topology, such as node centrality, can be extracted from the Hi-C matrix to yield further information about chromatin spatial organization.

The utility of network centrality allows one to identify 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, while closeness centrality is a global connectedness measure that characterizes the average distance of a given node to all 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. Google's Page-rank algorithm uses a variant of eigenvector centrality (Lohmann et al., 2010).

In this work we investigated dynamics of topological features of genome architecture and

explored how they varied with transcription during MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage. Sampling across a time course during reprogramming, we captured architecture by Hi-C, transcription by RNA-seq, and proteomics data. By combining different centrality measures we found important Hi-C features largely overlooked in previous studies, and this approach facilitated coordinated form-function analysis of chromatin conformation and gene expression in genome-wide data. Analyses of form-function dynamics revealed chromatin reorganization that occurred prior to changes in transcription. In this work, we introduce the concept of bifurcation to describe a critical transition from one cell identity to another. We detected a bifurcation in space-time 32 hours after activation of exogenous MYOD1 in fibroblasts that suggests a definitive transition into the myogenic lineage. Additionally we identified a core subset of myogenic genes that define this state. We further found robust synchronization of circadian gene expression, and determined that these genes are downstream targets of MYOD1, suggesting MYOD1 feedback onto circadian gene circuits. After the bifurcation, MYOG was associated with synchronization of a subset of important myogenic transcription factors. These findings support roles for MYOD1 and MYOG in entraining biological rhythms. Finally, our analysis of genomic regulatory elements such as chromatin remodeling genes, super enhancer regions and microRNAs provides additional clues toward understanding system-wide dynamics during reprogramming.

RESULTS

MYOD1-mediated Direct Reprogramming: Revisiting Weintraub

We converted primary human fibroblasts into the myogenic lineage using the transcription factor and master regulator MYOD1. In its native system, MYOD1 initiates the transcriptional program that turns muscle cell precursors into multinuclear muscle fibers. Fibroblasts were transduced with a lentiviral construct that expressed human MYOD1 fused with the tamoxifen-inducible mouse ER(T) domain (L-MYOD1) (Kimura et al., 2008). With 4-hydroxytamoxifen (4-OHT) treatment, transduced cells showed nuclear translocation of L-MYOD1 and morphological changes consistent with myogenic differentiation (Figure S1).

We then validated the activation of two key myogenic genes downstream of *MYOD1* (*MYOG* and *MYH1*) (Figure 1B). These results demonstrate successful conversion of fibroblasts into the myogenic lineage by L-MYOD1. Subsequent analyses were carried out on 4-OHT treated, transduced cells, sampling at 8 hour (hr) intervals for RNA-seq, small RNA-seq, and Hi-C analyses, and at 24 hr intervals for proteomics.

A Bifurcation Delineates Emergence of A New Cell Identity

A single structural feature of the Hi-C matrix is insufficient to characterize the relationship between the chromatin interaction profile and genome function (Babaei et al., 2015). We exploit network centrality to capture multiple structural features of the Hi-C matrix and enable efficient analysis of form-function dynamics during cellular reprogramming (Figure 2A, 2B, and STAR Methods for details on network centrality).

We first extracted low-dimensional representation of genome-wide form-function features based on 1 Mb binned Hi-C matrices and expression profiles (Figure 2C using Laplacian eigenmaps; see STAR Methods for details and Figure S3). The resulting form-function representations, fitted by minimum volume ellipsoid (MVE) (Figure 2C, S4, and STAR Methods for details), showed distinct configurations at different time points in general. However, we observed a similar form-function evolution at the initial time points prior to treatment with 4-OHT (−48 and 0 hrs). Similar behavior was observed at later time points during reprogramming (72 and 80 hrs). These results imply mathematically what may be expected logically; two stable basins exist: a fibroblast-like stage (early) and a muscle-like stage (late). By contrast, during the time interval $[8, 64]$ (hrs), the form-function ellipsoids at consecutive time steps are more distinguishable, indicating larger temporal variations of form and function. The temporal change of form and function can be quantitatively assessed using Kullback-Leibler (KL) divergence (STAR Methods), which yields the information 'distance' of form-function representations at any two time points. We found that there exists a bifurcation point at 32 hrs (Figure 2D) while showing the temporal difference score (TDS, see STAR Methods) of KL divergence, which reveals the information change at neighboring time steps. By contrasting reprogramming with data on human fibroblast proliferation over a

56-hr time course (Chen et al., 2015), we found that the detected bifurcation point (32 hrs) is aligned with a branching trajectory, corresponding to a bifurcation of order 1 (Borchert and Slade, 1981), with $P = 0.0048$ (Hotelling's T-Square test; see Figure 2E and STAR Methods). Taken together, we interpret from these data three phases for reprogramming: fibroblast, bifurcation, and myogenic. In support of a phase transition between the bifurcation and the myogenic phase, endogenous *MYOD1* and *MYOG* expression were first detected around the bifurcation point (Figure S1F).

This bifurcation was also observed at the gene-level in Hi-C data for key myogenic genes. We examined *MYOD1* and *MYOG*, chosen for their known importance in early activation of the myogenic program (Tapscott, 2005), and their transcriptional activation around the bifurcation point. We constructed base pair resolution contact maps (Chen et al., 2015) at the genomic locations of *MYOD1* and *MYOG*, with a 5kb buffer both upstream and downstream of the loci to include surrounding local conformation changes (Figure 2F, top; Figure S5C, top). We then binned these contact maps based on MboI cut sites, normalized the index values (reads per million, RPM), and determined the difference between successive time points (summation of element-wise absolute value differences, STAR Methods). This analysis revealed a pattern strikingly similar to what was found for genome-wide methods; chromatin form dynamics showed the largest amount of change early in reprogramming, with a local minimum in change observed at the bifurcation (Figure 2F, bottom; Figure S5D).

While a bifurcation was robustly detectable at different scales, A/B compartment switching was less distinct. We determined the A/B compartments according to the sign of the graph Laplacian Fiedler vector of the normalized 100 kb-binned Hi-C matrices (STAR Methods). A/B compartment switching was found to occur in 10.46% of the genome, with the most significant increase in A/B compartment switching occurring just after 4-OHT treatment (8 hrs). Interestingly, a local minimum in A/B compartment switching percentage occurs at our identified point of bifurcation (32 hrs), paralleling our finding of a local minimum of information divergence (Figure 2D and S5). Although the trajectory of A/B compartment switch reflects genome dynamics to some extent, with this analysis it is difficult

to recognize dynamical form-function relationships (Dixon et al., 2015).

A Direct Pathway from Fibroblasts to Myotubes

We next sought to examine the pathway into the myogenic lineage, to determine whether the data support transit through a myoblast-like stage or directly to a more differentiated myotube-like state. For this, 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 has shown that the boundaries of TADs remain stable between cell types (Dixon et al., 2015, 2012), however the dynamical TAD-level interactions and functional changes during cellular reprogramming are not well understood. We interpret the genome as a network of TADs (Figure 1C), where network vertices correspond to TADs, and edge weights are given by the interaction frequency between two TADs from Hi-C (retaining only interactions that exceed the 50th-percentile of inter-TAD contacts; see STAR Methods). The function associated with a TAD is characterized by the sum of RNA-seq values of the set of genes contained within the TAD defined region.

We applied network centrality analysis (STAR Methods) to extract 2D representation of dynamical form-function features at the TAD scale (Figure 3A). The resulting configuration of chromosomes was robust over time, but positions of TADs within a chromosome shifted. This can be seen by contrasting the fibroblast stage (prior to 4-OHT; −48 hr) with the subsequent reprogramming time points $(0, \ldots, 80$ hrs) (Figure 3B). 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 3C). We found that the identified TADs are of high gene density, and genes within them are highly expressed ($P < 0.001$; see STAR 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 in order to determine whether cells transitioned through a myoblast-like state. Gene sets were extracted from Gene Ontology (GO) (Table S1), 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 $P = 0.0029$ and 0.0191, respectively (Figure 3D, and STAR Methods). By contrast, the position shifts of TADs that contained myoblast or skeletal muscle genes were not statistically significant. These results imply a direct pathway from fibroblasts to myotubes during reprogramming across this 80-hr time course.

A direct pathway of reprogramming is further supported by expression analysis of three myogenic regulatory factors: *MYF5*, *MYOD1* and *MYOG* (Bentzinger et al., 2012; Weintraub et al., 1991). It is known from the hierarchy of transcription factors regulating progression through natural myogenic differentiation (Bentzinger et al., 2012) that *MYF5* is expressed in myoblasts, while *MYOD1* and *MYOG* are up-regulated in myotubes. In our data *MYF5* was not activated during reprogramming, while *MYOD1* and *MYOG* were expressed after the bifurcation point (Figure 3E and S1F).

Form Precedes Function

Given the cell-state trajectory, it is unclear whether L-MYOD1 induces a genome architecture change prior to its role in mediating muscle gene transcription, or vice versa. To answer this question, we studied the evolution of form and function separately under 5 GO-identified gene modules: fibroblast, myotube, skeletal muscle, cell cycle and circadian related genes (Figure 4A and Table S1). Form-function evolution is depicted by TDS (STAR Methods) built on the multi-centrality based structural features (form) and gene expression (function). We found clusters of genes (STAR Methods) that have similar form-function changes in terms of their TDS values. We then identified the time points that had dominant form or function changes observed in the network degree centrality and the RNA-seq RPKM values (Figure 4B and 4C). This suggests that the dynamical form-function relationships were significantly affected by particular subsets of genes (with peak TDS values) at particular time points. We next tracked the number of genes in each of 5 gene modules that had significant form-function changes over time with our TDS-based indicator. We observed that the significant form change occurs at 8 hrs while the function change does not occur until at least 16 hrs (in the vicinity of the bifurcation). Thus, form precedes

function, suggesting that chromatin structure change facilitates the orchestrated activation of transcriptional networks associated with cell differentiation.

To explore the unique characteristics of form-function relationships in human fibroblast-to-muscle reprogramming, we compared this dataset to data obtained via methodologically similar methods on proliferating fibroblasts over a 56-hr time course (Chen et al., 2015). These cells were cell cycle and circadian rhythm synchronized before introduction to full growth medium, with time point collection of RNA-seq and Hi-C every 8 hours. Given a gene module (focusing on fibroblast or muscle related genes, see Table S1), we observed that the pattern of form-function evolution during reprogramming is quite different from fibroblast proliferation (Figure 5A). We found that genes that are associated with significant form change at 8 hrs and function change at 32-40 hrs are in the majority ($> 30\%$) of gene modules during cellular reprogramming, while only a small portion of these genes ($< 5\%$) have the same significant form-function change during human fibroblast proliferation (Figure 5B). We extracted the genes (77 in fibroblast module and 72 in muscle module; see Table S2) that are responsible for significant form-function change during cellular reprogramming and those that are less active during fibroblast proliferation. This provides a set of backbone genes shown by its distinct form-function evolution (Figure 5C). The statistical significance of the identified backbone genes was evaluated as $P < 0.05$ by comparing with fibroblast proliferation (Figure 5D, and STAR Methods). A form-function space (3D) is then introduced to illustrate the dynamics of backbone genes during cellular reprogramming, where the x-y plane is composed of the first two principal components of centrality-based Hi-C features, and the z-axis represents gene expression (Figure 5E). We divided backbone genes into 3 groups (K-means), each of which represents a set of backbone genes that maintain similar form-function characteristics. We observed that projection onto the structural surface (x-y plane) is distinctly different between time points 0 and 8 hrs. This implies that form change is the dominant factor early in the cellular reprogramming. By contrast, from 24 to 32 hrs, the functional change plays a more important role shown by Figure 5F, namely, form precedes function.

We next introduce the concept of a 'portrait of 4DN', which provides a simple quantitative

assessment of form-function dynamical relationships at the chromosome level. Specifically, on a 2D plane, we designate one axis as a measure of form in terms of network connectivity (STAR Methods), and the other as a measure of function in terms of the average RNA-seq RPKM values. The portrait of 4DN is then described by a form-function domain, made up of 8 time points $[0, 56]$ (hrs) for each chromosome (Figure 6A). The form-function difference at the bifurcation point was observed by contrasting reprogramming with fibroblast proliferation (Figure 6B). We also noted that the centroid of the fitted form-function ellipsoid (MVE estimate, Figure 6A and STAR Methods) for each chromosome shifts between cellular reprogramming and fibroblast proliferation. Such a shift can be decomposed into the horizontal change and the vertical change, where the former describes the structural change, and the latter corresponds to the functional change. We found that most chromosomes involve significant structural change (86.4%) as compared to the functional change (13.8%) (Figure 6C). Furthermore, the area of the form-function ellipse is able to characterize the variance (uncertainty) of 4DN (Figure 6C). We observed that ellipsoids associated with cellular reprogramming have larger volumes than those under human fibroblast proliferation. This implies that cellular reprogramming leads to a more complex dynamical behavior, indicated by a space-time bifurcation and phase transition during this process (Figure 2D).

MYOD1-mediated Synchronization of Circadian Rhythms

We observed that upon MYOD1 nucleus translocation, the population of cells exhibit robust synchronization in circadian gene expression. Upon further inspection this finding can be interpreted as follows; a large portion of the core circadian gene network relies on E-box motif targets and transcription factors for control, the same motif that the bHLH protein MYOD1 targets (Ueda et al., 2002). This is further supported by the observation that known core circadian genes (STAR Methods and Table S1) with E-box targets displayed the most profound synchronization initially, starting with an uptick in gene expression post MYOD1 addition (Figure 7A-D). JTK_CYCLE confirmed this observation, as all E-box circadian genes were found to have a period of 24 hrs, with a maximum lag between any genes of 4 hrs (except for *CRY1*; See Table S4) (Hughes et al., 2010).

Interestingly, the subset of transcripts displaying oscillatory behavior was different pre- and post- our identified bifurcation point. *MYOD1* and *MYOG* expression began around our identified bifurcation point at 32 hrs and both transcripts displayed oscillatory expression. Additionally, circadian transcript oscillations dampened at time point 40 hr, corresponding with when the cells were given low-serum differentiation medium. Core clock trajectories are shown in Figure 7E. We examined the subset of transcripts that were found to be only oscillatory after the bifurcation point and synchronous (in phase or antiphase) with *MYOD1* and *MYOG* expression. Transcripts that were found to be oscillating in phase are potentially interacting in an excitatory fashion, and vice versa. Since both MYOD1 and MYOG are known transcription factors, we further investigated which newly oscillating transcripts may have motifs for either MYOD1 or MYOG binding sites in their promoter regions using MotifMap (Daily et al., 2011). For those genes that were found to oscillate only after the bifurcation point, 51 oscillating transcripts possessed upstream MYOG binding sites and were synchronous with MYOG. Similarly, 17 oscillating transcripts with MYOD1 binding sites were found to be synchronous with MYOD1.

We found 23 known transcription factors to oscillate after the bifurcation point at our selected significance. Six of these oscillating transcription factors were synchronous and targeted by MYOG. Only a single oscillating transcription factor, ELF3, was found to be targeted and synchronous by MYOD1. Trends over the entire experimental time series for these targeted transcription factors are shown in Figure 7F.

Several of the six oscillatory transcription factors targeted by MYOG or MYOD1 have been shown to be related to cell differentiation in literature. Numerous studies have implicated the important role of SOX15 in muscle differentiation (Meeson et al., 2007). GATA6 has been shown to regulate vascular smooth muscle development in several studies (Xie et al., 2015). ISL1 has been shown to interact with CITED2 to induce cardiac cell differentiation in mouse embryonic cells (Pacheco-Leyva et al., 2016). ELF3 is found to play a diverse role in several types of cell differentiation (Böck et al., 2014).

Changes in Regulatory Elements During Reprogramming

Here we highlight our analyses of early-stage chromatin remodelling gene expression dynamics, super enhancer dynamics, and microRNAs expression.

Examination of early stage RNA-seq data [-48, 16] (hrs) revealed endogenous mechanisms relevant to *MYOD1* transcriptional activation including muscle stage-specific markers and chromatin remodeling factors (See Figure S6). At the 16-hr time point, we found that the combined upregulation of *DES*, *MYL4*, *TNNT1* and *TNNT2* suggests a differentiation from a myoblast lineage to a skeletal muscle (Gard and Lazarides, 1980; Schiaffino et al., 2015). *EZH2* has been associated with both safe-guarding the transcriptional identity of skeletal muscle stem cells and with 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 hours ($P = 7.2 \times 10^{-5}$) 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 are necessary for reprogramming of cell fate to muscle lineage (Bentzinger et al., 2012; Ferrán et al., 2016; Jang et al., 2015).

We also investigated how muscle-related super enhancer-promoter (SE-P) interactions change over time throughout L-MYOD1-mediated reprogramming. To capture these dynamics, we extracted the Hi-C contact between skeletal muscle super enhancer regions and associated genes TSS (±1kb), as determined by Hnisz et al. (2013) (618 SE-P regions; STAR Methods). We observed that for these skeletal muscle SE-P Hi-C regions, the strongest amount of contact occured relatively early in the reprogramming process, peaking 16-24 hrs post-L-MYOD1 addition to the nucleus (Figure S7). Exact SE-P contact vs 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 introduction to the nucleus. Contact vs function trends for the top 36 upregulated genes are shown in Figure S7 (STAR Methods).

We measured the abundance of 2588 microRNA (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 down-regulated in expression levels over the time course relative to the baseline control (FDR < 0.05) (Table S3). 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 three miRNAs, plus miR-133b (FDR = 0.09), significantly increased expression levels after 4-OHT treatment (Figure S8). Their expression patterns were highly similar to that observed in mouse C2C12 cell differentiation (Rao et al., 2006). The identification of muscle-specific miRNAs, particularly miR-206 known to be skeletal muscle specific (McCarthy, 2008) and expressed greater than 1000 fold at the later time points than baseline (Figure S8), further supports L-MYOD1-mediated reprogramming of fibroblasts to myotubes. Notably, the cardiac-specific species miR-1-5p, miR-208a, and miR-208b (McCarthy, 2011) are not detected in our samples.

DISCUSSION

In this study, we provide analysis on L-MYOD1-mediated reprogramming of human fibroblasts into the myogenic lineage from a network and dynamical systems perspective. Distinct from previous studies on cellular reprogramming, we generated an enriched time-series dataset integrated with Hi-C, RNA-seq, miRNA, and proteomics data, over a 112-hr time course. 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 a bifurcation point at 32 hours after initiation of L-MYOD1 nuclear localization by 4-OHT treatment, and show form precedes function during this reprogramming regime. We also made a comparative study between the cellular reprogramming and the human fibroblast proliferation, and quantified how different these two processes were at different scales. Analysis on genomic regulatory elements such as enhancers, transcription factors, and microRNA provided additional understandings on biological mechanism of the cellular reprogramming.

Our data suggest a direct pathway of reprogramming from fibroblasts to myotubes that bypasses a myoblast intermediate and is associated with 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 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 was theorized in (Del Vecchio et al., 2017) that a reprogrammed bio-system with positive perturbation (i.e. overexpression of one or more specific transcription factors like MYOD1) would bypass the intermediate state and move directly towards 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.

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 introduction 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. Additionally, 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 post-bifurcation point, after which MYOG entrains the oscillations of a distinct subset of myogenic transcription factors. Therefore, MYOD1-mediated reprogramming and circadian synchronization are mutually coupled, as is the case in many other studies of the reprogramming of cell fate (Umemura et al., 2014; Wagner et al., 2014).

Our proposed bio- 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 prior to gene transcription during cellular reprogramming, i.e., form precedes function, and that dynamical nuclear reorganization plays a key role in defining cell identity. 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 ChIP-seq 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 structure from multiple views, and facilitates quantitative integration with gene expression. Accordingly, the detailed connections between network structure 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. Identifying an architectural function for transcription factors that is distinct from transcription would define a new molecular function with as yet an unknown role in developmental and cancer cell biology. We believe that investigating form-function dynamics will be key in understanding and advancing cell reprogramming strategies. This perspective may have broad translational impact spanning cancer cell biology and regenerative medicine.

AUTHOR CONTRIBUTIONS

I. R. conceived and supervised the study. H. C., S. R., W. M. and I. R. designed and performed the experiments. S. L., P.-Y. C., A. H. and I. R. contributed network centrality based analytic tools. S. L., H. C., S. R., L. S., G. H and I. R. performed computational analyses and interpreted the data. All authors 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.

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 Richard McEachin for support in processing Illumina sequence data with a standardized pipeline. We also 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 Program and the DARPA Deep-Purple and FunCC Program. We also acknowledge the seminal work of Mark Groudine and late Hal Weintraub, whose ideas continue to guide our thinking.

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.

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', *Proceedings of the National Academy of Sciences* **107**(44), 19090–19095.

Babaei, S., Mahfouz, A., Hulsman, M., Lelieveldt, B. P., de Ridder, J. and Reinders, M. (2015), 'Hi-C chromatin interaction networks predict co-expression in the mouse cortex', *PLoS Comput Biol* **11**(5), 1–21.

Bentzinger, C. F., Wang, Y. X. and Rudnicki, M. A. (2012), 'Building muscle: molecular regulation of myogenesis', *Cold Spring Harbor perspectives in biology* **4**(2), 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', *Developmental biology* **386**(2), 321–330.

Borchert, R. and Slade, N. A. (1981), 'Bifurcation ratios and the adaptive geometry of trees', *Botanical Gazette* **142**(3), 394–401.

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.

Chen, J., Hero, A. O. and Rajapakse, I. (2016), 'Spectral identification of topological domains', *Bioinformatics* **32**(14), 2151–2158.

Chen, P.-Y., Choudhury, S. and Hero, A. O. (2016), 'Multi-centrality graph spectral decompositions and their application to cyber intrusion detection', *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 4553–4557.

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.

Dekker, J., Marti-Renom, M. A. and Mirny, L. A. (2013), 'Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data', *Nature Reviews Genetics* **14**(6), 390–403.

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 Systems* pp. 109–120.

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**(7539), 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**(7398), 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', *Scientific reports* **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 biology* **16**(1), 180.

Gard, D. L. and Lazarides, E. (1980), 'The synthesis and distribution of desmin and vimentin during myogenesis in vitro', *Cell* **19**(1), 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**(4), 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', *Journal of biological rhythms* **25**(5), 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 & disease* **6**(8), 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**(10), 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 & development* **25**(8), 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', *Human molecular genetics* **17**(16), 2507–2517.

Kosak, S. T. and Groudine, M. (2004), 'Form follows function: the genomic organization of cellular differentiation', *Genes & development* **18**(12), 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**(5), 597–610.

Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.

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.

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**(4), 1–8.

McCarthy, J. J. (2008), 'Microrna-206: the skeletal muscle-specific myomir', *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1779**(11), 682–691.

McCarthy, J. J. (2011), 'The myomir network in skeletal muscle plasticity', *Exercise and sport sciences reviews* **39**(3), 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', *The EMBO journal* **26**(7), 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 Reports* **7**(6), 1037–1049.

Piegl, L. (1989), 'Modifying the shape of rational b-splines. part 1: curves', *Computer-Aided Design* **21**(8), 509–518.

Rajapakse, I. and Groudine, M. (2011), 'On emerging nuclear order', *The Journal of cell biology* **192**(5), 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', *Proceedings of the National Academy of Sciences* **103**(23), 8721–8726.

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.

Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A. and Reggiani, C. (2015), 'Developmental myosins: expression patterns and functional significance', *Skeletal muscle* **5**(1), 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 Journal* **280**(17), 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**(5), 861–872.

Tapscott, S. J. (2005), 'The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription', *Development* **132**(12), 2685–2695.

Ueda, H. R., Chen, W., Adachi, A., Wakamatsu, H. et al. (2002), 'A transcription factor response element for gene expression during circadian night', *Nature* **418**(6897), 534–539.

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', *Proceedings of the National Academy of Sciences* **111**(47), 5039–5048.

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.

Wagner, N., Alasibi, S., Peacock-Lopez, E. and Ashkenasy, G. (2014), 'Coupled oscillations and circadian rhythms in molecular replication networks', *The journal of physical chemistry letters* **6**(1), 60–65.

Weintraub, H. (1993), 'The myod family and myogenesis: redundancy, networks, and thresholds', *Cell* **75**(7), 1241–1244.

Weintraub, H., Davis, R. et al. (1991), 'The myod gene family: nodal point during specification of the muscle cell lineage', *Science* **251**(4995), 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', *Proceedings of the National Academy of Sciences* **86**(14), 5434–5438.

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

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', *Science signaling* **8**(376), 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 research* pp. 1–13.

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 research* **40**(8), 3419–3430.

MAIN FIGURE TITLES AND LEGENDS

Figure 1: Experimental design to explore cellular reprogramming with time-evolving genome architecture (form) and gene expression (function).

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

(B) Overview of Experiment.

(B1) *Top*: The cassette for myogenic reprogramming lenti-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 are samples transduced with L-MYOD1, no transduction (L4) or lenti-vector only (L5), L7 is RT- negative control, and L8 is water for PCR negative control. Two key MYOD1 downstream genes, *MYOG* & *MYH1* are activated by the expression of L-MYOD1. *GAPDH* is used as internal control, and *CDKN1A* (P21) is universally expressed.

(B2) Time course of MYOD1-mediated reprogramming. The time window outlined in green corresponds to time points at which both genome architecture and transcription are captured by Hi-C and RNA-seq.

(C) Scale-adaptive Hi-C matrices and gene expression. The considered scales include 1Mb, 100kb, TAD and gene level.

Figure 2: Detecting a bifurcation point in the form-function time series data.

(A) Overview of our network-based approach to explore form-function dynamics: feature extraction and integration, Laplacian eigenmaps for low-dimensional data representation, and evaluation of information divergence.

(B) Illustration of multi-centrality based structural features for Hi-C of 26 selected cell-cycle genes in CDC, CDK, CDKN and CCN gene families. The extracted centrality feature vectors are combined with function vectors (i.e., gene expression), leading to the form-function time series

data. *Top* (or *Bottom*): Network representation of Hi-C matrix at 24 hr (or -48 hr) with centrality measures: eigenvalue centrality, betweenness centrality, closeness centrality and local Fiedler vector centrality (LFVC) (STAR Methods). The node is named as node index plus gene name, and its size corresponds to the nodal centrality score. The edge width implies the Hi-C contact number. Nodes with the top five largest centrality values are highlighted in magenta. *Middle*: The gene-level Hi-C matrix and expression over -48, 0, and 24 hrs.

(C) 2D data representation of network centrality features based on Hi-C contact map and gene expression binned at 1 Mb scale.

(D) Identification of a bifurcation point. *Left*: Time-series Hi-C maps (bin indices 1148–1434 of chromosome 8 at 100 kb scale) and gene expression profiles at log-scale. *Middle*: TDS of information divergence at neighboring time points (STAR Methods). *Right*: A bi-stable landscape picture for cellular reprogramming.

(E) Branching trajectory with bifurcation point 32 hrs (*Top*), and P-value trajectory (*Bottom*).

(F) Genomic architecture of *MYOD1*. *Top*: First row depicts Hi-C matrices at base pair scale, where blue points are Hi-C 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 (Piegl, 1989) that fits 3D representation of MboI binned contact matrices using Laplacian eigenmaps. *Bottom*: Summation of the difference of all matrix elements between time points (STAR Methods). Green rectangle highlights the area around the bifurcation point.
See also Figure S1F, S2, S3, S4 and S5.

Figure 3: Form-function dynamics for networks of TADs during reprogramming.

(A) 2D representations of TAD-scale form-function features at time 0, 24, 48 and 80 hrs. Eigenvector, closeness and local Fiedler vector centrality were found to contribute the most to PC1, while RNA-seq, degree and betweenness centrality contribute the most to PC2 (STAR Methods). 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 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 (100kb-binned Hi-C) together with gene expression at time 0, 32 and 80 hrs. *Right*: P values of gene density and average gene expression (STAR Methods).

(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 (STAR Methods).

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

Figure 4: Dynamical form-function relationship under gene modules of interest.

(A) Outline for detecting form-function change. There are two major steps: calculating TDS (form or function), and identifying critical time points of form-function change under gene clusters.

(B) Characterization of form change for fibroblast related genes. *Left*: Heatmap of structural TDS for the largest gene cluster, showing the dominant form change at 8 hrs. *Right*: Degree centrality varies significantly from 0 to 8 hrs.

(C) Characterization of function change for fibroblast related genes. *Left*: Heatmap of functional TDS for the largest gene cluster, showing the dominant function change at 32 hrs. *Right*: RNA-seq RPKM (normalized over time) varies significantly from 24 to 32 hrs.

(D) Form-function change indicator for gene modules of interest.

Figure 5: Dynamics of backbone genes under fibroblast and muscle gene module.

(A) Form-function change indicators for gene modules of interest during cellular reprogramming and fibroblast proliferation, respectively.

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

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

(D) Statistical significance of backbone genes for form-function evolution during cellular reprogramming.

(E) 3D configuration of muscle-related backbone genes in form-function space from 0 to 8 hrs, highlights 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 purple square.

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

Figure 6: Portrait of 4DN under cellular reprogramming and human fibroblast proliferation.

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

(B) Shift of form-function domains of chromosomes at the bifurcation point (32 hr). Chromosomes 5, 12 and 13 yield the top three most significant changes.

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

Figure 7: Form and function dynamics of circadian E-box genes.

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

(B) Core circadian gene expression (RNA-seq) over time: Dexamethasone synchronization (B1) and L-MYOD1 synchronization (B2). 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 S1). Rows and columns correspond to core circadian genes, contacts are binary (i.e. any contact between genes at a given time are shown).

(D) Size of largest connected component (blue) and the corresponding Fiedler number (black), derived from the Hi-C contact map.

(E) Normalized gene expression (RPKM, cubic spline) highlighting oscillation dampening post-bifurcation and post-differentiation medium for select core circadian genes; MYOD1 and MYOG also shown.

(F) TFs that are targeted by MYOG or MYOD1 (ELF1), and are only oscillatory post-bifurcation point (STAR Methods).

(G) Conceptual diagram of biological rhythm entrainment during MYOD1-mediated reprogramming.

SUPPLEMENTAL FIGURE TITLES AND LEGENDS

Figure S1: Fluorescent micrographs showing immunocytochemistry dynamics of MYOD1 localization and the detection of MYH1 expression, and quantification of RNA and protein abundance; Related to Section "MYOD1-mediated Direct Reprogramming: Revisiting Weintraub".

(A-E) Left panels show DAPI staining of nuclei, middle panels show anti-MYOD1 staining (A-D) and anti-MYH1 staining (E), right panels are overlay of left and middle images, respectively.

(F) 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: Example of network representation and multi-centrality based structural features; Related to Figure 2.

(A) Network of selected 26 cell-cycle related genes (lying in categories of CDC, CDK, CDKN and CCN) at time −48, 8, and 24 hrs. Edge between two nodes indicates the existence of chromatin interaction between two genes. The edge width is proportional to Hi-C contact number. Each gene is named as node index plus gene name.

(B) Structural features based on eigenvector centrality, degree centrality, local fiedler vector centrality, betweenness, closeness, local clustering coefficient, multi-hop walk weights, and distance to genes CDKN1A and CCNB1. The network connectivity is characterized by Fiedler number (FN) equal to 0.252, 0.039 and 0.043, respectively.

Figure S3: 2D data representation of form-function features via PCA and Laplacian eigenmaps; Related to Figure 2.

(A) 2D manifolds embedded in form-function features based on 1Mb-scale Hi-C and gene expression over 12 time points, where x and y axis represent the first and second principal component of the projection matrix obtained from PCA (left) and Laplacian eigenmaps (right), respectively.

(B) Feature scores (STAR Methods) associated with PC1 and PC2 using PCA (left) and Laplacian eigenmaps (right), respectively. PC1 is dominated by structural features, while PC2 is

dominated by gene expression.

Figure S4: Ellipsoid estimate using MVE and sample covariance matrix based on 2D data representation obtained from Laplacian eigenmaps; Related to Figure 2.

(A) MVE estimate (STAR Methods) at time −48, 8 and 32 hrs.

(B) Fitted ellipsoid determined by sample covariance. Compared to MVE, the fitted ellipsoid does not necessarily enclose all the data points.

Figure S5: A/B Compartment switching summary and the genomic architecture of *MYOG*; Related to Figure 2.

(A) Percentage of genome-wide A/B compartment switch locations at each time point.

(B) Percentage of each chromosome that changes A/B compartmentalization at any point during direct reprogramming.

(C) Genomic architecture of *MYOG*, similar to Figure 2F.

(D) Summation of the difference of all matrix elements between time points (STAR Methods). Green rectangle highlights the area around the bifurcation point.

Figure S6: Early-phase expression dynamics of genes related to muscle cell terminal differentiation and chromatin remodeling; Related to Section "Changes in Regulatory Elements During Reprogramming"

(A) Genes encoding proteins involved in adult muscle function, including components of the contractile apparatus (DES, MYL4, TNNT1, TNN2), and EZH2, which is a repressor and is involved in myogenesis.

(B) Chromatin remodeling factors and master transcription factors act cooperatively with MYOD1 to drive proliferating human fibroblasts into muscle cells. They include ARID5A, part of the BAF47 muscle remodeling complex which 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.

Figure S7: Form and function of super enhancers and associated genes over time; Related to Section "Changes in Regulatory Elements During Reprogramming".

Top: Average Hi-C RPM contact between Potential super enhancer and associated gene TSS regions over time, as defined by Hnisz et al. (2013). *Bottom*: Top upregulated SE-P genes, log2(RPKM) (Blue) and SE-P Hi-C normalized contact (Red; see STAR Methods) over time.

Figure S8: Four muscle specific miRNAs significantly increased expression levels in the later time points relative to the baseline control; Related to Section "Changes in Regulatory Elements During Reprogramming"

X-axis corresponds to sampling time points, and y-axis shows log-scale differences at other time points compared to baseline (-48 hrs).

STAR★METHODS

The outline of detailed methods include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- METHOD DETAILS
  - Generation of a human *MYOD1* expressing construct
  - Cell culture, lenti-viral transduction, and induction of *MYOD1*
  - RNA-seq and small RNA-seq
  - Crosslinking of cells for Hi-C
  - Generation of Hi-C libraries for sequencing
  - Generation of Hi-C matrices
  - Reverse transcriptional polymerase chain reaction (RT-PCR) analysis
  - Immunocytochemistry analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Scale-adaptive gene expression
  - Scale-adaptive Hi-C matrix
  - Network representation of 4DN: graph Laplacian and Fiedler number
  - Structural feature extraction via network centrality measures
  - Integration of form and function
  - Data representation on low-dimensional non-linear manifolds
  - Fitting the data: minimum volume ellipsoid
  - Evaluation of information divergence

KEY RESOURCES TABLE

Please see a submitted separate table named by 'Key resource table'.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Corresponding Contact, Indika Rajapakse (indikar@umich.edu).

METHOD DETAILS

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, lenti-viral 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.

Lenti-viral 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 MYOD1 into the nucleus for myogenic reprogramming, we treated the cells transduced with lenti-hMYOD1-mER(T) by adding (Z)-4-Hydroxytamoxifen (4-OHT) (Cat# H7904, Sigma-Aldrich) to a final concentration of 1 µM to each flask in GM for two days. This treatment translocated the hMYOD1-mER(T) protein in the cytoplasm to the nucleus for MYOD1-meditated 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).

Crosslinking of cells for Hi-C

During a time course sampling, at each time point the cells in a T75 flask were washed with 10 mL PBS, and then incubated with 15 mL of 1% formaldehyde prepared in PBS at room temperature for exactly 10 min. To quench the crosslinking reaction, glycine of 2.5 M 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 scraped off with a scraper and transferred into 15 mL tubes. The crosslinked cells were collect with centrifugation at 800 x g for 10 min at 4° C. The cells collected were washed in 1 mL ice-cold PBS briefly, and spun down at 800 x g for 10 min at 4° C. After centrifugation, the supernatant was discarded 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 accord-

ing to the manufacturers 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 (`http://homer.salk.edu/homer/interactions/`) was used to develop the contact matrix with "makeTagDirectory" with the tbp 1 setting, and with "analyzeHiC" with the "-raw" and "-res 1000000" settings to produce the raw contact matrix at 1Mb resolution, or with the "-raw" and "-res 100000 settings to produce contact matrix at 100kb resolution.

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 (see Key Resources Table) 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 by converting number of interaction reads into values at fixed resolution bins (e.g., 100kb, 1Mb). However, RNA-seq data (RPKM) are generated

at gene level. For consistent analysis of form and function, we transform the RNA-seq data at gene level to its counterpart at bin level, namely,

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

where L_j is the length of gene j at base-pair unit, $\frac{L_j}{1000}$ is at kilobase-pair unit, L_{j,bin_i} is the length of the portion of gene j belonging to bin i, R_j signifies the RPKM value at 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 nearby loci in linear base-pair distance are more likely to be ligated 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 contact matrices by calculating the contact frequency of two genes; see details in (Chen et al., 2015, SI). Moreover, to construct TAD-scale contact matrices, we begin by normalizing both intra- and inter-chromosome Hi-C matrices at 100kb resolution, and then compute the density of genome contacts among 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

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

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 convert it to a sparse version by retaining only interactions that exceed the 50th-percentile of Hi-C

contact at TAD scale. This provides the backbone interaction network of TADs.

Network representation of 4DN: graph Laplacian and Fiedler number

Let $G_t = (V, E_t)$ denote a weighted undirected graph at time t, where V is a node set with cardinality $|V| = n$, and $E_t \subset \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ is an edge set at time t. The Hi-C matrix $\mathbf{H}_t$ can then be interpreted as an adjacency matrix corresponding to G_t, where $(i, j) \in E_t$ 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 $\mathbf{D}_t = \mathrm{diag}(\mathbf{H}_t\mathbf{1})$ is the degree matrix of G_t, $\mathbf{1}$ denotes the vector of all ones, and $\mathrm{diag}(\mathbf{x})$ signifies the diagonal matrix with diagonal vector $\mathbf{x}$. Given $\mathbf{L}_t$, the Fiedler number and the Fiedler vector is defined by the second smallest eigenvalue and its corresponding eigenvector. It is known from spectral graph theory (Chung, 1997) that G_t is connected (namely, there exists a path between every pair of distinct nodes) if and only if the Fiedler number is nonzero, and the entrywise signs of Fiedler vector encodes information on network partition.

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 provide essential topological characteristics of nodes (Newman, 2010). In what follows, we introduce the 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,

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

where $\mathrm{degree}(i, t)$ denotes the degree of node i at time t. We remark that $\mathrm{degree}(i, t)$ exhibits the spatial proximity (in terms of contact frequency) 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.

• Local Fiedler vector centrality (LFVC) (Chen, Choudhury and Hero, 2016). LFVC evaluates the structural importance of a node regarding the network connectivity, and thus is defined through the Fiedler vector,

$$\text{LFVC}(i,t) = \sum_{j \in \{j | (i,j) \in E_t, \forall j\}} ([\mathbf{y}_t]_i - [\mathbf{y}_t]_j)^2, \tag{3}$$

where $\mathbf{y}_t$ is the Fiedler vector of G_t. Since the coefficients of the Fiedler vector imply a powerful heuristic to partition the network into well-separated clusters (implying A/B compartments), LFVC characterizes the nodal significance on the network partition and connectivity.

• Closeness. Closeness is defined by the shortest-path distances of a node to all other nodes,

$$\text{closeness}(i,t) = \frac{1}{\sum_{j \in V, j \neq i} \rho_t(i,j)}, \tag{4}$$

where $\rho_t(i,j)$ denotes the shortest-path distance between node i and node j in the connected network G_t. The distance measure between two nodes is adopted as $1/[\mathbf{H}_t]_{ij}$ in the Hi-C context. The closeness implies how far one node is from the geometrical center of the network.

• Betweenness. Betweenness is the fraction of the number of shortest paths passing 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

$$\text{betweenness}(i,t) = \sum_{k \in V, k \neq i} \sum_{\substack{j \in V \\ j \neq i, j > k}} \frac{\sigma_{kj}(i,t)}{\sigma_{kj}(t)}, \tag{5}$$

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.

• Local clustering coefficient (LCC). . LCC of a node quantifies how close its neighbours are to being a complete graph,

$$\mathrm{LCC}(i,t) = \frac{2|\{(j,k)|(j,k) \in E_t, \forall j,k \in N_i(t)\}|}{|N_i(t)|(|N_i(t)|-1)}, \tag{6}$$

where $N_i(t)$ is the set of neighbors connected with node i, and $|\{(j,k)|(j,k) \in E_t, \forall j,k \in N_i(t)\}|$ denotes the number of edges in the neighborhood of node i. The LCC centrality characterizes the local (well-connected) topological feature.

• Weight of multi-hop walks. The h-hop walk weight of a node is given by the sum of edge weights associated with paths departing from this node and traversing through h edges, which can be computed in an iterative manner (Chen, Choudhury and Hero, 2016)

$$\begin{aligned} &\mathbf{w}_t^{(h+1)} = \mathbf{H}_t\mathbf{d}_t^{(h)} + \mathbf{A}_t\mathbf{w}_t^{(h)},\ \mathbf{d}_t^{(h)} = \mathbf{A}\mathbf{d}_t^{(h-1)}, \\ &\mathbf{d}_t^{(0)} = \mathbf{1},\ \mathbf{w}_t^{(1)} = \mathbf{H}_t\mathbf{1},\ h = 1,2,\ldots, \end{aligned} \tag{7}$$

where $\mathbf{w}_t^{(h)}$ denotes the vector of h-hop walk weights, and $\mathbf{A}_t$ is the binary Hi-C matrix with $[\mathbf{A}_t]_{ij} = 1$ if $[\mathbf{H}_t]_{ij} > 0$, and $[\mathbf{A}_t]_{ij} = 0$ if $[\mathbf{H}_t]_{ij} = 0$. It is often the case that the largest number of hops is chosen as the network diameter, which is the largest shortest-path hop count between any nodal pairs in all connected components of the network. We note that indirect interactions among nodes can be taken into account through h-hop walk.

• Distances to reference nodes (Chen, Choudhury and Hero, 2016). Given nodes of interest (e.g., those with maximal degrees and/or with high RNA-seq values), we can explore network distances of each node to the reference nodes as structural features. The use of network distances helps to avoid ambiguity of centrality measures due to the possibly high structural symmetry in networks.

Integration of form and function

The extracted centrality feature vectors can be then 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 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, the eigenvector centrality, and h-hop walk weights could be strongly correlated, and the replicates of RNA-seq data could also be strongly correlated. Therefore, data points given by rows of $\mathbf{X}$ are lying on or near 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

$$\mathbf{Ly} = \lambda \mathbf{Dy} \tag{8}$$

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 column of $\mathbf{Y}$ and the column 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

$$\text{FS}(i,j) = \frac{|Q(i,j)|}{\sum_i |Q(i,j)|}. \tag{9}$$

In Figure S3, we compare Laplacian eigenmaps with PCA, and demonstrate feature scores associated with the obtained 2D manifolds. We remark that the considered approach is also applicable to finding intrinsic non-linear manifolds of other dimensions (higher than 2D).

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 $X = \{\mathbf{x}_i \,|\, \mathbf{x}_i \in \mathbb{R}^m,\ i \in \{1, 2 \ldots, n\}\}$ denote the dataset of our 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 X can be parametrized as

$$W_{\mathbf{Q},\mathbf{b}} = \{\mathbf{x} \in \mathbb{R}^m \mid \|\mathbf{Q}\mathbf{x} - \mathbf{b}\|_2 \leq 1\}, \tag{10}$$

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 $E_{\mathbf{Q},\mathbf{b}}$ is given by $\mathbf{c} \coloneqq \mathbf{Q}^{-1}\mathbf{b}$, and $\mathbf{\Lambda} \coloneqq \mathbf{Q}^2$ since the ellipsoid (10) can be reformulated as $W_{\mathbf{Q},\mathbf{b}} = \{\mathbf{x} \in \mathbb{R}^m \mid (\mathbf{x} - \mathbf{c})^T \mathbf{\Lambda} (\mathbf{x} - \mathbf{c}) \Sigma \leq 1\}$. Finding the minimum volume ellipsoid can be then cast as a convex program

$$\begin{aligned} \underset{\mathbf{Q},\mathbf{b}}{\text{minimize}} \quad & \det(\mathbf{Q}^{-1}) \\ \text{subject to} \quad & \|\mathbf{Q}\mathbf{x}_i - \mathbf{b}\|_2 \leq 1,\ i = 1, 2, \ldots, n \\ & \mathbf{Q} \text{ is positive definite.} \end{aligned}$$

The MVE estimates the shape of the uncertainty ellipsoid for X, which is different from its sample covariance. The latter is the maximum likelihood estimate under the assumption of Gaussian distribution. We compare MVE with the ellipsoid determined by data covariance in Figure S4.

Evaluation of information divergence

We adopt Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) to measure the information divergence of form and function at different time points. Let f and g denote two distributions with dimension m and known mean and covariance, denoted by $(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$ and $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, respectively. The KL divergence between f and g under the given second-order statistics is calculated as

$$\mathrm{KL}(f, g) = \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\Sigma}_f) + (\boldsymbol{\mu}_g - \boldsymbol{\mu}_f)^T \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\mu}_g - \boldsymbol{\mu}_f) + \log\left(\frac{\det(\Sigma_g)}{\det(\Sigma_f)}\right) - m\right).$$

KL is not symmetric, and it's symmetrized version becomes

$$\mathrm{SKL}(f, g) = \frac{1}{2}(\mathrm{KL}(f, g) + \mathrm{KL}(g, f)).$$

The information divergence can be used to detect the bifurcation point during form-function evolution. In our analysis, f and g can be regarded as distributions of low-dimensional data, representing form-function features. The resulting second-order statistics are approximated by the center and the shape of the fitted ellipsoid using MVE.

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

$$\mathrm{TDS}(i,t) = \frac{\sum_{t'\in N_t} \mathrm{dist}([\mathbf{X}_t]_i, [\mathbf{X}_{t'}]_i)}{|N_t|}, \tag{11}$$

where N_t defines the time window around t, e.g., $N_t = \{t-1, t, t+1\}$ of size 3 or $N_t = \{t-1, t\}$ of size 2, and $\mathrm{dist}(\cdot)$ is a generic distance function between $[\mathbf{X}_t]_i$ and $[\mathbf{X}_{t'}]_i$, e.g., Euclidean distance or KL divergence. By dropping the node index, TDS in (11) can be used to assess the network-scale temporal difference,

$$\mathrm{TDS}(t) = \frac{\sum_{t'\in N_t} \mathrm{dist}(\mathbf{X}_t, \mathbf{X}_{t'})}{|N_t|}.$$

In our analysis, $\mathbf{X}_t$ represents either form-function feature matrix or it low-dimensional data representation.

Bifurcation: branching trajectory and statistical significance

To depict the branching trajectory at bifurcation, we studied the dataset of human fibroblast proliferation (Chen et al., 2015) and the dataset of the direct cellular reprogramming 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 is given by the principal subspace of form-function data at the first two time points for both proliferation and reprogramming (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 fitted ellipsoids (given by MVE estimates) over time. The trajectory of centroids is then smoothed

using the cubic spline. Lastly, we provided a statistical significance of the bifurcation, where P value is defined from the multivariate Hotelling's T-Square test associated with the null hypothesis that the centroids of the proliferation and reprogramming are identical at a given time point.

A/B compartment switching analysis

A/B compartments were identified through methods conceptually similar to that described in (Chen, Hero and Rajapakse, 2016). 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 HiC matrix). The entrywise sign of the Fiedler vector of the graph Laplacian associated with $\tilde{\mathbf{H}}$ is used to identify A/B compartments. This is calculated for each sample (12 time points) and 100-kb bins that change in Fiedler vector sign between consecutive time points (e.g. A to B, or B to A) are recorded. The percentage of A/B compartment changes between each time point, and the total percentage of bins that change A/B compartment between any consecutive time points is recorded.

Bifurcation identification at single gene level

Raw contacts from Hi-C 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 the absolute value difference 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, my-

otube, 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 S1).

Identification of gene clusters for form-function change indicator

Given genes of interest, we perform K-means (with K = 12) to their structural/functional temporal difference scores over the 80-hr time course. By merging clusters that have similar form-function changes and excluding clusters that only have subtle form-function changes, we obtain gene clusters associated with time points at which form and function maintain significant changes. The number of genes within each cluster is summarized in Figure 4D.

Statistical significance for temporal change of form and function

At TAD scale, the significance test is made by comparing the quantities of TADs of interest (e.g., gene density and gene expression in Figure 3C and average TADs' position shift in Figure 3D) to a random background distribution. The background distribution is generated by randomly selecting the same number of TADs to record the quantities of interest over 1000 trials. The probability of the resulting right-tailed event is used as P value. The same idea can be applied to the significance test built on form-function dynamics at other scales (e.g. gene level in Figure 5D).

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 identified bifurcation point (32 Hours) 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(RPKM) over time was calculated and sorted for each gene.

DATA AND SOFTWARE AVAILABILITY

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

SUPPLEMENTAL ITEM TITLES AND LEGENDS

Table S1

Title: Gene modules extracted from Gene Ontology (GO). Related to Figure 3, 4, 5, and 7.

Table S2

Title: Core subset of myogenic genes that steer cellular reprogramming. Related to Figure 5.

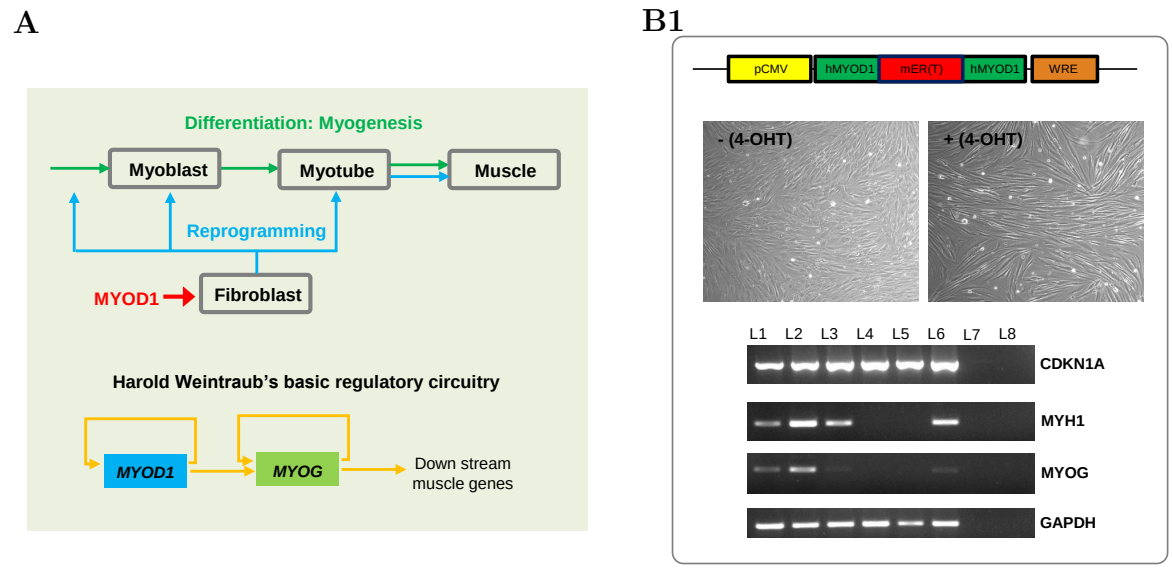
Table S3

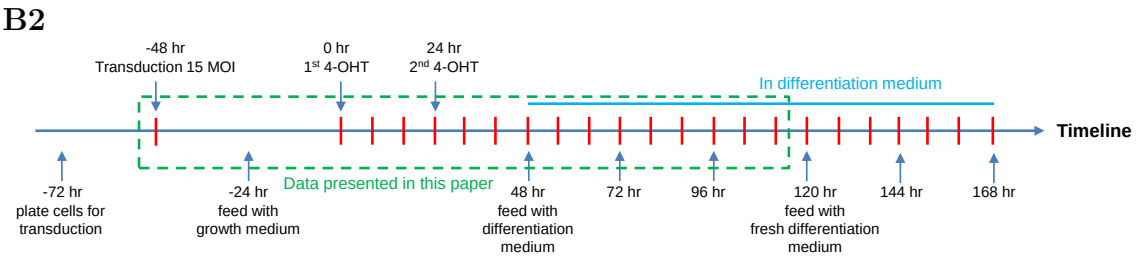
Title: List of miRNAs that significantly change expression level over the reprogramming time course. Related to Figure S8.

Table S4

Title: JTK output for E-box circadian genes. Related to Figure 7B2.

Main Figures 1-7





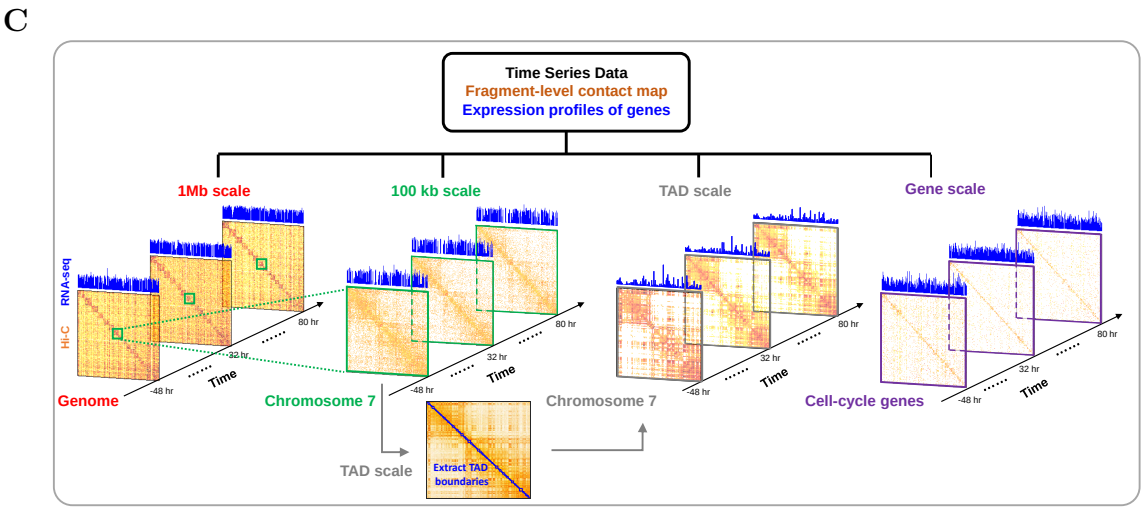


Figure 1

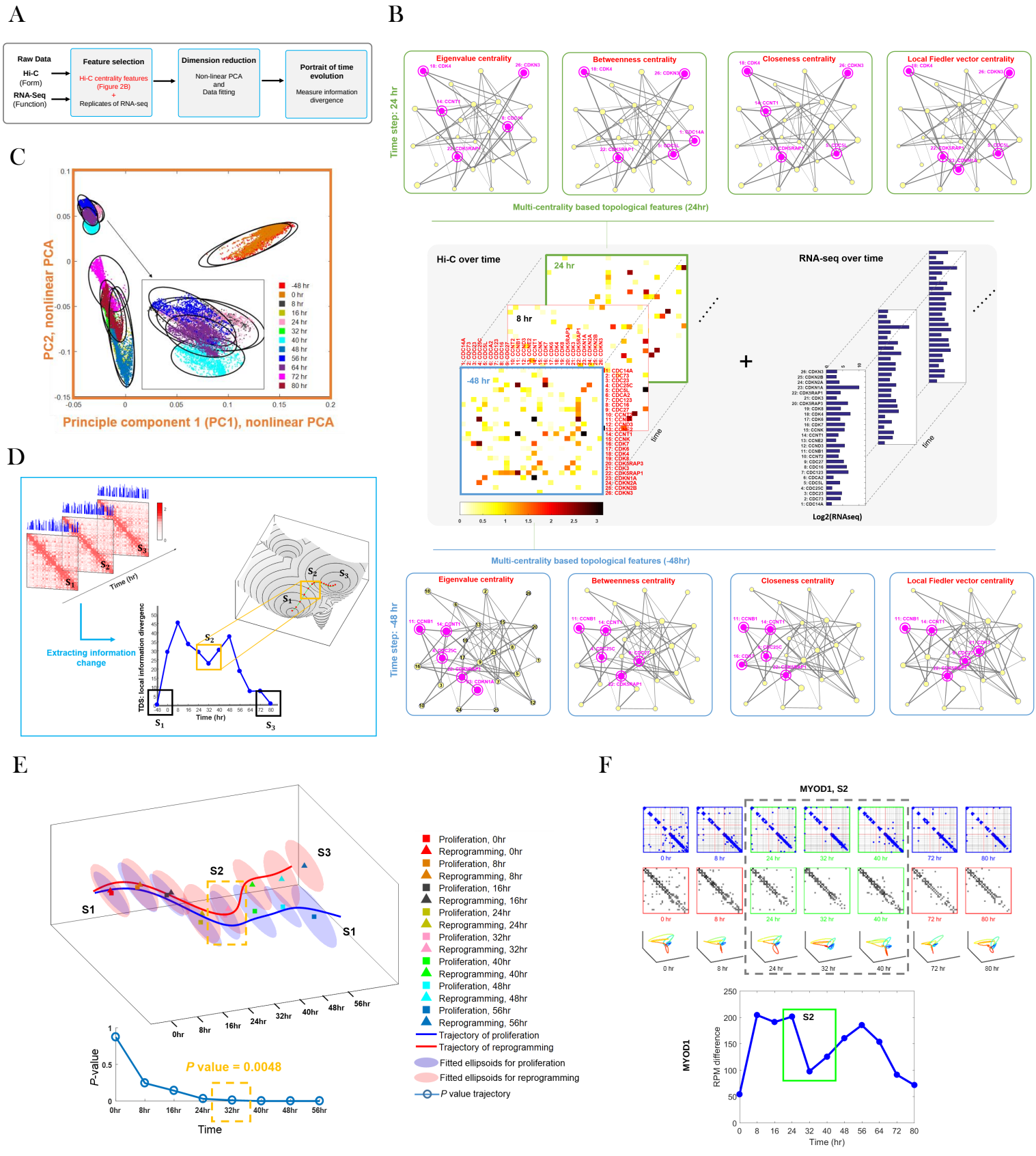


Figure 2

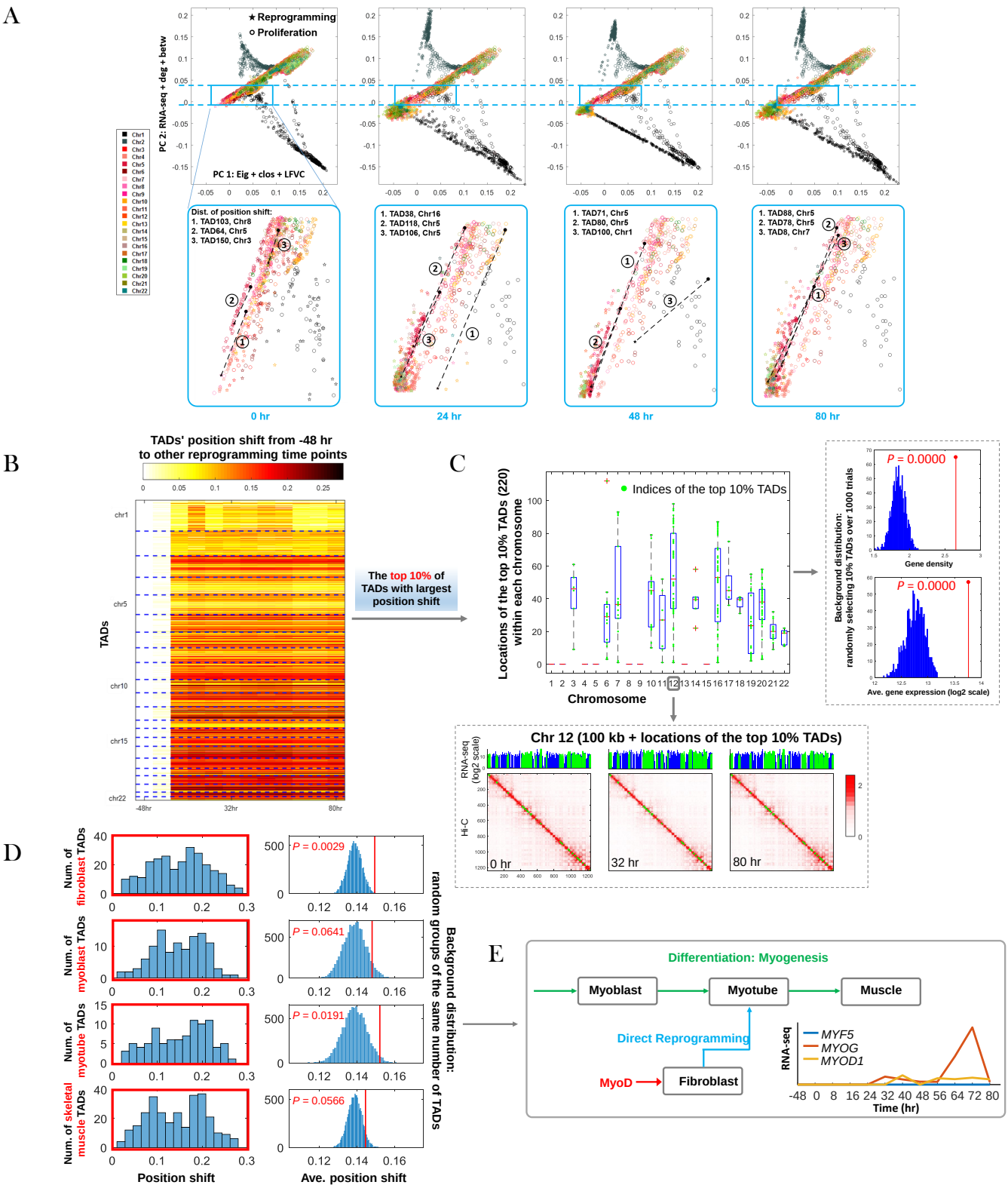


Figure 3

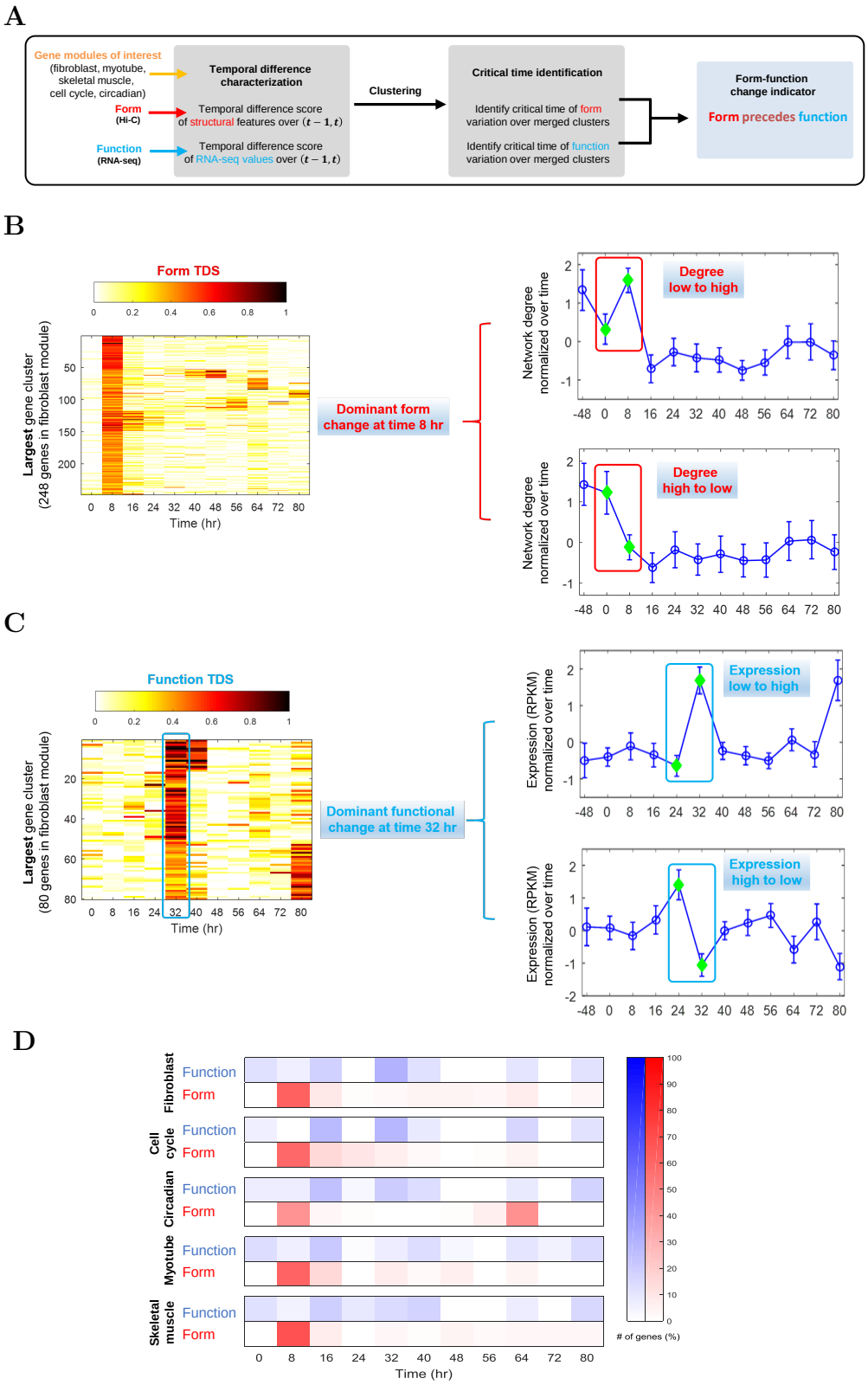


Figure 4

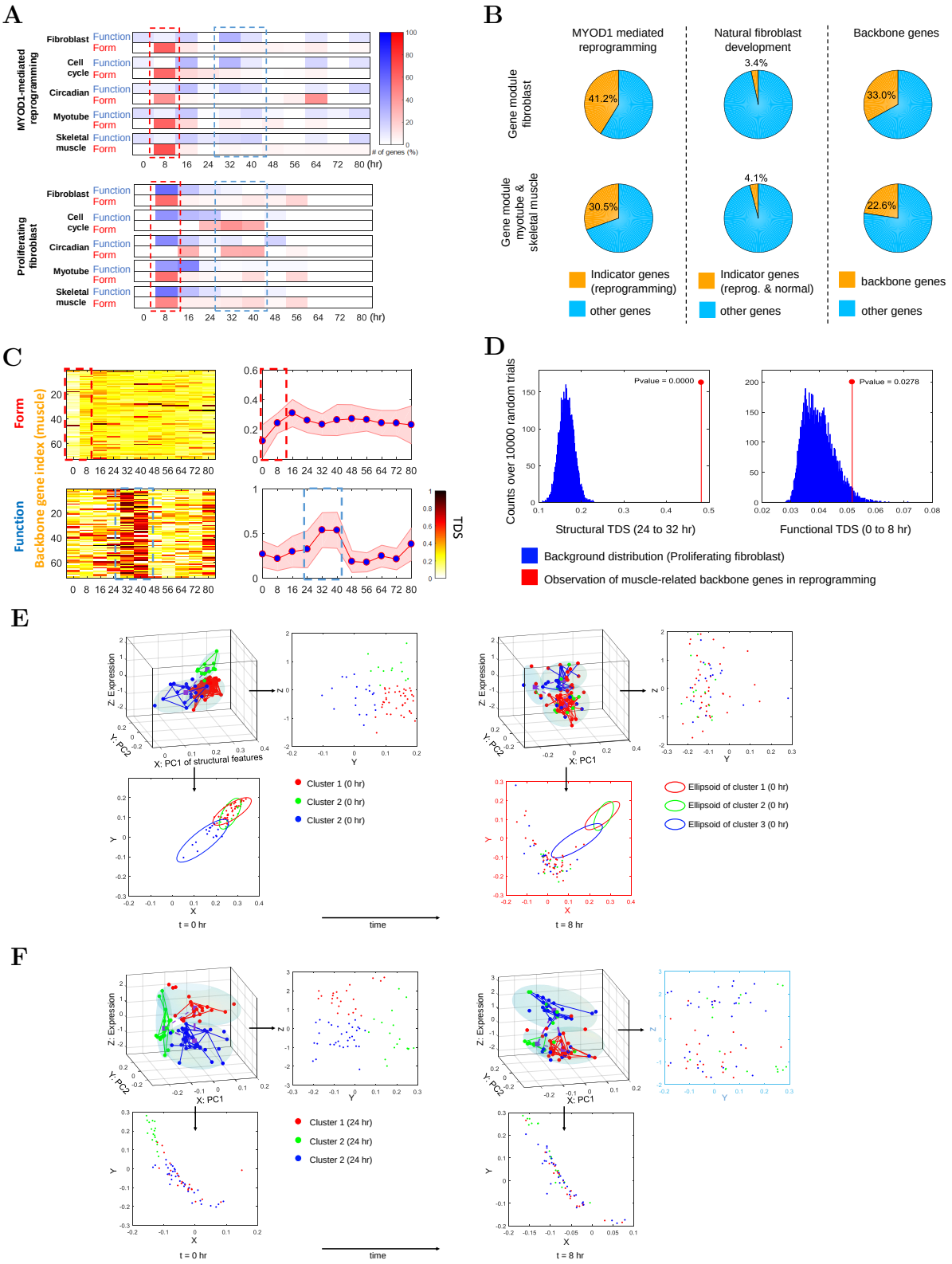


Figure 5

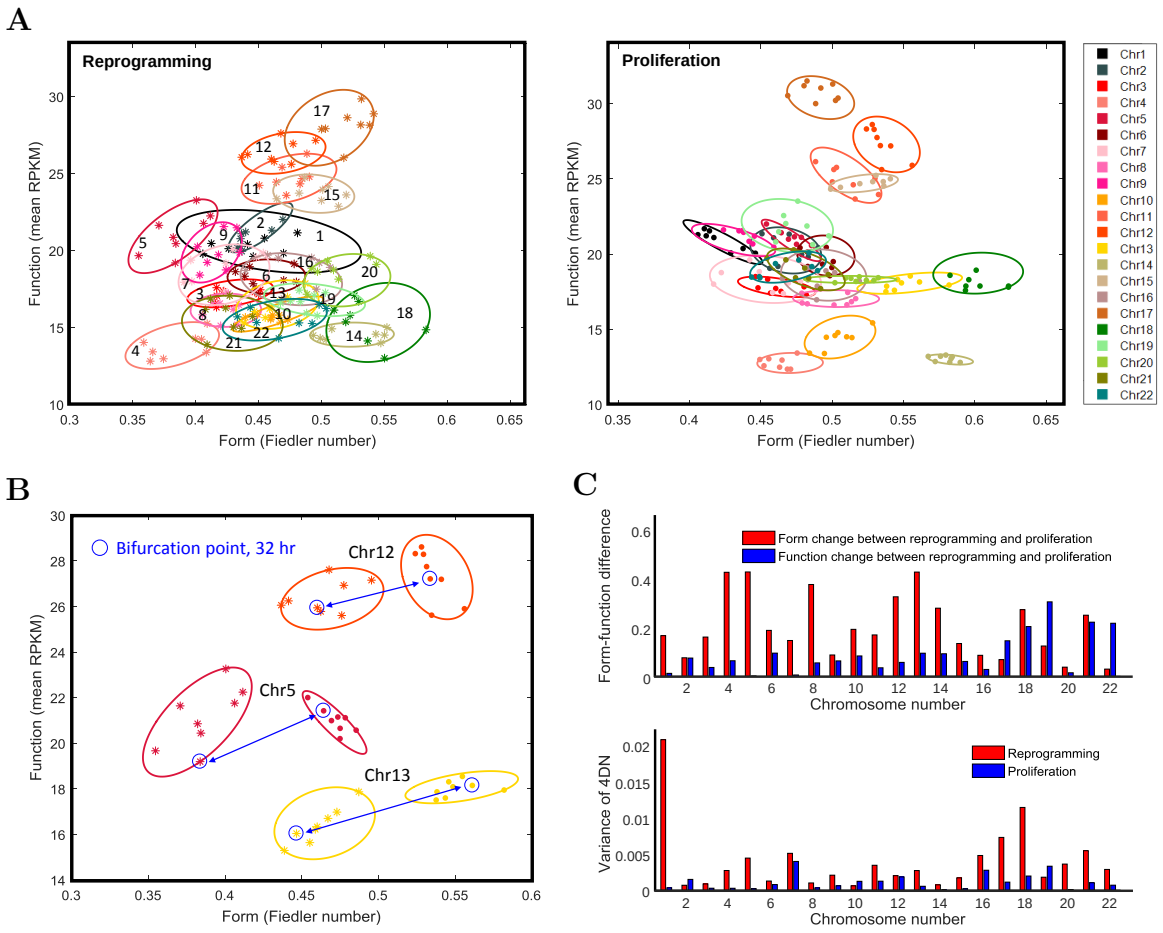


Figure 6

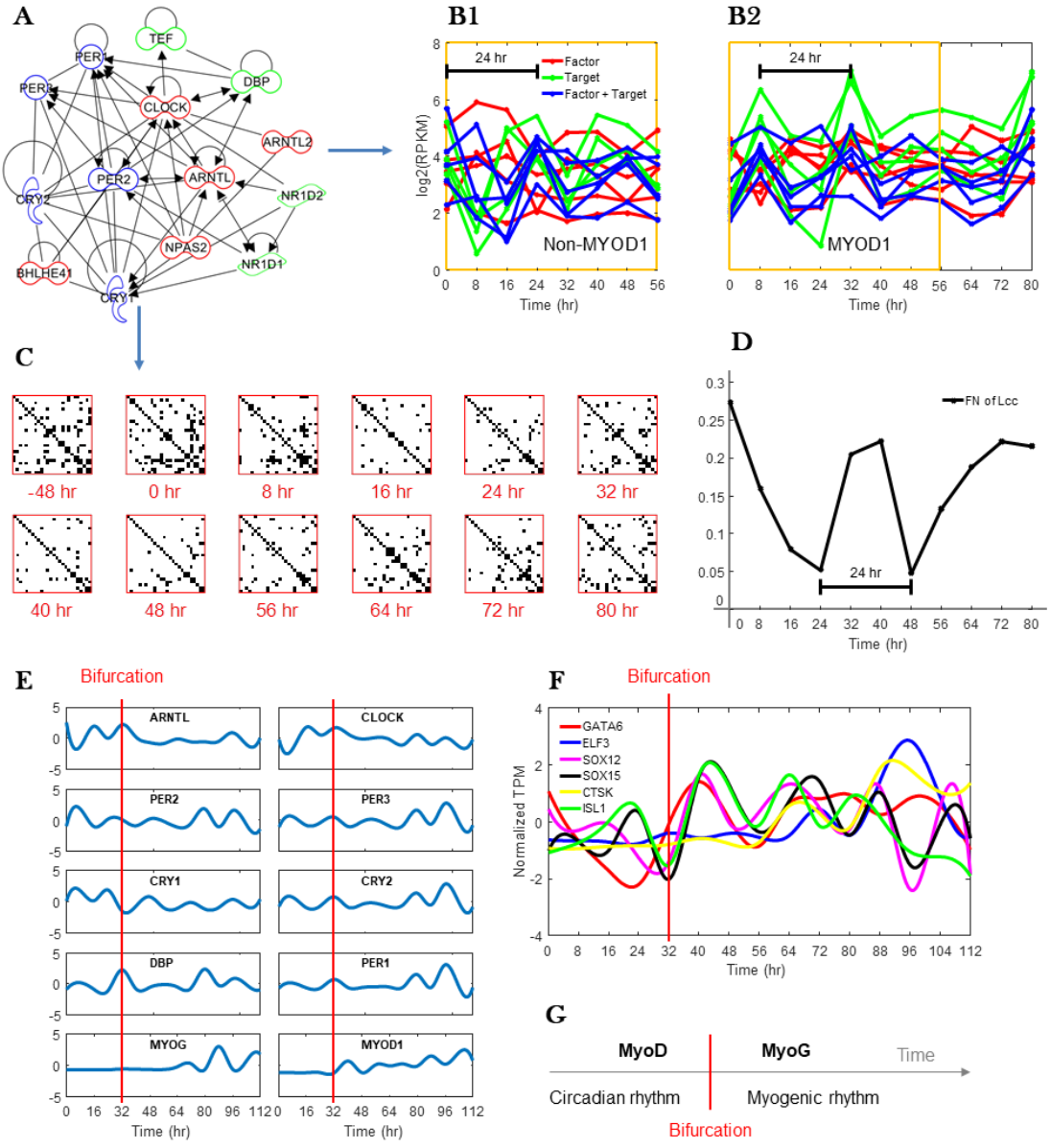


Figure 7

Supplemental Information: Figures S1-S8

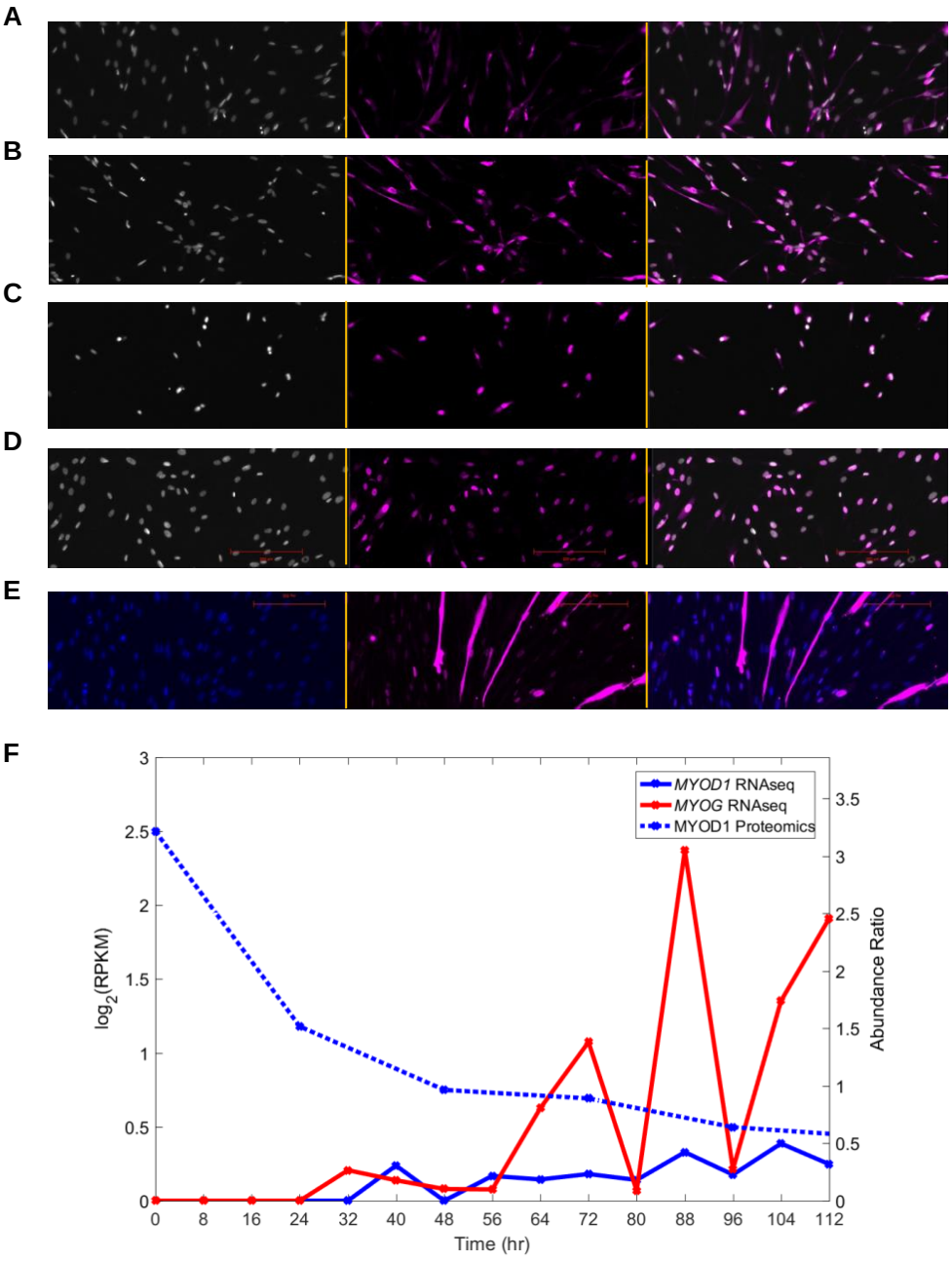
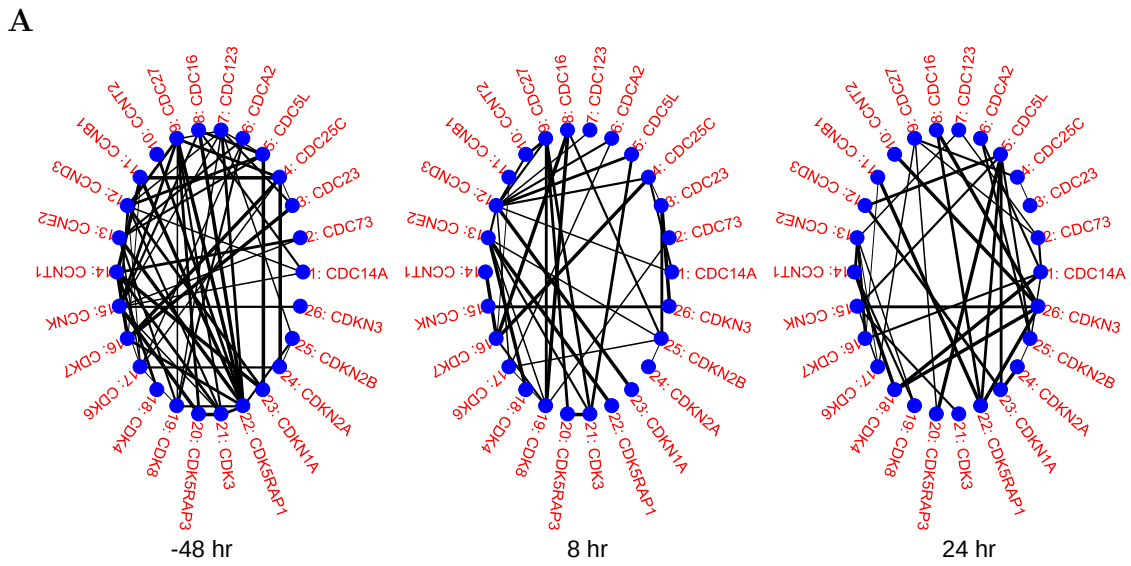


Figure S1



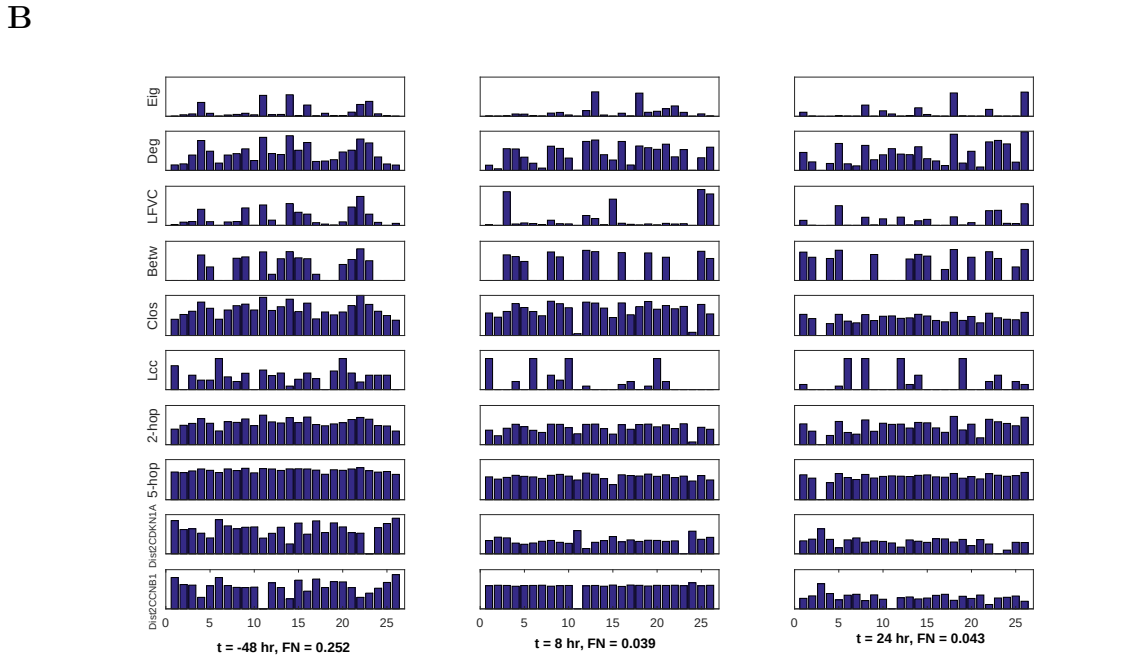


Figure S2

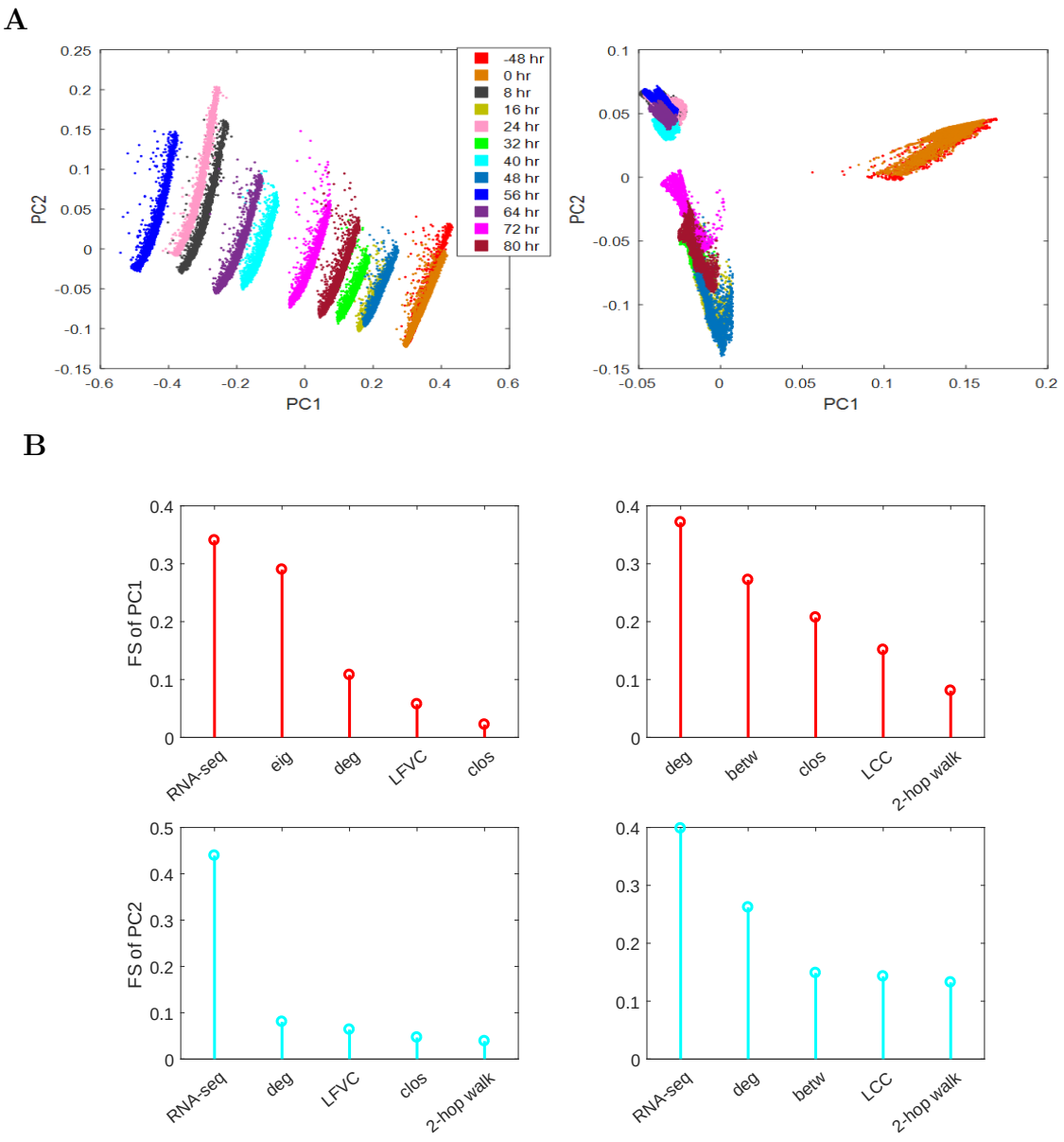
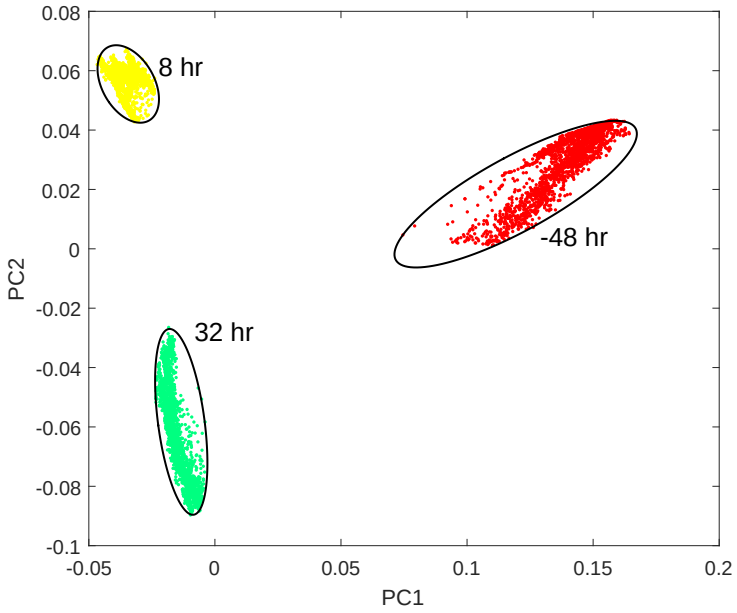


Figure S3

A



B

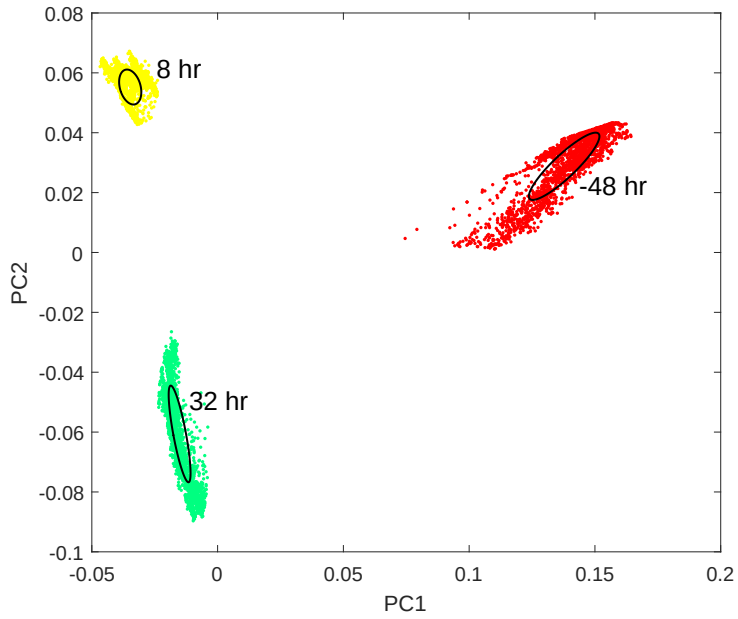


Figure S4

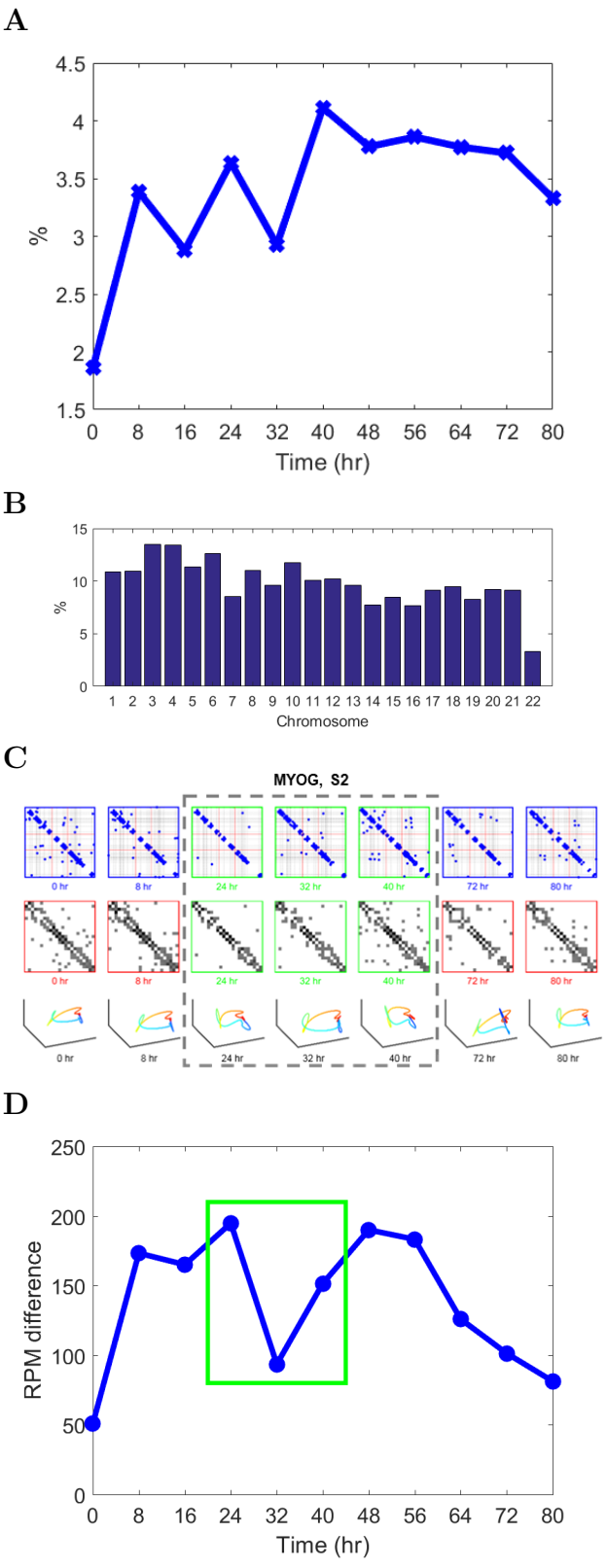


Figure S5

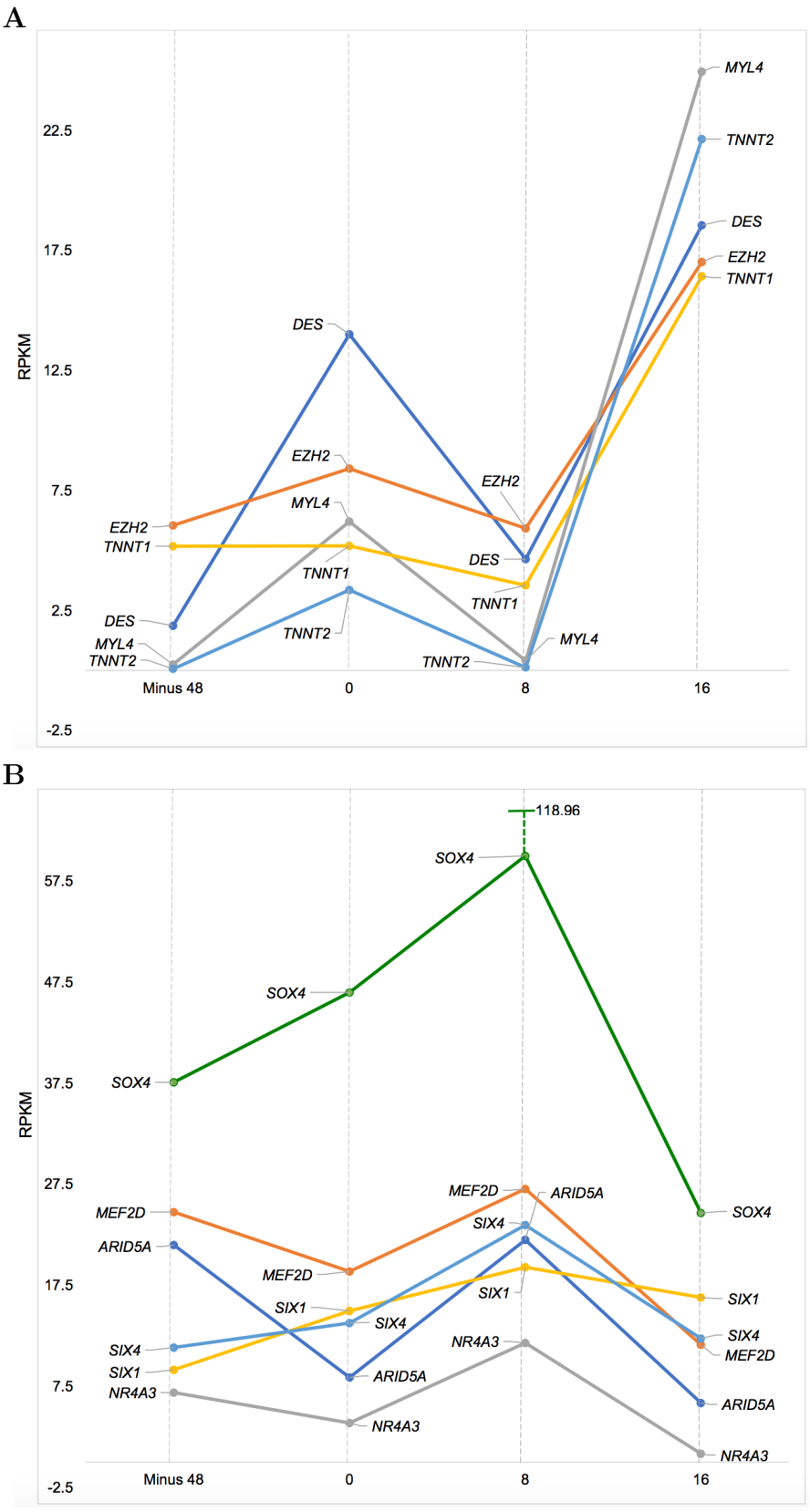


Figure S6

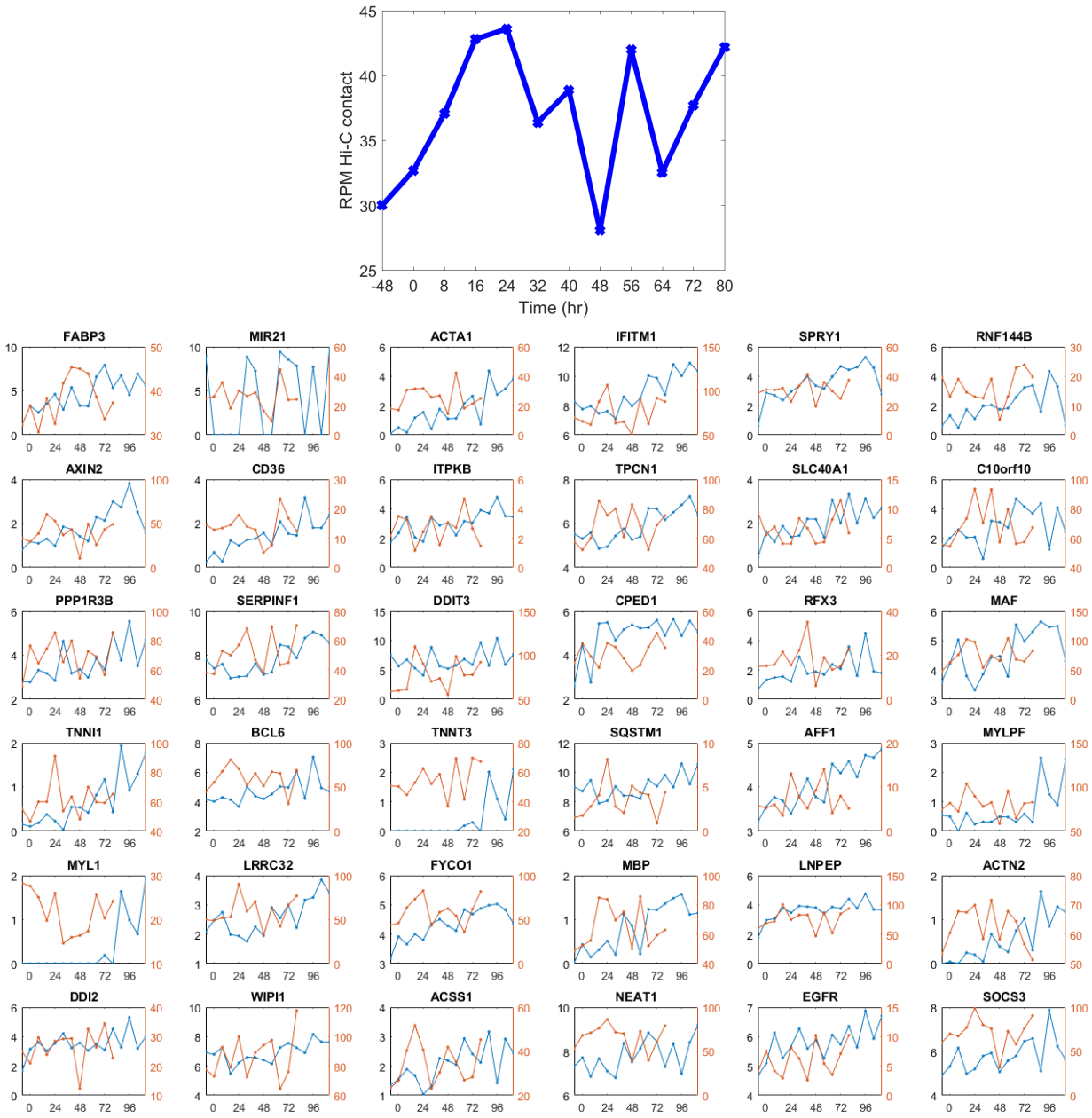


Figure S7

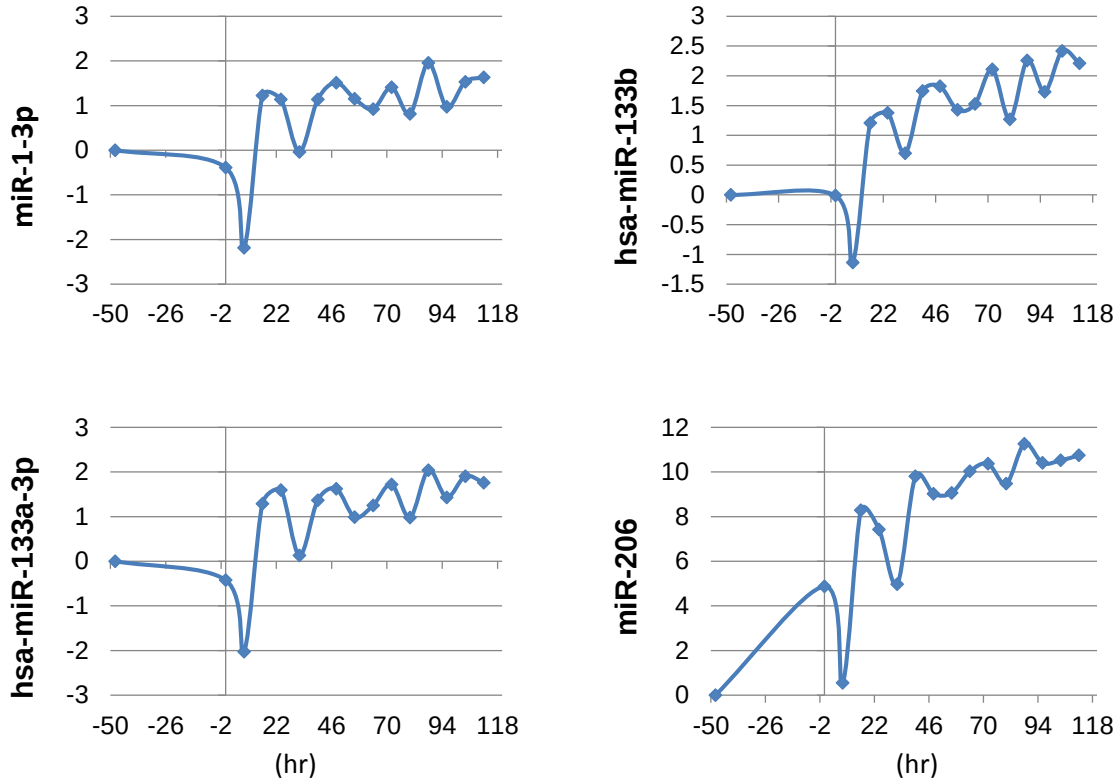


Figure S8