The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[2], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]

Defining the transcriptional dynamics of a temporal process such as cell differentiation is challenging owing to the high variability in gene expression between individual cells. Time-series gene expression analyses of bulk cells have difficulty distinguishing early and late phases of a transcriptional cascade or identifying rare subpopulations of cells, and single-cell proteomic methods rely on a priori knowledge of key distinguishing markers[1]. Here we describe Monocle, an unsupervised algorithm that increases the temporal resolution of transcriptome dynamics using single-cell RNA-Seq data collected at multiple time points. Applied to the differentiation of primary human myoblasts, Monocle revealed switch-like changes in expression of key regulatory factors, sequential waves of gene regulation, and expression of regulators that were not known to act in differentiation. We validated some of these predicted regulators in a loss-of function screen. Monocle can in principle be used to recover single-cell gene expression kinetics from a wide array of cellular processes, including differentiation, proliferation and oncogenic transformation.

Cellular processes such as proliferation, differentiation and reprogramming are governed by complex gene-regulatory programs. Progress through these processes is a function not only of time but also of cell-cell signaling and other stimuli. During differentiation, for example, each cell makes independent fate decisions by integrating a wide array of signals from other cells and executing a complex choreography of gene-regulatory changes. Thus individual cells can execute the same sequence of transcriptional changes over highly varying time scales. Unraveling the network of gene regulatory interactions remains a central goal of efforts to understand these processes.

Recently, several studies carried out at single-cell resolution revealed high cell-to-cell variation in the expression of most genes, even key developmental regulators, during the differentiation process[2–6]. Such high variability can complicate analysis of these experiments[7]. In general, averages of measurements from two or more distinct groups of data points can follow trends that qualitatively differ from the trend that describes each group, a phenomenon known as Simpson's paradox[8]. Such averaging artifacts can make factors that are correlated appear to be uncorrelated or even make positively correlated factors appear negatively correlated. As a population of cells captured at the same time may include many distinct intermediate differentiation states, considering only its average properties would mask trends occurring across individual cells. Solving this problem by experimental synchronization of cells or by stringent isolation of precursors at distinct stages is challenging and can sharply alter differentiation kinetics.

Computational analysis of gene expression data could help define biological progression between cellular states and reveal regulatory modules of genes that co-vary in expression across individual cells[9]. Previous analyses have used approaches from computational geometry[10,11] to order bulk cell populations from time-series microarray experiments by progress through a biological process independently of when the samples were collected. The recently developed SPD algorithm can resolve progression along multiple lineages arising from a progenitor cell type using supervised machine learning[12]. However, because these algorithms operate on bulk expression measurements, they are sensitive to mixture effects arising from Simpson's paradox and other averaging artifacts. Single-cell assays such as flow or mass cytometry[1], coupled with machine learning algorithms such as SPADE[13], can overcome these effects to reconstruct complex lineages and resolve intermediate stages of progress through differentiation. Coupled with SPADE, cytometry can track changes in up to 32 proteins to reconstruct complex cellular hierarchies of differentiation and reveal rare cell states. In principle, single-cell RNA-Seq could also be used to resolve cellular transitions during differentiation through temporal profiling of the entire transcriptome.

We hypothesized that ordering whole-transcriptome profiles of single cells with an unsupervised algorithm would improve temporal resolution during differentiation without a priori knowledge of marker genes. In essence, one RNA-Seq experiment would constitute a time series, with each cell representing a distinct time point along a continuum. Monocle is derived from a previous algorithm[10] for temporally ordering bulk microarray samples but extends it to accommodate single-cell variation and to allow for multiple cell fates stemming from a single progenitor cell type. Monocle

[1]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Harvard Stem Cell Institute, Harvard University, Cambridge, Massachusetts, USA. [4]Fluidigm Corporation, South San Francisco, California, USA. [5]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.L.R. (john_rinn@harvard.edu).

Received 10 November 2013; accepted 25 February 2014; published online 23 March 2014; doi:10.1038/nbt.2859

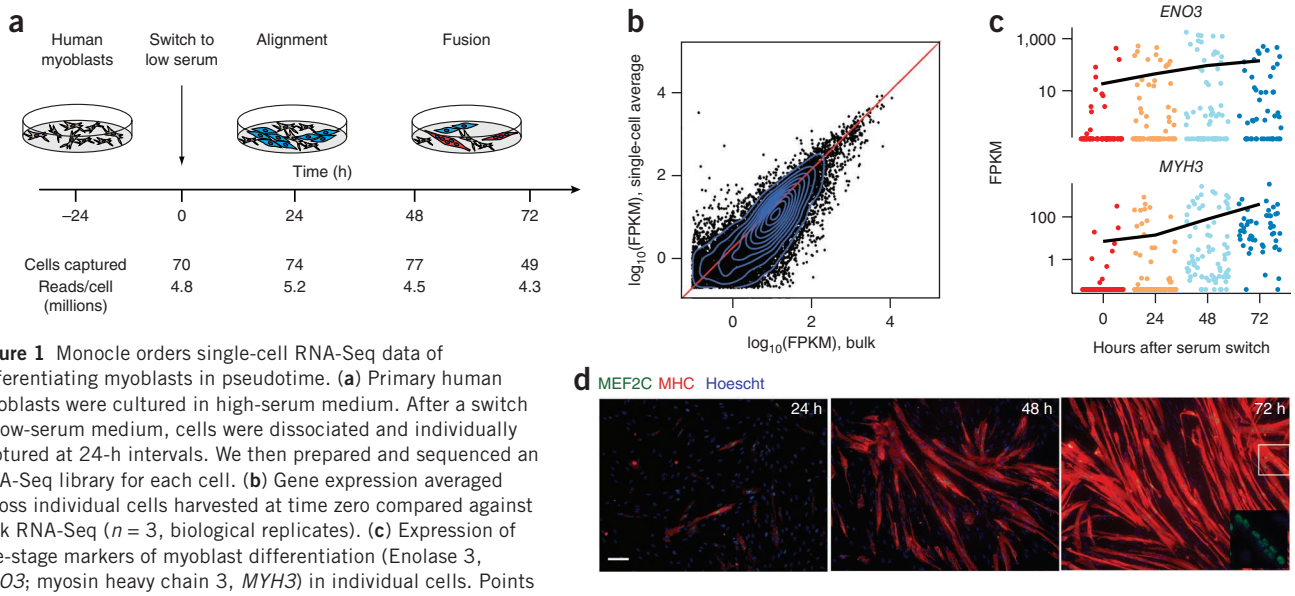

Figure 1 Monocle orders single-cell RNA-Seq data of differentiating myoblasts in pseudotime. (**a**) Primary human myoblasts were cultured in high-serum medium. After a switch to low-serum medium, cells were dissociated and individually captured at 24-h intervals. We then prepared and sequenced an RNA-Seq library for each cell. (**b**) Gene expression averaged across individual cells harvested at time zero compared against bulk RNA-Seq ($n = 3$, biological replicates). (**c**) Expression of late-stage markers of myoblast differentiation (Enolase 3, *ENO3*; myosin heavy chain 3, *MYH3*) in individual cells. Points are colored by time collected (0 h, red; 24 h, gold; 48 h, light blue; 72 h, dark blue). (**d**) Representative immunofluorescence staining at the moment of cell sampling of the indicated markers (myocyte enhancer factor 2C, MEF2C in green; myosin heavy chain, MYH2/MHC in red; Hoechst staining in blue; scale bar, 100 μm). Inset is a magnification of the boxed region, showing MEF2C only.

orders single-cell expression profiles in 'pseudotime'—a quantitative measure of progress through a biological process.

We began by investigating the single-cell transcriptome dynamics of skeletal myoblasts during differentiation. Skeletal myoblasts undergo a well-characterized sequence of morphological and transcriptional changes during differentiation[14]. Global expression and epigenetic profiles have reinforced the view that a small cohort of transcription factors (for example, MYOD, MYOG, MRF4 and MYF5) orchestrates these changes[15]. However, efforts to expand this set of factors and map the broader myogenic regulatory network have been hampered by the low temporal resolution of global expression measurements, with thousands of genes following a limited number of coarse kinetic trends[16]. We expanded primary human skeletal muscle myoblasts (HSMM) under high-mitogen conditions (GM) and induced differentiation by switching to low-serum medium (DM). We captured between 49 and 77 cells at each of four time points after the switch to DM using the Fluidigm C_1 microfluidic system. RNA from each cell was isolated and used to construct a single mRNA-Seq library per cell, which was then sequenced to a depth of ~4 million reads per library, resulting in a complete gene expression profile for each cell (**Fig. 1a** and **Supplementary Fig. 1**).

We first confirmed that the average of the expression profiles of single cells collected at the same time correlated well with bulk RNA-Seq libraries at those times, and moderately expressed genes were detectable (≥1 fragment per kb per million mapped fragments (FPKM)) in most individual cells (**Fig. 1b** and **Supplementary Figs. 2** and **3**). However, markers of mature myocytes were present at all time points after serum switch, and many other genes showed similar temporal heterogeneity (**Fig. 1c**). We speculated that the high variability in cell-to-cell gene expression levels was due to unsynchronized differentiation, with myoblasts, intermediate myocytes and mature myotubes residing in the same well. Indeed, large, multinucleated cells expressing *MYH2*, a marker of mature myotubes, were abundant after 72 h in DM, but these cells were present at lower frequency even at 24 h (**Fig. 1d**).

We developed Monocle (Online Methods and **Supplementary Source Code**) to informatically order the cells by their progress through differentiation rather than by the time they were collected, maximizing the transcriptional similarity between successive pairs of cells (**Fig. 2a**). First the algorithm represents the expression profile of each cell as a point in a high-dimensional Euclidean space, with one dimension for each gene. Second, it reduces the dimensionality of this space using independent component analysis[17]. Dimensionality reduction transforms the cell data from a high-dimensional space into a low-dimensional one that preserves essential relationships between cell populations but is much easier to visualize and interpret[18]. Third, Monocle constructs a minimum spanning tree (MST) on the cells, a previously developed approach now commonly used in other single-cell settings, such as flow or mass cytometry[1,13]. Fourth, the algorithm finds the longest path through the MST, corresponding to the longest sequence of transcriptionally similar cells. Finally, Monocle uses this sequence to produce a 'trajectory' of an individual cell's progress through differentiation.

As cells progress along a differentiation trajectory, they may diverge along two or more separate paths. After Monocle finds the longest sequence of similar cells, it examines cells not along this path to find alternative trajectories through the MST. It orders these subtrajectories and connects them to the main trajectory, and annotates each cell with both a trajectory and a pseudotime value. Monocle thus orders cells by progress through differentiation and can reconstruct branched biological processes, which might arise when a precursor cell makes cell fate decisions that govern the generation of multiple subsequent lineages. Importantly, Monocle is unsupervised and needs no prior knowledge of specific genes that distinguish cell fates, and is thus suitable for studying a wide array of dynamic biological processes.

Monocle decomposed myoblast differentiation into a two-phase trajectory and isolated a branch of nondifferentiating cells (**Fig. 2b**). The first phase of the trajectory was primarily composed of cells collected under high-mitogen conditions and that expressed markers of actively proliferating cells, such as *CDK1*, whereas the second mainly consisted of cells collected 24, 48 or 72 h after serum switch. Cells in the second phase were positive for markers of muscle differentiation such as *MYOG* (**Supplementary Fig. 4**). A tightly grouped third

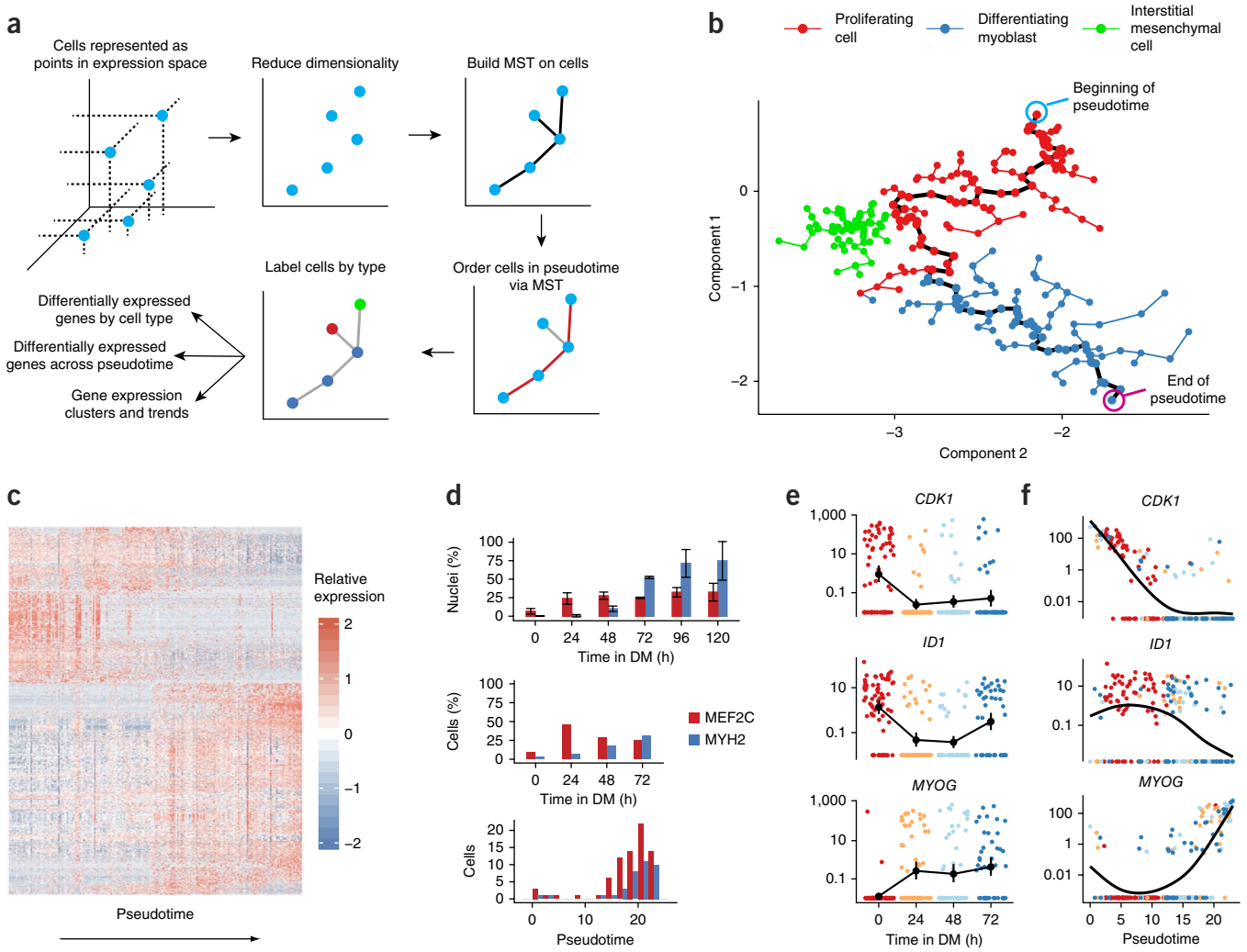

Figure 2 Monocle orders individual cells by progress through differentiation. (**a**) An overview of the Monocle algorithm. (**b**) Cell expression profiles (points) in a two-dimensional independent component space. Lines connecting points represent edges of the MST constructed by Monocle. Solid black line indicates the main diameter path of the MST and provides the backbone of Monocle's pseudotime ordering of the cells. (**c**) Expression for differentially expressed genes identified by Monocle (rows), with cells (columns) shown in pseudotime order. Interstitial mesenchymal cells are excluded. (**d**) Bar plot showing the proportion of MEF2C- and MYH2-expressing cells measured by immunofluorescence at the time of collection (top), RNA-Seq at the time of collection (middle) or RNA-Seq at pseudotime (bottom). MEF2C was considered detectably expressed at or above 100 FPKM, MYH2 at 1 FPKM. MEF2C exhibits a bimodal pattern of expression across the cells (not shown), and a threshold of 100 FPKM separates the modes. (**e**) Expression of key regulators of muscle differentiation, ordered by time collected (cyclin-dependent kinase 1, *CDK1*; inhibitor of DNA binding 1, *ID1*; myogenin, *MYOG*). (**f**) Regulators from **e**, ordered by Monocle in pseudotime. Points in **e**,**f** are colored by time collected (0 h, red; 24 h, gold; 48 h, light blue; 72 h, dark blue). Error bars, 2 s.d.

population of cells branched from the trajectory near the transition between phases. These cells lacked myogenic markers but expressed *PDGFRA* and *SPHK1*, suggesting that they are contaminating interstitial mesenchymal cells and did not arise from the myoblasts. Such cells were recently shown to stimulate muscle differentiation[19]. Monocle's estimates of the frequency and proliferative status of these cells were consistent with estimates derived from immunofluorescent stains against ANPEP (also known as CD13) and nuclear Ser10-phosphorylated histone H3 (**Supplementary Fig. 4**). Monocle thus enabled analysis of the myoblast differentiation trajectory without subtracting these cells by immunopurification, maintaining *in vitro* differentiation kinetics that resemble physiological cell crosstalk occurring in the *in vivo* niche.

To find genes that were dynamically regulated as the cells progressed through differentiation, we modeled expression of each gene as a nonlinear function of pseudotime. A total of 1,061 genes were dynamically regulated during differentiation (false discovery rate (FDR) < 5%; **Fig. 2c**). Cells positive for *MEF2C* and *MYH2*, early and late markers of differentiation, respectively, were present at expected frequencies as assayed by both immunofluorescence and RNA-Seq. Moreover, the pseudotime ordering of cells shows an increase in *MEF2C*$^+$ cells before the increase in *MYH2*$^+$ cells (**Fig. 2d**). Notably, genes that act at the early and late stages of muscle differentiation showed pseudotemporal kinetics that were highly consistent with expectations, with cell-cycle regulators active early in pseudotime and sarcomere components active later, confirming the accuracy of the ordering (**Supplementary Fig. 5**).

We next examined the pseudotemporal kinetics of a set of genes whose mouse orthologs are targeted by Myod, Myog or Mef2 proteins in C2C12 myoblasts[20] (**Supplementary Fig. 6**). The kinetics of these genes during differentiation were highly consistent with changes observed during mouse myogenesis, with nearly all significantly

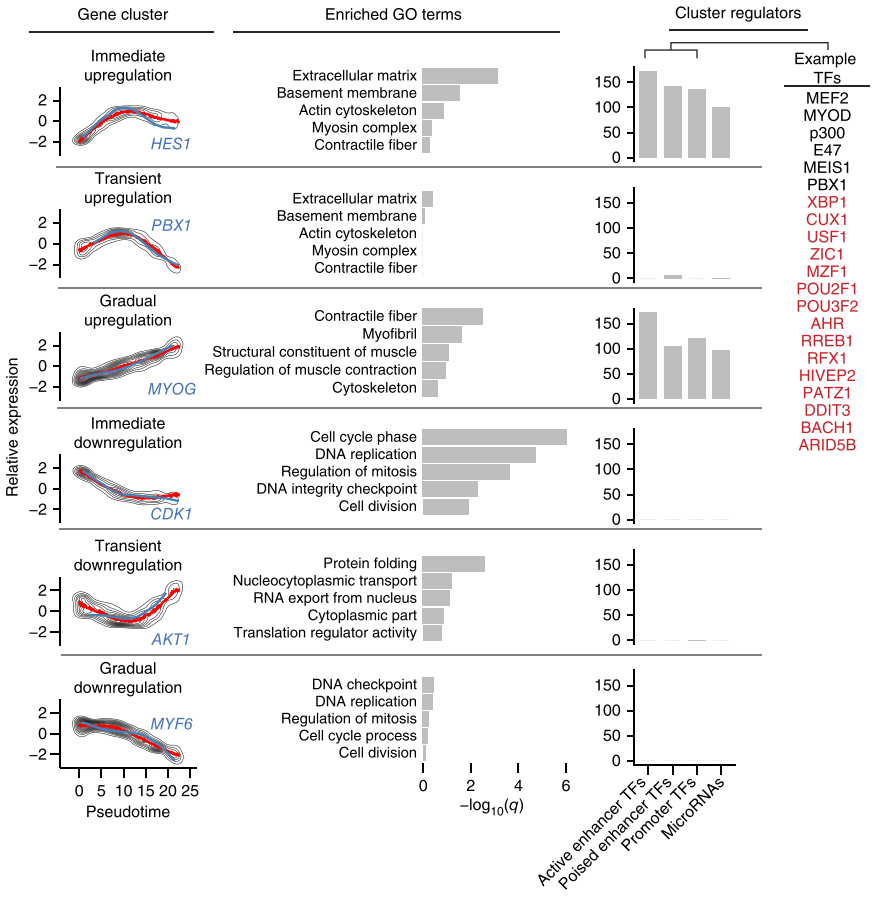

Figure 3 Pseudotime ordering of cells reveals genes activated or repressed early in differentiation, along with potential upstream regulators. Left, relative gene expression levels were *K*-medioids clustered. The mean expression for each cluster is shown in red, and an example gene with a known role in myogenesis from each cluster is highlighted in blue. Middle, selected Gene Ontology terms associated with genes in each cluster. Enrichment scores are shown as $-\log_{10}(q)$, where q is the significance of the enrichment after multiple testing (Online Methods). Right, number of transcription factors (TFs) with conserved binding site motifs in regulatory elements for genes in each cluster. Transcription factors are segregated according to the function of regulatory elements to which they bind. Examples are shown on the right, with known myogenic factors in black and factors without a known role in muscle differentiation in red.

dynamically regulated genes also differentially expressed during mouse myogenesis and vice versa. In contrast to the high resolution of pseudotime ordering, simply comparing gene expression levels between groups of cells collected on different days masked changes in key transcriptional regulators of myogenesis. For example, the pseudotime reordering of the cells showed switch-like inactivation of *ID1*, which is a critical event in muscle differentiation and leads to the activation of *MYOG*[15] (**Fig. 2e**,**f**). Thus, Monocle's ordering of cells by progress increases temporal resolution of transcriptional dynamics, pinpointing the timing of key regulatory events that govern differentiation.

We further assessed Monocle's robustness over different experimental designs by simulating experiments with fewer captured cells. Monocle placed subsets as small as 50 cells in pseudotemporal order highly similar (Spearman ≥ 0.8) to their relative order in the full data set. The algorithm retained the ability to detect dynamically regulated genes with high precision (≥95%) over all designs and with increasing recall as more of the cells were included (**Supplementary Fig. 7**).

We next grouped genes with similar trends in expression, reasoning that such groups might share biological functions and regulators. Clustering of genes according to direction and timing revealed six distinct trends (**Fig. 3**). Genes downregulated early or upregulated late in pseudotime were highly enriched for biological processes central to myogenesis, including cell-cycle exit and activation of muscle-specific structural proteins. However, the other clusters included many genes with broad roles in development, including mediators of cell-cell signaling, RNA export and translational control, and remodeling of cell morphology (**Supplementary Fig. 8**).

A time-series analysis of myoblast differentiation with bulk RNA-Seq identified up- and downregulated genes but did not identify the transient clusters or distinguish the early from late regulation visible with pseudotemporally ordered single cells (**Supplementary Fig. 9**). Furthermore, dynamic range of expression was compressed for most genes, confirming that failure to account for variability in progress through differentiation leads directly to the effects associated with Simpson's paradox. Pseudotemporal cell ordering thus decomposes the coarse kinetic trends produced by bulk RNA-Seq into distinct, sequential waves of transcriptional reconfiguration.

To identify factors driving myoblast differentiation, we performed a *cis*-regulatory analysis on genes in each pseudotemporal cluster. We identified *cis*-regulatory elements on the basis of DNase I hypersensitive sites in HSMM cells and HSMM-derived myotubes[21], classified them according to function on the basis of histone marks[22] and annotated them with conserved transcription factor binding sites. While downregulated genes were enriched at near-significant levels with binding sites for proteins that affect proliferation (for example, MAX, E2F and NMYC), nearly all significantly enriched motifs fell near upregulated genes. These genes were highly enriched for regulatory elements containing binding motifs for 175 transcription factors, including many well-known regulators of myogenesis, such as MYOD, MYOG, PBX1, MEIS1 and the MEF2 family (**Supplementary Fig. 10**). Some, but not all, of these factors were revealed by a regulatory element analysis performed using bulk RNA-Seq data, underscoring the power of increased pseudotemporal resolution of single-cell analysis (**Supplementary Fig. 11**). A similar analysis of microRNA target sites identified miR-1, miR-206, miR-133 and many others as regulators of genes activated during myogenesis (**Supplementary Fig. 12**). Of these, only miR-1 and miR-206 target sites were significantly enriched among genes found to be transiently upregulated and then sharply downregulated. This may suggest that miR-1 and miR-206, which are expressed at an intermediate stage of myoblast differentiation, may act to strongly suppress a subset of genes activated earlier.

Many of the transcription factors implicated by our *cis*-regulatory analysis as governing differentiation had no previously appreciated role in muscle development. To test potential functions of these factors, we performed an RNA interference–mediated loss-of-function

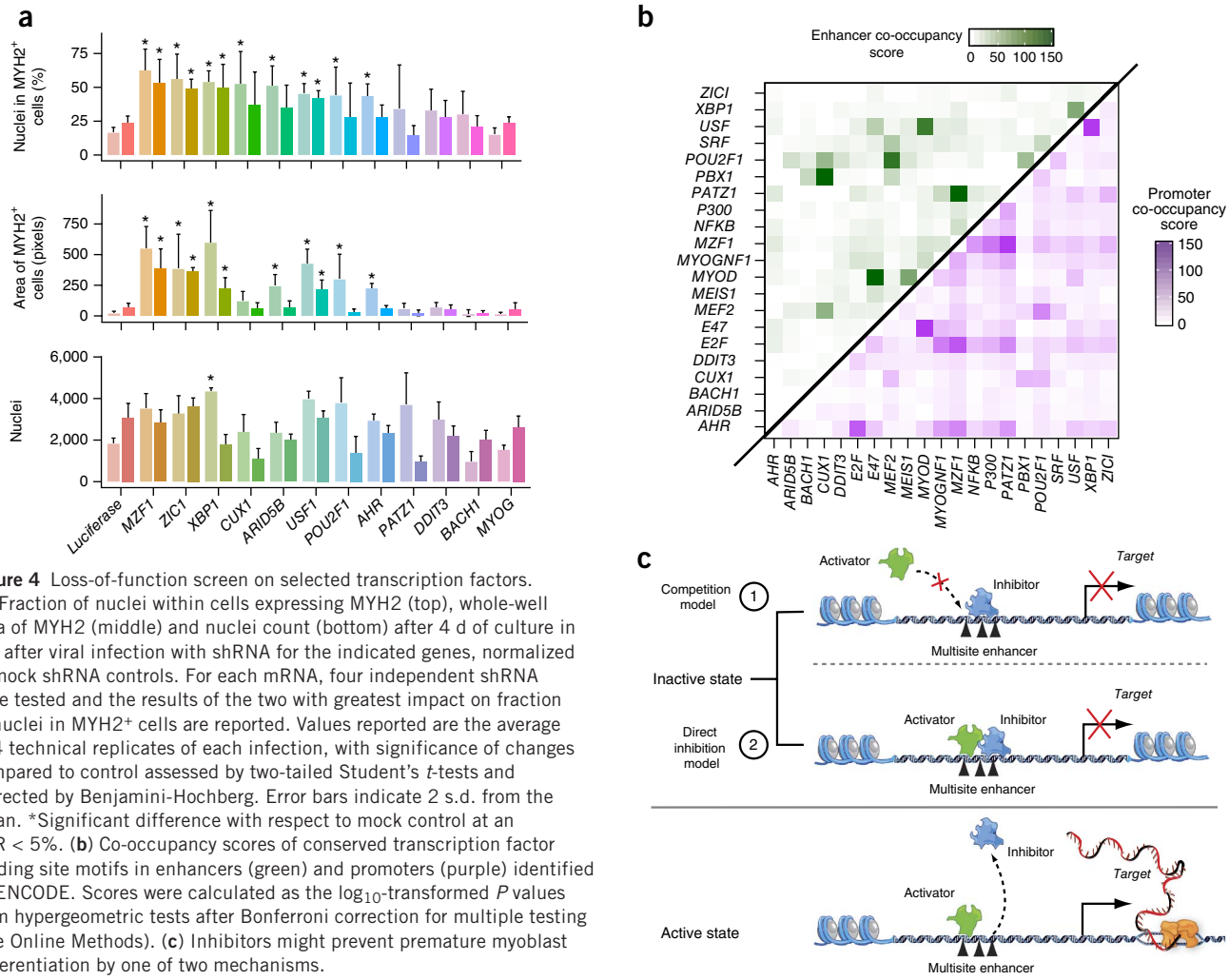

Figure 4 Loss-of-function screen on selected transcription factors. (**a**) Fraction of nuclei within cells expressing MYH2 (top), whole-well area of MYH2 (middle) and nuclei count (bottom) after 4 d of culture in DM after viral infection with shRNA for the indicated genes, normalized to mock shRNA controls. For each mRNA, four independent shRNA were tested and the results of the two with greatest impact on fraction of nuclei in MYH2$^+$ cells are reported. Values reported are the average of 4 technical replicates of each infection, with significance of changes compared to control assessed by two-tailed Student's *t*-tests and corrected by Benjamini-Hochberg. Error bars indicate 2 s.d. from the mean. *Significant difference with respect to mock control at an FDR < 5%. (**b**) Co-occupancy scores of conserved transcription factor binding site motifs in enhancers (green) and promoters (purple) identified by ENCODE. Scores were calculated as the $\log_{10}$-transformed *P* values from hypergeometric tests after Bonferroni correction for multiple testing (see Online Methods). (**c**) Inhibitors might prevent premature myoblast differentiation by one of two mechanisms.

screen for 11 candidates. Briefly, we infected proliferating myoblasts with lentiviruses carrying one of 44 distinct short hairpin RNAs targeting either one of these factors or a mock (non-targeting) control, followed by serum switch–induced differentiation for 5 d. We then measured the frequency and size of MYH2$^+$ cells by immunofluorescence and automated quantification. Cells infected with two or more independent shRNAs targeting *MZF1*, *ZIC1*, *XBP1* and *USF1* showed significantly altered (FDR < 5%) differentiation kinetics (**Fig. 4a**,**b** and **Supplementary Fig. 13**). *ZIC1*, *XBP1* and *USF1* showed significantly altered differentiation kinetics (**Fig. 4a**,**b** and **Supplementary Fig. 13**) when depleted with two or more independent hairpins (FDR < 5%).

Knockdown of *XBP1*, *USF1*, *ZIC1* and *MZF1* enhanced myotube formation, with larger myotubes containing a higher fraction of total nuclei than mock shRNA controls (**Fig. 4a**). Depletion of *CUX1*, *ARID5B*, *POU2F1* and *AHR* also increased differentiation efficiency, albeit less significantly. Whole-well counts of nuclei were similar between knockdowns and mock controls, indicating that enhanced differentiation was not simply a result of higher initial cell counts or increased proliferation (**Fig. 4a**). With the exception of *ZIC1*, forced overexpression did not substantially alter differentiation kinetics (data not shown).

Notably, several of these factors have binding motifs that are highly enriched in promoters and enhancers that also have motifs for known muscle regulators (**Fig. 4b**). For example, *USF1* motifs are enriched in enhancers that also have *MYOD* motifs. Together, these results confirm that the transcription factors identified as possible regulators in fact influence myoblast differentiation, and demonstrate the power of Monocle for identifying key differentiation genes.

This study demonstrates that Monocle can exploit the inherent temporal variability during differentiation to order individual cells according to progress without relying on known markers. This pseudotime ordering pinpoints key events in differentiation, such as the *ID1*/*MYOG* switch, that are masked both by conventional bulk cell expression profiling and by single-cell expression profiles ordered by time collected. The reordering resolves sequentially activated or repressed groups of genes that can be further scrutinized to reveal upstream regulators. The temporal resolution offered by hundreds of ordered cells might enable future efforts to computationally infer new gene-regulatory modules. For example, the enrichment of transiently upregulated genes for common miRNA target sites raises the question of whether those miRNAs are expressed later, curtailing what would have been higher levels of expression. Sequencing-based measurements of small RNAs and mRNAs from the same cell will provide answers to such systems-level questions. Moreover, single-cell analysis distinguishes cells of interest from contaminating cell types such as interstitial mesenchymal cells without experimental isolation that might disrupt cell-cell interactions important in the *in vivo* niche.

Because our approach requires no a priori knowledge of marker genes to reorder cells, it is suitable for discovering regulators and markers of poorly characterized temporal processes.

We identified eight previously unappreciated transcription factors that influence the course of myoblast differentiation, demonstrating the principle of pseudotemporal analysis and expanding the catalog of regulators in this well-studied system. Several of the eight factors reported here may normally repress differentiation by competing with promyogenic factors for these regulatory elements. Alternatively, these inhibitors may co-occupy regulatory elements with promyogenic factors, preventing transactivation of their targets (**Fig. 4c**). Previous studies in other contexts provide mechanistic data supporting both of these models. *USF1* inhibits *MyoD* autoactivation in *Xenopus* by competing with MyoD at the *MyoD* promoter through an alternative E-box[23]. Our results suggest that *USF1* may repress a broad array of targets via E-box competition. *CUX1* represses targets in several developmental contexts through binding site competition[24]. *XBP1* was recently reported to inhibit myoblast differentiation in mice[25], potentially through regulatory element competition. Further experiments in HSMM cells and myoblasts from other anatomic sites will be needed to confirm the mechanism of these factors.

While positive regulators of myogenesis have been well characterized, only a handful of inhibitors have been identified. The eight inhibitors reported here may shed light on how the balance of proliferation and differentiation is maintained during development and regeneration. Ordering the expression profiles of individual cells by biological progress is thus a useful tool for studying cell differentiation, and it could in principle be used to map regulatory networks that govern a much wider array of biological processes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. GEO: GSE52529. Monocle is available in **Supplementary Source Code** and at http://monocle-bio.sourceforge.net/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to S. Bordbar, C. Zhu, A. Wagers and the Broad RNAi platform for technical assistance and to M. Soumillon for discussions. C.T. was supported by a Damon Runyon Cancer Research Foundation Fellowship. D.C. was supported by a Human Frontier Science Program Fellowship. D.C. and T.S.M. were supported by the Harvard Stem Cell Institute. This work was supported by US National Institutes of Health grants 1DP2OD00667, P01GM099117 and P50HG006193-01. This work was also supported in part by the Single Cell Genomics initiative, a collaboration between the Broad Institute and Fluidigm Inc.

AUTHOR CONTRIBUTIONS

C.T. and D.C. conceived the strategy of ordering individual cells by developmental progress. C.T. designed and wrote Monocle and performed the computational analysis. D.C., C.T., J.G., P.P., S.L. and M.M. performed the experiments. D.C., C.T. and J.L.R. designed the study. C.T., D.C., J.G., N.J.L., K.J.L., T.S.M. and J.L.R. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Bendall, S.C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
2. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality inexpression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
3. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
4. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2**, 666–673 (2012).
5. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
6. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
7. Ramskold, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
8. Simpson, E.H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Series B Stat. Methodol.* **13**, 238–241 (1951).
9. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
10. Magwene, P.M., Lizardi, P. & Kim, J. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19**, 842–850 (2003).
11. Gupta, A. & Bar-Joseph, Z. Extracting dynamics from static cancer expression data. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **5**, 172–182 (2008).
12. Qiu, P., Gentles, A.J. & Plevritis, S.K. Discovering biological progression underlying microarray samples. *PLOS Comput. Biol.* **7**, e1001123 (2011).
13. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
14. Abmayr, S.M. & Pavlath, G.K. Myoblast fusion: lessons from flies and mice. *Development* **139**, 641–656 (2012).
15. Tapscott, S.J. The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. *Development* **132**, 2685–2695 (2005).
16. Tomczak, K.K. Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J.* doi:10.1096/fj.03-0568fje (2004).
17. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw.* 411–430 (2000).
18. Amir, E.-A.D. *et al.* visNe enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
19. Joe, A.W.B. *et al.* Muscle injury activates resident fibro/adipogenic progenitors that facilitate myogenesis. *Nat. Cell Biol.* **12**, 153–163 (2010).
20. Blais, A. An initial blueprint for myogenic differentiation. *Genes Dev.* **19**, 553–569 (2005).
21. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
22. Ernst, J. & Kellis, M. ChromHMM: automating chromatin state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
23. Lun, Y., Sawadogo, M. & Perry, M. Autoactivation of *Xenopus* MyoD transcription and its inhibition by USF. *Cell Growth Differ.* **8**, 275–282 (1997).
24. Sansregret, L. & Nepveu, A. The multiple roles of CUX1: insights from mouse models and cell-based assays. *Gene* **412**, 84–94 (2008).
25. Acosta-Alvear, D. *et al.* XBP1 controls diverse cell type- and condition-specific transcriptional regulatory networks. *Mol. Cell* **27**, 53–66 (2007).

ONLINE METHODS

The single cell ordering problem. As cells differentiate, they undergo a process of transcriptional reconfiguration, with some genes being silenced and others newly activated. While many studies have compared cells at different stages of differentiation, examining intermediate states has proven difficult for two reasons. First, it is often not clear from cellular morphology or established markers what intermediate states exist between, for example, a precursor cell type and its terminally differentiated progeny. Moreover, two cells might transit through a different sequence of intermediate stages and ultimately converge on the same end state. Second, even cells in a genetically and epigenetically clonal population might progress through differentiation at different rates *in vitro*, depending on positioning and physical contact with neighboring cells. Looking at average behavior in a group of cells is thus not necessarily faithful to the process through which an individual cell transits.

Here, we describe an unsupervised algorithm, Monocle, that computationally reconstructs the transcriptional transitions undergone by differentiating cells. It orders a mixed, unsynchronized population of cells according to progress through the learned process of differentiation. Because the population may actually differentiate into multiple separate lineages, Monocle allows the process to branch, and it can assign each cell to the correct sublineage. It subsequently identifies genes that distinguish different states and genes that are differentially regulated through time. Finally, it performs clustering on all genes to classify them according to kinetic trends. The algorithm is inspired by and extends one proposed by Magwene *et al.* to time-order microarray samples[10]. Monocle differs from previous work in three ways. First, single-cell RNA-Seq data differ from microarray measurements in many ways, and so Monocle must take special care to model them appropriately at several steps in the algorithm. Second, the earlier algorithm assumes that samples progress along a single trajectory through expression space. However, during cell differentiation, multiple lineages might arise from a single progenitor. Monocle can find these lineage branches and correctly place cells on them. Finally, Monocle also performs differential expression analysis and clustering on the ordered cells to help a user identify key events in the biological process of interest.

Consider the gene expression profile of each cell captured in the experiment as a vector in $\mathbb{R}^d$, where d is the number of genes detectably measured in the experiment. In the case of RNA-Seq, d might be equal to the number of genes in the organism's transcriptome. In a single-cell qPCR experiment, d might be 48 genes or fewer. The endpoints of differentiation will be separated in this space, and cells transiting between these endpoints might proceed along an arbitrarily complex (nonlinear) path (or paths). Typically, we do not know where the endpoints of differentiation reside in this space or what the path between them should look like. Moreover, the expression profile from each cell will contain some measurement error (for example, from sequencing-based sampling error), and mRNA levels will vary stochastically due to transcriptional noise, so cells will 'travel' through expression space in a noisy way.

We wish to (i) identify the (two or more) endpoints of the biological process, (ii) learn the shape of the path(s) between them, and (iii) accurately place cells along this path. The first two challenges together are referred to as the curve reconstruction problem. Formally, consider differentiation as a continuous, smooth function $\vec{f}(s) = [x_1(s), x_2(s), \ldots, x_d(s)]$. Each point in the image of the function is the state of the cell at s, which can be thought of as progress through the biological process, rather than time. That is, two cells might execute exactly the same sequence of changes as they differentiate but take different amounts of time to do so. A sample from $\vec{f}$ at progress s is a random vector $\vec{\rho}(s_0) = \vec{f}(s) + \vec{\delta}(s)$, where $\vec{\delta}(s)$ is a vector describing biological and technical noise. An experiment data set consists of a finite set of samples $V = \{\vec{\rho}(s_0), \vec{\rho}(s_1), \ldots, \vec{\rho}(s_n)\}$.

Estimating the geometry of $\vec{f}$ has been referred to as the curve reconstruction problem, and two classes of approaches have been proposed. The first uses polygonal reconstruction, which approximates the smooth curve with polygonal segments connecting the elements of V. The second, principal curve reconstruction, directly fits a smooth curve through V. While we consider the latter a promising avenue for ordering expression profiles captured by single-cell RNA-Seq, we focus on the former approach, as it extends very naturally to settings where more than two endpoints exist in the data.

A polygonal reconstruction of $\vec{f}$ (which is assumed to be smooth and twice-differentiable) from V is a graph that connects every pair of samples that are adjacent on $\vec{f}$ and no others. Adjacent points are those for which there is no point in V between them on the curve $\vec{f}$. In other words, there is no point of intermediate progress in V between adjacent points. If a sufficient number of points on $\vec{f}$ are sampled without error, an accurate polygonal reconstruction can be achieved by finding a traveling salesperson path (TSP) or a minimum spanning tree (MST) through V[26–28].

We now state the cell ordering problem formally. Suppose $\vec{f}(s)$ is a vector-valued function parameterized by progress s through a biological process and is sampled at a finite number of points with error. Let $V = \vec{\rho}(s_i)$ be the set of samples. Then:

Definition 1. A permutation π of the index set $\{1, 2, \ldots, n\}$ is an ordering of the points in V by progress if $\pi(i) \preccurlyeq \pi(j) \Rightarrow s_i \preccurlyeq s_j$ for all i,j in the index set.

The ordering problem is to find the progress-ordering permutation, π, given the data V (ref. 1).

An order of cells allows us to say that one cell precedes another as differentiation progresses, but it does not directly tell us how similar two cells are in terms of the process. Assuming differentiation follows a one-dimensional path embedded in a d-dimensional metric space, simply taking the distance in that space between two cells would be like saying that the distance between Hong Kong and New York is straight-line distance through the center of the Earth. Just as we must measure the distance New York to Hong Kong along the surface of the Earth (that is, the great circle, or geodesic distance), we must measure the distance between two cells along $\vec{f}$. This leads us to a definition for the distance between two cells in terms of differentiation:

Definition 2. Given a permutation π that orders cells by progress, specifying that $s_i \preccurlyeq s_j$, the pseudotime scale of the biological process is a function $\psi : V \to \mathbb{R}$. Denoting s_{i-1} as the cell that precedes s_i and s_0 as the element of V that is strictly preceded by no other, we construct ψ as

$$\psi_t(s_i) = \begin{cases} 0 & \text{if } i = 0 \\ \psi_t(s_{i-1}) + \| \vec{\rho}(s_i) - \vec{\rho}(s_{i-1}) \| & \text{if } i > 0 \end{cases}$$

That is, the ordering π induces a one-dimensional measure of progress in terms of transcriptional state during differentiation. This 'pseudotemporal' scale of differentiation is numerically arbitrary and will of course vary from experiment to experiment, but nevertheless is useful for downstream analysis.

Dimensionality reduction. Ordering cells by progress through a biological process can be formulated as the problem of finding a one-dimensional function $\vec{f}(s)$ embedded in $\mathbb{R}^d$. Although d might be the number of genes in the transcriptome, this need not be the case. It might make sense to ignore some subset of genes or even augment the space with some other quantitative dimensions based on per-cell measurements from another assay. Moreover, expression measurements for many genes will show large covariance with other genes in the experiment. Thus, it may make sense to reduce the dimensionality of the space before ordering the cells. In practice, dimensionality reduction not only can reduce the variability in pairwise cell-to-cell distance, it can also simplify interpreting cellular trajectories through expression space.

Monocle uses independent component analysis (ICA) to reduce the dimensionality of the expression data before ordering the cells. ICA was originally developed as a means of separating a set of mixed signals (for example, captured by a collection of microphones) into (statistically) independent sources. It has been widely adopted for image analysis and other types of signal processing tasks. The usefulness of ICA is often explained as a way of solving the 'cocktail party problem', where one has placed n microphones around the room to listen to several conversations among k cocktail partygoers. Each of the microphones will pick up a mixture of the conversations, and the signal recovered from each will vary depending on where in the room the microphone is relative to the partygoers. By separating the n mixed signals into k independent components, ICA aims to recover the k individual voices. If k is smaller than n, then ICA has also reduced the dimension of the original data.

Reducing dimensionality of single-cell expression data amounts to describing each cell in terms of abstract sources, which are hidden variables that describe a cell's state but which are reflected in observed gene expression values. More formally, the expression measurements of n genes in m cells can

be represented as an $n \times m$ matrix x. If those expression measurements are generated by a linear combination of k source signals, we can write

$$S = Ax$$

where S is a $k \times m$ matrix and A is an invertible weight matrix that transforms x into S.

In practice, a single-cell RNA-Seq experiment will detect expression for only a subset of genes in each cell; others will be detectably expressed but far too noisily or in too few cells to reliably use in downstream analysis. It is useful to preselect the genes that will be used for ordering the cells. One might select genes that are known markers of progress through differentiation, although this would undoubtedly introduce substantial bias into the analysis. A better approach might be to simply select all genes detectably expressed above a certain threshold in a certain fraction of cells. One might also only include genes that vary over a sufficiently large dynamic range, as including genes that do not appreciably change likely would just increase noise in the analysis.

We have found that selecting genes that are differentially expressed between groups of cells collected at different real times allows robust ordering (see "Differential expression analysis", below). However, in some experiments this might not be available (for example, because all the cells came from a single tissue or time point). In these cases, we recommend selecting genes on the basis of a minimum expression level and a minimum level of variance.

Ordering cells by progress. Monocle orders cells by progress through a biological process, resulting in an induced 'pseudotime' scale describing that process in transcriptional terms. The algorithm used here extends from one introduced by Magwene *et al.*[10], although there are some departures from their approach that cater to the nature of single-cell RNA-Seq data. We briefly outline the approach below.

Cells are described as points in $\mathbb{R}^d$, which might be a reduced expression space as described in the previous section. We first construct a weighted complete graph, where vertices represent cells and edges are weighted by the distance in $\mathbb{R}^d$ between the connected cells. Next, the algorithm finds the MST on the complete graph. If this graph has no branches, the algorithm returns the MST as the polygonal reconstruction of the biological process path $\vec{f}(s)$. Otherwise, the algorithm uses the MST to order the cells.

Branched trees could result from one of two phenomena. First, an MST on the cells with branches might result from a population of cells that is transiting along two or more independent biological (sub)processes. For example, undirected differentiation of human embryonic stem cells would produce cells of all three germ layers—ectoderm, mesoderm and ectoderm—as well as intermediate states along the way. We address this case in the following section. Alternatively, the branches might simply result from biological or technical noise, representing small deviations from a single biological process. In this case, we need to find an appropriate place for such cells in the pseudotemporal ordering.

Monocle deals with branches due to noise by constructing a PQ tree[29], which captures the set of paths through the cells that constitute good orderings. A PQ tree is a rooted, ordered tree in which ordered elements as represented as leaves and internal nodes encode legal orderings. Internal nodes are of one of two types: Q nodes, whose children are ordered (albeit possibly in reverse), and P nodes, whose children are not. Monocle follows the approach of Magwene *et al.*[10] to construct a PQ tree that compactly represents good orderings of the cells:

1. For a set of cells, calculate the MST.
2. Find the longest path through the MST, called its diameter path.
3. Create a PQ tree with a single empty Q node, Q_{Main}.
4. Vertices on the diameter path with degree greater than 2 are called indecisive. Find the indecisive backbone of the diameter path, the longest continuous subset of the vertices of the diameter path for which both endpoints are decisive.
5. Moving along the indecisive backbone, make each decisive vertex a child of Q_{Main}, so the decisive vertices are ordered by the Q node.
6. For each indecisive vertex, create a new P node, attach it to Q_{Main}, and make the children of the indecisive vertex children of the new P node.

For the indecisive nodes newly attached as children of P nodes, recursively apply the whole algorithm, creating new Q nodes for each child in the MST of each indecisive vertex, and so forth.

The PQ tree encodes a family of orderings, and specific orderings can be extracted quickly. Children of Q nodes will appear as a subsequence consistent with their order in their parent Q node. In particular, vertices along the first diameter path, which ideally hold most of the cells, will appear in that order. However, small branches off of this path, which appear in P nodes, might be permuted in ways that result in discontinuities in the expression space. Similarly, children of Q nodes might need to be emitted in reverse order to make smooth transitions in the final ordering of cells. Magwene *et al.*[10] do not address this situation, preferring to emit the PQ tree itself. Monocle always emits an ordering of cells, and thus it exhaustively searches orderings encoded by the PQ tree to find one that obeys its constraints and minimizes the total distance traveled by the resulting polygonal reconstruction in the embedding geometry $\mathbb{R}^d$, beginning at one end of the diameter path of the full MST. While this might result in superpolynomial running time and memory, in practice with real data, this procedure takes only a few seconds and small amount of RAM on a laptop because the number of cells in P nodes is typically very small relative to the cells in Q nodes.

Monocle thus emits an ordering of the cells that relies on the MST to 'sketch' the basic shape of the polygonal reconstruction and uses the PQ tree to handle small, noise-driven branches.

Per-cell expression profiles were calculated in this study using the Tuxedo suite of tools[30]. The reads for each cell were mapped with TopHat[31] 2.0.9 and Bowtie[32] 2.0.6 against build 19 of the human genome, downloaded through the UCSC genome browser. TopHat was provided with GENCODE[33] gene annotations (build version 17). Mapped reads were analyzed with Cuffdiff[34] 2.2 to generate per-cell expression profiles. Bulk RNA-Seq libraries were mapped using an identical workflow and analyzed with Cuffdiff 2.2 to generate differential gene expression calls. Downstream heatmaps and clustering were performed with the CummeRbund (http://compbio.mit.edu/cummeRbund/) R package.

The myoblast expression profiles were ordered using Monocle. Expression space was reduced to two dimensions using the fastICA[17] package. The initial space used as input before the ICA reduction was defined by a subset of genes, selected as follows. First, genes detectable in fewer than 50 cells at or above FPKM 1 were discarded. Next, the remaining genes were analyzed for differential expression using a Tobit-family[35] generalized linear model (GLM) through the VGAM package in R. A minimum FPKM value of 0.1 was used as the censoring threshold for the Tobit model. This analysis reported genes that were significantly differentially expressed between groups of cells harvested on different days. Only genes significant at an FDR < 0.01 (after Benjamini-Hochberg correction) were kept for ICA analysis. The input space for ICA was taken to be the standardized, log-transformed FPKM values of these genes. These selection criteria were picked to exclude genes that contribute only Gaussian noise to the input for ICA. That is, we aimed to provide ICA with genes that were (i) reliably and accurately measured by RNA-Seq and (ii) sufficiently dynamic over differentiation. We explored alternative means of selecting genes for ICA reduction, such as selecting on the basis of simple dynamic range or variance thresholds, and these methods produced qualitatively similar, but less accurate, orderings. Monocle orders cells on the basis of proximity in expression space. Intercell distances were calculated as the Euclidean distance in two-dimensional ICA space. These distances were used to construct the MST. The MST was then used to order the cells, allowing two distinct lineages to arise from the initial population of cells as described below.

Identifying branches in a biological process. During differentiation, progenitor cells make a series of decisions that restrict and specify which terminal cell type they will ultimately become. Many progenitors can generate multiple lineages, and in the extreme case, embryonic stem cells and induced pluripotent stem cells can produce any cell type found in the organism. In a single-cell RNA-Seq experiment, cells transiting through differentiation might be captured at these decision points. Monocle aims to reconstruct a branched process in the embedding geometry that recapitulates the subpaths cells take. This assumes that transitions through decision points are smooth

doi:10.1038/nbt.2859

and continuous in the embedding geometry, or rather, that transitions from progenitor to one of several committed cell types is marked by a smooth shift in transcriptional state, as opposed to an instantaneous shift to a distant part of the embedding geometry.

To handle this situation, we must refine our definitions of several key ideas. Differentiation must be expressed not as a single smooth, continuous vector-valued function, $\vec{f}(s)$, but rather as a piecewise smooth, continuous one. That is, the process begins with some initial path through the embedding geometry defined as before but reaches a certain point and splits into two or more segments. We write this function as $\vec{f}(b,s)$, where b selects a branch segment in the process and s is the progress along that branch. The initial segment is denoted b_0, and we denote the first cell on any branch b_i as having progress (b_i, s_0). A lineage is defined as an ordered set of branches b_i such that $b_i \prec b_j \; \forall i,j$ and, if b_i precedes and is adjacent to b_j, then the left open interval of $\vec{f}(b_j,s)$ begins at $\vec{f}(b_i,s_n)$, where s_n is the last state on the branch b_i.

An ordering of cells from a branched process can be represented as a tree rooted near $\vec{f}(b_0,s_0)$. In the unbranched case, an ordering means that $\pi(i) \preccurlyeq \pi(j) \Rightarrow s_i \preccurlyeq s_j$. For the branched case, we construct a tree on the index set, where if element i of the index set is the parent of j in the directed acyclic graph (DAG), then $\vec{f}(b_x,s_i) \preccurlyeq \vec{f}(b_y,s_j)$. If i has only a single child in the tree, then $b_x = b_y$. If, however, i has multiple children, then $b_x \prec b_y$, and the left open interval of $\vec{f}(b_y,s)$ begins at $\vec{f}(b_x,s_i)$.

Given the above definition for a branched ordering, the notion of pseudotime for a branched process can be readily updated as well:

$$\psi_t(b_x,s_i) = \begin{cases} 0 & \text{if } x = 0, i = 0 \\ \psi_t(\text{Parent}(b_x,s_i)) + \| \vec{\rho}((b_x,s_i)) - \vec{\rho}(\text{Parent}(b_x,s_i)) \| & \text{otherwise} \end{cases}$$

where Parent(b,s) denotes the parent of cell (b,s) in the ordering tree.

Each Q node is the parent of cells along an indecisive backbone that fixes their order in the embedding geometry. To build a branched ordering of the cells, each time Monocle adds a Q node to the tree, it also records the length of the corresponding indecisive backbone in the MST. When the PQ tree is complete, Monocle selects the k Q nodes with the longest indecisive backbones, where k is selected by the user and corresponds to the number of terminally differentiated cell types in the experiment. From this list, Monocle selects the Q node with the shortest backbone, prunes the corresponding subtree from the MST, and orders its cells using the above exhaustive procedure. Monocle then does the same for the remaining selected Q nodes.

The algorithm then reassembles the ordered subsets of the cells with a depth-first traversal of the PQ tree. The Q node with the longest backbone will always be Q_{Main}, the first one created. The second longest backbone will be found branching from the longest in the MST, and so on. The root cell (b_0,s_0) of the ordering tree is the first cell in the ordering of the longest backbone. The remaining cells under Q_{Main} are added in order, until the indecisive node that created the second longest backbone is reached, at which point the ordering tree branches, creating b_1 and b_2, with cells on the second longest backbone added as children in one branch, and the remaining cells on the longest backbone added to the other. This procedure is applied, creating branches in the ordering tree whenever one of the k longest backbone Q nodes is encountered, until all the cells have been added to the ordering tree.

Differential expression analysis. Monocle can identify genes and transcripts that are differentially expressed across distinct cell types or that change significantly as a function of pseudotime. Generalized additive models (GAMs) relate one or more predictor variables to a response variable as

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

where Y is a response variable, such as a particular gene's expression level, and the x_i's are predictor variables[36]. The function g is a link function, typically the identity or log function. The f_i's are nonparametric functions, such as cubic splines or some other smoothing function. Generalized additive models are similar to generalized linear models but allow testing of variables in response to a numerically estimated trend in the predictors, alleviating the burden of specifying their distribution. While this necessitates some approximations in downstream testing, it has proven to be highly effective in many settings, particularly when one wishes to model the response variable as a function of both categorical and continuous predictors.

Monocle models each gene's expression level across cells using a Tobit model[35]. That is, each gene's observable (log-transformed) expression level Y depends on a latent variable Y^*:

$$Y = \begin{cases} Y^* & \text{if } Y^* > \lambda \\ \lambda & \text{if } Y^* \leq \lambda \end{cases}$$

where λ is a detection threshold. The latent variable Y^* depends on the variables x_i, which might express the day on which each cell was collected, Monocle's pseudotime value for each cell, and so forth. The parameter λ is a user-specified value (FPKM = 0.1 by default).

Monocle's generalized additive model is thus

$$E(Y) = s(\psi_t(b_x,s_i)) + \varepsilon$$

where $\psi_t(b_x,s_i)$ denotes the assigned pseudotime of a cell and s is a cubic smoothing function with (by default) three effective degrees of freedom. The error term ε is normally distributed with a mean of zero. Testing for differential expression is performed with an approximate χ^2 likelihood ratio test. The GAM and associated testing functions are provided through the VGAM package[37].

Monocle also supports testing for differential expression between groups of cells collected on different days or otherwise categorically labeled in the experiment. In these tests, the GAM simply uses the categorical labels as predictor variables, with no smoothing.

In this study, pseudotemporally ordered myoblasts were analyzed for dynamically regulated genes using a GAM that described log-transformed FPKM values as dependent variables from the Tobit family, which varied as a smooth function of pseudotime. Smoothing was performed with a cubic spline with three effective degrees of freedom. A randomly selected set of 15 genes was manually assessed for goodness of fits using standard criteria (for example, Q-Q plots) and confirmed to be well-fit by both the pseudotemporal and day-collected Tobit models. Significance of pseudotime dependency was performed with an approximate likelihood ratio test (via the VGAM lrtest() function) against the reduced model of no pseudotime dependence. In all tests, genes with an FDR < 0.05 after Benjamini-Hochberg correction were considered pseudotemporally regulated.

Bulk RNA-Seq libraries were analyzed with Cuffdiff 2 to call differentially expressed genes. Bulk RNA-Seq libraries met the widely used assumption of increasing fragment count overdispersion as a function of increasing expression. Cuffdiff 2, used to assess changes in bulk RNA-Seq libraries, explicitly models overdispersion and includes it in statistical testing.

Clustering genes by pseudotemporal expression pattern. Once Monocle has fit a GAM for each gene, these models can be used to predict smooth response curves as a function of pseudotime. Standardizing these curves allows for efficient K-medioid clustering of all genes in a data set across pseudotime. Pairwise distances between genes x and y are calculated as

$$d(x,y) = 1 - \frac{\rho_{x,y}}{2}$$

where $\rho_{x,y}$ indicates the Pearson correlation of their response curves. Clusters correspond to genes that follow the same relative kinetic trends. Clustering based on the GAM response curves, rather than the raw data, produces, in practice, more coherent clusters with a lower root mean squared error with respect to the medioids and sharper kinetic trends, allowing analysis of a more diverse set of patterns.

In this study, clustering analysis was carried out on all detectably expressed genes, regardless of significance of pseudotemporal regulation. K-medioid clustering was performed on the predicted response of genes pseudotime GAM after log-transformation and standardization. Clustering was performed using the PAM package in R. Six clusters were generated, as this was the largest K that produced qualitatively distinct clusters without redundancy.

Primary human myoblast culture and treatment. Human skeletal muscle myoblasts (HSMM) derived from quadriceps biopsy (Lonza, catalog

#CC-2580, lot #257130: healthy, age 17, female, of European ancestry, body mass index 19) were expanded in 10% FBS SkGM-2 (Lonza) and differentiated for the indicated time points upon switch to MEM-alpha supplemented with 2% horse serum (Lifetech). All procedures were performed using HSMM within five passages from explant. Cells were verified to be mycoplasma negative before data collection (Promocell, catalog #PK-CA91-1024).

For HSMM RNAi experiments, individual shRNA hairpins (pLKO.1 backbone) for the transcription factor under study were obtained as lentiviral particles in supernatant from The RNAi Consortium (Broad Institute RNAi platform). Lentiviral particle production and titering were performed as previously described[38]. HSMM were infected in growth medium, expanded and collected after puromycin selection for knockdown validation, for the indicated time points for gene expression and immunofluorescence analyses. Sequences for shRNAs are listed in **Supplementary Table 1**.

Immunofluorescence analyses were performed on PFA-fixed plates, with antibodies raised against muscle MYH2 (clone MF-20, E-bioscience), MEF2C (Abcam cat. ab79436) and phospho-histone H3 (Millipore cat. 06570) according to Cacchiarelli *et al.*[39]. Whole well imaging was performed together with Hoechst staining using a Celigo S cytometer (Brooks Automation). The resulting images were analyzed with ImageJ to obtain information about total number of cells or nuclei, fractions of nuclei in MYH2- or CD13-positive cells, and total MYH2-positive cell area. All immunofluorescence experiments were performed in technical quadruplicate. Significance of changes in response to knockdown of target genes with respect to luciferase control was assessed with a pooled-variance *t*-test, with similar variance between control and target infections.

Cell capture and mRNA sequencing. HSMM in either growth or differentiation medium were dissociated, washed and resuspended in GM or DM containing 0.1 U/µl of RNaseOUT (Lifetech). Each time point was collected in three independent biological replicates for regular mRNA sequencing, while for single-cell mRNA sequencing the independent replicates were pooled in equal amounts.

For bulk RNA-Seq libraries, total RNA was extracted with Trizol and mRNA libraries were produced starting from 100 ng total RNA using the TruSeq mRNA-Seq library kit (Illumina) according to manufacturer's instructions. Briefly, mRNA library construction consisted in a first step of RNA poly(A) selection, followed by salt-mediated fragmentation. The fragmented RNA was then converted to a double-stranded cDNA by retrotranscription and second DNA strand synthesis. End-repair and 3′-adenylation were then performed to produce cDNA termini compatible for ligation of 5′ and 3′ adapters. Finally, 11–14 PCR cycles were performed to amplify the obtained libraries. Stringent DNA purification with Agencourt AMPure XP magnetic beads (Beckman Coulter) was performed after each step in the protocol, in particular twice after adaptor ligation and twice after PCR reaction, to minimize contamination by adapters in the library. Library QC and quantification were performed using Bioanalyzer DNA High Sensitivity (Agilent) and qubit High Sensitivity (Lifetech) assays, respectively. Sample size for conventional, bulk RNA-Seq libraries was fixed at 3 biological replicates, in accord with previous reports[34] that this design is sufficient to capture the vast majority of differentially expressed genes for *in vitro* differentiation experiments involving homogeneous starting cell populations.

For single-cell mRNA sequencing, dissociated cells were captured and processed with the C_1 Single-Cell Auto Prep System (Fluidigm) following manufacturer's protocol 100-5950. Starting with a suspension of cells at a concentration of approximately 250 cells/µl, up to 96 single cells are captured in each C_1 microfluidic device. In this study, we used one C_1 capture chip at 0, 24, 48 and 72 h after switching to differentiation medium, for a total of four independent captures. After imaging with a microscope to identify which sites have captured a single cell, processing of the cells occurs within the C_1 instrument to perform the steps of cell lysis, cDNA synthesis with reverse transcriptase, and PCR amplification of each cDNA library. The cDNA synthesis and PCR use reagents from the SMARTer Ultra Low RNA Kit for Illumina Sequencing (Clontech 634936). The SMARTer chemistry uses a strand-switching mechanism so that both the first and second strands of cDNA are synthesized in a single reaction. After harvest from the C_1 microfluidic device, each cDNA library is subjected to tagmentation (simultaneous fragmentation and tagging with sequencing adapters) using the Nextera XT DNA Sample Preparation Kit (Illumina FC-131-1096) as described in Fluidigm protocol 100-5950. PCR amplification of the tagmented cDNA uses Index Primers from the Nextera XT DNA Sample Preparation Index Kit (Illumina FC-131-1002). After PCR, the cDNA libraries from individual cells are pooled and purified using AMPure XP beads (Agencourt Bioscience Corp A63880) as described in Fluidigm protocol 100-5950.

All libraries (bulk and single-cell) were sequenced using 100-bp paired end sequencing on a HiSeq 2500 (Illumina), generating 10–20 million reads for each TruSeq library and 4 million reads for each C1 single cell library. Libraries that contained fewer than 1 million reads or for which less than 80% of fragments mapped to nonmitochondrial protein coding genes were excluded.

Benchmarking robustness of Monocle. Cross-validation of Monocle's ordering and differential analysis routines was performed as follows. First, subsets of the cells of varying size were selected at random, to profile performance for an experiment with 15% of the cells, 25%, 50% and so on. Each subset was independently ordered using the same procedure as described above. That is, genes were assessed for differential expression between cells in the subset collected in GM (day 0) and DM (days 1, 2 and 3). Genes significant at FDR < 1% were used to order the subset. The pseudotime values for these cells ordered by themselves were compared to the values when they were ordered as part of the full data set by Pearson correlation. Similarly, the overall concordance of pseudotime ordered expression values for the panel of markers genes from **Supplementary Figure 6** was calculated as the Pearson correlation of each gene's expression values under the two orderings. The correlation values for each gene were then averaged to calculate an overall concordance score for the subset. Finally, the genes were analyzed for statistically significant pseudotime-dependent changes in expression using the methods described above. Genes marked as significantly dynamically regulated when considering all the cells (as shown in **Supplementary Fig. 6**) were taken to be the true positives. The true positive, false positive, true negative and false negative calls for the subset were calculated and used to compute precision and recall values for each subset.

qPCR expression analysis. Total RNA from the HSMM differentiation time course and shRNA treatments was extracted with RNeasy (Qiagen). qPCR analyses to assess knockdown efficiency of shRNA treatments were performed by retrotranscribing 50 ng of total RNA with Superscript VILO (Lifetech) followed by SYBR Green amplification of 1–2 ng of the resulting cDNA (Roche). qPCR expression values were analyzed with PCR Miner[40]. shRNA knockdowns of target genes were compared with nontargeting (luciferase) control to calculate knockdown efficiency. Primers for qPCR assays are listed in **Supplementary Table 2**.

Regulatory sequence analysis. Regulatory regions of the genome were defined as ENCODE DNase I–hypersensitive 'hotspots'[21], downloaded through the ENCODE data portal of the UCSC genome browser. The "HSMM" and "HSMMtube" DNase I HS tracks were merged for further analysis using bedtools[41]. These regions were assigned a probable function by overlapping them with predicted regulatory roles produced by ChromHMM[22] for HSMM cells, which integrates ChIP-Seq histone modification data. Thus each hypersensitive site was classified as a promoter element, an active enhancer, an insulator and so on.

Competitive gene set tests. A competitive gene set test assesses the hypothesis that a given group of genes is ranked more highly (or less highly) than expected by chance in a larger list. All competitive gene set tests were performed with the geneSetTest() function of the limma package in R[42]. This function takes a ranked list of all genes along with a specified subset of genes and performs a Wilcoxon rank-sum test on the ranks of the subset. All *P* values are corrected for multiple testing according to the number of gene sets tested.

Transcription factor binding analysis. HSMM regulatory sequences were mined for transcription factor binding site enrichment by overlapping them with the "conserved transcription factor binding site" track available through UCSC. The regulatory elements and corresponding binding sites were then

doi:10.1038/nbt.2859

associated with their nearest gene using the closestBed utility of the bedtools package to create a group of genes potentially regulated by each transcription factor. These gene sets were then used in competitive gene set tests as described above to identify transcription factors whose potential targets are, for example, more highly enriched in a given pseudotemporal cluster than expected by chance under the null hypothesis.

Transcription factor co-occupancy scores were derived by counting the number of regulatory regions (for example, enhancers active in HSMMs) in which both factors have binding sites. These co-occupancy counts were then assessed for statistical significance by hypergeometric tests. The *P* values for these tests were corrected for multiple testing according to the number of pairs of transcription factors assessed, log-transformed and reported as the co-occupancy score. Multiple testing correction by Bonferroni was used to control for positive correlation between co-occupancy scores of two pairs of factors where one factor is the same.

26. De Figueiredo, L.H. & de Miranda Gomes, J. Computational morphology of curves. *Vis. Comput.* **11**, 105–112 (1994).
27. Giesen, J. Curve reconstruction in arbitrary dimension and the traveling salesman problem. *Proc. 8th Discrete Geometry and Computational Imagery Conference (DGCI)* 164–176 (1999).
28. Giesen, J. Curve reconstruction, the traveling salesman problem, and Menger's theorem on length. *Discrete Comput. Geom.* **24**, 577–603 (2000).
29. Booth, K.S. & Lueker, G.S. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. Syst. Sci.* **13**, 335–379 (1976).
30. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
31. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and genefusions. *Genome Biol.* **14**, R36 (2013).
32. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
33. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), S4.1–S4.9 (2006).
34. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
35. Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36 (1958).
36. Hastie, T. and Tibshirani, R. Generalized additive models, *Stat. Sci.* **1**, 297–318 (1986).
37. Yee, T.W. & Wild, C.J. Generalized Additive Models. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 481–493 (1996).
38. Moffat, J. *et al.* A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283–1298 (2006).
39. Cacchiarelli, D. *et al.* miR-31 modulates dystrophin expression: new implications for Duchenne muscular dystrophy therapy. *EMBO Rep.* **12**, 136–141 (2011).
40. Zhao, S. & Fernald, R.D. Comprehensive algorithm for quantitative real-time polymerase chain reaction. *J. Comput. Biol.* **12**, 1047–1064 (2005).
41. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Wu, D. & Smyth, G.K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).