Direct lineage conversion offers a new strategy for tissue regeneration and disease modelling. Despite recent success in directly reprogramming fibroblasts into various cell types, the precise changes that occur as fibroblasts progressively convert to the target cell fates remain unclear. The inherent heterogeneity and asynchronous nature of the reprogramming process renders it difficult to study this process using bulk genomic techniques. Here we used single-cell RNA sequencing to overcome this limitation and analysed global transcriptome changes at early stages during the reprogramming of mouse fibroblasts into induced cardiomyocytes (iCMs)1–4. Using unsupervised dimensionality reduction and clustering algorithms, we identified molecularly distinct subpopulations of cells during reprogramming. We also constructed routes of iCM formation, and delineated the relationship between cell proliferation and iCM induction. Further analysis of global gene expression changes during reprogramming revealed unexpected downregulation of factors involved in mRNA processing and splicing. Detailed functional analysis of the top candidate splicing factor, Ptbp1, revealed that it is a critical barrier for the acquisition of cardiomyocyte-specific splicing patterns in fibroblasts. Concomitantly, Ptbp1 depletion promoted cardiac transcriptome acquisition and increased iCM reprogramming efficiency. Additional quantitative analysis of our dataset revealed a strong correlation between the expression of each reprogramming factor and the progression of individual cells through the reprogramming process, and led to the discovery of new surface markers for the enrichment of iCMs. In summary, our single-cell transcriptomics approaches enabled us to reconstruct the reprogramming trajectory and to uncover intermediate cell populations, gene pathways and regulators involved in iCM induction.

Direct cardiac reprogramming that converts scar-forming fibroblasts into iCMs shows promise as an approach to replenish lost cardiomyocytes in diseased hearts1–4. Considerable efforts have been made to improve the efficiency and unravel the underlying mechanism5–15. However, it still remains unknown how the conversion of fibroblast into cardiomyocyte is achieved without following conventional cardiomyocyte specification and differentiation process. This is partly because the starting fibroblasts exhibit molecular heterogeneity that is mostly uncharacterized, and the reprogramming population contains fully, partially and unconverted cells. Traditional population-based genome-wide approaches are incapable of resolving this unsynchronized cell fate-switching process. Therefore, we leveraged the power of single-cell transcriptomics to better investigate the reprogramming of iCMs that is mediated by Mef2c, Gata4 and Tbx5.

Previous studies have indicated that a snapshot of an unsynchronized biological process can capture cells at different stages of the process16. Because the emergence of iCMs occurs as early as day 3 (refs 1, 11–15), we reasoned that day-3 reprogramming fibroblasts contain a wide spectrum of cells that are transitioning from fibroblast to iCM. We therefore performed single-cell RNA sequencing (RNA-seq) on day-3 cardiac fibroblasts infected with separate Mef2c, Gata4 and Tbx5 viral constructs (hereafter M + G + T) from seven independent experiments (for experimental design, see Extended Data Fig. 1), followed by a series of quality control steps (Extended Data Fig. 1, Methods and Supplementary Tables 1, 2). Extensive data normalization was performed to correct for technical variations and batch effects (Extended Data Figs 1, 2 and Methods). After comparing the entire set of single-cell RNA-seq data to bulk RNA-seq data of endogenous cardiac fibroblasts and cardiomyocytes that were obtained from parallel experiments, we detected a group of resident or circulating immune or immune-like cells (Extended Data Fig. 3) that were not included in the subsequent analyses.

Unsupervised hierarchical clustering and principal component analysis (PCA) on the remaining 454 non-immune cells revealed three gene clusters that account for most of the variability in the data: cardiomyocyte-, fibroblast- and cell-cycle-related genes (Fig. 1a, b and Extended Data Fig. 4a–c). On the basis of the expression of cell-cycle-related genes, the cells were grouped into cell-cycle-active (CCA) and cell-cycle-inactive (CCI) populations (Fig. 1a); this was confirmed by the molecular signature of the cells in their proliferation states (Extended Data Fig. 4d–g, proliferating or non-proliferating). Within CCA and CCI populations, hierarchical clustering further identified four subpopulations based on differential expression of fibroblast versus cardiomyocyte genes: fibroblasts, intermediate fibroblasts, pre-iCMs and iCMs (Fig. 1a). When plotted using PCA or t-distributed stochastic neighbour embedding analysis, a stepwise transcriptome shift from fibroblast to intermediate fibroblast to pre-iCM to iCM was evident (Fig. 1c and Extended Data Fig. 4h, i). We also analysed the reprogramming process as a continuous transition using SLICER (selective locally linear inference of cellular expression relationships)17, an algorithm for inferring nonlinear cellular trajectories (Fig. 1d, e). The trajectory built by SLICER suggested that fibroblasts, intermediate fibroblasts, pre-iCMs and iCMs form a continuum on the bottom CCI path, representing an iCM-reprogramming route. We further calculated the pseudotime for each cell on the trajectory by defining a starting fibroblast and measuring the distance of each single cell to the starting cell along the reprogramming route (Fig. 1e). We then examined the distribution of cells along the pseudotime line by plotting the ‘free energy’ (max(density)−density) of the trajectory and discovered a peak (lowest density) in pre-iCM state (Fig. 1f). These data suggest that the pre-iCM stage is an unstable cell state seeking to settle into a more stable state, such as the iCM state, consistent with the PCA and hierarchical clustering analysis showing that pre-iCMs express both cardiomyocyte and fibroblast markers as an intermediate cell type and our other experimental evidence (Fig. 1a–c and Extended Data Fig. 4j–o). To experimentally test the iCM route, we performed
designed four sets of experiments to address the relationship between iCMs and pre-iCMs were predominantly CCI (Fig. 1i). We therefore

Figure 1 | Single-cell RNA-seq reconstructs iCM reprogramming and identifies intermediate cell populations. a, Hierarchical clustering results of 454 single cardiac fibroblasts that were infected with M + G + T or that were mock- or DsRed-infected for 3 days with representative gene ontology terms of the three identified gene clusters underneath. ECM, extracellular matrix; Fib, fibroblast; H, high; iFib, intermediate fibroblast; L, low; M, medium; Pos. reg. of SMC prolif., positive regulation of smooth muscle cell proliferation. For P values, see Extended Data Fig. 4. b, c, PCA showing representative genes (b) or cell groups (c). PC, principal component. In b, cell cycle genes are shown in orange, cardiomyocyte markers in red and fibroblast markers in blue. d, e, Three-dimensional trajectory constructed by SLICER showing hierarchical clustering/PCA cell groups (d) or pseudotime (e). LLE, local linear embedding; NP, non-proliferating; Pro, proliferating. f, Free energy of the reprogramming process. g, h, Microarray of MGT- or LacZ-transduced cardiac fibroblasts from day 0 to 14 plotted as a PCA plot (g) or heat map (h) showing the mean expression of representative genes from a, b, CM, cardiomyocyte population-based gene expression profiling at reprogramming days 0, 3, 5, 7, 10 and 14 (Fig. 1g, h and Extended Data Fig. 4–v). PCA generated a pattern showing an oriented path during reprogramming (Fig. 1g and Extended Data Fig. 4p–v). Expression of the three main gene clusters selected from single-cell data showed consistent changes in population data (Fig. 1h and Extended Data Fig. 4t–v), supporting the SLICER trajectory.

By analysing CCA and CCI populations, we found that even though proliferative iCMs (CCA iCMS) were observed (Fig. 1a–d), iCMS and pre-iCMS were predominantly CCI (Fig. 1i). We therefore designed four sets of experiments to address the relationship between cell proliferation and iCM reprogramming by: (1) manipulating the expression of cell-cycle-related genes in fibroblasts that were lentivirally transduced with M + G + T (Extended Data Fig. 5a–p) or infected with a single doxycycline-inducible MGT construct (Extended Data Fig. 5q–s); (2) synchronizing the cell cycle of starting cardiac fibroblasts (Figs 1j–l); (3) transiently overexpressing large T antigen to accelerate cardiac fibroblast proliferation (Extended Data Fig. 5t–z); (4) establishing an immortalized cardiac fibroblast (CF) line CF-T (see Methods) before the initiation of iCM reprogramming (Fig. 1m–p). All four sets of experiments yielded consistent results showing that decreased proliferation or cell-cycle synchronization enhanced iCM reprogramming, whereas increased proliferation suppressed iCM generation.

We next examined the cellular composition of our isolated starting cardiac fibroblasts (see Methods) and identified five subpopulations (Fig. 2a, b, Extended Data Fig. 6a–i and Supplementary Discussion 1). To delineate how these subpopulations were reprogrammed, we applied hierarchical clustering calculated from the starting cardiac fibroblasts to those that had been transduced with M + G + T and determined the correlation of the expression of non-cardiomyocyte lineage markers to the status of reprogramming (Fig. 2c). Expression of both endothelial and epicardial genes was significantly decreased in all cells that were transduced with M + G + T, irrespective of the reprogramming status. However, fibroblast and myofibroblast and/or smooth muscle genes were suppressed in iCMs, but not in intermediate fibroblasts or pre-iCMs (Fig. 2c and Extended Data Fig. 6j, k); this finding was supported by experimental data tracking protein expression of representative markers along the reprogramming trajectory (Fig. 2d–f and Extended Data Fig. 6l, m). Therefore, we conclude that endothelial and epicardial genes can be readily suppressed, whereas fibroblast and myofibroblast and/or smooth muscle genes were gradually suppressed along the course of reprogramming. This differential suppression is consistent with the difference in the layer of origin among different cardiac cell lineages during development and suggests that recent (epigenetic) memories might be easier to be erased than ones that have been gained earlier in development. The progressive suppression
of fibroblast markers also indicates that there is a difference between iCM and induced pluripotent stem cell (iPS cell) reprogramming, because early downregulation of fibroblast markers, such as Thy1, is one of the hallmarks and prerequisites for iPS cell reprogramming to proceed18.

To understand the molecular cascades that underlie iCM induction, we performed nonparametric regression and K-medoid clustering (see Methods), and identified three major clusters of genes that are significantly correlated to and show similar trends during reprogramming (Extended Data Fig. 7a–d). Further analysis identified six smaller gene clusters with narrower variation across the trend and gene ontology analyses were performed for each cluster (Fig. 3a–g, Supplementary Table 3 and Supplementary Discussion 2). The largest cluster (cluster 1) that shows a trend of immediate and continuous downregulation of gene expression is enriched in gene ontology terms related to protein translation/biosynthesis, modification and transportation (Fig. 3b). Such changes are probably to balance for increased energy requirements during the cell-fate switch and/or to transit from a protein production and ‘secretion factory’ (a fibroblast) to an energy-consuming ‘power station’ (a cardiomyocyte). The downregulated genes in cluster 2 are enriched in gene ontology terms that suggest a late suppression of fibroblast genes and growth factors, whereas the upregulated genes in clusters 4 and 5 are enriched in gene ontology terms that indicate engagement in a metabolic shift and structural changes towards a cardiomyocyte fate (Fig. 3e, f).

Unexpectedly, we found that cluster 1 is also enriched in the gene ontology terms ‘mRNA splicing’, ‘mRNA processing’ and ‘RNA recognition motif’. This finding prompted us to interrogate the role of splicing factor(s) in iCM induction. We therefore used an inducible iCM cell line derived from mouse embryonic fibroblasts (iCMFs)19 to screen a short hairpin RNA (shRNA) library that targeted 26 splicing factors representing the most common splicing factor families20 and identified Ptbp1 as the top candidate that also showed differential expression in cardiac fibroblasts versus cardiomyocytes (Extended Data Fig. 7e–h). Notably, knockdown of Ptbp1 in various primary fibroblasts consistently resulted in a significant increase in reprogramming efficiency (Fig. 3h, i and Extended Data Fig. 8a–p), demonstrating that Ptbp1 is a general barrier to iCM induction. However, overexpression of Ptbp1 has minimal effects (Extended Data Fig. 8q–u). To understand how Ptbp1 silencing led to improved iCM reprogramming, we performed high-depth RNA-seq to analyse alternative splicing events of day-3 reprogramming cells with or without Ptbp1 expression. A total of 1,494 alternative splicing events were detected upon Ptbp1 knockdown, 97% of which were not induced by MGT alone (Extended Data Fig. 9a and Supplementary Tables 4, 5). Notably, calculation of the difference in the percentage of spliced-in (ΔPSI) suggested that alternative splicing events between reprogramming versus control fibroblasts and endogenous cardiomyocytes versus cardiac fibroblasts were in an opposite direction (negative association, ΔPSI = 0.008). Knockdown of Ptbp1 in reprogramming fibroblasts, however, induced a strong positive association (ΔPSI = 2.2 × 10⁻¹⁸), suggesting that Ptbp1 silencing together with MGT, but not MGT alone, shifted the splicing pattern from cardiac fibroblast towards cardiomyocyte (Fig. 3j, k). Furthermore, a higher percentage of exon-skipping events (63%) of the five known alternative splicing types was observed in MGT-infected cells upon Ptbp1 silencing (Fig. 3l). Motif analysis using the RNA map analysis and plotting server (rMAPS)21 showed that a CT-rich Ptbp1-binding motif was significantly enriched in exon-skipping exons compared to background exons (Fig. 3m). Notably, in exons that were included more often upon Ptbp1 knockdown, the motif was strongly enriched within 100 bp of the upstream intron (P < 1 × 10⁻⁷), whereas, in exons that were skipped more often upon Ptbp1 knockdown, the motif was less strongly enriched, but showed a broad peak at 50–200 bp in the downstream intron (P < 0.05). These data are consistent with the higher percentage of inclusion (69%) than skipping (31%) among exon-skipping events that were observed in Ptbp1 knockdown samples (Extended Data Fig. 9b), suggesting that Ptbp1 is a repressor of exon inclusion when bound to an upstream intron, and probably is a weaker repressor of exon skipping when bound to a downstream intron. Next we assessed the gene ontology terms of genes that were alternatively spliced upon Ptbp1 silencing (Fig. 3n, o and Extended Data Fig. 9c–i). In addition to altering the splicing patterns of genes related to cardiomyocyte...
Figure 3 | Ptbp1 as a barrier to iCM splicing repatterning. a–g. Six gene clusters were identified during reprogramming (a) with gene ontology analysis (b–g, false discovery rate (FDR) <0.05). The number of genes is shown in parentheses. h, i. Representative 20× immunocytochemistry images of cTNT and α-MHC–GFP (h) and quantification (i) of MGT-infected cardiac fibroblasts treated with shRNA against Ptbp1 (shPtbp1) or shRNA non-targeting control (shNT). m = 20 images. Scale bar, 200 μm. Data are mean ± s.e.m., two-sided Student’s t-test; ***P < 0.001. j–q. Splicing analyses of day-3 MGT-infected cardiac fibroblasts treated with shPtbp1 or shNT. j, k, Correlation between ΔPSI of cardiomycocytes versus cardiac fibroblasts, and ΔPSI of MGT versus LacZ (j) or MGT and shPtbp1 versus MGT and shNT (k). The trend line generated by linear regression and P values from a one-sided binominal test are shown. l, Number of detected alternative splicing events among the five alternative splicing types. AS, alternative splice; AS3; A5SS, alternative lineage and function (Fig. 3n). Ptbp1 silencing resulted in changes in the splicing pattern of 21 other splicing factors, suggesting that Ptbp1 knockdown might trigger a second wave of splicing changes by regulating the switching of the isoform of other splicing factors. Furthermore, we explored the potential downstream effects of Ptbp1-mediated re-patterning of splicing events (Supplementary Table 6 and Supplementary Discussion 3). DESeq2 (ref. 22) analyses of differentially expressed genes revealed that Ptbp1 knockdown enhanced the MGT-induced cardiac fibroblast to cardiomyocyte transcriptome shift by augmenting MGT-mediated changes (Fig. 3p and Extended Data Fig. 9i–n) and altering the expression of an additional set of cardiac and fibroblast lineage genes (Fig. 3q and Extended Data Fig. 9o).

To determine whether cardiac reprogramming is a rare and random event or a Me2c-, Gata4- and/or Tbx5-determined process, we plotted the expression of Me2c, Gata4, Tbx5 and M + G + T in each cell against the reprogramming pseudotime of that cell calculated by SLICER (Fig. 4a and Extended Data Fig. 9p). We found that the expression levels of Me2c, Gata4, Tbx5 and M + G + T are highly correlated with the reprogramming progress, despite the fact that their expression was not used in the generation of the trajectory. We also determined the mean expression levels of Me2c, Gata4 and Tbx5 and the mean ratio of expression (Me2c/Gata4, Me2c/Tbx5 and Gata4/Tbx5) in the fibroblast, intermediate fibroblast, pre-iCM and iCM populations along the reprogramming trajectory (Extended Data Fig. 9q–s). Consistent with our previous studies8,14,22, we observed higher levels of Me2c than Gata4 and Tbx5 in iCMs, further underscoring the importance of high Me2c expression in iCM induction.

To unravel the gene networks regulated by reprogramming factors, we navigated the relationship between the expression of a reprogramming factor and its downstream targets in each single cell. Using Tbx5 as an example, we calculated the Spearman correlation between Tbx5 expression and the expression of its downstream targets22,23 within each reprogramming cell (Fig. 4b, left). We then generated a correlation matrix for selected Tbx5 targets to determine their co-expression patterns (Fig. 4b, right). The correlation patterns suggest that Tbx5 acts by promoting cardiac function-related genes and by suppressing protein biosynthesis and non-cardiomyocyte lineages (Fig. 4b and Extended Data Fig. 9t, u).

Finally, we aimed to discover novel markers for targeting or enriching cell populations during iCM induction. To identify specific markers for each cell population along the reprogramming trajectory, we selected genes that were expressed significantly higher (for positive selection markers) or lower (for negative selection markers) in the cell population of interest than the other three populations (Tukey-adjusted P value <0.05 in pairwise comparisons after ANOVA; Extended Data Fig. 10a–f and Supplementary Table 7). Negative selection markers for iCMs appeared the most attractive as a supplement to cardiac positive selection markers. Among the top 20 negative markers for iCMs, we focused on four surface markers, Cd200, Cldc1, Tmsf1 and Vcam1 (Fig. 4c). Linear regression analysis suggests that the expression of these markers was highly anti-correlated with the reprogramming process and was barely detectable in iCMs (Fig. 4d). Further experimental validation confirmed that Cd200 was a negative selection marker (Fig. 4e, f), and knockdown of Cd200 did not affect reprogramming efficiency (Extended Data Fig. 10g–n).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
We have used single-cell transcriptomics analysis to gain insights into the heterogeneity of cells within an unsynchronized cardiac reprogramming system. The findings show promise for improving the efficiency and detection of iCM formation. We also anticipate that the experimental and analytical methods presented here, when applied in additional cell programming or reprogramming contexts, will yield crucial insights into cell fate determination and the nature of cell type identity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June 2016; accepted 22 September 2017.
Published online 25 October 2017.

1. Ieda, M. et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell 142, 375–386 (2010).
2. Jayawardena, T. M. et al. microRNA-mediated in vitro and in vivo direct reprogramming of cardiac fibroblasts to cardiomyocytes. Circ. Res. 110, 1465–1473 (2012).
3. Qian, L. et al. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. Nature 485, 593–598 (2012).
4. Song, K. et al. Heart repair by reprogramming non-myocytes with cardiac transcription factors. Nature 485, 599–604 (2012).
5. Dal-Pra, S., Hodgkinson, C. P., Mirotou, M., Kirste, I. & Dzau, V. J. Demethylation of H3K27 is essential for the induction of direct cardiac reprogramming by miR combo. Circ. Res. 120, 1403–1413 (2017).
6. Ma, H., Wang, L., Yin, C., Liu, J. & Qian, L. In vivo cardiac reprogramming using an optimal single polycistronic construct. Cardiovasc. Res. 108, 217–219 (2015).
7. Mathison, M. et al. In vivo cardiac cellular reprogramming efficacy is enhanced by angiogenic preconditioning of the infarcted myocardium with vascular endothelial growth factor. J. Am. Heart Assoc. 1, e005562 (2012).
8. Muraoka, N. et al. Chemo- and telomerase reprogramming in vivo in direct cardiac reprogramming. Circulation 135, 978–995 (2017).
9. Zhao, Y. et al. High-efficiency reprogramming of fibroblasts into cardiomyocytes requires suppression of pro-fibrotic signalling. Nat. Commun. 6, 8343 (2015).
10. Zhou, H., Dickson, M. E., Kim, M. S., Bassel-Duby, R. & Olson, E. N. Akt1/protein kinase B enhances transcriptional reprogramming of fibroblasts to functional cardiomyocytes. Proc. Natl Acad. Sci. USA 112, 11864–11869 (2015).
11. Ikhovits, J. L., Addis, R. C., Epstein, J. A. & Gearhart, J. D. Inhibition of TGFβ signaling increases direct conversion of fibroblasts to induced cardiomyocytes. PLoS ONE 9, e89678 (2014).
12. Liu, Z. et al. Re-patterning of H3K27me3, H3K4me3 and DNA methylation during fibroblast conversion into induced cardiomyocytes. Stem Cell Res. 16, 507–518 (2016).
13. Muraoka, N. et al. MiR-133 promotes cardiac reprogramming by directly repressing Snai1 and silencing fibroblast signatures. EMBO J. 33, 1565–1581 (2014).
14. Wang, L. et al. Stoichiometry of Gata4, MeF2c, and Tbx5 influences the efficiency and quality of induced cardiac myocyte reprogramming. Circ. Res. 116, 237–244 (2015).
15. Zhou, Y. et al. Bmi1 is a key epigenetic barrier to direct cardiac reprogramming. Cell Stem Cell 18, 382–395 (2016).
16. Shin, J. et al. Single-cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. Cell Stem Cell 17, 360–372 (2015).
17. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear gene networks from gene expression data. Genome Biol. 16, 6 (2015).
18. Polo, J. M. et al. A molecular roadmap of reprogramming somatic cells into iPSCs. Cell 151, 1617–1632 (2012).
19. Vaseghi, H. R. et al. Generation of an inducible fibroblast cell line for studying direct cardiac reprogramming. Genesis 54, 398–406 (2016).
20. Fu, X. D. & Ares, M. J. Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. 15, 689–701 (2014).
21. Park, J. W., Jung, S., Rouchka, E. C., Tseng, Y. T. & Xing, Y. rMAPS: RNA map analysis and plotting server for alternative exon regulation. Nucleic Acids Res. 44, W333–W338 (2016).
22. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data. Genome Biol. 15, 550 (2014).
23. Wang, L. et al. Improved generation of induced cardiomyocytes using a polycistronic construct expressing optimal ratio of Gata4, MeF2c and Tbx5. J. Vis. Exp. 105, e53426 (2015).
24. He, A. K., Gong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc. Natl Acad. Sci. USA 108, 5632–5637 (2011).
25. Waldron, L. et al. The cardiac Tbx5 interactome reveals a chromatin remodeling network essential for cardiac septation. Dev. Cell 36, 262–275 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank UNC AAC Core, HTSF Core, Flow Core for technical support. This study was supported by NIH HG00272 to J.F.P., NIH BD2K Fellowship (T32 CA201159) and NIH F31 Fellowship (HG008912) to J.D.W., NIH/NHLBI R00 HL109079 and American Heart Association (AHA) 15GRNT25530005 to J.L.H., AHA 13SDG17060010, Ellison Medical Foundation (EMF) AG-NS-1064-13, and NIH/NHLBI R01 HL128331 to L.Q., and gifts from H. McAllister and C. Sewell.

Author Contributions L.Q., Z.L. and L.W. conceived and designed the study, Z.L. and L.W. designed and performed single-cell RNA-seq; Z.L., L.W., Y.Z. and C.Y. prepared samples for microarray and bulk RNA-seq; L.W., Y.Z., H.M., H.R.V., C.Y., S.V., J.B.W., S.A. and M.Z. performed other experiments; Z.L., J.D.W. and J.F.P. performed data analysis and modelling, W.S. helped with statistical analysis. Z.L., J.D.W., J.L. and L.Q. wrote the manuscript, with extensive input from all authors. J.L. and L.Q. provided funding and overall supervision.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Published by Springer Nature. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.L. (jiangdong_liu@med.unc.edu) or L.Q. (li_qian@med.unc.edu).
METHODS

Mouse strains and plasmids. Transgenic CD1 mice that expressed α-MHC-promoter-driven GFP were described previously. All animal experiments conformed to the NIH guidelines (Guide for the Care and Use of Laboratory Animals) and UNC Qian Laboratory animal protocol 15.277.0. This protocol was approved by the University of North Carolina at Chapel Hill Institutional Animal Care and Use Committee (IACUC) that oversees the university’s animal care and use (NIH/PHS Animal Welfare Assurance Number: A3410-0; USDA Animal Welfare Assurance Number: S-5577-00; establishment registration number: 11-92004; AAALAC Institutional Number: 329). pMXs retroviral vectors containing mouse Gata4, Me2c, or Tbx5 were described previously. The empty pMXs and pMXs-puro retroviral vectors were purchased from Cell Biolabs and they contain a partial LacZ stuffer sequence and were therefore referred to as LacZ in this manuscript. pMXs-ΔRed and the polycistronic pMXs-puro-MGT were described previously. Lentiviruses containing Me2c, Gata4, or Tbx5 were cloned by replacing the GFP insert in pLenti-GFP-puro (Addgene 17448), Me2c, Gata4, or Tbx5 using BamHI and SalI. qTripZ-ΔT was cloned by removing the tet-on promoter and RFP in the qTripZ vector using SalI and MluI. The plasmid was described previously and all other shRNAs (pLKO.1-vector based, pHCh2, and single cells were captured with the C1 system. Bright field images were taken of each capture site. For experiments E3 and E5–7, day 3 + M + G + T-transduced cells. Four of the seven experiments (E3 and E5–7) contained cells treated with two different conditions in order to estimate the relative abundance of mouse mRNA between treatments. Specifically, for experiments E1, E2 and E4, cardiac fibroblasts transduced with M + G + T for 3 days were collected by trypsinization, stained with 7AAD or NearIR Live/Dead dye (Thermo Fisher Scientific), and FACS-sorted for live cells (negative for the Live/Dead dye). Pilot experiments showed an average diameter of 12.6 μm and a buoyancy of 7.5–2.5 cells (buoyancy buffer) of cardiac fibroblasts. Therefore, the sorted single-cell suspension (around 2,000 cells per μl) was loaded on a medium-sized (10–17 mm) microfluidic RNA-seq chip (C1 Single-Cell mRNA Seq IFC, 100–601, initially designed chips were used in E1–E3 and redesigned chips were used in E4–E7) and single cells were captured with the C1 system. Bright field images were taken of each capture site. For experiments E3 and E5–7, day 3 + M + G + T-transduced cells (E5 and E6) or untransduced (E3 and E7) cardiac fibroblasts were stained with the NearIR Live/Dead dye and 0.25–1 μm carboxyfluorescein succinimidyl ester (CFSE, Thermo Fisher Scientific), whereas the DsRed-transduced cardiac fibroblasts were stained with the NearIR Live/Dead dye only. Then, 12,000 CFSE- and NearIR-stained green fluorescent cells and 12,000 DsRed- and NearIR-stained red fluorescent cells were sequentially FACS-sorted into the same tube and mixed. For experiment E3, cell sorting was slightly different, and 700 of each of the CFSE single-positive cells (untransduced), DsRed single-positive cells (DsRed-transduced) and double-negative cells (from the DsRed-transduced wells but with no DsRed protein expression) were sorted into a single-cell suspension. After cell capture, fluorescent images of GFP and RFP channels as well as bright field pictures were taken.

Next, control RNA spike-ins were added into lysis mix A (see Fluidigm’s protocol), which were then loaded onto the IFC plate before cell lysis. Experiments E1 and E2 used the Ambion Array Control spike-ins (AM1780) that were included in the SMARTer kit. E1 used only spike 1, 4 and 7 according to Fluidigm’s protocol but at a concentration that is 100-fold higher than suggested, based on recommendations from the UNC Advanced Analytics Core (AAC) that provided the Fluidigm service. E2 used all 8 spike-ins contained in the kit at the following: 0.5 μl of other components to make lysis mix A: 10 pg/ml of spike 2 and with a 10-fold reduction for the next spike and so on. For E3, we used the Ambion spike-ins at half the concentration of those used in E2 and another spike-in, the External RNA Controls Consortium (ERCC) RNA spike-in Mix 1 (Ambion, Life Technologies) after an 80,000-fold dilution. For E4–7, only the Ambion spike-ins were used. After a 1.5 h of pre-plating and 1 cell/μl day 6. Day 0 sample spike-in was mixed with 10 μl of other components to make lysis mix A. Then cell lysis, reverse transcription and cDNA pre-amplification were performed on the chip according to Fluidigm’s standard protocol and the control RNA spike-ins were processed in parallel with cellular RNA. Differences in spike-ins added to each experiment reflected how the technology evolved over time during the progress of this project. To address the spike-ins issue, among others, we developed a pipeline described in the ‘Processing and normalization of single-cell RNA-seq data’ section to normalize and analyse all acquired useful data.

Illumina library preparation and sequencing. After in situ cDNA library preparation and the bright field images of each capture site (nest) on the chip were carefully examined. Forty-six empty nests, 30 nests with two or more cells, and 22 nests containing morphologically unhealthy cells out of 672 capture sites on seven chips were excluded from further analysis, resulting in 574 single-cell cDNA libraries. The size distribution and quality of cDNA libraries from each single cell were ensured by bioanalyzer. For E3 only, cDNA library concentrations

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Counts, cells in experiments that involved two treatments were classified as biological starting material in each cell. On the basis of the normalized DsRed ERCC spike-in and redesigned chip in E4), suggesting the existence of batch effects. By comparing biological replicate experiments, we found different mean total data Fig. 1g).

Mean probability of detection of all 12 spike-ins is 0.30, cells from that plate and this is the probability of detection for each spike-in (http://www-huber.embl.de/users/anders/HTSeq).

Gene expression was counted with Htseq-count using the union mode 30 Fig. 1d). This step removed 61 outliers from the 574 sequenced single cells, to spike-in or mouse genome for each single cell are detailed in Supplementary Information about the number of total reads and the percentages of reads mapped to each single cell by their unique Nextera barcode and sequencing reads without barcodes were received from the HTSF in.fastq format. For microarray and bulk RNA-seq samples, cellular RNA was extracted with TRIzol (microarray samples were further purified with the RNAeasy kit from Qiagen), and only samples with an RNA integrity number (RIN) above 8, as determined using a bioanalyzer, were further processed. Microarray samples were submitted to the HTSF for one-colour Cy-dye labelling and long oligo (60-mer) Agilent high-density microarrays. Bulk RNA-seq samples were prepared with the TrueSeq Stranded mRNA Library Prep Kit (Illumina). The barcoded Illumina libraries were pooled and submitted to the HTSF for sequencing. About 6 × 10⁸ 100-bp paired-end reads per sample were obtained and sequencing reads removed of Illumina indexes were received from HTSF in.fastq format.

Processing and normalization of single-cell RNA-seq data. The quality of sequencing results was first checked by FASTQC. Reads were high and no trimming was required. The raw reads were then mapped to the merged genome of mm10, ERCC, and E. coli K12 with TopHat2 using default settings. Information about the number of total reads and the percentages of reads mapped to spike-in or mouse genome for each single cell are detailed in Supplementary Table 1. Outliers showing high ratios of percentage reads mapped to spike-in to percentage reads mapped to mouse genome were removed (Extended Data Fig. 1d). This step removed 61 outliers from the 574 sequenced single cells, resulting in 513 high-quality single cells for analysis (Supplementary Table 2). Gene expression was counted with Htsq-count using the union mode 30 (http://www-huber.embl.de/users/anders/HTSeq).

Limit of detection of our single-cell RNA-seq was determined as previously described 29. In brief, the concentration of each ERCC spike-in in the lysis chamber was first calculated. For experiment E3, seven of the spike-ins were present at 1.24 molecules per chamber and were as follows: ERCC-00039, ERCC-00067, ERCC-00077, ERCC-00143 and ERCC-00150. For experiments E4-E7, five of the spike-ins were present at 1.24 molecules per chamber and were as follows: ERCC-00031, ERCC-00033, ERCC-00058, ERCC-00069 and ERCC-00134. The number of non-zero measurements of each spike-in was then counted. This number was divided by the total number of high-quality cells from that plate and this is the probability of detection for each spike-in at this concentration. Mean probability of detection of all 12 spike-ins is 0.30, consistent with previous findings 28 and suggesting single-molecule sensitivity of our experiments.

We developed a three-step normalization strategy in order to extract biologically meaningful information from all the single-cell RNA-seq data (Extended Data Fig. 1c). Firstly, we normalized mouse gene raw counts to each cell’s technical and biological size factors within each experiment using a previously described method 11. These two size factors account for technical variations within each experiment, such as amplification efficiency and differences in the amount of biologically relevant reads in each cell resulting from the normalized DsRed counts, cells in experiments that involved two treatments were classified as DsRed-transduced (E3R, E5R, E6R and E7R, expressing high levels of DsRed), or M + G + T-transduced (E5M, E6M) or untransduced cells (E3U, E7U; Extended Data Fig. 1g).

Secondly, we corrected for ‘batch effects’ that account for technical contributions to experiment-to-experiment variations due to different cell-capture efficiency, types/amounts of spike-ins and Fluidigm chips (Extended Data Fig. 1b), while preserving biological information, such as total mRNA abundance. By comparing biological replicate experiments, we found different mean total mRNA counts per cell (Extended Data Fig. 1b) that probably resulted from varying cell-capture efficiency per plate (68 sequenced cells in E5, 33% more compared to 51 cells in E6; Supplementary Table 2), various amounts of spike-ins used (100-fold more concentrated spike-ins in E1 than in E2) and different types of spike-ins and Fluidigm chips used (Fluidigm spike-in and previous chip in E2 and ERCC spike-in and redesigned chip in E4), suggesting the existence of batch effects. To determine whether different treatments affected mouse mRNA abundance in the cell, we also examined mean total mRNA reads from different treatments in the same experiment (Extended Data Fig. 1h). We found no difference in mean total mRNA counts between uninfected and DsRed-transduced cells (E3U versus E3R and E7U versus E7R) but 40% less counts in cells undergoing reprogramming (M + G + T-transduced, E5M, E6M) than DsRed-transduced cells (E5R, E6R), suggesting biological variations caused by treatment. Therefore, to retain mRNA abundance information while correcting for batch effects, we normalized each treatment to spike-in or mouse genome for each single cell by their unique Nextera barcode and sequencing reads without barcodes were received from the HTSF in.fastq format.
new hierarchical clustering for control cells in E3 was calculated using the four cell-lineage-related gene clusters but not the cell cycle genes identified in Fig. 2a. The calculated hierarchical clustering was very similar to that in Fig. 2a and was applied to reprogramming cells from E1 and E2 to generate Fig. 2c. For all correlation analyses, gene expression was always log-transformed before analysis. In Fig. 4a, d and Extended Data Fig. 9p, CCI cells were used. Linear regression was performed to obtain the regression coefficient (R value) and its corresponding P value (two-sided, α = 0.05). For correlation analysis of Tbx5 and its target genes in M–G + + T-transduced CCI cells and include sequences that were performed in B with the 'precomp' function using all of the 34,578 detected genes and the 3D plot was generated with the 'scatterplot3d' package in R (Extended Data Fig. 4p, q).

**Analysis of splicing.** We aligned the bulk RNA-seq data (100-bp, paired-end reads) to mm10 using Mapsplice version 2.1.4. To detect alternative splicing, we used rMAT35 version 3.2.5 with Ensembl GRCh38.82 gene annotations and the novelSS (novel splicing site) flag to identify unannotated splicing events. All other rMATs parameters were set to the default values. In Fig. 3j–l, n, o and Extended Data Fig. 9a–i, we used FDR < 0.05 and ∆PSI > 15 as cut-offs, resulting in 1,494 alternative splicing events for MGT and shPtp1 versus MGT and shNT, and 879 alternative splicing events for MGT versus LacZ (see Supplementary Tables 4, 5). In Fig. 3k, l to determine whether the direction (sign) of PSI change is consistent between two groups, we identified the overlapping alternative splicing events between the samples (69 overlapping events between MGT versus LacZ and cardiomyocytes versus cardiac fibroblasts, and 155 events between MGT and shPtp1 versus MGT and shNT, and cardiomyocytes versus cardiac fibroblasts). We then conducted a binomial test by first transforming the paired ∆PSI data into either +1 or −1 based on if their signs agreed or not, and then calculating the proportion of +1's and compare it to 50% using a one-sided binomial test. The results for ∆PSI (cardiomyocyte–cardiac fibroblasts) versus (MGT–LacZ) showed that for only 34.78%, the signs in these two groups agree with each other (P = 0.0077). Therefore, we conclude that there was enough statistical evidence to support that the directions (signs) in cardiomyocyte–cardiac fibroblasts and MGT–LacZ are different. The result for cardiomyocyte–cardiac fibroblasts versus shPtp1–shNT shows that for 83.22% the signs in two groups are the same (P = 2.2 × 10−16). Therefore we conclude that there is enough statistical evidence to support that the directions (signs) in cardiomyocyte–cardiac fibroblasts and shPtp1–shNT are the same. To plot the positional distribution of Ptp1-binding motifs, we used rMAPS version 1.0.5. The rMAPS tool can only be used for exon-skipping events and has a database of known binding motifs for RNA-binding proteins, including Ptp1. It considers exon-skipping events with FDR < 0.05 and ∆PSI > 5 as statistically significant and all others insignificant. Then it takes all exon-skipping events (both significant and insignificant) and uses the events that are not significant to create a background profile. We basically extracted the exon skipping events from the rMATs comparison of MGT and shPtp1 versus MGT and shNT and ran rMAPS on this list to generate Fig. 3m.

**Proliferation assays.** Lentiviruses were packaged by transfecting HEK293T cells with Lipofectamine 2000 as previously described22. For packaging shRNA viruses, a total of 10μg plasmids consisting of equivalent concentrations of each of the four or five shRNA targeting different regions of the gene were used. Lentiviral vectors expressing Mef2c, Gata4, Tbx5, inducible MGT (iMGT) and large T antigen were added to cultures. The expression of expressing Tet and the MGT construct under the control of a Tet-ON promoter. Therefore, we conclude that there was enough statistical evidence to support that the regression coefficients are significant. The regression coefficients were calculated in the linear hypothesis test. In Extended Data Fig. 5b–m, pMXs-puro-MGT was used for iCM induction. In the EdU–incorporation assay, cells were pulsed with 10μM of EdU for three days before staining with Alexa Fluor 647-labelled EdU of the Click-iT Plus Edu Alexa Fluor 647 Flow Cytometry Assay Kit (ThermoFisher Scientific, C10634). Propidium iodide (Life Technologies, P3566) staining was performed as previously described22. For IMGT induction, doxycycline was added at 1 μg ml−1. For all other experiments, doxycycline was added at 2 μg ml−1. Proliferation was measured by flow cytometry and the 1.1 multiplier of the 95% quantile for negative controls as a cut-off. Lastly, probe intensity data were log2-transformed and replicated probes for each gene were averaged and used for the analyses. As were performed in B with the 'precomp' function using all of the 34,578 detected genes and the 3D plot was generated with the 'scatterplot3d' package in R (Extended Data Fig. 4p, q).

**LETTER**
performed as previously described\(^\text{14}\) (anti-Ptbp1, Cell Signaling 8776, 1:500). Adult cardiac fibroblasts (AdCF) and tail-tip fibroblast (AdTTF) were isolated using the explant method as previously described\(^\text{15}\). Clone 271 was used for all shPtbp1-related experiments, except for the initial screen, whereas mixed clones of shCd200 viruses were used for shCd200-related experiments. Information on shRNAs are listed in Supplementary Table 8.

Immunofluorescence staining and flow cytometry were performed as previously described\(^\text{12}\). Primary antibodies were used at the following dilutions: rabbit anti-GFP (Invitrogen, A11122, 1:500), chicken anti-GFP (Abcam, ab13970, 1:1,500), anti-α-SMA (Sigma, A2547, 1:200), anti-SM22α (Abcam, ab14106, 1:200), anti-α-actinin (Sigma, A7811, 1:500), anti-Cx43 (Sigma, C6219, 1:200), APC-Thy1.2 (eBioscience, 17-0902-81, 1:100) and APC-Cd200 (Biolegend, 123809, 1:200). Images were captured using an EVOS FL Auto Cell Imaging System (Life Technologies). All images shown in this study were overlaid with Hoechst nuclear staining, except for live images. For quantification, 10–30 images from multiple repeated experiments were randomly taken at 10×, 20× or 40× magnification at the same exposure and then counted in a blinded way. For Extended Data Fig. 6g–i, 2 test. Application and results of these tests are described in detail in Methods.

Statistics. Unless otherwise stated, values are expressed as mean ± standard deviation (s.d.) or standard mean of error (s.e.m.) of multiple biologically independent samples. Statistical tests performed include Student’s t-test, one way ANOVA followed by post hoc correction, linear regression, Spearman correlation, Kolmogorov–Smirnov test, binomial test, likelihood ratio test and χ\(^2\) test. Application and results of these tests are described in detail in Methods and figure legends. Generally, *P* < 0.05 was considered statistically significant, **P** < 0.01 was considered highly significant and ***P*** < 0.001 was considered very highly significant. All data are representative of multiple repeated experiments.

Data availability. The RNA-seq data that support the findings of this study are available in the Gene Expression Omnibus (GEO) under the accession number GSE98571. Source Data for all figures are available in the online version of the paper.

26. Zheng, X. et al. Prolyl hydroxylation by EglN2 destabilizes FOXO3a by blocking its interaction with the USP9x deubiquitinase. Genes Dev. 28, 1429–1444 (2014).
27. Qian, L., Berry, E. C., Fu, J. D., Ieda, M. & Srivastava, D. Reprogramming of mouse fibroblasts into cardiomyocyte-like cells in vitro. Nat. Protoc. 8, 1204–1215 (2013).
28. Wu, A. R. et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11, 41–46 (2014).
29. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375 (2014).
30. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169 (2015).
31. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095 (2013).
32. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127 (2007).
33. Chen, C. et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE 6, e17238 (2011).
34. Han, H. et al. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. Nature 498, 241–245 (2013).
35. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386 (2014).
36. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).
37. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. Proc. Natl Acad. Sci. USA 111, E5593–E5601 (2014).
38. Pinto, A. R. et al. An abundant tissue macrophage population in the adult murine heart with a distinct alternatively-activated macrophage profile. PLoS ONE 7, e36814 (2012).
39. Whitfield, M. L. et al. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell 13, 1977–2000 (2002).
40. Gooding, C. & Smith, C. W. Tropomyosin exons as models for alternative splicing. Adv. Exp. Med. Biol. 644, 27–42 (2008).
41. Terenzi, F. & Ladd, A. N. Conserved developmental alternative splicing of muscleblind-like (MBNL) transcripts regulates MBNL localization and activity. RNA Biol. 7, 43–55 (2010).
42. Dube, D. K., McLean, M. D., Dube, S. & Poiesz, B. J. Translational control of tropomyosin expression in vertebrate hearts. Anat. Rec. (Hoboken) 297, 1585–1595 (2014).
Extended Data Figure 1 | See next page for caption.
Extended Data Figure 1 | Experimental design, analysis pipeline, quality control and normalization for single-cell RNA-seq.

**a**, Experimental workflow. Hearts were isolated from P1.5 neonatal mice and cells were dissociated by enzymatic digestion. Thy1$^+$ cells were then purified by MACS and plated overnight. The adherent cells (cardiac fibroblasts) were then transduced with retroviruses encoding the reprogramming factors Mef2c, Gata4 and Tbx5, DsRed or were left untransduced (mock). On day 3 after transduction, cells were trypsinized and Live/Dead stained. Additionally, for some experiments designed to examine the relative mouse RNA abundance in cells receiving different treatments, mock and M + G + T cells were labelled with a green cell tracker (CFSE), FACS-sorted for live cells, and then mixed at a designated ratio with FACS-sorted live DsRed$^+$ cells from a parallel DsRed transduction. The single-cell suspension was loaded onto a medium size chip (10–17 μm) and single cells were captured on a Fluidigm C1 machine. Bright field, and for some experiments, fluorescent images, were taken for all capture sites. Individual cDNA libraries for each cell were prepared in situ by reverse transcription with pre-amplification after adding RNA spike-in. Bright field and/or fluorescent images for each capture site were examined and libraries from nests with 0 or multiple cells were excluded from downstream analysis. Illumina libraries were then prepared for each cell, pooled, quality-checked and sequenced on Hiseq 2500.

**b**, Design of the seven independent single-cell RNA-seq experiments, including treatment, RNA spike-ins and Fluidigm chips used. **c**, Data analysis pipeline. Barcodes were trimmed from RNA-seq raw reads and the quality of these reads was confirmed with FASTQC. High-quality reads were mapped to the mm10 genome with TopHat2 and counted with Htseq count. Outliers were removed as described in **d**. The raw counts were normalized first to technical and biological size factors within each experiment using DEseq and then to experiment size factors calculated based on relative mouse mRNA abundance in cells receiving different treatments (**h**). Residual batch effects between experiments receiving the same treatment were removed using ComBat. Cell grouping and modelling were then performed using the normalized gene counts with PCA, hierarchical clustering, SLICER and more. The most important three quality control steps are labelled in red in a and c. The above strict quality control criteria ensured that only high-quality and biologically meaningful data from healthy single cells were analysed. **d**, For each of the seven single-cell experiments, the percentage of reads mapped to spike-in in each cell was plotted against percentage of reads mapped to mouse genome (left) or mouse mRNA (right) in that cell. Cells outside of the red circles are outliers. **e**, For each of the five single-cell experiments that contained ERCC spike-in, mean count numbers of each ERCC spike-in were plotted against their concentration in lysis mix A (see Fluidigm's protocol for details). Linear regression coefficients (R value) and corresponding P values (two-sided, α = 0.05) are shown. The results showed a dynamic range (~10$^3$) of ERCC concentration that covers the full spectrum of mouse gene expression levels. The high R values indicate a strong correlation between hypothetical molecular concentrations and measured gene counts in our experiments. **f**, Squared coefficients of variation (CV$^2$) were plotted against mean expression of ERCC spike-ins (left) or mouse genes (right) for experiments containing ERCC spike-in. **g**, DsRed counts in experiments E3 and E4–E7 plotted against Mef2c counts and/or total mouse mRNA counts after normalization for technical and biological size factors within each experiment (Methods). Cells in the four experiments were classified as DsRed-transduced (E3R, E5R, E6R, E7R), M + G + T-transduced (E5M, E6M) or untransduced cells (E3U, E7U) based on these plots. **h**, i, Normalization to experiment size factors to account for technical contributions to experiment-to-experiment variations, such as varied capture efficiency, while retain biological variations, such as differences in total mRNA abundance in cells receiving different treatments. **i**, Median total mouse mRNA counts were calculated for each treatment in each experiment and mean mRNA counts were compared between different experiments (E1, E2 and E4) or different treatments in one experiment (E3, E4–E7) with a two-sided Student’s t-test (α = 0.05). Experiment size factors were calculated based on the ratio of median mRNA counts between different treatments. After normalization to the experiment size factors, the median mRNA count was equal to 1,000,000 cells for uninfected and DsRed-transduced cells and 616,136 cells for M + G + T transduction. **j**, PCA of two biological replicates (E5 and E6) that have different sequencing depth per cell due to different capture efficiencies before (left) and after (right) normalization to experiment size factors. Top 400 PCA genes were used.
Extended Data Figure 2 | Normalization and outlier removal of single-cell RNA-seq data (continued). a–c, Removal of batch effects using ComBat on non-immune cells (described in Extended Data Fig. 3). PCA of all batches of M+G+T-transduced (a), DsRed-transduced (b) or uninfected (c) cells before (left) and after (right) ComBat normalization. d–g, After ComBat normalization, outliers of each treatment group were further removed by examination of the mean gene expression level of each cell (box plot, d) and PCA (e). Uninfected cells are shown as an example in d and e. A total of 454 healthy non-immune cells were further analysed. f, Pairwise comparison of mean mouse gene expression between different experiments and treatment conditions. Correlation coefficient calculated by linear regression is shown. g, Heat map coloured by correlation coefficient in f. The strongest correlation was seen within each treatment group. DsRed-infected and uninfected cells also showed strong intertreatment correlation. M+G+T-transduced cells showed relatively low correlation with DsRed-transduced and uninfected cells.
Extended Data Figure 3 | Single-cell analysis identified a subpopulation of immune-like cells. a–c, Data from 513 control or reprogramming single cardiac fibroblasts and bulk RNA-seq data of neonatal cardiac fibroblasts and cardiomyocytes were analysed with PCA. To identify groups of related cells and genes, the top 400 genes with highest weight in the first three principal components were then analysed by unsupervised hierarchical clustering (a) and PCA (b, c). Representative genes in each of the four gene clusters identified by hierarchical clustering are listed to the right of hierarchical clustering heat map. Notably, in addition to cardiomyocyte, fibroblast and cell-cycle genes, immune response genes were identified as the other major gene cluster. b, PCA loading plot showing four major gene clusters. c, PCA score plot. Both hierarchical clustering and PCA results showed that bulk cardiac fibroblast and cardiomyocyte data were very close in distance and both of them were clustered together with single cells expressing high levels of immune genes. The dashed line separates immune-like single cells and other single cells. d, Markers of the major immune cell lineages. e, Violin plots of the expression of major immune cell lineage markers in bulk cardiac fibroblasts, bulk cardiomyocytes, immune-like single cells and other single cells. f, Expression of the macrophage marker Cd14 and the dendritic cell marker Cd11c (also known as Itgax) in each immune-like cell showed that 42 cells express macrophage markers and 3 express dendritic cell markers, with 4 cells expressing both. These data suggest that the immune-like cells are probably cardiac resident immune cells, which also express cardiac fibroblast markers, such as Thy1. Although follow-up work to delineate the potential of these immune cells to be reprogrammed into iCMs will be of great interest, it is not the focus of this study. Therefore, for all subsequent analyses of the single-cell data, we focused on the non-immune cardiac fibroblasts.
Extended Data Figure 4 | See next page for caption.
Extended Data Figure 4 | Cell grouping, representative gene expression, and number of genes detected in single-cell RNA-seq data and population-based gene expression profiling during reprogramming. Related to Fig. 1a–h. a, PCA score plot showing variance of top 10 PCs. Related to Fig. 1a–c. b, Violin plots showing the expression levels of representative cardiac, fibroblast and cell-cycle genes in the seven cell groups identified by hierarchical clustering and PCA (Fig. 1a–c). c, Gene ontology analysis from Fig. 1a showing the P value of each gene ontology term. d–g, Determination of the proliferation status of each single cell using genes periodically expressed during the cell cycle that were identified in a previous report. d, A 3D LLE plot calculated based on the expression of these cell cycle genes in each of the 454 cells. e, Frequency of cells in LLE component 3. The dark red plane in d and the red dotted line in e indicate the threshold for proliferating (Pro) and non-proliferating (NP) cells. f, g, PCA plots as in Fig. 1c, but colour- and shape-coded for proliferating and non-proliferating (I) or CCA and CCI (g) cells. h, i, ISNE plots of all single cells colour- and shape-coded by hierarchical clustering and PCA cell groups (h) or proliferating and non-proliferating cell groups (i). The cells that were grouped as intermediate fibroblasts, pre-iCMs and iCMs constituted 30.6% (77 out of 252), 24.6% (62 out of 252) and 44.8% (113 out of 252) of all cells transduced with M + G + T, respectively. In contrast to previous population- and marker-based studies, our single-cell RNA-seq data suggests that the fate conversion from fibroblast to iCM occurs rapidly (approximately 3 days) with nearly 45% of the cardiac fibroblasts exhibiting transcriptomic signatures indicative of a cardiac fate. j, Live fluorescent images of day-5 MGT-transduced cardiac fibroblasts showing co-expression of α-MHC–GFP and Thy1 (surface labelling). Double-positive cells are labelled with an asterisk (*). All images were taken at 40× magnification. Scale bar, 100 μm. k, α-MHC–GFP+ Thy1+ and α-MHC–GFP– Thy1+ cells were FACS-sorted from day-7 MGT-transduced cardiac fibroblasts and expression of representative cardiac (Myl4 and Actc1) and fibroblast (Col3a1 and Postn) markers were determined by qRT–PCR. Day-7 mock-transduced cells were included as control. Data are mean ± s.d. n = 4 samples. One-way ANOVA followed by Bonferroni correction (two-sided). ***P < 0.01, **P < 0.001; ns, not significant. Myl4 and Actc1 expression increased 80–100-fold and reached approximately the same level as Gapdh in α-MHC–GFP+ Thy1+ cells compared to mock transduction. Expression level of the fibroblast marker Postn was maintained at a high level in GFP+ Thy1+ cells. For the other fibroblast marker, Col3a1, although its relative expression in GFP+ Thy1+ cells was decreased compared to mock-transduced and GFP+ Thy1+ cells, but its absolute expression was still high compared to Gapdh (around 1.4-fold of Gapdh). The data strongly support the existence of cardiomyocyte- and fibroblast-marker double-positive pre-iCM and suggest that the pre-iCM state represents an intermediate cell population that is transitioning from intermediate fibroblast to iCM or that is locked between intermediate fibroblast and iCM during reprogramming.

I. To determine whether iCMs may be differentiated from rare cardiac stem/progenitor cells, we plotted the expression of cardiac stem/progenitor cell markers in each of the hierarchical clustering and PCA single-cell groups using violin plots. All of these markers were nearly undetectable in fibroblasts, intermediate fibroblasts, pre-iCMs and iCMs, suggesting the direct conversion from cardiac fibroblast to iCM without going through a stem/progenitor cell stage. m, Distribution of gene expression levels in single cells. Data are mean ± s.e.m. n = 454 cells. The limit of gene detection was set to 1 based on this plot. n, Distribution of the number of genes detected in all, CCI or CCA single cells. Comparison of the distributions in CCI and CCA cells using a two-sample Kolmogorov–Smirnov test resulted in a one-sided P value of 5.248 × 10−11, suggesting that the number of genes in CCI is significantly smaller than in CCA. On the basis of this result, only CCI cells were used in o, Distribution of the number of genes detected in each CCI cell group. Analysis using a one-sided, two-sample Kolmogorov–Smirnov test (P values: 0.00521 for intermediate fibroblasts versus fibroblasts, 0.00481 for pre-iCMs versus intermediate fibroblasts and 1.104 × 10−6 for iCMs versus pre-iCMs) suggests that the number of genes expressed decreased when the cells adopted the iCM fate. This observation demonstrates a dynamic re-patterning of transcription machinery during reprogramming and is consistent with the hierarchical clustering analysis and experimental evidence that pre-iCMs co-expressed both cardiac and fibroblast markers, further indicating that the pre-iCM state constitutes an intermediate population during iCM reprogramming. p–v, Population-based gene expression profiling of reprogramming cardiac fibroblasts at day 0, 3, 5, 7, 10 and 14. p, q, Results from PCA analysis using all genes were similar to those using the top 400 genes (r–v). p, Scree plot of the top 10 PCs. q, A 3D PCA score plot. r–v, Analyses using the top 400 PCA genes. Related to Fig. 1g, h, r, Scree plot of top 10 PCs. s, PCA score plot using PC1 and PC3. t, Hierarchical clustering identified four major gene clusters: gradually upregulated during reprogramming (red, mainly cardiac genes), downregulated in MGT-transduced compared to LacZ-transduced (blue, mainly extracellular matrix (ECM) genes) and gradually upregulated (light grey) or downregulated (dark grey) in both LacZ and MGT cells (culture or viral effects, mainly ECM and immune-response genes). The results are consistent with the expression of representative genes selected from single-cell data (Fig. 1h), showing gradually increased expression of cardiomyocyte markers during reprogramming, first increased and then decreased expression of cell-cycle genes in both MGT and LacZ cells, suggesting the cells may be differentiated from rare cardiac stem/progenitor cells. u, Heat map showing loading (weight) of the genes in t in PC1, 2 and 3. Upregulated (cardiac) genes are highly weighted in PC1, and the other three gene clusters are highly weighted in PC2 and PC3. The results are consistent with s and Fig. 1g–v. Gene ontology analysis of the four gene clusters in t, showing gene ontology terms and their corresponding P values (listed on the right).
Extended Data Figure 5 | Inhibition of cell proliferation or cell-cycle synchronization promotes iCM reprogramming. Related to Fig. 1i–p.

a, Comparison of the ratio of CCA:CCI cells in the three treatment groups: uninfected, DsRed-infected and M + G + T-infected. Analysis using a χ² test suggests that proliferation states were not significantly different among the treatment groups at day 3.

b, c, Knockdown (KD) efficiency of shRNAs (b) or overexpression (OE) levels (c) of cell cycle-related genes were determined by qRT–PCR on day-4 lentiviral transduced cells. shNT, non-targeting control shRNA. Data are mean ± s.d. n = 3 samples.

d–i, Cell-cycle staging of cardiac fibroblasts simultaneously transduced with reprogramming factors and shRNA (e, g) or overexpression (f, h) constructs by propidium iodide (PI) staining. e, f, Flow cytometry histogram of propidium iodide staining intensity. g–i, Percentages of cells in G0/G1, S or G2/M phases were calculated based on e and f. i, Summary of g and h. j–m, Measurement of DNA synthesis in cardiac fibroblasts simultaneously transduced with reprogramming factors and shRNA (k) or overexpression (l) constructs by EdU incorporation assay followed by flow cytometry. dMFI, difference (Δ) in median EdU fluorescence intensity between EdU⁺ cells and EdU⁻ cells. m, Summary of k and l. Constructs that markedly decreased or increased cell proliferation were labelled in red in g–i, k–m and were used for experiments in n–s.

n–s, The effect of manipulation of cell proliferation through knockdown or overexpression of cell-cycle-related genes on iCM reprogramming. Reprogramming factors were introduced by lentiviral vectors instead of retroviral vectors to avoid retroviral infection bias of proliferating cells. Cardiac fibroblasts were simultaneously transduced with lentiviral Mef2c, Gata4 and Tbx5 (n–p) or inducible MGT (imGT, q–s) and lentiviral knockdown or overexpression constructs that markedly decreased (o, r) or increased (p, s) cell proliferation. Percentages of α-MHC–GFP⁺ and cTnT⁺ cells were quantified by flow cytometry. t–z, The effect of large T antigen transduction on iCM reprogramming. Cardiac fibroblasts were simultaneously transduced with reprogramming factors and lentiviral large T antigen. After 10 days, α-actinin⁺ cTnT⁺ cells were immunostained, imaged and quantified by counting randomly selected 20× fields from multiple repeated experiments (u–x). Both percentages of positive cells per field (v, x) and numbers of positive cells per field (w) were quantified. Percentages of cells showing a sarcomere structure in α-actinin⁺ cells were also quantified (y–z). The percentage of α-actinin⁺ cells that show sarcomere structures decreased from 50% to 0% upon large T transduction and accelerated proliferation. u, y, Representative images at 40× magnification with Hoechst nuclear staining. Scale bars, 100 μm.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Extended Data Figure 6 | Heterogeneity of the isolated cardiac fibroblasts (Thy1+ non-immune, non-cardiomyocyte cardiac cells) and stepwise suppression of non-cardiomyocyte lineages during iCM reprogramming. Related to Fig. 2. a–c, Limited transcriptome change by retrovirus transduction. To determine whether introduction of viruses could influence cellular identities of cardiac fibroblasts, molecular features of the uninfected and DsRed-transduced cells were compared and only 25 genes were differentially expressed (ANOVA, \( P < 0.05 \)), many of which were related to immune response (data not shown), suggesting that uninfected and virus-infected cardiac fibroblasts shared very similar gene expression profiles. a, PCA of the control cells from experiment E3 as shown in Fig. 2b, but colour-coded by treatment. The results showed that uninfected and DsRed-transduced cells were indistinguishable by PCA, suggesting limited global transcriptome changes by retroviral transduction. b, Violin plots showing the expression of representative cardiomyocyte, fibroblast and cell-cycle genes in uninfected- and DsRed-transduced cardiac fibroblasts. Retroviral transduction does not affect the expression of these genes. c, Same as a, but with all control cardiac fibroblasts from experiments E3, E5R, E6R and E7. On the basis of the results from a–c, we concluded that retroviral transduction does not influence cellular identities of cardiac fibroblasts and therefore we analysed control cardiac fibroblasts containing both uninfected and DsRed-transduced cells together in Fig. 2a, b, d. Gene ontology analysis from Fig. 2a showing the \( P \) value of each gene ontology term. e, Violin plots showing the expression of additional non-cardiomyocyte lineage markers in cardiac fibroblasts. Related to Fig. 2a, f, PCA analysis of control cardiac fibroblasts from all four experiments (E3, E5R, E6R and E7) with cells colour-coded by non-cardiomyocyte lineage groups. Related to Fig. 2b, g, Immunostaining of Thy1 and α-SMA, or Thy1 and CD31 of day-7 explant cardiac fibroblast cultures. Images were taken at 20× magnification. Scale bars, 200 μm. h, i, Representative flow cytometry plots (h) and quantification (i) of α-SMA\(^{-} \) and CD31\(^{-} \) cells in Thy1\(^{+} \) cells. There were 72.6% α-SMA\(^{-} \) and 9% Cd31\(^{-} \) cardiac fibroblasts, consistent with the single-cell RNA-seq data in Fig. 2a, showing a high percentage of cells expressing myofibroblast/smooth muscle markers and a low percentage of cells expressing endothelial markers. Data are mean ± s.d. j, Violin plots showing the expression levels of additional lineage markers. Related to Fig. 2c, k, Same as j, but using cells from experiments E4–E7. These experiments were performed using the redesigned Fluidigm medium chip as a repeat of experiments E1–E3 (Fig. 2c), which used the original Fluidigm medium chip. l, m, Tracking of protein expression of a myofibroblast/smooth muscle cell marker (α-SMA) by co-staining with α-MHC–GFP in cardiac fibroblasts during reprogramming at days 5, 7, 10, 14 and 21. l, Representative images at 40× magnification with Hoechst nuclear staining. Scale bar, 100 μm. m, Quantification of α-MHC–GFP\(^{+} \) and α-SMA high, low or negative cells. Data are mean ± s.e.m. The results showed that as reprogramming proceeded, protein expression of Thy1, SM22α (Fig. 2d–f) and α-SMA in α-MHC–GFP\(^{+} \) cells decreased over time, with no Thy1\(^{+}\)SM22α\(^{+}\), high α-SMA\(^{+}\) cells and around 50–60% of Thy1\(^{−}\)SM22α\(^{−}\)α-SMA\(^{−}\) cells on day 21 of reprogramming.
Extended Data Figure 7 | Identification of regulatory pathways involved in iCM reprogramming and screening of a shRNA library against major splicing factors during iCM induction. 

a–d, Three clusters of genes that significantly related to, and showed similar trends during, the reprogramming process were identified by nonlinear regression (see Methods). The number of genes included in each cluster is shown in parentheses. The solid line in each plot shows the overall trend of the cluster, and the grey colour indicates the 2D density of gene trends passing through each region of the plot.

b–d, Gene ontology analysis of genes in the three clusters showing gene ontology terms with FDR < 0.05.

e, f, Screening of a shRNA library of splicing factors for key regulators of iCM reprogramming. icMEFs were induced by doxycycline to express MGT and at the same time were transduced with lentiviruses encoding shRNA targeting various splicing factors. On day 3 after transduction, knockdown efficiency was determined by qRT–PCR (e; n = 6 samples from two independent experiments) and α-MHC–GFP+ cells were quantified by flow cytometry (f; n = 3 samples, data representative of three independent experiments). Data are mean ± s.d. Knockdown of Ptbp1 led to the highest fold increase in percentage of α-MHC–GFP+ cells compared to shNT. 

g, h, Ptbp1 expression in freshly isolated cardiac fibroblasts and cardiomyocytes was determined by qRT–PCR (g; mean ± s.e.m., n = 8 samples from two independent experiments) or western blotting (h).

e, g, Two-sided Student’s t-test. **P < 0.01, ***P < 0.001.
Extended Data Figure 8 | Manipulation of Ptbp1 through loss- and gain-of-function during iCM reprogramming. a, b, Ptbp1 knockdown efficiency of different shRNA clones in day-3 transduced mouse embryonic fibroblasts (MEFs) determined by qRT–PCR (a, mean ± s.e.m., data representative of three independent experiments) or western blotting (b). shPtbp1-271 showed the highest knockdown efficiency (>97%) and was used for subsequent experiments. c–u, Knockdown (shPtbp1, c–p) or overexpression (lentiviral OE-Ptbp1, q–u) of Ptbp1 in neonatal cardiac fibroblasts (neoCF) (c–h, i–l), AdCF (i–l) or AdTTF (m–p) when iCM reprogramming was induced by MGT (except in e–h, where M + G + T was used as a further confirmation). After 10 days (14 days for OE-Ptbp1), expression of cardiac markers was determined by immunostaining followed by imaging and blinded quantification (e, f, i, j, m, n, r, s) or flow cytometry (c, d, g, h, k, l, o, p, t). e, i, m, r, Representative 20× images with Hoechst nuclear staining. Scale bars, 200 μm. f, j, n, s, n = 10–20 images. Data are mean ± s.e.m. c, g, k, o, Representative flow cytometry plots. Percentages of cells are shown. d, h, l, p, t, Quantification of flow cytometry data measured in triplicate. Data are mean ± s.d. q, Ptbp1 overexpression was verified by qRT–PCR (mean ± s.d.). u, Expression levels of representative cardiac (left axis, Tnnt2, Actc1 and Ryr2) and fibroblast (right axis, Col3a1) markers were determined by qRT–PCR (mean ± s.d.). Mock, untransduced cardiac fibroblasts. Where appropriate, a two-sided Student’s t-test or one-way ANOVA followed by Bonferroni correction (two-sided) was performed. *P < 0.05, **P < 0.01, ***P < 0.001; ns, not significant.
Extended Data Figure 9 | See next page for caption.
Extended Data Figure 9 | Splicing re-patterning and transcriptome shift underlying shPtbp1-mediated enhancement of iCM reprogramming, and correlation of Mef2c, Gata4 and Tbx5 expression and reprogramming. a–n, Splicing analyses of day-3 reprogramming cells upon Ptbp1 silencing. Related to Fig. 3j–q. a, Number of overlapping and non-overlapping alternative splicing events identified between MGT versus LacZ and MGT and shPtbp1 versus MGT and shNT. The minimal overlap suggests that Ptbp1 knockdown caused extensive re-patterning of the splicing landscape during iCM induction. b, Number of exon-skipping events that skip (grey) or include (red) the exon more in MGT and shPtbp1 compared to MGT and shNT. c, Gene ontology analysis of alternative splicing genes between MGT and LacZ. d, e, A total of 138 alternative splicing events (83 genes) between MGT and shPtbp1 and MGT and shNT were reported by rMATS to be the most significant (and MGT and shNT and MGT and shNT). Tropomyosins are essential genes for muscle contraction and they are known to undergo extensive alternative splicing. The most studied tropomyosin alternative splicing events have been mutually exclusive exons and it is interesting to find two exon-skipping events of Tpm1 as the top 2 alternative splicing events upon Ptbp1 knockdown during reprogramming. AS, alternative splicing event; A3SS, alternative 3′ splicing site; ES, exon-skipping event; IR, intron retention. e, Gene ontology analysis of the most significant alternative splicing genes. f–i, Sashimi plots for the rest of the representative alternative splicing events in the blue-labelled gene ontology terms in Fig. 3n. The event shown in Fig. 3o was for Mbnl1, which is an essential splicing factor for cardiac function that switches isoforms during heart development. The event shown in h was an exon-skipping event in Tpm1 exon 3 (exon 2b in older literature), which was also the top event in d. This exon-3-skipped isoform of Tpm1 (Tpm1α) is the one enriched in cardiac and striated muscle cells, regulating the assembly and functionality of actin filament for contraction. j, Overlap of alternative splicing genes and differentially expressed genes between MGT and shPtbp1 versus MGT and shNT cells. k, Overlap of differentially expressed genes between MGT versus LacZ and MGT and shPtbp1 versus MGT and shNT. l, Percentage of overlapping differentially expressed genes in k, showing the same or opposite direction of changes. m–o, On the basis of the results shown in k, top gene ontology terms for differentially expressed genes between MGT and LacZ (m), overlapping differentially expressed genes (n) and differentially expressed genes (DEG) between MGT and shPtbp1 and MGT and shNT (o) are shown. p–s, Correlation between Mef2c, Gata4 and Tbx5 expression and reprogramming. Related to Fig. 4a. p, Correlation between the total expression of M, G, and T in individual cells and the SLICER-calculated reprogramming progress of each cell. Trend line and the correlation coefficient by linear regression are shown (P = 3.9 × 10−78, α = 0.05, two-sided). q, r, Expression levels of Mef2c, Gata4 and Tbx5 in fibroblasts, intermediate fibroblasts, pre-iCMs and iCMs plotted as mean ± s.e.m. (q) or violin plots to show distribution (r). s, Ratios of expression levels of Mef2c, Gata4 and Tbx5 in the four cell groups. Data are mean ± s.e.m. t, u, Spearman correlation between Mef2c, Gata4 and Tbx5 expression and the expression of 178 known and predicted splicing factors or 1,602 additional transcription factors. Genes with correlation coefficient >0.3 or <−0.3 with one or more of Mef2c, Gata4 and Tbx5 were selected and the intercorrelation matrix of the 17 selected splicing factors (t) and the 65 selected transcription factors (u) were calculated and plotted as a heat map. The splicing factors Mbnl1 and Rbms3 are strongly anti-correlated with expression of Mef2c, Gata4 and Tbx5, and Rbms20 is the only factor that is positively correlated with expression Mef2c, Gata4 and Tbx5 (P < 1 × 10−7) by two-sided Spearman correlation, α = 0.05. In u, two sets of genes, A and B, were found to be strongly anti-correlated with Mef2c, Gata4 and Tbx5 expression and that were also strongly co-expressed. These genes include Id1, Id2, Id3, Tcf21 and Foxp1 (P < 1 × 10−15 by two-sided Spearman correlation, α = 0.05) that might serve as ‘secondary’ key factors that further trigger the activation/inhibition of downstream cascades for successful conversion from fibroblasts to iCMs.
Extended Data Figure 10 | Putative markers for iCMs and pre-iCMs.
Analysis using an ANOVA identified 7,624 differentially expressed genes among fibroblasts, intermediate fibroblasts, pre-iCMs and iCMs. There were 954 and 285 candidates for negative and positive selection markers of iCMs and 55 candidates for positive markers of pre-iCMs. These candidates were expressed the lowest and highest in iCMs and highest in pre-iCMs, respectively. No gene passed the selection criteria for negative markers of pre-iCMs. Top candidates were selected by largest fold change in expression in the cell population of interest compared to the expression in fibroblasts.

**a**, Violin plots showing the expression of non-surface genes in top 30 candidates for negative markers of iCMs. Related to Fig. 4d. **b**–**e**, Top 30 candidates for positive selection markers of iCMs (b, c) or pre-iCMs (d, e). **b**, Fold change in gene expression in iCM/fibroblast (b) or pre-iCM/fibroblast (d). **c, e**, Violin plots of the same genes in four cell populations. **f**, Top 30 genes showing largest expression fold change in pre-iCMs and iCMs. **g**–**n**, Effect of Cd200 knockdown (g–j) or overexpression (k–n) on iCM reprogramming. Cardiac fibroblasts were untransduced (mock) or simultaneously transduced with MGT and lentiviral shNT or shCd200 or LacZ and OE-Cd200 for 14 days. Knockdown or overexpression efficiency was verified by qRT–PCR (g, k). Data are mean ± s.d. n = 3 samples. Percentages of α-MHC–GFP+ cTnT+ and Cx43+ cells were determined by immunostaining followed by imaging and blinded quantification (h, i, l, m) with representative 20× images in h, l. Scale bars, 200 μm. n = 20 (i) or 10 (m) images. Data are mean ± s.e.m. Percentages of α-MHC–GFP+, cTnT+ and double-positive cells were also quantified by flow cytometry (j, n). Data are mean ± s.d. n = 3 samples. Two-sided Student’s t-test was used. ***P < 0.001; ns, not significant.
Experimental design

1. Sample size
   Describe how sample size was determined.

   For single cell RNA-seq, the Fluidigm C1 system has a maximum capturing capacity of 96 cells per plate. Depending on the capture efficiency in that plate, we generally had 60-90 healthy single cells per plate. For bulk RNA-seq, sample size were two (biological replicates). For cell culture experiments such as immunostaining, flow cytometry, and qRT-PCR, sample size were usually three. For imaging analysis, 10 - 40 fields were counted.

2. Data exclusions
   Describe any data exclusions.

   Multimers and empty capture sites on the C1 plate were excluded for downstream cDNA library construction. After gene counting, each cell’s % reads mapped to spikein and % reads mapped to mouse mRNA were plotted and outliers are removed. After Combat normalization, the last few outliers were removed using the default "outlier removal" function in the SinGuLAR analysis package from Fluidigm, which was based on median gene expression level in each cell.

3. Replication
   Describe whether the experimental findings were reliably reproduced.

   The number of replication plates for scRNA-seq were: five plates for M+G+T-transduced cells (three full plates and two half plates), four plates for DsRed retroviral transduction control cells (four half plates), and two half plates for untransduced control cells. Experimental findings from the scRNA-seq data were reproducible. All other cell culture experiments were also repeated for usually three or more times. Conclusions were only made when results were consistent among replicate experiments.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.

   For scRNA-seq data, we intentionally mixed differently-treated samples in the sample plate (plate M4, M5, C1, and C2) for the purpose of randomization. For most cell culture experiments, at least one other postdoc repeated the experiment and verified the results. For important findings such as the role of Ptbp1 knocking-down on iCM reprogramming, all four participating postdocs have repeated the experiment and confirmed the finding.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   Blinding was widely used in the study. Data collection and analysis, such as immunostaining, image counting, RNA extraction, and qRT-PCR, were frequently performed by participants other than the experiment designer. During these data collection and analysis steps, the participants were routinely blinded to group allocation.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a

- **Confirmed**

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

- A statement indicating how many times each experiment was replicated

- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

- A description of any assumptions or corrections, such as an adjustment for multiple comparisons

- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted

- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

- Clearly defined error bars

*See the web collection on statistics for biologists for further resources and guidance.*

### Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

- fastqc/0.11.4, tophat/2.1.0, samtools/0.1.19, bedtools/2.24.0, bamtools/1.0.2, bowtie2/2.2.6, pico/5.04, htsq-cound/0.6.1p1, Combat, R, DESeq, DESeq2, SINGuLAR/3.5.2, Rtsne, prcomp, SLICER, LLE, rMATS/3.2.5, rMAPS/1.0.5, VGAM, FlowJo 7.6.1, Prism5

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

### Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

- No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

- rabbit anti-GFP (Invitrogen, A11122, 1:500), chicken anti-GFP (Abcam, ab13970, 1:1500), anti-αSMA (Sigma, A2547, 1:200), anti-SM22α (Abcam, ab14106, 1:200), anti-αActinin (Sigma, A7811, 1:500), anti-Cx43 (Sigma, C6219, 1:200), APC-Thy1.2 (ebioscience, 17-0902-81, 1:100), APC-Cd200 (Biolegend, 123809, 1:200), cy3-αSMA (Sigma, C6198, 1:500), APC-Thy1.1 (ebioscience, 17-0900-82, 1:100), PE-CD31 (Biolegend, 102408, 1:200), and anti-Ptbp1 (Cell Signaling 8776, 1:500).

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

- platE cells (Cell Biolabs), 293T cells (ATCC)

b. Describe the method of cell line authentication used.

- Virus packaging and transgene expression were successful.

c. Report whether the cell lines were tested for mycoplasma contamination.

- All cell lines were tested negative for mycoplasma contamination.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

- No commonly misidentified cell lines were used.
Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals
Provide details on animals and/or animal-derived materials used in the study.

Transgenic mice of CD1 background that contain αMHC promoter driven-GFP were used. Breeders of 2-6 months old were set up and P1.5 pups were collected for cardiac fibroblast or tail-tip fibroblast cell isolation. Mouse embryonic fibroblasts were isolated from E13.5 embryos.

Policy information about studies involving human research participants

12. Description of human research participants
Describe the covariate-relevant population characteristics of the human research participants.

This study does not involve human research participants.