CellNet: Network Biology Applied to Stem Cell Engineering

Patrick Cahan,[1,2,3,8] Hu Li,[4,8] Samantha A. Morris,[1,2,3,8] Edroaldo Lummertz da Rocha,[5,6,7] George Q. Daley,[1,2,3,9,*] and James J. Collins[5,9,*]

[1]Stem Cell Transplantation Program, Division of Pediatric Hematology and Oncology, Manton Center for Orphan Disease Research, Howard Hughes Medical Institute, Boston Children's Hospital and Dana Farber Cancer Institute, Boston, MA 02115, USA

[2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[3]Harvard Stem Cell Institute, Cambridge, MA 02138, USA

[4]Center for Individualized Medicine, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic College of Medicine, Rochester, MN 55905, USA

[5]Howard Hughes Medical Institute, Department of Biomedical Engineering and Center of Synthetic Biology, Boston University, Boston, MA 02215, USA

[6]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA

[7]Graduate Program in Materials Science and Engineering, Federal University of Santa Catarina, 88040-900 Florianópolis, Brazil

[8]Co-first author

[9]Co-senior author

*Correspondence: george.daley@childrens.harvard.edu (G.Q.D.), jcollins@bu.edu (J.J.C.)

http://dx.doi.org/10.1016/j.cell.2014.07.020

SUMMARY

Somatic cell reprogramming, directed differentiation of pluripotent stem cells, and direct conversions between differentiated cell lineages represent powerful approaches to engineer cells for research and regenerative medicine. We have developed CellNet, a network biology platform that more accurately assesses the fidelity of cellular engineering than existing methodologies and generates hypotheses for improving cell derivations. Analyzing expression data from 56 published reports, we found that cells derived via directed differentiation more closely resemble their in vivo counterparts than products of direct conversion, as reflected by the establishment of target cell-type gene regulatory networks (GRNs). Furthermore, we discovered that directly converted cells fail to adequately silence expression programs of the starting population and that the establishment of unintended GRNs is common to virtually every cellular engineering paradigm. CellNet provides a platform for quantifying how closely engineered cell populations resemble their target cell type and a rational strategy to guide enhanced cellular engineering.

INTRODUCTION

Transitions between cellular states are fundamental to development, physiology, and pathology. Directing state transitions in vitro is a current preoccupation of stem cell biology, as the derived cells can be used to investigate otherwise inaccessible cell types in development and disease, for drug screening, and for regenerative cell therapies. Dramatic cell-state transitions have been achieved in vitro and in vivo through the enforced expression of transcription factors. For example, differentiated somatic cells—fibroblasts (Takahashi and Yamanaka, 2006), keratinocytes (Aasen et al., 2008), peripheral blood (Loh et al., 2010; Staerk et al., 2010), and neural progenitors (Kim et al., 2009)—have been reprogrammed to pluripotent stem cells; fibroblasts have been converted to cells resembling myoblasts (Davis et al., 1987), motor neurons (Vierbuchen et al., 2010), cardiomyocytes (Ieda et al., 2010), hepatocytes (Huang et al., 2011; Sekiya and Suzuki, 2011), and blood progenitors (Szabo et al., 2010); B cells have been converted to macrophage-like cells (Xie et al., 2004); and exocrine pancreas cells have been converted to insulin-producing beta cells (Zhou et al., 2008). Furthermore, pluripotent stem cells can be coaxed to specific lineages through a combination of defined growth conditions and ectopic gene expression (Murry and Keller, 2008).

The widespread practice of cellular engineering has raised critical questions about the relationship of the derived cells to their native counterparts. To what extent does a cell population engineered in vitro resemble the corresponding target cell or tissue in both molecular and functional terms? While functional complementation via transplantation in live animals has been used to assess the ability of engineered cells to mimic the physiology of their native counterparts, such experiments are technically challenging, lack quantitative rigor, and provide limited insights when judging human tissue function in animal hosts. The molecular similarity of engineered populations is typically assessed by semiquantitative PCR, array-based expression profiling, or RNA sequencing followed by simple clustering analysis. However, such global analyses do not provide an intuitive or a quantitative means for diagnosing the deficiencies of engineered cells, nor do they provide a systematic approach to prioritize interventions to improve derivations of the desired populations.

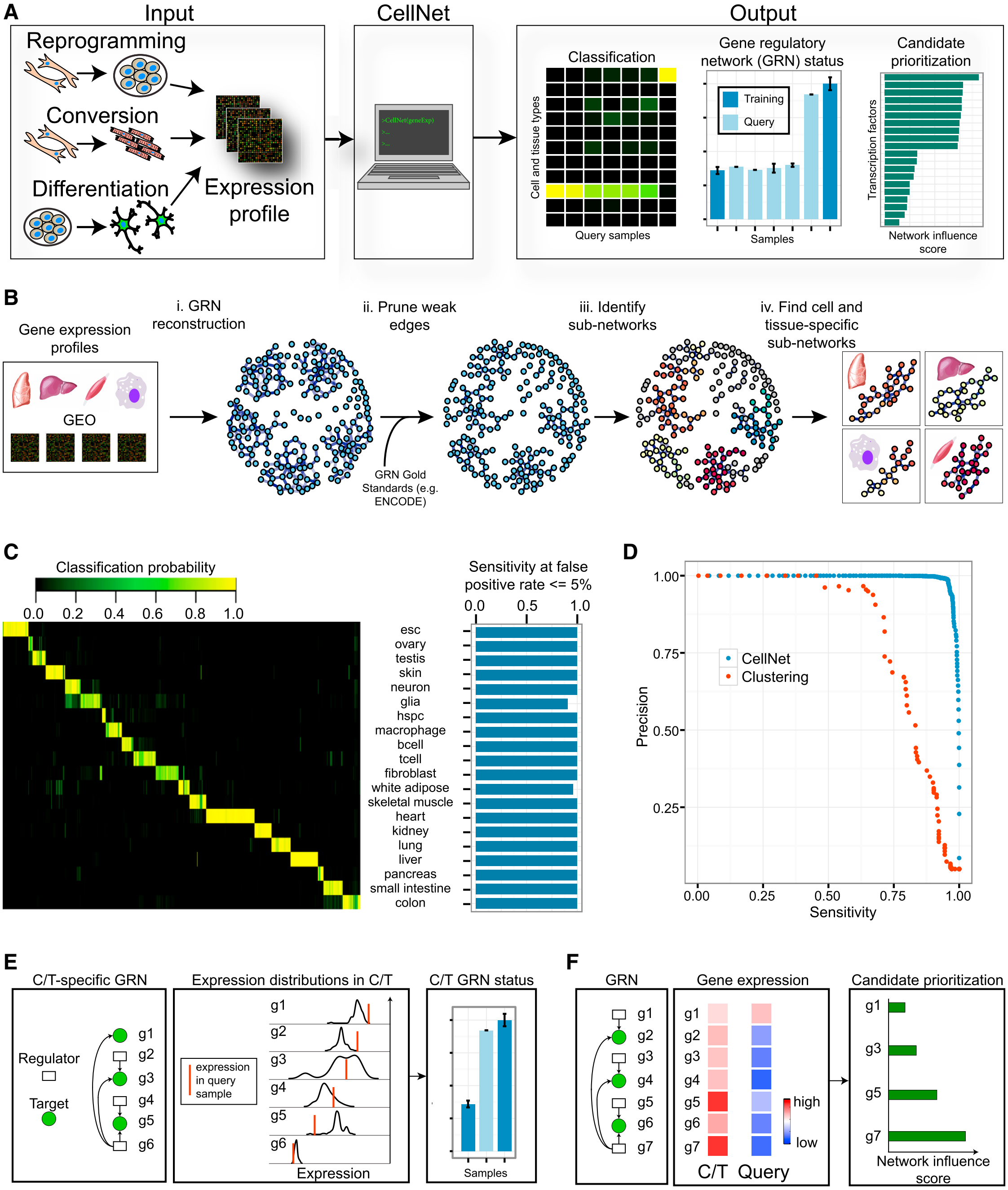

Figure 1. Construction and Validation of CellNet

(A) CellNet was designed to analyze gene expression profiles of mouse and human cell populations engineered by reprogramming to pluripotency, by direct conversion among somatic cell types, or by directed differentiation of pluripotent stem cells. CellNet can perform three types of analysis. First, CellNet calculates the probability that query samples express C/T-specific GRN genes to an extent that is indistinguishable from each cell and tissue type in the training data set. Second, CellNet measures the extent to which C/T GRNs are established in query samples relative to the corresponding C/T in the training data. Third, CellNet scores transcriptional regulators according to the likelihood that changing their expression will result in improved GRN establishment.

(B) CellNet is based on C/T-specific GRNs. C/T-specific GRNs were determined by first reconstructing a single GRN from a diverse panel of cell types and tissues and perturbations (i). Second, lower performing edges from the GRN were removed based on a comparison to a set of Gold Standard of regulatory relationships (ii). Finally, C/T-specific GRNs were identified by splitting the GRN into densely interconnected subnetworks (iii) followed by attribution to C/Ts based on gene set enrichment analysis (iv).

(C) Classification heatmap of an independent mouse validation data set. Binary classifiers were trained for each C/T using the C/T-specific GRN genes as predictors. Each row represents a C/T classifier, and each column represents a validation array. Higher classification scores indicate a higher probability that a

(legend continued on next page)

Here, we provide a network biology platform, CellNet, which assesses the fidelity of cell fate conversions and generates specific hypotheses aimed at improving derived cell populations. Our platform includes both novel and previously described components, which we outline below. We describe the construction of this platform for human and mouse cell and tissue types and use it to assess the results of 56 published attempts at reprogramming to pluripotency (most of which use the canonical reprogramming factors Oct4, Sox2, Klf4, and Myc), directed differentiation, and direct conversion of somatic cells. On the basis of these analyses, we have documented quantitatively that reprogramming is the most complete and successful of the various cell fate conversions; indeed, CellNet confirms that iPSC are virtually indistinguishable from ES cells in their faithful establishment of gene regulatory networks (GRNs). Further, we show that neurons and cardiomyocytes derived by directed differentiation of pluripotent stem cells more completely establish the target tissue- and cell-type GRNs than do neurons and cardiomyocytes directly converted from fibroblasts. Moreover, analysis of cardiomyocytes converted from cardiac fibroblasts in situ demonstrates that the in vivo environment provides selective and/or inductive signals that more completely establish heart GRNs. We also demonstrate that GRNs of the starting cell type are detectable in purified populations of both directed differentiation and in direct conversion experiments, and we show that the establishment of unintended GRNs is common to virtually every cellular engineering paradigm. Thus CellNet provides a platform for assessing and improving efforts at cellular engineering.

RESULTS

CellNet Construction and Validation

CellNet is predicated on the discovery of GRNs, which govern the steady-state expression program of a particular cell type as well as its transcriptional responses to environment, disease, and age. GRNs thus act as major molecular determinants of cell-type identity (Davidson and Erwin, 2006). We reasoned that measuring the establishment of cell and tissue (C/T)-specific GRNs in engineered populations would serve as both a robust metric of cellular identity and a tool to identify aberrant regulatory nodes. We designed CellNet to query gene expression profiles for the extent to which C/T GRNs are established, to classify input samples by their similarity to target cells and tissues, and to score transcriptional regulators according to their likelihood of improving the engineered population (Figure 1A).

In the ideal case, a comprehensive catalog of GRNs would be available from chromatin immunoprecipitation followed by sequencing (ChIP-seq) for all cell types and all transcription factors, and GRN establishment in engineered cells would be quantified by comparing their global transcription factor binding profiles to the reference set. Such data are currently unavailable, but over the past decade methods to reconstruct GRNs using genome-wide expression data have matured substantially (Marbach et al., 2012), and expression repositories have accumulated a wide array of biological perturbations (Lukk et al., 2010; Rung and Brazma, 2013), which are needed for accurate GRN reconstruction.

We developed a pipeline to reconstruct GRNs using 3,419 publicly available gene expression profiles of diverse cell types and tissues, including embryonic stem cells, ovary, testis, neurons, glia, skin, heart, skeletal muscle, fibroblasts, white adipose, kidney, endothelial cells, hematopoietic stem and progenitor cells (HSPC), B cells, T cells, macrophages, lung, liver, pancreas, small intestine, and colon (Figure 1B). We used the Affymetrix 430.2 and the HG133 plus 2 platforms to reconstruct mouse and human GRNs, respectively, because at present they represent the most comprehensive data set of cells and tissues and network perturbations. We performed a census of Gene Expression Omnibus (GEO) to determine the number of arrays that profiled well-annotated cell and tissue types. We found 20 mouse cell and tissue types that were represented by at least 100 expression profiles and 16 human cell and tissue types represented by at least 60 profiles (Table S1), as the performance of GRNs reconstructed from fewer than 60 profiles degrades substantially (Faith et al., 2007). To reconstruct GRNs, we first randomly selected an equivalent number of profiles from each GEO accession and each C/T, ensuring that each C/T was represented by a minimum of ten distinct conditions, which act as perturbations that enable GRN reconstruction. After integrating these samples into a single data set, we reconstructed GRNs using a modified version of the context likelihood of relatedness (CLR) algorithm, which uses a context-specific measure of correlation significance to infer direct regulatory relationships between transcriptional regulators and target genes (Faith et al., 2007). We first applied CLR to the complete data set to find regulatory relationships that span most tissues and cell types. Then, we used CLR to find regulatory relationships that are specific to cell and tissue types of common developmental origins, relationships that can be obscured when reconstructing GRNs from all C/Ts simultaneously (Novershtern et al., 2011). Then we combined the results into a single GRN per cell and tissue type.

query sample expresses the C/T GRN genes at a level indistinguishable from the same C/T in the training data. (Right) The sensitivity to accurately determine the source of the validation samples at a false positive rate $\leq 5\%$.

(D) Precision and sensitivity curves of CellNet (blue) and hierarchical clustering (red) based C/T classifiers on the validation data set. Precision is the number of true positives divided by the number of positive calls. Sensitivity is the number of true positives divided by the sum of the true positives and false negatives. Shown are the average of all mouse C/T classifiers computed across a range of classification scores (left to right).

(E) Using a gene expression profile to quantify GRN establishment or status. CellNet compares the expression of C/T GRN genes in the query sample to the distributions of gene expression in the C/T based on the training data and integrates this information with the importance of each gene to the classifier and the connectivity of each gene to arrive at GRN status, which is normalized to the GRN status of the C/T in the training data.

(F) Combining gene expression with GRNs to prioritize transcriptional regulators to improve cellular engineering. Given a query gene expression profile and a selected target C/T, CellNet computes a network influence score, which scores each transcriptional regulator based on its number of target genes, and extent of dysregulation of target genes and the regulator, weighted by the expression of the regulator in the C/T.

See Figure S1.

We then assessed the accuracy of regulatory relationships predicted by GRN reconstruction by comparison to three gold standards. The first was based on ENCODE-generated transcription factor binding data (Mouse ENCODE Consortium et al., 2012), the second on gene expression profiling of mouse embryonic stem cell lines modified for inducible expression of one of 94 transcription factors (Correa-Cerro et al., 2011; Nishikawa et al., 2007), and the third on ChIP-ChIP and ChIP-seq of 54 transcription factors in mESC (Xu et al., 2013; Zheng et al., 2010). The ability of our GRNs to predict true regulatory relationships, defined by ChIP-seq/ChIP-seq TF-binding sites, or by the genes that change in expression upon transcription factor gain or loss-of-expression, was as good or better than the state-of-the-art reported in a recent meta-analysis of algorithms for GRN reconstruction (Marbach et al., 2012) (Figure S1 available online). This result demonstrates that merging gene expression profiles from disparate experimental contexts yields high quality biological models, likely due to the wide diversity of perturbations to the underlying GRNs, regardless of the potential confounding batch effects common to microarray experiments. Then, we used the Gold Standards to select a threshold of strongly predicted regulatory relationships, resulting in 717,140 putative regulatory relationships among 20,722 target genes and 1,268 transcriptional regulators in mouse GRNs, and 536,897 putative regulatory relationships among 20,084 target genes and 1,566 transcriptional regulators in human GRNs. We have made the complete set of regulatory relationships that define these GRNs available for download at our companion website.

Next, we sought to define components of the GRN that are cell and tissue specific. Complex networks are comprised of communities or subnetworks in which some nodes (i.e., genes) are more densely interconnected to each other than to nodes in other subnetworks. Applying the InfoMap community-detection algorithm (Rosvall and Bergstrom, 2007), we found 1,787 mouse and 2,315 human subnetworks in our GRNs. By performing gene set enrichment analysis (Efron and Tibshirani, 2007), we found that 94 mouse and 76 human subnetworks were enriched in genes more highly expressed in a C/T (Figure 1B) and as expected, these subnetworks were highly enriched in Gene Ontology Biological Processes associated with specific cells or tissues (Table S2). For simplicity, we combined the subnetworks associated with each C/T into one general GRN per C/T and used these general GRNs except when stated below (Table S3).

For each C/T, we trained a binary classifier by using all of the genes in each C/T GRN in a randomly selected subset of the overall training data. To assess the performance of the classifiers, we applied the classifiers to an independent set of arrays and calculated the sensitivity and false positive rate. The classifiers performed well, reaching close to 100% sensitivity at false positive rates $\leq 5\%$ for all target C/Ts (Figure 1C and Figures S1E–S1H). Clustering is the most commonly used approach to determine relationships among samples based on expression profiling. To determine whether CellNet was better at classification than clustering, we also applied clustering to the validation data sets. At sensitivities greater than 63%, CellNet had substantially superior precision, defined as the proportion of classifications that are correct (Figure 1D). At sensitivities greater than 90%, CellNet achieved close to 100% precision. On the other hand, at these same sensitivities, clustering classifications were less than 25% precise, meaning that most of the calls made at this threshold by clustering were false positives. Our analysis implies that claims of cell engineering fidelity based on clustering alone can convey false impressions of similarity when indeed quite significant differences persist.

To help identify which subnetworks fail to be established or are aberrantly established in engineered cells, we devised a metric of GRN status given an expression profile (Figure 1E) that integrates the closeness of the expression of each gene in a GRN to its expected value, weighted by its importance to the network and its importance in the associated C/T classifier. Finally, to help prioritize transcriptional regulators for the improvement of cellular engineering, we devised a network influence score that integrates the expression level of the regulator in the target C/T, the extent of dysregulation of the regulator and its predicted targets in the query sample, and the number of predicted targets (Figure 1F).

We have made CellNet available as a web application (http://cellnet.hms.harvard.edu) where visitors can explore the GRNs, download our software to run locally, or upload expression data to be processed on our servers.

CellNet Limitations

We anticipated that the C/T type classifier might be limited in two aspects. First, tissues represent heterogeneous populations, whereas the goal of stem cell engineering is often the production of a highly pure preparation of a specific cell type. Second, the vast majority of our training data were derived from primary cells and tissues, whereas most attempts at engineering cell populations are performed in vitro. To address how tissue heterogeneity influences CellNet performance, we applied CellNet to primary neonatal cardiomyocytes purified on the basis of a reporter for alpha myosin heavy chain (aMHC) expression (Ieda et al., 2010), to nociceptor neurons dissected and purified based on Nav1.8 expression (Chiu et al., 2013), and to laser capture microdissected (LCM) human dopaminergic neurons (Zheng et al., 2010). Purified mouse cardiomyocytes classified exclusively as heart (Figure 2A) and reached greater than 90% heart GRN status (Figure 2B). Heart tissue includes fibroblasts, endocardium, neural, and vascular tissue in addition to cardiac muscle cells, but despite the heterogeneous nature of the heart training data, our classifier accurately determined the tissue of origin for purified cardiomyocytes. This analysis also explains why tissues like skeletal muscle, which shares functional features with heart (e.g., myosin-based contraction), and lung, which shares cell compositional features with the heart (e.g., substantial endothelial cell contributions), show higher heart GRN establishment levels than other cell types and tissues, because some members of the heart GRN are shared by skeletal muscle and endothelial cells (Figure 2B). When we applied CellNet to purified pain receptor neurons, we found that they classified exclusively as generic neurons (Figure 2C) (score = 0.792) and achieved a high neuron GRN status (mean = 91.6%) (Figure 2D). Finally, when we applied CellNet to LCM human dopaminergic neurons, we found that they classified highly as neurons (mean score = 0.887) (Figure 2E) and reached a high neuron GRN establishment (mean GRN status = 98.5%) (Figure 2F). Taken together, these results

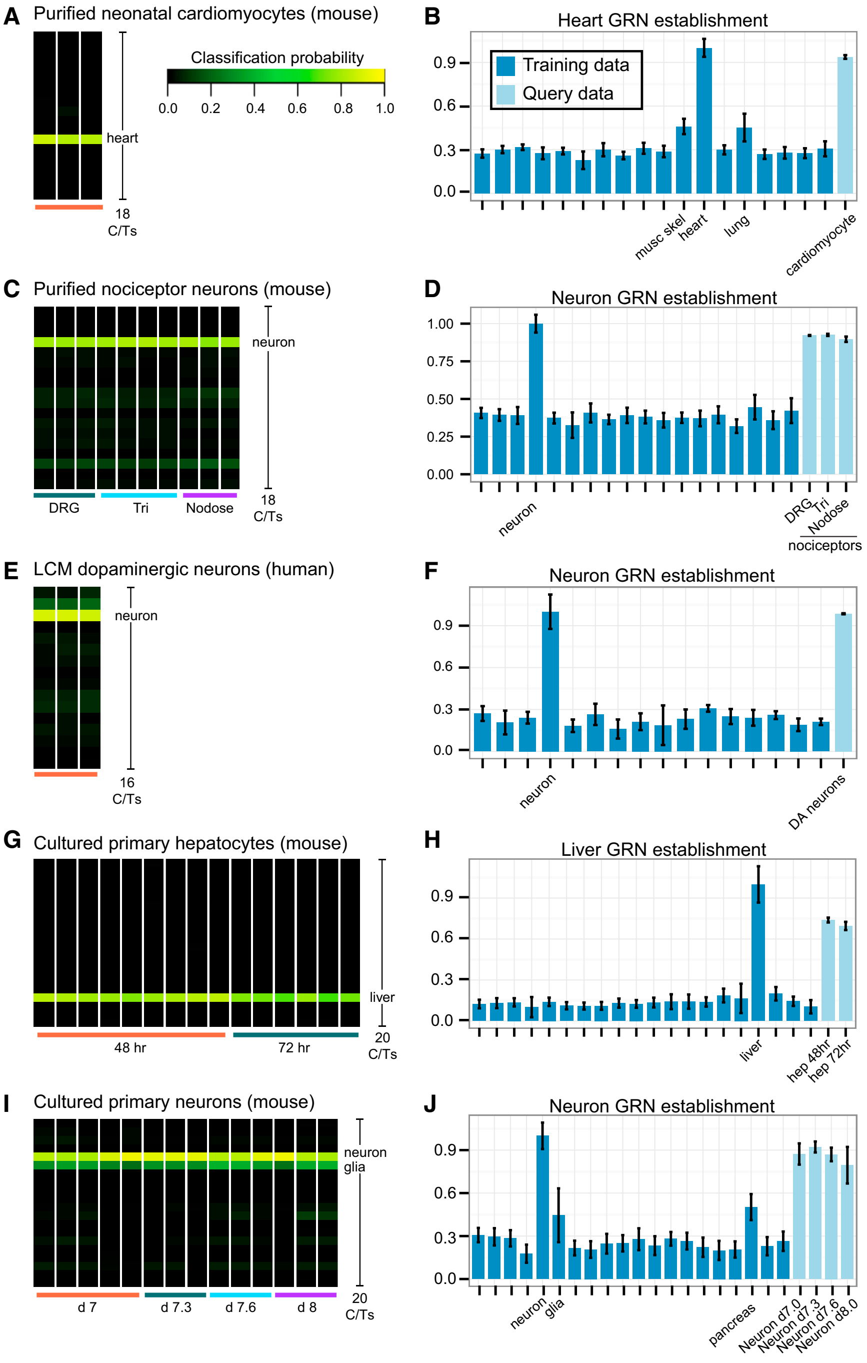

Figure 2. CellNet Analysis of Purified Primary and Cultured Cells

(A and B) Cell and tissue classification (A) and heart C/T GRN status (B) of primary neonatal cardiomyocytes in which RNA was harvested directly after aMHC-GFP positive cells were purified by FACS.

(C and D) Classification (C) and neuron GRN status (D) of dissected dorsal root (DRG), trigeminal (Tri), and nodose ganglia nociceptor neurons purified on the basis of Nav1.8 expression.

(E and F) Classification (E) and neuron GRN status (F) of laser capture micro-dissected human dopaminergic (DA) neurons.

(G and H) Classification (G) and liver GRN status (H) of primary hepatocytes cultured for 2-3 days.

(I) Cell and tissue characterization of cortical neurons cultured for 7-8 days. Cultured neurons are classified primarily as neurons and secondarily as glia, similar to the the validation neuron data in Figure 1B.

(J) Neuron C/T GRN establishment levels of cultured primary neurons. In all figures, dark blue bars represent the GRN status for the indicated C/T GRN in the training data, light blue represents the GRN status for the indicated C/T GRN in the query samples, and bars are standard deviations.

See Figure S2.

demonstrate that despite cell composition heterogeneity in the training data, CellNet is able to accurately classify specific cell types according to the tissue from which they are derived.

To address how the effects of cell culture influence the performance of CellNet, we analyzed the expression profiles of cortical neurons (Peng et al., 2012) and hepatocytes (Mao et al., 2011) that had been cultured in vitro prior to RNA extraction. We found that primary hepatocytes cultured for 3 days classified exclusively as liver (Figure 2G) and maintained a high liver GRN status (Figure 2H). Predictably, however, as a consequence of the known dedifferentiation of hepatocytes in culture (Strain, 1994), CellNet noted a modest decrease in the classification between 48 and 72 hr, which is reflected in decreased expression of mature hepatocyte markers (Figure S2). Similarly, neurons cultured for 8 days retained high neuron classification (mean = 0.870) and a high neuron GRN status (Figures 2I and 2J). The cultured neurons classified secondarily as glial cells (mean score = 0.325), consistent with a minor glia contamination in the cultured cells. The classifications did not substantially diminish for up to 8 days of culture, demonstrating that despite the dedifferentiation effects of cell culture, CellNet is sufficiently sensitive to detect the resident neural GRN. These results suggest that gene expression changes induced by cell culture, while modestly reducing the classification score, are not sufficient to abolish the classification of the primary cell type.

Engineered Neurons

To assess the relative fidelity of directed differentiation and direct conversion to neural cells types, we applied CellNet to engineered mouse neurons. To make the comparison between the two engineering paradigms as fair as possible, we focused our analysis on engineered cell populations that had the best target C/T classification among published examples, as these

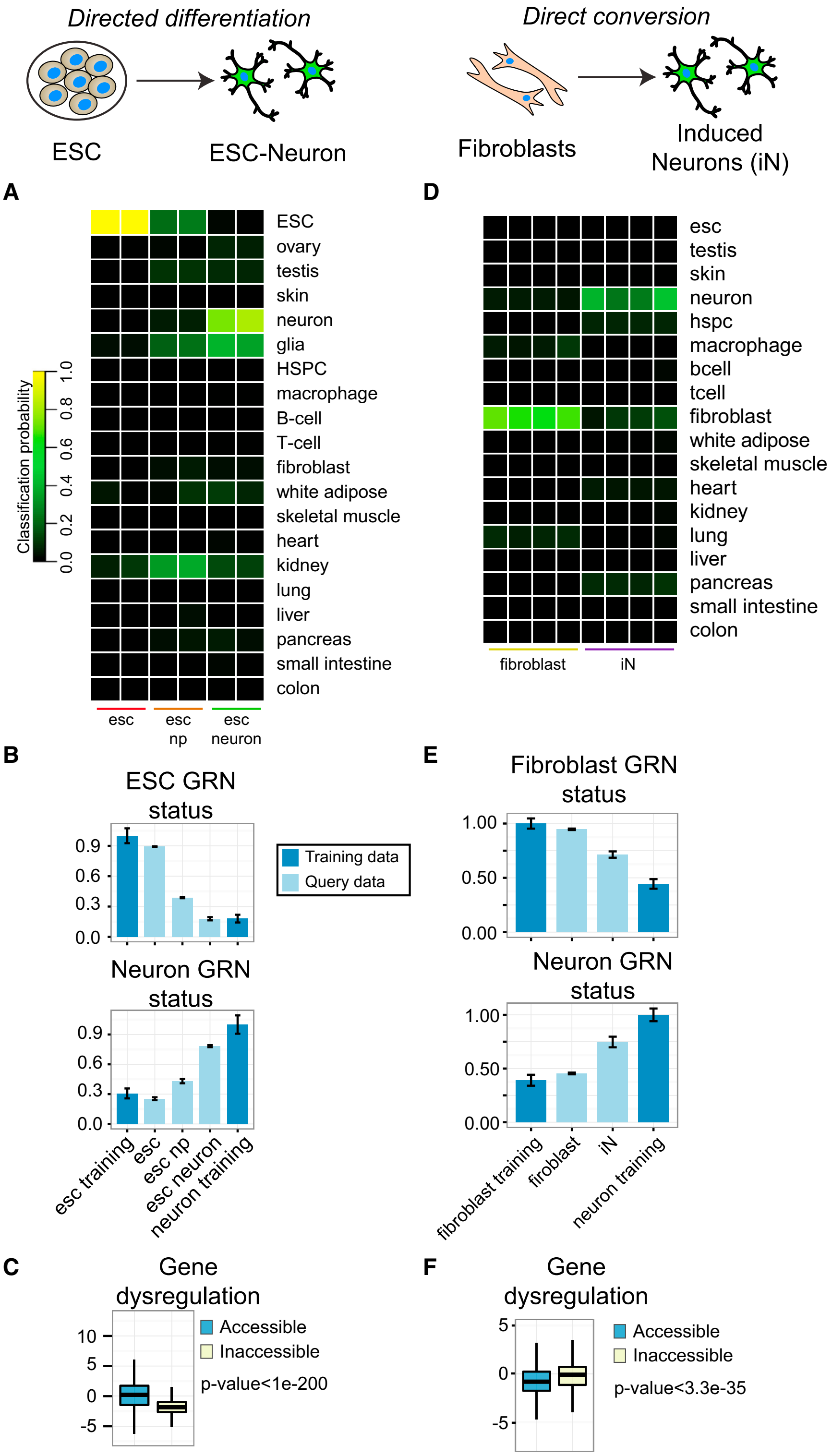

Figure 3. CellNet Analysis of Engineered Mouse Neurons

(A) C/T classification heatmap of the starting cell population (esc), intermediate neural progenitors (esc-np), and esc-derived neurons (esc-neuron).

(B) ESC (top) and neuron (bottom) GRN status in ESC, esc-np, and esc-neurons compared to GRN status of training ESC and neuron data. Error bars in all GRN status plots represent one standard deviation above and below the mean.

(C) Z scores of neural genes in esc-neurons. "Accessible" indicates genes with promoters that are DNase hypersensitive, whereas "Inaccessible" indicates genes with promoters that are not DNase hypersensitive.

(D) C/T classification heatmap of the starting cell population (fibroblasts) and induced neurons (iN).

(E) Fibroblast (top) and neuron (bottom) GRN status in fibroblasts and iNs compared to GRN status of training fibroblast and neuron data.

(F) Z scores of neural genes in iNs.

See Figure S3.

represent what is optimally possible to date (see Table S4 for list of all experiments analyzed and Table S5 for target C/T classifications). The ESC-derived cells (Arnold et al., 2013) achieved a high neuron classification (score = 0.773; Figure 3A), with high fidelity for both the silencing of the ESC GRN and the establishment of the neuron GRN (Figure 3B). Indeed, neurons directly differentiated from ESCs were nearly indistinguishable from native neuronal populations, further arguing that in vitro derived populations can be effectively classified against training data derived primarily from primary tissues.

The CellNet platform can be interrogated to test specific mechanistic hypotheses about what limits the robustness of cell fate conversions. For instance, we hypothesized that chromatin accessibility would dictate the fidelity of establishment of lineage-specific gene sets. We first mapped DNase hypersensitivity data from the Mouse ENCODE project to the promoters of genes that are more highly expressed in neural cells than in ESCs (i.e., "neural genes") (Mouse ENCODE Consortium et al., 2012). Then we asked whether the expression of neural genes with accessible promoters was more effectively established in ESC-derived neurons than neural genes with inaccessible promoters (Figure 3C). As predicted, we found that expression patterns for neural genes with accessible promoters in the ESC state were more readily established and closer to native neural populations than were genes lacking DNase hypersensitivity. This result indicates greater dysregulation for occluded promoters and implies that additional culture conditions that encourage the remodeling of chromatin must be explored during directed differentiation to fully establish the neural GRN.

We tested whether GRNs associated with lineages other than ESCs and neurons were detected in ESC-derived neurons. As expected, we detected modestly high glia classification scores in populations of ESC-derived neurons, which typically include glial contaminants, and consistent with the known glia constituents that contaminate the neuron training data (Figure 1C).

Next, we applied CellNet to the direct conversion of mouse fibroblasts to dopaminergic neurons via the ectopic expression of Ascl1, Nr4a2, and Lmx1a (Caiazzo et al., 2011). Induced neurons (iN) were classified as generic neurons (mean score = 0.340; Figure 3D), consistent with the functional characteristics of these cells (e.g., spontaneous action potentials resembling in vivo neurons and K+ induced monoamine release), but not as robustly as did ESC-neurons (mean score = 0.773). Despite the fact that iN were purified on the basis of TH-GFP positivity prior to RNA collection, they exhibited high fibroblast GRN status (Figure 3E) and a correspondingly high dysregulation of the fibroblast GRN regulators (Figure S3A), suggesting that the converted cells retain fibroblast identity. To explore possible mechanisms whereby neural genes remained silenced, we asked whether there was an association between promoter accessibility and dysregulation of neural genes. In contrast to the case of ESC-neurons, we found that neural genes with inaccessible promoters in fibroblasts tended to be less dysregulated than accessible genes, suggesting that nucleosome occlusion is not a major factor limiting establishment of endogenous GRNs in the directly converted population, consistent with Ascl1 acting as a pioneer factor that opens occluded chromatin (Wapinski et al., 2013). Taken together, this analysis suggests that directed conversion establishes a neuron GRN but fails to silence the fibroblast GRN, resulting in a hybrid cell type.

We tested whether GRNs and constituent subnetworks associated with lineages other than fibroblasts and neurons were established in iN (relative to fibroblasts or neurons). We found that the heart subnetwork 1 and pancreas subnetwork 1 were partially established (Figures S3B and S3C), and speculated that the aberrant GRNs arose because one or more of the conversion factors specifically targeted the aberrant subnetworks. To explore this possibility, we determined the enrichment of CellNet predicted targets of each conversion factor in each subnetwork, finding that indeed the pancreas subnetwork 1 was significantly enriched (p value = 8.56×10^{-6} by Chi-square test). Heart subnetwork 1 genes were not enriched as targets of any of the conversion factors, suggesting that they are indirect targets or that their regulatory relationships were not detected in the GRN reconstruction process.

Engineered Cardiomyocytes

Next, we applied CellNet to the best performing examples of directed differentiation and direct conversion to murine cardiomyocytes. First, we analyzed the directed differentiation of mouse ESCs to cardiomyocytes (Christoforou et al., 2008). In this experiment, ‘‘Cardiac progenitor cells’’ (ESC-CPC) carrying a GFP reporter for the expression of the cardiac-specific gene Nkx2.5 were FACS purified and analyzed after 5, 7, and 8 days of directed differentiation, while ESC-derived cardiomyocytes (ESC-CM) were isolated after 3 weeks by drug selection of ESCs carrying a neomycin resistance cassette under the control of an aMHC promoter. The ESC-CM classified as heart (mean score = 0.568) (Figure 4A), and to a lesser extent as fibroblast (mean score = 0.208). Furthermore, the ESC GRN was not completely silenced, nor was the heart GRN entirely established (Figure 4B); these cells scored lower than purified neonatal cardiomyocytes (Figure 2A), again indicating that the incomplete establishment of cardiomyocyte identity was not a consequence of inadequacy of the training data.

The activation of GRNs for alternate cell fates can reflect the infidelity of cell fate determination, population impurity, or intrinsic transcriptional features reflecting a shared developmental ontogeny. We found that the directed differentiation of ESCs to cardiomyocytes partially established HSPC subnetwork 1 in the ESC-CPC (Figure S4A). Given that the ESC-CPC were purified based on a Nkx2.5 reporter, and that a Nkx2.5+ population in the developing embryo includes a CD41+ population with definitive hemogenic potential (Nakano et al., 2013), the partial establishment of the HSPC GRN is consistent with a model where lateral plate mesoderm, the common developmental precursor to both the heart and definitive blood, emerges early in directed differentiation in vitro. We also found that the fibroblast GRN was well established in the ESC-CM (Figure S4B), consistent with the partial fibroblast classification (Figure 5A). Notably, the partial establishment of the fibroblast GRN was also detected in two other independent studies of ESC-CM (Figure S4C), which may result from either a latent developmental plasticity of in-vitro-derived ESC-CM or as an artifact of in vitro culture conditions.

We hypothesized that nucleosome occlusion contributes to the incomplete activation of heart-related genes during directed differentiation and found that heart genes lacking DNase hypersensitive sites in ESCs were more dysregulated in ESC-CMs than heart genes with accessible promoters (Figure 4C). The effect is not as pronounced as we found in ESC-neurons, but nonetheless supports the conclusion that the lack of accessibility of cardiac promoters prevents the faithful establishment of target GRNs in directly differentiated cells, and suggests that strategies that lead to more open, accessible chromatin during directed differentiation would enhance cellular engineering.

Next, we applied CellNet to the direct conversion of mouse cardiac fibroblasts to cardiomyocyte-like cells via the ectopic expression of Gata4, Mef2c, and Tbx5 (Ieda et al., 2010). Consistent with the demonstrated functional capacity of these samples (e.g., Ca^{2+} oscillations, spontaneous contraction, and intracellular action potential), the week 2 and week 4 cells were exclusively classified as heart (Figure 4D), albeit at a much lower overall classification score than ESC-derived CM (0.282 versus 0.568, respectively). The lower classification score for induced cardiomyocytes (iCMs) was not due to the failure to silence the residual fibroblast GRN (indeed, the fibroblast GRN was effectively decommissioned), but rather to an incomplete establishment of the heart GRN (Figure 4E). These data imply that the remaining barrier to achieving a more bona fide heart GRN status in iCM is the incomplete activation of heart related genes. Indeed, several transcription regulators had low network influence scores in iCM, suggesting that manipulating these factors might enhance direct conversion to cardiac fates (Figure S4D). These factors include heart factors that have been shown to

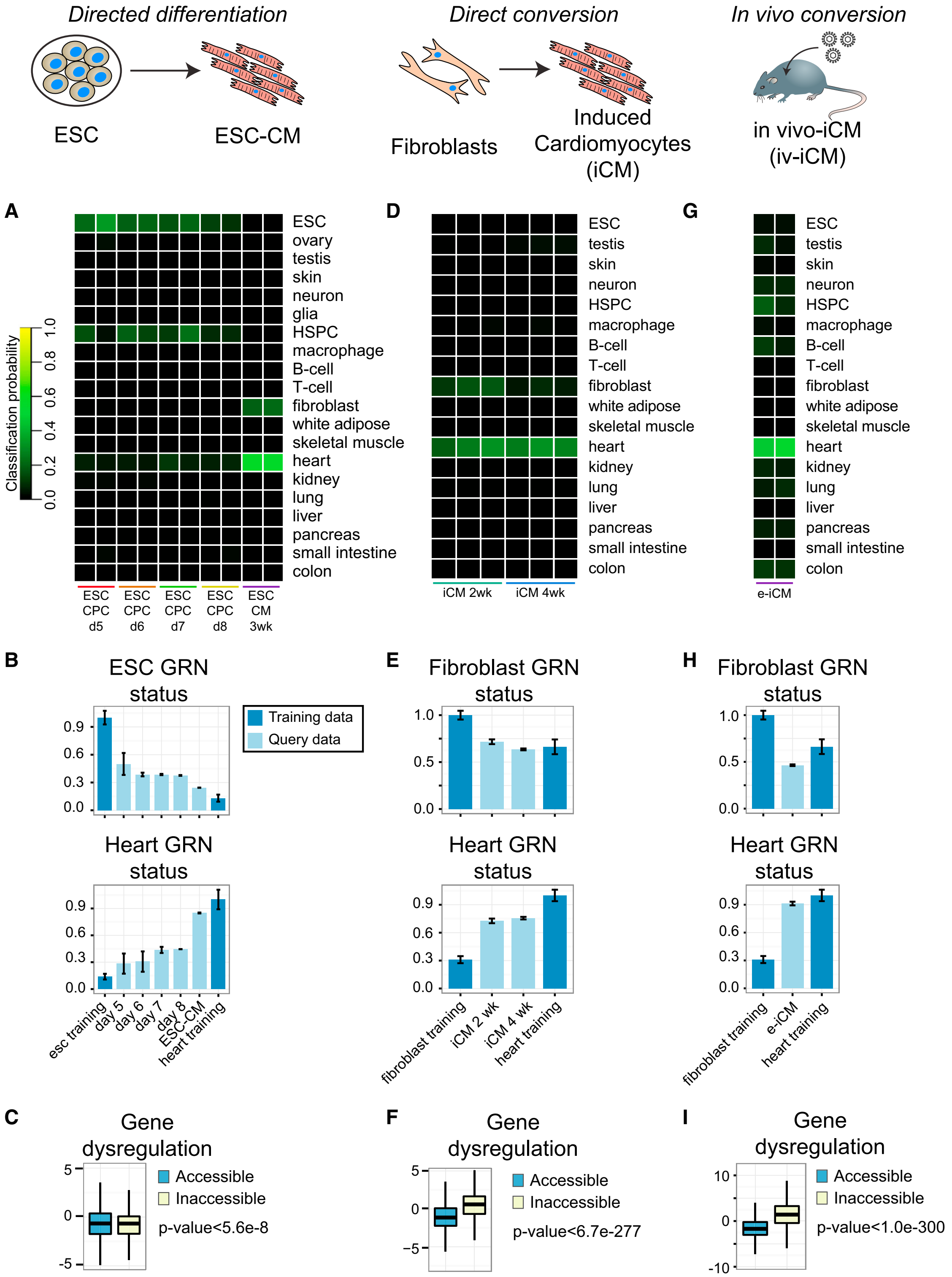

Figure 4. CellNet Analysis of Engineered Mouse Cardiomyocytes

(A) C/T classification heatmap of a time course of directed differentiation of ESC to cardiomyocytes. Cardiac progenitor cells (CPC), ESC-derived cardiomyocytes at 3 weeks of differentiation (ESC-CM).

(B) ESC (top) and heart (bottom) GRN status in CPC and ESC-CMs compared to GRN status of training ESC and heart data. Error bars in all GRN status plots represent one standard deviation above and below the mean.

(legend continued on next page)

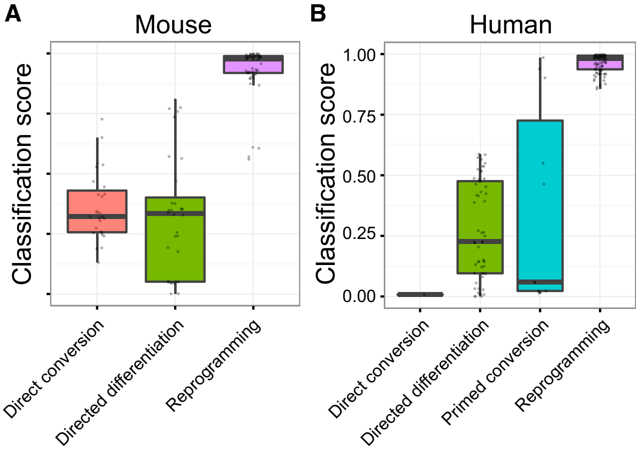

Figure 5. CellNet Analysis of 56 Distinct Cell Engineering Studies
CellNet classification score of target cell or tissue type in mouse (A) and human (B) cell engineering experiments. Only the most terminal (i.e., fully reprogrammed, or last time points of differentiation or conversion) samples profiled are included. See Figure S5.

enhance iCM induction (e.g., Myocd [Christoforou et al., 2013; Ieda et al., 2010], Nkx2-5 [Ieda et al., 2010], and Hand2 [Song et al., 2012]), and other heart-associated factors, which, to our knowledge, have not been tested in direct conversions (e.g., Gata6 and Tbx20). CellNet also identified novel factors predicted to enhance CM induction, including Phb, which is downregulated as a consequence of cardiac hypertrophy (Chowdhury et al., 2013), the sarcomere-associated transcription factor Ankrd1, and Lrpprc. Analyzing promoter accessibility, we found that heart genes (defined as genes expressed higher in heart than in fibroblasts) with accessible promoters in fibroblasts tended to be more dysregulated than genes with occluded promoters (Figure 4F), implying that the pioneering function of Gata4 to specifically open inaccessible chromatin is sufficient to activate expression of heart genes in directly converted populations (Ieda et al., 2010; Zaret and Carroll, 2011).

Next, we asked whether the inductive environment of the mouse heart improved the CellNet classification and heart GRN establishment of iCM that had been converted in situ from resident cardiac fibroblasts, subsequently purified, and profiled (iv-iCM) (Fu et al., 2013). Indeed, we found that the iv-iCM had substantially increased heart classification scores (mean = 0.498; Figure 4G) associated with a 90% heart GRN status and complete silencing of the fibroblast GRN (Figure 4H), consistent with the improved functionality of these cells (Qian et al., 2012). As in iCM converted in vitro, we found that heart genes with accessible promoters in fibroblasts remained more dysregulated than heart genes with inaccessible promoters, suggesting that expression of the Gata4 pioneer factor is particularly adept at engaging and activating critical heart genes (Figure 4I). Most importantly, CellNet analysis demonstrates that endogenous niches provide an enhanced milieu for direct conversion of fibroblasts to the target cell type.

CellNet Analysis of Engineered Populations

Finally, we applied CellNet to all reprogramming, directed differentiation, and direct conversion experiments published to date that had been profiled on platforms compatible with the CellNet training data (Table S4). This analysis included 22 studies of reprogramming to pluripotency, 15 of direct conversions (5 to neural lineages, 5 to cardiac lineages, 1 to endothelial cells, and 4 to hematopoietic lineages), and 16 of directed differentiation (5 to neural lineages, 4 to cardiac lineages, 2 to endothelial cells, 1 to fibroblasts, 2 to hepatocytes, and 2 to hematopoietic lineages). Relative to reprogramming, differentiation and conversion protocols show far less fidelity in generating populations that approach their target C/T, with a few exceptions (Figure 5, Table S5, and Figure S5). Notably, some “primed conversions,” so-called because they entail a transient ectopic expression of Pou5f1, achieved very high target C/T classifications (e.g., human induced endothelial cells and hepatocytes). Moreover, direction conversions between developmentally related cell types achieve higher target C/T classifications (e.g., pre-B cell to macrophage) than conversions among more distinct lineages (e.g., fibroblasts to hematopoietic progenitors).

DISCUSSION

A major goal of stem cell biology is the engineering of cells and tissues in vitro for applications in biomedical research, drug discovery, and therapeutic transplantation. However, metrics are lacking to define how closely engineered cell populations approach molecular and functional identity to target cell types and to identify and repair the critical gene expression differences between engineered cell types and their ideal target cell or tissue types. Here, we describe CellNet, a computational platform that determines the extent to which engineered cell populations have established GRNs that govern C/T identity. We have applied the platform to assess how well current cellular engineering techniques generate target cell populations. In a companion paper, we use CellNet to improve the conversion of B cells to macrophages and to reveal the unanticipated intestinal potential of induced hepatocyte-like cells (Morris et al., 2014). Several

(C) Z scores of heart genes in ESC-CMs. “Accessible” indicates genes with promoters that are DNase hypersensitive, whereas “inaccessible” indicates genes with promoters that are not DNase hypersensitive.
(D) Classification heatmap of the starting cell population (fibroblasts), and iCMs 2 and 4 weeks after ectopic expression of transgenes.
(E) Fibroblast (top) and heart (bottom) GRN status in fibroblasts and iCMs compared to GRN status of training fibroblast and heart data.
(F) Z scores of heart genes in iCMs.
(G) Classification heatmap of iCMs induced in vivo (iv-iCMs).
(H) Fibroblast (top) and heart (bottom) GRN status in fibroblasts and iv-iCMs compared to GRN status of training fibroblast and heart data.
(I) Z scores of heart genes in iv-iCMs.
See Figure S4.

insights from our analytical studies have implications for stem cell biology and cellular engineering.

First, CellNet indicates that reprogramming of somatic cells to pluripotency is the most faithful and complete cell-fate conversion among the many reported studies, providing a distinct and independent corroboration of other meta-analyses that confirms that iPSC and ESC are indistinguishable at the transcriptome level (Cahan and Daley, 2013; Newman and Cooper, 2010). Second, we found that neurons and cardiomyocytes derived by directed differentiation of pluripotent stem cells more completely establish the target tissue- and cell-type GRNs than do neurons and cardiomyocytes directly converted from fibroblasts. While ESC-derived cells appear superior to directly converted cells, even these cell populations lack full establishment of target C/T GRNs. The target C/T GRNs appear to be incompletely established because of promoter inaccessibility, as indicated by a lack of DNase hypersensitive sites, suggesting that developmental programs of in vitro differentiation fail to properly induce critical pioneer factors that activate essential gene expression programs. Surprisingly, according to CellNet analysis, the major barrier to direct conversion does not appear to be restricted access to the promoters or enhancers of target C/T genes, as these seem to be effectively activated by virtue of the powerful action of pioneer transcription factors included in the reprogramming cocktails. Instead, it appears that a failure to adequately silence donor cell GRNs and the expression of aberrant GRNs resulting in hybrid cell types represent the main molecular deficiencies of directly converted cell populations. These observations suggest that conversions between closely related cell types, which consequently minimize the “epigenetic distance” between donor and target cell, allows the most high fidelity fate conversions, as observed for lineage and stage conversions within the hematopoietic system (Di Tullio et al., 2011; Doulatov et al., 2013; Riddell et al., 2014).

Third, analysis of cardiomyocyte populations induced in situ from resident cardiac fibroblasts demonstrates that the native tissue environment provides additional selective and/or inductive signals to more completely establish the heart GRN. We anticipate that this finding is likely to hold for most engineered cell types.

Fourth, we have found that GRNs of the starting cell type are detectable in purified populations of both directed differentiation and in direct conversion experiments. In ESC-CM, the residually expressed members of the ESC-GRN are associated with ESC cell-cycle regulation and their impact on the tumor forming potential and developmentally immature phenotype of ESC-derived cells may be considerable. Directly induced neurons retain fibroblast GRNs, which may impede the normal range of neuron behavior. Our work has identified specific regulators of these residual GRNs that are candidates to be targeted in future studies in an attempt to extinguish features of the ESC cell cycle that persist in ESC-CMs, or to silence the residual fibroblast GRN in induced neurons.

Finally, we have discovered that the establishment of unintended GRNs is common to virtually every cellular engineering paradigm, either during the conversion or in the final products. For example, an unexpected GRN establishment occurs during the directed differentiation of ESC to cardiomyocytes, during which the HSPC GRN is partially established. This may reflect in vitro development of an aberrant cell population or may reflect development of a common mesodermal intermediate, which may in fact represent a signature of shared developmental ontogeny that is reflected in transient GRN expression detected in time-course experiments by CellNet. An example of unintended GRN establishment in the final cell product is the partial establishment of pancreas and heart subnetworks in induced neurons. It is possible that the iN transgenes target sufficient numbers of pancreas genes to modestly increase the expression of their cognate targets, and we speculate that these transcriptional regulators may prove useful in engineering pancreas cell types. By suppressing the transcriptional regulators of the pancreas GRNs, it may be possible to prevent their establishment and thereby improve the efficiency and fidelity of the iN conversions. Taken together, these results confirm that off-target transcriptional events of cellular engineering are frequent. More investigations are needed to determine how unintended targeting can either facilitate the conversion, lead to less efficient protocols or result in populations with hybrid GRNs.

We acknowledge that our current CellNet platform is limited in several ways, which suggest important avenues for future improvements. First, whereas mice and humans consist of hundreds of individual cell types, the limitations of the data available for training CellNet, which requires at least 60 gene arrays per cell or tissue type to achieve optimal performance of the GRN reconstruction, provides adequate resolution for only 20 target cell and tissue types in mouse and 16 in human. Moreover, the GRNs were reconstructed based largely on tissues rather than purified cell types. For example, the training data for the “neuron” C/T encompasses dissected brain and multiple types of neurons. Consequently, CellNet compares engineered populations to a generic “neuron,” and at present cannot distinguish between distinct neuronal subtypes. These limitations notwithstanding, CellNet performs admirably as a metric for assessing engineered cell similarities to a wide number of experimentally and medically important target cell and tissue types, and as demonstrated in the accompanying paper (Morris et al., 2014), can effectively suggest hypotheses for dysregulated transcription factors that can be further manipulated to enhance the identity and function of engineered populations. Our long-term goal is to employ single-cell RNA sequencing data to resolve the issue of tissue heterogeneity and to expand the platform to train on numerous distinct cell types.

Second, the training data are largely comprised of in vivo profiles, whereas most cellular engineering samples to date are derived in vitro. This limitation appears to be of minimal consequence because of the strong predictive power of C/T GRNs, which are not so confounded by signatures of cell culture as to compromise their utility. When we compared primary cultured cells and lines for which expression profiles were available (cardiomyocytes, neurons, and hepatocytes), we found that CellNet matched the cultured cells with their in vivo counterparts with high fidelity. Ideally, however, we will need to profile cultured primary cells for many more medically relevant cell types to be able to formally distinguish generic cell culture artifacts from defects of cellular engineering.

Finally, as sequencing costs fall, the increased use of RNA-seq is likely to soon replace microarrays as the transcriptomic platform of choice. To that end, we are working to extend CellNet so that it can be applied directly to RNA-seq data. CellNet is also amenable to incorporation of additional types of global epigenomic data, such as analysis of DNA methylation and histone marks, and accumulating experience with whole-genome location analysis of transcription factor binding, such that future versions of CellNet will have improved resolution and predictive value.

EXPERIMENTAL PROCEDURES

Training CellNet

To generate the complete training data sets, GEO was queried for expression profiles on the Affymetrix Mouse 430.2, Illumina8v2, HG133 plus 2, MoGene 1.0, and HuGene 1.0 platforms. Affymetrix microarrays not passing the GNUse quality control metric were not used to train CellNet. Probe intensities from raw .CEL files were background corrected and summarized into probeset values. Then, values of probesets mapping to the same gene were averaged. Each array was normalized by dividing each gene expression value by the total gene expression per array. To reconstruct cell type and tissue-specific GRNs, gene expression profiles from GEO representing 16–20 cell and tissue types with at least 10 perturbations per C/T were used as input to CLR (Faith et al., 2007). The resulting expression matrix was quantile normalized to create a data set for input to the CLR GRN reconstruction algorithm. Samples were then limited to those as defined in Extended Experimental Procedures, prior to re-executing CLR. An optimal regulatory relationship score was selected based on the Gold Standard assessment and used to remove lower performing edges from the GRN. C/T-specific GRNs were identified by first splitting up the large single GRN into subnetworks using InfoMap (Rosvall and Bergstrom, 2007). Subnetworks were attributed to specific C/Ts by determining which subnetworks were more highly represented in gene set enrichment analysis of each C/T versus all other C/Ts. Gene set enrichment analysis was performed on each subnetwork to identify C/T-enriched subnetworks, and subnetworks enriched in the same C/Ts were merged into single C/T GRNs. For each platform, a single Random Forest classifier was trained for each C/T (Breiman, 2001). Initially, approximately 50% of arrays were randomly selected and used to train each Random Forest classifier. The performance of these classifiers was evaluated by classifying the remaining, independent arrays, which were not used to train the classifier and were from different studies. The performance of these classifiers is presented in Figure 1B and Figures S1E–S1H. To improve the overall power of the classifiers, we then used all arrays to train the classifiers used in the final version of CellNet.

GRN Status

We reasoned that the extent to which a GRN is established in a sample is reflected in the expression of the genes in the GRN such that the GRN genes should fall within a range of expression observed in the corresponding C/T in the training data. We formalized this notion as a GRN status metric, defined in comparison to the complete training data set. The status of C/T GRN in a query sample is defined as the weighted mean of the Z scores of the genes in the GRN, where the Z score is defined in reference to the expression distribution of each gene in a C/T. The GRN status score can be weighted by the absolute expression level of each gene in a C/T so that genes more highly expressed have more influence on the GRN status (default) and/or by the importance to the Random Forest classifier (default):

$$RGS(x) = \sum_{i=1}^{n} \left(Zscore(gene)_{C/T} * gene\ weight_{GRN} \right)$$

where n = number of genes in the GRN and *gene weight* is proportional to the expression of the gene in C/T and, optionally, derived from the importance of the gene to the Random Forest classifier. Zscore(gene) is determined in reference to the expression distribution of the gene in C/T samples in the complete training data set. We transform and normalize the raw GRN status score so that higher GRN status indicates a GRN status more similar to the endogenous C/T as:

$$GS(x) = 1,000 - \frac{RGS(x)}{\left(\sum_j^k RGS(y)\right) \Big/ k}$$

where $\left(\sum_j^k RGS(y)\right)/k$ is the average RGS of C/T samples in the complete training data.

Network Influence Score

To estimate the importance of transcriptional regulators to dysregulated GRNs, we devised the following network influence score (NIS):

$$NIS(\mathrm{TR}) = \sum_{i=1}^{n} \left(Zscore(target)_{C/T} * weight_{target} \right) + n * Zscore(TR)_{C/T} * weight_{TR}$$

where TR = transcriptional regulator and n = number of genes in the GRN. By default, CellNet weights the target and TR by their mean expression in the C/T samples of the complete training data set when searching for inappropriately silenced factors.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.07.020.

AUTHOR CONTRIBUTIONS

P.C., S.A.M., and H.L. designed, performed, and interpreted analyses and wrote the paper. E.L.d.R. performed analysis and helped to write the paper. J.J.C. and G.Q.D. oversaw the project and helped to design the study and write the paper.

ACKNOWLEDGMENTS

J.J.C. is supported by NIH grant R24DK092760 and the HHMI. G.Q.D. is supported by grants from the NIH (Progenitor Cell Biology Consortium UO1-HL100001, R24DK092760, and P50HG005550), the Ellison Medical Foundation, Doris Duke Medical Foundation, and Boston Children's Hospital Stem Cell Program, and is an affiliate member of the Broad Institute. P.C. is supported by NIDDK (K01DK096013) and received support from NHLBI (T32HL066987 and T32HL007623). S.A.M. is supported by a Young Investigator Award from Alex's Lemonade Stand Foundation. H.L. is supported by Mayo Clinic Center for Individualized Medicine and Mayo Clinic Center for Regenerative Medicine. E.L.d.R. is supported by National Council for Scientific and Technological Development and the program Science Without Borders (CNPq, Brazil).

Received: January 14, 2014
Revised: May 28, 2014
Accepted: July 17, 2014
Published: August 14, 2014

REFERENCES

Aasen, T., Raya, A., Barrero, M.J., Garreta, E., Consiglio, A., Gonzalez, F., Vassena, R., Bilić, J., Pekarik, V., Tiscornia, G., et al. (2008). Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat. Biotechnol. *26*, 1276–1284.

Arnold, P., Schöler, A., Pachkov, M., Balwierz, P.J., Jørgensen, H., Stadler, M.B., van Nimwegen, E., and Schübeler, D. (2013). Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. Genome Res. *23*, 60–73.

Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.

Cahan, P., and Daley, G.Q. (2013). Origins and implications of pluripotent stem cell variability and heterogeneity. Nat. Rev. Mol. Cell Biol. *14*, 357–368.

Caiazzo, M., Dell'Anno, M.T., Dvoretskova, E., Lazarevic, D., Taverna, S., Leo, D., Sotnikova, T.D., Menegon, A., Roncaglia, P., Colciago, G., et al. (2011). Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. Nature *476*, 224–227.

Chiu, I.M., Heesters, B.A., Ghasemlou, N., Von Hehn, C.A., Zhao, F., Tran, J., Wainger, B., Strominger, A., Muralidharan, S., Horswill, A.R., et al. (2013). Bacteria activate sensory neurons that modulate pain and inflammation. Nature *501*, 52–57.

Chowdhury, D., Tangutur, A.D., Khatua, T.N., Saxena, P., Banerjee, S.K., and Bhadra, M.P. (2013). A proteomic view of isoproterenol induced cardiac hypertrophy: prohibitin identified as a potential biomarker in rats. J. Transl. Med. *11*, 130.

Christoforou, N., Miller, R.A., Hill, C.M., Jie, C.C., McCallion, A.S., and Gearhart, J.D. (2008). Mouse ES cell-derived cardiac precursor cells are multipotent and facilitate identification of novel cardiac genes. J. Clin. Invest. *118*, 894–903.

Christoforou, N., Chellappan, M., Adler, A.F., Kirkton, R.D., Wu, T., Addis, R.C., Bursac, N., and Leong, K.W. (2013). Transcription factors MYOCD, SRF, Mesp1 and SMARCD3 enhance the cardio-inducing effect of GATA4, TBX5, and MEF2C during direct cellular reprogramming. PLoS ONE *8*, e63577.

Correa-Cerro, L.S., Piao, Y., Sharov, A.A., Nishiyama, A., Cadet, J.S., Yu, H., Sharova, L.V., Xin, L., Hoang, H.G., Thomas, M., et al. (2011). Generation of mouse ES cell lines engineered for the forced induction of transcription factors. Sci Rep *1*, 167.

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796–800.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987–1000.

Di Tullio, A.A., Vu Manh, T.P., Schubert, A.A., Castellano, G.G., Månsson, R.R., and Graf, T.T. (2011). CCAAT/enhancer binding protein alpha (C/EBP(alpha))-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. Proc. Natl. Acad. Sci. USA *108*, 17016–17021.

Doulatov, S., Vo, L.T., Chou, S.S., Kim, P.G., Arora, N., Li, H., Hadland, B.K., Bernstein, I.D., Collins, J.J., Zon, L.I., and Daley, G.Q. (2013). Induction of multipotential hematopoietic progenitors from human pluripotent stem cells via respecification of lineage-restricted precursors. Cell Stem Cell *13*, 459–470.

Efron, B., and Tibshirani, R. (2007). On testing the significance of sets of genes. Ann. Appl. Stat. *1*, 107–129.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. *5*, e8.

Fu, J.-D., Stone, N.R., Liu, L., Spencer, C.I., Qian, L., Hayashi, Y., Delgado-Olguin, P., Ding, S., Bruneau, B.G., and Srivastava, D. (2013). Direct Reprogramming of Human Fibroblasts toward a Cardiomyocyte-like State. Stem Cell Rep. *1*, 235–247.

Huang, P., He, Z., Ji, S., Sun, H., Xiang, D., Liu, C., Hu, Y., Wang, X., and Hui, L. (2011). Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. Nature *475*, 386–389.

Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G., and Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell *142*, 375–386.

Kim, J.B., Sebastiano, V., Wu, G., Araúzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. Cell *136*, 411–419.

Loh, Y.-H., Hartung, O., Li, H., Guo, C., Sahalie, J.M., Manos, P.D., Urbach, A., Heffner, G.C., Grskovic, M., Vigneault, F., et al. (2010). Reprogramming of T cells from human peripheral blood. Cell Stem Cell *7*, 15–19.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010). A global map of human gene expression. Nat. Biotechnol. *28*, 322–324.

Mao, T., Shao, M., Qiu, Y., Huang, J., Zhang, Y., Song, B., Wang, Q., Jiang, L., Liu, Y., Han, J.-D.J., et al. (2011). PKA phosphorylation couples hepatic inositol-requiring enzyme 1alpha to glucagon signaling in glucose metabolism. Proc. Natl. Acad. Sci. USA *108*, 15852–15857.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G.; DREAM5 Consortium (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., George, Q., and Daley, G.Q. (2014). Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet. Cell *158*, this issue, 889–902.

Mouse ENCODE Consortium, Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Murry, C.E., and Keller, G. (2008). Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell *132*, 661–680.

Nakano, H., Liu, X., Arshi, A., Nakashima, Y., van Handel, B., Sasidharan, R., Harmon, A.W., Shin, J.-H., Schwartz, R.J., Conway, S.J., et al. (2013). Haemogenic endocardium contributes to transient definitive haematopoiesis. Nat Commun *4*, 1564.

Newman, A.M., and Cooper, J.B. (2010). Lab-specific gene expression signatures in pluripotent stem cells. Cell Stem Cell *7*, 258–262.

Nishikawa, S.-I., Jakt, L.M., and Era, T. (2007). Embryonic stem-cell culture as a tool for developmental cell biology. Nat. Rev. Mol. Cell Biol. *8*, 502–507.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

Peng, Z.F., Chen, M.J., Manikandan, J., Melendez, A.J., Shui, G., Russo-Marie, F., Whiteman, M., Beart, P.M., Moore, P.K., and Cheung, N.S. (2012). Multifaceted role of nitric oxide in an in vitro mouse neuronal injury model: transcriptomic profiling defines the temporal recruitment of death signalling cascades. J. Cell. Mol. Med. *16*, 41–58.

Qian, L., Huang, Y., Spencer, C.I., Foley, A., Vedantham, V., Liu, L., Conway, S.J., Fu, J.-D., and Srivastava, D. (2012). In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. Nature *485*, 593–598.

Riddell, J., Gazit, R., Garrison, B.S., Guo, G., Saadatpour, A., Mandal, P.K., Ebina, W., Volchkov, P., Yuan, G.-C., Orkin, S.H., et al. (2014). Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. Cell *157*, 549–564.

Rosvall, M., and Bergstrom, C.T. (2007). An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. USA *104*, 7327–7331.

Rung, J., and Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nat. Rev. Genet. *14*, 89–99.

Sekiya, S., and Suzuki, A. (2011). Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature *475*, 390–393.

Song, K., Nam, Y.-J., Luo, X., Qi, X., Tan, W., Huang, G.N., Acharya, A., Smith, C.L., Tallquist, M.D., Neilson, E.G., et al. (2012). Heart repair by reprogramming non-myocytes with cardiac transcription factors. Nature *485*, 599–604.

Staerk, J., Dawlaty, M.M., Gao, Q., Maetzel, D., Hanna, J., Sommer, C.A., Mostoslavsky, G., and Jaenisch, R. (2010). Reprogramming of human peripheral blood cells to induced pluripotent stem cells. Cell Stem Cell *7*, 20–24.

Strain, A.J. (1994). Isolated hepatocytes: use in experimental and clinical hepatology. Gut *35*, 433–436.

Szabo, E., Rampalli, S., Risueño, R.M., Schnerch, A., Mitchell, R., Fiebig-Comyn, A., Levadoux-Martin, M., and Bhatia, M. (2010). Direct conversion of human fibroblasts to multilineage blood progenitors. Nature *468*, 521–526.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035–1041.

Wapinski, O.L., Vierbuchen, T., Qu, K., Lee, Q.Y., Chanda, S., Fuentes, D.R., Giresi, P.G., Ng, Y.H., Marro, S., Neff, N.F., et al. (2013). Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell *155*, 621–635.

Xie, H., Ye, M., Feng, R., and Graf, T. (2004). Stepwise reprogramming of B cells into macrophages. Cell *117*, 663–676.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford) *2013*, bat045.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227–2241.

Zheng, B., Liao, Z., Locascio, J.J., Lesniak, K.A., Roderick, S.S., Watt, M.L., Eklund, A.C., Zhang-James, Y., Kim, P.D., Hauser, M.A., et al.; Global PD Gene Expression (GPEX) Consortium (2010). PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. Sci. Transl. Med. *2*, 52ra73.

Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J., and Melton, D.A. (2008). In vivo reprogramming of adult pancreatic exocrine cells to β-cells. Nature *455*, 627–632.