nature genetics

A predictive computational framework for direct reprogramming between human cell types

Owen J L Rackham[1,2,12], Jaber Firas[3–5,12], Hai Fang[1], Matt E Oates[1], Melissa L Holmes[3–5], Anja S Knaupp[3,5], The FANTOM Consortium[6], Harukazu Suzuki[7,8], Christian M Nefzger[3–5], Carsten O Daub[7–9], Jay W Shin[7,8], Enrico Petretto[2], Alistair R R Forrest[7,8,10], Yoshihide Hayashizaki[8,11], Jose M Polo[3–5] & Julian Gough[1]

Transdifferentiation, the process of converting from one cell type to another without going through a pluripotent state, has great promise for regenerative medicine. The identification of key transcription factors for reprogramming is currently limited by the cost of exhaustive experimental testing of plausible sets of factors, an approach that is inefficient and unscalable. Here we present a predictive system (Mogrify) that combines gene expression data with regulatory network information to predict the reprogramming factors necessary to induce cell conversion. We have applied Mogrify to 173 human cell types and 134 tissues, defining an atlas of cellular reprogramming. Mogrify correctly predicts the transcription factors used in known transdifferentiations. Furthermore, we validated two new transdifferentiations predicted by Mogrify. We provide a practical and efficient mechanism for systematically implementing novel cell conversions, facilitating the generalization of reprogramming of human cells. Predictions are made available to help rapidly further the field of cell conversion.

We now know that it is possible to switch the phenotype of one somatic cell type to another. This epigenetic rewiring process can be artificially managed and even reversed with the use of transcription factors[1]. The best known example is the reprogramming of somatic cells into induced pluripotent stem (iPS) cells by the introduction of four exogenous factors (Oct3/Oct4, Sox2, c-Myc and Klf4)[2,3]. Previous and subsequent reports have demonstrated that other cell types can also be obtained by direct transdifferentiation using the same strategy[4–9]. These discoveries came about through a process of exhaustive testing of large sets of plausible transcription factors. With roughly 2,000 different transcription factors[10–12] and approximately 400 unique cell types in humans[13], the space of possible sets is very large (>10^{11} combinations of three factors across 400 cell types), and discovery will advance slowly using an educated trial-and-error approach. There are a number of existing algorithms that identify transcription factors that might assist in cell-to-cell conversions, considering both epigenetic factors and transcription factor activation[14–17]. More recently, approaches such as CellNet[18,19] and a new entropy-based method[20] have provided larger-scale predictions of transcription factors for conversion into many cell types, as well as showing experimentally that these predictions are able to control cell identity. Herein we present a comprehensive atlas of predictions for a large number of human cell conversions, which we have implemented using a network-based computational framework (Mogrify) applied to the FANTOM5 data sets[21], which include ~300 different cell and tissue types. We have shown that we are able to independently recover (via prediction) human conversion factors that were previously discovered experimentally, and, more notably, we have predicted and validated two new conversions.

To predict the sets of transcription factors required for each cell conversion, we identify the transcription factors that are not only differentially expressed between cell types but also exert a regulatory influence on other differentially expressed genes in the local network (**Fig. 1a**). A single score that captures the differential expression for every gene in every cell type is defined by combining the log-transformed fold change in expression and adjusted *P* value relative to a background (**Fig. 1b**). The regulatory influence of each transcription factor in each cell type is calculated by performing a weighted sum of the differential expression scores over the known interactome (as defined by the STRING database[22] and Motif Activity Response Analysis (MARA)[23]; **Fig. 1c**). This sum is weighted by two factors: (i) the directness of the regulation, that is, the number of intermediates between the transcription factor and a downstream gene, and (ii) the specificity, that is, the number of other genes also regulated by the upstream transcription factor.

[1]Department of Computer Science, University of Bristol, Bristol, UK. [2]Program in Cardiovascular and Metabolic Disorders, Duke–National University of Singapore Medical School, Singapore. [3]Department of Anatomy and Developmental Biology, Monash University, Clayton, Victoria, Australia. [4]Australian Regenerative Medicine Institute, Monash University, Clayton, Victoria, Australia. [5]Development and Stem Cells Program, Monash Biomedicine Discovery Institute, Monash University, Clayton, Victoria, Australia. [6]A list of members and affiliations appears in the **Supplementary Note**. [7]RIKEN Omics Science Center, Yokohama, Japan (ceased to exist as of 1 April 2013 owing to reorganization). [8]Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan. [9]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. [10]Harry Perkins Institute of Medical Research, Queen Elizabeth II Medical Centre and Centre for Medical Research, University of Western Australia, Nedlands, Western Australia, Australia. [11]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan. [12]These authors contributed equally to this work. Correspondence should be addressed to O.J.L.R. (owen.rackham@duke-nus.edu.sg), J.M.P. (jose.polo@monash.edu) or J.G. (julian.gough@bristol.ac.uk).

Received 15 September 2015; accepted 16 December 2015; published online 18 January 2016; doi:10.1038/ng.3487

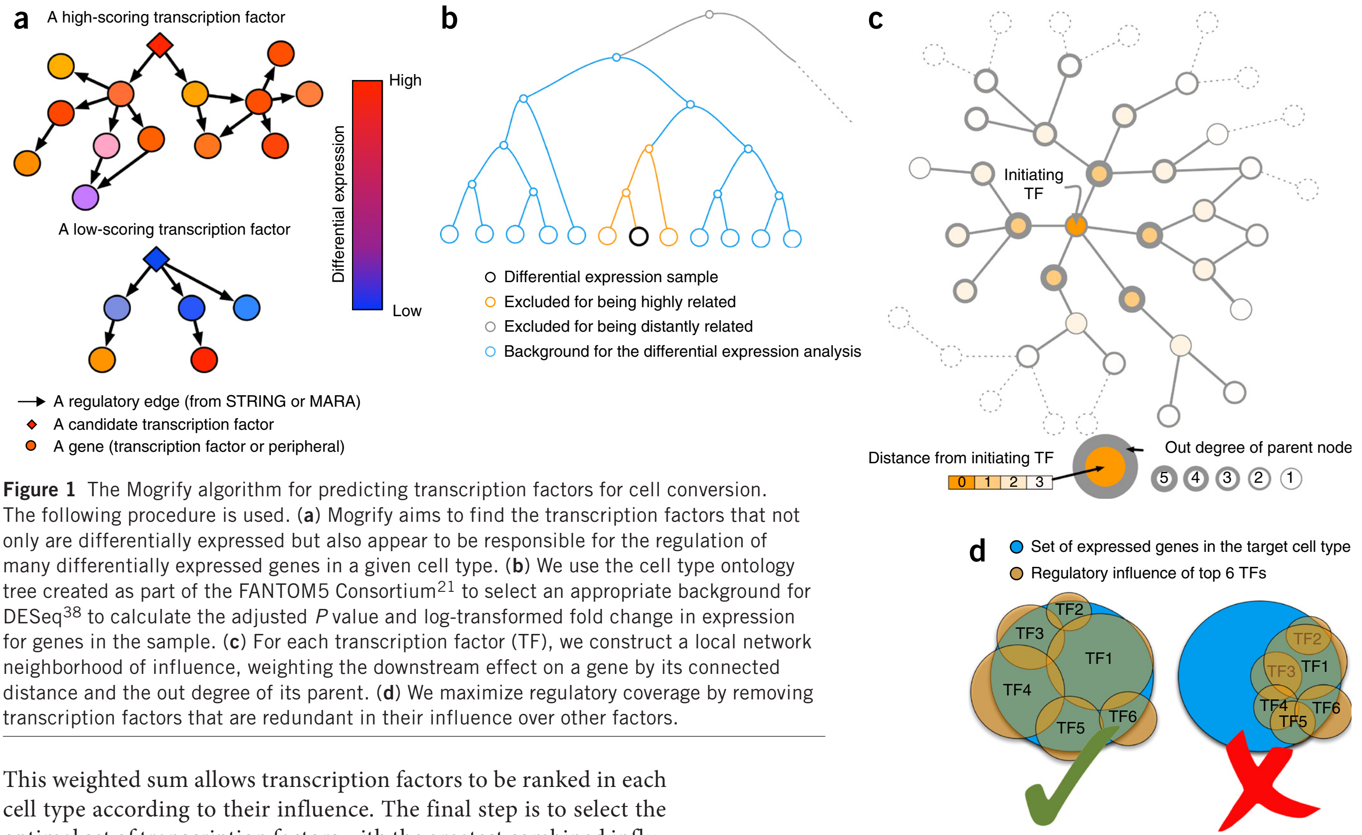
a
A high-scoring transcription factor
A low-scoring transcription factor
Differential expression
High
Low
A regulatory edge (from STRING or MARA)
A candidate transcription factor
A gene (transcription factor or peripheral)
b
Differential expression sample
Excluded for being highly related
Excluded for being distantly related
Background for the differential expression analysis
c
Initiating TF
Distance from initiating TF
0 1 2 3
Out degree of parent node
5 4 3 2 1
d
Set of expressed genes in the target cell type
Regulatory influence of top 6 TFs
TF1
TF2
TF3
TF4
TF5
TF6

Figure 1 The Mogrify algorithm for predicting transcription factors for cell conversion. The following procedure is used. (**a**) Mogrify aims to find the transcription factors that not only are differentially expressed but also appear to be responsible for the regulation of many differentially expressed genes in a given cell type. (**b**) We use the cell type ontology tree created as part of the FANTOM5 Consortium[21] to select an appropriate background for DESeq[38] to calculate the adjusted *P* value and log-transformed fold change in expression for genes in the sample. (**c**) For each transcription factor (TF), we construct a local network neighborhood of influence, weighting the downstream effect on a gene by its connected distance and the out degree of its parent. (**d**) We maximize regulatory coverage by removing transcription factors that are redundant in their influence over other factors.

This weighted sum allows transcription factors to be ranked in each cell type according to their influence. The final step is to select the optimal set of transcription factors with the greatest combined influence over genes differentially expressed in the target cell type in comparison to the donor cell type. This is done by adding transcription factors to the set in order of rank by differential influence, omitting those that do not increase the influence of the set, until the combined influence reaches 98% of expressed target cell genes (**Fig. 1d** and Online Methods). Biologically speaking, Mogrify identifies transcription factors that control the parts of the regulatory network most responsible for the identity of the target cell type.

To assess the predictive power of Mogrify, we first determined how Mogrify performed against well-known, previously published human cell conversions. These predictions should not be considered as perfect combinations but as positive reference points useful for comparison. In almost every case, Mogrify predicted the complete set of transcription factors previously demonstrated to work but sometimes used an upstream transcription factor in lieu of the published factor (**Fig. 2**).

For example, it is known that human fibroblasts can be converted into iPS cells by introducing *OCT4* (also known as *POU5F1*), *SOX2*, *KLF4* and *MYC*[3] or *OCT4*, *SOX2*, *NANOG* and *LIN28* (ref. 24). Mogrify predicted *NANOG*, *OCT4* and *SOX2* as the top three transcription factors for this conversion, a combination that has also been experimentally validated[25]. Seminal work by the Graf laboratory demonstrated that the conversion of B cells and fibroblasts into macrophage-like cells is possible through expression of CEBPα and PU.1 (also known as SPI1)[26,27], which Mogrify perfectly predicted. For the conversion of human dermal fibroblasts into cardiomyocytes, the closest reliable sample was cellularly heterogeneous heart tissue; nevertheless, Mogrify's predicted list included four of the five transcription factors (or similar factors) used in the human conversion of these cells[28]. There are reports in the literature of transdifferentiations of various cell types into neurons in both mouse and human (**Supplementary Table 1**). The sets of transcription factors used vary, probably owing to the heterogeneity and complexity of neurons; however, factors common to all experiments[29] were predicted by Mogrify (**Supplementary Table 2**). Finally, between human fibroblasts and hepatocytes, Mogrify predicted a combination of transcription factors highly similar to that required for conversion and maturation (**Fig. 2**)[6,30,31]. Using the

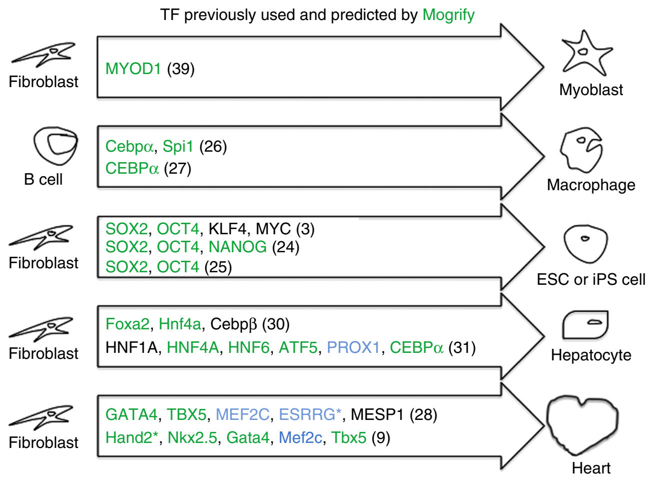

Figure 2 Mogrify predictions for some of the known transdifferentiations that are published in the literature. Transcription factors that Mogrify correctly identifies from the published lists are highlighted. Samples are grouped using the FANTOM cell ontology[21]. For each publication, the transcription factors in the initial maximum-coverage set are shown in green and the transcription factors in the overall predicted Mogrify set are shown in blue. For instance, the transdifferentiation of a fibroblast into a myoblast[39] required only MYOD1, which was identified by Mogrify. An asterisk indicates that a closely related transcription factor is predicted. Numbers in parentheses denote corresponding references. ESC, embryonic stem cell.

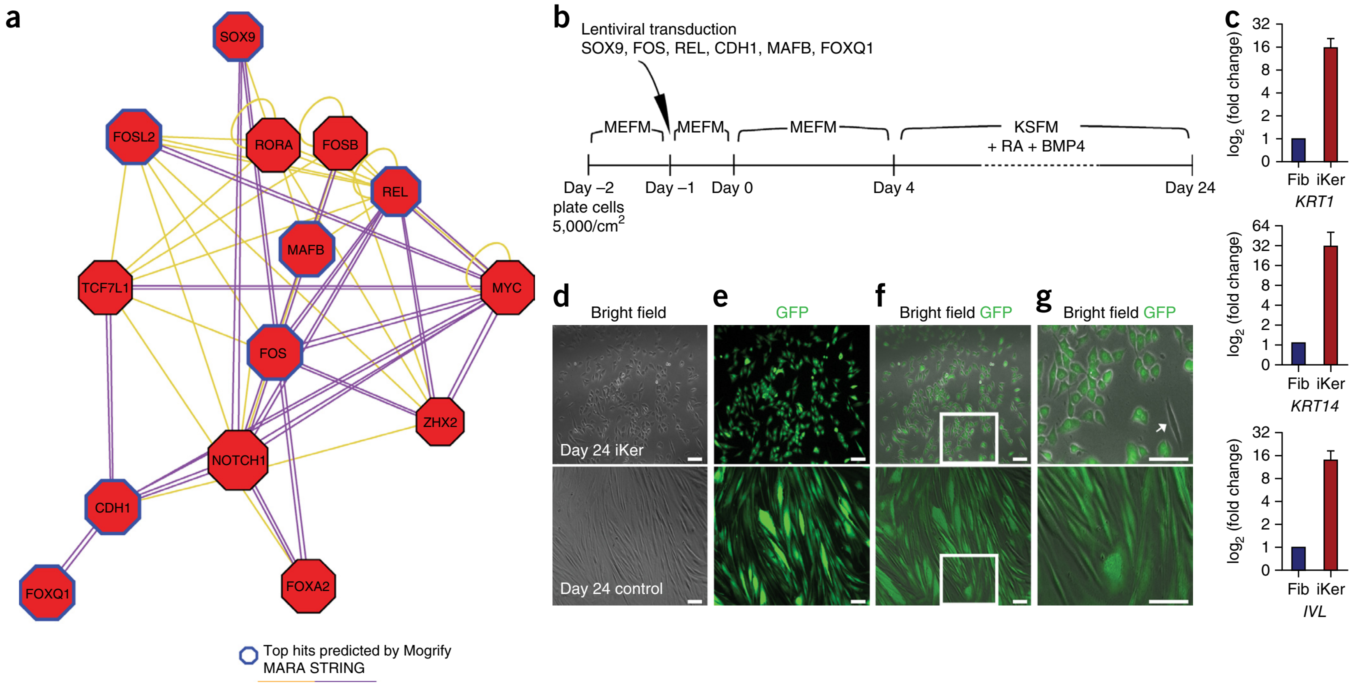

Figure 3 Induction of keratinocytes from dermal fibroblasts. (**a**) The transcription factor network predicted by Mogrify to be involved in the transdifferentiation of dermal fibroblasts into keratinocytes. (**b**) An outline of the method used for the transdifferentiation assay. MEFM, mouse embryonic fibroblast medium; KSFM, keratinocyte serum-free medium; RA, retinoic acid. (**c**) qPCR analysis of the indicated markers in cells collected at days 12–16 during transdifferentiation. All values are from experimental replicates and are given relative to gene expression in dermal fibroblasts (Fib) ($n = 3$; error bars, s.e.m.). (**d**–**g**) Bright-field and GFP images at day 24 showing the cobblestone morphology of transdifferentiated cells (top) and fibroblast morphology of GFP$^+$ control cells (bottom). The box in **f** refers to an area zoomed in **g**. The arrow in **g** indicates an untransduced cell that maintained its fibroblastic morphology. Scale bars, 50 μm.

conversions shown in **Figure 2**, we assessed the ability of Mogrify, CellNet and the entropy-based approach of D'Alessio *et al.*[20] to recover these known factors. The average recovery rate of published transcription factors was 84% for Mogrify, 31% for CellNet and 51% for the method of D'Alessio *et al.* (see **Supplementary Fig. 1** for details). In six of the ten conversions shown in **Figure 2**, Mogrify recovered 100% of the required transcription factors; if Mogrify had been used in the original studies, the experiments could have been a success the first time. CellNet and the method of D'Alessio *et al.* each only recovered all the factors for one of the ten conversions. It is important to note that the conversions proposed by CellNet and the method of D'Alessio *et al.* may also work, as the published conversions represent only one positive example of success.

To demonstrate empirically the predictive capabilities of Mogrify, we conducted two new cell conversions using human cells. The first was a conversion of human fibroblast to keratinocyte (iKer) cells: for this conversion, cells were transduced with lentiviruses encoding *FOXQ1*, *SOX9*, *MAFB*, *CDH1*, *FOS* and *REL*, predicted by Mogrify (**Fig. 3a**,**b** and **Supplementary Table 3**). By day 16 after transduction, the keratinocyte-associated markers keratin 1, keratin 14 and involucrin were markedly upregulated in the transdifferentiated cells (**Fig. 3c**). Moreover, within 3 weeks, the majority of transduced cells exhibited cobblestone morphology, a classic characteristic displayed by keratinocytes. Adjacent untransduced, GFP-negative cells or control cells transduced with GFP-only viruses maintained their fibroblast morphology (**Fig. 3d**–**g**). This morphological and molecular characterization of the reprogrammed cells indicates that Mogrify successfully predicts the transcription factors necessary to induce the conversion of human fibroblasts into keratinocyte-like cells. The second conversion was of adult human keratinocyte (HEKa) cells into microvascular endothelial cells (iECs): for this conversion, we selected *SOX17*, *TAL1*, *SMAD1*, *IRF1* and *TCF7L1* from the seven transcription factors suggested by Mogrify (**Fig. 4a** and **Supplementary Table 4**). These five transcription factors are predicted to regulate ~92% of the required genes for iECs. Once these transcription factors were overexpressed in HEKa cells, we determined that the cells needed to be kept in their media until day 4 (**Fig. 4b**). We used FACS analysis to follow the kinetics of cell reprogramming, using the well-established endothelial marker CD31 (**Fig. 4c**). By day 14 after transduction, we detected that more than 2% of the infected cells had upregulated CD31 levels and, by day 18, almost 10% had upregulated CD31 levels. At this point, we isolated CD31$^+$ cells and evaluated the expression of endothelial-associated genes (*CD31* (*PECAM1*), *CDH5* (encoding VE-cadherin) and *VEGFR2* (*KDR*)) by quantitative PCR (qPCR), which showed a clear reactivation of all the genes assessed (**Fig. 4d**). Finally, we performed immunofluorescence to verify the morphology and expression of the transdifferentiated cells. Only cells transduced with viruses encoding the predicted transcription factors—and not control cells—presented the right morphology and expressed CD31 and VE-cadherin on the cell surface (**Fig. 4e**–**j**). This morphological and molecular characterization of the reprogrammed cells indicates the successful transition of human keratinocytes into human endothelial-like cells.

There have been several reports suggesting that the Yamanaka factors can initiate transdifferentiation without traversing the pluripotent state (reviewed in ref. 1). These findings have recently been challenged by Hochedlinger and Hanna[32,33]. We observe that Mogrify did not predict the use of Yamanaka factors for the transdifferentiations

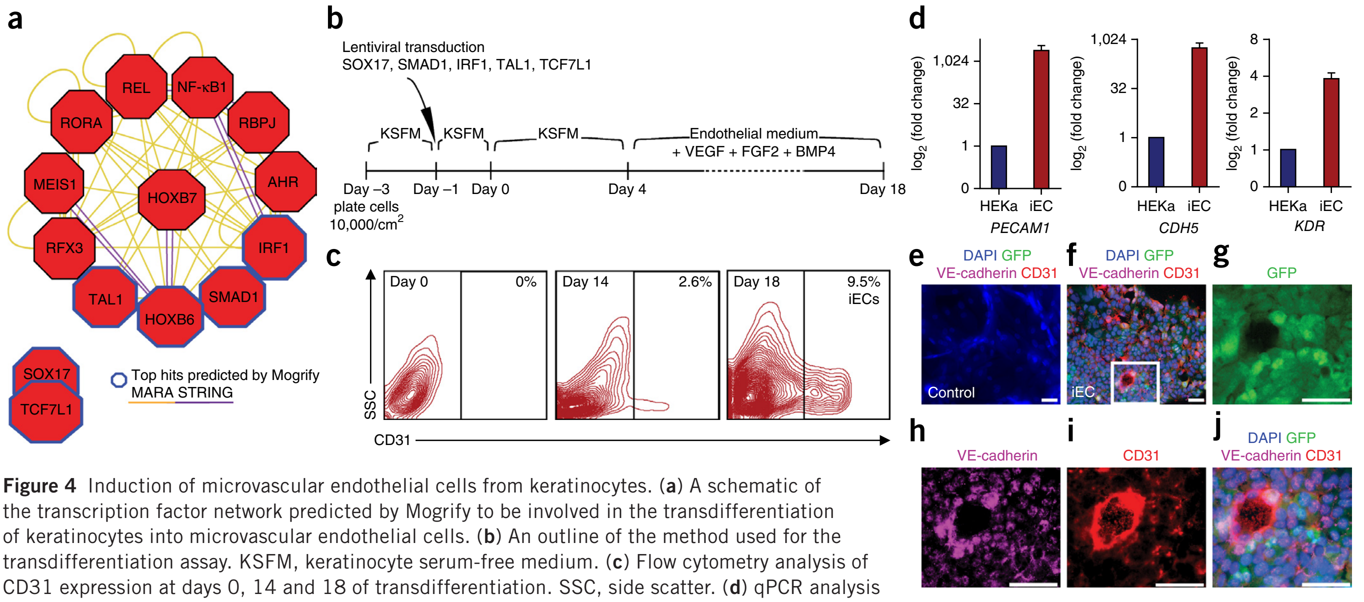

Figure 4 Induction of microvascular endothelial cells from keratinocytes. (**a**) A schematic of the transcription factor network predicted by Mogrify to be involved in the transdifferentiation of keratinocytes into microvascular endothelial cells. (**b**) An outline of the method used for the transdifferentiation assay. KSFM, keratinocyte serum-free medium. (**c**) Flow cytometry analysis of CD31 expression at days 0, 14 and 18 of transdifferentiation. SSC, side scatter. (**d**) qPCR analysis of the indicated expression markers in CD31$^+$ cells collected at day 18 of transdifferentiation. All values are from experimental replicates and are given relative to gene expression in keratinocytes (n = 3; error bars, s.e.m.). (**e**–**j**) Immunofluorescence analysis of the endothelial markers CD31 and VE-cadherin at day 18 for vector-free control cells (**e**) and transdifferentiating cells (**f**). The box refers to an area zoomed in **g**–**j**. Scale bars, 50 μm.

mentioned in this report (except in conversion to iPS cells). Mogrify prediction, however, is based on the source and target regulatory networks and does not have the capacity to detect factors only transiently expressed during the reprogramming process.

Since Conrad Waddington depicted the first epigenetic landscape, several attempts have been made to produce a more representative cellular landscape[34–36], but these efforts have focused on one or two cell types and are based on path-integral quasipotentials, mechanistic modeling or probability landscapes. We hypothesized that the identification of differences in all-against-all comparisons of transcription factor networks as determined by Mogrify in combination with transcriptional profiles would allow the creation of a three-dimensional landscape representing human cell types (**Supplementary Fig. 2**). The landscape places cell types that are molecularly similar close together in the *x*-*y* plane and adjusts the height (*z* direction) according to how likely a cell type is to be a good starting cell source (see the Online Methods for details). Interestingly, we observed that different stem cells were positioned at the highest locations in the landscape. This may suggest that the transcriptional networks of cells at the highest points in landscapes are controlled by fewer transcription factors and that the more differentiated a cell becomes (in valleys), the more transcription factors are needed to fine-tune the transcriptional network.

Having mapped the landscape of human cell types in terms of naturally occurring states and the transitions between them, we note that core control sets of transcription factors that describe the individual cell types are captured, even though the primary aim of Mogrify is to predict transcription factors for cellular conversions. We believe that these sets could aid researchers in determining the role of different transcription factors in their favorite cell type. In practice, Mogrify provides a substantial advance over the strategies currently being applied in laboratories for cell reprogramming, helping in the prediction of transcription factors whose overexpression will induce directed cell conversion. Mogrify has been precalculated on conversions between all possible combinations of the 307 FANTOM5 tissue or cell types resulting in 93,942 directed conversions, with the results provided online (see URLs) via an interface for guiding experimentation and exploring the cellular landscape. Although it is likely that some trial and error will still be involved for some conversions, Mogrify provides a starting point and systematic means to explore new conversions in human cell types. Because Mogrify incorporates a transcription factor redundancy step, it is able to give a finite set of factors as a prediction for the cell conversion, which has greater usefulness than just the ranking of all transcription factors. Although Mogrify has taken advantage of the rich FANTOM5 data, MARA and STRING, these databases have their own limitations, which will impart some restrictions on Mogrify's predictions. For example, the FANTOM5 data are in some instances limited to a few replicates and there is possible heterogeneity in some samples. MARA relies on known DNA-binding motifs to estimate binding to target genes, knowledge that is incomplete. STRING is incomplete in other ways, but future increases in the abundance of empirical data on transcription factor interactions and binding in diverse cell types will help to improve Mogrify. It should be noted that Mogrify as well as other methods finds positive regulators of the target cell and does not interrogate the extinction of the source cell signature. This may result in less faithful conversions, mitigated by the downregulation of source genes that has been observed after the introduction of core target transcription factors (for example, in Polo *el al.*[37]). Mogrify predictions will not guarantee conversion but will certainly aid in the development of transdifferentiation protocols. Other players such as noncoding RNAs, small molecules, epigenetic factors and signaling pathways provide a rich source of improvements for the future. At present, the major challenge to progress in the field of reprogramming is in increasing the number of successful cell conversions. That is what this resource makes possible, paving the way for the routine manipulation of cells, an understanding of the processes involved and the immediate translation of any breakthroughs in the clinical delivery techniques under heavy development in academia and industry.

URLs. Mogrify, http://www.mogrify.net/.

METHODS

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

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

ACKNOWLEDGMENTS

We would like to thank all members of the FANTOM5 Consortium for contributing to the generation of samples and analysis of the data set and thank GeNAS for data production. J.G. and O.J.L.R. were supported by grants from the Biotechnology and Biological Sciences research council and the Japanese Society for the Promotion of Science. J.M.P. was supported by a Silvia and Charles Senior Medical Viertel Fellowship, the Metcalf award from the National Stem Cell Foundation of Australia, National Health and Medical Research Council of Australia (NHMRC) project grant APP1085302 and the Australia Research Council's special initiative Stem Cells Australia. FANTOM5 was made possible by a Research Grant for the RIKEN Omics Science Center from MEXT to Y.H. and a grant of Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from MEXT, Japan, to Y.H.

AUTHOR CONTRIBUTIONS

O.J.L.R. and J.G. initiated the project on the basis of discussions with Y.H. about FANTOM5. J.M.P. led the experimental contribution and helped further develop the Mogrify algorithm. J.F. performed all the experimental validations with contributions from M.L.H., A.S.K. and C.M.N. O.J.L.R. performed the data analysis and interpretation, with significant input from J.G. in the early stages of the work. O.J.L.R., J.M.P. and J.G. prepared the manuscript with input from all named authors at various stages. M.E.O., E.P. and H.F. provided help and advice for technical aspects of the implementation. H.S. and J.W.S. were involved in early discussion on cell conversion concepts. A.R.R.F. and C.O.D. were involved in FANTOM5 management. A.R.R.F. coordinated the collection of the primary cells and tissues profiled in FANTOM5.

COMPETING FINANCIAL INTERESTS

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

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Firas, J., Liu, X., Lim, S.M. & Polo, J.M. Transcription factor–mediated reprogramming: epigenetics and therapeutic potential. *Immunol. Cell Biol.* **93**, 284–289 (2015).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
3. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
4. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).
5. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).
6. Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* **14**, 394–403 (2014).
7. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390–393 (2011).
8. Pfisterer, U. *et al.* Direct conversion of human fibroblasts to dopaminergic neurons. *Proc. Natl. Acad. Sci. USA* **108**, 10343–10348 (2011).
9. Addis, R.C. *et al.* Optimization of direct fibroblast reprogramming to cardiomyocytes using calcium activity as a functional measure of success. *J. Mol. Cell. Cardiol.* **60**, 97–106 (2013).
10. Wilson, D., Charoensawan, V., Kummerfeld, S.K. & Teichmann, S.A. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* **36**, D88–D92 (2008).
11. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
12. Fulton, D.L. *et al.* TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* **10**, R29 (2009).
13. Vickaryous, M.K. & Hall, B.K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
14. Heinäniemi, M. *et al.* Gene-pair expression signatures reveal lineage control. *Nat. Methods* **10**, 577–583 (2013).
15. Lang, A.H., Li, H., Collins, J.J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Comput. Biol.* **10**, e1003734 (2014).
16. Crespo, I. & Del Sol, A. A general strategy for cellular reprogramming: the importance of transcription factor cross-repression. *Stem Cells* **31**, 2127–2135 (2013).
17. Davis, F.P. & Eddy, S.R. Transcription factors that convert adult cell identity are differentially polycomb repressed. *PLoS One* **8**, e63407 (2013).
18. Morris, S.A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889–902 (2014).
19. Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).
20. D'Alessio, A.C. *et al.* A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports* doi:10.1016/j.stemcr.2015.09.016 (23 October 2015).
21. FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
22. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808–D815 (2013).
23. FANTOM Consortium. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* **41**, 553–562 (2009).
24. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
25. Huangfu, D. *et al.* Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat. Biotechnol.* **26**, 1269–1275 (2008).
26. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).
27. Rapino, F. *et al.* C/EBPα induces highly efficient macrophage transdifferentiation of B lymphoma and leukemia cell lines and impairs their tumorigenicity. *Cell Reports* **3**, 1153–1163 (2013).
28. Fu, J.-D. *et al.* Direct reprogramming of human fibroblasts toward a cardiomyocyte-like state. *Stem Cell Reports* **1**, 235–247 (2013).
29. Zou, Q. *et al.* Direct conversion of human fibroblasts into neuronal restricted progenitors. *J. Biol. Chem.* **289**, 5250–5260 (2014).
30. Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**, 386–389 (2011).
31. Kogiso, T., Nagahara, H., Otsuka, M., Shiratori, K. & Dowdy, S.F. Transdifferentiation of human fibroblasts into hepatocyte-like cells by defined transcriptional factors. *Hepatol. Int.* **7**, 937–944 (2013).
32. Bar-Nur, O. *et al.* Lineage conversion induced by pluripotency factors involves transient passage through an iPSC stage. *Nat. Biotechnol.* **33**, 761–768 (2015).
33. Maza, I. *et al.* Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors. *Nat. Biotechnol.* **33**, 769–774 (2015).
34. Qiu, X., Ding, S. & Shi, T. From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLoS One* **7**, e49271 (2012).
35. Bhattacharya, S., Zhang, Q. & Andersen, M.E. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Syst. Biol.* **5**, 85 (2011).
36. Flöttmann, M., Scharp, T. & Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.* **3**, 216 (2012).
37. Polo, J.M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151**, 1617–1632 (2012).
38. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
39. Lattanzi, L. *et al.* High efficiency myogenic conversion of human fibroblasts by adenoviral vector–mediated *MyoD* gene transfer. An alternative strategy for *ex vivo* gene therapy of primary myopathies. *J. Clin. Invest.* **101**, 2119–2128 (1998).

ONLINE METHODS

Mogrify consists of a number of steps, which are outlined below and then described in more depth in the following sections.

1. Collect expression data for each gene (x) in each sample (s).
2. Calculate differential expression against a tree-based background for each gene in each sample and then combine the log-transformed fold change (L_x^s) and adjusted P value (P_x^s) to generate a gene score (G_x^s).
3. For each transcription factor (x) in each sample, calculate the network score (N^S) by performing a weighted sum of the gene scores over two different subnetworks (N_x^s $_{\mathrm{MARA}}$ and N_x^s $_{\mathrm{STRING}}$) centered on each transcription factor.
4. Rank transcription factors on the basis of a combination of the G_x^s and N_x^s scores.
5. Calculate the set of transcription factors for a conversion between any two cell types on the basis of comparisons of the ranked lists from the cell types.
6. Remove transcriptionally redundant transcription factors from the lists.
7. Create a cell conversion landscape by arranging the cell types on a two-dimensional plane according to their required transcription factors and add height on the basis of the average coverage of the required genes that are directly regulated by the transcription factors selected.

Step 1: expression data taken from the FANTOM5 data set. Mogrify uses 700 libraries of clustered CAGE (cap analysis of gene expression) tags, which provide transcription start site (TSS) locations. These are mapped to their corresponding genes (provided by the FANTOM5 Consortium[21]). These data are used to create tag counts for each gene in each library. In total, there are 15,878 distinct genes (of which 1,408 are transcription factors) expressed with at least 20 TPM (tags per million) in at least one sample. (See the website for more details on the libraries analyzed.)

Step 2: tree-based differential expression. Calculating differential expression is a common problem when analyzing biological data, and a number of techniques exist to do this. We elected to use DESeq[38] for this work, as it performs well in benchmark evaluations, allows analysis of some non-replicated data sets and has a short runtime. To calculate differential expression, it is necessary to identify two groups: the set of samples in which you wish to identify differential expression and the background for comparison. The problem of selecting the correct background is important. The inclusion of too many irrelevant samples can reduce the statistical power of the test. The inclusion of too narrow a range or too few samples in the background makes it impossible to tell which genes are truly differentially expressed. One solution is to perform an exhaustive calculation of pairwise tests between each of the cell types. This approach has two problems: first, it is very computationally expensive and, second, it does not identify the genes that are differentially expressed between a sample and an average background but rather identifies ones that are differentially expressed specifically between two samples. For Mogrify, we are interested in the genes that are important for a given cell type in all situations and hence compare against a collection of samples. To do this, we implemented a tree-based background selection method based on the FANTOM5 cell ontology[21] (**Fig. 1b**). The principle of this approach is to exclude cell types whose ontologies are very close to that of the cell type of interest while including others that are near in the tree to the background. This was achieved by selecting a point near the top of the tree that would act as the breaking point. Samples in the same clade as the cell type being analyzed were removed, and those not in the same clade but still below the breaking point were included. The result of this process is a set of samples that is broad enough to give reliable results but narrow enough that the statistical power is kept at a manageable level. It is possible that, if a transcription factor exists whose expression correlates highly with the choice of background, this factor will go undetected for the cell in question, but this tradeoff is justified by the advantages outlined above.

This tree-based background selection approach for DESeq is run on all FANTOM5 libraries (grouped by replicates), generating a log-transformed fold change in expression and a false discovery rate (FDR)-adjusted P value for each gene in each sample. Because the background is non-uniform, the results of each differential expression calculation are not directly comparable; hence, for the remaining steps, these figures are used to rank transcription factors in each sample and it is the rankings that are compared.

Because we are only interested in identifying transcription factors with a high level of influence, we convert the log-transformed fold changes in expression and FDR-adjusted P values into a single positive score (G_x^s) using the following equation

$$G_x^s = \left|L_x^s\right|\left(-\log_{10} P_x^s\right) \qquad (1)$$

where L_x^s is the log-transformed fold change in the expression of gene x in sample s and P_x^s is the adjusted P value for gene x in sample s. The formula ensures that genes with a high log-transformed fold change in expression and a low adjusted P value score very highly and vice versa. This is applied to every gene in each sample, creating a matrix of 700 samples × 15,878 genes for differential expression.

Step 3: calculate a transcription factor's network-based sphere of influence. To assess the importance of each transcription factor, its effect on its local neighborhood is calculated using two sources of network information: the STRING database[22] and MARA[23]. These two databases contain different types of interactions. MARA provides protein-DNA interactions for transcription factors with known binding sites in the promoter regions of a gene. This represents a low-level, directed regulatory network of interactions. STRING is a meta-database of interactions that contains various types of interactions, including protein-protein, protein-DNA and protein-RNA interactions, as well as biological pathways. This provides a view of the interactions taking place that both directly and indirectly affect gene expression.

To calculate influence, a weighted sum of gene influences (from step 2) is performed over a transcription factor's local network neighborhood. This local network is constrained to a maximum of three edges, and the effect of each node diminishes the further from the seed transcription factor it is located and depending on the out degree of its parent (**Fig. 1c**). Distance weighting is used so that genes that are increasingly distant from direct regulation have less of an impact on the score. Edge weighting is used to compensate for highly ubiquitous transcription factors and prevent them from receiving artificially high scores by regulating a large number of barely differentially expressed genes. We consider that a transcription factor that is regulating ten genes that have $G_x^s = 100$ to be more important than a transcription factor that is regulating 1,000 genes that have $G_x^s = 1$.

The equation to perform this weighted sum is

$$N_{x,n}^s = \sum_{r \in V_x}^{s} G_r^s . \frac{1}{L_{r,n}} . \frac{1}{O_{r,n}} \qquad (2)$$

where $r \in V_x$ is each gene (r) in the set of nodes (V_x) that make up the local subnetwork of transcription factor x, $L_{r,n}$ is the level (or number of steps) r is away from x in network n and $O_{r,n}$ is the out degree of the parent of r in network n. This is performed over both the MARA and STRING networks, resulting in two transcription factor influence lists ($N_{x,\mathrm{MARA}}^s$ and $N_{x,\mathrm{STRING}}^s$).

Step 4: rank the transcription factors on the basis of the results of steps 2 and 3. The results of steps 2 and 3 are three ranked transcription factor lists for each sample based on G_x^s, $N_{x,\mathrm{MARA}}^s$ and $N_{x,\mathrm{STRING}}^s$. To obtain the final ranking of each transcription factor in each sample, its rank in each of the three lists is added together. Ranks are limited to a maximum of 100, as we observed that beyond the top 100 transcription factors the remaining regulatory influence was very small. If a transcription factor does not appear in a particular list, then it is given a score of 100. The result of this is a single ranked list of transcription factors for each cell type; those with the lowest score or rank are predicted to facilitate a cell conversion.

Step 5: compute all pairwise experimental comparisons to create predictions. To predict the set of transcription factors for a given conversion, the ranked lists from the source and target cell types are compared. If a transcription factor from the target cell type list is already expressed in the source target (at greater than 20 TPM), then it is removed from the list.

doi:10.1038/ng.3487

Step 6: remove transcriptionally redundant transcription factors. Once the final ranking is complete, regulatory redundancy is removed. This is achieved by comparing the lists of genes that the transcription factors directly regulate. For a given transcription factor, if there is a higher-ranking transcription factor that regulates over 98% of the genes that it would regulate, then this transcription factor is removed. This means that the resulting predictions include transcription factors that are diverse in their regulatory spheres of influence. This cutoff was chosen empirically to minimize the number of factors predicted while maximizing network coverage (see **Supplementary Fig. 3** for details).

Step 7: create a cell reprogramming landscape based on steps 1–6. To create the reprogramming landscape, we calculate the *x* and *y* coordinates independently of the *z* coordinate. To reduce the complexity of the landscape, we average the gene expression profiles of individual samples grouped by the cell ontology provided by FANTOM5. The result of this is a set of 314 ontologies that contain at least three samples each from which we have the average gene expression. The *x* and *y* coordinates are calculated by multidimensional scaling (MDS) of these profiles. The result of MDS is a projection of the data where the distance between points is maintained from the multidimensional reality to two-dimensional reduction. As a result, two points that are close together in the *x*-*y* plane of the landscape have similar expression profiles and, as such, represent similar cell types. The *z* axis of the landscape is calculated by considering the regulatory coverage of the top eight Mogrify-predicted transcription factors. For every conversion, we look at the set of genes that are expressed in the ontology and the number of these that are directly regulated by each transcription factor. We calculate the area under the curve (AUC) of the cumulative coverage for the top eight transcription factors normalized by the maximum possible AUC to retrieve a value between 0 and 1 for each ontology as the height. As such, a height of 1 represents an ontology where all of the required genes are directly regulated by the top ranked transcription factor and a height of 0 represents an ontology where none of the top eight transcription factors directly regulate any of the required genes. The *x*, *y* and *z* values are then used in the R package plot3D to generate the landscape using the image2D and persp3D packages. The different stem cells at the highest locations were found with a gene set enrichment score of 0.41 and a *P* value of 0.011.

Benchmarking. To compare the performance of Mogrify with that of other methods, a benchmarking experiment was carried out. First, we assessed the effect on performance of using the complete Mogrify algorithm in comparison to the MARA, STRING and differential expression components alone. Second, a comparison with CellNet and the method of D'Alessio *et al.* was carried out. These are the only other techniques that currently provide a means to calculate transcription factor sets for a wide variety of cell types. To carry out comparisons, the sets of transcription factors from the published conversions shown in **Figure 2** were used as true positives. The benchmarking consisted of assessing the performance of each technique in recovering these transcription factors using the following steps.

Step 1. For each conversion, identify the number of transcription factors to consider. Mogrify is the only method to provide a set of transcription factors rather than a ranked list of all transcription factors; because the object is to compare other methods to Mogrify, the information generated by Mogrify on the number of factors to use is applied to the other methods, that is, no method is allowed to use more factors than the other methods. For example, for the conversion between B cells and macrophage, Mogrify predicted that eight transcription factors should be adequate, so for all methods the top eight transcription factors were used for comparison.

Step 2. For each method, determine whether the correct transcription factors have been predicted. For each published set of transcription factors, the predictions from each method are compared and two statistics are extracted: first, the recovery rate for the published transcription factors (which is 100% if all of the published factors are contained in the predicted set) and, second, the average rank of the published factors (for each correctly identified transcription factor, the ranks are summed and divided by the total number of correctly identified transcription factors).

The results from these two steps can be found in **Supplementary Tables 5** and **6**, and a summary of the comparison of Mogrify to CellNet and the method of D'Alessio *et al.* can be found in **Supplementary Figure 1**.

To extract the results for CellNet, we used publicly available data sets for fibroblasts (Gene Expression Omnibus (GEO), GSE14897) and B cells (GSE65136) as the starting point and used the web interface for CellNet (http://cellnet.hms.harvard.edu/) to provide predictions for each of the conversions in **Figure 2**. D'Alessio *et al.* provide ranked sets of transcription factors for many cell types, and these ranked lists were used for the comparison.

Lentivirus generation. For lentivirus generation, human embryonic kidney (HEK) 293T cells (Sigma) were cultured in T-75 flasks. Once they reached 90–95% confluence, they were transfected with a -lv165 vector expressing CDH1, FOS, FOXQ1, IRF1, MAFB, REL, SMAD1, SOX9, SOX17, TAL1 or TCF7L1 from the *EEF1A1* promoter and IRES2-eGFP (GeneCopoeia), together with the second-generation packaging plasmids psPAX2 and pMD2.G from the Trono laboratory (Addgene), using LTX Lipofectamine (Invitrogen) transfection reagent. Virus-containing supernatants were collected at 24 and 36 h after transfection and concentrated with ultracentrifugation filters (Millipore). Viral concentrates were then stored at −80 °C. Titrations were based on eGFP expression, as determined by flow cytometry. The cell line used in these experiments tested negative for mycoplasma contamination.

Cell culture. Before their use in experiments, human adult epidermal keratinocytes (HEKa cells; Gibco) and human dermal fibroblasts (HDFs; Gibco) were plated at 2.5 × 10^3 cells/cm^2 and passaged at least three times. HEKa cells were cultured in KSFM (Gibco) that contained 10% HKGS (Gibco) and 1% penicillin-streptomycin (Gibco). In contrast, HDFs were cultured in medium 106 (Gibco) that contained 10% LSGS (Gibco) and 1% penicillin-streptomycin. Cells were then frozen in liquid nitrogen for later use. For the transdifferentiation of keratinocytes into endothelial cells, cells were thawed and seeded at 2.5 × 10^3 cells/cm^2 and grown until they reached 90% confluence. They were then reseeded at 5.0 × 10^3 cells/cm^2 and grown for 2 d in KSFM before being infected with concentrated lentiviral particles encoding IRF1, SMAD1, SOX17, TAL1 and TCF7L1 in the presence of polybrene (Millipore) in KSFM. After the addition of viruses (12–24 h), the medium was replaced with fresh KSFM. At day 4, the medium was replaced with human endothelial serum-free medium (Gibco) with 1% penicillin-streptomycin containing human VEGF (50 ng/μl; PeproTech), human BMP4 (20 ng/μl; PeproTech) and human FGF2 (20 ng/μl; PeproTech). For the transdifferentiation of fibroblasts into keratinocytes, cells were seeded at 2.5 × 10^3 cells/cm^2 and grown until they reached 90% confluence. They were then reseeded at 2.5 × 10^3 cells/cm^2 and grown for 24 h in MEFM before being transduced with lentiviral particles encoding CDH1, FOS, FOXQ1, MAFB, REL and SOX9 in the presence of polybrene in MEFM for 24 h. At day 4, the medium was replaced with KSFM containing 1% penicillin-streptomycin, 1 μM retinoic acid and 25 ng/ml human BMP4 (R&D). Fresh medium was added at least once every 2 d throughout all of the experiments. Each of the experiments was repeated three or four times.

Flow cytometry. At various time points, transdifferentiating cells were dissociated with 0.25% trypsin-EDTA (Gibco) for 3 min at 37 °C. Cells were then prepared for flow cytometry analysis or sorting. They were incubated with APC-conjugated antibody to human CD31 (17-0319-41, eBioscience; 1:200 dilution) at 4 °C for 15 min, washed with DPBS (Gibco), centrifuged at 300*g* for 7 min and then resuspended in medium containing propidium iodide (Sigma-Aldrich). An LSR-II analyzer (BD Bioscience) and the Influx cell sorter (BD Biosciences) were used for data analysis and cell sorting, respectively.

Quantitative PCR. Total RNA was extracted using the RNeasy Micro kit (Qiagen) following the manufacturer's instructions. Extracted RNA was reverse transcribed into cDNA using the Superscript III kit (Invitrogen). Real-time qPCR reactions were set up in triplicate using Brilliant II SYBR Green QPCR Mastermix (Stratagene) and then run on the 7500 Real-Time PCR System (**Supplementary Table 7**).

doi:10.1038/ng.3487

Immunofluorescence. Cells were fixed with 4% paraformaldehyde in DPBS at room temperature for 10 min. There was no need to permeabilize cells as the markers of interest are expressed on the cell surface. Cells were blocked with 5% donkey serum in DPBS for 30 min and then incubated with primary antibodies (goat polyclonal antibody to CD31, sc-1506, Santa Cruz Biotechnology (1:100 dilution) and rabbit polyclonal antibody to VE-cadherin, ab33168, Abcam (1:1,000 dilution)) overnight at 4 °C. The next day, cells were incubated with secondary antibodies (donkey anti-goat conjugated to Alexa Fluor 555, Invitrogen (1:2,000 dilution) and donkey anti-rabbit conjugated to Alexa Fluor 647, Invitrogen (1:2,000 dilution)) for 2 h at room temperature. Finally, cells were overlaid with 4′,6-diamidino-2-phenylindole (DAPI; Life Technologies; 1:1,000 dilution) for 1 min. All images were acquired using the inverted Nikon Eclipse T*i* epifluorescence microscope with the Nikon Digital Sight DS-U2 camera and were processed and analyzed using Fiji software.

doi:10.1038/ng.3487