Mogrify: Factors For Direct Reprogramming Between All Cell Types

Owen Rackham*[1,2], Hai Fang[2], Matt E. Oates [2] , the FANTOM consortium, Harukazu Suzuki[3,4], Jay W. Shin [3,4], Carsten O. Daub [3,4,5], Alistair R.R. Forrest [3,4] and Julian Gough[2]

1. *Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital, London, UK*
2. *Dept. of Computer Science, University of Bristol, Bristol, UK*
3. *RIKEN Omics Science Center, Yokohama, Kanagawa, 230-0045 Japan$*
4. *RIKEN Center for Life Science Technologies, Division of Genomic Technologies, Yokohama, Kanagawa, 230-0045 Japan*
5. *Department of Biosciences and Nutrition, Karolinska Institutet, NOVUM S-14183 Huddinge, Stockholm, Sweden*

§RIKEN Omics Science Center ceased to exist as of April 1st, 2013, due to RIKEN reorganization

***Corresponding author**

Address: Dr Owen Rackham, Clinical Sciences Centre, Hammersmith Hospital, London

Email: owen.rackham@imperial.ac.uk

Abstract

It is known that over-expression of sets of endogenous transcription factors in somatic cells can induce stem-cell-like pluripotency or trans-differentiation. These discoveries relied on exhaustive testing of large sets of transcription factors, an approach that is both inefficient and unscalable.

Here we present a novel network-based method (Mogrify) that combines gene expression data with regulatory network information to identify targeted sets of transcription factors to induce cell conversion between any two cell types. We validate Mogrify by recovering known reprogramming factors for all published cell conversions as well as providing further experimental evidence. Further to this, novel transcription factors for published conversions are identified, but critically transcription factor sets to induce novel cell conversions are presented, culminating in a trans-differentiation landscape of human cell types.

Mogrify is available for every conversion between the 705 different human cell-types in the FANTOM5 dataset and made freely available to the community via a web interface. (Mogrify.net).

username:editor password:FANTOM which will be removed upon publication

Introduction

Cells are the minimum unit of life, containing all of the information required for survival and the ability to embody this information such that they can react to their environment and change themselves accordingly. Multicellular organisms implement a division of labour between cell types. Different cell types are specialized for certain tasks, allowing the organism as a whole to be more efficient but ultimately removing the ability of any single cell type to survive in the absence of the others. Within a multicellular organism, the difference between any two cell types starts with the genes that are expressed. A large number of genes are required for all cell types, known as housekeeping genes, but more specialized genes are required/expressed in a subset of all possible cell types. Each cell type has its own stable programmatic state. These different expression programs are come about as a result of cellular differentiation, a genetic re-wiring process that was generally thought to act only in one direction.

We now know that this process can be artificially managed and even reversed via the introduction of exogenous transcription factors (TFs). A set of four transcription factors (Oct3/4, Sox2, c-Myc and Klf4) introduced into fibroblasts is enough to induce a conversion into an induced pluripotent stem cell (iPS)[1]. This discovery opened the possibility that all cellular phenotypes could be reprogrammed artificially. Further reports demonstrated that different sets of transcription factors introduced into the same starting cell type could increase the efficiency of the conversion to a stem cell[2], or lead to directed trans-differentiation into other cell types including: myoblasts[3], neurons[4–6] , hepatocytes[7,8] and cardiomyocytes[9].

These discoveries came about through a process of exhaustive testing of large sets of transcription factors combined with expert knowledge. The field, which initially showed a great deal of promise, has now stalled in its attempt to find sets of transcription factors for further conversions into new cell types. With roughly 2000 different TFs[10–12] and approximately 400 unique cell types in humans[13], the space of possible sets is very large and impractical to explore using the current approach. Therefore there is a clear need for computational framework to guide experimentation.

In this work, we propose and implement a network-based computational technique, validated against current knowledge and experimental results. The purpose of this technique is to

identify the required TFs for any cell conversion between any two cell types or tissues from the FANTOM5 project. We show that we are able to independently recover via prediction, the factors that were previously discovered experimentally for successfully converting cell types. We validate the technique using a high-throughput experimental screen and also provide examples of novel transitions for future testing. In each case literature support for each of the predicted factors is provided and a computational *'reprogramming'* landscape describing the relationships between each of the cell types in the FANTOM5 dataset is produced.

This work is part of the FANTOM5 project[14]. Data downloads, genomic tools and co-published manuscripts have been summarized at http://fantom.gsc.riken.jp/5/top/

Results and Discussion

Mogrify ranks TFs for a given cell type based on their predicted influence over the local regulatory network (see figure 1 for a summary and methods for details). This ranking is achieved by considering the differential expression of the TF as well as its downstream target genes, hence quantifying the TF's influence on regulation within the cell.

Once important TFs for each type have been found, a pairwise comparison of source and target cell types can be performed in order to calculate the factors that should be used to induce a cell conversion. During this phase TFs that are expressed with greater than 20 tags per million reads (TPM) in both target and source cell type as well as TFs that provide redundant regulation (i.e. if a higher ranking TF already regulates 98% of the same genes) are excluded.

Biologically speaking, this identifies TFs which regulate the genes that are most important to a given cell type in terms of expression and regulatory influence. Most other related approaches in a similar context do not take the local network of a TF into account. By doing this, Mogrify is able to identify TFs that sit at the top of activity cascades, as well as those TFs whose own expression level may be low but whose regulatory influence is high, for instance, as a result of post-translational modification.

In each trans-differentiation Mogrify predicts a ranked list of transcription factors. We have chosen not to apply an arbitrary cut-off to the number of factors in this list, instead leaving this choice to the experimental group exploring a conversion. We observe that the likely number of factors needed (and indeed number of possible sets) is different depending on the conversion and as a result experimental investigation is required to inform this choice.

Prediction of known reprogramming factors for cell conversion

The FANTOM5 project has used Cap Analysis of Gene Expression (CAGE) to generate a sequencing based expression atlas that covers a broad range of primary cell types, tissues and cell lines[14]. Within the collection, digital expression profiles are available for multiple donor and target cell types for which successful trans-differentiations have been previously published. We are able to show that Mogrify can independently predict the factors that induce these trans-differentiations based on the FANTOM 5 data and network information (see figure 3 for a summary).

Fibroblast to Embryonic Stem Cells

It is known that human fibroblasts can be converted to iPS cells by introducing OCT4, SOX2, NANOG and LIN28[15]. Impressively Mogrify predicts NANOG, OCT4 and SOX2 as the top 3 TFs with equal frequency for this conversion. We note that, although we do not find LIN28 here, the original publication found that LIN28 was not essential for reprogramming but improved the efficiency. Mogrify did not assign a high rank to KLF4, which was used in the original conversion to iPs cells[1], as it was predicted to be redundant to higher-ranking TFs. This agrees with the later work by Yu *et al*[15] showing that KLF4 is not required.

Fibroblast to Cardiomyocytes

Mouse fibroblasts have been converted to cardiomyocytes using TBX5, GATA4 and MEF2C[9]. For the same conversion, Mogrify predicts GATA4, GATA6, HAND1, NKX2-5 and TBX5 as the top 5 TFs. This set includes two of the three TFs used in the reprogramming experiment with the third, MEF2C, being ranked 8th. It was shown by Zhou et al[16] that MEF2C was not required in order to induce cardiac marker genes, reporting that TBX5, GATA4 and transcriptional co-activator MYOCD (not detectable by Mogrify) were the most efficient set of TFs. NKX2-5 was included in the starting set of 14 TFs in the original experiment[9]. GATA6 is known to act with GATA4 in early heart development[17] while HAND1 has been shown to facilitate cardiomyocyte proliferation rather than differentiation[18]. A close homolog of HAND1 (HAND2) has been used alongside GATA4, TBX5 and two microRNAs in order to perform the same conversion.

Fibroblast to Hepatocytes

Two successful cell conversions between mouse fibroblast and hepatocyte have been reported. The first paper reported the use of three TFs (GATA4, HNF4A and FOXA3)[8] and an additional knockdown of CKDN2A. It was followed by a different combination that used just two TFs (HNF4A and one of FOXA1, FOXA2 or FOXA3)[7]. Mogrify results for this conversion (figure 2) identified FOXA2 and HNF4A in rank positions first and second.

Interestingly Mogrify also predicts two nuclear receptor genes (NR1H4 and NR5A2). These are ligand dependent TFs that are required for liver development. They are also important in the early stages differentiation[19] when they regulate genes involved in pluripotency such as OCT4[20]. Two other genes of significant interest are ONECUT1 and ATF5, both high scoring in the Mogrify predictions. In particular ONECUT1 is known to stimulate, amongst other things, FOXA2 expression and is important in hepatic differentiation[21]. This conversion has never been successful in human; it is possible that the additional factors identified by Mogrify (in human data) may be enough to make the conversion possible.

Converting to Neurons

There are a number of reports in the literature for trans-differentiations from various cell types to neurons in both mouse and human (see table 1). The set of transcription factors used for these experiments was not the same, but some are shared. For instance ASCL1, MYT1L, NEUROD1, NEUROG2 and BRN2 (also known as *POU3F2)* occur in at least 4 of the 7 published conversions. The set of the 9 highest-scoring TFs predicted by Mogrify for a conversion from dermal fibroblasts to neurons is: *CUX2, SOX2, ZNF238, SOX9, FOSB, RFX4, BRN2, ARNT2 and MEF2C.* This set includes BRN2 (shared by experiments) but both ASCL1 and MYT1L are absent. The reason they are absent is that Mogrify is designed to remove TFs from the list of predictions where their regulatory influence is redundant to a higher-ranked TF. In this instance NEUROD1, ASCL1, NEUROG2 and MYTL1 would be ranked third, fifth, sixth and 14th respectively but were removed due to redundancy to one of the top 4 predictions (*CUX2, SOX2, ZNF238 and SOX9*). It has been shown in other work looking at the roles of SOX2, NEUROD1 and NEUROG1 that SOX2 is sufficient for commitment to neural lineages[22,23]. Further to this, it has been demonstrated that SOX2 up-regulates the expression of NEUROD1, which in turn down-regulates the expression of SOX2 in a negative-feedback loop during brain development[24]. CUX2 has been reported as the notch effector gene that regulates interneuron development by activating proneural TFs such as ASCL1, OLIG3 and NEUROG2[25]. Thus, it would appear that CUX2 and SOX2 are higher order regulators for the neuron lineage that in turn activate those genes used in the published cell conversions.

Fibroblast to Macrophage

Trans-differentiation between fibroblasts and macrophages is facilitated by SPI1 with an increase in efficiency on addition of CEBPA[26]. By comparing these known factors with the Mogrify predictions (see figure 2) we see that SPI1 (also known as PU.1) is predicted as the top factor followed by MEF2A and MITF. It has previously been shown that MEF2A coordinates the behaviour of the CEBP family of TFs during differentiation in pigs[27]. This is

another example of Mogrify identifying higher order regulators that are known to coordinate the behaviour of identified reprogramming genes.

Comparison to other techniques

There are few published techniques for producing ranked lists of TFs for a given cell type. However, as part of FANTOM5, two other approaches were applied to the data: Motif Activity Response Analysis (MARA), which looks for correlation in expression for all of the genes which contain a transcription factors binding site and a gene expression enrichment metric, which looks at how specific the expression of a transcription factor is to each cell type[14]. A comparison to these techniques gives some indication of Mogrify's performance. The results from ranking the required TFs for five known conversions are shown in Table 2.

Although Mogrify does not perfectly predict the known TFs (i.e. always being ranked at the top), it performs much better than both of the other techniques in all of the 5 well known trans-differentiations, with the exception of MEF2C in the heart example; MARA ranks it in 1st position and Mogrify in 8th position. In this test, Mogrify would have predicted the correct factors in the top ten for each of the known conversions and in the majority of cases in the top 3 ranked positions.

Since Mogrify combines network information from STRING and MARA as well as using gene expression data to calculate the TF ranking it is useful to compare the results of using each individual information source separately (see table 2). The results show that the combination is much more powerful than using each source separately. This is likely to be because of the different aspects of biology that each source reflects. MARA provides TF-DNA interactions, STRING many interactions/associations between proteins, and the gene expression data provides the specificity of each gene. When these data sources are combined a much more complete picture of the cell identity is achieved than only considering each data source alone. Whilst no weighting of the techniques is applied in Mogrify the predictions made in each case are converted to ranks and these are compared.

Validation with experimental results

Attempts have been made to experimentally test sets of TFs for cell conversion in a high throughput way. For instance in Shin *et al.*[28] the authors developed a screening technique that attempts to randomly test the ability of different sets of TFs at converting fibroblasts to monocytes. These results provide a dataset linking different combinations of TFs with the number of marker genes activated. We assume that the number of expressed marker genes is correlated with the success of the conversion. As a result the Spearman’s rank correlation

between the scores provided by each technique for a given set of TFs and the number of marker genes expressed in the experiment provides a comparison between techniques and a measure of their ability via experimental validation (See figure 3).

The results show that Mogrify predictions correlate more closely with the marker gene expression (0.84) than transcription factor specificity (0.73). These correlations are based on combinations of a limited number of TFs that were selected by the authors of Shin *et al.*[28]. Since increasing the size of the TF set is correlated with the number of marker genes expressed, figure 3 also includes the correlation that can be achieved using just the number of TFs for comparison.

Reprogramming Landscape

Several attempts have been made to produce a cellular landscape[29–31]. These attempts have focused on one or two cell types and attempt to create a landscape based on path-integral quasi-potentials, mechanistic modeling or probability landscapes. In this work we define a landscape based on the Mogrify predictions for TFs required for over-expression (see figure 4). Samples are grouped using the cell ontology terms provided by Forrest *et al*[14]. The ontology terms group samples at various levels in terms of their similarity, for example all fibroblast samples or all in-vitro samples. To create a landscape, the ontology terms are arranged using a multidimensional scaling of their average gene expression profiles. As a result, samples with similar expression profiles are close together in the x-y plane. The height of the landscape (shown using colour in figure 4a but rendered as textured landscape on the cover image and in more detail in supplementary figure 1) is calculated by first calculating which genes need to be activated in the target cell type and then calculating the proportion of these that appear in the local neighbourhood of each of the top ranking TFs (see methods). As such, target cell type where the top ranking TFs regulate most or all of the required genes are assumed to be those where TFs can have the greatest influence. As a result, areas on the landscape that are `high' are those that should be the easiest to convert.

The landscape suggests that the two classes of cells that should be the easiest to convert into are embryonic stems cells and cells from the immune system. Cell types that can be found close to the `*embryonic stem cell ridge'* are predicted to be good potential sources of cells for iPS conversion. These cell types include fibroblasts, amniotic membrane cells, salivary acinar cells, blood/immune cells (in particular T-cells) and neuron/ neural stem cells. Each of these cell types with the exception of salivary acinar cells has already been the subject of study into the production of iPS[32–36]. Although these experiments do not

conclusively prove that Mogrify finds the easiest donors from which to reach the stem-cell-like state, they do confirm that the top predictions are viable.

Untested cell conversions

The landscape also highlights as yet untested cell conversions that Mogrify predicts to be the most plausible. Two such cell conversions are described below (many more can be found by exploring www.Mogrify.net).

The first example of a putative cell conversion is between endothelial cells of the vein and astrocytes, for which the top four predicted TFs were SOX2, SOX9, SNAI2 and ARNT2. According to the landscape, these cell types have similar expression profiles and the predictions contain TFs that regulate many of the required genes. Both SOX2 and SOX9 are already known to promote differentiation towards astrocytes and inhibit neurogenesis during normal brain development[37,38].

The second promising putative conversion is between lymphatic vessel fibroblast and renal mesangial cells. This conversion is predicted to require PAX8, PAX2, HOXA7 and FOXD1. It has been previously shown that PAX8 and PAX2 are both required for the mesenchymal-epithelial transitions during nephric duct formation[39]. FOXD1 modulates the formation of discrete sections in the kidney, with FOXD1 negative mutant mice having no discrete kidney cell type zones and pelvic fused kidneys[40].

Online resource for cell conversion

Mogrify has been run on conversions between all possible combinations of FANTOM libraries resulting in 393,792 pairwise conversions. These are provided to the community online with an interface designed to facilitate their exploration (www.Mogrify.net), with the intention of making the predictions and algorithm as accessible as possible to researchers in the field. Website users can choose to explore the results in three different ways: by selecting a particular conversion of interest; by selecting a particular TF of interest; or lastly by exploring the landscape of all predictions. For instance a user who is interested in cell conversions between fibroblast and retinal pigment epithelial cells could select this transition from the main page and would be presented with a list of predicted TFs, a network showing how these TFs are related and a graph showing the regulatory coverage of the required genes (see supplementary figure 4). It is then possible to enter the set of TF clones that the user has at their disposal, and Mogrify will display the optimal set of TFs from those available. If there are experimental indications that some TFs are not favourable, this can be entered into the website allowing Mogrify to update its predictions for that conversion on the

strength of those results. So Mogrify can be used as a practical tool to both begin and iterate an experimental programme based on available materials and outcomes.
There is a space on every prediction page on the Mogrify website where users are invited to leave comments, so that discussion between groups working on (or expert in) the same or similar conversions can share their knowledge or progress.

Conclusions

This is the first predictor of its kind, accurately predicting TFs whose over-expression will induce directed cell conversion. The extent of the data collected by FANTOM5 also means that a huge number of predictions can be made, allowing for the first time, a cell reprogramming landscape to be drawn. The predictions are provided online via an interface to guide experimentation and for exploration of the cellular landscape. It is hoped that this will stimulate many more successful trans-differentiations. Whilst it is highly unlikely that all of the predictions made by Mogrify are correct, it provides an intelligent system for taking advantage of the rich FANTOM data. The field of cell conversion is already exploring small-molecule[41] and RNA[42,43] induced cell conversions and the inclusion of these data types is an immediate priority for release in Mogrify. At present the major challenge to progress the field is in increasing the number of successful cell conversions so that a better understanding of the process can be achieved. This resource will play a role in allowing that to happen, consequently enabling refinement in the method.

References

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–76 (2006).

2. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).

3. Choi, J. *et al.* MyoD converts primary dermal fibroblasts, chondroblasts, smooth muscle, and retinal pigmented epithelial cells into striated mononucleated myoblasts and multinucleated myotubes. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 7988–92 (1990).

4. Ambasudhan, R. *et al.* Direct reprogramming of adult human fibroblasts to functional neurons under defined conditions. *Cell Stem Cell* **9,** 113–8 (2011).

5. Qiang, L. *et al.* Directed conversion of Alzheimer's disease patient skin fibroblasts into functional neurons. *Cell* **146,** 359–71 (2011).

6. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463,** 1035–41 (2010).

7. Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475,** 390–3 (2011).

8. Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475,** 386–9 (2011).

9. Ieda, M. *et al.* Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell* **142,** 375–386 (2010).

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,** D88–92 (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,** 252–63 (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,** 425–55 (2006).

14. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).

15. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).

16. Zhou, L., Liu, Y., Lu, L., Lu, X. & Dixon, R. A. F. Cardiac gene activation analysis in mammalian non-myoblasic cells by Nkx2-5, Tbx5, Gata4 and Myocd. *PLoS One* **7,** e48028 (2012).

17. Xin, M. *et al.* A threshold of GATA4 and GATA6 expression is required for cardiovascular development. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 11189–94 (2006).

18. Risebro, C. A. *et al.* Hand1 regulates cardiomyocyte proliferation versus differentiation in the developing heart. *Development* **133,** 4595–606 (2006).

19. Fayard, E., Auwerx, J. & Schoonjans, K. LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. *Trends Cell Biol.* **14,** 250–60 (2004).

20. Gu, P. *et al.* Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.* **25,** 3492–505 (2005).

21. Pierreux, C. E., Vanhorenbeeck, V., Jacquemin, P., Lemaigre, F. P. & Rousseau, G. G. The transcription factor hepatocyte nuclear factor-6/Onecut-1 controls the expression of its paralog Onecut-3 in developing mouse endoderm. *J. Biol. Chem.* **279,** 51298–304 (2004).

22. Puligilla, C., Dabdoub, A., Brenowitz, S. D. & Kelley, M. W. Sox2 induces neuronal formation in the developing mammalian cochlea. *J. Neurosci.* **30,** 714–22 (2010).

23. Ring, K. L. *et al.* Direct reprogramming of mouse and human fibroblasts into multipotent neural stem cells with a single factor. *Cell Stem Cell* **11,** 100–9 (2012).

24. Kuwabara, T. *et al.* Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nat. Neurosci.* **12,** 1097–105 (2009).

25. Iulianella, A., Sharma, M., Vanden Heuvel, G. B. & Trainor, P. A. Cux2 functions downstream of Notch signaling to regulate dorsal interneuron formation in the spinal cord. *Development* **136,** 2329–34 (2009).

26. Feng, R. *et al.* PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 6057–62 (2008).

27. Zhao, X. *et al.* Comparative analyses by sequencing of transcriptomes during skeletal muscle development between pig breeds differing in muscle growth rate and fatness. *PLoS One* **6,** e19774 (2011).

28. Shin, J. W. *et al.* Establishment of single-cell screening system for the rapid identification of transcriptional modulators involved in direct cell reprogramming. *Nucleic Acids Res.* **40,** e165 (2012).

29. 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).

30. Flöttmann, M., Scharp, T. & Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.* **3,** 216 (2012).

31. 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).

32. Seki, T. *et al.* Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. *Cell Stem Cell* **7,** 11–4 (2010).

33. Mack, A. A., Kroboth, S., Rajesh, D. & Wang, W. B. Generation of induced pluripotent stem cells from CD34+ cells across blood drawn from multiple donors with non-integrating episomal vectors. *PLoS One* **6,** e27956 (2011).

34. Ono, M. *et al.* Generation of induced pluripotent stem cells from human nasal epithelial cells using a Sendai virus vector. *PLoS One* **7,** e42855 (2012).

35. Miyoshi, K. *et al.* Generation of human induced pluripotent stem cells from oral mucosa. *J. Biosci. Bioeng.* **110,** 345–50 (2010).

36. Chou, B.-K. *et al.* Efficient human iPS cell derivation by a non-integrating plasmid from blood cells with unique epigenetic and gene expression signatures. *Cell Res.* **21,** 518–29 (2011).

37. Bani-Yaghoub, M. *et al.* Role of Sox2 in the development of the mouse neocortex. *Dev. Biol.* **295,** 52–66 (2006).

38. Stolt, C. C. *et al.* The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes Dev.* **17,** 1677–89 (2003).

39. Bouchard, M., Souabni, A., Mandler, M., Neubüser, A. & Busslinger, M. Nephric lineage specification by Pax2 and Pax8. *Genes Dev.* **16,** 2958–70 (2002).

40. Levinson, R. S. *et al.* Foxd1-dependent signals control cellularity in the renal capsule, a structure required for normal renal development. *Development* **132,** 529–39 (2005).

41. Hou, P. *et al.* Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science* **341,** 651–4 (2013).

42. Guan, D., Zhang, W., Liu, G.-H. & Belmonte, J. C. I. Switching cell fate, ncRNAs coming to play. *Cell Death Dis.* **4,** e464 (2013).

43. Warren, L. *et al.* Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. *Cell Stem Cell* (2010). doi:10.1016/j.stem.2010.08.012

44. Suzuki, H. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* **41,** 553–62 (2009).

45. Von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31,** 258–61 (2003).

46. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* **99,** 248–56 (2012).

47. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol.* **11,** 220 (2010).

48. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

49. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14,** R95 (2013).

Acknowledgements

FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y.H.. We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the dataset and thank GeNAS for data production.

This work was supported by Research Grants from MEXT to RIKEN CLST and from MEXT to RIKEN PMI.

The authors would like to thank Boris Lenhard and Nathan Harmston for their kind help in hosting the webpages.

Author Contributions:

O.J.L.R. and J.G. were the major drivers of the novel research directions discussed in this paper. O.J.L.R. performed the data analysis and interpretation, with significant input by J.G. in the early stages of the work. O.L.J.R. prepared the manuscript with J.G. and A.R.R.F. making contributions to its structure and refinement. M.E.O and H.F. provided help and advice with technical aspects of the implementation.

H.S. and J.W.S. were involved in the cell conversion concepts. A.R.R.F. and C.O.D were involved in the FANTOM5 concepts and management.

Online Methods

The basic hypothesis behind Mogrify (outlined in figure 5) comes from the fact that each cell contains the same DNA sequence. This means that the underlying network of potential interactions is the same for each cell. It follows then that the difference between two cell types is defined by which parts of this underlying network are being used.

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

1. Collect expression data for each gene (x) in each sample (s).

2. Calculate the differential expression against a tree-based background for each gene in each sample then combine the log fold change (L_x^s) and adjusted P- value (P_x^s) to a gene score (G_x^s).

3. For each TF ($_x$) in each sample calculate the network score (N^S) by performing a weighted sum of gene scores over two different sub networks (N_{xMARA}^S and $N_{xSTRING}^S$) centered on each TF.

4. Rank TFs based on a combination of G_x^s and N_x^s scores.

5. Calculate the set of transcription factors for a conversion between any two cell types based on comparisons of ranked lists from each cell type.

6. Remove transcriptionally redundant TFs from the lists.

7. Create a cell conversion landscape by arranging the cell types on a 2D plane based on their required TFs and add a height based on the average coverage of the required genes that are directly regulated by the TFs selected.

Step 1: Expression data taken from FANTOM5 dataset.

Mogrify uses 705 libraries (187 tissue libraries and 518 primary cells) of clustered CAGE tags, which provide the TSS locations. These are mapped to their corresponding genes (provided by the FANTOM5 consortium[14]). This data is used to create tag counts for each gene in each library. In total there are 15,878 distinct genes (of which 1408 are TFs) expressed with at least 20 TPM in at least one sample. (See www.Mogrify.net for more details of the libraries analysed).

Step 2: Tree-based differential expression

Calculating differential expression is a common problem when analysing biological data and a number of techniques exist to do this (for a review see[46,47]). We elected to use DESeq[48] for this work as it performs well in benchmark evaluations[49], it allows analysis of some non-replicated datasets and has a short runtime. In order to calculate differential expression, it is necessary to identify two groups; the set of samples you wish to identify differential expression in and the background to compare against. The problem of selecting the correct background is important. Too many irrelevant samples can reduce the statistical power of the test. Too narrow 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: firstly it is very computationally expensive and secondly it does not reveal the genes that are differentially expressed between a sample and an average background, but rather 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 against a collection of samples. In order to do this we implemented a tree-based background selection method based on the FANTOM5 cell ontology[14] (see figure 6). The principle of this approach is to exclude cell types whose ontologies are very close whilst including others that are near in the tree to the background. This was achieved by picking a point near to the top of the tree that would act as the breaking point. Samples in the same clade as the cell type being analysed were removed and those not in the same clade, but still below this point, were included. The result

of this 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.

This tree-based background selection for DEseq is run on all FANTOM 5 libraries (grouped by replicates) creating log-fold changes and FDR adjusted p-values for each gene in each sample. Because there is non-uniform background, the results of each differential expression calculation are not directly comparable, hence for the remaining steps, these figures are used to rank genes in each sample and it is the rankings that are compared.

Since we are only interested in identifying TFs with a high level of influence, we convert the log fold change and FDR adjusted P-values to a single positive score (G_x^s) using the following equation:

$$\text{Eq 1: } G_x^s = |L_x^s|(-\log_{10} P_x^s)$$

where

- L_x^s is the log-fold change of gene x in sample s.
- P_x^s is the adjusted p-value of gene $x$in sample s.

The formula ensures that those genes with high log-fold changes and a low adjusted P-value score very highly and vice versa. This is applied to every gene in each sample creating a 705 sample by 15878 genes matrix of differential expression.

Step 3: Calculate a TF's network-based sphere of influence.

In order to assess the importance of each TF, its effect on its local neighborhood is calculated using two sources of network information: the STRING database[62] and Motif Activity Response Analysis (MARA)[44]. These two techniques, described below, contain different types of interactions.

MARA provides Protein-DNA interactions between TFs 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 contain various types of interactions including PROTEIN-PROTEIN, PROTEIN-DNA, PROTEIN-RNA as well as biological pathways. This provides a view of the interactions that takes place both directly and indirectly affecting gene expression.

In order to calculate the influence, a weighted sum of gene influences (from step 2) is performed over a transcription factors local network neighbourhood. This local network is constrained to a maximum of 3 edges and the effect of each node diminishes the further

from the seed TF it is located and depending on the out degree of its parent. (see figure 7 for a description).

The equation to perform this weighted sum is:

Equation 2: $N^{s}_{x,n} = \sum^{s}_{r \epsilon V_x} G^{s}_{r} \cdot \frac{1}{L_{r,n}} \cdot \frac{1}{O_{r,n}}$

where:

- $x \epsilon V_x$ is each gene (r) in the set of nodes (V_x) that make up the local sub-network of TF x.
- $L_{r,n}$ is the level (or number of steps) r is away from x in the network n.
- $L_{r,n}$ is the degree of the parent of r in the network n.

This is performed over both the MARA and STRING networks resulting in two TF-influence lists ($N^{s}_{x,MARA}$ and $N^{s}_{x,STRING}$).

Step 4: Rank the TFs based on the results of Step 2 and 3.

The result of steps 3 and 4 are three ranked TF lists for each sample based on G^{s}_{x}, $N^{s}_{x,MARA}$ and $N^{s}_{x,STRING}$. To get the final ranking of each TF in each sample, it's rank in each of the three lists is added together. Ranks are limited to a maximum of 100 as we found empirically that after the top 100 TFs the remaining regulatory influence was very small. If a TF doesn't appear in a particular list then it is given a score of 100. The result of this is a single ranked list of TFs for each cell type; those with the lowest score/rank are those predicted to facilitate a cell conversion.

Step 5: Compute all pairwise experiment comparisons to create predictions

In order to predict the set of TFs for a given conversion the ranked lists from the source and target cell type are compared. If a TF from the target cell type list is already expressed in the source target (greater than 20 TPM) then it is removed from the list.

Step 6: Remove transcriptionally redundant TFs.

Once the final ranking is complete, regulatory redundancy is removed. This is achieved by comparing the lists of genes that each of the TFs directly regulates. For a given TF, if there exists a higher-ranking TF that regulates over 98% of the genes that it would regulate, then it is removed. This means that the resulting predictions include TFs that are diverse in their regulatory sphere of influence.

Step 7: Create a cell reprogramming landscape based on steps 1-6.

The reprogramming landscape is produced in two phases. Firstly the x-y coordinate of a cell type is defined by clustering expression profiles in 2D space using a multi dimensional scaling algorithm. This results in cell types with similar average gene expression profiles clustering close to each other within the x-y plane.

The z value (height) for each cell type is the calculated using number of required genes that are regulated by the set of TFs predicted for each conversion. This means that cell types that have many genes that have no known regulatory control will be lower in the landscape as they represent the cell types that have the least confidence of being reached.

Experimental design for screening experiment

Fibroblast cells were transduced with viral vectors containing 18 monocyte-enriched transcriptional modulators. The stochastic nature of viral infection ensured that different sets of TFs entered each cell. The resulting cells were cultured before being FACS sorted using two monocyte specific cell surface markers (CD14 and HLA-DR). Following this, a nested-single-cell-polymerase chain reaction was used to identify the sets of TFs that were acquired by each cell, as well as the presence/absence of a set of 17 monocyte marker genes.