Bioinformatics, 2017, 1–7
doi: 10.1093/bioinformatics/btx205
Advance Access Publication Date: 7 April 2017
Original Paper

OXFORD

Systems biology

Systemic evaluation of cellular reprogramming processes exploiting a novel R-tool: *eegc*

Xiaoyuan Zhou[1,2], Guofeng Meng[2], Christine Nardini[1,3,4,*] and Hongkang Mei[2,*]

[1]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China and University of Chinese Academy of Sciences, Beijing, China, [2]Computational and Modeling Sciences, Platform Technologies and Science China, Shanghai, GSK, [3]CNR IAC "Mauro Picone", Via dei Taurini 19, Roma, Italy and [4]Personal Genomics S.r.l, Strada Le Grazie 15, Verona, Italy

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 15, 2016; revised on March 31, 2017; editorial decision on April 3, 2017; accepted on April 5, 2017

Abstract

Motivation: Cells derived by cellular engineering, i.e. differentiation of induced pluripotent stem cells and direct lineage reprogramming, carry a tremendous potential for medical applications and in particular for regenerative therapies. These approaches consist in the definition of lineage-specific experimental protocols that, by manipulation of a limited number of biological cues—niche mimicking factors, (in)activation of transcription factors, to name a few—enforce the final expression of cell-specific (marker) molecules. To date, given the intricate complexity of biological pathways, these approaches still present imperfect reprogramming fidelity, with uncertain consequences on the functional properties of the resulting cells.

Results: We propose a novel tool *eegc* to evaluate cellular engineering processes, in a *systemic* rather than marker-based fashion, by integrating transcriptome profiling and functional analysis. Our method clusters genes into categories representing different states of (trans)differentiation and further performs functional and gene regulatory network analyses for each of the categories of the engineered cells, thus offering practical indications on the potential lack of the reprogramming protocol.

Availability and Implementation: *eegc* R package is released under the GNU General Public License within the Bioconductor project, freely available at https://bioconductor.org/packages/eegc/.

Contact: christine.nardini.rsrc@gmail.com or hongkang.k.mei@gsk.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the generation of induced pluripotent stem cells (iPSCs) originally described by Takahashi and Yamanaka (2006), numerous functional cell types (from epithelial to cardiac to central nervous system cells) can be obtained by engineered differentiation processes (Kamao *et al.*, 2014; Lian *et al.*, 2013; Shi *et al.*, 2012). The rapid development of this technology has been paralleled by research on lineage conversion (a.k.a. cell-reprogramming) to achieve the conversion of one cell type (*original* somatic cell) into another (*induced* cell) mimicking a different *target* primary cell. This was pioneered by the identification by Davis and colleagues of MyoD, the transcription factor (TF) capable to drive cell conversion from fibroblast into myoblast (Davis *et al.*, 1987), and further expanded to other cell lineage conversions with overexpression or ablation of lineage-specific TFs (Graf and Enver 2009; Xu *et al.* 2015). The enforced expression of TFs has been used to drive cell fate conversion either by direct differentiation from iPSCs or by conversion between cell lineages, based on the down-regulation of the original cell genes' expression and up-regulation of target cell gene expression (Ieda *et al.*, 2010; Xie *et al.*, 2004). Still, incomplete reprogramming

remains an issue, owing to the persistence of genes of the original cell or by the silencing of cell-specific genes in the target cells, ultimately leading to immature induced cells (Feng *et al.*, 2008; Marro *et al.*, 2011; Morris and Daley 2013).

This still challenging issue motivates computational biologists to work mostly at the improvement of two stages of the reprogramming: (i) better selection of the targets to manipulate and (ii) identification of the causes for failure, to modify the input of the reprograming protocols. As broadly across biology, latest approaches explore solutions that move from a reductionist to a systemic focus.

For the first stage, after the early manual selection of TFs, Mogrify has been developed from high-throughput experiments to predict, from 173 human cell types and 134 tissues, the best candidate TFs driving cell fate conversion based on gene expression data and regulatory network information (Rackham *et al.*, 2016).

Similarly, for the second stage, after the initial comparison of induced versus target primary cells transcriptional profiles, with a focus on marker genes—i.e. genes that are highly expressed either in original somatic or in the target primary cells (Sandler *et al.*, 2014; Szabo *et al.*, 2010)—attention has been turned to the more extended and large plethora of genes interconnected and downstream of such TFs. CellNet (Morris *et al.*, 2014) has been developed to enhance the estimate of successful cell fate conversion, by analyzing transcriptomics based on gene regulatory network (GRN). However, CellNet allows input limited to microarray data, quickly being overwhelmed by more precise RNA-seq data.

Along these lines, we here offer our contribution to the analysis of incomplete reprogramming: our approach is designed to analyze gene expression data (from any platform: microarray and RNA-seq), to isolate the genes that undergo significant expression changes in the (trans)differentiation process, to then categorize them into progressive stages of the reprogramming process. Further, it exploits networks and functional analyses for the evaluation of the success of the process.

The different stages are designed based on an intuitive dynamic progression of states from *inactive* (no change from the original cell), to *insufficient* (partial desired (in)activation) to final *successful* (in)activation, with the addition of two extreme situations, *reverse* (expression opposite to the expected one) and *over* (beyond the expected levels of (in)activation). With this characterization and further exploration of each of these five classes by functional annotation and systemic GRN analysis, our approach evaluates in a systemic fashion (affected biological functions and pathways, but also topologically relevant TFs) the impact of imperfect expression values (*insufficient*, *inactive*, *over* and *reverse*) and suggests potential molecules to be manipulated by the reprogramming process, in consideration of the usability (functionality) of the induced cells.

2 Materials and methods

2.1 Data

The design of our tool was motivated by the evaluation of the lineage conversion protocol published by Sandler *et al.* (2014), where dermal microvascular endothelial cells (DMEC, *original* cells) were reprogrammed to hematopoietic cells with multipotent progenitor activity (rEC-hMPP, *induced* cells) via the induction of TFs (FOSB, GFI1, RUNX1 and SPI1, globally referred to with the acronym FGRS) and a phenocopy of microenviromental niches, to finally mimic purified $Lin^{-}CD34^{+}$ cord blood cells (CB, *target* cells). Transcriptomic profiles of the three types of cells (DMEC, rEC-hMPP and CB) screened by RNA-sequencing and quantified in FPKM (Fragments Per Kilobase of transcript per Million mapped reads) were downloaded from GEO with accession number GSE57662. Genes expressed in less than 40% samples were filtered out and FPKMs were log2 transformed after adding a pseudo-value of 2 to avoid infinite values.

2.2 Differential gene identification and categorization by *eegc*

Differentially expressed genes (DEGs) are computed in each pairwise comparison between the original, induced and target cells with *limma* R package (Smyth, 2004). Significance is defined by fold change ≥ 2 and false discovery rate (FDR) ≤ 0.01 to correct for multiple hypothesis testing within each list (*omic* data). This choice is adequate to also control the overall error rate descending from testing three genes lists (see Supplementary Material for details). DEGs are categorized into five categories and namely: *Inactive*, *Insufficient* and *Successful*, representing the genes unchanged, insufficiently changed and successfully modified in the reprogramming process, respectively. In addition, two more categories are defined: *Reverse*, indicating the genes differentially expressed in a direction opposite to the expected one, and *Over* for genes that are overly expressed in the induced cells in comparison to the target cells. To formally define these categories, we exploit the patterns that are differential across the three comparisons (Table 1), with the definition of the *expression difference* (*ED*) as the difference of the gene expression in each comparison and the *ED* ratio as the ratio of *EDs* between two arms. The necessity of five classes is motivated by the observation that *Inactive* and *Successful ED* ratios are, conveniently, centered around 0 and 1, however, they cover a relatively wide range of values, with queues overlapping with the *Over* and *Insufficient* categories for *Successful* genes, and with *Reverse* and *Insufficient* for *Inactive* genes (see also Results in Fig. 2 for a graphical output). To gain an accurate and practical categorization allowing to highlight the genes that need attention in the engineering process, *Inactive* and *Successful* genes boundaries were set more stringently around the intuitive peaks of 0 and 1, by shrinking the *ED* ratio boundaries to the 5th and 95th quantile of the ED-ranked *Successful* and *Inactive* genes (named *operational ranges*). In each category, further analysis is done by separating up- from down-regulated genes, leading to 10 genes categories.

2.3 GO and KEGG functional annotation by *eegc*

Functional annotation is performed by embedding in *eegc* the R package *clusterProfiler* (Yu *et al.*, 2012) with functional enrichment analyses on gene ontology (GO) (Ashburner *et al.*, 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012).

2.4 Cell/tissue-specific analysis by *eegc*

We exploited for this the Gene Enrichment Profiler database (Benita *et al.*, 2010) containing expression profiles of ~12 000 NCBI GeneID entries across 126 primary human cells/tissues (C/Ts) clustered in 30 groups. The database provides custom enrichment scores for the genes in the 126 C/Ts. As a consequence, genes are attributed, with different enrichment scores, to more than one C/T. Thus we applied the *SpeCond* R package (Cavalli, 2009) to identify genes specific unique to the 126 C/Ts. The statistical significance of the tissue specificity was assessed by hypergeometric tests.

Table 1. Gene categorization

Category	*ED* patterns			*ED* ratio	Operational ranges of *ED* ratio	Regulation pattern from original cell to target cell	
	Target, original (*CB, DMEC*)	Induced, original (*rEChMPP, DMEC*)	Induced, target (*rEChMPP, CB*)				
Reverse	√/×	√	√	<0		*Up* *Down*	*Reverse.Up* *Reverse.Down*
Over				>1		*Up* *Down*	*Over.Up* *Over.Down*
Inactive	√	×	√	∼0	(Q^{5th} *ED* ratio, Q^{95th} *ED* ratio) (−0.39, 0.50)	*Up* *Down*	*Inactive.Up* *Inactive.Down*
Insufficient	√	√	√	0∼1		*Up* *Down*	*Insufficient.Up* *Insufficient.Down*
Successful	√	√	×	∼1	(Q^{5th} *ED* ratio, Q^{95th} ED ratio) (0.28, 1.31)	*Up* *Down*	*Successful.Up* *Successful.Down*

Results of the pair-wise comparisons among original (DMEC), induced (rEChMPP) and target (CB) cells were classified into five categories named *Reverse*, *Over*, *Inactive*, *Insufficient* and *Successful* based on *ED* patterns and *ED* ratios. Each category is separated into *Up* and *Down* (expression variation).

Note: √ represents differential, × represents non-differential states identified by *limma* gene expression differential analysis. Values in italics and parenthesis indicate the specific boundaries values in our exemplar analysis (see Results).

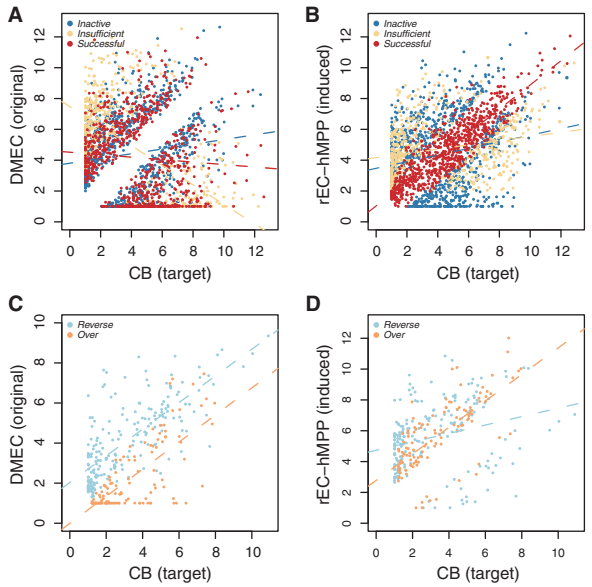

Fig. 1. Expression profile (FPKM in log2 scale) of (**A**, **C**) the original endothelial cells (DMEC) versus CB target cells and of (**B**, **D**) the induced rEC-hMPPs versus CB, Each gene category is fitted to a linear model. *Successful* genes changed the Pearson's correlation from −0.071 between DEMC and CB to 0.898 between rEC-hMPP and CB, while *Inactive* genes showed virtually no change of correlation from 0.140 to 0.208, *Insufficient* genes went from −0.626 to 0.233. The *Reverse* genes reduced the correlation from 0.741 (between DMEC and CB) to 0.275 (between rEC-hMPP to CB) and *Over* genes presented a slightly change in correlation from 0.723 to 0.708

2.5 GRN based evaluation by *eegc*

From each of the 16 networks defined in CellNet we isolated the C/T-specific TFs and their corresponding down-stream targets (TGs, defined in CellNet in 16 C/T-specific GRNs), into 1455 TF-TG gene sets. The 16 C/T-specific gene sets and the 1455 TF-TG gene sets were used to generate two types of enrichment analyses: one gene-based and one TF-based. The first uses directly the 16 C/T gene sets, to offer an enrichment analysis complementary to the former one (Section 2.4, based on coherent expression levels) including regulatory elements that may not share the same expression profile. The second selects, within each of the 16 C/T-specific TF-TG sets, only the TFs (hereafter *relevant* TFs) with: (i) highly significant enrichment for their C/T (FDR ≤ 0.01) and (ii) top (50) betweenness centrality (Koschutzki and Schreiber, 2008) computed by the *igraph* R package (Csaridi and Nepusz, 2006).

2.6 Confirmatory DNA methylation analysis

Based on the interpretation of the results of *eegc* on Sandler *et al.* dataset, an additional, custom, analysis was run to explore the potential epigenetic causes of the observed cell lineage conversion. DNA methylation data for hematopoietic progenitor cells in cord blood (CD34+HPCs), taken as proxies for the target CD cells, were downloaded from ArrayExpress (www.ebi.ac.uk/arrayexpress) with accession no. E-MTAB-487, and quantile normalized Beta values ranging from 0 (unmethylated) to 1 (completely methylated) (Bocker *et al.*, 2011). The average Beta value was calculated among the seven HPC samples for each of the 27578 CpG dinucleotides.

Coherence between gene expression and methylation was tested for each of the 10 gene categories according to the finding that, during differentiation, hypomethylation positively correlates with gene over-expression (Han *et al.*, 2012; 2014). Hypomethylation and hypermethylation were defined as Beta value ranging from 0 to 0.2 and 0.8 to 1, respectively (Du *et al.*, 2010). Special attention was given to the 65 vascular and hematopoietic specific genes reported in Sandler's paper (Sandler *et al.*, 2014 *marker* genes) and to the *relevant* TFs defined in Section 2.5.

3 Results and discussion

3.1 Characterization by gene categories

2770, 3645 and 2003 differentially expressed genes were identified in the rEC-hMPP to DMEC, CB to DMEC and rEC-hMPP to CB comparisons, respectively, and classified into the *Inactive*, *Insufficient*, *Successful* categories in Figure 1A, B and *Reverse*, *Over* categories in Figure 1C, D.

Lineage conversion can be thought of, in general, as a progressive change of expression from original- to target-specific genes and, for this conversion in particular, as a change towards the up-

regulation of hematopoietic genes (target cells specific) and down-regulation of endothelial cells (original cells specific).

Based on the expression profiles of the vascular and hematopoietic *marker* genes only (hierarchical clustering, Pearson's correlation, Sandler *et al.*, 2014), it can be concluded that rEC-hMPPs are closer to the CBs target cell than the DMEC original cell. However, a different perspective, based on Pearson's correlation measurement of all differential genes (across the five categories), helps understanding the imperfect final result of this experiment, as it highlights that not all the *marker* genes are successfully induced (Supplementary Table S1), in particular, some hematopoietic (TEK, SOX17, ECE1, ENG) and vascular (JUNB, KLF2) markers did not reach the expected expression level (*Insufficient*), and other were dis-regulated, such as ETS1 and F13A1 in the *Over* and *Reversed* categories.

To offer additional insight into the success of the cellular engineering process, we calculated and compared the proportions of genes in each category among all the categorized genes with the assumption that a high proportion of *Successful* genes would reflect a better (trans)differentiation. Results show that the *Inactive* genes dominate (Fig. 2), indicating an incomplete conversion, supported by the results of the additional assays made by the authors, both *in vitro* and *in vivo*, confirming that the obtained rEC-hMPP cells could further effectively differ into myeloid (erythrocytes, megakaryocytes, monocytes, macrophages) and lymphoid lineages (B cells, nature killer cells), but only negligibly into the T-lymphoid progeny, representing an important limitation with respect to the properties of the target cells.

3.2 Functional evaluation of cellular programming

The limited fidelity of the reprogramming in the ability to convert into the T-cell progeny is confirmed by the GO functional enrichment analyses performed with *eegc*. Indeed, the significant GO terms enriched by the 10 major categories are grouped into 6 clusters (Fig. 3, Supplementary Table S2). Among those, cluster 1–3 refer to endothelial cell related functions enriched by down-regulated genes, while cluster 4–6 explicitly refer to terms related with T-cell differentiation, with up-regulated *Successful*, *Inactive* and *Insufficient* genes being enriched for these categories. This suggests that not all of the genes contributing to these functions (GO categories) are (sufficiently) activated (owing to many *Inactive* and *Insufficient* genes). Interestingly, none of the Successfully down-regulated genes is enriched in T-cell differentiation related GO terms, suggesting the lack of a mechanism to silence the biological cues that impede T-cell differentiation.

These results are further confirmed in the sister KEGG pathways analysis (Supplementary Fig. S1 and Supplementary Table S3), indicating that up-regulated genes, mostly in the *Inactive* category, are enriched in immune related pathways and particularly in T-cell related pathways, such as hsa04660: T cell receptor (TCR) signaling pathway (Supplementary Fig. S2), and hsa04064: NF-κB signaling pathway, in line with the important role of TCR signaling in T cell lineage development from lymphoid precursors, T cell activation under antigen stimulations and trigger of the downstream NF-κB signaling in a TCR-to-NF-κB cascade, also involved in the differentiation of T cells (Berg, 2012; Suman Paul, 2013). Again, down-regulated genes are enriched in hsa04015: RAP1 signaling pathway that controls cell–cell and cell-matrix interactions (Supplementary Fig. S2) confirming the absence of mechanisms involving gene silencing (down-regulation) of T-cells differentiation.

Besides these, we also noticed that *Reversed.Down* and *Over.Up* genes were specifically enriched in cell development or morphogenesis related GO terms (Supplementary Table S2) and KEGG pathways (Supplementary Table S3). By this observation, we can speculate that the genes over or reversely up-regulated in the induced cells are incline to regulate cell growth and maintain homeostasis despite the forced mechanisms induced by reprogramming, creating biomolecular resilience to the expected lineage conversion.

3.3 Tissue specific analysis of each gene category

To deepen these observations we exploited the 126 C/T specific gene sets provided in *eegc* to run enrichment analysis.

In particular, we selected 10 C/T groups, related to *Hematopoietic* thus representing the induced cells: 'stem cells', 'Myeloid', 'B cells' and 'T cells'; and to *Endothelial* thus representing the original cells: 'Endothelial CD105+', 'Lung', 'Kidney', 'Thyroid', 'Heart' and

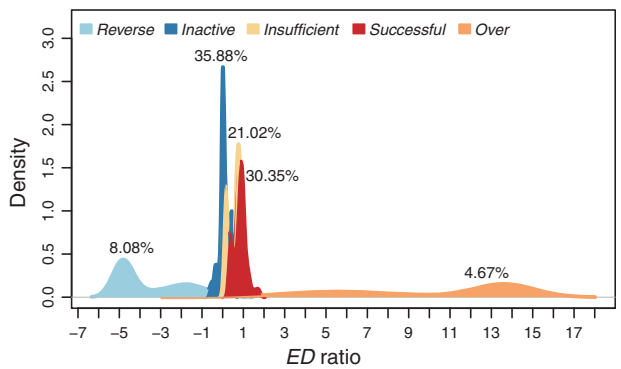

Fig. 2. The proportion of genes in each category among genes in all represented by *ED* ratios. Extreme lower or higher ratios given by the *Reserved* or *Over* genes are narrowed to their median values, respectively, to make the ratios on *x*-axis readable

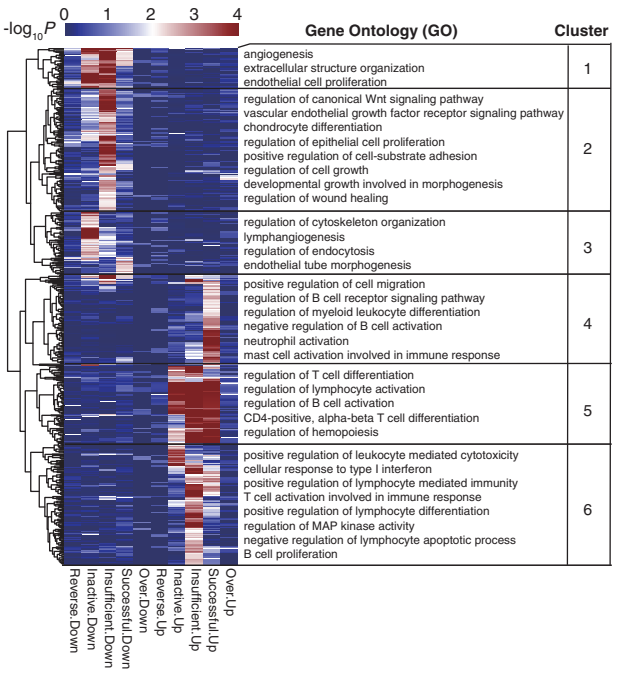

Fig. 3. Functional enrichment. Representative GO terms significantly enriched by 10 gene categories based on a log transformed corrected *P*-value (FDR $\leq$ 0.01, manually removing redundant functional terms). FDRs lower than 10^{-4} were adjusted to 10^{-4} for a better visualization in heatmaps

'Uterus' (Fig. 4A, Supplementary Fig. S3 and Supplementary Table S4). The largest part of the differentially up-regulated genes belongs to the hematopoietic group, and, in confirmation of the previous functional analysis, *Successful.Up, Insufficient.Up* and *Inactive.Up* are included, with only a small set of *Over.Up*. Not surprisingly, T-cells enrichment involves mainly the *Inactive.Up* category. Finally, and again, the *Successful.Down* category includes typically endothelial genes, confirming that genes down-regulation for successful reprogramming involves functions associated to the cells of origin, or, symmetrically, that down-regulation in the reprogrammed cells is not easily achieved on genes that are typically hematopoietic.

3.4 Gene regulatory analysis

Similarly to the gene-based tissue specific analysis (Sections 2.4 and 3.3) the results of the network-based tissue specific enrichment (Section 2.5) confirm that genes in the *Down* categories are mostly specific to the original endothelial cell, while genes in the *Up* categories refer mostly to the induced rEC-hMPPs derived blood cells, including T and B cells and macrophages, and particularly, *Insufficient* and *Inactive* up-regulated genes are significantly enriched in T cell (Fig. 4B, C, Supplementary Table S5).

The TF-TG enrichment analysis shows that significantly enriched TFs are clustered into two groups (*Down* and *Up* gene categories) and specific to endothelial cells and hematopoietic cells, shown in Figure 4D for the *Successful*, *Insufficient*, *Inactive* and *Reverse* genes and Supplementary Table S6 for the remaining ones. Coherently with previous findings, none of T cell-specific TFs is *Successful*, although CellNet, lists the same TFs for B and T cells, limiting the resolution of the suggestions for the reprogramming process improvement. This clarifies once more the need for multiple enrichment approaches as they are offered in *eegc*.

As shown in Figure 4D, the successfully over- or under-expressed genes are regulated by more *Successful* TFs (red labeled) while the *Inactive* or *Insufficient* genes failed to be properly expressed because they are largely regulated by *Inactive* TFs (dark blue labeled).

At a closer look, TFs SPI1, RUNX1, FOSB and GFI1, whose transcription were enforced to trigger the reprogramming process, are as expected in the *Successful.Up* (SPI1, RUNX1, and FOSB), despite some of their TGs falling into the *Insufficient.Up* category (Supplementary Table S6), reflecting a relatively successful regulation by these TFs during reprogramming. However, GFI1 shows a large variance of expression across samples and thus was not selected in the differential analysis for categorization. In this case, most of its TGs were in the *Inactive.Down* and *Inactive.Up* genes, which represents a possible relevant cause of the limits of the reprogramming process.

Besides the FGRS, we noted that *Inactive* genes were enriched in the TF-TG sets, raising the possibility to discover alternative/additional TFs for reprogramming improvement. The endothelial and T cell-specific TFs (relevant for the analysis of this specific dataset) are significantly present in the categories *Inactive.Down* (Supplementary Table S7) and *Inactive.Up* (Supplementary Table S8), respectively.

For the endothelial TF-TG set *Inactive.Down* genes are enriched in SOX17-TGs, which means these genes should be down-regulated by SOX17 while they are not. Not surprisingly SOX17 itself is categorized as *Inactive.Down*. Functionally, SOX17, if properly down-regulated, would allow endothelial-to-hematopoietic transition (EHT) in synergy with RUNX1 (Lizama *et al.*, 2015).

In the T-cell TF-TG set, we observe the failure to activate: (i) BCL11B, critical for T lymphocytes survival and early T-cell development via Notch signaling pathway (Liu *et al.*, 2010; RothenBerg, 2012); (ii) TCF7 a direct Notch target required from the Early T-cell precursor stage (RothenBerg, 2012) and (iii) TBX21 which directs T-helper1 (Th1) lineage commitment (Michael and

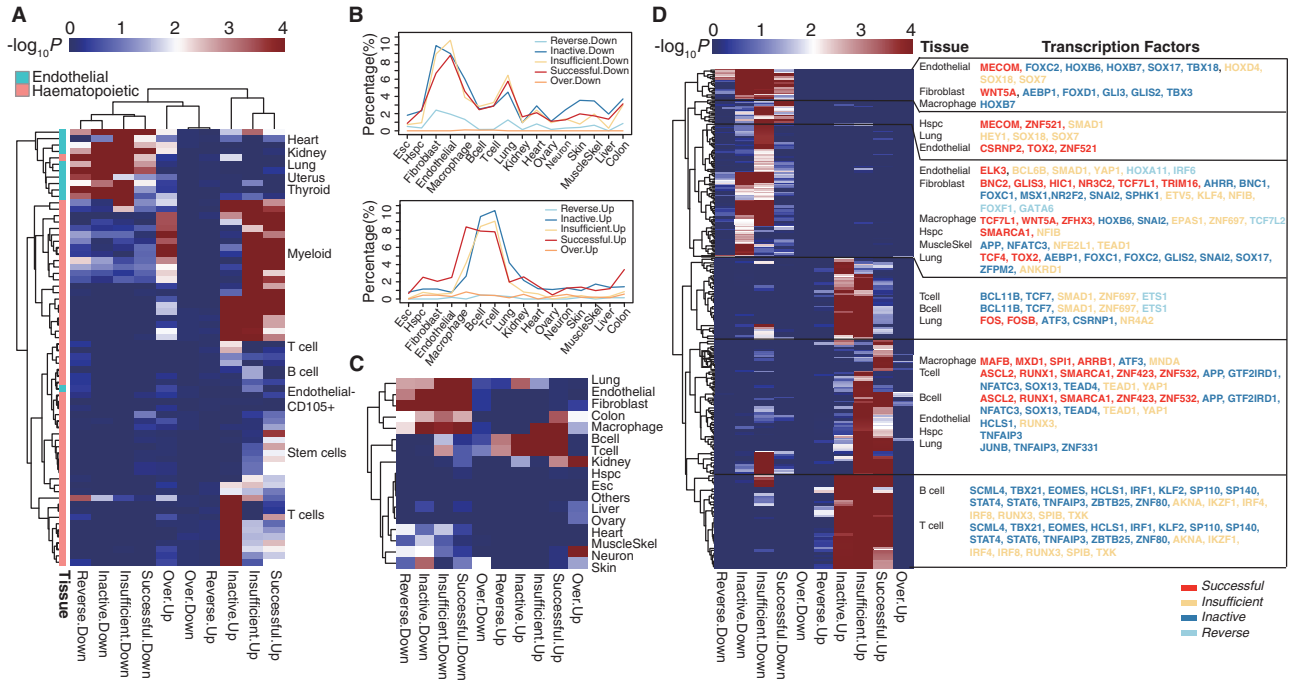

Fig. 4. Cell/tissue specificity and TF enrichment by the ten gene categories. (**A**) Enrichment in endothelial- and hematopoietic-specific cell and tissue gene sets identified from the Gene Enrichment Profiler database. (**B**) Percentages in 16 CellNet cell/tissue-specific gene sets. (**C**) Enrichment in 16 CellNet cell/tissue-specific gene sets. (**D**) Enrichment in 1455 CellNet TF-TG sets. TFs were selected by a FDR ≤ 0.01 and categorized in *Successful, Insufficient, Inactive* and *Reverse* categories with color labeled

Michael, 2000; Szabo *et al.*, 2015). Also an *Inactive.Up* gene KLF2, whose absence would make T cell prone to apoptosis (Pearson *et al.*, 2008), contributes to the inability to obtain viable T-cells.

The two TFs TBX21 and TCF7, whose inactivation potentially explains the functional immaturity of the induced cell, were also predicted by Mogrify (Rackham *et al.*, 2016) among the TFs able to drive cell conversion from microvascular endothelial cells to blood cells, corroborating *eegc* suggestion to focus or include these genes in the reprogramming protocol.

3.5 Coherence between expression and DNA methylation

The results produced by *eegc* allow to confirm that several types of anomalies occurring in the process involve not only TFs, but also, importantly, genes assumed to be markers of the target cells, impacting on all four 'imperfect' categories (*Reverse, Inactive*, *Insufficient* and *Over*). There are overall two recurring messages emerging from these analyses across the four results provided by *eegc*: (i) successfully up-regulated genes are not backed by a sufficient number of genes involved in the completion of the same functions (presence of *Insufficient*, *Inactive* genes) and (ii) successfully down-regulated genes are markers of pertain to functions that are typical of the original cell, not of the target cell. The latter seems to suggest that while the reprogramming process permits to preserve the original functions that are needed also in the target cell, it cannot perform the silencing of functions needed to allow complete reprogramming, as it is the case for T-cells differentiation.

As other studies have described lineage conversion achieved by the combination of TFs with epigenetic regulators such as chromatin modifiers (Takeuchi and Bruneau, 2009) or by a deficiency of DNA methyltransferase Dnmt1 in mice (Dhawan *et al.*, 2011), we explored this additional layer of information to shed light on the connections with the epigenomic cellular makeup in order to give workable directions to experimentalists designing the protocol.

In particular, literature supports the negative correlation between a gene's expression and the methylation of its promoters, and symmetrically between demethylation and higher expression (Jones, 2012). In particular, during cellular reprogramming several studies report the coherence between hypomethylation and overexpression of the target cell-specific TFs or of marker genes (Han *et al.*, 2012; 2014).

Supplementary Table S9 shows indeed that cell-specific TFs and markers genes in the *Successful.Up* category are perfectly in line with this expectation: all expectedly hypomethylated genes are also overexpressed. Conversely, the reverse is not true (*Successful.Down* genes are not hypermethylated). In our results, successfully down-regulated genes pertain in general to functions associated to the cell of origin, i.e. they are not functionally involved in the reprogramming process.

Focusing on the genes that failed to be (in)activated, we observe that they have a coherent methylation state with gene expression in the target cell (Supplementary Table S10) and that they include T cell-specific *relevant* TFs and *marker* genes. Thus, incoherence between expected and observed methylation-expression patterns is an indicator of 'distance' from the target cell. In particular, the *Inactive.Up* BCL11B, TCF7 and TBX21 *relevant* TFs, participating the T cell development, are hypomethylated in cord blood samples but fail to be overexpressed in the induced cells; the same holds for the *Inactive.Up* marker gene KLF2 specifically expressed in hematopoietic progenitors. Coherently, RUNX1 showed perfectly matched gene expression and methylation and was successfully up-regulated as a necessary TF to promote EHT.

Experimental validation is required to know whether this depends on the ability of the engineering process to mimic hypomethylation by acting on TF overexpression or reversely to mimic hypermethylation to control gene down-regulation.

As a final observation we specifically searched for the activity of methyltransferases, responsible for maintenance and *de novo* DNA methylation in human (Bestor, 2000). DNA-methyltransferase 1 (encoded by gene DNMT1) and DNA-methyltransferase 3 alpha and beta (encoded by genes DNMT3A and DNMT3B, respectively) did not show significant differences between the original DMEC cells and the target CB cells (Supplementary Table S11), and hence could not be classified into any of the categories in Table 1. However, we observed an expression increase of DNMT1 and DNMT3B (fold change >2, Supplementary Table S11) in the induced cells compared to original cells, which indicates that changes of methylation patterns are indeed elicited by the reprogramming protocol.

Overall we confirm, as Sandlers *et al.* observed experimentally, that the niche environment and the four TFs were not sufficient to make a complete reprogramming allowing progenitors to differentiate into mature blood cells. Incrementally, the results of our package recommend a list of candidate TFs (among which BCL11B, TBX21, TCF7, KLF2 in Supplementary Table S10) whose selection is driven by a mixture of heterogeneous criteria, to guarantee that further improvement of cellular engineering protocol take into better account the complex interplay among transcriptional actors.

The results of *eegc* also suggest, for their systemic nature, to move beyond the transcriptional level and our final investigation into the epigenomic layer confirms that mixed techniques including not only TFs forced overexpression, but also induction of hypomethylation has to be taken in consideration.

Acknowledgements

We would particularly thank Lian Liu for the statistical guidance; thank Kang An, Han Xu, Ming-hsun Ho, Yanyan Zhang, Yuan Li, Xiaotao Wang and Jianfei Wang in GSK for valuable discussion on the method development.

Funding

This work was supported by National Natural Science Foundation of China (NSFC) n.31171277.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Benita,Y. *et al.* (2010) Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. *Blood*, **115**, 5376–5384.

Berg,L.J. (2012) Signaling pathways that regulate T cell development and differentiation. *J. Immunol.*, **189**, 5487–5488.

Bestor,T.H. (2000) The DNA methyltransferases of mammals. *Hum. Mol. Genet.*, **9**, 2395–2402.

Bocker,M.T. *et al.* (2011) Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood*, **117**, E182–E189.

Cavalli,F. (2009) SpeCond: condition specific detection from expression data. R package version 1.28.0.

Csaridi,G. and Nepusz,T. (2006) The igraph software package for complex network research, InterJournal, Complex Systems, 1695.

Davis,R.L. *et al*. (1987) Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell*, **51**, 987–1000.

Dhawan,S. *et al*. (2011) Pancreatic beta cell identity is maintained by DNA methylation-mediated repression of Arx. *Dev. Cell*, **20**, 419–429.

Du,P. *et al*. (2010) Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Feng,R. *et al*. (2008) PUA and C/EBP alpha/beta convert fibroblasts into macrophage-like cells. *Proc. Natl. Acad. Sci. USA*, **105**, 6057–6062.

Graf,T. and Enver,T. (2009) Forcing cells to change lineages. *Nature*, **462**, 587–594.

Han,D.W. *et al*. (2012) Direct reprogramming of fibroblasts into neural stem cells by defined factors. *Cell Stem Cell*, **10**, 465–472.

Han,J.K. *et al*. (2014) Direct conversion of adult skin fibroblasts to endothelial cells by defined factors. *Circulation*, **130**, 1168–1178.

Ieda,M. *et al*. (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, **142**, 375–386.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet*., **13**, 484–492.

Kamao,H. *et al*. (2014) Characterization of human induced pluripotent stem cell-derived retinal pigment epithelium cell sheets aiming for clinical application. *Stem Cell Rep*., **2**, 205–218.

Kanehisa,M. *et al*. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*., **40**, D109–D114.

Koschutzki,D., and Schreiber,F. (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul. Syst. Bio*., **2**, 193–201.

Lian,X. *et al*. (2013) Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nat. Protoc*., **8**, 162–175.

Liu,P.T. *et al*. (2010) Critical roles of Bcl11b in T-cell development and maintenance of T-cell identity. *Immunol. Rev*., **238**, 138–149.

Lizama,C.O. *et al*. (2015) Repression of arterial genes in hemogenic endothelium is sufficient for haematopoietic fate acquisition. *Nat. Commun*., **6**, 7739.

Marro,S. *et al*. (2011) Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. *Cell Stem Cell*, **9**, 374–382.

Michael,L.D. and Michael,J.B. (2000) Notch signaling in T cell development. *Curr. Opin. Immunol*., **12**, 166–172.

Morris,S.A. and Daley,G.Q. (2013) A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Res*., **23**, 33–48.

Morris,S.A. *et al*. (2014) Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell*, **158**, 889–902.

Pearson,R. *et al*. (2008) Kruppel-like transcription factors: a functional family. *Int. J. Biochem. Cell Biol*., **40**, 1996–2001.

Rackham,O.J.L. *et al*. (2016) A predictive computational framework for direct reprogramming between human cell types. *Nat. Genet*., **48**, 331–335.

Rothenberg,E.V. (2012) Transcriptional drivers of the T-cell lineage program. *Curr. Opin. Immunol*., **24**, 132–138.

Sandler,V.M. *et al*. (2014) Reprogramming human endothelial cells to haematopoietic cells requires vascular induction. *Nature*, **511**, 312–318.

Shi,Y. *et al*. (2012) Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat. Protoc*., **7**, 1836–1846.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, Article3.

Suman Paul,B.C.S. (2013) A new look at TCR signaling to NF-κB. *Trends Immunol*., **34**, 269–281.

Szabo,E. *et al*. (2010) Direct conversion of human fibroblasts to multilineage blood progenitors. *Nature*, **468**, 521. U191.

Szabo,S.J. *et al*. (2015) A novel transcription factor, T-bet, directs Th1 lineage commitment. *J. Immunol*., **194**, 2961–2975.

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.

Takeuchi,J.K. and Bruneau,B.G. (2009) Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. *Nature*, **459**, 708–U112.

Xie,H.F. *et al*. (2004) Stepwise reprogramming of B cells into macrophages. *Cell*, **117**, 663–676.

Xu,J. *et al*. (2015) Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell*, **16**, 119–134.

Yu,G. *et al*. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *omics*, **16**, 284–287.