A general strategy for cellular reprogramming: the importance of transcription factor cross-repression

Isaac Crespo[1] and Antonio del Sol[1,*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-Belval, University of Luxembourg, L-1511 Luxembourg, Luxembourg
[*]Corresponding author: Antonio del Sol (antonio.delsol@uni.lu)

Keywords: Cellular reprogramming, differentiation, dedifferentiation, transdifferentiation, network stability, cross repression, cross-antagonistic motif, retroactivity, positive circuit.

Author contributions

I. C. and A. dS. conceived the idea for the paper. I. C. wrote software, performed the experiments and analyzed the data. I. C. and A. dS. contributed to writing the paper. A. dS. coordinated and supervised the project.

Abstract

Transcription factor cross-repression is an important concept in cellular differentiation. A bistable toggle switch constitutes a molecular mechanism that determines cellular commitment and provides stability to transcriptional programs of binary cell fate choices. Experiments support that perturbations of these toggle switches can interconvert these binary cell fate choices, suggesting potential reprogramming strategies. However, more complex types of cellular transitions could involve perturbations of combinations of different types of multistable motifs. Here we introduce a method that generalizes the concept of transcription factor cross-repression to systematically predict sets of genes, whose perturbations induce cellular transitions between any given pair of cell types. Furthermore, to our knowledge, this is the first method that systematically makes these predictions without prior knowledge of potential candidate genes and pathways involved, providing guidance on systems where little is known. Given the increasing interest of cellular reprogramming in medicine and basic research, our method represents a useful computational methodology to assist researchers in the field in designing experimental strategies.

Introduction

The central role of transcription factor cross-repression determining cell fate is one of the most important concepts emerged from years of lineage differentiation research[1-4]. In its simplest formulation, two regulators that negatively influence each other establish a bistable “toggle switch”, readily explaining the two mutual exclusive cell fate outcomes. More complicated schemes also include transcription factors auto-regulation and

antagonistic cross-regulation of target genes. Several examples of these binary cell fate choice mechanisms have emerged in the last ten years[5-14]. Integration of this knowledge can be represented in a binary decision tree from embryonic stem cells (ES cells) to differentiated cells passing by different progenitors[1] (see figure 1). This tree defines distinct paths between different cell types in a Waddington's landscape[15-17], where different cell types can be interpreted as steady stable states of cellular gene regulatory networks termed as attractors. Cross-repression motifs not only determine binary decisions in the tree, but based on their bistable behavior, characterized by mutually exclusive gene expression states; they also play a key role in the stability of each possible cell fate. Furthermore, experimental evidences have demonstrated that perturbations of genes belonging to these motifs are able to trigger transitions between these binary cell fate choices[18,19]. Indeed, although attractor's stability is determined by a regulatory core comprised of one or several interconnected positive feedback loops, known as positive circuits[20], these cross-antagonistic motifs are shown to be localized on the top of the hierarchical organization of the set of positive circuits, whose attractor states change from one binary cell choice to the other. Hence these motifs constitute master switches between binary cell fate choices (intralineage transdifferentiation). The strategy of perturbing top positive circuits in such hierarchical organization can be extended to transitions between any given pairs of cellular phenotypes even if they are not derived from a direct common progenitor. In particular, these transitions can include other types of cellular reprogramming, i.e. the transition of a differentiated cell to another cell type, either to a progenitor cell (dedifferentiation) or to another differentiated cell type coming from a different progenitor cell (interlineage transdifferentiation). In these cases, a more

47 complex set of positive circuits with mutually exclusive gene expression stable states
48 could determine these transitions. This strategy leads to the identification of a small
49 number of genes (reprogramming determinants) triggering the transitions between
50 different cellular phenotypes. Indeed, in the last decade several labs have experimentally
51 demonstrated that despite differences of cell types in the expression of thousands of
52 genes, perturbation of few reprogramming determinants are usually able to trigger
53 cellular transitions from one stable cellular phenotype to another[21-23] . Nevertheless, these
54 experiments[24,25] have relied on a brute force search of effective cocktails of transcription
55 factors to achieve desired cellular transitions, and therefore, due to the combinatorial
56 complexity of this problem, they constitute a time and resource consuming strategy.
57 Hence, this fact together with the increasing interest in cellular reprogramming urge to
58 develop strategies to systematically identify optimal combinations of reprogramming
59 determinants capable of inducing cellular transitions. A number of computational models
60 aiming at understanding cell fate and reprogramming have been proposed in literature[24-
61 29]. They attempt to model the dynamic behavior of specific parts of the gene regulatory
62 network (GRN) that govern the dynamics of a larger network. Although these models
63 give some insights into the relevant network motifs in cell fate decisions, they are usually
64 quite complex, relying on large number of input parameters and constraints, and only
65 consider small fractions of previously known genes to model the regulatory mechanism,
66 and most importantly, they do not provide a systematic platform to identify key
67 regulatory motifs that guarantee cellular stability and are likely to be involved in the
68 transitions between different stable cellular states. One step forward in this direction is
69 the methodology developed by Chang and co-workers [25] to test, compare and rank

different recipes based on their simulated efficiency and fidelity to reprogram somatic cells to iPS in a model that considers certain level of stochasticity. However, this methodology lacks any strategy to look for better combinations or to improve the efficiency and fidelity and relies on a preliminary list of candidate genes both for the network reconstruction process and the selection of combinations to test.

Here we propose a cellular transition-dependent method that identifies candidates for reprogramming determinants by focusing on stability motifs in gene regulatory networks. Given that the approach does not require a preliminary list of candidates, it can be applied to biological systems without prior knowledge on it. Our method initially searches for differentially expressed positive circuits (DEPCs), for which the expression levels of their genes change between two different cellular phenotypes. Further, a hierarchical organization of these circuits is analyzed in order to identify master regulatory positive circuits, which directly or indirectly regulate the states of the other DEPCs.

Finally, given the stochastic nature of molecular interactions and abundances in gene regulatory networks affecting cellular reprogramming efficiency and fidelity, we use a previously introduced network topological characteristic termed retroactivity[30], which positively correlates with expression noise[31], in order to detect combinations of genes in master regulatory DEPCs that are more affected by expression noise and need to be controlled in order to minimize information loss during signal transmission in gene regulatory networks. These gene combinations are the best candidates for reprogramming determinants according to our model.

We selected three representative biological examples of cellular reprogramming with experimental information on reprogramming determinants inducing effective transitions

between cellular phenotypes in order to assess the applicability of our method. These examples are the transdifferentiation from T-helper lymphocyte Th2 to Th1 (intralineage transdifferentiation), from myeloid to erythroid cells (interlineage transdifferentiation), and from fibroblast to hepatocyte (distant interlineage transdifferentiation). In the Th2-Th1 example, we identified GATA3 and T-bet as potential inducers of Th2 to Th1 T-helper transdifferentiation, which is in full agreement with previously reported experimental observations[32,33]. Our results showed that cells committed to become megakaryocytes or erythrocytes in the erythroid lineage can be reprogrammed to the myeloid lineage and become granulocytes or macrophages by perturbation of a single reprogramming determinant, i.e. the activation of GATA1. This induced transition has been experimentally validated[19]. Finally, the application of our method to the example of fibroblast to hepatocyte reprogramming allowed us to detect combinations of reprogramming determinants that induce this cellular transition. Among these detected combinations, the combined activation of HNF4 and FOXA2 has been experimentally validated by the work of Sekiya and Suzuki published in 2011[34].

In conclusion, here we propose, to our knowledge, the first method that systematically identifies combinations of genes (reprogramming determinants), which are potentially capable of inducing transitions between specific pairs of cellular phenotypes, without prior knowledge of possible candidates for reprogramming determinants. Our method generalizes the principle of transcription factor cross-repression in binary lineage decisions in the sense that it searches for master regulatory positive circuits, which contribute to the stability of cellular gene regulatory networks, and whose genes are differentially expressed with respect to specific pairs of cellular phenotypes.

Perturbations of combinations of genes belonging to these circuits that swap their steady stable states are expected to induce transitions between these phenotypes. We believe that considering the increasing interest of the research community in using cellular reprogramming in the establishment of cell disease models and regenerative medicine, our method constitutes a useful computational protocol that aims to assist researchers in the field in designing experimental strategies.

Results

A popular framework for conceptualizing and describing cellular transitions is that of the landscapes proposed by Waddington[15-17], where cellular phenotypes may be seen as stable steady states (termed as attractors) of GRNs represented as wells separated by the so-called epigenetic barriers. These barriers are established by those elements stabilizing GRNs in their attractors. Given that cellular reprogramming implies a transition between two cellular stable transcriptional programs (two attractors of the GRN), it is necessary that the corresponding GRN was at least bi-stable. The presence of positive circuits or positive feed-back loops (the sign of a circuit is defined by the product of the signs of its edges, being activation positive and inhibition negative) in a GRN is a necessary condition for the existence of at least two attractors (multi-stability)[20]. Hence, some of the positive circuits constitute the stability elements of the GRN. In particular, there are positive circuits whose genes are differentially expressed between two given attractors. By swapping the states of these circuits it should be possible to induce transitions from one attractor to another, similarly to how transitions between cell types derived from a common progenitor cell can be induced by swapping the states of cross-repression

motifs. Given the stochastic nature of molecular interactions in GRNs, perturbations of different combinations of genes belonging to these positive circuits can trigger these transitions with different efficacy.

Description of the method

Here we propose a method to design reprogramming protocols based on the topological relationship between the elements involved in the stabilization of specific attractors. The hierarchical organization analysis of strongly connected components (SCCs) formed by one or more DEPCs allows us to identify combinations of genes belonging to master regulatory DEPCs that should be perturbed in order to directly or indirectly target all DEPCs and consequently to induce specific cellular transitions. Finally, we select among these combinations of genes those with highest interface out-degree that refers to the number of genes that are directly regulated by them. The reason for this step is to minimize the retroactivity effect on master regulatory circuits[30,31], which considers the increased time response of these circuits after noise or external perturbations. This allows us to minimize the expression noise due to retroactivity contextualized to the specific cellular transition under study. In other words, we select combinations of genes participating in more transcriptional regulation events in order to minimize DEPCs time response and the stochastic behavior of GRN under perturbation, and therefore to minimize information loss during signal transmission. This strategy allows us to narrow down a huge combinatorial searching problem to a set of minimal combinations that constitutes alternative reprogramming protocols and the output of our method.

The method can be described with the following three steps, which are shown in figure 2:

1. Detecting master regulatory SCCs.
2. Determining master regulatory DEPCs for each master regulatory SCC.

3. Detecting reprogramming determinant genes within master regulatory circuits.

Detecting master regulatory SCCs

In order to detect master regulatory SCCs or clusters of DEPCs that should be independently perturbed it is necessary to detect and list all positive circuits or positive regulatory feed-back loops. We also need to identify network attractors corresponding to the two phenotypes of the cellular transition under interest. Once we have this information we proceed to determine, among the entire set of positive circuits, which are DEPCs for this specific cellular transition, meaning that the expression levels of their genes change between involved cellular phenotypes. These DEPCs can be clustered forming SCCs, and these SCCs (if there is more than one) can be interconnected. In order to detect which are the SCCs that should be independently perturbed to guarantee that all DEPCs are reached by the perturbation signal, we analyze the hierarchical organization of SCCs formed by DEPCs. It is worth stressing that this hierarchical organization is cellular transition dependent since it is based on positive circuits that change between initial and final cellular phenotypes (See methods for details about the circuit's detection, attractor computation and hierarchical analysis).

Determining the master regulatory DEPCs for each master regulatory SCC

DEPC with higher degree interface is considered the master regulatory circuit of each specific SCC. The degree interface of a circuit is the count of genes directly regulated by genes belonging to the circuit. These DEPCs master regulators should be independently perturbed in order to induce the desired cellular transition, and minimal combinations of genes able to target all master regulatory DEPCs equal in number to the number of such DEPCs. In other words, the perturbation of one gene per master regulatory DEPCs is required. Since different minimal combinations (equal in number) can arise from this

procedure, we aim to select the best combinations according to retroactivity contribution criteria. It is worth stressing that despite the degree interface could be calculated for any circuit in the GRN, the method only pay attention on those genes that belong to DEPCs when comparing two attractors, given that they are the ones that are going to be destabilized and re-stabilized in the original and final attractor respectively.

Detecting reprogramming determinant genes

Identification of genes belonging to DEPCs master regulators with maximum gene degree interface, means that they are the most regulatory genes, and therefore main responsible for DEPCs retroactivity. This set of genes constitutes the reprogramming determinants. If more than one combination of reprogramming determinant candidates equal in number of genes and interface out-degree, all of them are considered reprogramming determinants according to our model, and they constitute alternative solutions.

Application of the method to three illustrative biological examples

We selected three different biological examples of cellular reprogramming in order to illustrate and validate the applicability of our method as generalization of transcription factor cross-repression concept in illustrative biological cases. These examples provide an experimental validation of the identified sets of reprogramming determinants as effective inducers of transitions between cellular phenotypes. The Th2-Th1 and Myeloid-Erythroid examples are based on GRNs previously published by Mendoza et al. [35] and Krumsiek et al. and Dore et al.[36,37], respectively. These two networks were constructed to describe the differentiation process of the corresponding human cell types. We showed that the appropriate perturbations of these networks allow inducing transdifferentiation

between cell types with the same cellular precursor. The mouse Fibroblast-hepatocyte reprogramming example illustrates the case of a cellular transition between two cell types that do not share the same direct cellular precursor. In this case we reconstructed a literature based GRN of differentially expressed genes between both cell types[38]. This network was contextualized by an iterative network pruning described in the methods section and previously published[39]. This contextualized network is specific for the cellular transition under study, and therefore suitable to describe input-output relationships or network response under specific perturbations for a given initial network stable state (stable expression pattern).

The networks for the three examples were enriched when it was possible with information about miRNAs interactions experimentally validated and publicly available[40,41]. Details about GRN for these three biological examples are included in methods section and supplements.

Th2-Th1

T lymphocytes are classified as either T helper cells or T cytotoxic cells. T helper cells take part in cell- and antibody-mediated immune responses and they are sub-divided in Th0 (precursor) and effector Th1 and Th2 cells depending on the array of cytokines that they secrete[42]. T-helper differentiation network determines the fate of the T-Helper lineage [35], with three different attractors corresponding with the three different phenotypes (Th0, Th1 and Th2). We applied our method on a GRN previously published [35] , which represents the regulatory mechanisms determining T-helper basic types. This network includes T-bet and GATA-3 forming a cross-repression motif responsible for the differentiation either to Th1 or to Th2 from a common precursor (Th0). We applied our

method in order to detect reprogramming determinants for the Th2-Th1 transdifferentiation. The SCCs hierarchy analysis followed by the maximum retroactivity criteria allowed us to identify one master regulatory SCC with one master regulatory DEPC (named as circuit 16 in figure 3a and supplements) among five DEPCs of this specific cellular transition. Circuit 16 corresponds to the positive feed-back loop formed by GATA-3, T-bet, SOCS-1, IL-4R and STAT-6. The interface out-degree of this circuit is 11, resulting of the sum of interface out-degree of all genes belonging to it. Within this DEPC master regulator there are two genes with equal contribution to the circuit degree interface: GATA-3 and T-bet have a degree interface of 4. According to the methodology presented here both GATA-3 and T-bet constitute independent reprogramming determinants, by inactivation and activation respectively. The predicted capability of T-bet to induce the transition from Th2 to Th1 is in full agreement with reported experimental results[18]. To our knowledge, there is no experimental evidence of either the capability or incapability of GATA3 to induce the transition from Th2 to Th1 when inactivated.

It is worth mentioning that the cross-repression motif responsible for the binary cell decision between Th1 and Th2 from the precursor Th0 is embedded in the master regulatory SCC, and the detected master regulatory DEPC, named as circuit 16, is composed of the two genes forming the cross-repression motif. This example illustrates how a motif responsible for cell fate decision can also participate in the derived cellular phenotypes stabilization and how its proper perturbation can trigger transitions between them.

Myeloid-Erythroid

Within the hematopoiesis there are several binary decisions from multipotent stem cells to different type of blood cells. One of these decisions, the one determining if multipotent stems cells become erythroid (later erythrocytes and megakaryocytes) or myeloid precursor cells (later macrophages and granulocytes) requires the participation of the transcription factor cross-repression motif including GATA-1 and PU.1. As it is shown in figure 3a, the application of our method on a GRN previously published[36,37], containing this motif embedded and connected with other multi-stable motifs allowed us to identify GATA-1 as a reprogramming gene able to induce the transition from myeloid to erythroid precursor cells. This finding is in full agreement with the experimental results obtained by Heyworth et al. [19], where the authors reported that myeloid precursors infected with an inducible form of GATA-1 generated erythroid colonies when GATA-1 was induced. In figure 3 b it is shown that in this example we found a single master regulatory circuit, named as Circuit 12, with an interface out-degree of 8, which is formed by the mutual inhibition between GATA-1 and PU.1. In this particular case we obtained two possibilities with identical gene degree interface of 4: activation of GATA-1 and inhibition of PU.1. The activation of GATA-1 refers to the experiment performed by Heyworth et al. [19].[19]. To our knowledge there is no experimental evidence to support that the inhibition of PU.1 is neither able nor unable to produce the same effect yet. As in the previous example, here we observe how a cross-repression motif not only participates in binary cell fate decision, but also can be exploited to re-specify the cellular commitment in cells sharing the same precursor,

Fibroblast-Hepatocyte

Normally, hepatocytes differentiate from hepatic progenitor cells to form the liver during the regular development. However, hepatic programs can also be activated in different cells under particular stimuli or fusion with hepatocytes. The transition from mouse fibroblasts to hepatocyte-like cells induced by the perturbation of specific combinations of transcription factors has been previously reported by several authors[34,38]. As it is shown in the table included in figure 3 c, in this case the SCCs hierarchical analysis allowed us to identify two master regulatory SCCs, one including circuit 2 (including NR5A2 and FOXA2) and one including circuits 0, 7 and 4 (including genes AGT, PPARGC1A, UCP2 and HNF4A). Within the latter SCC, the DEPC, named as circuit 0, is the one with the highest interface out-degree of 20. Then, we proceeded to identify reprogramming determinants by targeting both master regulatory circuits. Within circuit 2, the gene that contributes the most to the circuit retroactivity is FOXA2, with an interface out-degree of 5. Within the circuit 0, HNF4A is the one with the highest contribution to the circuit retroactivity with an interface out-degree of 9. Therefore, the final combination of reprogramming determinants is HNF4A and FOXA2. Both genes should be activated to trigger the transition from fibroblast to hepatocyte. This result is supported by the work of Sekiya and Suzuki published in 2011[34]. These authors experimentally validated three different combinations of two transcription factors able to induce the transition from mouse fibroblast to hepatocyte, including HNF4A and FOXA2. This cellular transition constitutes a good example of reprogramming cells without a common direct precursor (interlineage transdifferentiation).

Details about attractors, circuits and genes interface out-degree o for the three biological examples are included in the supplements.

Discussion

Cellular reprogramming, including the conversion of one differentiated cell type to another (trans-differentiation) or to a more immature cell (dedifferentiation), constitutes an invaluable tool for studying cellular changes during development and differentiation, and has an enormous relevance for regenerative medicine and disease modeling. Although, substantial progress has been made in developing experimental reprogramming techniques, to date the scientific community is still faced with challenges such as the identification of optimal sets of genes whose repression and/or activation are capable of reprogramming one cell type to another (reprogramming determinants), and the elucidation of molecular changes and relevant pathways involved in these transitions (9). Furthermore, there is currently no methodology able to systematically predict reprogramming determinants that could guide the design of cellular reprogramming experiments. The development of computational models of transcriptional regulation that underlies cellular transitions would help to predict these reprogramming determinants. Moreover, the analysis of gene regulatory network properties has allowed the identification of functionally relevant motifs of interactions that could play a role in cellular transitions. In particular, transcription factor cross-antagonism has been described as a mechanism that plays a key role in cell fate decisions. A bistable toggle switch constitutes a molecular cross-repression motif that determines cellular commitment and provides stability to gene regulatory networks underlying transcriptional programs of binary decision cell choices. Experimental evidences indicate that flipping the stable states of these toggle switches produces interconversion between binary decision choices. Nevertheless, interlineage transdiferentiation and dedifferentiation

could involve perturbation of combinations of cross-repression motifs together with other multistable motifs. Here we propose a method, which considers the connectivity of these different multistable motifs, in order to systematically identify sets of reprogramming determinants able to induce transitions from differentiated cells to other cell types, either to progenitor cells (dedifferentiation) or to other differentiated cell types (transdifferentiation). Our strategy rests on the identification of a subset of all network positive circuits (necessary condition for network multistability), whose genes are differentially expressed between the cellular states involved in these. We termed this subset as differentially expressed positive circuits (DEPC). Further, a hierarchical organization of these circuits allows us to detect master regulatory positive circuits, which directly or indirectly regulate the states of the other DEPCs. By focusing on genes belonging to these master regulatory circuits, we dramatically reduced the number of possible combinations of reprogramming determinants.

However, some of these gene combinations in master regulatory DEPCs are more influenced by expression noise, affecting signal transmission in gene regulatory networks, and consequently decreasing reprogramming efficiency and fidelity. This is due to the fact that they are participating in a bigger number of regulations, so a limited concentration of the gene product has to interact with several targets a part from the one that closes the DEPC. In other words, the gene product has to distribute to different regulated targets, so the probability that the DEPC signal feed-back is broken by chance is higher (neglecting considerations about different molecular affinities that are assumed similar). Hence, in order to increase signal transmission our method proposes these gene combinations as reprogramming determinants. It is worth mentioning that we have

considered in our model some of the important events influencing reprogramming efficiency and fidelity, such as the role of noise in network dynamics and the regulatory interactions played by miRNAs. However, other factors, such as epigenetic modifications that block activation of certain genes can affect the expected network behavior after specific perturbations. Furthermore, it has been experimentally shown that epigenetic modifications can prevent cellular reprogramming reversibility in some cases [43]. In addition, our model does not take into account different delays in time response of distinct regulatory interactions. Nevertheless, given that the purpose of our method is the identification of reprogramming determinants, rather than a detailed description of network dynamics, we consider that our model provides reasonable predictions. More accurate predictions shall require addressing these considerations in the future.

Interestingly, despite there was no methodological constraint or theoretical limitation to prevent that genes non-transcription factor are reprogramming determinants, to date, in a blind application of the method, TFs always came up as reprogramming determinants.

It is worth mentioning that applicability of the method presented here is restricted to cellular transitions between stable states or stable expression patterns and constraint by the availability of information to reconstruct the corresponding GRN, as it is explained in more detailed in methods' section.

Thus, our method constitutes the first strategy that systematically provides lists of combinations of reprogramming determinants for cellular reprogramming events involving two given cellular phenotypes without prior knowledge on potential candidates and pathways involved. Due to that, the method is easily exportable to different biological systems, providing guidance even without having expertise in a biological

process. In particular, this method is suitable for cellular transdifferentiation, especially when transitions occur between different cellular lineages. Indeed, interlineage transdifferentiation involves significant changes in several molecular mechanisms that increase the complexity of this type of reprogramming, and therefore hinders the prediction of reprogramming determinants.

Hence, given the increasing interest in various applications of cellular reprogramming in medicine and basic research, our method represents a useful computational methodology to assist researchers in the field in designing experimental strategies, especially when very little about a specific biological system is known.

Methods

Networks reconstruction

Among the selected biological examples, Th2-Th1 and Myeloid-Erythroid reprogramming illustrate the case of transdifferentiation between two cell types sharing a direct common precursor. We based our analysis on previously published GRNs describing the regular differentiation process of T-helper and cell fate decisions during hematopoiesis[35-37]. These two published network were enriched with miRNA interactions experimentally validated and publicly available in two different databases: TransmiR[40] and miRTarBase[41], including information about miRNA regulatory genes and miRNA regulated genes respectively. Only miRNA forming closed loops with network genes and, therefore, able to affect the stability of the network were included (see table 1).

The Fibroblast-Hepatocyte reprogramming example illustrates a distant (interlineage) cellular transdifferentiation. Therefore, no canonical previously published network can be exploited to detect the reprogramming determinants. Such reprogramming requires the reconstruction of a GRN contextualized to this specific cellular transition.

Given that the final goal is to induce the transition from one specific cell phenotype to one another, the network is constructed based on changing elements between these two states, i. e., differentially expressed genes (DEG) between these two conditions or cell types obtained from microarray experiments. We scanned the literature and collected 24 genes known to play a relevant role in liver development and function and differentially expressed when comparing fibroblasts and hepatocytes according to previous works [44-47] . We proceed to try to connect these genes using interactions obtained from literature harvested from the entire PubMed. For this specific purpose we used the information contained in the ResNet mammalian database from Ariadne Genomics (http://www.ariadnegenomics.com/). The ResNet database includes biological relationships and associations, which have been extracted from the biomedical literature using Ariadne's MedScan technology[48,49]. More specifically, we included interactions annotated in the ResNet mammalian database in the category of Expression, PromotorBinding and Regulation. In the Expression category interactions indicates that the regulator changes the protein level of the target, by means of regulating its gene expression or protein stability. In the PromotorBinding category interactions indicates that the regulator binds the promotor of the target. Finally, in the Regulation category interactions indicates that the regulator changes the activity of the target. Similar resources for network reconstruction are the IPA tool of Ingenuity Systems

(http://www.ingenuity.com/) and the Transfac tool (http://www.biobase-international.com).

Once we had a raw GRN from literature, we proceed to remove interactions inconsistent with expression data by an iterative network pruning. These removals represent interactions apparently not active in the biological context under study. It should be taken into account that interactions from literature usually come from different biological contexts as cell types, tissues or even species. This network pruning allows us to reduce the amount of "false" interactions and to obtain a contextualized network. The algorithm applied for this network pruning[39] was originally conceived to predict missing expression values in gene regulatory network, but could be applied to contextualize the network when all the expression values in two given cellular phenotypes or stable transcriptional programs are known. Basically, the algorithm exploits the consistency between predicted and known stable states from experimental data to guide the iterative network pruning that contextualizes the network to the biological conditions under which the expression data were obtained. This process implies the booleanization of cellular phenotypes coming from experimental expression data; genes considered as up-regulated and down-regulated for a given p-value (usually < 0.05 for a regular t-test) are assumed as "1" and "0" respectively. This is due to the fact that a Boolean model is assumed to compute network attractors. An evolutionary algorithm, more specifically an estimation of distributions algorithm (EDA) [50] samples the probability distribution of positive feedback loops or positive circuits and individual interactions within the subpopulation of the best-scored networks at each iteration of the pruning algorithm. The resulting contextualized network is based not only on previous knowledge about local connectivity

but also on a global network property (stability) providing robustness in predictions (the remaining set of interactions) against noisy sources of information and network incompleteness. Despite we tried to enrich this network with miRNA interactions as we did in the two previous examples, none miRNA involved in regulatory loops or circuits with genes differentially expressed were found experimentally validated for mouse. More details about network reconstruction process for the Fibroblast-Hepatocyte reprogramming example are included in the supplementary information.

Main properties of these three biological examples GRN are shown in table 2.

Network transformation in a directed acyclic graph (DAG)

The first step of the method, named as "Detecting master regulatory SCCs" in results section, requires the hierarchical analysis of a subnetwork of the complete GRN including only DEPCs and all genes and interactions connecting them. This subnetwork contains positive feed-back loops, so it should be transformed in order to be able to analyze its hierarchy. The transformation of this subnetwork of connected DEPCs in a DAG was performed by contraction of DEPCs strongly connected, i e, SCCs of differentially expressed genes, in single super-nodes. This network transformation allows the hierarchical analysis of the network following the method described by Jothi et al. [51], resulting in the location of SCCs at different levels of hierarchy with the subsequent identification of master regulators SCCs on the top of the hierarchy pyramid.

During the application of this network transformation to the three examples included in this work we also forced the method to work on differentially expressed negative circuits (DENC) instead of DEPCs to illustrate the failure of the method when a wrong stability element is considered. Interestingly, we could not found any single DENC in none of the

three examples, despite the relative abundance of negative circuits in the three GRNs (17, 11 and 11 for Th2-Th1, Myeloid-Erythroid, and Fibroblast-Hepatocyte respectively, whereas the corresponding number of positive circuits are 29, 25 and 19). Consequently, it was not possible to perform the network transformation in a DAG and the subsequent hierarchical analysis because there was no SCC of negative circuits to analyze. This finding is consistent with the role of positive circuits or positive feed-back loops as cornerstone of multi-stable behavior in networks of interacting elements.

Circuits' detection

The Johnsons algorithm [52] was implemented to detect all elementary feedback circuits in the network. A feedback circuit is a path in which the first and the last nodes are identical. A path is elementary if no node appears twice. A feedback circuit is elementary if no node but the first and the last appears twice. Once we have all elementary feedback circuits, we select positive feedback circuits, or feedback circuits for which the difference between the number of activating edges and the number of inhibiting edges is even. Both elementary feedback circuit detection, positive feedback circuits sorting and DEPFCs detection were implemented in Perl.

Attractor computation

We assumed a Boolean model to compute attractors with a synchronous updating scheme [53] and using our own implementation[39] of the algorithm described by Garg *et al.*, 2007 [54]. The logic rule applied by default is the following: if none of its inhibitors and at least one of its activators is active, then a gene becomes active; otherwise the gene is inactive. If different regulatory rules are known for specific genes, this knowledge can be included in the model. Results in the attractor computation were consistent with the results obtained

using previously published software to compute attractors in Boolean systems (Boolnet [55], GenYsis[54]).

Minimal input data for the method usage and limitations

Given that our methodology considers transitions between attractor states, it requires the availability of expression data of stable cellular phenotypes. In addition, if the GRN has been experimentally validated and its attractors are consistent with the cellular phenotypes under study, our methodology is readily to be applied. Otherwise, the GRN has to be reconstructed from publicly available data, and therefore the applicability of our methodology could be limited by the availability of information. In this case, the reliability of the resulting GRN can be estimated by evaluation of how well the stable states of this network coincide with the experimental expression data. We usually assumed a threshold of 70 % to consider a GRN worth to be processed. For instance, in the Fibroblast-Hepatocyte example after the network contextualization process, the attractor computation of the resulting GRN revealed a matching with the expression data of 76 % for both conditions (fibroblast and hepatocytes), meaning that 76 % of gene expression values in the network are well predicted for these two conditions. The remaining 24 % of the gene expression values are not well predicted due to two different possibilities: incompleteness of the network or wrong assumed regulatory rules in specific cases. It is worth noticing that our method for contextualizing GRNs rests on removal of inconsistent regulatory interactions rather than on the addition of new interactions, and therefore the possibility of adding new predicted interactions could improve the description of the expression data. This is a very interesting and very relevant point, and despite it is out of the scope of the present work, and the fact that it

507 constitutes a challenging computational problem, it should be definitely pursued in order
508 to improve our methodology**.**
509
510 **References**

511 1 Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587-594,
512 doi:10.1038/nature08533 (2009).
513 2 Cantor, A. B. & Orkin, S. H. Hematopoietic development: a balancing act. *Curr*
514 *Opin Genet Dev* **11**, 513-519 (2001).
515 3 Graf, T. Differentiation plasticity of hematopoietic cells. *Blood* **99**, 3089-3101
516 (2002).
517 4 Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell
518 biology. *Cell* **132**, 631-644, doi:10.1016/j.cell.2008.01.025 (2008).
519 5 Arinobu, Y. Reciprocal activation of GATA-1 and PU.1 marks initial
520 specification of hematopoietic stem cells into myeloerythroid and myelolymphoid
521 lineages. *Cell stem cell* **1**, 416-427 (2007).
522 6 Iwasaki, H. & Akashi, K. Myeloid lineage commitment from the hematopoietic
523 stem cell. *Immunity* **26**, 726-740 (2007).
524 7 Zhou, L. TGF-[bgr]-induced Foxp3 inhibits TH17 cell differentiation by
525 antagonizing ROR[ggr]t function. *Nature* **453**, 236-240 (2008).
526 8 Laslo, P. Multilineage transcriptional priming and determination of alternate
527 hematopoietic cell fates. *Cell* **126**, 755-766 (2006).
528 9 Frontelo, P. Novel role for EKLF in megakaryocyte lineage commitment. *Blood*
529 **110**, 3871-3880 (2007).

10 Hwang, E. S., Szabo, S. J., Schwartzberg, P. L. & Glimcher, L. H. T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. *Science* **307**, 430-433 (2005).

11 Heins, N. Glial cells generate neurons: the role of the transcription factor Pax6. *Nature Neurosci.* **5**, 308-315 (2002).

12 Kajimura, S. Regulation of the brown and white fat gene programs through a PRDM16/CtBP transcriptional complex. *Genes Dev.* **22**, 1397-1409 (2008).

13 Niwa, H. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**, 917-929 (2005).

14 Ralston, A. & Rossant, J. Genetic regulation of stem cell origins in the mouse embryo. *Clin. Genet.* **68**, 106-112 (2005).

15 Waddington, C. H. (Macmillan Publishers Limited. All rights reserved, 1957).

16 Kauffman, S. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437-467 (1969).

17 Kauffman, S. (Macmillan Publishers Limited. All rights reserved, 1993).

18 Szabo, S. J. *et al.* A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655-669 (2000).

19 Heyworth, C., Pearson, S., May, G. & Enver, T. Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *Embo J* **21**, 3770-3781, doi:10.1093/emboj/cdf368 (2002).

20 Thomas, R., Thieffry, D. & Kaufman, M. DYNAMICAL BEHAVIOR OF BIOLOGICAL REGULATORY NETWORKS .1. BIOLOGICAL ROLE OF FEEDBACK LOOPS AND PRACTICAL USE OF THE CONCEPT OF THE

LOOP-CHARACTERISTIC STATE. *Bull. Math. Biol.* **57**, 247-276, doi:10.1007/bf02460618 (1995).

21 Lukk, M. *et al.* A global map of human gene expression. *Nat Biotech* **28**, 322-324, doi:http://www.nature.com/nbt/journal/v28/n4/abs/nbt0410-322.html#supplementary-information (2010).

22 Muller, F.-J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nat Meth* **8**, 315-317, doi:http://www.nature.com/nmeth/journal/v8/n4/abs/nmeth.1580.html#supplementary-information (2011).

23 Dudley, J. T., Tibshirani, R., Deshpande, T. & Butte, A. J. Disease signatures are robust across tissues and experiments. *Molecular systems biology* **5**, doi:http://www.nature.com/msb/journal/v5/n1/suppinfo/msb200966_S1.html (2009).

24 Ding, S. & Wang, W. Recipes and mechanisms of cellular reprogramming: a case study on budding yeast Saccharomyces cerevisiae. *BMC Syst Biol* **5**, 50, doi:10.1186/1752-0509-5-50 (2011).

25 Chang, R., Shoemaker, R. & Wang, W. Systematic search for recipes to generate induced pluripotent stem cells. *PLoS Comput Biol* **7**, e1002300, doi:10.1371/journal.pcbi.1002300 (2011).

26 Chickarmane, V. & Peterson, C. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS One* **3**, e3478, doi:10.1371/journal.pone.0003478 (2008).

27 Flottmann, M., Scharp, T. & Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front Physiol* **3**, 216, doi:10.3389/fphys.2012.00216 (2012).

28 MacArthur, B. D., Please, C. P. & Oreffo, R. O. Stochasticity and the molecular mechanisms of induced pluripotency. *PLoS One* **3**, e3086, doi:10.1371/journal.pone.0003086 (2008).

29 Wang, J., Xu, L., Wang, E. & Huang, S. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* **99**, 29-39, doi:10.1016/j.bpj.2010.03.058 (2010).

30 Del Vecchio, D., Ninfa, A. J. & Sontag, E. D. Modular cell biology: retroactivity and insulation. *Molecular systems biology* **4**, 161, doi:10.1038/msb4100204 (2008).

31 Chalancon, G. *et al.* Interplay between gene expression noise and regulatory network architecture. *Trends Genet* **28**, 221-232, doi:10.1016/j.tig.2012.01.006 (2012).

32 Lee, H. J. *et al.* GATA-3 induces T helper cell type 2 (Th2) cytokine expression and chromatin remodeling in committed Th1 cells. *J Exp Med* **192**, 105-115 (2000).

33 Hwang, E. S., Szabo, S. J., Schwartzberg, P. L. & Glimcher, L. H. T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. *Science* **307**, 430-433, doi:10.1126/science.1103336 (2005).

34 Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390-393, doi:10.1038/nature10263 (2011).

35 Mendoza, L. A network model for the control of the differentiation process in Th cells. *Biosystems* **84**, 101-114, doi:10.1016/j.biosystems.2005.10.004 (2006).

36 Krumsiek, J., Marr, C., Schroeder, T. & Theis, F. J. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PLoS One* **6**, e22649, doi:10.1371/journal.pone.0022649 (2011).

37 Dore, L. C. & Crispino, J. D. Transcription factor networks in erythroid cell and megakaryocyte development. *Blood* **118**, 231-239, doi:10.1182/blood-2011-04-285981 (2011).

38 Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**, 386-389, doi:10.1038/nature10116 (2011).

39 Crespo, I., Krishna, A., Le Bechec, A. & del Sol, A. Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. *Nucleic Acids Res* **41**, e8, doi:10.1093/nar/gks785 (2013).

40 Wang, J., Lu, M., Qiu, C. & Cui, Q. TransmiR: a transcription factor–microRNA regulation database. *Nucleic Acids Res* **38**, D119-D122, doi:10.1093/nar/gkp803 (2010).

41 Hsu, S.-D. *et al.* miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res*, doi:10.1093/nar/gkq1107 (2010).

42 Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nat Rev Immunol* **2**, 933-944, doi:10.1038/nri954 (2002).

43 Jopling, C., Boue, S. & Izpisua Belmonte, J. C. Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nat Rev Mol Cell Biol* **12**, 79-89, doi:10.1038/nrm3043 (2011).

44 Kyrmizi, I. *et al.* Plasticity and expanding complexity of the hepatic transcription factor network during liver development. *Genes Dev* **20**, 2293-2305, doi:10.1101/gad.390906 (2006).

45 Zaret, K. S. Genetic programming of liver and pancreas progenitors: lessons for stem-cell differentiation. *Nat Rev Genet* **9**, 329-340, doi:10.1038/nrg2318 (2008).

46 Schrem, H., Klempnauer, J. & Borlak, J. Liver-enriched transcription factors in liver function and development. Part I: the hepatocyte nuclear factor network and liver-specific gene expression. *Pharmacol Rev* **54**, 129-158 (2002).

47 Schrem, H., Klempnauer, J. & Borlak, J. Liver-enriched transcription factors in liver function and development. Part II: the C/EBPs and D site-binding protein in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation. *Pharmacol Rev* **56**, 291-330, doi:10.1124/pr.56.2.5 (2004).

48 Novichkova, S., Egorov, S. & Daraselia, N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* **19**, 1699-1706 (2003).

49 Daraselia, N. *et al.* Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604-611, doi:10.1093/bioinformatics/btg452 (2004).

50 Armananzas, R. *et al.* A review of estimation of distribution algorithms in bioinformatics. *BioData Min* **1**, 6, doi:10.1186/1756-0381-1-6 (2008).

51 Jothi, R. *et al.* Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular systems biology* **5**, 294, doi:10.1038/msb.2009.52 (2009).

52 Johnson, D. B. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing,* **4**, 77-84 (1975).

53 Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* **24**, 1917-1925, doi:10.1093/bioinformatics/btn336 (2008).

54 Garg, A., Xenarios, I., Mendoza, L. & DeMicheli, G. Vol. 4453 *Lecture Notes in Computer Science* (eds Terry Speed & Haiyan Huang) 62-76 (Springer Berlin / Heidelberg, 2007).

55 Mussel, C., Hopfensitz, M. & Kestler, H. A. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26**, 1378-1380, doi:10.1093/bioinformatics/btq124 (2010).

Acknowledgements

Funding for open access charge: Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg.

Author contributions

I. C. and A. dS. conceived the idea for the paper. I. C. wrote software, performed the experiments and analyzed the data. I. C. and A. dS. contributed to writing the paper. A. dS. coordinated and supervised the project.

Additional information

Supplementary Information accompanies this paper on Supplementary_file_1.docx and Supplementary_file_2.xlsx.

Competing financial interests: The authors declare no competing financial interests.

Figure Legends

Figure 1| Cell identity cascading landscape representing the cellular transcriptional program. Paths between pluripotent and differentiated cells, representing cellular differentiation process pass through stable expression profiles corresponding to multipotent progenitors. Binary cell fate decisions at multipotent rogenitor level are characterized by cross-repression motifs of competing transcription factors. Transdifferentiation between somatic cells are divided in those sharing a direct precursor cell (intra-lineage transdifferentiation), where cross-repression motifs, which determine cell fate decision, play a key role in stabilizing binary cell decisions and transitions between them; and those without a direct precursor (inter-lineage transdifferentiation), characterized by a more complex molecular mechanism underlying cellular transitions. Blue and red colors in cross-repression motifs and GRN stability core represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation

and **'-|'** represents inhibition or negative regulation.

Figure 2| Design of cellular reprogramming protocol in three steps. a) Detecting master regulatory strongly connected components (SCCs). In this first step, those positive circuits or positive feed-back loops in the gene regulatory network (GRN) whose genes change their expression levels between two cellular phenotypes are selected from the population of network circuits. These differentially expressed positive circuits (DEPCs) form SCCs. A hierarchical analysis in the space of these SCCs allows us to determine master regulatory SCCs. SCC 1 and 2 are located on the top of the hierarchy of the represented toy network without displaying connectivity between them. These SCCs should be independently perturbed to guarantee that the perturbation signal reaches every DEPC in the GRN. **b)** Detecting master regulatory DEPCs. Within each master regulatory SCC, a master regulatory DEPC is determined based on a retroactivity score (interface out-degree) or, in other words, based on the number of genes directly regulated by this circuit. The master regulatory DEPC is the one with the highest interface-out degree. In this toy example, Circuit 1 (composed by genes 'a', 'b' and 'c') is the master regulatory DEPC of the SCC 1, and Circuit 1 (composed by genes 'p' and 'o') of SCC 2 is the other master regulatory DEPC . These master regulatory DEPCs are colored in red in the retroactivity ranking table. **c)** Detecting reprogramming determinants. Once the master regulatory DEPCs have been determined, the selection of final reprogramming determinants is based on maximizing the sum of individual gene interface out-degrees included in the combination. In this toy example, gene 'a' is the one with highest retroactivity within the Circuit 1 of the SCC 1. Similarly, gene 'p' has the highest

interface out-degree in its respective circuit and SCC. Therefore, the reprograming determinants are 'a' and 'b' (both should be perturbed to induce the hypothetical cellular transition). Blue and red colors in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

Figure 3| Reprogramming determinants in three illustrative biological examples. a) Th2-Th1 reprogramming. Activation of T-bet and, alternatively, inhibition of GATA-3 are predicted as effective perturbations to induce this cellular transition. **b)** Cellular reprogramming from myeloid to erythroid cells. Both, activation of GATA-1 or inhibition of PU.1 are predicted as independently able to induce this cellular transition. **c)** Cellular reprograming from fibroblast to hepatocyte. In this particular case no single gene is able to induce the cellular transdifferentiation according to our predictions. On the other hand, combined activation of HNF4A and FOXA2 is predicted as an effective combination of reprogramming determinants. Blue and red colors in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

674 **Tables**

	miRNA	Interaction
Th2-Th1	**1. mir-145**	• **IFN-B -> mir-145** • **mir-145 -| STAT1**
Myeloid-Erythroid	**1. mir-34a** **2. mir-155**	• **mir-34A -| PU.1** • **CEBPA -> mir-34A** • **mir-155 -| FLI1** • **PU.1 -> mir-155** • **mir-155 -| PU.1**

675

676 **Table 1| miRNAs included in the biological examples. '->'** represents activation and **'-|'**

677 represents inhibition.

678

	Genes	Interactions	Activations	Inhibitions	miRNA
Th2-Th1	**24**	**38**	**28**	**10**	**1**
Myeloid-Erythroid	**13**	**34**	**19**	**15**	**2**
Fibroblast-Hepatocyte	**27**	**56**	**46**	**10**	**0**

679

680 **Table 2| Main properties of the gene regulatory networks of the three biological examples**

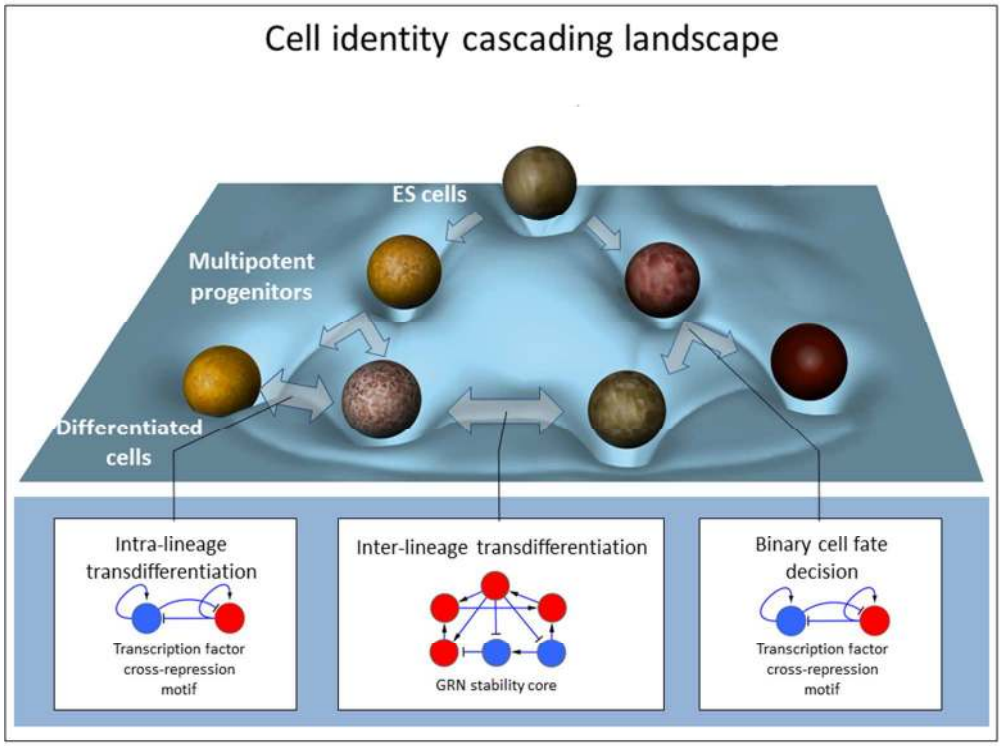


Figure 1| Cell identity cascading landscape representing the cellular transcriptional program. Paths between pluripotent and differentiated cells, representing cellular differentiation process pass through stable expression profiles corresponding to multipotent progenitors. Binary cell fate decisions at multipotent rogenitor level are characterized by cross-repression motifs of competing transcription factors. Transdifferentiation between somatic cells are divided in those sharing a direct precursor cell (intra-lineage transdifferentiation), where cross-repression motifs, which determine cell fate decision, play a key role in stabilizing binary cell decisions and transitions between them; and those without a direct precursor (inter-lineage transdifferentiation), characterized by a more complex molecular mechanism underlying cellular transitions.
Blue and red colors in cross-repression motifs and GRN stability core represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. '->' represents activation or positive regulation and '-|' represents inhibition or negative regulation.

254x190mm (96 x 96 DPI)

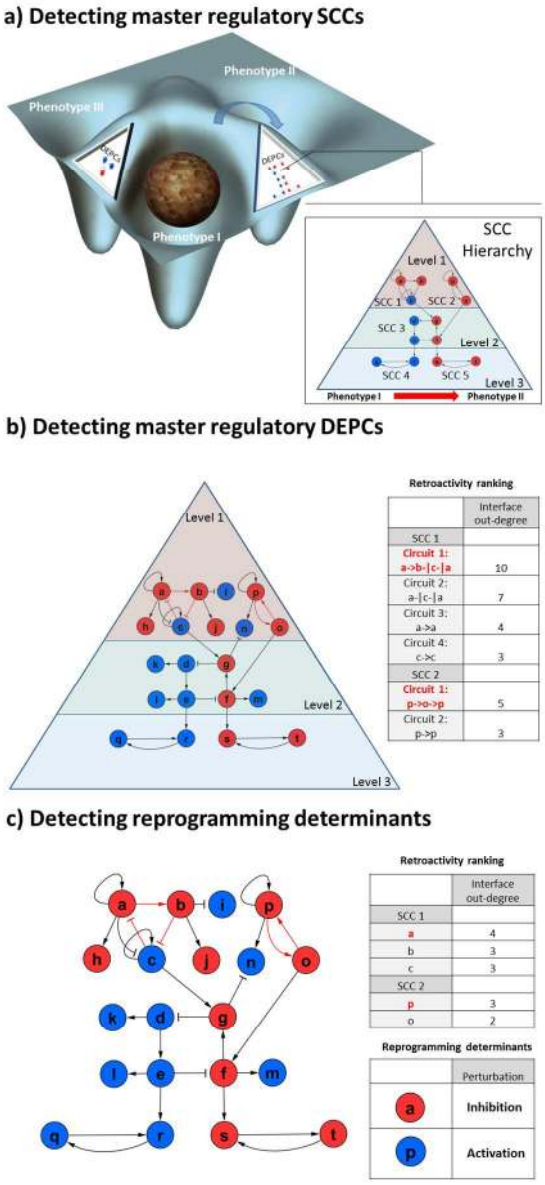
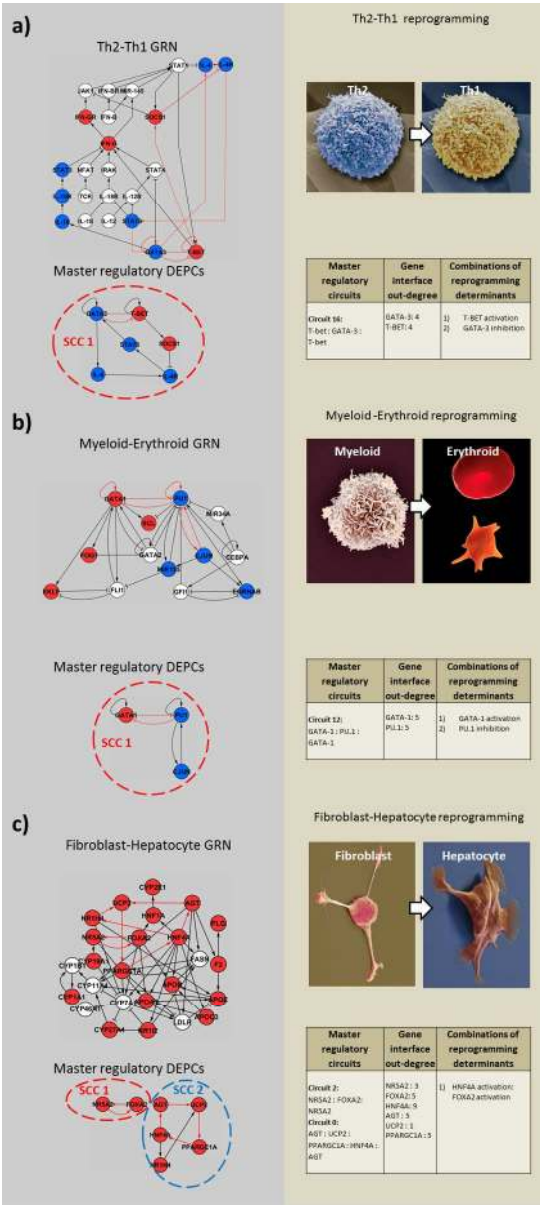


Figure 2| Design of cellular reprogramming protocol in three steps. a) Detecting master regulatory strongly connected components (SCCs). In this first step, those positive circuits or positive feed-back loops in the gene regulatory network (GRN) whose genes change their expression levels between two cellular phenotypes are selected from the population of network circuits. These differentially expressed positive circuits (DEPCs) form SCCs. A hierarchical analysis in the space of these SCCs allows us to determine master regulatory SCCs. SCC 1 and 2 are located on the top of the hierarchy of the represented toy network without displaying connectivity between them. These SCCs should be independently perturbed to guarantee that the perturbation signal reaches every DEPC in the GRN. b) Detecting master regulatory DEPCs. Within each master regulatory SCC, a master regulatory DEPC is determined based on a retroactivity score (interface out-degree) or, in other words, based on the number of genes directly regulated by this circuit. The master regulatory DEPC is the one with the highest interface-out degree. In this toy example, Circuit 1 (composed by genes 'a', 'b' and 'c') is the master regulatory DEPC of the SCC 1, and Circuit 1 (composed by genes 'p' and 'o') of SCC 2 is the other master regulatory DEPC . These master regulatory DEPCs are colored

in red in the retroactivity ranking table. c) Detecting reprogramming determinants. Once the master regulatory DEPCs have been determined, the selection of final reprogramming determinants is based on maximizing the sum of individual gene interface out-degrees included in the combination. In this toy example, gene 'a' is the one with highest retroactivity within the Circuit 1 of the SCC 1. Similarly, gene 'p' has the highest interface out-degree in its respective circuit and SCC. Therefore, the reprograming determinants are 'a' and 'b' (both should be perturbed to induce the hypothetical cellular transition). Blue and red colors in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. '->' represents activation or positive regulation and '-|' represents inhibition or negative regulation.
254x557mm (96 x 96 DPI)



254x568mm (96 x 96 DPI)

A general strategy for cellular reprogramming: the importance of transcription factor cross-repression

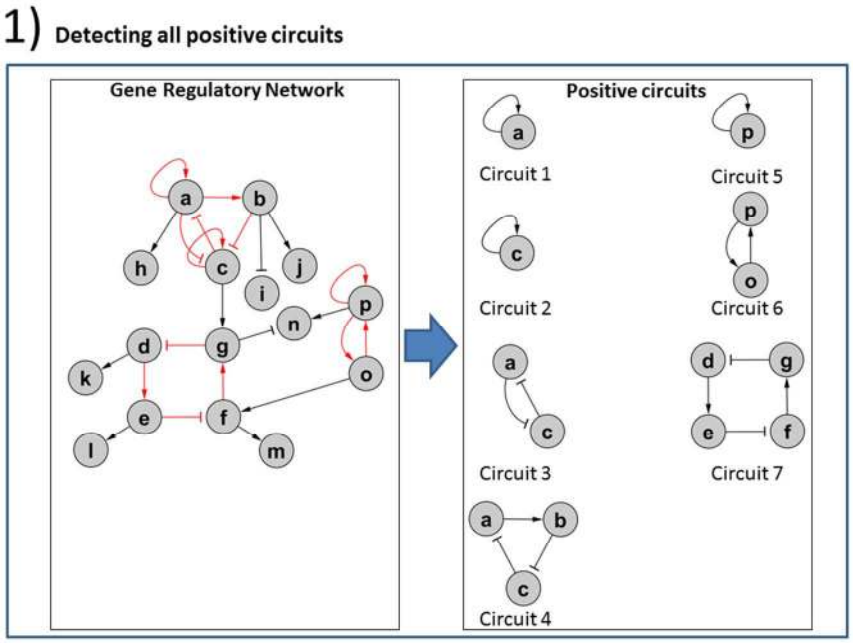
Isaac Crespo[1] and Antonio del Sol[1,*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-Belval, University of Luxembourg, L-1511 Luxembourg, Luxembourg

[*]Corresponding author: Antonio del Sol (antonio.delsol@uni.lu)

Supplementary information

Design of cellular reprogramming protocols in seven steps

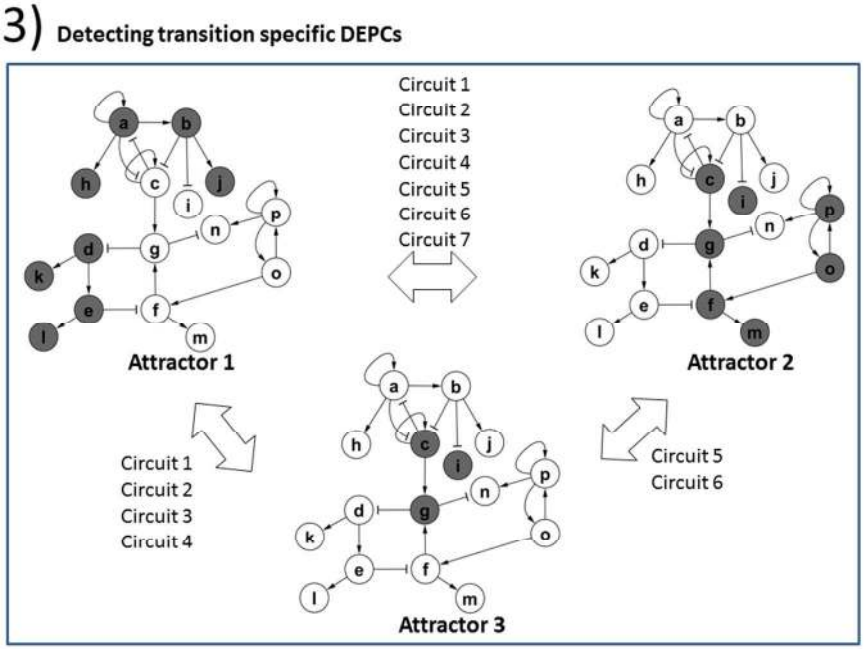


Supplementary figure 1| Positive circuit's detections. Seven positive circuits or positive feed-back loops (the sign of a circuit is defined by the product of the signs of its edges, being activation positive and inhibition negative) are present in this illustrative toy network. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

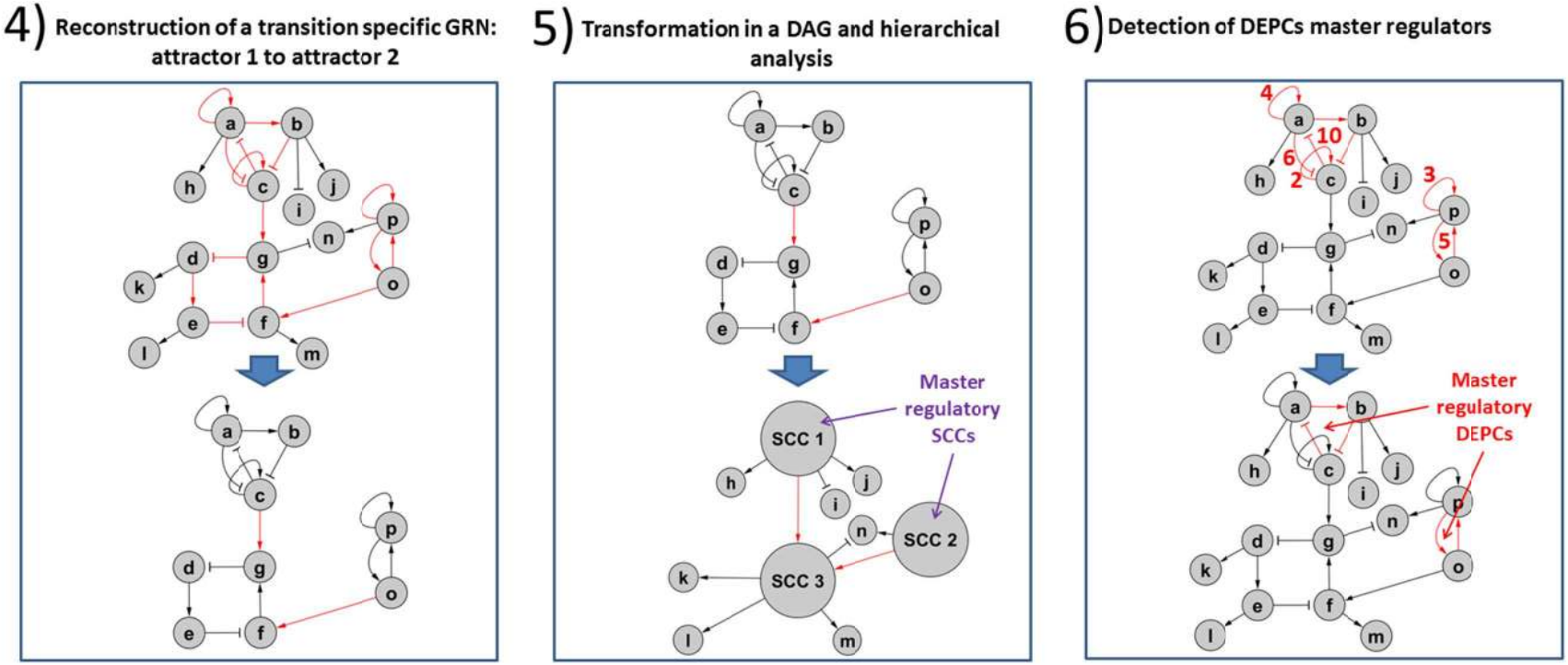
2) Computing network attractors

	Attractor 1	Attractor 2	Attractor 3
a	1	0	0
b	1	0	0
c	0	1	1
d	1	0	0
e	1	0	0
f	0	1	0
g	0	1	1
h	1	0	0
i	0	1	1
j	1	0	0
k	1	0	0
l	1	0	0
m	0	1	0
n	0	0	0
o	0	1	0
p	0	1	0

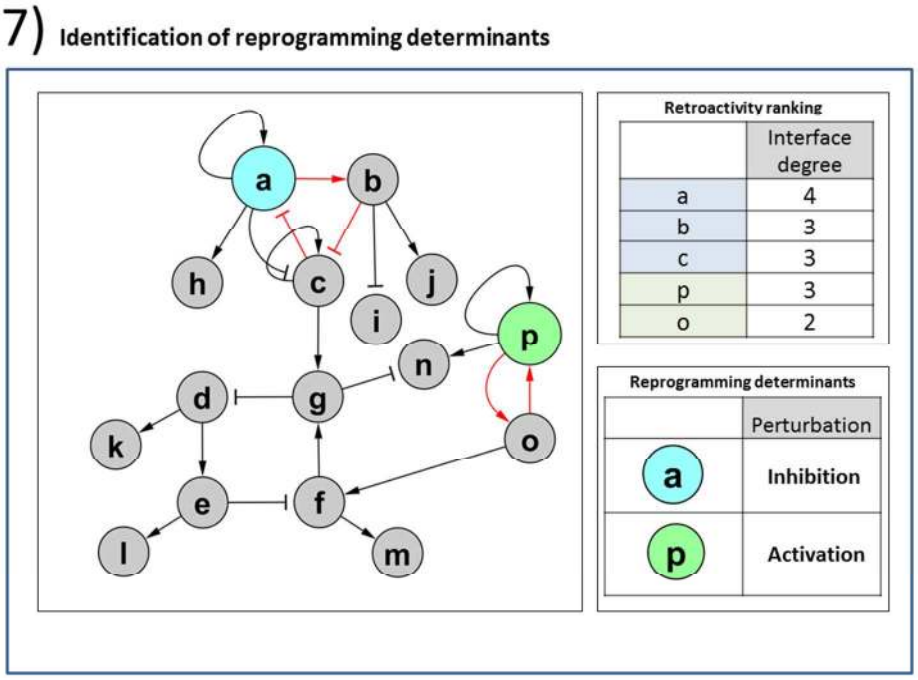
Supplementary figure 2| Network attractors computation. We assumed a Boolean model to compute attractors with a synchronous updating scheme. In such a representation '0' represents Down-regulation and '1' represents Up-regulation.



Supplementary figure 3| Transition specific DEPCs detection. Differentially expressed positive circuits (DEPCs) are those for which the expression levels of their genes change between two different attractors corresponding to two different cellular phenotypes. White and grey colors stand for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation. Transition between Attractor 1 and 2 requires the change of all positive circuits in the network. Therefore, for this specific transition all positive circuits are DEPCs. Notice that not all genes in the network are changing; gene 'n' is 'inactive' in Attractor 1 and 2.



Supplementary figure 4|Detecting master regulatory DEPCs. Detection of master regulatory DEPCs requires reconstruction of a transition specific subnetwork (Attractor 1 to Attractor 2 is represented) including only DEPCs for this specific transition and connections between them (step 4). In step 5 those DEPCs of the previously obtained subnetwork that are forming SCCs are contracted in a single supernode. The hierarchical analysis of such contracted subnetwork allows us to identify master regulatory SCCs (SCC 1 and SCC 2 in the figure). Within each master regulatory SCCs, the DEPC with highest interface out-degree (red numbers in the figure) is identified as master regulatory DEPCs (step 6); circuits 4 and 6 are the master regulatory DEPCs of this example. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.



Supplementary figure 5|Identification of reprogramming determinants. Identification of genes belonging to DEPCs master regulators with maximum gene interface out-degree. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

SOURCE	INTERACTION	TARGET	PMID
GATA3	->	GATA3	
GATA3	->	IL-10	
GATA3	->	IL-4	
GATA3	-|	STAT4	
GATA3	-|	T-BET	
IFN-B	->	IFN-BR	
IFN-B	->	MIR-145	20382746
IFN-BR	->	STAT1	
IFN-G	->	IFN-GR	
IFN-G	->	MIR-145	20382746
IFN-GR	->	JAK1	
IL-10	->	IL-10R	
IL-10R	->	STAT3	
IL-12	->	IL-12R	
IL-12R	->	STAT4	
IL-18	->	IL-18R	
IL-18R	->	IRAK	
IL-4	->	IL-4R	
IL-4R	->	STAT6	
IRAK	->	IFN-G	
JAK1	->	STAT1	
MIR-145	-|	STAT1	20098684
NFAT	->	IFN-G	
SOCS1	-|	IL-4R	
SOCS1	-|	JAK1	
STAT1	-|	IL-4	
STAT1	->	SOCS1	
STAT1	->	T-BET	
STAT3	-|	IFN-G	
STAT4	->	IFN-G	
STAT6	->	GATA3	
STAT6	-|	IL-12R	
STAT6	-|	IL-18R	
T-BET	-|	GATA3	
T-BET	->	IFN-G	
T-BET	->	SOCS1	
T-BET	->	T-BET	
TCR	->	NFAT	

		Th0	Th1	Th2
GATA3	:	0	1	0
IFN-B	:	0	0	0
IFN-BR	:	0	0	0
IFN-G	:	0	0	1
IFN-GR	:	0	0	1
IL-10	:	0	1	0
IL-10R	:	0	1	0
IL-12	:	0	0	0
IL-12R	:	0	0	0
IL-18	:	0	0	0
IL-18R	:	0	0	0
IL-4	:	0	1	0
IL-4R	:	0	1	0
IRAK	:	0	0	0
JAK1	:	0	0	0
MIR-145	:	0	0	1
NFAT	:	0	0	0
SOCS1	:	0	0	1
STAT1	:	0	0	0
STAT3	:	0	1	0
STAT4	:	0	0	0
STAT6	:	0	1	0
T-BET	:	0	0	1
TCR	:	0	0	0

Circuit 0	Circuit 6	Circuit 15	Circuit 16	Circuit 17
GATA3 -> GATA3	IL-4 -> IL-4R	T-BET -> T-BET	T-BET -| GATA3	T-BET -> SOCS1
	IL-4R -> STAT6		GATA3 -| T-BET	SOCS1 -| IL-4R
	STAT6 -> GATA3			IL-4R -> STAT6
	GATA3 -> IL-4			STAT6 -> GATA3
				GATA3 -| T-BET

GENE	INTERFACE OUTDEGREE
GATA3	4
IL-4	1
IL-4R	1
SOCS1	1
STAT6	1
T-BET	4

SOURCE	INTERACTION	TARGET	PMID
CEBPA	->	MIR-34A	20889924
CEBPA	->	GFI1	
CEBPA	->	PU1	
CEBPA	->	CEBPA	
CJUN	->	MIR-155	21515911
CJUN	->	EGR-NAB	
CJUN	->	PU1	
EGR-NAB	-|	GFI1	
EKLF	-|	FLI1	
FLI1	->	GATA1	
FLI1	-|	EKLF	
FOG1	-|	GATA2	
GATA1	->	FOG1	
GATA1	->	SCL	
GATA1	->	FLI1	
GATA1	->	EKLF	
GATA1	-|	PU1	
GATA1	-|	GATA2	
GATA1	->	GATA1	
GATA2	-|	PU1	
GATA2	->	GATA1	
GATA2	->	GATA2	
GFI1	-|	PU1	
GFI1	-|	EGR-NAB	
MIR-155	-|	PU1	6688
MIR-155	-|	FLI1	2313
MIR-34A	-|	PU1	20598588
PU1	->	MIR-155	21730352
PU1	->	CJUN	
PU1	-|	SCL	
PU1	->	EGR-NAB	
PU1	-|	GATA2	
PU1	->	PU1	
PU1	-|	GATA1	

		1	2	3	4	5	6	7
SCL	:	1	1	0	0	1	0	1
EGRNAB	:	0	0	1	0	0	0	0
MIR34A	:	0	1	0	0	0	1	1
PU1	:	0	0	1	0	0	0	0
FOG1	:	1	1	0	0	1	0	1
GFI1	:	0	1	0	0	0	1	1
CJUN	:	0	0	1	0	0	0	0
GATA2	:	0	0	0	0	0	0	0
CEBPA	:	0	1	0	0	0	1	1
MIR155	:	0	0	1	0	0	0	0
GATA1	:	1	1	0	0	1	0	1
EKLF	:	1	1	0	0	0	0	0
FLI1	:	0	0	0	0	1	0	1

Circuit 0	Circuit 3	Circuit 11	Circuit 12
CJUN -> PU1	PU1 -> PU1	GATA1 -> GATA1	GATA1 -| PU1
PU1 -> CJUN			PU1 -| GATA1

GENE	INTERFACE OUTDEGREE
CJUN	3
GATA1	5
PU.1	5

SOURCE	INTERACTION	TARGET	Type
AGT	->	CYP11A1	Expression
AGT	->	F2	Expression
AGT	->	FASN	Expression
AGT	->	LDLR	Expression
AGT	->	UCP2	Expression
APOA1	->	APOE	Expression
APOA1	->	LDLR	Expression
APOB	->	FASN	Expression
APOC3	->	APOA1	Expression
APOC3	->	APOB	Expression
APOE	->	APOB	Expression
APOE	->	CYP11A1	Expression
CYP11A1	->	CYP1B1	Expression
CYP19A1	-|	CYP7A1	Expression
CYP1A1	->	CYP1B1	Expression
CYP1B1	->	CYP1A1	Expression
CYP27A1	->	CYP11A1	Expression
CYP7A1	-|	CYP11A1	Expression
CYP7A1	-|	CYP27A1	Expression
CYP7A1	-|	CYP46A1	Expression
CYP7A1	->	LDLR	Expression
F2	->	APOE	Expression
F2	->	PLG	Expression
FASN	->	PPARGC1A	Expression
FOXA2	->	APOA1	PromoterBinding
FOXA2	->	APOB	Expression
FOXA2	->	CYP7A1	PromoterBinding
FOXA2	->	HNF1A	PromoterBinding
FOXA2	->	NR5A2	Expression
HNF1A	->	CYP2E1	Expression
HNF4A	->	AGT	PromoterBinding
HNF4A	->	APOA1	PromoterBinding
HNF4A	->	APOB	PromoterBinding
HNF4A	->	APOC3	PromoterBinding
HNF4A	->	CYP7A1	PromoterBinding
HNF4A	->	FASN	PromoterBinding
HNF4A	->	HNF1A	PromoterBinding
HNF4A	->	NR1H4	PromoterBinding
HNF4A	->	NR1I2	PromoterBinding
HP	->	F2	Expression
LDLR	-|	APOB	Expression
LDLR	-|	APOE	Expression
NR1H4	->	APOE	Expression
NR1H4	-|	CYP7A1	PromoterBinding
NR1H4	->	UCP2	Expression
NR1I2	->	CYP27A1	PromoterBinding
NR1I2	-|	CYP7A1	Expression
NR1I2	->	FASN	Expression
NR5A2	->	APOA1	PromoterBinding

NR5A2	->	CYP19A1	PromoterBinding
NR5A2	->	FOXA2	PromoterBinding
PPARGC1A	->	CYP11A1	Expression
PPARGC1A	->	CYP7A1	Expression
PPARGC1A	-|	FASN	Expression
PPARGC1A	->	HNF4A	Expression
PPARGC1A	-|	LDLR	Expression
UCP2	->	PPARGC1A	Expression

MedLine Reference
15887230:4, 1457080:2, 19221003:4, 7895901:4, 8175981:7, 10579324:1051, 2294118:100, 18308844:1147
10669632:2, 10691100:4, 11893585:8, 10079099:1152, 9564040:1233, 18678787:1115
19109942:6, 16203876:5, 11463364:1117, 15113941:1064, 16373953:1094, 15113738:1152
11564720:10, 16203876:1167, 17043664:1114, 17478552:1200, 18618022:1193
19109942:6, 20118013:6, 20814019:5, 19762685:1183
6712971:2, 20534134:2, 8413767:5, 9861782:8, 12773300:1227, 10488084:1283, 15066991:1187, 16207713:1347, 12
2151255:5, 12606523:1243
17277197:1345
18678879:1, 15109267:5, 10893424:7, 15649902:3, 9012660:6, 9988739:2, 12747582:2, 16051671:1237, 16338932:1
9012660:2, 18509206:1199, 8613278:1066
2344296:0, 17848837:2, 11013310:7, 8696954:11, 12551940:1199, 15755832:1114, 12518038:1043, 11714857:1228,
12401891:1197
7744798:8
12933663:1250
8783816:6, 19460354:3, 17324381:5, 16484233:1235, 19047483:1339, 17511620:1185, 12376470:1146, 15142886:11
19460354:3, 19690180:1154, 12970067:1173, 17511620:1185, 12376470:1146, 15297627:1319, 15142886:1139, 158
9804849:1200, 9804848:1200
18621681:1044
14870923:7, 8891849:5, 14512880:11, 12588950:1251, 18621681:1044, 12597773:1048, 11701475:1097
18621681:1044
11907135:1088, 16609145:1069, 12213890:1147, 19815588:1198, 17456796:1247, 11397693:1140, 11788471:1121
16683250:3, 15145976:1256, 11145944:1124
8089219:5, 2151408:5, 15894352:0, 4262519:6, 6968239:3, 14699093:1277, 9345041:1125, 12724354:1215, 7749851
19719788:0, 17446185:1241, 19029118:1191
7961760:5, 9512550:1106, 11714848:1391
11244563:100
16492670:1151, 15358835:1133
11875061:7, 9685261:3, 11805192:1179, 11904435:1041, 16912278:1176, 19074951:1195, 19417011:1210, 1046857
11595170:8, 14699589:1482, 11927588:1246
8810289:55
15067378:9, 21298017:4, 20684663:5, 10574924:1172, 14672953:1186, 12145290:1289
9012660:1114, 8613278:1191, 16140878:1120, 15123688:1053, 10085149:1032, 9202083:1220, 9153249:1114
18510493:3, 8344962:5, 20007910:1147, 16140878:1120, 10085149:1067, 16301212:1320, 8613278:1275, 9012660:1
9592157:2, 16223942:4, 20938723:8, 9760243:8, 11802721:1, 8760876:2, 9592159:1133, 14766742:1346, 10551874:
15322103:7, 16488887:1193, 10627496:1267, 11907135:1259, 21245926:1307, 17145766:1236, 16492670:1274, 176
15310732:1176, 16800817:100
9812974:6, 10967120:8, 8735941:7, 1734282:3, 16670373:1273, 11679424:1044, 15141028:1179, 10606640:1046, 97
14729567:1160, 15146238:1302, 16051671:1072, 16603721:1096
12774017:7, 18305375:4, 10691738:6, 12601364:8, 16912278:1137, 17764444:1202, 14657421:1504, 16051671:1331
2069574:100
8732781:5, 18272520:1, 10683382:8, 15145984:5, 8187218:4, 11904390:100, 20028946:1215, 16537968:1264, 16396
18369154:3, 15888448:8, 20005821:4, 3956506:3, 10940295:4, 18497424:1207, 11060356:1473, 20686698:1277, 166
12454263:1058, 12954636:1055, 15102878:1438
20699090:14, 18820241:1, 15694933:5, 18499494:4, 12897188:11, 17823457:4, 16168958:2, 19056864:5, 12006384:
15980055:1245
17088262:5, 17456796:1230
15629111:1, 15331348:2, 12202460:1616, 17456796:1082, 11893771:1145, 12393840:1058, 11248085:1200, 150392
20185760:1118
15218078:4

20607599:9, 20214950:10, 21273442:9, 19762543:1257, 14593077:1138, 11927588:1069, 16109788:1075, 16357189
10799577:3, 11145965:1168, 17075876:1139, 15143151:1057, 15614783:2201, 12972592:1263
21108604:4, 19389810:1104, 20133449:1150
14522988:7, 15329387:1, 15331348:8, 15576845:1237, 16051671:1072, 17636037:1101, 16037564:1168
14729567:1252
12107181:7, 19208857:5, 17636037:1187, 18664618:1318, 16603721:1285, 16825189:1297, 16885156:1207, 171457
19322023:6
21034559:6, 9163473565e7552cff17c5039d90b0c5

!951361:1063, 17303773:1131

079, 8613278:1327 <more data available...>

11024044:1228, 19667110:1123 <more data available...>

139, 11577022:1073
33926:1081, 11577022:1073

.:1262

8:1044, 10051618:1237

1128, 15123688:1053, 8995295:1265 <more data available...>
1055, 10085149:1150 <more data available...>
36037:1220, 19389810:1192, 12815072:1219 <more data available...>

'92724:1047, 18184923:1139 <more data available...>

6637:1544, 11734567:1070 <more data available...>
644710:1199, 18045818:1137 <more data available...>

1069

87:1294, 15322103:1176, 19174369:1264 <more data available...>

:1058

66:1238, 20133449:1220

		1	2	3	4
FOXA2	:	1	0	0	0
CYP19A1	:	1	0	0	0
CYP1A1	:	1	1	1	0
NR1I2	:	1	1	0	0
LDLR	:	0	0	0	0
CYP7A1	:	0	0	0	0
UCP2	:	1	1	0	0
CYP46A1	:	1	1	1	1
AGT	:	1	1	0	0
NR1H4	:	1	1	0	0
CYP1B1	:	1	1	1	0
CYP2E1	:	1	1	0	0
F2	:	1	1	0	0
APOA1	:	1	1	0	0
NR5A2	:	1	0	0	0
APOE	:	1	1	0	0
PLG	:	1	1	0	0
APOB	:	1	1	0	0
PPARGC1A	:	1	1	0	0
APOC3	:	1	1	0	0
CYP11A1	:	1	1	0	0
CYP27A1	:	1	1	0	0
HNF1A	:	1	1	0	0
FASN	:	0	0	0	0
HNF4A	:	1	1	0	0
HP	:	0	0	0	0

Circuit 0	Circuit 2	Circuit 7	Circuit 14
AGT -> UCP2	FOXA2 -> NR5A2	HNF1A -> UCP2	HNF4A -> NR1H4
UCP2 -> PPARGC1A	NR5A2 -> FOXA2	UCP2 -> PPARGC1A	NR1H4 -> UCP2
PPARGC1A -> HNF4A		PPARGC1A -> HNF4A	UCP2 -> PPARGC1A
HNF4A -> AGT		HNF4A -> HNF1A	PPARGC1A -> HNF4A

GENE	INTERFACE OUTDEGREE
AGT	5
FOXA2	5
HNF4A	9
NR1H4	3
NR5A2	3
PPARGC1A	5
UCP2	1

A general strategy for cellular reprogramming: the importance of transcription factor cross-repression

Isaac Crespo[1] and Antonio del Sol[1,*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, L-4362 Esch-Belval, University of Luxembourg, L-1511 Luxembourg, Luxembourg
[*]Corresponding author: Antonio del Sol (antonio.delsol@uni.lu)

Keywords: Cellular reprogramming, differentiation, dedifferentiation, transdifferentiation, network stability, cross repression, cross-antagonistic motif, retroactivity, positive circuit.

Abstract

Transcription factor cross-repression is an important concept in cellular differentiation. A bistable toggle switch constitutes a molecular mechanism that determines cellular commitment and provides stability to transcriptional programs of binary cell fate choices. Experiments support that perturbations of these toggle switches can interconvert these binary cell fate choices, suggesting potential reprogramming strategies. However, more complex types of cellular transitions could involve perturbations of combinations of different types of multistable motifs. Here we introduce a method that generalizes the concept of transcription factor cross-repression to systematically predict sets of genes, whose perturbations induce cellular transitions between any given pair of cell types. Furthermore, to our knowledge, this is the first method that systematically makes these predictions without prior knowledge of potential candidate genes and pathways involved, providing guidance on systems where little is known. Given the increasing interest of cellular reprogramming in medicine and basic research, our method represents a useful computational methodology to assist researchers in the field in designing experimental strategies.

Introduction

The central role of transcription factor cross-repression determining cell fate is one of the most important concepts emerged from years of lineage differentiation research[1-4]. In its simplest formulation, two regulators that negatively influence each other establish a bistable “toggle switch”, readily explaining the two mutual exclusive cell fate outcomes. More complicated schemes also include transcription factors auto-regulation and

antagonistic cross-regulation of target genes. Several examples of these binary cell fate choice mechanisms have emerged in the last ten years[5-14]. Integration of this knowledge can be represented in a binary decision tree from embryonic stem cells (ES cells) to differentiated cells passing by different progenitors[1] (see figure 1). This tree defines distinct paths between different cell types in a Waddington's landscape[15-17], where different cell types can be interpreted as steady stable states of cellular gene regulatory networks termed as attractors. Cross-repression motifs not only determine binary decisions in the tree, but based on their bistable behavior, characterized by mutually exclusive gene expression states; they also play a key role in the stability of each possible cell fate. Furthermore, experimental evidences have demonstrated that perturbations of genes belonging to these motifs are able to trigger transitions between these binary cell fate choices[18,19]. Indeed, although attractor's stability is determined by a regulatory core comprised of one or several interconnected positive feedback loops, known as positive circuits[20], these cross-antagonistic motifs are shown to be localized on the top of the hierarchical organization of the set of positive circuits, whose attractor states change from one binary cell choice to the other.Hence these motifs constitute master switches between binary cell fate choices (intralineage transdifferentiation). The strategy of perturbing top positive circuits in such hierarchical organization can be extended to transitions between any given pairs of cellular phenotypes even if they are not derived from a direct common progenitor. In particular, these transitions can include other types of cellular reprogramming, i.e. the transition of a differentiated cell to another cell type, either to a progenitor cell (dedifferentiation) or to another differentiated cell type coming from a different progenitor cell (interlineage transdifferentiation). In these cases, a more

complex set of positive circuits with mutually exclusive gene expression stable states could determine these transitions. This strategy leads to the identification of a small number of genes (reprogramming determinants) triggering the transitions between different cellular phenotypes. Indeed, in the last decade several labs have experimentally demonstrated that despite differences of cell types in the expression of thousands of genes, perturbation of few reprogramming determinants are usually able to trigger cellular transitions from one stable cellular phenotype to another[21-23] . Nevertheless, these experiments[24,25] have relied on a brute force search of effective cocktails of transcription factors to achieve desired cellular transitions, and therefore, due to the combinatorial complexity of this problem, they constitute a time and resource consuming strategy. Hence, this fact together with the increasing interest in cellular reprogramming urge to develop strategies to systematically identify optimal combinations of reprogramming determinants capable of inducing cellular transitions. A number of computational models aiming at understanding cell fate and reprogramming have been proposed in literature[24-29]. They attempt to model the dynamic behavior of specific parts of the gene regulatory network (GRN) that govern the dynamics of a larger network. Although these models give some insights into the relevant network motifs in cell fate decisions, they are usually quite complex, relying on large number of input parameters and constraints, and only consider small fractions of previously known genes to model the regulatory mechanism, and most importantly, they do not provide a systematic platform to identify key regulatory motifs that guarantee cellular stability and are likely to be involved in the transitions between different stable cellular states. One step forward in this direction is the methodology developed by Chang and co-workers [25] to test, compare and rank

different recipes based on their simulated efficiency and fidelity to reprogram somatic cells to iPS in a model that considers certain level of stochasticity. However, this methodology lacks any strategy to look for better combinations or to improve the efficiency and fidelity and relies on a preliminary list of candidate genes both for the network reconstruction process and the selection of combinations to test.

Here we propose a cellular transition-dependent method that identifies candidates for reprogramming determinants by focusing on stability motifs in gene regulatory networks. Given that the approach does not require a preliminary list of candidates, it can be applied to biological systems without prior knowledge on it. Our method initially searches for differentially expressed positive circuits (DEPCs), for which the expression levels of their genes change between two different cellular phenotypes. Further, a hierarchical organization of these circuits is analyzed in order to identify master regulatory positive circuits, which directly or indirectly regulate the states of the other DEPCs.

Finally, given the stochastic nature of molecular interactions and abundances in gene regulatory networks affecting cellular reprogramming efficiency and fidelity, we use a previously introduced network topological characteristic termed retroactivity[30], which positively correlates with expression noise[31], in order to detect combinations of genes in master regulatory DEPCs that are more affected by expression noise and need to be controlled in order to minimize information loss during signal transmission in gene regulatory networks. These gene combinations are the best candidates for reprogramming determinants according to our model.

We selected three representative biological examples of cellular reprogramming with experimental information on reprogramming determinants inducing effective transitions

between cellular phenotypes in order to assess the applicability of our method. These examples are the transdifferentiation from T-helper lymphocyte Th2 to Th1 (intralineage transdifferentiation), from myeloid to erythroid cells (interlineage transdifferentiation), and from fibroblast to hepatocyte (distant interlineage transdifferentiation). In the Th2-Th1 example, we identified GATA3 and T-bet as potential inducers of Th2 to Th1 T-helper transdifferentiation, which is in full agreement with previously reported experimental observations[32,33]. Our results showed that cells committed to become megakaryocytes or erythrocytes in the erythroid lineage can be reprogrammed to the myeloid lineage and become granulocytes or macrophages by perturbation of a single reprogramming determinant, i.e. the activation of GATA1. This induced transition has been experimentally validated[19]. Finally, the application of our method to the example of fibroblast to hepatocyte reprogramming allowed us to detect combinations of reprogramming determinants that induce this cellular transition. Among these detected combinations, the combined activation of HNF4 and FOXA2 has been experimentally validated by the work of Sekiya and Suzuki published in 2011[34].

In conclusion, here we propose, to our knowledge, the first method that systematically identifies combinations of genes (reprogramming determinants), which are potentially capable of inducing transitions between specific pairs of cellular phenotypes, without prior knowledge of possible candidates for reprogramming determinants. Our method generalizes the principle of transcription factor cross-repression in binary lineage decisions in the sense that it searches for master regulatory positive circuits, which contribute to the stability of cellular gene regulatory networks, and whose genes are differentially expressed with respect to specific pairs of cellular phenotypes.

Perturbations of combinations of genes belonging to these circuits that swap their steady stable states are expected to induce transitions between these phenotypes. We believe that considering the increasing interest of the research community in using cellular reprogramming in the establishment of cell disease models and regenerative medicine, our method constitutes a useful computational protocol that aims to assist researchers in the field in designing experimental strategies.

Results

A popular framework for conceptualizing and describing cellular transitions is that of the landscapes proposed by Waddington[15-17], where cellular phenotypes may be seen as stable steady states (termed as attractors) of GRNs represented as wells separated by the so-called epigenetic barriers. These barriers are established by those elements stabilizing GRNs in their attractors. Given that cellular reprogramming implies a transition between two cellular stable transcriptional programs (two attractors of the GRN), it is necessary that the corresponding GRN was at least bi-stable. The presence of positive circuits or positive feed-back loops (the sign of a circuit is defined by the product of the signs of its edges, being activation positive and inhibition negative) in a GRN is a necessary condition for the existence of at least two attractors (multi-stability)[20]. Hence, some of the positive circuits constitute the stability elements of the GRN. In particular, there are positive circuits whose genes are differentially expressed between two given attractors. By swapping the states of these circuits it should be possible to induce transitions from one attractor to another, similarly to how transitions between cell types derived from a common progenitor cell can be induced by swapping the states of cross-repression

motifs. Given the stochastic nature of molecular interactions in GRNs, perturbations of different combinations of genes belonging to these positive circuits can trigger these transitions with different efficacy.

Description of the method

Here we propose a method to design reprogramming protocols based on the topological relationship between the elements involved in the stabilization of specific attractors. The hierarchical organization analysis of strongly connected components (SCCs) formed by one or more DEPCs allows us to identify combinations of genes belonging to master regulatory DEPCs that should be perturbed in order to directly or indirectly target all DEPCs and consequently to induce specific cellular transitions. Finally, we select among these combinations of genes those with highest interface out-degree that refers to the number of genes that are directly regulated by them. The reason for this step is to minimize the retroactivity effect on master regulatory circuits[30,31], which considers the increased time response of these circuits after noise or external perturbations. This allows us to minimize the expression noise due to retroactivity contextualized to the specific cellular transition under study. In other words, we select combinations of genes participating in more transcriptional regulation events in order to minimize DEPCs time response and the stochastic behavior of GRN under perturbation, and therefore to minimize information loss during signal transmission. This strategy allows us to narrow down a huge combinatorial searching problem to a set of minimal combinations that constitutes alternative reprogramming protocols and the output of our method.

The method can be described with the following three steps, which are shown in figure 2:

1. Detecting master regulatory SCCs.
2. Determining master regulatory DEPCs for each master regulatory SCC.

3. Detecting reprogramming determinant genes within master regulatory circuits.

Detecting master regulatory SCCs

In order to detect master regulatory SCCs or clusters of DEPCs that should be independently perturbed it is necessary to detect and list all positive circuits or positive regulatory feed-back loops. We also need to identify network attractors corresponding to the two phenotypes of the cellular transition under interest. Once we have this information we proceed to determine, among the entire set of positive circuits, which are DEPCs for this specific cellular transition, meaning that the expression levels of their genes change between involved cellular phenotypes. These DEPCs can be clustered forming SCCs, and these SCCs (if there is more than one) can be interconnected. In order to detect which are the SCCs that should be independently perturbed to guarantee that all DEPCs are reached by the perturbation signal, we analyze the hierarchical organization of SCCs formed by DEPCs. It is worth stressing that this hierarchical organization is cellular transition dependent since it is based on positive circuits that change between initial and final cellular phenotypes (See methods for details about the circuit's detection, attractor computation and hierarchical analysis).

Determining the master regulatory DEPCs for each master regulatory SCC

DEPC with higher degree interface is considered the master regulatory circuit of each specific SCC. The degree interface of a circuit is the count of genes directly regulated by genes belonging to the circuit. These DEPCs master regulators should be independently perturbed in order to induce the desired cellular transition, and minimal combinations of genes able to target all master regulatory DEPCs equal in number to the number of such DEPCs. In other words, the perturbation of one gene per master regulatory DEPCs is required. Since different minimal combinations (equal in number) can arise from this

199 procedure, we aim to select the best combinations according to retroactivity contribution
200 criteria. It is worth stressing that despite the degree interface could be calculated for any
201 circuit in the GRN, the method only pay attention on those genes that belong to DEPCs
202 when comparing two attractors, given that they are the ones that are going to be
203 destabilized and re-stabilized in the original and final attractor respectively.

204 Detecting reprogramming determinant genes

205 Identification of genes belonging to DEPCs master regulators with maximum gene
206 degree interface, means that they are the most regulatory genes, and therefore main
207 responsible for DEPCs retroactivity. This set of genes constitutes the reprogramming
208 determinants. If more than one combination of reprogramming determinant candidates
209 equal in number of genes and interface out-degree, all of them are considered
210 reprogramming determinants according to our model, and they constitute alternative
211 solutions.

212 Application of the method to three illustrative biological examples

213 We selected three different biological examples of cellular reprogramming in order to
214 illustrate and validate the applicability of our method as generalization of transcription
215 factor cross-repression concept in illustrative biological cases. These examples provide
216 an experimental validation of the identified sets of reprogramming determinants as
217 effective inducers of transitions between cellular phenotypes. The Th2-Th1 and Myeloid-
218 Erythroid examples are based on GRNs previously published by Mendoza et al. [35] and
219 Krumsiek et al. and Dore et al.[36,37], respectively. These two networks were constructed to
220 describe the differentiation process of the corresponding human cell types. We showed
221 that the appropriate perturbations of these networks allow inducing transdifferentiation

between cell types with the same cellular precursor. The mouse Fibroblast-hepatocyte reprogramming example illustrates the case of a cellular transition between two cell types that do not share the same direct cellular precursor. In this case we reconstructed a literature based GRN of differentially expressed genes between both cell types[38]. This network was contextualized by an iterative network pruning described in the methods section and previously published[39]. This contextualized network is specific for the cellular transition under study, and therefore suitable to describe input-output relationships or network response under specific perturbations for a given initial network stable state (stable expression pattern).

The networks for the three examples were enriched when it was possible with information about miRNAs interactions experimentally validated and publicly available[40,41]. Details about GRN for these three biological examples are included in methods section and supplements.

Th2-Th1

T lymphocytes are classified as either T helper cells or T cytotoxic cells. T helper cells take part in cell- and antibody-mediated immune responses and they are sub-divided in Th0 (precursor) and effector Th1 and Th2 cells depending on the array of cytokines that they secrete[42]. T-helper differentiation network determines the fate of the T-Helper lineage [35], with three different attractors corresponding with the three different phenotypes (Th0, Th1 and Th2). We applied our method on a GRN previously published [35] , which represents the regulatory mechanisms determining T-helper basic types. This network includes T-bet and GATA-3 forming a cross-repression motif responsible for the differentiation either to Th1 or to Th2 from a common precursor (Th0). We applied our

method in order to detect reprogramming determinants for the Th2-Th1 transdifferentiation. The SCCs hierarchy analysis followed by the maximum retroactivity criteria allowed us to identify one master regulatory SCC with one master regulatory DEPC (named as circuit 16 in figure 3a and supplements) among five DEPCs of this specific cellular transition. Circuit 16 corresponds to the positive feed-back loop formed by GATA-3, T-bet, SOCS-1, IL-4R and STAT-6. The interface out-degree of this circuit is 11, resulting of the sum of interface out-degree of all genes belonging to it. Within this DEPC master regulator there are two genes with equal contribution to the circuit degree interface: GATA-3 and T-bet have a degree interface of 4. According to the methodology presented here both GATA-3 and T-bet constitute independent reprogramming determinants, by inactivation and activation respectively. The predicted capability of T-bet to induce the transition from Th2 to Th1 is in full agreement with reported experimental results[18]. To our knowledge, there is no experimental evidence of either the capability or incapability of GATA3 to induce the transition from Th2 to Th1 when inactivated.

It is worth mentioning that the cross-repression motif responsible for the binary cell decision between Th1 and Th2 from the precursor Th0 is embedded in the master regulatory SCC, and the detected master regulatory DEPC, named as circuit 16, is composed of the two genes forming the cross-repression motif. This example illustrates how a motif responsible for cell fate decision can also participate in the derived cellular phenotypes stabilization and how its proper perturbation can trigger transitions between them.

Myeloid-Erythroid

Within the hematopoiesis there are several binary decisions from multipotent stem cells to different type of blood cells. One of these decisions, the one determining if multipotent stems cells become erythroid (later erythrocytes and megakaryocytes) or myeloid precursor cells (later macrophages and granulocytes) requires the participation of the transcription factor cross-repression motif including GATA-1 and PU.1. As it is shown in figure 3a, the application of our method on a GRN previously published[36,37], containing this motif embedded and connected with other multi-stable motifs allowed us to identify GATA-1 as a reprogramming gene able to induce the transition from myeloid to erythroid precursor cells. This finding is in full agreement with the experimental results obtained by Heyworth et al. [19], where the authors reported that myeloid precursors infected with an inducible form of GATA-1 generated erythroid colonies when GATA-1 was induced. In figure 3 b it is shown that in this example we found a single master regulatory circuit, named as Circuit 12, with an interface out-degree of 8, which is formed by the mutual inhibition between GATA-1 and PU.1. In this particular case we obtained two possibilities with identical gene degree interface of 4: activation of GATA-1 and inhibition of PU.1. The activation of GATA-1 refers to the experiment performed by Heyworth et al. [19].[19]. To our knowledge there is no experimental evidence to support that the inhibition of PU.1 is neither able nor unable to produce the same effect yet. As in the previous example, here we observe how a cross-repression motif not only participates in binary cell fate decision, but also can be exploited to re-specify the cellular commitment in cells sharing the same precursor,

Fibroblast-Hepatocyte

290 Normally, hepatocytes differentiate from hepatic progenitor cells to form the liver during
291 the regular development. However, hepatic programs can also be activated in different
292 cells under particular stimuli or fusion with hepatocytes. The transition from mouse
293 fibroblasts to hepatocyte-like cells induced by the perturbation of specific combinations
294 of transcription factors has been previously reported by several authors[34,38]. As it is
295 shown in the table included in figure 3 c, in this case the SCCs hierarchical analysis
296 allowed us to identify two master regulatory SCCs, one including circuit 2 (including
297 NR5A2 and FOXA2) and one including circuits 0, 7 and 4 (including genes AGT,
298 PPARGC1A, UCP2 and HNF4A). Within the latter SCC, the DEPC, named as circuit 0,
299 is the one with the highest interface out-degree of 20. Then, we proceeded to identify
300 reprogramming determinants by targeting both master regulatory circuits. Within circuit
301 2, the gene that contributes the most to the circuit retroactivity is FOXA2, with an
302 interface out-degree of 5. Within the circuit 0, HNF4A is the one with the highest
303 contribution to the circuit retroactivity with an interface out-degree of 9. Therefore, the
304 final combination of reprogramming determinants is HNF4A and FOXA2. Both genes
305 should be activated to trigger the transition from fibroblast to hepatocyte. This result is
306 supported by the work of Sekiya and Suzuki published in 2011[34]. These authors
307 experimentally validated three different combinations of two transcription factors able to
308 induce the transition from mouse fibroblast to hepatocyte, including HNF4A and
309 FOXA2. This cellular transition constitutes a good example of reprogramming cells
310 without a common direct precursor (interlineage transdifferentiation).
311 Details about attractors, circuits and genes interface out-degree o for the three biological
312 examples are included in the supplements.

Discussion

Cellular reprogramming, including the conversion of one differentiated cell type to another (trans-differentiation) or to a more immature cell (dedifferentiation), constitutes an invaluable tool for studying cellular changes during development and differentiation, and has an enormous relevance for regenerative medicine and disease modeling. Although, substantial progress has been made in developing experimental reprogramming techniques, to date the scientific community is still faced with challenges such as the identification of optimal sets of genes whose repression and/or activation are capable of reprogramming one cell type to another (reprogramming determinants), and the elucidation of molecular changes and relevant pathways involved in these transitions (9). Furthermore, there is currently no methodology able to systematically predict reprogramming determinants that could guide the design of cellular reprogramming experiments. The development of computational models of transcriptional regulation that underlies cellular transitions would help to predict these reprogramming determinants. Moreover, the analysis of gene regulatory network properties has allowed the identification of functionally relevant motifs of interactions that could play a role in cellular transitions. In particular, transcription factor cross-antagonism has been described as a mechanism that plays a key role in cell fate decisions. A bistable toggle switch constitutes a molecular cross-repression motif that determines cellular commitment and provides stability to gene regulatory networks underlying transcriptional programs of binary decision cell choices. Experimental evidences indicate that flipping the stable states of these toggle switches produces interconversion between binary decision choices. Nevertheless, interlineage transdiferentiation and dedifferentiation

could involve perturbation of combinations of cross-repression motifs together with other multistable motifs. Here we propose a method, which considers the connectivity of these different multistable motifs, in order to systematically identify sets of reprogramming determinants able to induce transitions from differentiated cells to other cell types, either to progenitor cells (dedifferentiation) or to other differentiated cell types (transdifferentiation). Our strategy rests on the identification of a subset of all network positive circuits (necessary condition for network multistability), whose genes are differentially expressed between the cellular states involved in these. We termed this subset as differentially expressed positive circuits (DEPC). Further, a hierarchical organization of these circuits allows us to detect master regulatory positive circuits, which directly or indirectly regulate the states of the other DEPCs. By focusing on genes belonging to these master regulatory circuits, we dramatically reduced the number of possible combinations of reprogramming determinants.

However, some of these gene combinations in master regulatory DEPCs are more influenced by expression noise, affecting signal transmission in gene regulatory networks, and consequently decreasing reprogramming efficiency and fidelity. This is due to the fact that they are participating in a bigger number of regulations, so a limited concentration of the gene product has to interact with several targets a part from the one that closes the DEPC. In other words, the gene product has to distribute to different regulated targets, so the probability that the DEPC signal feed-back is broken by chance is higher (neglecting considerations about different molecular affinities that are assumed similar). Hence, in order to increase signal transmission our method proposes these gene combinations as reprogramming determinants. It is worth mentioning that we have

359 considered in our model some of the important events influencing reprogramming
360 efficiency and fidelity, such as the role of noise in network dynamics and the regulatory
361 interactions played by miRNAs. However, other factors, such as epigenetic modifications
362 that block activation of certain genes can affect the expected network behavior after
363 specific perturbations. Furthermore, it has been experimentally shown that epigenetic
364 modifications can prevent cellular reprogramming reversibility in some cases [43]. In
365 addition, our model does not take into account different delays in time response of
366 distinct regulatory interactions. Nevertheless, given that the purpose of our method is the
367 identification of reprogramming determinants, rather than a detailed description of
368 network dynamics, we consider that our model provides reasonable predictions. More
369 accurate predictions shall require addressing these considerations in the future.
370 Interestingly, despite there was no methodological constraint or theoretical limitation to
371 prevent that genes non-transcription factor are reprogramming determinants, to date, in a
372 blind application of the method, TFs always came up as reprogramming determinants.
373 It is worth mentioning that applicability of the method presented here is restricted to
374 cellular transitions between stable states or stable expression patterns and constraint by
375 the availability of information to reconstruct the corresponding GRN, as it is explained in
376 more detailed in methods' section.
377 Thus, our method constitutes the first strategy that systematically provides lists of
378 combinations of reprogramming determinants for cellular reprogramming events
379 involving two given cellular phenotypes without prior knowledge on potential candidates
380 and pathways involved. Due to that, the method is easily exportable to different
381 biological systems, providing guidance even without having expertise in a biological

process. In particular, this method is suitable for cellular transdifferentiation, especially when transitions occur between different cellular lineages. Indeed, interlineage transdifferentiation involves significant changes in several molecular mechanisms that increase the complexity of this type of reprogramming, and therefore hinders the prediction of reprogramming determinants.

Hence, given the increasing interest in various applications of cellular reprogramming in medicine and basic research, our method represents a useful computational methodology to assist researchers in the field in designing experimental strategies, especially when very little about a specific biological system is known.

Methods

Networks reconstruction

Among the selected biological examples, Th2-Th1 and Myeloid-Erythroid reprogramming illustrate the case of transdifferentiation between two cell types sharing a direct common precursor. We based our analysis on previously published GRNs describing the regular differentiation process of T-helper and cell fate decisions during hematopoiesis[35-37]. These two published network were enriched with miRNA interactions experimentally validated and publicly available in two different databases: TransmiR[40] and miRTarBase[41], including information about miRNA regulatory genes and miRNA regulated genes respectively. Only miRNA forming closed loops with network genes and, therefore, able to affect the stability of the network were included (see table 1).

The Fibroblast-Hepatocyte reprogramming example illustrates a distant (interlineage) cellular transdifferentiation. Therefore, no canonical previously published network can be exploited to detect the reprogramming determinants. Such reprogramming requires the reconstruction of a GRN contextualized to this specific cellular transition.

Given that the final goal is to induce the transition from one specific cell phenotype to one another, the network is constructed based on changing elements between these two states, i. e., differentially expressed genes (DEG) between these two conditions or cell types obtained from microarray experiments. We scanned the literature and collected 24 genes known to play a relevant role in liver development and function and differentially expressed when comparing fibroblasts and hepatocytes according to previous works [44-47] . We proceed to try to connect these genes using interactions obtained from literature harvested from the entire PubMed. For this specific purpose we used the information contained in the ResNet mammalian database from Ariadne Genomics (http://www.ariadnegenomics.com/). The ResNet database includes biological relationships and associations, which have been extracted from the biomedical literature using Ariadne's MedScan technology[48,49]. More specifically, we included interactions annotated in the ResNet mammalian database in the category of Expression, PromotorBinding and Regulation. In the Expression category interactions indicates that the regulator changes the protein level of the target, by means of regulating its gene expression or protein stability. In the PromotorBinding category interactions indicates that the regulator binds the promotor of the target. Finally, in the Regulation category interactions indicates that the regulator changes the activity of the target. Similar resources for network reconstruction are the IPA tool of Ingenuity Systems

427 (http://www.ingenuity.com/) and the Transfac tool (http://www.biobase-
428 international.com).
429 Once we had a raw GRN from literature, we proceed to remove interactions inconsistent
430 with expression data by an iterative network pruning. These removals represent
431 interactions apparently not active in the biological context under study. It should be taken
432 into account that interactions from literature usually come from different biological
433 contexts as cell types, tissues or even species. This network pruning allows us to reduce
434 the amount of "false" interactions and to obtain a contextualized network. The algorithm
435 applied for this network pruning[39] was originally conceived to predict missing expression
436 values in gene regulatory network, but could be applied to contextualize the network
437 when all the expression values in two given cellular phenotypes or stable transcriptional
438 programs are known. Basically, the algorithm exploits the consistency between predicted
439 and known stable states from experimental data to guide the iterative network pruning
440 that contextualizes the network to the biological conditions under which the expression
441 data were obtained. This process implies the booleanization of cellular phenotypes
442 coming from experimental expression data; genes considered as up-regulated and down-
443 regulated for a given p-value (usually < 0.05 for a regular t-test) are assumed as "1" and
444 "0" respectively. This is due to the fact that a Boolean model is assumed to compute
445 network attractors. An evolutionary algorithm, more specifically an estimation of
446 distributions algorithm (EDA) [50] samples the probability distribution of positive
447 feedback loops or positive circuits and individual interactions within the subpopulation of
448 the best-scored networks at each iteration of the pruning algorithm. The resulting
449 contextualized network is based not only on previous knowledge about local connectivity

but also on a global network property (stability) providing robustness in predictions (the remaining set of interactions) against noisy sources of information and network incompleteness. Despite we tried to enrich this network with miRNA interactions as we did in the two previous examples, none miRNA involved in regulatory loops or circuits with genes differentially expressed were found experimentally validated for mouse. More details about network reconstruction process for the Fibroblast-Hepatocyte reprogramming example are included in the supplementary information.

Main properties of these three biological examples GRN are shown in table 2.

Network transformation in a directed acyclic graph (DAG)

The first step of the method, named as "Detecting master regulatory SCCs" in results section, requires the hierarchical analysis of a subnetwork of the complete GRN including only DEPCs and all genes and interactions connecting them. This subnetwork contains positive feed-back loops, so it should be transformed in order to be able to analyze its hierarchy. The transformation of this subnetwork of connected DEPCs in a DAG was performed by contraction of DEPCs strongly connected, i e, SCCs of differentially expressed genes, in single super-nodes. This network transformation allows the hierarchical analysis of the network following the method described by Jothi et al. [51], resulting in the location of SCCs at different levels of hierarchy with the subsequent identification of master regulators SCCs on the top of the hierarchy pyramid.

During the application of this network transformation to the three examples included in this work we also forced the method to work on differentially expressed negative circuits (DENC) instead of DEPCs to illustrate the failure of the method when a wrong stability element is considered. Interestingly, we could not found any single DENC in none of the

three examples, despite the relative abundance of negative circuits in the three GRNs (17, 11 and 11 for Th2-Th1, Myeloid-Erythroid, and Fibroblast-Hepatocyte respectively, whereas the corresponding number of positive circuits are 29, 25 and 19). Consequently, it was not possible to perform the network transformation in a DAG and the subsequent hierarchical analysis because there was no SCC of negative circuits to analyze. This finding is consistent with the role of positive circuits or positive feed-back loops as cornerstone of multi-stable behavior in networks of interacting elements.

Circuits' detection

The Johnsons algorithm [52] was implemented to detect all elementary feedback circuits in the network. A feedback circuit is a path in which the first and the last nodes are identical. A path is elementary if no node appears twice. A feedback circuit is elementary if no node but the first and the last appears twice. Once we have all elementary feedback circuits, we select positive feedback circuits, or feedback circuits for which the difference between the number of activating edges and the number of inhibiting edges is even. Both elementary feedback circuit detection, positive feedback circuits sorting and DEPFCs detection were implemented in Perl.

Attractor computation

We assumed a Boolean model to compute attractors with a synchronous updating scheme [53] and using our own implementation[39] of the algorithm described by Garg *et al.*, 2007 [54]. The logic rule applied by default is the following: if none of its inhibitors and at least one of its activators is active, then a gene becomes active; otherwise the gene is inactive. If different regulatory rules are known for specific genes, this knowledge can be included in the model. Results in the attractor computation were consistent with the results obtained

using previously published software to compute attractors in Boolean systems (Boolnet [55], GenYsis[54]).

Minimal input data for the method usage and limitations

Given that our methodology considers transitions between attractor states, it requires the availability of expression data of stable cellular phenotypes. In addition, if the GRN has been experimentally validated and its attractors are consistent with the cellular phenotypes under study, our methodology is readily to be applied. Otherwise, the GRN has to be reconstructed from publicly available data, and therefore the applicability of our methodology could be limited by the availability of information. In this case, the reliability of the resulting GRN can be estimated by evaluation of how well the stable states of this network coincide with the experimental expression data. We usually assumed a threshold of 70 % to consider a GRN worth to be processed. For instance, in the Fibroblast-Hepatocyte example after the network contextualization process, the attractor computation of the resulting GRN revealed a matching with the expression data of 76 % for both conditions (fibroblast and hepatocytes), meaning that 76 % of gene expression values in the network are well predicted for these two conditions. The remaining 24 % of the gene expression values are not well predicted due to two different possibilities: incompleteness of the network or wrong assumed regulatory rules in specific cases. It is worth noticing that our method for contextualizing GRNs rests on removal of inconsistent regulatory interactions rather than on the addition of new interactions, and therefore the possibility of adding new predicted interactions could improve the description of the expression data. This is a very interesting and very relevant point, and despite it is out of the scope of the present work, and the fact that it

519 constitutes a challenging computational problem, it should be definitely pursued in order
520 to improve our methodology**.**
521
522 **References**
523
524
525 **Acknowledgements**
526 Funding for open access charge: Luxembourg Centre for Systems Biomedicine (LCSB),
527 University of Luxembourg.
528
529 **Author contributions**
530 I. C. and A. dS. conceived the idea for the paper. I. C. wrote software, performed the
531 experiments and analyzed the data. I. C. and A. dS.contributed to writing the paper. A.
532 dS. coordinated and supervised the project.
533 **Additional information**
534 Supplementary Information accompanies this paper on Supplementary_file_1.docx and
535 Supplementary_file_2.xlsx.
536 **Competing financial interests:** The authors declare no competing financial interests.
537
538 **Figure Legends**

Figure 1| Cell identity cascading landscape representing the cellular transcriptional program. Paths between pluripotent and differentiated cells, representing cellular

differentiation process pass through stable expression profiles corresponding to multipotent progenitors. Binary cell fate decisions at multipotent rogenitor level are characterized by cross-repression motifs of competing transcription factors. Transdifferentiation between somatic cells are divided in those sharing a direct precursor cell (intra-lineage transdifferentiation), where cross-repression motifs, which determine cell fate decision, play a key role in stabilizing binary cell decisions and transitions between them; and those without a direct precursor (inter-lineage transdifferentiation), characterized by a more complex molecular mechanism underlying cellular transitions. Blue and red colors in cross-repression motifs and GRN stability core represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

Figure 2| Design of cellular reprogramming protocol in three steps. a) Detecting master regulatory strongly connected components (SCCs). In this first step, those positive circuits or positive feed-back loops in the gene regulatory network (GRN) whose genes change their expression levels between two cellular phenotypes are selected from the population of network circuits. These differentially expressed positive circuits (DEPCs) form SCCs. A hierarchical analysis in the space of these SCCs allows us to determine master regulatory SCCs. SCC 1 and 2 are located on the top of the hierarchy of the represented toy network without displaying connectivity between them. These SCCs should be independently perturbed to guarantee that the perturbation signal reaches every DEPC in the GRN. **b)** Detecting master regulatory DEPCs. Within each master

regulatory SCC, a master regulatory DEPC is determined based on a retroactivity score (interface out-degree) or, in other words, based on the number of genes directly regulated by this circuit. The master regulatory DEPC is the one with the highest interface-out degree. In this toy example, Circuit 1 (composed by genes 'a', 'b' and 'c') is the master regulatory DEPC of the SCC 1, and Circuit 1 (composed by genes 'p' and 'o') of SCC 2 is the other master regulatory DEPC . These master regulatory DEPCs are colored in red in the retroactivity ranking table. **c)** Detecting reprogramming determinants. Once the master regulatory DEPCs have been determined, the selection of final reprogramming determinants is based on maximizing the sum of individual gene interface out-degrees included in the combination. In this toy example, gene 'a' is the one with highest retroactivity within the Circuit 1 of the SCC 1. Similarly, gene 'p' has the highest interface out-degree in its respective circuit and SCC. Therefore, the reprograming determinants are 'a' and 'b' (both should be perturbed to induce the hypothetical cellular transition). Blue and red colors in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

Figure 3| Reprogramming determinants in three illustrative biological examples. a) Th2-Th1 reprogramming. Activation of T-bet and, alternatively, inhibition of GATA-3 are predicted as effective perturbations to induce this cellular transition. **b)** Cellular reprogramming from myeloid to erythroid cells. Both, activation of GATA-1 or

inhibition of PU.1 are predicted as independently able to induce this cellular transition. **c)** Cellular reprograming from fibroblast to hepatocyte. In this particular case no single gene is able to induce the cellular transdifferentiation according to our predictions. On the other hand, combined activation of HNF4A and FOXA2 is predicted as an effective combination of reprogramming determinants. Blue and red colors in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

539 **Tables**

	miRNA	Interaction
Th2-Th1	**1. mir-145**	• **IFN-B -> mir-145** • **mir-145 -| STAT1**
Myeloid-Erythroid	**1. mir-34a** **2. mir-155**	• **mir-34A -| PU.1** • **CEBPA -> mir-34A** • **mir-155 -| FLI1** • **PU.1 -> mir-155** • **mir-155 -| PU.1**

540

541 **Table 1| miRNAs included in the biological examples. '->'** represents activation and **'-|'**

542 represents inhibition.

543

	Genes	Interactions	Activations	Inhibitions	miRNA
Th2-Th1	**24**	**38**	**28**	**10**	**1**
Myeloid-Erythroid	**13**	**34**	**19**	**15**	**2**
Fibroblast-Hepatocyte	**27**	**56**	**46**	**10**	**0**

544

545 **Table 2| Main properties of the gene regulatory networks of the three biological examples.**

546

547

548 1 Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587-594,

549 doi:10.1038/nature08533 (2009).

2 Cantor, A. B. & Orkin, S. H. Hematopoietic development: a balancing act. *Curr Opin Genet Dev* **11**, 513-519 (2001).

3 Graf, T. Differentiation plasticity of hematopoietic cells. *Blood* **99**, 3089-3101 (2002).

4 Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631-644, doi:10.1016/j.cell.2008.01.025 (2008).

5 Arinobu, Y. Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell stem cell* **1**, 416-427 (2007).

6 Iwasaki, H. & Akashi, K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**, 726-740 (2007).

7 Zhou, L. TGF-[bgr]-induced Foxp3 inhibits TH17 cell differentiation by antagonizing ROR[ggr]t function. *Nature* **453**, 236-240 (2008).

8 Laslo, P. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126**, 755-766 (2006).

9 Frontelo, P. Novel role for EKLF in megakaryocyte lineage commitment. *Blood* **110**, 3871-3880 (2007).

10 Hwang, E. S., Szabo, S. J., Schwartzberg, P. L. & Glimcher, L. H. T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. *Science* **307**, 430-433 (2005).

11 Heins, N. Glial cells generate neurons: the role of the transcription factor Pax6. *Nature Neurosci.* **5**, 308-315 (2002).

12 Kajimura, S. Regulation of the brown and white fat gene programs through a PRDM16/CtBP transcriptional complex. *Genes Dev.* **22**, 1397-1409 (2008).

13 Niwa, H. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**, 917-929 (2005).

14 Ralston, A. & Rossant, J. Genetic regulation of stem cell origins in the mouse embryo. *Clin. Genet.* **68**, 106-112 (2005).

15 Waddington, C. H. (Macmillan Publishers Limited. All rights reserved, 1957).

16 Kauffman, S. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437-467 (1969).

17 Kauffman, S. (Macmillan Publishers Limited. All rights reserved, 1993).

18 Szabo, S. J. *et al.* A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655-669 (2000).

19 Heyworth, C., Pearson, S., May, G. & Enver, T. Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *Embo J* **21**, 3770-3781, doi:10.1093/emboj/cdf368 (2002).

20 Thomas, R., Thieffry, D. & Kaufman, M. DYNAMICAL BEHAVIOR OF BIOLOGICAL REGULATORY NETWORKS .1. BIOLOGICAL ROLE OF FEEDBACK LOOPS AND PRACTICAL USE OF THE CONCEPT OF THE LOOP-CHARACTERISTIC STATE. *Bull. Math. Biol.* **57**, 247-276, doi:10.1007/bf02460618 (1995).

21 Lukk, M. *et al.* A global map of human gene expression. *Nat Biotech* **28**, 322-324, doi:http://www.nature.com/nbt/journal/v28/n4/abs/nbt0410-322.html#supplementary-information (2010).

22 Muller, F.-J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nat Meth* **8**, 315-317,

doi:http://www.nature.com/nmeth/journal/v8/n4/abs/nmeth.1580.html#supplementary-information (2011).

23 Dudley, J. T., Tibshirani, R., Deshpande, T. & Butte, A. J. Disease signatures are robust across tissues and experiments. *Molecular systems biology* **5**, doi:http://www.nature.com/msb/journal/v5/n1/suppinfo/msb200966_S1.html (2009).

24 Ding, S. & Wang, W. Recipes and mechanisms of cellular reprogramming: a case study on budding yeast Saccharomyces cerevisiae. *BMC Syst Biol* **5**, 50, doi:10.1186/1752-0509-5-50 (2011).

25 Chang, R., Shoemaker, R. & Wang, W. Systematic search for recipes to generate induced pluripotent stem cells. *PLoS Comput Biol* **7**, e1002300, doi:10.1371/journal.pcbi.1002300 (2011).

26 Chickarmane, V. & Peterson, C. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS One* **3**, e3478, doi:10.1371/journal.pone.0003478 (2008).

27 Flottmann, M., Scharp, T. & Klipp, E. A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front Physiol* **3**, 216, doi:10.3389/fphys.2012.00216 (2012).

28 MacArthur, B. D., Please, C. P. & Oreffo, R. O. Stochasticity and the molecular mechanisms of induced pluripotency. *PLoS One* **3**, e3086, doi:10.1371/journal.pone.0003086 (2008).

29 Wang, J., Xu, L., Wang, E. & Huang, S. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* **99**, 29-39, doi:10.1016/j.bpj.2010.03.058 (2010).

619 30 Del Vecchio, D., Ninfa, A. J. & Sontag, E. D. Modular cell biology: retroactivity and
620 insulation. *Molecular systems biology* **4**, 161, doi:10.1038/msb4100204 (2008).
621 31 Chalancon, G. *et al.* Interplay between gene expression noise and regulatory network
622 architecture. *Trends Genet* **28**, 221-232, doi:10.1016/j.tig.2012.01.006 (2012).
623 32 Lee, H. J. *et al.* GATA-3 induces T helper cell type 2 (Th2) cytokine expression and
624 chromatin remodeling in committed Th1 cells. *J Exp Med* **192**, 105-115 (2000).
625 33 Hwang, E. S., Szabo, S. J., Schwartzberg, P. L. & Glimcher, L. H. T helper cell fate
626 specified by kinase-mediated interaction of T-bet with GATA-3. *Science* **307**, 430-433,
627 doi:10.1126/science.1103336 (2005).
628 34 Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells
629 by defined factors. *Nature* **475**, 390-393, doi:10.1038/nature10263 (2011).
630 35 Mendoza, L. A network model for the control of the differentiation process in Th cells.
631 *Biosystems* **84**, 101-114, doi:10.1016/j.biosystems.2005.10.004 (2006).
632 36 Krumsiek, J., Marr, C., Schroeder, T. & Theis, F. J. Hierarchical differentiation of
633 myeloid progenitors is encoded in the transcription factor network. *PLoS One* **6**, e22649,
634 doi:10.1371/journal.pone.0022649 (2011).
635 37 Dore, L. C. & Crispino, J. D. Transcription factor networks in erythroid cell and
636 megakaryocyte development. *Blood* **118**, 231-239, doi:10.1182/blood-2011-04-285981
637 (2011).
638 38 Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by
639 defined factors. *Nature* **475**, 386-389, doi:10.1038/nature10116 (2011).

640 39 Crespo, I., Krishna, A., Le Bechec, A. & del Sol, A. Predicting missing expression values
641 in gene regulatory networks using a discrete logic modeling optimization guided by
642 network stable states. *Nucleic Acids Res* **41**, e8, doi:10.1093/nar/gks785 (2013).
643 40 Wang, J., Lu, M., Qiu, C. & Cui, Q. TransmiR: a transcription factor–microRNA
644 regulation database. *Nucleic Acids Res* **38**, D119-D122, doi:10.1093/nar/gkp803 (2010).
645 41 Hsu, S.-D. *et al.* miRTarBase: a database curates experimentally validated microRNA–
646 target interactions. *Nucleic Acids Res*, doi:10.1093/nar/gkq1107 (2010).
647 42 Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nat Rev Immunol*
648 **2**, 933-944, doi:10.1038/nri954 (2002).
649 43 Jopling, C., Boue, S. & Izpisua Belmonte, J. C. Dedifferentiation, transdifferentiation and
650 reprogramming: three routes to regeneration. *Nat Rev Mol Cell Biol* **12**, 79-89,
651 doi:10.1038/nrm3043 (2011).
652 44 Kyrmizi, I. *et al.* Plasticity and expanding complexity of the hepatic transcription factor
653 network during liver development. *Genes Dev* **20**, 2293-2305, doi:10.1101/gad.390906
654 (2006).
655 45 Zaret, K. S. Genetic programming of liver and pancreas progenitors: lessons for stem-cell
656 differentiation. *Nat Rev Genet* **9**, 329-340, doi:10.1038/nrg2318 (2008).
657 46 Schrem, H., Klempnauer, J. & Borlak, J. Liver-enriched transcription factors in liver
658 function and development. Part I: the hepatocyte nuclear factor network and liver-specific
659 gene expression. *Pharmacol Rev* **54**, 129-158 (2002).
660 47 Schrem, H., Klempnauer, J. & Borlak, J. Liver-enriched transcription factors in liver
661 function and development. Part II: the C/EBPs and D site-binding protein in cell cycle

control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation. *Pharmacol Rev* **56**, 291-330, doi:10.1124/pr.56.2.5 (2004).

48 Novichkova, S., Egorov, S. & Daraselia, N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* **19**, 1699-1706 (2003).

49 Daraselia, N. *et al.* Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604-611, doi:10.1093/bioinformatics/btg452 (2004).

50 Armananzas, R. *et al.* A review of estimation of distribution algorithms in bioinformatics. *BioData Min* **1**, 6, doi:10.1186/1756-0381-1-6 (2008).

51 Jothi, R. *et al.* Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular systems biology* **5**, 294, doi:10.1038/msb.2009.52 (2009).

52 Johnson, D. B. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing,* **4**, 77-84 (1975).

53 Garg, A., Di Cara, A., Xenarios, I., Mendoza, L. & De Micheli, G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics* **24**, 1917-1925, doi:10.1093/bioinformatics/btn336 (2008).

54 Garg, A., Xenarios, I., Mendoza, L. & DeMicheli, G. Vol. 4453 *Lecture Notes in Computer Science* (eds Terry Speed & Haiyan Huang) 62-76 (Springer Berlin / Heidelberg, 2007).

55 Mussel, C., Hopfensitz, M. & Kestler, H. A. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26**, 1378-1380, doi:10.1093/bioinformatics/btq124 (2010).

685