NETWORK ANALYSIS TO MODEL DISEASES AND CELLULAR REPROGRAMMING FOR THERAPEUTIC INTERVENTION

Isaac CRESPO

PhD- FSTC-2013-30
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on 15/11/2013 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Isaac Crespo
Born on 21 April 1976 in Madrid (Spain)

NETWORK ANALYSIS TO MODEL DISEASES AND CELLULAR REPROGRAMMING FOR THERAPEUTIC INTERVENTION

Dissertation defense committee

Dr Antonio del Sol, dissertation supervisor
Professor, Université du Luxembourg

Dr Reinhard Schneider
Université du Luxembourg

Dr Rudi Balling, Chairman
Professor, Université du Luxembourg

Dr Noel Buckley
Professor, King's College London

Dr Angela Bithell
Lecturer, University of Reading

Contents

Abstract

Applications of network analysis to the study of disease can be divided into two main categories: disease description, including characterisation, diagnosis and prognosis; and disease treatment, including drug target discovery and cellular reprogramming, together with its applications to regenerative medicine. In this dissertation, I will critically discuss some research projects on which I have been working during my PhD program. In correspondence with the two aforementioned categories, these projects can be broken down into two different blocks of content, with the common goal of acquiring insights into the study of disease.

In the first block of contents, corresponding with Chapter 2, I will explain and discuss novel strategies for network-based analysis and modelling which have been applied for disease description and characterisation in different case-studies, namely the metabolic syndrome, prion disease and the epithelial to mesenchymal transition in breast cancer. Indeed, these projects exploited the evolutionary conservation of motifs of regulatory interactions and consistency between computed and experimentally validated expression so as to reconstruct dynamical models and create a network-based characterisation of the corresponding systems.

With regards the second block of content, corresponding with Chapter 3, I explain and discuss novel computational methods which have been developed during my PhD program to address the task of the artificial induction of cellular reprogramming; something with a wealth of potential applications when it comes to the creation of disease models and in the field of regenerative medicine.

Within the general conclusion discussion focuses on the fact that, although the methodology explained in this work was developed in the context of disease study, one may find the application of some of these ideas and strategies fitting for other problems. Indeed, the same principles applied to detect driver genes capable of changing the cell phenotype when perturbed can also be applied to control biological living systems for basic research or industrial purposes. These principles could also be potentially extended to higher level systems than the cellular level (tissue or cell population level).

Acknowledgments

Without the support of a number of people and institutions, it would not have been possible to write this dissertation. It is my pleasure to have the opportunity to express my gratitude to some of them here.

For my academic achievements, I would like to acknowledge my supervisor, Prof. Dr. Antonio del Sol for the opportunity to join his team and his constant support and guidance. I can say with certainty that he provided me with all I needed to do a good job and his door was always open to listening new proposals; only my own skills and working capacity limited my productivity. In addition, three years of regular brain storming and critical scientific discussion constituted an invaluable training period beyond tangible things such as academic achievements or publications.

I would also like to acknowledge the help of the members of the evaluation committee, Prof. Dr. Rudi Balling and Prof. Dr. A. B. (Guus) Smit. Their opinions as external observers regarding

the projects and topics finally addressed during my PhD were invaluable. Their suggestions have helped to define my training as a PhD student and the resulting PhD thesis.
Given that this thesis relates to the interdependence among elements, it would be inconsistent if the collaborative efforts which also contributed to the shape were not recognised. I would like to express my gratitude to our collaborators in Luxembourg (Dr. Michèle Moes, Dr. Antony Le Béchec, Prof. Dr. Evelyne Friederich, Dr. Jochen Scheneider and Dr. Alessandro Michelucci), in Japan (Dr. Hiroaki Kitano) and in the United Kingdom (Prof. Dr. Noel J. Buckley, Dr. Angela Bithell and Jannis Kalkitsas) as well as to the rest of my co-workers and co-authors.
I would also like to thank Luxembourg and its institutions, particularly the Luxembourg Centre for Systems Biomedicine (LCSB) and the University of Luxembourg, for funding and hosting me for three years. It has been a privilege to witness the first steps of LCSB and the work of all the staff members, headed by Prof. Dr. Rudi Balling, and to participate in the development of LCSB as an important international research centre. Both Luxembourg and the people I have met here will have a special place in my heart forever.
My peers have also been a great source of support. The scientific discussions, especially within the Computational Biology Group, have definitely contributed not only to improving the quality of the research work included in this dissertation but also to my enjoyment of Luxembourg.
On a social note, I would like to thank my family for their continuing support. Without them I would have lost focus. In particular I would like to thank Ana, my partner, for her understanding and patience, and publicly apologise to her for the uncountable hours during which I was not in a very communicative mode (in the best case). Whether it was being stuck with a problem or a software bug, this is something which has happened very frequently over the last three years.

Chapter 1. Introduction and overview of the research

1.1. Introduction

Network analysis and disease: background

Disease refers to any condition associated with dysfunction of an organism's normal homeostasis. There are four main types of disease: pathogenic disease, deficiency disease, hereditary disease and physiological disease. What all of these have in common is the highly complex multifactorial nature of the causes of disease pathogenesis. These factors include genetic variation, epigenetic modifications and genome-environment interactions. Indeed, it is not only the identification of factors but also elucidation of how they interact with each other which is essential to understanding the complex and adaptive behaviour of an organism in a pathological condition.

Network analysis applied to disease study constitutes an example of a systems biology approach to answering biological questions. Systems biology represents an integrative strategy which attempts to understand the higher-level operating principles of living organisms [1,2]. Systems biology employs tools developed in physics, mathematics and computer science such as graph theory, nonlinear dynamics, control theory and modelling of dynamic systems with a view to solving questions related to the complexity of living systems. This complexity term refers to the concept developed in Complexity theory [2,3,4]; a discipline of physics which

focusses on the so called emergent properties. These emergent properties are intrinsic properties of the system which cannot be inferred from the separated analysis of its constitutive elements. Both Complexity theory and Systems biology are deeply rooted in the not always intuitive idea that the whole is more than the sum of its parts. For instance, emergent properties in the central nervous system, such as the processes of learning, memory and emotions cannot be fully understood even with a detailed analysis of single neurons; it is necessary to consider these neurons integrated altogether in a network of interactions in order to understand how they perform these processes. From the mathematical point of view, a network or graph is the application of a set in itself, i.e., a collection of elements of the set and the binary relations between these elements. As a result, networks or graphs are topological rather than geometrical objects, with the most important feature the adjacency relationships between points. Within the context of Systems Biology, a network can be defined as any interconnected group of elements (system) which shares information. For instance, gene regulatory networks (GRNs) refers to a collection of genes in a cell which interact with each other through their RNA and protein expression products, thereby regulating the expression levels of genes in the cell. Network analysis provides a framework through which to deal with the complexity of biological systems both in physiological and pathological conditions. The application of graph theory's formalised language, as well as that of vector algebra and nonlinear dynamics makes it possible to describe and predict the behaviour of biological systems and to explore strategies which can be used to intervene in these systems in order to reestablish normal homeostasis or to delay disease progression. Network representation of biological systems encodes information regarding regulatory and signaling pathways which connect constituent parts like proteins, DNA, RNA or metabolites, in a compact way which can be directly visualised and analysed either by a human being or a computer. This analysis includes both the network topology [5] and dynamics [5,6]. The first refers to the pattern of connectivity between elements of the network in a static snapshot, whilst the latter refers to the evolution of the network in time, assuming a given dynamical model. Changes observed in the network in time can be changes in node states (representing variation in concentrations or level of activity depending on the network) as well as changes in the network topology when certain interactions only occur under specific conditions which change during the described process. These network changes can be triggered by the disease factors mentioned above (genetic, epigenetic or environmental), namely disease pathogenesis characterised by both the direct effect of the pathogen agent (if any) and the network response. From this perspective, diseases can be viewed as specific types of network perturbation [7].

Disease as perturbation of cellular systems

Cells should be able to adapt to changing conditions, including environmental variation, such as temperature, pH or chemicals, as well as noise in biochemical processes, such as transcription and translation, which are inherently stochastic. Consequently, the underlying networks of interactions which rule cellular processes have developed, during evolution, a certain level of complexity. This complexity is used to provide cells with the flexibility to respond and adapt to new situations, as well as the robustness to preserve the integrity of cellular functions despite internal fluctuations in gene expression, protein and metabolites concentration and

environmental conditions variation. This combination of flexibility and robustness means that cells are able to exist in a discrete number of stable cellular phenotypes which correspond to stable states (attractor states) of the dynamic system on the basis of an underlying network of regulatory interactions. The existence of attractor states and their association with cellular phenotypes has been experimentally verified [8]. It is also worth mentioning that, despite the fact that thousands of genes characterise a specific cellular phenotype, an increasing number of examples show that the perturbation of few genes (1 to 5) effectively induce cellular reprogramming [9,10,11,12], i.e., induce cells to transit from one attractor to another. A popular framework for conceptualising and describing changes in cellular phenotypes is that of the landscape proposed by Waddington [13,14,15], where cellular phenotypes may be seen as attractors of gene regulatory networks represented as wells separated by the so-called epigenetic barriers. It is worth mentioning that for Waddington the meaning of the term "epigenetic" was different from that currently used by molecular biologists to describe covalent chromatin and DNA modifications; it referred to something closer to the physicist's concept of an "epigenetic state"[16]; a systems-level stable state which arises from genetic interactions. These barriers are established by those elements stabilising the network in their attractors. The process of changing the transcriptional program from one cellular phenotype to another corresponds to trajectories between wells in the gene expression state space described by Waddington's landscape and with the sequence of transient states followed when moving from one attractor to another in a network based dynamical model (see Figure 1).

In the context of disease, a dual robustness against internal or external fluctuations as well as fragility against specific perturbations can explain the formation of certain diseases [17,18]. Within the gene regulatory network landscape, diseases can be conceptualised as the result of induced transitions to a pre-existing disease attractor or the effect of changing the landscape with the creation of new disease attractors where the system remains. Although these two ideas may initially seem conceptually very different, they converge on the concept of disease as an aberrant stable cellular phenotype pointing to the idea that the same robustness which prevents the system from exiting a healthy state (something which sounds reasonable to be evolutionary conserved) provides the system with a certain resilience to emerge from the disease state; after all, the same molecular mechanisms are operating when it comes to the configuration of the general landscape, including both health and disease states. The concept of pre-existing disease attractors has been proposed in the context of cancer study as cancer attractors [19,20]. This concept constitutes an explanation of some interesting issues, such as the fact that there is no "continuum" of tumor transcriptomes, but instead a discrete set of cancer cells phenotypes, or the fact that oncogenesis recapitulates ontogenesis. These issues are not easily explained solely by the accumulation of random mutations scattered throughout the genome. These silent disease attractors resemble sleeping pseudogenes which can eventually be reactivated by sporadic mutations or triggering factors.

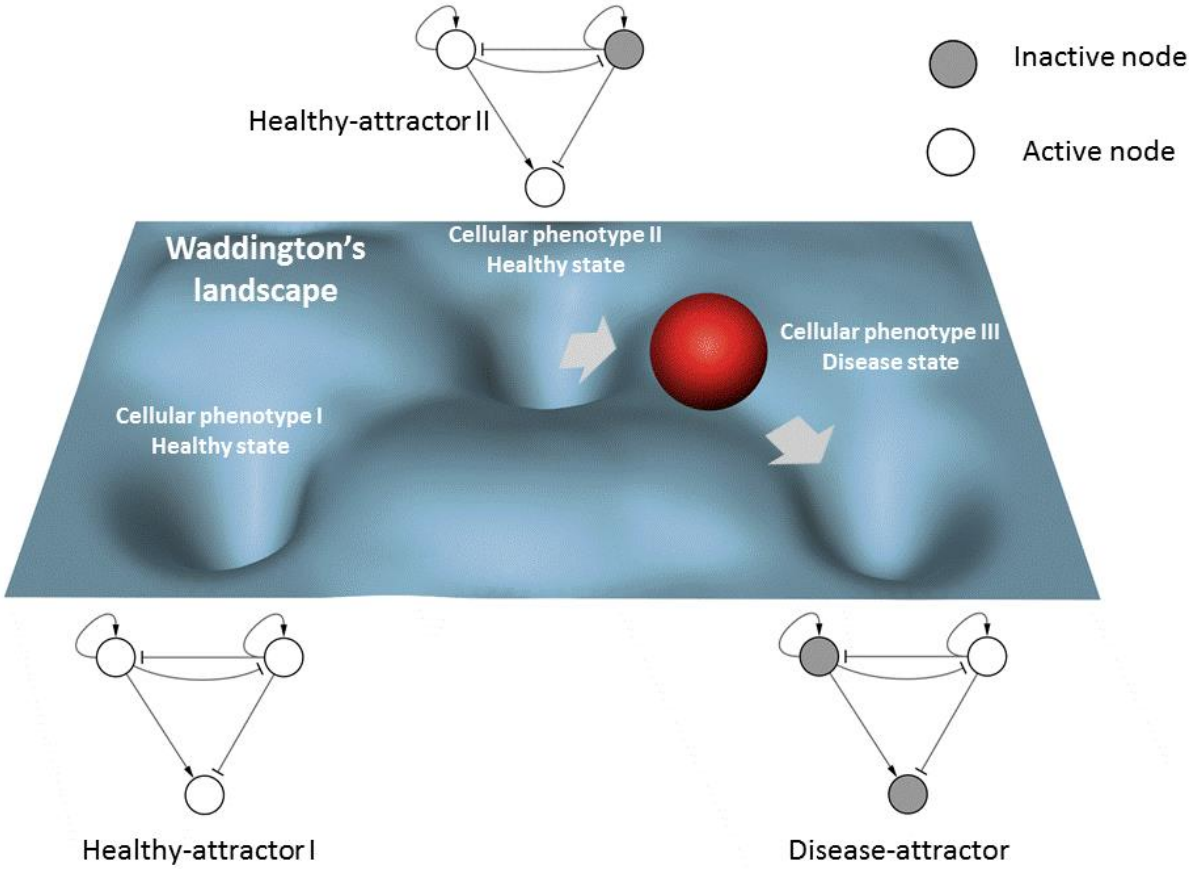


Figure 1 | Biological processes can be defined as transitions between stable states and are frequently described using Waddington's landscape or network-based models. Within Waddington's landscape, wells represent stable steady states (usually termed attractors) whilst separation between wells represents the so called "epigenetic barriers". When assuming a dynamical network-based model, such as Boolean based on logic functions or continuous based on ordinary differential equations, we can describe the evolution of the network state in time, and network attractors are self-maintained network states or states consistent with the regulatory logic. Disease can be described with these models as attractors where the system fall when intrinsic or extrinsic perturbation let it go out from the basal or healthy stable state.

Network analysis and disease: previous works

Applications of network analysis in disease study can be divided into two main categories: disease description; and disease treatment. Within the category of disease description we find several areas including the identification and characterisation of new disease genes, network based disease classification and identification and study of disease related subnetworks and their properties. Within the category of disease treatment we find a systems approach to simple- and multi-target drug design and cellular reprogramming for cell therapy. It is worth mentioning that there are also a handful of pioneering works using network analysis to address

the study of interactions between different diseases, resulting in, for example, comorbidity effects, instead of focusing on the study of a single disease. Following this idea, Goh and co-workers [21] reconstructed two networks: in the first, diseases as nodes were linked by mean of common implicated genes, whilst with the second, disease related genes were linked through diseases. These two networks provide a theoretic framework to reveal some of the common genetic origins of diseases. The authors found the existence of disease-specific functional modules and also discovered that the majority of human disease-related genes are not-essential and do not encode hub proteins. A selection based model was proposed to explain differences between essential and disease genes. Park and coworkers explored the impact of cellular networks on disease comorbidity by integrating information on cellular interactions, disease-genes relationship and population-level Medicare date. They found correlations between disease comorbidity and the structure of cellular networks [22].

Identification and characterisation of new disease genes

Whilst many different strategies have been employed for the investigation of disease related genes, all of them rest on two basic assumptions: first, genes involved in disease pathogenesis are not randomly positioned in the network. They tend to be in a central network position, clustering together and exhibiting high connectivity [21,23,24]; and second, the network neighbour of a disease-causing gene is likely to be involved in the same or similar disease [21,25,26,27]. These two assumptions arose following a network analysis of the first eukaryotic gene regulatory network available (yeast), thus showing that genes associated with a particular phenotype or function (especially those essential for this function), such as growth or response to genotoxic agents, have special network properties [28,29,30].

Genes up-regulated in lung squamous cancer tissues were reported highly connected and central in a human PPI network [23]. The authors justified this finding by suggesting that these genes are essential for tumor growth, and as such share network properties of essential genes.

Network analysis of genes included in a comprehensive census of human cancer genes [31] and revealed that genes related to cancer tend to cluster in the network [24]. In addition, such genes tended to participate in more clusters when compared to non-cancer genes.

Network analysis based on information pertaining to human genetic diseases in the database of Online Mendelian Inheritance in Man [32] showed that disease genes exhibit a tendency to interact with each other and to be co-expressed in the same tissue [21]. Interestingly, the study found that most disease genes are nonessential and have no tendency toward higher degrees in the PPI network. Discrepancy between this finding and the work in lung cancer described above [23] could be due to the fact that the first study investigated disease in general instead of focussing solely on cancer. Despite this discrepancy, it should be noted that the overall conclusion of these works is that genes related with a particular function or cellular phenotype, including disease pathogenesis, are not randomly distributed in the network; they tend to cluster and to be positioned centrally in the network.

Scanning topological properties can be useful when it comes to narrowing down the list of disease related genes when no prior knowledge is available. In cases where certain causative

genes have been identified in genetically heterogeneous diseases, another network analysis based strategy for disease related genes discovery can be used. Oti and coworkers developed a strategy centred around finding disease-causative genes when only some of the causative genes are known and for other genetic factors only locus information is available [25]. They showed that predictions based on connectivity between known causative genes and candidates were better than random selection of genes at the same locus (10-fold).

Based on the same principles, 2006 saw Franke and coworkers publish an algorithm to rank a set of candidate disease-causing genes in multiple susceptibility loci for further sequence or association analysis [27]. They constructed a network of computationally predicted relations as well as known (from literature) molecular interactions. The network was used to score candidate genes based on their interactions, assuming that for a given disease causative genes will be involved in only a few distinct biological pathways. This assumption implies the clustering of genes from different susceptibility loci, resulting in shorter network distances between disease genes than expected by chance (random selection of genes in the network).

Information pertaining to related diseases can also be exploited in the search for disease-causing genes. Along similar lines, and published in 2007, Lage and coworkers developed a strategy [26] consisting of ranking each candidate gene according to what they called the phenotype similarity score of diseases associated with the gene and its neighbours in a PPI network. Simply put, if a gene and/or its neighbours in the network are associated with a disease responsible for a specific phenotypic outcome, it is likely to be involved in other diseases with similar or identical phenotypic outcome.

In general, the concept that genes or proteins close to one another in a network are likely to cause similar diseases is becoming increasingly important when it comes to the detection of disease-related or –causing genes. Current approaches are based on mapping a set of candidates on a physical or functional network. Approaches with no reliance on prior knowledge of disease genes have yet to be developed.

Study of disease related subnetworks and their properties

The identification of individual disease genes or proteins in a global network makes it possible to define disease-related subnetworks and to study their properties. Much more than just a list of disease related genes or proteins, such disease subnetworks provide a hypothesis as to the molecular complexes, signaling pathways and other regulatory mechanisms (for instance, stable motifs or molecular switches) involved in the disease pathogenesis.

In order to gain an insight into Huntington's disease, Goehler and coworkers [33] constructed a PPI subnetwork around HTT, given that mutations in this gene produce the protein aggregation which causes Huntington's disease. Each direct interaction with HTT in the network was tested for its potential capability to enhance HTT aggregation. Indeed, it by this means that GIT1 was identified, with additional tests subsequently verifying its role in disease progression.

With the aim of exploring endotoxin inflammatory response in human leukocytes, Calvano and coworkers [34] reconstructed a network integrating information from literature and expression

data. The study of this response network allowed for the identification of new endotoxin-responsive modules and changes in the transcriptional program suppressing mitochondrial energy production and protein synthesis machinery.

Lim and coworkers [35] used Y2H to reconstruct a PPI to uncover ataxia-causing genes and genetic modifiers in humans. An original network of around 23 proteins involved in inherited ataxias was expanded, using information from literature, to ~3500 proteins and ~7000 interactions. Following network analysis, the authors found that the mean distance between ataxia-causing proteins was much lower than in a network constructed around 30 disease proteins independent of phenotypes. This demonstrated the utility of such analysis to detect newly involved genes.

Ghazalpour and coworkers [36] reconstructed a gene co-expression network integrating expression data from the livers of a panel of mice and genetic marker data from the same individuals for 22 different physiological traits. They found several co-expression modules enriched in genes with loci which had strong associations with a specific physiological trait, yielding a matrix of module/trait associations.

Indeed, one common trait seen across all of these works is the integration of previously known disease related genes with physical or functional interactions coming either from literature or experiments (or both). Analysis of these networks offers a mechanistic hypothesis about disease pathogenesis in terms of affected signaling cascades, metabolic pathways and/or molecular complexes. Such analysis can also aid in the description and explanation of the interplay between genetic and environmental factors influencing disease process packaging across a global network in a reduced number of functional modules.

Network-based classification of disease

A major challenge in the study of diseases has been to identify subsets of biomarkers with the highest predictive ability [37]. One strategy used to address this problem consists of applying network-based approaches, and focussing the search of disease biomarkers on specific network regions (for instance network modules or pathways). Network-based approaches can be applied both to separate “cases” from “controls” in case-control studies, to classify samples in disease states and to distinguish between very similar diseases. These approaches can also provide an improved prediction accuracy as has been shown in several works. Something which all of these works have in common is that they are based on the notion that informative genes fall under the same network neighbourhoods.

Following this principle, Ma and coworkers [38] developed a method to identify genes which are predictive of Alzheimer disease. Their method combined gene expression and protein association data to score the relationship of genes with a given disease class based not only on the association between the gene’s expression level and this class but also on the association of its network neighbours. The authors reported a better performance than when only expression data was used.

Network-based disease classifiers can also be used for disease prognosis. In order to predict metastasis in breast cancer, Chuang and coworkers [39] developed a method based on the expression levels of genes belonging to subnetworks (constructed according to network topology properties) of a human protein-protein interaction network. They found that subnetwork markers were more reproducible and showed higher accuracy when it came to predicting metastasis than individual marker genes selected without network information.

In a very similar approach to the previous one, Efroni and coworkers [40] developed a method designed to classify cancer samples into phenotypic disease states. The main difference from that described above [39] was that instead of subnetworks extracted from a protein-protein interaction network, they used curated pathways extracted from databases. After applying this method, authors found a small collection of pathways which distinguish the phenotypes with high accuracy.

In addition to protein-protein interaction networks, gene regulatory networks have also been used to classify disease states. Taking one example, Tuck and coworkers [41] developed a method based not only on the expression of a gene but also on its corresponding transcription factor, meaning that for a given sample a transcriptional interaction feature is considered "active" if the gene and the transcription factor are co-expressed. The authors showed in this work that the network based classification compared favourably with gene-expression-based classification.

Simple- and multi-target drug design

When considering a complex disease as a robust system but with certain fragility at specific points, network analysis based approaches to drug design provide an interesting alternative when it comes to looking for a suitable drug target. Indeed, these approaches maximise the desire effect and minimise secondary effects based on network topology or dynamical properties [42].

The development of a multi-target drug strategy based on network analysis rests on the assumption that, given the natural robustness and redundancy in pathways of biological networks, the perturbation of the network at multiple points should yield much better results than the traditional single-target strategy. Identifying alternatives ways in which to achieve the desired network perturbation allows us to avoid as many essential genes as possible, thus increasing the synergistic performance and decreasing side effects [43,44,45].

Cellular reprogramming

Cellular reprogramming may be applied both to the study (by means of creating cellular disease models) and to the treatment (within the context of cell therapy and regenerative medicine) of disease. The latter refers to the regeneration of damaged tissues and organs in the body by replacing damaged tissue and/or by inducing the body's own repair mechanisms even in the case of tissues without this natural capability. Conceptually speaking, regenerative medicine based on cellular reprogramming can be used in conjunction with gene therapy in the sense that reprogrammed cells introduced into the organism, and coming either from the patient

(autologous transplantation) or from another donor (allogeneic), can possibly be genetically modified.

Interaction between different diseases

Network analysis has been used to study interactions between different diseases resulting in, for example, comorbidity effects, instead of focusing on the study of a single disease. Goh and co-workers [21] reconstructed two networks: in the first, diseases as nodes were linked by mean of common implicated genes, whilst in the second, disease related genes were linked through diseases. These two networks provide a theoretic framework which facilitates the unveiling of some of the common genetic origins of diseases. The authors discovered the existence of disease-specific functional modules, and also found that the majority of human disease-related genes are not-essential and do not encode hub proteins. A selection based model was proposed to explain differences between essential and disease genes. Park and coworkers explored the impact of cellular networks on disease comorbidity through the integration of information on cellular interactions, the disease-genes relationship and population-level Medicare date. They found correlations between disease comorbidity and the structure of cellular networks [22].

Transcriptional regulation and transcriptional regulation modeling

Despite the fact that both physiological and pathological processes can be described and analysed at different levels (transcriptomics, proteomics, metabolomics, etc.) in a separated or integrated manner, the collection of works included in this dissertation is focussed on the transcriptional level. As a result of this we decided to include a brief summary to remind the reader of the main transcriptional regulatory mechanisms.

Transcription, or the process of creating a complementary RNA copy of a sequence of DNA, is a process regulated in eukaryotes through combinatorial interactions between several transcription factors, thus allowing for a sophisticated response to multiple environmental conditions. This regulation results in different transcription patterns which can be used to characterise physiological and pathological conditions at a cellular or tissue level. Assuming that the transcriptional machinery is always ready, these transcription patterns finally rest on the access to general and specific transcription factors genes in the DNA chain, as well as on the availability of the necessary molecules to be able to assemble the pre-initiation complex. Indeed, this anchors the DNA polymerase to the promotor region. More specifically, the assembly of the pre-initiation complex is determined by chromatin packing as well as by several regulatory elements.

The chromatin packing is affected by different types of histone modifications, namely methylation, phosphorylation, acetylation (and other acylations), glycosylation and ubiquitylation. Some of these histone modifications (like acetylation or phosphorylation) destabilise the union DNA-histone, thus creating repulsions between them, whereas others (methylation and combinations of methylation and acetylation) are used to tag specific DNA regions for protein recognition (by mean of the so called histones code). In addition to the local level of chromatin packing and nucleosome positioning, both of which have been proposed as

transcription regulatory elements, the nuclear organisation of chromatin in a three-dimensional space has also been proposed. This would involve a number of nuclear caves with high transcriptional activity involving DNA sequences from different regions of the genome (even different chromosomes) and other compact nuclear locations with low transcriptional activity [46].

Regulatory elements can be classified into two main categories: cis-regulatory elements and trans-regulatory elements. Cis-regulatory elements are DNA specific sequences, including enhancers, silencers and the TATA box. Trans-regulatory elements are elements which bind cis-regulatory elements or other trans-regulatory elements, namely transcription factors (TF), activators (bind enhancers), repressors (bind silencers), co-activators and basal factors (see Figure 2).

The binding of trans-regulatory elements can also be conditioned by different sorts of molecular modifications, affecting both cis- and trans-regulatory elements, namely DNA methylation (which prevent the binding of the polymerase machinery) and different types of protein modifications.

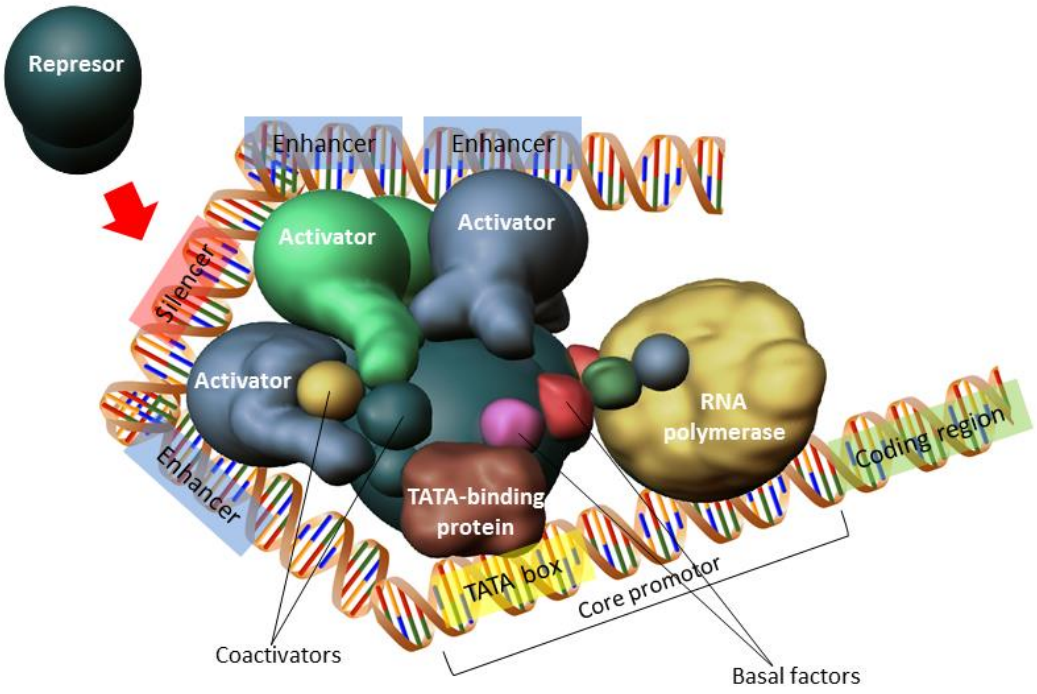


Figure 2 | Transcriptional regulation Cis- and Trans- regulatory elements. Cis-regulatory elements like enhancers, silencers or TATA box are DNA regions, whereas Trans-regulatory

elements are molecules that bind Cis-regulatory elements (activators, repressors and TATA-binding proteins) or other Trans-regulatory elements (coactivators and basal factors).

Apart from the regulatory elements described above, we must also take into account the regulatory effect of non-coding RNAs, including miRNAS and lncRNAS. miRNAs bind enzymes, thus resulting in the formation of effector complexes which can cleave mRNA, block mRNA from being translated or accelerate its degradation. Indeed, these non-coding RNAs essentially down-regulate specific genes [47]. lncRNAS have been proposed as guidance or scaffold for many proteins responsible for DNA and histone modifications, indirectly conditioning the access to target genes [48].

Modeling transcriptional regulation: gene regulatory networks

When modelling the transcriptional regulation of specific genes, two types of theoretical elementary system can be defined, depending on how they respond to regulatory signals (see Figure 3):

- Inducible systems - An inducible system is off (not expressed) unless a certain molecule is present (called an inducer) which allows for gene expression. The molecule is said to "induce expression". The manner in which this happens is dependent on the control mechanisms and can involve several of the regulatory elements described above.
- Repressible systems - A repressible system is on except in the presence of a certain molecule (called a repressor) which suppresses gene expression. The molecule is said to "repress expression".

The manner in which the system responds is dependent on the control mechanisms and can involve one or several of the regulatory elements described above. Whilst the simplest systems only include one inducer or repressor, more than one can exist and they can have synergistic effects or compete for the same regulatory mechanism.

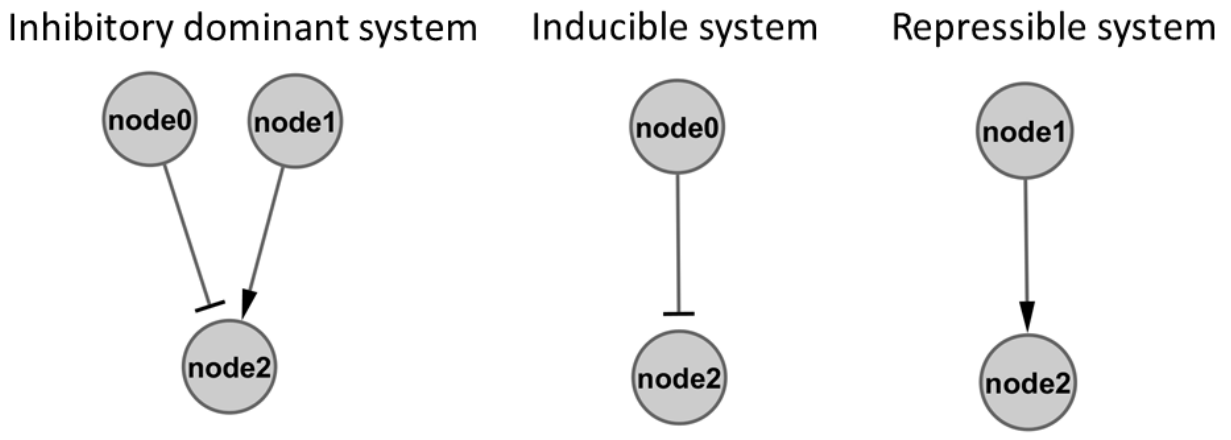
When modeling the global cellular transcriptional regulation using a systems approach, all of these aforementioned systems can be integrated into a single GRN. Depending on the events we are trying to describe, nodes and edges can represent different things. Nodes can represent the concentration of different molecular species, including genes, RNAs (mRNA, miRNA and lncRNA) and proteins. In addition to this, they can also represent composed forms of information, such as for example “level of activity”, as well as integrating information about the concentration of a specific protein and its state of activity (which can depend on different molecular events like the protein folding or the phosphorylation level). Edges can represent different types of interactions, which can in turn be broadly classified based on directionality in directed and undirected, and based on knowledge regarding the nature of the effect in signed and unsigned.

Undirected networks do not provide a representation of cause-effect relationships, but only an association which usually describes a physical interaction. The latter is the case for the example of protein-protein interaction networks, where nodes represent proteins and edges physical contacts. On the other hand, directed networks describe cause-effect relationships, thus meaning that an observer can determine the regulator and the regulated for each edge in the

network. If the nature of the action of the regulators is known these networks are referred to as signed, and each interaction in the network is either positive or negative. For instance, a positive/negative interaction in a GRN, where nodes represent concentrations of mRNA, means that an increase in the concentration of the regulator will result in an increase/decrease in the level of the regulated gene.

It is worth mentioning that, depending on the process we want to describe, both nodes and edges can be merged into super-nodes and super-edges respectively, thus representing a higher order of information. For example, one single super-node could represent a regulatory complex with activators, co-activators, basal factors and TATA binding protein, or a given super-edge could represent an entire signaling pathway. Similarly, one single node could be split into two nodes, representing, for instance, the active and inactive state of a protein.

All GRNs included in this dissertation are directed and signed; positive and negative interactions within these networks are termed “activations” and “inhibitions” respectively, whilst the corresponding regulators are termed “activators” and “inhibitors”. Within these GRNs, a given node can have more than one regulator, being the corresponding interactions either of the same or a different sign. When modelling the dynamics of GRNs, or the evolution of the network in time, decisions regarding how different interactions will influence the regulation of a specific node are made depending on the type of dynamical model. For instance, if the dynamical model assumed is a discrete model (for example, Boolean), then we must define a logic gate establishing the regulatory rules for each gene in the network. For instance, a system which is commonly assumed by default is the so called inhibitory dominant system. The rule assumed here is the following: for a given regulated gene to be considered as ‘active’, all the inhibitors should be ‘inactive’ and at least one of the activators should be ‘active’. If there are no activators but only inhibitors, the system is considered as a pure repressible system (see definition above) and is going to be ‘active’ unless any of the inhibitors (or more than one) is active. If the dynamical model assumed is continuous, the state of a given node is usually defined by a set of ordinary differential equations (ODEs) and a variable number of parameters (depending on the complexity of the model) must be taken into account. In essence, these parameters are going to define the strength of a specific interaction, whilst the resulting regulatory effect of a combination of different interactions with eventually different strengths is due to competition between the effect of activators and inhibitors. On certain occasions, the strength of the edges is displayed in the graph and such networks are called weighted. In contrast, when there are no weights for interactions, the network is called un-weighted.



Logic function for node2
node2 = node1 AND NOT node0

Truth-table

node0	node1	node2
0	0	**0**
0	1	**1**
1	0	**0**
1	1	**0**

Logic function for node2
node2 = NOT node0

Truth-table

node0	node2
0	**1**
1	**0**

Logic function for node2
node2 = node1

Truth-table

node1	node2
0	**0**
1	**1**

Figure 3 | Logic models to describe the regulation of three different systems. Truth tables include a Boolean representation of the state of the nodes, with "0" and "1" representing inactive and active states respectively. The logic operator AND establishes a condition which is true (true equates to active) when both elements involved are active. The logic operator OR establishes a condition which is true when at least one of the elements involved is active. Otherwise the condition is falso (inactive).

1.2. Motivation and outline of this dissertation.

Over the past fifteen years, molecular cell biology has transitioned from a descriptive science into a quantitative science which systematically measures cellular dynamics on different levels of genome, transcriptome, proteome and metabolome. Along with this transition emerges systems biology, which aims to unravel the complexity principles of living systems through the integration of experimental data into qualitative or quantitative models and computer simulations. Using a more holistic perspective instead of the traditional reductionism, systems biology approaches have been applied not only to the study of normal biology but also to biomedical research focused on the study of disease. Within the context of systems biology, a network can be defined as any interconnected group of elements (system) which shares information. Network analysis has been applied to identify new disease related genes and

pathways, as well as to detect and classify diseases. We also find network analysis approaches to simple- and multi-target drug design for disease treatment.

However, disease study based on network analysis approaches relies on the network reconstruction itself, as well as the detection of malfunctioning parts of the network associated with a specific disorder. Indeed, these disease-related subnetworks are strongly context dependent. There are several works which have aimed to identify subnetworks related to specific biological functions based on modules, motifs (overrepresented patterns of connectivity) or stability elements detection. However, none of these pioneering works have provided a mechanistic detailed explanation regarding how these disease-related network elements operate in order to produce disease pathogenesis. This lack of detail when it comes to the description of the mechanisms involved is precisely due to the limitations of different sources of information. On the one hand, experiments, for example gene expression data, provide information which is strongly contextualised to the experimental conditions, although a huge amount of data is usually required in order to statistically validate cause-effect relations (represented as directed and signed network interactions); purely experimental information based approaches explore an extremely large space but ignore a wealth of data pertaining to interactions previously described in the literature. On the other hand, literature based networks are too disconnected from experimental data to be able to describe input-output relationships and usually merge interactions described in different biological contexts (cell types, tissues and even species). Whilst there are a few published methods which have integrated experimental and literature information to elucidate signaling pathways [49] or to assess reconstructed GRN [50], none have addressed the context of disease study or disease treatment.

The precise underlying motivation for this dissertation is to establish how to integrate different information resources (literature, databases and experiments) to improve the current computational methods for network reconstruction and analysis within the context of disease study in order to be able to perform more accurate descriptions and reliable predictions. The contribution of this dissertation to the study of disease using network analysis can be split into two main categories: a) disease description: study of disease related sub-networks and their properties; b) disease treatment: designing strategies for cellular reprogramming.

Disease description: study of disease related sub-networks and their properties in three different diseases.

Disease description constitutes the second chapter of this dissertation, with three sections, 2.1, 2.2 and 2.3, corresponding to the study of three different diseases using network analysis.

Within this thematic block three different biological systems related to disease states were analysed using systems approaches/network analysis, namely prion disease, metabolic syndrome and the epithelial to mesenchymal transition in the context of breast cancer progression. A common characteristic of these three analyses is that they are focussed on the transcriptional level. Moreover, they are primarily concerned with changing elements at the

transcriptional level when comparing health and disease states. Another common characteristic of these three analyses is the central role of network stability and how this stability can be exploited to describe disease. Such disease description, including transitions between stable states, could be potentially applied to make reliable predictions about disease progression and its mechanisms, as well as to design therapeutic strategies.

In the three biological examples the identification of the regulatory core implies the integration of information about gene interactions coming from literature with experimental expression data. The main similarity of the three approaches used rests on the fact that networks are constituted by differentially expressed genes. In addition, the source of interactions has been previously published in works available in the literature. However, this is with the exception of the EMT case study, which provides a new miRNA interaction to enrich a regulatory core previously known from the literature. It is precisely the strategy used to determine the regulatory core which makes the difference between these three approaches. In the case of the metabolic syndrome we exploited the evolutionary conservation of specific overrepresented patterns of connectivity known as motifs, with special stability properties for an assumed dynamical model (multistable) so as to identify a regulatory core. In the case of the prion disease we exploited the local consistency of the experimental expression data with the expected values of the network stable states assuming a given dynamical model. Finally, we worked on the enrichment of a regulatory core for EMT previously known from the literature. The predictions performed computationally on this regulatory core support the idea of a missing interaction between a well-known driver gene for EMT (SNAI1) and a specific miRNA (miRNA-203) which had not been previously reported. This prediction was experimentally validated providing a more complete knowledge about the mechanism involved in this cellular transition.

Disease treatment: designing strategies for cellular reprogramming

The disease treatment part constitutes Chapter 3 of this dissertation. There are three sections in this chapter. The first section corresponds to the development of a computational method with which to contextualise GRNs to a specific biological context, with the integration of knowledge from literature and experiments. This contextualisation is a necessary condition to be able to model and predict the network response to perturbations in a given biological context. The remaining two sections of Chapter 3 refer to the development of a computational method to design recipes for cellular reprogramming. The second section of the chapter (3.2) proposes certain special stability elements of the network (the so called differentially expressed positive circuits or DEPCs) as the elements which should be perturbed to induce specific cellular transitions. Within this section we showed five biological examples to illustrate the general applicability of our methodology to different cellular transitioning systems harvested from literature with known effective reprogramming recipes. Finally, we showed the experimental validation of a reprogramming recipe predicted by our method to dedifferentiate astrocytes to neural progenitor cells (NPCs):

The third section of Chapter 3 (3.3) provides a strategy through which to minimise the number of genes and maximise the chance of effectiveness when it comes to a combinations of genes

predicted to be capable of inducing a desired cellular transition based on an expansion of the method described in Section 3.2. However, the concept of retroactivity is introduced as criteria with which to rank alternative combinations of reprogramming determinants based on the predicted stochastic behaviour of potentially relevant stability elements.

Finally, this dissertation ends with a general conclusions section which summarizes the main findings presented in this work as well as a discussion of the limitations of the developed methodology which could constitute a natural continuation in the future.

Chapter 2. Disease description and characterisation using network analysis and computational models

2.1 Detecting the disease regulatory core by identification of multistable network motifs: metabolic syndrome case study

This section refers to the work published in Cell Death & Disease in 2011 entitled "PPARγ population shift produces disease-related changes in molecular networks associated with metabolic syndrome"[51]. This publication was the result of a collaboration between the Computational Biology (W Jurkowski, K Roomp, I Crespo and A del Sol), contributing with the analysis, and the Medical Translational Research groups (Jochen Schneider), contributing with the biological interpretation of the results, both belonging to Luxembourg Centre for Systems Biomedicine (LCSB).

2.1.1 Introduction

The metabolic syndrome describes a cluster of metabolic abnormalities encompassing elevated fasting glucose concentrations, increased waist circumference, increased triglycerides, low HDL cholesterol levels and high blood pressure. In humans, this occurs with a prevalence of 20–25% worldwide according to the IDF (International Diabetes federation) and is on the rise, particularly in the elderly population, and even more alarming, even in children and adolescents. Total and cardiovascular mortality is increased in the metabolic syndrome, and the risk of developing overt type II diabetes is increased fivefold. From a pathophysiological point of view, the metabolic syndrome is widely held to be caused by central adiposity which can lead to insulin resistance under given genetic and environmental circumstances.
Peroxisome proliferator-activated receptor gamma (PPARγ) is a key regulator of adipocyte differentiation and has an important role in metabolic syndrome. Phosphorylation of the receptor's ligand-binding domain at serine 273 has been shown to change the expression of a large number of genes implicated in obesity. The difference in gene expression seen when comparing wild-type phosphorylated with mutant non-phosphorylated PPARγ may have important consequences for the cellular molecular network, the state of which can be shifted from a healthy to a stable diseased state.

2.1.2 Results

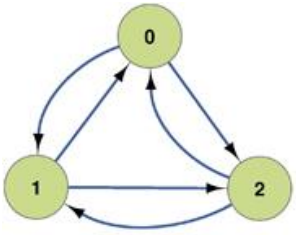
Identification of the disease regulatory core

In order to study the effects of modified PPARγ gene regulation, we reconstructed an interaction network (see methods for details) consisting of 235 DEGs, of which 152 were up-regulated in the phosphorylated state (wild type) whilst 48 were up-regulated in the mutated state (S273A mutant). We call this network the 'global' network. We subsequently searched a sub-network which makes an important contribution to the stability of the global network (the regulatory core).

It is worth mentioning that interactions included in this 'global' network were obtained from literature from different biological contexts, including different cells, tissues and organisms. The reason for this proceeding is that otherwise the lack of genome-wide context-specific information makes it impossible to reconstruct a gene regulatory network; the assumption made here is that it is worth gaining information through such a strategy despite the noise which can be introduced. As a result of this, interactions between DEGs are not as contextualised to the specific experimental conditions as the expression data is. In other words, some of the interactions could not be active in the specific conditions under study.

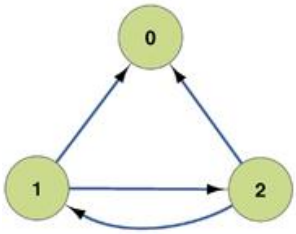
To address the problem of the lack of network specificity we looked for evolutionary conserved patterns of connectivity or network motifs. These motifs are statistically overrepresented patterns of connectivity, thus meaning it is unlikely that their constitutive interactions are there by chance. Network motifs and their properties have been studied since they were defined in 2002 [52]. At least they may reflect a framework in which particular functions are achieved efficiently. One remarkable characteristic of these motifs is that they tend to aggregate in motif clusters largely overlapping with known biological functions [53]. We decided to look for motifs with a very specific property: the capacity to exist in at least two stable states. These stable states should match with experimental expression data. This proceeding sought to integrate information from evolutionary conservation and experiments.

We first searched for motifs with multi-stable modes before selecting the top 10 statistically significant motif types consisting of 2, 3, 4 or 5 nodes using a z value-based ranking, all with a P-value of 0.01 (only 10 of 1000 randomly generated networks contained motifs at a higher frequency).



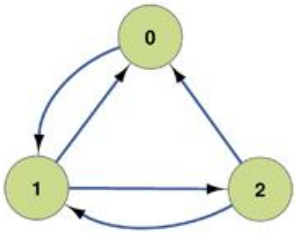
Node	A1	A2
0	1	0
1	1	0
2	1	0

n=2 Z=12.76 p=0.001



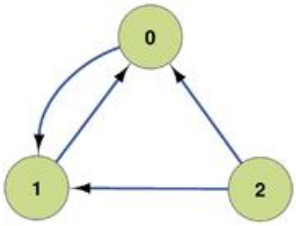
Node	A1	A2
0	1	0
1	1	0
2	1	0

n=33 Z=6.44 p<0.001



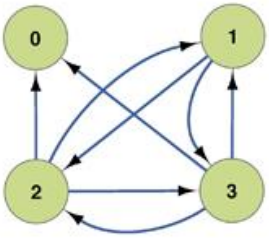
Node	A1	A2
0	1	0
1	1	0
2	1	0

n=10 Z=3.98 p<0.001



Node	A1	A2
0	1	0
1	1	0
2	0	0

n=7 Z=3.89 p<0.001



Node	A1	A2
0	1	0
1	1	0
2	1	0
3	1	0

n=3 Z=29.75 p<0.001

Figure 4 | Multi-stable motifs with matching gene expression values: n is the number of motifs of a given type found in the core network. A1 is attractor state 1, and A2 is attractor state 2.

We examined the expression profile of the genes in each motif to ensure that the expression profile was consistent in all members of the identified motif: among the top ranked 2-, 3-, 4- or 5-node motifs which exhibited multi-stable behaviour, we identified a number of 3- and 4-node motifs which were consistent with expression values (Figure 4). We thus identified 39 genes involved in 55 switches, forming a single cluster, which we call the ‘core’ network (Figure 5).

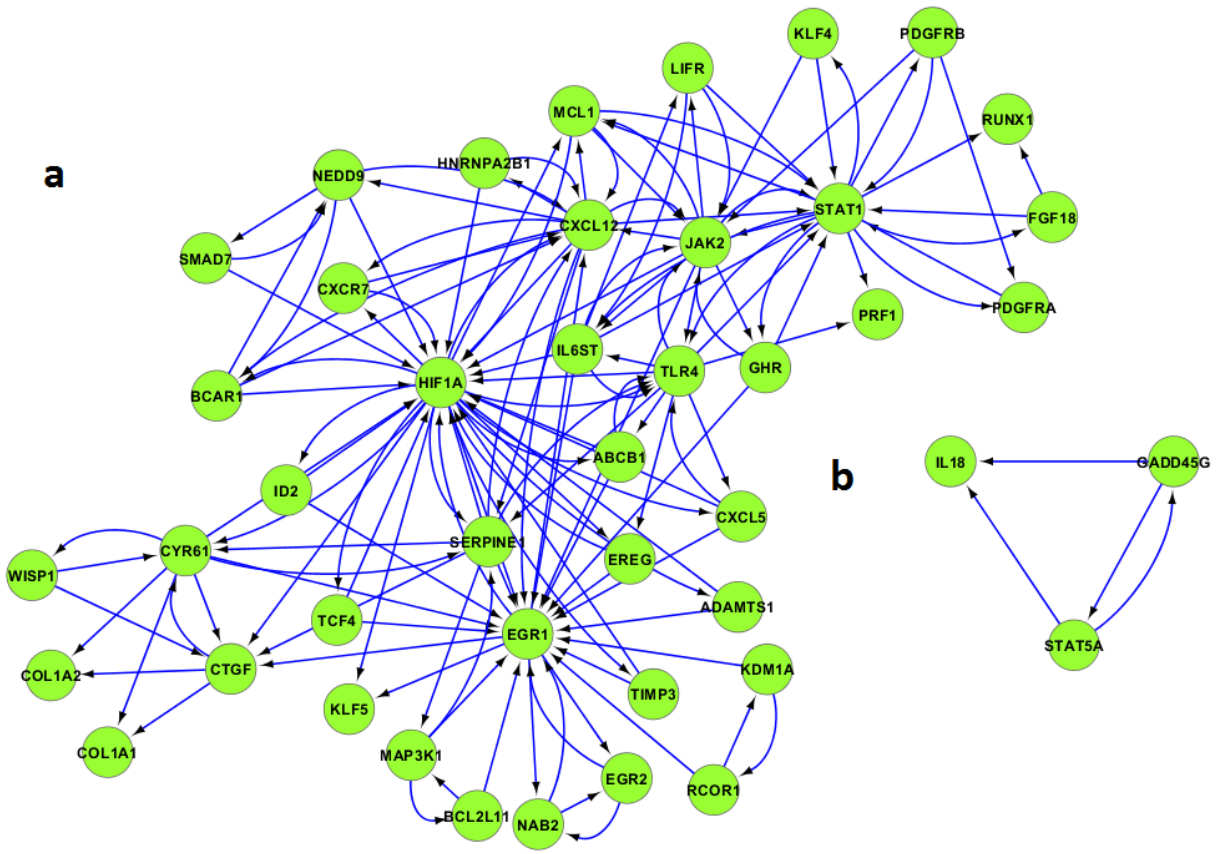


Figure 5 | The core network. (a) This network consists of 39 nodes and 55 bi-stable motifs. (b) Example of bi-stable switch motif, existing in ON/ON/ON or OFF/OFF/OFF states.

Analysis of the regulatory core

Although we forced all constitutive motifs in the regulatory core to be multi-stable and consistent with expression data, this does not guarantee that all together they still remain multi-stable and consistent with information from experiments. In light of this, we proceeded to analyse the core network stability.

Computation of the stability of the core network cluster reveals two attractors, in which all genes are either in an ON (up-regulated) or an OFF (down-regulated) state, which exactly match the expression pattern. The two most commonly found motifs are positive forward loop motifs with OFF/OFF/OFF changed to the ON/ON/ON states during the course of the disease. Consistent with this, the in silico perturbation of some of the core cluster's nodes triggers a transition from the OFF to the ON attractor, although the opposite transition (from ON to OFF) is not triggered by the perturbation of any node in the core. In all, 34 nodes in the cluster are capable of triggering the transition from OFF to ON when they are perturbed. In contrast, five nodes (COL1A1 (collagen type I-a1), COL1A2 (collagen type I-a2), kruppel-like factor 5 (KLF5), perforin-1 and runt-related transcription factor 1) constitute a set of genes which could potentially be in the ON state, taking part in different processes without the consequent activation of the cluster. We can see in Figure 6 that perturbation of KLF5, involved in a motif

with early growth response 1 (EGR1), does not trigger the transition of the cluster from OFF to ON.

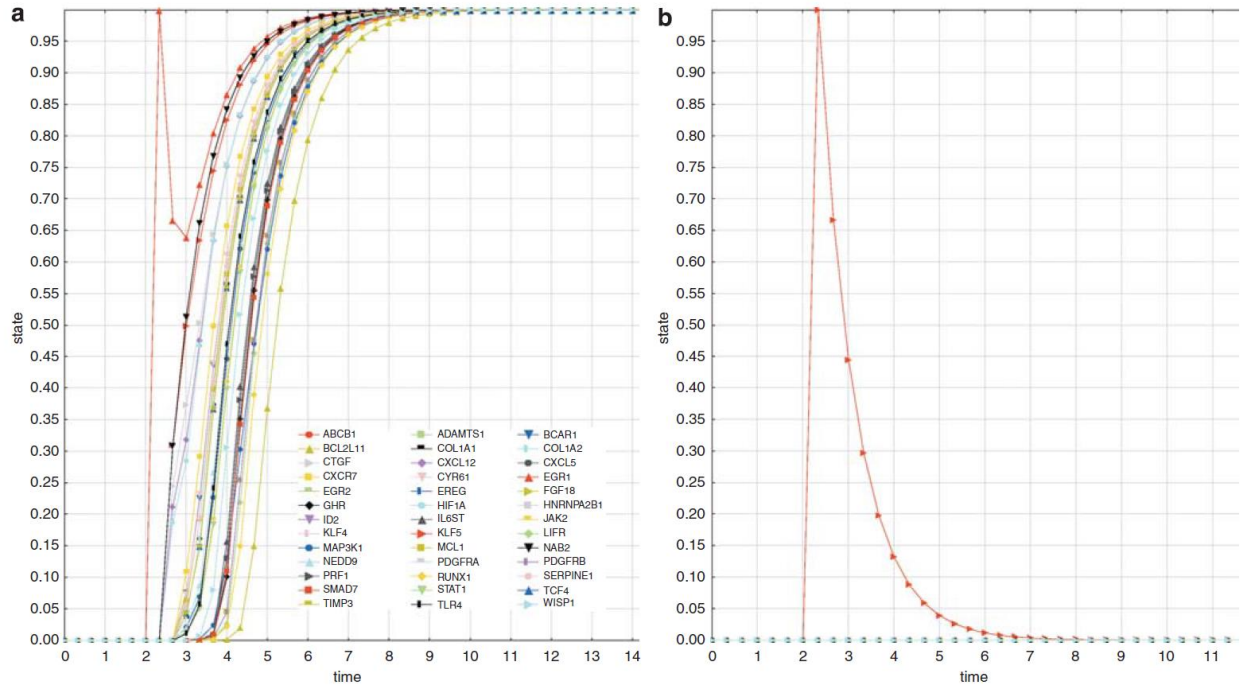


Figure 6| Two examples of core genes perturbations, with and without an effect on the core network. a) Perturbation of EGR1: all genes in the cluster change from the OFF to the ON state and remain there. b) Perturbation of KLF5: only the gene in question changes its state temporarily from OFF to ON before quickly returning to the original state. Both EGR1 and KLF5 occur within the same motif.

The main conclusion of this analysis is that, according to the assumed model, the regulatory core is easily put in a disease state but hardly reversed to the original healthy state. Additionally, simulations showed that certain genes in the core can be up-regulated without inducing the system to the disease state, although other genes work as disease inductors when they become active, namely COL1A1, COL1A2, KLF5, PRF1 and RUNX1.

Furthermore, we compared properties of genes in the core network with remaining genes in the global network ([51] Supplementary Table 1). Specificity of inter-gene interactions may be reflected in the network modularity. Using the Newman–Girvan algorithm, we identified 11 clusters, of which clusters 1 and 2 are mainly occupied by core genes. In order to identify intercluster connectors, we calculated the participation coefficient which was, on average, significantly different when comparing nodes in the core and global networks. The median participation in the core and global networks was 0.45 and 0, respectively; the distributions in the two groups differed significantly (Mann–Whitney–Wilcoxon W=2272, n1=44, n2=191, P-value=1.66e-07).

The betweenness centrality of a gene is a centrality measure which is proportional to the number of shortest paths between genes in the which that go through the gene in question. High betweenness centrality corresponds to a high level of inter-node communication, and is

therefore an appropriate measure for highlighting which genes link different molecular processes and pathways. Median betweenness centrality in the core and global networks were 243 and 0, respectively; the distributions in the two groups differed significantly (Mann–Whitney–Wilcoxon W=1430, n1=44, n2=191, P-value=3.65e-14). We identified a group of genes, which act as potential hubs and exist in the core network (with a betweenness centrality score ≥0.10 in the core network, as well as ≥2900 in the global network, making them the top four ranking genes in both cases), namely HIF1A, EGR1, STAT1 and CXCL12.

2.1.3 Discussion

We found that a group of differentially expressed genes are involved in bi-stable switches and form a network cluster of motifs considered as a regulatory core in the metabolic syndrome; the state of which changes with disease progression. These findings support the idea that bi-stable switches may represent a mechanism through which the core gene network can be locked into a diseased state and perturbations can be efficiently propagated to more distant regions of the network.

Analysis of the betweenness centrality showed that genes with the highest betweenness centrality were HIF1A (hypoxia-inducible factor 1a-subunit), EGR1, STAT1 (signal transducers and activators of transcription 1) and CXCL12 (chemokine (C–X–C motif) ligand 12) and that they are all overexpressed in the case of phosphorylated PPARγ (the putative disease state).

We examined these genes in more detail with regard to their association with adipogenesis, for which PPARγ is the master regulator under any condition. Through this we sought to confirm that our core network is consistent with experimental studies. We found that this was indeed the case: the role of the transcription factor HIF1A in adipocyte differentiation has been described previously [54]. EGR1 functions as a transcriptional regulator; the expression of which is rapidly induced during the differentiation of murine 3T3-L1 adipocytes [55]. STAT1 also acts as a transcription activator, which is rapidly activated in the 3T3-L1 adipocyte cell culture model [56] and CXCL12 is a chemokine, which demonstrates a significant increase in expression in differentiating 3T3-L1 pre-adipocytes.

It is worth mentioning that, although the network represents a cluster of events which have resulted from a PPARγ population shift, the connection between the shift and the changes produced downstream still remains unclear and it is not represented in our network. We hope that future research will clarify whether or not PPARγ is integrated into the regulatory core we detected or if this core is an independent module directly regulated by PPARγ or by means of indirect pathways.

2.1.4 Methods

Network reconstruction

We extracted DEGs from the results of gene expression analysis experiments by Choi et al., [57] in which PPARγ-null mouse embryonic fibroblasts were transfected with wild-type PPARγ (phosphorylated) or the S273A PPARγ mutant (not phosphorylated). The cutoffs for selecting

the DEGs were uncorrected P-values ≤0.008 and corrected false discovery rate P-values ≤0.15. This resulted in 577 DEGs, which we used in our subsequent analysis.

Subsequently, the ResNet mammalian database from Ariadne Genomics (http://www.ariadnegenomics.com/) was used to construct an interaction network of 235 DEGs with directed and signed interactions (the 'global' network). This database includes biological relationships and associations, which have been extracted from the biomedical literature using Ariadne's MedScan technology. MedScan processes sentences from PubMed abstracts and produces a set of regularised logical structures representing the meaning of each sentence. The ResNet mammalian database stores information harvested from the entire PubMed, including more than 715 000 relations for 106 139 proteins, 1220 small molecules, 2175 cellular processes and 3930 diseases. The focus of this database is solely humans, mice and rats.

Motif detection

Motif detection was performed using the FANMOD algorithm [58] for the global network and limited to motifs consisting of 2, 3, 4 or 5 nodes each. Each resultant topology was analysed in comparison with 1000 separately randomised versions of the initial network. Using both the original and the randomised versions, z-scores and P-values could be calculated for all motifs discovered in the original network.

The z-score of a motif is the original frequency minus the random frequency divided by S.D. The P-value of a motif is the number of random networks in which it occurred more often than in the original network, divided by the total number of random networks. Of the bi-stable switch motifs we identified, we also examined the expression profile of the genes in each motif to ensure that the expression profile was consistent in all motif members. We constructed a network (the 'core' network) consisting of 39 DEGs which occur in bi-stable motifs, which are significantly overrepresented in the global network, and whose genes are in a stable state, matching the experimental expression values.

Stability analysis

To compute the attractors of the core network, we used the program SQUAD (www.enfin.org) [59]. The program converts the network into a continuous dynamical system based on ordinary differential equations. In the absence of detailed kinetic parameters, the program interpolates a sigmoid curve between the states completely ON and completely OFF for each node. SQUAD first calculates the steady states found in a discrete dynamical system (Boolean model with asynchronous updating scheme) and then uses these states as a guide to localise the steady states in the continuous model.

Perturbations were also simulated using SQUAD. Each perturbation is a single pulse which changes the state of the node from 0 to 1 in the OFF attractor, or from 1 to 0 in the ON attractor as initial states of the system.

Modularisation and betweennes centrality analysis

The DEGs of the core and global networks were divided into groups by means of topological modularisation using the Newman–Girvan algorithm. The participation (P) of a node in intra-modular communication is calculated as follows:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{K_S^i}{k_i} \right)^2$$

where N_M=number of modules, K_s=number of connections with module s, k_i=total degree of module k. Analysis of betweenness centrality was calculated with igraph library in R comparing the core and global networks. The participation coefficient and betweenness centrality results for the core and global network groups were compared with the Mann–Whitney–Wilcoxon test in order to verify the statistical significance of the differences.

2.2 Detecting the disease regulatory core by identification of strongly connected components in a gene regulatory sub-network of locally consistent elements according to the connectivity pattern and experimental expression data in both health and disease states: prion disease case study

This section refers to a work published in BMC Systems Biology in 2012 entitled "*Gene regulatory network analysis supports inflammation as a key neurodegeneration process in prion disease*" [60].

2.2.1 Introduction

Prions are unique among transmissible, disease-causing agents in that they replicate through the conformational conversion of physiological forms of prion protein (PrP^C) to disease-specific PrP^{Sc} isoforms. They result in transmissible neurodegenerative diseases with common neuropathological features in mammals, including prion aggregation and accumulation, synaptic degeneration, microglia and astrocytes activation and neurons death. The activation of immune cells in the brain is believed to be one of the earliest events in prion disease development, where misfolded PrionSc protein aggregates are thought to act as irritants leading to a series of events which culminate in neuronal cell dysfunction and death. However, the role of these events in prion disease remains a matter of debate. Several hypotheses have been put forward to explain prion disease pathogenesis, such as PrP^C loss-of-function, PrP^{Sc} gain-of-toxic function, endoplasmatic reticulum stress, activation of autophagy and/or apoptotic death pathways, and chronic brain inflammation induced by misfolded protein and neuronal injury [61]. However, as yet there are no studies which have addressed the main initiator and/or propagator of the disease [62].
We have used a computational network analysis based on known gene expression data to address this complex question. In order to elucidate the mechanisms leading from abnormal protein deposition to neuronal injury, we have reconstructed a gene regulatory network and performed a detailed network analysis of genes differentially expressed in several mouse prion models.

2.2.2 Results

Identification of the disease regulatory core

We built a gene regulatory network based on the differentially expressed genes reported by Hwang et al. [63]. The functional relationships, based on gene expression and found in the literature, include interactions described in different biological contexts as different cells, tissues and organisms. As in the metabolic syndrome case example explained above, the resulting network is not contextualised to the specific experimental conditions under study, and thus we pruned the network based on the local inconsistency of expression data. If the expression data of a couple of regulator and regulated genes was inconsistent given the sign of the interaction connecting them, then this interaction was removed. After this pruning we obtained a resulting network which we called the global network, consisting of 106 genes which are differentially expressed during prion infection (all up-regulated), connected with 169 functional relations (all activations). Following this, we proceeded to identify the regulatory core of the global network responsible for the network stability. Network dynamics are regulated by the structure of the network through the flow of information by way of feed-forward and feed-back loops. When we looked for network structures with an exchange of information, we found a unique strongly connected component (SCC) consisting of 16 genes. The hallmark of such a structure is that, thanks to specific connectivity, the information can flow from one gene to any other in the structure following at least one path. This mutual influence between any pair of genes belonging to the SCC makes this structure relevant in terms of information exchange, and therefore potentially determinant of the network's stability. The SCC is mainly regulatory in nature with only 6 incoming functional relations. This SCC constitutes the regulatory core (see figure 7), and its regulatory impact extends up to 74 genes, meaning that the states of these 74 genes depend on the state of the SCC. This design resembles the "Medusa model" described by Kauffman [64], in which a set of genes represents a regulatory head, whilst the remaining genes represent arms controlled by the head.

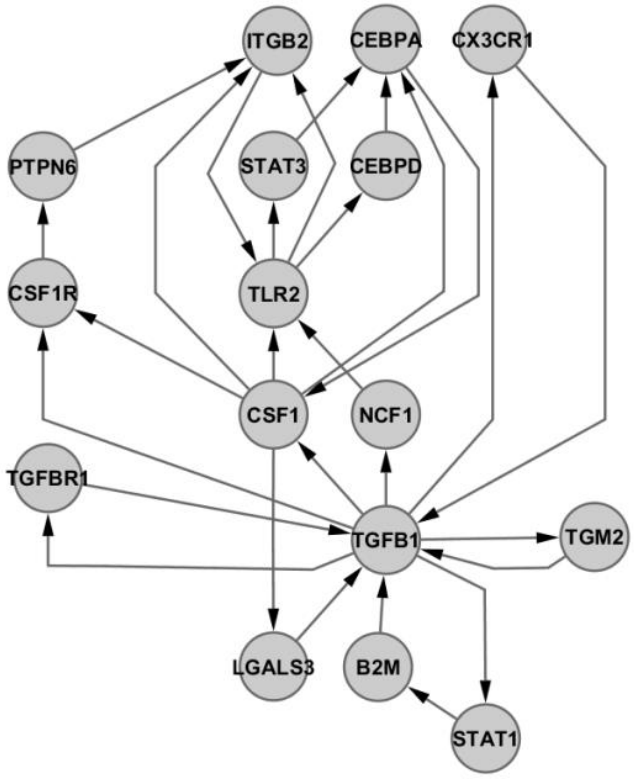


Figure 7 | Regulatory core composed of a single strongly connected component with sixteen nodes and twenty-eight regulatory interactions. There always exists a path (following the direction of the arrows) from any gene to any other gene within the regulatory core. Such reachability means that changes in the state of any gene can potentially affect the rest of the regulatory core.

Analysis of the disease regulatory core

We then carried out stability analysis of the regulatory core using a Boolean dynamical model with a synchronous updating scheme to compute network stable states. Two stable states were found for the regulatory core; one with all nodes "off" and one with all nodes "on" corresponding to health and disease states respectively. An *in silico* perturbation analysis demonstrated that core genes are individually capable of triggering the transition and that the network remains locked once the diseased state is reached. In contrast with this, no single gene repression is capable of reversing the disease state to the original health state, meaning that, due to this particular network topology, once the disease state is reached there is a strong resilience to leave it.

In order to assess the connectivity importance of genes belonging to the regulatory core we applied three different measures: network fragmentation, betweenness centrality and participation coefficient. This helped with the identification of genes which play the role of so called communication hubs (mediators of interactions between other, more peripheral genes). Fragmentation is a measure used to assess overall network connectivity and may be helpful in determining the impact of a sub-network on global topology. The fragmentation analysis of the

global network produced the following results. The mean of the giant component size for 1000 randomized removals of 16 nodes was 81.02 nodes (standard deviation 8.29), while it was only 38.00 nodes in the case of SCC node removal. The difference between these values is 5.18 times the standard deviation of the random removal values. This indicates that the size of the biggest set of connected nodes was reduced dramatically when we removed the nodes of the SCC instead of a random selection of 16 nodes. These results underlined the relevant role of the SCC as a connectivity element of the global network.

The betweenness centrality analysis of the global network showed that it is scarcely interconnected. There was a small group of central genes (mostly belonging to SCC) which had a much larger number of peripheral genes in the network connected to them. Six genes could be considered highly central (normalised betweenness > 1): TGFB1, CSF1, TLR2, CEBPA, LGALS3 and STAT3. In total, 25 genes were not peripheral (i.e. they mediated at least one gene connection). There was a significant difference when comparing the betweenness centrality of genes participating in the SCC and the genes in the rest of the network. Median betweenness centrality in the SCC and global networks were 123 and 0, respectively; the distributions in the two groups differed significantly (Mann–Whitney Wilcoxon W= 163, n1 = 16, n2 = 90, p-value = 1.406e-10) supporting the central role of the regulatory core in the global network. It should be noted that the betweenness centrality was more sensitive than other topological features such as degree or clustering coefficient to data incompleteness (missing genes or interactions). This is because it depends on the global network structure [65,66].

Having identified the hubs, we asked the question of whether the strong connectivity occurs between genes involved in common or distinct biological processes. Modules (clusters of genes sharing functional or topological properties) in the network were distinguished by assigning the pathological prion disease processes (derived from gene ontology annotations, described by Hwang et al.) to genes constituting the network core. Four modules were considered: disease-causing prion protein (PrPSc) replication and accumulation, immune response, neuronal cell death and other functions (genes which could not be assigned to any of the previous groups). Inter-modular participation is a measure used for the identification of genes which link different biological processes and this measure was calculated for all module members. Three groups can be distinguished according to node role (see materials and methods): (1) one connector hub with high inter-modular participation (P > 0.60) and significant within-module connectivity (z >2.5) at the same time highly central (normalized centrality >1): TGFB1; (2) satellite connectors, (genes with weak connectivity to other nodes of same function but with high ratio of connections to other modules) which share high centrality (normalized centrality >1): CSF1, TLR2, LGALS3 and STAT3; (3) less high central satellite connectors (positive normalized centrality): CEBPD, STAT1 and B2M; (4) other non-central but inter-module participative genes which are regulated by the SCC and are associated with a different functional category than the regulated gene or are regulating genes of other functions: CASP1, CLU, TGFBR2, P2RX7, NFATC1, CXCL10, CCND1, CYBB, AIF1 and GFAP. As expected, most of the selected hubs and connectors are parts of SCC supporting its assumed role as transition-driver. An example of this simulated transition is shown in Figure 8 where the perturbation (activation) of TLR2 induces the transition from healthy to disease state.

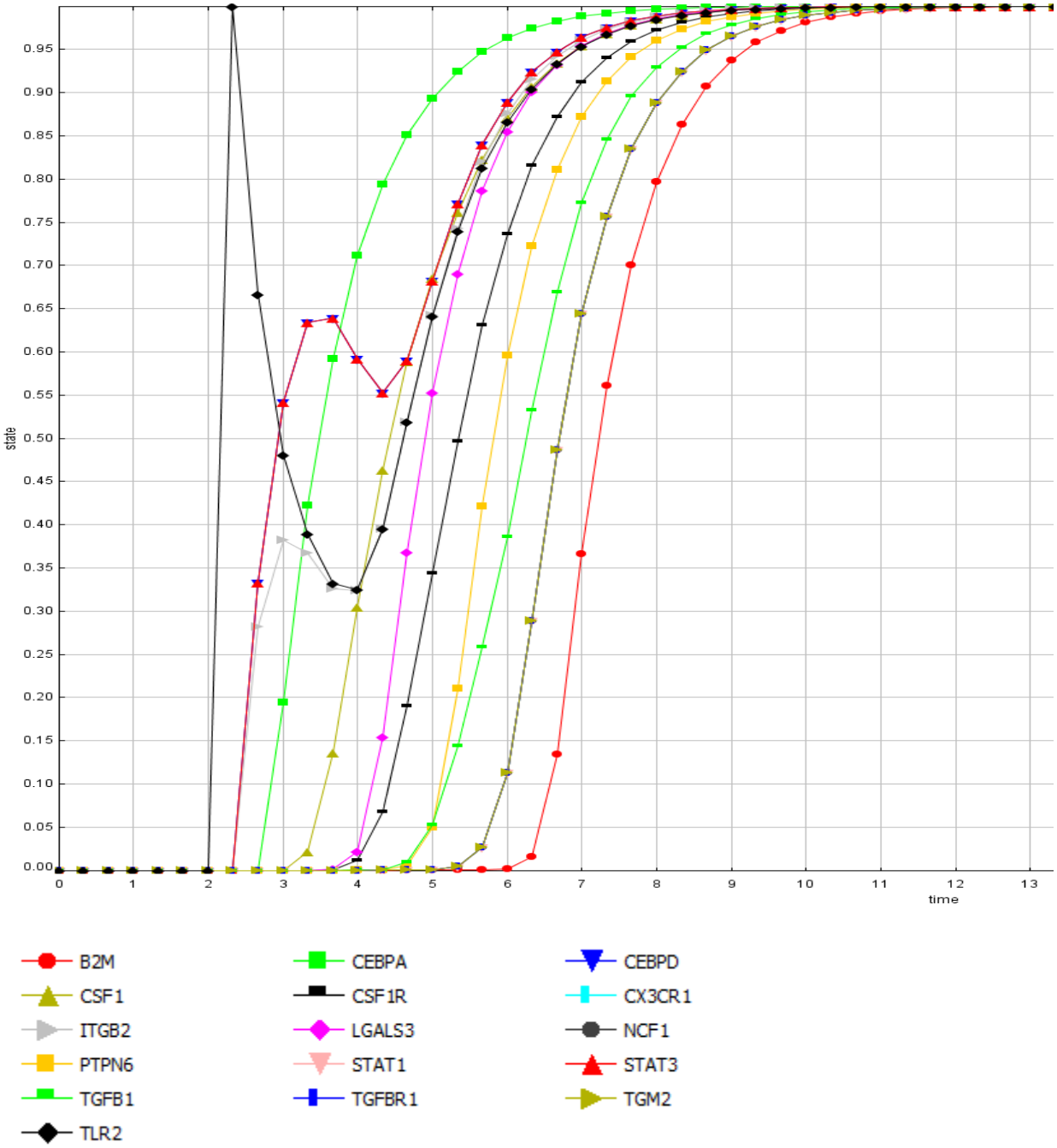


Figure 8 | Perturbation analysis of genes in the SCC. Perturbation of the TLR2 gene (black diamond), and its effect on the other genes of the SCC. The Y-axis represents the "level of activity"; Y-axis:0 indicates the "off" state, 1 indicates the "on" state. TLR2 is capable of triggering the transition from the "off" (healthy) to the "on" (disease) stable state for all genes in the SCC. The simulations were performed assuming a continuous dynamical system where the initial states are the attractors previously computed in a discrete model (Boolean). The X-axis represents "time" in arbitrary units.

Functional analysis

We have categorised the genes of the core network with regard to the four pathological features described by Hwang et al. [63] (Figure 9, Table 1). While no genes from the pathological feature category synaptic degeneration were found in the core network, it should be noted that only one of the 333 DEGs in the original mouse study was a member of this category. A potential sequence of reactive changes has been proposed by Hwang et al. and we have been able to identify genes in our SCC and core network at each stage in this proposed sequence. Transcriptome analysis suggested that one of the first changes was the activation of the complement pathways: the complement factor C3 is located in the core network. In addition, pattern recognition receptors (PRRs) and other receptors may potentially recognize PrPSc: ITGB2 and TLR2 in the SCC; CD14, CD68, FCGR2B, TREM2 in the core network. Subsequently, the complement complexes and PRRs may be responsible for stimulating the production of cytokines (CSF1 in the SCC and CXCL10 in the core network) and growth factors (TGFB1 in the SCC). They may also activate astrocytes, indicated by the increased expression of the glial marker GFAP in the core network. Cytokines released by microglia and astrocytes then lead to the activation of endothelial cells, which would stimulate the migration of leukocytes from the blood, followed by their differentiation into microglia (leukocyte extravasation), involving ITGB2, NCF1, TGFBR1, TGFB1 in the SCC and TGFBR2, ITGAX, CYBB in the core. The up-regulation of CSF1 (in the SCC) suggests that mononuclear leukocytes (blood monocytes) may be converted into microglia upon entering the brain.

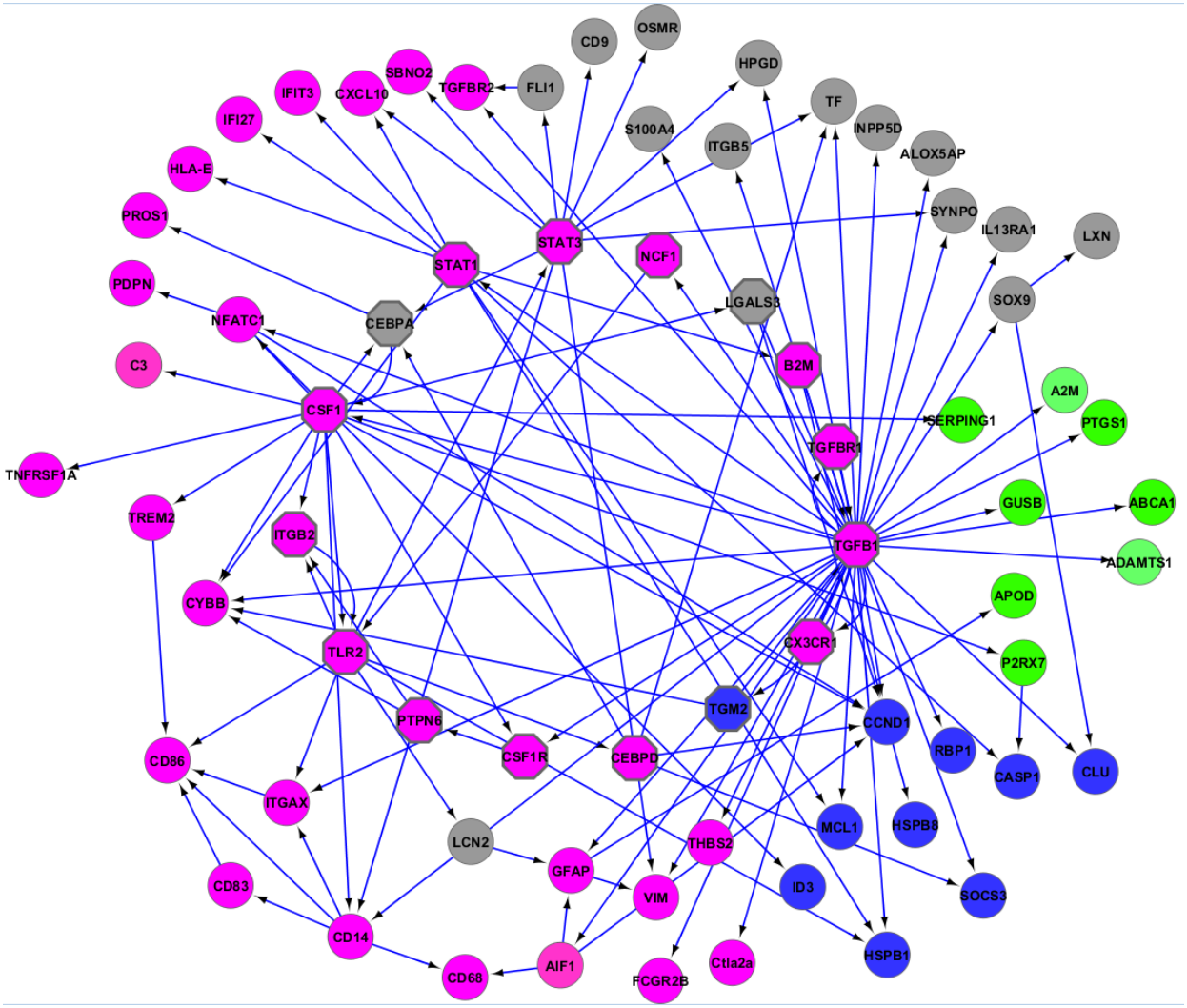


Figure 9 | Functional analysis of core network with pathological features.
Genes associated with PrP^{Sc} replication and accumulation are in green, with nerve cell death in blue, and immune response (including, microglia/astrocyte activation, leukocyte extravasation, general immune response) in pink. Other genes are indicated in grey. SCC genes are indicated as octagons.

2.2.3 Discussion

As result of this analysis we found a master regulatory core of 16 genes related to immune response controlling other genes involved in prion protein replication and accumulation, and neuronal cell death. This regulatory core determines the existence of two stable states which are consistent with the transcriptome analysis which compared prion infected with uninfected mouse brains.

In conclusion, we hypothesise that this locking may be the cause of the sustained immune response observed in prion disease. Our analysis supports the hypothesis that sustained brain

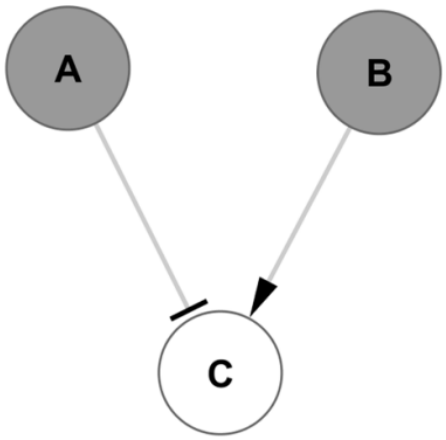
inflammation is the main pathogenic process leading to neuronal dysfunction and loss, which in turn leads to clinical symptoms in prion disease.

Inflammation response could be an overrepresented functional category in a great number of diseases. In the particular case of neurodegenerative diseases belonging to the class of protein misfolding diseases, it is well established that neuroinflammation plays a role. What we consider remarkable, and what constitutes our main finding, is the key role played by neuroinflammation in the specific case of prion disease, connecting different functional modules and constituting a switch which allows the network to reach a self-maintained disease state, once triggering factors (protein deposition and the formation of amyloid plaques) initiate the process. According to our simulations, the special topology which connects neuroinflammation elements (a cluster of positive feed-back loops or SCC) makes the regulatory core sensitive, under perturbation, to easy transition from inactive (healthy) to active (diseased) states, although it becomes very stable once the active state is reached.

In order to experimentally test the role of the genes of the identified master regulatory core in prion disease, we envisage inoculating $Prion^{Sc}$ into unconscious mice which are lacking one of the 16 SCC genes. Subsequently, the mice would be analysed pathologically with regard to neuronal function and death, as well as for DEGs for which the network analysis would be repeated. This analysis would be particularly interesting for factors whose explicit role has never been demonstrated in prion pathogenesis (*TLR2*, *NCF1*, *PTPN6)*.

It is worth mentioning that we assumed by default an inhibitory dominant system to define the regulatory rules in the assumed Boolean model. Whilst this simplified scheme can still capture certain regulatory behaviours which do not follow the inhibitory dominant system, it fails to capture others. This depends on the connectivity between regulators and regulated genes, but also on the state of the genes. The network pruning based on local inconsistency of expression data in network stable states partially overcomes this failure. With this said however, it is very important to stress the fact that this restricts conclusions regarding the network topology to stable states, thus meaning that predictions pertaining to dynamics in transient states should be avoided. In Figure 10 we can see examples of two different regulatory logic gates compatible with network topology and gene states (Figure 10 a) as well as those compatible only after the removal of a specific network interaction (Figure 10 b). In addition, this network pruning based on the local inconsistency of expression data relies on the assumption that expression data is essentially correct, and does not consider the possibility of noisy data. In order to address this issue we developed a computational method to contextualise gene regulatory networks based not on local consistency but on global consistency using an evolutionary algorithm which allows for the consideration of certain expression levels as wrong lectures if the global consistency is increased. This is very convenient when it comes to dealing with noisy information. The latter computational method is fully explained in Chapter 3.

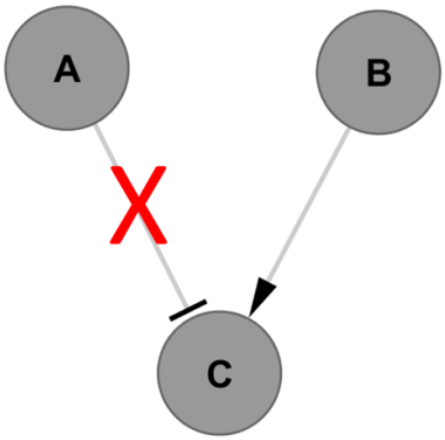
a)



Regulatory logic gate to define de state of C

Real regulatory logic gate	Assumed regulatory logic gate
C = B OR NOT A	C = B AND NOT A

b)



Regulatory logic gate to define de state of C

Real regulatory logic gate	Assumed regulatory logic gate
C = B OR NOT A	C = B AND NOT A

Figure 10| Example to illustrate two different regulatory logics compatible for some specific cases but not for another's. a) Real regulatory logic gate says that C is up-regulated if B is up-regulated or A is down-regulated, whereas the regulatory logic assumed by default in this work says that C is up-regulated if B is up-regulated and B is down-regulated. The network topology and gene states are compatible with both logics. b) In this case the assumed regulatory logic gate is not consistent with network topology and gene states; during the network pruning process the inhibition of A over C is going to be removed in order to make it consistent. The main problem with this is that we eventually lose real interactions. Nodes in grey and white

represent genes up-regulated or “ON” and down-regulated or “OFF” respectively. “T” and normal arrows represent inhibitions and activation respectively. A red cross represents the interaction removal during the pruning process.

2.2.4 Methods

Network reconstruction

The procedure for the network reconstruction consists of the following steps: a) Obtaining a list of differentially expressed genes. b) Connecting these genes using expression regulatory interactions from literature. c) Pruning the network based on local consistency in expression data of connected genes.

a) Obtaining a list of differentially expressed genes. A list of 333 DEGs was extracted from the results of gene expression analysis experiments performed by Hwang et al. [63]. These DEGs were found in all five prion-wild type mouse combinations in the study. They constitute the subset of genes which were differentially expressed in the five cases.

b) Connecting differentially expressed genes using gene regulatory interactions described in literature. The ResNet mammalian database from Ariadne Genomics http://www.ariadnegenomics.com/) was used to construct a gene regulatory network. The ResNet database includes biological relationships and associations, which have been extracted from the biomedical literature using Ariadne's MedScan technology [67,68]. MedScan processes sentences from PubMed abstracts Crespo et al. BMC Systems Biology 2012, 6:132 Page 9 of 12 http://www.biomedcentral.com/1752-0509/6/132 and produces a set of regularized logical structures representing the meaning of each sentence. The ResNet mammalian database stores information harvested from the entire PubMed, including over 715,000 relations for 106,139 proteins, 1220 small molecules, 2175 cellular processes and 3930 diseases. The focus of this database is solely on humans, mice and rats.
We used the list of differentially expressed genes to build a gene regulatory network without including any additional genes not found in microarray experiments resulting in a raw connected graph of 125 nodes and 255 interactions of known effect (positive or negative). In order to build the network we included only literature evidence of gene expression regulation (directed and signed interactions), thus meaning it is smaller and sparser than it would have been if all possible known interactions had been included (i.e. undirected protein-protein interactions, indirect interactions).

c) Pruning the network base on local consistency in expression data of connected genes. The expression patterns of the DEGs were checked in the Prion Database (http://prion.systemsbiology.net) to compare with topology and associations’ logic leading to the removal of 15 inconsistent nodes and 81 edges. Additionally, the discovery of few errors in the text mining process led us to further validate the network. To avoid false associations we took all sentences used by Pathway Studio (Ariadne Genomics) to determine gene associations and searched for co-occurrence of specific words: modifiers of sentence meaning, indicating

increased risk of false interpretation. During the next phase we conducted a manual verification of highly uncertain sentences and found two clearly incorrect associations: CD86 –+ > TGFB1 and CEBPA –+ > CASP8. In summary, we obtained a final graph of 106 nodes and 169 edges which we used for fragmentation analysis in this paper.

Determining the core regulatory network

Given that only genes with incoming interactions are relevant to the stability analysis, we had to identify genes involved in regulatory feed-back loops, or circuits, as well as genes regulated by them. For the first task we looked for strongly connected components (SCCs) in the raw network using Binom plugin [69] in Cytoscape [70]. An SCC is a network of nodes, where each node can be accessed directly or indirectly from every other node within the network. Simply put, there exists a path from each node in the network to every other node. Due to the specific connectivity in a SCC, the information can flow from one node to any other in the structure following at least one path. Such a path must respect the sense of the interactions (otherwise the component is weakly rather than strongly connected). Therefore, the state of any node in the SCC can directly or indirectly affect the state of any other node. This mutual influence between any pair of nodes within the SCC indicates that the SCC may be a relevant stability-related structure. We obtained a single SCC with 16 nodes.
Following this, we expanded these cores iteratively by adding first neighbours regulated by the SCC until no further neighbours could be added. This yielded a network consisting of the SCC and genes which are directly or indirectly regulated by genes in the SCC. The core network, including 74 nodes and 125 interactions (all nodes with incoming interactions) was used for the centrality analysis.

Stability analysis

For the stability analysis we used the SQUAD software package [50], creating a discrete dynamical system which allowed us to identify all the stable states of the system with an asynchronous updating scheme [51] using a binary decision diagram based algorithm [52]. Subsequently, a continuous dynamical system was created to identify the stable states in this continuous model which are located near to the stable states of the discrete system. Indeed, this is in accordance with the method described by Mendoza et al., 2006 [53], where the stable states of a Boolean model are taken as initial conditions in the continuous model.
Gene perturbations were simulated in the continuous model whilst the expression values of specific genes were changed, putting them in “1” and “0” when they were in the opposite value originally, i.e., “0” and “1” respectively.

Network properties

Fragmentation, betweenness centrality and inter-modular participation measurements were employed in order to compare the properties of SCC genes with other genes in the network, and to determine key genes which might be potential candidates for experimental validation.

In order to test the importance of the SCC in the network's connectivity, we examined the fragmentation effect of removing the 16 nodes belonging to the SCC in comparison with the fragmentation effect of 1000 different randomised removals of 16 nodes in the global network of 106 nodes. The giant component was the largest connected sub-graph found in the network for the given fragmentation and thus is a good measure with which to evaluate such fragmentation [53,71].

Betweenness centrality was computed for all genes in the network. The higher the value, the more central the gene is in the network of reference, i.e. other genes are more likely to be connected along the pathway involving these genes [71].

Modules in the global network were defined by functional and pathological process annotation of genes. The participation coefficient P is a measure quantifying intermodular connections of genes. For any gene in question, if P is greater than 0 and if the odds of inter-modular degree to total degree of the gene are less than 1, then it has to have at least one connection within its own and neighbouring modules. Together with measure of within module connectivity, participation allows us to define a node's role in the network ranging from most influential global hub to peripheral node (global hub, connector hub, provincial hub, kinless node, satellite connector, peripheral node and ultra-peripheral node). Such genes connect various functional pathways and may therefore be considered key regulators of cellular processes [56].

Functional analysis

Hwang *et al.* described four pathological features, which were derived from GO attributes: (1) PrP^{Sc} replication and accumulation, (2) microglia/astrocyte activation (which we refer to as immune response), (3) synaptic degeneration and (4) neuronal cell death. We mapped these pathological features onto the nodes in our core network and examined how the genes in our network may relate to disease progression.

Biological function		Genes
PrP^{Sc} replication and accumulation		A2M, ABCA1, ADAMTS1, APOD, PTGS1, SERPING1
Immune response		
	Complement activation: complement system	C3
	Complement activation: coagulation & kallikrein system	PDPN, PROS1
	Pattern recognition and other receptors	CD14, CD68, **ITGB2**, FCGR2B, TREM2, **TLR2**
	Microglia/astrocyte activation related	GFAP, **PTPN6**, **STAT1**, **STAT3**, THBS2, TNFRSF1A, VIM
	Cytokine, chemokine and growth factor related	**CSF1**, **CSF1R**, CXCL10, **CX3CR1**
	Leukocyte extravasation	CYBB, ITGAX, **NCF1**, **TGFB1**, **TGFBR1**, TGFBR2
	Other immune response	AIF1, **B2M**, CD83, CD86, **CEBPA**, Ctla2a, HLA-E, IFI27, IFIT3, NFATC1, SBNO2
Cell death		CASP1, CCND1, CLU, HSPB1, HSPB8, ID3, MCL1, RBP1, SOCS3, **TGM2**
Other		ALOX5AP, CD9, **CEBPD**, FLI1, GUSB, HPGD, IL13RA1, INPP5D, ITGB5, LCN2, **LGALS3**, LXN, OSMR, P2RX7, S100A4, SOX9, SYNPO, TF

Table 1 | Summary of the genes and their functional categories.

2.3 Expanding the disease regulatory core through the identification of missing regulatory elements: Epithelial to Mesenchymal Transition case study

This section refers to the work published in PLoS one in 2012 entitled "*A Novel Network Integrating a miRNA-203/SNAI1 Feedback Loop which Regulates Epithelial to Mesenchymal Transition" [72].* This publication was the result of a collaborative effort between the Cytoskeleton and Cell Plasticity lab, Life Sciences Research Unit-FSCT, University of Luxembourg (M Moes, ALe Béchec, C Laurini, A Halavatyi, G Vetter and E Friederich), contributing with the experimental part, and the Computational Biology group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg (I Crespo and A del Sol), contributing with the Computational modeling and analysis.

2.3.1 Introduction

The majority of human cancer deaths are caused by metastasis. The metastatic dissemination is initiated by the breakdown of epithelial cell homeostasis. During this phenomenon, referred to as epithelial to mesenchymal transition (EMT), cells change their transcriptional program, leading to phenotypic and functional alterations. The challenge of understanding this dynamic process resides in unraveling regulatory networks involving master transcription factors (e.g. SNAI1/2, ZEB1/2 and TWIST1) and microRNAs.

The importance of miRNAS as regulators in EMT has been recently pointed out by several works, and particularly the two clusters of the miR-200 epithelial marker family: miR-200b/200a/429 (miR-200b) and miR-200c/141 (miR-200c) [73,74]. The miR-200s regulate EMT through a double negative feedback loop with the ZEB factors, which, depending on the relative levels of miR-200 and ZEB, can induce the transition from epithelial- to mesenchymal-like states and back [75,76].
In addition, the transcription factor SNAI1, which plays a key role during the early stages of EMT, activates the expression of ZEB factors in a context-dependent manner [77,78].
However, the way in which these transcription factors and miRNAs interact is still a matter of debate. Indeed, there remains no integrated view describing how these transcription factors and miRNAs contribute together to conform molecular switches which rule the transition between epithelial and mesenchymal states. The dynamic properties of such networks are notably affected through feedback loops, eventually involving miRNAs and transcription factors, which act as toggle switches [79]. Previous versions of models of the regulatory core involving SNAI1, ZEB factors and miRNA-200 seemed to be incomplete, given that the dynamical behaviour was not consistent with two stable states observed in living cells, one for the epithelial and one for the mesenchymal phenotypes.
In this work we investigated the participation of other microRNAs in this regulatory core, and particularly those potentially associated with SNAI1 given that the up-regulation of this factor is capable of inducing the transition from epithelial to mesenchymal states.

2.3.2 Results

Expanding the disease regulatory core

Looking for miRNA candidates to enrich the regulatory core

In order to identify miRNAs participating in the regulatory mechanisms involving SNAI1, we analysed our time-resolved miRNA microarray data (GEO accession: GSE35074) of EMT induced by the perturbation (up-regulation) of SNAI1 in a breast carcinoma cell line.

At an established EMT state, 61 miRNAs were differentially expressed ([72] Table S1). Among those, 29 miRNAs were repressed and potentially regulated by the transcriptional repressor SNAI1. We combined these experimental results with miRNA expression signature analyses of four published datasets of epithelial and mesenchymal NCI60 cancer cell lines *(*Figure 1A and [72] Table S2, [80,81,82,83]). We then calculated expression correlations with the miR-200 epithelial marker family ([72] Table S3).
Interestingly, these analyses highlighted miR-203, whose expression was down-regulated in our EMT model and mesenchymal cancer cell lines, as being highly correlated to the expression of the miR-200s.

A

miRNA_id	MCF7-SNAI1	NCI60 Cell line panel			
	EMT model	Park et al.	Liu et al.	Blower et al.	Sokilde et al.
hsa-miR-203	DOWN	DOWN	DOWN	DOWN	DOWN
hsa-miR-200c	DOWN	DOWN	DOWN	DOWN	DOWN
hsa-miR-200b	DOWN	DOWN	DOWN	DOWN	DOWN
hsa-miR-200a	DOWN	DOWN	DOWN		DOWN
hsa-miR-141		DOWN	DOWN	DOWN	DOWN
hsa-miR-375		DOWN	DOWN	DOWN	
hsa-mir-7			DOWN	DOWN	DOWN
hsa-miR-429	DOWN		DOWN		DOWN
hsa-mir-215			DOWN	DOWN	DOWN
hsa-mir-205			DOWN	DOWN	DOWN
hsa-mir-194			DOWN	DOWN	DOWN
hsa-mir-192			DOWN	DOWN	DOWN

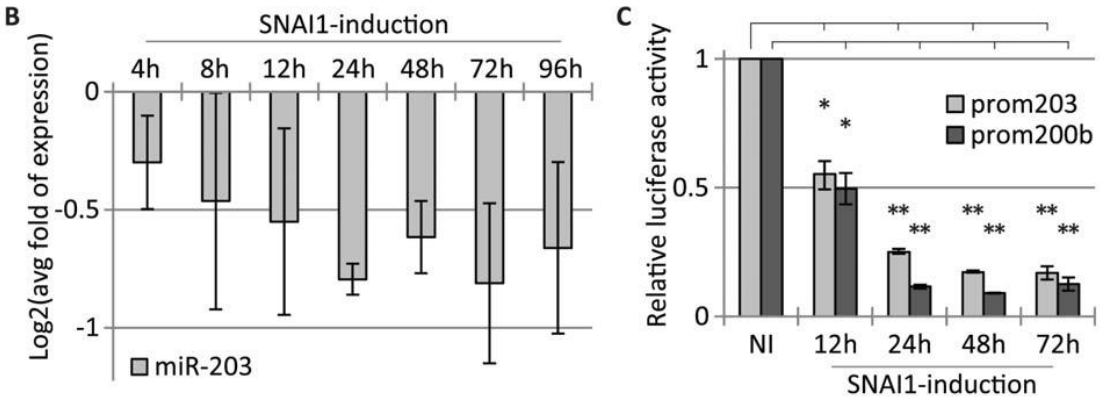


Figure 11 | Large-scale analysis of miRNA expression signatures, and miR-203 expression during SNAI1 induction in MCF7-SNAI1 cells. A) List of miRNAs found to be down-regulated, in at least three studies, in the large-scale analysis. B) qRT-PCR analyses of miR-203 expression levels normalised to U44 expression and expression levels in non-induced cells. C) Relative luciferase activity of miR-203 and miR-200b promoter constructs in non-induced (NI) and SNAI1-induced cells. Data are normalised to ''NI'' (*, p,0.05; **, p,0.01).

Interestingly, integrating a hypothetical feedback loop miR203/SNAI1 and the well-characterized miR200/ZEB feedback loops into a SNAI1-orchestrated EMT core network resulted in a model with a dynamical behaviour consistent with what was expected. Dynamic simulation revealed the existence of two stable states for this network and showed that the miR203/SNAI1 loop plays a crucial role in the switch from an epithelial to a mesenchymal state as well as in the stabilisation of the core network in these two states. These findings support previous studies [8,84] which showed the key role of feedback loops in network stability and determination of cell fate and plasticity.

In light of this, we decided to explore, in the wet lab, the regulation of miR-203 and miR-200 family members through SNAI1, as well as their integration into regulatory networks governing epithelial cell plasticity.

Experimental validation of new interactions

Regulation of SNAI1 on miR-203

In order to evaluate the effect of SNAI1 on miR-203 expression during SNAI1 induction we performed qRT-PCR analyses. MiR-203 was continuously repressed upon SNAI1 induction, similar to the miR-200b cluster. In addition, the relative luciferase activity of both miR-200b and miR203 promotor constructs when comparing non-induced and SNAI1-induced cells showed a significant decrease in promotor activity upon 12 h. These findings are consistent with the reduced promotor activity of miR-203 and miR-200c cluster by SNAI1 induction reported in other cell lines, namely HEK293T [85]and HCT116 [85,86]. Altogether these data suggest that SNAI1 regulates the expression of both miR-203 and miR-200.

Regulation of miR-203 on SNAI1

Following this, we investigated the role of miR-203 in relation with SNAI1 expression in breast carcinoma cells. Given that in MCF7-SNAI1 cells SNAI1 lacks its natural 3' UTR [87], these cells are not suitable to study the possible effect of miR-203 regulating the expression level of SNAI1. The mesenchymal breast cancer cell line HTB129 presents high levels of endogenous SNAI1 and low levels of miR-203 when compared to epithelial MCF7 cells [81,88]. HTB129 cells stably transfected with miR-203 (HTB129-miR203) exhibited a significant decrease in SNAI1 mRNA. HTB129-miR203 cells lost their typical fibroblastic, dispersed phenotype and acquired a more compact and cohesive appearance, more according to an epithelial. HTB129-miR203 cells further lost approximately 25% of their migratory and 15% of their invasive capacity. By performing MTT proliferation and Annexin V apoptosis assays we excluded the possibility that observed inhibitions were due to decreased cell proliferation and/or programmed cell death.

These results showed that miR-203 significantly reduces SNAI1 expression and promotes epithelial-like features such as a more cohesive phenotype and reduced motility. The next question was whether miR-203 directly or indirectly regulates SNAI1.

In silico analysis predicted two binding sites for miR-203, although none for miR-200 family members, within the 39' UTR of the SNAI1 mRNA (microRNA.org, August 2010 Release) [89]. The ability of miR-203 to directly target SNAI1 was evaluated by luciferase reporter assays in MDA231 cells, using SNAI1-39UTR reporter constructs - wild type or those lacking the predicted miR-203 target sites. Overexpression of miR-203 in MDA231 cells reduced the activity of the wild type SNAI1-39UTR, but not the mutant construct. Further, in agreement with in silico predictions, miR-200a and miR-200c (miR-200a/c), representing both seed sequences found within the miR-200 family, did not repress wild type SNAI1-39UTR reporter activity. These results indicate that miR-203, but not the miR-200s, directly regulates SNAI1 expression, thus linking miR-203 and SNAI1 in a double negative feedback loop and suggesting convergent yet non identical roles for these miRNAs in the regulation of SNAI1-orchestrated processes.

Integration of miR-2003/SNAI1 in an EMT regulatory core network

The mutual inhibition between miR-203 and SNAI1 forming a regulatory positive feed-back loop was integrated into the SNAI1 centered regulatory core with 15 interactions (2 activations and

13 inhibitions), 4 genes and 2 miRNAs (see figure12). This regulatory core includes interactions previously described in literature as the miR200/ZEB feedback loops [75], the activation of ZEB1/2 by SNAI1 [90,91,92], the inhibition of miR-203 by ZEB1/2 [85] and the repression of ZEB2 by miR-203[93]. In addition E-cadherin (CDH1), which is directly repressed by the SNAI1 and ZEB factors [77], was added to the regulatory core as an epithelial target gene.

Dynamic simulation of our core network revealed two stable states corresponding to the epithelial and mesenchymal phenotypes (“E” and ‘‘M’’ respectively) as described in literature ([72] Data S1) [1]. Transition probability between these two stable states further attributed a high robustness to both states ([72] Data S1), implicating that the regulatory core network is unlikely to spontaneously switch between these states without external stimulus. Importantly, the simulation of an up-regulation of SNAI1 triggered the transition from state ‘‘E’’ to ‘‘M’’ (Fig. 3A, B). Following this, and in order to show the importance of the miR203/SNAI1 feedback loop on the network dynamics, we performed an ‘edgetic’ (edge-specific genetic) perturbation, by removing the ‘‘miR-203 on SNAI1’’ interaction [94]. Interestingly, the dynamic simulation of the edge-altered regulatory core network revealed a single stable state ‘‘Eea’’ (edge-altered state ‘‘E’’) ([72] Data S1). Thus, the feedback regulation ‘‘miR-203 on SNAI1’’ is crucial for switching from an epithelial to a mesenchymal state and in stabilising the core network in both states.

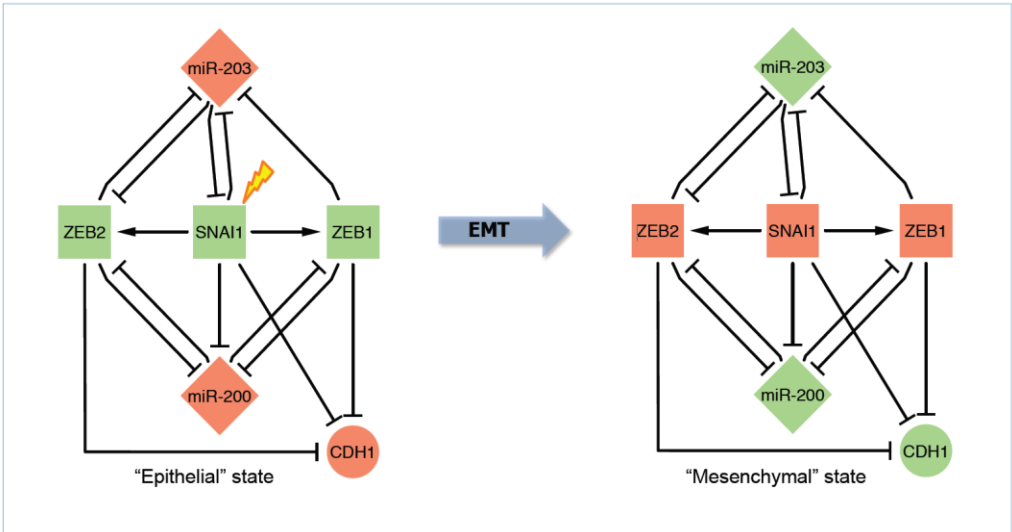


Figure 12 | EMT core network integrating the miR203/SNAI1 and miR200/ZEB double negative feedback loops. A) The top panel corresponds to the core network integrating described interactions between miR-203, miR-200s (miR-200), SNAI1, ZEB1, ZEB2 and E-cadherin (CDH1). The bottom panels show the stable states ‘‘E’’ and ‘‘M’’ obtained after dynamic analyses. B) In silico up-regulation of SNAI1 in a continuous dynamic system of the EMT core network. The state of SNAI1 is changed from ‘‘0’’ to ‘‘1’’ at time point 2 (arbitrary units of time), during two units of time.

Diamonds represent miRNAs, squares transcription factors, and circles coding-genes other than transcription factors. Red and green colours stand for up-regulated and down-regulated expression levels, respectively. Edges represent an interaction between two actors, either activation (arrow) or inhibition (blunt arrow). The ‘‘lightning’’ indicates a SNAI1 up-regulation triggering the transition from state ‘‘E’’ to ‘‘M’’ (red arrow).

The analysis of the dynamics of the resulting expanded regulatory core network indicated that it could function as a robust switch controlling early steps of EMT and epithelial homeostasis, further emphasising the importance of bistable feedback loops in determining cell plasticity [8,84].

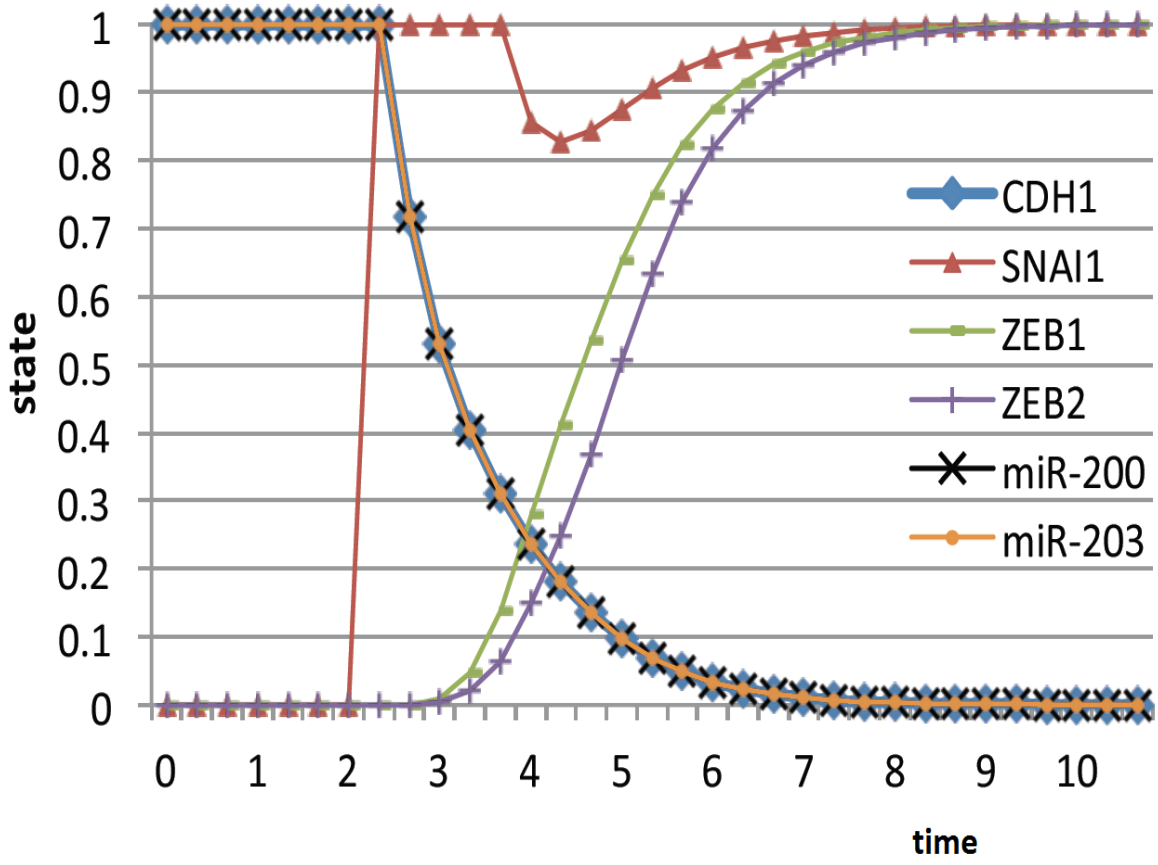


Figure 13 | EMT core network integrating the miR203/SNAI1 and miR200/ZEB double negative feedback loops. A) The top panel corresponds to the core network integrating described interactions between miR-203, miR-200s (miR-200), SNAI1, ZEB1, ZEB2 and E-cadherin (CDH1). The bottom panels show the stable states ‘‘E’’ and ‘‘M’’ obtained after dynamic analyses. B) In silico up-regulation of SNAI1 in a continuous dynamic system of the EMT core network. The state of SNAI1 is changed from ‘‘0’’ to ‘‘1’’ at time point 2 (arbitrary units of time), during two units of time. Diamonds represent miRNAs, square transcription factors, and circles coding-genes other than transcription factors. Red and green colours stand for up-regulated and down-regulated expression levels, respectively. Edges represent an interaction between two actors, either activation (arrow) or inhibition (blunt arrow). The ‘‘lightning’’ indicates an SNAI1 up-regulation triggering the transition from state ‘‘E’’ to ‘‘M’’ (red arrow).

The dynamical behaviour observed in simulation was consistent with experimental expression data, in contrast with previous versions of this SNAI1 centered regulatory core without the participation of miR-203. Obviously, the present regulatory core network is embedded into a larger network with multiple molecular actors such as SNAI2 and TWIST, the expression of which is interconnected in a context-dependent manner and regulated by various pathways [77,95,96]. Despite this connectivity and given the consistency between the stable states of the regulatory core and the expression profile of both epithelial and mesenchymal cellular phenotypes, the present version of the regulatory core can still be considered and analysed as a stability element which contributes to the general stability of the entire gene regulatory element. This core network will provide a good starting point from which to further study key regulatory circuits underlying EMT, such as TGF-b signaling which plays an important role in transient cancer cell invasion.

2.3.3 Methods

Stability and perturbation analysis

In order to simulate the dynamical behaviour of the system, we assumed a continuous model using the SQUAD program [59]. SQUAD assigns the same kinetic parameters for all regulations by default when this information is not available, as is the case in the present work. In addition, SQUAD limits the search for attractors using as initial states only the attractors found in a previous step in a discrete (Boolean) model with an asynchronous updating scheme [97]. Once these stable states from the discrete model are introduced as initial conditions in the continuous model the system is updated to its own stable states (usually slightly different than those found with the Boolean model). The probability of transition between the stable states was calculated in the discrete model after modelling the stochastic behaviour of the system using GenYsis, [97], whilst the resulting transition matrix showed that the probability of the system to spontaneously switch between the stable states after introducing noise according to this model is very low. In other words, the robustness of the system under stochasticity is very high; a common feature in biological networks and one which is consistent with what is known about epithelial and mesenchymal cellular phenotypes. To simulate perturbations in the continuous system the state of the specific nodes we wanted to perturb were changed from the current to the desired state for a given number of arbitrary units of time (defined by the user). The values of these states ranged from 0 to 1; they should be considered normalized. For instance, when the system is in epithelial phenotype the value of SNAI1 is close to 0. If we proceed to change this state to 1 for three arbitrary units of time we induce the transition to the mesenchymal phenotype, so once the perturbation finishes the system converges in a stable state with the state of SNAI1 close to 1.

Epithelial/mesenchymal miRNA expression signature study

The NCI60 panel was analysed using t-test analysis and the same classification as described in Park et al. [81]. miRNAS with a p-value lower or equal to 0.01 were considered differentially expressed. Expression levels (up- or down-regulated) correspond to the sign of the difference

between the averages of log-intensity values of mesenchymal cells and the average of log-intensity values of epithelial cells: sign($\log(I_M)$-$\log(I_E)$).

miRNA microarray analysis

miRNA microarray design, protocols and data are available at NCBI's Gene Expression through GEO Series accession number GSE35074. The established state of EMT is considered to have been reached after 72 h to 96 h of SNAI1 induction. This state refers to the ''late EMT stage' previously defined by the analysis of transcriptional events as well as phenotypic changes occurring upon SNAI1 induction in the MCF7-SNAI1 EMT model [87]. Averaged expression values for each time point (72 h and 96 h) were calculated, taking into account only replicates which have moduli of log-ratios ≥ 0.5 and t-test p-values ≤ 0.01 (according to LCSciences data processing).

Vector constructs

For exogenous miR-203 expression, the hsa-miR-203 stem-loop sequence (MI0000283) 2200/+192 relative to the first and last nucleotide of the stem-loop, was synthesised and cloned into BglII/HindIII sites of the pSUPER.retro.puro vector (pSUPERmiR-203) (OligoEngine) (DNA2.0). The Hsa-miR-203 promoter region [85] was synthesised and cloned into a pGL3-basic reporter using KpnI/HindIII sites (DNA2.0). Wild type human GAPDH- and SNAI1-3' UTR, as well as mutant SNAI1-39UTR lacking the predicted miR-203 binding sites, were synthesized and cloned into the psiCHECKTM-2 (Promega) vector at XhoI/NotI sites (DNA2.0).
Mir-200b promoter construct, pGL3miR200b/200a/429 (2321/+120) has been previously described [98].

Cell lines

''Tet-Off'' MCF7-SNAI1 cells expressing human SNAI1 upon removal of tetracycline from the culture medium have also been previously described [87,99]. The human breast cancer cell lines HTB129 and MDA231 (also known as MDA-MB-231 or HTB-26), purchased from the ATCC, were maintained in RPMI1640 and Leibovitz culture media (Lonza), respectively, supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. HTB129 cells stably expressing miR-203 were generated by pSUPER-miR203 vector transfection and puromycin selection. Cells stably transfected with the empty pSUPER.retro.puro vector served as a control.

Epifluorescence staining of cells

In order to reveal and illustrate the cell phenotype, DNA and F-actin were stained with DAPI (MPBiochemicals) and Phallo504 (Invitrogen), respectively. Cells were analysed by epifluorescence microscopy (Leica DMRX microscope). Images were acquired with a linear CCD camera (Micromax) and analysed with Metaview software (Universal Imaging Corporation Ltd).

RNA extraction and real-time quantitative PCR (qRT-PCR)

Total RNA was extracted using Trizol as recommended by the manufacturer (Invitrogen). RNA quality and concentration were evaluated spectroscopically using a NanoDrop 2000c instrument (ThermoScientific). Reverse transcription and qRT-PCR quantification of miRNA and mRNA were carried out as previously described [87,100], whilst U44 and GAPDH served as internal references, respectively. Oligonucleotides used in this study are listed in [72] Table S4.

Luciferase reporter assays

Indicated cell lines were plated in 6-well plates and transfection was carried out using Lipofectamine 2000 (Invitrogen). For promoter reporter assays, cells were cotransfected with a pGL3-promoter construct (600 ng) and a pRL-TK reference plasmid (5 ng) (Promega). For 39UTR reporter assays, cotransfection was realised with 90 ng 39UTR-psiCHECKTM-2 constructs and a total of 75 pmol Pre-mirTM miRNA Precursor Molecules (Ambion). After 24 h of incubation cells were lysed, and firefly and Renilla luciferase activities were measured with a FluoStar Optima instrument (BMG LABTECH) using the Dual-Luciferase Reporter Assay System (Promega). All reporter assays are shown as relative luciferase activities, normalised to controls.

Cell migration assay

Cell migration was evaluated using Ibidi culture inserts according to the manufacturer's protocol (Ibidi). Cells were seeded into the Culture-Inserts and grown overnight to confluency. After removal of the insert a 500 mm cell-free gap was created. Phase contrast images of the same gap fields were captured at 0 h and 24 h of incubation using an inverted light microscope (Leica DMIL) with camera (Leica DFC360 FX). Gap closure was quantified using ImageJ software (NIH).

Cell invasion assay

56104 cells were seeded in 2% FBS medium before being placed onto Transwell plates coated with 50 mg of extracellular matrix proteins (ECM gel E1270, Sigma). 10% FBS medium was added to the lower chamber as chemoattractant. After 24 h, cell invasion was quantified using the MTT assay (Sigma).

Statistical analysis

Assays were performed in technical triplicates and repeated in at least three biological replicates. Presented data are mean 6 SEM of three biological replicates. The paired t-test was used to estimate p-values. For the 39UTR reporter assays the one-tailed paired t-test was used to check for a potential decrease in relative luciferase activity. $P < 0.05$ was considered to be statistically significant. For qRTPCR assays, Log2-transformed mean fold changes (averaged over three biological replicates) are presented. Error bars are the SEM recalculated using the standard method for error propagation.

Chapter 3. Disease treatment and cellular reprogramming

Introduction

A wealth of exciting opportunities arises from the process known as “cellular reprogramming”. Through this process cells can be induced to change from one phenotype to another, including the possibility of modelling complex diseases (Alzheimer, Parkinson, schizophrenia, autism or blood disorders) using stem cells coming from more available cells of patients and applications in regenerative medicine. Regenerative medicine refers to the regeneration of damaged tissues and organs in the body by replacing damaged tissue and/or by inducing the body’s own repair mechanisms even in the case of tissues without this natural capability. The ability to control cellular programs opens new avenues in the field of regenerative medicine, allowing both to culture specific cell types in vitro for a posterior implantation in the patient or to induce cellular reprogramming in local cells from abundant and self-renewal sources (for example fibroblasts) to a missing or desired cell type (for example, missing dopaminergic neurons in Parkinson’s disease or lost cardiomyocytes after a heart attack). The development of technology for regenerative medicine has resulted in increasingly novel approaches to more deeply investigate the fundamental bases of cell identity and to better understand the natural differentiation process, including the identification of genes and regulatory elements which rule these processes.

Indeed, certain pioneers have attempted to detect genes responsible for cell fate and differentiation without previous knowledge regarding the underlying gene regulatory network [101]. However, the detection of molecular switches able to trigger transitions between different cellular phenotypes relies on the completeness and accuracy of knowledge pertaining to the underlying gene regulatory network. Network reconstruction methods can be broadly divided into two main categories: literature based and experimental based methods. Networks inferred purely from experimental data and those assembled from the literature have different limitations. In the first case, a wealth of data regarding interactions previously described in literature is ignored. On the other hand, literature-based networks are too disconnected from experimental conditions to be able to describe input-output relationships, such as cellular responses under specific biological stimuli or mechanisms which determine specific stable (long term) expression patterns. This is due to the fact that certain interactions are strongly dependent on the biological context; a context defined by both the intracellular and extracellular environment. Moreover, literature based networks usually merge interactions described in a different well defined biological context, such as cell types, tissues or even organisms, with the hope that the trade-off between the enrichment in information and noise addition is worth it. This assumption can be true for some network analysis which is based on network topology and which is very robust against noise. However, the resulting networks usually cannot describe molecular switches where one ‘false’ interaction can make a big difference. Neither purely literature based nor purely experimental data based methods provide the level of detail and accuracy needed for reliable predictions regarding how to induce cellular transitions with efficiency and fidelity.

In order to address the task of cellular reprogramming and particularly within the context of disease treatment, we developed a computational method to combine information from

literature with experimental expression data in order to reconstruct gene regulatory networks contextualized to the specific biological conditions under which the experiments were performed. This method exploits a general network property (the network stability) to guide an iterative network topology optimisation, and is fully described in the following section.

3.1 Contextualizing gene regulatory networks to specific biological conditions guided by the consistency between computed network stable states and experimental expression data

This section refers to part of the work published in Nucleic Acid Research in 2012 entitled "*Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimisation guided by network stable state"[102].* This paper provides a rigorous description of a novel algorithm with which to predict missing expression values in gene regulatory networks. As an intermediate step, the algorithm performs an iterative network pruning on a raw literature based gene regulatory network in order to make the topology as consistent as possible with the experimental expression data. It is precisely this network contextualisation which constitutes the main focus of the present section.

3.1.1 Introduction

The wealth of experimental data from high-throughput technologies in different areas of biology, and especially at a transcriptomics level, allows us to incorporate such data as networks of interactions. These networks can be reconstructed based on knowledge resources such as literature or specific databases (e.g. KEGG, Reactome, Transfac) or purely from experimental data by inferring interactions between genes from their co-expression patterns [103] or mutual information [104]. Literature-based networks are usually reconstructed by merging interactions from a different biological context (like different cell types, tissues or even organism) in an attempt to include all the information relevant to capturing the essential events and describing a particular biological system. The resulting network is usually noisy, with false interactions that are not active in the specific biological context under study. Network analyses focussed on topological features are robust against these 'false' interactions and these networks find their utility. However, detecting the molecular mechanisms involved in complex processes like cellular reprogramming is something which requires more accurate networks, given that a single incorrect interaction can make a big difference in both global and local network dynamics, thus hindering the identification of molecular switches and a proper strategy with which to perturb them and induce desired cellular transitions.

In this section we present a computational method which uses network stability to guide an iterative network pruning of literature-based network interactions. These interactions are apparently not active in the biological context under study according to expression data. This pruning is driven by the compatibility between predicted and experimentally verified stable gene expression patterns. Hence, it is reasonable to assume that interactions removed by pruning are not present in these steady states, given that they are inconsistent with expression data. Once these interactions have been removed, the resulting contextualized network can be analysed to detect molecular switches stabilizing transcriptional programs. This can in turn facilitate the prediction of missing expression values or the validation of specific expression values from noisy experimental data (see Figure 14). In addition, depending on how well the experimental expression data is explained by the network, predictions performed on this

network are more or less reliable, meaning that the matching between experimental expression data and computed attractors can be considered an indicator of network completeness and the goodness of the assumed regulatory logic rules, i.e., interactions between regulators. For validation purposes, and given that our method relies on network stability analysis of different cellular conditions, we selected four examples of transitions between different cellular phenotypes. In these cases we assumed that cellular phenotypes correspond to stable steady states of GRNs describing these processes. The analysed examples include: i) HL60-neutrophil differentiation (HL60), ii) epithelial to mesenchymal transition (EMT) and iii) mesodermal progenitor cells differentiation to osteoblasts (MPC). The method performance was tested in these examples, showing its predictability power.

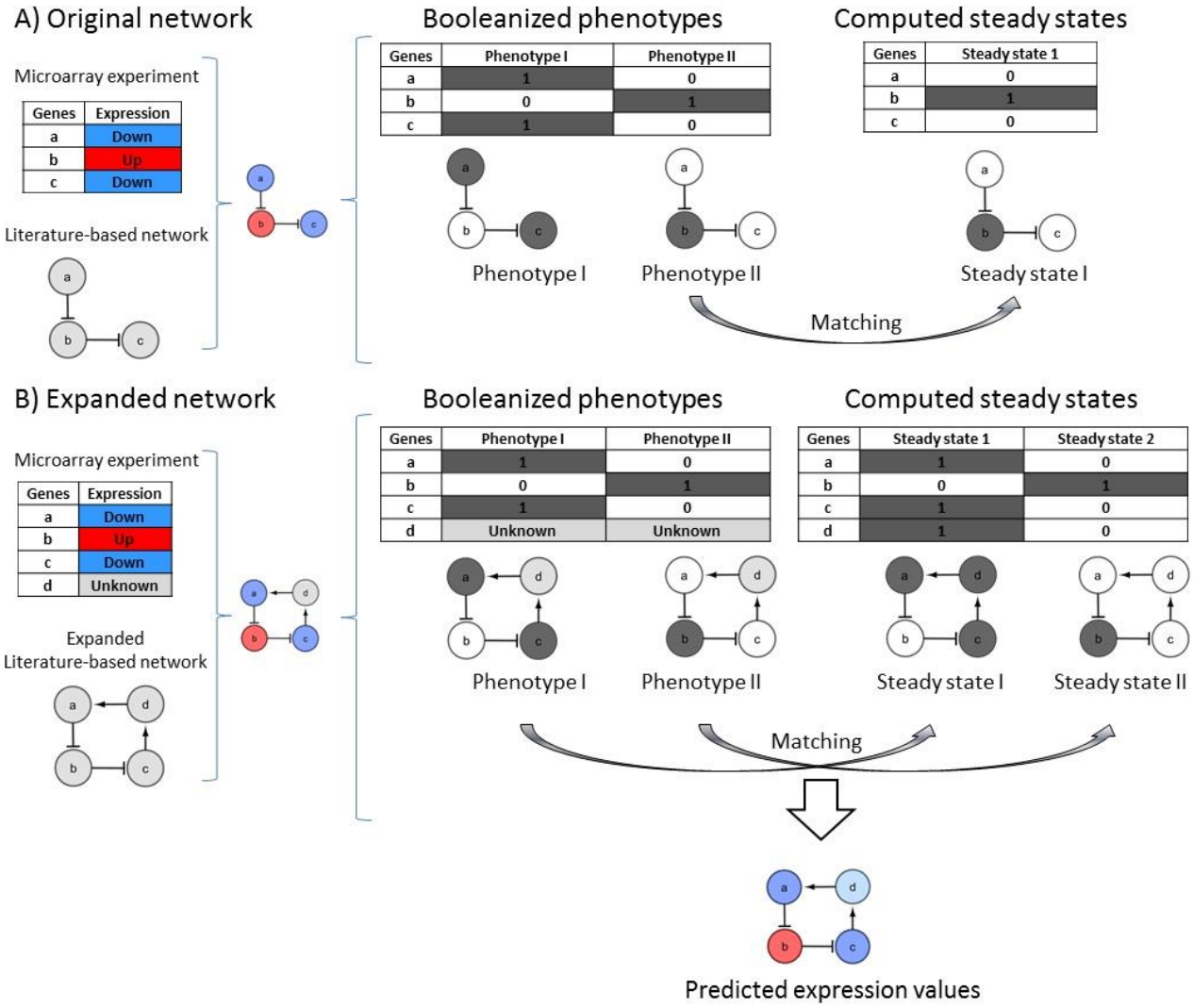


Figure 14 |Network contextualisation to predict missing expression values in order to expand and original gene regulatory network. (A) Gene regulatory network with three genes and two inhibitions. Two Booleanised phenotypes are generated from a microarray experiment. Nodes

in blue and red represent genes down and up-regulated, respectively, according to microarray experiments. Nodes in grey and white represent genes ON (1) and OFF (0), respectively, in the Booleanized phenotypes. The attractor computation of this network in a dynamical Boolean system with a synchronous updating scheme which provides only one steady state corresponding to the phenotype 2. (B) Gene regulatory network with four genes, two inhibitions and two activations. Only the expression values of genes a, b and c are known, while the expression value of gene d (in pale grey) is missing. Nodes in blue and red represent genes down and up-regulated, respectively, according to microarray experiments. Nodes in dark grey and white represent genes ON (1) and OFF (0), respectively, in steady states (attractors) computed according to a Boolean dynamical model. Gene d is predicted as down-regulated (in pale blue).

Comparing previous approaches for inference of regulatory and signaling networks

An important characteristic of our method is that it explores a reduced search space due to the fact that only interactions previously reported in literature can be included in the network. Methods purely based on experimental data [104,105,106] rely on a large amount of data in order to statistically validate network interactions and explore larger search spaces since interactions are not constrained by literature information. In some cases, literature-based methods can also deal with large search spaces, especially when additional interactions can be added and/or regulatory logic rules are flexible [49] [50].

A clear advantage of the method which we present here is that only a single experiment is required: a microarray experiment comparing two stable states of a biological system. Whilst other approaches combine literature information with experimental data, these require a significant number of perturbation experiments, i.e. different combinations of inputs and outputs [50] [49]. In order to be able to train the model, these methods require perturbation experiments targeting different starting points in the network including combinations of perturbations to solve the cross-talking between different pathways in the graph until the entire network is covered (see Figure 15). The main difference between our method and that developed by Irit Gat-Viks and co-workers [50] lies with the fact that although the approach was conceived to be able to analyse directed and signed networks with regulatory feed-backs, the algorithm updates the state of the network transforming the original graph in an acyclic graph. The algorithm then computes the local consistency between the state of a given node and its regulators, starting from a perturbed node and following a topological ordering on the graph's nodes (which exists, since the graph is acyclic). Starting from different parts of the network and constructing the corresponding acyclic graphs, the algorithm finally covers the entire network. In order to be able to train the model using this strategy perturbation experiments are essential, and specifically experiments targeting different starting points in the network including combinations of perturbations to solve the cross-talking between different pathways in the graph (see Figure 15). As a result of this, several perturbation experiments are required. In the particular case of the work published by Saez-Rodriguez et al. [49] the same explanation can be applied; given that they modelled the response of certain elements of a signalling network under perturbation, or in other words the cross-talking between different

paths in a signalling pathway. Indeed, several combinations of input perturbations up-stream have to be experimentally tested in order to gauge the output response down-stream.

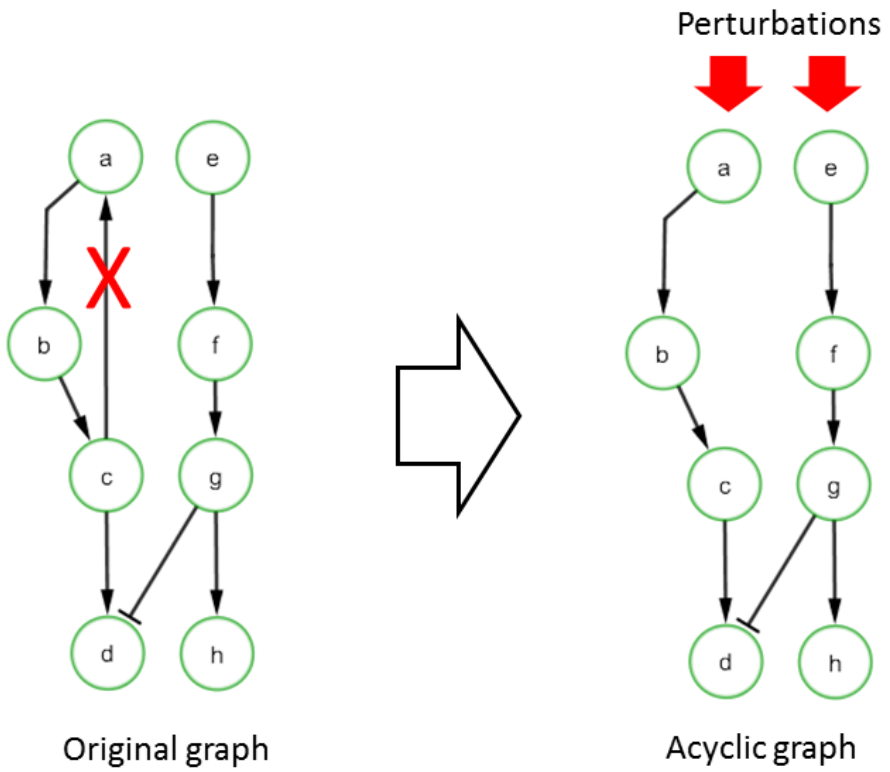


Figure 15 | Small example to illustrate the methodology applied by Irit Gat-Viks et al. The removal of specific edges means that the graph can be transformed into an acyclic graph. Node states are updated starting from different parts of the network (nodes 'a' and 'e') and following the topological ordering on the graph's nodes. Whilst the separated perturbation of 'a' and 'e' is required, so to is the combined perturbation so as to infer the regulatory mechanism of the node 'd', which constitutes a cross-talking between two pathways.

Thus, in the specific case of the work published by Irit Gat-Viks et al. [2] the strategy was tested on the lysine byosynthesis pathway on yeast, with the data set of perturbations comprising five groups of high-throughput experiments: a) expression profiles in nitrogen depletion medium after ten different periods of incubation [3]. b) Expression profiles in amino acid starvation after 5 different periods of time [3]. c) Expression profiles of His and Leu starvation and various GCN4 perturbations [4]. d) Protein and mRNA profiles of wild type strain YPD and minimal media [5]. e) 80 growth sensitivity phenotypes measured for each of a collection of ten gene-deletion mutant strains in eight conditions: Lys, Trp and Thr starvation, three minimal media and YPG conditions [6].

In the particular case of Saez-Rodriguez et al. [49], the biological system selected to validate their approach was hepatocellular carcinoma cells exposed to one of seven cytokines in the presence or absence of seven small-molecule kinase inhibitors. They measured the states of 16 intracellular signaling proteins both before and 30 minutes after they had been exposed to ligands. In addition, they also performed the experiments with combinations of two ligands in the presence and absence of small-molecule inhibitors for different protein kinases.

Another remarkable difference between the work published by Irit Gat-Viks et al. [50] and the approach we present pertains to the complete confidence which this previous method has in the experimental data. The assumption is that this information is always correct, with an adjustment of the regulation functions which define the state of a specific node based on the states of its parents in an acyclic graph. If after this process certain discrepancies or local inconsistencies still remain, the model is refined by the addition of a novel regulatory hypothesis (with interactions not described in the literature) using a learning algorithm. With the method we propose, a local inconsistency could be accepted if the global consistency of the computed network state and experimental expression data is increased; a strategy which is suitable for dealing with noisy expression data. We omit interactions which have not previously been described in the literature to refine the model, and only work on the contextualisation of networks with sufficient connectivity to explain missing expression values; the completeness of the network and the suitability of the logic assumed regulatory rules can be estimated with the percentage of matching gene expression between computed attractors and experimental data.

Another important feature of the method presented here is that it provides a strategy through which to increase the match with expression data using an evolutionary algorithm that considers the probability distribution of positive circuits and individual edges in an iterative process. Indeed, this means that it is not necessary to exhaustively explore the entire search space as in previously published works such as Layek et al. [107], who also exploited the attractors of the system. In this work the authors proposed a method with which to infer regulatory networks using a priori information regarding biological pathways and the concordance between network attractors and experimental data. In their method, they integrated information from pathways described in the literature to create a family of possible networks. They then verified whether or not the experimentally observed stable states agreed with computed attractors of the family of possible networks. Following this they selected the top networks (different alternative networks could fit expression data). If the match was not good they could question the validity of the pathway information. This means that the stable states distribution data can be used to assess the accuracy of the pathway information, although there is no method to improve this match. Here, we distinguish our method by providing a systematic technique to improve the match between experimental and computed steady states.

Finally, it is worth noting that our method exploits global network information, i.e. network stability, whereas several other methods have relied on local network information, such as pairwise gene expression covariation [104,105], or response to perturbations of specific genes [108]. Hence, our method represents a good compromise between robustness in predictions

(provided by the utilisation of a general property which takes into account all the information contained in the network) and the amount of required experimental information.

3.1.2 Results

Principle of the approach

The method proposed here enables us to generate optimized pruned networks from literature based on the knowledge of experimental stable gene expression. This approach involves searching for the optimal populations of solutions of pruned networks using an estimation of distribution algorithm (EDA). Indeed, in this way the method overcomes the limitations of classic optimisation techniques, which try to improve a single solution by exploring a limited portion of the solution space. This allows for the detection of alternative pruned network solutions caused by the multiplicity in network connectivity, thus in turn increasing the probability of achieving a global optimum which best fits the theoretical gene expression values to the experimental ones. It is worth noting that the full agreement between experimental and predicted gene expression values is limited by a lack of information pertaining to network connectivity.

Our method is designed to infer the gene regulatory network which determines stable cellular phenotypes with known expressions values. Assuming that cellular phenotypes correspond with stable states as proposed by several authors [8,109,110], these cellular phenotypes should correspond with steady states of the underlying gene regulatory network. The list of differentially expressed genes provided by the expression data analysis is transformed to generate two Booleanized phenotypes. Following this, our method generates alternative configurations of the original network in order to select those with attractors, computed with a Boolean model, with the best fit to Booleanized phenotypes. This population of alternative configurations of the original raw network is improved by iteratively sampling the probability distribution of positive circuits. This is a necessary condition for multi-stability [111] and individual interactions by means of an EDA, a sort of evolutionary algorithm which is described in the next section.

Estimation of distribution algorithm

EDAs are evolutionary search algorithms which can be applied to high-dimensional optimisation problems and have been applied to several bioinformatic problems [112]. EDAs use a set of selected solutions to create a probabilistic model which guides the search/optimisation process. Compared to other evolutionary algorithms, they avoid premature convergence of solutions, due to the modeling of the probability distribution over many iterations. Within the population of solutions, different patterns of connectivity between genes may be represented as probabilities. This knowledge in terms of probability is used to sample new solutions.

Depending on the complexity of the probabilistic models used to capture the interdependencies between the variables, EDAs can be divided into univariate and multivariate approaches. Univariate EDAs assume that all variables are independent and factorize the joint probability of the selected solutions as the product of univariate marginal probabilities. Multivariate EDAs factorize the joint probability distribution using statistics, and observing

more than one variable at a time. More specifically, it is the possibility of defining the interdependencies between variables which constitutes the main advantage of EDAs when compared with genetic algorithms.

Here we propose an EDA to perform and iteratively prune a literature based-network using populations of alternative pruned networks which are scored and selected using expression data. These selected highest scored pruned networks are used to generate the next population of alternative pruned networks successively until the fulfillment of the stop criteria. For each iteration of the algorithm, the new population of pruned networks is generated by sampling the probability distribution of positive circuits and individual interactions found in the best pruned networks of the previous population. Each pruned network is scored by comparing their predicted steady states with a Booleanized representation of the experimental expression data. Given that the scoring of the pruned networks is based on stability, a property which rests on the global topology of the network makes it impossible for us to assess each interaction separately. In our method, the dependencies between variables (interactions) are captured using information relating to the network topology. Specifically, we treat all of the interactions belonging to positive circuits as a unique entity, considering that this entity is present if and only if all of its interactions are also present or, in other words, if the circuit is complete.

In the expanded gene regulatory network (Figure 14), the contribution of each interaction to generate two steady states cannot be assessed separately due to the fact that all of these interactions are necessary to close the loop and produce a bi-stable behaviour. These interactions are not independent from the stability point of view.

Algorithm steps

In order to search for a set of alternative optimised pruned networks to explain the experimental expression data, the following algorithm was implemented in four steps (Figure 16):

1. Generation of an initial population of pruned networks. During this step the first population of pruned networks is generated through the random removal of interactions from the original literature-based network. The only constraint we introduce is that all networks are forced to include at least one positive circuit (a necessary condition for multi-stability). This positive circuit is randomly selected from the pool of all positive circuits in the original literature based network. The population size is defined by the user; a larger population size increases the likelihood of achieving global optimum, although it also increases computational expense and, in general, requires more iteration to converge with one or multiple solutions.

2. Selection of best-scored pruned networks. Each pruned network is scored using the objective function (described below) whilst a defined number of best-scored pruned networks are also selected. The user can define this selection number. In the examples included in this dissertation, and in order to illustrate the method, we used a selection number which represented 50 % of the population size.

3. Termination criteria. The algorithm verifies the fulfillment of the stop criteria (defined by the user): either the maximum number of iterations is reached, or all the scores in the population of pruned networks are higher than a defined value (e.g. 80%). If these criteria are not fulfilled the algorithm proceeds by generating the next population of pruned networks.

4. Generation of next pruned network population. The next population of pruned networks is created by sampling the probability distributions of each positive circuit and individual interaction; both of which are calculated from the best-scored pruned network selection. This makes it possible to decide whether or not circuits and individual interactions are included in the new pruned networks. In other words, taking into consideration the top scoring pruned networks, we assess the number of times that one specific positive circuit appears, creating a background for random generation of the next population. For example, assume that a hypothetical set of ten pruned networks has been selected due to their high scores, and that one specific circuit is present in seven of these ten pruned networks. In this case, the probability of the circuit is 0.7 and when we generate the next population of pruned networks, on average 70% of the new networks will have this circuit. Once the circuits are sampled we follow the same sampling on individual interactions in order to model interactions which are not present in selected circuits. Additionally, in order to retain the best scoring networks we implement elitism - pruned networks with the best scores within the subset of selected pruned networks are directly transferred to the next generation. The algorithm also introduces a certain amount of noise during the optimisation process by sampling the truncated probability of both circuits and interactions. These probability distribution values are truncated to 0.2, in the case of frequencies lower than 0.2, and to 0.8 in the case of frequencies higher than 0.8. This strategy avoids convergence (all pruned networks with or without a specific circuit or interaction) by chance, and enables efficient optimisation for smaller population sizes.

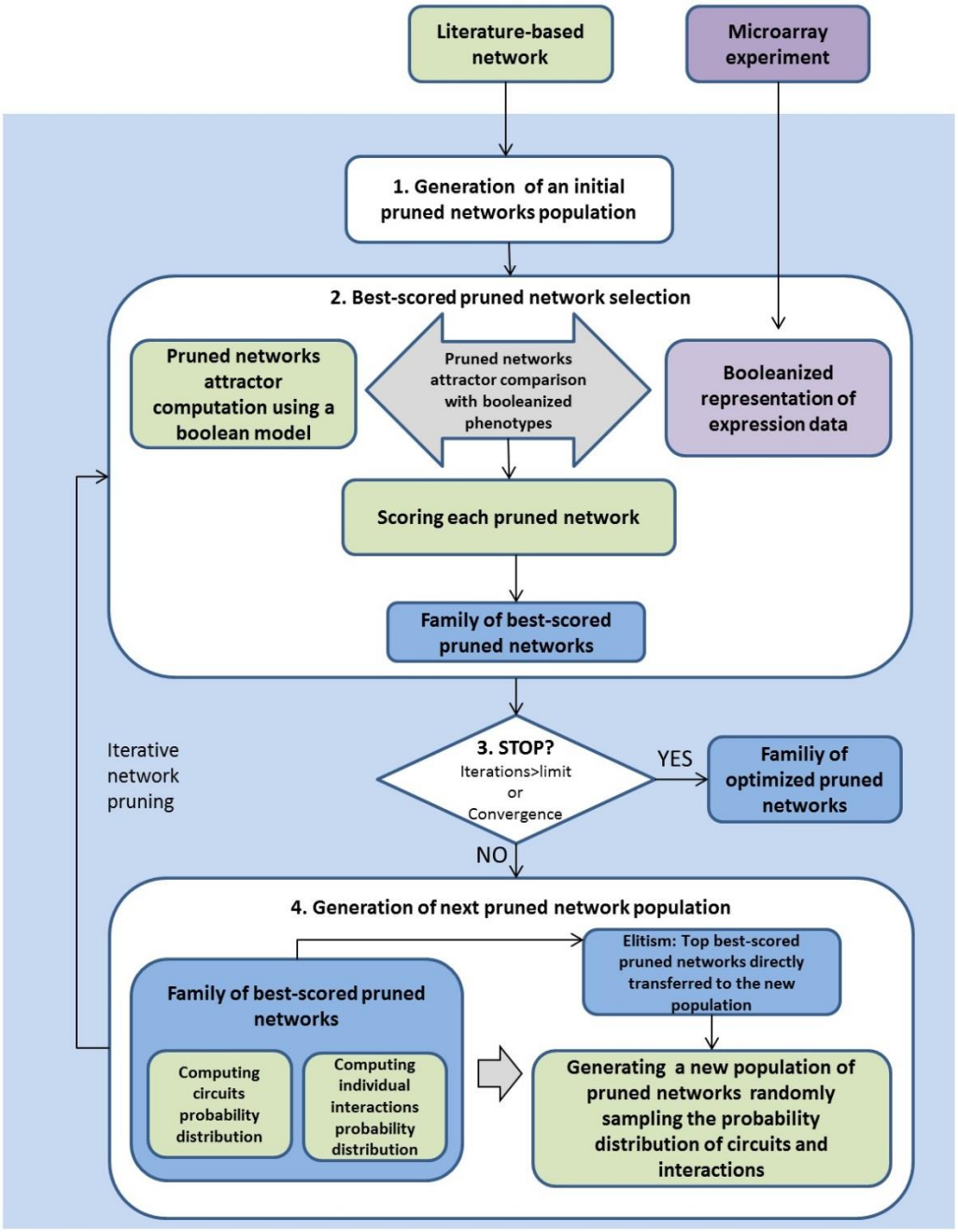


Figure 16 | Iterative network pruning using an estimation of distributions algorithm.

The objective function

The objective function assesses the match between predicted steady states and a Booleanized representation of the experimental expression data, assigning a score to each sampled pruned network (n). This score S uses the normalized Hamming distance (h) to compare N Boolean gene expression values (σ) between all the calculated steady states (α) of a pruned network and

the two known phenotypes (ϕ1 and ϕ2) defined by the expression data. In this way it is possible to identify the two best-matching phenotype/steady state couples (ϕα1 and ϕα2). Finally, the pruned network score (from 0 to 1) is defined as:

$$S_n = \left(1 - \frac{(h_{\varphi\alpha 1} + h_{\varphi\alpha 2})}{2}\right),$$

$$\text{with } h_{\varphi\alpha} = \frac{1}{N}\sum_{i=1}^{N}(\sigma_i{}^{\varphi} - \sigma_i{}^{\alpha})^2$$

A possible extension of the current method could consider not only the existence of steady states, but also cyclic stable states, which would in turn require the existence of negative circuits. Such an extended method could be applied to gene regulatory network inference in biological systems with oscillatory behaviour, such as cell cycles.

Validation of the method

Resulting optimized GRNs can be assessed based on their capacity to explain experimental expression data which is known a priori and to predict unknown expression values.

In order to validate the method we selected three illustrative examples, namely HL60-neutrophils differentiation, EMT and MPC-osteoblast differentiation. We applied the algorithm to these examples using different training sets (different in size and composition) and compared the distribution of network scores generated by the algorithm with the distribution of scores corresponding to a population of randomly generated expression patterns. The underlying idea of this cross-correlation study was to demonstrate that our strategy, which aims to predict missing expression values (for different training and predicted sets) actually performs better than random predictions. Results demonstrated that the P-values for the similarity of the two distributions were very low for the three examples, thus stressing the statistical significance of predictions obtained by our algorithm. Figure 17 shows the cumulative frequency distributions of the scores for each example, thus illustrating that pruned networks tends to have steady states which efficiently describe cellular phenotypes.

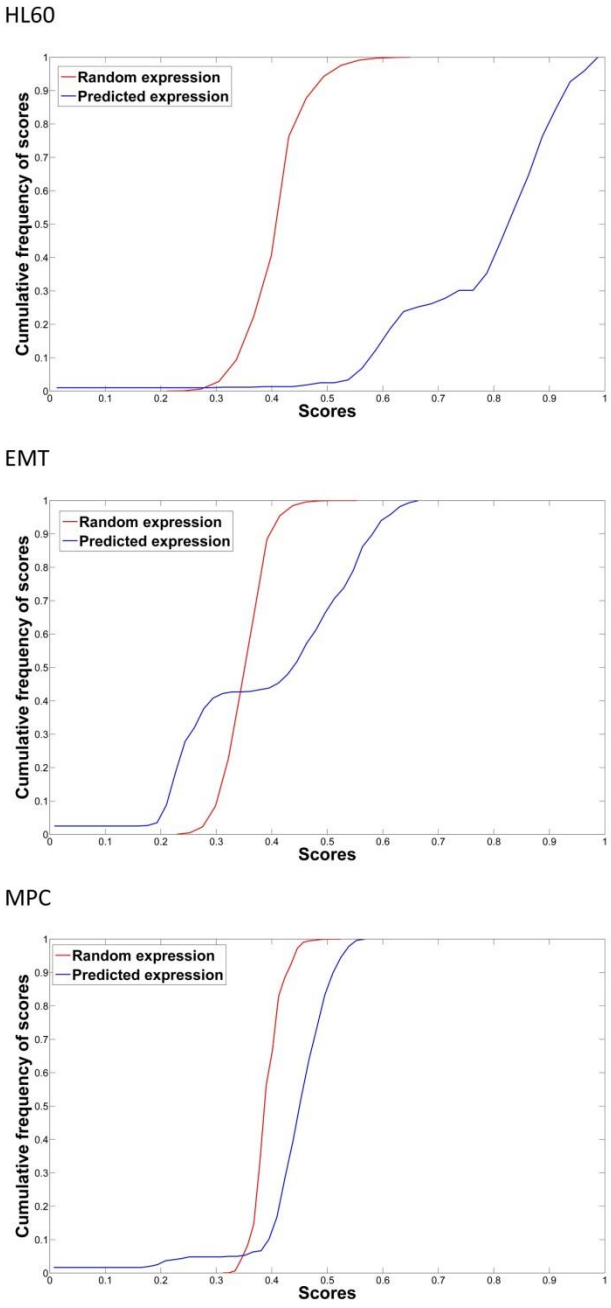


Figure 17 | The cumulative frequency distribution of the scores, which indicate similarity to the experimental phenotypes, applying the algorithm for the HL60 (top), EMT (middle) and MPC (bottom) networks. The above plot shows that for example, 10% of the highest scoring expression patterns are above 0.47, 0.39, and 0.42 for HL60, EMT and MPC networks, respectively. The corresponding scores using the optimized networks increased to 0.94, 0.59 and 0.52. Evidently, for less complex networks such as HL60, the prediction performance increases. The corresponding P-values for the Welch's t-test of the hypothesis that both distributions have the same true mean are 2.2e_16, 2.812e_9 and 2.2e_16.

More specifically, 10% of the highest scoring random expression patterns had scores above 0.47, 0.39 and 0.42 for HL60, EMT and MPC networks, respectively. The corresponding scores using the optimised networks increased to 0.94, 0.59 and 0.52, whilst the corresponding P-values using the Welch's t-test were $2.2e^{-16}$, $2.812e^{-9}$ and $2.2e^{-16}$. Results demonstrated that the p-values pertaining to the similarity of the two distributions were very low for the three examples, thus stressing the statistical significance of predictions obtained by our algorithm.

3.1.3 Discussion

Here we propose a method which enables us to contextualize literature based GRNs to specific biological conditions by exploiting experimental expression data corresponding to stable expression patterns. This approach, which searches for optimal populations of solutions of pruned networks, overcomes the limitations of classic optimisation techniques which try to improve a single solution by exploring a limited portion of the solution space. This allows for the detection of alternative pruned network solutions caused by the multiplicity in network connectivity. This multiplicity in turn increases the probability of achieving a global optimum which best fits theoretical gene expression values to the experimental ones. It is worth noting that the full agreement between experimental and predicted gene expression values is limited by a lack of information regarding network connectivity.

In order to validate our method, we used optimized networks to predict missing expression values in a cross-correlation study using different sets of expression values to train the model and to be predicted in three selected examples (HL60-neutrophil differentiation, Epithelial-Mesenchymal transition, Mesodermal progenitor-Osteoblast differentiation). Results showed a consistency between predicted and experimentally validated gene expression values higher than expected by chance.

In summary, the presented method constitutes a useful tool for the literature based contextualisation of GRNs. The resulting contextualized network can be used to curate experimental gene expression data, data analysis, modelling and prediction. A possible extension of the current method could consider not only the existence of stable steady states (fixed points), but also cyclic stable states, which would require the existence of negative circuits. Such an extended method could be applied to GRN inference in biological systems with oscillatory behaviour, such as cell cycles.

3.1.4 Methods

Computation of attractors

In order to compute the attractors, we model the network as a dynamical system using a deterministic rule-based approach or, more specifically, a Boolean dynamical model. Other possible dynamic models include continuous models, which have the benefit of being easily compared to quantitative experimental data [113], and discrete models with more than two possible values [114]. However, since the continuous models would have to be studied numerically as opposed to analytically, the computation of attractors becomes computationally expensive. Furthermore, biological regulatory processes are such that the graph of rate of expression between a regulated gene as a function of its regulator, commonly exhibits a sharp sigmoid curve, which can be approximated to a Boolean switch-like behaviour [115].

Within this Boolean dynamical model the network is created as a graph, which is directed and signed, in order to represent positive or negative regulation. Nodes represent genes whilst edges denote regulation. Each node has an associated value of "1" or "0", encoding the

activation/presence or inactivation/absence respectively. The logic functions which encode the regulation for each specific node are represented using the disjunctive normal form representation, which uses only AND, OR and NOT operators [116]. Given the regulators (activators and inhibitors) for each node, the Boolean function is evaluated using rules proposed by [117]: if none of its inhibitors and at least one of its activators are active, then a gene becomes active; otherwise, the gene is inactive. Finally, we use a synchronous updating scheme [118], whereby all genes in the network update their expression levels simultaneously in each time step. We use the synchronous updating scheme as it facilitates computation due to the smaller state space, and yet preserves the generic qualitative properties of the network [119]. An alternative updating scheme, which we did not investigate, would be the asynchronous scheme. This has a much larger state space, leading to a higher complexity of computing attractors [118]. With this synchronous updating scheme all of the genes are updated from one step to the next at the same time.

Using the set of Boolean functions for each node and synchronous updating, we then computed the attractors of the network, i.e. the set of states towards which a dynamical system evolves over time. The attractors were computed using an efficient method to model the network dynamics using Reduced Order Binary Decision Diagrams (ROBDD or in short BDD). This was done due to their compact representation of Boolean functions and the ease of computing complex Boolean operations [117]. More details regarding attractor computation are included in the Supplementary information.

Computation of circuits

The Johnsons algorithm [120] was implemented as a perl program 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 other than the first and last appears twice.

Cross-correlation study

In our biological examples, and in order to statistically validate predicted expression values, we compared the distribution of network scores generated by our optimized pruned networks from multiple training sets, with the distribution of scores corresponding to a population of randomly generated expression patterns (Figure 6). This population of random expression patterns was generated by randomly assigning one of the following values for each gene in the network: up-regulated, down-regulated, invariant-up and invariant-down. These values correspond to genes that in a Booleanized model change from 0 to 1 and from 1 to 0 in the first two cases. These genes also remain invariant in 1 and 0 for the latter two cases, respectively. Once we had assigned values to all genes, expression patterns were scored using the Booleanized phenotypes from experimental data. This scoring scheme is identical to that used during the optimisation process, thus reflecting the match between the random expression pattern and the experimental Booleanized phenotypes. We repeated the process 10000 times, obtaining a population of random expression patterns with the respective scores. Following this, we compared the population with the population of optimised pruned networks (30

alternative pruned networks, which constitute the last population of optimized pruned networks after the last iteration of the algorithm) for 20 different and randomly selected training and predicted sets of genes. We used different training sets to perform this cross-validation, since not all possible training sets are equally predictive due to the fact that not all genes are equally informative according to our method. For example, highly connected genes are, generally speaking, more informative than genes with few interactions. Preliminary tests showed that the optimal percentage of genes for which gene expression values can be predicted was 35%, based on the expression values of the remaining 65%. For instance, in the HL60 network, which includes 18 genes, 12 genes were used to predict the expression values of the remaining 6 genes for 20 different combinations of training and predicted genes. Following this, we scored the match between predictions and expression data using the same scoring process as in the pruning, although in this case all 18 genes were taken into consideration since some of the computed expression values of the training set genes could mismatch with experimental expression values. Therefore, we had a population of 600 pruned networks (30x20) with the corresponding scores. We then proceeded to compare this population with the randomly generated population of scored expression patterns. This comparison was made in order to show that gene expression predicted values were better than those predicted by chance.

Further, we used the Welch's t-test to estimate the similarity between the predicted and randomly generated populations of gene expression values scores.

Examples: **network reconstruction**

Cellular differentiation is a process central to our understanding of the nature of multicellular living systems, their stability in a changing environment, and how such systems fail in diseases. The relationship between attractors and cellular phenotypes has been proposed by several authors [109,110] [8]. Given that our method rests on the stability analysis of the system, we decided to work with cellular differentiation networks, adopting the assumption that cellular phenotypes correspond to steady states or attractors of gene regulatory networks which rule the differentiation process.

We chose three cellular differentiation processes as examples to illustrate the method: (i) HL60-neutrophil, (ii) mesodermal progenitor cells (MPCs) -osteoblast and (iii) the epithelial-mesenchymal transition (EMT) gene regulatory network.

The procedure for the network reconstruction was identical in the three examples and consisted of the following steps: a) obtaining a list of differentially expressed genes between two classes corresponding to long term expression patterns or cellular phenotypes. b) Connecting these genes using expression regulatory interactions from literature. c) Determining regulatory cores and the genes regulated by them.

a) Obtaining a list of differentially expressed genes.

   <u>HL60-neutrophils differentiation</u>

The multipotent promyelocytic leukemia cell line HL60 was originally isolated by Dr. Steven Collins from an acute promyelocytic leukemia (APL) patient [121]. The HL60 system was used by Huang *et al.* (2005) [8] to demonstrate the correspondence between cell fates and high-dimensional attractor states of the underlying network. In order to reconstruct the HL60-neutrophil differentiation gene regulatory network we used a set of genes composed of genes differentially expressed between HL60 cells (precursor or phenotype 1) and neutrophils (phenotype 2), differentiation induced by dimethyl sulfoxide (DMSO) in the experiment performed by F. Mollinedo and coworkers [122].

Epithelial-mesenchymal transition

Epithelial cancer cells are capable of transiting from an epithelial to a mesenchymal state, a key step towards the formation of metastasis. The EMT master transcription regulator SNAI1 (human snail) triggers a transcriptional program leading the transition from epithelial to mesenchymal. In the experiment performed by [87] this transition was triggered by the induced expression of SNAI1. In the case of the EMT, we used a set of differentially expressed genes between epithelial and mesenchymal cells obtained from an experiment performed by Vetter *et al.* [87] whereby the transition was triggered by the induced expression of SNAI1.

Mesodermal progenitor cells-osteoblast differentiation

Single human bone marrow-derived mesodermal progenitor cells (MPCs) differentiate into osteoblasts, chondrocytes, adipocytes, myocytes and endothelial cells. In the experiment performed by Qi *et al.* [123], MPCs were induced to differentiate into osteoblasts, cells involved in bones formation, through the addition of dexamethasone, ascorbic acid and β-glycerophosphate to the cell cultures.

A microarray analysis performed afterwards yielded a list of differentially expressed genes between osteoblasts and MPCs.

b) Connecting differentially expressed genes using gene regulatory interactions described in literature.

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 [67,68]. MedScan processes sentences from PubMed abstracts and produces a set of regularized logical structures representing the meaning of each sentence. We selected only the interactions included in the ResNet mammalian database in the category of expression, thus indicating that the regulator changes the protein level of the target, by means of regulating its gene expression or protein stability. Following this step we obtained 3 raw networks which were reduced, removing irrelevant nodes for the stability analysis, i.e., nodes without incoming edges.

c) Determining regulatory cores and genes regulated by them.

Given that only genes with incoming interactions are relevant to the stability analysis, we had to identify genes involved in regulatory feed-back loops, or circuits, as well as genes regulated by them. In order to accomplish the first task we looked for strongly connected components (SCCs) in the raw network using Binom plugin [69] in Cytoscape [70]. Afterwards we obtained one regulatory core for each example consisting of the strongly connected components and the connections between them. These connections could simply just interactions between one SCC and another, or paths which involve interactions and genes. Figure 18 displays the regulatory core of the HL60-netrophils differentiation network. The activation of CR1 by IRF1 constitutes an example of a connection through a simple interaction. The path IL1B->S100A8->IL8 constitutes the other example, with a path connecting the two SCC. Following this, we expanded these cores iteratively by initially adding neighbours regulated by the regulatory core until no further neighbors could be added. In the possible scenario of multiple disconnected regulatory cores, the network pruning and predictions were performed independently. After this step we obtained 3 gene regulatory networks (all nodes with incoming interactions).

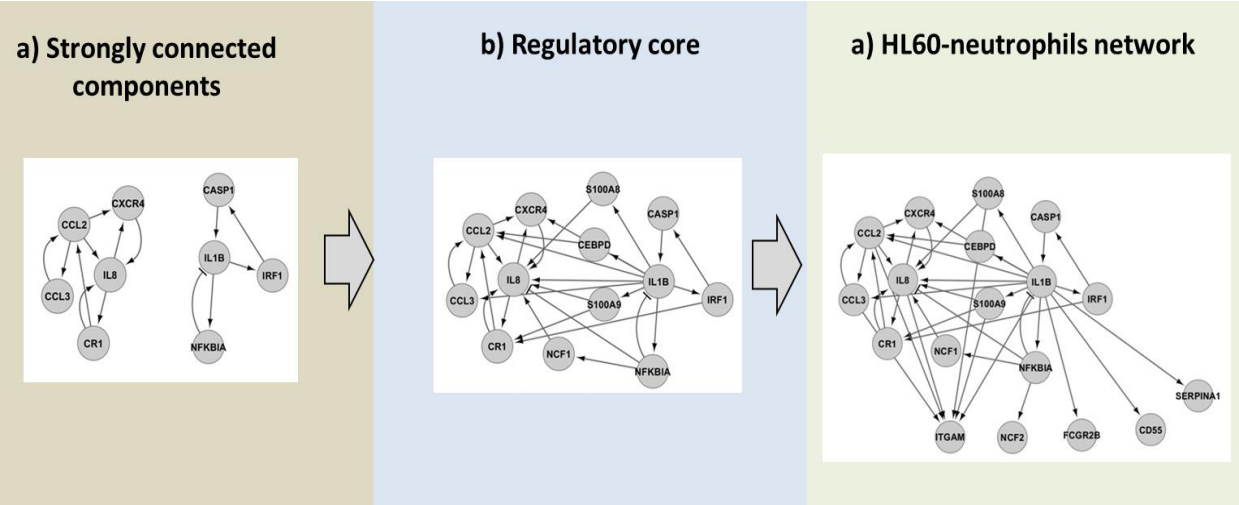


Figure 18 | Strongly connected components, regulatory core and the final HL60-neutrophils gene regulatory network. The regulatory core is composed of two different strongly connected components (with five and four genes each) and the connections between them. HL60-neutrophils gene regulatory network is composed of this regulatory core and five more genes directly regulated by the regulatory core.

	HL60-neutrophils	EMT	MPC-osteoblast
Number of nodes	18	46	67
Number of edges	39	129	123
Activations	37	92	72
Inhibitions	2	37	51

Table 2| Gene regulatory networks of three biological examples: HL60-neutrophil differentiation network, Epithelial-mesenchymal transition network, MPC-osteoblast differentiation network. Information about number of nodes and edges in table XXX.

HL60-neutrophil differentiation network

We were able to reconstruct a gene regulatory network with 18 genes and 38 interactions representing positive or negative effect over gene expression (see Figure 2A and Supplementary file II). The regulatory core includes two SCC.

Epithelial-mesenchymal transition network

In this case we obtained a gene regulatory network with 46 genes and 129 interactions representing a positive or negative effect over gene expression (see Figure 2B and Supplementary file II). The regulatory core includes one single SCC.

Mesodermal progenitor cells-osteoblast differentiation network

The resulting network in this case includes 67 genes and 123 interactions representing positive or negative effects over gene expression (see Figure 2C and Supplementary file II). The regulatory core includes four SCC.

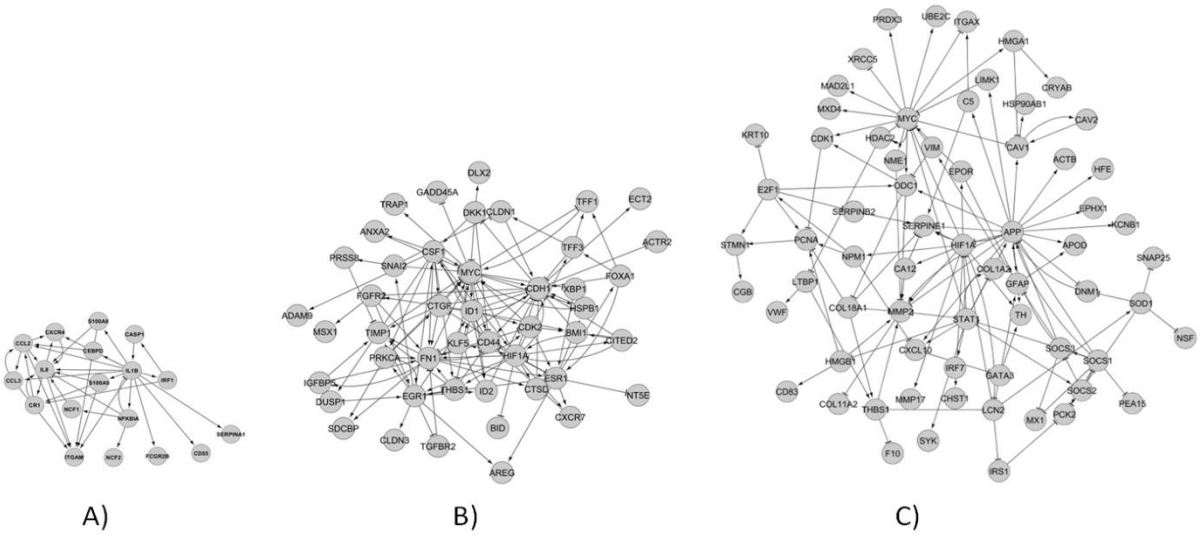


Figure 19 | Gene regulatory networks of three biological examples. 3A) HL60-neutrophil differentiation network, 3B) Epithelial-mesenchymal transition network, 3C) MPC-osteoblast differentiation network.

3.2 Detecting Cellular Reprogramming Determinants by Differential Stability Analysis of Gene Regulatory Networks

This section refers to the work published in BMC Systems Biology in 2013 entitled "*Detecting Cellular Reprogramming Determinants by Differential Stability Analysis of Gene Regulatory Networks*". In addition, it also refers to the unpublished work on the astrocytes to neural progenitor cells (NPCs) dedifferentiation as experimental validation of the described methodology to design recipes for cellular reprogramming. The experimental validation constitutes a collaborative effort between three parts: Department of Neuroscience Institute of Psychiatry King's College London (Dr. Angela Bithell, Jannis Kalkitsas and Prof. Dr. Noel J Buckley), contributing with the perturbation experiments performance and analysis, The Experimental Neurobiology group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg (Dr. Alessandro Michelucci), contributing with the microarray experiments, and the Computational Biology group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, contributing with the computational modelling and analysis [20].

3.2.1 Introduction

During classical cellular differentiation cells lose phenotypic plasticity until they become fully differentiated. Certain differentiated cells have the remarkable ability to be converted into different cell types via a process termed developmental redirection or cellular reprogramming. Both processes are carefully orchestrated by the activation and repression of specific sets of genes. Knowledge relating to these activations and repressions can be integrated as networks of interactions, thus allowing us to describe biological processes in general and transitions between network states and cellular reprogramming in particular, as transitions between stable

steady states (termed as attractors) of these networks. The relationship between cellular phenotypes and attractors has been proposed by several authors [8,109,110], whilst the last few years have witnessed the experimental validation of a number of examples showing that only few genes can induce transitions between cellular phenotypes [124,125,126]. The main motivation of this work is the identification of combinations of key genes, termed reprogramming determinants (RDs), which are effective in such transitions when perturbed and therefore with practical applications for cellular reprogramming. Although there are a number of approaches in the literature which can help to predict effective cocktails of transcription factors for cellular reprogramming [127,128] they either require a list of candidate genes to narrow down the combinatorial problem or are based on computational brute force to simulate network response under perturbation of combinations of genes. Indeed, the latter strategy becomes prohibitive when increasing the number of genes in the network.

Here we present a computational method with which to identify, without preliminary selection of candidate genes, reduced subsets of reprogramming determinant genes which can induce transitions between cellular phenotypes when perturbed. The method relies on the expression profiles of two stable cellular phenotypes and a topological analysis of differential stability elements of the gene regulatory network. It represents a useful framework which can assist researchers in the field of cellular reprogramming, particularly with regard to the design of experimental strategies. It has potential applications both in regenerative medicine, disease modelling and basic research.

3.2.2 Results

The method presented was conceived to design recipes for cellular reprogramming without the need for a prior list of candidates. This was achieved through the combination of expression profiles, network topology and stability analysis. The algorithm dramatically reduces the huge search space constituted by all possible combinations of genes by focussing on genes involved in the stability of the gene regulatory network (GRN). It rests on experimental expression profiles and the identification of differential stability elements for two given cellular phenotypes involved in a specific cellular transition. Such identification allows for destabilisation of the initial cell transcriptional program and stabilisation of the final one by perturbing molecular switches which define the epigenetic barrier between two given attractors. Stable cellular phenotypes are part of a large space of all available cellular states. At the transcriptional level, they represent stable expression patterns or transcriptional programs. The existence of multiple attractors in a GRN requires the presence of positive feedback loops or positive circuits (including an even number of inhibitions) [111]. However, not all positive circuits in the network are involved in network multistability; those whose constitutive genes cannot be in a coherent stable state according to the connectivity of the circuit (due to the connectivity of the circuit with the rest of the network) are not contributing to the stabilisation of the cellular program as they are not stabilized by it themselves. Moreover, there are some positive circuits which contribute to the stabilisation of one particular attractor but not another. Our method relies precisely on the identification of these differential stability elements between two given attractors and the design of perturbation protocols able to target all of them. We termed this subset of positive circuits differentially expressed positive circuits

(DEPCs), given that all of their constitutive genes change their expression values between two given attractors of the GRN. At the same time, these expression values should match with two attractors of the circuit when considered in isolation, i.e., the circuit is stable itself so it can stabilize other elements in the network. Indeed, experiments have shown that few driver genes are able to lead cellular systems from one stable phenotype to another [124,125,126]. This fact prompted us to add a final step to look for minimal combinations of reprogramming determinant genes capable of directly or indirectly targeting all DEPCs, given that our method also identified suboptimal (in number) protocols of perturbation. Analysis of a large number of randomly generated gene regulatory networks showed that these minimal sets of driver genes were always able to trigger transitions between all pairs of attractors. Further, we selected five different biological examples of cellular reprogramming in order to validate the applicability of our method. Moreover, we applied our method and experimentally validated a predicted recipe to dedifferentiate astrocytes to NPC. A Double inhibition of a combination of predicted reprogramming determinants was demonstrated as being capable of inducing a cellular transition from mature (non-proliferative) astrocytes to a NPC-like phenotype in morphology, expression of proliferation markers and epigenetic signature. This validated recipe has no previous references in the literature and constitutes a novel strategy used to induce this cellular transition.

These examples provide an experimental validation of the identified sets of RD genes as effective inducers of transitions between cellular phenotypes. The first five biological examples essentially re-discover known recipes and illustrate that our methodology is generally applicable to different cellular systems. The astrocytes dedifferentiation example constitutes the experimental validation of a novel predicted recipe and nicely illustrates that our strategy does not rest on previous knowledge regarding driver genes or reprogramming determinants but can find novel effective combinations when applied blindly to a given cellular system.

Description of the approach

Cellular phenotypes are characterized by stable expression programs at the transcriptional level. The underlying GRN can be conceptualised and described as a Waddington landscape [13,14,15], where cellular phenotypes corresponding to network attractors are represented as wells separated by barriers (see figure 20). These barriers are established by those network elements which stabilise GRNs in their attractors.

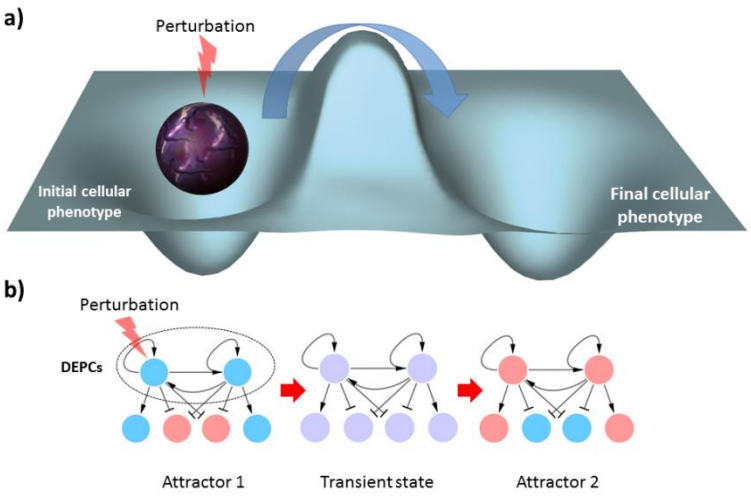
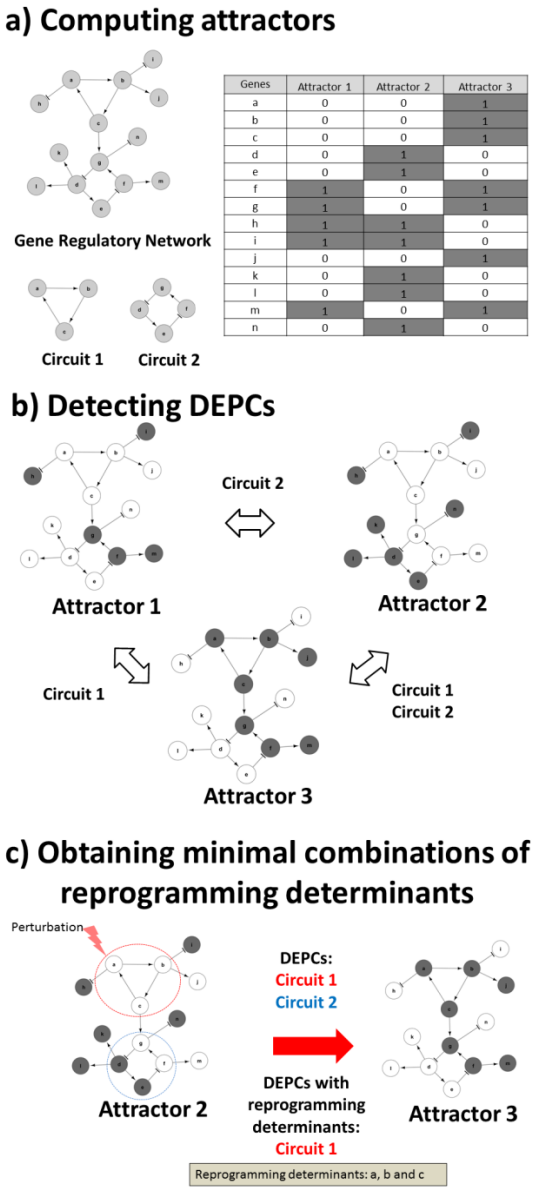


Figure 20| Waddington's landscape versus network based representation. a) Cell transcriptional program landscape representing two attractors and the epigenetic barrier between them. This conceptual figure represents a cell stabilized in an initial cellular phenotype and how a hypothetical perturbation can destabilize the cellular program and make cells exceed the barrier and fall down in a final cellular phenotype. This cellular reprogramming is represented as a blue arrow from the initial to the final attractor. b) Cellular reprogramming as transitions between network states. Differentially expressed positive circuits (DEPCs) are perturbed to induce the transition from Attractor 1 to Attractor 2 passing by a transient state. This transient state can be considered as a "short" term changing expression pattern until the system reaches an attractor. Regular arrows represent activation and T-arrows represent inhibitions. Blue and red nodes represent inactive and active genes respectively in attractors. Violet nodes represent transient states.

The method presented here takes a GRN and the expression profiles of the cellular phenotypes involved in the desired transition as input, and gives minimal combinations of RDs as an output. Therefore, the appropriate perturbation of these genes results in the cellular transition. Since stable cellular phenotypes can be considered as attractors of GRNs, cell fate and cellular reprogramming involve transitions between these attractors, and as such our method looks for combinations of genes able to destabilize a specific initial attractor and stabilize the final one in response to the appropriate perturbation. This strategy allows us to narrow down a huge combinatorial searching problem to a set of minimal combinations which constitutes alternative reprogramming protocols and the output of our method. The method can be described in three steps (see Figure 21: 1) computing GRN attractors 2) detecting DEPCs 3) obtaining minimal combinations of RD genes targeting the DEPCs which (de-)stabilizes the initial and final attractor states, respectively.



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

Figure 21 | Differential stability analysis: recipes for cellular reprogramming in three steps. a) Computing attractors. Network stability is analysed assuming a Boolean model and a synchronous updating scheme. Genes in "1" are active or "ON" and genes in "0" are inactive or "OFF" and are represented in grey and white respectively. **b) Detecting DEPCs.** A positive circuit is considered a DEPC if all of its constitutive genes change their expression values between two given attractors of the GRN. **c) Obtaining minimal combinations of reprogramming determinants.** Both Circuit 1 and Circuit 2 are DEPCs, although Circuit 2 is regulated by Circuit 1; any perturbation of Circuit 1 which is capable of moving it to a different attractor is also going to change the state of Circuit 2. Simulations showed that genes in Circuit 2 did not have to be perturbed to achieve transition from Attractor 2 to Attractor 3. Therefore, minimal combinations of reprogramming determinants are any individual gene of Circuit 1, i.e., genes "a", "b" or "c". Regular arrows represent activation and T-arrows represent inhibitions.

1) Computing attractors of the network: Attractors are calculated with a Boolean model of the GRN where values of "1" and "0" represent up and down regulated genes respectively, and assuming an inhibitory dominant rule (see methods for details).

2) Detecting DEPCs: At first, all positive circuits are detected (see methods section for details). Indeed, information about their constitutive genes as well as the expression profiles of the attractors involved in the cellular transition (initial and final) are combined, yielding a list of DEPCs. For a positive circuit to be DEPC it hast to fulfill two requirements: a) all of their constitutive genes change between the two attractors (i.e., they are differentially expressed), and b) the states of the circuit in both the initial and final phenotypes should match attractors of the circuit when considered in isolation; (i.e., only circuits in a stable state are considered as differential stability elements).

3) Obtaining minimal combinations of RD genes targeting all DEPCs. We looked for the minimal combination of genes which were able to directly or indirectly target all DEPCs. For this purpose the algorithm initially looked for combinations of genes with the requirement that there should be at least one gene for each DEPCs (see methods). This strategy leads to combinations with genes belonging to multiple DEPCs with the consequent reduction of the required number of genes. Following this, and as a final step, the algorithm determined which DEPCs did not have to be directly perturbed (see Figure 21c) by simulating the network response (according to the model assumed to compute attractors) under perturbation of the minimal combination of genes but the gene belonging to specific DEPCs one at a time. Through this we were able to reduce the final number of RDs removing genes targeting DEPCs which are regulated by others.

Validation of the approach in 1000 randomly generated networks

In order to validate this strategy, we applied our method to 1000 randomly generated and different sized networks, but with the same topological properties of a well-characterized gene regulatory network of E. coli. As a result of our analysis we obtained the following conclusions: a) Between any two given attractors we always obtained at least one DEPC; and b) perturbation of minimal combinations of genes which include DEPCs between pairs of attractors always succeeded triggering transitions between these states (see Figure 1 as example). Further, we calculated the percentage of RD genes which can trigger transitions between all calculated attractors. As is shown in Figure 3, interestingly, on average only 14% of the genes from the whole network was sufficient to bring about these transitions. In addition, on average, a maximum of 4 genes and a minimum of 1 gene was sufficient to bring about transitions from one attractor to another.

Application of the approach to five biological examples with effective recipes for transdifferentiation known from literature

We demonstrated the efficacy and the general applicability of the current protocol using five different biological examples of cellular reprogramming. These examples provided an experimental confirmation of the identified combinations of RD genes as effective inducers of transitions between stable cellular phenotypes. The T-helper and EMT examples were based on previously published GRNs [72,113]. In the latter case we expanded the original network with

the addition of a novel double-negative feed-back with miRNA34A, which has been recently published [129]. For the remaining examples (HL60, iHEP, iCM) we used text-mining techniques to construct gene regulatory networks by inferring gene-gene associations between genes found by previous studies to be differentially expressed [9,10,122]. Consequently, these networks were pruned in order to maximize matching between gene expression profiles and gene states found by our network dynamics simulation. This procedure allowed us to contextualize the networks to the biological conditions under which the experiments were performed [130]. More details about the network reconstruction and contextualisation processes are included in the methods section below. Detected driver genes and transitions between known phenotypes are shown in Table 1 for each example.

T-helper

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 are sub-divided into Th0 (precursor) and effector Th1, Th2, Th17 and Treg cells. A T-helper differentiation network determining the fate of the lineage has been proposed previously [Mendoza 2010]. Here we focussed on the transition between Th2 and Th1 phenotypes. We detected T-bet and GATA3 as independent RDs for Th2-Th1 (see Figure 5a) and Th1-Th2 respectively. These predictions are in full agreement with previously published experiments [131] [132,133].

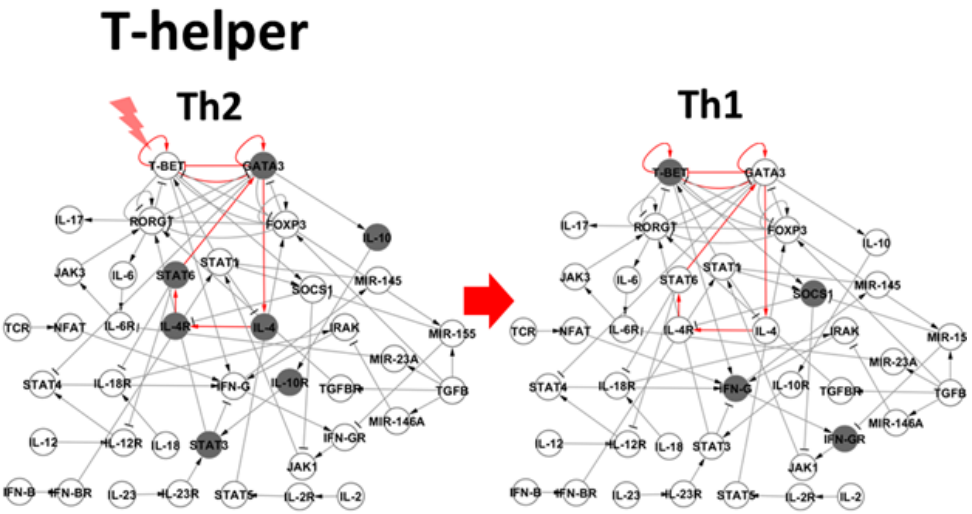


Figure 22 | Simulated perturbations performed assuming a Boolean model succeeded in triggering the transition. These results are consistent with previously published experimental perturbations. Genes in "ON" and "OFF" are represented in grey and white respectively. Regular arrows represent activation and T-arrows represent inhibitions. Perturbed DEPCs are coloured in red.

EMT

A transient phenomenon referred to as epithelial to mesenchymal transition (EMT) occurs during regular embryonic development and as a part of the metastatic cascade initiated by the breakdown of epithelial cell homeostasis in carcinomas. During the Epithelial to mesenchymal transition (EMT), cells change their genetic and transcriptomic program, thus leading to phenotypic and functional alterations. These alterations include the loss of epithelial features like cell-cell adhesions and cell polarity and the gain of cell motility and mesenchymal and stem-like properties. EMT can be initiated by multiple pathways converging in the activation of EMT inducers. The EMT example shows that SNAI1 is a triggering gene for the transition from epithelial to mesenchymal (see Figure 5b), which has been validated by experimental perturbation of this gene [72].

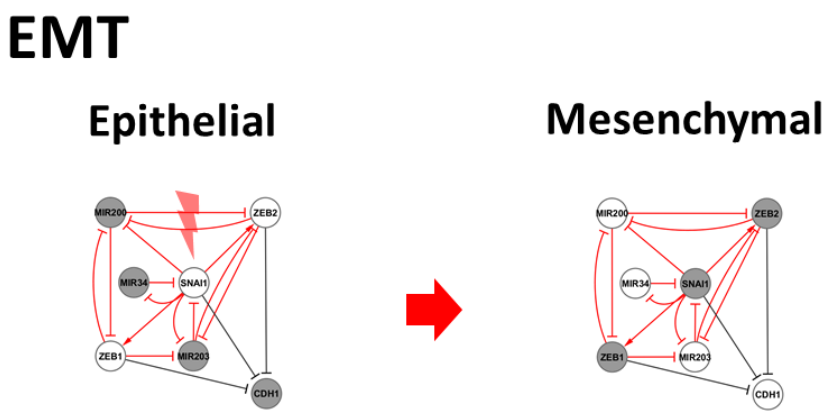


Figure 23 | Simulated perturbations performed assuming a Boolean model succeeded in triggering the transition. These results are consistent with previously published experimental perturbations. Genes in "ON" and "OFF" are represented in grey and white respectively. Regular arrows represent activation and T-arrows represent inhibitions. Perturbed DEPCs are coloured in red.

HL60

The multipotent promyelocytic leukemia cell line HL60 was originally isolated by Dr. Steven Collins from an acute promyelocytic leukemia (APL) patient [121]. The multipotent promyelocytic leukemia cell line HL60 can be stimulated, thus differentiating it into neutrophils using different chemical agents including granulocyte macrophage colony-stimulating factor (GM-CSF)[134], DMSO[135], all-trans-retinoic acid (ATRA) [136], 1,25-dihydroxyvitamin D3[137], and 12-O-tetradecanoylphorbol 13-acetate (TPA)[138]. Nevertheless, the way in which these chemical agents act at the gene regulatory level to induce the transition remains a relevant question when it comes to understanding the underlying mechanisms of

differentiation or reprogramming. Applying our method to the HL60 example allowed us to detect IRF1 as a triggering gene when it comes to inducing the differentiation from HL60 to neutrophil (see Figure 5c). This result is consistent with previous experimental findings [139].

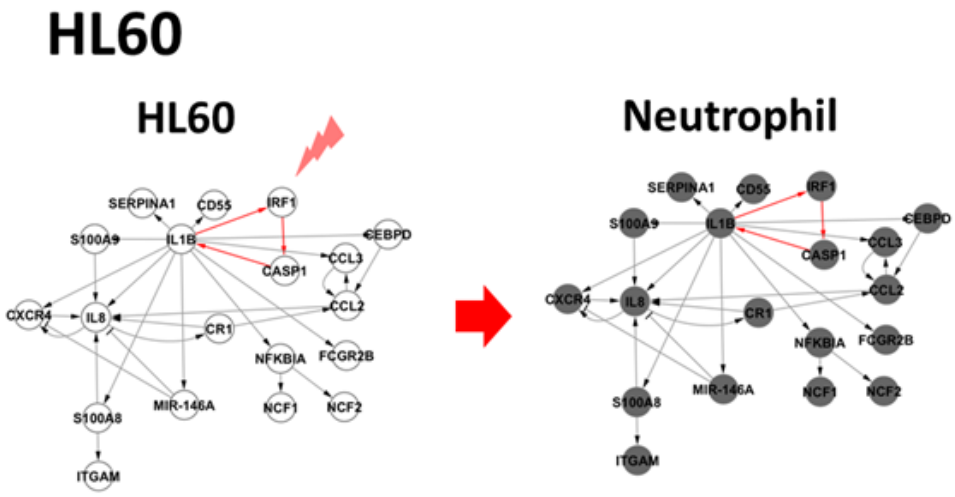


Figure 24 | Simulated perturbations performed assuming a Boolean model succeeded in triggering the transition. These results are consistent with previously published experimental perturbations. Genes in "ON" and "OFF" are represented in grey and white respectively. Regular arrows represent activation and T-arrows represent inhibitions. Perturbed DEPCs are coloured in red.

iHEP

Normally, hepatocytes differentiate from hepatic progenitor cells to form the liver during regular development. However, hepatic programs can also be activated in different cells under particular stimuli or fusion with hepatocytes. The transition from human fibroblasts to hepatocyte-like cells (iHEP) induced by the perturbation of specific combinations of transcription factors has been previously reported by Sekiya & Suzuki, 2011 [140]. In the iHEP example we found several minimal combinations capable of triggering the transition from fibroblast to hepatocyte. Among these minimal combinations, the combined perturbation (activation) of HNF4A and FOXA2 has been experimentally validated [140] (see Figure 5d).

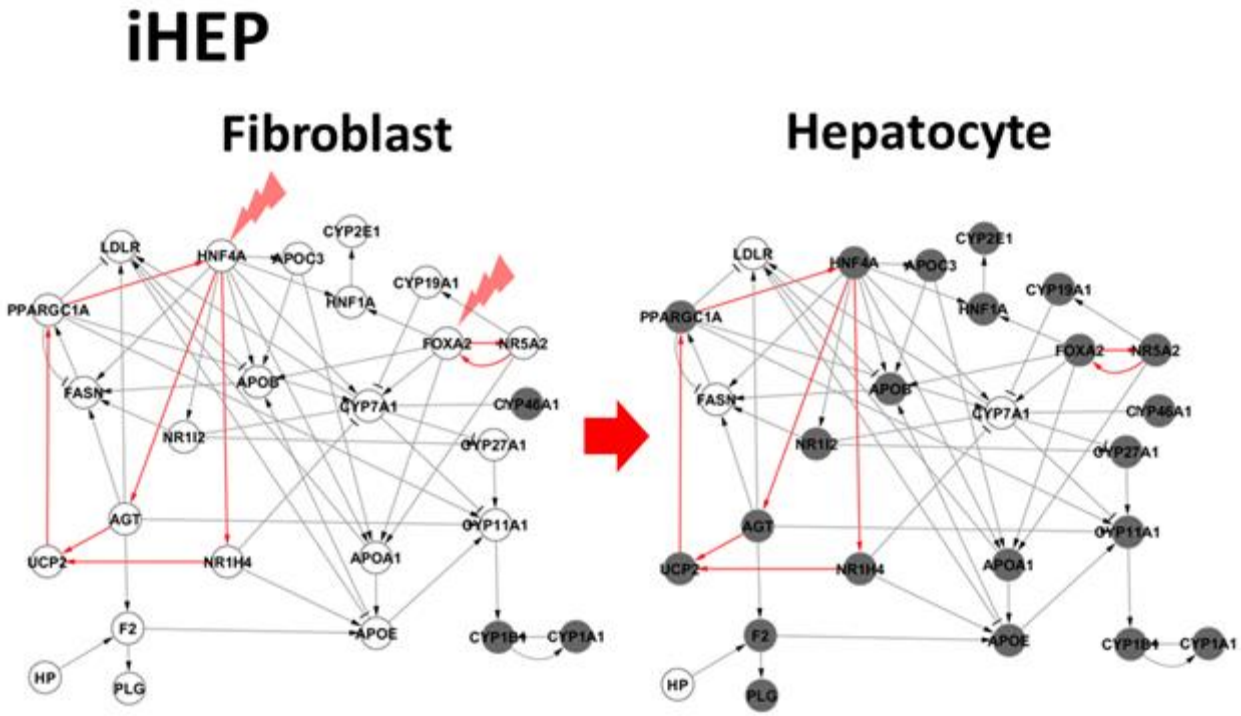


Figure 25 | Simulated perturbations performed assuming a Boolean model succeeded in triggering the transition. These results are consistent with previously published experimental perturbations. Genes in "ON" and "OFF" are represented in grey and white respectively. Regular arrows represent activation and T-arrows represent inhibitions. Perturbed DEPCs are coloured in red.

iCM

In the postnatal heart during the regular development, a large pool of existing fibroblasts was directly reprogrammed to an alternative fate as cardiomyocytes. To date, no single master regulator of direct cardiac reprogramming has been identified, although the combined perturbation of three developmental transcription factors (GATA4, MEF2C and TBX5) has been proposed and validated experimentally as a rapid and efficient way in which to induce this transition [10]. Our method found that when GATA4 and MEF2C are perturbed separately or in combination (see Figure 5e) they are able to trigger the transition from fibroblast to induced cardiomyocyte (iCM), thus indicating the important role played by these genes in this cellular transition. This finding is partially consistent with the experiment performed in [10], where GATA4 and METF2C in combination with TBX5 were simultaneously perturbed to achieve this cellular transition. Thus, our results propose the hypothesis that either GATA4 or METF2C are

individually capable of triggering this transition. To our knowledge, this prediction has not been experimentally validated in fibroblast-cardiomyocyte transition, although it has been reported that GATA4 is capable of reprogramming mesenchymal stromal and P19 cells [141] into cardiomyocytes [142,143].

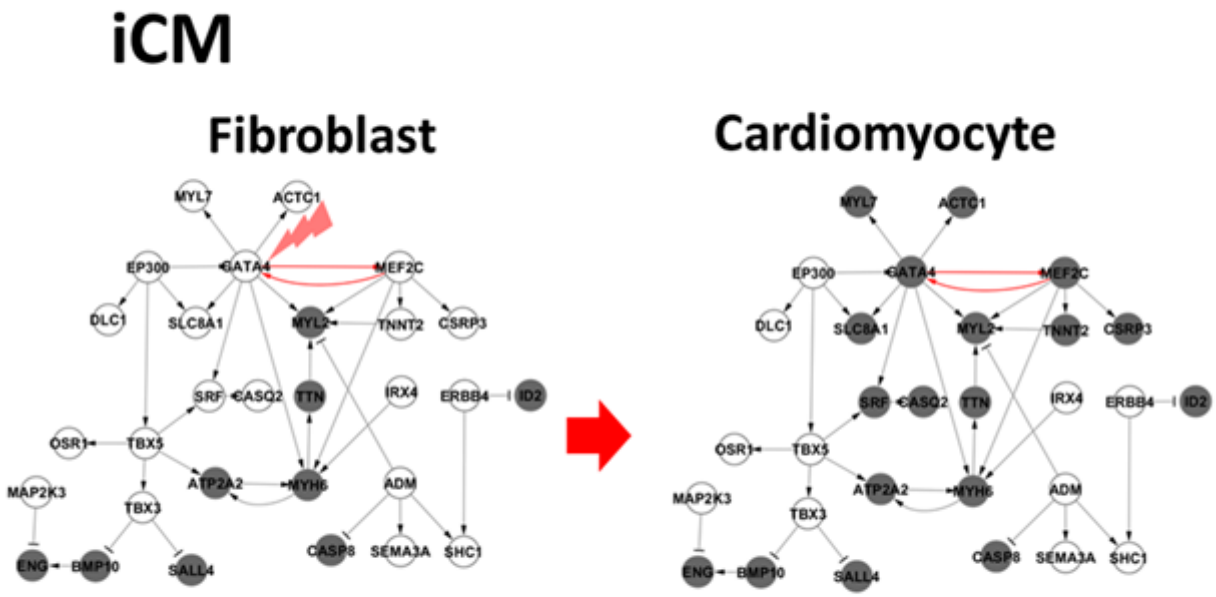


Figure 26 |Simulated perturbations performed assuming a Boolean model succeeded in triggering the transition. These results are consistent with previously published experimental perturbations. Genes in "ON" and "OFF" are represented by grey and white respectively. Regular arrows represent activation and T-arrows represent inhibitions. Perturbed DEPCs are coloured in red.

These examples nicely illustrate the experimental validation of our predicted driver genes, and therefore the utility of our method.

Example	Transitions	DEPFCs	Minimal combinations of reprogramming determinant genes
T-helper	Th2-Th1	4	GATA3, T-bet
EMT	Epithelial-Mesenchymal	12	SNAI1, ZEB2, MIR203
HL60	HL60-Neutrophil	1	IL1B, CASP1, IRF1
iHEP	Fibroblast-Hepatocyte	2	FOXA2:PPARGC1A, NR5A2:UCP2, HNF1A:PPARGC1A, HNF4A:NR5A2, NR5A2:PPARGC1A, FOXA2:HNF4A, HNF1A:UCP2, AGT:NR5A2, AGT:FOXA2, FOXA2:UCP2, AGT:HNF1A, HNF1A:HNF4A
iCM	Fibroblast-Cardiomyocyte	2	GATA4, MEF2C

Table 3| Minimal combinations of reprogramming determinant genes obtained after the application of our method in five different biological examples for specific transition between attractors corresponding to cellular phenotypes. Alternative combinations of reprogramming determinant genes are separated by commas. Combinations of reprogramming determinants genes perturbed in Figures 22,23, 24, 25 and 26 are in bold.

Networks	Genes	miRNAs	Interactions	Activations	Inhibitions	Positive circuits	Negative circuits
T-helper	36	4	71	47	24	108	108
EMT	4	3	17	2	15	12	0
HL60	18	1	30	28	2	2	0
iHEP	26	0	57	47	10	12	18
iCM	29	0	37	31	6	2	0

Table 4| Number of genes, miRNA interactions and circuits of the five biological examples are shown.

T-helper	EMT	HL60	iHEP	iCM
FOXP3 -> MIR-155	MIR200 -| ZEB1	MIR-146A -| CXCR4	None	None
IFN-G -> MIR-145	MIR200 -| ZEB2	MIR-146A -| IL8		
MIR-145-| STAT1	MIR203 -| SNAI1			
MIR-146A -| IRAK	MIR203 -| ZEB2			
MIR-155 -| IFN-GR	MIR34 -| SNAI1			
MIR-155 -| SOCS1	SNAI1 -| MIR200			
MIR-23A -| IL-6R	SNAI1 -| MIR203			
TGFB -> MIR-146A	SNAI1 -| MIR34			
TGFB -> MIR-155	ZEB1 -| MIR200			
TGFB -> MIR-23A	ZEB1 -| MIR203			
	ZEB2 -| MIR200			
	ZEB2 -| MIR203			

Table 5| Interactions with miRNAs included in the examples.

Experimental validation of a novel recipe to dedifferentiate astrocytes to neural stem cells.

During central nervous system development neural stem cells (NSCs) differentiate to neural progenitor cells (NPCs), which in turn differentiate to neurons and glia (astrocytes and oligodendrocytes). However, in early postnatal periods, parenchymal astrocytes are able to re-acquire the potential and characteristics of NPCs following injury. This raises the question of how this potential acquisition is regulated and whether mature astrocytes could represent a potential cell source to replace lost neurons.

We studied reprogramming processes of a system composed of CTX12 NPCs and two different astrocyte populations differentiated from CTX12 using two different protocols. We called them FBS-astrocytes and BMP4-astrocytes. Indeed, they are morphologically and functionally different, with FBS-astrocytes being less mature (phenotype closer to NPCs) and sensitive to FGF2 and EGF, which induces the dedifferentiation to NPCs. In contrast, BMP4-astrocytes are more mature and cannot be dedifferentiated by adding FGF2 and EGF to the culture media. We performed microarray experiments comparing these three cell types and reconstructed a GRN

connecting DEGs by means of links representing interactions between expression values. These interactions can be different in nature, including promotor binding interactions (with positive or negative effect), miRNA effects and direct regulations (including protein modifications). Assuming a Boolean model, we optimised the match between computed and experimental attractors using the network contextualisation algorithm briefly described in the methods of this section and fully detailed in Section 3.1 of this dissertation. The resulting contextualized network has three attractors corresponding with the three cellular phenotypes. We then proceeded to detect the reprogramming determinants, applying our strategy to this contextualized network for all possible transitions, namely, differentiation from CTX12 to FBS-astrocytes and to BMP4-astrocytes, the two corresponding dedifferentiation and the transdifferentiation from BMP4- to FBS- astrocytes and vice versa. Experimental testing was used for four combinations of predicted reprogramming determinants corresponding to four different cellular transitions, namely BMP4- and FBS- astrocytes' dedifferentiation and BMP4- and FBS- astrocytes' transdifferentiation in both senses (see Table 6).

Transitions	Minimal combinations of reprogramming determinants	Action required
CTX12 -> Astrocytes-BMP4	HMOX1:VEGFA VEGFA:TP53	Activation: Activation Activation: Inhibition
CTX12 -> Astrocytes-FBS	HMOX1 TP53	Activation Inhibition
Astrocytes-BMP4 -> CTX12	HMOX1:VEGFA VEGFA:TP53	Inhibition: Inhibition Inhibition: Activation
Astrocytes-FBS -> CTX12	HMOX1 TP53	Inhibition Activation
Astrocytes-BMP4 -> Astrocytes-FBS	VGFA TIMP2	Inhibition Activation
Astrocytes-FBS -> Astrocytes-BMP4	VGFA TIMP2	Activation Inhibition

Table 6| Predicted reprogramming determinant with the corresponding required action

As a result of these experiments we found that the dual selective inhibition of HMOX1 and VEGFA using SnPP and SU5416 respectively together with a permissive media with FGF2 and EGF led to a robust dedifferentiation of 'mature' (non-proliferative) astrocytes. This was in parallel with underlying transcriptional and epigenetic changes appropriate for this transition (see Figure 27). This conclusion rests on the expression of specific pluripotency/NPC and astrocytes markers (see table 6) as well as on the evaluation of the proliferation rate (2-tailed student's t-test, equal variance, $p<0.05$ $n=8$) of cells grown in permissive media with and without a reprogramming determinants inhibitor. This permissive media alone (without the inhibition of HMOX1 and VEGFA) showed in controls to be incapable of causing this differentiation by itself (unlike in the case of FBS-astrocytes). Finally, in order to asses epigenetic changes that reprogramming could confer to the resulting population of cells, a

ChIP-qPCR experiment was performed in order to analyse histone marks associated with activate chromatin (H3K4me3, H3K9ac) and repressive chromatin (H3K27me3) at promotor regions of two genes, CCNB1 and SOX2. Despite thorough analysis of the data, it corresponds only to a single biological replicate and more replication is required. Indeed, the results suggest a more open chromatin configuration when cells are grown in the presence of HMOX1 and VEGFA inhibitors (see Figure 27).

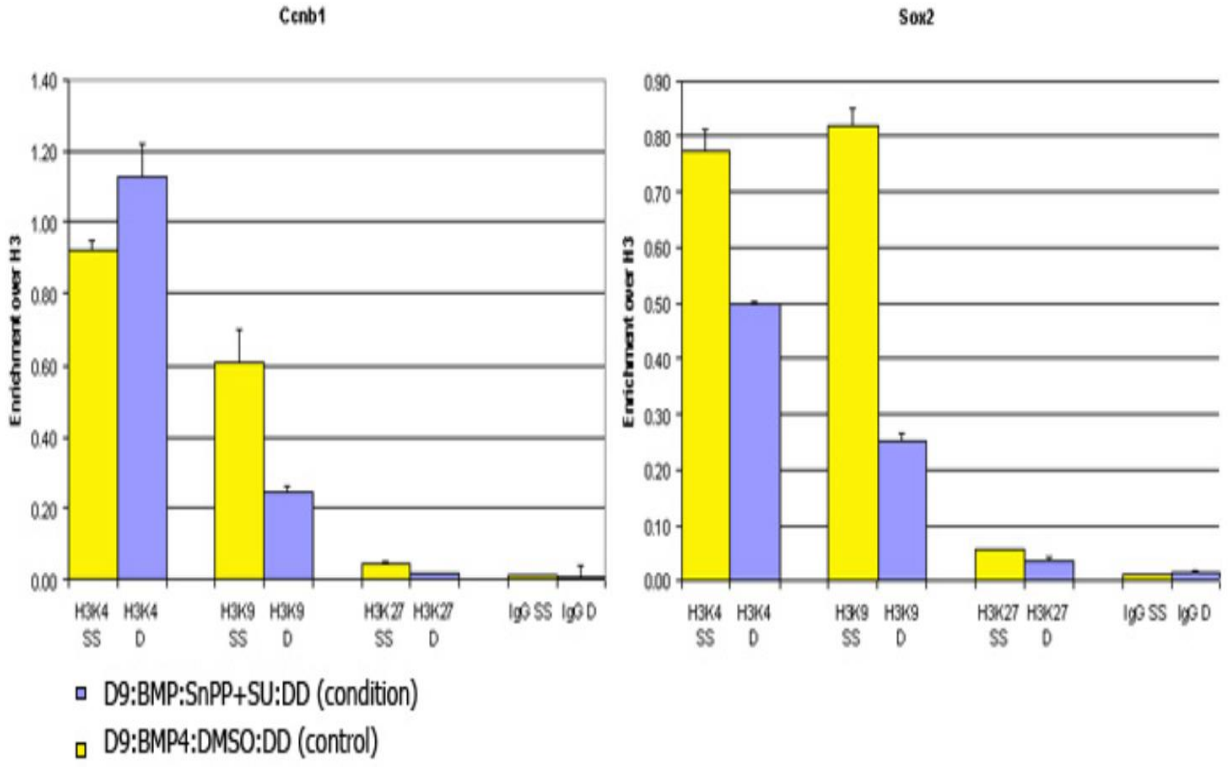


Figure 27| Dedifferentiation of BMP4-astrocytes is accompanied by epigenetic changes. Graphs show enrichment of activating (H3K4me3, H3K9ac) and repressive (H3K27me3) histone marks relative to H3 with driver (Day 9:BMP4-astrocytes:Sn+SU:DD) conditions at D9. Abbreviations: SS, SnPP+SU5416;D, DMSO control. Data are the average from technical triplicates in a single biological replicate (n=1). Error bars show standard deviations.

The other three experimental tests failed or showed large variability, possibly due to both the incompleteness of the GRN and/or the assumption of incorrect regulatory mechanisms for specific genes in the network. The model could be improved using, for instance, epigenetic information or time series to elucidate details about regulatory mechanisms or to infer missing interactions.

Condition	Ki67	Uhrf1	Thrsp	Aqp4	Gfap
Day 0: CTX12	1,00	1,00	1,00	1,00	1,00
Day 3: BMP4-astrocytes	0,01	0,05	17,88	1501,98	2471,33
Day 6: BMP4-astrocytes:DD*	0,20	0,34	9,11	1314,05	1227,60

Day 6: BMP4-astrocytes:DMSO	0,00	0,07	11,22	1020,57	605,93
Day 6:BMP4_astrocytes:Sn+SU	0,00	0,06	11,24	667,39	396,68
Day 9:BMP4-astrocytes:DMSO:DD	0,43	0,42	6,95	1504,72	992,59
Day 9:BMP4-astrocytes:Sn+SU:DD	0,77	0,61	5,30	381,03	175,37

Table 7| Changes in astrocytes and NPC-specific gene expression in BMP4 to NPC transitions. Samples were analysed for astrocyte specific markers (Thrsp, Aqp4 and Gfap) and NPC/proliferation specific markers (Uhrf and mKi67). Expression of Uhrf1 and Ki67 was significantly higher on Day 9: BMP4-astrocytes: SnPP+SU:DD compared to Day 3:BMP4 astrocytes ($p<0.05$ n=3, equal variance student's t-text).

Thus, our results proposed the hypothesis that the double inhibition of HMOX1 and VEGFA should induce a change in the transcriptional program when perturbed, thus leading to a CTX12-like cellular phenotype. Indeed, the experimental validation supports this prediction, indicating a cellular transition toward the desired phenotype (NPC CTX12). Figure 28 shows a schematic representation of the dedifferentiation experiment from a population of astrocytes previously differentiated using BMP4. The addition of HMOX1 and VEGFA inhibitors to a permissive media with EGF and FGF2 induces a cellular transition to NPC-like cells.

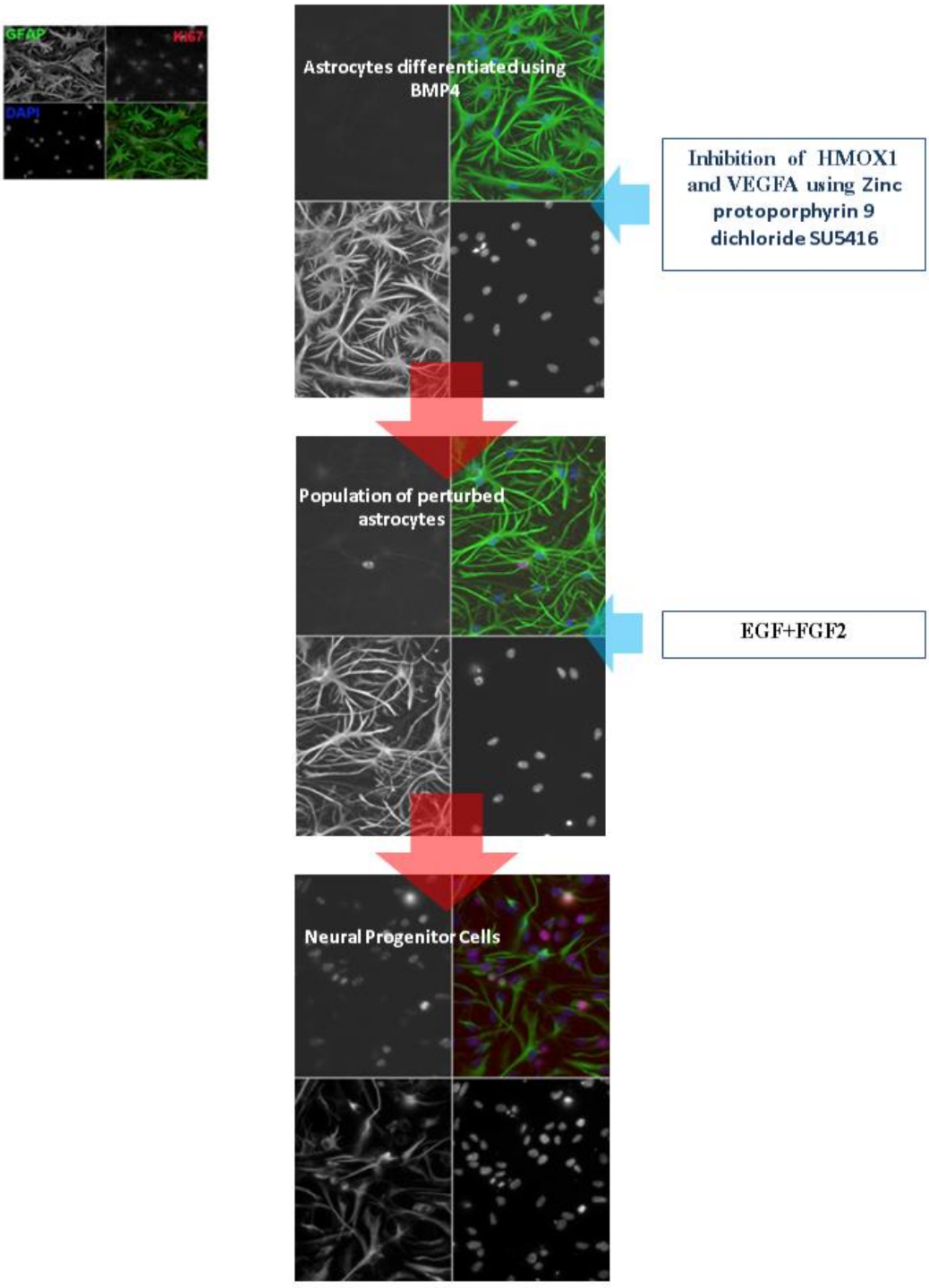


Figure 28| Schematic representation of the experiment with the successful dedifferentiation from astrocytes to neural progenitor cells. GFAP, Ki67 and DAPI are markers for astrocytes, proliferation and nucleus respectively. The remaining cell (bottom-right) merges all markers together.

3.2.3 Discussion

The method provides a strategy through which to induce transitions between cellular phenotypes. This makes it possible to explore the stability landscape, eventually with alternative combinations of perturbed genes with the subsequent differences in trajectories. This strategy directly addresses three major problems in cell reprogramming: a) Safety in the reprogramming process, avoiding undesired turnings which often lead to cancer; among the alternative solutions, some combinations of RD genes inducing risky transitions too close to a tumorigenic profile can be avoided and safe transitions selected [144]; b) Efficiency. The reduced set of alternative experimentally testable solutions makes it possible to find more efficient strategies through which to induce cellular transitions; c) The potentially incomplete reprogramming or the appearance of aberrant phenotypes (for instance, no effective equivalence between iPSC and ESC). Such alternative phenotypes could be detected as additional attractors in the stability landscape and be taken into account so as to obtain the desired transitions.

DEPC detection relies on attractor computation assuming a Boolean model, which is relatively simple and does not require parameter selection for a given topology. However, although this model does not take into account detailed cellular information, such as the strength of regulatory interactions and continuous gene expression values, it does preserve the regulatory logic which rules the flow of information in gene regulatory networks. This in turn makes it possible to roughly describe stable cellular phenotypes and to detect combinations of genes triggering transitions between them. In cellular reprogramming, genes of interests are predominantly transcription differentially regulated transcription factors (TFs) (i.e., either they are up- or down- regulated). Moreover, while maintaining the differentiated stable phenotypes, the expression levels of these TFs are kept stable. Even though multiple genes show dynamic changes during these transitions, most of them are involved in metabolic regulations, cell cycles, circadian rhythms, and various cellular responses to environmental stimuli. As a result, the detailed expression levels of such genes will not affect the transitions and hence, do not need to be considered. In addition, including such detailed information prohibitively increases computational requirements. It can also dangerously generate large numbers of false-positives which are only coincidental to cellular states due to experimental timing and conditions. Given that we are not interested in a detailed description of the regulatory mechanism we consider a Boolean model suitable for our purposes, but not for the elucidation of transient states. As a limitation of our method we should mention that transitions involving cyclic stable states are not yet considered but are subject to possible extension of the method presented here. Modelling transitions between cyclic attractors could be applied to identify driver genes in biological systems with oscillatory behaviour.

Regenerative medicine, where the goal is to replace or regenerate damaged or lost human cells, is a rapidly growing research area [145]. However, current therapies which focus on tissue regeneration are significantly impeded by our limited understanding of how to reprogram cells towards specific cellular populations. Hence, cellular reprogramming, including the conversion of one differentiated cell type to another (trans-differentiation) or to a more immature cell (dedifferentiation), has a high relevance for regenerative medicine and disease modelling [146].

Indeed, there is an increasing amount of experimental evidence to show that only a few key genes, known as reprogramming determinants (RDs), are required for the orchestration of the complex regulatory events which occur during reprogramming. Although substantial progress has been made in developing experimental reprogramming techniques, to date there is no protocol which is capable of systematically predicting combinations of RDs which can trigger transitions or of tackling the problem of low reprogramming efficiency and fidelity.

Here we provide a framework which acts as a guide when it comes to designing protocols to induce transitions between cellular phenotypes, providing effective cellular reprogramming protocols (including protocols for differentiation, dedifferentiation, trans-differentiation and pluripotency recovery). This work thus represents a major potential advance in the way we uncover RDs and the pathways involved in cellular reprogramming, with enormous scope for regenerative medicine across diverse tissue- and cell-types. The general applicability of our method was demonstrated using five illustrative examples corresponding with different cellular systems. In addition, its capacity to predict novel reprogramming recipes was demonstrated by the experimental validation of predictions performed on astrocytes induced to dedifferentiate in an in vitro system.

3.2.3 Methods

Randomly generated networks

One thousand sub-networks of size 20-40 genes were randomly extracted from the E. coli K12 transcriptional network of RegulonDB (http://regulondb.ccg.unam.mx/) using GeneNetWeaver [147] using greedy neighbour selection and including self-regulations. These sub-networks, preserving the topological features of the original K12 transcriptional network, were used as synthetic networks for validation of the current hypothesis.

Network reconstruction for the five examples with effective recipes for transdifferentiation known from literature and the BMP4-astrocytes dedifferentiation example GRN

We selected five biological examples to illustrate the applicability and utility of the method. T-helper and EMT examples are based on previously published networks. For the remaining three examples, i.e., HL60, iHEP and iCM, we reconstructed our own GRN. The main topological properties of the five final networks are shown in Table 2. References for the interactions of each case are included in the supplementary information. The procedure for the network reconstruction consisted of the following steps:

1. Obtaining a list of differentially expressed networks. In order to reconstruct the HL60-neutrophil differentiation gene regulatory network we used a set of genes composed of genes differentially expressed between HL60 cells and neutrophils, differentiation induced by dimethyl sulfoxide (DMSO) in the experiment performed by F. Mollinedo *et al.* [122]. The fibroblast-hepatocyte and fibroblast-cardiomyocyte networks were constructed using genes found differentially expressed in the experiments performed by Huang *et al.* [9] and Ieda *et al.* [10] respectively. These sets of differentially expressed genes were obtained after the performance of a T-test and selection of genes with a p-value < 0.05.

2. Connecting these genes using expression regulatory interactions from literature. 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 [67,68]. MedScan processes sentences from PubMed abstracts and produces a set of regularized logical structures which represent the meaning of each sentence. The ResNet mammalian database stores information harvested from the entire PubMed, including over 715,000 relations for 106,139 proteins, 1220 small molecules, 2175 cellular processes and 3930 diseases. The focus of this database is solely on humans, mice and rats. We selected only the interactions included in the ResNet mammalian database in the categories of Expression, PromotorBinding and Regulation. In the Expression category interactions indicate 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 indicate that the regulator binds the promotor of the target. Finally, in the Regulation category interactions indicate that the regulator changes the activity of the target. Following this step we obtained 3 raw networks which were reduced, removing irrelevant nodes for the stability analysis, i.e., nodes without incoming edges.
3. Network enrichment with experimentally validated miRNA interaction. GRNs were enriched when possible using miRNA interactions experimentally validated and publicly available in two different databases: TransmiR[148] and miRTarBase[149]. Included was information about miRNA regulatory genes and miRNA regulated genes respectively. The only miRNA included were those forming closed loops with network genes, which means that they are able to affect the stability of the network (see Table 1). This was with the exception of those forming a negative circuit with gene target and regulator of this miRNA. The reasoning behind this is that the dynamics of such a regulatory motif are not well described in a Boolean representation and can introduce a confusing factor to the model; in a Boolean system these motifs generate oscillatory behaviour, although it is known that in reality these dynamics strongly depend on kinetic parameters and the consequent time response with very different effects [150,151,152]. We decided not to introduce noise into the model, with the assumption that some regulatory effects could be missing (for example, an increased time response of specific genes under perturbation with the consequent delay in reaching an attractor). On the other hand, a Boolean representation is quite robust when it comes to describing stable steady states or fixed points (termed attractors in this paper) and is suitable for our purposes. We introduce only miRNA interactions capable of changing gene states in the existent attractors or generating new ones. Figure 29 presents examples of miRNAs finally not included in the HL60 model.
4. Contextualizing the network by pruning based on an optimisation algorithm.
This algorithm is fully described in Section 3.1, and as such is only briefly explained in the following paragraph so as to refresh the general concepts. In order to contextualize the network to the biological conditions under which the expression data was obtained we applied an algorithm [130] which exploits the consistency between predicted and known stable states from experimental data to guide an iterative network pruning. The

algorithm was conceived to predict missing expression values in gene regulatory networks, but could be applied to contextualize the network when all the expression values in two attractors are known. The method assumes a Boolean model to compute attractors of networks and an evolutionary algorithm is used to iteratively prune the networks. The evolutionary algorithm samples the probability distribution of positive circuits and individual interactions within the subpopulation of the best-pruned networks at each iteration. The resulting contextualize network is based not only on previous knowledge about local connectivity, but also on a global network property (stability). Given that this contextualisation is based on the stability of the network, no assessment can be performed on interactions not involved in this stability. As a result of this, and as an initial step before the contextualisation, genes without incoming edges were iteratively removed until only strongly connected components and genes regulated by them remained in the network.

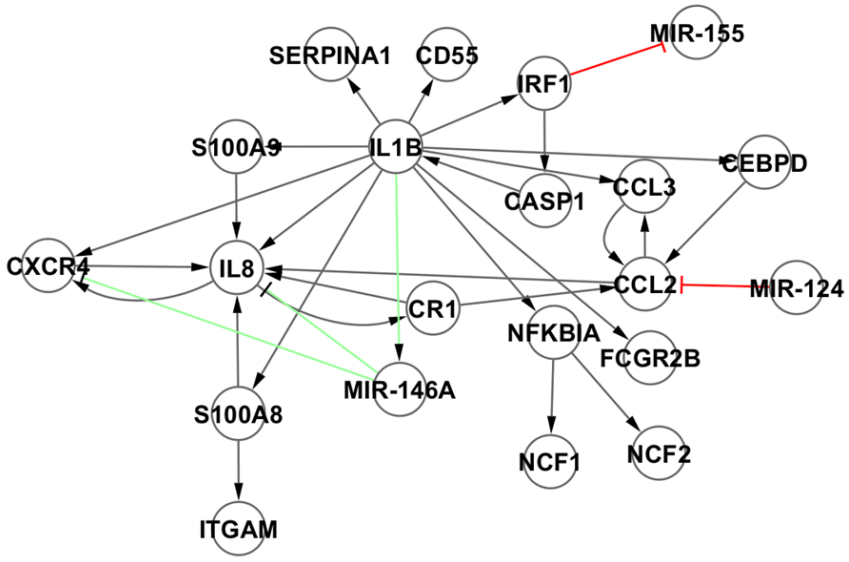


Figure 29 | HL60 gene regulatory network. miRNA interactions included and not included are represented in green and red respectively. MIR-146A has incoming and outgoing connections with DEPCs. Both MIR-155 and MIR-124 were removed due to their lack of outgoing and incoming interactions with DEPCS respectively.

The T-helper and EMT examples are based on previously published gene regulatory networks [72,153]. With the latter we expanded the original network with the addition of a recently published novel double-negative feed-back with miRNA34A [129].

Attractor computation

Attractor computation was performed assuming a discrete dynamical model or, more specifically, a Boolean model, with the application of a synchronous updating scheme [118] which updates all gene states simultaneously at each step until the system reaches an attractor. For this purpose we used our own implementation [130] written in C++ of the algorithm described by Garg *et al.*, 2007 [117]. 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. Regulatory logic of Boolean models for the five biological examples is included in the supplementary information.

Circuit detection

We implemented the Johnsons algorithm [120] in order to detect all elementary circuits in the network. A circuit is a path in which the first and the last nodes are identical. A path is elementary if no node appears twice. A circuit is elementary if no node but the first and the last appears twice. Once we had all of the elementary circuits, we selected positive feedback circuits, or circuits for which the difference between the number of activating edges and the number of inhibiting edges was even. Both elementary circuit detection and positive circuits sorting scripts were implemented in Perl.

Description of the algorithm to find minimal combinations of RDs

This algorithm can be described in three steps:

1. Detection of the gene represented the most within DEPCs. This gene was added to the growing minimal combination of RDs.
2. Marking DEPCs including this gene as targeted.
3. Checking if there are untargeted DEPCs remaining. If this is the case, the algorithm goes back to the step 1. If there is no untargeted DEPC left, the algorithm finishes at this point, and the current list of genes constitutes a minimal combination of RDs.

It is worth mentioning that eventually there are genes drawing in a number of targeted circuits. If this is the case the algorithm split the computation in different branches which provide different alternative minimal combinations or RDs.

Experimental part

Astrocytes differentiation

CTX12s were plated at $0.5x10^5/Cm^2$ in a normal growth medium (NGM), which is a modification of Sato's medium (see Table 8). The following day (D0) they are washed two times in DMEM:F12 treated for three days using BMP4- (Sato's + 20 ng/ml BMP4, Preprotech) or FBS- (10 % FBS in Sato's) differentiation medium for BMP4- and FBS-astrocytes respectively. Media were changed after 1 and 2 days. At D3 astrocytes populations were washed again two times in DMEM:F12 and were considered ready for perturbation experiments or microarray experiments. For transdifferentiation experiments the population of cells was exposed to

reprogramming determinants for 3 days (to D6). For dedifferentiation experiments, after D6 cells were washed again two times with DMEM:F12 and maintained in dedifferentiation conditions (NGM) for three more days (to Day 9).

Component	Final Concentration	Company
DMEM:F12		Invitrogen
Apotransferrin	5µg/ml	SCIPAC
Sodium selenite	130µl	Sigma
Progesterone	60ng/ml	Sigma
Putrescine	16µl/ml	Sigma
Insulin	5µg/ml	Sigma
BSA	100µl/ml	Sigma
Pen/Strep	1x	Sigma
L-Glutamine	2mM	Sigma
Glucose	5.6mg/ml	Sigma
T3	300ng/ml	Sigma
T4	400ng/ml	Sigma
Satos Medium+4-OHT+EGF+FGF		
EGF	20ng/ml	Peprotech
FGF2	10ng/ml	Peprotech
4-OHT	100nM	Sigma

Table 8| Sato's medium with growth factors and 4-OHT.

Microarray experiments

Microarray analyses were performed on CTX12, BMP4- and FBS- astrocytes using Illumina Single Color array for mice with the reference 8V20R011278551A. Expression data was analysed comparing CTX12 vs. BMP4- and FBS- astrocytes and BMP4- vs. FBS- astrocytes. The selected test for this analysis was an unpaired t-test whilst the selected cut-off was 0.01 and 1 for p-value and Fold change respectively (multiple testing correction used: Benjamini-Hochberg).

Immunocytochemistry

Coverslips from specific experimental time-points and conditions (D0, D3, D6, D9) were fixed and processed for immunocytochemistry. Coverslips were washed with 1x PBS before fixation in 4% paraformaldehyde for 10 min before being washed three times in 1xPBS. Cells were permeabilised with 0.1% TX-100 (TritonX-100, Sigma) in PBS for 5 min at room temperature (RT). Coverslips were then incubated in 30µl of primary antibodies in 10% normal goat serum (NGS) in PBS for 1 h at RT: anti---Ki---67 (rabbit IgG, 1:1000, Abcam) and anti-GFAP (mouse IgG1, 1:500, Millipore). They were then washed and incubated in secondary antibodies in 10% NGS in

PBS for 30mins at RT as follows: goat anti-rabbit IgG AlexaFluor 594 and anti-mouse IgG1 AlexaFluor 12 488 (both 1:1000, Molecular Probes). After washing, nuclei were DAPI-counterstained and mounted on Prolong Gold Molecular Probes).

Fluorescence cell counting and imaging, Ki-67 ratio statistics

Coverslips were counted using a Zeiss AxioImager Z1 fluorescence microscope, 63x objective, AxioCam MRM3 camera and AxioVision v.4.8.1.0 software. Five or more separate random fields were chosen per coverslip, with 300-500 cells counted per coverslip. Specific immunopositive cells were counted as a percentage of total number of DAPI-positive cells. Some markers (including GFAP) were also used to assess morphology changes. A student's t-test was applied when comparing Ki-67 positivity for each condition, assuming equal variance confirmed by F-test for all conditions.

Total RNA preparation and cDNA synthesis

Total RNA was extracted using RNAsy kit (Quiagen) according to the manufacturer's instructions and including an 'on-column' DNase step. RNA was eluted in 30 µl of RNase-free water whilst quality and concentration were analysed using a Nanodrop ND1000 spectrophotometer. The synthesis of cDNA was performed as follows: 1-2 µg of RNA was mixed with 1 µl of random primers and 1 µl Oligo dT primers (both Omega) in 17, 12 µl and incubated for 5 min at 70 °C. Samples were placed onto ice, and a 7.88 µl of the mix was added to each sample containing: 5µl 5x RT Buffer, 1.25 µl dNTP mix (10mM each) 1 µl M-MLV (200U) (Promega M3682), 0.63 µl RNasin (25U) (Promega). The 25 µl reactions were incubated for 1 hour at 37 °C, inactivated at 95 °C for 5 min and put on ice before being diluted to 100 µl with RNAse-free water and stored at -20°C.

Gene expression analysis by quantitative PCR

Real-time PCR was performed to obtain gene expression data for six genes, three astrocyte specific markers (THRSP, AQP4 and GFAP) [154,155], two NPC/proliferation specific markers (UHRF and MKI67)[154] and a house-keeping gene (GPADH). Gene analysis was performed using the Pfaffl method [156] with GAPDH as a reference gene. Relative expression levels comparison to CTX12 was performed using a Student's t-test, assuming equal variance. The solution for PCR consisted of 200-500nM forward and reverse primers (see Table 9) 10 µl (2x) iQ SYBR green mix (Biorad), 2 µl cDNA sample (water for controls) and HPLC-grade water to 20 µl. All conditions were run in triplicates in 96-well plates on Chrome4 real-time PCR thermocycler with MJ OpticonMonitor 3.1 software using the following program: 3 min at 94 °C, 45 cycles of 95 °C 30s, 60 °C 30s and 72°C 30s.Gene

Genes	Forward sequence	Reverse sequence
AQP4	GCTGTFATTCCAAACGAACTG	ATGATAACTGCGGGTCCAAA
GAPDH	TGCGACTTCAACAGCACTC	CTTGCTCAGTGTCCTTGCTG
GFAP	GAGAAAGGTTGAATCGCTGG	CGCTGTGAGGTCTGGCTT
KI67	TTCCTTCAGCAAGCCTGAG	GTATTAGGAGGCAAGTT
THRSP	TGACAGGGCAGGTTCTGTAG	CTCGGGGTCTTCATCAGTCT
UHRF1	ATGTGTGTGGTGGGCGCGAG	GAGTCAGTGCGGCAGCTGGG

Table 9| Primers list.

Chromatin Immunoprecipitation

Following cells' digestion and chromatin extraction, 10 µg was immunoprecipitated (ChIP) in lo-bind tubes with protease inhibitors (PI, Roche) and 1 µg antibody (H3, H3K4me3, H3K27me3, H3K9ac and rabbit IgG, all Abcam) in modified RIPA buffer in 500 µl volume for 1 hour at 4°C rotating before addition of 25 µl Protein G Dynabeads (Invitrogen) overnight. Beads were washed as follows (rotating every 2 min): 2x 800 µl wash buffer 1 (50mM NaCl, 2mM EDTA, 1% TX-100, 0.1% SDS), 1x 800 µ l wash buffer 2 (10mM Tris pH 8.1, 150mM NaCl, 1 mM EDTA, 1% NP40, 1% Na deoxycholate, 250 mM LiCl) and 2x 800 µ l TE. Chromatin was eluted and de-crosslinked in 100 µl Elution Buffer with 1µl RNase A and 4µl 5M NaCl (10% input samples were treated similarly) for 4 hours at 65°C and treated with proteinase K. Beads were removed and ChIP DNA was purified using a QIAquick PCR purification kit. Samples were run afterwards on a MyiQ Real-time PCR detection system (Biorad) with the same conditions described in the previous section. Each reaction had: 500 nM forward and reverse primers for SOX2 and CCNB1 (see Table 10), 10 µl 2x iQ SYBR green mix (Biorad), 2 µl DNA (water for controls) and water to complete the mix up to 20 µl. Conditions were run in triplicates with a standard curve of known genomic DNA concentration. Bio-rad iQ5 software was used for this analysis.

ChIP primers	Forward sequence	Reverse sequence
CCNB1	TACGACGGAGGTTTTATGG	GCAAGTTTCCACCCAAATCTT
SOX2	TCAGGAGTTGTAAGCAGA	CGGGCTCCAAACTTCTCTC

Table 10| ChIP primers.

3.3 Cell Identity and network stability: generalizing the transcription factor cross-antagonism concept and developing strategies for cellular reprogramming

This section refers to the work published on Stem Cells in 2013 entitled "*A general strategy for cellular reprogramming: the importance of transcription factor cross-repression".*

3.3.1 Introduction

The central role of transcription factor cross-antagonism in determining cell fate is one of the most important concepts to have emerged from years of lineage differentiation research [157,158,159,160]. In its simplest formulation, two regulators which negatively influence each other establish a bistable "toggle switch", readily explaining the two mutual exclusive cell fate outcomes. More complicated schemes also include transcription factor auto-regulation and antagonistic cross-regulation of target genes (see Figure 30).

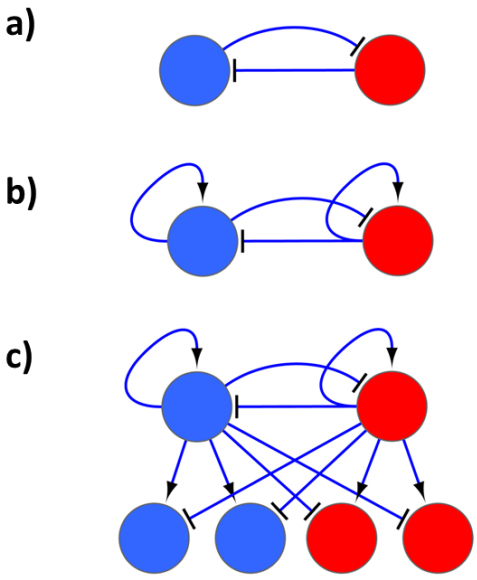


Figure 30| Cross-antagonistic motifs. A) Mutual inhibition B) Mutual inhibition and autoactivation C) Mutual inhibition and autoactivation, with cross-regulation downstream.

Several examples of these binary cell fate choice mechanisms have emerged in the last ten years [161,162,163,164,165,166,167,168,169,170]. 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[157] (see Figure 31). This tree defines distinct paths between different cell types in a Waddington's landscape[13,14,15], whereby different cell types can be interpreted as steady stable states of cellular gene regulatory networks termed attractors. Cross-antagonistic motifs not only determine binary decisions in the tree, but based on their bistable behaviour, characterized by mutually exclusive gene expression states, they also play a key role in the stability of binary cell fate choices or cell types derived from a common direct

ancestor. Furthermore, experimental evidence has demonstrated that perturbations of genes belonging to these motifs which swap their steady stable states are able to trigger transitions between these binary cell fate choices [131,171]. Indeed, although an attractor's stability is determined by a regulatory core comprised of one or several interconnected positive circuits [111], these cross-antagonistic motifs have been shown to be localized on the top of the hierarchical organisation of the set of positive circuits, whose steady stable 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 organisation can be extended to transitions between any given pairs of steady stable cellular phenotypes, even if they are not canonical cell fate choices. 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 (dedifferentiation) or to another differentiated cell type coming from a different primary progenitor (interlineage transdifferentiation). In these cases, a more complex set of positive circuits with or without 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, during the last decade several labs have experimentally demonstrated that despite cell type differences in the expression of thousands of genes, perturbation of few reprogramming determinants is usually able to trigger cellular transitions from one stable cellular phenotype to another [124,125,126] . Nevertheless, these experiments have relied on a brute force search of effective cocktails of transcription factors to achieve desired cellular transitions. As a result, and 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 indicates an urgent need to develop strategies for the systematic identification of optimal combinations of reprogramming determinants capable of inducing cellular transitions. Here we propose a cellular transition-dependent method which identifies candidates for reprogramming determinants by focussing on stability motifs in gene regulatory networks. 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 organisation of these circuits is analysed 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 the abundance of gene regulatory networks affecting cellular reprogramming efficiency and fidelity, we used a previously introduced network topological characteristic termed retroactivity [172]. This positively correlates with expression noise [173], and was used in order to detect combinations of genes in master regulatory DEPCs which were 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, according to our model, the best candidates for reprogramming determinants.

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 [132,133]. Our results showed that cells committed to becoming megakaryocytes or erythrocytes in the erythroid lineage can be reprogrammed to the myeloid lineage and become granulocytes or macrophages through the 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 which 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, which was published in 2011 [140].

In conclusion, here we propose, to our knowledge, the first method to systematically identify combinations of genes (reprogramming determinants), which are potentially capable of inducing transitions between specific pairs of cellular phenotypes, even without prior knowledge of possible candidates. Our method generalizes the principle of transcription factor cross-antagonisms in binary lineage decisions in the sense that it searches for master regulatory positive circuits (which can eventually be a cross-antagonistic motif). These circuits contribute to the stability of cellular gene regulatory networks, and have genes which are differentially expressed with respect to specific pairs of cellular phenotypes. Perturbations of combinations of genes belonging to these circuits which 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 for the establishment of cell disease models and regenerative medicine, our method constitutes a useful computational protocol which aims to assist researchers in the field of experimental strategy design.

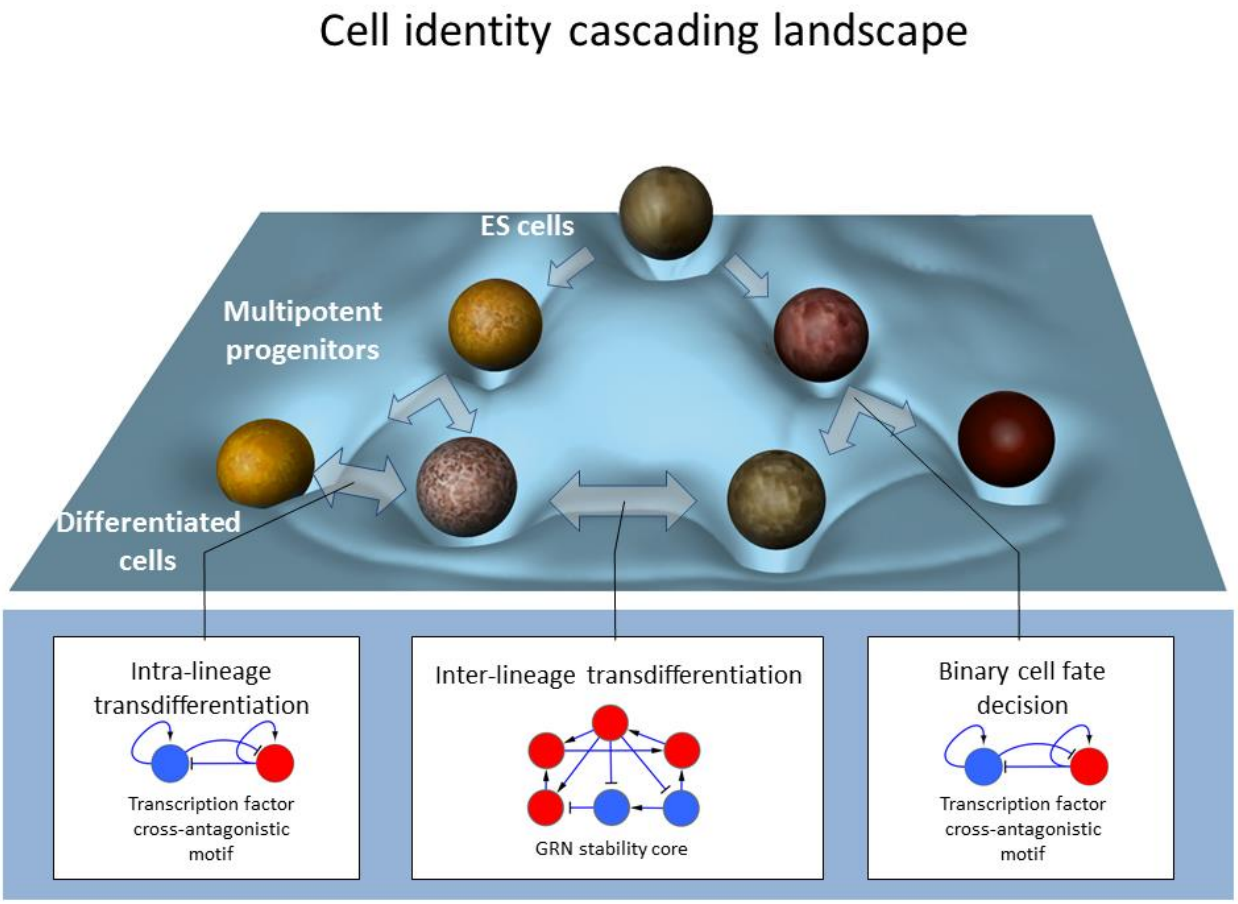


Figure 31| Cell identity cascading landscape representing the cellular transcriptional program. Paths between pluripotent and differentiated cells, representing cellular differentiation processes pass through stable expression profiles corresponding to multipotent progenitors. Binary cell fate decisions at multipotent progenitor level are characterized by cross-repression motifs of competing transcription factors. Transdifferentiation between somatic cells is divided between 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; in addition, those without a direct precursor (inter-lineage transdifferentiation), are 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.

3.3.2 Results

A popular framework through which to conceptualise and describe cellular transitions is that of the landscapes proposed by Waddington [13,14,15], 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 was necessary for the corresponding GRN to be 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) [111]. 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, similar 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 stabilisation of specific attractors. The hierarchical organisation analysis of strongly connected components (SCCs) formed by one or more DEPCs allows us to identify combinations of genes belonging to master regulatory DEPCs which should be perturbed in order to directly or indirectly target all DEPCs and to consequently induce specific cellular transitions. Finally, we selected among these combinations of genes those with the highest interface out-degree. This refers to the number of genes which are directly regulated by them. The reason for this step was to minimize the retroactivity effect on master regulatory circuits [172,173], which considers the increased time response of these circuits after noise or external perturbations. This allowed us to minimize the expression noise due to retroactivity contextualized to the specific cellular transition under study. In other words, we selected combinations of genes participating in more transcriptional regulation events in order to minimise DEPCs time response and the stochastic behaviour of GRN under perturbation. Indeed, this allowed us to minimize information loss during signal transmission. This strategy also allowed us to narrow down a huge combinatorial searching problem to a set of minimal combinations which constitute alternative reprogramming protocols and the output of our method.

The method can be described with the following seven steps:

1. Detecting all positive circuits in the GRN.
2. Computing network attractors.
3. Detecting transition specific DEPCs.
4. Reconstruction of transition specific GRN.
5. Transformation of GRNs in a directed acyclic graph (DAG) and hierarchical analysis.
6. Detection of DEPCs' master regulators.

7. Identification of reprogramming determinants.

In order to illustrate the different steps of the method we constructed a toy GRN including 16 nodes and 23 interactions (see Figure 32).

Detecting all positive circuits in the GRN

In order to detect master regulatory SCCs or clusters of DEPCs which should be independently perturbed, it was necessary to detect and list all positive circuits or positive regulatory feed-back loops. 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) were detected with different sizes ranging from 1 (self-loop) to 4.

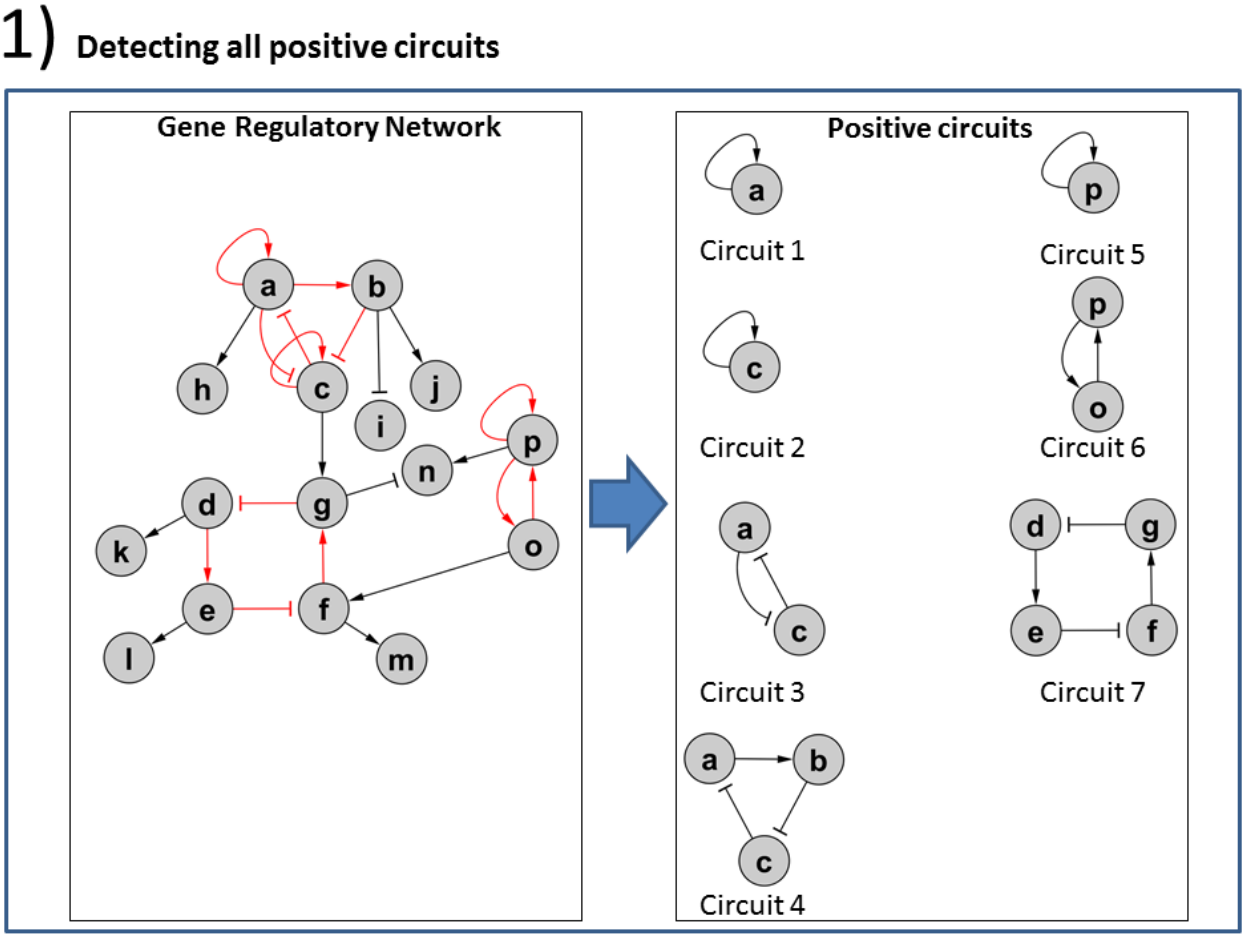


Figure 32| Positive circuit's detections. Seven positive circuits or positive feed-back loops (the sign of a circuit is defined by the product of the signs of its edges, being activation positive and inhibition negative) are present in this illustrative toy network. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

Computing network attractors

For the computation of network attractors we assumed a Boolean dynamical model with a synchronous updating scheme. This model allowed us to test three different attractors; all of which are represented in Figure 33.

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

Figure 33| 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.

Detecting transition specific DEPCs

Once we had information about attractors and circuits we proceeded to determine, among the entire set of positive circuits, which were DEPCs for specific cellular transitions, meaning that the expression levels of their genes change between involved cellular phenotypes. Figure 34 shows DEPCs between specific attractors.

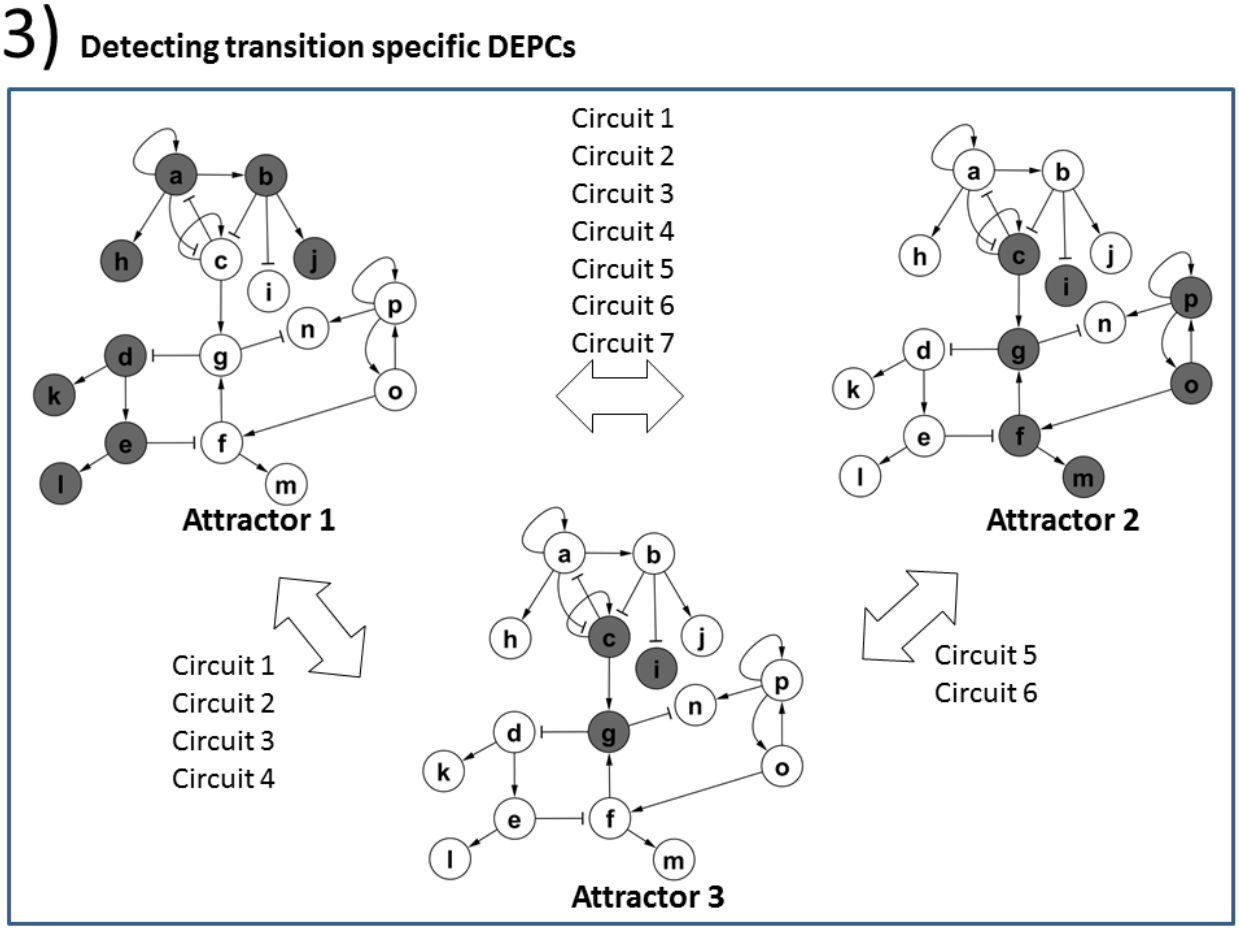


Figure 34| Transition specific DEPCs detection. Differentially expressed positive circuits (DEPCs) are those for which the expression levels of their genes change between two different attractors corresponding to two different cellular phenotypes. White and grey colours stand for down-regulation and 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 Attractors 1 and 2.

Reconstruction of transition specific GRN

Detection of master regulatory DEPCs requires the reconstruction of a transition specific subnetwork (Attractor 1 to Attractor 2 is represented in Figure 35) including only DEPCs for this specific transition and connections between them.

4) **Reconstruction of a transition specific GRN: attractor 1 to attractor 2**

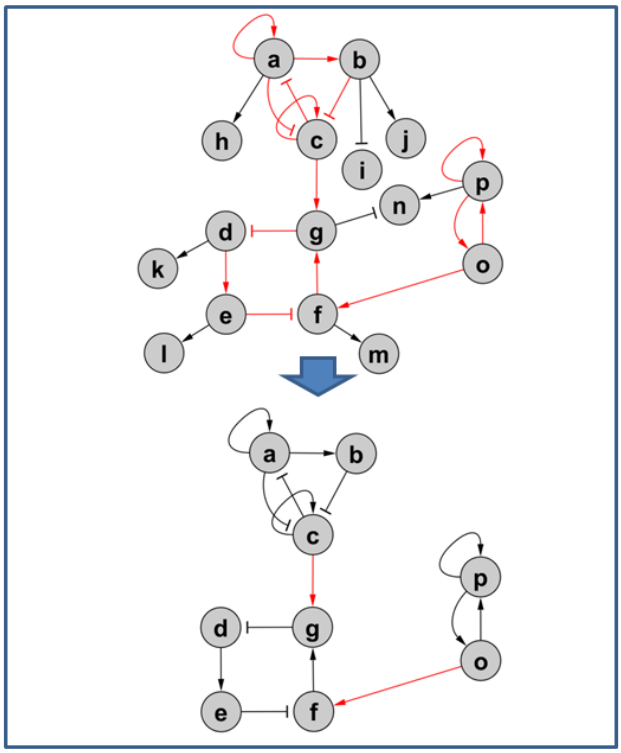


Figure 35| Reconstruction of attractor 1 to attractor 2 transition specific subnetwork. This network includes all differentially expressed positive circuits and the connections between them.

In step 5 those DEPCs of the previously obtained subnetwork which are forming SCCs are contracted in a single supernode. The hierarchical analysis of such a contracted subnetwork allows us to identify master regulatory SCCs (SCC 1 and SCC 2 in the figure). Within each master regulatory SCC, the DEPC with the 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.

GRN's transformation in a DAG and hierarchical analysis

DEPCs can be clustered, and form SCCs. These SCCs (if there is more than one) can be interconnected. In order to detect which SCCs should be independently perturbed to guarantee that all DEPCs are reached by the perturbation signal, we analysed the hierarchical analysis of the DAG resulting from collapsing SCCs in single super-nodes. It is worth stressing that this hierarchical organisation is cellular transition dependent since it is based on positive circuits which change between initial and final attractors (corresponding to cellular phenotypes).

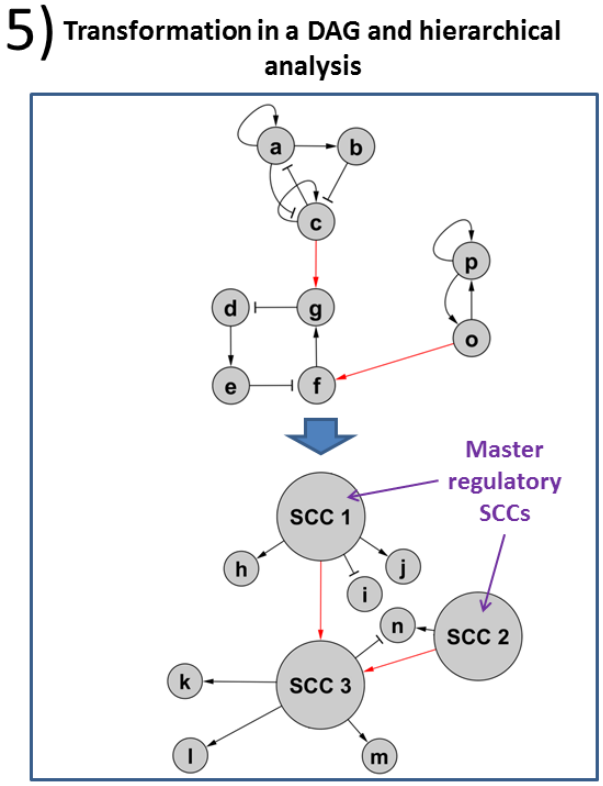


Figure 36 | Transformation of the transition specific network in a direct acyclic graph (DAG) and hierarchical analysis. In step 5 those DEPCs of the previously obtained sub-network which are forming SCCs are contracted in a single super-node. The hierarchical analysis of such contracted sub-network allows us to identify master regulatory SCCs labeled as SCC 1 and SCC 2 in the figure.

Determining the master regulatory DEPCs for each master regulatory SCC

DEPCs with higher degree interfaces were 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 to the number of such DEPCs. In other words, the perturbation of one gene per master regulatory DEPC 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. Figure 38 shows the retroactivity values for each gene within the two master regulatory SCCs.

6) Detection of DEPCs master regulators

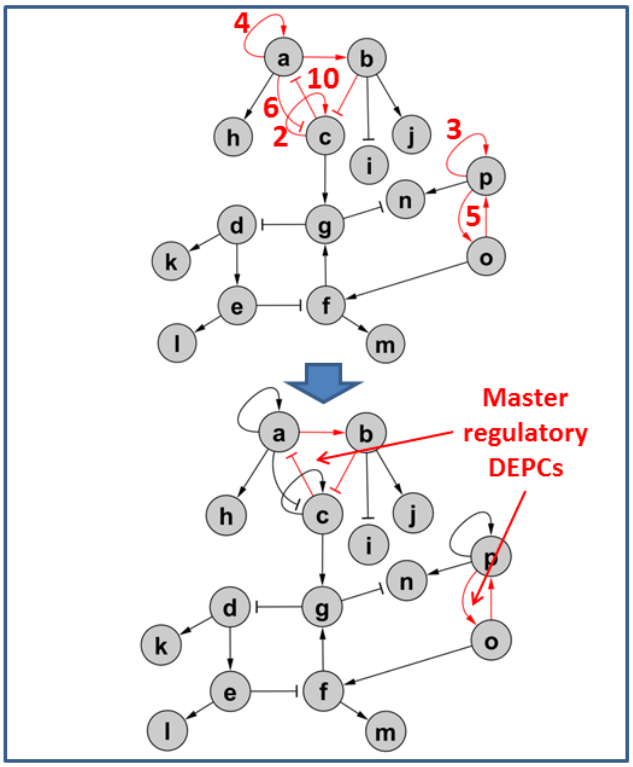


Figure 37| Detection of master regulatory differentially expressed positive circuits (DEPCs) based on the retroactivity criteria.

Detecting reprogramming determinant genes

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

7) Identification of reprogramming determinants

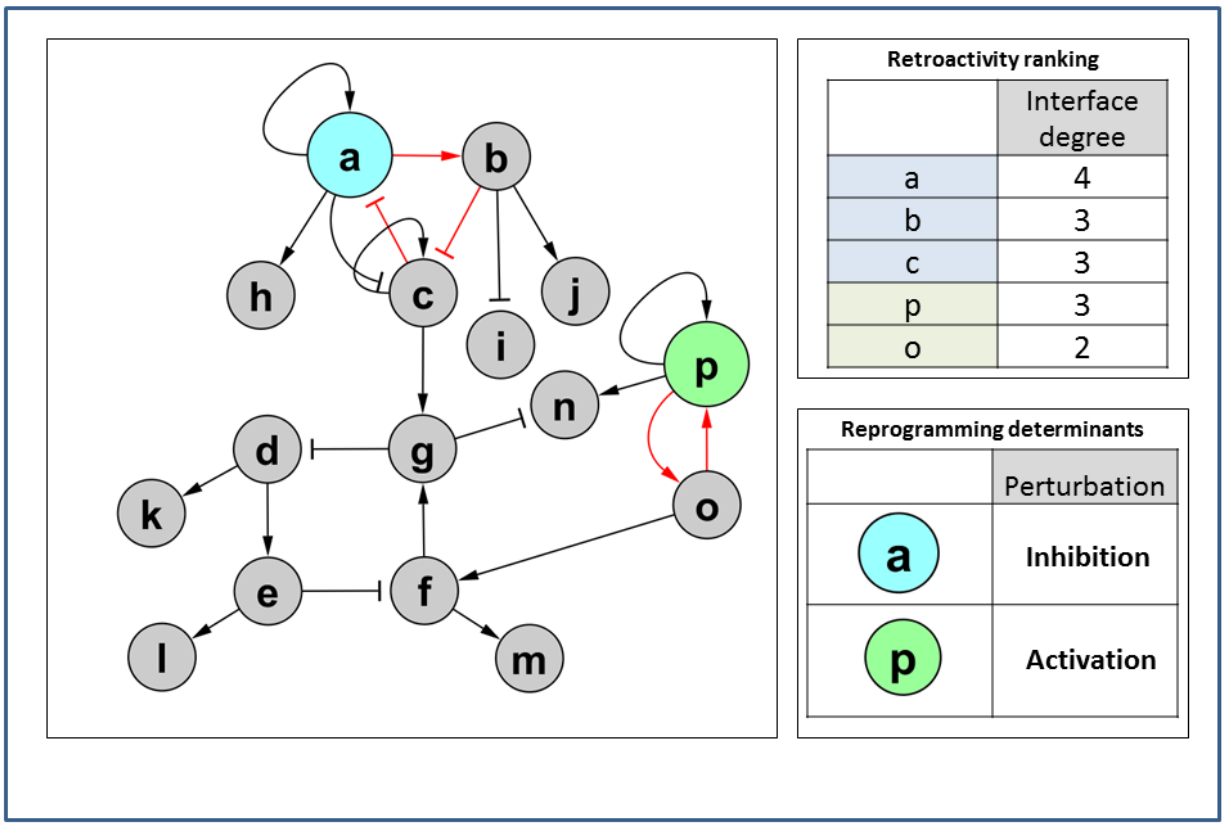


Figure 38 | Identification of reprogramming determinants. Identification of genes belonging to DEPCs master regulators with maximum gene interface out-degree. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

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 a generalisation of the 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.* [113], Krumsiek *et al.* and Dore *et al.* [174,175] 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 make it possible to induce 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 which 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 [9]. This network was contextualized by an iterative network pruning described in the methods section and previously published [102]. This contextualized network is specific to 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 possible, with information about miRNAs' interactions from experimentally validated and publicly available information [148,149].

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 are sub-divided into Th0 (precursor) and effector Th1 and Th2 cells depending on the array of cytokines which they secrete [176]. A T-helper differentiation network determines the fate of the T-Helper lineage [113], with three different attractors corresponding to the three different phenotypes (Th0, Th1 and Th2). We applied our method to a previously published GRN [113], 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 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 from 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 [131]. 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. Moreover, the detected master regulatory DEPC, knwon 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 stabilisation and how its proper perturbation can trigger transitions between them.

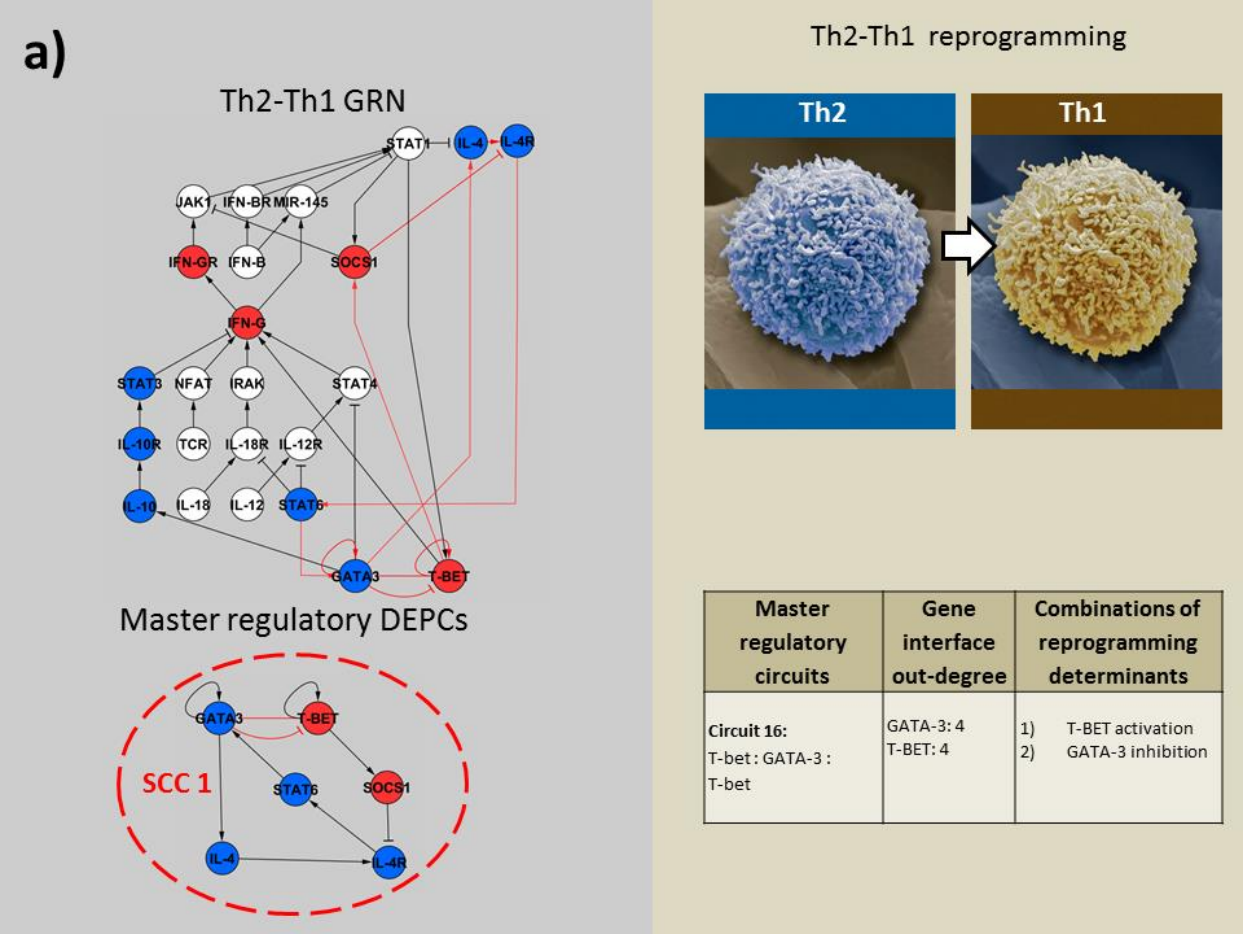
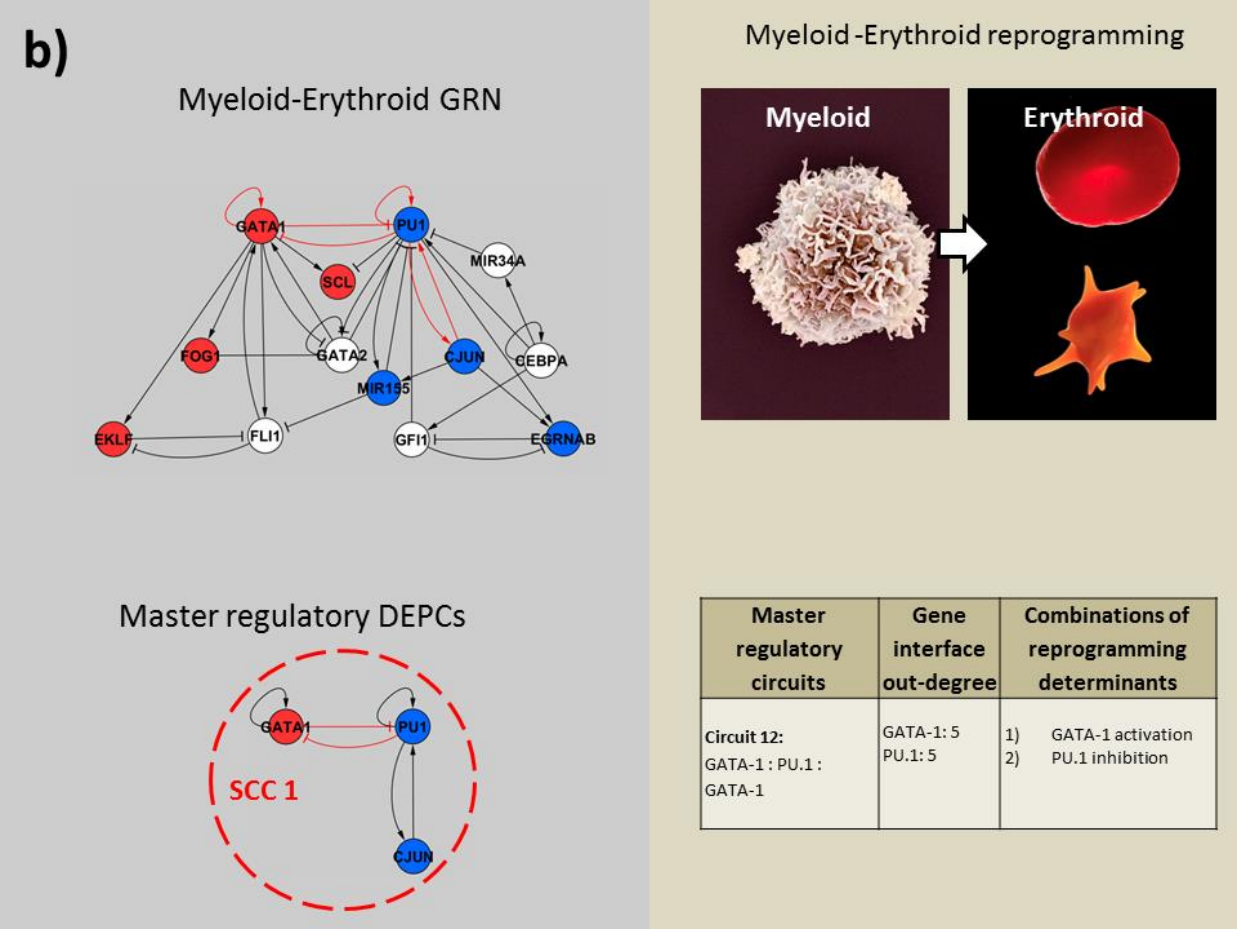


Figure 39 | Th2-Th1 reprogramming. Activation of T-bet and, alternatively, inhibition of GATA-3 are predicted as effective perturbations to induce this cellular transition. Blue and red colours in network nodes represent mutually excluding expression states for a given pair of cellular phenotypes, standing for down-regulation and up-regulation respectively. **'->'** represents activation or positive regulation and **'-|'** represents inhibition or negative regulation.

Myeloid-Erythroid

Within the hematopoiesis there are several binary decisions from multipotent stem cells to different types of blood cells. One of these decisions, namely that determining whether 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 is shown in Figure 3a, our method was applied to a previously published GRN [174,175], containing this motif embedded and connected with other multi-stable motifs. This allowed us to identify GATA-1 as a reprogramming gene capable of inducing the transition from myeloid to erythroid precursor cells. This finding is in full agreement with the experimental results obtained by Heyworth *et al.* [171], where the authors reported that myeloid precursors infected with an inducible form of GATA-1 generated erythroid colonies when GATA-1 was induced. Figure 3 b shows that in this example we found a single master regulatory circuit, known 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 an 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][171]. To our knowledge there is no experimental evidence to indicate that the inhibition of PU.1 is either able or 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 can also be exploited to re-specify the cellular commitment in cells sharing the same precursor.



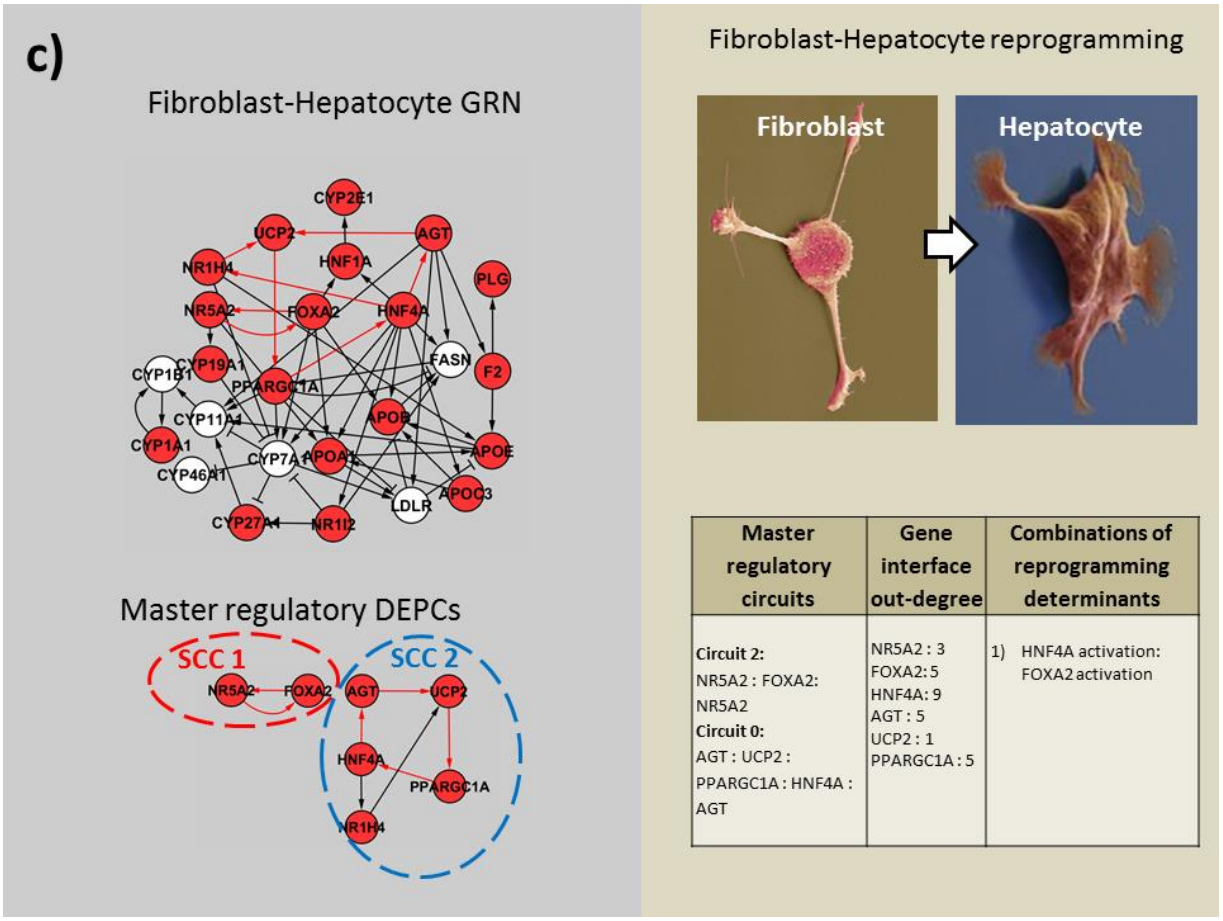
Master regulatory circuits	Gene interface out-degree	Combinations of reprogramming determinants
Circuit 12: GATA-1 : PU.1 : GATA-1	GATA-1: 5 PU.1: 5	1) GATA-1 activation 2) PU.1 inhibition

Figure 40 | 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. 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.

Fibroblast-Hepatocyte

Hepatocytes normally differentiate from hepatic progenitor cells to form the liver during 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 [9,140]. As 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. The first comprised circuit 2 (including NR5A2 and FOXA2) whilst the second comprised circuits 0, 7 and 4 (including genes AGT, PPARGC1A, UCP2 and HNF4A). Within the latter SCC, the DEPC, termed circuit 0, is the one with the highest interface out-degree of 20. Following this, we proceeded to identify reprogramming determinants by targeting both master regulatory circuits. Within circuit 2, the gene which made the most

significant contribution to the circuit retroactivity was FOXA2, with an interface out-degree of 5. Within the circuit 0, HNF4A was the one with the highest contribution to the circuit retroactivity with an interface out-degree of 9. Therefore, the final combination of reprogramming determinants was 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 [140]. 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).



Master regulatory circuits	Gene interface out-degree	Combinations of reprogramming determinants
Circuit 2: NR5A2 : FOXA2: NR5A2 **Circuit 0:** AGT : UCP2 : PPARGC1A : HNF4A : AGT	NR5A2 : 3 FOXA2: 5 HNF4A: 9 AGT : 5 UCP2 : 1 PPARGC1A : 5	1) HNF4A activation: FOXA2 activation

Figure 41| Cellular reprogramming from fibroblast to hepatocyte. In this particular case no single gene was able to induce the cellular transdifferentiation according to our predictions. On the other hand, combined activation of HNF4A and FOXA2 was 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.

3.3.3 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. It also has an enormous relevance for regenerative medicine and disease modelling. 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), as well as the elucidation of molecular changes and relevant pathways involved in these transitions (9). Furthermore, there is currently no methodology which is capable of systematically predicting reprogramming determinants which could guide the design of cellular reprogramming experiments. The development of computational models of transcriptional regulation which underlies cellular transitions would help to predict these reprogramming determinants. Moreover, the analysis of gene regulatory network properties has allowed for the identification of functionally relevant motifs of interactions which could play a role in cellular transitions. In particular, transcription factor cross-antagonism has been described as a mechanism which plays a key role in cell fate decisions. A bistable toggle switch constitutes a molecular cross-repression motif which determines cellular commitment and provides stability to gene regulatory networks underlying transcriptional programs of binary decision cell choices. Experimental evidence indicated 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 which are capable of inducing 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 transitions. We termed this subset differentially expressed positive circuits (DEPC). Furthermore, a hierarchical organisation of these circuits allowed us to detect master regulatory positive circuits, which directly or indirectly regulate the states of the other DEPCs. By focussing 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 were 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 larger number of regulations, and thus a limited concentration of the gene product must interact with several targets apart from that which closes the DEPC. In other words, the gene product has to distribute to different regulated targets, thus bringing about a higher probability that the DEPC signal feed-back is broken by chance (neglecting considerations about different molecular affinities which are assumed to be 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 which block the activation of certain genes can affect the expected network behaviour after specific perturbations. Furthermore, it has been experimentally shown that epigenetic modifications can prevent cellular reprogramming reversibility in certain cases [177]. In addition, our model did not take into account different delays in time response of distinct regulatory interactions. Nevertheless, given that the purpose of our method was the identification of reprogramming determinants, rather than a detailed description of network dynamics, we feel that our model provides reasonable predictions. More accurate predictions would necessitate the addressing of these considerations in the future.

Thus, our method constitutes the first strategy to systematically provide lists of combinations of reprogramming determinants for cellular reprogramming events involving two given cellular phenotypes without prior knowledge on potential candidates and pathways involved. As a result of this, the method is easily exportable to different biological systems, providing guidance even without the presence of 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, thus increasing the complexity of this type of reprogramming and therefore hindering 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 of experimental strategy design, especially when very little is known about a specific biological system.

3.3.3 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 hematopoiesi [113,174,175]. These two published networks were enriched with miRNA interactions experimentally validated and publicly available in two different databases: TransmiR [148] and miRTarBase [149], including information about miRNA regulatory genes and miRNA regulated genes respectively. Indeed, the only miRNA included were those forming closed loops with network genes as they are able to affect the stability of the network (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 was to induce the transition from one specific cell phenotype to another, the network was constructed based on changing elements between these two states, i.e., differentially expressed genes (DEG) between the two conditions or cell types obtained from microarray experiments. Genes of transcription factors belonging to cross-repression motifs were included among these DEGs (eventually, more than one motif). We proceeded to try to connect these DEGs 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 [67,68].

Once we had raw GRN from the literature, we proceeded to remove interactions which were 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 the literature usually come from different biological contexts to those of cell types, tissues or even species. This network pruning allowed us to reduce the number of “false” interactions and to obtain a contextualised network. The algorithm applied for this network pruning (see Chapter 3 Section 3.1 [102]) was originally conceived to predict missing expression values in gene regulatory networks. However, it could also be applied to contextualize the network when all the expression values in two given cellular phenotypes or stable transcriptional programs are known. 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. Although we tried to enrich this network with miRNA interactions, as we did in the two previous examples, none of the miRNA involved in regulatory loops or circuits with differentially expressed genes were experimentally validated for mice. More details about the network reconstruction process for the Fibroblast-Hepatocyte reprogramming example are included in the supplementary information.

Main properties of these three biological examples of GRN are shown in Table 11.

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

Table 11| miRNAs included in the biological examples. **'->'** represents activation and **'-|'** represents inhibition.

	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**

Table 12| Main properties of the gene regulatory networks of the three biological examples

Network transformation in a directed acyclic graph (DAG)

The first step of the method, termed "Detecting master regulatory SCCs" in the results section, requires the hierarchical analysis of a subnetwork of the complete GRN including only DEPCs as well as all genes and interactions connecting them. This subnetwork contains positive feed-back loops, so it should be transformed in order to be able to analyse its hierarchy. The transformation of this subnetwork of connected DEPCs in a DAG was performed by contraction of strongly connected DEPCs, i e, SCCs of differentially expressed genes, in single supernodes. This network transformation allows for the hierarchical analysis of the network following the method described by Jothi *et al.* [178], 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.

Circuit's detection

The Johnson's algorithm [120] 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 had all of the elementary feedback circuits, we selected positive feedback circuits, or feedback circuits for which the difference between the number of activating edges and the number of inhibiting edges was 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 [118] and used our own implementation [102] of the algorithm described by Garg *et al.*, 2007 [117]. 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 from the attractor computation were consistent with the results obtained using previously published software to compute attractors in Boolean systems (Boolnet [179], GenYsis [117]).

Chapter 4. Conclusions

In this dissertation, we applied various systems biology approaches to investigate cellular activation dynamics and reprogramming in the context of disease description and treatment. Focussed on the transcriptional level, we explored different strategies through which to integrate information from experiments and literature and to reconstruct the gene regulatory network underlying disease pathogenesis. We also sought to detect regulatory cores responsible for the stable expression profiles characterizing cellular disease phenotypes. In addition, we developed two computational methods which constitute a pipeline to design perturbation recipes to induce gene regulatory networks to transit from one stable state to another or, in other words, to reprogram a cell from one initial stable expression pattern to another. These two computational methods find utility in the context of disease treatment with potential applications in drug target discovery and cellular reprogramming.

More specifically, the second chapter focussed on the identification of the core of regulation of three different GRNs corresponding to three different diseases, namely metabolic syndrome (Section 3.1), prion disease (section 3.2) and EMT (Section 3.3) in the context of breast cancer. The main findings of these three case study systems are the strategies developed to identify cores of regulation within the respective GRNs, namely the search of network motifs (evolutionary conserved) with specific dynamical properties (multistability), exploiting local consistency of expression values to curate the network in an automated way and to detect afterwards stability cores (clusters of strongly connected components). Finally, there is the

simulation of perturbation experiments to identify missing interactions in incomplete models of transcriptional regulation.

Secondly, the application of these approaches to specific systems and analysis of the resulting gene regulatory cores allowed us to detect relevant genes, processes, and to identify novel interactions. In the case of the metabolic syndrome HIF1A, EGR1, STAT1 and CXCL12 were identified as key genes within the detected regulatory core; findings which are supported by the previously reported association of these genes with adipocyte's differentiation and activation. In the case of prion disease several genes involved in neuroinflammation were found to play a key role in the stabilisation of the disease state, thus stressing the key role of neuroinflammation in disease progression. Finally, the inconsistency between the simulated and the experimentally known behaviour of epithelial cells under perturbation of SNAI1 pointed out gaps in the current knowledge of regulatory mechanisms participating in the cellular transition to mesenchymal phenotypes (EMT). An miRNA-203/SNAI1 feedback loop was simulated in silico and experimentally validated. This finding has contributed to increased knowledge of the transcriptional regulation operating in EMT.

The third chapter focussed on the identification of cellular reprogramming determinants capable of inducing changes in transcriptional programs from one initial cellular phenotype to another (which could be from healthy to diseased cellular phenotypes or from one abundant cellular phenotype to another without self-renewal). For this purpose two novel computational methods were developed and are described in this chapter. The first (Section 3.1) refers to an algorithm to contextualize literature based GRNs using experimental expression data. The algorithm allows for the pruning of noisy networks, thus making them more suitable to describe input-output relationships within the biological conditions under which the expression data was obtained. This is something which is compulsory when it comes to modelling cellular transitions and predicting driver genes or reprogramming determinants. The second computational method refers to an algorithm designed precisely to detect these reprogramming determinants or combinations of genes able to induce cellular transitions when perturbed. The complete development of the method was split into two separated works (Sections 3.2 and 3.3) published in BMC Systems Biology and Stem Cells respectively [180]. Within the first one the concept of differentially expressed positive circuits (DEPCs) was proposed as the stability element to be targeted in order to induce specific cellular transitions. Within the second one we expanded the methodology, developing a strategy to minimize the number of genes to perturb and to maximize the chance of successfully inducing the transition by means of considering and indirectly measuring noise, termed as retroactivity. This strategy allowed us to select for perturbation those genes and positive circuits which were more predisposed to failure when transferring the signal to the regulatory mechanisms which keep its expression level in the desired state, i.e., its feed-back loops, due to stochastic events. These two computational methods sequentially applied to a given gene expression data-set constitute our knowledge of the first pipeline which allows for the design of recipes for cellular reprogramming without a previous list of candidates both for the network reconstruction and predictions. This pipeline can be applied to guide experimental design and as a predictive tool for hypothesis generation.

While here we focussed on the expression data at transcriptional level, we want to emphasize that there are certain considerations which we have not been able to take into account. More specifically, epigenetic modifications capable of operating on both DNA and protein level might dramatically change the network topology, with consequent effects on the regulatory mechanisms. They are partially considered by the network contextualisation described in Section 3.1 for the cases of epigenetic modifications which do not change between the two conditions considered (for example, health and disease states). However, this contextualisation does not take into account epigenetic modification, which changes during the described process. The impact of such modifications has not been assessed and information regarding them has not been included or modelled in any of the illustrative examples described in this dissertation. As a result of this, the efficiency of the predicted reprogramming determinants is constrained by potential mechanisms to lock gene expression (for instance, DNA methylation) and probably limited to a cell subpopulation with incomplete locking (due to stochastic events) which should be sorted by our perturbations. More research on this direction should be carried out in order to elucidate the quantitative impact of this limiting factor, and to find the most reliable way in which to model this effect. Finally, this is the best strategy with which to overcome the limitations imposed by these mechanisms and to achieve cellular reprogramming with efficiency and fidelity.

Although the methodology explained in this work was developed in the context of disease study, one may find these ideas applicable to other problems. For example, the strategy described in chapter three refers to cellular reprogramming in the context of transitions from disease to normal states or to transitions from an abundant source of cells with self-renewal to another cell without regeneration capability within the context of regenerative medicine. Despite this however, the same principle can be applied to perform control on biological living systems for basic research or industrial purposes, and could be potentially extended to higher level systems than the cellular level (tissue or cell population level).

List of figures

List of tables

Curriculum vitae

The author was born in Madrid, Spain, on April 21, 1976. In September 1994 he started his studies Veterinary Medicine at Universidad Complutense de Madrid (UCM) and he obtained both his bachelor's degree in Veterinary medicine as well as his master's degree in Animal Medicine and Health in 1997 and 2000 respectively. After seven years working as Veterinary clinician in 2007 he started his studies Animal Genetics at UCM and he obtained his master's degree in Veterinary Science in 2009. In 2009 he started his studies in Bioinformatics and Computational Biology at UCM and he obtained his master's degree in January 2011.

In December 2010 he started his postgraduate research at Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, under supervision of Prof. dr. A. del Sol. The results of this research are described in this thesis.

References

1. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annual review of genomics and human genetics 2: 343-372.
2. Kitano H (2002) Systems biology: a brief overview. Science 295: 1662-1664.
3. LV B (1976) General Systems Theory: Foundations, Development, Applications. New York: George Braziller.
4. Bar-Yam Y (1997) Dynamics of Complex Systems: Addison-Wesley.
5. Barabasi A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101-113.
6. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8: 450-461.
7. del Sol A, Balling R, Hood L, Galas D (2010) Diseases as network perturbations. Current Opinion in Biotechnology 21: 566-571.

8. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. Physical review letters 94: 128701.
9. Huang P, He Z, Ji S, Sun H, Xiang D, et al. (2011) Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. Nature 475: 386-389.
10. Ieda M, Fu JD, Delgado-Olguin P, Vedantham V, Hayashi Y, et al. (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell 142: 375-386.
11. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, et al. (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463: 1035-1041.
12. Muller FJ, Schuppert A (2011) Few inputs can reprogram biological networks. Nature 478: E4; discussion E4-5.
13. Waddington CH (1957) The Strategy of the Genes. Macmillan Publishers Limited. All rights reserved.
14. Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22: 437-467.
15. Kauffman S (1993) Origins of Order: Self-organization and Selection in Evolution. Macmillan Publishers Limited. All rights reserved.
16. Brock A, Chang H, Huang S (2009) Non-genetic heterogeneity [mdash] a mutation-independent driving force for the somatic evolution of tumours. Nat Rev Genet 10: 336-342.
17. Kitano H, Oda K, Kimura T, Matsuoka Y, Csete M, et al. (2004) Metabolic syndrome and robustness tradeoffs. Diabetes 53 Suppl 3: S6-S15.
18. Kitano H (2003) Cancer robustness: tumour tactics. Nature 426: 125.
19. Huang S, Ernberg I, Kauffman S (2009) Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. Seminars in cell & developmental biology 20: 869-876.
20. Creixell P, Schoof EM, Erler JT, Linding R (2012) Navigating cancer network attractors for tumor-specific therapy. Nature biotechnology 30: 842-848.
21. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proceedings of the National Academy of Sciences of the United States of America 104: 8685-8690.
22. Park J, Lee DS, Christakis NA, Barabasi AL (2009) The impact of cellular networks on disease comorbidity. Molecular systems biology 5: 262.
23. Wachi S, Yoneda K, Wu R (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 21: 4205-4208.
24. Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291-2297.
25. Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clinical genetics 71: 1-11.
26. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature biotechnology 25: 309-316.
27. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. American journal of human genetics 78: 1011-1025.
28. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42.
29. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD (2004) Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 101: 18006-18011.
30. Shachar R, Ungar L, Kupiec M, Ruppin E, Sharan R (2008) A systems-level approach to mapping the telomere length maintenance gene circuitry. Molecular systems biology 4: 172.

31. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. (2004) A census of human cancer genes. Nature reviews Cancer 4: 177-183.
32. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research 30: 52-55.
33. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, et al. (2004) A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. Molecular cell 15: 853-865.
34. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. Nature 437: 1032-1037.
35. Lim J, Hao T, Shaw C, Patel AJ, Szabo G, et al. (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125: 801-814.
36. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, et al. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS genetics 2: e130.
37. Sotiriou C, Piccart MJ (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nature reviews Cancer 7: 545-553.
38. Ma X, Lee H, Wang L, Sun F (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics 23: 215-221.
39. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Molecular systems biology 3: 140.
40. Efroni S, Schaefer CF, Buetow KH (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PloS one 2: e425.
41. Tuck DP, Kluger HM, Kluger Y (2006) Characterizing disease states from topological properties of transcriptional regulatory networks. BMC bioinformatics 7: 236.
42. Kitano H (2007) A robustness-based approach to systems-oriented drug design. Nature reviews Drug discovery 6: 202-210.
43. Simko GI, Gyurko D, Veres DV, Nanasi T, Csermely P (2009) Network strategies to understand the aging process and help age-related drug design. Genome medicine 1: 90.
44. Hase T, Tanaka H, Suzuki Y, Nakagawa S, Kitano H (2009) Structure of protein interaction networks and their implications on drug design. PLoS computational biology 5: e1000550.
45. Zimmermann GR, Lehar J, Keith CT (2007) Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug discovery today 12: 34-42.
46. Cavalli G, Misteli T (2013) Functional implications of genome topology. Nature structural & molecular biology 20: 290-299.
47. Wu L, Belasco JG (2008) Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. Molecular cell 29: 1-7.
48. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329: 689-693.
49. Saez-Rodriguez J, Alexopoulos LG, Epperlein J, Samaga R, Lauffenburger DA, et al. (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. Molecular systems biology 5: 331.
50. Gat-Viks I, Tanay A, Shamir R (2004) Modeling and analysis of heterogeneous regulation in biological networks. Journal of computational biology : a journal of computational molecular cell biology 11: 1034-1049.
51. Jurkowski W, Roomp K, Crespo I, Schneider J, Del Sol A (2011) PPARγ population shift produces disease-related changes in molecular networks associated with metabolic syndrome. Cell death & disease 2: e192.

52. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. Science 298: 824-827.
53. Dobrin R, Beg QK, Barabasi AL, Oltvai ZN (2004) Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. BMC bioinformatics 5: 10.
54. Yun Z, Maecker HL, Johnson RS, Giaccia AJ (2002) Inhibition of PPAR gamma 2 gene expression by the HIF-1-regulated gene DEC1/Stra13: a mechanism for regulation of adipogenesis by hypoxia. Developmental cell 2: 331-341.
55. Boyle KB, Hadaschik D, Virtue S, Cawthorn WP, Ridley SH, et al. (2009) The transcription factors Egr1 and Egr2 have opposing influences on adipocyte differentiation. Cell death and differentiation 16: 782-789.
56. Tenney R, Stansfield K, Pekala PH (2005) Interleukin 11 signaling in 3T3-L1 adipocytes. Journal of cellular physiology 202: 160-166.
57. Choi JH, Banks AS, Estall JL, Kajimura S, Bostrom P, et al. (2010) Anti-diabetic drugs inhibit obesity-linked phosphorylation of PPARgamma by Cdk5. Nature 466: 451-456.
58. Wernicke S, Rasche F (2006) FANMOD: a tool for fast network motif detection. Bioinformatics 22: 1152-1153.
59. Di Cara A, Garg A, De Micheli G, Xenarios I, Mendoza L (2007) Dynamic simulation of regulatory networks using SQUAD. BMC bioinformatics 8: 462.
60. Crespo I, Roomp K, Jurkowski W, Kitano H, Del Sol A (2012) Gene regulatory network analysis supports inflammation as a key neurodegeneration process in prion disease. BMC systems biology 6: 132.
61. Soto C (2003) Unfolding the role of protein misfolding in neurodegenerative diseases. Nature reviews Neuroscience 4: 49-60.
62. Soto C, Satani N (2010) The intricate mechanisms of neurodegeneration in prion diseases. Trends in molecular medicine.
63. Hwang D, Lee IY, Yoo H, Gehlenborg N, Cho JH, et al. (2009) A systems approach to prion disease. Molecular systems biology 5: 252.
64. Kauffman S (2004) A proposal for using the ensemble approach to understand genetic regulatory networks. Journal of theoretical biology 230: 581-590.
65. Sanz J, Cozzo E, Borge-Holthoefer J, Moreno Y (2012) Topological effects of data incompleteness of gene regulatory networks. BMC systems biology 6: 110.
66. de Silva E, Thorne T, Ingram P, Agrafioti I, Swire J, et al. (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biology 4: 39.
67. Novichkova S, Egorov S, Daraselia N (2003) MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 19: 1699-1706.
68. Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, et al. (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 20: 604-611.
69. Zinovyev A, Viara E, Calzone L, Barillot E (2008) BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. Bioinformatics 24: 876-877.
70. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.
71. Newman M, Barabási A-L, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton, New Jersey: Princeton University Press.
72. Moes M, Le Bechec A, Crespo I, Laurini C, Halavatyi A, et al. (2012) A novel network integrating a miRNA-203/SNAI1 feedback loop which regulates epithelial to mesenchymal transition. PloS one 7: e35440.
73. Bracken CP, Gregory PA, Khew-Goodall Y, Goodall GJ (2009) The role of microRNAs in metastasis and epithelial-mesenchymal transition. Cellular and molecular life sciences : CMLS 66: 1682-1699.

74. Nicoloso MS, Spizzo R, Shimizu M, Rossi S, Calin GA (2009) MicroRNAs--the micro steering wheel of tumour metastases. Nature reviews Cancer 9: 293-302.
75. Brabletz S, Brabletz T (2010) The ZEB/miR-200 feedback loop--a motor of cellular plasticity in development and cancer? EMBO reports 11: 670-677.
76. Gregory PA, Bracken CP, Bert AG, Goodall GJ (2008) MicroRNAs as regulators of epithelial-mesenchymal transition. Cell cycle 7: 3112-3118.
77. Thiery JP, Acloque H, Huang RY, Nieto MA (2009) Epithelial-mesenchymal transitions in development and disease. Cell 139: 871-890.
78. Peinado H, Olmeda D, Cano A (2007) Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? Nature reviews Cancer 7: 415-428.
79. Shiraishi T, Matsuyama S, Kitano H (2010) Large-scale analysis of network bistability for human cancers. PLoS computational biology 6: e1000851.
80. Blower PE, Verducci JS, Lin S, Zhou J, Chung JH, et al. (2007) MicroRNA expression profiles for the NCI-60 cancer cell panel. Molecular cancer therapeutics 6: 1483-1491.
81. Park SM, Gaur AB, Lengyel E, Peter ME (2008) The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes & development 22: 894-907.
82. Liu H, D'Andrade P, Fulmer-Smentek S, Lorenzi P, Kohn KW, et al. (2010) mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. Molecular cancer therapeutics 9: 1080-1091.
83. Sokilde R, Kaczkowski B, Podolska A, Cirera S, Gorodkin J, et al. (2011) Global microRNA analysis of the NCI-60 cancer cell panel. Molecular cancer therapeutics 10: 375-384.
84. Gordon AJ, Halliday JA, Blankschien MD, Burns PA, Yatagai F, et al. (2009) Transcriptional infidelity promotes heritable phenotypic change in a bistable gene network. PLoS biology 7: e44.
85. Wellner U, Schubert J, Burk UC, Schmalhofer O, Zhu F, et al. (2009) The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. Nature cell biology 11: 1487-1495.
86. Burk U, Schubert J, Wellner U, Schmalhofer O, Vincan E, et al. (2008) A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. EMBO reports 9: 582-589.
87. Vetter G, Le Bechec A, Muller J, Muller A, Moes M, et al. (2009) Time-resolved analysis of transcriptional events during SNAI1-triggered epithelial to mesenchymal transition. Biochemical and biophysical research communications 385: 485-491.
88. Batlle E, Sancho E, Franci C, Dominguez D, Monfar M, et al. (2000) The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nature cell biology 2: 84-89.
89. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. Nucleic acids research 36: D149-153.
90. Dave N, Guaita-Esteruelas S, Gutarra S, Frias A, Beltran M, et al. (2011) Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition. The Journal of biological chemistry 286: 12024-12032.
91. Guaita S, Puig I, Franci C, Garrido M, Dominguez D, et al. (2002) Snail induction of epithelial to mesenchymal transition in tumor cells is accompanied by MUC1 repression and ZEB1 expression. The Journal of biological chemistry 277: 39209-39216.
92. Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, et al. (2008) A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. Genes & development 22: 756-769.

93. Saini S, Majid S, Yamamura S, Tabatabai L, Suh SO, et al. (2011) Regulatory Role of mir-203 in Prostate Cancer Progression and Metastasis. Clinical cancer research : an official journal of the American Association for Cancer Research 17: 5287-5298.
94. Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, et al. (2009) Edgetic perturbation models of human inherited disorders. Molecular systems biology 5: 321.
95. Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, et al. (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. Proceedings of the National Academy of Sciences of the United States of America 107: 15449-15454.
96. Thuault S, Tan EJ, Peinado H, Cano A, Heldin CH, et al. (2008) HMGA2 and Smads co-regulate SNAIL1 expression during induction of epithelial-to-mesenchymal transition. The Journal of biological chemistry 283: 33437-33446.
97. Garg A, Mohanram K, Di Cara A, De Micheli G, Xenarios I (2009) Modeling stochasticity and robustness in gene regulatory networks. Bioinformatics 25: i101-109.
98. Bracken CP, Gregory PA, Kolesnikoff N, Bert AG, Wang J, et al. (2008) A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. Cancer research 68: 7846-7854.
99. Yatskou M, Novikov E, Vetter G, Muller A, Barillot E, et al. (2008) Advanced spot quality analysis in two-colour microarray experiments. BMC research notes 1: 80.
100. Saumet A, Vetter G, Bouttier M, Portales-Casamar E, Wasserman WW, et al. (2009) Transcriptional repression of microRNA genes by PML-RARA increases expression of key cancer proteins in acute promyelocytic leukemia. Blood 113: 412-421.
101. Heinaniemi M, Nykter M, Kramer R, Wienecke-Baldacchino A, Sinkkonen L, et al. (2013) Gene-pair expression signatures reveal lineage control. Nature methods.
102. Crespo I, Krishna A, Le Bechec A, del Sol A (2013) Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. Nucleic acids research 41: e8.
103. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, et al. (2010) The transcriptional network for mesenchymal transformation of brain tumours. Nature 463: 318-325.
104. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics 7 Suppl 1: S7.
105. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308: 523-529.
106. Gat-Viks I, Shamir R (2003) Chain functions and scoring functions in genetic networks. Bioinformatics 19 Suppl 1: i108-117.
107. Layek RK, Datta A, Dougherty ER (2011) From biological pathways to regulatory networks. Molecular bioSystems 7: 843-851.
108. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, et al. (2008) Models from experiments: combinatorial drug perturbations of cancer cells. Molecular systems biology 4: 216.
109. Kauffman SA (1969) J Theor Biol 22, 437.
110. Kauffman SA (1993) The Origins of Order. Oxford University Press, New York.
111. Thomas R, Thieffry D, Kaufman M (1995) DYNAMICAL BEHAVIOR OF BIOLOGICAL REGULATORY NETWORKS .1. BIOLOGICAL ROLE OF FEEDBACK LOOPS AND PRACTICAL USE OF THE CONCEPT OF THE LOOP-CHARACTERISTIC STATE. Bulletin of Mathematical Biology 57: 247-276.
112. Armananzas R, Inza I, Santana R, Saeys Y, Flores JL, et al. (2008) A review of estimation of distribution algorithms in bioinformatics. BioData mining 1: 6.

113. Mendoza L (2006) A network model for the control of the differentiation process in Th cells. Bio Systems 84: 101-114.
114. Garg A, Mendoza L, Xenarios I, DeMicheli G (2007) Modeling of multiple valued gene regulatory networks. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference 2007: 1398-1404.
115. Thomas R (1998) Laws for the dynamics of regulatory networks. The International journal of developmental biology 42: 479-485.
116. E M (1970) Schaum's Outline of Boolean Algebra and Switching Circuits. New York.
117. Garg A, Xenarios I, Mendoza L, DeMicheli G (2007) An Efficient Method for Dynamic Analysis of Gene Regulatory Networks and in silicoGene Perturbation Experiments. Research in Computational Molecular Biology. In: Speed T, Huang H, editors: Springer Berlin / Heidelberg. pp. 62-76.
118. Garg A, Di Cara A, Xenarios I, Mendoza L, De Micheli G (2008) Synchronous versus asynchronous modeling of gene regulatory networks. Bioinformatics 24: 1917-1925.
119. Speed T HH (2007) Regulatory Networks and in-silico Gene Perturbation Experiments. Research in Computational Molecular Biology Springer Berlin / Heidelberg; 2007: 62-76: Lecture Notes in Computer Science], 4453: 62-76.
120. Johnson DB (1975) Finding all the elementary circuits of a directed graph. SIAM Journal on Computing, 4: 77-84.
121. Gallagher R, Collins S, Trujillo J, McCredie K, Ahearn M, et al. (1979) Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. Blood 54: 713-733.
122. Mollinedo F, Lopez-Perez R, Gajate C (2008) Differential gene expression patterns coupled to commitment and acquisition of phenotypic hallmarks during neutrophil differentiation of human leukaemia HL-60 cells. Gene 419: 16-26.
123. Qi H, Aguiar DJ, Williams SM, La Pean A, Pan W, et al. (2003) Identification of genes responsible for osteoblast differentiation from human mesodermal progenitor cells. Proceedings of the National Academy of Sciences of the United States of America 100: 3305-3310.
124. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, et al. (2010) A global map of human gene expression. Nat Biotech 28: 322-324.
125. Muller F-J, Schuldt BM, Williams R, Mason D, Altun G, et al. (2011) A bioinformatic assay for pluripotency in human cells. Nat Meth 8: 315-317.
126. Dudley JT, Tibshirani R, Deshpande T, Butte AJ (2009) Disease signatures are robust across tissues and experiments. Molecular systems biology 5.
127. Chang R, Shoemaker R, Wang W (2011) Systematic search for recipes to generate induced pluripotent stem cells. PLoS computational biology 7: e1002300.
128. Ding S, Wang W (2011) Recipes and mechanisms of cellular reprogramming: a case study on budding yeast Saccharomyces cerevisiae. BMC systems biology 5: 50.
129. Siemens H, Jackstadt R, Hunten S, Kaller M, Menssen A, et al. (2011) miR-34 and SNAIL form a double-negative feedback loop to regulate epithelial-mesenchymal transitions. Cell cycle 10: 4256-4271.
130. Crespo I, Krishna A, Le Bechec A, Del Sol A (2012) Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states. Nucleic acids research.

131. Szabo SJ, Kim ST, Costa GL, Zhang X, Fathman CG, et al. (2000) A novel transcription factor, T-bet, directs Th1 lineage commitment. Cell 100: 655-669.
132. Lee HJ, Takemoto N, Kurata H, Kamogawa Y, Miyatake S, et al. (2000) GATA-3 induces T helper cell type 2 (Th2) cytokine expression and chromatin remodeling in committed Th1 cells. The Journal of experimental medicine 192: 105-115.
133. Hwang ES, Szabo SJ, Schwartzberg PL, Glimcher LH (2005) T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. Science 307: 430-433.
134. Tomonaga M, Golde DW, Gasson JC (1986) Biosynthetic (recombinant) human granulocyte-macrophage colony-stimulating factor: effect on normal bone marrow and leukemia cell lines. Blood 67: 31-36.
135. Collins SJ, Ruscetti FW, Gallagher RE, Gallo RC (1978) Terminal differentiation of human promyelocytic leukemia cells induced by dimethyl sulfoxide and other polar compounds. Proceedings of the National Academy of Sciences of the United States of America 75: 2458-2462.
136. Breitman TR, Selonick SE, Collins SJ (1980) Induction of differentiation of the human promyelocytic leukemia cell line (HL-60) by retinoic acid. Proceedings of the National Academy of Sciences of the United States of America 77: 2936-2940.
137. McCarthy DM, San Miguel JF, Freake HC, Green PM, Zola H, et al. (1983) 1,25-dihydroxyvitamin D3 inhibits proliferation of human promyelocytic leukaemia (HL60) cells and induces monocyte-macrophage differentiation in HL60 and normal human bone marrow cells. Leukemia research 7: 51-55.
138. Rovera G, Santoli D, Damsky C (1979) Human promyelocytic leukemia cells in culture differentiate into macrophage-like cells when treated with a phorbol diester. Proceedings of the National Academy of Sciences of the United States of America 76: 2779-2783.
139. Shen M, Bunaciu RP, Congleton J, Jensen HA, Sayam LG, et al. (2011) Interferon regulatory factor-1 binds c-Cbl, enhances mitogen activated protein kinase signaling and promotes retinoic acid-induced differentiation of HL-60 human myelo-monoblastic leukemia cells. Leukemia & lymphoma 52: 2372-2379.
140. Sekiya S, Suzuki A (2011) Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. Nature 475: 390-393.
141. McBurney MW (1993) P19 embryonal carcinoma cells. The International journal of developmental biology 37: 135-140.
142. Li H, Zuo S, Pasha Z, Yu B, He Z, et al. (2011) GATA-4 promotes myocardial transdifferentiation of mesenchymal stromal cells via up-regulating IGFBP-4. Cytotherapy 13: 1057-1065.
143. Hu DL, Chen FK, Liu YQ, Sheng YH, Yang R, et al. (2010) GATA-4 promotes the differentiation of P19 cells into cardiac myocytes. International journal of molecular medicine 26: 365-372.
144. Miura K, Okada Y, Aoi T, Okada A, Takahashi K, et al. (2009) Variation in the safety of induced pluripotent stem cell lines. Nature biotechnology 27: 743-745.
145. DeWitt N (2008) Regenerative medicine. Nature 453: 301-301.
146. Cherry AB, Daley GQ (2012) Reprogramming cellular identity for regenerative medicine. Cell 148: 1110-1122.
147. Schaffter T, Marbach D, Floreano D (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27: 2263-2270.
148. Wang J, Lu M, Qiu C, Cui Q (2010) TransmiR: a transcription factor–microRNA regulation database. Nucleic acids research 38: D119-D122.
149. Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang W-C, et al. (2010) miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic acids research.

150. Rosenfeld N, Elowitz MB, Alon U (2002) Negative autoregulation speeds the response times of transcription networks. Journal of molecular biology 323: 785-793.
151. Koide T, Hayata T, Cho KW (2005) Xenopus as a model system to study transcriptional regulatory networks. Proceedings of the National Academy of Sciences of the United States of America 102: 4943-4948.
152. Locke JC, Kozma-Bognar L, Gould PD, Feher B, Kevei E, et al. (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. Molecular systems biology 2: 59.
153. Mendoza L, Pardo F (2010) A robust model to describe the differentiation of T-helper cells. Theory in biosciences = Theorie in den Biowissenschaften 129: 283-293.
154. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. The Journal of neuroscience : the official journal of the Society for Neuroscience 28: 264-278.
155. Lovatt D, Sonnewald U, Waagepetersen HS, Schousboe A, He W, et al. (2007) The transcriptome and metabolic gene signature of protoplasmic astrocytes in the adult murine cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 27: 12255-12266.
156. Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. Nucleic acids research 29: e45.
157. Graf T, Enver T (2009) Forcing cells to change lineages. Nature 462: 587-594.
158. Cantor AB, Orkin SH (2001) Hematopoietic development: a balancing act. Current opinion in genetics & development 11: 513-519.
159. Graf T (2002) Differentiation plasticity of hematopoietic cells. Blood 99: 3089-3101.
160. Orkin SH, Zon LI (2008) Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132: 631-644.
161. Arinobu Y (2007) 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.
162. Iwasaki H, Akashi K (2007) Myeloid lineage commitment from the hematopoietic stem cell. Immunity 26: 726-740.
163. Zhou L (2008) TGF-[bgr]-induced Foxp3 inhibits TH17 cell differentiation by antagonizing ROR[ggr]t function. Nature 453: 236-240.
164. Laslo P (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. Cell 126: 755-766.
165. Frontelo P (2007) Novel role for EKLF in megakaryocyte lineage commitment. Blood 110: 3871-3880.
166. Hwang ES, Szabo SJ, Schwartzberg PL, Glimcher LH (2005) T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. Science 307: 430-433.
167. Heins N (2002) Glial cells generate neurons: the role of the transcription factor Pax6. Nature Neurosci 5: 308-315.
168. Kajimura S (2008) Regulation of the brown and white fat gene programs through a PRDM16/CtBP transcriptional complex. Genes Dev 22: 1397-1409.
169. Niwa H (2005) Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. Cell 123: 917-929.
170. Ralston A, Rossant J (2005) Genetic regulation of stem cell origins in the mouse embryo. Clin Genet 68: 106-112.
171. Heyworth C, Pearson S, May G, Enver T (2002) Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. The EMBO journal 21: 3770-3781.

172. Del Vecchio D, Ninfa AJ, Sontag ED (2008) Modular cell biology: retroactivity and insulation. Molecular systems biology 4: 161.
173. Chalancon G, Ravarani CN, Balaji S, Martinez-Arias A, Aravind L, et al. (2012) Interplay between gene expression noise and regulatory network architecture. Trends in genetics : TIG 28: 221-232.
174. Krumsiek J, Marr C, Schroeder T, Theis FJ (2011) Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. PloS one 6: e22649.
175. Dore LC, Crispino JD (2011) Transcription factor networks in erythroid cell and megakaryocyte development. Blood 118: 231-239.
176. Murphy KM, Reiner SL (2002) The lineage decisions of helper T cells. Nature reviews Immunology 2: 933-944.
177. Jopling C, Boue S, Izpisua Belmonte JC (2011) Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. Nature reviews Molecular cell biology 12: 79-89.
178. Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, et al. (2009) Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular systems biology 5: 294.
179. Mussel C, Hopfensitz M, Kestler HA (2010) BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. Bioinformatics 26: 1378-1380.
180. Crespo I, Del Sol A (2013) A General Strategy for Cellular Reprogramming: The Importance of Transcription Factor Cross-Repression. Stem cells.