UNIVERSITÉ DU LUXEMBOURG

PhD-FSTC-2019-60
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 12/09/2019 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *BIOLOGIE*

by

Gaia ZAFFARONI
Born on 30 June 1990 in Segrate (Italy)

INTEGRATIVE APPROACH TO PREDICT SIGNALLING PERTURBATIONS FOR CELLULAR TRANSITIONS: APPLICATION TO REGENERATIVE AND DISEASE MODELS

Dissertation defence committee

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

Dr Dunja Knapp
Center for Regenerative Therapies, Dresden
Technische Universität Dresden

A-Prof. Dr Enrico Glaab
Assistant professor, Université du Luxembourg

Prof. Dr Thomas Sauter, Chairman
Professor, Université du Luxembourg

Affidavit

I hereby confirm that the PhD thesis entitled “Integrative approach to predict signalling perturbations for cellular transitions: application to regenerative and disease models” has been written independently and without any other sources than cited.

Luxembourg, 08/08/2019

Gaia Zaffaroni

Dedication

E quindi uscimmo a riveder le stelle.

Thence we came out, and saw again the stars.

Dante Alighieri (Inferno XXXIV, 139)

Index

List of Figures

List of Tables

Summary

Transitions between cellular states are involved in all kind of biological events: cells differentiate during development, assume different phenotypes because of experimental or pathological conditions, and change their identity during cellular reprogramming. The possibility to induce specific cellular transitions represents a great opportunity for disease treatment and regenerative medicine. Cellular states are maintained by gene regulatory networks, and the manipulation of their master regulators can trigger cellular transitions. Modulating signal transduction is a convenient way to obtain such transitions avoiding transfer of genetic material, with substantial benefit in terms of clinical safety. While multiple methods separately consider gene regulatory networks for cellular transitions, and the role of signalling pathways in biological processes, no approach so far integrates the two regulatory layers in order to identify signalling perturbations that can induce desired cellular transitions.

This thesis presents methods for the prediction of signalling molecules or pathways that regulate the gene regulatory network (GRN) of a given cellular state and induce its transition to the desired state. The overall approach consists in the integration of signalling pathways and GRNs to model how signalling cues act on the transcription factors (TFs) that sit at the interface between the two regulatory layers, and cause changes in the gene expression program. To this end, gene expression data was used to estimate the probability of signalling molecules to activate and inhibit interface TFs, combined with in silico perturbations of the Boolean GRN underlying the desired cellular conversion.

This approach was systematically applied to the prediction of signalling molecules and pathways involved in cellular transitions, with particular focus on reprogramming and differentiation cases. The predictions obtained consistently recapitulated experimental perturbations and literature knowledge. Additionally, the methods proposed outperformed available tools for the prediction of both signalling molecules and pathways.

To show the applicability of this approach to disease treatment in an animal model, signalling perturbations were predicted for the reversal of cirrhotic liver state to healthy in rat.

Experimental results confirmed that the activation of the angiopoietins receptor Tie2 produced favorable changes in the gene expression of TFs that play an important role in cirrhosis. Finally, this approach was applied to the analysis of limb regeneration upon amputation in the salamander *Ambystoma mexicanum*. Using time series gene expression data, predictions were generated to retrieve the signals activated and inhibited along the regeneration process. Literature evidence connected predicted pathways and proteins to specific regeneration stages, clarifying their relation to wound healing, blastema formation and differentiation.

1 Introduction

The Greek philosopher Heraclitus has been quoted as saying “change is the only constant in life”. At the cellular level, this is certainly true: transitions between different states during development bring the embryo to form a complete organism, with specialized cells, tissues and organs. Changes in cellular state are also required to adapt to new conditions and response to external stimuli, and are detrimental when associated with pathologic conditions. The ability of inducing specific cellular transitions is a fundamental tool for disease treatment and regenerative medicine. Cellular states are modelled as gene expression profiles determined by gene regulatory networks. These models have been used for the selection of transcriptional cellular determinants before, but their manipulation raises safety concerns for therapeutic use. On the other hand, chemical compounds, small molecules and growth factors can trigger cellular transitions by acting on the signalling network. Signalling has so far been studied in isolation from the gene regulatory network that determines a cell’s state, but their integration needs to be achieved in order to computationally select signalling perturbations that can induce desired cellular transitions.

In this chapter, cellular transitions are described and their relevance is delineated in Section 1.1. The use of gene regulatory networks as models for cellular states will be introduced (Section 1.2). The role of signalling in relation to cellular state, and the computational tools available for its analysis are presented in Section 1.3. Finally, an introduction is given for the animal models used in this thesis, which are used to get insight into disease and regeneration (Section 1.4).

1.1 Cellular transitions for disease treatment and regenerative medicine

A cellular transition, also referred to as cellular conversion, is here defined as the shift of a cell from an initial state, to a desired cellular state. Depending on the nature of the difference between the initial and final state, the transitions can be classified into two broad categories: cell state transitions and cell fate transitions.

Cell fate transitions concern changes in cellular identity, including differentiation, transdifferentiation and reprogramming. Following the seminal work of (Takahashi and Yamanaka 2006), the induction of cell fate transitions has found many applications in multiple areas of biomedical research. On one hand, the possibility of deriving cells that are difficult to obtain from patients in the lab has revolutionized the field of disease modelling by overcoming the limitations associated with animal disease models, allowing the study of monogenic, chromosomal and complex disorders (Avior, Sagi, and Benvenisty 2016). Examples of cell types obtained in vitro for disease modelling are neurons (Chang et al. 2018; Hong Li et al. 2018), cardiomyocytes (Brandão et al. 2017) and hepatocyte-like cells (Parafati et al. 2018).

On the other hand, the induction of cell fate transitions has a potential application in regenerative medicine, which has the objective of restoring the functionality of tissues or organs by the replacement of damaged cells with induced cells, present in situ or transplanted. Extensive effort has been put towards in vivo reprogramming to obtain hepatocytes (Song et al. 2016), β-cells (Q. Zhou et al. 2008), cardiomyocytes (Qian et al. 2012), and others, as reviewed in (Srivastava and DeWitt 2016). The transplantation of cells and tissues obtained in vitro is already used in therapy, for example for the treatment of epidermolysis bullosa (Hirsch et al. 2017), or ADA-SCID, a rare immunodeficiency syndrome (Kuo and Kohn 2016).

In both disease modelling and regenerative medicine, the experimental challenge is to induce cells to change their identity in a controlled manner. This can be obtained by inducing a pluripotent state first and then deriving more differentiated cells from it, as in the case of induced pluripotent stem cells (iPSCs) stimulation and differentiation, or by direct transdifferentiation of the initial cells to the desired type (J. Xu, Du, and Deng 2015).

Cell state transitions can be defined as the phenotypic conversion within a same cell type, such as transitions between healthy and disease states, between treatment with small molecules or growth factors and control state, or between different cell culture conditions. The interest in inducing these transitions is manifold. The insurgence of a disease can be considered at the cellular level as a cell state transition, as cellular identity might be unaffected, but gene expression, signalling activity, metabolic state, and function can be compromised. In this sense, a successful cure is one that induces the transition between the disease and healthy cellular state. In other cases, cellular transitions are a consequence of experiments performed to clarify the effect of perturbations on the cellular state. For instance, the development of therapeutic drugs can be extremely expensive, so the initial assessment of a compound activity and toxicity is often relying on cell-based assays, which compare the initial and the treated cellular state to assess the effects of chemical compounds (Michelini et al. 2010; O'Brien et al. 2006). Similarly, cellular transitions are also induced when performing experiments in order to understand drugs mechanism of action, and in general to study the mechanism of cellular biological processes. The discovery of new drugs for a disease or the proteins involved in a particular phenotype can therefore be reframed as the identification of factors that can trigger a desired cellular conversion.

Overall, the induction of desired cellular transitions is of clinical interest, as it allows to revert diseased cells to normal counterparts, or to derive desired cells and organs for cell replacement therapies. Computational methods can be instrumental to the application of cellular transitions in therapy by reducing the cost and time required for the experimental discovery of the factors that are able to trigger the desired transition.

1.2 Gene regulatory network models for cellular transitions

Conrad Waddington introduced in 1957 the intuitive representation of cellular differentiation as a ball moving in an “epigenetic landscape”, an inclined and rugged surface with multiple valleys and stable points (Waddington 1957). Stem and progenitor cells descend following valleys towards a more differentiated cell type, and on the way encounter forking points, which represent metastable cellular states that are forced to adopt one of few possible states by undergoing transcriptional changes. The topography of the “epigenetic landscape” is thus defined by the coordinated expression of genes emerging from the interaction of transcriptional regulators, termed gene regulatory network.

1.2.1 Gene regulatory networks as underlying programs defining cellular states

The differences in gene expression observed between two cellular states might span thousands of genes, but each gene expression program (and thus the cell state) is maintained by a gene regulatory network (GRN) containing only several genes. GRNs are used as models that represent the transcriptional interactions occurring among genes in a particular cellular state. While the potential interactions between transcriptional regulators and their gene targets are encoded in the genomic sequence, the interactions taking place in any specific cellular condition are determined by the expression of the regulators (S. Huang 2012) (Figure 1.1).

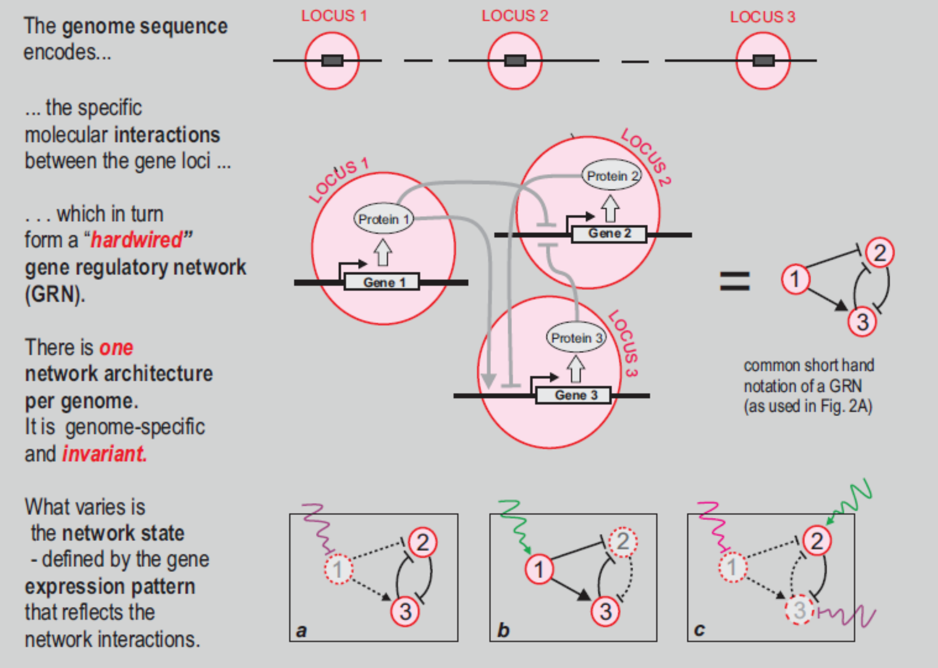

The genome sequence encodes...
LOCUS 1
LOCUS 2
LOCUS 3
... the specific molecular interactions between the gene loci ...
. . . which in turn form a "hardwired" gene regulatory network (GRN).
There is one network architecture per genome. It is genome-specific and invariant.
What varies is the network state - defined by the gene expression pattern that reflects the network interactions.
Protein 1
Gene 1
Protein 2
Gene 2
Protein 3
Gene 3
=
1
2
3
common short hand notation of a GRN (as used in Fig. 2A)
a
b
c

Figure 1.1: The interactions between transcriptional regulators are defined from the genomic sequence, as is the DNA sequence that defines if a transcriptional regulator can bind the regulatory region and modulate the expression of other genes. However in different cellular conditions, each regulator can be expressed or not because of external factors, resulting in different portions of the overall GRN being active. Figure from (S. Huang 2012).

There are multiple modelling formalisms used to represent gene regulatory networks. Boolean modelling was first introduced by Kauffman in the 1970s and is one of the most widely used approaches. Transcriptional regulators activity was shown to follow a sigmoid or "switch" function, which can be approximated with a step function resulting in two discrete states (H.-C. Chen et al. 2004). Consistently with this observation, in the Boolean framework genes can take one of two states (active or inactive), and their state is defined through logic functions by the state of their regulators (Glass and Stuart A. Kauffman 1973; S. Kauffman 1969). Gene expression profiles can be reduced to vectors of Boolean gene states, which are used to represent cellular states. Among all possible Boolean states there are some that are stable, meaning that if no perturbation intervenes on the GRN, the states of each gene remain constant with updates of the network state. These stable states in the network state space are called attractors. Kauffman suggested that each stable cellular state or cell type can be associated with an attractor (Stuart A Kauffman 1993), so that cellular transitions correspond to shifts between network attractors (S. Huang 1999). Boolean modelling allows to capture coarse-grained properties of large GRNs, while also simplifying the construction of models from gene expression data (S. Huang 1999). For example, the Boolean modelling formalism allows to represent experimentally observed phenotypes and predict the outcome of novel perturbations. The manual reconstruction of Boolean models connecting signalling and transcriptional regulators allowed to clarify the role of specific proteins in the context of Th cell differentiation (Naldi et al. 2010), terminal differentiation of B cells (Méndez and Mendoza 2016), and mesoderm specification in Drosophila (Mbodj et al. 2016). Boolean GRN models recovered known phenotypes during the differentiation of myeloid progenitors (Krumsiek et al. 2011) and correctly predicted the outcome of perturbations applied to mouse embryonic stem cells (H. Xu et al. 2014) and to the hematopoietic system (Moignard et al. 2015).

1.2.2 GRN models for cellular transitions

The conversion of fibroblasts to myoblasts by over-expression of the transcription factor (TF) Myod (Davis, Weintraub, and A B Lassar 1987), and the induction of pluripotent stem cells from fibroblasts with the activation of only four TFs (Takahashi and Yamanaka 2006), opened the way to the manipulation of TF expression for the purpose of inducing cellular conversions, which has become the most widely used strategy for lineage reprogramming and generation of pluripotent cells. A number of computational methods exists for identifying the TFs that can trigger cellular reprogramming or differentiation using gene expression data. Given a GRN, it is possible to identify its master regulators and shortlist TFs that can trigger changes in the GRN state, resulting in cellular transitions. Multiple features are used in order to identify such master regulators (Hartmann, Ravichandran, and Sol 2019), among which:

- Gene expression of the TFs in the GRN
- Network topology and network motifs, such as positive circuits and strongly connected components, which play an important role in shaping the GRN attractor space
- GRN attractor states that can be reached with in silico perturbations of GRN components and their combinations

Effective computational methods usually combine these features for the prediction of TFs triggering cellular transitions. For example, in SeesawPred the normalized ratio difference (Okawa et al. 2016) used to prioritize TFs that show significant change in daughter cells compared to progenitors is combined with the identification of strongly connected components (Hartmann, Okawa, et al. 2018, see Appendix). CellNet compares the gene expression of a query dataset to the reference expression of genes in cell/tissue specific GRNs, and indicates which genes would bring the query closer to the reference by topological measures (Cahan et al. 2014). Mogrify selects TFs that are differentially expressed and jointly regulate all other differentially expressed genes, therefore taking into account the topology of the GRN in its prediction (Rackham et al. 2016). In (Lang et al. 2014), the authors use an attractor-based landscape model to calculate the “predictivity” of TFs for a given cellular state, which is then combined with their expression. Additionally, other computational methods exist that do not consider GRNs but identify transcription factors that are uniquely expressed in the desired cell type, which are to be activated in order to induce the corresponding cellular transition (D’Alessio et al. 2015).

1.2.3 Induction of cellular transitions with chemical compounds

The expression of master regulators is normally manipulated with the introduction in the cells of exogenous DNA coding for the TF protein required. Multiple vectors and strategies exist in order to obtain efficient and reliable expression of the desired proteins. Despite these efforts, in general these approaches raise clinical safety concerns related to the integration of exogenous DNA in the cellular genome, which can lead to insertional mutagenesis and tumour formation (Ben-David and Benvenisty 2011; Okita, Ichisaka, and Yamanaka 2007). Alternative DNA delivery systems that do not require integration, using both viral and non-viral vehicles, have recently been developed (Hardee et al. 2017; Makhija et al. 2018; Uludag, Ubeda, and Ansari 2019). However, these techniques still show very limited efficiency and are therefore unsuitable for clinical application (Haridhasapavalan et al. 2019).

Small molecules and chemical compounds are routinely used to alter the cells environment, in order to reproduce in vitro pathological conditions, or study the cellular response to drugs and other stimuli. In recent years, their use for cell reprogramming and differentiation has emerged as a promising strategy for in vitro and in vivo applications (Federation, Bradner, and Meissner 2014; Heng Li and Homer 2010; Qin, Zhao, and Fu 2017) because they provide a valid strategy for cost-effective, transient, controllable induction of cellular transitions (De et al. 2017; J. Xu, Du, and Deng 2015; Pesaresi, Sebastian-Perez, and Cosma 2019). Thus, the use of chemical compounds is preferable for the generation of cells in clinical applications (Takeda et al. 2018) (Figure 1.2).

The combinations of compounds, small molecules and growth factors that trigger cellular transitions are defined through phenotype- and target-based screenings. Phenotype-based screenings consist in performing multiple assays in order to identify molecules that can induce the desired phenotype, without requiring a priori knowledge of the cellular transition considered. On the other hand, target-based screenings use such information to restrict the analysis to regulators of pathways involved in the cellular transition (De et al. 2017). Thus, the identification of chemical factors and small molecules inducing cellular transitions is usually lengthy, expensive and labour intensive (Cao et al. 2016; Y. Tang and Cheng 2017). The

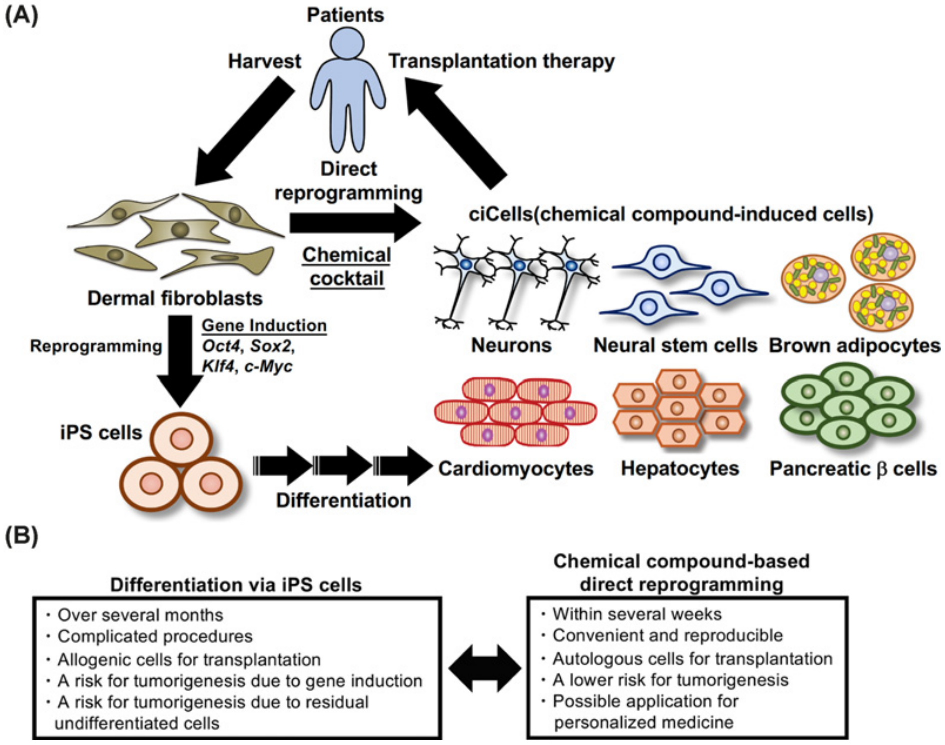

(A)
Patients
Harvest
Transplantation therapy
Direct reprogramming
Chemical cocktail
ciCells(chemical compound-induced cells)
Dermal fibroblasts
Reprogramming
Gene Induction
Oct4, Sox2, Klf4, c-Myc
Neurons
Neural stem cells
Brown adipocytes
iPS cells
Cardiomyocytes
Hepatocytes
Pancreatic β cells
Differentiation
(B)
Differentiation via iPS cells
・Over several months
・Complicated procedures
・Allogenic cells for transplantation
・A risk for tumorigenesis due to gene induction
・A risk for tumorigenesis due to residual undifferentiated cells
Chemical compound-based direct reprogramming
・Within several weeks
・Convenient and reproducible
・Autologous cells for transplantation
・A lower risk for tumorigenesis
・Possible application for personalized medicine

Figure 1.2: Schematic depiction of the derivation of cells of interest from patient fibroblasts by either transdifferentiation or reprogramming to induced pluripotent stem cells (iPSCs) followed by differentiation. The limitations observed in the use of iPSCs and corresponding advantages of transdifferentiation are mentioned. Figure from (Takeda et al. 2018).

development of computational methods for the prediction of non-transcriptional perturbations that induce cellular transitions is therefore desirable.

1.3 Signalling network models

Small molecules and chemical compounds induce cell fate transitions mainly by acting either on signalling pathways, on the cellular metabolism, or by altering the status of the chromatin, therefore modifying the gene expression program of the cells. Signalling events play a major role in cellular transitions. During development, the concerted action of combinations of signalling pathways over time dictates the formation of tissues and organs (Basson 2012; Perrimon, Pitsouli, and Shilo 2012). Multiple signalling pathways are also involved in maintaining tissue homeostasis (Biteau, Hochmuth, and Jasper 2011). In pathological conditions, the deregulation of signalling pathways induces inflammation, immune response, and disease specific phenotypes.

The signalling network consists of proteins and small molecules that can interact activating and inhibiting each other, resulting in the transmission of external cues to the nucleus, where the signals are integrated and the cellular response is defined. While signal transduction is a probabilistic process that has low probability of emerging over a background of unspecific protein-protein interactions, multiple strategies are used to ensure robust and correct transmission of signals, such as compartmentalization, multimerisation, and integration of multiple signals (Ladbury and Arold 2012). Among all signal transduction paths, there are some stereotypical ones that have been well described and act as functional units, that were termed canonical signalling pathways. The concept of canonical signalling has with time been put in discussion, as more and more evidence is found that alternative signal transduction paths exist and play important roles in many cellular contexts (Meyerovich et al. 2016; Ohta et al. 2016; Regan et al. 2017; Voloshanenko et al. 2018). However, canonical pathways still represent functional entities useful for the interpretation of high-throughput experiments such as gene expression analysis, proteomics or phosphoproteomics data, where significant differences between two conditions might be present for thousands of genes or proteins, and pathway enrichment analysis methods are applied to identify which are the biological functions associated with them.

Despite their central role in the regulation of all kinds of biological processes, the dynamics of signalling events remain difficult to study experimentally because signalling pathways involve many different types of interactions among proteins (M. J. Lee and Yaffe 2016): phosphorylation and other post-transcriptional modifications, complex formation, compartmentalization, transport. Phosphorylation is the main strategy used for signal transduction in many canonical pathways (Ardito et al. 2017) and results primarily in transcriptional regulation (Needham et al. 2019). Phosphorylation can happen on different amino acids of a protein (phosphosites) controlling the activity state of the protein by inducing changes in protein 3D structure, cellular surface charge, binding to protein partners (M. J. Lee and Yaffe 2016). However, current phosphoproteomics techniques do not allow an absolute, proteome-wide quantification of phosphorylation events (Needham et al. 2019). In fact, in any given phosphoproteomics experiment only 40 to 60% of the phosphosites existing across the proteome are captured (Invergo and Beltrao 2018; Vlastaridis et al. 2017). Additionally, only 3% of all human identified phosphosites have known functional role (Needham et al. 2019).

Given the limitations and the constraints that characterize phosphoproteomics experiments, many studies have used gene expression data to get insights into signalling events. The numerous attempts to assess the level of agreement between transcriptomics, proteomics and phosphoproteomics data reached discordant conclusions, with some studies finding moderate correlation, while others only reported qualitative similarities among the measurements (Gnad, Wallin, et al. 2016; Kandasamy et al. 2016; Olsen et al. 2010; Pines et al. 2011; Richter et al. 2015; Rotival et al. 2015; Blevins et al. 2019). Overall, transcription rates, protein abundance and phosphorylation of protein residues cannot generally be considered equivalent or informative of each other. On the other hand, cellular response to signalling perturbations has been shown to change across cellular populations according to their state, and in particular to the abundance of specific signalling proteins, prior to perturbation (Niepel et al. 2017; Strasen et al. 2018).

1.3.1 Use of gene expression in signalling analysis

Multiple computational approaches exist that aim at explaining how the transition between two gene expression profiles is associated with signalling events: either in the context of drug profiling, where the methods are used to predict drugs mechanism of action, or in diseases, an approach particularly useful to identify dysregulated signalling pathways in cancer. Two broad classes of computational methods exist, namely GRN-free and GRN-based approaches. GRN-free methods map gene expression data on signalling pathways in order to define whether they are activated or inhibited, and focus mostly on differentially expressed genes (DEGs). GRN-based methods, on the other hand, consider also transcriptional regulation interactions in their analysis and in particular try to explain expression changes in groups of genes with their upstream signalling regulators.

Starting with Connectivity Map (Lamb et al. 2006), a number of methods for the identification of cellular perturbations have been developed. These methods (Lamb 2007; Parikh et al. 2010; Schubert et al. 2018) compare the signature of a query perturbation with a database of gene expression signatures, obtained perturbing the signalling network in a controlled manner (with drugs and chemical compounds, signalling pathways inhibitors and activators, overexpression or knock-down of signalling proteins), and select candidate perturbations based on the similarity of signatures. Their predictions are limited to the perturbations present in their corresponding databases and rely on signatures from a limited set of cell types or lines.

Another class of GRN-free methods is represented by pathway enrichment approaches that use transcriptomics data to infer which signalling pathways are associated with different cellular conditions. They comprise methods based on over-representation of DEGs in signalling pathways (Sartor, Leikauf, and Medvedovic 2009; Subramanian, Tamayo, et al. 2005), and approaches taking into account pathway topology or cross-talk (Dutta, Wallqvist, and Reifman 2012; Massa, Chiogna, and Romualdi 2010; Naderi Yeganeh and Mostafavi 2017; Tarca et al. 2009). Such methods have been successfully applied to the understanding of multiple diseases, with a particular focus on cancer (Dutta, Wallqvist, and Reifman 2012; Sebastian-Leon et al. 2014). Additionally, they were used to study differentiation trajectories and reconstruct the role of signalling pathways during development (Dutkowski and Ideker 2011). Because canonical pathways are highly variable depending on the database used (Kirouac et al. 2012; Türei, Korcsmáros, and Saez-Rodriguez 2016) and subject to extensive crosstalk (Schaefer et al. 2009), a different class of methods tries to address this problem by predicting sets of proteins corresponding to functional units, termed sub-pathways (Amadoz et al. 2015; Han et al. 2015; Haynes et al. 2013; Hidalgo et al. 2017; Martini et al. 2013). It is important to notice that pathways are not predictive of the transcriptional state of their targets (Housden and Perrimon 2014), therefore these methods give limited insight on the effects of signalling events on the GRN.

Efforts towards the integration of the GRN in the modelling framework have been limited to far. Various studies have presented manually-curated integrated models for individual systems. For example in (Peng et al. 2010) the signalling activity of IKK proteins on NF-κB is captured using both transcriptional activity and signalling activity predictions resulting from manually curated models. (Zañudo and Albert 2015) presents a method to identify reprogramming factors in manually curated Boolean models containing both signalling and transcriptional interactions. In (Mbodj et al. 2016) a Drosophila development model built manually is iteratively refined to match literature knowledge and experimental data. Finally, Yachie-Kinoshita et al. 2018 generated a manually curated Boolean network model for pluripotent stem cells where pathways are represented as single nodes that can act on effector TFs and regulate the GRN. Available general GRN-based methods use ordinary differential equations (ODEs) to model the expression level of each gene as a function of the expression of its regulators. Signalling molecules that are involved in the perturbation response are selected based on the fact that their expression alone cannot explain the expression of the genes they regulate (Balwierz et al. 2014; Cotton et al. 2015; Noh, Shoemaker, and Gunawan 2018; Osmanbeyoglu et al. 2014). These approaches need to extract parameters from a large amount of expression data experiments, i.e. the same cell type treated with multiple compounds or other type of perturbations. This limits their applicability to only a few well studied cell types.

Gene expression data is also used routinely to discover drugs mechanism of action (MoA), identifying their targets and downstream effectors. There are different modelling approaches used for this purpose, and these methods might make use of signalling networks, or be network-free. DeMAND (Woo et al. 2015) is considered the state of the art method for the identification of signalling proteins that are involved in the gene expression response to drug treatment. It considers both signalling and transcriptional networks, in the form of regulons. Regulons are in this case bipartite networks where each protein is connected to the genes that it regulates at the level of transcription, signalling or complex formation. A signalling protein is predicted as drug MoA if the expression of the genes that it regulates is significantly perturbed by the drug treatment. Methods for the inference of causal signalling networks that can explain how signalling events result in gene expression changes follow a similar strategy. In particular, methods have been developed to reconstruct the path followed by signal transduction from the site of perturbation to transcription factors in order to identify drug mode of action (Melas et al. 2015) or signalling rewiring caused by genetic mutations in cancer (Y.-A. Kim, Wuchty, and Przytycka 2011; Paull et al. 2013). The output of these methods are signalling sub-networks that connect the perturbation site to the genes showing de-regulated expression. While these methods reconstruct signalling networks that are specific for the perturbation applied, being it a ligand or a drug, they only consider superficially how signalling triggers changes in gene expression, because they do not consider the gene expression changes obtained indirectly through the interplay of TFs in the GRN.

1.4 Application to animal models of disease and regeneration

1.4.1 Liver cirrhosis

Liver cirrhosis is the fourth most common cause of death in central Europe and causes more than one million deaths per year worldwide (Tsochatzis, Bosch, and Burroughs 2014). It is defined as an advanced stage of liver fibrosis caused by chronic inflammatory injury of the liver tissues, which might be caused by hepatitis B or C, alcohol abuse or non-alcoholic liver diseases (Schuppan and Afdhal 2008). Inflammation induces portal and perivascular fibroblasts and quiescent stellate cells (HSCs) to transdifferentiate and activate into myofibroblasts. These cells are normally associated with wound healing: they deposit collagen to strengthen the extracellular matrix and contract the edges of the wound. Generally when the injury subsides, myofibroblasts undergo apoptosis and the fibrosis is resolved (Scott L Friedman 2008; Hinz et al. 2007). This does not occur in chronic inflammation conditions and myofibroblasts continue to proliferate, synthetize excess collagen, and limit the activity of interstitial metalloproteinases (MMPs) that could degrade the type I collagen prevalent in fibrotic liver (Scott L Friedman 2008; Schuppan and Afdhal 2008) (Figure 1.3).

In cirrhosis, hepatic angiogenesis is closely associated with the fibrotic process. The liver wound healing process is characterized by the expression of proteins and growth factors that have pro-fibrotic and pro-angiogenetic role such as PDGF, VEGF, FGF and TGF-β1, and proteins involved in the remodelling of the extra-cellular matrix (ECM), such as β-catenin, ephrins, integrins and other adhesion molecules (Fernández et al. 2009). Deregulated angiogenesis causes the distortion of hepatic vasculature resulting in increased hepatic

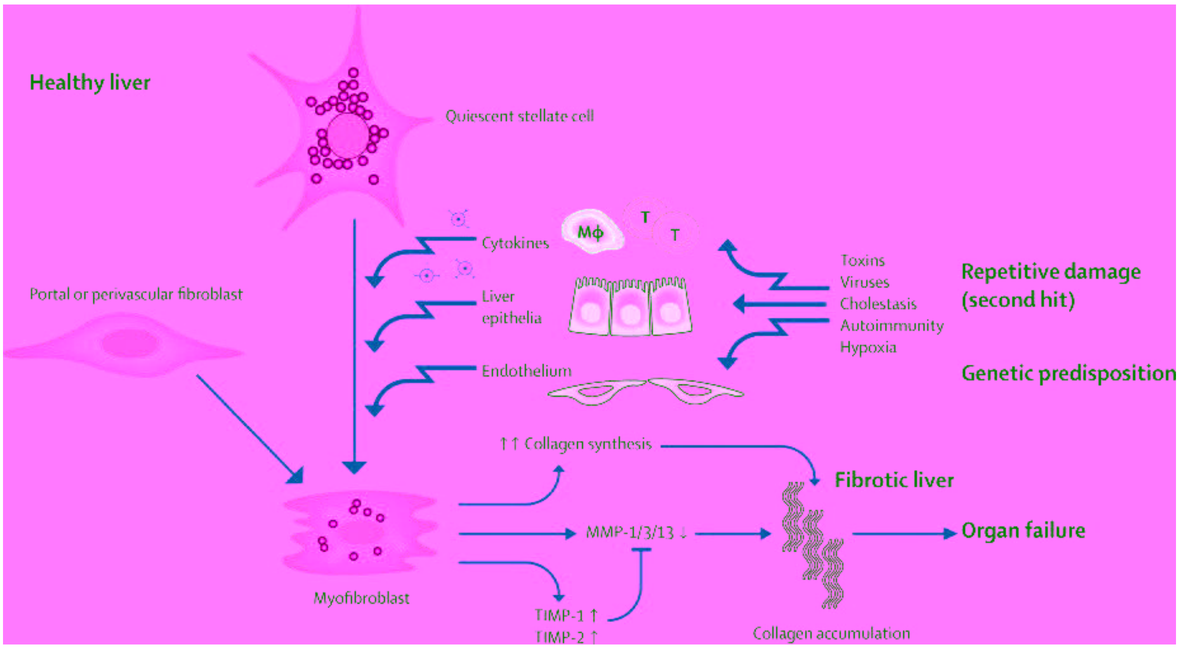


Figure 1.3: Figure from (Schuppan and Afdhal 2008) showing how chronic inflammation causes accumulation of myofibroblasts, which produce excessive amounts of collagen and inhibit metalloproteinases (MMPs) resulting in liver fibrosis.

vascular resistance and portal hypertension, with consequent failure of the hepatic functions. The hypoxic conditions induced by portal fibrosis further induce neo-angiogenesis through the activation of hypoxia-inducible factors (Elpek 2015).

The recommended treatment of cirrhosis depends on its clinical stage, but current therapies aim at avoiding the progression of the disease to advanced stages, when liver transplant is the only available treatment (Tsochatzis, Bosch, and Burroughs 2014). Some therapeutic strategies have been tested in order to control new blood vessel formation and promote the regression of portal hypertension (Fernández et al. 2009). Additionally, sorafenib and sunitinib proved effective anti-angiogenic factors in experimental models, also showing anti-fibrotic effects (Mejias et al. 2009; Tugues et al. 2007). Thus, therapies acting against both angiogenesis and fibrosis should be considered. Inhibition of VEGF activity however has been associated previously with negative effects (H. X. Chen and Cleck 2009), and new therapeutic strategies are needed. Classically, rats have been used as animal model for cirrhosis, as their response to fibrotic stimuli is more robust as compared to wild-type mice

(Bissell 2011). Carbon tetrachloride ($\mathrm{CCl_4}$) is a hepatic toxin widely used to induce liver injury. Upon repeated insult, the rat liver develops progressive fibrosis and finally cirrhosis (Y. Liu et al. 2013), showing great similarity with human liver cirrhosis. Additionally, the fibrotic process is reversed upon withdrawal of the toxin, allowing to study fibrosis reversibility (Y. Liu et al. 2013).

1.4.2 Axolotl limb regeneration

Regeneration is defined as the reactivation of developmental processes outside the development stages in order to restore missing tissues or organs. Humans have very limited regenerative capabilities, and liver is the only organ showing large scale regeneration, albeit compensatory: the cells present in the liver proliferate to compensate the missing liver lobes, but they maintain their differentiated state (Abu Rmilah et al. 2019). True regeneration occurs in humans only at the distal portion of the digits (DOUGLAS 1972). Conversely, some animals have higher regeneration potential and undergo epimorphic regeneration: the differentiated structures present at the injury site de-differentiate, giving rise to a structure called blastema, and differentiate again in the completely re-established body part. Examples of animals with regenerative capabilities include planarians (Reddien and Alvarado 2004), zebrafish (Gemberling et al. 2013), echinoderms (Ben Khadra et al. 2017), and amphibians. In particular, urodeles (newts and salamanders) show regeneration of limbs (Stocum 2017) and other parts of their bodies, such as spinal cord (Chernoff et al. 2003), skin (Yokoyama et al. 2018), retina (Mitashov 1996), and heart (Garcia-Gonzalez and Morrison 2014).

The neotenic salamander *Ambystoma mexicanum* is a classical model of regeneration that can regrow limbs, tail, heart, liver, among others (Stocum and Cameron 2011). The limb regeneration process upon amputation consists of 3 main phases: wound healing, blastema formation, and limb re-development (Figure 1.4). These phases are common to all vertebrate appendage regeneration (Miller, Johnson, and Whited 2019), and require multiple cellular transitions.

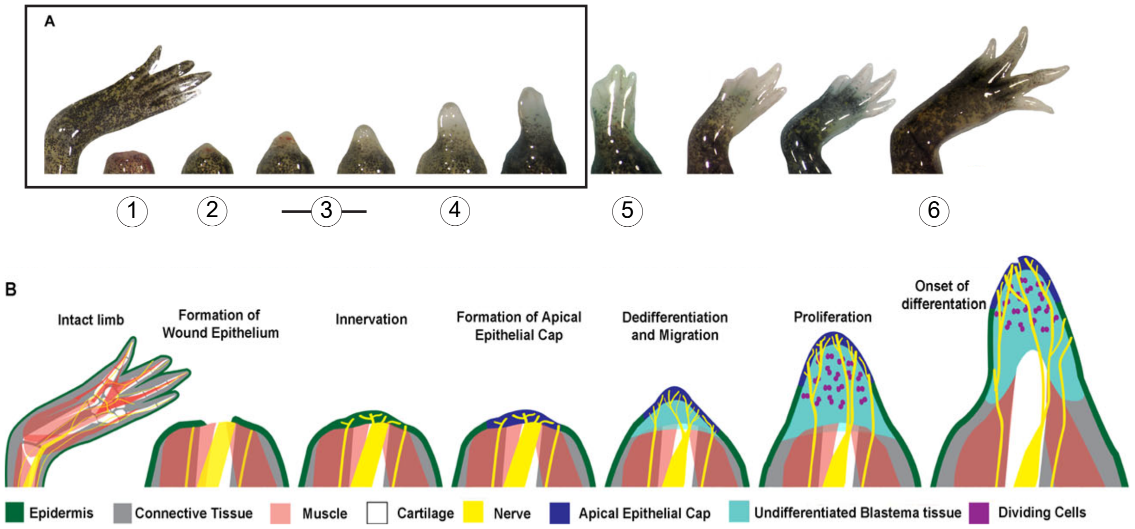


Figure 1.4: **A)** Following amputation the wound epithelium seals the wound site (1), followed by blastema formation (2) and proliferation (3). Blastema cells differentiate (4), tissues are patterned and continue to grow (5) until a perfect copy of the original limb is obtained (6). **B)** Cell type composition of the regenerating stump in the initial phases of the regeneration (inset corresponding to the area selected in **A**). Adapted from (C. McCusker, Bryant, and David M. Gardiner 2015).

Wound healing The initial healing phase is common to every kind of injury, and is shared by other animals too. It consists in the formation of the wound epithelium, which covers the affected area within hours from the injury, and derives from keratinocytes (C. McCusker, Bryant, and David M. Gardiner 2015; Stocum and Cameron 2011). This epithelium is histologically different from mature epidermis. It lacks of a collagen basal lamina, and can thus interact with other cells present in the stump. During this initial phase, migration of fibroblasts to the wound site can be observed, alongside their active proliferation (Endo, Bryant, and David M. Gardiner 2004). The initial proliferative response, observed also in myotubes and neural stem cells, has been associated with the release of MARCKS-like protein by the wound epithelium (Sugiura et al. 2016).

Blastema establishment When nerves are present at the injury site, the wound epithelium thickens and its interaction with the mesenchymal cells is impeded, interfering with scar-free healing and pushing the process towards regeneration (Makanae, Hirata, et al. 2013). This regeneration-specific structure is called apical epithelium cap (AEC) and is a structure essential for regeneration (Thornton 1957). The signalling molecules used by nerves and AEC are still to be completely characterized, but their interplay stimulates fibroblasts migration to the injury site and dedifferentiation in blastema cells, which then proliferate and accumulate in situ, while keratinocytes become non-proliferative and then dedifferentiate (C. McCusker, Bryant, and David M. Gardiner 2015). Fibroblast growth factors (FGF) signalling has been proven to regulate fibroblasts dedifferentiation (Makanae, Hirata, et al. 2013), while Msx1 seems to be involved in myotube-to-myoblast conversion (Antos and Tanaka 2010), but many other factors are potentially involved. For example, both the absence of nerves and the inhibition of macrophage signalling can impede regeneration (C. McCusker, Bryant, and David M. Gardiner 2015).

It was shown that dermal fibroblasts, cartilage and bone, Schwann cells, and muscle cells all contribute to blastema and regeneration. The blastema itself is thus a heterogeneous aggregation of proliferating dedifferentiated cells from these tissues, together with mesenchymal stem cells (Kragl and Tanaka 2009). Cells from different progenitors are localized in different subregions in the blastema, and have different differentiating potential (Kragl and Tanaka 2009; Eugen Nacu et al. 2013).

Limb re-development Once the blastema is established, a new limb with the same dimension and functionality of the amputated one is formed. In order to do that, genes are expressed in a pattern that is very similar to that responsible for the limb bud development at the embryo stage (Knapp et al. 2013). Lineage tracing studies have demonstrated that most cell types present in the blastema are programmed to differentiate to the same cell type they are derived from: epithelial cells, muscle cells, nerves, and Schwann cells de-differentiate during blastema formation, and differentiate again during the formation of the regenerated limb (Kragl and Tanaka 2009). However, for cells of the connective tissues (lateral plate mesoderm origin) the same is not true, and any of them can arise from a common progenitor in the blastema (Gerber et al. 2018). These progenitor cells express Prrx1, a gene that is expressed in the developing limb bud mesenchyme (Nohno et al. 1993) and is used as blastema connective tissue progenitor marker gene (Satoh, Makanae, et al. 2011).

For a correct patterning of the new limb, positional information brought by cells not implicated in the blastema but surrounding it, is necessary (Endo, Bryant, and David M. Gardiner 2004). In particular, it is necessary to have enough positional diversity for the blastema cells to proliferate and give rise to the missing part, gaining more positional identity going from basal to distal positions (C. McCusker, Bryant, and David M. Gardiner 2015). Some signalling pathways involved in the process of patterning have been identified, with retinoic acid inducing cells proximalization by the induction of the proximal determinant genes Meis1 and Meis2, and FGF inhibiting the action of retinoic acid signalling (Mercader et al. 2000).

Intensive research has gone in clarifying the mechanisms involved in each regeneration step. Nonetheless, many questions remain open or lack detailed understanding. Among them, there are two main themes. First of all, the signals secreted by nerves are still not completely characterized, and their role in multiple aspects of the induction, proliferation and differentiation of blastema cells is still incompletely understood (Farkas and James R. Monaghan 2017). Secondly, it has become evident in recent years that the immune system response after amputation is a major determinant in the differences observed between regenerating animals, which undergo scar-less healing and subsequent regeneration, and non-regenerating animals (James W. Godwin and N. Rosenthal 2014; James W Godwin, Pinto, and N. A. Rosenthal 2013). In this regard, reactive oxygen species (ROS) have been observed to induce the activation of cell death pathways required for the regeneration. However, mechanisms to counteract excessive ROS-induced cellular stress are expected to be in place, in order to protect blastema cells from dying. How a balance between the two requirements is achieved is still an open question (Love et al. 2013; Tseng et al. 2007; Wang et al. 2015; F. Zhang, R. Liu, and J. Zheng 2016). Additionally, the adaptive immune system has been shown to clean up the amputation site in the initial phases, but its role in later stages of the regeneration is still undefined (James W Godwin, Pinto, and N. A. Rosenthal 2013).

1.5 Summary

In summary, the ability to induce controlled cellular transitions is important in the context of disease modelling and treatment, and is necessary for the application of regenerative medicine to pathological conditions. Additionally, as cellular conversions are involved in complex processes such as the case of regeneration, it grants the possibility to understand and reproduce such biological mechanisms.

While computational models for the prediction of transcriptional determinants of cellular states exist, no such method is available to date for the prediction of signalling determinants. Acting on the signalling network with chemical compounds, small molecules or growth factors is particularly interesting because their use is more suited for therapeutical applications and is safer compared to TF-based experimental strategies. The role of the signalling network on cellular state is so far analysed independently of the GRN that determines such state. Thus, this thesis focuses on integrating the signalling and transcriptional regulatory layer in an ensemble approach for the induction of cellular transitions.

2 Aims and scope of thesis

The controlled induction of cellular transitions is of interest for disease treatment, regenerative medicine, and disease modelling. Recently, the possibility of inducing such transitions by acting on the signalling network garnered more attention, but computational methods for directing the identification of efficient protocols are needed to ease the temporal and financial burden of experimental trial-and-error discovery. This thesis proposes predictive methods that use gene expression data to predict signalling perturbations that can induce cellular transitions.

The aims of this work are:

- Devise a general approach for the integration of the signalling and transcriptional regulatory networks, in order to model the effect of signalling molecules and pathways on GRNs. The two regulatory layers will be modelled separately to take into account their specificities, and then integrated at the level of interface TFs, which are TFs that are regulated by signalling and regulate the GRN. The shift between cellular states will be represented by a transition-specific GRN, while signalling events will be modelled following a probabilistic approach, in order to account for the inherent stochastic nature of signal transduction and the uncertainty associated with use of gene expression data as a measure of signalling activity.

- Develop predictive methods for cellular transitions. Prioritize signalling molecules according to their potential for inducing the desired changes in the GRN, giving particular importance to the specificity of their action. Additionally, the predicted molecules will be associated with signalling pathways to facilitate the interpretation of results in the context of novel cellular transitions.

- Systematically evaluate the methods on large numbers of datasets and consistently compare predictions with literature knowledge to assess their relevance to the cellular transitions analysed. The performance of the methods presented will also be benchmarked against available tools for the prediction of both signalling molecules and pathways.

- Apply the methods developed to the analysis of animal models of disease and regeneration. Firstly, this integrative approach will be applied to the reversal of a pathological condition such as cirrhosis, verifying that the candidate signalling perturbations act on the GRN in predicted ways. Then, it will be used to identify signalling events involved in salamander limb regeneration, a complex phenomenon which requires cellular transitions at multiple stages.

2.1 Originality

The work presented in this thesis aims at identifying any signalling perturbations that can induce the transition between an initial and desired cellular state by triggering a shift in the gene expression program of the cells. This makes this work conceptually different from previous studies that aim at describing which signalling pathways or molecules are involved in the observed cellular response to stimuli.

The effective integration of signal transduction and transcriptional regulation allows to model the action of signalling pathways on an underlying GRN. This goal is achieved by combining available approaches for GRN and signalling pathways analysis in a new framework, where the signalling and transcriptional regulatory mechanisms are brought together and integrated at the level of interface TFs, thus permitting the selection of the most suitable perturbations for the induction of a desired cellular transition.

From the methodological point of view, while computational methods that consider either GRN or signalling network models are available, the number of approaches trying to integrate the two regulatory levels is limited. These attempts fall under two main categories: either they rely on a vast number of data samples, or on manual curation of the integrated networks in order to accurately capture the effect of signalling cues on gene expression, as described in Section 1.2.2. The approach presented here, on the contrary, has limited data requirements and uses widely available transcriptomics data to ensure general applicability and flexibility.

3 Materials and Methods

The main product of this thesis is an approach for the prediction of signalling perturbations that can induce cellular conversions. Signalling pathways are used to transmit information from the outside of a cell to the inside in order for the cell to respond to it. Signal transduction takes place by the interaction of proteins and their exchange of information in the form of phosphorylation and other covalent modifications, formation of complexes, modulation of protein localization and stability (M. J. Lee and Yaffe 2016). The signal finally reaches transcription factors activating or inhibiting them, and therefore exerting control on the cellular gene expression program.

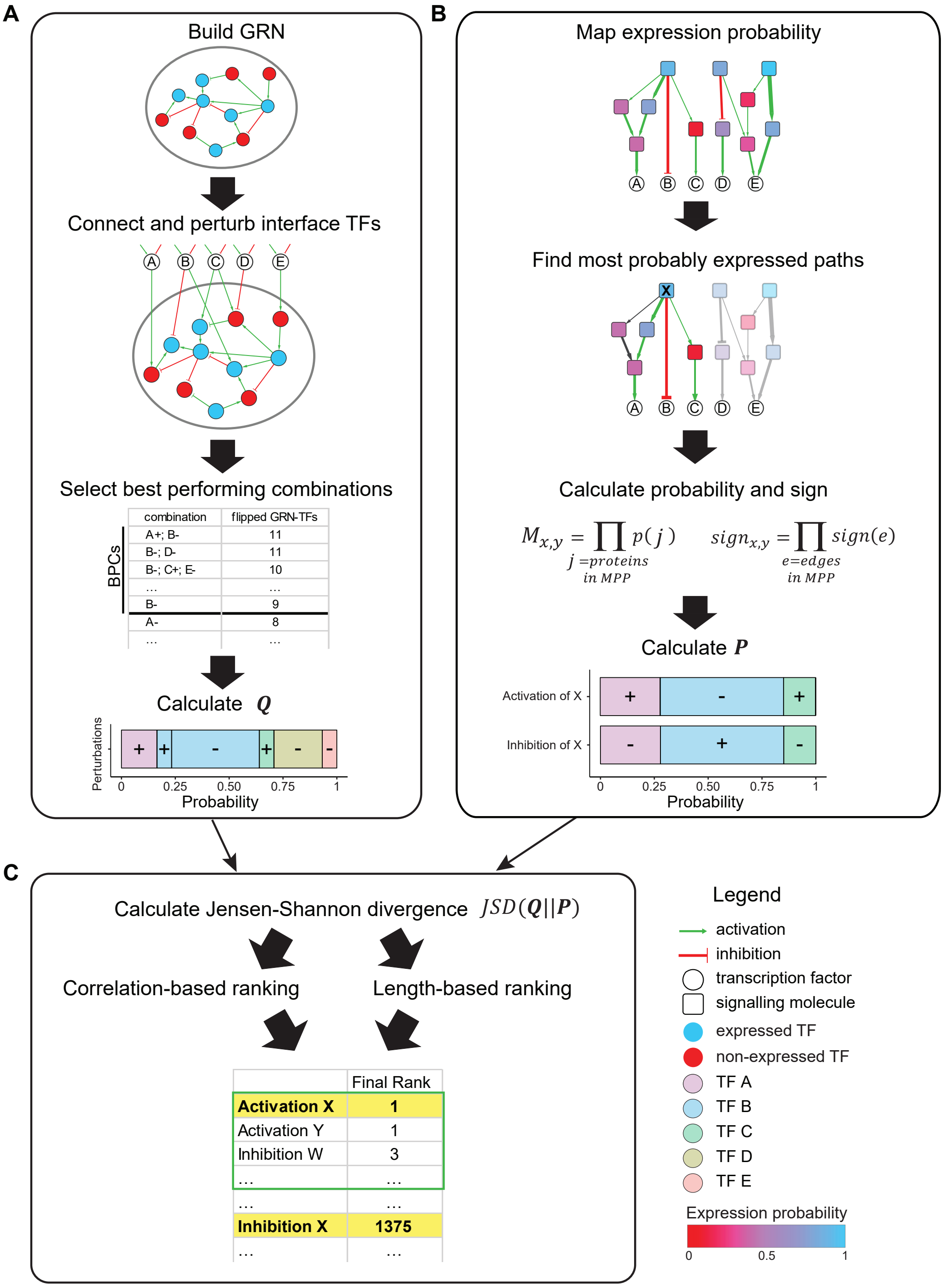
The proposed framework consists of two regulatory layers: the GRN reflecting the interplay of TFs involved in the cellular transition regulating each other's expression, and the signalling network containing all interactions used for canonical signal transduction. The GRN is modelled as a Boolean network; the signalling network is modelled by a probabilistic approach. The difference in modelling approaches is necessary because of the different levels of uncertainty associated with the use of gene expression data in order to capture different cellular processes. mRNA abundance has been successfully used to represent transcriptional activity (H.-C. Chen et al. 2004), and in particular it is widely accepted that a Boolean representation of TFs activity is able to capture the overall state of large regulatory networks (S. Huang 1999). The use of mRNA abundance as a proxy for protein signalling activity however is less straightforward, as gene expression is only weakly correlated with protein abundance (Pines et al. 2011), and this is in turn not linearly correlated with signalling activity (Richter et al. 2015). Additional uncertainty is present in signal transduction modelling because of its stochastic nature, as the physical interaction of the correct proteins in the cytoplasm against a background of unspecific protein-protein interactions defines successful transduction (Ladbury and Arold 2012).

The two network models communicate through a layer of TFs that are directly regulated by canonical signalling pathways, and can regulate other TFs in the GRN. These TFs are here termed *interface TFs*, and represent the point where any signal coming from the exterior of the cell is de-coded and exerts its action on the cell, by affecting its gene expression program.

3.1 Prediction of signalling molecules for cellular transition

In order to predict which signalling perturbations are able to trigger a desired cellular transition, a method is developed that identifies which interface TFs can best induce the GRN changes required for the cellular transition desired, and independently estimates the effect of signalling molecules on the interface TFs. This two pieces of information are then combined to prioritize signalling molecules (including plasma membrane receptors, intermediate signalling proteins, small molecules used in canonical pathways) that specifically act on the most effective interface TFs, limiting un-specific effects. This first predictive method is called INCanTeSIMO (Integrated Network approach for Cellular Transitions through SIgnalling MOlecules).

INCanTeSIMO requires gene expression profiles of the initial and desired states, without relying on a high number of replicates or time-series measurements. It is also independent from other type of data, such as protein abundance or phosphoproteomics data. This allows its application to any cellular conversion, including ones that have not been achieved experimentally.



combination	flipped GRN-TFs
A+; B-	11
B-; D-	11
B-; C+; E-	10
…	…
B-	9
A-	8
…	…

$$M_{x,y} = \prod_{\substack{j = proteins \\ in\ MPP}} p(j) \qquad sign_{x,y} = \prod_{\substack{e = edges \\ in\ MPP}} sign(e)$$

$JSD(\boldsymbol{Q}||\boldsymbol{P})$

	Final Rank
Activation X	**1**
Activation Y	1
Inhibition W	3
…	…
…	…
Inhibition X	**1375**
…	…

Figure 3.1: Overview of the method for the prediction of signalling molecules inducing cellular transitions. **(A)** The Boolean GRN containing TFs changing their state during the cellular transition is connected to interface TFs. All combinations of up to four interface TFs are assigned a fixed state and the effect of these perturbations on the initial state is assessed. Combinations having the top three flipping scores (including ties) are selected as the best performing combinations (BPCs). The frequency of each interface TF state (activated +, inhibited –) across the BPCs is used to calculate the probability distribution Q of each TF state of causing GRN state changes. **(B)** The gene expression probability of each protein is mapped onto the signalling network and defines the signalling interactions probability. The most probably expressed paths (MPPs) connecting each signalling molecule X to each interface TF are selected. MPPs probability and sign are used to estimate the probability distribution P of successfully activating or inhibiting the interface TFs by activating or inhibiting X. Correlation-based and length-based probabilities are calculated for each couple of signalling molecule and interface TF (see Methods). **(C)** The probability distributions Q and P are compared using Jensen-Shannon divergence (JSD). Perturbations of each signalling molecule are ranked according to this measure. The best ranking that each molecule perturbation obtains across the correlation-based and length-based rankings is used as its overall rank. Finally, molecule perturbations ranking in the first fraction of the final ranking are selected as candidates for the cellular transition.

The method comprises three steps. Initially, the expression state of TFs is Booleanized and used to represent their activity. TFs showing differential activity in the initial and desired cellular state are connected by transcriptional interactions from previous knowledge, giving a transition-specific Boolean GRN.The two cellular states are modelled as separate point state attractors of the same Boolean network. In this framework, experimental induction of the cellular conversion corresponds to the induction of the transition from the initial to the final state attractor, which must be achieved by regulating the state of interface TFs. To select the most effective interface TFs for this scope, exhaustive perturbations are performed in silico (Figure 3.1 A).

Independently, the effect of activating or inhibiting each molecule of the signalling network on the downstream interface TFs is estimated. The minimal condition for signalling interactions to occur is the presence of the proteins involved in the interaction. The protein availability can be estimated from gene expression data, as previously done by other methods (Efroni, Schaefer, and Buetow 2007; Sebastian-Leon et al. 2014). The probability of signal to travel from a signalling molecule to interface TFs is the probability of all the molecules in the path connecting them to be present at the same time. In particular, we assume that the signal will travel along the most probable path available (Figure 3.1 B). Additional assumptions are:

a) genes correlated in gene expression are more likely working together as a signalling functional unit (R. Huang, Wallqvist, and Covell 2006), thus the probability of their interaction is higher;

b) the length of the path influences its probability, because the longer a signalling path, the more chances of off-target interactions (cross-talk) taking place.

These criteria are taken into account in calculating the probability of the most probable path connecting a signalling molecule to its downstream targets, while the sign of the effect is given by the sign of the interactions present in the path (Figure 3.1 B).

Finally, for each signalling molecule the Jensen-Shannon divergence is calculated between the probability of acting on interface TFs and the likelihood of the interface TFs to induce the required GRN state (Figure 3.1 C). Signalling molecules are ranked according to the divergence, which privileges signalling molecules acting specifically on effective interface TFs compared to molecules with unspecific effects.

3.1.1 Gene expression state and probability

All the datasets analyzed in this thesis contain gene expression data obtained by microarray. For microarray generated on popular platforms, it is possible to apply frozen normalization approaches and estimate the expression probability using a background distribution (a strategy here referred to as *frmaBool*). Raw .CEL files were normalized with frozen-RMA (R package *fRMA* (Matthew N. McCall, Bolstad, and Rafael A. Irizarry 2010)) and assigned an expression state using Gene Expression Barcode (Matthew N McCall, Jaffee, et al. 2014; Matthew N McCall, Uppal, et al. 2011). The barcode approach assumes that normalized and log-transformed expression values subject to all possible conditions, cell types, tissues and other biological variables, follow a distribution specific for each probeset in a microarray. This distribution is approximated as a mixture model of a Gaussian distribution modelling values corresponding to non-expression, and a uniform distribution corresponding to expression (Matthew N McCall, Uppal, et al. 2011). The assumption of expressed values following a uniform distribution was developed in the analysis of cancer samples, where disruption of regulation causes the expression values to vary across wide ranges without necessarily informing a biological phenotype (Parmigiani et al. 2002). In this work, the aim is to differentiate between expression values that correspond to cross-hybridization, and values that represent real mRNA presence, so the shape of the expressed distribution was kept as simple as possible. The uniform distribution was considered as it does not require parameters and is defined as spanning form the mean value of the non-expressed distribution to the value 15, which is the maximum value obtained after fRMA normalization. The non-expressed distribution for each probeset in each microarray platform, on the other hand, is supported by two kinds of evidence:

a) *S. cerevisiae* mRNA: the intensity observed when non-specific transcripts hybridize with the probes is considered as a measure of the cross-hybridization, and therefore belongs to the non-expressed distribution (Matthew N McCall, Uppal, et al. 2011).

b) negative controls: it is assumed that each gene will be expressed only in a fraction of the conditions in which its expression is measured (e.g. only in some tissues or cell types), so its overall distributions will present multiple modes, and the lowest intensity mode is expected to correspond to lack of expression (Zilliox and Rafael A Irizarry 2007).

For each gene, the probeset with highest variance was selected. TFs were assigned Boolean state 1 (expressed) if their assigned value had probability lower than 0.05 of belonging to the non-expressed distribution $N(\mu, \sigma)$, and state 0 otherwise. The values μ and σ are made available as R Bioconductor packages for microarray platforms Affymetrix Human Genome U133A, Affymetrix Human Genome U133 Plus 2.0, Affymetrix Human Genome U133A 2.0, Affymetrix Human Gene 1.0 ST Array, Affymetrix Mouse Gene 1.0 ST Array and Affymetrix Mouse Genome 430 2.0 Array (Matthew N McCall, Jaffee, et al. 2014).

In this work, an expression probability value was calculated for each gene based on its most variable probeset x, as follows:

$$p(x) = \frac{\frac{1}{2} f_{e(x)}}{\frac{1}{2} f_{e(x)} + \frac{1}{2} f_{n(x)}}$$

where f_e is the probability density function (pdf) of x in the uniform distribution $U(\sigma, 15)$, and f_n is the pdf of x in $N(\mu, \sigma)$. Each gene was assigned the maximum expression probability obtained across all replicates of the initial cellular state.

Intuitively, this formulation assigns expression probability 0 to any expression value $\leq \mu$, which is the most frequent non-expressed value, and probability 1 to any value that does not belong to the non-expressed distribution, independently of the actual value. This agrees with the idea of not assuming that gene expression levels correspond to protein abundance levels, a correlation that has been proven non trivial (Edfors et al. 2016; Washburn et al. 2003).

To generalize the previous method to other microarray platforms and potentially RNA-seq data, colleagues in the Computational Biology group developed a background-independent approach to calculate expression probability, termed *geneDE*. Among the differentially expressed genes (DEGs) between two cellular states, some genes can be expected to pass from the expressed to the non-expressed state, or vice versa. Therefore, the minimum values and maximum values observed for each DEG were collected in two separate distributions. These empirical expressed and not-expressed distributions were then used to calculate the expression probability according to the formula:

$$p(x) = \frac{\frac{1}{2} f_{max}(x)}{\frac{1}{2} f_{max}(x) + \frac{1}{2}(1 - f_{min}(x))}$$

where $f_{max}(x)$ is the empirical cumulative density function (ecdf) of x in the maximum values distribution and $f_{min}(x)$ is the ecdf of x in the minimum values distribution. Because some DEGs undergo a change in their expression levels but do not change their state, the two distributions overlap. The intersection point between the two distributions is used as a threshold to assign Boolean state: the TFs with average expression value across replicates lower than the threshold were assigned Boolean state 0, and 1 otherwise.

3.1.2 GRN reconstruction

TFs that were assigned opposite Boolean state in the initial and final gene expression profile were assumed to have differential activity in the two states. They were used as seed for the GRN representing the transition between the two phenotypes. Transcriptional regulatory interactions among human TFs and other transcriptional regulators as defined in Animal TFDB 2.0 (L. Zhang, Ng, and S. Li 2015) were collected from MetaCore from Clarivate Analytics in March 2017. They correspond to interactions labelled with the mechanisms “Transcriptional regulation”, “Influence on Expression” and “Regulation”; their effect was “Activation”, “Inhibition” or “Unspecified”.

The literature-supported interactions among the selected TFs were collected in an initial network, which was then optimized using a previously developed method (Crespo et al. 2013). Briefly, this algorithm assumes that the cellular states are represented by separate point attractors in the Boolean network state space. A genetic algorithm that randomly removes interactions and assigns signs to interactions with unspecified effect is used to make the attractors state of the network compatible with the gene expression profiles. The simulation of the network state follows a synchronous update scheme, where all nodes of the network are updated at each simulation step. The update of each node follows a majority rule, meaning that the state of the node at the following step is defined by the summation of all active inhibitions and activations that it receives. The fitness of the resulting network is assessed by the similarity of the point attractors obtained with the initial and final Booleanized gene expression state.

A network solution was chosen randomly among the ones that maintain the more edges and have the least mismatches between simulated data (attractors) and experimental data (Booleanized gene expression data). Datasets where the resulting GRN contained less than 10 connected TFs were excluded from the study, because they were usually associated with perturbations not targeting signalling molecules.

3.1.3 Likelihood of interface TFs to induce the required gene expression changes

Signalling cues exert an effect on the GRN by controlling the activity of TFs that belong to signalling pathways and therefore act as signalling effectors. Because signal transduction is a fast process, acting in a matter of seconds (Kanshin et al. 2015), but the production of TF protein from the DNA requires dozens of minutes (Shamir et al. 2016), we consider as available signalling effectors only TFs that are already expressed in the initial cellular state. In the same vein, the signal will travel along paths composed of proteins already present in the cell when the cue is applied. Therefore, interface TFs are here defined as TFs that are expressed at the initial time point, belong to canonical pathways which are expressed (there is a path composed of expressed genes connecting the TF to any 0-indegree node of the pathway), and can regulate TFs present in the GRN (*GRN-TFs*).

Interface TFs are the only way in which signalling cues can act on the GRN in this model, and their likelihood of inducing the desired changes in the GRN depends on their activity state and on the GRN itself. Usually, the expression of an interface TFs does not imply it is also active, as interface TFs are directly regulated by signalling pathways. Only when an interface TF was initially expressed, but is not expressed in the desired cellular state, we assume that its expression was associated with activity. In this case, the interface TF is part of the GRN, and the only perturbation that can be applied to it is inhibition. For all other interface TFs, both activation and inhibition are in principle possible.

To estimate the effect of each interface TF state on the GRN, in silico simulations were performed, starting from the initial cell Boolean state. The state of interface TFs singularly or in combinations of up to four was fixed, and the state of the GRN updated synchronously following a majority logic rule, until convergence to a fixed-point attractor state. The perturbations were ranked by flipping score, which was defined as the number of GRN-TFs that assume the final cell Boolean state after simulation. Datasets for which the best flipping score did not represent the 40% of the GRN-TFs were discarded from further analysis.

In order to prioritize signalling molecules that induce the most effective GRN perturbations, combinations of interface TFs not showing any synergistic effects, meaning that the GRN-TFs they affected were the same as the ones affected by the separate components of the combination, were discarded. In this way, the scoring of signalling molecules favors more specific candidates, because the molecules are required to act on few effective interface TFs. The combinations of interface TFs obtaining the three best flipping scores, including ties, were selected (best performing combinations, *BPCs*).

3.1.4 Most probable paths calculation

75 canonical signalling pathways were selected from MetaCore in July 2017, and merged together in a single signalling network. All interactions reported in MetaCore are the result expert manual curation of full text papers from literature. The nodes present in the network represent signalling entities (single proteins or functional complexes), and the interactions are directed and signed when possible. Edges corresponding to “Technical” or “Unspecified” effect, and “Technical”, “Transcriptional Regulation”, “Influence on Expression”, “Catalysis” and “Transport” mechanisms were removed. Further manual curation was also applied to interactions involving TFs known to act as complexes but represented in MetaCore as separate functional nodes. Edges that were associated with literature not supporting specifically the direct interaction between a TF and a second molecule were removed (Tables in Appendix 7.1 and 7.2). The final signalling network contained 2496 nodes and 6876 edges. Half of the nodes belong to more than one pathway, showing the high frequency of cross-talk events taking place (Figure 3.2).

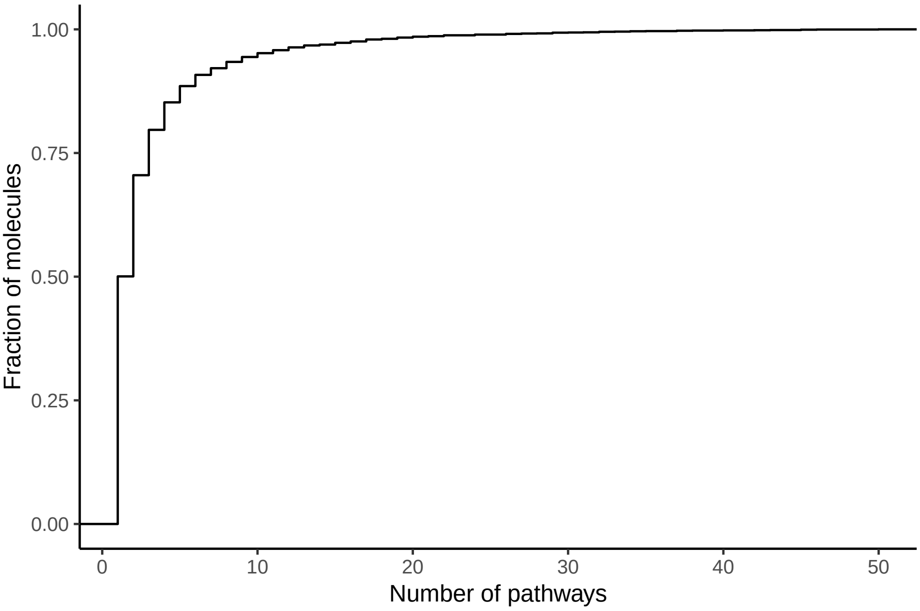


Figure 3.2: Overview on the number of canonical pathways to which each signalling molecules belong. 50% of the molecules only belong to one pathway, but others belong to dozens of them.

Signal transduction along any path requires the presence of the proteins that form it, which is itself dependent on the expression of the corresponding genes. Therefore, the paths followed by signalling cascades are assumed to depend on the availability of the corresponding proteins, and the overall effect of a signalling molecule on interface TFs is approximated by the most probably expressed path (*MPP*) connecting them. Only directed paths were considered, meaning that all the interactions follow the same direction, from molecule to interface TF.

The probability of a path between signalling molecule x and interface TF y is defined as the probability of all the components of the path being expressed, therefore it is the product of the expression probability of each of the genes in the path. The probability of the MPP $M_{x,y}$ is defined as:

$$M_{x,y} = p\left(MPP_{x \to y}\right) = \max p(x \to \cdots \to y)$$

where $p(x \to \cdots \to y) = \Pi_j p(j)$ and $p(j)$ is the expression probability of the intermediate signalling molecule j in the path connecting x to y.

Two variations of this approach were considered:

- The longer the path connecting two nodes in the network, the higher is chances that unspecific or non functional interactions occur. Therefore, the probability of a path was corrected by its length (measured by the number of interactions l present in the path) by multiplying it by a factor $e^{(-l)}$ (Jaeger et al. 2014), so that the longer the path, the lower its probability. The resulting probability was re-normalized across all interface TFs.

- Independently, co-expression has been observed among members of the same functional modules (R. Huang, Wallqvist, and Covell 2006). Assuming that genes with the same functional properties have common transcriptional regulation, the probability of interactions between genes that are co-expressed across CMap datasets (absolute Pearson correlation $>$0.7 and sign of the correlation matching the sign of the interaction) was increased. MPPs were calculated using these modified probabilities.

These two corrections were applied independently to the MPP discovery (see Supplementary Information and Figure S4). The distribution of the relative probability with which the same signalling molecule x acts on all interface TFs ($\boldsymbol{P}_x$) was obtained by:

$$P_{x,y} = \frac{M_{x,y}}{\sum_{i \in \text{interfaceTFs}} M_{x,i}}$$

Finally, the sign of each MPP is calculated as:

$$\text{sign}_{x,y} = \prod_{e=edges \in MPP} \text{sign}(e)$$

If the sign is positive, activating (inhibiting) x results in the activation (inhibition) of TF y with probability $P_(x,y)$. If the sign is negative, activating (inhibiting) of x results in inhibition (activation) of y with the same probability. Because the MPPs are obtained separately with correlation-corrected and length-corrected probabilities, two $\boldsymbol{P}_x$ and signs exist for each molecule x.

3.1.5 Ranking of signalling molecule perturbations

The probability of each signalling molecule to reach each interface TFs is compared to the likelihood of the interface TFs to induce the desired changes in the GRN state. A high frequency of a specific activated or inhibited interface TFs among the BPCs suggests that it consistently alters the state of the GRN, thus for each Boolean state of each interface TF (TFs-state pair s), the frequency F_s in all BPCs was calculated as:

$$F_s = \frac{k_s}{k}$$

where k is the total number of BPCs and k_s is the number of combinations where s is present. The frequencies were normalized across all TF-state pairs, resulting in the probability distribution $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \frac{F_i}{\sum_{i \in F} F_i}$$

This distribution is compared to the probability of each signalling molecule to act on the interface TFs $\boldsymbol{P}_x$ by Jensen-Shannon divergence:

$$JSD\left(P_x \| Q\right) = \frac{1}{2} D\left(P_x \| M\right) + \frac{1}{2} D(Q \| M)$$

where $M = \frac{1}{2}\left(P_x + Q\right)$ and $D(X \| Y) = \sum_i X(i) \log \frac{X(i)}{Y(i)}$ (Kullback-Leibler divergence). Both the activation and inhibition of each signalling molecule are considered. Their divergence score differ because they have opposite effects on the downstream interface TFs and thus different effectiveness in changing the GRN state. If for example, the activation of molecule x activates TF y, which is frequently present in BPCs, the divergence score will be low. Inhibition of x, however, inhibits TF y, which might not be present in the BPCs. Thus, the probability of reaching TF y is the same, but the resulting ranking is different for the two perturbations applied to molecule x.

The signalling molecules were sorted by their divergence, where lower divergence corresponds to better ranking, and assigned rank $R(x)$:

$$R(x) = \min \mathrm{rank}_{v \in P_x}(x, v)$$

where $\mathrm{rank}(x, v)$ is the rank obtained by the molecule x in the variant v of the MPP calculation, using either correlation- or length- corrected probabilities.

Different cut-offs, defined as fractions of the maximum R value present in the final ranking, were considered (from 1 to 10%). Signalling molecules whose rank was lower than the cutoff were selected as candidates for the induction of the cellular transition considered. The prediction was considered successful if at least one of direct targets of the experimental perturbation was selected among the candidates, similar to (K. Chen and Keaney 2012). The optimal cut-off was selected as the one having maximum improvement compared to random chance of success across the datasets from CMap. This was calculated as the probability of selecting at least one perturbation target in a randomly chosen list of the same size, by using one-sided hypergeometric test.

3.1.6 Functional measures

The list of candidate signalling molecules was tested for functional and topological features. First, the enrichment in Gene Ontology (GO) biological process terms was calculated for both candidate and non-candidate signalling molecules with the use of R package *gProfilerR*. The GO terms associated to the targets of the experimental perturbations were collected, and their overlap with enriched terms in the candidate and non-candidate molecules list was calculated across datasets. The distributions obtained were compared by one-sided Wilcoxon test (p-value $<$.05).

The distance of candidate signalling molecules from the direct targets of experimental perturbations was calculated along the signalling network, taking into account the directionality of the interactions. Each candidate was assigned an out-distance and an in-distance, depending if the perturbation targets were reached by following edges downstream or upstream. The smallest of the two distances was then selected, and the average distance of each candidate to all reachable perturbation targets was calculated. The average distance from perturbation targets was calculated in the same way for non-candidate molecules. The distributions of average distances for candidate and non-candidate signalling molecules were compared by Wilcoxon test with 100’000 Monte Carlo replicates (p-value $<.05$).

3.2 Prediction of signalling pathways for cellular transitions

In the case of novel systems or transitions, interpretation of the list of single molecule candidates can be complicated. As development, differentiation and other biological processes are driven by the concerted action of signalling pathways, one could argue that the prediction of canonical pathways is particularly relevant for these applications. The majority of the existing pathways analysis methods look at differential activity state between two cellular contexts, while only very few search for pathways that mediate the transition between these contexts (Y.-A. Kim, Wuchty, and Przytycka 2011; Melas et al. 2015; Paull et al. 2013). These methods aim at finding the most suitable network to connect a causal set of genes, such as disease or mutated genes and known drug targets, to a target set, composed of DEGs. The limited available information on the paths effectively used means that the evaluation of these methods is done by comparison of the causal networks obtained with canonical pathways. We therefore extended our method from the prediction of single signalling molecules, to the prediction of the activation or inhibition of canonical pathways that can drive cellular transition. This extension is here termed *INCAnTeSIMO_path*.

3.2.1 Pathways prediction

For each signalling pathway present in the signalling network, the inhibitors of the pathway were identified as the nodes that have both incoming and outgoing inhibitory interactions. This is because in the canonical pathways considered inhibitors are associated with their own inhibitors, so that the overall outcome of the pathway is activatory. Then, two separate signalling gene sets were prepared for each pathway: one activatory set with all active effectors of the pathway and the inhibition of its inhibitors; and one inhibitory set containing the inhibition of its effectors and the activation of its inhibitors.

As the candidates were ranked according to their specificity in inducing the desired GRN perturbations, it is not expected that all signalling molecules belonging to a pathway perform equally. However, if molecules that are particularly influential in a pathway are predicted, the whole pathway can be expected to induce similar effects on the GRN and therefore induce the same cellular transitions. Network centrality measures express how important particular nodes in a graph are to its connectivity, and multiple definitions are used (Newman 2010).

Here, the signalling gene sets were tested for significance in the list of candidate signalling molecules according to the concept of source/sink centrality (Naderi Yeganeh and Mostafavi 2019). Source/sink centrality is a topological property of network nodes that depends on the number of directed paths inside a signalling pathway that are incoming and outgoing of each node, and the length of such paths. It is calculated by (Naderi Yeganeh and Mostafavi 2019):

$$C_{SSC}(v) = C_{source}(v) + \beta C_{sink}(v) = \sum_{w_j\text{:vu-walk of G}} \alpha^{|w_j|} + \beta \sum_{w_j\text{:uv-walk of G}} \alpha^{|w_j|}$$

where v is the node in the pathway graph G considered, u is any other node in G, w_j is an incoming or outgoing path connecting v to u and $|w_j|$ is its length, α is a dampening factor that decreases the contribution of longer paths to the overall measure (here α=0.1), and β indicates the relative importance of the source and sink components (here β=1) in the centrality score. $C_{SSC}(v)$ represents the importance of the node v as sender or receiver of information in the pathway considered.

In the original implementation, an enrichment score is obtained by calculating the aggregated importance of DEGs, and a statistical significance is assigned using a bootstrap sampling approach. Here, the aggregated importance of signalling molecule candidates obtained from INCAnTeSIMO U is calculated by:

$$Agg(U) = \prod_{u_i \in U} C_{SSC}(u_i)$$

The probability of observing a higher aggregate score for a randomly selected subset of

G is used as the p-value for each pathway.

3.3 Comparison with existing methods

The methods proposed here were compared with computational approaches developed previously to predict signalling molecules or pathways related to the differences between two cellular states. None of them was specifically developed to identify signalling perturbations to induce the conversion between an initial and final cellular state. CMap (Subramanian, Narayan, et al. 2017) and DeMAND (Woo et al. 2015) are intended for the prediction of perturbations on single signalling molecules, while pathway enrichment methods considered were over-representation analysis, GSEA (Subramanian, Tamayo, et al. 2005), SPIA (Tarca et al. 2009) and CADIA (Naderi Yeganeh and Mostafavi 2019).

3.3.1 Single molecule prediction

Initially, differential expression analysis was compared with INCAnTeSIMO. Differential gene expression between the initial and final expression profiles was calculated with the R package *limma*. Genes showing absolute log fold change (lfc) $> log_2(1.5)$ and having Benjamini-Hochberg (BH)-adjusted p-value $<$.05 were considered differentially expressed. If replicates were not present, the lfc cut-off alone was applied. Signalling molecules were ranked according to decreasing absolute lfc values to generate their ranking by differential expression.

Connectivity map (CMap) (Subramanian, Narayan, et al. 2017) allows to query a gene expression signature against a database composed of chemical and genetic perturbations applied to different cellular types. It uses the ranking of genes according to log fold change to match known perturbation profiles to the query, which are then returned as predictions. DEGs were selected (BH-adjusted value$<$.05 or absolute lfc $> log_2(1.5)$ if replicates were missing). DEGs were sorted separately for up- or down-regulation by decreasing absolute lfc, and up to 150 genes for each class were submitted to the Batch query functionality of CMap L1000 query (Subramanian, Narayan, et al. 2017)), using the option sig_fastgutc_tool. CMap perturbations were considered correct if any of the direct experimental perturbation, its direct targets, or alternative drugs acting on the same targets were assigned a connectivity score (tau)$>$90 in the summary results across all cell lines. Datasets with less than 10 DEGs or raising errors during submission were discarded.

DeMAND (Woo et al. 2015) is a method for predicting genes associated with the mode of action of a drug or compound. It scores each gene based on how significantly the expression of its targets is dysregulated following the application of a drug. It relies on context-specific regulatory networks representing both direct and indirect transcriptional regulation. Given its requirements for numerous replicates, DeMAND could not be applied to datasets used to evaluate the proposed method, so the opposite strategy was followed. The method presented here was applied to the GEO13 datasets present in the original DeMAND study (Woo et al. 2015). Only Affymetrix datasets were considered, using either the barcode strategy presented in Section 3.1.3, or the MAS5.0 detection call approach described later in Section 3.4. Genes assigned FDR$\leq$.1 were considered predicted. Successful predictions were defined as the ones that recover perturbation direct targets as determined from STITCH, DrugBank and the original study.

3.3.2 Pathway prediction methods

Pathway enrichment in DEGs tests if the genes associated to a particular pathway are over-represented in the list of DEGs. Gene set enrichment analysis (GSEA) (Subramanian, Tamayo, et al. 2005) calculates a pathway-level score by considering not only if DEGs belong to a pathway, but also their log fold change. Therefore, it tests if the genes belonging to a pathway are enriched in the ordered list of DEGs, either at the lowest lfc (meaning among the down-regulated genes), or in the highest lfc values (meaning they are up-regulated). MetaCore pathway enrichment was calculated by one-sided hypergeometric test (BH-adjusted p-value$<$.05). Pathway enrichment and GSEA were applied to the signed signalling gene sets corresponding to MetaCore signalling pathways defined in Section 3.2.1 using the R package *clusterProfiler* (Yu et al. 2012).

SPIA (Tarca et al. 2009) analysis combines classical enrichment analysis with expression changes that are propagated across the topology of KEGG pathways. DEGs were selected

(BH-adjusted value$<$.05 or absolute lfc$>$log2(1.5) if replicates were missing). KEGG signalling pathways were scored using R package *SPIA*. If any of the significant pathways contained direct targets of the perturbation, the prediction was considered successful. Datasets where no DEGs were found were discarded.

Causal disturbance analysis (CADIA) (Naderi Yeganeh and Mostafavi 2019) considers not only the topology of KEGG pathways, but also the directionality of interactions. In it, the concept of source/sink centrality is combined with classic pathway over-representation analysis (ORA) to derive a pathway enrichment score based on DEGs. The R package *CADIA* was used to calculate a causal disturbance score (cadia) for each pathway, using parameters α=0.1 and β=1. Pathways with cadia$\leq$0.05 were considered as candidate pathways. Pathways enriched in aggregate importance of the DEGs were also selected ($P_{SSC} \leq 0.05$), and are referred to as SSC DEG.

INCAnTeSIMO_path was compared also with alternative approaches using INCAnTeSIMO candidate molecules to predict signalling pathways, namely CADIA and over-representation analysis (ORA) on MetaCore pathways. Regarding ORA, each signalling gene set was tested for enrichment in the single molecule candidates against a background composed of all the signalling molecules connected to the GRN. Then, the gene sets was tested for enrichment in the single molecule candidates against a background composed of all the molecules in the complete signalling network. The two backgrounds can vary significantly depending on the GRN, so this ensures that the over-represented signalling gene sets are as a whole able to act on the GRN, and their components can induce the desired cell fate conversion. Both tests were performed using one-sided Fisher exact test with FDR correction, and gene sets enriched for both tests (FDR$\leq$0.05) were retained. In CADIA, the original method proposed in (Naderi Yeganeh and Mostafavi 2019) was adapted to use candidate molecules instead of DEGs and predict signed MetaCore pathways instead of KEGG pathways (cadia score$\leq$0.05).

3.4 Application to a cirrhotic animal model

Expression data for healthy liver of male Wistar rats was extracted from GEO dataset GSE71201. Cirrhosis was induced in 10 male Wistar rats by exposure to inhalation of CCl_4, as previously described (Tsuchida and Scott L. Friedman 2017) and in accordance to the criteria of the investigation and ethics committee of the Hospital Clínic Universitari and the University of Barcelona. Five cirrhotic rats were treated with 10 mg/kg of CVX-060 (Pfizer, Inc., New York, NY, USA) diluted in 500 μl of saline solution and injected intravenously via the tail vein, once a week for 4 weeks. Gene expression of CCl_4 and CCl_4+CVX-060 -treated livers was obtained with microarray (Affymetrix GeneChip Rat Genome 230.2 Array). These experimental procedures were performed by collaborators at the Hospital Clínic in Barcelona. Two replicates for each treatment were kept after quality control and PCA visualization.

For the prediction of signalling molecules, gene expression data of the cirrhotic and healthy liver was used. Gene expression probability was assigned to each gene according to $1 - p$, p being the p-value obtained from Affimetrix MAS5.0 detection call (Affymetrix 2002). If the expression probability $\geq$0.94 (corresponding to a “marginal” or “present” call from MAS5.0), the gene was considered expressed, and not-expressed otherwise.

The predicted GRN state after CVX-060 treatment was obtained by selecting all interface TFs activated or inhibited by the activation of Tie2 with probability higher than zero, according to the calculation of the MPPs. The BPCs containing only combinations of such interface TFs were selected, and the GRN-TFs that change their state in response to them are expected to change upon CVX-060 application. The GRN state predicted after the inhibition of Ang2 was obtained in the same way.

3.5 Application to an animal regeneration model

This section presents methods specific for the application of the method to the prediction of pathways involved at each time step of a limb regeneration time series in the salamander *Ambystoma mexicanum*. The following procedures were performed by collaborators at the

Research institute of Molecular Pathology (IMP), at the Center for regenerative Therapies Dresden, and in the Computational Biology research group at LCSB.

3.5.1 Experimental procedures

Time-course includes time points: 0, 1, 3, 5, 7, 10 and 14 days following upper-arm amputation. Tissue between 0.5 mm behind the amputation plane and the tip of the blastema was collected. Connective tissue progenitors were specifically and irreversibly labeled during the limb bud development via Cre-induced recombination of a reporter construct. Connective tissue specificity is achieved by the Prrx1 –limb specific enhancer that controls the expression of the Cre recombinase in the transgenic axolotls. Upon Cre activation using drug Tamoxifen, a DNA-cassette is removed from the reporter-transgene allowing the expression of the red fluorescent protein Cherry. Cherry+ cells were selected using fluorescence-activated cell sorting. RNA was extracted and reverse transcribed. The resulting cDNA was then transcribed into labeled complementary RNA (cRNA) which was hybridized to custom Agilent 2x400K oligonucleotide microarrays.

3.5.2 Data processing

Probes showing low gene expression correlation across replicates were excluded from further analysis. When multiple probes were mapping to the same gene, the mean of the highly correlated probes (Pearson correlation $>$.8) was assigned to the gene. Differential expression analysis was performed with the R package *limma* for each couple of consecutive time points (i.e. D1 vs. D0, D3 vs D1, etc.), and genes were considered differentially expressed with absolute log fold change higher than $log_2(1.5)$ and p-value corrected for false discovery rate (Benjamini and Hochberg 1995) lower than 0.05. The Boolean state and expression probability was calculated at each time point with the geneDE approach described in Section 3.1.3.

3.5.3 GRN reconstruction

An initial GRN for axolotl was inferred from the microarray data using five different GRN inference tools: CLR (Faith et al. 2007), TIGRESS (Haury et al. 2012), PLSNET (Guo et al. 2016), GENIE3 (Huynh-Thu et al. 2010) and Pearson's correlation coefficient. The inference was performed for TFs differentially expressed and changing their Boolean state at any time interval. All tools were used with default parameters. The results of the different methods were combined by keeping interactions that were ranked among the best 10% of all interactions, by at least four of the five used tools. The sign of the interaction was assigned according to the results of Pearson's correlation: if the correlation is negative, the interaction is deemed inhibitory, otherwise activatory. For each time interval, a subnetwork was extracted containing only TFs that are DE and changing their Boolean state in that time interval. Each subnetwork was optimized as mentioned in Section 3.1.2, resulting in GRNs specific for each time step of the regeneration time course. In order to confirm that TFs relevant to the biological processes involved in regenerations are captured during the GRN inference process, enrichment in Gene Ontology terms related to Biological Processes was tested. The R package *topGO* (Alexa, Rahnenfuhrer, and Lengauer 2006) was applied using algorithm "weight01" and Fisher test statistics.

3.6 Databases

3.6.1 Perturbation targets

The protein targets for drugs and small molecules used in the experiments were obtained from STITCH ((Szklarczyk et al. 2016), v5.0, accessed in October 2017, with experimental evidence confidence >0.4). The effect of the perturbations on their targets (activation, inhibition, or unknown effect) was extracted from DrugBank ((Wishart et al. 2018), accessed in October 2017) and MetaCore from Clarivate Analytics.

3.6.2 Pathway targets

The experimental perturbations collected were rarely directed to the perturbation of complete canonical pathways. Therefore, the pathways that should be recapitulated by predictions were not defined beforehand. To do so, the pathways that contained at least 25% of the direct perturbation targets were selected. Additionally, if any of the direct targets only belong to one canonical pathway, that was also retained as positive pathway. This procedure was applied to both Metacore pathways, as defined in 3.2.1 and to KEGG pathways.

3.6.3 Datasets

Datasets where gene expression data was measured by microarray before and after the application of a signalling perturbation were collected across different databases. In particular, all datasets present in Connectivity Map (build 02, (Lamb et al. 2006)) generated on the platform Affymetrix Human Genome U133A 2.0 Array were considered. Then, manually selected experiments where a single perturbation was applied were collected from ArrayExpress and Gene Expression Omnibus. All experiments whose perturbation targeted molecules absent from the signalling network or disconnected from interface TFs were removed, as well as chemically undefined perturbations (culturing with other cell types, ROS and other cellular stressors, use of serum, etc.). The analysis was then restricted to datasets where the perturbation applied targeted up to 30 signalling molecules present in the signalling network, in order to test the methodology developed on well-characterized signalling perturbations and avoid obtaining correct predictions by chance. In particular, experiments related to cell differentiation and reprogramming of non-cancerous cell types were used to test the prediction of both signalling molecules and pathways to induce cellular transitions (Table in Appendix 7.3).

3.6.4 Phosphoproteomics datasets

Phosphoproteomics experiments where a single perturbation was applied to the cells were collected. They were paired with gene expression datasets with closely matching initial

Table 3.1: Phosphoproteomics datasets considered. The log fold change used in the original study to define differentially phosphorylated (DP) proteins is reported (DP lfc). The number of interface TFs that are connected to the GRN (selected) and DP is reported for comparison with the number of selected or DP interface TFs. The enrichment of the selected interface TFs in DP proteins was calculated by one-sided Fisher test.

Dataset	**Cell type**	**Perturbation**	**DP lfc**	**sel. DP iTFs**	**sel. iTFs**	**DP iTFs**	**p-value**
D'Souza et al. 2014	HaCaT	TGF-β	1	18	59	51	**2.36E-04**
Sharma et al. 2014	HeLa	EGF	1	19	30	139	**3.24E-03**
Wilkes et al. 2015	MCF7	EGFR inhibition [EGFR2]	1	16	60	52	**4.13E-03**
Gnad, Doll, et al. 2016	HCT116	MAPK inhibition [GDC0973 (1μM)]	log2(3)	3	29	15	1.08E-01
Rudolph et al. 2016	MCF7	EGF	2.38	1	27	2	1.43E-01
Wierer et al. 2013	MCF7	estradiol	log2(1.5)	0	57	3	1.00E+00

conditions and perturbation applied. When possible, the same cell type was perturbed with the same chemical compound in both phosphoproteomics and transcriptomics datasets, and the delay before measurements was comparable. Otherwise, a delay of up to 48 hours was considered. Closely related cell lines and equivalent perturbations (acting on the same protein targets) were accepted. The list of differentially phosphorylated (DP) proteins were obtained directly from the original phosphoproteomics studies, when available, or extracted by repeating the analysis as described by the authors (Table 3.1). Each protein was assigned the highest log fold change observed for any of its phosphosites.

3.6.5 Availability

The methods were implemented as a Snakemake pipeline (Koster and Rahmann 2012), consisting of Matlab and R scripts, and was made available at https://git-r3lab.uni.lu/gaia.zaffaroni/INCanTeSIMO. Microarray data generated in the context of the application to cirrhosis is available in Gene Expression Omnibus under accession number

GSE122822. The analysis of all datasets was performed on the UL HPC platform (Varrette et al. 2014).

4 Results

In this chapter, the performance of INCanTeSIMO and INCanTeSIMO_path is assessed across numerous datasets obtained from different sources. Additionally, their predictions in examples of cellular differentiation and reprogramming are compared with literature knowledge. Finally, they were applied to the analysis of animal models of disease and regeneration.

4.1 Comparison of MPPs with phosphoproteomics data

Protein regulation in signalling pathways takes many forms (M. J. Lee and Yaffe 2016), but phosphorylation takes a prominent role as the primary mechanism used to transmit signal in the cytoplasm. For this reason, phosphoproteomics experiments are used to infer protein activity in response to signalling cues (Invergo and Beltrao 2018). Proteins showing significant changes in their phosphorylation state are expected to play an active role in the transmission of the signal, so that signalling paths responding to a stimulus show an enrichment in differentially phosphorylated (DP) proteins compared to paths that are not involved. Under this assumption, the most probably expressed paths (MPPs) identified with INCanTeSIMO were tested for enrichment in phosphorylation changes.

Previously published experiments, measuring gene expression and phosphoproteomics data before and after a specific perturbation was applied, were collected (Table 3.1). For each of them, the direct perturbation targets were defined from databases, and the interface TFs available were identified as described in the Methods section 3.6. The interface TFs selected were significantly enriched in DP TFs, when the number of DP TFs was high (Table 3.1). The MPPs between each target and interface TF pairs were computed using both the correlation- and length-corrected probability as defined previously.

Signalling proteins can contain multiple phosphosites, which are amino acids that can be modified by the addition of a phosphate group. Phosphorylation of each site might or not have a functional role, by modifying the way the protein can interact with its partners (M. J. Lee and Yaffe 2016). At the moment, the annotation of functional sites is available for a limited number of proteins, while for the vast majority of proteins this information is still not available (Invergo and Beltrao 2018). Thus, any significant change in the phosphorylation of a site was assumed to be functional, and a protein was considered DP if any of its phosphosite resulted DP in the original publications (Table 3.1).

The frequency of DP proteins in each MPP was compared to up to 100 randomly selected simple paths connecting the same source and target signalling nodes (corresponding to perturbation target and interface TF, t-test with p-value < 0.05) (Figure 4.1A). Overall the majority of the MPPs for each dataset had more DP proteins than other possible paths, for both methods used to define MPPs (Figure 4.1B), even if some particular target-interface TF combination was connected by paths that did not contain any phosphorylated protein (Figure 4.1C).

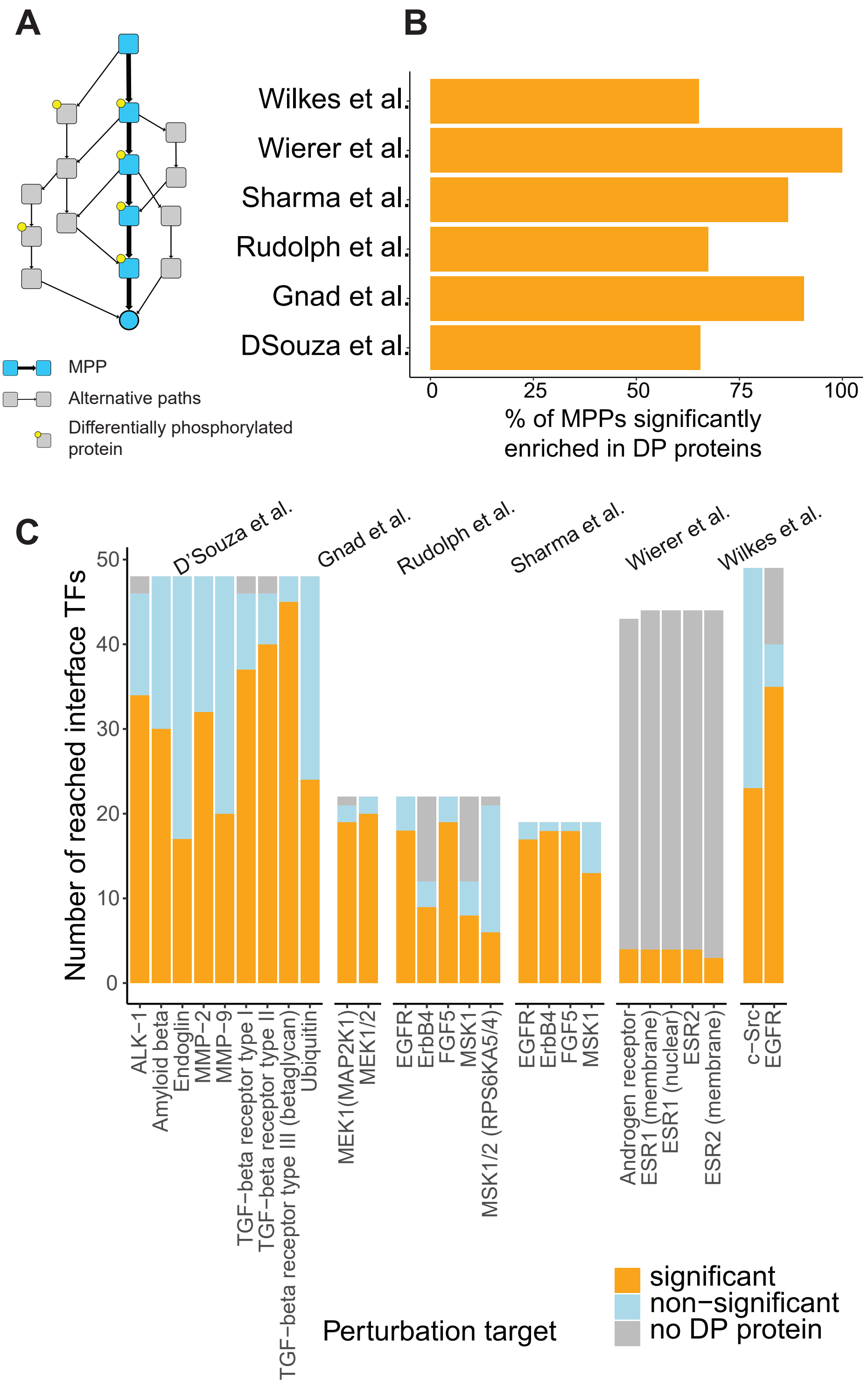

A
MPP
Alternative paths
Differentially phosphorylated protein
B
Wilkes et al.
Wierer et al.
Sharma et al.
Rudolph et al.
Gnad et al.
DSouza et al.
0
25
50
75
100
% of MPPs significantly enriched in DP proteins
C
Number of reached interface TFs
0
10
20
30
40
50
D'Souza et al.
Gnad et al.
Rudolph et al.
Sharma et al.
Wierer et al.
Wilkes et al.
ALK−1
Amyloid beta
Endoglin
MMP−2
MMP−9
TGF−beta receptor type I
TGF−beta receptor type II
TGF−beta receptor type III (betaglycan)
Ubiquitin
MEK1(MAP2K1)
MEK1/2
EGFR
ErbB4
FGF5
MSK1
MSK1/2 (RPS6KA5/4)
EGFR
ErbB4
FGF5
MSK1
Androgen receptor
ESR1 (membrane)
ESR1 (nuclear)
ESR2
ESR2 (membrane)
c−Src
EGFR
Perturbation target
significant
non−significant
no DP protein

Figure 4.1: Enrichment of MPPs in differentially phosphorylated (DP) proteins. **(A)** MPPs are defined by both correlation-based and length-based probabilities from the perturbation targets to all interface TFs. For both methods, the fraction of DP proteins in each MPP connecting each target-interface TF pair is compared by t-test to the number of DP proteins in up to 100 randomly selected simple paths connecting the same pair. **(B)** Average fraction of MPPs significantly enriched in DP proteins compared to alternative simple paths, per dataset (P-value $<$ 0.05). **(C)** Breakdown of the results for each perturbation target. Orange: number of interface TFs for which the fraction of DP proteins in the MPP is significantly higher than in other simple paths. Light-blue: interface TFs reached with MPPs that show no significant difference in DP enrichment; grey: interface TFs that are connected to the perturbation target with paths (both MPPs and alternative paths) that do not contain DP proteins. The same results (panels B and C) were obtained with both correlation-based and length-based MPPs.

These results show that using MPPs to represent the signal transduction from signalling molecules to interface TFs is in agreement with observed phosphorylation patterns. This suggests that MPPs can be used as approximation of paths used for signal transduction.

4.2 GRN perturbation results

In general, a limited number of TFs was observed to change their Boolean state between the initial and final cellular state, so the size of the GRNs representing the cellular conversion was moderate (on average 24 TFs, Figure 4.2A). The number of interface TFs selected for each dataset was highly variable, ranging between 12 and 151 and averaging 39 TFs. The best perturbation in each dataset was able to change the Boolean state of 71% of the GRN TFs on average.

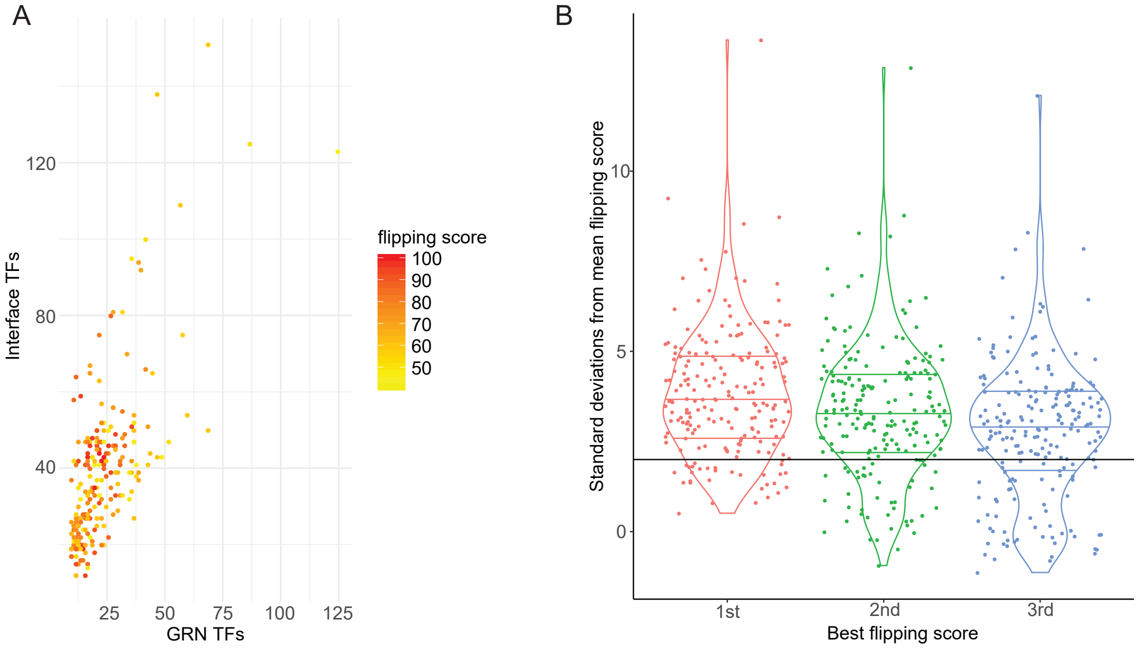


Figure 4.2: Effect of the in silico perturbations of combinations of interface TFs on the state of the GRN. **(A)** The percentage of GRN-TFs that change their state in the best combination overall is represented. **(B)** The performance of interface TF combinations that obtain the first, second and third flipping scores, compared to the average flipping score obtained with all the combinations tested, represented by distance measured in number of standard deviations from the average. The black line represents 2 standard deviations from the average.

The best performing combinations (BPCs) were selected by considering the combinations obtaining the best three flipping scores. This corresponds to combinations that are overall more than 2 standard deviations away from the average flipping scores obtained in each dataset (Figure 4.2B).

4.2.1 Effect of the number of targets on the effect of interface TFs on the GRN state

The influence of the number of targets of an interface TFs on its frequency among the BPCs was analysed. There was a general tendency of interface TFs with more targets to be present among the BPCs in more datasets (Figure 4.3). However, no single interface TF was present in the BPCs of all datasets, and even interface TFs with more than 100 targets were present in the BPCs of as low as 10-20% of all datasets.

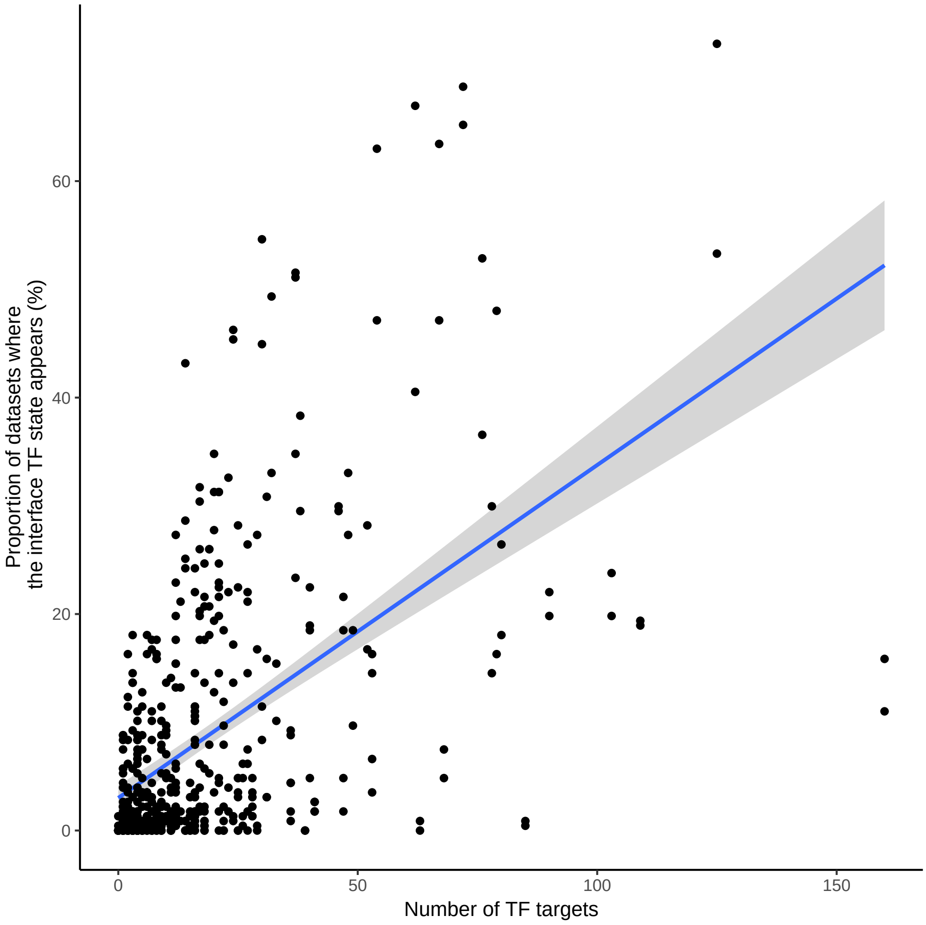


Figure 4.3: Effect of the number of transcriptional GRN-TFs that are targets of each interface TF (x-axis) on the frequency of an interface TF state among the BPCs of multiple datasets (y-axis). 538 interface TF states are present across the 228 datasets considered.

This result is not surprising. The expression of the interface TFs in the initial cellular state and the expression state of the cell in general influence the set of interface TFs considered. Additionally, each cellular transition is assigned a different GRN, containing a distinct set of differentially expressed TFs. Therefore, the same interface TF will have different targets in each dataset, and different effectiveness in changing the state of the corresponding GRN. As a result, the presence of an interface TF among the BPCs cannot be estimated a priori.

4.2.2 Synergy among interface TFs

Before selecting the BPCs, combinations of interface TFs that did not show any synergistic effect were discarded. The rationale behind this choice is to limit the number of interface TFs that the ideal signalling molecule should act on, in order to prioritize specific GRN perturbations. In fact, small molecules or drugs with a limited number of targets are generally preferred to compounds that have unspecific effects on cells. Selecting BPCs that showed some synergistic effect, the number of interface TFs to perturb was reduced (Figure 4.4), while the GRN state changes they are predicted to induce was maximised.

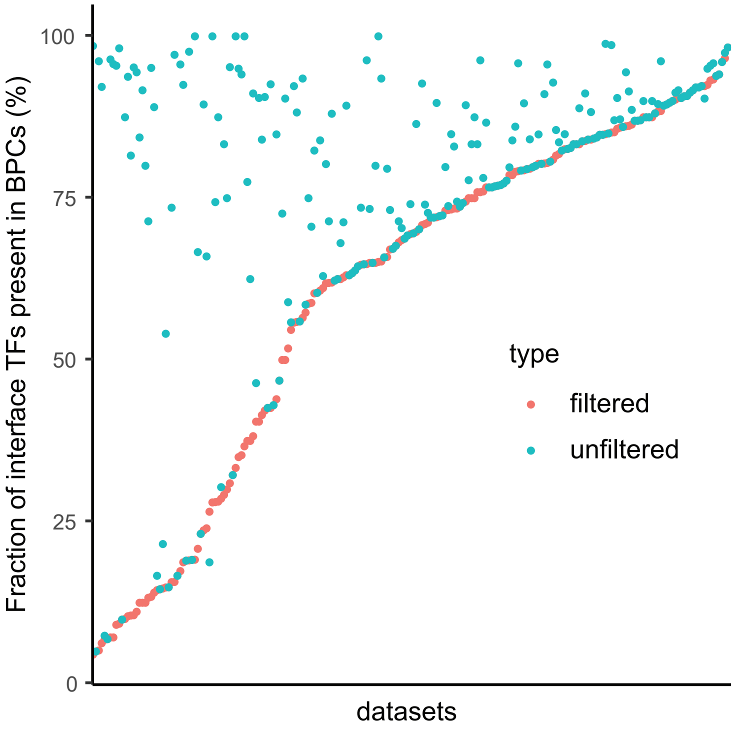


Figure 4.4: Effect of filtering for synergistic effect on the number of interface TFs in the BPCs. The fraction of interface TFs connected to the GRN that are present in the BPCs is shown for each dataset. BPCs are selected by a) considering all combinations with the three best flipping scores ("unfiltered", blue); b) by considering the combinations with synergistic activity of the involved TFs, so that the combination is affecting the state of more GRN-TFs that the sum of its components ("filtered", red). Apart from two cases, the filtered BPCs contain an equal or lower fraction of interface TFs compared to the unfiltered BPCs.

The success rate of INCanTeSIMO in predicting direct targets of the experimentally applied perturbation was 61% when ranking signalling molecules by their probability of activating the interface TFs present in synergistic BPCs. In order to test if this filtering had a positive influence on the overall method performance, the success rate was calculated when all BPCs, without filtering for synergistic combinations, were considered. The performance obtained was 56%. The difference bewteen the two was not large, but nonetheless showed that predictions improved when only synergistic combinations of interface TFs were considered.

It must be noticed that focusing on synergistic combinations did not result in selecting “rare” interface TFs, defined as TFs that are present in less than 1% of the BPCs of a certain cellular transition. In fact, BPCs composed of only rare TFs represented only 0.006% of the BPCs across all datasets. This indicated that BPCs are composed mostly of ”common” interface TFs, while it hardly happened that a specific combination of otherwise non-effective interface TFs resulted in large GRN state changes. Enumerating all combinations of interface TF states is unfeasible, but this pattern was assumed to exist in higher order combinations too, so that almost exclusively ”common” TFs would be present in BPCs. Thus, the frequency of interface TF across BPCs was used as a measure for its effectiveness in changing the GRN state as desired, independently from the perturbation size considered.

4.3 Overall performance in the prediction of signalling molecules for cellular transitions

INCanTeSIMO was applied to single drug perturbations applied to cell lines, obtained from CMap and ArrayExpress. After quality controls, 228 datasets (193 from CMap, 35 from ArrayExpress), corresponding to cellular transitions, were analysised. The signalling molecules were ranked by calculating the Jensen-Shannon divergence between their probability of acting on interface TFs, and the likelihood of interface TFs to induce the desired GRN state transitions (see Materials and Methods). For each dataset, 1400-1500 signalling molecules were tested for both their activation or inhibition, for a total of around 3000 potential signalling perturbations. The ranking prioritized molecules that reach with high probability and specificity interface TFs that performed well in the in silico perturbations of the GRN.

Initially, the ranking obtained with INCanTeSIMO was compared with simple differential expression analysis: genes were sorted by decreasing absolute lfc, an increasing part of the ranking was selected and the datasets for which perturbation targets were found in the selection were counted. Ranking by differential expression did not prioritize perturbation targets, which were only found after a big portion of the ranking was selected (Figure 4.5A). The ranking generated by INCanTeSIMO performed better: to select at least one correct perturbation target in 50% of the datasets, 920 molecules should be selected according to differential gene expression, and 236 are necessary with INCanTeSIMO. This result confirms that differential expression is not informative on the role of genes in signal transduction, and therefore the use of more complex approaches is required to extract meaningful predictions from gene expression data.

The success rate of a method predicting signalling molecules for cellular transitions was defined as the fraction of datasets for which a correct prediction was obtained. As each cellular transition was obtained experimentally to generate the data analysed, a prediction was considered correct if at least one of the known targets of the experimental perturbation appeared in the top ranked molecules. At different ranking cut-offs, INCanTeSIMO success rate was better than the expected by random selection of the same number of molecules (Figure 4.5B). Cut-off=0.06 was used for following analyses because the gain of performance of INCanTeSIMO compared to random selection was maximum, with successful predictions in 139 out of 228 datasets (61%)(Complete table in Appendix 7.3). At this cut-off, the method correctly predicted perturbation targets in 115/193 CMap examples (60%, versus random success rate of 40%), 6/10 datasets for non-cancer cell lines selected from ArrayExpress, 5/6 datasets with matched phosphoproteomics data, and 13/19 cell type transitions datasets (discussed below).

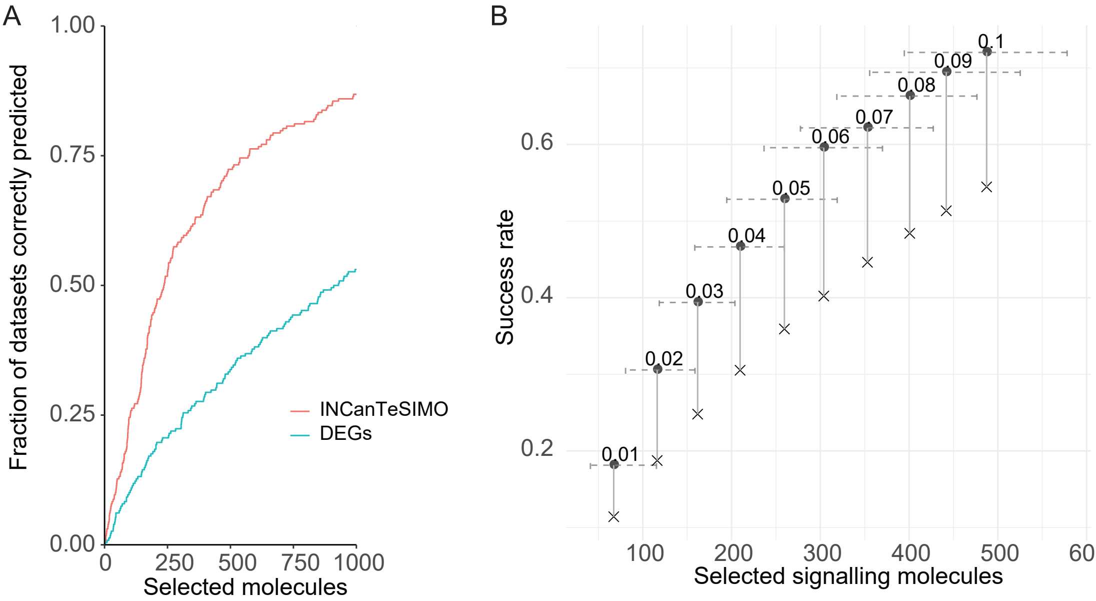


Figure 4.5: Overall performance of INCanTeSIMO in predicting direct perturbation targets. **(A)** Fraction of datasets with at least one perturbation target correctly predicted, across increasing selection of ranked molecules. The proteins were ranked either by INCanTeSIMO, or by differential expression analysis (log fold change). **(B)** Variation of success rate and number of selected signalling molecules, at different ranking cut-offs. Circles=INCanTeSIMO success rate, X=success rate for random selection of the same number of molecules. Horizontal error bars: 5th and 95th percentile of selection size. To the same fraction of ranking selected correspond variable set sizes because of ties in the ranking. The selection of 6% of the ranking gives the best improvement compared to random performance. Only CMap datasets where used to define this cut-off.

4.3.1 Factors influencing the performance

The probability of finding at least a direct perturbation target among the candidate molecules depends on their number. Therefore, datasets were divided in classes according to the number of targets that the applied perturbation (compound or protein) has across the databases used (see Methods section 3.6.1). The performance of INCanTeSIMO was assessed in each of the classes obtained by comparing the observed success rate to the frequency at which a target is expected to be selected by random selections of the same size (Figure 4.6). The method was significantly better than random in datasets with 1 to 10 known perturbation targets (p-value=1.27e-04 for datasets with 1-5 targets, and 9.40e-06 for 6-10). These datasets represent the 74% of all tested datasets and also correspond to the most interesting application cases, as the use of target-specific drugs or proteins is preferred for the controlled induction of cellular transitions.

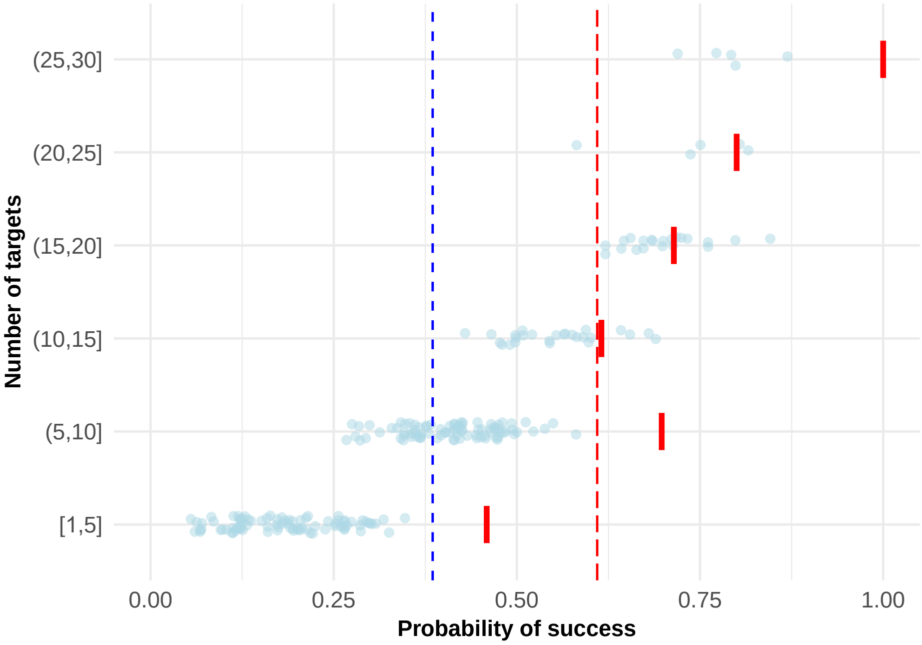


Figure 4.6: Prediction of signalling molecules compared to expectation, according to the number of known perturbation targets. The datasets were divided in classes according to the number of direct perturbation targets, and the expected success rate was calculated for each dataset depending on the number of targets and the number of candidates selected by INCanTeSIMO (light blue dots). The fraction of datasets in each class where INCanTeSIMO predicted correctly at least one perturbation target is depicted with short red bars. The overall success rate of INCanTeSIMO (red dotted line, 61%) and the overall expected success rate (blue dotted line, 39%) are reported.

The number of DEGs between in initial and final cellular states was significantly higher for datasets in which INCanTeSIMO obtained correct predictions (Figure 4.7, p-value=5.032e-06). This result suggests that the method is particularly suited for cellular transitions that require extensive gene expression changes, not only for the TFs in the GRN, but also of the genes that they regulate. In other words, the bigger the differences between the initial and final gene expression state, the more likely it is that perturbations of the GRN via signalling perturbations can effectively induce the transition. On the contrary, when only limited differences are present between the initial and desired gene expression profile, acting on the GRN might not be the most effective strategy, and punctual regulation of the expression of single genes might be more indicated.

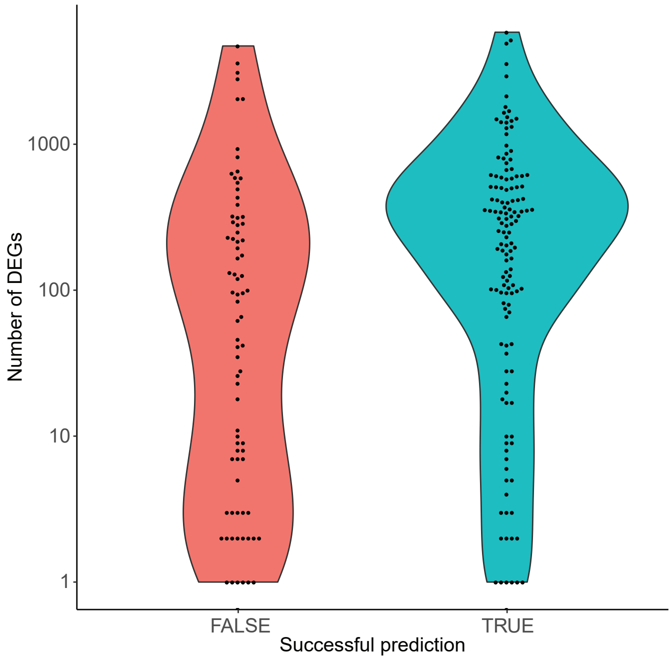


Figure 4.7: Success of the signalling molecule predictions with respect to the difference between initial and final cellular states. In datasets where INCanTeSIMO obtained successful predictions, the number of differentially expressed genes between the two gene expression profiles was significantly higher than in the datasets where the predictions were not correct (one-sided Mann-Whitney test, p-value = 5.032e-06).

The vast majority of the perturbation-target pairs present in the analysed datasets do not have a defined sign (Figure 4.8). For the 18% of the pairs that do have signs, the perturbation plays predominantly an inhibitory role on its targets (69%). Among the predictions obtained with INCanTeSIMO there was no bias in the sign of the candidate molecules, but the predominance of inhibitors among the perturbations was correctly reflected in the higher percentage of inhibitory molecules correctly predicted (69%). Nonetheless, errors in the predicted sign might occur in the predictions. This can be explained by the fact that multiple equally probable paths with opposite signs might exists between two nodes in the signalling network, but only one MPP is selected as representative of the effect of a signalling molecule on an interface TFs, thus determining the selection of the activation or inhibition of the same molecule.

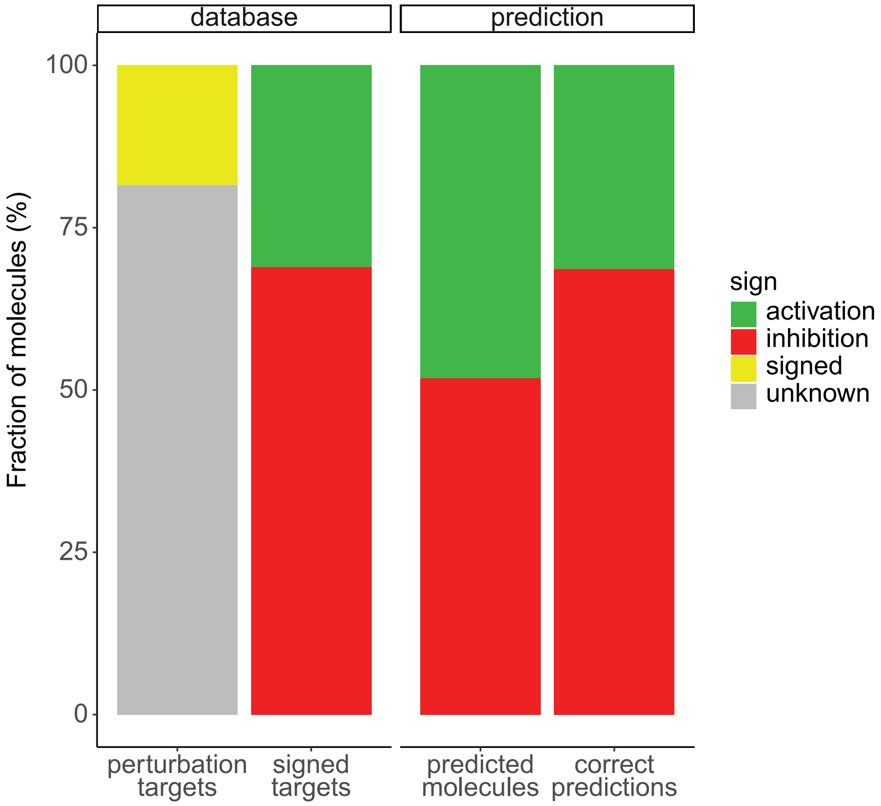


Figure 4.8: Effect of experimental perturbations on their protein targets. The vast majority of the drug-target pairs present in the datasets analysed and contained in the signalling network ("perturbation targets") is unsigned across multiple databases. Of the known interactions ("signed targets"), 69% are inhibitory. In the predictions obtained with INCanTeSIMO ("predicted molecules"), there is balance between activations and inhibitions, and the correct predictions show the same fraction of inhibitions as observed across all known interactions.

The effect of signalling molecules on the interface TFs are calculated using both length-based and correlation-based interaction probabilities. The two measures are moderately predictive taken separately: both of them predict around 43% of the datasets correctly at cut-off=0.06. The combination of the rankings obtained with the two strategies however resulted in a limited number of ties and better predictions overall. The influence of the two approaches on the predictions is similar: among the datasets associated with correct predictions mentioned previously (61% of the total 228 datasets), 37% would be correctly predicted with any of the two probability strategies, 31% require length-based probabilities, and the remaining 31% stems from correlation-based probabilities. Thus, there is not a single strategy that is best for the inference of signal transduction paths, as it seems that both co-expression of the proteins that need to interact, and the number of interactions overall, play a role in the choice of signal transduction paths used.

4.3.2 Functional and topological properties of candidate molecules

Apart from the perturbations experimentally tested, the perturbation of other signalling molecules might trigger the desired cellular transitions. For example, molecules that are involved in the regulation of the same biological processes as the experimental perturbation, or molecules that are positioned in their vicinity in the signalling network. Therefore, the sets of predicted candidate signalling molecules were tested for their relatedness to direct perturbation targets in terms of functional annotation and distance from the experimental targets.

Functional analysis was performed by Gene Ontology (GO) analysis. The biological process terms associated with direct perturbation targets were defined as ”target terms”. Separately, the GO terms enriched among the candidate signalling molecules were collected, as well as the terms enriched among non-candidate molecules. The presence of target terms among the two sets of enriched terms was considered. Target terms were overrepresented more frequently among the candidate signalling molecules than among the non-candidate ones (one-sided Wilcoxon test, p-value<2.2e-16, Figure 4.9A). Thus, candidate molecules were involved in the same biological processes as the direct perturbation targets more than signalling molecules discarded by INCanTeSIMO.

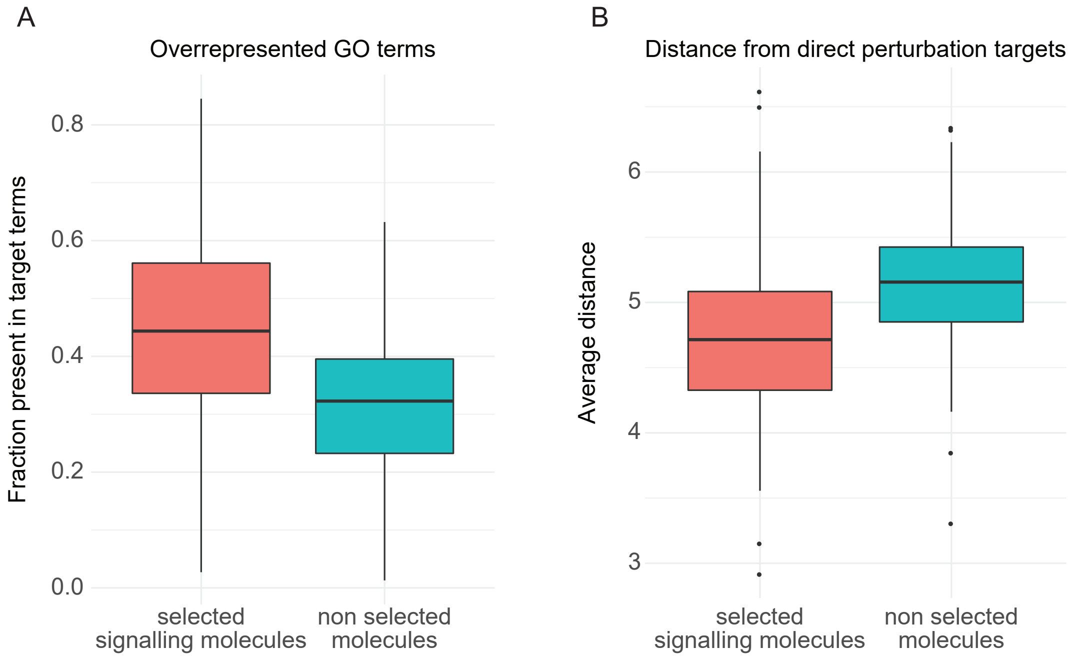


Figure 4.9: Functional and topological properties of candidate molecules. **(A)** Percentage of functional terms mapping to the perturbation targets also enriched in the signalling molecules selected as candidates by INCanTeSIMO, or the molecules not selected. The candidates have significant higher portion of enriched functional terms shared with the perturbation targets. **(B)** The average distance of the selected molecules and non-selected molecules from the direct perturbation targets show that candidates predicted by INCanTeSIMO are located significantly closer to the true perturbation than non predicted molecules.

To study where candidate signalling molecules were located in the signalling network, the distance from each molecule in the network to the perturbation target was defined as the minimum number of interactions with same direction required to connect two nodes. The average distance of candidate molecules and non-candidate molecules from the perturbation targets showed significant difference (one-sided Wilcoxon test, p-value$<$2.2e-16, Figure 4.9B). For the majority of the datasets, the distances were significantly shorter for candidate molecules (74% of datasets), longer in 2% of the cases, and comparable in the remaining 24%. This result indicates that candidate molecules are not distributed randomly on the signalling network, but are gathered in the region where the experimental perturbation exerts its function.

In summary, the candidate molecules selected by the proposed method are involved in the same biological processes and pathways as perturbation targets more than non-candidate molecules and DEGs. Their placing in the network is also not casual but shows similarity to perturbation targets. These results suggest that perturbation of candidate signalling molecules, even if not yet experimentally tested, are likely to induce the desired cellular conversion in the same way that known perturbations do.

4.4 Performance in the prediction of signalling pathways for cellular transitions

To better interpret the prediction of signalling molecules obtained with INCanTeSIMO, a method for the prediction of canonical pathways, INCanTeSIMO_path, was developed. Because candidates are selected by INCanTeSIMO based on their specific activity on the interface TFs, it is expected that only some members of a pathway will be predicted, while others acting with less specificity will not be prioritized. Therefore, over-representation of the members of a pathway among the candidates is not necessary for a pathway to be relevant to the desired cellular transition. Instead, the pathways were ranked according to their enrichment in causal disturbance, calculated according to source/sink centrality (see Methods section 3.2.1). Briefly, this method selects pathways based on the fact that highly influential nodes in their signal transduction are predicted as candidates by INCanTeSIMO.

To evaluate the performance of INCanTeSIMO_path, positive pathways were determined for each dataset from the direct targets of the experimental perturbation. In particular, MetaCore signalling pathways that contain at least 25% of the direct perturbation targets were collected. Additionally, perturbation targets that only belong to one pathway were selected, and such pathways were added to the positive ones (see Methods section 3.2.1). The activation and inhibition of a pathway were considered separately. Finally, 166 datasets among the 228 analysed had at least one positive MetaCore pathway, with a median of 3 positive pathways per dataset.

INCanTeSIMO_path correctly predicted positive pathways in 58% of the cellular transitions ($p.value \leq 0.05$), compared to an expected success rate of 47% for the random selection of the same number of pathways. The probability of selecting at least one positive pathway by chance is affected by the number of positive pathways, therefore the datasets were classified according to the number of positive pathways, and the probability of random success in each class was compared to the performance obtained by INCanTeSIMO_path (Figure 4.10). The method was better than random selection for datasets with 2 to 4 positive pathways (p-values 3.7e-2, 5.4e-2, 5.2e-2), which represent 65% of all datasets, and comparable to random selection in the other cases.

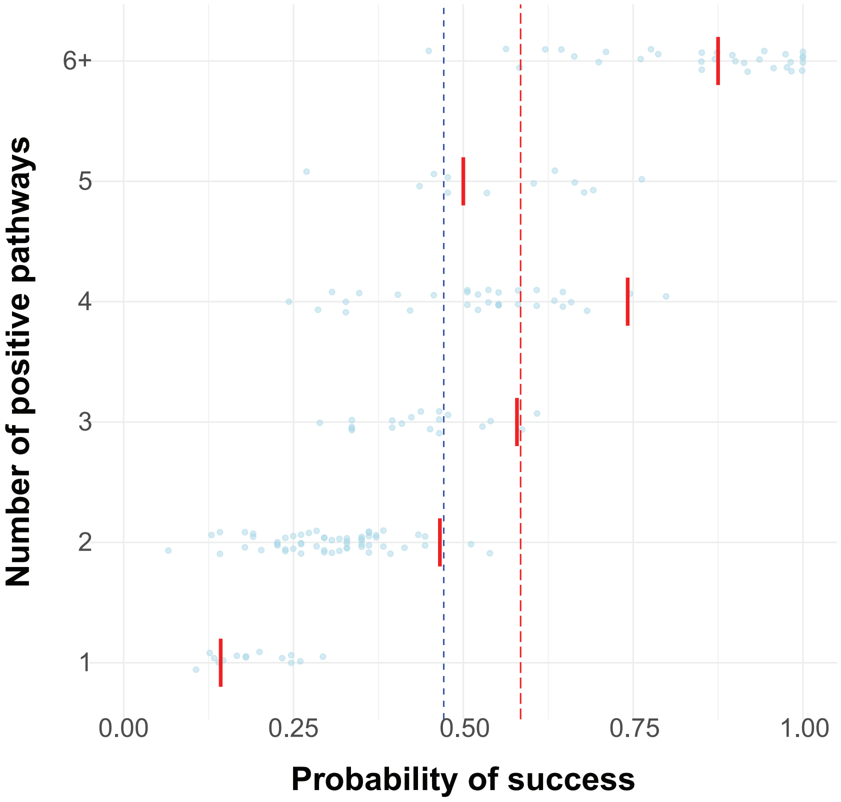


Figure 4.10: Performance of INCanTeSIMO_path in relation to the number of positive pathways.The datasets were divided in classes according to the number of positive pathways present, and the expected success rate was calculated for each dataset depending on the number of pathways and the number of candidates selected by INCanTeSIMO_path (light blue dots). The fraction of datasets in each class where INCanTeSIMO predicted correctly at least one of the positive pathways is depicted with short red bars. The overall success rate of INCanTeSIMO (red dotted line, 58%) and the overall expected success rate (blue dotted line, 47%) are reported.

4.5 Comparison to other methods

No method exists to date similar to this work in terms of application or modelling strategy. The approach presented was therefore compared to computational tools that are widely used to analyse signalling events using gene expression data. As mentioned in the Methods section 3.3, they are methods that predict both signalling molecules (Connectivity Map, DeMAND) and entire pathways (SPIA, GSEA, CADIA).

4.5.1 Comparison to single molecule prediction tools

The performance obtained with INCanTeSIMO was compared to Connectivity Map and DeMAND. These two methods use DEGs but otherwise follow quite different strategies for the prediction of potential perturbations. Connectivity Map takes advantage of an extensive database of gene expression signatures generated by known perturbations, and searches for similarities with the query expression profile. DeMAND, on the other hand, calculates the enrichment of deregulated genes among the targets of each potential perturbation, defined by a context-specific regulatory network.

Connectivity Map was applied to all 228 datasets. Both INCanTeSIMO and Connectivity Map had a success rate of 61% on the datasets associated with a prediction (Figure 4.11A), however in some datasets there were not enough DEGs for Connectivity Map to give a result, so predictions were available in 144 datasets only. The majority of the datasets analysed are already present in the compendium of gene expression signatures used for Connectivity Map for generating predictions. A closer look at the cell fate transition cases, which were obtained from independent data sources, revealed that Connectivity Map correctly predicted perturbations in only 25% of the cases, while INCanTeSIMO succeeded in 75% of them (Figure 4.11B). Thus, Connectivity Map could not be successfully applied to novel cellular transitions. INCanTeSIMO on the contrary showed consistent performance across the different types of datasets.

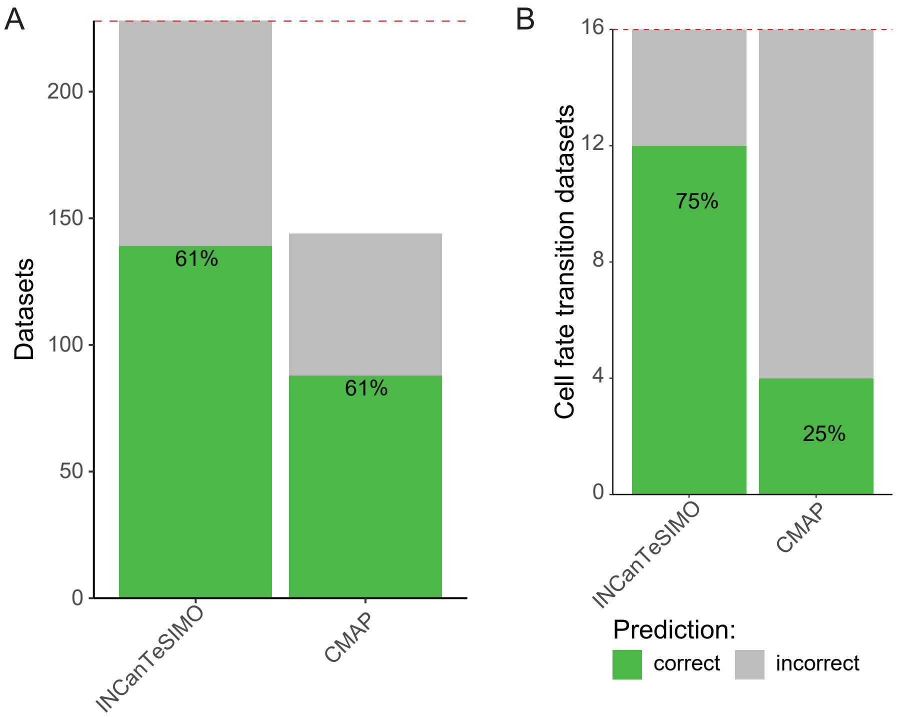


Figure 4.11: Performance of INCanTeSIMO compared to Connectivity Map. **(A)** Number and proportion of datasets in which INCanTeSIMO and Connectivity Map correctly predict at least one of the known perturbation targets. INCanTeSIMO generated predictions for all 228 datasets, while Connectivity Map failed to analyse 84 datasets. **(A)** Number and proportion of cell fate transitions in which INCanTeSIMO and Connectivity Map obtained correct predictions. These datasets are a subset of the 228 analysed overall.

DeMAND (Woo et al. 2015), which is one of the few GRN-based tools available, could not be compared with INCanTeSIMO in the same manner, because of its requirement for at least six gene expression data replicates per condition. Therefore, INCanTeSIMO was applied to eight datasets used for DeMAND's benchmarking and compatible with our method. Across these drug-induced cellular transitions, both INCanTeSIMO and DeMAND correctly predicted perturbation targets in six datasets (Figure 4.12). The fact that INCanTeSIMO obtained a comparable performance to a method which requires substantially more data, while also indicating if the candidate molecules should be activated or inhibited (correct sign predicted in 5/6 datasets), suggests that it is a flexible method suitable for a wide range of applications.

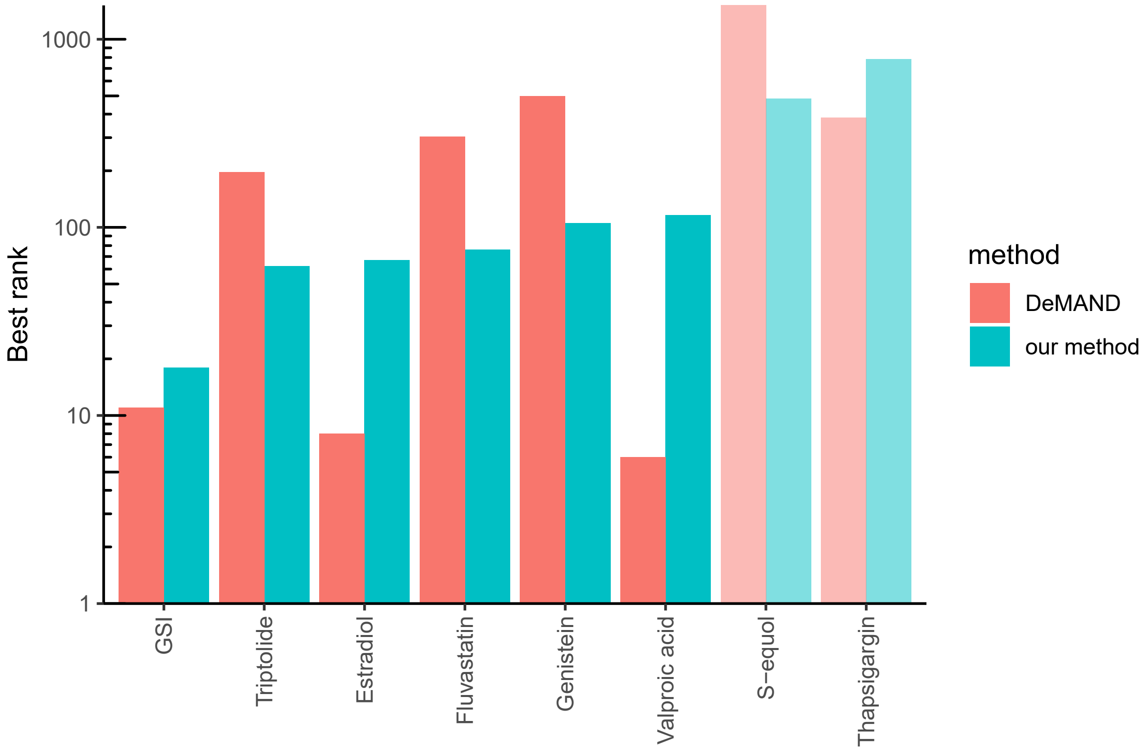


Figure 4.12: Performance of INCanTeSIMO and DeMAND on the eight compound perturbation datasets that could be analysed with both methods. Both tools predicted direct perturbation targets for 6 datasets and were not successful in predicting targets of (S)-equol and thapsigargin.

4.5.2 Comparison to pathway prediction tools

The performance of INCanTeSIMO_path in predicting signalling pathways for the induction of cellular transitions was compared to existing approaches that analyse differential pathway activity by focusing on DEGs. There exist three main classes of methods: over-representation analysis (ORA), functional class scoring (FCS), and topological analysis (Khatri, Sirota, and Butte 2012). Here representative methods for each classes were selected: enrichment by hypergeometric test (ORA), gene set enrichment analysis (GSEA) (FCS), and the methods SPIA and CADIA, which combine ORA and topological measures (see Methods for a brief description of each method). Causal network inference methods were not applied to this analysis because they consider reduced sets of pathways or TFs (Catlett et al. 2013; Parikh et al. 2010), or have limited applications, as in the case of cancer-specific methods (Y.-A.

Kim, Wuchty, and Przytycka 2011; Paull et al. 2013) and tools assuming that the signalling perturbation is known (Melas et al. 2015).

As mentioned in the Methods section 3.2.1, the “positive” pathways for each dataset were determined from the direct perturbation targets. SPIA and CADIA by default predict KEGG pathways, while all other methods can be adapted to predict MetaCore canonical pathways, so both MetaCore and KEGG positive pathways were determined for each dataset. The 166 datasets that had both KEGG and MetaCore positive pathways were considered for assessment of pathway analysis: each dataset was assigned up to 66 MetaCore positive pathways (median 3), and up to 77 KEGG positive pathways (median 5).

The performance of each method was quantified as the number of datasets in which at least one positive pathway was predicted as significant according to the method-specific significance cut-off. However, not all methods generated predictions for each dataset, as reported in Figure 4.13A. In fact, methods that use DEGs as input for the prediction failed to predict pathways in a major portion of the datasets (from 41% for SPIA to 73% for CADIA). This was due to the quantity of DEGs present in each dataset, which varied greatly but was generally low (0 to 5855, median 96.5). In comparison, the number of candidate signalling molecules generated by INCanTeSIMO was fairly high and stable (196 to 476 candidates in each dataset, median 304.5), allowing the methods that used them as input data to calculate pathway enrichments in all datasets. The only method that obtained correct predictions in more than 50% of the datasets was INCanTeSIMO_path (Figure 4.13A). However, this result was dependent on the number of pathways predicted by each method, which was higher for INCanTeSIMO_path than all other methods (Figure 4.13B). To evaluate the goodness of the ranking obtained with each method, the number of pathways that needed to be selected from the ranking in order to pick positive pathways was calculated across all 166 datasets with all the tested methods (Figure 4.14A).

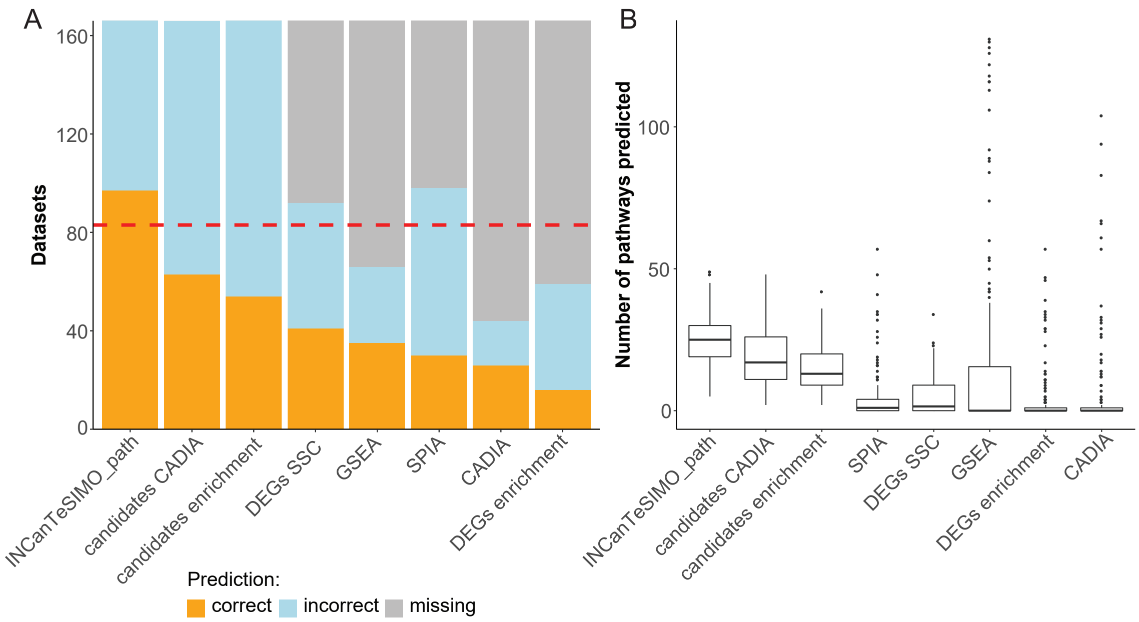


Figure 4.13: Overview of the results for pathway prediction methods. **(A)** Performance of the methods for pathway prediction. A prediction is considered correct if at least one of the positive pathways is predicted, and incorrect otherwise. Cases where there was no prediction are reported as missing. The red dotted line represents 50% of all datasets (in total 166). **(B)** Number of pathways predicted by each method, according to the significance level recommended in each method.

The results clearly showed that using the signalling candidates obtained from INCanTeSIMO resulted in a better ranking of canonical pathways. The selection of the best 19 pathways according to INCanTeSIMO_path was sufficient to obtain correct predictions in half of the datasets (”candidates SSC”), applying pathway enrichment to the candidate molecules (”candidates enrichment”) required 25.5 pathways to obtain successful predictions in 50% of the datasets, and applying CADIA on MetaCore pathways using the candidate molecules (”candidates CADIA”) required 24 pathways for the same result. This performance was consistently better than the methods based on DEGs (Figure 4.14A, the ranking were significantly different according to Kolmogorov-Smirnov test, p.value<1e-13 for all DEG-based methods). In fact, 60.5 pathways were required using SPIA, 140 in pathway enrichment, while for CADIA and GSEA even considering all KEGG or MetaCore pathways respectively was not sufficient to obtain correct predictions in 50% of the datasets.

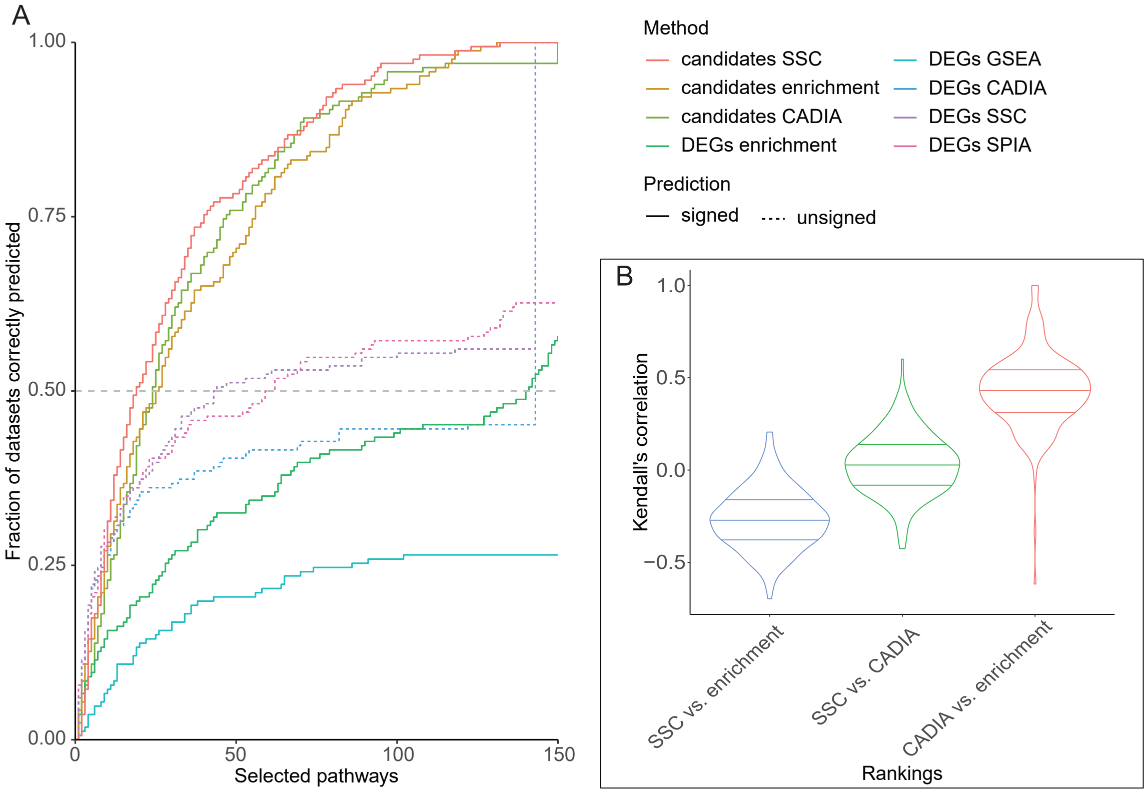


Figure 4.14: Overview of the results for pathway prediction methods. **(A)** Fraction of datasets where at least one positive pathway is correctly predicted for each of the pathway prediction methods. Candidate SSC corresponds to the ranking used in INCanTeSIMO_path (source/sink centrality enrichment across the candidate molecules obtained from INCanTeSIMO). Because some tools did not return any prediction in some datasets, considering all ranked pathways does not allow to reach success rate =1. **(B)** Correlation between the rankings obtained with the different methods using INCanTeSIMO candidate molecules as input for the prediction of pathways. SSC corresponds to the approach implemented in INCanTeSIMO_path. The rankings obtained with enrichment or CADIA are highly concordant, while SSC obtains rankings that are quite different from both.

Therefore, INCanTeSIMO_path predicted more signalling pathways, but also ranked them more correctly, thus obtaining a better performance compared to DEG-based methods. The performance at increasing number of selected pathways of INCanTeSIMO_path was comparable with the one obtained by calculating pathway enrichment or CADIA enrichment using the candidate molecules (Kolmogorov-Smirnov test, p.value=0.11 in both cases). However, INCanTeSIMO_path required slightly less pathways to be selected in order to obtain a successful prediction in 50% of the datasets, and was therefore the best method among the tested. Additionally, the ranking of predicted pathways was different between INCanTeSIMO_path and the other candidate-based methods, and similar between pathway enrichment and CADIA (Figure 4.14B), proving that the similar performance of INCanTeSIMO_path to the other candidate-based methods was not generated by the similarity of the methods overall. In fact, the definition of positive pathways based on the number of perturbation targets they contain, gives an advantage to pathway enrichment methods, and particularly enrichment in candidate molecules (assuming they are direct perturbation targets). The fact that INCanTeSIMO_path performed better than these methods proved that source/sink centrality is an effective alternative metric for the prediction of signalling pathways that can control the GRN state.

In summary, the prediction of pathways inducing cellular transitions is a complex problem in which no method could achieve high performance. Using the candidates obtained from INCanTeSIMO resulted in better predictions compared to DEG-based methods, while also suggesting the activation or inhibition of signalling pathways. INCanTeSIMO_path is the only method taking into account explicitly the importance of specific molecules in the context of signal transduction in signalling pathways combined with the causal role of signalling on GRNs, instead of an over-representation criterion. It showed the best success rate among all methods and the best ranking overall, suggesting that the integrative approach followed is the suitable for modelling the interplay between cell signalling and GRNs in the context of cellular transitions.

4.6 Cell fate transition examples

The ability to induce the transition between different cell types opens the door to advances in regenerative medicine, as it could be used to replace damaged tissues and organs, both ex and in vivo. INCanTeSIMO was applied to datasets where single growth factors or chemical compounds induced changes in cellular identity. Compared to other datasets, here the GRNs associated to the cellular conversion were larger (on average 38 vs 23 GRN TFs), and the method achieved a better performance. Direct targets of the experimental perturbation were found among the candidate signalling molecules in 75% of all datasets (Table 4.1).

4.6.1 Differentiation

Nine differentiation cases were considered. Human mesenchymal stromal cells differentiate into chondrocyte when treated with either BMP2 or TGF-β3 (Mrugala et al. 2009). INCanTeSIMO predicted the activation of BMP receptor 2 and TGFBR3 when applied to the BMP2-treated gene expression data, and the activation and inhibition of other members of the TGF-β protein superfamily, which plays a key role in chondrocyte differentiation, when analysing TGF-β3-treated gene expression.

INCanTeSIMO also correctly predicted targets for the differentiation of hematopoietic stem/progenitor cells to erythroid and megakaryocytic precursors (Zini et al. 2012), for the terminal differentiation of neonatal keratinocytes (Q. T. Tran et al. 2012), and for the induction of hepatoblasts differentiation towards hepatocyte-like cells (Ogawa et al. 2013). The differentiation of myeloid-derived suppressor cells into M2-like macrophages (Wang et al. 2015) and of intestinal stem cells into secretive progenitor cells (T.-H. Kim et al. 2014) were also correctly associated with experimentally perturbed signalling molecules.

4.6.2 Cell activation and maintenance

Among the cellular conversions considered there were cases of specification of a new cellular fate. The activation of pre-adipocytes to primed adipocytes was correctly associated with the activation of DAX1, a nuclear receptor for steroid hormones (Tomlinson et al. 2010). The specification of mesenchymal stem cells towards the subendothelial murate cell fate is activated by TGF-β1 treatment and impeded by bFGF (Sacchetti et al. 2007). In accordance with these observations, our method predicted bFGF targets and regulators as involved in both processes.

Table 4.1: Signalling molecules predicted in cell fate transitions

Type	Initial cell type	Perturbation	Final cell type	Ref.	Best rank	Predicted direct targets	notes
D	hMSC	BMP2	chondrocytes	(Mrugala et al. 2009)	**10**	ALK-2 Chordin__inh BMP receptor 2 Noggin__inh Ectodin__inh	TGF-β3 targets predicted: TGF-beta receptor type III (betaglycan)
D		TGF-β3			**38**	Endoglin__inh	BMP2 targets predicted: Chordin | Noggin | Ectodin | PTCH1__inh
D	HSPC	valproic acid	Erythroid and megakaryocytic precursors	(Zini et al. 2012)	**5**	HDAC9__inh HDAC2	
D	NHEK	density-induced differentiation, treated with	terminally differentiated keratinocytes	(Tran et al. 2012)	**1**	ErbB4 MSK1	
D	hepatoblasts	cAMP	hepatocyte-like cells	(Ogawa et al. 2013)	**143**	Protein kinase G1	
D	mMDSC (Myeloid-derived suppressor cells)	R848	tumoricidal M1-like macrophages	(Wang et al. 2015)	-		TLR7__inh | TLR8__inh (opposite sign)
		PAM3	immunosuppressive M2-like macrophage		**78**	TLR1	
D	Intestinal stem cells (ISC)	Atoh1 inhibition	enterocyte progenitors	(Kim et al. 2014)	-		
		dibenzazepine	secretive progenitors		**56**	NOTCH1 (NICD)__inh	
D	Dermomyotome	CHIR99021	myotome	(Nakajima et al. 2018)	-		
A	pre-adipocytes	dexamethasone	primed pre-adipocytes	(Tomlinson et al. 2010)	**93**	DAX1	
A	mesenchymal stem cells	bFGF	non-HME cells	(Sacchetti et al. 2007)	**24**	Casein kinase II, alpha chains Casein kinase II, alpha' chain (CSNK2A2)	
A		TGF-β1	subendothelial mural cell fate		**103**	Ubiquitin	bFGF targets predicted: Syndecan-3__inh | Casein kinase II, alpha' chain (CSNK2A2)__inh | S100B__inh
M	hES-T3	activin A + bFGF	-	(Tsai et al. 2010)	**86**	ALK-4	Protocols comparison: MEF feeder
M			-		**15**	ALK-4 ALK-7 ALK-2__inh	Protocols comparison: feeder-free
R	MEFs (Mouse Embryonic Fibroblasts)	SB-431542	Astrocytes	(Tian et al. 2016)	**41**	TGF-beta 1__inh	
R	Mouse embryonic fibroblasts	CHIR99021 + RepSox + Forskolin + valproic acid	cardiomyocytes	(Fu et al. 2015)	**166**	RepSox: JNK1(MAPK8)__inh	
R	Adult fibroblasts	SP600125 + SB202190 + Go6983	hMSC	(Lai et al. 2017)	**89**	SP600125 : p38beta (MAPK11)__inh; JAK3__inh; MSK1__inh SB202190: p38beta (MAPK11)__inh; p38alpha (MAPK14)__inh	Go6983: cPKC (conventional) (opposite sign)

Maintenance of cell identity sometime also requires supplementing the culture medium with growth factors or compounds. Maintenance of hESC-T3 is achieved with standardized protocols (MEF feeder or feeder-free) or in conditioned medium with the addition of activin A (Tsai et al. 2010). Comparing the latter with established protocols, INCanTeSIMO correctly predicted the activation and inhibition of multiple activin receptors.

4.6.3 Transdifferentiation

The transdifferentiation of mouse embryonic fibroblasts into astrocytes is obtained experimentally with SB-431542 (Tian et al. 2016), an inhibitor of the TGF- β type I receptor kinase activity (Laping 2002). Correctly, it was predicted that the inhibition of TGF-β1 would induce this cellular transition. Additionally, INCanTeSIMO was tested on two cases of transdifferentiation that were obtained with the combination of multiple chemical compounds.

The direct conversion of mouse embryonic fibroblasts to cardiomyocytes can be induced with the minimal combination of four distinct compounds (CHIR99021, RepSox, Forskolin, and valproic acid). INCanTeSIMO only predicted a direct target of RepSox, but further inspection into the candidate signalling molecules revealed the activation of Axin, which is a target of both GSK3, which is inhibited by CHIR99021, and of G-protein alpha-s, a target of Forskolin. On the other hand, no direct or indirect target of valproic acid were among the candidates.

Human dermal fibroblasts are converted in mesenchymal stem cells with the minimal combination of SP600125, SB202190 and Go6983 (Lai et al. 2017). In this context, INCanTeSIMO correctly predicted three targets of SP600125 (the inhibition of p38, JAK3 and MSK1) and two SB202190 targets (the inhibition of p38 in its α and β forms). Regarding Go6983, our method predicted the activation, instead of the inhibition, of protein kinases C.

In summary, signalling molecule perturbations inducing cell fate conversions were consistently captured by INCanTeSIMO. Notably, the method was able to predict alternative perturbations for the same conversion, as in the case of mesenchymal stromal cells differentiated into chondrocytes, and mutually exclusive perturbations, with the prediction of activation or inhibition of bFGF signalling molecules to induce or inhibit the specification of subendothelial murate cell fate in mesenchymal cells. This suggests that among the predicted candidates are recapitulated both experimentally known perturbations, but also other biologically relevant ways of inducing the same cellular conversion.

4.6.4 Pathway-level predictions

As mentioned previously, in order to assess the performance of INCanTeSIMO_path the positive pathways for each dataset were determined from the direct perturbation targets, and the method could correctly predict them in 58% of the datasets. In order to confirm the overall good quality of the predictions, literature evidence was collected that could clarify if pathways that are not related to the experimentally applied perturbation, but are predicted by INCanTeSIMO_path, could induce the cellular transitions studied. In general, the pathways predicted were previously implicated in the cellular conversion analysed, as shown in Table 4.2. A few examples are discussed here:

- While not containing targets of the experimental perturbation, the inhibition of NOTCH signalling pathway was reported previously to improve the differentiation of hMSCs to chondrocytes (Sun et al. 2018). Other correct predictions for this transition include the activation of ossification (Su et al. 2018), response to hypoxia (Kanichai et al. 2008; Koay and Athanasiou 2008) and androgen receptor (S.-s. C. Huang et al. 2013) pathways.

- The differentiation of dermomyotome into myotome has been previously associated with the activation of insulin (Pirskanen, Kiefer, and Hauschka 2000) and inhibition of BMP signalling (Reshef, Maroto, and A. B. Lassar 1998), which contain the targets for the perturbation experimentally tested (CHIR99021). Among our predictions, also ERK5 signalling (Carter et al. 2009; Delfini et al. 2009) and the inhibition of TGF-β signals (J. Zhou and Sears 2018) have were supported by literature evidence.

Table 4.2: Top pathways predicted by INCanTeSIMO_path for cell fate transitions

Type	Transition	Perturbation	Bool	# pathways predicted	Predicted pathway	Sign	Rank	Contains perturbation targets	Literature
D	hMSC -> chondrocytes	BMP2	frmaBool	19	IL-2	+	1	no	
					T cell receptor	+	2	no	
					MIF	-	3	no	10.1016/j.stemcr.2016.07.003
					ossification	+	18	yes	10.3390/ijms19082343
D	hMSC -> chondrocytes	TGF-beta 3	frmaBool	15	response to hypoxia	+	1	no	[1] 10.1016/j.joca.2008.04.007; [2] 10.1002/jcp.21446
					synaptogenesis	+	2	no	
					NOTCH	-	3	no	10.1038/s12276-018-0151-9
					Androgen receptor cross-talk	+	13	yes	10.1016/j.scr.2013.06.001
D	dermomyotome -> myotome	CHIR99021	geneDE	18	ERK5	+	1	no	[1] 10.1016/j.ydbio.2009.05.544; [2] 10.1242/jcs.045757
					neurogenesis	-	2	no	
					antigen presentation	+	3.5	no	
					TGFb	-	3.5	no	10.1002/dvdy.24681
					insulin	+	5	yes	10.1006/dbio.2000.9784
					BMP in cardiac development	-	14	yes	10.1101/gad.12.3.290
D	mMDSC (Myeloid-derived suppressor cells)-> tumoricidal M1-like macrophages	R848	geneDE	19	Innate inflammatory response	-	1	no	
					IL-10	+	2.5	no	10.1189/jlb.3A0414-210R
					JAK/STAT	+	2.5	no	10.3389/fimmu.2018.00608
					response to RNA viral infection	-	4	no	
					Th cell differentiation	+	10	yes	10.3389/fimmu.2019.00219
D	mMDSC->immunosuppressive M2-like macrophage	PAM3	geneDE	10	Skeletal muscle development	+	1	no	
					Wnt	+	2	no	10.1073/pnas.0509703103
					TGFb	+	3	no	10.18632/oncotarget.10561
D	Intestinal stem cells (ISC) -> enterocyte progenitors	Atoh1 inhibition	geneDE	19	Phagosome in antigen presentation	-	1	no	10.1155/2017/7970385
					phagocytosis	-	2	no	
					death receptors	+	3.5	no	
					response to RNA viral infection	-	3.5	no	
					EMT	+	9	yes	10.15252/embj.201591517
D	ISC -> secretive progenitors	dibenzazepine	geneDE	9	Phagosome in antigen presentation	-	1	no	10.1155/2017/7970385
					Neutrophil activation	-	2	no	
					phagocytosis	-	3.5	no	10.1007/s00427-012-0422-8

					amphoterin	+	3.5	no	
D	HSPC -> erythroid and megakaryocytic precursors	valproic acid	frmaBool	24	death receptors	+	2.5	no	10.1002/jcp.25967
					neurogenesis	-	2.5	no	
					NK cell cytotoxicity	-	2.5	no	
D	NHEK -> terminally differentiated keratinocytes	density-induced differentation + EGF	frmaBool	11	T cell receptor	-	1	no	
					protein C	+	2	no	10.4331/wjbc.v5.i2.169
					NK cell cytotoxicity	-	3	no	14760888
D	hepatoblasts -> hepatocyte-like cells	cAMP	frmaBool	19	ossification	-	1	no	
					MIF	+	2	no	
					antigen presentation	+	3	no	
T	MEFs (Mouse Embryonic Fibroblasts)->Astrocytes	SB-431542	geneDE	23	insulin	-	1	no	
					IFN-gamma	+	2	no	10.3389/fimmu.2015.00539
					antigen presentation	+	3	no	10.1016/j.stemcr.2018.09.015
					angiogenesis	-	15	yes	10.1074/jbc.M109.006551
A	MSC -> non-HME cells	bFGF	frmaBool	38	MIF	+	2	no	
					Phagosome in antigen presentation	-	2	no	10.1002/stem.406
					amphoterin	+	2	no	
A	MSC -> subendothelial mural cell fate	TGFb	frmaBool	35	neuropeptides	-	3	no	
					NOTCH	-	3	no	
					Neuromuscular junction	-	3	no	
					B cell receptor	-	3	no	
					IL-2	+	3	no	
A	pre-adipocytes -> primed pre-adipocytes	dexamethasone	frmaBool	28	ossification	-	2	no	10.1038/cdd.2015.168
					response to RNA viral infection	+	2	no	10.1210/en.2009-1140
					neurogenesis	+	2	no	
					MIF	-	10	yes	10.1038/emm.2015.26
					Neutrophil activation	+	15	yes	
					Th cell differentiation	-	16	yes	10.4049/jimmunol.1001269
					Neutrophil activation	-	18	yes	

- Myeloid derived suppressors cells (MDSCs) can differentiate towards M1- or M2-like macrophages. INCanTeSIMO_path correctly predicted that IL-10 and JAK/STAT signalling are involved in M1 differentiation (Bayik, Tross, and Klinman 2018; Beury et al. 2014) and it associated TGF-β and Wnt pathways to M2-like differentiation (Pukrop et al. 2006; F. Zhang, R. Liu, and J. Zheng 2016).

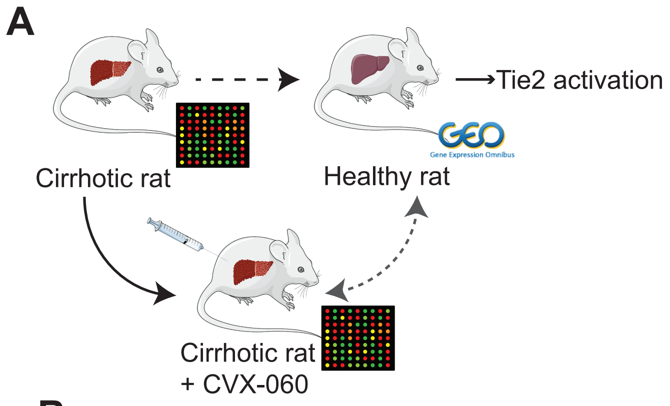
Overall, many of the pathways predicted, though not associated with the experimental perturbations, could be involved in the cellular transitions studied. These results show that INCanTeSIMO_path is able to predict pathways involved in cellular transitions independently of how the transition was induced in the experiments that generated the gene expression data used for the prediction.

4.7 Applications to animal models of disease and regeneration

The methods presented here were applied to the prediction of signalling molecules and pathways involved in the transition from a pathological to a healthy state in a rat model of cirrhosis, obtained with the treatment with CCl_4. Secondly, the molecules and pathways involved in the regeneration of amputated salamander limbs were identified by applying the present methods to time series gene expression data.

4.7.1 Reversion of cirrhotic state in rat liver

Currently liver transplantation is the only effective therapy available for cirrhotic patients, and new therapeutic strategies are urgently needed, also in the prospect of avoiding further advancing of the disease towards hepatocellular carcinoma (HCC). Wistar rats treated with CCl_4 represent a classical disease model for cirrhosis. Gene expression data from cirrhotic and healthy rat liver was used to predict which signalling perturbations could induce the shift between diseased and healthy tissue state (Figure 4.7.1A).



B

GRN TF	Booleanized cirrhotic data	Booleanized healthy data	Predicted state after ideal perturbation	Predicted CVX-060 treatment state	Booleanized CVX-060 treated data
EBF1	1	0	0	0	0
EGR3	1	0	0	0	0
PRDM1	1	0	0	0	0
RUNX3	1	0	0	0	0
ANKRD1	1	0	0	0	1
FGF2	1	0	0	0	1
FOSL1	1	0	0	0	1
FOSL2	1	0	0	0	1
GREB1	1	0	0	0	1
GRHL1	1	0	0	0	1
RELB	1	0	0	0	1
RXRA	1	0	0	0	1
TP63	1	0	0	1	0
POU2F2	1	0	0	1	1
SP7	0	1	1	1	1
MYBL2	0	1	1	1	0
NR0B1	0	1	1	1	0
REST	0	1	1	1	0
VDR	0	1	1	1	0
RARG	0	1	0	0	1
HOXA1	0	1	0	0	0
MYOG	0	1	0	0	0
NEUROG1	0	1	0	0	0
PBX1	0	1	0	0	0
SALL4	0	1	0	0	0
SOX2	0	1	0	0	0

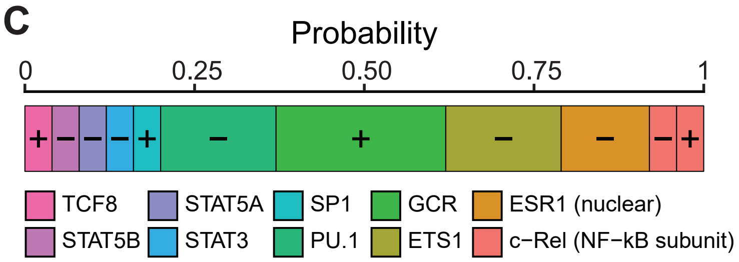


Figure 4.15: Prediction of signalling perturbations for the reversal of the cirrhotic phenotype in rat liver. **A)** Overview of the approach followed: liver gene expression data of cirrhotic rats were generated by collaborators, and used as the initial cellular state. Gene expression data for the desired healthy state was obtained from GEO. Once the activation of Tie2 was predicted by INCanTeSIMO, cirrhotic rats were treated with its agonist CVX-060, gene expression data was generated from treated liver and compared with the healthy state. **B)** The state of the TFs in the GRN in the cirrhotic, healthy and CVX-060 treated samples. The ideal perturbation state refers to the state that the GRN TFs can reach if any of the BPCs is applied. The predicted CVX-060 treated state is the state the GRN-TFs can have if the BPCs composed only of interface TF states induced by the activation of Tie2, according to INCanTeSIMO (using correlation-based MPPs). Green background shows when a state is matching the desired healthy state. **C)** Interface TFs present in the BPCs and their relative probability of inducing the desired changes on the GRN. +: the activation of the interface TFs has desired effect on the GRN; –: its inhibition is effective. The two states are not mutually exclusive, as seen in c-Rel (NF-kB subunit).

Gene expression of whole cirrhotic livers was quantified with microarray experiments, and compared with healthy liver data obtained from public databases (see Materials and Methods section 3.4). The GRN model representing the cellular transition between cirrhotic and healthy state (disease GRN) consisted of 26 TFs (Figure 4.7.1B), and 106 interface TFs were available for perturbation. After exhaustive in silico perturbation, the BPCs were predicted to affect the state of 19 GRN-TFs (Figure 4.7.1B), and were composed of 10 interface TFs (Figure 4.7.1C).

Signalling molecules were ranked using INCanTeSIMO. Among the candidate molecules many proteins are known to be involved in different aspects of cirrhosis, or other liver diseases such as fibrosis, fatty liver disease and HCC. As mentioned in the Introduction (1.4.1), fibrosis, inflammation and blood vessel structure are fundamental biological processes involved in the development of liver cirrhosis.

Among the predictions obtained, there was the inhibition of proteins associated with fibrosis (e.g. CHIP, AP-1, CBP, MDM2) and matrix metallopeptidases responsible for matrix remodelling, and the activation of ESR2 which is known for its antifibrogenic role (B. Zhang et al. 2018). Additionally, multiple innate immune response proteins, including interleukins, were correctly predicted (W.-C. Zhou 2014). Among the canonical pathways, BMP signalling was predicted as inhibited by INCanTeSIMO_path. Different BMP proteins act differently on hepatic stellate cells (HSCs), BMP2 and BMP4 in particular potentiate their transdifferentiation and have a profibrogenic effect (Herrera, Addante, and Sánchez 2017).

Table 4.3: Pathways predicted for the reversal of the cirrhotic state. Only pathways with more than 3 candidates are presented.

Pathway	Sign	p-value	Candidates	candidates	Literature	role
BMP	-	0.01	11	CBP__inh; ALK-1__inh; CDK4; BMP15__inh; BMPR1B__inh; BMP4__inh; BMP receptor 2__inh; BMP6__inh; BMP7__inh; GDF5__inh; RUNX2__inh	29295498	HSC transdifferentiation
ossification	-	0.0015	14	CBP__inh; C/EBPbeta__inh; CSF1__inh; BMPR1B__inh; BMP4__inh; BMP receptor 2__inh; BMP6__inh; BMP7__inh; GDF5__inh; IGF-2__inh; TWIST1__inh; GSK3 beta; Dsh__inh; RUNX2__inh; EGFR__inh	"	"
cartilage	-	0.044	7	BMPR1B__inh; BMP4__inh; BMP receptor 2__inh; BMP6__inh; GDF5__inh; RUNX2__inh; CTGF__inh	"	"
insulin	-	0.0115	7	IRS-1__inh; JAK2__inh; Tuberin; Insulin receptor__inh; Furin__inh; H-Ras__inh; GSK3 beta	10.1530/JOE-15-0409	wrong sign
Nitric oxide	-	0.021	5	GM-CSF receptor__inh; JAK2__inh; H-Ras__inh; NO intracellular__inh; GM-CSF__inh	26027855	HSC activation and proliferation
AR cross-talk	-	0.025	15	CBP__inh; MDM2__inh; Androgen receptor; EGF__inh; ERK2 (MAPK1)__inh; JAK2__inh; Tuberin; Caveolin-1__inh; PAK6__inh; H-Ras__inh; GSK3 beta; Dsh__inh; EGFR__inh; HB-EGF__inh; Neuregulin 1__inh; IL-6 receptor__inh	10.1002/hep.26135	fibrosis and HSC activation
ESR2	+	0.0325	6	ESR2; MMP-26__inh; CHIP__inh; EGF__inh; Corticoliberin__inh; Insulin receptor__inh; IGF-2__inh; H-Ras__inh; EGFR__inh; OT-NPI__inh	28884481	HSC activation and proliferation
NADPH	-	0.045	9	Pitx2__inh; ERK2 (MAPK1)__inh; Caspase-3; Tuberin; H-Ras__inh; GSK3 beta; NF-AT1(NFATC2)__inh; H(,2)O(,2) intracellular__inh; p67-phox__inh	15915457	HSC activation and proliferation
antigen presentation	+	0.0355	9	NFYA; CIITA; CD86; ICAM3; JAK2__inh; TNF-R1; NFKBIE; NF-kB2 (p100); NKG2C		

Nitric oxide and NADPH signalling are also predicted as inhibited, in accordance to the role of oxidative stress in HSC activation and proliferation (Adachi et al. 2005; Iwakiri and M. Y. Kim 2015)(Table 4.3).

Members of angiopoietins signalling, which is a key pathway in blood vessel normalization, are also present among the predictions. Signalling through Angiopoietin 1 (Ang1)-Tie2 is known to stabilize blood vessels, while Angiopoietin 2 (Ang2) acts as a context-dependent antagonist of Ang1, decreasing its effect and giving rise to immature blood vessels (Fagiani and Christofori 2013). Higher expression and activity of Ang2 has been associated with cirrhotic conditions. In agreement with this characteristics, INCanTeSIMO predicted the activation of Angiopoiein 1 and 4, the inhibition of Angiopoietin 2 and 3, and the activation of Tie2 (ranking 24th among all signalling molecules) (Table in Appendix 7.4).

The activation of Tie2 is predicted to induce activation of interface TFs SP1 and ETS1, and the inhibition of GCR, STAT5A and B, ESR1, and PU.1 (Figure 4.16), resulting in a GRN state that matches partially the healthy liver state (Figure 4.7.1B). This prediction was validated by treating cirrhotic rats with CVX-060, a specific inhibitor of Ang2 which improves Tie2 activity, and generating gene expression data for the whole liver. The GRN-TFs EBF1, EGR3, PRDM1, RARG, RUNX3, SP7, and TP63 were observed to change their expression state to match the healthy counterpart (Figure 4.7.1B). Literature evidence confirmed that many of these TFs are indeed involved in liver cirrhosis and blood vessel stabilization.

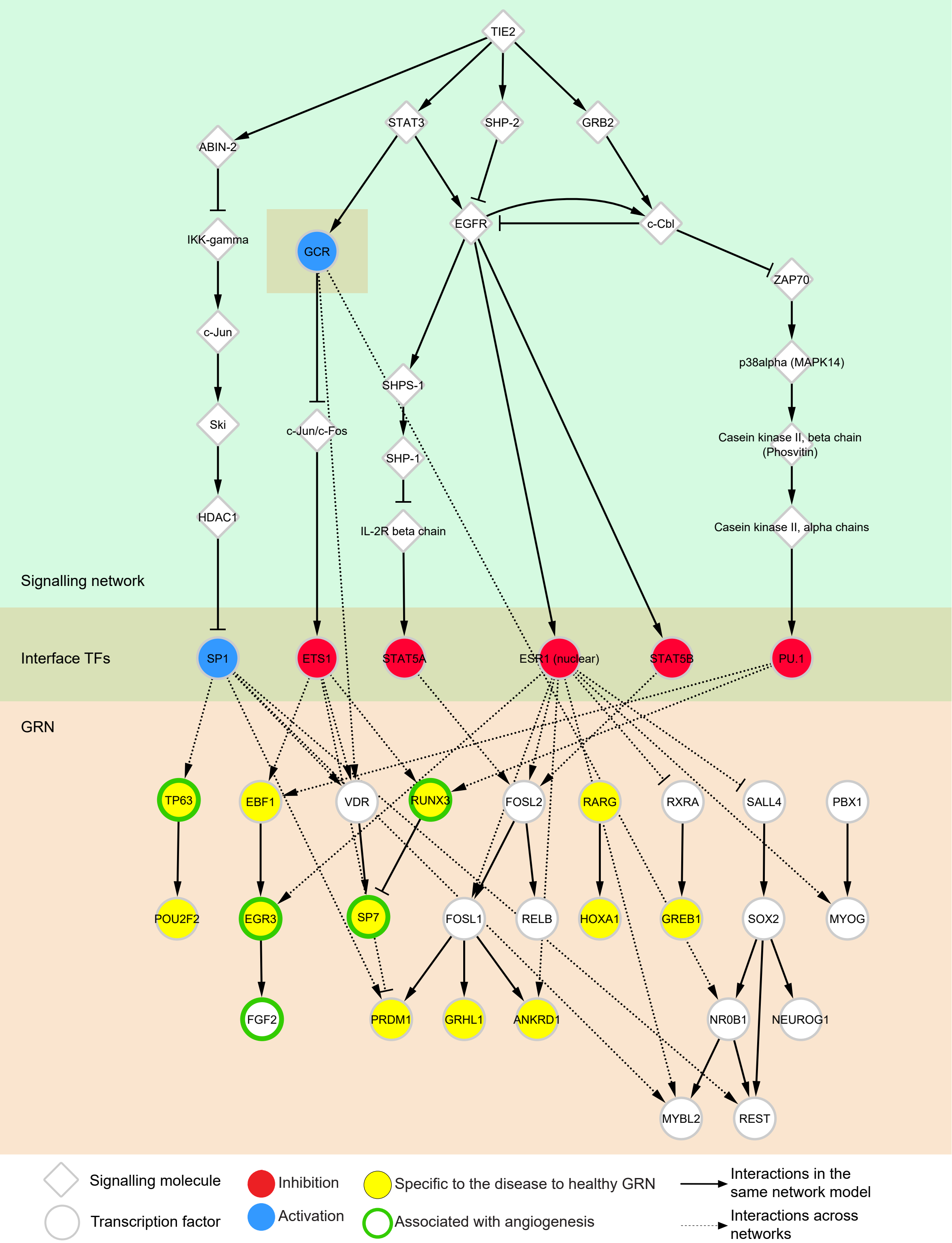


Figure 4.16: Integrated signalling and gene regulatory network for cirrhosis. The application of CVX-060 specifically activates Tie2, which in turn activates (blue circles) or inhibits (red circles) interface TFs following the MPPs depicted. The interface TFs act on the disease-vs-healthy GRN (disease GRN). In the GRN, white TFs are common between this GRN and the treated-to-healthy GRN (treatment GRN), and yellow TFs are specific for the disease GRN. Association of the TFs with the regulation of angiogenesis is represented with a green border.

EGR3 is a pro-inflammatory factor that can stimulate fibrosis when overexpressed, and whose suppression has been shown to attenuate TGF-β signalling and consequently fibrogenesis (Fang et al. 2013). It also plays an essential role in VEGF signalling in angiogenesis (D. Liu, D. Ghosh, and Lin 2008). Runx3 is expressed in embryonic liver and regulates fetal hematopoiesis (Bruijn and Dzierzak 2017). Its inhibitions causes organogenesis defects in mice, and excessive intrahepatic angiogenesis (J.-M. Lee et al. 2013). SP7 and TP63 have been implicated in VEG-mediated angiogenesis (Farhang Ghahremani, Goossens, and Haigh 2013; W. Tang et al. 2012). Retinoic acid signalling through RARG was shown to reverse hepatic stellate cell activation and fibrosis (Panebianco et al. 2017). PRDM1 and EBF1 do not have a reported function in angiogenesis or cirrhosis to date, but have been connected to other liver diseases: PRDM1 expression has recently been implicated with HCC (N. Li et al. 2019), and EBF1 was observed to be differentially methylated in nonalcoholic fatty liver disease cirrhosis (Gerhard et al. 2018). They could therefore represent novel therapeutic targets.

To further consolidate the results obtained, a second GRN representing the transition between the treated and healthy states was built (treatment GRN). It contained 30 GRN TFs, of which 14 in common with the disease-to-healthy GRN (disease GRN). The 12 TFs unique to the disease GRN were localized in network-specific modules, that contained TFs implicated in vascular growth, including EGR3, RUNX3, SP7 and TP63 discussed above (Figure 4.16). Modules that were shared between the two GRNs did not contain TFs associated to angiogenesis, confirming that by activating Tie2 the expression state of TFs regulating angiogenic processes was reverted to its healthy counterpart.

A previous functional study showed that inhibition of Ang2 and activation of Tie2 signalling through CVX-060 treatment are beneficial to the stability of intrahepatic blood vessels and cause a reduction of liver inflammatory infiltrate, overall improving the fibrotic condition in cirrhotic rats (Pauta et al. 2015). The current analysis complements these previous results by providing insights in the gene expression changes occurring during the treatment.

Still, activation of Tie2 does not completely revert the disease phenotype, suggesting that multiple facets of the disease should be targeted together to induce a complete return to healthy expression profile and cellular state. For this, a combination of different candidate molecules might be considered. For example, inhibition of endothelin ranks 17th among the predictions (Table in Appendix 7.4). Endothelin-1 overexpression has been associated with the pathological activation of HSCs, which are one of the main cells responsible for collagen expression in the liver, and is also connected to exacerbated vasoconstriction leading to intrahepatic vascular dysfunction (Tsuchida and Scott L. Friedman 2017).

4.7.2 Salamander limb regeneration

The approach developed here was applied to the prediction of signalling molecules and pathways that drive the regeneration process in salamander limbs. In particular, the analysis considered the initial stages of the regeneration up to 14 days post amputation (dpa), comprising the response to amputation, formation of the wound epithelium and subsequently of the blastema. Gene expression data of cells from the mesenchymal lineage, expressing Prrx1, was collected at amputation, 1, 3, 5, 7, 10, and 14 dpa.

A regeneration-specific GRN was built around the TFs that change their state between any two consecutive time points during the time course, and then interval-specific GRNs were extracted from it. GO enrichment showed that in each GRN there were TFs implicated in the regulation of processes related to signalling, cellular differentiation and embryonic development in agreement with expectations (Table in Appendix 7.5). For each of these intervals, the signalling molecules able to induce the GRN to change from its initial to the final state were predicted (see Methods section 3.5). The analysis of the interval between day 3 and day 5 did not identify interface TFs perturbations able to change the state of the GRN substantially ($>$40% of the GRN-TFs) and this interval was therefore discarded.

Overall, the predictions revealed signalling molecules and pathways related to wound healing up to 3 dpa, followed by cellular migration and de-differentiation around 5-7 dpa, and finally cellular proliferation and re-differentiation (Table 4.4).

Table 4.4: Selected signalling molecules and pathways predicted for the different stages of salamander limb regeneration.

Signalling entity	**Time intervals**	**Role**	**Literature evidence**
NADPH and ROS signalling (both signs)	0-1d	Cellular activation and proliferation	(Al Haj Baddar, Chithrala, and Voss 2019)
HDAC2 inhibition	0-1d	Wound healing	(Taylor and Beck 2012)
Phagocytosis	0-1d, 10-14d	Wound healing	(James W Godwin, Pinto, and N. A. Rosenthal 2013)
HHs, Smoothened	0-1d, 7-14d	Cellular proliferation and migration	(B. Singh et al. 2015; B. N. Singh et al. 2012)
Wnt pathway	0-3d	Wound healing	(D. Zhang et al. 2009)
p38/JNK	0-3d, 7-14d	Wound healing, EMT	(Sader et al. 2019)
Bcl-2	0-3d,10-14d	apoptosis	(Bucan et al. 2018)
FGF receptors	0-10d	Fibroblasts de-differentiation, blastema formation	(Makanae, Mitogawa, and Satoh 2014)
ERK/MEK	0-14d	blastema formation, blastema differentiation	(Owlarn et al. 2017; Suzuki et al. 2007; Tasaki et al. 2011; Maximina H Yun, Phillip B Gates, and Jeremy P Brockes 2014)
PI3K/AKT	0-14d	blastema formation	(Suzuki et al. 2007)
GDF5	0-1d, 7-14d	Blastema formation	(Makanae, Hirata, et al. 2013)
Retinoic acid receptors	1-3d, 5-7d	Apical epidermal cap, skeletal patterning and differentiation	(J. R. Monaghan et al. 2012; Nguyen et al. 2017)
C/EBPβ	1-3d, 7-10d	macrophage functionality	(James W Godwin, Pinto, and N. A. Rosenthal 2013; Ruffell et al. 2009)
Neuregulin 1	1-3d	Blastema formation	(Farkas, Freitas, et al. 2016)
p53 inhibition	1-7d	Blastema formation	(M. H. Yun, P. B. Gates, and J. P. Brockes 2013)
ErbB2-3	5-7d, 10-14d	cellular migration	(Rojas-Muñoz et al. 2009)

Table 4.4: Selected signalling molecules and pathways predicted for the different stages of salamander limb regeneration.

Signalling entity	**Time intervals**	**Role**	**Literature evidence**
Thrombin	5-7d	Cell cycle re-entry	(Imokawa and Jeremy P Brockes 2003; Maximina H Yun, Phillip B Gates, and Jeremy P Brockes 2014)
FGF8	5-7d	Cellular proliferation	(Eugeniu Nacu et al. 2016)
MMPs	7-10d	Blastema induction	(Satoh, Makanae, et al. 2011)
BMPR1B	7-10d	Chondrocyte differentiation	(Kotzsch et al. 2009)
TNF-α	10-14d	Blastema induction	(Nguyen-Chi et al. 2017)
p53	7-14d	Blastema differentiation	(M. H. Yun, P. B. Gates, and J. P. Brockes 2013)

In the initial time point (0 to 1 dpa), the predicted candidates include signals that have been specifically associated with the initial steps of regeneration: p38/JNK signalling, which has a central role in wound closure and EMT (Sader et al. 2019), ERK/MEK and PI3K/AKT signalling, which are involved in the initiation of regeneration and blastema formation in *X. laevis* and planarians (Owlarn et al. 2017; Suzuki et al. 2007; Tasaki et al. 2011). Bcl-2 is among the predictions, in agreement with the reported importance of Bcl-2 family proteins in the regulation of apoptosis in the initial phases of limb regeneration (Bucan et al. 2018). The activation of Wnt signalling, which is generally associated with enhanced would healing (D. Zhang et al. 2009), was also predicted.

Between 1 and 3 dpa, INCanTeSIMO predicted the activation of C/EBPβ, which has been shown to play a fundamental role in regeneration by regulating macrophage functionality (James W Godwin, Pinto, and N. A. Rosenthal 2013; Ruffell et al. 2009). The predicted inhibition of NK cell cytotoxicity and T cell receptor signalling agree with the hypothesis that inhibiting the lysis of progenitor populations might be necessary for successful regeneration in salamander (James W. Godwin and N. Rosenthal 2014).

Following 5 dpa, the activation of ErbB proteins and Src/FAK signalling is predicted, in accordance with their role in cellular migration during vertebrate regeneration (Makanae and Satoh 2012; Rojas-Muñoz et al. 2009). Additionally, the predicted activation of the kallikrein-kinin signalling pathway, which includes coagulation factors and thrombin, is correctly associated to the initiation of the regenerative process (Imokawa and Jeremy P Brockes 2003).

At the next time step, between 7 and 10 dpa, there is the prediction of multiple matrix metalloproteinases (MMP-2, 9, 13, 14). Importantly, they have been implicated with blastema induction and regulation of Prrx1 expression in *A. mexicanum* (Satoh, Makanae, et al. 2011). Multiple proteins belonging to the hedgehog pathway (Hedgehog, Smoothened, PKA, GLI3) are predicted at this stage, in concordance with the observation that HH signalling is not necessary for the dedifferentiation of mature cells, but is required for their proliferation and migration (B. Singh et al. 2015; B. N. Singh et al. 2012). Insulin signalling has also been associated with this stage (Stocum and Cameron 2011).

In the limb bud stage (10 to 14 dpa), TNF-α and the downstream NF-κB signalling are predicted. TNF-α has been previously shown to play a primary role in the activation of blastema cells in zebrafish (Nguyen-Chi et al. 2017). ERK signalling is predicted also at this stage, in accordance with its activity in inducing blastema cells differentiation observed in planarians (Tasaki et al. 2011).

Limb regeneration is a process that strongly depends on combination of signalling inputs received by the blastema cells. Previous studies indicate that there is no single signal responsible for the establishment of the regenerative program, but is the right combination of different factors that leads to such biologic process. The guidance role is mainly associated with dermis and peripheral nerves (Endo, Bryant, and David M. Gardiner 2004; Satoh, David M. Gardiner, et al. 2007).

Dermis The apical epithelial cap (AEC), the most distal region of the wound epidermis, was shown to regulate the blastema growth by expressing several growth factors that stimulate cellular migration, proliferation, and changes in gene expression profiles. Among these factors are the retinoic acid and fibroblasts growth factors (FGFs) and Wnt signalling molecules.

FGFs are known to have a role in dedifferentiation of fibroblasts and in blastema formation, but they are not sufficient to induce limb patterning (Makanae, Mitogawa, and Satoh 2014). BMP2, BMP7 and growth and differentiation factor 5 (GDF5) work in coordination with FGFs to induce the formation of blastema-like structures, by attracting fibroblasts to the wound site (Makanae, Hirata, et al. 2013; Makanae, Mitogawa, and Satoh 2014). Retinoic acid, on the other hand, plays a multifunctional role both in AEC (J. R. Monaghan et al. 2012) and in limb patterning (Catherine D McCusker and David M Gardiner 2014). Wnt signalling is involved in wound healing and positional information in the regenerating limb (S. Ghosh et al. 2008; D. Zhang et al. 2009).

Nerves While the development of embryonic limbs is nerve-independent, regeneration in adult salamanders depends on the innervation of the amputated limb. Limb denervation limits blastema cells proliferation, an effect mediated by the interaction of the wound epithelium with the regenerating nerve. The presence of nerves is necessary for scar-free wound healing, but a higher nerve signalling is necessary (and sufficient) for blastema formation (C. McCusker, Bryant, and David M. Gardiner 2015). Regenerating axons are thought to produce factors that play a key role in cell proliferation. A number of factors have been identified that are present in blastema cells, can rescue the regeneration in denervated limbs, and whose inhibition hinders regeneration. They are anterior gradient protein (AG) (Kumar and Jeremy P. Brockes 2012), BMP2, BMP7, FGF2 and FGF8 (Makanae, Mitogawa, and Satoh 2014), Neuregulin 1 (Farkas, Freitas, et al. 2016), and transferrin (Mescher et al. 1997). Combinations of these factors have been shown to effectively substitute the effect of innervation on the regenerative process. Treatment of skin wounds with BMP7, FGF2 and FGF8 induces blastema formation (Makanae, Mitogawa, and Satoh 2014), and the addition of FGF2, FGF8, and BMP2, followed by RA treatment, is able to mimic the signalling cues brought to limb wounds by nerve and positional identity, inducing the regeneration of complete limbs (Vieira and Catherine D. McCusker 2019).

With the exception of AG protein and transferrin, not present in the signalling network considered, these factors associated with both dermis and nerves signalling were consistently recovered during the analysis: Wnt and retinoic acid signalling were selected at initial time points, Neuregulin 1 was identified between day 1 and 3, while FGF and BMP signals were predicted for later stages.

Multiple other predictions could be confirmed by literature review. Both the activation and inhibition of the NADPH and ROS signalling pathway were selected at 0 to 1 dpa, supporting the idea that ROS are necessary for inducing regeneration, but at the same time some counteracting processes might be in action in order to protect the progenitor cells from excessive cellular stress (Miller, Johnson, and Whited 2019). Finally, as with immune response, phagocytosis was predicted as activated initially, then inhibited at 5-7 dpa, and activated again after day 10. Together with he activation of C/EBPβ at two different time points, these results suggest that immune response mechanisms might play multiple roles along the regenerative process.

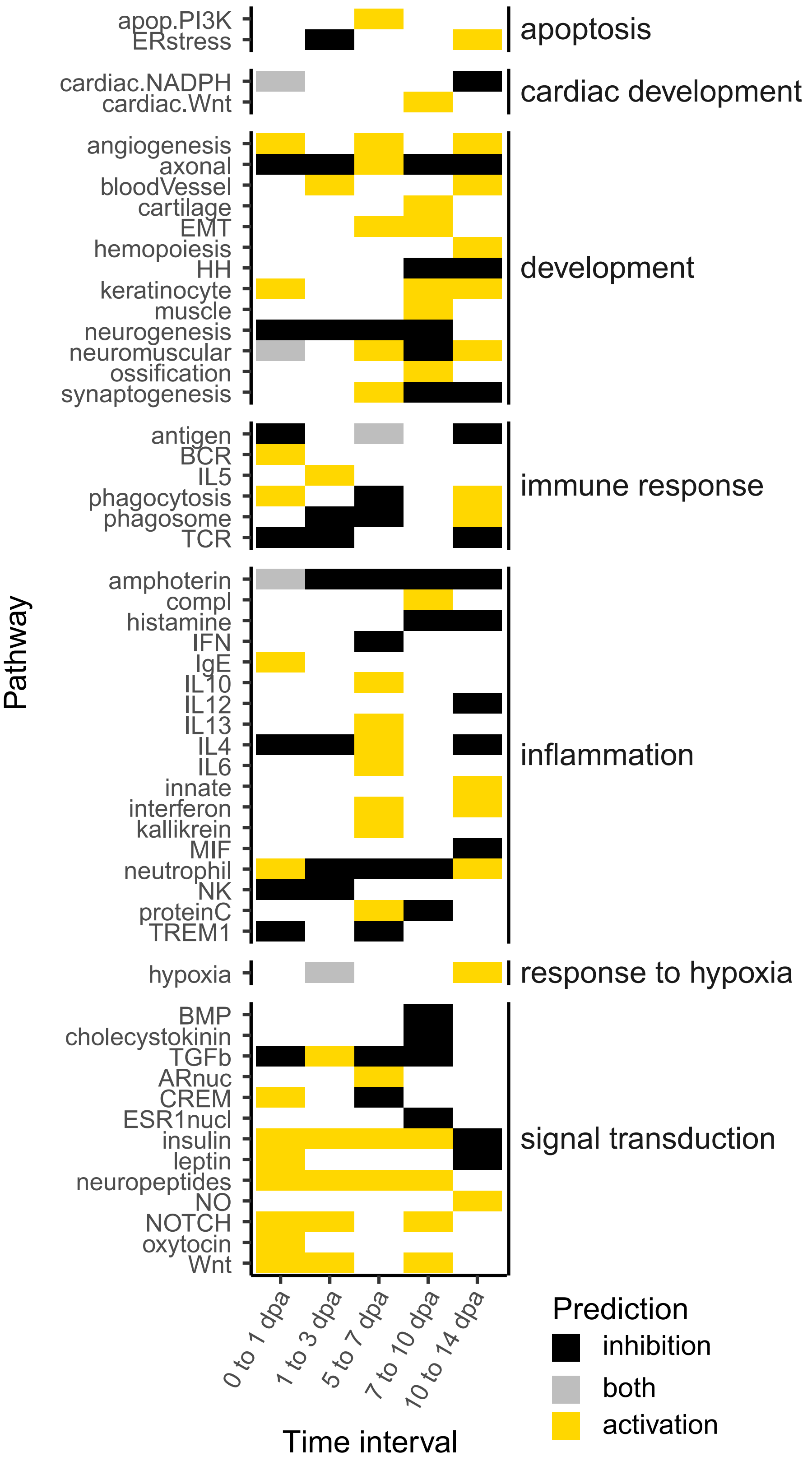


Figure 4.17: Signalling pathways predicted by INCanTeSIMO for each time interval along the regenerative process. Complete pathway names are in the Appendix 7.1.

In summary, INCanTeSIMO and INCanTeSIMO_path predictions could identify known pathways and proteins involved in the initial stages of salamander limb regeneration. Regarding signalling pathways, canonical pathways involved in signal transduction were predicted all along the regeneration process, and multiple pathways involved in development were predicted starting from the fifth day post amputation 4.17. In some cases, candidate proteins had been associated with regeneration in other animal models such as newts, zebrafish or planarians. For other proteins, it was observed that changes in their gene expression or protein abundance occur at specific regeneration stages, but their functional role has to be determined yet. These proteins constitute novel candidates for experimental validation in *A. mexicanum*. As it is well established that some of the signals involved in the regenerative process are not secreted by the analysed blastema cells (Prrx1+), an additional refinement of the candidates for experimental validation could consist in identifying the molecules expressed by specific surrounding cell types, such as macrophages, neurons or epidermal cells of the AEC.

5 Discussion and perspectives

Signalling pathways exert their functions by transmitting stimuli and inducing cellular responses to them. In particular, cell signalling can regulate the cellular gene expression program by modulating the activity of transcription factors. The regulatory activity of signalling on GRNs can be harnessed in order to trigger the shift between cellular states for the purpose of disease treatment or regenerative medicine.

Disease treatment in general involves the use of drugs that control the activity of specific proteins or small molecules, improving cellular functionality. Often, drugs targets are involved in signal transduction and their identification relies on the identification of differentially expressed pathways or of upstream regulators of genes that show differential expression. Regenerative medicine consists of multiple approaches for the restoration of functional tissues and organs. Thanks to a combination of technological advances and improved understanding of cellular fates determination, it is now possible to obtain almost any cell type for transplantation from easily accessible cells, such as fibroblasts (J. Xu, Du, and Deng 2015). The possibility of inducing cellular transitions with small molecules overcomes the safety risks connected to DNA delivery methods and promises to be less expensive, non immunogenic, and easily optimized (Pesaresi, Sebastian-Perez, and Cosma 2019).

The issue of identifying which chemical compounds can induce the intended cellular transitions has been so far addressed by leveraging knowledge on the desired cellular type and signalling pathways related to it, or by screening chemical compounds for the desired effect (De et al. 2017). Both strategies are inefficient, and thus the development of computational methods for the selection of signalling targets is desirable. The approach presented here is a first attempt at addressing this issue in a systematic way.

Available methods for the prediction of drivers of cellular conversions predict transcription factors that are involved in the cellular identity desired or play a key role in the gene expression program of the initial or desired cell type (Cahan et al. 2014; D'Alessio et al. 2015; Rackham et al. 2016; Okawa et al. 2016). On the other hand, computational methods using gene expression data that focus on signalling either pinpoint signalling pathways that are activated or inhibited by a perturbation (being it a compound, or a disease condition), or identify the specific molecules targeted by such perturbations (mode of action proteins in the case of drugs, or disease-specific genes), without considering the regulatory role of signalling on gene expression.

Thus, the primary contribution of this thesis is the introduction of a generally applicable approach to predict which signalling perturbations can induce the transition between an initial and a desired cellular gene expression state. Compared to existing tools, this approach might predict the target of an experimentally applied perturbation, but aims more generally at identifying any signalling perturbation that could induce the same shift in gene expression program. In other words, its purpose is not to describe how cells respond to stimuli, but to find which signalling stimuli could induce them to reach a desired state.

5.1 Integration of signalling and transcriptional networks

The framework used in this work differs from existing approaches by explicitly modelling the interaction between signalling networks and GRN. The two regulatory layers are represented using different formalisms, reflecting the inherent differences in terms of mechanism of action, time-scale, and uncertainty associated with gene expression data.

The gene expression regulation was represented using a Boolean network model. The GRN consists of the TFs that change their expression Boolean state between the initial and desired cellular states, with the aim of capturing the cellular shift required for the desired cellular transition. The inference of the correct GRN topology for each cellular transition takes advantage of transcriptional interactions manually curated from literature (Crespo et al. 2013). This general strategy is suitable for any cellular transition because it contextualizes the network of potential transcriptional interactions to the gene expression profile of the initial and target cellular states, without requiring large amounts of data, manual curation, or parameter estimation.

Gene expression levels have been used previously as measures of signalling activity, in methods that differ from the approach proposed here in terms of aims and integration with transcriptional regulation. For instance, in PATHiWAYS (Sebastian-Leon et al. 2014), signalling pathways were integrated with cellular functions such as cell cycle, growth, proliferation, survival or apoptosis, and differential activation of pathways estimated from gene expression data was used to interpret disease mechanisms.

Because the aim of this work is to induce transitions between stable cellular states, it can be assumed that the proteins present in the initial stable state are actively expressed and not inherited from previous cellular states, so their gene expression should be detectable. Thus, the probability of a gene being expressed in the initial cellular state was used to define which signalling proteins might be exploited for signal transduction, resulting in the definition of most probably expressed signalling paths (MPPs). The inferred MPPs were consistently supported by phosphorylation state changes in cellular transitions where phosphoproteomics data was available. This result, although limited to six examples, indicated that it is reasonable to use gene expression data to analyse signalling events, in the absence of perturbation or phosphoproteomics data.

5.1.1 Prediction of signalling molecules for cellular transitions

At the interface between the signalling and GRN models, lie TFs that are regulated by signalling pathways and regulate the expression of other genes, among which the TFs present in the GRN. This is where the integration between the two regulatory layers takes place. In INCanTeSIMO, the perturbations of interface TFs that best cause a shift from the initial to desired GRN states are selected, and the initial signalling state is used to define which signalling molecules are most likely to induce them following MPPs. Therefore, signalling molecules are ranked according to their probability of inducing the desired GRN state. This is in contrast to other GRN-based approaches such as DeMAND (Woo et al. 2015) that only take into account the topology of the GRN, without considering collective changes in TF expression induced by signalling cues. Specificity of action guides the selection of signalling molecules: molecules that act on few effective interface TFs are preferred to others that have an indiscriminate action.

Small molecules and chemical compounds are increasingly used to induce cellular transitions, however their mechanism of action remains in large part poorly understood, with 60% of existing approved drugs not having annotated targets and only half of all known drug-target pairs associated with inhibiting or activating effect (Sawada et al. 2018). For this reason, this work did not focus on identifying chemical compounds or small molecules, and instead INCanTeSIMO was developed with the objective of considering each signalling molecule in the network as equally capable of inducing the desired cellular transitions.

In the light of this incomplete knowledge, the success rate obtained by INCanTeSIMO in predicting direct targets of experimental perturbations (61%) is satisfactory. Furthermore, the method outperformed other available approaches, such as Connectivity Map and DeMAND. In comparison to Connectivity Map, the main advantage of INCanTeSIMO is the independence from pre-collected gene expression signatures, which severely affected Connectivity Map's performance in the analysis of novel cellular transitions. With regards to DeMAND, INCanTeSIMO obtained comparable performance while requiring significantly less data samples, and it also specifies if the predicted molecules should be activated or inhibited. Thus, the approach followed here offers the best balance between the amount of data required as input and the performance and type of predictions obtained.

5.1.2 Pathway perturbations for cellular transitions

Given that for many signalling molecules there might be no activators and inhibitors reported, it is beneficial to also predict signalling pathways triggering cellular conversions. In particular, as chemical inhibitors and activators for specific members of each pathway are well known (Lis, Kuzawińska, and Bałkowiec-Iskra 2014; Moreira, Fernandes, and Ramos 2007; Tamm et al. 2000; F. H. Tran and J. J. Zheng 2017), it is useful to predict pathways with the expectation that any perturbation applied to them will induce the cellular conversion required. Numerous pathways prediction tools exist, which aim at identifying signalling pathways involved in disease, cellular response to stimuli, cellular conversions etc. However, the majority of these approaches select pathways based on their gene expression differences between two compared conditions. This leads to uncertainty as to whether the observed expression change is the outcome or the driver of a switch in gene regulatory program.

INCanTeSIMO_path identifies pathways by calculating the aggregate source/sink centrality of INCanTeSIMO candidate molecules, under the assumption that if molecules that are central to a pathway are able to induce the cellular conversion required, the other components of the pathway can exert the same effect. Source/sink centrality (Naderi Yeganeh and Mostafavi 2019) measures the importance of molecules in a pathway based on how often they are used as the initial or final point in directed signal transduction paths. Thus, the pathway score used in INCanTeSIMO_path is not a measure of over-representation of the components of a signalling pathway in a set of significant genes, as in enrichment-based tools, but takes into account the role of the candidate molecules in the pathway, according to its topology and directionality. INCanTeSIMO_path showed better performance than other methods based on differential expression inputs or enrichment measures, and its predictions could be systematically confirmed by literature review.

One important caveat in predicting signalling pathways is their loose definition: canonical signalling pathways are highly variable depending on the database used (Kirouac et al. 2012; Türei, Korcsmáros, and Saez-Rodriguez 2016) and subject to extensive crosstalk (Schaefer et al. 2009). Thus, the prediction of some pathways according to source/sink centrality might present artefacts if candidate molecules are shared between multiple signalling pathways, and results could change according to the pathway database used. The predictions obtained should be interpreted taking into account which candidate molecules caused the prediction of a particular pathway.

5.2 Advantages of this approach

As mentioned previously, the use of manually curated transcriptional interactions for the inference of GRN allows to analyse cellular transitions for which few single expression profiles are available per cellular state. Depending on the type of data, which defines which approach for gene expression Booleanization and probability estimation can be applied, as low as one sample per state is sufficient for obtaining predictions. Additionally, as only the initial signalling state is considered (by estimating the probability of expression of signalling molecules in the initial cellular state), cellular transitions that have not yet been obtained in vitro can be analysed too, by comparing primary cell states. Thus, this approach is extremely flexible and generally applicable to any kind of biological process where cellular transitions are involved.

In the same vein, and unlike ad hoc models for specific cell types, no prior biological knowledge on the cellular transition desired is required as input. This was most evident in the application of INCanTeSIMO to a liver cirrhosis model in rat, resulting in the prediction of many signalling molecules associated with the disease, and among them the activation of Tie2, a receptor for angiopoietins involved in the stabilization of blood vessels (Fagiani and Christofori 2013). The GRN TFs expected to change upon Tie2 activation were identified via simulations, and literature review revealed their involvement in angiogenesis. Experimental observations confirmed that some of these TFs shifted towards their healthy expression state upon treatment with an agonist of Tie2, which has previously been shown to improve blood vessels stability (Pauta et al. 2015). Therefore, while INCanTeSIMO only required the disease and healthy gene expression profiles as input, it was able to identify signalling molecules and corresponding TFs related to the same specific biological process, also involved in the disease, and whose perturbation improved the pathological state.

Overall, novel predictions obtained with INCanTeSIMO and INCanTeSIMO_path were regularly related both to experimentally validated perturbations and to the cellular transition considered. Importantly, the methods correctly predicted signalling targets for cellular differentiation and reprogramming, which represent a key aspect of regenerative medicine. Similarly the predictions obtained for the initial steps of limb regeneration in the salamander *A. mexicanum* consistently recapitulated existing knowledge and previous observations, capturing the signalling cues that are known to come from dermis, nerves and other cells and act on the Prrx1-expressing connective tissue cells. In summary, the integration of signalling network and GRN, by explicitly modelling the regulatory activity of signalling on transcription, allowed capturing consistently signalling molecules and pathways that can trigger the cellular transition desired.

5.3 Limitations

The current implementation is computationally intensive. For this reason, combinations of only up to 4 interface TFs could be exhaustively perturbed in silico. Although there are hardly any synergistic effects among low-efficiency TFs in the combination of this size inducing significant change of the GRN state, it cannot be discarded that this might happen in higher order combinations. Other computationally intensive tasks include the calculation of the most probable paths connecting each signalling molecule to all interface TFs.

This approach cannot be applied to the analysis of very similar cellular states, or in general cellular conversions that show changes in gene expression for TFs that are not interacting among them to form a connected GRN of reasonable size (at least 10 TFs). From the biological point of view, the inference of small, disconnected GRNs suggests that the differences between initial and desired cellular states cannot be recapitulated by the propagation of a signalling perturbation across a GRN model. Rather, the differences in gene expression might be caused indirectly by perturbations which act at the level of other biological processes (e.g. metabolic pathways), and result in gene expression changes by affecting isolated TFs and their regulation of non-TF targets. With the same principle, datasets where no perturbation of the interface TFs can induce changes in at least 40% of the GRN TFs cannot be reliably analysed with this approach.

5.3.1 Issues with validation

A major issue in testing computational methods is the validation of obtained predictions. In this case, the methods were applied to known cellular transitions associated with a specific single signalling perturbation, being it a growth factor, chemical compound or small molecule applied to the cells. As mentioned previously, the majority of drugs approved for clinical use do not have known targets (Sawada et al. 2018), which also suggests that even drugs with associated targets might have unknown targets. Among the known and unknown targets, it is unclear which ones are drivers of the cellular response observed. Therefore, only perturbations with at least one known target were considered, and a successful prediction was defined as the selection of at least one signalling molecule or pathway directly targeted by the perturbation.

Another source of uncertainty is the fact that multiple alternative perturbations might induce the same cellular conversion. As first step into considering this possibility, the similarity of predicted perturbations to the experimentally tested ones was assessed. This analysis showed that there are functional and topological similarities among the predictions and the experimentally applied perturbations. However, unrelated perturbations might also induce the same cellular state, either because they share the same downstream effectors, or because they can act on different parts of the GRN that regulate each other resulting in the same GRN state. This became evident in the pathway analysis of cell fate transitions: through literature review, the potential of inducing the cellular transition was confirmed extensively for pathways not affected by the experimental perturbation.

The increasing number of signalling perturbations able to induce cellular transitions will allow to validate more extensively candidate pathways or molecules predicted by computational methods, but ultimately there is a need for large-scale screenings of perturbations that can induce cellular transitions in healthy cell types, expanding from the cancer cell line perturbation repositories currently available (Xiao et al. 2015; Subramanian, Narayan, et al. 2017).

5.3.2 Applicability to different data types

The expression state and probability play a central role in the analysis, as both the GRN and the signalling network are contextualized to the cellular transitions using them. Their estimation should therefore be performed as accurately as possible. All gene expression data used in this work was obtained by microarray. For specific microarray platforms, the datasets available in public repositories are so abundant that reliable estimation of the expression values range is possible for each gene separately (Matthew N. McCall, Bolstad, and Rafael A. Irizarry 2010)(frmaBool strategy). For datasets generated using Affymetrix chips, the MAS5.0 algorithm can estimate if a gene is expressed above background noise levels. However, to obtain a general approach for Booleanization, a data-driven strategy was devised (geneDE).

While the geneDE Booleanization approach cannot account for different dynamic expression ranges among genes and requires replicates for each cellular state, it relies on well established differential expression analysis pipelines and is extremely flexible. This strategy can readily be applied to RNA sequencing (RNA-seq) data. An alternative and more precise approach would be to use large databases of RNA-seq experiments, homogeneously pre-processed, to define gene-specific expression distributions. Multiple such databases already exist for human, mouse and rat (Lachmann et al. 2018; Söllner et al. 2017). A strategy similar to the one used in RefBool (Jung et al. 2017) could be applied to determine if genes are expressed or not and to calculate the probability of expression for each gene, allowing the use of INCanTeSIMO on RNA-seq data.

5.4 Outlook

5.4.1 Heterogeneity in the cellular response to signals

Cell-to-cell differences in the response to the same stimulus are ubiquitous and caused by pre-existing differences among cells (Selimkhanov et al. 2014; Toettcher, Weiner, and Lim 2013). These differences can be explained with the existence of clusters of different cellular functional states (subpopulations) (Snijder et al. 2009), which are observed in both signalling dynamics and gene expression patterns (Lane et al. 2017). The existence of different subpopulations implies that the same signalling stimulus might induce the cellular conversion desired only in portions of the cells, resulting in low conversion efficiency.

Using single cell RNA sequencing data, it is possible to identify sub-populations of cells with different functional state. This in turn could be used to define which specific signalling cues are required by each of them in order to induce the desired cellular transition, resulting in overall higher conversion efficiency.

The approach presented here could be adapted to use the gene expression patterns of each subpopulation separately. The expression of a gene across the cells in a subpopulation could help define if it is expressed or not, and its probability of expression. Then, a subpopulation-specific GRN could be reconstructed with the use of GRN inference methods developed for bulk (Faith et al. 2007; Huynh-Thu et al. 2010; Margolin et al. 2006) or single cell data (Aibar et al. 2017; Chan, Stumpf, and Babtie 2017). Finally, the paths most likely connecting signalling molecules and interface TFs could be identified, integrating in the calculation the gene expression correlation of the proteins involved in the paths in each subpopulation. This would result in different interface TFs and signalling paths available for signal transduction in each subpopulation, and thus in specialized predictions.

5.4.2 Relation with other mechanisms regulating gene expression

Cellular conversions can be induced also by acting on the cells metabolism or epigenetic landscape. The methods presented are not suitable in their current form for the discovery of this type of interventions, as metabolic or epigenetic regulation are not taken into account. However, they could be integrated in the current modelling framework.

Metabolism can regulate gene expression by exerting direct and indirect control on chromatin (Hong Li et al. 2018). Additionally, the energetic balance of the cells can influence gene expression, and the induction of glycolysis or autophagy was shown to improve reprogramming of cells to iPSCs in combination with TFs (T. Chen et al. 2011; Zhu et al. 2010). Thus, the metabolic regulation of the GRN can be represented as mediated by the signalling pathways sensing the metabolic state of the cells, such as AMPK (Burkewitz, Y. Zhang, and Mair 2014), mTOR (Kennedy and Lamming 2016), autophagy and hypoxia pathways, and could be improved by modelling of chromatin role on gene expression.

Recently, the development of CRISPR/Cas9 systems for epigenetic modification have simplified and made accessible targeted epigenetic editing (Jeffries 2018). Remaining challenges for this approach are the delivery of the machinery necessary for these systems inside the cells, and the possibility of regulating multiple target genes at the same time. The cell epigenetic landscape can also be regulated by modulating the activity of DNA methyltransferases or histone-modifying enzymes like acetyltransferases, deacetylases, and methyltransferases with chemical compounds. This strategy was already used in disease treatment (Wouters and Delwel 2016) and cellular reprogramming (E. Li and Davidson 2009; Shi et al. 2008; K. A. Tran et al. 2015). Further advancements in the understanding of pioneer transcription factors, which are able to bind the DNA and exert their function even when the chromatin is not accessible (Iwafuchi-Doi and Zaret 2014), might in the future lead to the development of general methods for the prediction of combinations of TFs and chemical compounds that can trigger desired cellular conversions.

Epigenetic editing (via targeted systems, pioneer factors or enzymes modulation) might be included in the scope of the present method by including in the GRN model the epigenetic state of the TFs. Provided the availability of epigenetic data such as DNA methylation, histone acetylation or methylation, the epigenetic profile of the initial and desired cellular state could be delineated and compared. In particular, considering the accessibility and activity state of the regulatory regions of the TFs in the GRN, it would be possible to predict if changes to these properties are needed to obtain the desired gene expression profile, and which compounds or TFs are suitable for the purpose of inducing them.

5.4.3 Prediction of signal combinations

An important limitation of the methods at present is that they only estimate the effect of single candidate molecules or pathways, whereas combinations of them normally reach higher cellular conversion efficiency (Cao et al. 2016; Kunisada et al. 2012). Indeed, established experimental protocols for the induction of cellular conversions often consist of multiple molecules applied in cocktails (Lai et al. 2017).

An initial attempt to predict combinations of molecules according to their addictive activity on interface TFs proved unsuccessful. This suggests that in reality, a key point for prediction of combinations will be understanding the interplay of different signal transduction paths. Signalling pathways are not isolated components, but are embedded in a signalling network where protein-protein interactions are free to occur and are stochastic events (Ladbury and Arold 2012), giving rise to extensive cross-talk opportunities.

A strategy that takes into account such cross-talk at the level of signalling network is necessary. Different molecules might use the same path to act on common downstream molecules, reinforcing the activation or inhibition of the interface TFs downstream, or might be redundant. Furthermore, different signalling molecules might have a synergistic effect and induce the activation or inhibition of interface TFs in ways that are not deducible from their individual effects (Housden and Perrimon 2014).

In order to extend the current approach to account for signalling cross-talk, it would be necessary to consider not only the probability and sign of the most probably expressed paths between signalling molecules and interface TFs, but also the intermediate molecules used in such paths. To discriminate between redundant, additive and synergistic effects, however, extensive knowledge on the logic rules governing the activation or inhibition of signalling molecules is required. While this knowledge might be available for canonical signalling paths, non-canonical signal transduction has been shown repeatedly to be a prominent regulator of gene expression (Meyerovich et al. 2016; Ohta et al. 2016; Regan et al. 2017; Voloshanenko et al. 2018) and is not as thoroughly documented so far. A possible alternative might be to define general logic rules governing classes of molecules, for example based on their mode of action or protein super-family.

5.5 Conclusion

The ability to induce transitions between cellular states is necessary for disease modelling and regenerative medicine. Cellular states are recapitulated by gene expression patterns, which are maintained by GRNs. In turn, changes in GRN state are reflected in differences in cellular functionality and possibly cellular identity. Not only transcriptional regulators, but also signalling events can regulate the state of the GRN by modulating the activity of interface TFs. The use of chemical compounds and small molecules targeting signalling pathways in order to induce cellular transitions is gaining more attention because of its advantages in terms of reproducibility, controllability, and safety compared to other strategies. However, the discovery of compounds for cellular transitions still relies on screenings or guidance from existing knowledge, and would benefit from predictive computational methods.

At present, existing computational methods focus separately on GRN or signalling networks. A number of tools exist for the prediction of instructive transcriptional factors that can induce cellular transitions. Additionally, methods that focus on the signalling network to describe how changes in gene expression affect signal transduction are also available. A reduced number of methods consider the action of signalling cues on their downstream TFs, but only manually curated and ad hoc models exist that integrate signalling and transcriptional networks. The aim of this work was to develop a general approach to analyse the effect that signalling cues have on the GRN by using gene expression data. In conclusion, the significant points of this thesis are:

- *Gene expression probability can reasonably approximate signal transduction*. Transcriptomics and phosphoproteomics data show at best a moderate correlation. However, the expression of proteins that constitute a signalling path is the first condition for signal transduction to happen along such path. It was observed that the most probably expressed paths connecting perturbation targets and interface TFs tend to show significantly more phosphorylation changes as compared to other paths, suggesting that paths preferentially used for signal transduction can be inferred using gene expression data.

- *Integration of regulatory layers can be obtained with gene expression data*. The proposed approach consisted in the integration of two distinct models for the transcriptional and signalling regulatory layers. The stochasticity of signal transduction and its effect on the interface TFs were taken into account by the use of a probabilistic approximation of signal transduction, and the role of interface TFs in determining the state of the GRN was estimated by in silico perturbations of a Boolean network model.

- *Signalling perturbations for cellular transitions can be consistently predicted with this approach*. The prediction of candidate molecules is based on the specificity with which signalling molecules are predicted to activate and inhibit the interface TFs that regulate the GRN state, calculated by Jensen-Shannon's divergence. The results obtained showed better performance compared to other methods in recovering experimental perturbations, and functional and topological characteristics shared with targets of such perturbations. Signalling pathways were predicted by calculating the aggregate importance of candidate molecules, expressed in terms of source/sink centrality, in each pathway. Again, the performance obtained was superior to previously published methods. Candidate molecules and pathways are associated with a sign, to further guide the design of experiments for the induction of the desired cell state shift.

- *Previous knowledge was consistently recapitulated in the predictions*. Extensive literature review confirmed that the molecules and pathways predicted for the induction of multiple and diverse cellular transitions. The approach was applied to multiple examples of cellular differentiation and reprogramming, but also the reversal of disease phenotypes and to limb regeneration in animal models, confirming its general scope and general applicability.
- *The effect of signalling on the GRN was validated experimentally*. The application of this method to the prediction of signalling molecules for the treatment of cirrhosis resulted in the selection of CVX-060, an agonist of the receptor Tie2. As predicted, it induced changes in the expression of GRN TFs involved in cirrhotic processes.

In conclusion, the methods presented here represent a useful addition to the existing computational tools for cellular conversions. They are generally applicable tools that can direct the identification of signalling molecules and pathways for the induction of desired cellular transitions, such as the induction of cellular differentiation or reprogramming and the reversal of pathological phenotypes, without needing biological knowledge of the cellular conversion studied to be given as input. The potential applications of this approach include disease treatment and regenerative medicine.

6 References

Abu Rmilah, Anan et al. (May 2019). “Understanding the marvels behind liver regeneration”. In: *Wiley Interdisciplinary Reviews: Developmental Biology* 8.3, e340.

Adachi, Tohru et al. (June 2005). “NAD(P)H oxidase plays a crucial role in PDGF-induced proliferation of hepatic stellate cells”. In: *Hepatology* 41.6, pp. 1272–1281.

Affymetrix, Inc. (2002). “Statistical Algorithms Description Document”.

Aibar, Sara et al. (Nov. 2017). “SCENIC: single-cell regulatory network inference and clustering”. In: *Nature Methods* 14.11, pp. 1083–1086.

Al Haj Baddar, Nour W., Adarsh Chithrala, and S. Randal Voss (Feb. 2019). “Amputation-induced reactive oxygen species signaling is required for axolotl tail regeneration”. In: *Developmental Dynamics* 248.2, pp. 189–196.

Alexa, A., J. Rahnenfuhrer, and T. Lengauer (July 2006). “Improved scoring of functional groups from gene expression data by decorrelating GO graph structure”. In: *Bioinformatics* 22.13, pp. 1600–1607.

Amadoz, Alicia et al. (2015). “Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity.” In: *Scientific reports* 5.November, p. 18494.

Antos, Christopher L. and Elly M. Tanaka (2010). “Vertebrates that regenerate as models for guiding stem cells”. In: *Advances in Experimental Medicine and Biology* 695, pp. 184–214.

Ardito, Fatima et al. (Aug. 2017). “The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review)”. In: *International Journal of Molecular Medicine* 40.2, pp. 271–280.

Avior, Yishai, Ido Sagi, and Nissim Benvenisty (Mar. 2016). “Pluripotent stem cells in disease modelling and drug discovery”. In: *Nature Reviews Molecular Cell Biology* 17.3, pp. 170–182.

Balwierz, Piotr J. et al. (May 2014). “ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs”. In: *Genome Research* 24.5, pp. 869–884.

Basson, M Albert (June 2012). “Signaling in Cell Differentiation and Morphogenesis”. In: *Cold Spring Harbor Perspectives in Biology* 4.6, a008151–a008151.

Bayik, Defne, Debra Tross, and Dennis M. Klinman (Mar. 2018). “Factors Influencing the Differentiation of Human Monocytic Myeloid-Derived Suppressor Cells Into Inflammatory Macrophages”. In: *Frontiers in Immunology* 9, p. 608.

Ben Khadra, Yousra et al. (Oct. 2017). “An integrated view of asteroid regeneration: tissues, cells and molecules”. In: *Cell and Tissue Research* 370.1, pp. 13–28.

Ben-David, Uri and Nissim Benvenisty (Apr. 2011). “The tumorigenicity of human embryonic and induced pluripotent stem cells”. In: *Nature Reviews Cancer* 11.4, pp. 268–277.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological).*

Beury, Daniel W. et al. (Dec. 2014). “Cross-talk among myeloid-derived suppressor cells, macrophages, and tumor cells impacts the inflammatory milieu of solid tumors”. In: *Journal of Leukocyte Biology* 96.6, pp. 1109–1118.

Bissell, D. Montgomery (Sept. 2011). “Therapy for hepatic fibrosis: Revisiting the preclinical models”. In: *Clinics and Research in Hepatology and Gastroenterology* 35.8-9, pp. 521–525.

Biteau, Benoit, Christine E. Hochmuth, and Heinrich Jasper (Nov. 2011). “Maintaining Tissue Homeostasis: Dynamic Control of Somatic Stem Cell Activity”. In: *Cell Stem Cell* 9.5, pp. 402–411.

Blevins, William R. et al. (Dec. 2019). “Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker’s yeast”. In: *Scientific Reports* 9.1, p. 11005.

Brandão, Karina O et al. (Sept. 2017). “Human pluripotent stem cell models of cardiac disease: from mechanisms to therapies”. In: *Disease Models & Mechanisms* 10.9, pp. 1039–1059.

Bruijn, Marella de and Elaine Dzierzak (Apr. 2017). “Runx transcription factors in the development and function of the definitive hematopoietic system”. In: *Blood* 129.15, pp. 2061–2069.

Bucan, Vesna et al. (Aug. 2018). “Identification of axolotl BH3-only proteins and expression in axolotl organs and apoptotic limb regeneration tissue”. In: *Biology Open* 7.8, bio036293.

Burkewitz, Kristopher, Yue Zhang, and William B Mair (July 2014). “AMPK at the nexus of energetics and aging.” In: *Cell metabolism* 20.1, pp. 10–25.

Cahan, Patrick et al. (Aug. 2014). “CellNet: Network Biology Applied to Stem Cell Engineering”. In: *Cell* 158.4, pp. 903–915.

Cao, Nan et al. (2016). “Conversion of human fibroblasts into functional cardiomyocytes by small molecules.” In: *Science (New York, N.Y.)* 1502.April, aaf1502.

Carter, E. J. et al. (Sept. 2009). “MEK5 and ERK5 are mediators of the pro-myogenic actions of IGF-2”. In: *Journal of Cell Science* 122.17, pp. 3104–3112.

Catlett, Natalie L et al. (Dec. 2013). “Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data”. In: *BMC Bioinformatics* 14.1, p. 340.

Chan, Thalia E., Michael P.H. Stumpf, and Ann C. Babtie (Sept. 2017). “Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures”. In: *Cell Systems* 5.3, 251–267.e3.

Chang, Kuo-Hsuan et al. (Sept. 2018). “Modeling Alzheimer’s Disease by Induced Pluripotent Stem Cells Carrying APP D678H Mutation”. In: *Molecular Neurobiology*.

Chen, H.-C. et al. (Aug. 2004). “Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle”. In: *Bioinformatics* 20.12, pp. 1914–1927.

Chen, Helen X. and Jessica N. Cleck (Aug. 2009). “Adverse effects of anticancer agents that target the VEGF pathway”. In: *Nature Reviews Clinical Oncology* 6.8, pp. 465–477.

Chen, Kai and John F Keaney (Oct. 2012). “Evolving concepts of oxidative stress and reactive oxygen species in cardiovascular disease.” In: *Current atherosclerosis reports* 14.5, pp. 476–83.

Chen, Taotao et al. (Oct. 2011). “Rapamycin and other longevity-promoting compounds enhance the generation of mouse induced pluripotent stem cells”. In: *Aging Cell* 10.5, pp. 908–911.

Chernoff, Ellen A.G. et al. (Feb. 2003). “Urodele spinal cord regeneration and related processes”. In: *Developmental Dynamics* 226.2, pp. 295–307.

Cotton, Travis B. et al. (2015). “Discerning mechanistically rewired biological pathways by cumulative interaction heterogeneity statistics”. In: *Scientific Reports* 5, p. 9634.

Crespo, Isaac et al. (Jan. 2013). “Detecting cellular reprogramming determinants by differential stability analysis of gene regulatory networks.” In: *BMC systems biology* 7, p. 140.

D’Alessio, Ana C. et al. (Oct. 2015). “A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity”. In: *Stem Cell Reports* 5.5, pp. 763–775.

D’Souza, Rochelle C. J. et al. (2014). “Time-resolved dissection of early phosphoproteome and ensuing proteome changes in response to TGF-β”. In: *Science Signaling* 7.335, rs5–rs5.

Davis, R L, H Weintraub, and A B Lassar (Dec. 1987). “Expression of a single transfected cDNA converts fibroblasts to myoblasts.” In: *Cell* 51.6, pp. 987–1000.

De, Debojyoti et al. (2017). “Small molecule-induced cellular conversion”. In: *Chemical Society Reviews* 46.20, pp. 6241–6254.

Delfini, Marie-Claire et al. (Sept. 2009). “The timing of emergence of muscle progenitors is controlled by an FGF/ERK/SNAIL1 pathway”. In: *Developmental Biology* 333.2, pp. 229–237.

DOUGLAS, B. S. (May 1972). “CONSERVATIVE MANAGEMENT OF GUILLOTINE AMPUTATION OF THE FINGER IN CHILDREN”. In: *Journal of Paediatrics and Child Health* 8.2, pp. 86–89.

Dutkowski, Janusz and Trey Ideker (2011). “Protein Networks as Logic Functions in Development and Cancer”. In: *PLoS Computational Biology* 7.9, e1002180.

Dutta, Bhaskar, Anders Wallqvist, and Jaques Reifman (2012). “PathNet: A tool for pathway analysis using topological information”. In: *Source Code for Biology and Medicine* 7.1, p. 10.

Edfors, Fredrik et al. (Oct. 2016). “Gene-specific correlation of RNA and protein levels in human cells and tissues”. In: *Molecular Systems Biology* 12.10, p. 883.

Efroni, Sol, Carl F. Schaefer, and Kenneth H. Buetow (May 2007). “Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis”. In: *PLoS ONE* 2.5. Ed. by Nick Monk, e425.

Elpek, Gülsüm Özlem (Mar. 2015). “Angiogenesis and liver fibrosis”. In: *World Journal of Hepatology* 7.3, p. 377.

Endo, Tetsuya, Susan V. Bryant, and David M. Gardiner (2004). “A stepwise model system for limb regeneration”. In: *Developmental Biology* 270, pp. 135–145.

Fagiani, Ernesta and Gerhard Christofori (Jan. 2013). “Angiopoietins in angiogenesis”. In: *Cancer Letters* 328.1, pp. 18–26.

Faith, Jeremiah J et al. (Jan. 2007). “Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles”. In: *PLoS Biology* 5.1. Ed. by Andre Levchenko, e8.

Fang, Feng et al. (Oct. 2013). “Early Growth Response 3 (Egr-3) Is Induced by Transforming Growth Factor-β and Regulates Fibrogenic Responses”. In: *The American Journal of Pathology* 183.4, pp. 1197–1208.

Farhang Ghahremani, Morvarid, Steven Goossens, and Jody J Haigh (May 2013). “The p53 family and VEGF regulation: “It’s complicated””. In: *Cell Cycle* 12.9, pp. 1331–1332.

Farkas, Johanna E., Polina D. Freitas, et al. (Aug. 2016). “Neuregulin-1 signaling is essential for nerve-dependent axolotl limb regeneration”. In: *Development* 143.15, pp. 2724–2731.

Farkas, Johanna E. and James R. Monaghan (Jan. 2017). “A brief history of the study of nerve dependent regeneration”. In: *Neurogenesis* 4.1, e1302216.

Federation, Alexander J., James E. Bradner, and Alexander Meissner (2014). “The use of small molecules in somatic-cell reprogramming”. In: *Trends in Cell Biology* 24.3, pp. 179–187. arXiv: NIHMS150003.

Fernández, Mercedes et al. (2009). “Angiogenesis in liver disease”. In: *Journal of Hepatology* 50.3, pp. 604–620.

Friedman, Scott L (2008). “Hepatic stellate cells: protean, multifunctional, and enigmatic cells of the liver.” In: *Physiological reviews* 88.1, pp. 125–72.

Garcia-Gonzalez, Claudia and Jamie Ian Morrison (Feb. 2014). “Cardiac regeneration in non-mammalian vertebrates”. In: *Experimental Cell Research* 321.1, pp. 58–63.

Gemberling, Matthew et al. (Nov. 2013). “The zebrafish as a model for complex tissue regeneration.” In: *Trends in genetics : TIG* 29.11, pp. 611–20.

Gerber, Tobias et al. (2018). “Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration”. In: *Science* 0681.September, eaaq0681.

Gerhard, Glenn S. et al. (Dec. 2018). “Differentially methylated loci in NAFLD cirrhosis are associated with key signaling pathways”. In: *Clinical Epigenetics* 10.1, p. 93.

Ghosh, Sukla et al. (Mar. 2008). “Analysis of the expression and function of Wnt-5a and Wnt-5b in developing and regenerating axolotl (Ambystoma mexicanum) limbs”. In: *Development, Growth & Differentiation* 50.4, pp. 289–297.

Glass, Leon and Stuart A. Kauffman (Apr. 1973). “The logical analysis of continuous, non-linear biochemical control networks”. In: *Journal of Theoretical Biology* 39.1, pp. 103–129.

Gnad, Florian, Sophia Doll, et al. (2016). “Phosphoproteome analysis of the MAPK pathway reveals previously undetected feedback mechanisms”. In: *Proteomics* 16.14, pp. 1998–2004.

Gnad, Florian, Jeffrey Wallin, et al. (2016). “Quantitative phosphoproteomic analysis of the PI3K-regulated signaling network”. In: *Proteomics* 16.14, pp. 1992–1997.

Godwin, James W. and Nadia Rosenthal (Jan. 2014). “Scar-free wound healing and regeneration in amphibians: Immunological influences on regenerative success”. In: *Differentiation* 87.1-2, pp. 66–75.

Godwin, James W, Alexander R Pinto, and Nadia A Rosenthal (June 2013). “Macrophages are required for adult salamander limb regeneration.” In: *Proceedings of the National Academy of Sciences of the United States of America* 110.23, pp. 9415–20.

Guo, Shun et al. (Dec. 2016). “Gene regulatory network inference using PLS-based methods”. In: *BMC Bioinformatics* 17.1, p. 545.

Han, Heonjong et al. (Jan. 2015). “TRRUST: a reference database of human transcriptional regulatory interactions.” en. In: *Scientific reports* 5, p. 11432.

Hardee, Cinnamon et al. (Feb. 2017). “Advances in Non-Viral DNA Vectors for Gene Therapy”. In: *Genes* 8.2, p. 65.

Haridhasapavalan, Krishna Kumar et al. (Feb. 2019). “An insight into non-integrative gene delivery approaches to generate transgene-free induced pluripotent stem cells”. In: *Gene* 686, pp. 146–159.

Hartmann, András, Satoshi Okawa, et al. (Dec. 2018). “SeesawPred: A Web Application for Predicting Cell-fate Determinants in Cell Differentiation”. In: *Scientific Reports* 8.1, p. 13355.

Hartmann, András, Srikanth Ravichandran, and Antonio del Sol (July 2019). “Modeling Cellular Differentiation and Reprogramming with Gene Regulatory Networks”. In: *Computational Stem Cell Biology*. Vol. 17 Suppl 1, pp. 37–51.

Haury, Anne-Claire et al. (Jan. 2012). “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.” In: *BMC systems biology* 6.1, p. 145.

Haynes, Winston A et al. (Jan. 2013). “Differential expression analysis for pathways.” In: *PLoS computational biology* 9.3, e1002967.

Herrera, Blanca, Annalisa Addante, and Aránzazu Sánchez (Dec. 2017). “BMP Signalling at the Crossroad of Liver Fibrosis and Regeneration”. In: *International Journal of Molecular Sciences* 19.1, p. 39.

Hidalgo, Marta R. et al. (Jan. 2017). “High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes”. In: *Oncotarget* 8.3, pp. 5160–5178.

Hinz, Boris et al. (June 2007). “The Myofibroblast”. In: *The American Journal of Pathology* 170.6, pp. 1807–1816.

Hirsch, Tobias et al. (Nov. 2017). “Regeneration of the entire human epidermis using transgenic stem cells”. In: *Nature* 551.7680, pp. 327–332.

Housden, Benjamin E. and Norbert Perrimon (Oct. 2014). “Spatial and temporal organization of signaling pathways”. In: *Trends in Biochemical Sciences* 39.10, pp. 457–464.

Huang, Ruili, Anders Wallqvist, and David G. Covell (2006). “Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute’s anticancer screen”. In: *Genomics* 87.3, pp. 315–328.

Huang, Shao-shan Carol et al. (Jan. 2013). “Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling.” In: *PLoS computational biology* 9.2, e1002887.

Huang, Sui (Aug. 1999). “Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery”. In: *Journal of Molecular Medicine* 77.6, pp. 469–480.

Huang, Sui (Feb. 2012). “The molecular and mathematical basis of Waddington’s epigenetic landscape: A framework for post-Darwinian biology?” In: *BioEssays* 34.2, pp. 149–157.

Huynh-Thu, Vân Anh et al. (Jan. 2010). “Inferring regulatory networks from expression data using tree-based methods.” In: *PloS one* 5.9, e12776.

Imokawa, Yutaka and Jeremy P Brockes (May 2003). “Selective Activation of Thrombin Is a Critical Determinant for Vertebrate Lens Regeneration”. In: *Current Biology* 13.10, pp. 877–881.

Invergo, Brandon M. and Pedro Beltrao (Aug. 2018). “Reconstructing phosphorylation signalling networks from quantitative phosphoproteomic data”. In: *Essays In Biochemistry* 0.June, EBC20180019.

Iwafuchi-Doi, Makiko and Kenneth S. Zaret (Dec. 2014). “Pioneer transcription factors in cell reprogramming”. In: *Genes & Development* 28.24, pp. 2679–2692.

Iwakiri, Yasuko and Moon Young Kim (Aug. 2015). “Nitric oxide in liver diseases”. In: *Trends in Pharmacological Sciences* 36.8, pp. 524–536.

Jaeger, Savina et al. (2014). “Causal Network Models for Predicting Compound Targets and Driving Pathways in Cancer.” In: *Journal of biomolecular screening* 19.5, pp. 791–802.

Jeffries, Matlock A. (Nov. 2018). “Epigenetic editing: How cutting-edge targeted epigenetic modification might provide novel avenues for autoimmune disease therapy”. In: *Clinical Immunology* 196, pp. 49–58.

Jung, Sascha et al. (2017). “Prediction of Chromatin Accessibility in Gene-Regulatory Regions from Transcriptomics Data”. In: *Scientific reports* Under revi.May, pp. 1–10.

Kandasamy, Richard K. et al. (Oct. 2016). “A time-resolved molecular map of the macrophage response to VSV infection”. In: *npj Systems Biology and Applications* 2.August, p. 16027.

Kanichai, Manoj et al. (Sept. 2008). “Hypoxia promotes chondrogenesis in rat mesenchymal stem cells: A role for AKT and hypoxia-inducible factor (HIF)-1α”. In: *Journal of Cellular Physiology* 216.3, pp. 708–715.

Kanshin, Evgeny et al. (Feb. 2015). “A Cell-Signaling Network Temporally Resolves Specific versus Promiscuous Phosphorylation”. In: *Cell Reports* 10.7, pp. 1202–1214.

Kauffman, S.A. (Mar. 1969). “Metabolic stability and epigenesis in randomly constructed genetic nets”. In: *Journal of Theoretical Biology* 22.3, pp. 437–467.

Kauffman, Stuart A (1993). *The origins of order: Self-organization and selection in evolution*. OUP USA.

Kennedy, Brian K and Dudley W Lamming (June 2016). “The Mechanistic Target of Rapamycin: The Grand ConducTOR of Metabolism and Aging.” In: *Cell metabolism* 23.6, pp. 990–1003.

Khatri, Purvesh, Marina Sirota, and Atul J. Butte (2012). *Ten years of pathway analysis: Current approaches and outstanding challenges*.

Kim, Tae-Hee et al. (Feb. 2014). “Broadly permissive intestinal chromatin underlies lateral inhibition and cell plasticity.” In: *Nature* 506.7489, pp. 511–515. arXiv: `NIHMS150003`.

Kim, Yoo-Ah, Stefan Wuchty, and Teresa M. Przytycka (Mar. 2011). “Identifying Causal Genes and Dysregulated Pathways in Complex Diseases”. In: *PLoS Computational Biology* 7.3. Ed. by Markus W. Covert, e1001095.

Kirouac, Daniel C et al. (2012). “Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks”. In: *BMC Systems Biology* 6.1, p. 29.

Knapp, Dunja et al. (2013). “Comparative Transcriptional Profiling of the Axolotl Limb Identifies a Tripartite Regeneration-Specific Gene Program”. In: *PLoS ONE* 8.5.

Koay, E.J. and K.A. Athanasiou (Dec. 2008). “Hypoxic chondrogenic differentiation of human embryonic stem cells enhances cartilage protein synthesis and biomechanical functionality”. In: *Osteoarthritis and Cartilage* 16.12, pp. 1450–1456.

Koster, J. and S. Rahmann (Oct. 2012). “Snakemake–a scalable bioinformatics workflow engine”. In: *Bioinformatics* 28.19, pp. 2520–2522.

Kotzsch, Alexander et al. (Apr. 2009). “Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5–type I receptor specificity”. In: *The EMBO Journal* 28.7, pp. 937–947.

Kragl, Martin and Elly M. Tanaka (Aug. 2009). “Axolotl (Ambystoma mexicanum) limb and tail amputation”. In: *Cold Spring Harbor Protocols* 4.8, pdb.prot5267.

Krumsiek, Jan et al. (2011). “Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network”. In: *PLoS ONE* 6.8.

Kumar, Anoop and Jeremy P. Brockes (2012). “Nerve dependence in tissue, organ, and appendage regeneration”. In: *Trends in Neurosciences* 35.11, pp. 691–699.

Kunisada, Yuya et al. (Mar. 2012). “Small molecules induce efficient differentiation into insulin-producing cells from human induced pluripotent stem cells”. In: *Stem Cell Research* 8.2, pp. 274–284.

Kuo, Caroline Y and Donald B Kohn (May 2016). “Gene Therapy for the Treatment of Primary Immune Deficiencies”. In: *Current Allergy and Asthma Reports* 16.5, p. 39.

Lachmann, Alexander et al. (Dec. 2018). “Massive mining of publicly available RNA-seq data from human and mouse”. In: *Nature Communications* 9.1, p. 1366.

Ladbury, John E. and Stefan T. Arold (2012). “Noise in cellular signaling pathways: Causes and effects”. In: *Trends in Biochemical Sciences* 37.5, pp. 173–178.

Lai, Pei-Lun et al. (Dec. 2017). “Efficient Generation of Chemically Induced Mesenchymal Stem Cells from Human Dermal Fibroblasts”. In: *Scientific Reports* 7.1, p. 44534.

Lamb, Justin (Jan. 2007). “The Connectivity Map: a new tool for biomedical research”. In: *Nature Reviews Cancer* 7.1, pp. 54–60.

Lamb, Justin et al. (Sept. 2006). “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.” In: *Science (New York, N.Y.)* 313.5795, pp. 1929–35.

Lane, Keara et al. (Apr. 2017). “Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation”. In: *Cell Systems* 4.4, 458–469.e5.

Lang, Alex H. et al. (Aug. 2014). “Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes”. In: *PLoS Computational Biology* 10.8. Ed. by Alexandre V. Morozov, e1003734.

Laping, N. J. (July 2002). “Inhibition of Transforming Growth Factor (TGF)-beta 1-Induced Extracellular Matrix with a Novel Inhibitor of the TGF-beta Type I Receptor Kinase Activity: SB-431542”. In: *Molecular Pharmacology* 62.1, pp. 58–64.

Lee, Jong-Min et al. (May 2013). “Abnormal liver differentiation and excessive angiogenesis in mice lacking Runx3”. In: *Histochemistry and Cell Biology* 139.5, pp. 751–758.

Lee, Michael J and Michael B Yaffe (June 2016). “Protein Regulation in Signal Transduction”. In: *Cold Spring Harbor Perspectives in Biology* 8.6, a005918.

Li, Enhu and Eric H Davidson (June 2009). “Building developmental gene regulatory networks.” In: *Birth defects research. Part C, Embryo today : reviews* 87.2, pp. 123–30.

Li, Heng and Nils Homer (Sept. 2010). “A survey of sequence alignment algorithms for next-generation sequencing.” In: *Briefings in bioinformatics* 11.5, pp. 473–83.

Li, Hong et al. (Oct. 2018). “Modeling Parkinson’s Disease Using Patient-specific Induced Pluripotent Stem Cells”. In: *Journal of Parkinson’s Disease* 8.4, pp. 479–493.

Li, Na et al. (Aug. 2019). “PRDM1 levels are associated with clinical diseases in chronic HBV infection and survival of patients with HBV-related hepatocellular carcinoma”. In: *International Immunopharmacology* 73, pp. 156–162.

Lis, Krzysztof, Olga Kuzawińska, and Ewa Bałkowiec-Iskra (Dec. 2014). “Tumor necrosis factor inhibitors - state of knowledge.” In: *Archives of medical science : AMS* 10.6, pp. 1175–85.

Liu, Dawei, Debashis Ghosh, and Xihong Lin (Jan. 2008). “Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models.” In: *BMC bioinformatics* 9, p. 292.

Liu, Yan et al. (Mar. 2013). “Animal models of chronic liver diseases”. In: *American Journal of Physiology-Gastrointestinal and Liver Physiology* 304.5, G449–G468.

Love, Nick R et al. (Feb. 2013). “Amputation-induced reactive oxygen species are required for successful Xenopus tadpole tail regeneration”. In: *Nature Cell Biology* 15.2, pp. 222–228.

Makanae, Aki, Ayako Hirata, et al. (Sept. 2013). “Nerve independent limb induction in axolotls.” In: *Developmental biology* 381.1, pp. 213–26.

Makanae, Aki, Kazumasa Mitogawa, and Akira Satoh (Dec. 2014). “Co-operative Bmp- and Fgf-signaling inputs convert skin wound healing to limb formation in urodele amphibians.” In: *Developmental biology* 396.1, pp. 57–66.

Makanae, Aki and Akira Satoh (2012). “Early Regulation of Axolotl Limb Regeneration”. In: *Anatomical Record* 295.March, pp. 1566–1574.

Makhija, Harshyaa et al. (Sept. 2018). “A novel λ integrase-mediated seamless vector transgenesis platform for therapeutic protein expression”. In: *Nucleic Acids Research* 46.16, e99–e99.

Margolin, Adam A et al. (Jan. 2006). “ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.” In: *BMC bioinformatics* 7 Suppl 1, S7.

Martini, Paolo et al. (Jan. 2013). “Along signal paths: an empirical gene set approach exploiting pathway topology.” In: *Nucleic acids research* 41.1, e19.

Massa, Maria Sofia, Monica Chiogna, and Chiara Romualdi (2010). “Gene set analysis exploiting the topology of a pathway.” In: *BMC systems biology* 4, p. 121.

Mbodj, Abibatou et al. (2016). “Qualitative Dynamical Modelling Can Formally Explain Mesoderm Specification and Predict Novel Developmental Phenotypes”. In: *PLOS Computational Biology* 12.9, e1005073.

McCall, Matthew N., Benjamin M. Bolstad, and Rafael A. Irizarry (2010). “Frozen robust multiarray analysis (fRMA)”. In: *Biostatistics* 11.2, pp. 242–253.

McCall, Matthew N, Harris A Jaffee, et al. (Jan. 2014). “The Gene Expression Barcode 3.0: improved data processing and mining tools”. In: *Nucleic Acids Research* 42.D1, pp. D938–D943.

McCall, Matthew N, Karan Uppal, et al. (Jan. 2011). “The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes.” In: *Nucleic acids research* 39.Database issue, pp. D1011–5.

McCusker, Catherine D and David M Gardiner (June 2014). “Understanding positional cues in salamander limb regeneration: implications for optimizing cell-based regenerative therapies.” In: *Disease models & mechanisms* 7.6, pp. 593–9.

McCusker, Catherine, Susan V. Bryant, and David M. Gardiner (Apr. 2015). “The axolotl limb blastema: cellular and molecular mechanisms driving blastema formation and limb regeneration in tetrapods”. In: *Regeneration* 2.2, pp. 54–71.

Mejias, Marc et al. (Apr. 2009). “Beneficial effects of sorafenib on splanchnic, intrahepatic, and portocollateral circulations in portal hypertensive and cirrhotic rats”. In: *Hepatology* 49.4, pp. 1245–1256.

Melas, Ioannis N et al. (2015). “Identification of drug-specific pathways based on gene expression data: application to drug induced lung injury.” In: *Integrative biology : quantitative biosciences from nano to macro* 7.8, pp. 904–20.

Méndez, Akram and Luis Mendoza (2016). “A Network Model to Describe the Terminal Differentiation of B Cells”. In: *PLOS Computational Biology*, pp. 1–26.

Mercader, N et al. (2000). “Opposing RA and FGF signals control proximodistal vertebrate limb development through regulation of Meis genes.” In: *Development (Cambridge, England)* 127.18, pp. 3961–3970.

Mescher, Anthony L. et al. (Dec. 1997). “Transferrin is necessary and sufficient for the neural effect on growth in amphibian limb regeneration blastemas”. In: *Development, Growth and Differentiation* 39.6, pp. 677–684.

Meyerovich, Kira et al. (Mar. 2016). “The non-canonical NF-κB pathway is induced by cytokines in pancreatic beta cells and contributes to cell death and proinflammatory responses in vitro”. In: *Diabetologia* 59.3, pp. 512–521.

Michelini, Elisa et al. (Sept. 2010). “Cell-based assays: fuelling drug discovery”. In: *Analytical and Bioanalytical Chemistry* 398.1, pp. 227–238.

Miller, Bess M., Kimberly Johnson, and Jessica L. Whited (2019). “Common themes in tetrapod appendage regeneration: a cellular perspective”. In: *EvoDevo* 10.1, p. 11.

Mitashov, V I (Aug. 1996). “Mechanisms of retina regeneration in urodeles.” In: *The International journal of developmental biology* 40.4, pp. 833–44.

Moignard, Victoria et al. (2015). “Decoding the regulatory network of early blood development from single-cell gene expression measurements”. In: *Nature Biotechnology* advance on.3.

Monaghan, J. R. et al. (2012). “Gene expression patterns specific to the regenerating limb of the Mexican axolotl”. In: *Biology Open* 1, pp. 937–948.

Moreira, Irina Sousa, Pedro Alexandrino Fernandes, and Maria João Ramos (Mar. 2007). “Vascular endothelial growth factor (VEGF) inhibition–a critical review.” In: *Anti-cancer agents in medicinal chemistry* 7.2, pp. 223–45.

Mrugala, Dominique et al. (Mar. 2009). “Gene expression profile of multipotent mesenchymal stromal cells: Identification of pathways common to TGFbeta3/BMP2-induced chondrogenesis.” In: *Cloning and stem cells* 11.1, pp. 61–76.

Nacu, Eugeniu et al. (2016). “FGF8 and SHH substitute for anterior–posterior tissue interactions to induce limb regeneration”. In: *Nature* 1, pp. 1–16.

Nacu, Eugen et al. (Feb. 2013). “Connective tissue cells, but not muscle cells, are involved in establishing the proximo-distal outcome of limb regeneration in the axolotl.” In: *Development (Cambridge, England)* 140.3, pp. 513–8.

Naderi Yeganeh, Pourya and M. Taghi Mostafavi (2017). “Use of Structural Properties of Underlying Graphs in Pathway Enrichment Analysis of Genomic Data”. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB ’17*. August. New York, New York, USA: ACM Press, pp. 279–284.

Naderi Yeganeh, Pourya and M. Taghi Mostafavi (2019). “Causal Disturbance Analysis: A Novel Graph Centrality Based Method for Pathway Enrichment Analysis”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* PP.March, pp. 1–1.

Naldi, Aurélien et al. (Jan. 2010). “Diversity and plasticity of Th cell types predicted from regulatory network modelling.” In: *PLoS computational biology* 6.9, e1000912.

Needham, Elise J. et al. (2019). “Illuminating the dark phosphoproteome”. In: *Sci. Signal.* 12.565, eaau8645.

Newman, Mark (2010). *Networks: an introduction*. Oxford University Press, p. 772.

Nguyen-Chi, Mai et al. (Aug. 2017). “TNF signaling and macrophages govern fin regeneration in zebrafish larvae”. In: *Cell Death & Disease* 8.8, e2979–e2979.

Nguyen, Matthew et al. (2017). “Retinoic acid receptor regulation of epimorphic and homeostatic regeneration in the axolotl”. In: *Development* January, dev.139873.

Niepel, Mario et al. (2017). “Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling”. In: *Nature Communications* 8.1.

Noh, Heeju, Jason E. Shoemaker, and Rudiyanto Gunawan (2018). “Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection”. In: *Nucleic Acids Research* 46.6.

Nohno, Tsutomu et al. (July 1993). “A Chicken Homeobox Gene Related to Drosophila paired Is Predominantly Expressed in the Developing Limb”. In: *Developmental Biology* 158.1, pp. 254–264.

O’Brien, P. J. et al. (Sept. 2006). “High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity measured in a novel cell-based model using high content screening”. In: *Archives of Toxicology* 80.9, pp. 580–604.

Ogawa, S. et al. (2013). “Three-dimensional culture and cAMP signaling promote the maturation of human pluripotent stem cell-derived hepatocytes”. In: *Development* 140.15, pp. 3285–3296.

Ohta, Sho et al. (June 2016). “BMP regulates regional gene expression in the dorsal otocyst through canonical and non-canonical intracellular pathways”. In: *Development* 143.12, pp. 2228–2237.

Okawa, Satoshi et al. (2016). “A Generalized Gene-Regulatory Network Model of Stem Cell Differentiation for Predicting Lineage Specifiers.” In: *Stem cell reports* 7.3, pp. 307–315.

Okita, Keisuke, Tomoko Ichisaka, and Shinya Yamanaka (July 2007). “Generation of germline-competent induced pluripotent stem cells”. In: *Nature* 448.7151, pp. 313–317.

Olsen, Jesper V et al. (2010). “Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis.” In: *Science signaling* 3.104, ra3.

Osmanbeyoglu, Hatice U. et al. (Nov. 2014). “Linking signaling pathways to transcriptional programs in breast cancer”. In: *Genome Research* 24.11, pp. 1869–1880.

Owlarn, Suthira et al. (Dec. 2017). “Generic wound signals initiate regeneration in missing-tissue contexts”. In: *Nature Communications* 8.1, p. 2282.

Panebianco, Concetta et al. (Aug. 2017). “Senescence in hepatic stellate cells as a mechanism of liver fibrosis reversal: a putative synergy between retinoic acid and PPAR-gamma signalings”. In: *Clinical and Experimental Medicine* 17.3, pp. 269–280.

Parafati, Maddalena et al. (Sept. 2018). “A nonalcoholic fatty liver disease model in human induced pluripotent stem cell-derived hepatocytes, created by endoplasmic reticulum stress-induced steatosis”. In: *Disease Models & Mechanisms* 11.9, p. dmm033530.

Parikh, Jignesh R. et al. (2010). “Discovering causal signaling pathways through gene-expression patterns”. In: *Nucleic Acids Research* 38.SUPPL. 2, pp. 109–117.

Parmigiani, Giovanni et al. (2002). “A statistical framework for expression-based molecular classification in cancer”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64.4, pp. 717–736.

Paull, Evan O. et al. (Nov. 2013). “Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)”. In: *Bioinformatics* 29.21, pp. 2757–2764.

Pauta, Montse et al. (2015). “Overexpression of angiopoietin-2 in rats and patients with liver fibrosis. Therapeutic consequences of its inhibition”. In: *Liver International* 35.4, pp. 1383–1392.

Peng, S C et al. (2010). “Computational modeling with forward and reverse engineering links signalling network and genomic regulatory responses: NF-kappaB signalling-induced gene expression responses in inflammation”. In: *BMC Bioinformatics* 11.

Perrimon, Norbert, Chrysoula Pitsouli, and Ben-Zion Shilo (2012). “Signaling mechanisms controlling cell fate and embryonic patterning.” In: *Cold Spring Harbor perspectives in biology* 4.8, a005975.

Pesaresi, Martina, Ruben Sebastian-Perez, and Maria Pia Cosma (Mar. 2019). “Dedifferentiation, transdifferentiation and cell fusion: in vivo reprogramming strategies for regenerative medicine”. In: *The FEBS Journal* 286.6, pp. 1074–1093.

Pines, Alex et al. (Dec. 2011). *Global Phosphoproteome Profiling Reveals Unanticipated Networks Responsive to Cisplatin Treatment of Embryonic Stem Cells*.

Pirskanen, Asta, Julie C. Kiefer, and Stephen D. Hauschka (Aug. 2000). “IGFs, Insulin, Shh, bFGF, and TGF-β1 Interact Synergistically to Promote Somite Myogenesis in Vitro”. In: *Developmental Biology* 224.2, pp. 189–203.

Pukrop, T et al. (Apr. 2006). “Wnt 5a signaling is critical for macrophage-induced invasion of breast cancer cell lines”. In: *Proceedings of the National Academy of Sciences* 103.14, pp. 5454–5459.

Qian, Li et al. (May 2012). “In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes”. In: *Nature* 485.7400, pp. 593–598.

Qin, Hua, Andong Zhao, and Xiaobing Fu (2017). “Small molecules for reprogramming and transdifferentiation”. In: *Cellular and Molecular Life Sciences*.

Rackham, Owen J L et al. (Jan. 2016). “A predictive computational framework for direct reprogramming between human cell types”. In: *Nature Genetics* January, pp. 1–8.

Reddien, Peter W. and Alejandro Sánchez Alvarado (Nov. 2004). “FUNDAMENTALS OF PLANARIAN REGENERATION”. In: *Annual Review of Cell and Developmental Biology* 20.1, pp. 725–757.

Regan, Joseph L. et al. (Dec. 2017). “Non-Canonical Hedgehog Signaling Is a Positive Regulator of the WNT Pathway and Is Required for the Survival of Colon Cancer Stem Cells”. In: *Cell Reports* 21.10, pp. 2813–2828.

Reshef, R., M. Maroto, and A. B. Lassar (Feb. 1998). “Regulation of dorsal somitic cell fates: BMPs and Noggin control the timing and pattern of myogenic regulator expression”. In: *Genes & Development* 12.3, pp. 290–303.

Richter, Erik et al. (Mar. 2015). “A Multi-Omics Approach Identifies Key Hubs Associated with Cell Type-Specific Responses of Airway Epithelial Cells to Staphylococcal Alpha-Toxin”. In: *PLOS ONE* 10.3. Ed. by Antonino Passaniti, e0122089.

Rojas-Muñoz, Agustin et al. (Mar. 2009). “ErbB2 and ErbB3 regulate amputation-induced proliferation and migration during vertebrate regeneration”. In: *Developmental Biology* 327.1, pp. 177–190.

Rotival, Maxime et al. (2015). “Integrating Phosphoproteome and Transcriptome Reveals New Determinants of Macrophage Multinucleation”. In: *Molecular & Cellular Proteomics* 14.3, pp. 484–498.

Rudolph, Jan Daniel et al. (Dec. 2016). “Elucidation of Signaling Pathways from Large-Scale Phosphoproteomic Data Using Protein Interaction Networks”. In: *Cell Systems* 3.6, 585–593.e3.

Ruffell, Daniela et al. (Oct. 2009). “A CREB-C/EBP cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair”. In: *Proceedings of the National Academy of Sciences* 106.41, pp. 17475–17480.

Sacchetti, Benedetto et al. (2007). “Self-Renewing Osteoprogenitors in Bone Marrow Sinusoids Can Organize a Hematopoietic Microenvironment”. In: *Cell* 131.2, pp. 324–336.

Sader, Fadi et al. (Dec. 2019). “Epithelial to mesenchymal transition is mediated by both TGF-β canonical and non-canonical signaling during axolotl limb regeneration”. In: *Scientific Reports* 9.1, p. 1144.

Sartor, Maureen A, George D Leikauf, and Mario Medvedovic (Jan. 2009). “LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data.” In: *Bioinformatics (Oxford, England)* 25.2, pp. 211–7.

Satoh, Akira, David M. Gardiner, et al. (2007). “Nerve-induced ectopic limb blastemas in the axolotl are equivalent to amputation-induced blastemas”. In: *Developmental Biology* 312.1, pp. 231–244.

Satoh, Akira, Aki Makanae, et al. (July 2011). “Blastema induction in aneurogenic state and Prrx-1 regulation by MMPs and FGFs in Ambystoma mexicanum limb regeneration”. In: *Developmental Biology* 355.2, pp. 263–274.

Sawada, Ryusuke et al. (Dec. 2018). “Predicting inhibitory and activatory drug targets by chemically and genetically perturbed transcriptome signatures”. In: *Scientific Reports* 8.1, p. 156.

Schaefer, Carl F. et al. (2009). “PID: The pathway interaction database”. In: *Nucleic Acids Research* 37.SUPPL. 1, pp. 674–679.

Schubert, Michael et al. (Dec. 2018). “Perturbation-response genes reveal signaling footprints in cancer gene expression”. In: *Nature Communications* 9.1, p. 20.

Schuppan, Detlef and Nezam H Afdhal (Mar. 2008). “Liver cirrhosis”. In: *The Lancet* 371.9615, pp. 838–851.

Sebastian-Leon, P et al. (2014). “Understanding disease mechanisms with models of signaling pathway activities”. In: *BMC Syst Biol* 8.1, p. 121.

Selimkhanov, Jangir et al. (Dec. 2014). “Systems biology. Accurate information transmission through dynamic biochemical signaling networks.” In: *Science (New York, N.Y.)* 346.6215, pp. 1370–3.

Shamir, Maya et al. (Mar. 2016). “SnapShot: Timescales in Cell Biology”. In: *Cell* 164.6, 1302–1302.e1.

Sharma, Kirti et al. (Sept. 2014). “Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling”. In: *Cell Reports* 8.5, pp. 1583–1594.

Shi, Yan et al. (Nov. 2008). “Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Oct4 and Klf4 with Small-Molecule Compounds”. In: *Cell Stem Cell* 3.5, pp. 568–574.

Singh, Bhairab N et al. (Nov. 2012). “Hedgehog and Wnt coordinate signaling in myogenic progenitors and regulate limb regeneration”. In: *Developmental Biology* 371.1, pp. 23–34.

Singh, Bhairab et al. (June 2015). “Hedgehog Signaling during Appendage Development and Regeneration”. In: *Genes* 6.2, pp. 417–435.

Snijder, Berend et al. (2009). “Population context determines cell-to-cell variability in endocytosis and virus infection”. In: *Nature* 461.7263, pp. 520–523.

Söllner, Julia F. et al. (Dec. 2017). “An RNA-Seq atlas of gene expression in mouse and rat normal tissues”. In: *Scientific Data* 4, p. 170185.

Song, Guangqi et al. (June 2016). “Direct Reprogramming of Hepatic Myofibroblasts into Hepatocytes In Vivo Attenuates Liver Fibrosis”. In: *Cell Stem Cell* 18.6, pp. 797–808.

Srivastava, Deepak and Natalie DeWitt (2016). “In Vivo Cellular Reprogramming: The Next Generation”. In: *Cell* 166.6, pp. 1386–1396.

Stocum, David L. (Aug. 2017). “Mechanisms of urodele limb regeneration”. In: *Regeneration* 4.4, pp. 159–200.

Stocum, David L. and Jo Ann Cameron (2011). “Looking proximally and distally: 100 years of limb regeneration and beyond”. In: *Developmental Dynamics* 240.February, pp. 943–968.

Strasen, Jette et al. (2018). “Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels”. In: *Molecular Systems Biology* 14.1, e7733.

Su, Peihong et al. (Aug. 2018). “Mesenchymal Stem Cell Migration during Bone Formation and Bone Diseases Therapy”. In: *International Journal of Molecular Sciences* 19.8, p. 2343.

Subramanian, Aravind, Rajiv Narayan, et al. (Nov. 2017). “A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles”. In: *Cell* 171.6, 1437–1452.e17.

Subramanian, Aravind, Pablo Tamayo, et al. (Oct. 2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–50.

Sugiura, Takuji et al. (2016). “MARCKS-like protein is an initiating molecule in axolotl appendage regeneration”. In: *Nature*.

Sun, Junkui et al. (Sept. 2018). “Notch ligand Jagged1 promotes mesenchymal stromal cell-based cartilage repair”. In: *Experimental & Molecular Medicine* 50.9, p. 126.

Suzuki, Makoto et al. (Apr. 2007). “Transgenic Xenopus with prx1 limb enhancer reveals crucial contribution of MEK/ERK and PI3K/AKT pathways in blastema formation during limb regeneration”. In: *Developmental Biology* 304.2, pp. 675–686.

Szklarczyk, Damian et al. (Jan. 2016). “STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data”. In: *Nucleic Acids Research* 44.D1, pp. D380–D384.

Takahashi, Kazutoshi and Shinya Yamanaka (Aug. 2006). “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.” English. In: *Cell* 126.4, pp. 663–76.

Takeda, Yukimasa et al. (2018). “Chemical compound-based direct reprogramming for future clinical applications.” In: *Bioscience reports* 38.3.

Tamm, I et al. (May 2000). “Expression and prognostic significance of IAP-family genes in human cancers and myeloid leukemias.” In: *Clinical cancer research : an official journal of the American Association for Cancer Research* 6.5, pp. 1796–803.

Tang, Wanjin et al. (Jan. 2012). “Transcriptional Regulation of Vascular Endothelial Growth Factor (VEGF) by Osteoblast-specific Transcription Factor Osterix (Osx) in Osteoblasts”. In: *Journal of Biological Chemistry* 287.3, pp. 1671–1678.

Tang, Yuewen and Lin Cheng (Apr. 2017). “Cocktail of chemical compounds robustly promoting cell reprogramming protects liver against acute injury”. In: *Protein & Cell* 8.4, pp. 273–283.

Tarca, Adi Laurentiu et al. (Jan. 2009). “A novel signaling pathway impact analysis”. In: *Bioinformatics* 25.1, pp. 75–82.

Tasaki, Junichi et al. (June 2011). “ERK signaling controls blastema cell differentiation during planarian regeneration”. In: *Development* 138.12, pp. 2417–2427.

Taylor, Amy J. and Caroline W. Beck (Sept. 2012). “Histone deacetylases are required for amphibian tail and limb regeneration but not development”. In: *Mechanisms of Development* 129.9-12, pp. 208–218.

Thornton, Charles Stead (Mar. 1957). “The effect of apical cap removal on limb regeneration in Amblystoma larvae”. In: *Journal of Experimental Zoology* 134.2, pp. 357–381.

Tian, E et al. (July 2016). “Small-Molecule-Based Lineage Reprogramming Creates Functional Astrocytes”. In: *Cell Reports* 16.3, pp. 781–792.

Toettcher, Jared E., Orion D. Weiner, and Wendell A. Lim (Dec. 2013). “Using Optogenetics to Interrogate the Dynamic Control of Signal Transmission by the Ras/Erk Module”. In: *Cell* 155.6, pp. 1422–1434.

Tomlinson, Julianna J. et al. (Jan. 2010). “Insulin Sensitization of Human Preadipocytes through Glucocorticoid Hormone Induction of Forkhead Transcription Factors”. In: *Molecular Endocrinology* 24.1, pp. 104–113.

Tran, Freddi Huan and Jie J Zheng (Apr. 2017). “Modulating the wnt signaling pathway with small molecules”. In: *Protein Science* 26.4, pp. 650–661.

Tran, Khoa A. et al. (May 2015). “Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways”. In: *Nature Communications* 6.1, p. 6188.

Tran, Quynh T. et al. (Apr. 2012). “EGFR regulation of epidermal barrier function”. In: *Physiological Genomics* 44.8, pp. 455–469.

Tsai, Zong Yun et al. (2010). “Identification of microRNAs regulated by activin A in human embryonic stem cells”. In: *Journal of Cellular Biochemistry* 109.1, pp. 93–102.

Tseng, Ai-Sun et al. (Jan. 2007). “Apoptosis is required during early stages of tail regeneration in Xenopus laevis”. In: *Developmental Biology* 301.1, pp. 62–69.

Tsochatzis, Emmanuel A., Jaime Bosch, and Andrew K. Burroughs (May 2014). “Liver cirrhosis”. In: *The Lancet* 383.9930, pp. 1749–1761.

Tsuchida, Takuma and Scott L. Friedman (May 2017). “Mechanisms of hepatic stellate cell activation”. In: *Nature Reviews Gastroenterology & Hepatology* 14.7, pp. 397–411.

Tugues, Sònia et al. (Dec. 2007). “Antiangiogenic treatment with Sunitinib ameliorates inflammatory infiltrate, fibrosis, and portal pressure in cirrhotic rats”. In: *Hepatology* 46.6, pp. 1919–1926.

Türei, Dénes, Tamás Korcsmáros, and Julio Saez-Rodriguez (Nov. 2016). “OmniPath: guidelines and gateway for literature-curated signaling pathway resources.” In: *Nature methods* 13.12, pp. 966–967.

Uludag, Hasan, Anyeld Ubeda, and Aysha Ansari (June 2019). “At the Intersection of Biomaterials and Gene Therapy: Progress in Non-viral Delivery of Nucleic Acids”. In: *Frontiers in Bioengineering and Biotechnology* 7, p. 131.

Varrette, S et al. (2014). “Management of an Academic HPC Cluster: The UL Experience”. In: *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*. Bologna, Italy: IEEE, pp. 959–967.

Vieira, Warren A. and Catherine D. McCusker (2019). “Hierarchical pattern formation during amphibian limb regeneration”. In: *BioSystems* 183.May, p. 103989.

Vlastaridis, Panayotis et al. (2017). “Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes”. In: *GigaScience* 6.2, pp. 1–11.

Voloshanenko, Oksana et al. (Dec. 2018). “β-catenin-independent regulation of Wnt target genes by RoR2 and ATF2/ATF4 in colon cancer cells”. In: *Scientific Reports* 8.1, p. 3178.

Waddington, C H (1957). *The strategy of the genes: a discussion of some aspects of theoretical biology*. Tech. rep.

Wang, Daifeng et al. (Apr. 2015). “Loregic: a method to characterize the cooperative logic of regulatory factors.” In: *PLoS computational biology* 11.4, e1004132.

Washburn, Michael P et al. (Mar. 2003). “Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.6, pp. 3107–12.

Wierer, Michael et al. (June 2013). “PLK1 Signaling in Breast Cancer Cells Cooperates with Estrogen Receptor-Dependent Gene Transcription”. In: *Cell Reports* 3.6, pp. 2021–2032.

Wilkes, Edmund H et al. (2015). “Empirical inference of circuitry and plasticity in a kinase signaling network”. In: *Pnas*, pp. 1423344112–.

Wishart, David S et al. (Jan. 2018). “DrugBank 5.0: a major update to the DrugBank database for 2018”. In: *Nucleic Acids Research* 46.D1, pp. D1074–D1082.

Woo, Jung Hoon et al. (July 2015). “Elucidating Compound Mechanism of Action by Network Perturbation Analysis”. In: *Cell* 162.2, pp. 441–451. arXiv: 15334406.

Wouters, Bas J and Ruud Delwel (Jan. 2016). “Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia”. In: *Blood* 127.1, pp. 42–52.

Xiao, Yun et al. (Sept. 2015). “Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes”. In: *Scientific Reports* 5.1, p. 10889.

Xu, Huilei et al. (Aug. 2014). “Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells.” In: *PLoS computational biology* 10.8, e1003777.

Xu, Jun, Yuanyuan Du, and Hongkui Deng (Feb. 2015). “Direct Lineage Reprogramming: Strategies, Mechanisms, and Applications”. In: *Cell Stem Cell* 16.2, pp. 119–134.

Yachie-Kinoshita, Ayako et al. (2018). “Modeling signaling-dependent pluripotency with Boolean logic to predict cell fate transitions”. In: *Molecular Systems Biology* 14.1, e7952. arXiv: 1705.11170.

Yokoyama, Hitoshi et al. (Aug. 2018). “Skin regeneration of amphibians: A novel model for skin regeneration as adults”. In: *Development, Growth & Differentiation* 60.6, pp. 316–325.

Yu, Guangchuang et al. (May 2012). “clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters”. In: *OMICS: A Journal of Integrative Biology* 16.5, pp. 284–287.

Yun, M. H., P. B. Gates, and J. P. Brockes (Oct. 2013). “Regulation of p53 is critical for vertebrate limb regeneration”. In: *Proceedings of the National Academy of Sciences* 110.43, pp. 17392–17397.

Yun, Maximina H, Phillip B Gates, and Jeremy P Brockes (July 2014). “Sustained ERK activation underlies reprogramming in regeneration-competent salamander cells and distinguishes them from their mammalian counterparts.” English. In: *Stem cell reports* 3.1, pp. 15–23.

Zaffaroni, Gaia et al. (Apr. 2019). “An integrative method to predict signalling perturbations for cellular transitions”. In: *Nucleic Acids Research*, pp. 1–16.

Zañudo, Jorge G T and Réka Albert (2015). “Cell Fate Reprogramming by Control of Intracellular Network Dynamics”. In: *PLoS Computational Biology* 11.4, pp. 1–24. arXiv: 1408.5628.

Zhang, Bin et al. (Mar. 2018). “Estrogen receptor β selective agonist ameliorates liver cirrhosis in rats by inhibiting the activation and proliferation of hepatic stellate cells”. In: *Journal of Gastroenterology and Hepatology* 33.3, pp. 747–755.

Zhang, D.L. et al. (Jan. 2009). “Effect of Wnt signaling pathway on wound healing”. In: *Biochemical and Biophysical Research Communications* 378.2, pp. 149–151.

Zhang, Fan, Runsheng Liu, and Jie Zheng (2016). “Sig2GRN : a software tool linking signaling pathway with gene regulatory network for dynamic simulation”. In: *BMC Systems Biology* 10.Suppl 4.

Zhang, Lu, Yen Kaow Ng, and ShuaiCheng Li (Nov. 2015). “Reconstructing directed gene regulatory network by only gene expression data”. English. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 163–170.

Zhou, Jia and Renee L Sears (May 2018). “Bioinformatics Approaches to Stem Cell Research”. In: *Current Pharmacology Reports*.

Zhou, Qiao et al. (Oct. 2008). “In vivo reprogramming of adult pancreatic exocrine cells to β-cells”. In: *Nature* 455.7213, pp. 627–632.

Zhou, Wen-Ce (2014). “Pathogenesis of liver cirrhosis”. In: *World Journal of Gastroenterology* 20.23, p. 7312.

Zhu, Saiyong et al. (Dec. 2010). “Reprogramming of Human Primary Somatic Cells by OCT4 and Chemical Compounds”. In: *Cell Stem Cell* 7.6, pp. 651–655.

Zilliox, Michael J and Rafael A Irizarry (Nov. 2007). “A gene expression bar code for microarray data”. In: *Nature Methods* 4.11, pp. 911–913.

Zini, Roberta et al. (Dec. 2012). “Valproic acid triggers erythro/megakaryocyte lineage decision through induction of GFI1B and MLLT3 expression.” In: *Experimental hematology* 40.12, 1043–1054.e6.

7 Appendices

Table 7.1: Canonical pathways present in MetaCore from Clarivate Analytics included in the signalling network.

Class	Pathway	Short name
Signal transduction	Androgen receptor nuclear signaling	ARnuc
Signal transduction	Androgen receptor signaling cross-talk	ARcross
Signal Transduction	BMP and GDF signaling	BMP
Signal Transduction	Cholecystokinin signaling	cholecystokinin
Signal transduction	CREM pathway	CREM
Signal transduction	ERBB-family signaling	ERBB
Signal transduction	ESR1-membrane pathway	ESR1membr
Signal transduction	ESR1-nuclear pathway	ESR1nucl
Signal transduction	ESR2 pathway	ESR2
Signal transduction	Insulin signaling	insulin
Signal transduction	Leptin signaling	leptin
Signal transduction	Neuropeptide signaling pathways	neuropeptides
Signal transduction	Nitric oxide signaling	NO
Signal transduction	NOTCH signaling	NOTCH
Signal transduction	Oxytocin signaling	oxytocin
Signal Transduction	TGF-beta, GDF and Activin signaling	TGFb
Signal transduction	WNT signaling	Wnt
Inflammation	Amphoterin signaling	amphoterin
Inflammation	Complement system	compl
Inflammation	Histamine signaling	histamine
Inflammation	IFN-gamma signaling	IFN
Inflammation	IgE signaling	IgE
Inflammation	IL-10 anti-inflammatory response	IL10
Inflammation	IL-12,15,18 signaling	IL12
Inflammation	IL-13 signaling pathway	IL13

Inflammation	IL-2 signaling	IL2
Inflammation	IL-4 signaling	IL4
Inflammation	IL-6 signaling	IL6
Inflammation	Inflammasome	inflammasome
Inflammation	Innate inflammatory response	innate
Inflammation	Interferon signaling	interferon
Inflammation	Jak-STAT Pathway	JAKSTAT
Inflammation	Kallikrein-kinin system	kallikrein
Inflammation	MIF signaling	MIF
Inflammation	Neutrophil activation	neutrophil
Inflammation	NK cell cytotoxicity	NK
Inflammation	Protein C signaling	proteinC
Inflammation	TREM1 signaling	TREM1
Response to hypoxia and oxidative stress		hypoxia
Immune response	Antigen presentation	antigen
Immune response	BCR pathway	BCR
Immune response	IL-5 signalling	IL5
Immune response	Innate immune response to RNA viral infection	RNAviral
Immune response	Phagocytosis	phagocytosis
Immune response	Phagosome in antigen presentation	phagosome
Immune response	T helper cell differentiation	Th
Immune response	TCR signaling	TCR
Immune response	Th17-derived cytokines	Th17
Apoptosis	Anti-Apoptosis mediated by external signals by Estrogen	apop-estrogen
Apoptosis	Anti-Apoptosis mediated by external signals via MAPK and JAK/STAT	apop-MAPK
Apoptosis	Anti-apoptosis mediated by external signals via NF-kB	apop-NFkB
Apoptosis	Anti-Apoptosis mediated by external signals via PI3K/AKT	apop-PI3K
Apoptosis	Apoptosis stimulation by external signals	apop-external
Apoptosis	Death Domain receptors & caspases in apoptosis	death

Apoptosis	Endoplasmic reticulum stress pathway	ERstress
Development	Blood vessel morphogenesis	bloodVessel
Development	Regulation of angiogenesis	angiogenesis
Development	Cartilage development	cartilage
Development	Hedgehog signaling	HH
Development	Regulation of telomere length	telomere
Development	Regulation of epithelial-to-mesenchymal transition	EMT
Development	Keratinocyte differentiation	keratinocyte
Development	Melanocyte development and pigmentation	melanocyte
Cardiac development	BMP-TGF beta signaling	cardiac-BMP
Cardiac development	FGF-ErbB signaling	cardiac-FGF
Cardiac development	Role of NADPH oxidase and ROS	cardiac-NADPH
Cardiac development	Wnt-beta-catenin, Notch, VEGF, IP3 and integrin signaling	cardiac-Wnt
Development	Hemopoiesis, Erythropoietin pathway	hemopoiesis
Development	Skeletal muscle development	muscle
Development	ERK5 in cell proliferation and neuronal survival	ERK5
Development	Neurogenesis in general	neurogenesis
Development	Neurogenesis-Axonal guidance	axonal
Development	Neurogenesis-Synaptogenesis	synaptogenesis
Development	Neuromuscular junction	neuromuscular
Development	Ossification and bone remodeling	ossification

Table 7.2: Signalling interactions manually removed from the signalling network after literature review

from	**to**	**Effect**	**Mechanism**
ADAM17	ErbB4 = ErbB4(CTF)	Activation	Cleavage
gamma-Secretase complex	ErbB4(CTF) = ErbB4(ICD)	Activation	Cleavage
AKT(PKB)	SMAD7	Inhibition	Binding
TGF-beta receptor type I	SMAD7	Activation	Binding
SNAIL1	ID1	Activation	Binding
SMAD4	BRG1	Activation	Binding
ESR1 (nuclear)	PPAR-gamma	Inhibition	Binding
PPAR-gamma	STAT3	Inhibition	Binding
SMAD3	MYOG	Inhibition	Binding
PKC-alpha	PPAR-alpha	Inhibition	Phosphorylation
SMAD4	HNF4-alpha	Activation	Binding

Table 7.3: List of datasets analysed bu INCanTeSIMO. Type is either cm = datasets extracted from ConnectivityMap build 02 (Lamb et al. 2006), ae = datasets manually extracted from ArrayExpress corresponding to cell state transitions, pp = datasets for which phosphoproteomics data was available, or cf = datasets corresponding to cell fate transitions. Subfix is the internal subfix used for analysis. t = time after perturbation (h). NP = number of direct perturbation targets present in the signalling network. BR = ranking obtained by the best of them, NS = number of signalling molecules selected at 6% of the maximum rank (number of candidate molecules), success = correct prediction of any perturbation target among the selected candidates (T=true,F=false), NC = number of correctly predicted direct perturbation targets.

Type	GEO id	perturbation	t	cell type	array platform	NP	BR	NS	success	NC
cm		valproic acid	6	MCF7	HG-U133A	12	685	299	F	0
cm		alpha-estradiol	6	MCF7	HG-U133A	40	20	308	T	1
cm		tamoxifen	6	MCF7	HG-U133A	41	13	272	T	6
cm		dexverapamil	6	MCF7	HG-U133A	8	316	233	F	0
cm		sulindac	6	MCF7	HG-U133A	13	282	252	F	0
cm		tacrolimus	6	MCF7	HG-U133A	20	25	320	T	1
cm		rofecoxib	6	MCF7	HG-U133A	7	9	350	T	1
cm		celecoxib	6	MCF7	HG-U133A	4	74	299	T	1
cm		SC-58125	6	MCF7	HG-U133A	3	268	293	F	0
cm		tanespimycin	6	MCF7	HG-U133A	18	84	317	T	2
cm		rofecoxib	6	MCF7	HG-U133A	7	159	254	T	1
cm		indometacin	6	MCF7	HG-U133A	50	115	257	T	2
cm		prednisolone	6	MCF7	HG-U133A	37	88	305	T	1
cm		thalidomide	6	MCF7	HG-U133A	9	151	278	T	1
cm		genistein	6	MCF7	HG-U133A	22	43	387	T	1
cm		genistein	6	MCF7	HG-U133A	22	10	267	T	3
cm		fludrocortisone	6	MCF7	HG-U133A	2	158	257	T	1
cm		NU-1025	6	MCF7	HG-U133A	7	129	281	T	1
cm		acetylsalicylic acid	6	MCF7	HG-U133A	34	195	196	F	0
cm		LY-294002	6	MCF7	HG-U133A	22	13	253	T	5
cm		sirolimus	6	MCF7	HG-U133A	19	110	300	T	1
cm		LY-294002	6	MCF7	HG-U133A	22	38	311	T	5
cm		trichostatin A	6	MCF7	HG-U133A	18	23	236	T	3
cm		trichostatin A	6	MCF7	HG-U133A	18	86	287	T	1

cm		diclofenac	6	MCF7	HG-U133A	18	284	294	F	0
cm		nifedipine	6	MCF7	HG-U133A	6	123	293	T	1
cm		felodipine	6	MCF7	HG-U133A	5	318	273	F	0
cm		valproic acid	6	MCF7	HG-U133A	12	6	292	T	1
cm		valproic acid	6	MCF7	HG-U133A	12	39	356	T	1
cm		valproic acid	6	MCF7	HG-U133A	12	79	275	T	1
cm		LY-294002	6	HL60	HG-U133A	22	2	312	T	2
cm		sirolimus	6	HL60	HG-U133A	19	32	301	T	2
cm		fulvestrant	6	MCF7	HG-U133A	7	198	356	F	0
cm		rosiglitazone	6	HL60	HG-U133A	15	330	280	F	0
cm		troglitazone	6	HL60	HG-U133A	6	143	288	T	1
cm		raloxifene	6	ssMCF7	HG-U133A	39	86	295	T	1
cm		tamoxifen	6	MCF7	HG-U133A	41	136	330	T	1
cm		tanespimycin	6	MCF7	HG-U133A	18	114	274	T	2
cm		genistein	6	MCF7	HG-U133A	22	84	293	T	1
cm		raloxifene	6	HL60	HG-U133A	39	163	357	T	1
cm		wortmannin	6	HL60	HG-U133A	28	320	371	F	0
cm		sirolimus	6	ssMCF7	HG-U133A	19	67	402	T	2
cm		alpha-estradiol	6	ssMCF7	HG-U133A	40	27	262	T	1
cm		wortmannin	6	ssMCF7	HG-U133A	28	44	292	T	3
cm		valproic acid	6	HL60	HG-U133A	12	99	343	T	1
cm		nordihydroguaiaretic acid	6	ssMCF7	HG-U133A	3	866	298	F	0
cm		thioridazine	6	MCF7	HG-U133A	26	130	268	T	1
cm		haloperidol	6	MCF7	HG-U133A	36	69	333	T	5
cm		tanespimycin	6	MCF7	HG-U133A	18	6	328	T	2
cm		LY-294002	6	PC3	HG-U133A	22	36	283	T	3
cm		rosiglitazone	6	PC3	HG-U133A	15	49	374	T	1
cm		troglitazone	6	PC3	HG-U133A	6	142	390	T	1
cm		tanespimycin	6	PC3	HG-U133A	18	1	336	T	1

cm		arachidonic acid	6	MCF7	HG-U133A	10	1	228	T	2
cm		oligomycin	6	MCF7	HG-U133A	2	1137	327	F	0
cm		arachidonic acid	6	MCF7	HG-U133A	10	120	225	T	1
cm		trichostatin A	6	PC3	HG-U133A	18	6	240	T	1
cm		monorden	6	PC3	HG-U133A	12	27	239	T	1
cm		tanespimycin	6	PC3	HG-U133A	18	2	282	T	3
cm		indometacin	6	PC3	HG-U133A	50	3	350	T	5
cm		prochlorperazine	6	MCF7	HG-U133A	19	7	238	T	3
cm		valproic acid	6	PC3	HG-U133A	12	44	294	T	1
cm		LY-294002	6	PC3	HG-U133A	22	77	268	T	1
cm		troglitazone	6	PC3	HG-U133A	6	13	269	T	1
cm		monorden	6	PC3	HG-U133A	12	34	210	T	3
cm		fluphenazine	6	MCF7	HG-U133A	21	13	314	T	2
cm		iloprost	6	SKMEL5	HG-U133A	9	62	261	T	1
cm		LY-294002	6	SKMEL5	HG-U133A	22	149	476	T	1
cm		SC-58125	6	SKMEL5	HG-U133A	3	313	237	F	0
cm		tanespimycin	6	ssMCF7	HG-U133A	18	113	288	T	1
cm		nordihydroguaiaretic acid	6	ssMCF7	HG-U133A	3	153	322	T	1
cm		geldanamycin	6	MCF7	HG-U133A	7	163	264	F	0
cm		resveratrol	6	MCF7	HG-U133A	49	56	288	T	6
cm		thalidomide	6	MCF7	HG-U133A	9	125	279	T	1
cm		NU-1025	6	MCF7	HG-U133A	7	82	253	T	1
cm		geldanamycin	6	MCF7	HG-U133A	7	44	357	T	2
cm		pentamidine	6	MCF7	HG-U133A	4	843	279	F	0
cm		resveratrol	6	PC3	HG-U133A	49	85	230	T	2
cm		alpha-estradiol	6	PC3	HG-U133A	40	172	250	F	0
cm		genistein	6	PC3	HG-U133A	22	152	240	F	0
cm		fulvestrant	6	PC3	HG-U133A	7	127	395	T	2
cm		alpha-estradiol	6	MCF7	HG-U133A	40	3	270	T	2

cm		colforsin	6	HL60	HG-U133A	30	203	204	F	0
cm		naltrexone	6	HL60	HG-U133A	4	1245	358	F	0
cm		astemizole	6	HL60	HG-U133A	13	191	306	F	0
cm		gallamine triethiodide	6	HL60	HG-U133A	9	429	345	F	0
cm		nomifensine	6	HL60	HG-U133A	6	229	296	F	0
cm		nalbuphine	6	HL60	HG-U133A	6	1844	298	F	0
cm		spironolactone	6	HL60	HG-U133A	4	444	356	F	0
cm		terfenadine	6	HL60	HG-U133A	5	134	297	T	1
cm		mianserin	6	HL60	HG-U133A	24	244	287	F	0
cm		pirenzepine	6	HL60	HG-U133A	7	269	342	F	0
cm		thioproperazine	6	HL60	HG-U133A	8	1335	307	F	0
cm		pindolol	6	HL60	HG-U133A	6	586	335	F	0
cm		trichostatin A	6	HL60	HG-U133A	18	94	323	T	1
cm		thalidomide	6	HL60	HG-U133A	9	843	306	F	0
cm		tiratricol	6	HL60	HG-U133A	38	158	350	T	1
cm		tranylcypromine	6	HL60	HG-U133A	2	Inf	304	F	0
cm		flufenamic acid	6	HL60	HG-U133A	10	23	329	T	2
cm		trichostatin A	6	HL60	HG-U133A	18	106	285	T	4
cm		xylometazoline	6	HL60	HG-U133A	10	436	330	F	0
cm		nimesulide	6	HL60	HG-U133A	5	525	310	F	0
cm		oxymetazoline	6	HL60	HG-U133A	14	356	293	F	0
cm		tolfenamic acid	6	HL60	HG-U133A	2	736	303	F	0
cm		labetalol	6	HL60	HG-U133A	15	24	263	T	1
cm		oxybutynin	6	HL60	HG-U133A	6	21	347	T	1
cm		clonidine	6	HL60	HG-U133A	9	507	252	F	0
cm		cinnarizine	6	HL60	HG-U133A	14	157	370	T	2
cm		spiperone	6	HL60	HG-U133A	18	151	358	T	1
cm		trichostatin A	6	HL60	HG-U133A	18	235	316	F	0
cm		pimozide	6	HL60	HG-U133A	22	311	353	F	0

cm		mepacrine	6	HL60	HG-U133A	3	808	315	F	0
cm		clomipramine	6	HL60	HG-U133A	20	325	308	F	0
cm		mifepristone	6	HL60	HG-U133A	37	218	271	F	0
cm		alprenolol	6	HL60	HG-U133A	14	4	316	T	1
cm		fluphenazine	6	HL60	HG-U133A	21	144	338	T	1
cm		ketotifen	6	HL60	HG-U133A	5	62	337	T	1
cm		methapyrilene	6	HL60	HG-U133A	5	48	354	T	2
cm		dobutamine	6	HL60	HG-U133A	16	36	311	T	3
cm		betamethasone	6	HL60	HG-U133A	39	54	301	T	1
cm		ketanserin	6	HL60	HG-U133A	12	8	365	T	2
cm		zidovudine	6	HL60	HG-U133A	11	11	366	T	1
cm		desipramine	6	HL60	HG-U133A	18	437	302	F	0
cm		hemicholinium	6	HL60	HG-U133A	4	507	280	F	0
cm		phenylpropanolamine	6	HL60	HG-U133A	8	461	284	F	0
cm		metergoline	6	HL60	HG-U133A	20	42	297	T	1
cm		clenbuterol	6	HL60	HG-U133A	10	219	299	F	0
cm		maprotiline	6	HL60	HG-U133A	13	73	370	T	2
cm		dosulepin	6	HL60	HG-U133A	6	732	271	F	0
cm		resveratrol	6	HL60	HG-U133A	49	15	283	T	3
cm		budesonide	6	HL60	HG-U133A	7	82	318	T	1
cm		chloroquine	6	HL60	HG-U133A	4	731	278	F	0
cm		bromperidol	6	HL60	HG-U133A	5	179	326	T	1
cm		etamivan	6	HL60	HG-U133A	8	38	358	T	2
cm		cyclizine	6	HL60	HG-U133A	5	245	295	F	0
cm		trichostatin A	6	HL60	HG-U133A	18	7	288	T	5
cm		tubocurarine chloride	6	HL60	HG-U133A	28	121	301	T	1
cm		dihydroergocristine	6	HL60	HG-U133A	20	1198	306	F	0
cm		papaverine	6	HL60	HG-U133A	9	389	324	F	0
cm		tetrahydroalstonine	6	HL60	HG-U133A	4	913	343	F	0

cm		harmine	6	HL60	HG-U133A	16	7	278	T	1
cm		cytisine	6	HL60	HG-U133A	26	287	362	F	0
cm		atropine	6	HL60	HG-U133A	17	30	307	T	2
cm		physostigmine	6	HL60	HG-U133A	11	1637	257	F	0
cm		berberine	6	HL60	HG-U133A	11	488	317	F	0
cm		trichostatin A	6	HL60	HG-U133A	18	65	333	T	4
cm		quipazine	6	HL60	HG-U133A	35	472	253	F	0
cm		sulfathiazole	6	PC3	HG-U133A	2	1053	303	F	0
cm		amiloride	6	PC3	HG-U133A	5	250	366	F	0
cm		trichostatin A	6	PC3	HG-U133A	18	153	343	T	1
cm		levodopa	6	PC3	HG-U133A	3	649	271	F	0
cm		thioridazine	6	PC3	HG-U133A	30	92	330	T	7
cm		captopril	6	PC3	HG-U133A	17	210	344	F	0
cm		diflunisal	6	PC3	HG-U133A	1	Inf	416	F	0
cm		lidocaine	6	PC3	HG-U133A	16	1	256	T	2
cm		naloxone	6	PC3	HG-U133A	13	300	289	F	0
cm		bromocriptine	6	PC3	HG-U133A	12	41	252	T	4
cm		amoxapine	6	PC3	HG-U133A	12	17	219	T	2
cm		dipyridamole	6	PC3	HG-U133A	3	504	314	F	0
cm		edrophonium chloride	6	PC3	HG-U133A	2	740	329	F	0
cm		cyproheptadine	6	PC3	HG-U133A	21	41	293	T	2
cm		ciprofloxacin	6	PC3	HG-U133A	14	53	286	T	2
cm		famotidine	6	PC3	HG-U133A	2	46	313	T	1
cm		loperamide	6	PC3	HG-U133A	28	266	286	F	0
cm		trichostatin A	6	PC3	HG-U133A	18	65	284	T	5
cm		danazol	6	PC3	HG-U133A	9	50	350	T	1
cm		perphenazine	6	PC3	HG-U133A	19	8	353	T	3
cm		paclitaxel	6	PC3	HG-U133A	9	1118	332	F	0
cm		lisuride	6	PC3	HG-U133A	25	57	220	T	1

cm		sulfathiazole	6	HL60	HG-U133A	2	2127	304	F	0
cm		sulpiride	6	HL60	HG-U133A	5	614	279	F	0
cm		amiloride	6	HL60	HG-U133A	5	297	319	F	0
cm		pyrimethamine	6	HL60	HG-U133A	4	777	329	F	0
cm		dicycloverine	6	HL60	HG-U133A	6	106	323	T	1
cm		thioridazine	6	HL60	HG-U133A	30	90	353	T	1
cm		captopril	6	HL60	HG-U133A	17	2	318	T	1
cm		diflunisal	6	HL60	HG-U133A	1	Inf	339	F	0
cm		apomorphine	6	HL60	HG-U133A	16	29	342	T	1
cm		naloxone	6	HL60	HG-U133A	13	321	283	F	0
cm		bromocriptine	6	HL60	HG-U133A	12	17	247	T	1
cm		amoxapine	6	HL60	HG-U133A	12	233	279	F	0
cm		loxapine	6	HL60	HG-U133A	24	267	320	F	0
cm		dipyridamole	6	HL60	HG-U133A	3	803	308	F	0
cm		edrophonium chloride	6	HL60	HG-U133A	2	906	335	F	0
cm		ciprofloxacin	6	HL60	HG-U133A	14	308	342	F	0
cm		famotidine	6	HL60	HG-U133A	2	64	381	T	1
cm		loperamide	6	HL60	HG-U133A	28	493	346	F	0
cm		trichostatin A	6	HL60	HG-U133A	18	118	324	T	3
cm		haloperidol	6	HL60	HG-U133A	36	61	262	T	3
cm		perphenazine	6	HL60	HG-U133A	19	182	287	F	0
cm		methotrexate	6	HL60	HG-U133A	2	42	266	T	1
cm		paclitaxel	6	HL60	HG-U133A	9	1215	278	F	0
cm		lisuride	6	HL60	HG-U133A	25	30	264	T	4
ae	GSE10778	VEGF	1	HUVEC	HG-U133A	7	268	272	F	0
ae	GSE10778	EGF	1	HUVEC	HG-U133A	11	356	311	F	0
ae	GSE11367	IL-17	6	VSMC	HG-U133_Plus_2	1	47	401	T	1
ae	GSE14419	ZymosanA	3	macrophage	HG-U133A_2	2	171	292	T	1

ae	GSE30242	mometasone furoate	6	lung fibroblast	HG-U133A	3	83	273	T	1
ae	GSE35830	TGF-B3	10	Ect1 ectocervical epithelial cell	HG-U133_Plus_2	6	2	408	T	2
ae	GSE27313	Wnt3a	6	MSC	HG-U133_Plus_2	4	118	282	T	1
ae	GSE16450	IFN-a	6	BE(2)-C	HG-U133_Plus_2	2	97	243	T	1
ae	GSE41683	dexamethasone	24	preadipocytes	HuGene-1_0-st	4	785	339	F	0
ae	GSE32217	EGF	48	keratinocyte	HuGene-1_0-st	11	425	262	F	0
pp	GSE6462	EGF	6	MCF7	HG-U133A_2	11	91	357	T	2
pp	GSE6783	EGF	8	HeLa	HG-U133A	11	142	325	T	2
pp	GSE11710	TGF-B	6	HaCaT	HG-U133_Plus_2	15	72	308	T	1
pp	GSE18232	cobimetinib	48	HCT116	HG-U133A	2	1214	289	F	0
pp	GSE11506	estradiol	3	MCF7	HG-U133_Plus_2	7	78	331	T	1
pp	GSE6521	PD-168393	1	MCF7	HG-U133_Plus_2	3	130	334	T	1
cf	GSE10315	BMP2	24	MSC	HG-U133_Plus_2	19	10	396	T	5
cf	GSE10315	BMP2	504	MSC	HG-U133_Plus_2	19	374	306	F	0
cf	GSE10315	TGF-B3	24	MSC	HG-U133_Plus_2	6	38	402	T	1
cf	GSE10315	TGF-B3	504	MSC	HG-U133_Plus_2	6	21	345	T	1
cf	GSE31283	valproic acid	48	HSC	HG-U133A	12	5	227	T	2
cf	GSE19393	dexamethasone	48	preadipocytes	HG-U133_Plus_2	4	93	356	T	1
cf	GSE19393	dexamethasone	48	preadipocytes	HG-U133_Plus_2	4	255	362	F	0
cf	GSE6460	FGF2	168	MSC	HG-U133_Plus_2	12	25	322	T	2
cf	GSE6460	TGF-B	168	MSC	HG-U133_Plus_2	15	171	259	T	1
cf	GSE32217	EGF	48	keratinocyte	HuGene-1_0-st	11	1	252	T	2
cf	GSE39157	cAMP	-	hepatoblast	HuGene-1_0-st	10	143	272	T	1
cf	GSE16910	activin A	-	hES-T3	HG-U133_Plus_2	10	86	422	T	1
cf	GSE16910	activin A	-	hES-T3	HG-U133_Plus_2	10	15	354	T	3
cf	GSE57032	R848	168	mMDSC	OpArray Human 35K	2	2358	311	F	0
cf	GSE57032	PAM3	168	mMDSC	OpArray Human 35K	3	78	329	T	1

cf	GSE51398	inhibition of Atoh	-	ISC	Mouse430A_2	9	529	370	F	0
cf	GSE51398	dibenzazepine	-	ISC	Mouse430A_2	10	56	359	T	1
cf	GSE98147	CHIR99021	-	dermomyotome	HuGene-1_0-st	2	641	306	F	0
cf	GSE69924	SB-431542	576	MEF	Mouse430_2	4	1421	372	F	0

Table 7.4: Signalling molecules predicted for the reversal of the cirrhotic state in rat liver. The molecules obtaining rank up to 50 across correlation-based and length-based predictions are reported.

Signalling molecule	**Rank**	**Gene symbols**	**Receptor**
NFYA	1	NFYA	
ESR2	1	ESR2	yes
ADA3-like protein	1	TADA3	
AHR	1	AHR	yes
MMP-26__inh	1	MMP26	
CHIP__inh	1	STUB1	
c-Jun/c-Fos__inh	6	FOS;JUN	
CBP__inh	7	CREBBP	
IFN-alpha__inh	7	IFNA1;IFNA10;IFNA13;IFNA14;IFNA16;IFNA17;IFNA2;IFNA21;IFNA4;IFNA5;IFNA6;IFNA7;IFNA8	
CIITA	8	CIITA	
p90Rsk	8	RPS6KA1;RPS6KA2;RPS6KA3	
NALP12	8	NLRP12	
Somatotropin__inh	9	GH1	
MDM2__inh	10	MDM2	
IRF1__inh	10	IRF1	
Pitx2__inh	11	PITX2	
SAP__inh	11	SH2D1A	
CD86	12	CD86	yes
CTLA-4	12	CTLA4	
IL-2R beta chain__inh	12	IL2RB	yes
AP1G1__inh	12	AP1G1	
AP-1 beta subunit__inh	12	AP1B1	
AP complex 2 medium (mu) chain__inh	12	AP2M1	
AP1M1__inh	12	AP1M1	
Beta-adaptin 2__inh	12	AP2B1	

IFN-beta__inh	13	IFNB1	
p53	14	TP53	
Prolactin__inh	15	PRL	
SDF-1__inh	16	CXCL12	
Endothelin-1__inh	17	EDN1	
ECE2__inh	17	ECE2	
p70 S6 kinase2	19	RPS6KB2	
IRS-1__inh	19	IRS1	yes
MMP-2__inh	19	MMP2	
TIMP1	20	TIMP1	
ClO('-) intracellular__inh	21		
TIRAP (Mal)__inh	21	TIRAP	yes
Androgen receptor	22	AR	yes
ADAR1	22	ADAR	
double-stranded RNA__inh	22		
GHR__inh	23	GHR	yes
TIE2	24	TEK	yes
L-Carnitine cytoplasm	24		
Glucocorticoids intracellular	24		
Angiopoietin 3__inh	24	ANGPTL1	
VEGF-A	26	VEGFA	
MMP-19	26	MMP19	
PTP-1B	26	PTPN1	yes
NPX1	26	NPTX1	
Matrilysin (MMP-7)	26	MMP7	
IGFBP7/8__inh	26	CTGF;IGFBP7	
Pleiotrophin (OSF1)__inh	26	PTN	
Thrombopoietin__inh	27	THPO	
C/EBPbeta__inh	30	CEBPB	
IL-22RA2	31	IL22RA2	yes

IL-22__inh	31	IL22	
Angiopoietin 1	32	ANGPT1	
Angiopoietin 4	32	ANGPT4	
Angiopoietin 2__inh	32	ANGPT2	
C1q	33	C1QA;C1QB;C1QC	
Calreticulin__inh	33	CALR	
N-Acetyl-D-glucosamine intracellular__inh	33		
MBL2__inh	33	MBL2	
D-Mannose extracellular region__inh	33		
C4BP alpha__inh	33	C4BPA	
C1qRp__inh	33	CD93	yes
PKC-beta	35	PRKCB	
ACES	35	ACHE	
COLQ	35	COLQ	
Adenosine A2a receptor__inh	38	ADORA2A	yes
p120GAP	39	RASA1	
GM-CSF receptor__inh	40	CSF2RA;CSF2RB	yes
c-MPL__inh	40	MPL	yes
FGF1__inh	40	FGF1	
Heparin extracellular region__inh	40		
IL-3 receptor__inh	40	CSF2RB;IL3RA	yes
M-CSF receptor__inh	43	CSF1R	yes
IL-3__inh	43	IL3	
IFN-kappa__inh	44	IFNK	
IFN-alpha/beta receptor__inh	44	IFNAR1;IFNAR2	yes
IFN-omega__inh	44	IFNW1	
CSF1__inh	44	CSF1	
ICAM3	45	ICAM3	

GATA-4	46	GATA4	
IL-15__inh	47	IL15	
IMP extracellular region__inh	47		
Detralfate extracellular region__inh	47		
JAB1__inh	48	COPS5	
CISH	49	CISH	
ULBP2__inh	49	ULBP2	
RAET1G__inh	49	RAET1G	
ICOS-L__inh	49	ICOSLG;LOC102723996	
RAET1E__inh	49	RAET1E	
Prolactin receptor__inh	50	PRLR	yes
Lactogen__inh	50	CSH2	

Table 7.5: Top 20 GO terms enriched in the TFs contained in the GRN specific for each time step. dpa=days post amputation

0 to 1 dpa	1 to 3 dpa	5 to 7 dpa	7 to 10 dpa	10 to 14 dpa
positive regulation of MAP kinase activity	positive regulation of nucleic acid-templated transcription	histone modification	cation transport	regulation of interleukin-12 production
supramolecular fiber organization	cell cycle	dephosphorylation	extracellular matrix organization	blood coagulation
nuclear export	ribosome biogenesis	in utero embryonic development	heart valve morphogenesis	positive regulation of protein metabolic process
organic hydroxy compound transport	membrane organization	positive regulation of MAP kinase activity	smooth muscle cell differentiation	response to mechanical stimulus
negative regulation of protein modification process	ribonucleoprotein complex assembly	nucleic acid phosphodiester bond hydrolysis	negative regulation of apoptotic process	cellular process involved in reproduction in multicellular organism
response to lipopolysaccharide	transmembrane receptor protein tyrosine kinase signaling pathway	positive regulation of organelle organization	maintenance of location	negative regulation of protein binding
histone modification	protein ubiquitination	negative regulation of translation	cell differentiation	regulation of GTPase activity
positive regulation of intracellular protein transport	ubiquitin-dependent protein catabolic process	female pregnancy	trabecula morphogenesis	myeloid cell development

Table 7.5: Top 20 GO terms enriched in the TFs contained in the GRN specific for each time step. dpa=days post amputation

0 to 1 dpa	1 to 3 dpa	5 to 7 dpa	7 to 10 dpa	10 to 14 dpa
protein-containing complex subunit organization	positive regulation of cell growth	embryonic limb morphogenesis	regulation of transcription from RNA polymerase II promoter in response to hypoxia	osteoclast differentiation
negative regulation of locomotion	negative regulation of protein modification by small protein conjugation or removal	ribonucleoprotein complex assembly	ventricular cardiac muscle cell differentiation	regulation of cytokine biosynthetic process
plasma membrane bounded cell projection assembly	RNA localization	extrinsic apoptotic signaling pathway	dephosphorylation	cellular response to external stimulus
regulation of protein binding	regulation of protein ubiquitination	positive regulation of cell proliferation	humoral immune response	chromatin organization
response to mechanical stimulus	response to purine-containing compound	positive regulation of MAPK cascade	calcineurin-mediated signaling	alpha-beta T cell differentiation
protein import into nucleus	G protein-coupled receptor signaling pathway	response to hydrogen peroxide	outflow tract morphogenesis	response to lipopolysaccharide
positive regulation of epithelial to mesenchymal transition	response to drug	actin filament organization	response to fluid shear stress	regulation of cysteine-type endopeptidase activity involved in apoptotic process

Table 7.5: Top 20 GO terms enriched in the TFs contained in the GRN specific for each time step. dpa=days post amputation

0 to 1 dpa	**1 to 3 dpa**	**5 to 7 dpa**	**7 to 10 dpa**	**10 to 14 dpa**
viral process	negative regulation of hydrolase activity	regulation of blood pressure	heart formation	non-canonical Wnt signaling pathway
positive regulation of cell adhesion	protein stabilization	inflammatory response	renal tubule morphogenesis	positive T cell selection
cellular response to nutrient	regulation of signaling receptor activity	tissue remodeling	response to tumor necrosis factor	lymphocyte apoptotic process
negative regulation of lipid localization	microtubule cytoskeleton organization	multicellular organism development	negative regulation of cell migration	regulation of toll-like receptor signaling pathway
regulation of smooth muscle cell proliferation	histone H3 acetylation	activation of protein kinase activity	antigen receptor-mediated signaling pathway	regulation of cell size

7.1 Published papers

András Hartmann, Satoshi Okawa, et al. (Dec. 2018). “SeesawPred: A Web Application for Predicting Cell-fate Determinants in Cell Differentiation”. In: *Scientific Reports* 8.1, p. 13355

Gaia Zaffaroni et al. (Apr. 2019). “An integrative method to predict signalling perturbations for cellular transitions”. In: *Nucleic Acids Research*, pp. 1–16