UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2021-069
The Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 09/09/2021 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Menglin ZHENG
Born on 09 April 1990 in Fujian (China)

COMPUTATIONAL GUIDED FRAMEWORKS TO IDENTIFY SIGNALLING PERTURBATIONS FOR CELLULAR TRANSITIONS: APPLICATION TO CELLULAR CONVERSION, DISEASE AND REGENERATION

Dissertation defence committee

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

Dr Jorge Goncalves
Professor, Université du Luxembourg

Dr Jens Christian Schwamborn, Chairman
Professor, Université du Luxembourg

Dr Oscar Millet
CIC bioGUNE-BRTA (Center for Cooperative Research in Biosciences)

Dr Jan Rehwinkel
Professor, University of Oxford

Acknowledgements

I have been engaged in bioinformatics since my undergraduate degree. I enjoy and immerse myself in the process of addressing biological problem with big data. Therefore, I decided to continue master and PhD study after my undergraduate. I am honored to have the opportunity to complete my PhD in LCSB, where I met a lot of wonderful people, and they gave me a lot of help and encouragement. Here, I would like to express my gratitude to them.

First of all, I would like to express my sincere thanks to my supervisor, Prof. Dr. Antonio del Sol, for his guidance and help over the past four years. In countless brainstorming sessions, he taught me how to think critically, how to come up with questions and find ways to solve them. I think this is also the most important thing at the doctoral level.

Secondly, I would like to thank my CET members, Prof. Dr. Jorge Goncalves and Dr. Feng He, for evaluating my academic achievements and giving very valuable suggestions in every CET meeting.

I also would like to thank my collaborators, Prof. Dr. Hongkui Deng, Prof. Dr. Elly Tanaka and Prof. Dr. Jan Rehwinkel, for their collaboration and contribution, making my research completer and more persuasive.

Many thanks go to my dear colleagues in Computational Biology group. They create a lovely atmosphere and gave me a lot of pleasant time and academic support. Especially, Dr. Satoshi Okawa gave me a lot of feedbacks and advices when I got stuck in my study.

Furthermore, I would like to thank my family and friends, especially my husband. I am grateful for their selfless support and encouragement to get me out of the predicament during my PhD study.

Finally, I would like to thank LCSB for providing an amazing academic platform and thank Fond National de la Recherche Luxembourg for financial support for my PhD.

Affidavit

I hereby confirm that the PhD thesis entitled “Computational guided frameworks to identify signalling perturbations for cellular transitions: application to cellular conversion, disease and regeneration” has been written independently and without any other sources than cited.

Luxembourg, 08/07/2021

Menglin Zheng

Table of Contents

List of Figures

List of Tables

Summary

Transition of cellular identity or phenotype holds great promise for the clinical applications, such as regenerative medicine and disease treatment. Studies have shown that cellular transitions can be induced by perturbation of a handful of key transcription factors (TFs). However, TF-based cellular transitions require transfer of genetic materials, which have raised safety concerns and limit their translation into clinical applications. Replacement of the direct manipulation of TFs with signalling perturbations offers a more controlled and safer way to accomplish such transitions. Nevertheless, the identification of optimal signalling perturbations relies on cell-based phenotypic or pathway-based screenings of chemical libraries, which is lengthy, costly and labour intensive. Therefore, a systematic guidance from the computational perspective to identify signalling perturbation for cellular transitions is desirable.

To date, several computational methods have been developed to predict signalling perturbations that are responsible for the observed gene expression dysregulations. Some of these methods infer external stimuli from perturbation databases whose perturbations result in similar transcriptional signatures as desired. However, the databases in these methods mainly involve perturbation datasets for cancer cells, which are not suitable for the non-cancers study due to the extensive rewiring of signalling pathways in cancer cells. Another group of methods try to infer causal upstream signaling perturbations resulting in observed dysregulated genes directly or identify dysregulated pathways or sub-pathways by using some enrichment measures given the dysregulated genes. These methods use gene expression as proxy for signalling activity due to the scarcity of protein activity measures. Therefore, it is unclear whether signalling activity can be reflected without considering the presence of post-translational modifications.

In this thesis, a wide range of perturbation datasets were manually collected and compiled to construct databases consisting solely of non-cancer cells. Computational methods that integrated the perturbation database with a network-based model were proposed. One method named SiPer was developed to identify chemical compounds specifically targeting given sets of TFs to induce the desired change of cellular state. The method was applied to chemical-based cellular conversions and recapitulated the compounds used in existing protocols, including conversions between cell types, functional cell subtypes and phenotypic states. Moreover, by applying this method, we successfully developed a novel and efficient protocol to drive the conversion of hepatic progenitors into functional human induced

hepatocytes. Another method we devised, termed ChemPert, was designed to predict chemical compounds to induce cellular transition from the given initial state to the given final state. ChemPert was applied to prioritize chemical compounds to revert pathologic phenotypes of various diseases, including age-related diseases and infectious diseases. A considerable number of state-of-the-art therapeutics alongside potentially novel candidates were predicted, underling the potential of this method for drug discovery. Lastly, in order to identify signalling pathways and proteins to induce cellular transitions for the organisms lacking substantial prior knowledge of perturbation, a de novo inference method was developed by combining signalling and gene regulatory networks. This method was applied to the analysis of salamander limb regeneration and the predicted signaling cues along regeneration process were substantially consistent with the literature.

Taken together, these computational methods proposed in this thesis provide a systematic guidance to identify signalling perturbations for designing new experimental strategies for gene therapies and *in vitro* cell engineering for cell transplantation, as well as for designing therapeutic strategies for disease treatment.

1 Introduction

Cellular transition is triggered by a conserved set of inducing signals, transcriptional regulators and downstream effectors and happens across a broad range of physiological and pathological conditions. During development, cells differentiate from the committed progenitors to the terminally differentiated cells that carry out their specialized function. This differentiation occurs constantly to ensure stable tissue mass and function, which is a process of natural tissue generation or rejuvenation. Cellular transition is also required to adapt the external influences such as natural fluctuations exerted by the niche and disturbance under pathologic conditions. One extreme example of this is the regeneration in some organisms, where the cells at the position of wound dedifferentiate into progenitors to form blastema after loss of tissue. Then, these cells undergo proliferation, patterning and differentiation to regenerate the lost tissue. Therefore, the induction of desired cellular transition is an appealing therapeutic strategy for the treatment of disease, including regenerative medicine and drug discovery.

Cellular transition is driven by different regulatory layers, including signalling, transcriptional and epigenetic layers. A milestone of cellular transition for cell engineering is that the somatic cells are converted into pluripotent stem cells by ectopic expression of four TFs (Takahashi and Yamanaka, 2006). After that, a breakthrough number of cellular conversions by manipulating TFs, including direct conversions between somatic cells are reported. However, these conversions that involve the transfer of genetic materials, have raised safety concerns, which limit their translation into therapeutic applications. Alternatively, the usage of chemical compounds acting on signalling network not only addresses these concerns, but also offers an easily controlled and effective strategy for cellular conversion. On the other hand, signalling proteins are normally the targeting points of drugs to induce the reversion of disease phenotype to its healthy counterpart. In regeneration, external fluctuation causing by injury stimulates sets of signalling pathways to trigger changes in the downstream gene regulatory network (GRN) in order to induce defined cellular decisions to regenerate or repair the target tissue. Therefore, identifying signalling perturbations that can induce cellular transitions holds great promise to design strategy for regenerative medicine and discover drugs for disease treatment efficiently. However, the discovery of signalling cues for cellular reprogramming or reversing disease phenotype mainly accomplished by experimental testing by trial and error, which requires a large amount of resource and time, and burdens by a low success rate. Over the years, a number of computational tools have been developed for identifying siganlling elements based on different concepts and principles, which provide

valuable and complementary guidance along with experimental investigation. Nevertheless, still some significant limitations exist in the developed methods and need to be addressed for being generally applicable.

The remainder of this chapter describes the current knowledge about cellular transitions, as well as reviews existing methods for the identification of signalling cues. Specifically, the types of cellular transitions and their relevant applications are introduced in Section 1.1. Followed by delineating the regulatory layers of cellular transitions in Section 1.2, Section 1.3 provides a detailed review of current exiting methods to identify singalling perturbation and discusses their limitations.

1.1 Categories and applications of cellular transitions

In this thesis, the cellular transition is defined as the conversion of a cell from an initial cellular state to a desired target state. This ranges from cell (sub)type conversion, to cell phenotypic state changes, such as different functional or morphological states. Here, we broadly classify the cellular transitions into two categories: cell fate transitions and cell phenotypic state transitions, depending upon the different characteristics of initial and target states.

Cell fate transitions also refer to cell (sub)type conversion, including differentiation, transdifferentiation and reprogramming, which focus on the changes in cellular identity. In 1957, Conrad Waddington came up with the epigenetic landscape model, which intuitively represented the cellular differentiation as ball rolling down from a hill into valleys. The stem and progenitor cells descend along distinct differentiation routes toward more differentiated cells, and finally reach terminally specialized cells. While this epigenetic landscape model was used as explanation for cellular transitions and thought to be unidirectional for many years, studies on cellular plasticity later showed that cellular development process can be reversed and should be multidirectional. One breakthrough work is the direct cellular conversion from fibroblasts into myoblasts by forced ectopic expression of one transcription factor MYOD by Davis et al. in 1987. This study achieved the transdifferentiation that one differentiated cell type was directly converted to another without going through a pluripotent cellular state. Subsequent transdifferentiation studies further changed the perception of one-way cellular transitions, such as the direct cell type conversions from fibroblasts to neurons (Vierbuchen et al., 2010), cardiomyocytes (Ieda et al., 2010) and hepatocytes (Huang et al., 2011), as well as conversions between cell subtypes like T cells (Youngblood et al., 2017) and neurons (Okawa et al., 2018). Another remarkable study reprogrammed somatic cells into induced pluripotent

stem cells (iPSCs) by using four TFs, pioneered by Yamanaka et al. in 2006 (Takahashi and Yamanaka, 2006). After the induction of iPSCs, many protocols were developed to generate various cell types from iPSCs, such as neurons (Chambers et al., 2009; Dimos et al., 2008), adipocytes (Taura et al., 2009a), vascular cells (Taura et al., 2009b) and hepatocytes (Si-Tayeb et al., 2010). These findings open up a new avenue for therapeutic strategy in regenerative medicine.

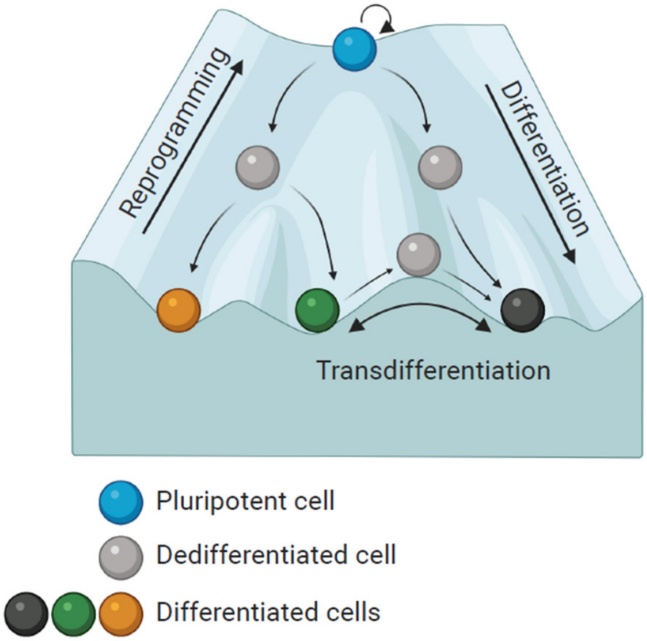

Figure 1. 1 Description of cell fate transition on Waddington landscape.

Three types of cell fate transition can be distinguished from the Waddington landscape, including differentiation, reprogramming and transdifferentiation. Pluripotent cells can be differentiated to distinct terminally differentiated cells. Reversely, the totally differentiated somatic cells can also be converted to pluripotent cells, which is known as reprogramming. In addition, the somatic cells can switch directly between cell lineages, bypassing the pluripotent states. Figure taken from (Granados et al., 2020)

The induction of desired cell fate transitions provides valuable resources for multiple clinical applications. First, the derivation of iPSCs sparks widespread enthusiasm to develop models for human disorders as a replacement of the animal models, due to nonnegligible physiological and genetic differences between human and animals (Avior et al., 2016). In addition, the own advantages of iPSCs: patient-specific primary cell lines, indefinite self-renewal and high capacity of differentiation, allow them as suitable models for the study of a wide range of diseases. For example, the neurons derived from iPSCs of Alzheimer's disease (AD) patient facilitated the exploration of cellular and molecular mechanisms underlying AD

with live human neurons (Chang et al., 2019). Furthermore, cell fate transition significantly promotes the development of regenerative medicine, aiming at replacing or impairing the damaged cells with regenerated cells to restore the normal function of cells or tissues. There are mainly two different strategies for regenerative medicines: *in vivo* generation and *ex vivo* transplantation (Figure 1.2). The first strategy is to convert the resident cells within injured tissues into desired cell types in situ to impair the damaged tissues by external stimuli. A few of organisms, such as planarian flatworms, Xenopus tadpoles, zebrafish and urodele amphibians (salamanders and newts) are able to regenerate their tissues, whereas the regeneration ability is significantly attenuated in mammals. However, great efforts have been put on *in vivo* reprogramming for the treatment of disease or injury, such as the *in vivo* generation of hepatocytes to ameliorate liver fibrosis (Song et al., 2016) and conversion of astrocytes into neurons with potential to repair injured spinal cord (Su et al., 2014), garnering attention for the potential of tissue regeneration in mammals. The transplantation of cells or tissues follows the strategy of injecting healthy/modified cells into damaged tissues of patient to restore the functionality of cells and tissues (Sampogna et al., 2015). The transplantation of in vitro induced cells has been applied in clinics, such as the treatment of junctional epidermolysis bullosa (Hirsch et al., 2017), and primary immune deficiencies (Kuo and Kohn, 2016).

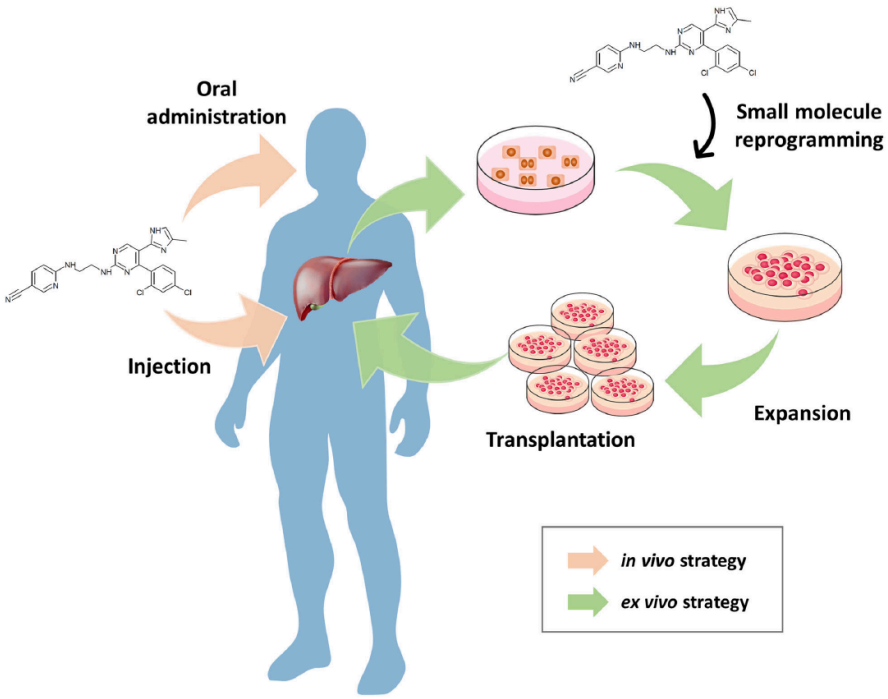

Figure 1. 2 Two strategies for regenerative medicine.

An ex vivo strategy can generate the desired cells through in vitro reprogramming, and these cells are then expanded and transplanted into patients (green arrow). The in vivo strategy is to

generate desired cells by in vivo reprogramming while external stimuli are given, for example, oral administration or injection of small molecules for patients (pink arrow). Figure taken from (Kim et al., 2020).

Cell phenotypic state transitions focus on the changes of cellular phenotypes under different conditions, while the cellular identity is unaffected. One representative example is transition between healthy and disease states. The induction of a disease can be considered as a transition between phenotypic states at the cellular level, since cellular identity might maintain, whereas signalling activity, epigenetic and metabolic states, gene expression and function can be compromised. Cell phenotypic state transition is also a strategy for cells to ensure the tissue maintenance. Especially, tissue-specific stem cells, such as hematopoietic stem cells (Bernitz et al., 2016; Takizawa et al., 2011), neural stem cells (Codega et al., 2014) and muscle stem cells (Cheung et al., 2012), convert between quiescent and active states to maintain the tissue homeostasis and repair tissues after injury. Specifically, stem cells in quiescent state remain out of the cell cycle while maintain the divide capacity and can reversibly convert to active state in which they re-enter the cell cycle and generate new differentiated cells in response to stimuli. Moreover, phenotypic state transition also occurs under different culture conditions. For example, mouse embryonic stem cells were refined from a prime state to a more plastic naïve state by adding two inhibitors (MEK and GSK3β inhibitor) in the culture medium (Ying et al., 2008). Compared to cell fate conversion, the characterization of cell phenotypic state conversion is more challenge due to the scarcity of efficient approaches to isolate and purify these cells. Fortunately, the development of single cell RNA-seq (scRNA-seq) technology opens up a new strategy to track the changes between different phenotypic state within one cell type in the transition process. Because scRNA-seq captures the heterogeneity of single cell in high-resolution views, eliminating confounding effects in traditional cell isolation.

The interest of deriving these phenotypic conversions is multifaceted. For disease case, the cellular transition from the healthy state to the disease state triggers the dysfunction of cells. In contrast, reverting the disease state to the healthy counterpart can be used as an indication for disease treatment. In this sense, cellular phenotypic transitions are indicators to evaluate the effect of perturbation experiments on the cellular states. For example, it is very essential to evaluate the drug candidates in terms of efficacy and toxicity in the early stage of drug discovery. Therefore, comparing the initial and the resultant cellular state to examine the influences of chemical compounds significantly eases the burden of high-cost and low successful rate during drug development (Michelini et al., 2010). On the other hand, the mechanism of drug action, as well as the general mechanism of cell biological processes can

also be learned by inducing cell phenotypic transitions. Therefore, the discovery of new drugs or druggable targets for a certain disease can be redefined as the identification of factors that can trigger the desired cell transitions.

Taken together, the derivation of desired cellular transitions is of clinical importance, as it allows to derive desired cells and tissues for cell replacement therapies, or to revert disease phenotypes to their healthy counterparts.

1.2 Regulators for cellular transitions

Cellular transition is a complex and dynamic process modulating by multiple layers of regulators, including transcriptional factors, epigenetic modifications, metabolic processes and signalling pathways. In this section, we will discuss these regulators and introduce the reasons that we focus on signalling pathways for cellular transitions.

1.2.1 Transcription factors play key roles in cellular transitions

Conversion of cellular identity or phenotype has been shown to be induced by perturbation of a handful of key TFs. TFs cooperate to establish a transcriptional network close to the target cell type/state so that induce the desired cellular transition. The landmark study is the reprogramming of somatic cells into iPSCs by reintroducing fours pluripotency TFs (Takahashi and Yamanaka, 2006). In agreement, subsequent studies also use the same strategy to convert somatic cells into another lineage cells from the same or another embryonic germ layer. For example, fibroblasts originating from mesoderm have been converted into cardia cells in the same layer, neural cells from ectoderm, hepatic cells from endoderm by forced expression of lineage-specific TFs and pioneer factors (Figure 1.3). The discovery of key TFs normally relies on experimental testing of large sets of potential TFs exhaustively.

Recently, a number of computational methods have been developed to identify the TFs inducing cellular conversions. Some of these methods rely on GRNs by considering different topological characteristics to identify the master regulators of the network. For example, CellNet compares the expression of genes in cell- or tissue-specific GRN between query and reference datasets, in order to infer the TFs that can bring the query arrive at GRN by topological charaterization (Cahan et al., 2014). Another method, Mogrify, predicts TFs that are not only differentially expressed between initial and target cell types but also regulate other differentially expressed genes in the cell type-specific network (Rackham et al., 2016). In addition, some other methods identify TFs that are uniquely expressed in the target cell type

without considering GRNs (D'Alessio et al., 2015). More recently, Okawa et al. propose another method, TransSyn, which enables the identification of TFs that determine the subpopulation identity by considering transcriptional synergy (Okawa et al., 2018).

Despite the development of TF-based cellular transitions, most protocols are not appropriate for clinical translation due to safety concerns. The desired TFs are delivered into cells relying on the use of viral vectors and then the viral vectors are integrated into the genome of host cells. This raises high risk of tumour formation, insertional mutagenesis and other unexpected off-target effects (Ben-David and Benvenisty, 2011; Yamanaka, 2020). Some non-invasive delivery strategies have been developed recently, such as Sendai virus (Miyamoto et al., 2018), single guide RNA (Chakraborty et al., 2014; Liu et al., 2018). However, the clinical application of these strategies is hindered by low efficiency (Haridhasapavalan et al., 2019). Alternatively, chemical compounds, including small molecules, cytokines and growth factors, have been explored and successfully induce cellular transitions. Details will be introduced in followed section.

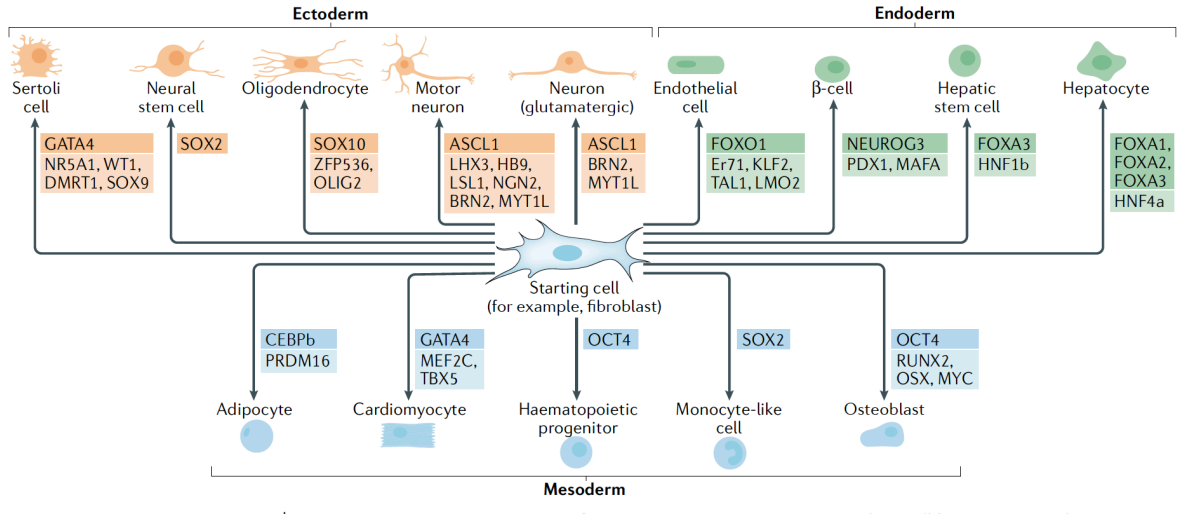

Figure 1. 3 Examples of cellular transitions induced by transcription factors.

TFs can convert somatic cells into another lineage cells from the same or another embryonic germ layer. Figure taken from (Wang et al., 2021).

1.2.2 Chemical compounds for cellular transitions

Chemical compounds are often used to change the cell environment, so as to understand the cellular response to external stimuli, reproduce pathological conditions in vitro, or treat disease by reverting pathological phenotypes. Recently, using chemical compounds for cellular

conversion have made remarkable progress (Federation et al., 2014; Li and Ding, 2010; Xu et al., 2008), since chemical -based cellular transition is nonimmunogenic, easily synthesized, and cost-effective (Federation et al., 2014; Hu et al., 2015). More importantly, they avoid the manipulation of genetic materials and are more convenient and can be easily controlled by changing the combinations and concentrations. Therefore, they are not only used to generate desired cell types in vitro, but also used as drugs to induce endogenous regenerations in patients or revert disease phenotypes (Figure 1.2).

Chemical compounds induce cellular transitions mainly by acting on metabolic processes, epigenetic modifications and signalling pathways, and in turn modifying the transcriptional programmes (Figure 1.4). Specifically, metabolic modulation is important for cellular transition, as the conversion between cell fates/states along with metabolic switches due to the differences in metabolism between initial and desired cells. For example, chemical compounds inducing glycolysis and autophagic metabolisms are found to promote the course of reprogramming (Chen et al., 2011; Zhu et al., 2010). In addition, epigenetic modifications, including DNA methylation and histone modifications, such as methylation, acetylation and phosphorylation, contribute the low efficiency of cellular transitions. Therefore, chemical compounds overcoming these barriers can also improve the cellular transitions (Chen et al., 2013; Esteban et al., 2010; Shi et al., 2008).

More importantly, in the process of cellular transition, many signalling pathways are involved and play a major role. For example, the maintenance of the pluripotency and self-renewal in pluripotent stem cells is coordinated by many signalling pathways and therefore chemical compounds targeting these signalling pathways derive the formation of iPSCs (Qin et al., 2017). Moreover, to maintain homeostasis of tissues, multiple signalling pathways modulate tissue-specific stem cell transitions between quiescent and active states (Biteau et al., 2011). During development, the generation of tissues and organs is also dedicated by the combination of signalling pathways over time (Basson, 2012). Under pathological conditions, the dysregulation of signalling pathways can trigger immune response, inflammation and other disease-related phenotypes. Due to the variety of signalling pathways, a wide range of chemical compounds targeting distinct signalling pathways have been identified to promote different kinds of cellular conversions. Chemical compound CHIR99021 acting on Wnt signalling is involved in different cellular conversion cocktails (Hou et al., 2013; Hu et al., 2015; Li et al., 2011). Since epithelial-mesenchymal transition (EMT) is a roadblock during reprogramming and TGF-β signalling pathway induces (EMT), the inhibitors of TGF-β, such as SB431542,

Repsox and A83-01, have been widely used in various contexts (Boyer et al., 2006; Christman, 2002; Mikkelsen et al., 2008). More other chemical compounds targeting signalling pathways like FGF, Hedgehog, JAK-STAT, MAPK/ERK and so on, can be found in reviews (De et al., 2017; Federation et al., 2014; Qin et al., 2017). Chemical compounds targeting signalling pathways are also widely used for disease treatment. For example, recent studies show that AMPK, SIRT1and mTOR signalling pathways are crucial for aging and drugs acting on them lead to new interventions for age-related diseases (Yu et al., 2021). Therefore, identification of optimal chemical compounds acting upon the desired signalling pathway to induce cellular transition gains increasing attention in the field of regenerative medicine and disease treatment.

To date, the traditional methods employed to identifying the optimal chemical compounds that control desired cellular transitions mainly rely on cell-based phenotypic or pathway-based screenings of chemical libraries (Lyssiotis et al., 2011). Cell-based phenotypic screenings are high-throughput and focus on identifying compounds that induce desired phenotype by performing different assays and without considering the prior knowledge of the cellular transitions. Pathway-based screenings are undertaken within a small group of compounds with known mechanism of action and determine how they regulate specific pathways involved in the cellular transitions. However, the elucidation of the molecular mechanism of compounds is usually very complicated as they may have multiple, non-relevant targets. Moreover, the experimental identification of chemical is costly and time consuming. In this regard, a systematically guidance from computational perspective to identify chemical compounds or signalling targets to induce cellular transitions is desirable. Therefore, we proposed computational methods in this thesis to identify signalling perturbations to induce the cellular transitions.

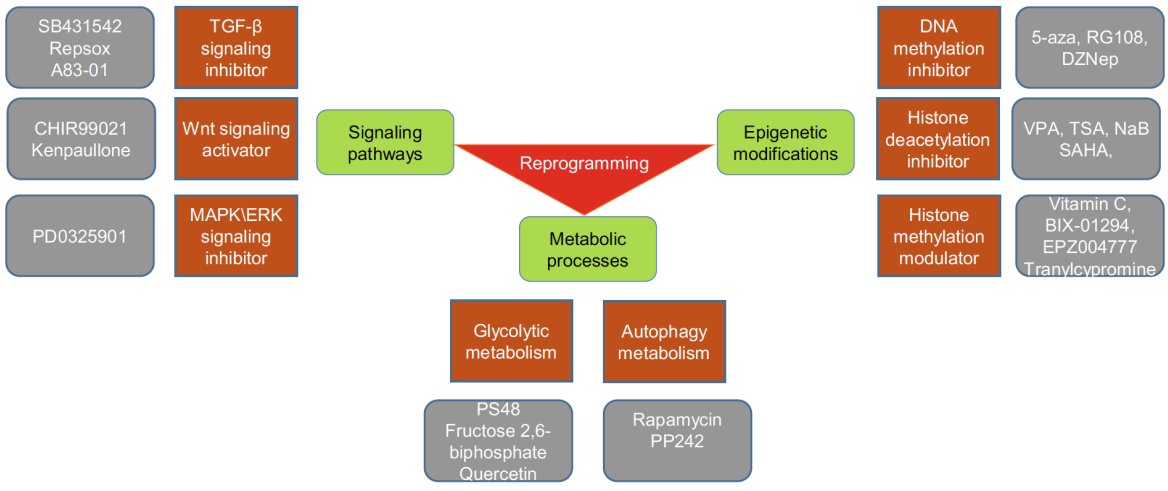

Figure 1. 4 Three major mechanisms of chemical compounds for cellular transitions.

Chemical compounds induce cellular transitions mainly by modulating metabolism, regulating epigenetic barriers and signalling pathways. Figure taken from (Qin et al., 2017).

1.3 Existing methods for the identification of signalling perturbation

Given the importance of signalling pathways in different biological processes, multiple computational methods have been developed to predict signalling perturbations that are responsible for the observed gene expression dysregulations, including signalling proteins, signalling pathways and chemical compounds. The algorithms of these methods can be broadly grouped into two classes, compendium-based inference and de novo prediction.

The compendium-based algorithms rely on pre-compiled gene expression profiles to infer the signalling perturbations whose gene expression signatures are similar with query samples. In 2006, Lamb et al. propose the first general compendium-based approach that connects chemical perturbations with cellular transcriptional responses (Lamb et al., 2006). In 2017, Subramanian et al. present a second generation of perturbation signature compendium (including chemical and genetic perturbations), L1000, with more than 1000-fold scale up of the CMap across different cell lines (Subramanian et al., 2017). Later, this compendium is widely used as a resource for the development of different computational methods (Musa et al., 2017). SPEED is another compendium-based method that constructs a database including gene expression datasets with single pathway perturbation and identified signalling pathways that cause similar regulatory patterns as desired. In a similar manner, PROGENy also infers pathway activity from gene signatures in a wide range of perturbations focusing on cancer cell lines.

Another group of methods infer signalling perturbations using different algorithms without considering the prior knowledge. Some of de novo algorithms try to identify the dysregulated pathways by using the enrichment measures on the dysregulated genes between different samples directly (Sartor et al., 2009; Subramanian et al., 2005). On the other hand, some other methods predict the pathways not only consider the differentially expressed genes but also take the interaction and regulatory information embedded in the pathway topology into account (Dutta et al., 2012; Massa et al., 2010; Tarca et al., 2009). These methods have been widely used to understand various diseases, especially cancers, as well as investigate the roles of signalling pathways in the course of development. Moreover, another group of de nove

predictions attempt to infer the causal upstream signaling perturbations resulting in the observed dysregulated genes by connecting signalling network with downstream de-regulated genes (Fakhry et al., 2016; Krämer et al., 2014). More recently, the development of scRNA-seq technique allows researchers to dissect the cell-cell interactions through the intercellular signal transductions. Browaeys et al. propose a scRNA-seq based computational tool for identifying extracellular ligands that could induce observed gene expression changes in both healthy and disease states (Browaeys et al., 2020). Jung et al. also develop an approach to predict signalling proteins which regulate the dysregulated inflammatory response by modelling intercellular communications in a hyperinflammatory condition (Jung et al., 2021).

1.3.1 Limitations of existing methods

In compendium-based methods, the databases mainly involve perturbation datasets for cancer cells. As we known, the signalling pathways and transcriptional logics in cancer cells are significantly different from those of non-cancer cells and hence the databases are not suitable for identifying chemical compounds for non-cancer cells. While the compendium-based methods have an advantage in making accurate predictions for chemical compounds present in the databases, they are limited to de novo prediction. In addition, current compendium-based methods only compare the similarity of gene signature without considering the initial cellular state. However, the downstream response and the affinities between chemical compounds and protein targets are also determined by the initial cellular states besides perturbations. Therefore, it is also essential to consider the initial cellular state while identifying signalling perturbations.

Signal transduction involves many different kinds of interactions between proteins, such as phosphorylation and other post-translational modifications. Due to the scarcity of direct measurements of signalling activity, the methods of de novo predictions use gene expression as a proxy for protein activity. However, it is unclear that signalling activity can be reflected without considering the presence of post-translational modifications. Moreover, due to the extensive crosstalk in signalling pathways, circumventing the activation/inhibition of undesired genes is also essential. Even though some de novo prediction methods try to predict the singalling perturbations inducing the dysregulation of downstream genes, they do not attempt to ensure the predictions specifically acting on desired genes, while minimizing the off-target effects.

In addition, the existing computational methods mainly aim at identifying signalling events either inducing diseases, especially cancer, or elucidating the mechanism of action for

drugs. No computational method is developed to identify signalling perturbations to induce cell reprogramming or transdifferentation.

Therefore, it is desirable to develop more comprehensive computational methods that address the limitations of current methods to predict signalling perturbations.

2 Aims and scope of thesis

The generation of desired cellular transition is of clinical interest, including drug discovery, disease modelling and regenerative medicine. Studies have shown that cellular transitions can be induced by ectopic expression of a small set of TFs. However, TF-based cellular transitions have raised safety concerns and therefore their translation into clinical applications is limited. Alternatively, the usage of chemical compounds to induce signalling perturbations not only addresses these concerns, but also offers an easily controlled and cost-effective strategy for cell transitions. The conventional approaches to select optimal chemical compounds rely on exhaustive screening of a large scale of compounds or require the prior knowledge about molecular mechanism of compounds, which are either inefficient and resource intensive, or not always available. Therefore, the development of computational methods to identify chemical compounds or signalling targets are required. Existing computational methods for signalling pathways have their own advantages and limitations in different aspects (Section 1.3). More comprehensive computational methods that addresses the limitations of existing methods, particularly the methods able to predict signalling determinants for cellular conversions, are desirable. In this thesis, three different integrative computational methods are proposed depending on the aim of corresponding study.

2.1 Aims of the thesis

Aim 1. Develop a scRNA-seq based method to predict chemical compounds and corresponding signalling protein targets acting on specific sets of TFs to induce cellular conversions, including conversions between cell types, functional cell subtypes or distinct phenotypic states within a same cell type. This involves the construction of a database by manually collecting and compiling a large scale of perturbation datasets consisting solely of non-cancer cells. Since the aim of this method is to predict chemical compounds specifically targeting the desired set of TFs, the duration of perturbations in this database requires within six hours to ensure that the desired TFs are the initial response of perturbations and effected by the perturbation specifically. This database is integrated into a network-model using scRNA-seq data to ensure the predictions are specific to the initial cellular state. The method will be validated by a set of benchmarking datasets and existing chemical-based protocols of cellular conversion. Moreover, the method will be applied to develop a protocol for the maturation of hepatocytes from hepatic progenitors, aiming at generating human induced hepatocytes that

resemble the functionality of primary hepatocytes and solve the problem of abnormal lipid metabolism in previous protocol.

Aim 2. Develop an integrative computational method to predict chemical compounds and corresponding signalling proteins to induce cellular transition from the given initial state to the desired target state. The method will be applied to revert the non-cancer disease phenotypes to their healthy counterparts. This also requires the development of a database focusing on non-cancer cells. However, since the aim of this method is to predict chemical compounds to induce the transition of two states, the database will be extended to any two cellular states before and after perturbation, rather than within six hours. The performance of this method will be compared with the results of using cancer perturbation databases as well as other existing methods. Finally, the method will be applied to different kinds of diseases, including age-related diseases and infectious disease, verifying its general applicability of identifying drugs for disease treatment.

Aim 3. Develop an integrative method that integrates signalling network and GRN to model the effect of signalling pathways and proteins on GRNs. The main application of this method is to identify signalling pathways and proteins that can induce salamander limb regeneration. Due to the lack of prior knowledge of gene regulatory network for axolotls, raw GRNs for salamander limb regeneration are required to developed by using time serious microarrary data. These GRNs will be combined with canonical signalling pathways to identify signalling cues involved in cellular transitions at different stages of salamander limb regeneration.

2.2 Originality

The proposed computational methods in this thesis integrate multiple resources, aiming at identifying signalling perturbations to induce cellular transitions for regenerative medicine and disease treatment. Previous approaches either use perturbation databases focusing on cancer cells or infer signalling transduction by using gene expression directly, limiting their suitability for the study of cellular transition for non-cancer cells and the predictive power. In this thesis, an important advancement is the development of manually curated, large-scale perturbation databases consisting solely of non-cancer cells, which provide valuable resources for the study of non-cancer signalling perturbation. The present methods first infer potential signalling proteins from the perturbation databases to narrow down the candidates, and then further predict the initial state-specific candidates by integrating the network model using gene expression of the initial cellular state. This framework can be general and flexible to apply to

reverse disease phenotypes and induce cellular conversion, such as reprogramming, differentiation and transdifferentiation, in human, mouse and rat. To study the organism whose prior knowledge of perturbation is not widely available, another de novo prediction method is proposed in this thesis. This method integrates signalling and transcriptional regulatory networks to predict signalling pathways that induce the regeneration of salamander limb. In this method, we propose a model simulating the perturbations of signalling pathways that can induce the transitions between GRN states corresponding the initial and desired cellular states, which is conceptually different from the existing methods.

3 Materials and methods

The details of materials and methods are presented in the Results section together with each manuscript. Here, the materials and methods for each manuscript is briefly summarized.

In "*A single cell-based computational platform for cell engineering using chemical compounds*", a computational method, SiPer, relying on scRNA-seq, is designed to identify chemical compounds and their corresponding signalling protein targets that can activate/inhibit specific sets of query TFs to induce cellular conversion. In this study, a non-cancer cell perturbation database with duration not larger than six hours and signalling network are constructed. The method first pre-selects a list of signalling protein candidates from the built-in database. Subsequently, the candidates that are specific to the initial cellular state are further identified by integration of signalling network and scRNA-seq data using an algorithm that combines signalling entropy and probabilistic model. Finally, the chemical compounds targeting the predicted signalling proteins are identified. This method requires the set of desired TFs and scRNA-seq of the initial cellular state given by users as input.

In "*A database-driven computational method to identify chemical compounds reverting disease phenotype*", a computational method, ChemPert, is developed to predict chemical compounds to induce the cellular transition from the initial state to the target state. ChemPert is applied to revert the disease phenotypes to their healthy counterparts. Similar as SiPer, ChemPert also integrates a normal cell perturbation database and a network model, whereas the datasets in the database of ChemPert is much larger than SiPer. In addition, ChemPert uses enrichment analysis in the network model due to the usage of bulk RNA-seq. The algorithm of ChemPert is also composed of three major steps, 1) pre-selection of candidate signalling proteins from the perturbation database of transcriptional signatures, 2) network-based modelling to predict signalling proteins that are specific to the initial cellular state, 3) identification of chemical compounds which target the predicted signalling proteins.

In "*An integrative network model to predict signalling pathways for salamander limb regeneration*", a time series microarray dataset specific to connective tissue cells of salamander limb is generated. This dataset is used to predict signalling pathways whose perturbation can induce the shift of GRN state and in turn induce the regeneration of salamander limb based on our integrative network method. Specifically, the method integrates two distinct models for the signalling and transcriptional regulatory layers. These two layers are connected by the interface TFs. The signal transduction and its effect on the interface TFs are mimicked by using a

probabilistic model. The key interface TFs that determine the state of the GRN is identified by simulating the perturbations in silico on a Boolean network model.

4 Results

4.1 Manuscript 1: A single cell-based computational platform for cell engineering using chemical compounds

4.1.1 Preface

Conversion of cellular identity or phenotype has been shown to be induced by perturbation of a handful of key TFs. Replacement of the direct manipulation of TFs with chemical compounds acting on signalling pathways offers a more controlled and safer way to accomplish such conversion without the transfer of genetic material, which holds great promise for regenerative medicine. Nevertheless, prioritizing chemical compounds that specifically target relevant TFs remains a challenge. Here, we develop a scRNA-seq based computational method that systematically predicts chemical compounds specifically targeting desired sets of TFs to induce cellular conversions. This integrates a compendium of 5591 transcriptomics data of normal cell perturbations with a network model. SiPer recapitulated the experimentally validated chemical compounds in diverse cellular conversion experiments. Moreover, by applying SiPer, we successfully developed a novel and efficient protocol to drive the conversion of hepatic progenitors into functional human induced hepatocytes.

In this work, I developed the computational method and performed computational analysis and prediction. The experiment about the generation of hepatocytes was carried out by Dr. Bingqing Xie.

4.1.2 Manuscript

A single cell-based computational platform for cell engineering using chemical compounds

Menglin Zheng[1,6], Bingqing Xie[2,6], Satoshi Okawa[1,3], Soon Yi Liew[2], Hongkui Deng[2,*], Antonio del Sol[1,4,5,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Esch-sur-Alzette, L-4367 Belvaux, Luxembourg;

[2] School of Basic Medical Sciences, State Key Laboratory of Natural and Biomimetic Drugs, Peking University Health Science Center and the MOE Key Laboratory of Cell Proliferation and Differentiation, College of Life Sciences, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100191, China;

[3] Integrated BioBank of Luxembourg, Dudelange L-3555, Luxembourg;

[4] CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 801 Building, 48160 Derio, Spain;

[5] IKERBASQUE, Basque Foundation for Science, Bilbao 48013, Spain;

[6] These authors contribute equally: Menglin Zheng, Bingqing Xie;

[*] Correspondence: Antonio del Sol (Antonio.delSol@uni.lu) or Hongkui Deng (hongkui_deng@pku.edu.cn)

Abstract

Chemical-based engineering of functional cell populations by single-cell technologies holds great promise for safe and precise strategies for cell therapies. Nevertheless, development of chemical-based cell conversion protocols currently requires large-scale screening of chemical compounds, which is time- and labour intensive. In order to address this bottleneck, we develop a single-cell RNA-sequencing based platform (SiPer) to systematically identify chemical compounds specifically targeting desired sets of transcription factors to guide the engineering of specific cell populations. This platform integrates the first large compendium of chemical perturbations on non-cancer cells with a gene network model. We show that SiPer is generally applicable to diverse cell engineering examples, recapitulating known chemical compounds and their corresponding protein targets. Importantly, using chemical compounds predicted by SiPer, we develop a highly efficient protocol for the generation of functional human induced hepatocytes (hiHeps). These results demonstrate that SiPer provides a valuable resource to efficiently identify chemical compounds for single-cell technology-based cell engineering.

Introduction

The generation of functionally specific cells by using cellular conversion protocols is of clinical interest, providing a valuable resource for cell transplantation and in-vivo cellular conversion as therapeutic strategies. However, several challenges need to be addressed for achieving optimal cell engineering, such as the accurate characterization of cell populations (including cell types, cell subtypes and phenotypic states) and the identification of cell conversion factors. Single-cell RNA-sequencing (scRNA-seq) technologies allowing for the discrimination of different cell populations have made it possible to overcome these roadblocks to achieve more precise cell engineering. Due to the significant amount of scRNA-seq data produced in the last decade, experimental researchers are increasingly interested in engineering cell populations identified by scRNA-seq(Francesconi et al., 2019; Liu et al., 2017). It has been observed that a small set of specific transcription factors (TFs) can be sufficient to induce cellular conversion(Morris and Daley, 2013). In order to facilitate a wide range of cell engineering experiments, a handful of computational frameworks have been developed to identify cellular conversion TFs based on bulk transcriptomics data(Cahan et al., 2014; D'Alessio et al., 2015; Rackham et al., 2016) and scRNA-seq data (Okawa et al., 2018). However, TF-based protocols that involve transfer of genetic material, have raised safety concerns, such as unwanted long-term expression of the delivered TFs and the possibility for insertional mutagenesis, which limits their translation into clinical applications(Cieślar-Pobuda et al., 2017). Alternatively, replacing TFs with non-integrative means(Rivetti di Val Cervo et al., 2021; Shi et al., 2017), for example, chemical compounds to drive signalling perturbations, not only addresses these concerns, but also offers an easily controlled and cost-effective strategy for cell engineering(Hou et al., 2013; Xu et al., 2015). However, chemical-induced cellular conversion protocols currently rely on exhaustive trial-and-error testing of a large scale of compounds, which is both inefficient and resource intensive. A systematic guidance is needed for facilitating compound-based cell engineering, especially the generation of functional cell subtypes or distinct phenotypic states of a same cell type, which is crucial for successful clinical application of engineered cells.

Here, we present SiPer (SIgnalling proteins and chemical PERturbagens), a single-cell RNA-sequencing (scRNA-seq) based computational platform that identifies signalling proteins and chemical perturbagens (including small molecules, drugs and cytokines) for targeting specific sets of desired TFs to induce conversion of cell populations. Importantly, optimal perturbagens should be specific to initial cellular state and specifically target only desired sets of TFs to minimize the off-target effects(Federation et al., 2014; Strasen et al., 2018). Therefore,

SiPer integrates an initial cell-specific network model with a manually collected compendium of 5591 experimentally generated transcriptional signatures of normal cells within six hours before and after chemical perturbations. The collected compendium constitutes an important resource to infer the effect of chemical perturbations on downstream TFs, incorporating prior knowledge of distinct perturbations in a variety of cell types. Notably, this compendium consists solely of normal cell type perturbations, as disease cells, such as cancer cells, often exhibit significant differences in the signalling mechanisms compared to normal cells(Sharma and Petsalaki, 2019). As a result, SiPer is able to identify optimal perturbagens targeting user-defined sets of TFs for cell engineering. Alternatively, SiPer can compute differentially expressed TFs (DETFs) between the initial and target cell populations and use them as query TFs for perturbagen prediction.

We evaluated the predictive power of SiPer in 200 chemical perturbation datasets and showed that a considerable number of experimentally validated chemical compounds and their protein targets were accurately predicted. Furthermore, SiPer was applied to known chemical-induced cell conversion examples, correctly identifying perturbagens and their corresponding protein targets. For example, SiPer was applied to phenotypic conversion of human embryonic stem cells (hESCs) from a lineage primed state to a more plastic naïve state. Besides two classical inhibitors (GSK3β and MEK inhibitor), additional specific chemical compounds, such as adenylyl cyclase activators, LCK/SRC inhibitors to maintain hESCs in a stable naïve state were also recapitulated by SiPer(Theunissen et al., 2014). The key factors, such as ROCK inhibitors and PDGFR inhibitors, were predicted by SiPer for the oligodendrocyte differentiation(Marques et al., 2016; Pedraza et al., 2014). Moreover, SiPer identified signalling molecules of PPARγ agonist for the conversion of brown adipocytes from human dermal fibroblasts (HDFs), which plays a critical role in adipogenesis for both white and brown adipocytes(Takeda et al., 2017). The glucocorticoid receptor agonist, phosphodiesterase inhibitor, and cyclooxygenase inhibitor were also predicted by SiPer to direct the differentiation of adipocytes from mesenchymal stem cells (MSCs)(Pittenger et al., 1999). SiPer is freely available as a web application at https://siper.uni.lu and can easily be applied to any cell system.

Finally, we applied SiPer to drive the differentiation of hepatic progenitors into hiHeps. A previously well-characterized protocol comprised of a two-chemical cocktail based on the unspecified hepatocyte culture medium (H2C) (Xie et al., 2019). However, long-term cultured hiHeps in this protocol gradually presented excessive lipid accumulation, hindering its further development. To identify a new, fully defined formulation, SiPer was applied and

experimentally tested. Validation of SiPer predictions in a defined Williams' E medium quickly and effectively resulted in a final chemical cocktail that was able to generate hiHeps that resembled primary hepatocytes (PHs), while successfully circumventing the excessive lipid accumulation in hiHeps as seen with H2C, allowing their potential use in future cell transplantation.

These results demonstrate that SiPer is a valuable resource for funnelling effort towards the establishment of high-quality chemical cellular conversion protocols, enabling the design of new experimental strategies for gene therapies and scRNA-seq-based cell engineering for disease modelling and cell transplantation.

Results

Development and evaluation of SiPer

Perturbing a relatively small set of specific TFs is usually sufficient to induce desired changes in cellular identity/phenotype(Morris and Daley, 2013). Herein, SiPer is designed to identify chemical perturbagens and their corresponding protein targets in the intracellular signalling network (signalling proteins) that can activate/inhibit any set of query TFs for engineering cell populations (Figure 4.1A). Query TFs can be known cell conversion TF cocktails provided by users. If query TFs are unknown, SiPer performs differential expression analysis between scRNA-seq data of initial and target cell populations and uses the DETFs as input (Figure 4.1A).

Normal cell perturbation compendium (NCPC): The inference of exact signalling paths acting on specific TFs will require not only the gene expression (protein abundance) information, but also protein activity measures such as phosphorylation levels. However, phosphorylation data before and after cell perturbation is hardly available for a wide range of perturbagens. Therefore, in order to infer the relationship between signalling perturbations and TFs, we exhaustively collected 5591 transcriptional profiles in response to chemical perturbations within six hours across 134 different types of normal cells and tissues from three species, human, mouse and rat (Table S1). To ensure specificity between query TFs and signalling perturbations and minimize the off-target effects, only DETFs of datasets within six hours after perturbation were considered as initial transcriptional response as suggested by Lamb et al (2006) (Lamb et al., 2006). These profiles were derived from 2386 unique signalling perturbagens, which covered 2072 TFs in both activation (up) and inhibition (down) directions with no significant bias towards either of them (Figure 4.2A). Notably, unlike the previous databases of perturbation-based transcriptional signatures(Lamb et al., 2006; Schubert et al., 2018; Subramanian et al., 2017), NCPC solely consists of transcriptomics data of normal

cells/tissues in order to exclude cancer cells, whose signalling pathways and transcriptional regulatory networks are known to exhibit significant rewiring from the normal counterparts(Sharma and Petsalaki, 2019). NCPC contains two parts, Perturb-reTFs, a compendium of response transcriptional signatures (i.e., DETFs) to each signalling perturbagen and Perturb-targets, a compendium of the protein targets of each signalling perturbagen (Figure 4.1B). Thus, instead of attempting to infer exact signalling paths that could reach a query set of TFs, SiPer finds signalling proteins in NCPC whose perturbation has been shown to result in the up/down-regulation of similar sets of TFs (See Methods and Figure 4.1C).

Gene network modelling: The affinities of signalling proteins to chemical perturbations are often highly specific to the initial cellular state(Hodos et al., 2018; Strasen et al., 2018). In this regard, SiPer leverages scRNA-seq technologies that allow for characterization of heterogeneity of cell populations, which can aid in precise cell engineering. To prioritize signalling proteins that are specific to the initial cellular state, SiPer combines scRNA-seq data of the initial cell population and the prior knowledge network (PKN) (Methods and Figure 4.1C), which consists of two layers, the upstream signalling networks and downstream TF-TF interactions. These two layers of networks were combined by interface TFs, which mediate the signal transduction from cytoplasm to the nucleus. This resulted in complete PKNs including 6633 genes and 96137 interactions for human, and 6825 genes and 109789 interactions for mouse and rat (Table S2). Then, SiPer divides the final set of candidate signalling proteins into different functional groups based on Reactome signalling pathways. For each group of signalling proteins, SiPer identifies optimal perturbagens based on the similarity between their target proteins and the candidate signalling proteins (see Methods and Figure 4.1C). In addition, SiPer annotates the functional mechanism of perturbagens and automatically separates perturbagens into different groups based on their similarities in target proteins, which allows users to design an optimal combination of perturbagens targeting distinct functional groups.

The performance of SiPer was optimized (Supplementary notes 3, 4) and assessed by the accuracy in identifying correct signalling proteins and perturbagens using 200 chemical perturbation datasets (see Methods and Table S3). Here, DETFs were considered as the transcriptional response to perturbagens and were used as query TFs. A considerable number of experimentally used perturbagens and corresponding protein targets were correctly identified (Figures 4.2B, C, Figures S4.1A, B). Importantly, when NCPC was replaced with a CMap compendium that includes perturbations in cancer cells (see Methods)(Subramanian et al., 2017), the performance of SiPer significantly decreased (Figures 4.2B, C, Figures S4.1A, B). To investigate the reason for this decreased performance, we identified 1956 perturbagens

by which both normal and cancer cell types were stimulated (Table S4). For each of these perturbagens, a hierarchical clustering analysis was performed based on DETFs to evaluate whether the normal/cancer cell types were clustered together. For each perturbagen, the fraction of cells correctly clustered to their respective class (i.e., normal or cancer) was significantly higher than mis-clustered ones (one-sided Wilcoxon test, *P-value <2.22e-16*) (Figures 4.2D). Indeed, the transcriptional responses of normal cells were distinct from cancer cell counterparts (Figures 4.2D, Table S4). This result indicates that it is essential to construct a normal cell-specific perturbation compendium, such as NCPC, for accurate prediction of signalling perturbagens for conversion between normal cell populations. Furthermore, we investigated the robustness of SiPer to different parameters of the algorithm. First, the number of cells in input scRNA-seq data did not significantly affect the accuracy of SiPer (Figures 4.2E). In addition, the overall performance was maintained across datasets with numbers of query TFs less than 100 (Figures 4.2F). The decreased accuracy with a large number of query TFs is presumably because of the lack of specificity of target signalling proteins affecting a large number of downstream TFs. Moreover, we also demonstrated that SiPer was robust to the change in NCPC and in the PKN and did not show significant bias toward/against cell types in NCPC (Figures S4.1E). To evaluate whether prediction of SiPer is specific to query TFs, SiPer was run on the same number of randomly selected TFs. The performance dramatically decreased compared to the original performance, which implicates that the predicted signalling proteins are specific to the query TFs (Figures S4.1E).

SiPer accurately predicts perturbagens for conversion of phenotypic states/cell subtypes

SiPer was first applied to the conversion between different phenotypic states of the same cell type, where the conversion chemical cocktails are experimentally validated (Table 4.1 and Table S5). Predicted signalling proteins exhibited high specificity to the respective query TFs (Figure 4.2G) and allowed for the identification of perturbagens that target these proteins (Table 4.1). For the conversion of lineage primed embryonic stem cells (ESCs) to a more plastic naïve state in both human and mouse, SiPer was able to predict both perturbagens and signalling proteins for this conversion event given the conversion TFs NANOG, KLF2, TFCP2L1, NR5A2 and KLF4 for human (Takashima et al., 2014) and Esrrb, Nanog, Klf2, Klf4 and Tfcp2l1 for mouse reported by previous studies (Dunn et al., 2014; Ivanova et al., 2006; Martello et al., 2012; Niwa et al., 2009; Ye et al., 2013). For instance, the two classical inhibitors (2i) CHIR99021 (GSK3β inhibitor) and PD0325901 (MEK inhibitor) or their

corresponding signalling protein targets(Ying et al., 2008; Zimmerlin et al., 2016) were correctly identified. In addition to these two inhibitors, SiPer predicted the receptor of LIF (LIFR) in both mouse ESCs (mESCs) and hESCs, which is an important cytokine required for sustaining ESCs self-renewal (Niwa et al., 1998). Furthermore, since the 2i is not sufficient to maintain a stable naïve state of hESCs (Zimmerlin et al., 2017), studies have reported additional chemical compounds for the maintenance of the ground state of naïve hESCs, such as forskolin(Park et al., 2018) and WH-4-023 (Theunissen et al., 2014), which were identified by SiPer as well. Moreover, Theunissen et al. (2014) have shown that activin A enhanced the efficiency of naïve hESCs conversion from the primed state (Theunissen et al., 2014). Although SiPer did not predict activin A directly, its target proteins ACVR2A, ACVR2B and BMPR2 were identified. SiPer network visualization of the underlying putative signalling cascades also identified CTNNB1, the central effector of WNT signalling stimulated by CHIR99021(Zimmerlin et al., 2016). In addition, FOXO1 and SMAD2 that are essential for mediating the signal to the query TFs were also predicted (Figure 4.2H), consistent with their key role in maintaining pluripotency (Sakaki-Yumoto et al., 2013; Zhang et al., 2011). The presence of adipocyte differentiation genes, such as FOXO1 and STAT1, as key signalling effectors can also partially explain why the predicted signalling proteins showed relative high specificity to differentiation of MSCs into adipocytes (Figure 4.2G). Hematopoietic stem cells (HSCs) are functionally heterogeneous and we applied SiPer to identify key signalling proteins and perturbagens that return mouse active HSCs to a quiescent state to avoid HSC exhaustion with scRNA-seq data discriminating these two states (Kowalczyk et al., 2015). SiPer predicted a MEK inhibitor U0126 that has been reported to revert active HSCs to quiescence (Baumgartner et al., 2018). In addition, high expression of retinoic acid and low CDK4/6 levels have been shown to be crucial for the maintenance of HSC quiescence (Cabezas-Wallscheid et al., 2017; Fukushima et al., 2019; Matsumoto et al., 2011). SiPer identified retinoic acid receptor (RAR) agonists and CDK inhibitors and corresponding targets Cdk4 and Cdk6. SiPer was also applied to the differentiation of hair follicle stem cells (HFSCs) from telogen to anagen transition (Yang et al., 2017). SiPer identified inhibition of TGFβ1 and activation of TGFβ2 as well as their targets Tgfbr1 and Tgfbr2 with unknown effects, consistent with the previous finding that TGFβs plays distinct roles during hair follicle development (Foitzik et al., 1999; Oshimori and Fuchs, 2012). Apart from the conversion between phenotypic states, we also applied SiPer to the conversion of cell subtypes. Okawa et al. (2018) identified a set of TFs that can convert human hindbrain neuroepithelial cells (hNES) to medial floor plate midbrain progenitors (hProgFPM), allowing for differentiating into dopaminergic neurons

rapidly(Okawa et al., 2018). We used these TFs as input for SiPer and predicted an HMG-CoA reductase inhibitor, GSK3 inhibitors as well as a Wnt/β-catenin inhibitor. These compounds have been shown as important factors for midbrain dopamine neuron differentiation (Castelo-Branco et al., 2004; Chung et al., 2009; Rhim et al., 2015). In another example, using scRNA-seq data of oligodendrocyte precursor cells (OPCs) and Myelin-forming oligodendrocytes (MFOLs) within oligodendrocytes (Marques et al., 2016), SiPer identified ROCK inhibitors that can promote the myelin formation from OPCs (Pedraza et al., 2014), and PDGFR inhibitors, whose target Pdgfra has been known as a marker of OPCs (Marques et al., 2016).

SiPer accurately predicts known perturbagens involved in cell type conversion

We further investigated the predictive power of SiPer in broad cell type conversions, including cell reprogramming and differentiation, for which experimentally validated chemical conversion cocktails are reported (Table 4.2 and Table S5). Specifically, given the conversion TFs LHX6, DLX5, FOXG1, ASCL1 and SOX2 for GABAergic neurons from human fibroblast reported by Colasante et al (2015) (Colasante et al., 2015), multiple essential chemical compounds and corresponding protein targets reported in previous study were captured by SiPer, including CHIR99021, Pifithrin-α, LDN193189 and forskolin (Dai et al., 2015). Moreover, a NOTCH-independent role of RBPJ in the neuronal specialization into GABAergic neurons has been reported, which was recapitulated by SiPer network visualization where RBPJ acts as a key regulator of the target TFs in the absence of NOTCH (Figure S4.1D) (Hori et al., 2008; Komine et al., 2011). Zhang et al. (2016) developed a chemical protocol to drive the generation of neural stem cell-like cells from mouse fibroblasts (Zhang et al., 2016a). Using Pou3f4, Sox2, Klf4, c-Myc, and Tcf3 as input TFs for SiPer (Han et al., 2012), we identified GSK3β inhibitors, RAR agonists and autography modulators as well as the receptor of basic fibroblast growth factor (Fgfr4), which were signalling pathways targeted by chemical compounds used by Zhang et al. (2016). Takeda et al. (2017) demonstrated that the chemical cocktails which contained a BMP inhibitor Dorsomorphin, an adenylyl cyclase activator forskolin and a PPARγ agonist Rosiglitazone could direct brown adipocytes conversion from HDFs (Takeda et al., 2017). Consistently, the protein targets of Dorsomorphin (FKBP1A and KDR), forskolin (ADRB2) and rosiglitazone (PPARG, RXRB, ADRA1A and ADRA1B) were all predicted by SiPer. For the induction of cardiomyocytes from mouse embryonic fibroblasts (MEFs), histone deacetylase inhibitors, MAPK/ERK inhibitors and RAR agonists, as well as the signalling protein targets of CHIR99021(Gsk3a and Gsk3b) and BayK 8644 (Htr2a, Adora2a and Adora2b) were predicted by SiPer (Fu et al., 2015; Park et al., 2015). Next, we

applied SiPer to identify perturbagens and key signalling proteins in the context of cellular differentiation (Table 4.2 and Table S5). Sun et al. (2007) and Gao et al. (2008) reported that nicotinamide in high glucose with or without retinoic acid can induce human MSC differentiation into pancreatic progenitors (Gao et al., 2008; Sun et al., 2007). SiPer identified the protein target PPARD for nicotinamide and chemical compound AHPN, which also acts as an RAR agonist. The chemical compound cocktails composed of 3-isobutyl-1-methylxanthine, indomethacin, dexamethasone and insulin have been shown to induce the adipogenic differentiation of human MSCs (Pittenger et al., 1999). The signalling protein targets of these four chemical compounds were predicted by SiPer, including ADORA1 and CFTR for 3-isobutyl-1-methylxanthine, NR3C1 and NR3C2 for indomethacin, IGF1R for insulin and PTGS1 for dexamethasone. In agreement with this, a glucocorticoid receptor agonist dexamethasone and a COX inhibitor DUP-697 were identified as perturbagens. Besides the predicted proteins, SiPer network visualization further identified several pro-adipogenic genes, such as CTNNB1 (Chen et al., 2020), RAC1 (Kunitomi et al., 2020), PPP1CB (Cho et al., 2015) and CREB1 (Zhang et al., 2004) (Figure S4.1E), indicating that they might regulate the differentiation TFs in a coordinated fashion.

In summary, using the set of experimentally validated conversion TFs or DETFs as input, SiPer was able to identify a considerable amount of experimentally validated perturbagens and corresponding target signalling proteins in the wide range of cell engineering cases (Table S5). This result demonstrates that the current challenge of replacing TFs with chemical compounds in cellular conversion protocols, especially conversions between phenotypic states and cell subtypes, could be systematically addressed by SiPer, potentially opening up new therapeutic strategies.

Experimental validation of predicted perturbagens for hepatic maturation

Having evaluated the predictive abilities of SiPer, we applied SiPer to cellular conversion from hepatic progenitors to functional hepatocytes. Lineage reprogramming is a direct method of generating functional cells in vitro, however, using this method to generate cells with functional maturity similar to their bona fide counterparts remains challenging. To this end, by mimicking the natural tissue regeneration route, we established a two-step lineage reprogramming strategy to generate functional hiHeps (Xie et al., 2019): human fibroblasts were first reprogrammed into hepatic progenitors and then, in a second step, induced into functionally mature hepatocytes that resemble primary hepatocytes. In the second step, a two-chemical combination of forskolin and SB431542 (2C) was identified to be crucial to inducing, capturing and

maintaining the functional state of hiHeps. However, in the established protocol, hiHeps cultured in unspecified hepatocyte culture medium based 2C (H2C) long-term begin to gradually show excessive accumulation of lipid droplets (Figure 4.3A). To overcome this, we tested other basal media and found that hiHeps cultured in 2C formulated with the well-defined and widely used Williams' E medium (W2C) does not exhibit abnormal lipid metabolism (Figure 4.3A). However, W2C-cultured hiHeps are functionally immature compared to primary hepatocytes (Figure 4.3B). Thus, we applied SiPer to predict additional perturbagens that may activate key hepatic transcription factors and enhance functional maturation of hiHeps from hepatic progenitors in the Williams' E medium.

From a list of key TFs associated with functional maturity in PHs, a set of eight differentially expressed TFs in PHs compared to hepatic progenitors (log2 fold change > 4) were selected as query TFs for SiPer analysis (Figure 4.3C). A fraction of cell population with hepatic progenitor markers, denoted as human hepatic progenitor-like cells (hHPLCs) (Xie et al., 2019), was identified by scRNA-seq during the first step of the engineering protocol and this data was used for the SiPer analysis. Notably, the predicted candidates included an adenylyl cyclase activator and a TGFβ inhibitor, which are the two pathways targeted in our 2C maturation medium by forskolin and SB431542 respectively. Apart from forskolin and SB431542, additional thirteen chemical compounds of ten different functional mechanism groups were identified (Table S6).

We performed "W2C+1" test of the thirteen candidates to evaluate their effect on hiHeps maturation. qPCR results showed that seven candidates targeting five pathways: hydrocortisone and dexamethasone (glucocorticoid activators), Bio (a GSK3 inhibitor), progesterone (a progesterone receptor agonist), Trichostatin A (TSA) and valproic acid (VPA) (HDAC inhibitors), and U0126 (a MEK inhibitor) could upregulate the expression of hepatic TFs and functional genes compared to W2C (Figure 4.3D). The combined effect of adding the five compounds targeting five different pathways (hydrocortisone, progesterone, Bio, U0126 and TSA) to W2C further enhanced expression of key hepatic genes (Figure S4.2A). To find the essential combination of the seven compounds, omitting each factor determined forskolin, SB431542 and hydrocortisone to be indispensable for hepatic functional gene expression (Figure S4.2B). Thus, hiHeps cultured in W2C+hydrocortisone (W3C) were further evaluated (Figure 4.3E).

Morphological examination revealed that hydrocortisone containing W3C cultured cells showed typical hepatocyte morphology (Figure 4.3F), without presenting signs of abnormal lipid metabolism (Figure 4.3G). Global hierarchical clustering revealed that W3C-

cultured hiHeps were clustered closely with H2C-cultured hiHeps and PHs (Figure 4.3H). Moreover, the expression levels of hepatic TFs and key hepatic functional genes were more similar between W3C-cultured hiHeps and PHs (Figure 4.3I) than between W2C-cultured hiHeps and PHs, which was further confirmed by qPCR analysis (Figure 4.3J). In addition, the expression levels of the query TFs along with other key hepatic TFs were also upregulated (Figure S4.2C). Importantly, W3C supported long-term culture of hiHeps up to at least 28 days with stably maintained albumin secretion at levels similar to PHs (Figure 4.3K), good hepatocyte morphology and cell survival, whereas W2C-cultured hiHeps showed low albumin secretion with gradual cell death, and H2C-cultured hiHeps showed excessive lipid accumulation (Figure S4.2D). In addition, W3C supported glycogen synthesis and lipoprotein uptake of hiHeps (Figure 4.3L) and drug-metabolizing activity of CYP3A4 and CYP1A2 - key hepatocyte functions - to similar levels to that of PHs (Figure 4.3M). Collectively, the data showed that SiPer effectively identified perturbagens to induce hiHeps maturation in a Williams' E medium. Importantly, the new formulation was identified through a very simplified, straightforward experimental process due to the integration of SiPer in the workflow.

Discussion

The derivation of functionally specific cell populations is highly valuable in disease modelling, drug discovery and regenerative medicine (Shi et al., 2017). However, the clinical application of engineered cell populations via integrative strategies (e.g., delivery of conversion TFs by integrating retroviral or lentiviral vectors) is limited by the safety and ethical concerns (Hockemeyer and Jaenisch, 2016; Shi et al., 2017; Valenti et al., 2019). It is more desirable to perturb target TFs by less invasive means, such as chemical compounds. However, since downstream target TFs of chemical compounds for each cell population are largely unknown, the establishment of chemical-induced protocols for engineering specific cell population remains a challenge. Indeed, bulk RNA-seq data blurs the heterogeneity and asynchrony of cell populations, which hinders the efficiency and precision of cellular conversion, especially for highly similar cell populations at the transcriptional level. A systematic guidance taking advantage of scRNA-seq data can narrow down the scope of candidate chemical compounds and significantly improve the efficiency of experimental processes. Thus, in this study, we developed SiPer, a scRNA-seq based computational platform that predicts perturbagens specifically targeting sets of TFs, either provided by the user or derived from differential expression analysis, to direct desired cell population conversion. Our systematic assessment corroborated that SiPer is accurate and robust in identifying desired perturbagens and

corresponding protein targets in chemical perturbation benchmarking datasets. The high accuracy of SiPer is attributed to the pre-compiled transcriptome signature of perturbation compendium (NCPC). NCPC enables SiPer to select candidate signalling proteins based on prior knowledge, circumventing the requirement of protein activity measures, which are not always available for a wide range of perturbagens. Notably, NCPC is the first large-scale perturbation compendium consisting solely of normal cells, which provides an important resource to learn signalling perturbations of normal cell types. Furthermore, use of scRNA-seq data allows SiPer to be applied to any novel cell populations identified by scRNA-seq. Indeed, many known signalling proteins and chemical compounds for conversion between cell populations with minor transcriptional differences, such as different cellular phenotypic states and cell subtypes, were correctly identified by SiPer. We also applied SiPer to various kinds of cell type conversion examples, including reprogramming and differentiation, which accurately recapitulated experimentally validated perturbagens and corresponding protein targets.

Moreover, we demonstrated the capability of SiPer with experimental validation. A recently reported protocol (H2C) has enabled the successful generation of hiHeps (Xie et al., 2019). However, this protocol employs an unspecified culture medium, and long term cultured hiHeps in this condition showed excessive lipid accumulation, which hinders its use for practical applications, including drug screening and cell transplantation. To this end, the application of SiPer using scRNA-seq data of hHPLCs successfully predicted useful perturbagens, facilitating the formulation of a new maturation medium to generate functional hiHeps from hepatic progenitors in the defined Williams' E medium. The resulting new medium formulation, comprising of rationally guided additives in a fully specified basal medium that were identified after only two rounds of qPCR assays, is able to induce functional hiHeps that are similar to PHs without abnormal lipid accumulation (Figures 4.3D-K). Importantly, hydrocortisone predicted by SiPer was found to be crucial for hepatic progenitor maturation. As a known additive in the mainly unspecified hepatocyte culture medium (HCM), hydrocortisone could be the key component contributing to the improved maturation efficiency of the previously established HCM based 2C over Williams' E based 2C, verifying the predictive power of SiPer. Notably, integrating SiPer predictions into the workflow dramatically simplified the experimental design, ultimately resulting in a straightforward, streamlined experimental process.

In summary, SiPer constitutes a valuable computational platform to facilitate the design of efficient strategies for scRNA-seq based cell engineering using chemical compounds, which

holds great promise for both basic cell research and regenerative medicine. Users can easily apply SiPer to any cellular system by accessing the web interface.

Material and Methods

Construction of normal cell perturbation compendium

Collection and compilation of transcriptome profiles (Perturb-reTFs) Transcriptome profiles (bulk RNA-seq and scRNA-seq) of signalling perturbations across diverse non-cancer cells and tissues in human, mouse and rat were collected from databases Gene Expression Omnibus (GEO) (Barrett et al., 2013), ArrayExpress (Kolesnikov et al., 2015) and LINCS L1000 (only non-cancer cell lines) (Subramanian et al., 2017) (Table S1). Specifically, the keywords commonly used in perturbation studies, such as 'time series', 'response', 'treat', 'perturb', 'presence' and 'effect', were used to search for the datasets in GEO and ArrayExpress. To minimize the off-target effects of signalling perturbations, we only considered datasets where gene expression profiling was conducted within six hours after perturbation, which are considered as initial transcriptional response as suggested by Lamb et al (2006) (Lamb et al., 2006). Identification of DETFs for each dataset was performed as the following procedure. If available, the original submitter-supplied processed data from GEO and ArrayExpress were used for further analysis. Otherwise, the pre-processing including background correction and normalization for bulk RNA-seq was performed with R package limma (v3.38.3) (Ritchie et al., 2015). For the datasets from LINCS L1000, quantile-normalized gene expression profiles were used. Differential expression analysis for bulk RNA-seq datasets was carried out by using R package limma. The genes with Benjamini-Hochberg (BH) adjusted p-value<0.05 and absolute fold change >1.5 were considered as differentially expressed genes (DEGs) for the datasets with at least three replicates. When the replicates were less than three, only the fold change criterion was used. As for scRNA-seq datasets, pre-processing including genes and cells filtering was performed as described in original studies (Table S3). The DEGs for scRNA-seq were identified by using Wilcoxon Rank Sum test in R package Seurat (v3.2.0) (Stuart et al., 2019) with BH adjusted p-value<0.05. DETFs were identified from DEGs based on TF database AnimalTFDB 2.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB2/), including TFs, transcription co-factors and chromatin remodelling factors (Zhang et al., 2015). The expression levels of DETFs were booleanized according to their up-/down-regulation comparing to unperturbed control samples (up-regulation=1, down-relugation=-1). In addition, mouse and rat gene symbols were converted into human homologous gene symbols with R package

Biomart (Durinck et al., 2009). This compendium contained a list of perturbagens and their corresponding response TFs with Boolean value (-1 and 1) and was denoted as Perturb-reTFs.

Identification of signalling protein targets of perturbagens (Perturb-targets) The direct signalling protein targets of perturbagens were collected from database DrugBank (www.drugbank.ca) (Wishart et al., 2018), Drug Repurposing Hub (www.broadinstitute.org/repurposing) (Corsello et al., 2017) and STITCH v5.0 (http://stitch.embl.de) (Szklarczyk et al., 2016). For STITCH, we only kept the protein targets with experiment and database evidence and confidence >0.4. The targets of ligands were identified from manually curated ligand-receptor pairs from Ramilowski et al (Ramilowski et al., 2015). The effect (mode of action) of perturbagens was treated as unknown if two databases reported contradictory effects (e.g.one database reported inhibition, another reported activation) or all databases reported unknown. Otherwise, we kept the effect as inhibition or activation if at least one database reported so. The effects of activation, inhibition and unknown were assigned with values 1, -1 and 2 respectively. This collection of perturbagens and their direct protein targets was denoted as Perturb-targets.

Construction of prior knowledge network

The PKN consists of two layers, the upstream signalling interactomes and downstream TF-TF interactions. These two layers of networks were integrated by interface TFs, which mediate the signal transduction from cytoplasm to the nucleus. The signalling interactomes were constructed by integrating ReactomeFI (Wu et al., 2010) with Omnipath (Türei et al., 2016) databases. The TFs that have no outgoing edge in the signalling network and whose upstream are signalling proteins were defined as interface TFs. These interface TFs connect the upstream signalling interactomes to TF-TF interactions which were manually curated interactions among TFs from MetaCore database from Clarivate Analytics (Supplementary Table 2). In our study, the TFs that have direct interactions with interface TFs were defined as first layer TFs (denoted as 1stTFs). If the query TFs were not located in first layer (non-1stTFs), we identified the 1stTF that can specifically target non-1stTFs as proxy of non-1stTFs (details see Supplementary Note 1).

Collection and pre-processing of benchmarking datasets

Each benchmarking dataset consists of the following three data sources, 1) bulk or single cell transcriptome profiles before and after single chemical perturbation to identify DETFs, 2) scRNA-seq of initial cellular state and 3) a perturbagen with defined target proteins used for

the data in 1). The identification of DETFs for all benchmarking datasets as query TFs is same as the compilation of transcriptome profiles in NCPC as described above. To identify corresponding scRNA-seq of initial cellular state perturbation datasets profiled by bulk RNA-seq, we collected scRNA-seq datasets matching the initial cellular state of bulk perturbation datasets and use these scRNA-seq-bulk-DETF pairs as validation datasets. To ensure that the scRNA-seq expression profile resembles the matching bulk RNA-seq expression profile, the Spearman correlation between bulk RNA-seq of initial cellular state and all collected scRNA-seq datasets was calculated (Supplementary Note 2). Then bulk perturbation datasets whose initial cellular state showed the highest correlation with matching scRNA-seq were maintained as validation datasets, which resulted in 200 benchmarking datasets (Table S3).

Identification of candidates of signalling proteins from NCPC

SiPer first selects candidates of signalling proteins from NCPC based on a set of query TFs with the expression direction information (i.e., up- or down- regulation). The query TFs can be obtained by performing differential expression analysis with SiPer given the scRNA-seq data of initial and target cell populations or provided by user directly. Notably, the identification of DETFs as input is the same as the description in the above section (*Construction of normal cell perturbation compendium*) and we only consider up to top 100 DETFs based on the absolute fold-change to ensure the accuracy of SiPer (Figure 4.2D). Given the set of query TFs, SiPer calculates the similarity between query and each reference in Perturb-reTFs of NCPC by a modified Jaccard similarity coefficient defined as:

$$J(Q,R) = \frac{\sum_{i=1}^{|Q \cap R|} I(Q_i, R_i)}{|Q \cup R|}$$

with indicator function:

$$I(Q_i, R_i) = \begin{cases} 1, if\, Q_i * R_i = 1 \\ 0, if\, Q_i * R_i = -1 \end{cases}$$

where Q and R are query TFs and reference of response TFs in Perturb-reTFs respectively. We modified the Jaccard similarity coefficient by adding an indicator function that assigns a value to the common TFs based on whether their effects are consistent. If the common TFs have the same effect (both inhibition/activation), 1 is assigned, and 0 otherwise. This modified Jaccard similarity coefficient ensures SiPer to consider the number as well as the effects of common TFs between the query and reference. Then perturbagens in Perturb-reTFs are ranked in descending order based on their modified Jaccard similarity coefficient. In addition, the z-score measuring the number of standard deviations from the mean modified Jaccard similarity

coefficient across all references in Perturb-reTFs is calculated for each perturbagen. The perturbagens with z-score >2 are selected for the further analysis. The details of the parameter optimization and selection are described in Supplementary Note 3. Next, the signalling protein targets of each selected perturbagen are retrieved from the Perturb-targets of NCPC and ordered by their frequencies. The top 40% of signalling proteins are considered as candidate signalling proteins (Supplementary Note 3). Furthermore, the "activation" sign is assigned to candidate signalling proteins when more activation effects are reported in NCPC than inhibition effects and vice versa. The "unknown" effect is assigned if all reported effects are unknown.

Prediction of signalling proteins with network model

Candidate signalling proteins identified from NCPC are further filtered to predict the final set of signalling proteins by taking into account the gene expression state of starting cells and PKN among signalling proteins and TFs. The algorithm consists of two major steps: 1) estimation of efficiency of a signalling protein to transmit the signal to a downstream TF, and 2) identification of the most efficient signalling proteins specifically targeting a query set of TFs. In step 1), the efficiency is estimated by two measurements, reachability and specificity as described below.

Reachability scRNA-seq data is able to measure the gene expression heterogeneity among individual cells within the starting cell population. We assume that two genes that are directly connected in the PKN and exhibit similar expression patterns at high expression levels across all single cells have higher potential to transmit the signal between them. In another word, the strength of interaction between two signalling proteins is proportional to the dot product of their expression levels, i.e.,

$$w_{ij} \sim Ei \cdot Ej$$

where Ei and Ej are expression vectors of two genes across cells in scRNA-seq data, w_{ij} is the interaction strength between protein i and j. Since the distribution of w_{ij} is heavily skewed, a log transformation is performed on w_{ij} to make its distribution relatively symmetrical,

$$logw_{ij} = log2(w_{ij} + 1)$$

Then the distribution is normalized by the maximum value as,

$$normalized\ w_{ij} = \frac{logw_{ij}}{max(logW)}$$

Once the interactions strengths (i.e. edge weights) are assigned to the network, the weighted shortest path for each pair of signalling protein and downstream TF is identified using

Dijkstra's algorithms implemented in R package igraph. For this computation, the inverse of edge weight is used as the distance measure between two proteins. The reachability score of each shortest path is then defined as,

$$R_{s\to t} = \Pi_{e\in sp_{s\to t}} Inverse(normalized\ w_e)$$

where s, t represent a signalling protein and a TF respectively, e is an edge in the shortest path from s to t. The reachability score measures to what extent the TF t can receive the signal from signalling protein s via shortest path $sp_{s\to t}$.

Specificity We introduce another measure contributing to the efficiency of signal transduction, specificity, based on the concept of signalling entropy. It states that if a node in a shortest path interacts with all its neighbours with an equal strength, then this node has an equal probability of transmitting the signal to all the neighbours. In this scenario, this node has the maximum uncertainty as to whether it transmits the signal through the shortest path. The probability of signal being transduced from signalling protein i to j is defined as

$$p_{ij} = \frac{w_{ij}}{\sum_{k\in N(i)} w_{ik}}$$

where w_{ij} is the interaction strength between i and j, $N(i)$ is all neighbours of protein i. The uncertainty of signal transmission from a node via the shortest path is measured by signalling entropy,

$$S_i = -\sum_{j\in N(i)} p_{ij} log p_{ij}$$

To avoid the influence of the number of neighbours, the entropy is normalized by the maximum entropy,

$$normalized\ S_i = \frac{S_i}{\max S}$$

where the $\max S$ is assumed when all neighbours have an equal probability, i.e., when the signalling entropy is $-\sum_{j=1}^{n} \frac{1}{n} log \frac{1}{n} = logn$, where n is the number of neighbours for i. The entropy of a shortest path from a signalling protein to a certain TF can be computed by

$$S_{s\to t} = \sum_{n\in sp_{s\to t}} normalized\ S_n$$

where n is proteins in the shortest path from s to t. The smaller the value of entropy $S_{s\to t}$, the more specific the signal transmitting from s to t. The specificity is defined as the exponential inverse of $S_{s\to t}$,

$$reverted\ S_{s\to t} = 1/exp\ (S_{s\to t})$$

Efficiency Based on the reachability and specificity described above, the efficiency (EF) of signal transition from a signalling protein s to TF t is computed by,

$$EF_{s\to t} = R_{s\to t} * reverted\ S_{s\to t}$$

In addition, SiPer aims at identifying the signalling proteins that have the maximum effect on a specific set of TFs, while minimizing the effect on non-target TFs. To this end, we calculate the efficiency score between a candidate signalling protein to every 1stTFs. This set of efficiency scores is then normalized by the sum of all efficiency scores. In an ideal case, a signalling protein targets only on the specific query set of TFs but no other TFs, i.e., the ideal efficiency score vector has a non-zero value only in the indices of target TFs and 0 in all the others. SiPer then calculates the distance between real and ideal efficiency score vectors by considering them as discrete probability distributions whose divergence is measured by JSD

$$JSD(\mathrm{P}, Q) = \frac{1}{2}D(\mathrm{P}, M) + \frac{1}{2}D(Q, M)$$

where P, Q are real and ideal efficiency score vectors, respectively. $M = \frac{1}{2}(\mathrm{P} + Q)$ and D is Kullback-Leibler divergence as

$$D(X, Y) = \sum_i X(i) log \frac{X(i)}{Y(i)}$$

The lower the JSD value, the higher the similarity between the two efficiency score vectors, which in turn implies the higher specificity of the effect of signalling protein to target TFs. Once JSD value is calculated for all candidate signalling proteins inferred from NCPC, the signalling proteins ranked on the top 40-percentile and JSD value not equal to 1 are selected as final candidates. If the number of top 40-percentile of signalling proteins are less than 100, only the ones with JSD value less than 1 are selected (Supplementary Note 4).

Identification of perturbagens targeting predicted signalling proteins

Apart from the prediction of signalling proteins, SiPer further predicts the perturbagens targeting predicted signalling proteins. Noteworthily, besides the perturbagens presented in compendium Perturb-reTFs, we collected 6154 perturbagens with defined target proteins reported in public databases. This enables SiPer to identify even perturbagens that have not been previously used for perturbation-based gene expression profiling. In addition, the perturbagens were categorized into different groups based on their usability in different research areas, including cellular reprogramming, cell cycles, metabolism, immunology, as well as FDA approved drugs. Users of SiPer can use this information to further narrow down candidate perturbagens, depending on the specific application fields.

To predict perturbagens, SiPer first performs over-representation analysis (ORA) on predicted signalling proteins based on the Reactome signalling database using R package WebGestaltR (v0.4.4) (Liao et al., 2019) with the default parameter settings. Since enriched signalling pathways have different extent of overlapping in terms of signalling proteins, which will result in replicate occurrence of signalling proteins across groups. Therefore, we used ReCiPa (Vivar et al., 2013) to merge enriched signalling pathways with high level of redundancy by setting the parameters max_overlap and min_overlap to 0.75 and 0.1 respectively (details about the algorithm can be found in original paper). Predicted signalling proteins are then separated into different functional groups based on the merged signalling pathways. To identify the perturbagens for each group of signalling proteins, a modified Jaccard similarity coefficient is used,

$$J(Q,R) = \frac{\sum_{i=1}^{|Q \cap R|} I(Q_i, R_i)}{|Q \cup R|}$$

with indicator function,

$$I(Q_i, R_i) = \begin{cases} 1, if Q_i * R_i = 1 \\ 0.5, if |(Q_i * R_i)| = 2 \\ 0.25, if Q_i * R_i = 4 \\ 0, if Q_i * R_i = -1 \end{cases}$$

where Q is a group of signalling proteins and R is targets of a perturbagen. The three modes of action (inhibition, activation and unknown) between a perturbagen and its targets are presented by values -1, 1, 2 respectively. If the common proteins between Q and R have same effect, 1 is assigned, and 0 otherwise. If the mode of common proteins is unknown in Q or R, or both of them, 0.5 and 0.25 is assigned, respectively. This modified Jaccard similarity coefficient ensures SiPer to identify the perturbagens targeting predicted signalling proteins with consistent modes as predicted. Finally, SiPer selects perturbagens with top 10 Jaccard similarity coefficients in each functional category and further merges the perturbagens into different groups based on their target similarity using ReCiPa.

Visualization of the paths between perturbagens and query sets of TFs

To further demonstrate the biological insight into predicted perturbagens of users' interest and query TFs, SiPer has a build-in function to visualize the signalling cascades between predicted perturbagens, the predicted targets of perturbagens, intermediate signalling proteins and query TFs. The visualization is achieved by using Cytoscape (v3.8.0) (Shannon et al., 2003) that accesses per R package Rcy3 (v2.2.9) (Gustavsen et al., 2019).

Performance evaluation of SiPer

Since elucidating substrates of chemical compounds with differential expression analysis has been proposed (Ganter et al., 2005; Wolpaw et al., 2011), we evaluated the capability of SiPer to predict correct signalling proteins and perturbagens by comparing with results of DEGs and randomization. In addition, we developed another Perturb-reTFs compendium from both normal and disease cells (mainly cancers) collected from the CMap database (Subramanian et al., 2017) and applied SiPer to the benchmarking datasets using this compendium. The performance of SiPer to predict the protein targets of perturbagens from a specific set of TFs was assessed by the number of correctly predicted (true positive (TP)) signalling protein targets of perturbagens reported in the public databases. The sensitivity computed by the ratio of TP signalling protein predictions to the total number of reported protein targets of perturbagens was used to evaluate the proportion of correct prediction. We did not use measures that consider false positive predictions, since the protein targets of chemical compounds in public databases are not complete and a query set of TFs can be targeted not only by the reported signalling protein targets but also by other signalling proteins in the signalling network. Indeed, predicted signalling proteins not reported in the databases constitute potentially novel perturbations for targeting the query set of TFs. The average sensitivity of SiPer across benchmarking datasets with different cut-off were calculated. To further investigate the property of predicted signalling proteins that were not direct targets of perturbagens, we examined whether these predicted signalling proteins were functionally related to the perturbagens apart from as direct targets. We performed ORA against Reactome signalling pathways with the top 40% of SiPer predicted signalling proteins. We considered the predicted signalling proteins whose enriched pathways with at least one of the direct protein targets were functionally related to the perturbagens. For the prediction of perturbagens, the success rate was considered, which measures the fraction of datasets in which experimentally used perturbagen was predicted. Besides the true perturbagens, the potential perturbagens which had at least one target protein also targeted by the experimentally used perturbagen were also investigated. To examine whether the transcriptional responses of normal cells were distinct from cancer cell counterparts, we identified 1956 perturbagens by which both normal and cancer cell types were stimulated (Table **S**4). For each of these perturbagens, we performed a hierarchical clustering analysis based on DETFs and evaluated whether the normal/cancer cell types were clustered together.

To assess whether SiPer is generally applicable to diverse chemical-based cellular conversion, we applied SiPer to various kinds of cellular conversion examples. Specifically,

we collected examples where experimentally validated conversion TFs or the scRNA-seq data of initial and final cell populations are available in a same study to identify DETFs. Moreover, the perturbagens or essential signalling proteins and pathways for these conversions need to have been experimentally validated in these examples (Table S5).

Evaluation of robustness of SiPer

The performance of SiPer relies on the datasets collected in NCPC. The robustness of SiPer to changes in NCPC was evaluated by randomly removing 10% of datasets in Perturb-reTFs and applying SiPer to the benchmarking datasets. This procedure was repeated 100 times and the average performance was computed. The sensitivity of SiPer to the presence of cell types in NCPC was also investigated by, for each benchmarking dataset, removing all the datasets of the same cell type from NCPC. Furthermore, we tested if SiPer was capable of identifying the target proteins when the correct perturbagen was completely absent in Perturb-reTFs. This was achieved by removing for each benchmarking dataset, the datasets with the same perturbagen as the benchmarking dataset and then performing the SiPer analysis. Moreover, the robustness of SiPer to the PKN was evaluated by randomly removing 10% of interactions in the PKN with 100 iterations. To investigate the predicted signalling proteins are specific to query TFs, SiPer was run on the same number of randomly selected TFs. In addition, we also examined the robustness of SiPer to the number of cells in input scRNA-seq data by randomly selected different number of cells (ranging from 100 to 1000, with 100 increment) and the expression profiles of the selected cells were used. Finally, the impact of the number of query TFs was evaluated by splitting the benchmarking datasets into different groups based on the number of DETFs. The performances of datasets in different groups were compared.

Human primary hepatocyte isolation and cell culture

The present study was approved by the Clinical Research Ethics Committee of China-Japan Friendship Hospital (Ethical approval No: 2009-50) and Stem Cell Research Oversight of Peking University (SCRO201103-03), and it was conducted according to the principles of the Declaration of Helsinki.

Human primary hepatocytes were isolated from human donor livers with informed consent. Briefly, liver tissues were perfused with collagenase IV and dispase (Sigma-Aldrich) until the tissues were incompact and separated with tweezers. Hepatocytes were washed 3 times with HCM^{TM} (Lonza), plated in collagen-coated plates and cultured in HCM^{TM} in the first 12 hours. PHs were then cultured in William's E medium containing 2% B27 (Gibco), 1%

GlutaMAX, and 20 μM forskolin, 10 μM SB431542, 0.5 μM IWP2, 0.1 μM LDN193189 and 5 μM DAPT.

Hepatic progenitors were generated by lineage reprogramming from human fibroblasts (Xie et al., 2019) and cultured in hepatic expansion medium (50% DMEM/F12 and 50% William's E Medium supplemented with 1% PS, 2% B27 (without VA), 5 mM Nicotinamide, 200 μM 2-phospho-L-ascorbic acid (pVc), 3 μM CHIR99021, 5 μM SB431542, 0.5 μM Sphingosine-1-phosphate (S1P), 5 μM Lysophosphatidic acid (LPA), 50 ng/mL EGF and 2 μg/mL doxycycline).

To generate hiHeps, hepatic progenitors were cultured until compactly confluent on 0.2 mg/mL Matrigel-coated plates and then were treated with chemicals formulated with William's E Medium or HCM for 7-10 days unless otherwise specified. Information of chemicals used for hiHeps maturation test were listed in Supplementary Table 6.

Gene expression analysis

Total RNA was isolated by Direct-zol RNA Miniprep (ZYMO RESEARCH) and then reverse-transcribed with TransScript First-Strand cDNA Synthesis SuperMix (TransGen Biotech). RT-qPCR was performed using KAPA SYBR® FAST Universal qPCR Mix (KAPA Biosystems) on a BIO-RAD CFX384TM Real-time System. The quantified values were normalized against the input determined by housekeeping genes (RPL13A or RRN18S). The RT-qPCR primer sequences were provided in Table S6.

Albumin ELISA, PAS staining, LDL uptake and oil red O staining

Secretion of human albumin was measured using the Human Albumin ELISA Quantitation kit (Bethyl Laboratory) according to the manufacturer's instructions. The PAS staining system was purchased from Sigma-Aldrich. Cultures were fixed with 4% paraformaldehyde (DingGuo) and stained according to the manufacturer's instructions. For the LDL uptake assay, hiHeps were incubated with 10 μg/mL DiI-Ac-LDL (Invitrogen) for 4 h and 1 μg/mL Hoechst 33342 (Thermo Fisher Scientific) for 30 min at 37 °C and then washed 3 times before imaging using fluorescence microscopy. Lipid detection was performed with a Lipid (Oil Red O) Staining Kit (Sigma) according to the manufacturer's instructions.

Measurements of drug-metabolizing activity of CYP450s

The methods for measuring the CYP450 activity were described previously (Xie et al., 2019). Briefly, one 500 μL reaction contained 2.5×10^5 cells and the indicated substrates (Testosterone for CYP3A4, Phenacetin for CYP1A2). After incubation for 15–30 min at 37 °C in an orbital shaker, the reactions were stopped by the addition of sample aliquots to tubes containing triple the volume of quenching solvent (methanol) and were frozen at –80 °C. Isotope-labeled reference metabolites (6β-Hydroxytestosterone-[D7] for CYP3A4, Acetomidophenol-[13C2, 15N] for CYP1A2) were used as internal standards. The concentration of products (6β-Hydroxytestosterone for CYP3A4, Acetaminophen for CYP1A2) was analysed by ultraperformance liquid chromatography-tandem mass spectrometry (UPLC/MS/MS). UPLC/MS/MS analyses were performed using an ACQUITY H-Class UPLC System (Waters) coupled to a Sciex API5500Q-trap Mass Spectrometer (SCIEX). The analytical column was an ACQUITY UPLC® BEH C18 1.7 μm 2.1 × 50 mm. hiHeps were cultured for around 17 days in different culture medium to detect drug-metabolizing activities.

RNA Sequencing and Bioinformatics Analysis

For bulk RNA sequencing, total RNA was exacted using Direct-zol RNA Miniprep (ZYMO RESEARCH). RNA sequencing libraries were prepared using the NEBNext UltraTM RNA Library Prep kit for Illumina (NEB, USA) following the manufacturer's recommendations. The fragmented and randomly primed 150-bp paired-end libraries were sequenced on Illumina Hiseq 4000 platform. The generated sequencing reads were mapped against the human genome build hg19 using STAR aligner (Dobin et al., 2013), and the read counts for each gene were calculated using featureCounts. Gene expression was normalized by DESeq2 (v1.22.2) with variance-stabilizing transformation (VST) (Love et al., 2014), and the low expression genes with total counts across all samples less than 10 were excluded. Unsupervised hierarchical clustering of RNA-seq data was conducted by the factoextra package in R. Heatmaps were generated by the pheatmap and gplots R packages.

For single cell RNA sequencing, hHPLCs were harvested based on the criteria in order of completed epithelial conversion of whole-well cells, hepatic progenitor markers (ALB, CK18, CK8, EPCAM, HNF1B, DLK1), and response to the maturation medium to generate functional hepatocytes. hHPLCs were further resuspended at 1×10^6 cells per milliliter in 1 × PBS with 0.04% BSA. Then, cell suspensions (300-1000 living cells per microliter determined by trypan blue staining) were loaded on a Chromium Single Cell Controller (10x Genomics) to generate single-cell gel beads in emulsion (GEMs) by using Single Cell 30 Library and Gel Bead Kit (10x Genomics). Captured cells were lysed and the released RNA were barcoded through

reverse transcription in individual GEMs. Barcoded cDNAs were pooled and cleanup by using DynaBeads MyOne Silane Beads (Invitrogen). Single-cell RNA-seq libraries were prepared using Single Cell 30 Library Gel Bead Kit (10x Genomics) following the manufacture's introduction. Sequencing was performed on an Illumina HiSeq X Ten with pair end 150bp (PE150).

Raw FASTQ files of scRNA-seq data of hepatic progenitors were processed with CellRanger software (v3.1.0), including alignment, filtering, barcode counting and unique molecular identifiers (UMI) counting. Reads were aligned with GRCh38 genomes. Quality control was performed with R package Seurat (v3.2.0) by removing genes expressed in fewer than three cells and low-quality cells with less than 500 genes expressed (Stuart et al., 2019). In addition, cells expressed more than 6000 genes and with more than 10% mitochondrial counts were also excluded.

The pre-processed scRNA-seq of hepatic progenitors and a set of differentially expressed TFs in PHs compared to hepatic progenitors (log2 fold change > 4) were selected as inputs for SiPer analysis. The list of perturbagen candidates returned by SiPer were manually separated into different functional mechanism groups based on their protein targets and functional annotations. Finally, we selected 1-2 perturbagens in the groups that have at least three perturbagens.

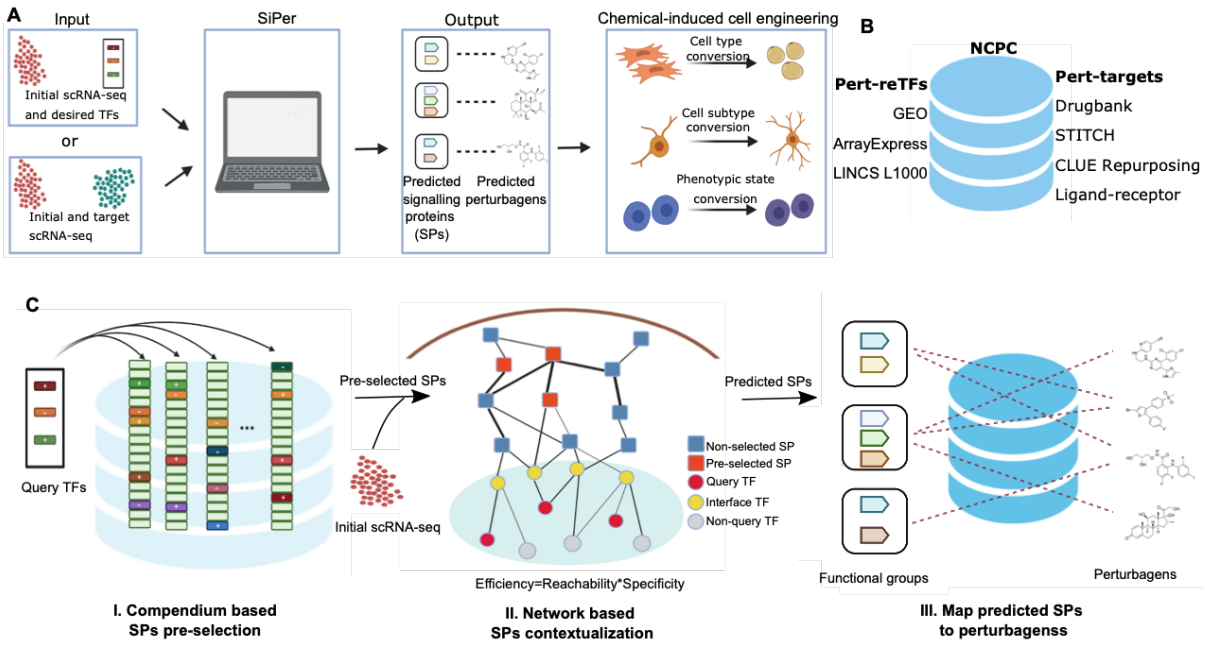

Figure 4. 1 Schematic outline of SiPer.

A, The workflow of SiPer. The input of SiPer can be initial scRNA-seq and desired TFs provided by the user, or initial and target scRNA-seq. Given input, SiPer identifies chemical perturbagens and corresponding signalling protein targets (SPs) to direct cell population engineering, including cell type, cell subtype and phenotypic state conversion.
B, NCPC is composed of two compendia, Perturb-reTFs and Perturb-targets. Perturb-reTFs contains perturbagens and their response TFs identified by transcriptomics data collected from GEO, ArraryExpress and LINCS L1000. Perturb-targets contains perturbagens and their signalling protein targets from Drugbank, STITCH, CLUE Repurposing and manually curated ligand-receptor interactions.
C, SiPer contains three major stages, 1) pre-selection of candidate signalling proteins from the built-in NCPC by using only query TFs, which can be identified by differential expression analysis or provided by user, 2) network-based modelling to predict signalling proteins specifically targeting the query set of TFs using scRNA-seq of initial cell type/state, 3) identification of perturbagens which target the predicted signalling proteins.

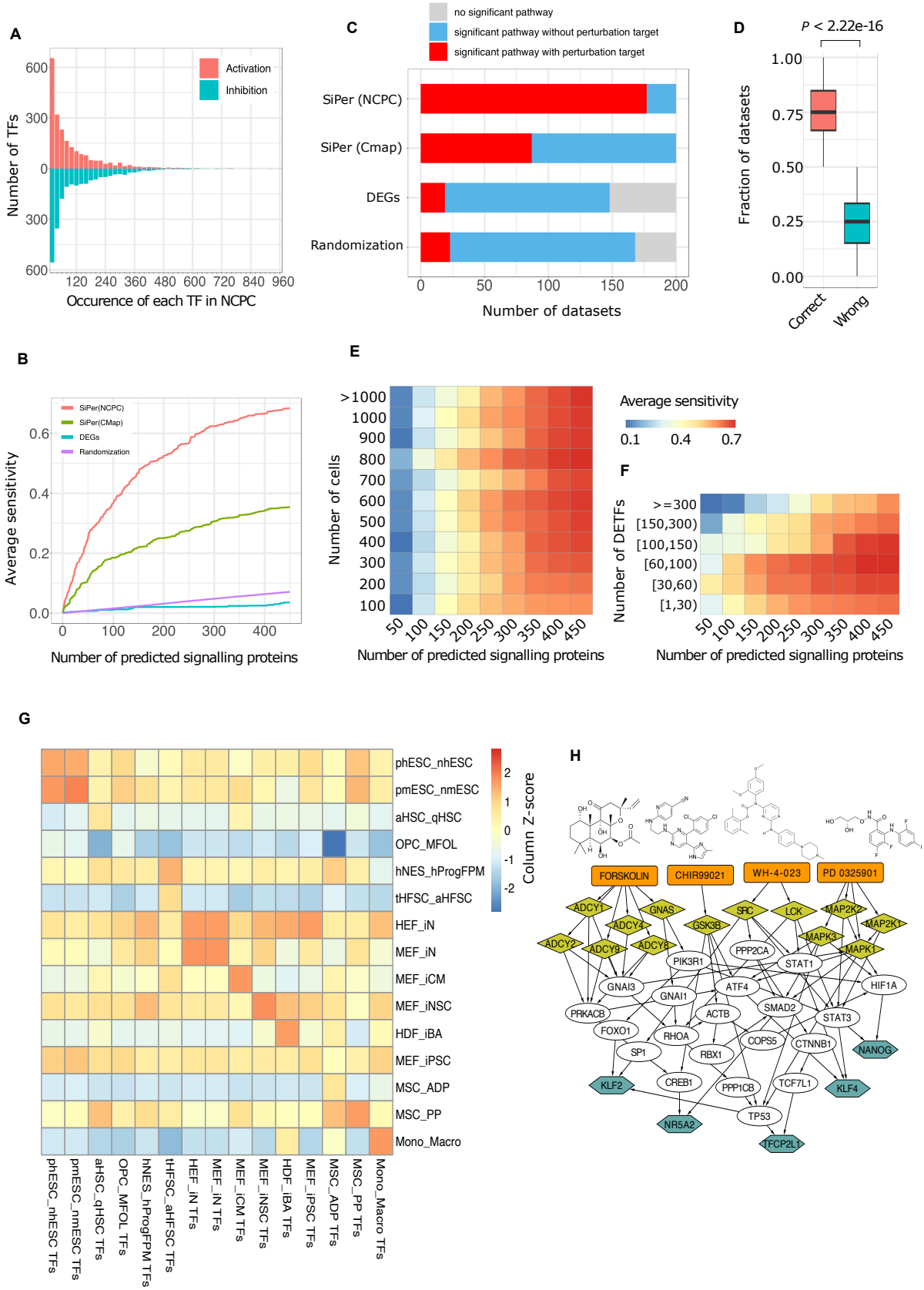

Figure 4. 2 Evaluation and application of SiPer to cellular conversion.

A, Frequency of response TFs in NCPC, including inhibited and activated TFs.

B, Average sensitivity of protein targets across 200 benchmarking datasets at different thresholds for predicted signalling proteins.
C, Number of datasets in which signalling pathways enriched in predicted signalling proteins contain at least one protein target of true perturbagens (red), no protein target of true perturbagens (blue) or no significantly enriched pathway (grey). In this way, we examined whether predicted signalling proteins which are not direct targets of correct perturbagens were nevertheless functionally related to the direct protein targets. This result suggests that non-target proteins predicted by SiPer were more functionally relevant to the correct perturbagen than those predicted by other approaches. Performance was compared among: SiPer using NCPC (SiPer(NCPC)), SiPer using CMap-based Perturb-reTFs (SiPer(CMap)), signalling proteins based on DEGs (DEGs) and random selection of signalling proteins (Randomization) for **B**-**C**.
D, The fraction of datasets correctly clustered to their corresponding class (i.e., normal or cancer) (one-sided Wilcoxon test, *P-value <2.22e-16*).
E, Robustness evaluation of SiPer with respect to input cell number by randomly selecting cells in input scRNA-seq data.
F, Robustness evaluation of SiPer with respect to DETF number by splitting the benchmarking datasets into different subsets based on the number of DETFs.
G, SiPer's efficiency score matrix denoting the regulatory potential of signalling proteins to downstream TFs. Each cell in heatmap represents the average efficiency score of top 20 predicted signalling proteins to corresponding query TF sets for different cellular conversion cases from original cell to target cell. For example, "phESC_nhESC" means primed hESC converts to naïve hESC. Abbreviations: phESC/pmESC: primed hESC/mESC; nhESC/nmESC: naïve hESC/mESC; aHSC/qHSC: active/quiescent HSC; tHFSC/aHFSC: telogen/anagen HFSC; iN: induced GABAergic neuron; iBA: induced brown adipocyte; ADP: adipocyte; PP: pancreatic progenitor; Mono: monocyte; Macro: macrophage.
H, SiPer network visualization of putative signalling cascades for phenotypic state conversion of primed ESCs to naïve state between predicted perturbagens (orange rectangle), predicted signalling protein targets (yellow diamond), intermediate signalling proteins (white ellipse) and query TFs (blue hexagon).

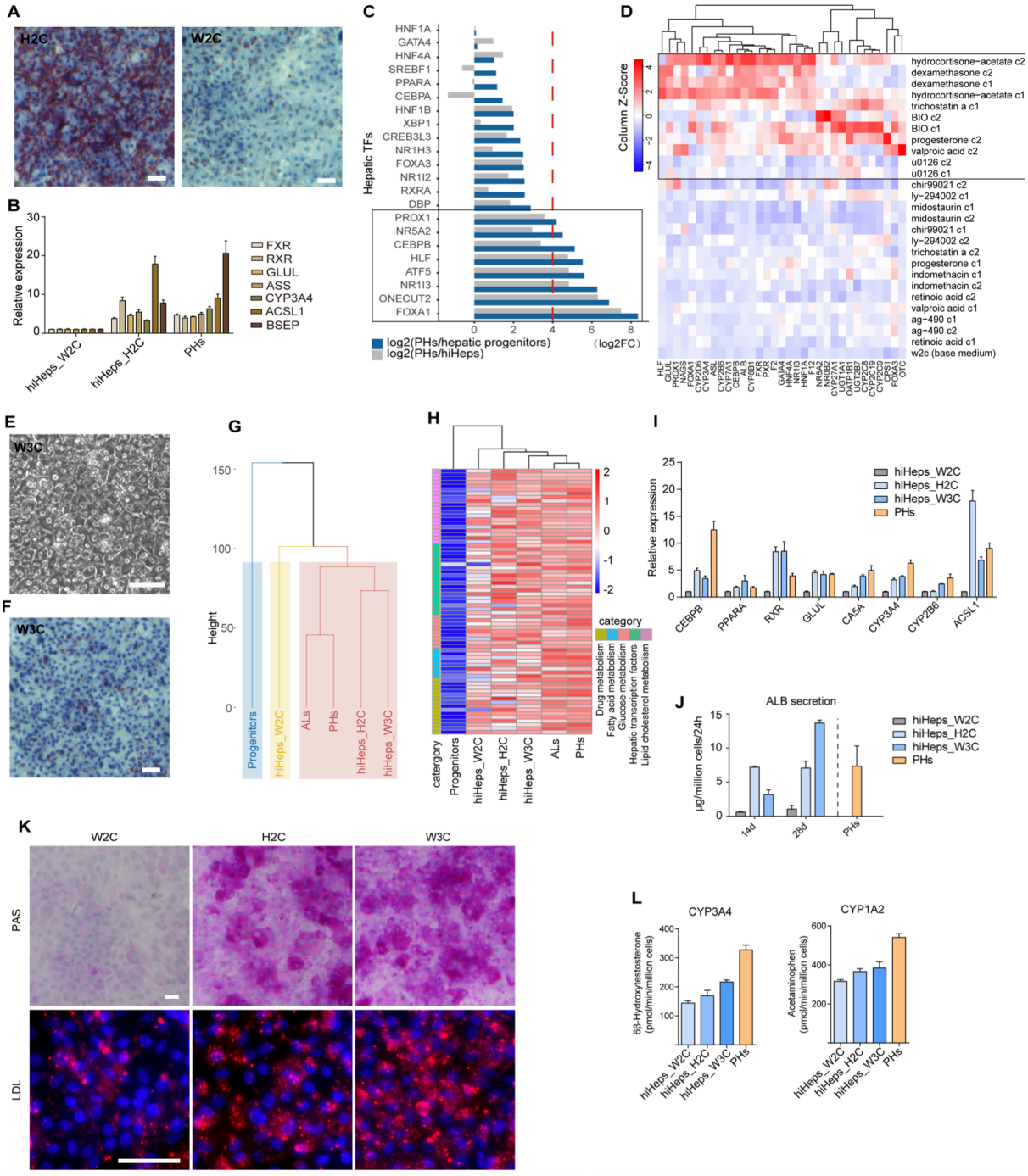

Figure 4. 3 Generation of functional hiHeps by applying SiPer predicted chemical perturbagens.

A, Oil-red staining of lipid synthesis and accumulation in hiHeps cultured in H2C and W2C at day 14 of culture. Scale bar = 50um.

B, qRT-PCR analysis showing gene expression of key hepatocyte markers of hiHeps (n=6) cultured in H2C and W2C compared with primary human hepatocytes (PHs, n=4). Relative expression normalized to W2C. Data are $mean \pm MSE$.

C, Expression fold change of top differentially expressed hepatic TFs (DETFs) in primary human hepatocytes (PHs) vs. hepatic progenitors (blue columns), and PHs vs. hiHeps cultured in H2C (grey columns). The top 8 DETFs were used as query TFs for SiPer analysis.

D, Heatmap of gene expression of key hepatocyte markers in hiHeps cultured in W2C and "W2C+1" condition in which all predicted candidates were screened at two different concentrations (c1 and c2) by qPCR analysis. Seven hits targeting 5 pathways are indicated in the black box.

E, Representative bright field images of hiHeps cultured in W3C. Scale bar = 50um.

F, Oil-red staining of lipid synthesis and accumulation in hiHeps cultured in W3C at day 14 of culture. Scale bar = 50um.

G, Hierarchical clustering of global gene expression of hepatic progenitors, hiHeps cultured in W2C, H2C, and W3C, primary human hepatocytes (PHs), and adult liver tissues (ALs) by RNA-seq analysis.

H, Heatmap of gene expression profile of hepatic transcription factors and functional hepatocyte genes involved in drug metabolism, fatty acid metabolism, glucose metabolism and lipid cholesterol metabolism in hepatic progenitors, hiHeps cultured in W2C, H2C, W3C, primary human hepatocytes (PHs), and adult liver tissues (ALs). Panel of genes analysed listed in Table S6.

I, qRT-PCR analysis of gene expression of key hepatocyte markers in hiHeps (n=6) cultured in W2C, H2C, W3C and primary hepatocytes (PHs, n=4). Relative expression was normalized to hiHeps_W2C. Data are $mean \pm MSE$.

J, ALB secretion of hiHeps (n=3) cultured in W2C, H2C and W3C at day 14 and 28 of culture, and ALB secretion of primary hepatocytes (PHs, n=6) by ELISA. Data are $mean \pm MSE$.

K, Analysis of key hepatic functions of hiHeps cultured in W2C, H2C and W3C: PAS staining of glycogen synthesis (upper panel) and low-density lipoprotein (LDL) uptake staining (lower panel). Scale bar = 50um.

L, UPLC/MS/MS (Ultra performance liquid chromatography - tandem mass spectrometer analysis) of drug-metabolizing activity of CYP3A4 and CYP1A2 of hiHeps cultured in W2C, H2C and W3C, and PHs. n=3. Data are $mean \pm MSE$.

Start cell	End cell	Query TFs	Perturbagens	Protein targets	TF source	Perturba-gens references
Primed hESC	Naïve hESC	NANOG KLF2 LKF4 NR5A2 TFCP2L1	PD0325901 CHIR99021 Forskolin WH-4-023 Y-27632 LIF Activin A	MAP2K1 GSK3B ADCY5 LCK LRRK2 LIFR ACVR2A	(Takashima et al., 2014)	(Park et al., 2018; Theunissen et al., 2014)
Primed mESC	Naïve mESC	Esrrb Nanog Klf2 Klf4 Tfcp2l1	PD0325901 CHIR99021 LIF	Map2k2 Map2k1 Gsk3b Gsk3a Lifr	(Dunn et al., 2014; Ivanova et al., 2006; Martello et al., 2012; Niwa et al., 2009; Ye et al., 2013)	(Ying et al., 2008)
Active HSC	Quiescent HSC	DETFs	U0126 RGB286638 AHPN	Map2k1 Mapk14 Cdk4 Cdk6 Rela Sp1	(Kowalczyk et al., 2015)	(Baumgartner et al., 2018; Cabezas-Wallscheid et al., 2017; Fukushima et al., 2019; Matsumoto et al., 2011)
Telogen HFSC	Anagen HFSC	DETFs	D4476 GW788388	Tgfb1 Tgfbr1 Tgfbr2	(Yang et al., 2017)	(Foitzik et al., 1999; Oshimori and Fuchs, 2012)
hNES	hProgFPM	OTX2 LMX1A FOXA2 LIN28A SOX1 RFX4 HMGA1	Simvastatin Indirubin-3-monoxime Sulindac	H2AFX HRAS GSK3A CDK1 CDK3 CTNNB1	(Okawa et al., 2018)	(Castelo-Branco et al., 2004; Chung et al., 2009; Rhim et al., 2015)
OPC	MFOL	DETFs	GSK429286a Regorafenib	Rock1 Fgfr1 Fgfr2 Pdgfra Pdgfrb	(Marques et al., 2016)	(Marques et al., 2016; Pedraza et al., 2014)

Table 4. 1 Results of cell phenotypic state/subtype conversion examples obtained by SiPer, including protein targets and perturbagens.

Start cell	End cell	Query TFs	Perturbagens	Protein targets	TF source	Perturba-gens references
HEF	GABAer-gic neuron	FOXG1 SOX2 ASCL1 DLX5 LHX6	CHIR99021 PD0325901 Pifithrin-α LDN193189 SB431542 Forskolin	GSK3B MAP2K1 TP53 BMPR1BA CVR1 ADRB2	(Colasante et al., 2015)	(Dai et al., 2015)
MEF	NSC	Pou3f4 Sox2 Klf4 Myc Tcf3	CHIR99021 LDN193189 SMER28 Retinoic acid bFGF	Gsk3b Fkbp1a Rara Rarg Fgfr4	(Han et al., 2012)	(Zhang et al., 2016a)
HDF	Brown adipocyte	CEBPB MYC	Dorsomorph-in Forskolin Rosiglitazone LDN193189 SB431542	FKBP1A KDR ADRB2 PPARG ADRA1A	(Kishida et al., 2015)	(Takeda et al., 2017)
MEF	Cardiomy-ocyte	Hand2 Nkx2-5 Gata4 Mef2c Tbx5	Valproic acid CHIR99021 PD0325901 SC1 TTNPB BayK 8644	Hdac1 Hdac2 Mapk1 Gsk3b Htr2a Adora2a	(Addis et al., 2013)	(Fu et al., 2015; Park et al., 2015)
MSC	Pancreatic progenitor	PDX1 NKX6-1 NEUROD1 NEUROG3 MAFA FOXA2	Nicotinamide Retinoic acid EGF	PARP1 RARG RARG ERBB3 EGFR	(Wang et al., 2011)	(Gao et al., 2008; Sun et al., 2007)
MSC	Adipocyte	PPARG RXRA CEBPA STAT1 SREBF1 FOXO1 EBF1 IRF4	Dexametha-sone Indomethacin Insulin 3-isobutyl-1-methylxanth-ine	NR3C1 NR3C2 PTGS2 IGF1R RB1 ADORA1 CFTR	(Jimenez et al., 2007; Kim and Spiegelman, 1996; Lin et al., 2008; Linhart et al., 2001; Nakae et al., 2003; Nielsen et al., 2008; Stephens et al., 1999; Vasanthakumar et al., 2015)	(Pittenger et al., 1999)

Table 4. 2 Results of cell type conversion examples obtained by SiPer, including protein targets and perturbagens.

Code availability

SiPer was implemented in R and code repository is available from Gitlab (https://git-r3lab.uni.lu/menglin.zheng/SiPer). The web application was developed with PAWS framework and is available at https://siper.uni.lu .

Data availability

Bulk RNA-seq data and scRNA-seq data generated in this study have been deposited to Gene Expression Omnibus GSE162908 and GSE162909, respectively.

Reference

Addis, R.C., Ifkovits, J.L., Pinto, F., Kellam, L.D., Esteso, P., Rentschler, S., Christoforou, N., Epstein, J.A., and Gearhart, J.D. (2013). Optimization of direct fibroblast reprogramming to cardiomyocytes using calcium activity as a functional measure of success. J Mol Cell Cardiol *60*, 97-106.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res *41*, D991-995.

Baumgartner, C., Toifl, S., Farlik, M., Halbritter, F., Scheicher, R., Fischer, I., Sexl, V., Bock, C., and Baccarini, M. (2018). An ERK-Dependent Feedback Mechanism Prevents Hematopoietic Stem Cell Exhaustion. Cell Stem Cell *22*, 879-892.e876.

Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F.B., Pastor-Flores, D., Roma, L.P., Renders, S., Zeisberger, P.*, et al.* (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. Cell *169*, 807-823.e819.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903-915.

Castelo-Branco, G., Rawal, N., and Arenas, E. (2004). GSK-3beta inhibition/beta-catenin stabilization in ventral midbrain precursors increases differentiation into dopamine neurons. J Cell Sci *117*, 5731-5737.

Chen, M., Lu, P., Ma, Q., Cao, Y., Chen, N., Li, W., Zhao, S., Chen, B., Shi, J., Sun, Y.*, et al.* (2020). CTNNB1/β-catenin dysfunction contributes to adiposity by regulating the cross-talk of mature adipocytes and preadipocytes. Sci Adv *6*, eaax9605.

Cho, Y.L., Min, J.K., Roh, K.M., Kim, W.K., Han, B.S., Bae, K.H., Lee, S.C., Chung, S.J., and Kang, H.J. (2015). Phosphoprotein phosphatase 1CB (PPP1CB), a novel adipogenic activator, promotes 3T3-L1 adipogenesis. Biochem Biophys Res Commun *467*, 211-217.

Chung, S., Leung, A., Han, B.S., Chang, M.Y., Moon, J.I., Kim, C.H., Hong, S., Pruszak, J., Isacson, O., and Kim, K.S. (2009). Wnt1-lmx1a forms a novel autoregulatory loop and controls midbrain dopaminergic differentiation synergistically with the SHH-FoxA2 pathway. Cell Stem Cell *5*, 646-658.

Cieślar-Pobuda, A., Knoflach, V., Ringh, M.V., Stark, J., Likus, W., Siemianowicz, K., Ghavami, S., Hudecki, A., Green, J.L., and Łos, M.J. (2017). Transdifferentiation and reprogramming: Overview of the processes, their similarities and differences. Biochim Biophys Acta Mol Cell Res *1864*, 1359-1369.

Colasante, G., Lignani, G., Rubio, A., Medrihan, L., Yekhlef, L., Sessa, A., Massimino, L., Giannelli, S.G., Sacchetti, S., Caiazzo, M.*, et al.* (2015). Rapid Conversion of Fibroblasts into Functional Forebrain GABAergic Interneurons by Direct Genetic Reprogramming. Cell Stem Cell *17*, 719-734.

Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M.*, et al.* (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med *23*, 405-408.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M.*, et al.* (2015). A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. Stem Cell Reports *5*, 763-775.

Dai, P., Harada, Y., and Takamatsu, T. (2015). Highly efficient direct conversion of human fibroblasts to neuronal cells by chemical compounds. J Clin Biochem Nutr *56*, 166-170.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
Dunn, S.J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naïve pluripotency. Science *344*, 1156-1160.
Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184-1191.
Federation, A.J., Bradner, J.E., and Meissner, A. (2014). The use of small molecules in somatic-cell reprogramming. Trends Cell Biol *24*, 179-187.
Foitzik, K., Paus, R., Doetschman, T., and Dotto, G.P. (1999). The TGF-beta2 isoform is both a required and sufficient inducer of murine hair follicle morphogenesis. Dev Biol *212*, 278-289.
Francesconi, M., Di Stefano, B., Berenguer, C., de Andrés-Aguayo, L., Plana-Carmona, M., Mendez-Lago, M., Guillaumet-Adkins, A., Rodriguez-Esteban, G., Gut, M., Gut, I.G.*, et al.* (2019). Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. Elife *8*.
Fu, Y., Huang, C., Xu, X., Gu, H., Ye, Y., Jiang, C., Qiu, Z., and Xie, X. (2015). Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. Cell Res *25*, 1013-1024.
Fukushima, T., Tanaka, Y., Hamey, F.K., Chang, C.H., Oki, T., Asada, S., Hayashi, Y., Fujino, T., Yonezawa, T., Takeda, R.*, et al.* (2019). Discrimination of Dormant and Active Hematopoietic Stem Cells by G(0) Marker Reveals Dormancy Regulation by Cytoplasmic Calcium. Cell Rep *29*, 4144-4158.e4147.
Gao, F., Wu, D.Q., Hu, Y.H., Jin, G.X., Li, G.D., Sun, T.W., and Li, F.J. (2008). In vitro cultivation of islet-like cell clusters from human umbilical cord blood-derived mesenchymal stem cells. Transl Res *151*, 293-302.
Gustavsen, J.A., Pai, S., Isserlin, R., Demchak, B., and Pico, A.R. (2019). RCy3: Network biology using Cytoscape from within R. F1000Res *8*, 1774.
Han, D.W., Tapia, N., Hermann, A., Hemmer, K., Höing, S., Araúzo-Bravo, M.J., Zaehres, H., Wu, G., Frank, S., Moritz, S.*, et al.* (2012). Direct reprogramming of fibroblasts into neural stem cells by defined factors. Cell Stem Cell *10*, 465-472.
Hockemeyer, D., and Jaenisch, R. (2016). Induced Pluripotent Stem Cells Meet Genome Editing. Cell Stem Cell *18*, 573-586.
Hodos, R., Zhang, P., Lee, H.C., Duan, Q., Wang, Z., Clark, N.R., Ma'ayan, A., Wang, F., Kidd, B., Hu, J.*, et al.* (2018). Cell-specific prediction and application of drug-induced gene expression profiles. Pac Symp Biocomput *23*, 32-43.
Hori, K., Cholewa-Waclaw, J., Nakada, Y., Glasgow, S.M., Masui, T., Henke, R.M., Wildner, H., Martarelli, B., Beres, T.M., Epstein, J.A.*, et al.* (2008). A nonclassical bHLH Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling. Genes Dev *22*, 166-178.
Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K.*, et al.* (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science *341*, 651-654.
Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533-538.
Jimenez, M.A., Akerblad, P., Sigvardsson, M., and Rosen, E.D. (2007). Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. Mol Cell Biol *27*, 743-757.

Kim, J.B., and Spiegelman, B.M. (1996). ADD1/SREBP1 promotes adipocyte differentiation and gene expression linked to fatty acid metabolism. Genes Dev *10*, 1096-1107.
Kishida, T., Ejima, A., Yamamoto, K., Tanaka, S., Yamamoto, T., and Mazda, O. (2015). Reprogrammed Functional Brown Adipocytes Ameliorate Insulin Resistance and Dyslipidemia in Diet-Induced Obesity and Type 2 Diabetes. Stem Cell Reports *5*, 569-581.
Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T.*, et al.* (2015). ArrayExpress update--simplifying data submissions. Nucleic Acids Res *43*, D1113-1116.
Komine, O., Nagaoka, M., Hiraoka, Y., Hoshino, M., Kawaguchi, Y., Pear, W.S., and Tanaka, K. (2011). RBP-J promotes the maturation of neuronal progenitors. Dev Biol *354*, 44-54.
Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res *25*, 1860-1872.
Kunitomi, H., Oki, Y., Onishi, N., Kano, K., Banno, K., Aoki, D., Saya, H., and Nobusue, H. (2020). The insulin-PI3K-Rac1 axis contributes to terminal adipocyte differentiation through regulation of actin cytoskeleton dynamics. Genes Cells *25*, 165-174.
Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N.*, et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929-1935.
Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res *47*, W199-w205.
Lin, Y.F., Jing, W., Wu, L., Li, X.Y., Wu, Y., Liu, L., Tang, W., Long, J., Tian, W.D., and Mo, X.M. (2008). Identification of osteo-adipo progenitor cells in fat tissue. Cell Prolif *41*, 803-812.
Linhart, H.G., Ishimura-Oka, K., DeMayo, F., Kibe, T., Repka, D., Poindexter, B., Bick, R.J., and Darlington, G.J. (2001). C/EBPalpha is required for differentiation of white, but not brown, adipose tissue. Proc Natl Acad Sci U S A *98*, 12532-12537.
Liu, Z., Wang, L., Welch, J.D., Ma, H., Zhou, Y., Vaseghi, H.R., Yu, S., Wall, J.B., Alimohamadi, S., Zheng, M.*, et al.* (2017). Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. Nature *551*, 100-104.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.
Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Mendanha Falcão, A., Xiao, L., Li, H., Häring, M., Hochgerner, H., Romanov, R.A.*, et al.* (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science *352*, 1326-1329.
Martello, G., Sugimoto, T., Diamanti, E., Joshi, A., Hannah, R., Ohtsuka, S., Göttgens, B., Niwa, H., and Smith, A. (2012). Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. Cell Stem Cell *11*, 491-504.
Matsumoto, A., Takeishi, S., Kanie, T., Susaki, E., Onoyama, I., Tateishi, Y., Nakayama, K., and Nakayama, K.I. (2011). p57 is required for quiescence and maintenance of adult hematopoietic stem cells. Cell Stem Cell *9*, 262-271.
Morris, S.A., and Daley, G.Q. (2013). A blueprint for engineering cell fate: current technologies to reprogram cell identity. Cell Res *23*, 33-48.
Nakae, J., Kitamura, T., Kitamura, Y., Biggs, W.H., 3rd, Arden, K.C., and Accili, D. (2003). The forkhead transcription factor Foxo1 regulates adipocyte differentiation. Dev Cell *4*, 119-129.

Nielsen, R., Pedersen, T.A., Hagenbeek, D., Moulos, P., Siersbaek, R., Megens, E., Denissov, S., Børgesen, M., Francoijs, K.J., Mandrup, S.*, et al.* (2008). Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. Genes Dev *22*, 2953-2967.
Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes Dev *12*, 2048-2060.
Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118-122.
Okawa, S., Saltó, C., Ravichandran, S., Yang, S., Toledo, E.M., Arenas, E., and Del Sol, A. (2018). Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. Nat Commun *9*, 2595.
Oshimori, N., and Fuchs, E. (2012). Paracrine TGF-β signaling counterbalances BMP-mediated repression in hair follicle stem cell activation. Cell Stem Cell *10*, 63-75.
Park, G., Yoon, B.S., Kim, Y.S., Choi, S.C., Moon, J.H., Kwon, S., Hwang, J., Yun, W., Kim, J.H., Park, C.Y.*, et al.* (2015). Conversion of mouse fibroblasts into cardiomyocyte-like cells using small molecule treatments. Biomaterials *54*, 201-212.
Park, T.S., Zimmerlin, L., Evans-Moses, R., and Zambidis, E.T. (2018). Chemical Reversion of Conventional Human Pluripotent Stem Cells to a Naïve-like State with Improved Multilineage Differentiation Potency. J Vis Exp.
Pedraza, C.E., Taylor, C., Pereira, A., Seng, M., Tham, C.S., Izrael, M., and Webb, M. (2014). Induction of oligodendrocyte differentiation and in vitro myelination by inhibition of rho-associated kinase. ASN Neuro *6*.
Pittenger, M.F., Mackay, A.M., Beck, S.C., Jaiswal, R.K., Douglas, R., Mosca, J.D., Moorman, M.A., Simonetti, D.W., Craig, S., and Marshak, D.R. (1999). Multilineage potential of adult human mesenchymal stem cells. Science *284*, 143-147.
Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., Shin, J.W.*, et al.* (2016). A predictive computational framework for direct reprogramming between human cell types. Nat Genet *48*, 331-335.
Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B.*, et al.* (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun *6*, 7866.
Rhim, J.H., Luo, X., Xu, X., Gao, D., Zhou, T., Li, F., Qin, L., Wang, P., Xia, X., and Wong, S.T. (2015). A High-content screen identifies compounds promoting the neuronal differentiation and the midbrain dopamine neuron specification of human neural progenitor cells. Sci Rep *5*, 16237.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.
Rivetti di Val Cervo, P., Besusso, D., Conforti, P., and Cattaneo, E. (2021). hiPSCs for predictive modelling of neurodegenerative diseases: dreaming the possible. Nat Rev Neurol, 1-12.
Sakaki-Yumoto, M., Liu, J., Ramalho-Santos, M., Yoshida, N., and Derynck, R. (2013). Smad2 is essential for maintenance of the human and mouse primed pluripotent stem cell state. J Biol Chem *288*, 18546-18560.
Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun *9*, 20.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Sharma, S., and Petsalaki, E. (2019). Large-scale datasets uncovering cell signalling networks in cancer: context matters. Curr Opin Genet Dev *54*, 118-124.
Shi, Y., Inoue, H., Wu, J.C., and Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. Nat Rev Drug Discov *16*, 115-130.
Stephens, J.M., Morrison, R.F., Wu, Z., and Farmer, S.R. (1999). PPARgamma ligand-dependent induction of STAT1, STAT5A, and STAT5B during adipogenesis. Biochem Biophys Res Commun *262*, 216-222.
Strasen, J., Sarma, U., Jentsch, M., Bohn, S., Sheng, C., Horbelt, D., Knaus, P., Legewie, S., and Loewer, A. (2018). Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels. Mol Syst Biol *14*, e7733.
Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e1821.
Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K.*, et al.* (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437-1452.e1417.
Sun, B., Roh, K.H., Lee, S.R., Lee, Y.S., and Kang, K.S. (2007). Induction of human umbilical cord blood-derived stem cells with embryonic stem cell phenotypes into insulin producing islet-like structure. Biochem Biophys Res Commun *354*, 919-923.
Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res *44*, D380-384.
Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W.*, et al.* (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254-1269.
Takeda, Y., Harada, Y., Yoshikawa, T., and Dai, P. (2017). Direct conversion of human fibroblasts to brown adipocytes by small chemical compounds. Sci Rep *7*, 4304.
Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L.*, et al.* (2014). Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. Cell Stem Cell *15*, 471-487.
Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods *13*, 966-967.
Valenti, M.T., Serena, M., Carbonare, L.D., and Zipeto, D. (2019). CRISPR/Cas system: An emerging technology in stem cell research. World J Stem Cells *11*, 937-956.
Vasanthakumar, A., Moro, K., Xin, A., Liao, Y., Gloury, R., Kawamoto, S., Fagarasan, S., Mielke, L.A., Afshar-Sterle, S., Masters, S.L.*, et al.* (2015). The transcriptional regulators IRF4, BATF and IL-33 orchestrate development and maintenance of adipose tissue-resident regulatory T cells. Nat Immunol *16*, 276-285.
Vivar, J.C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. Omics *17*, 414-422.
Wang, H.W., Lin, L.M., He, H.Y., You, F., Li, W.Z., Huang, T.H., Ma, G.X., and Ma, L. (2011). Human umbilical cord mesenchymal stem cells derived from Wharton's jelly differentiate into insulin-producing cells in vitro. Chin Med J (Engl) *124*, 1534-1539.

Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z.*, et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-d1082.
Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol *11*, R53.
Xie, B., Sun, D., Du, Y., Jia, J., Sun, S., Xu, J., Liu, Y., Xiang, C., Chen, S., Xie, H.*, et al.* (2019). A two-step lineage reprogramming strategy to generate functionally competent human hepatocytes from fibroblasts. Cell Res *29*, 696-710.
Xu, J., Du, Y., and Deng, H. (2015). Direct lineage reprogramming: strategies, mechanisms, and applications. Cell Stem Cell *16*, 119-134.
Yang, H., Adam, R.C., Ge, Y., Hua, Z.L., and Fuchs, E. (2017). Epithelial-Mesenchymal Micro-niches Govern Stem Cell Lineage Choices. Cell *169*, 483-496.e413.
Ye, S., Li, P., Tong, C., and Ying, Q.L. (2013). Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. Embo j *32*, 2548-2560.
Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.
Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res *43*, D76-81.
Zhang, J.W., Klemm, D.J., Vinson, C., and Lane, M.D. (2004). Role of CREB in transcriptional regulation of CCAAT/enhancer-binding protein beta gene during adipogenesis. J Biol Chem *279*, 4471-4478.
Zhang, M., Lin, Y.H., Sun, Y.J., Zhu, S., Zheng, J., Liu, K., Cao, N., Li, K., Huang, Y., and Ding, S. (2016). Pharmacological Reprogramming of Fibroblasts into Neural Stem Cells by Signaling-Directed Transcriptional Activation. Cell Stem Cell *18*, 653-667.
Zhang, X., Yalcin, S., Lee, D.F., Yeh, T.Y., Lee, S.M., Su, J., Mungamuri, S.K., Rimmelé, P., Kennedy, M., Sellers, R.*, et al.* (2011). FOXO1 is an essential regulator of pluripotency in human embryonic stem cells. Nat Cell Biol *13*, 1092-1099.
Zimmerlin, L., Park, T.S., Huo, J.S., Verma, K., Pather, S.R., Talbot, C.C., Jr., Agarwal, J., Steppan, D., Zhang, Y.W., Considine, M.*, et al.* (2016). Tankyrase inhibition promotes a stable human naïve pluripotent state with improved functionality. Development *143*, 4368-4380.
Zimmerlin, L., Park, T.S., and Zambidis, E.T. (2017). Capturing Human Naïve Pluripotency in the Embryo and in the Dish. Stem Cells Dev *26*, 1141-1161.

4.1.3 Supplementary Information

Supplementary Figures

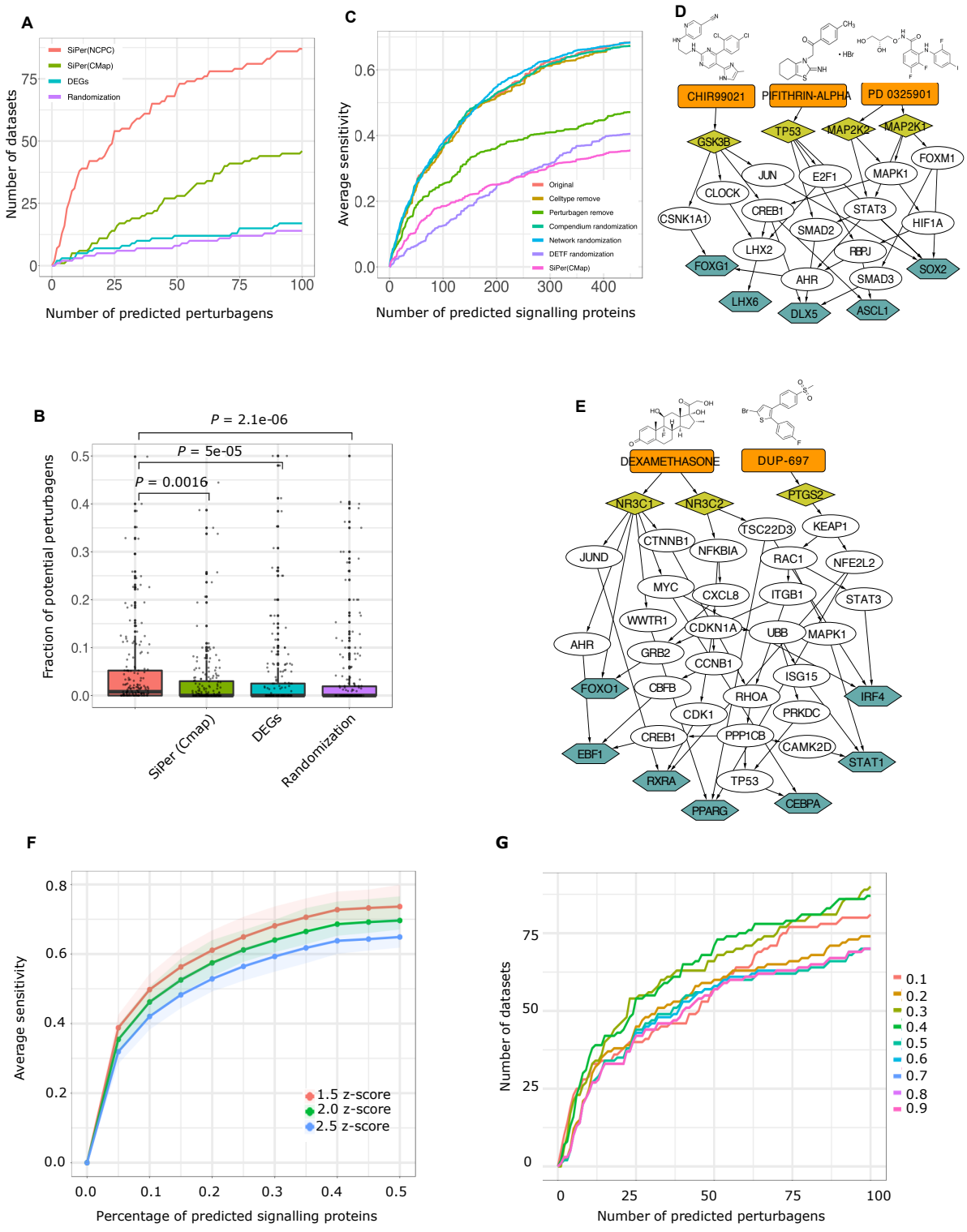

Figure S4. 1 Evaluation and parameter tuning of SiPer.

A, Number of datasets for which true perturbagens are predicted at different thresholds for predicted perturbagens in each functional group.

B, Fraction of perturbagens with at least one protein target (denoted as potential perturbagens) among predicted perturbagens. SiPer predicted a significantly higher fraction of potential perturbagens than CMap-based Perturb-reTFs, DEGs and randomization. P-values were calculated by one-sided Wilcoxon test. Performance was compared among: SiPer using NCPC (SiPer(NCPC)), SiPer using CMap-based Perturb-reTFs (SiPer(CMap)), signalling proteins based on DEGs (DEGs) and random selection of signalling proteins (Randomization) for **A**-**B**.
C, Robustness analysis of SiPer in terms of average sensitivity tested by removing NCPC datasets with same cell type as test dataset (Celltype remove), removing NCPC datasets with same perturbagen as test dataset, random removal of 10 % NCPC datasets , random removal of 10% interactions in PKN, randomly selecting same number of TFs as the number of test DETFs, and replacing the NCPC compendium with CMap-based compendium.
D-E, SiPer network visualization of putative signalling cascades between predicted perturbagens (orange rectangle), predicted signalling protein targets (yellow diamond), intermediate signalling proteins (white ellipse) and query TFs (blue hexagon). **D,** Reprogramming into GABAergic neurons from HEFs, **E,** differentiation of MSCs to adipocytes.
F, Parameter tuning for the z-score and percentage used in the stage of pre-selection of candidate signalling proteins from NCPC.
G, Parameter tunning for the fraction of final candidate signalling proteins used for the prediction of perturbagens.

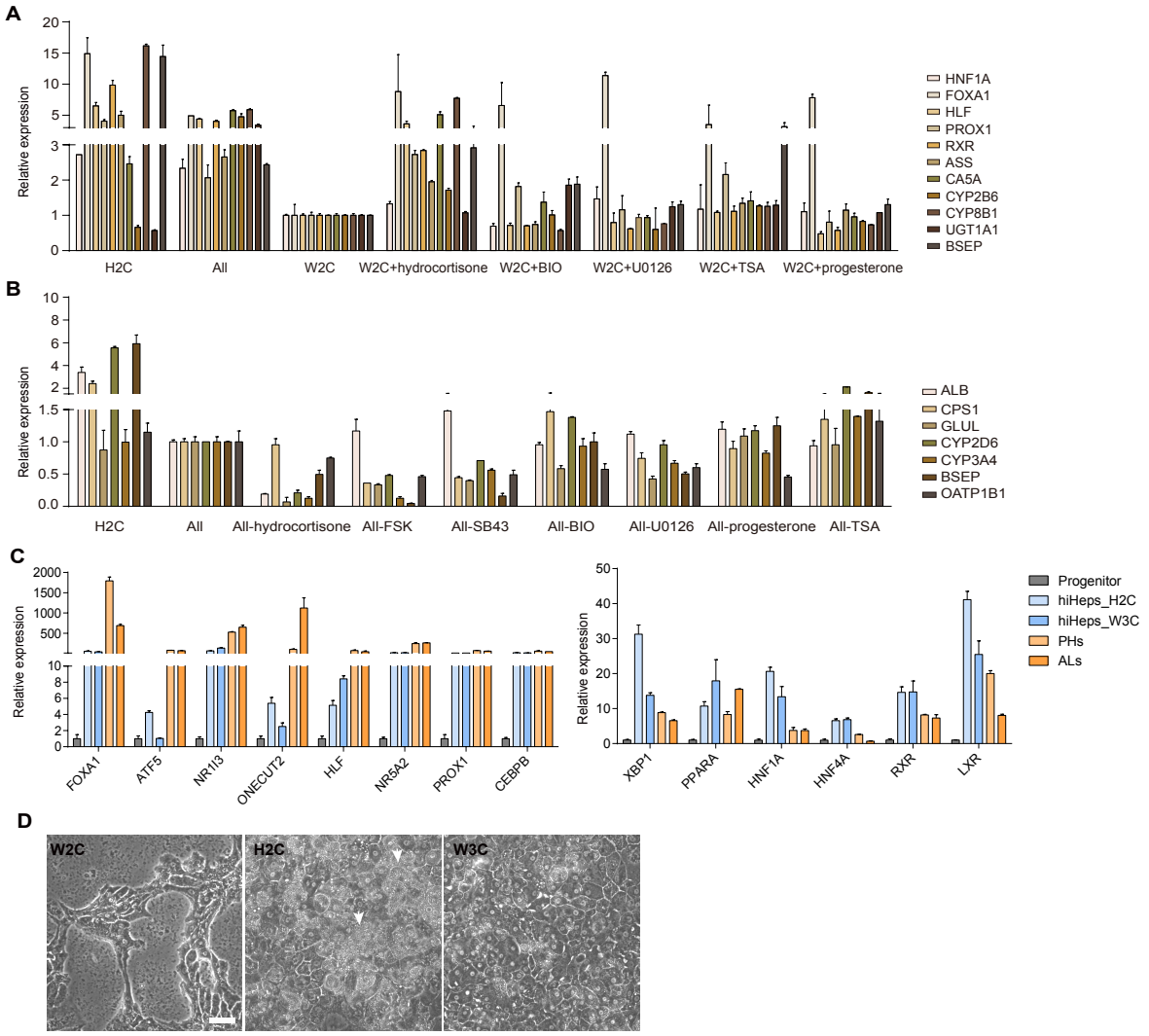

Figure S4. 2 Characterization of hiHeps cultured in chemical compounds predicted by SiPer.

A, Gene expression of key hepatocyte markers in hiHeps cultured in H2C, All (W2C+hydrocortisone, BIO, U0126, TSA and progesterone), W2C and "W2C+1" condition by qRT-PCR analysis. Relative expression normalized to hiHeps_W2C. n=2. Data are $mean \pm MSE$.

B, Gene expression analysis of key hepatocyte markers in hiHeps cultured in H2C, All (W2C+hydrocortisone, BIO, U0126, TSA and progesterone) and single molecule omitted ("All-1") conditions by qRT-PCR analysis. Relative expression normalized to hiHeps_All. n=2. Data are $mean \pm MSE$.

C, qRT-PCR analysis of gene expression of query hepatic transcription factors (left) and other key hepatic transcription factors (right) in hepatic progenitors(n=4), hiHeps cultured in H2C (n=6), W3C (n=6), PHs (n=2) and ALs (n=2). Data are $mean \pm MSE$.

D, Representative bright field images of hiHeps cultured in W2C, H2C, and W3C at day 28 of culture. Arrows indicate lipid accumulation in H2C-cultured hiHeps. Scale bar = 50um.

Supplementary Notes

Supplementary Note 1. Identification of 1stTF as proxy for the query non-1stTFs

Even though the 1stTFs can cover the majority of TFs, some of the query TFs can be located in non-first layer. In this case, we identified the 1stTF as proxy for the query non-1stTFs. First, in order to ensure the TF interactions are specific to the initial cellular state, we used scRNA-seq of the initial cellular state to contextualize the prior knowledge TF interaction network. The TFs in the "hairball" network were considered as "expressed" if their expression value were non-zero in more than 50% of cells. Otherwise, they were treated as "not expressed". The edges of the prior knowledge TF network were kept if their edge signs were "Unspecified" or consistent with the state of the two connected TFs. For example, if the edge sign was "Activation", the interaction will be kept if the connected TFs are both "expressed" or "not expressed". If the edge sign was "Inhibition", the interaction will be maintained if the states of the two connected TFs are either "expressed and not expressed" or "not expressed and expressed". After the contextualization of the TF interaction network, the reachability and the specificity of 1stTFs to the query non-1stTFs were examined. We counted the minimum path length from each 1stTF to the maximum number of the query non-1stTFs and then calculated the number of TFs which could be reached and effected by this 1stTF. In an ideal case, the proxy of query non-1stTFs only has an effect on the specific non-1stTFs but not on other TFs. SiPer then calculates the distance between the real and ideal effected TFs by JSD, as was also performed for the identification of signalling proteins,

$$JSD(\mathrm{P}, Q) = \frac{1}{2}D(\mathrm{P}, M) + \frac{1}{2}D(Q, M)$$

where P, Q are real and ideal effected TF vectors, respectively. $M = \frac{1}{2}(\mathrm{P} + Q)$ and D is Kullback-Leibler divergence as

$$D(X, Y) = \sum_i X(i) log \frac{X(i)}{Y(i)}$$

The lower the JSD value, the higher the specificity of the 1stTF to target the query non-1stTFs. Finally, the 1stTF with smallest JSD value was selected as a proxy for the query non-1stTFs.

Supplementary Note 2. Identification of scRNA-seq of initial cellular state for perturbation datasets profiled by bulk RNA-seq

In order to validate SiPer, we collected perturbation datasets profiled by bulk RNA-seq as well. We collected 239 bulk perturbations in our NCPC compendium, with existing scRNA-seq of corresponding initial cellular state from other studies (62 collected scRNA-seq). To ensure the

scRNA-seq dataset matches corresponding initial cellular state as much as possible, spearman correlation was calculated between the scRNA-seq datasets and bulk RNA-seq of initial cellular state of perturbation dataset. First, the common genes among the collected scRNA-seq datasets for human and mouse were identified separately, which resulted in 8721 and 6525 common genes for human and mouse respectively. For each bulk RNA-seq of perturbation dataset, the spearman correlations with all scRNA-seq data were calculated by using the average expression values of intersected genes between this bulk data and scRNA-seq common genes. We only kept the bulk perturbations whose initial cell type was consistent with that of highest correlation scRNA-seq. For example, the dataset of E-TABM-786 and corresponding scRNA-seq data from GSE114952 were maintained, because the MEFs stimulated with Mitomycin and Bleomycin in E-TABM-786 had highest correlation with scRNA-seq of MEFs from GSE114952. Accordingly, 165 bulk perturbations were kept for benchmarking. The details of bulk and scRNA-seq datasets and their Spearman correlation value and final selected benchmarking datasets can be found in Table S3.

Supplementary Note 3. Parameter tuning for perturbagens and signalling proteins selection in NCPC

In order to determine a set of suitable candidates of signalling proteins from NCPC, it is essential to optimize the thresholds for perturbagens and signalling proteins selection in Stage 1 of SiPer. We employed bootstrapping for the parameter tuning as follows. We randomly selected 100 datasets from Perturb-reTFs. For each dataset, the response TFs were used as query and the modified Jaccard indices between the set of query TFs and all references in Perturb-reTFs database except this dataset itself were calculated. To determine the number of selected perturbagens, the z-score measuring the number of standard deviations from the mean modified Jaccard similarity coefficient across all references in Perturb-reTFs was calculated for each perturbagen. The perturbagens with z-score larger than threshold were considered. We set the threshold of z-score to 1.5, 2.0 and 2.5 respectively. Once the perturbagens were selected, the signalling protein targets of each selected perturbagen were retrieved from the Perturb-targets of NCPC and ordered by their frequencies. To optimize the number of signalling proteins for subsequent analysis, the thresholds ranging from 5% to 50% with 5% increment were used. In total, there were 3*10 pairs of thresholds for each dataset and the average sensitivity among the 100 selected datasets were computed for each threshold pair. Finally, the above procedure was repeated for 1000 times and the average sensitivity among these 1000 repetition were calculated. The results showed that the sensitivity between Z-score 2 and 1.5 at

each rank cut-off was relatively closer than that of Z-score 2 and 2.5 (Figure S4.1F). In addition, the sensitivity increased very slightly after rank cut-off of 40%. Therefore, to ensure the higher significance as well as to maintain more real targets for following analyses, we only considered the perturbagens with Z-score larger than 2 and the signalling proteins ranking on the top 40% as candidates of signalling proteins from NCPC for further analysis.

Supplementary Note 4. Parameter tuning of signalling proteins selection for final perturbagen prediction

Apart from the prediction of signalling proteins, SiPer further predicts the perturbagens targeting predicted signalling proteins (Stage 3 of SiPer). In order to determine the optimal number of signalling proteins used for the prediction of perturbagens, different percentiles of predicted signalling proteins of benchmarking datasets, ranging from 10% to 90% with 10% increment, were selected. The selected signalling proteins were separated into different groups by performing ORA based on the Reactome signalling database. For each threshold of the percentile of predicted signalling proteins, the number of datasets for which true perturbagens were predicted at different cutoffs of predicted perturbagens was investigated (Figure S4.1G). The results show that there is no significant difference between different percentiles, but the signalling proteins rank on the top 40-percentile have the best performance. Therefore, we use the top 40-percentile as the threshold to select the signalling proteins for perturbagen prediction. In addition, to ensure the number of final selected signalling proteins being enough to identify enriched signalling pathways, we considered all the siganlling proteins with JSD value less than 1 if the top percentile of signalling proteins were less than 100. Notably, this parameter could be further optimized with more benchmarking datasets in the future.

4.2 Manuscript 2: A database-driven computational method to identify chemical compounds reverting disease phenotype

4.2.1 Preface

Given the desired set of TFs, SiPer predicts chemical compounds specifically targeting these TFs to induce cellular conversion by the integration of the perturbation database focusing on non-cancer cells and the network model. However, the sets of TFs that can induce the change of cellular states are not always available. Therefore, instead of targeting specific sets of TFs, another method, ChemPert, is developed to predict chemical compounds that can induce the change of cellular states by integrating a perturbation database with a network-based model. The database of ChemPert includes any two cellular states before and after perturbation, which results in more than 10-fold scale up of database in SiPer. The method was applied to revert pathologic phenotypes to their healthy counterparts for different kinds of non-cancer diseases, including aged-related diseases and infectious diseases. A considerable number of clinical or pre-clinical applied chemical compounds in corresponding disease was captured by this method, demonstrating the usability of ChemPert for drug discovery.

In this study, I manually collected and compiled the perturbation datasets to construct the database and developed the whole pipeline of the method. I also performed the comparison between ChemPert and the other existing methods and finally applied the method to different diseases.

4.2.2 Manuscript

A database-driven computational method to identify chemical compounds reverting disease phenotype

Menglin Zheng[1], Satoshi Okawa[1,2], Antonio del Sol[1,3,4,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Esch-sur-Alzette, L-4367 Belvaux, Luxembourg;

[2] Integrated BioBank of Luxembourg, Dudelange L-3555, Luxembourg;

[3] CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 801 Building, 48160 Derio, Spain;

[4] IKERBASQUE, Basque Foundation for Science, Bilbao 48013, Spain;

[*] Correspondence: Antonio del Sol (Antonio.delSol@uni.lu)

Abstract

The standard experimental methods employed for the identification of optimal drugs are costly and resource consuming. Several computational methods have been developed to screen candidates in silico in an unbiased manner based on perturbation databases. However, the databases of these methods are not suitable for the study of non-cancer diseases since they mainly focus on cancer cells that exist extensive rewiring of signalling pathways. Here, we constructed a large scale of perturbation database consisting solely of transcriptional signatures from non-cancer cells. This database was combined with a network model for the prediction of chemical compounds to revert the cellular disease phenotypes to their healthy counterparts. The method consistently outperformed other existing computational methods. Moreover, it was applied to different kinds of non-cancer diseases, including aged-related diseases and infectious diseases. A considerable number of the chemical compounds applied in current state-of-the-art therapeutics were recapitulated. Therefore, the proposed method is of great utility in identifying drugs for disease treatment systematically.

Introduction

Gene expression profile is postulated to be a surrogate for disease phenotype where reversion of the profile in disease condition to its healthy counterpart is an indication of therapeutic efficacy. Phenotypic screening as a conventional strategy for drug discovery relies on the exhaustive trial-and-error testing of a large number of compounds, which is time-consuming and labour-intensive (Eder et al., 2014). In this regard, computational approaches taking advantage of large-scale data, especially the widely available transcriptome data, have been developed for drug discovery. These methods can be broadly grouped into two classes, de novo prediction and signature matching inference. The de novo algorithms attempt to infer causal upstream signalling protein targets of chemical compounds whose perturbation result in observed dysregulated genes (Browaeys et al., 2020; Fakhry et al., 2016; Krämer et al., 2014; Parikh et al., 2010). These methods use gene expression as a proxy for protein activity without considering the presence of post-translational modifications. This limits the accuracy of these methods in predicting druggable signalling protein targets due to the poor knowledge of signalling activity. The signature matching methods, instead of inferring the activation of pathway directly, rely on pre-compiled transcriptomic signatures of given chemical perturbations to infer the chemical compounds associated with query signatures (Lamb et al., 2006; Schubert et al., 2018; Subramanian et al., 2017). While the latter approach has an advantage in making accurate predictions for samples present in the compendium, it is not suitable for the prediction of drugs outside the compendium and for recognising subtle differences in cell states that may result in distinct downstream responses. In addition, the current existing compendia of gene expression profiles analysed after cell perturbations are all based on cancer cells or mixing of normal and cancer cells (Lamb et al., 2006; Subramanian et al., 2017; Xiao et al., 2015). In cancer cells, the signal transduction pathways and transcriptional logics are significantly different from those of non-cancer cells (Sharma and Petsalaki, 2019). Therefore, the compendia are not suitable for identifying drugs for non-cancer disease.

To address these limitations, here we present an integrative method, ChemPert, that combines a manually curated database of transcriptional signatures and a network-based model for the prediction of perturbagens (including small molecules, drugs, cytokines and growth factors) to revert the cellular disease phenotypes to their healthy counterparts. Notably, our database solely consists of manually collected transcriptional signatures of non-cancer cells before and after chemical perturbations. Moreover, unlike the existing methods that predict chemical compounds directly from a database, ChemPert predicts signalling proteins from the

perturbation database and then identifies potential chemical compounds targeting these proteins. This allows ChemPert to identify more potential chemical compounds out the scope of the perturbation database. Since the affinities between chemical compounds and protein targets are also determined by initial cellular states besides perturbagens, the expression level of initial cellular state is further integrated into a network model. The integration of a network model aims at ensuring the predicted signalling protein targets and subsequently corresponding perturbagens are specific to the initial cellular states.

We showed that, using our database purely consisting of non-cancer perturbation datasets to predict perturbagens led to significantly better performance than the ones including cancer datasets, underscoring that the transcriptional signature database in this study serves as an important resource. Our benchmarking also revealed that integrating a network model considering initial cellular state with database-driven inference further prioritized the cellular state-specific signalling protein targets. We compared our method with other existing methods and showed that our method significantly outperforms in recapitulating both signalling proteins and perturbagens. Furthermore, ChemPert was applied to different types of diseases including age-related and infectious diseases, and captured a considerable number of perturbagens used in clinics and animal models to revert disease phenotype.

In summary, these results demonstrate the efficacy of our method in identifying potential drugs to revert non-cancer disease phenotype. ChemPert can easily be applied to different disease indications and is freely accessible at https://siper.uni.lu/chempert, requiring only differential expression genes and the gene expression profile of the initial cell state. Therefore, we believe that our method is valuable for clincal and pharmaceutical research in prioritization of drugs for disease treatment.

Results

Overview of ChemPert algorithm

ChemPert identifies perturbagens to convert cell disease state to normal representation. The algorithm of ChemPert is composed of three major stages (Figures 4.4A-C), 1) pre-selection of candidate signalling proteins from the curated perturbation database of transcriptional signatures, 2) network-based modelling to predict signalling proteins specific to the initial cellular state, 3) identification of perturbagens which target the predicted signalling proteins.

Stage 1. ChemPert identifies signalling proteins in database whose perturbation result in similar transcriptional signatures as required. Specifically, ChemPert obtains query transcriptional signatures from DEGs (see Methods) and then a modified Jaccard index

between transcriptional signatures of query and references in the database is computed (Figure 4.4A). The perturbagens with high value of the modified Jaccard index are selected and their corresponding signalling protein targets are identified from our build-in perturbagen target database (Figure 4.4A). The signalling proteins are ranked and the top ones are selected based on the sum of Jaccard index of their corresponding perturbagens (see Methods).

Stage 2: The candidate signalling proteins identified in Stage 1 do not consider the initial cellular state. However, the affinities of signalling proteins to chemical perturbations are often highly cell type-specific (Hodos et al., 2018; Strasen et al., 2018). Therefore, in Stage 2, ChemPert further predicts the signalling proteins that are specific to the initial cellular state by integrating signalling networks and expression profile of the initial cell state. First, we postulate that the active interface transcription factors (iTFs) transmit the signal from signalling pathway to gene regulatory network and in turn trigger dysregulation of downstream genes. Accordingly, ChemPert identifies the active iTFs that are enriched for DEGs based on TF regulons (Figure 4.4B) (Garcia-Alonso et al., 2019). Subsequently, to identify the signalling proteins from stage 1 whose perturbation signal can be transferred to these active iTFs, we hypothesise that the signal transduction is not possible to be accomplished through the paths whose intermediate signalling proteins are not available. Therefore, we estimate the protein availability from the gene expression profile and perform enrichment analysis on the paths for each pair of signalling protein and downstream iTF (Figure 4.4B). As a consequence, the iTFs that have the possibility to receive signal for each signalling protein are identified. Furthermore, to ensure that the predicted signalling proteins specifically act on the active iTFs and have a minimized effect on the non-active ones, ChemPert computes Jensen-Shannon divergence (JSD) for each signalling protein (see Methods). High JSD value indicates that the signalling protein can not specifically target the active iTFs, which are removed from the signalling protein candidates (Figure 4.4B).

Stage 3: The top 200 predicted signalling proteins from Stage 2 are used to identify perturbagens. ChemPert predicts perturbagens whose targets are enriched for top ranking signalling proteins by using aREA with our build-in target database (Figure 4.4C).

The composition of non-cancer cell perturbation database and prior knowledge network (PKN)

In order to infer the relationship between the signalling perturbation and downstream transcriptional signatures, we exhaustively collected and compiled transcriptome profiles of chemical perturbations from public resources (see Methods). This resulted in a database consisting of around 63000 transcriptional signatures derived from 2779 unique perturbagens

(Table S7). Most of the perturbagens (~70%) have relatively low occurrence frequency (Figure 4.5A) and covered 2133 unique transcriptional regulators in both activation (up) and inhibition (down) directions with no significant bias towards either of them (Figure 4.5B). Importantly, our database solely consists of transcriptomics data of chemical perturbations across 146 unique normal cell types/lines/tissues, since signalling pathways and transcriptional regulatory networks of cancer cells are known to exhibit significant rewiring, which are different from the normal counterparts (Sharma and Petsalaki, 2019). As shown in Figure 4.5C, cell type neurons, cell line HA1E and tissue liver have the largest number of datasets among all cell types, cell lines and tissues in our database, respectively. Notably, majority of the perturbations (~97%) in our database have duration not larger than 24 hours, which can reflect the transcriptional response of perturbation more specifically (Figure 4.5D).

The PKN used in the network-based prediction of signalling proteins is constructed by integrating ReactomeFI (Wu et al., 2010) with Omnipath (Türei et al., 2016), which resulted in the PKN composing of 8845 nodes. In the PKN, 1190 iTFs whose upstream are signalling proteins and have no downstream targets were identified. The path length for each signalling protein to iTFs was determined by using the majority length of shortest paths from one signalling protein to all iTFs. Signalling proteins with path length 3 occupied the highest number (Figure 4.5E).

Integration of non-cancer database and network model significantly increases the predictive power of ChemPert

To investigate whether our non-cancer database can increase the predictive power for normal cell perturbation, we used two other databases: 1) CMap database alone, which mainly contain perturbation datasets of cancer cells (Subramanian et al., 2017), 2) combining our database with CMap database. We randomly selected the normal cell perturbation datasets from the databases as benchmarking datasets and these datasets were removed from the databases afterwards for fair comparison. We applied ChemPert to benchmarking datasets using three databases. The result showed that a better performance was consistently obtained when the original database was used, especially comparing to the predictions from CMap database only (Figure 4.6A). To ensure that the better performance was caused by the difference between normal and cancer cells rather than the unique perturbagens from our database, we constructed two databases solely consisting of normal or cancer cells with the same perturbagens. Consistently, using our normal cell perturbation database showed higher average sensitivity than using the cancer one (Figure 4.6B). Importantly, the signalling proteins of some

perturbagens can only be predicted by using the normal cell perturbation database (Table S8). For example, 3M003, a TLR agonist, was not included in both databases. Its protein targets (TLR7, TLR8) were predicted and ranked at the top of the predictions when ChemPert used the normal cell perturbation database, whereas they were not predicted completely using the cancer cell perturbation database. We found that the transcriptional signatures from the dataset perturbed by 3M003 are similar with those from the datasets perturbed by another TLR agonist, resiquimod, on normal cells, but not on cancer cells (Figure 4.6C). These results indicate that we cannot accurately infer transcriptional signatures of normal cells using cancer counterparts, and therefore it is essential to construct a normal cell-specific perturbation database to predict the signalling perturbations specific for non-cancer cells.

In addition, the affinities of protein targets to external stimuli are often cell state-specific. To this end, ChemPert further filtered out the potential false positive signalling proteins from the database by integrating the signalling network with gene expression level of initial cellular state to ensure that the predictions are specific to the initial cellular state. Following the network model, the ranks of true positive signalling proteins increased significantly (one-sided Wilcoxon test, *p-value = 9.197e-07*) (Figure 4.6D).

Comparison with existing computational methods

To evaluate the performance of ChemPert, a comprehensive comparison between ChemPert and other computational methods that are commonly used to infer signalling proteins or perturbagens from DEGs (CMap query, NicheNet, DeMAND and QuaternaryProd) was performed (Browaeys et al., 2020; Fakhry et al., 2016; Lamb et al., 2006; Woo et al., 2015). Specifically, DeMAND and QuaternaryProd are both network-based methods and can only predict signalling proteins whose perturbations cause downstream gene dysregulation. CMap is designed to predict perturbagens existing in its compendium by identifying the ones that have similar signatures as the input given by user. NicheNet is able to predict ligands only by modelling the relationship between ligands and downstream genes. Due to the different capabilities of these methods, first we compared ChemPert with DeMAND and QuaternaryProd in terms of predicting signalling proteins. The average sensitivity in identifying the direct protein targets of the correct perturbagens across all the benchmarking datasets was evaluated. The results revealed that ChemPert significantly outperformed DeMAND and QuaternaryProd consistently (Figure 4.7A). Indeed, ChemPert had an 18-fold and 9-fold higher average sensitivity compared to DeMAND and QuaternaryProd in the top 100 predicted signalling proteins (the average sensitivity is 0.54, 0.03 and 0.06 for ChemPert,

DeMAND and QuaternaryProd, respectively). Next, ChemPert was further compared with CMap in identifying perturbagens. Averagely, only around 10% of the datasets whose perturbagens were identified by CMap, but our method obtained the true perturbagens in 68% of the datasets (Figure 4.7B). Moreover, the ranks of experimentally used perturbagens that were detected by each method were examined. The median rank of true perturbagens in ChemPert was 29, which was much higher compared to the median rank in CMap (105) (Figure 4.7C). Although CMap only provided the predicted perturbagens, we further identified the protein targets of its predicted perturbagens from the public databases. The predicted protein targets were ranked based on their frequency and compared with other methods. While the average sensitivity of CMap was also systematically better than DeMAND and QuaternaryProd, ChemPert still had significantly better performance than CMap (one-sided Wilcoxon test, *p-value=1.071e-13* in terms of sensitivity in top 100 predictions) (Figure 4.7A). Since ChemPert is capable of predicting different kinds of perturbagens, including drugs and ligands, we further compared ChemPert with NicheNet in terms of ligand predicting. In the top 100 predicted perturbagens, almost 60% of datasets whose experimentally used ligand were identified by ChemPert, but NicheNet predicted the true ligands in only around 45% of datasets (Figure 4.7D). Importantly, the ranks of true ligands in ChemPert were significantly higher than those in NicheNet (one-sided Wilcoxon test, *p-value < 2.2e-16*) (Figure 4.7E). In conclusion, these results revealed a superior performance of ChemPert in predicting signalling proteins and perturbagens.

Recapitalizing perturbagens involved in age-related diseases

The method was first applied to age-related diseases, and a considerable number of predicted perturbagens ranked at the top 30 have been employed in clinics or validated experimentally for the treatment of the corresponding diseases (Table 4.3 and Table S9). For example, a substantial number of literatures have shown that the PI3K/AKT/mTOR pathway plays a crucial role in cartilage degradation and can be used as a therapeutic target for the clinical treatment of osteoarthritis (OA) (Pal et al., 2015; Sun et al., 2020a). Our method was able to identify the PI3K and mTOR inhibitors NVP-BEZ235 and Wortmannin ranking at 1 and 7 respectively. In addition, other top-ranking perturbagens predicted by our method, such as celastrol (Liu et al., 2020), succinate (Yang et al., 2015), fucoxanthin(Ha et al., 2021), flavopiridol (Haudenschild et al., 2019), SAHA (Makki and Haqqi, 2016), erlotinib(Brooks, 2013)and gefitinib(Sun et al., 2018), were all reported to be promising drugs to treat or prevent OA. Another application of our method in aging-associated disease is atherosclerosis, and 12

out of the top 30 predicted perturbagens have been investigated as potential therapeutic strategies against atherosclerosis (Table 4.3). We identified a heat shock protein 90 (HSP90) inhibitor geldanamycin, which has been shown to attenuate inflammation in atherosclerosis (Madrigal-Matute et al., 2010). HU-211, which is one of the cannabinoid-based drugs (Klein and Newton, 2007; Mach et al., 2008) and COX-2 inhibitor celecoxib (Pang et al., 2019; Papageorgiou et al., 2016), were demonstrated to have beneficial effect in atherosclerosis. The predicted rutin was also reported to ameliorate diabetic atherosclerosis burden by suppressing the premature senescence of cells (Li et al., 2018b). In addition, we also successfully predicted a certain number of drugs to treat neurodegenerative disorder diseases, such as Parkinson's disease (PD) and Alzheimer's disease (AD). Specifically, the drug coumarin that has been used together with levodopa to treat PD (Stefanachi et al., 2018), and niacin that has been demonstrated to attenuate neuro-inflammatory response in PD (Chong et al., 2021; Fukushima, 2005; Giri et al., 2019), were predicted and ranked at the top one and two respectively. We also examined other age-related diseases like sarcopenia and type2 diabetes using our method and captured certain number of drugs with clinical or preclinical support, which are shown in Table 4.3 and Table S9 detailly.

Accurately capturing perturbagens for infectious diseases

The method was further applied to different infectious diseases to identify anti-infection drugs. First, given the transcriptional signatures and expression data of influenza A-infected lung tissues, the method identified a certain number of antiviral drugs ranking at the top 30 that have been reported to inhibit the replication of influenza A viruses, such as geldanamycin, mitomycin C, alvocidib, thapsigargin and oroxylin A (Table 4.4 and Table S10). Specifically, we predicted geldanamycin ranking at the top one which was shown to suppress the replication of the viruses both in vitro and in vivo via inhibiting Hsp90 (Taechowisan et al., 2020). The influenza viral replication is also blocked upon the treatment with Mitomycin C or alvocidib by inhibiting viral RNA synthesis. Thapsigargin and oroxylin A both can promote the secretion of interferon to induce host antiviral defences. Given the predictive power of our method in identifying drugs against influenza A viruses, we extended the application of the method to COVID-19 infection. Interestingly, many of the drugs predicted to treat COVID-19 were overlapping with those against influenza A, such as geldanamycin, chalcone, loratadine, thapsigargin and oroxylin A (Table 4.4). These drugs were also suggested by other studies to be as therapeutic options for COVID-19. Indeed, studies have shown that anti-influenza virus drugs hold promise for the treatment of COVID-19 (Indari et al., 2021; Wang et al., 2020) and

they share infection pathways (Barh et al., 2020). In addition, some other predicted drugs, like moxifloxacin and imiquimod, were currently only reported to show potential for COVID-19 using different mechanisms. Moxifloxacin has been shown to have a strong interaction with COVID-19 potential drug target, main protease protein (Mpro), preventing the COVID-19 replication (Marciniec et al., 2020). Imiquimod, a toll-like receptor (TLR) 7 agonist, is capable of triggering both the innate and acquired immune response and provides an effective therapeutic approach (Angelopoulou et al., 2020). Regarding the drugs against Epstein-Barr virus (EBV) infection, our method identified usnic acid, which has been shown to exhibit inhibition effect on EBV activation. Moreover, cyclin-dependent kinase (CDK) inhibitor alsterpaullone and protein kinase C (PKC) inhibitor staurosporine that induce apoptosis of EBV-infected cells were also predicted (Goswami et al., 2012; Watanabe et al., 2020). Studies have revealed that Ephrin receptor A2 (EphA2) interacts with EBV entry proteins to induce the fusion of EBV into epithelial cells (Chen et al., 2018; Zhang et al., 2018). ALW-II-41-27, an inhibitor of EphA2 was predicted and this implicates that inhibiting EphA2 could be an effective therapeutic strategy against EBV.

Conclusion

In this study, we developed an integrative method, ChemPert, that predicts perturbagens acting on desired genes to revert the non-cancer disease phenotype by integrating a manually curated perturbation database with a network-based model. One of the challenges to infer exact signalling paths targeting specific genes is the scarcity of protein activity data, such as protein phosphorylation measurements. Thus, instead of predicting the exact signalling paths, ChemPert infers signalling proteins from the precompiled perturbation data compendium, whose perturbations result in the dysregulation of similar transcriptional signatures to the query ones. Although similar strategies have been taken in previous studies (Lamb et al., 2006; Schubert et al., 2018; Subramanian et al., 2017), their compendia were mainly derived from cancer cells. In addition, a bottleneck for the experimental use of chemical compounds is off-target effects, i.e., the chemical compound could lead to activation/inhibition of undesired downstream genes. Herein, ChemPert attempts to identify the chemical compounds activating/inhibiting a specific set of iTFs, while minimizing the effect on other undesired ones. This distinguishes ChemPert from the other existing computational methods which rely on DEGs for signalling perturbation inference. As a consequence, ChemPert does not require post-translational data, and only requires DEGs and RNA-seq of initial cellular state as input,

allowing its application to shift the cellular disease states to their healthy counterparts for any non-cancer disease with chemical compounds.

Our results showed that ChemPert was capable of consistently predicting signalling proteins and perturbagens used in benchmarking datasets. Notably, using our transcriptional signature database for ChemPert had significantly better performances than those of using the databases including cancer cells. Furthermore, we compared ChemPert with other existing methods and showed that ChemPert superiorly outperformed in predicting signalling proteins as well as perturbagens. Finally, we showed the applicability of ChemPert in identifying potential perturbagens (drugs) for various diseases, such as aged-related diseases and infectious diseases. A considerable number of clinical or preclinical drugs for these diseases were recapitulated by ChemPert.

In summary, we believe that ChemPert is a useful tool that rapidly narrows down the number of candidates in an unbiased manner. This facilitates the identification of druggable signalling protein targets as well as drugs for a wide range of diseases, holding great promise on drug discovery.

Methods

Construction of non-cancer cell chemical perturbation database

In this study, we constructed a database depicting the relationship between external stimuli, intermediate signalling protein target of stimuli and downstream transcriptional signatures. First, in order to demonstrate the relationship between external perturbation and effected gene signatures, we manually curated a large scale number of chemical perturbation datasets focusing on normal cell types/lines or tissues in human, mouse and rat. We first collected transcriptome profiles of chemical perturbations from Gene Expression Omnibus (GEO) (Barrett et al., 2013) and ArrayExpress (Kolesnikov et al., 2015). The datasets were pre-processed, including background correction and normalization, either from the original studies or using limma package (v3.38.3) (Ritchie et al., 2015). In addition, we also extracted the chemical perturbation datasets of non-cancer cell lines from database LINCS L1000 at the Level 3, where quantile-normalization was performed (Subramanian et al., 2017).The transcriptional signatures of each chemical perturbation were obtained by performing differential expression analysis using limma package. The genes with Benjamini-Hochberg (BH) adjusted p-value<0.05 and absolute fold change >1.5 were considered as DEGs compared to un-perturbed control samples when the sample replicates were larger than two. Otherwise, only the fold change was used as criterion. The differential transcriptional regulators based on

TF database AnimalTFDB 2.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB2/) (Zhang et al., 2015), including transcription factors (TFs), transcriptional co-factors and chromatin remodelling factors, were identified from DEGs and considered as transcriptional signatures. Furthermore, these transcriptional signatures were assigned with Boolean value 1 and -1, which represented up-regulation and down-regulation after perturbation respectively. Furthermore, gene symbols of mouse and rat were converted to human homologous gene symbols with R package Biomart (v2.38.0) (Durinck et al., 2009).

In addition, the direct signalling protein targets of perturbagens were retrieved from the database Drug Repurposing Hub (www.broadinstitute.org/repurposing) (Corsello et al., 2017), DrugBank (www.drugbank.ca) (Wishart et al., 2018), and STITCH v5.0 (http://stitch.embl.de) (Szklarczyk et al., 2016). In STITCH, only the targets with a confidence value larger than 0.4 were kept, along with the experiment and database evidence. The receptor targets of ligands were identified from manually curated ligand-receptor pairs from Ramilowski et al. (Ramilowski et al., 2015). The effects of perturbagens on protein targets, including activation, inhibition and unknown, were assigned with value 1, -1 and 2 respectively.

Identification of signaling protein candidates from database

Given a set of query transcriptional signatures between two states of cells, we first identify the potential signalling protein targets from the database whose perturbation can induce similar transcriptional signatures as the query ones. More specifically, the similarity between query transcriptional signatures and those of each perturbation dataset (e.g., reference) in the database is calculated by using a modified Jaccard similarity coefficient as:

$$J(Q,R) = \frac{\sum_{i=1}^{|Q \cap R|} I(Q_i, R_i)}{|Q \cup R|}$$

with indicator function:

$$I(Q_i, R_i) = \begin{cases} 1, if\, Q_i * R_i = 1 \\ 0, if\, Q_i * R_i = -1 \end{cases}$$

where Q and R are query and reference transcriptional signatures in the database respectively. In order to ensure the consistent effect of a transcriptional signature between the query and the reference, we modified the Jaccard similarity coefficient by adding an indicator function. If the transcriptional signature has the same effect (both inhibition/activation), then 1 is assigned, and 0 otherwise. The perturbagens are ranked based on the similarity coefficient in descending

order. Only the highly confident perturbagens with z-score of similarity coefficient larger than 3.5 are selected for the further analysis. Next, we retrieve the signalling protein targets of each selected perturbagen from databases and order the signalling proteins based on the sum of their corresponding perturbagens. The effects of signalling proteins are reported based on the majority effects of their perturbagens. For example, value 1 is assigned to the candidate signalling protein when more predicted perturbagens have activation effect on it. Value 2 is assigned to a signalling protein only if all of its predicted perturbagens have unknown effect on it.

Identification of short paths between signalling proteins and interface TFs

The prior knowledge signalling protein network is constructed by integrating ReactomeFI (Wu et al., 2010)with Omnipath (Türei et al., 2016),which are two comprehensive databases including several other signaling pathway resources. The TFs whose upstream are signalling proteins and have no downstream targets are defined as interface TFs. They transmit the signal from cytoplasm to the nucleus and connect the signalling pathway network with the downstream gene regulatory network. Based on the signalling protein network, the short paths from one signalling protein to the interface TFs (iTFs) are identified as follows: First, the shortest path lengths from each signalling proteins to all interface TFs are calculated using unweighted breadth-first algorithm implemented in R package igraph. Subsequently, the path length with the largest number of interface TFs can be reached by this signalling protein is considered as the final path length. We use this final path length to calculate all the possible short paths from this signalling proteins to the interface TFs within this length. This procedure is employed for all signalling proteins in the prior knowledge network.

Identification of cell type-specific signaling proteins based on a network-model

To ensure that the predicted signalling proteins are specific to the initial cellular state, the method takes into account the gene expression level of initial cells by integration of a network model. The network model mainly consists of three steps. Initially, the active interface TFs are predicted using database DoRothEA v2 (Garcia-Alonso et al., 2019), which integrates different resources of TF-target interactions (TF regulons). The TF regulons and the ranked DEGs based on fold change or p-value are passed to msviper function of Viper package (v1.18.1) (Alvarez et al., 2016) to carry out aREA. The interface TFs are ranked based on p-value of aREA analysis in ascending order and the top maximum 50 interface TFs are selected as active ones until the targets of this set of TFs cover half of the DEGs. Then, we perform Gene Set

Enrichment Analysis with Package fgsea (v1.10.1) (Sergushichev, 2016) on the short paths for each signalling protein to the downstream interface TFs. Specifically, the genes are ranked according to the mean expression level of the initial cellular state in descending order. Then we examine if the intermediate signalling proteins of all short paths for each pair of signalling protein and interface TF are enriched for the highly expressed genes. To this end, we only keep the interface TFs with adjusted p-value less than 0.25 for each signalling protein, which indicates that these TFs have the possibility to be reached by this protein. In an ideal case, the signalling protein can specifically induce DEGs should reach all active interface TFs rather than the non-active ones. Therefore, we use vectors to denote the targeted interface TFs for ideal case and each signalling protein. We calculate the similarity of the two vectors by considering them as discrete probability distributions whose divergence is measured by JSD

$$JSD(P,Q) = \frac{1}{2}D(P,M) + \frac{1}{2}D(Q,M)$$

where P is the vector of desired active interface TFs in the ideal case and Q is the vector of interface TFs targeted by a signalling protein. $M = \frac{1}{2}(P+Q)$ and D is Kullback-Leibler divergence as

$$D(X,Y) = \sum_i X(i) log \frac{X(i)}{Y(i)}$$

The higher the JSD value, the lower the similarity between the two vectors. Finally, the predicted signalling proteins from the database are further filtered if they have a JSD value equal to 1.

Identification of perturbagens targeting the predicted signalling proteins

Given a list of ranked signalling protein candidates, the perturbagens targeting these proteins are derived from our database as follows: Similar to the identification of the active interface TFs, the perturbagen-target interactions are converted into a regulon-like class so that they can be passed to msviper function together with the top 200 ranked signalling proteins using aREA. The predicted perturbagens are ranked based on the normalized enriched score (NES) and the ones with false discovery rate less than 0.05 are kept.

Comparison with other computational methods

To evaluate the predictive capacity of ChemPert, a comprehensive comparison with other four existing methods that are developed for inferring signalling perturbation is performed. The detail implementation of each method is described as follows.

Connectivity map query Query tool in Connectivity map (CMap) searches for chemical and genetic perturbations which have closest signatures to input signatures given by user (Lamb et al., 2006). To apply the CMap query to identify signalling proteins, first, top 150 DEGs (ordered by absolute fold change) are submitted to the CMap query computing tool "sig_fastutc_tool" (https://clue.io/query#l1000batch). Only perturbagens with CMap connectivity score (tau) larger than 95 as recommended are selected and then their protein targets are further identified from the public databases (see above section "Construction of non-cancer cell chemical perturbation database"). These protein targets are then ranked by frequency. CMap requires the input datasets with at least10 DEGs.

DeMAND DeMAND is a computational tool aiming at predicting mechanism of action (MoA) of a compound, given gene expression profiles before and after perturbation and a molecular interaction network (Woo et al., 2015). The faster version of DeMAND (DeMANDfast) is run with default parameters using our PKN, since the context specific network is not available. DeMANDfast ranks DEGs based on a p-value which measurs how significantly the DEGs can dysregulate their downstream genes after their perturbation. DeMAND requires the input dataset with at least three replicates for both before and after perturbation.

NicheNet NicheNet learns cellular communication by modelling the linking of ligands to downstream genes with the integration of expression profiles and molecular interaction networks (Browaeys et al., 2020). Given a set of genes of interest, NicheNet predicts the ligands which trigger their expression change. The genes with non-zero values in at least one sample are considered as background genes and the DEGs are selected as genes of interest. The potential ligands are chosen when at least one of their corresponding targets (i.e. receptors) are expressed in at least one sample. The ligands are ranked based on Pearson correlation coefficient between the predicted and the real transcriptional responses. Notably, NicheNet is unable to predict any non-ligand perturbagens.

QuaternaryProd Given a set of DEGs, QuaternaryProd identifies upstream regulators by performing causal reasoning with a statistitical test based on network (Fakhry et al., 2016). Here, the causal relation engine with Quaternary Dot Product scoring statistic over the human STRINGdb is performed as suggested by the authors. Gene symbols for the mouse datasets are

converted into human homologous Entrez IDs. The default parameter values are used but the log fold change threshold. Here, log2(1.5) is used to ensure agreement with other methods. QuaternaryProd requires datasets with at least two replicates for both before and after perturbation samples.

Due to the different requirement of each method, first, we randomly select ten groups of 250 datasets from our databases which fulfilled the requirements of all methods as benchmarking datasets for the comparison among ChemPert, CMap, DeMAND and QuaternaryProd (i.e., the datasets with at least three replicates for each condition and at least 10 DEGs). Since NicheNet is capable of predicting ligands only, another 10 groups of 250 datasets with ligand perturbation are randomly chosen for the comparison between ChemPert and NicheNet. Notably, the benchmarking datasets are removed from the database when we implement ChemPert. The average sensitivity (true-positive rate) for identifying the direct protein targets of the experimentally used perturbagens is calculated for each group of datasets, as a function of the number of top predicted signalling proteins to measure the capability of predicting signalling proteins. In terms of predicting perturbagens, the fraction of datasets whose experimentally used perturbagens being predicted is computed, as a function of the number of top predicted perturbagens.

Disease	Perturbagen	Rank	Reference
Osteoarthritis	NVP-BEZ235 Celastrol Wortmannin Succinate Fucoxanthin Flavopiridol Adenosine SAHA Erlotinib Gefitinib	1 3 7 11 15 18 21 23 24 29	(Sun et al., 2020a) (Sun et al., 2020a) (Liu et al., 2020)32201950 (Pal et al., 2015) (Yang et al., 2015) (Ha et al., 2021) (Haudenschild et al., 2019) (Corciulo et al., 2017) (Makki and Haqqi, 2016) (Brooks, 2013) (Sun et al., 2018)
Atherosclerosis	Geldanamycin; HU-211; Chalcone; DHMEQ; Daidzein; Alvocidib; Rutin; Equol; 3-methyladenine; Indomethacin; PDBu; Celecoxib	2; 6; 11; 13; 14; 15; 17; 19; 23; 25; 26; 28	(Abeyrathna and Su, 2015; Anthony et al., 1998; Burleigh et al., 2005; Chiba et al., 2006; Dai et al., 2016; Kim et al., 2015; Klein and Newton, 2007; Li et al., 2018b; Mach et al., 2008; Madrigal-Matute et al., 2010; Pang et al., 2019; Papageorgiou et al., 2016; Rostam et al., 2018; Zhang et al., 2016b)
Parkinson's disease	Coumarin; Niacin; Andrographolide; 2-Arachidonoylglycerol; Arbutin; Cucurbitacin-E;	1; 2; 3; 5; 7; 9;	(Stefanachi et al., 2018) (Chong et al., 2021; Fukushima, 2005; Giri et al., 2019) (Geng et al., 2019) (Mounsey et al., 2015)

	Papaverine; Sphingosine-1-phosphate; Chloramphenicol; Colchicine; Melatonin; Caffeine; Norepinephrine	10; 12; 17; 20; 24; 29; 30	(Ding et al., 2020) (Arel-Dubeau et al., 2014) (Leem et al., 2020) (Motyl and Strosznajder, 2018) (Han et al., 2019) (Salama and Arias-Carrión, 2011; Salama et al., 2012) (Mayo et al., 2005) (Hong et al., 2020; Postuma et al., 2012) (Espay et al., 2014)
Type2 diabetes	Quercetin; Chalcone; Digoxin; Celastrol; DHMEQ; Myristic acid	15; 16; 17; 18; 22; 30	(Dhanya et al., 2017) (Rocha et al., 2020) (Spigset and Mjörndal, 1999) (Han et al., 2016) (Saisho et al., 2008) (Takato et al., 2017)
Alzheimer's disease	Imatinib; Sphingosine-1-phosphate ; Dopamine; Andrographolide ; Brucine; Arbutin; Niacin; Papaverine; Quinine Hydrochloride	2 ; 3 7 8 9 13 16 17 18	(Kumar et al., 2019) (Chua et al., 2020; Czubowicz et al., 2019; He et al., 2021) (Pan et al., 2019) (Rivera et al., 2016) (Dastan et al., 2019) (Gharagozloo et al., 1999) (Morris et al., 2004) (Ahmadzadeh, 2014; Tamada et al., 2019) (Eyal, 2018; Schiffman et al., 1990)
Sarcopenia	Succinate; Colchicine; Niacin;	2; 11; 19;	(Fogarty et al., 2020) (Huang et al., 2020)

	Adenosine	25	(Pirinen et al., 2020) (Gnad et al., 2020)

Table 4. 3 The top 30 predictions of ChemPert with literature evidences for aged-related diseases.

Disease	Perturbagen	Rank	Reference
Influenza A	Geldanamycin	1	(Wang et al., 2017)
	Mitomycin	2	(Nayak and Rasmussen, 1966)
	Chalcone	7	(Dao et al., 2011)
	Loratadine	8	(Trukhan et al., 2016)
	Alvocidib	11	(Wang et al., 2012)
	Celastrol	12	(Khalili et al., 2018)
	Daidzein	19	(Chung et al., 2015)
	Thapsigargin	21	(Goulding et al., 2020)
	Oroxylin a	24	(Jin et al., 2018)
	15D-PGJ2	25	(Huang et al., 2019)
	Usnic-acid	27	(Shtro et al., 2015)
	Gentamicin	29	(Sun et al., 2020b)
COVID19	Geldanamycin	1;	(Barh et al., 2020)
	Mitoxantrone	3;	(Lokhande et al., 2020) (Hamoda et al., 2021) (Vijayakumar et al., 2020) (Hou et al., 2021) (Berretta et al., 2020) (Kapoor et al., 2020) (Xing et al., 2020) (Caruso et al., 2020) (Kandwal and Fayne, 2020) (Al-Beltagi et al., 2021) (Nguyen et al., 2021) (Baghaki et al., 2020) (Gao et al., 2021) (Shahzad and Willcox, 2020) (Talasaz et al., 2021) (Marciniec et al., 2020) (Angelopoulou et al., 2020)
	Ilimaquinone	4;	
	Chalcone	7;	
	Loratadine	8;	
	Equol	9;	
	Cucurbitacin-i	10;	
	Alvocidib	11;	
	Celastrol	12;	
	Olomoucine	15;	
	Thapsigargin	16;	
	Daidzein	17	
	Celecoxib	18	
	Oroxylin a	20	
	15D-PGJ2	22	
	MetoproloL	25	
	Moxifloxacin	28	
	Imiquimod	30	

Epstein-Barrvirus	Usnic-acid Alsterpaullone Staurosporine lALW-II-41-27	7; 16; 20; 30	(Shtro et al., 2015) (Watanabe et al., 2020) (Yee et al., 2011) (Goswami et al., 2012) (Cao et al., 2021)

Table 4. 4 The top 30 predictions of ChemPert with literature evidences for infectious diseases.

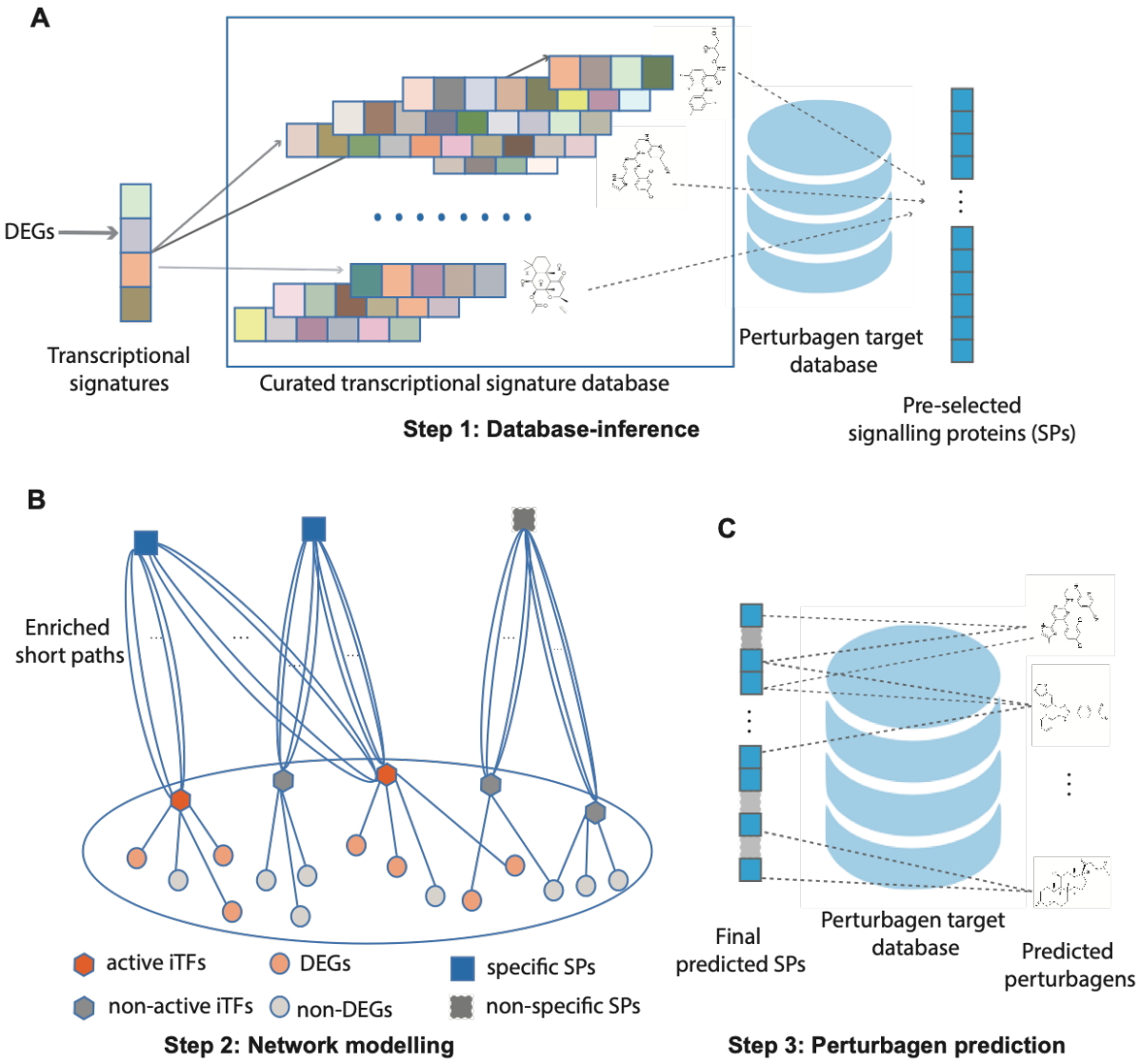

Figure 4. 4 Schematic outline of ChemPert.

A, Pre-selection of candidate signalling proteins from database. First, the perturbagens from the database whose transcriptional signatures are similar with query transcriptional signatures from DEGs are identified. Then, the signalling protein targets of the top similar perturbagens are identified and ranked based on similarity score. The top ranked signalling proteins are selected as candidates for network modelling.
B, Network-based modelling to filter out signalling candidates that are not specifically targeting the set of active iTFs. Whether the signalling protein targets the iTF or not is determined by gene set enrichment analysis for short paths. The iTFs that are enriched for DEGs are defined as active iTFs.
C, Identification of perturbagens which target the predicted signalling proteins.

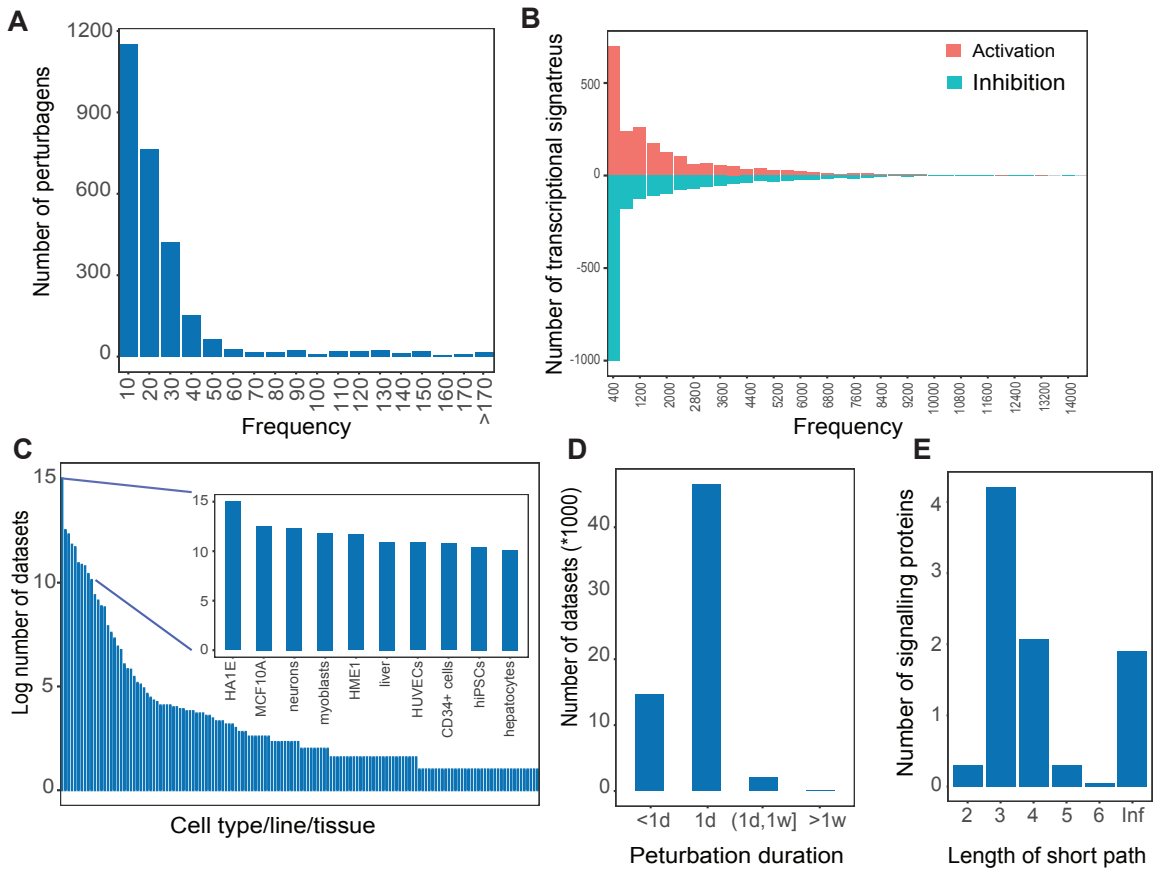

Figure 4. 5 The components of normal cell chemical perturbation database.

A, Frequency of perturbagens in database. X-axis represents the frequency of perturbagen, and y-axis means the number of perturbagen with corresponding frequency.
B, Frequency of transcriptional signatures in database, including inhibited and activated ones. X-axis represents the frequency of signature and y-axis means the number of signatures with corresponding frequency.
C, Distribution of datasets for cell types/lines/tissues in the database. Y-axis means log2(number+1) for each cell type/line/tissue.
D, Distribution of datasets for different perturbation durations.
E, Distribution of short path length for signalling proteins.

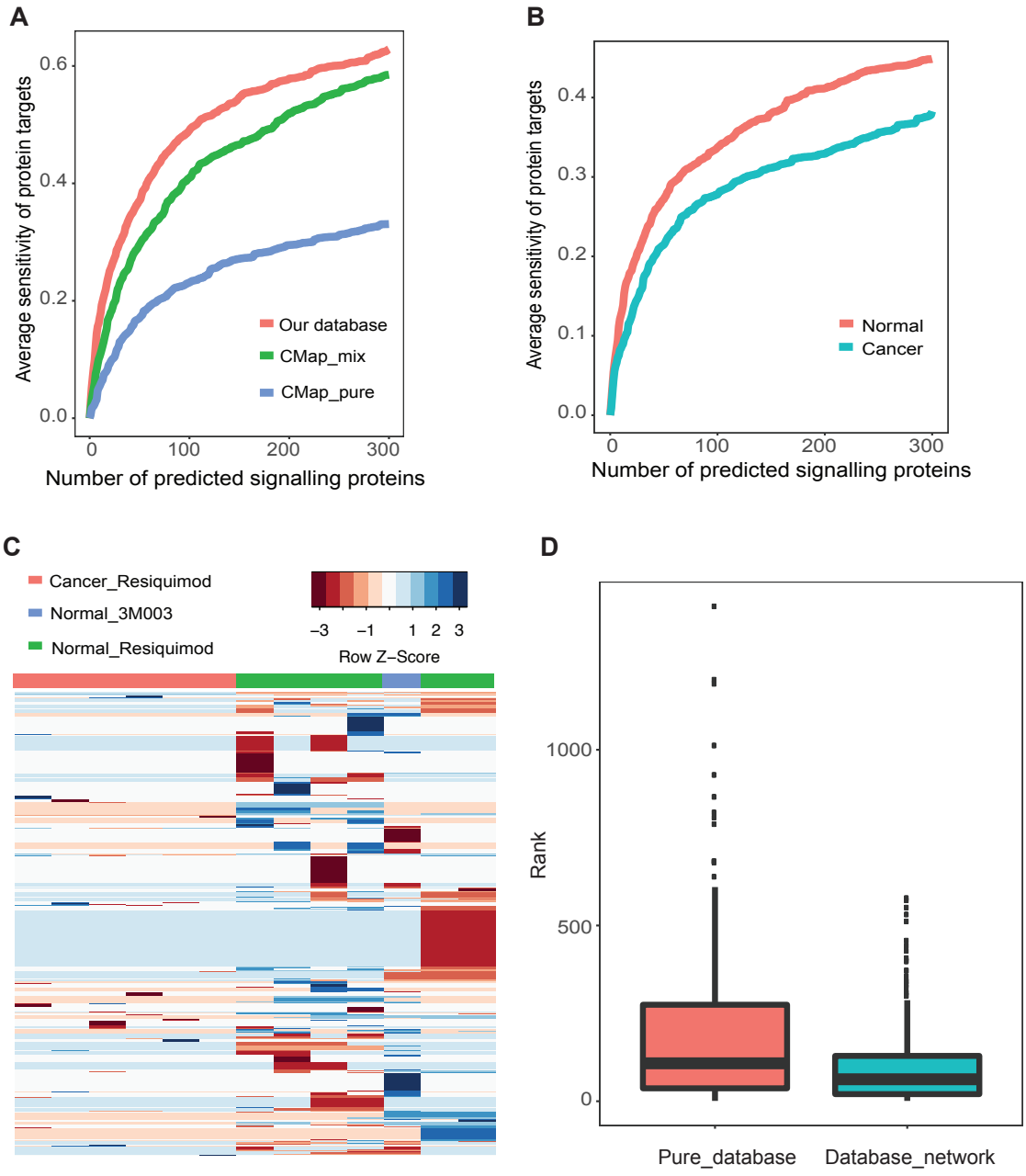

Figure 4. 6 Evaluation of ChemPert.

A, Average sensitivity across benchmarking datasets at different thresholds for predicted signalling proteins using different perturbation databases.
B, Average sensitivity across benchmarking datasets at different thresholds for predicted signalling proteins by using two databases with same perturbagens solely consisting of normal or cancer cells (subsets of our database and CMap).
C, Clustering of transcriptional signatures among resiquimod perturbation of cancer cells, 3M003 and resiquimod perturbation of normal cells.
D, Rank of true positive signalling proteins predicted by ChemPert using solely database inference or integration database with network model.

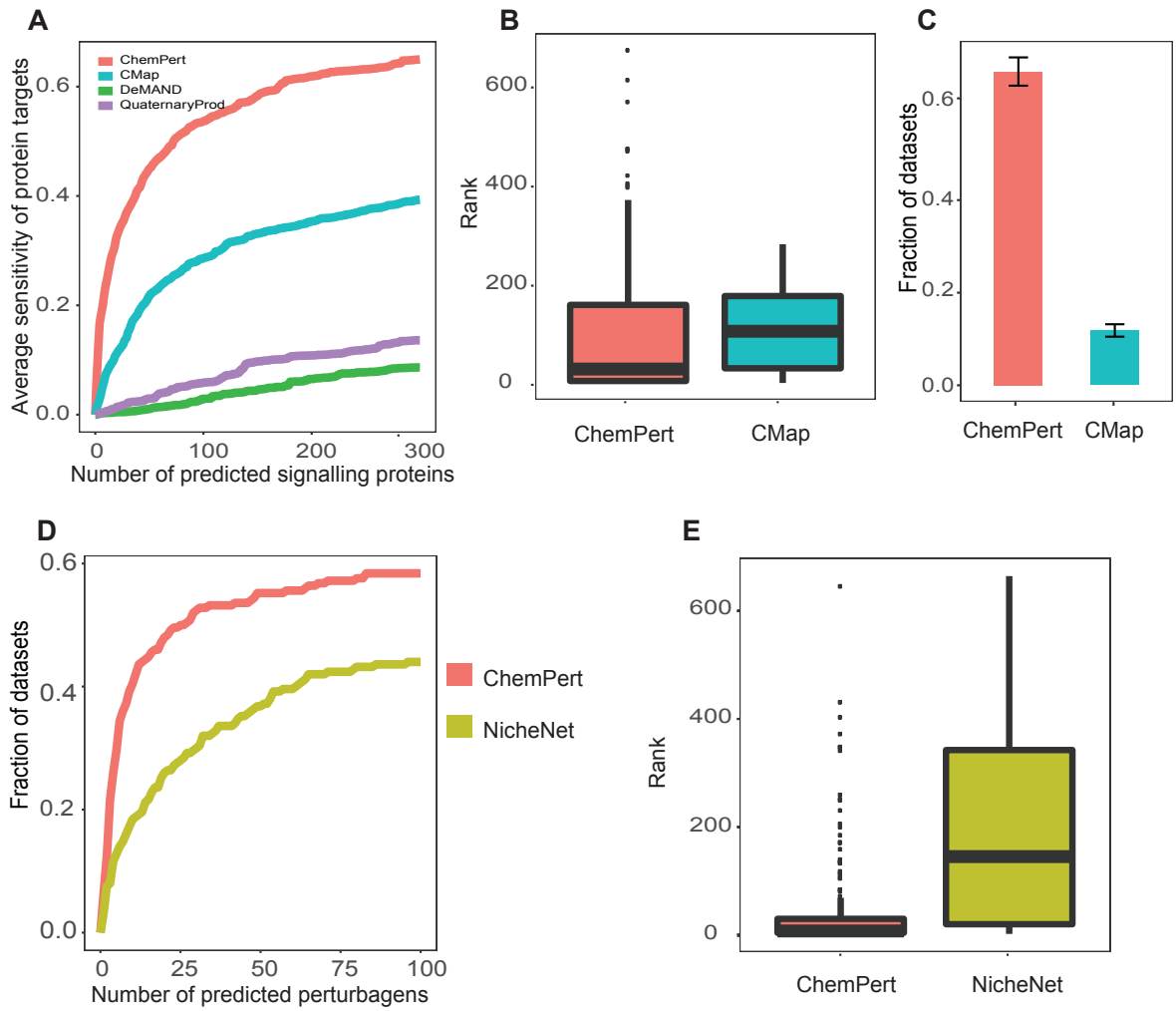

Figure 4. 7 Comparison with existing computational methods.

A, Average sensitivity across benchmarking datasets at different thresholds for predicted signalling proteins using different methods.
B, Rank of experimentally used perturbagens in benchmarking datasets among predicted perturbagens obtained by ChemPert or CMap.
C, Fraction of benchmarking datasets whose experimentally used perturbagens were predicted by ChemPert or CMap.
D, Fraction of benchmarking datasets whose experimentally used perturbagens were predicted by ChemPert or CMap at different thresholds.
E, Rank of experimentally used ligands in benchmarking datasets among predicted perturbagens obtained by ChemPert or NicheNet.

Code availability

ChemPert was implemented in R and is available from Gitlab (https://git-r3lab.uni.lu/menglin.zheng/chempert).

References

Abeyrathna, P., and Su, Y. (2015). The critical role of Akt in cardiovascular function. Vascul Pharmacol *74*, 38-48.

Ahmadzadeh, A. (2014). Papaverine increases human serum albumin glycation. J Biol Phys *40*, 97-107.

Al-Beltagi, S., Preda, C.A., Goulding, L.V., James, J., Pu, J., Skinner, P., Jiang, Z., Wang, B.L., Yang, J., Banyard, A.C.*, et al.* (2021). Thapsigargin Is a Broad-Spectrum Inhibitor of Major Human Respiratory Viruses: Coronavirus, Respiratory Syncytial Virus and Influenza A Virus. Viruses *13*.

Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet *48*, 838-847.

Angelopoulou, A., Alexandris, N., Konstantinou, E., Mesiakaris, K., Zanidis, C., Farsalinos, K., and Poulas, K. (2020). Imiquimod - A toll like receptor 7 agonist - Is an ideal option for management of COVID 19. Environ Res *188*, 109858.

Anthony, M.S., Clarkson, T.B., and Williams, J.K. (1998). Effects of soy isoflavones on atherosclerosis: potential mechanisms. Am J Clin Nutr *68*, 1390s-1393s.

Arel-Dubeau, A.M., Longpré, F., Bournival, J., Tremblay, C., Demers-Lamarche, J., Haskova, P., Attard, E., Germain, M., and Martinoli, M.G. (2014). Cucurbitacin E has neuroprotective properties and autophagic modulating activities on dopaminergic neurons. Oxid Med Cell Longev *2014*, 425496.

Baghaki, S., Yalcin, C.E., Baghaki, H.S., Aydin, S.Y., Daghan, B., and Yavuz, E. (2020). COX2 inhibition in the treatment of COVID-19: Review of literature to propose repositioning of celecoxib for randomized controlled studies. Int J Infect Dis *101*, 29-32.

Barh, D., Tiwari, S., Weener, M.E., Azevedo, V., Góes-Neto, A., Gromiha, M.M., and Ghosh, P. (2020). Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19. Comput Biol Med *126*, 104051.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res *41*, D991-995.

Berretta, A.A., Silveira, M.A.D., Cóndor Capcha, J.M., and De Jong, D. (2020). Propolis and its potential against SARS-CoV-2 infection mechanisms and COVID-19 disease: Running title: Propolis against SARS-CoV-2 infection and COVID-19. Biomed Pharmacother *131*, 110622.

Brooks, M.B. (2013). Erlotinib and gefitinib, epidermal growth factor receptor kinase inhibitors, may treat non-cancer-related tumor necrosis factor-α mediated inflammatory diseases. Oncologist *18*, e3-5.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods *17*, 159-162.

Burleigh, M.E., Babaev, V.R., Patel, M.B., Crews, B.C., Remmel, R.P., Morrow, J.D., Oates, J.A., Marnett, L.J., Fazio, S., and Linton, M.F. (2005). Inhibition of cyclooxygenase with indomethacin phenethylamide reduces atherosclerosis in apoE-null mice. Biochem Pharmacol *70*, 334-342.

Cao, Y., Xie, L., Shi, F., Tang, M., Li, Y., Hu, J., Zhao, L., Zhao, L., Yu, X., Luo, X.*, et al.* (2021). Targeting the signaling in Epstein-Barr virus-associated diseases: mechanism, regulation, and clinical study. Signal Transduct Target Ther *6*, 15.

Caruso, F., Singh, M., Belli, S., Berinato, M., and Rossi, M. (2020). Interrelated Mechanism by Which the Methide Quinone Celastrol, Obtained from the Roots of Tripterygium wilfordii, Inhibits Main Protease 3CL(pro) of COVID-19 and Acts as Superoxide Radical Scavenger. Int J Mol Sci *21*.

Chen, J., Sathiyamoorthy, K., Zhang, X., Schaller, S., Perez White, B.E., Jardetzky, T.S., and Longnecker, R. (2018). Ephrin receptor A2 is a functional entry receptor for Epstein-Barr virus. Nat Microbiol *3*, 172-180.
Chiba, T., Kondo, Y., Shinozaki, S., Kaneko, E., Ishigami, A., Maruyama, N., Umezawa, K., and Shimokado, K. (2006). A selective NFkappaB inhibitor, DHMEQ, reduced atherosclerosis in ApoE-deficient mice. J Atheroscler Thromb *13*, 308-313.
Chong, R., Wakade, C., Seamon, M., Giri, B., Morgan, J., and Purohit, S. (2021). Niacin Enhancement for Parkinson's Disease: An Effectiveness Trial. Front Aging Neurosci *13*, 667032.
Chua, X.Y., Chai, Y.L., Chew, W.S., Chong, J.R., Ang, H.L., Xiang, P., Camara, K., Howell, A.R., Torta, F., Wenk, M.R.*, et al.* (2020). Immunomodulatory sphingosine-1-phosphates as plasma biomarkers of Alzheimer's disease and vascular cognitive impairment. Alzheimers Res Ther *12*, 122.
Chung, S.T., Huang, Y.T., Hsiung, H.Y., Huang, W.H., Yao, C.W., and Lee, A.R. (2015). Novel daidzein analogs and their in vitro anti-influenza activities. Chem Biodivers *12*, 685-696.
Corciulo, C., Lendhey, M., Wilder, T., Schoen, H., Cornelissen, A.S., Chang, G., Kennedy, O.D., and Cronstein, B.N. (2017). Endogenous adenosine maintains cartilage homeostasis and exogenous adenosine inhibits osteoarthritis progression. Nat Commun *8*, 15019.
Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M.*, et al.* (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med *23*, 405-408.
Czubowicz, K., Jęśko, H., Wencel, P., Lukiw, W.J., and Strosznajder, R.P. (2019). The Role of Ceramide and Sphingosine-1-Phosphate in Alzheimer's Disease and Other Neurodegenerative Disorders. Mol Neurobiol *56*, 5436-5455.
Dai, S., Wang, B., Li, W., Wang, L., Song, X., Guo, C., Li, Y., Liu, F., Zhu, F., Wang, Q.*, et al.* (2016). Systemic application of 3-methyladenine markedly inhibited atherosclerotic lesion in ApoE(-/-) mice by modulating autophagy, foam cell formation and immune-negative molecules. Cell Death Dis *7*, e2498.
Dao, T.T., Nguyen, P.H., Lee, H.S., Kim, E., Park, J., Lim, S.I., and Oh, W.K. (2011). Chalcones as novel influenza A (H1N1) neuraminidase inhibitors from Glycyrrhiza inflata. Bioorg Med Chem Lett *21*, 294-298.
Dastan, Z., Pouramir, M., Ghasemi-Kasman, M., Ghasemzadeh, Z., Dadgar, M., Gol, M., Ashrafpour, M., Pourghasem, M., Moghadamnia, A.A., and Khafri, S. (2019). Arbutin reduces cognitive deficit and oxidative stress in animal model of Alzheimer's disease. Int J Neurosci *129*, 1145-1153.
Dhanya, R., Arya, A.D., Nisha, P., and Jayamurthy, P. (2017). Quercetin, a Lead Compound against Type 2 Diabetes Ameliorates Glucose Uptake via AMPK Pathway in Skeletal Muscle Cell Line. Front Pharmacol *8*, 336.
Ding, Y., Kong, D., Zhou, T., Yang, N.D., Xin, C., Xu, J., Wang, Q., Zhang, H., Wu, Q., Lu, X.*, et al.* (2020). α-Arbutin Protects Against Parkinson's Disease-Associated Mitochondrial Dysfunction In Vitro and In Vivo. Neuromolecular Med *22*, 56-67.
Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184-1191.
Eder, J., Sedrani, R., and Wiesmann, C. (2014). The discovery of first-in-class drugs: origins and evolution. Nat Rev Drug Discov *13*, 577-587.
Espay, A.J., LeWitt, P.A., and Kaufmann, H. (2014). Norepinephrine deficiency in Parkinson's disease: the case for noradrenergic enhancement. Mov Disord *29*, 1710-1719.
Eyal, S. (2018). The Fever Tree: from Malaria to Neurological Diseases. Toxins (Basel) *10*.

Fakhry, C.T., Choudhary, P., Gutteridge, A., Sidders, B., Chen, P., Ziemek, D., and Zarringhalam, K. (2016). Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics *17*, 318.
Fogarty, M.J., Marin Mathieu, N., Mantilla, C.B., and Sieck, G.C. (2020). Aging reduces succinate dehydrogenase activity in rat type IIx/IIb diaphragm muscle fibers. J Appl Physiol (1985) *128*, 70-77.
Fukushima, T. (2005). Niacin metabolism and Parkinson's disease. Environ Health Prev Med *10*, 3-8.
Gao, J., Ding, Y., Wang, Y., Liang, P., Zhang, L., and Liu, R. (2021). Oroxylin A is a severe acute respiratory syndrome coronavirus 2-spiked pseudotyped virus blocker obtained from Radix Scutellariae using angiotensin-converting enzyme II/cell membrane chromatography. Phytother Res *35*, 3194-3204.
Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res *29*, 1363-1375.
Geng, J., Liu, W., Gao, J., Jiang, C., Fan, T., Sun, Y., Qin, Z.H., Xu, Q., Guo, W., and Gao, J. (2019). Andrographolide alleviates Parkinsonism in MPTP-PD mice via targeting mitochondrial fission mediated by dynamin-related protein 1. Br J Pharmacol *176*, 4574-4591.
Gharagozloo, P., Lazareno, S., Popham, A., and Birdsall, N.J. (1999). Allosteric interactions of quaternary strychnine and brucine derivatives with muscarinic acetylcholine receptors. J Med Chem *42*, 438-445.
Giri, B., Belanger, K., Seamon, M., Bradley, E., Purohit, S., Chong, R., Morgan, J.C., Baban, B., and Wakade, C. (2019). Niacin Ameliorates Neuro-Inflammation in Parkinson's Disease via GPR109A. Int J Mol Sci *20*.
Gnad, T., Navarro, G., Lahesmaa, M., Reverte-Salisa, L., Copperi, F., Cordomi, A., Naumann, J., Hochhäuser, A., Haufs-Brusberg, S., Wenzel, D.*, et al.* (2020). Adenosine/A2B Receptor Signaling Ameliorates the Effects of Aging and Counteracts Obesity. Cell Metab *32*, 56-70.e57.
Goswami, R., Gershburg, S., Satorius, A., and Gershburg, E. (2012). Protein kinase inhibitors that inhibit induction of lytic program and replication of Epstein-Barr virus. Antiviral Res *96*, 296-304.
Goulding, L.V., Yang, J., Jiang, Z., Zhang, H., Lea, D., Emes, R.D., Dottorini, T., Pu, J., Liu, J., and Chang, K.C. (2020). Thapsigargin at Non-Cytotoxic Levels Induces a Potent Host Antiviral Response that Blocks Influenza A Virus Replication. Viruses *12*.
Ha, Y.J., Choi, Y.S., Oh, Y.R., Kang, E.H., Khang, G., Park, Y.B., and Lee, Y.J. (2021). Fucoxanthin Suppresses Osteoclastogenesis via Modulation of MAP Kinase and Nrf2 Signaling. Mar Drugs *19*.
Hamoda, A.M., Fayed, B., Ashmawy, N.S., El-Shorbagi, A.A., Hamdy, R., and Soliman, S.S.M. (2021). Marine Sponge is a Promising Natural Source of Anti-SARS-CoV-2 Scaffold. Front Pharmacol *12*, 666664.
Han, J., Kim, S.J., Ryu, M.J., Jang, Y., Lee, M.J., Ju, X., Lee, Y.L., Cui, J., Shong, M., Heo, J.Y.*, et al.* (2019). Chloramphenicol Mitigates Oxidative Stress by Inhibiting Translation of Mitochondrial Complex I in Dopaminergic Neurons of Toxin-Induced Parkinson's Disease Model. Oxid Med Cell Longev *2019*, 4174803.
Han, L.P., Li, C.J., Sun, B., Xie, Y., Guan, Y., Ma, Z.J., and Chen, L.M. (2016). Protective Effects of Celastrol on Diabetic Liver Injury via TLR4/MyD88/NF-κB Signaling Pathway in Type 2 Diabetic Rats. J Diabetes Res *2016*, 2641248.

Haudenschild, D.R., Carlson, A.K., Zignego, D.L., Yik, J.H.N., Hilmer, J.K., and June, R.K. (2019). Inhibition of early response genes prevents changes in global joint metabolomic profiles in mouse post-traumatic osteoarthritis. Osteoarthritis Cartilage *27*, 504-512.
He, Q., Ding, G., Zhang, M., Nie, P., Yang, J., Liang, D., Bo, J., Zhang, Y., and Liu, Y. (2021). Trends in the Use of Sphingosine 1 Phosphate in Age-Related Diseases: A Scientometric Research Study (1992-2020). J Diabetes Res *2021*, 4932974.
Hodos, R., Zhang, P., Lee, H.C., Duan, Q., Wang, Z., Clark, N.R., Ma'ayan, A., Wang, F., Kidd, B., Hu, J.*, et al.* (2018). Cell-specific prediction and application of drug-induced gene expression profiles. Pac Symp Biocomput *23*, 32-43.
Hong, C.T., Chan, L., and Bai, C.H. (2020). The Effect of Caffeine on the Risk and Progression of Parkinson's Disease: A Meta-Analysis. Nutrients *12*.
Hou, Y., Ge, S., Li, X., Wang, C., He, H., and He, L. (2021). Testing of the inhibitory effects of loratadine and desloratadine on SARS-CoV-2 spike pseudotyped virus viropexis. Chem Biol Interact *338*, 109420.
Huang, D.D., Yan, X.L., Fan, S.D., Chen, X.Y., Yan, J.Y., Dong, Q.T., Chen, W.Z., Liu, N.X., Chen, X.L., and Yu, Z. (2020). Nrf2 deficiency promotes the increasing trend of autophagy during aging in skeletal muscle: a potential mechanism for the development of sarcopenia. Aging (Albany NY) *12*, 5977-5991.
Huang, S., Jiang, L., Cheon, I.S., and Sun, J. (2019). Targeting Peroxisome Proliferator-Activated Receptor-Gamma Decreases Host Mortality After Influenza Infection in Obese Mice. Viral Immunol *32*, 161-169.
Indari, O., Jakhmola, S., Manivannan, E., and Jha, H.C. (2021). An Update on Antiviral Therapy Against SARS-CoV-2: How Far Have We Come? Front Pharmacol *12*, 632677.
Jin, J., Chen, S., Wang, D., Chen, Y., Wang, Y., Guo, M., Zhou, C., and Dou, J. (2018). Oroxylin A suppresses influenza A virus replication correlating with neuraminidase inhibition and induction of IFNs. Biomed Pharmacother *97*, 385-394.
Kandwal, S., and Fayne, D. (2020). Repurposing drugs for treatment of SARS-CoV-2 infection: computational design insights into mechanisms of action. J Biomol Struct Dyn, 1-15.
Kapoor, N., Ghorai, S.M., Kushwaha, P.K., Shukla, R., Aggarwal, C., and Bandichhor, R. (2020). Plausible mechanisms explaining the role of cucurbitacins as potential therapeutic drugs against coronavirus 2019. Inform Med Unlocked *21*, 100484.
Khalili, N., Karimi, A., Moradi, M.T., and Shirzad, H. (2018). In vitro immunomodulatory activity of celastrol against influenza A virus infection. Immunopharmacol Immunotoxicol *40*, 250-255.
Kim, N.Y., Kohn, J.C., Huynh, J., Carey, S.P., Mason, B.N., Vouyouka, A.G., and Reinhart-King, C.A. (2015). Biophysical induction of vascular smooth muscle cell podosomes. PLoS One *10*, e0119008.
Klein, T.W., and Newton, C.A. (2007). Therapeutic potential of cannabinoid-based drugs. Adv Exp Med Biol *601*, 395-413.
Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T.*, et al.* (2015). ArrayExpress update--simplifying data submissions. Nucleic Acids Res *43*, D1113-1116.
Krämer, A., Green, J., Pollard, J., Jr., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics *30*, 523-530.
Kumar, M., Kulshrestha, R., Singh, N., and Jaggi, A.S. (2019). Expanding spectrum of anticancer drug, imatinib, in the disorders affecting brain and spinal cord. Pharmacol Res *143*, 86-96.
Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N.*, et al.* (2006). The Connectivity Map: using gene-

expression signatures to connect small molecules, genes, and disease. Science *313*, 1929-1935.
Leem, Y.H., Park, J.S., Park, J.E., Kim, D.Y., Kang, J.L., and Kim, H.S. (2020). Papaverine inhibits α-synuclein aggregation by modulating neuroinflammation and matrix metalloproteinase-3 expression in the subacute MPTP/P mouse model of Parkinson's disease. Biomed Pharmacother *130*, 110576.
Li, Y., Qin, R., Yan, H., Wang, F., Huang, S., Zhang, Y., Zhong, M., Zhang, W., and Wang, Z. (2018). Inhibition of vascular smooth muscle cells premature senescence with rutin attenuates and stabilizes diabetic atherosclerosis. J Nutr Biochem *51*, 91-98.
Liu, D.D., Zhang, B.L., Yang, J.B., and Zhou, K. (2020). Celastrol ameliorates endoplasmic stress-mediated apoptosis of osteoarthritis via regulating ATF-6/CHOP signalling pathway. J Pharm Pharmacol *72*, 826-835.
Lokhande, K.B., Doiphode, S., Vyas, R., and Swamy, K.V. (2020). Molecular docking and simulation studies on SARS-CoV-2 M(pro) reveals Mitoxantrone, Leucovorin, Birinapant, and Dynasore as potent drugs against COVID-19. J Biomol Struct Dyn, 1-12.
Mach, F., Montecucco, F., and Steffens, S. (2008). Cannabinoid receptors in acute and chronic complications of atherosclerosis. Br J Pharmacol *153*, 290-298.
Madrigal-Matute, J., López-Franco, O., Blanco-Colio, L.M., Muñoz-García, B., Ramos-Mozo, P., Ortega, L., Egido, J., and Martín-Ventura, J.L. (2010). Heat shock protein 90 inhibitors attenuate inflammatory responses in atherosclerosis. Cardiovasc Res *86*, 330-337.
Makki, M.S., and Haqqi, T.M. (2016). Histone Deacetylase Inhibitor Vorinostat (SAHA) Suppresses IL-1β-Induced Matrix Metallopeptidase-13 Expression by Inhibiting IL-6 in Osteoarthritis Chondrocyte. Am J Pathol *186*, 2701-2708.
Marciniec, K., Beberok, A., Pęcak, P., Boryczka, S., and Wrześniok, D. (2020). Ciprofloxacin and moxifloxacin could interact with SARS-CoV-2 protease: preliminary in silico analysis. Pharmacol Rep *72*, 1553-1561.
Mayo, J.C., Sainz, R.M., Tan, D.X., Antolín, I., Rodríguez, C., and Reiter, R.J. (2005). Melatonin and Parkinson's disease. Endocrine *27*, 169-178.
Morris, M.C., Evans, D.A., Bienias, J.L., Scherr, P.A., Tangney, C.C., Hebert, L.E., Bennett, D.A., Wilson, R.S., and Aggarwal, N. (2004). Dietary niacin and the risk of incident Alzheimer's disease and of cognitive decline. J Neurol Neurosurg Psychiatry *75*, 1093-1099.
Motyl, J., and Strosznajder, J.B. (2018). Sphingosine kinase 1/sphingosine-1-phosphate receptors dependent signalling in neurodegenerative diseases. The promising target for neuroprotection in Parkinson's disease. Pharmacol Rep *70*, 1010-1014.
Mounsey, R.B., Mustafa, S., Robinson, L., Ross, R.A., Riedel, G., Pertwee, R.G., and Teismann, P. (2015). Increasing levels of the endocannabinoid 2-AG is neuroprotective in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine mouse model of Parkinson's disease. Exp Neurol *273*, 36-44.
Nayak, D.P., and Rasmussen, A.F., Jr. (1966). Influence of mitomycin C on the replication of influenza viruses. Virology *30*, 673-683.
Nguyen, T.T.H., Jung, J.H., Kim, M.K., Lim, S., Choi, J.M., Chung, B., Kim, D.W., and Kim, D. (2021). The Inhibitory Effects of Plant Derivate Polyphenols on the Main Protease of SARS Coronavirus 2 and Their Structure-Activity Relationship. Molecules *26*.
Pal, B., Endisha, H., Zhang, Y., and Kapoor, M. (2015). mTOR: a potential therapeutic target in osteoarthritis? Drugs R D *15*, 27-36.
Pan, X., Kaminga, A.C., Wen, S.W., Wu, X., Acheampong, K., and Liu, A. (2019). Dopamine and Dopamine Receptors in Alzheimer's Disease: A Systematic Review and Network Meta-Analysis. Front Aging Neurosci *11*, 175.

Pang, Y., Gan, L., Wang, X., Su, Q., Liang, C., and He, P. (2019). Celecoxib aggravates atherogenesis and upregulates leukotrienes in ApoE(-/-) mice and lipopolysaccharide-stimulated RAW264.7 macrophages. Atherosclerosis *284*, 50-58.
Papageorgiou, N., Zacharia, E., Briasoulis, A., Charakida, M., and Tousoulis, D. (2016). Celecoxib for the treatment of atherosclerosis. Expert Opin Investig Drugs *25*, 619-633.
Parikh, J.R., Klinger, B., Xia, Y., Marto, J.A., and Blüthgen, N. (2010). Discovering causal signaling pathways through gene-expression patterns. Nucleic Acids Res *38*, W109-117.
Pirinen, E., Auranen, M., Khan, N.A., Brilhante, V., Urho, N., Pessia, A., Hakkarainen, A., Kuula, J., Heinonen, U., Schmidt, M.S.*, et al.* (2020). Niacin Cures Systemic NAD(+) Deficiency and Improves Muscle Performance in Adult-Onset Mitochondrial Myopathy. Cell Metab *31*, 1078-1090.e1075.
Postuma, R.B., Lang, A.E., Munhoz, R.P., Charland, K., Pelletier, A., Moscovich, M., Filla, L., Zanatta, D., Rios Romenets, S., Altman, R.*, et al.* (2012). Caffeine for treatment of Parkinson disease: a randomized controlled trial. Neurology *79*, 651-658.
Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B.*, et al.* (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun *6*, 7866.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.
Rivera, D.S., Lindsay, C., Codocedo, J.F., Morel, I., Pinto, C., Cisternas, P., Bozinovic, F., and Inestrosa, N.C. (2016). Andrographolide recovers cognitive impairment in a natural model of Alzheimer's disease (Octodon degus). Neurobiol Aging *46*, 204-220.
Rocha, S., Ribeiro, D., Fernandes, E., and Freitas, M. (2020). A Systematic Review on Anti-diabetic Properties of Chalcones. Curr Med Chem *27*, 2257-2321.
Rostam, M.A., Shajimoon, A., Kamato, D., Mitra, P., Piva, T.J., Getachew, R., Cao, Y., Zheng, W., Osman, N., and Little, P.J. (2018). Flavopiridol Inhibits TGF-β-Stimulated Biglycan Synthesis by Blocking Linker Region Phosphorylation and Nuclear Translocation of Smad2. J Pharmacol Exp Ther *365*, 156-164.
Saisho, Y., Hirose, H., Horimai, C., Miyashita, K., Takei, I., Umezawa, K., and Itoh, H. (2008). Effects of DHMEQ, a novel nuclear factor-kappab inhibitor, on beta cell dysfunction in INS-1 cells. Endocr J *55*, 433-438.
Salama, M., and Arias-Carrión, O. (2011). Colchicine as a promising drug for Parkinson's disease. Med Hypotheses *76*, 150.
Salama, M., Ellaithy, A., Helmy, B., El-Gamal, M., Tantawy, D., Mohamed, M., Sheashaa, H., Sobh, M., and Arias-Carrión, O. (2012). Colchicine protects dopaminergic neurons in a rat model of Parkinson's disease. CNS Neurol Disord Drug Targets *11*, 836-843.
Schiffman, S.S., Clark, C.M., and Warwick, Z.S. (1990). Gustatory and olfactory dysfunction in dementia: not specific to Alzheimer's disease. Neurobiol Aging *11*, 597-600.
Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun *9*, 20.
Sergushichev, A.A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv, 060012.
Shahzad, S., and Willcox, M. (2020). Immuno-pathogenesis of nCOVID-19 and a possible host-directed therapy including anti-inflammatory and anti-viral prostaglandin (PG J(2)) for effective treatment and reduction in the death toll. Med Hypotheses *143*, 110080.
Sharma, S., and Petsalaki, E. (2019). Large-scale datasets uncovering cell signalling networks in cancer: context matters. Curr Opin Genet Dev *54*, 118-124.

Shtro, A.A., Zarubaev, V.V., Luzina, O.A., Sokolov, D.N., and Salakhutdinov, N.F. (2015). Derivatives of usnic acid inhibit broad range of influenza viruses and protect mice from lethal influenza infection. Antivir Chem Chemother *24*, 92-98.
Spigset, O., and Mjörndal, T. (1999). Increased glucose intolerance related to digoxin treatment in patients with type 2 diabetes mellitus. J Intern Med *246*, 419-422.
Stefanachi, A., Leonetti, F., Pisani, L., Catto, M., and Carotti, A. (2018). Coumarin: A Natural, Privileged and Versatile Scaffold for Bioactive Compounds. Molecules *23*.
Strasen, J., Sarma, U., Jentsch, M., Bohn, S., Sheng, C., Horbelt, D., Knaus, P., Legewie, S., and Loewer, A. (2018). Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels. Mol Syst Biol *14*, e7733.
Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K.*, et al.* (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437-1452.e1417.
Sun, H., Wu, Y., Pan, Z., Yu, D., Chen, P., Zhang, X., Wu, H., Zhang, X., An, C., Chen, Y.*, et al.* (2018). Gefitinib for Epidermal Growth Factor Receptor Activated Osteoarthritis Subpopulation Treatment. EBioMedicine *32*, 223-233.
Sun, K., Luo, J., Guo, J., Yao, X., Jing, X., and Guo, F. (2020a). The PI3K/AKT/mTOR signaling pathway in osteoarthritis: a narrative review. Osteoarthritis Cartilage *28*, 400-409.
Sun, Y., He, Z., Li, J., Gong, S., Yuan, S., Li, T., Ning, N., Xing, L., Zhang, L., Chen, F.*, et al.* (2020b). Gentamicin Induced Microbiome Adaptations Associate With Increased BCAA Levels and Enhance Severity of Influenza Infection. Front Immunol *11*, 608895.
Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res *44*, D380-384.
Taechowisan, T., Samsawat, T., Puckdee, W., and Phutdhawong, W.S. (2020). Antiviral activity of geldanamycin and its derivatives against influenza virus. Journal of Applied Pharmaceutical Science *10*, 113-120.
Takato, T., Iwata, K., Murakami, C., Wada, Y., and Sakane, F. (2017). Chronic administration of myristic acid improves hyperglycaemia in the Nagoya-Shibata-Yasuda mouse model of congenital type 2 diabetes. Diabetologia *60*, 2076-2083.
Talasaz, A.H., Kakavand, H., Van Tassell, B., Aghakouchakzadeh, M., Sadeghipour, P., Dunn, S., and Geraiely, B. (2021). Cardiovascular Complications of COVID-19: Pharmacotherapy Perspective. Cardiovasc Drugs Ther *35*, 249-259.
Tamada, K., Nakajima, S., Ogawa, N., Inada, M., Shibasaki, H., Sato, A., Takasawa, R., Yoshimori, A., Suzuki, Y., Watanabe, N.*, et al.* (2019). Papaverine identified as an inhibitor of high mobility group box 1/receptor for advanced glycation end-products interaction suppresses high mobility group box 1-mediated inflammatory responses. Biochem Biophys Res Commun *511*, 665-670.
Trukhan, D.I., Mazurov, A.L., and Rechapova, L.A. (2016). [Acute respiratory viral infections: Topical issues of diagnosis, prevention and treatment in therapeutic practice]. Ter Arkh *88*, 76-82.
Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods *13*, 966-967.
Vijayakumar, B.G., Ramesh, D., Joji, A., Jayachandra Prakasan, J., and Kannan, T. (2020). In silico pharmacokinetic and molecular docking studies of natural flavonoids and synthetic indole chalcones against essential proteins of SARS-CoV-2. Eur J Pharmacol *886*, 173448.
Wang, C., Liu, P., Luo, J., Ding, H., Gao, Y., Sun, L., Luo, F., Liu, X., and He, H. (2017). Geldanamycin Reduces Acute Respiratory Distress Syndrome and Promotes the Survival of Mice Infected with the Highly Virulent H5N1 Influenza Virus. Front Cell Infect Microbiol *7*, 267.

Wang, S., Zhang, J., and Ye, X. (2012). [Protein kinase inhibitor flavopiridol inhibits the replication of influenza virus in vitro]. Wei Sheng Wu Xue Bao *52*, 1137-1142.
Wang, X., Cao, R., Zhang, H., Liu, J., Xu, M., Hu, H., Li, Y., Zhao, L., Li, W., Sun, X.*, et al.* (2020). The anti-influenza virus drug, arbidol is an efficient inhibitor of SARS-CoV-2 in vitro. Cell Discov *6*, 28.
Watanabe, T., Sato, Y., Masud, H., Takayama, M., Matsuda, H., Hara, Y., Yanagi, Y., Yoshida, M., Goshima, F., Murata, T.*, et al.* (2020). Antitumor activity of cyclin-dependent kinase inhibitor alsterpaullone in Epstein-Barr virus-associated lymphoproliferative disorders. Cancer Sci *111*, 279-287.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z.*, et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-d1082.
Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R.*, et al.* (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441-451.
Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol *11*, R53.
Xiao, Y., Gong, Y., Lv, Y., Lan, Y., Hu, J., Li, F., Xu, J., Bai, J., Deng, Y., Liu, L.*, et al.* (2015). Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. Sci Rep *5*, 10889.
Xing, J., Shankar, R., Drelich, A., Paithankar, S., Chekalin, E., Dexheimer, T., Chua, M.S., Rajasekaran, S., Tseng, C.K., and Chen, B. (2020). Analysis of Infected Host Gene Expression Reveals Repurposed Drug Candidates and Time-Dependent Host Response Dynamics for COVID-19. bioRxiv.
Yang, L., Zhang, J., and Wang, G. (2015). The effect of sodium hyaluronate treating knee osteoarthritis on synovial fluid interleukin -1β and clinical treatment mechanism. Pak J Pharm Sci *28*, 407-410.
Yee, J., White, R.E., Anderton, E., and Allday, M.J. (2011). Latent Epstein-Barr virus can inhibit apoptosis in B cells by blocking the induction of NOXA expression. PLoS One *6*, e28506.
Zhang, H., Li, Y., Wang, H.B., Zhang, A., Chen, M.L., Fang, Z.X., Dong, X.D., Li, S.B., Du, Y., Xiong, D.*, et al.* (2018). Ephrin receptor A2 is an epithelial cell receptor for Epstein-Barr virus entry. Nat Microbiol *3*, 1-8.
Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res *43*, D76-81.
Zhang, T., Hu, Q., Shi, L., Qin, L., Zhang, Q., and Mi, M. (2016). Equol Attenuates Atherosclerosis in Apolipoprotein E-Deficient Mice by Inhibiting Endoplasmic Reticulum Stress via Activation of Nrf2 in Endothelial Cells. PLoS One *11*, e0167020.

4.3 Manuscript 3: An integrative network model to predict signalling pathways for salamander limb regeneration

4.3.1 Preface

Stimulating the endogenous regeneration for injured tissues and organs is an appealing therapeutic strategy of regeneration medicine. Some lower vertebrates, like salamanders, can readily regenerate their lost tissues, but adult mammals cannot. Therefore, understanding and identifying the key mechanisms underlying the regeneration process in such organisms with regeneration capacity holds significant promise for achieving regenerative repair in mammals. In this study, we attempt to systematically identify signalling pathways that can induce salamander limb regeneration. Unlike the well-studies organisms, such as human, mouse and rat, the prior knowledge of salamander, including perturbation datasets and gene regulatory network, are hardly available. Therefore, the integration of perturbation database and network model is not suitable for the study of salamander. Here, we developed a de novo prediction approach that integrates signalling transduction and transcriptional regulation to model the effect of signalling perturbations on the underlying gene regulatory network to induce defined cellular decisions in regeneration process.

In this work, I pre-processed the microarray dataset. Using the time serious microarray, I constructed and analysed the regeneration-specific gene regulatory network. Dr. Gaia Zaffaroni performed the prediction of signalling pathways. We wrote the manuscript for this project together.

4.3.2 Manuscript

An integrative network model to predict signalling pathways for salamander limb regeneration

Menglin Zheng[1,6], Gaia Zaffaroni[,6], Satoshi Okawa[1,3], Dunja Knapp [2], Elly M Tanaka[2], Antonio del Sol[1,4,5,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Esch-sur-Alzette, L-4367 Belvaux, Luxembourg;

[2] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany;

[3] Integrated BioBank of Luxembourg, Dudelange L-3555, Luxembourg;

[4] CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 801 Building, 48160 Derio, Spain;

[5] IKERBASQUE, Basque Foundation for Science, Bilbao 48013, Spain;

[6] These authors contribute equally: Menglin Zheng, Gaia Zaffaroni;

[*] Correspondence: Antonio del Sol (Antonio.delSol@uni.lu)

Abstract

Understanding the regeneration mechanisms in salamanders facilitates the development of regenerative medicine to restore the functionality of damaged cells and tissues. However, due to the enormously complexity of regeneration, experimentalists have to dissect it largely by trial and error, which is lengthy and costly. Here, we developed a computational method that integrated signalling network and gene regulatory network (GRN) to predict the signalling pathways altering the GRN to induce defined cellular activities. This method was applied to predict key signalling pathways whose perturbations induce salamander limb by using time serious microarray data of connective tissue cells. The predictions recapitulated the signalling clues related to different stages of limb regeneration, including would healing, cellular migration, dedifferentiation and patterning, which aids experimentalists in understanding regeneration systematically.

Introduction

Regeneration is a process that the remaining cells are capable of restoring and re-growing the missed or injured tissues under fluctuation or injury. Different species have different levels of ability to regenerate. Only limited number of species are able to regenerate the complete lost tissues, such as planarian (Reddien and Sánchez Alvarado, 2004), arthropods (Suzuki et al., 2019), echinoderms (Wilkie, 2001) and amphibians (Brockes et al., 2001). Especially, salamander is the only tetrapod that can completely regenerate not only its functional limbs but also other tissues, such as its retina, spinal cord and heart. Therefore, it is an ideal model organism to study tissue regeneration. Regeneration takes place at the position of amputation where a wide range of cell types, such as dermis and interstitial cells, are migrated, harboured and dedifferentiated into progenitor to form blastema (Muneoka et al., 1986). The cells in blastema are further stimulated by signals from overlying ectoderm and injured nerves in niche to undergo proliferation and patterning to achieve the tissue re-development. To date, studies have demonstrated that the regeneration process is blocked completely by inhibition of some canonical signalling pathways, including FGF, retinoic acid and TGFB (Lévesque et al., 2007; Mercader et al., 2000). However, it is still unclear how the niche of an amputated limb stimulates sets of signalling pathways to trigger changes in the downstream gene regulatory network (GRN) in order to induce defined cellular activities such as the initiation of proliferation, migration, and the dedifferentiation of a mature connective cell into a blastema stem cell with embryonic limb progenitor properties.

Due to the complexity in regeneration, currently experimentalists dissect it mainly by trial and error, which is time- and resource-intensive. To our knowledge, no existing computational method is designed to aid experimentalists in addressing this challenge systematically. Here, we developed a computational method to model and predict the effect of cell-niche interactions on target cells by integrating signalling network with GRNs. We applied this method to predict signalling proteins and pathways whose perturbation could induce regeneration systematically, allowing us to understand salamander limb regeneration. In particular, the analysis considered the initial stages of the regeneration up to 14 days post amputation (dpa), comprising the response to amputation, the wound epithelium formation and subsequently the blastema formation. Gene expression data of limb-connective tissue cells from the mesenchymal lineage, expressing Prrx1, was collected at time point 1, 3, 5, 7, 10, and 14 dpa. We constructed a global regeneration-specific GRN by using the TFs that are differentially expressed and change their states between any two consecutive time points of the time-serious data. Subsequently, and each interval specific GRN was extracted. Gene Ontology (GO) enrichment analysis was

performed on each interval-specific GRN. GO enrichment showed that each GRN containing TFs in the regulation of processes related to initial response to stimuli, signal transduction, proliferation and embryonic development in accordance with expectations. For each of these intervals, the signalling molecules and pathways that were able to induce the shifting of GRN from its initial to the target state were predicted with literature evidence. 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. In addition, some novel signaling pathways and molecules that have not been reported in previous studies were also predicted, which could potentially have novel functions in axolotl limb regeneration and are primary candidates for experimental follow-up studies.

In summary, this study is not only valuable for understanding the salamander limb regeneration but shall also pave the way for a comparative analysis of regeneration across species. It is worth stressing that the proposed computational method can be applied to other systems to identify upstream signalling molecules/pathways whose perturbation can induce desired downstream GRN change to drive the cellular transition.

Method and material

Generation of time series microarray data for salamander limb regeneration

We generated time-series transcriptome data of the blastema-specific cells, deriving from the connective tissue, which involved key information on which parts of the limb to regenerate. The time-course includes time points: 0, 1, 3, 5, 7, 10 and 14 days following the amputation of upper-arm. 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 the drug Tamoxifen, a DNA-cassette is removed from the reporter-transgene allowing the expression of the red fluorescent protein Cherry. Cherry+ cells (tagged to the Prrx1 promoter, denoted as Prrx1+ 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.

Preprocessing of the microarray data

To ensure high transcript quantification, the probes with a low correlation between the replicates were excluded (Pearson's correlation <0.7). For the gene with multiple mapping probes, mean value of probes that have high correlation (Pearson's correlation >0.8) was assigned to the gene. Differential expression analysis was performed between the consecutive time points (i.e., D1 vs D0, D3 vs D1) with R package limma. The genes with absolute log fold change (lfc) larger than log2(1.5) and Benjamini and Hochberg (BH) adjusted p-value less than 0.05 were considered as differentially expressed genes (DEGs). The differential expressed TFs (DETFs) were extracted from the DEGs based on the annotation-available database AnimalTFDB 2.0 (Zhang et al., 2015). Furthermore, the Boolean state and expression probability for each gene at each time point was assigned and calculated according to the *Data booleanization and probability calculation* section.

Data booleanization and probability calculation

Gene expression data was booleanized to identify genes that are expressed or non-expressed following a data-driven, platform-independent approach. We reasoned that compared to non-DEGs, the DEGs between two cellular states can be more likely to shift either from expressed to non-expressed states or the other way around. Especially, we expected that the maximum value of a DEG across replicates was most likely to be expressed and its minimum value was most likely to be non-expressed. Therefore, the minimum and maximum expression values of each DEG were collected into two separate distributions. According to these empirical expressed and non-expressed distributions, the expression probability of a gene was calculated as the ratio:

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

where $f_{max}(x)$ is the empirical probability density function (epdf) of x in the maximum-value distribution and $f_{min}(x)$ epdf of x in the minimum-value distribution. The intersection point of the two distributions, which has the minimum misclassified value, was selected as a threshold to discretize the data (Figure 4.8A). The genes with the mean expression value across replicates larger than the threshold were assigned as Boolean state 1, otherwise state 0.

Inference of GRNs for salamander limb regeneration

To construct the regeneration-specific GRNs for salamander, in this study, five different network inference methods were applied: TIGRESS (Haury et al., 2012), CLR (Faith et al., 2007) , GENIE3 (Huynh-Thu et al., 2010), PLSNET (Guo et al., 2016) and Pearson's correlation. This is due to the fact that the integration of multiple methods has been shown to improve the overall performance of the inference (Marbach et al., 2012). These methods implemented different principles, such as mutual information, linear regression, random forest and correlation. Default parameters for each tool were used. The results of these five methods were integrated by keeping the interactions that were ranked at the top 10% of interactions in minimum 4 methods. The signs of edges were assigned based on Pearson's correlation. If the correlation value is positive, the edge is considered an activation, otherwise inhibition.

The TFs that were differential expressed and had their Boolean states changed in any consecutive time points were treated as seed TFs to infer the GRN. First, we inferred a global regeneration-specific GRN with all the seed TFs across the entire time course. Then, for each time interval, a subnetwork was extracted from the global network including only the seed TFs of the corresponding time interval. These subnetworks were further contextualized to the Booleanized gene expression profiles of each time interval by using a genetic algorithm developed by our group (Crespo and Del Sol, 2013), resulting in interval-specific GRNs. Briefly, this algorithm assumes that each cellular phenotype is a stable steady-state attractor in a Boolean network, and iteratively prunes edges to ensure the network is consistent with the Booleanized gene expression data.

To access whether the inferred networks capture the biological processes relevant to regeneration, we first performed GO analysis considering the topology on the whole network across all time intervals by using the plugin tool BiNGO of Cytoscape (version 3.5.1), where the hypergeometric test was carried out and the terms with BH adjusted p-value larger than 0.05 were selected as significantly overrepresented terms (Maere et al., 2005; Shannon et al., 2003). BiNGO gives us an overview on the distribution of GO terms across the whole process. Furthermore, to gain a deeper insight on GO enrichment in each time interval, we implemented R package TopGo using Fisher's exact test for each subnetwork (Alexa et al., 2006).

Canonical signalling pathways

In order to identify the signalling pathways and molecules altering downstream GRN to trigger the transitions of cellular states, we focused on canonical signalling pathways rather than

inferring signalling interactions derived from different sources, such as protein sequence motifs and protein-protein interactions (PPIs). This is due to the fact that the curated database knowledge is usually the most reliable (Linding et al., 2007) and the canonical signalling pathways is easier to benchmark during the validation of the method compared to the predicted signalling networks. In terms of salamander limb regeneration, this is also reasonable since currently most of the reported molecular factors for regeneration are canonical members and they will most certainly play a key role in the blastema formation. To this end, 75 canonical signalling pathways were merged in a single signalling network selected from MetaCore from Clarivate Analytics in July 2017, resulting in 2496 nodes and 6876 edges. The nodes in the network represent entities, including single proteins and protein complexes. Only edges representing signalling interactions were included in the network.

Prioritization of signalling molecules for salamander regeneration

In our study, we modelled the GRNs underlying the phenotypic differences between the two cellular states in a Boolean network. Here, each time point represents one cellular state. In order to prioritize the signalling candidates that could induce transitions between two consecutive time points, we applied INCAnTeSIMO (Zaffaroni et al., 2019). Briefly, the GRN representing the cellular state transition for each time interval was connected to the signalling network through interface TFs. The perturbations on the interface TFs were simulated in-silico since the interface TFs act as signalling effectors which transmit the signal from cytoplasm to the nucleus. These interface TFs were perturbed exhaustively and in combinations of up to 4 TFs at a time by fixing their Boolean states and synchronously updating the Boolean state of the network following a majority logic rule until it converged to a stable-state attractor. The interface TFs alone or combinations that can induce minimum 40% of TFs flipping their Boolean states in the GRN were kept. The combinations of interface TFs with the best flipping scores (including ties) were used to calculate the likelihood of each interface TF to induce the desired shift in gene expression program. Subsequently, the signalling molecules targeting on these best interface TFs were identified. The availability of each node of the signalling network was calculated based on expression probability as described above. The most probable path between each signalling molecule and the interface TF was selected to represent the probability and the sign with which the signalling molecule influences the interface TF activity state. Finally, signalling molecules connected to the GRN were ranked by comparing their probabilities of activating/inhibiting the interface TFs with the likelihood of the interface TFs

that have effects on the GRN state by calculating Jensen-Shannon divergence (JSD). The top 6% of the ranking was selected as single molecule candidates (Figure 4.8C).

Prediction of signalling pathways for salamander regeneration

In novel transitions or systems, it is complicated to explain a list of single molecules. However, one could argue that the predicted canonical pathways are particularly relevant for many biological processes, such as development and proliferation, since these processes are driven by the concerted action of signalling pathways. Therefore, we further predicted the canonical signalling pathways underlying the predicted signalling candidates. To infer signalling pathways that trigger desired change of GRN, it is not reasonable to simply perform gene set enrichment analysis using the predicted signalling molecules. Because it is not expected that all signalling molecules perform equally in a pathway. Nevertheless, if the predicted signalling molecules are particularly influential in a pathway, we can assume that this pathway can induce similar effects as the molecules on the GRN and in turn induce the same cellular transitions. According to the topology of network, the network centrality indicates the importance of particular nodes in a network to its connectivity. To this end, our method used the concept of source/sink centrality (SSC) to perform pathway enrichment analysis on the list of predicted signalling molecules (Yeganeh and Mostafavi, 2020). SSC of each node was calculated by considering the number of directed paths in a signalling pathway that pass through this node, including incoming and outgoing, as well as the length of these paths. It was calculated as follows:

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

where v is the considered node in the pathway graph G, u is any other node in G, w_j is an incoming or outgoing path connecting v to u, α is a dampening factor (here α =0.1) that decreases by the length of paths $|w_j|$, and β (here $\beta = 1$) leverages the importance of the source and sink components in the centrality score. $C_{ssc}(v)$ determines the importance of the node v as sender and receiver of a signal in the considered pathway.

In the original study, an enrichment score was obtained by calculating the aggregated C_{ssc} of DEGs, and the statistical significance was computed by using a bootstrap (Yeganeh and Mostafavi, 2020). Here, the aggregated C_{ssc} of signalling molecule candidates V obtained from our method was calculated by:

$$Agg(V) = \prod_{v_i \in V} C_{ssc}(v_i)$$

The probability of observing a higher aggregate score for a randomly selected subset of G is used as the p-value for each pathway.

Results

Overview of the method

In this study, we present an integrative network-based model that combines signalling and transcriptional network to identify signalling molecules and pathways that can induce the transition from an initial to a target cellular state. The method only requires the gene expression profiles of the initial and target cellular states and therefore can be easily applied to any cellular transitions. Here, our method was employed to help understand the process of salamander limb regeneration. Specifically, DEGs are identified and used to calculate Boolean state and the expression probability of each gene (Figure 4.8A). Subsequently, the GRN is constructed by integrating five different GRN inference methods and then contextualized to the Boolean states of TFs in the GRN (Figure 4.8B). In order to predict the signalling molecules and pathways whose perturbations induce the GRN transition as expected, the signalling network is connected to the downstream GRN with interface TFs. Since the interface TFs transmit the perturbation signal from signalling network to GRN, we mimic the perturbations on interface TFs and update the GRN synchronously using Boolean network. As a consequence, the likelihood of each interface TF to induce the desired GRN state changes is obtained and compared to the probability of each signalling molecule to activate and inhibit the interface TFs. This allows to rank single signalling molecule. The top-ranking ones are selected, and their topological properties are calculated and used for enrichment in each signalling pathway (Figure 4.8C).

GO enrichment analysis on inferred network

The transcriptomics data of regenerating axolotl limb published to date were either a mixture of heterogeneous cell types in the entire limb, including those that do not undergo cell fate conversion during regeneration (Knapp et al., 2013), or lack a sufficient sequencing depth to accurately quantify low-abundance genes, such as TFs (Gerber et al., 2018). Therefore, we generated time-series microarray data of Prrx1+ cells between 0 and 14 dpa, which enabled us to discern the gene expression dynamics specific to this cell type during the course of limb regeneration. To gain insights into the transcriptional regulation among TFs, we constructed a

raw GRN by inferring potential interactions based on various measures for statistical dependency between each pair of TFs (see Methods). This resulted in a global GRN consisting of 672 TFs and 3360 interactions. To investigate whether the constructed GRN captured the biological processes related to the limb regeneration, we first performed GO enrichment analysis on the global GRN considering the topology of the network. The resultant GO term network revealed three major clusters (Figure 4.9A). One cluster gathered GO terms related to response to external stimulus and stress (Figure 4.9D), which have been implicated in the initiation of limb regeneration in previous study (Darnet et al., 2019). In particular, the response to reactive oxygen species (ROS) was presented, which has been also reported as the earliest indication to induce the regeneration for limb and tail in Xenopus and salamanders (Al Haj Baddar et al., 2019; Love et al., 2013). Another large group of GO terms related to regulation of cellular process was observed (Figure 4.9B), including terms for tissue development, such as, muscle development and ossification, correctly associated with the blastema differentiation into new muscle, cartilage and bone during the regeneration process (Stock et al., 2003). Positive regulation of Wnt and TGF-β pathways was also found in this cluster. Wnt pathway is known to play distinct roles in limb regeneration, including blastema formation, blastema cell differentiation and proliferation and orchestrating tissue organization (Caubit et al., 1997; Wehner et al., 2014; Yokoyama et al., 2007). TGF-β pathway also plays spatially and temporally various roles in regeneration, such as the formation of wound epithelium and bud structure, and proliferation (Ho and Whitman, 2008). The third group of GO terms was specifically enriched with the function of limb development, such as limb/appendage morphogenesis and mesoderm development (Figure 4.9C). Previous study has revealed that limb development shares similar basic mechanisms involved in limb regeneration (Muneoka and Bryant, 1982). Thus, these three clusters were in accordance with the three stages occurring in regeneration: wound healing, blastema formation, and limb patterning (Knapp et al., 2013), and this implicated the reliability of our constructed network.

Next, each time interval specific subnetwork was also examined by performing GO enrichment analysis. In the initial time interval (0 – 1 dpa), there were multiple GO terms associated with immune response and response to stimulus (Figure 4.10A). Between 1 and 7 dpa, GO terms for chromatin organization, signal transduction and cellular proliferation were enriched (Figures 4.10B-C). After 5 dpa, the GO terms related to tissue development and morphogenesis were enriched (Figures 4.10D-E). This confirmed that each subnetwork was

consistent with the limb regeneration temporally. The full list of enriched GO terms for each time interval is shown in (Table S11).

Prediction of signalling molecules and pathways for salamander limb regeneration

Given the interval specific GRNs, our computational method was applied to identify signalling molecules and pathways for limb regeneration in each time interval (Table S12). Many of our predictions have been shown to play key roles in regeneration in previous studies. Overall, across the time series, the method identified signalling factors mainly related to wound healing up to 3 dpa, followed by cellular migration, de-differentiation and proliferation relevant pathways between 5 pda and 7 pda and finally cellular re-differentiation (Table 4.5). For the interval between 3 pda and 5 pda, there were no interface TFs that were able to induce substantial flipping of TF states (>40% of the GRN-TFs) in the GRN and therefore this interval was discarded. Details of the predicted signalling candidates were depicted temporally as follows.

At the beginning of amputation (0 to 3 dpa), a set of the predicted candidates included signals that have been reported to have association with the initiation of regeneration. Specifically, the method predicted p38/JNK signalling, which has an essential role in wound closure and epithelial to mesenchymal transition (EMT) (Sader et al., 2019). ERK/MEK and PI3K/AKT signallings were also predicted, consistent with their roles in the initiation of regeneration and blastema formation reported in *X. laevis* and planarians (Owlarn et al., 2017; Suzuki et al., 2007; Tasaki et al., 2011). In addition, our method predicted Bcl-2 family proteins, which have been shown to be important in the initial stages of limb regeneration by regulating apoptosis (Bucan et al., 2018). The predicted activating of Wnt signalling has been shown to promote the wound healing (Zhang et al., 2009), in accordance with the previous finding that the initial phase of regeneration exhibited the wound healing process (Knapp et al., 2013).

Between 1 and 3 dpa, our method identified the inhibition of T cell receptor signalling and NK cell cytotoxicity, in agreement with the hypothesis about the necessity of inhibiting the lysis of progenitor populations for salamander regeneration (Godwin and Rosenthal, 2014). The predicted activation of C/EBPβ has been shown to play an important role in regeneration by regulating the function of macrophage (Godwin et al., 2013; Ruffell et al., 2009).

After 5 dpa, the activation of the kallikrein-kinin signalling pathway was predicted, including coagulation factors and thrombin, in accordance with their roles in the initiation of the regenerative process (Imokawa and Brockes, 2003). In addition, our method identified the

activation of ErbB proteins and Src/FAK signalling, which are correctly associated to cellular migration during vertebrate regeneration (Makanae and Satoh, 2012; Rojas-Muñoz et al., 2009).

In the next stage (7-10 dpa), multiple matrix metalloproteinases (MMPs-2, 9, 13, 14) were predicted, which are required to induce blastema formation and regulate the expression of Prrx1 in A. mexicanum (Satoh et al., 2011). The hedgehog pathway as well as Wnt signalling were predicted at this time interval, in agreement with the observation that they coordinate and play roles in proliferation and migration for regeneration (Singh et al., 2012; Singh et al., 2015). In addition, the predicted insulin signalling was also reported to be required in this stage (Stocum and Cameron, 2011).

In the limb bud stage, between 10 and 14 dpa, ERK signalling was predicted, which has been shown to induce the differentiation of blastema cells (Tasaki et al., 2011).TNF-α and NF-κB signallings were predicted, which are consistent with the activation of blastema cells in zebrafish (Nguyen-Chi et al., 2017).

From the view of function, the predicted singalling pathways mainly can be categorized into five major classes: immune response, inflammation, signal transduction, development and apoptosis (Figure 4.11). Many immune system-related singalling pathways, including interleukin family signalling, were predicted across the entire time course, especially in the first and last time intervals. Consistently, a previous study has revealed that the initial phase of regeneration exhibited the wound healing process, where immune cells secrete various signaling molecules, such as cytokines, growth factors and chemokines (Knapp et al., 2013). Apart from the immune cells with the functions of protecting injured tissue from infection, macrophages play important roles in anti-inflammatory to facilitate regeneration. In our result, we identified phagocytosis which has been shown to be related to limb regeneration by hindering fibrosis formation (Godwin et al., 2013). We also predicted a series of signal transduction pathways, such as WNT, NOTCH, and ERBB singalling pathways, which have been known to be required for limb regeneration in regulation of various tissue development, such as apical ectodermal ridge, spinal cord and cartilage (Beck et al., 2003; Fisher et al., 2007; Kawakami et al., 2006). From the class of development, we predicted hedgehog, EMT, angiogenesis and ossification pathways, which mainly appeared after 5 dpa. These development-related singalling pathways have been proven with functions of axial patterning, cellular migration and tissue development (Chiang et al., 1996; Dolan et al., 2018; Hankenson et al., 2011; Sader et al., 2019). Finally, our method also captured singalling pathways related to apoptosis at the beginning of amputation (0-1 dpa). Tseng et al. have performed an

experiment to inhibit the caspase-dependent apoptosis, which triggered axon mis-pattern in regeneration, in order to prove that apoptotic event is required for regeneration (Tseng et al., 2007).

Taken together, our newly developed computational method predicted many signalling pathways and molecules, which have been shown to be relevant to limb regeneration, and are therefore primary candidates for experimental follow-ups to gain a better understandings of this biological process.

Conclusion

In this study, we proposed a network-based model with the integration of signalling and transcriptional networks to identify signalling molecules and pathways that can induce the transition from an initial to a target cellular state. Here, we employed Boolean models, as they represent a simple yet powerful approach to modelling GRNs. Although continuous models can model more precise gene expression dynamics, it requires a large amount (>100 samples) of transcriptome data or kinetic parameter inference, which is not widely available currently. While numerous computational methods have been developed to identify signalling pathways, our method conceptually differs from these methods, since none of them combines signalling pathways with cell type specific GRNs. Therefore, they are unable to systematically describe and predict key signalling pathways inducing cellular transitions by altering the GRN state as our method does.

The method was applied to identify intracellular signalling pathways to induce defined cellular decisions such as the initiation of proliferation, migration and patterning during salamander limb regeneration. The transcriptomics data of regenerating axolotl limb published to date were derived from a mixture of many different cell types, including those not participating in regeneration. This likely masked the gene expression changes in the important target cell type: the connective tissue cells. To address this limitation, time-course microarray data of blastema specific cell type (connective tissue cells) following upper-arm amputation up to 14 dpa was generated. By using this time-series microarray dataset, a raw GRN for salamander limb regeneration was constructed by the integration of different network inference methods. We could not use interactions reported in literature for this system, as currently no curated molecular interaction database for axolotls exists and obtaining molecular interactions among orthologues to other, more well-studied species (e.g., mouse and human) may lose interactions specific to Prrx1+ cells in regenerating axolotl limb. The reconstructed GRNs correctly captured the three stages occurring in regeneration, including wound healing,

blastema formation, and limb patterning. By applying our method, the signalling molecules and canonical siganlling pathways that are able to induce the transition of GRN for each time interval were predicted with literature evidences. Notably, our computational method takes advantage of well-conserved canonical signalling pathways, and it can readily be applied to non-model organisms (e.g., axolotls) where no database for curated molecular interactions exists. Overall, the predictions up to 3 dpa were related to wound healing, including immune response and response to stimuli. Around 5-7 dpa, signalling molecules and pathways relevant to cellular migration and de-differentiation were captured. Predictions related to cellular proliferation and re-differentiation were identified after 7 pda. Some novel predictions that could potentially have novel functions in axolotl limb regeneration are primary candidates for experimental follow-up studies.

In conclusion, we proposed a computational method to identify upstream signalling molecules/pathways whose perturbation can induce desired downstream GRN change and in turn trigger cellular transitions. Our method aids experimentalists in understanding salamander limb regeneration, and can also be easily applied to any other biological processes.

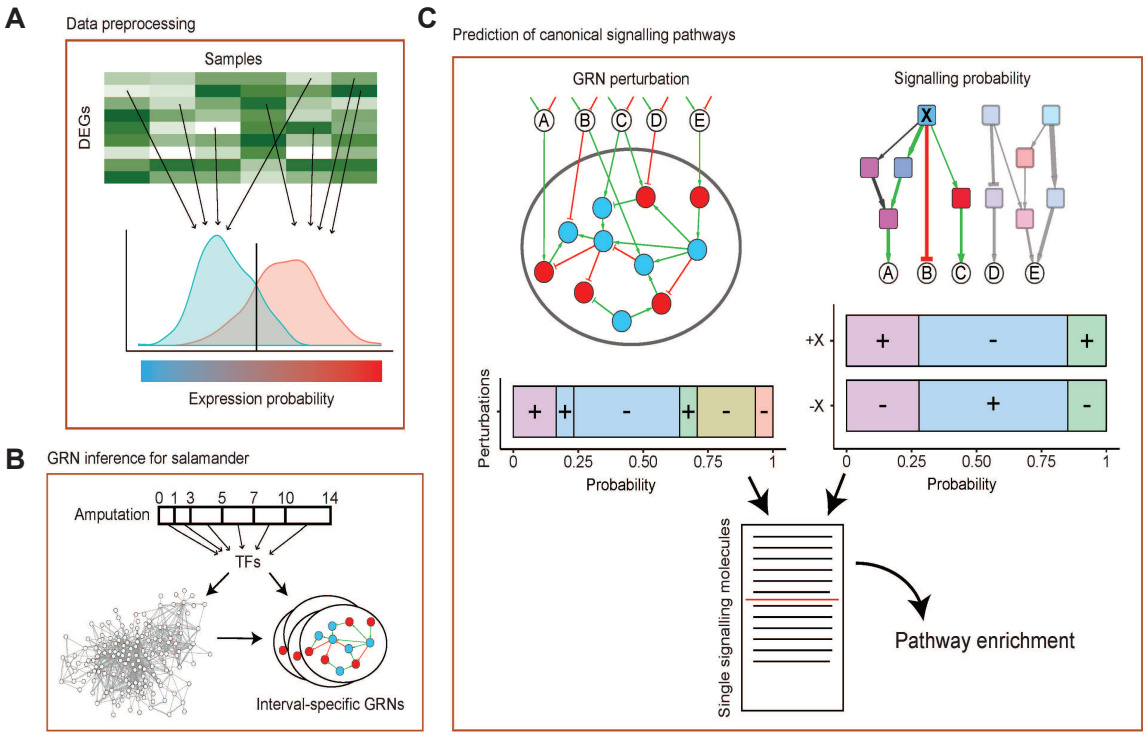

Figure 4. 8 Schematic outline of the method.

A, Gene expression data is first submitted to perform differential expression analysis. The maximum and minimum expression values for each DEGs are selected and formed the overall expressed and not-expressed probability distributions. These distributions define the Boolean state and the expression probability of each gene.
B, A global GRN is first constructed by using the time-series microarray datasets. Interval specific subnetworks containing TFs of corresponding time interval are extracted from the global GRN.
C, Exhaustive perturbations of the interface TFs connecting the GRN to signalling pathways are used to define the probability of each interface TF to induce the desired GRN state changes. The expression probability of each gene is mapped on the signalling network and used to define the probability of each signalling molecule to activate or inhibit each interface TFs. The two probability distributions are compared by Jensen-Shannon divergence and result in a ranking of signalling molecules. The final pathway predictions are obtained by enrichment analysis.

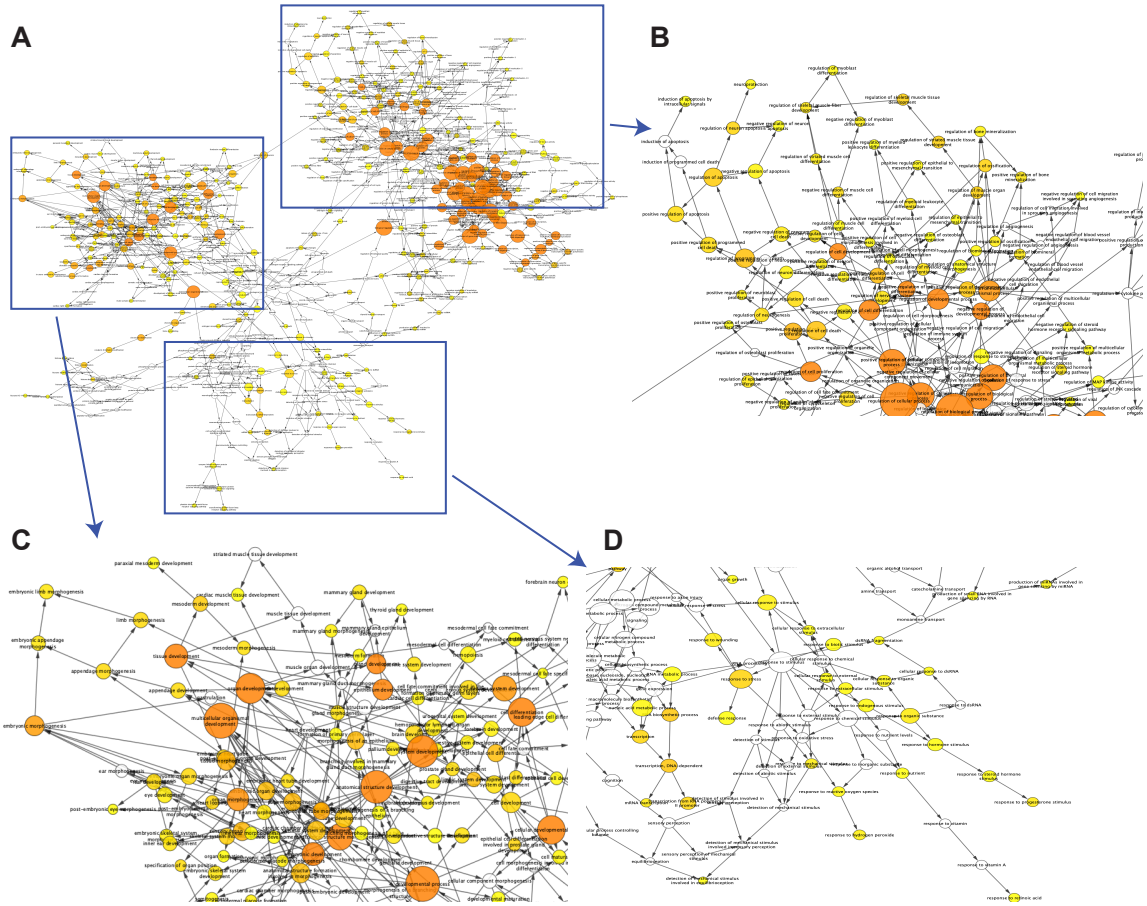

Figure 4. 9 GO term network across entire time course.

Enrichment of biological process (BP) GO terms for all TFs across time course included in the global GRN. GO terms with BH adjusted p-value 0.05 are shown. The color and the size represent the p-values and number of genes belonging to a certain GO category, respectively. There are mainly three modules in the network.
A, Complete GO terms distribution on the global constructed GRN.
B, Details of GO terms cluster related to regulation of cellular process.
C, Details of GO terms cluster related to limb development.
D, Details of GO terms cluster related to response to external stimulus and stress.

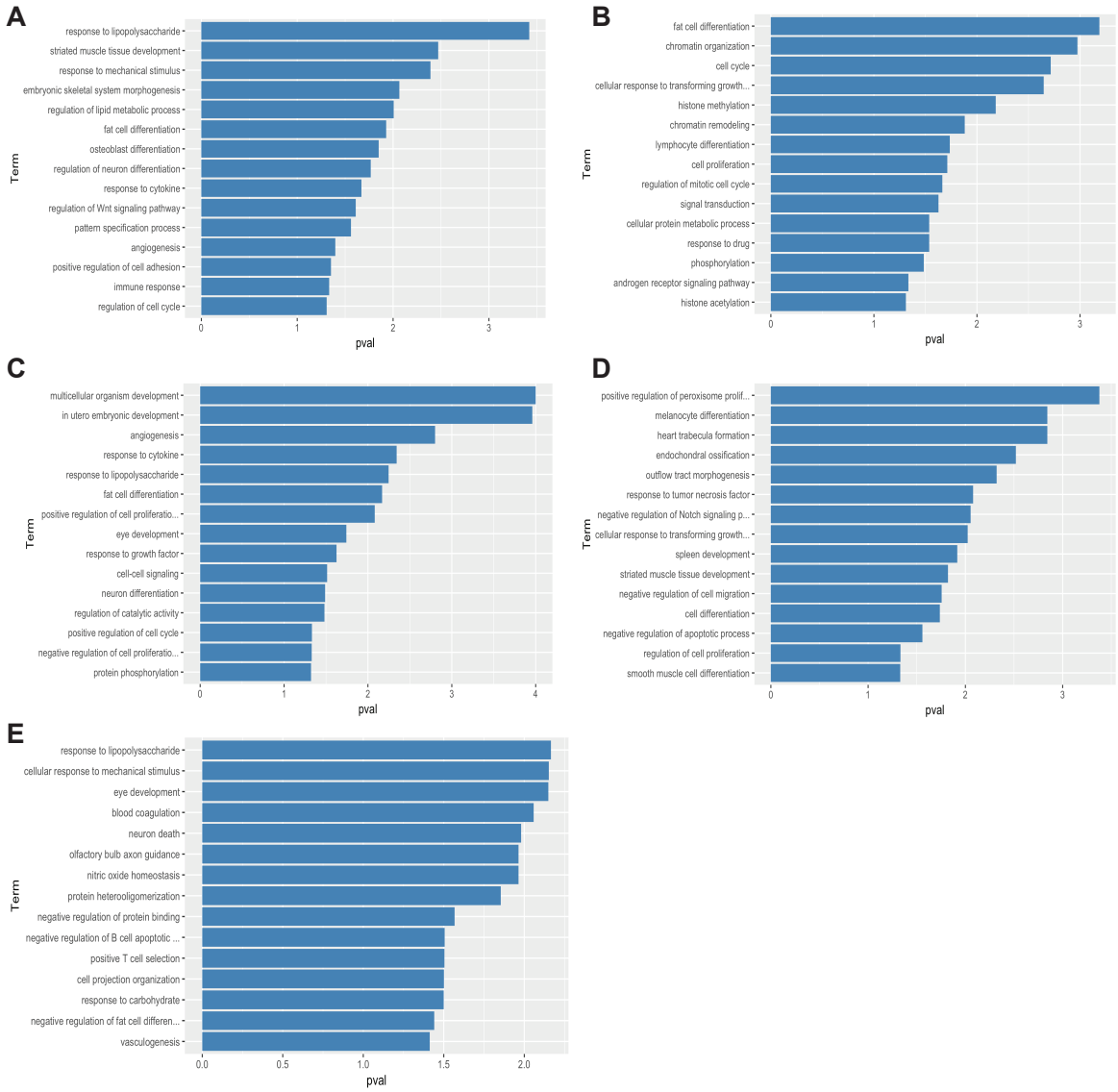

Figure 4. 10 The top 15 GO terms for each time interval.

The GO enrichment analysis is performed on each subnetwork and the GO terms are selected with p-values less than 0.05. The x-axis value is negative log10 transformation(-log10(pval)). A, 0-1 dpa; B, 1-3 dpa; C, 5-7 dpa; D, 7-10 dpa; E, 10-14 dpa. Full list of the GO terms are shown in Table S11.

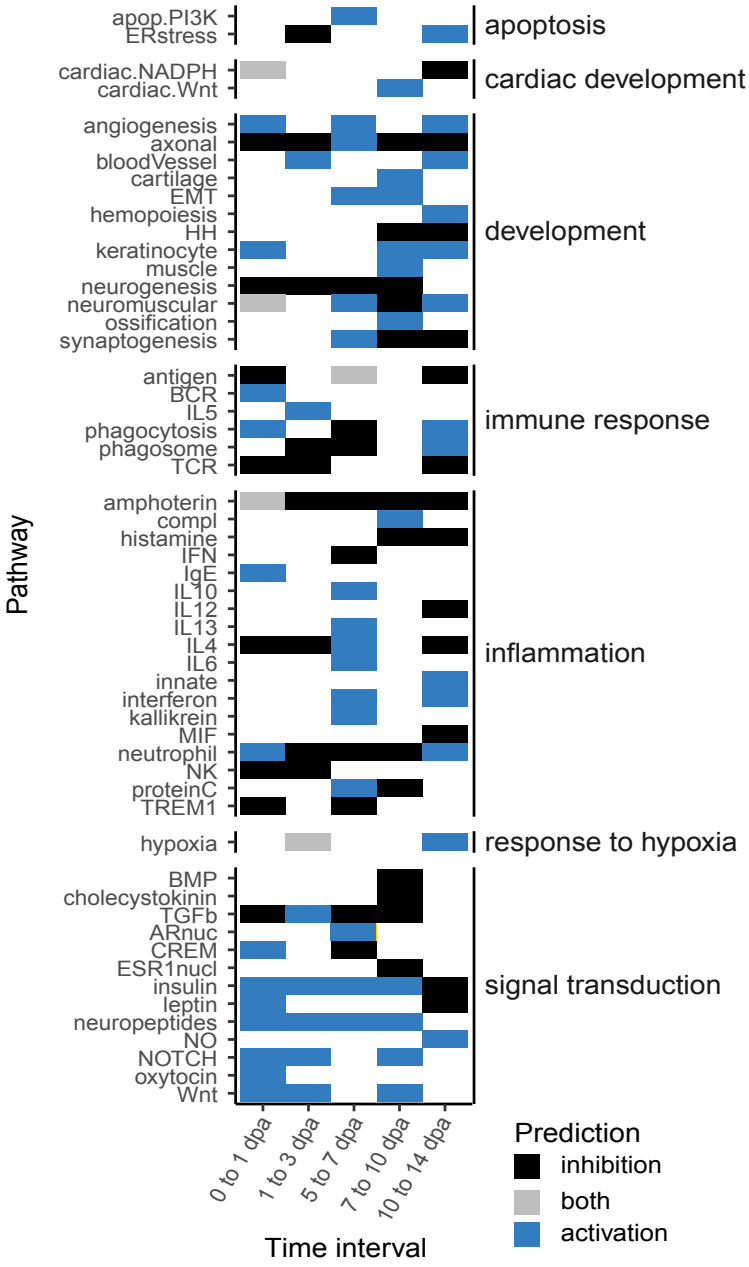

Figure 4. 11 The categories of predicted signalling pathways for each time interval along the regeneration processes.

The full list of predictions is shown in Table S12 and the full pathway names are listed in Table S13.

Signalling entity	Time interval	Role	Literature evidence
NADPH and ROS signalling	0-1dpa	Cellular activation and proliferation	(Al Haj Baddar et al., 2019)
Phagocytosis	0-1dpa, 10-14dpa	Wound healing	(Godwin et al., 2013)
HHs, Smoothened	0-1dpa, 7-14dpa	Cellular activation and proliferation	(Singh et al., 2012; Singh et al., 2015)
Wnt pathway	0-3dpa, 7-10dpa	Wound healing	(Zhang et al., 2009)
p38/JNK	0-3dpa, 7-14dpa	Wound healing, EMT	(Sader et al., 2019)
Bcl-2	0-3dpa, 10-14dpa	Apoptosis	(Bucan et al., 2018)
FGF receptors	0-10dpa	Fibroblasts de-differentiation, blastema formation	(Makanae et al., 2014)
ERK/MEK	0-14dpa	Fibroblasts de-differentiation, blastema formation	(Owlarn et al., 2017; Suzuki et al., 2007) (Tasaki et al., 2011; Yun et al., 2014)
PI3K/AKT	0-14dpa	Blastema formation	(Suzuki et al., 2007)
GDF5	0-1dpa, 7-14dpa	Blastema formation	(Makanae et al., 2013)
Retinoic acid receptors	1-3dpa, 5-7dpa	Apical epidermal cap, skeletal patterning and differentiation	(Monaghan et al., 2012; Nguyen et al., 2017)
C/EBPβ	1-3dpa, 7-10dpa	Macrophage functionality	(Godwin et al., 2013; Ruffell et al., 2009)
Neuregulin 1	1-3dpa	Blastema formation	(Farkas et al., 2016)
p53	1-14dpa	Blastema formation	(Yun et al., 2013)
ErbB2-3	5-7dpa, 10-14dpa	Cellular migration	(Rojas-Muñoz et al., 2009)
Thrombin	5-7dpa	Cell cycle re-entry	(Imokawa and Brockes, 2003; Yun et al., 2014)
FGF8	5-7dpa	Cellular proliferation	(Nacu et al., 2016)
MMPs	7-10dpa	Blastema induction	(Satoh et al., 2011)
BMPR1B	7-10dpa	Chondrocyte differentiation	(Kotzsch et al., 2009)

TNF-α	10-14dpa	Blastema induction	(Nguyen-Chi et al., 2017)

Table 4. 5 The predicted signalling molecules and pathways for salamander limb regeneration in the different time intervals and corresponding literature evidences.

Reference

Al Haj Baddar, N.W., Chithrala, A., and Voss, S.R. (2019). Amputation-induced reactive oxygen species signaling is required for axolotl tail regeneration. Dev Dyn *248*, 189-196.
Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600-1607.
Beck, C.W., Christen, B., and Slack, J.M. (2003). Molecular pathways needed for regeneration of spinal cord and muscle in a vertebrate. Dev Cell *5*, 429-439.
Brockes, J.P., Kumar, A., and Velloso, C.P. (2001). Regeneration as an evolutionary variable. J Anat *199*, 3-11.
Bucan, V., Peck, C.T., Nasser, I., Liebsch, C., Vogt, P.M., and Strauß, S. (2018). Identification of axolotl BH3-only proteins and expression in axolotl organs and apoptotic limb regeneration tissue. Biol Open *7*.
Caubit, X., Nicolas, S., and Le Parco, Y. (1997). Possible roles for Wnt genes in growth and axial patterning during regeneration of the tail in urodele amphibians. Dev Dyn *210*, 1-10.
Chiang, C., Litingtung, Y., Lee, E., Young, K.E., Corden, J.L., Westphal, H., and Beachy, P.A. (1996). Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. Nature *383*, 407-413.
Crespo, I., and Del Sol, A. (2013). A general strategy for cellular reprogramming: the importance of transcription factor cross-repression. Stem Cells *31*, 2127-2135.
Darnet, S., Dragalzew, A.C., Amaral, D.B., Sousa, J.F., Thompson, A.W., Cass, A.N., Lorena, J., Pires, E.S., Costa, C.M., Sousa, M.P.*, et al.* (2019). Deep evolutionary origin of limb and fin regeneration. Proc Natl Acad Sci U S A *116*, 15106-15115.
Dolan, C.P., Dawson, L.A., and Muneoka, K. (2018). Digit Tip Regeneration: Merging Regeneration Biology with Regenerative Medicine. Stem Cells Transl Med *7*, 262-270.
Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol *5*, e8.
Farkas, J.E., Freitas, P.D., Bryant, D.M., Whited, J.L., and Monaghan, J.R. (2016). Neuregulin-1 signaling is essential for nerve-dependent axolotl limb regeneration. Development *143*, 2724-2731.
Fisher, M.C., Clinton, G.M., Maihle, N.J., and Dealy, C.N. (2007). Requirement for ErbB2/ErbB signaling in developing cartilage and bone. Dev Growth Differ *49*, 503-513.
Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M., Nowoshilow, S., Kageyama, J., Khattak, S.*, et al.* (2018). Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science *362*.
Godwin, J.W., Pinto, A.R., and Rosenthal, N.A. (2013). Macrophages are required for adult salamander limb regeneration. Proc Natl Acad Sci U S A *110*, 9415-9420.
Godwin, J.W., and Rosenthal, N. (2014). Scar-free wound healing and regeneration in amphibians: immunological influences on regenerative success. Differentiation *87*, 66-75.
Guo, S., Jiang, Q., Chen, L., and Guo, D. (2016). Gene regulatory network inference using PLS-based methods. BMC Bioinformatics *17*, 545.
Hankenson, K.D., Dishowitz, M., Gray, C., and Schenker, M. (2011). Angiogenesis in bone regeneration. Injury *42*, 556-561.
Haury, A.C., Mordelet, F., Vera-Licona, P., and Vert, J.P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC Syst Biol *6*, 145.
Ho, D.M., and Whitman, M. (2008). TGF-beta signaling is required for multiple processes during Xenopus tail regeneration. Dev Biol *315*, 203-216.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*.
Imokawa, Y., and Brockes, J.P. (2003). Selective activation of thrombin is a critical determinant for vertebrate lens regeneration. Curr Biol *13*, 877-881.
Kawakami, Y., Rodriguez Esteban, C., Raya, M., Kawakami, H., Martí, M., Dubova, I., and Izpisúa Belmonte, J.C. (2006). Wnt/beta-catenin signaling regulates vertebrate limb regeneration. Genes Dev *20*, 3232-3237.
Knapp, D., Schulz, H., Rascon, C.A., Volkmer, M., Scholz, J., Nacu, E., Le, M., Novozhilov, S., Tazaki, A., Protze, S*., et al.* (2013). Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. PLoS One *8*, e61352.
Kotzsch, A., Nickel, J., Seher, A., Sebald, W., and Müller, T.D. (2009). Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5-type I receptor specificity. Embo j *28*, 937-947.
Lévesque, M., Gatien, S., Finnson, K., Desmeules, S., Villiard, E., Pilote, M., Philip, A., and Roy, S. (2007). Transforming growth factor: beta signaling is essential for limb regeneration in axolotls. PLoS One *2*, e1227.
Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jørgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K*., et al.* (2007). Systematic discovery of in vivo phosphorylation networks. Cell *129*, 1415-1426.
Love, N.R., Chen, Y., Ishibashi, S., Kritsiligkou, P., Lea, R., Koh, Y., Gallop, J.L., Dorey, K., and Amaya, E. (2013). Amputation-induced reactive oxygen species are required for successful Xenopus tadpole tail regeneration. Nat Cell Biol *15*, 222-228.
Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.
Makanae, A., Hirata, A., Honjo, Y., Mitogawa, K., and Satoh, A. (2013). Nerve independent limb induction in axolotls. Dev Biol *381*, 213-226.
Makanae, A., Mitogawa, K., and Satoh, A. (2014). Co-operative Bmp- and Fgf-signaling inputs convert skin wound healing to limb formation in urodele amphibians. Dev Biol *396*, 57-66.
Makanae, A., and Satoh, A. (2012). Early regulation of axolotl limb regeneration. Anat Rec (Hoboken) *295*, 1566-1574.
Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat Methods *9*, 796-804.
Mercader, N., Leonardo, E., Piedra, M.E., Martínez, A.C., Ros, M.A., and Torres, M. (2000). Opposing RA and FGF signals control proximodistal vertebrate limb development through regulation of Meis genes. Development *127*, 3961-3970.
Monaghan, J.R., Athippozhy, A., Seifert, A.W., Putta, S., Stromberg, A.J., Maden, M., Gardiner, D.M., and Voss, S.R. (2012). Gene expression patterns specific to the regenerating limb of the Mexican axolotl. Biol Open *1*, 937-948.
Muneoka, K., and Bryant, S.V. (1982). Evidence that patterning mechanisms in developing and regenerating limbs are the same. Nature *298*, 369-371.
Muneoka, K., Fox, W.F., and Bryant, S.V. (1986). Cellular contribution from dermis and cartilage to the regenerating limb blastema in axolotls. Dev Biol *116*, 256-260.
Nacu, E., Gromberg, E., Oliveira, C.R., Drechsel, D., and Tanaka, E.M. (2016). FGF8 and SHH substitute for anterior-posterior tissue interactions to induce limb regeneration. Nature *533*, 407-410.

Nguyen, M., Singhal, P., Piet, J.W., Shefelbine, S.J., Maden, M., Voss, S.R., and Monaghan, J.R. (2017). Retinoic acid receptor regulation of epimorphic and homeostatic regeneration in the axolotl. Development *144*, 601-611.
Nguyen-Chi, M., Laplace-Builhé, B., Travnickova, J., Luz-Crawford, P., Tejedor, G., Lutfalla, G., Kissa, K., Jorgensen, C., and Djouad, F. (2017). TNF signaling and macrophages govern fin regeneration in zebrafish larvae. Cell Death Dis *8*, e2979.
Owlarn, S., Klenner, F., Schmidt, D., Rabert, F., Tomasso, A., Reuter, H., Mulaw, M.A., Moritz, S., Gentile, L., Weidinger, G.*, et al.* (2017). Generic wound signals initiate regeneration in missing-tissue contexts. Nat Commun *8*, 2282.
Reddien, P.W., and Sánchez Alvarado, A. (2004). Fundamentals of planarian regeneration. Annu Rev Cell Dev Biol *20*, 725-757.
Rojas-Muñoz, A., Rajadhyksha, S., Gilmour, D., van Bebber, F., Antos, C., Rodríguez Esteban, C., Nüsslein-Volhard, C., and Izpisúa Belmonte, J.C. (2009). ErbB2 and ErbB3 regulate amputation-induced proliferation and migration during vertebrate regeneration. Dev Biol *327*, 177-190.
Ruffell, D., Mourkioti, F., Gambardella, A., Kirstetter, P., Lopez, R.G., Rosenthal, N., and Nerlov, C. (2009). A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. Proc Natl Acad Sci U S A *106*, 17475-17480.
Sader, F., Denis, J.F., Laref, H., and Roy, S. (2019). Epithelial to mesenchymal transition is mediated by both TGF-β canonical and non-canonical signaling during axolotl limb regeneration. Sci Rep *9*, 1144.
Satoh, A., makanae, A., Hirata, A., and Satou, Y. (2011). Blastema induction in aneurogenic state and Prrx-1 regulation by MMPs and FGFs in Ambystoma mexicanum limb regeneration. Dev Biol *355*, 263-274.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Singh, B.N., Doyle, M.J., Weaver, C.V., Koyano-Nakagawa, N., and Garry, D.J. (2012). Hedgehog and Wnt coordinate signaling in myogenic progenitors and regulate limb regeneration. Dev Biol *371*, 23-34.
Singh, B.N., Koyano-Nakagawa, N., Donaldson, A., Weaver, C.V., Garry, M.G., and Garry, D.J. (2015). Hedgehog Signaling during Appendage Development and Regeneration. Genes (Basel) *6*, 417-435.
Stock, S.R., Blackburn, D., Gradassi, M., and Simon, H.G. (2003). Bone formation during forelimb regeneration: a microtomography (microCT) analysis. Dev Dyn *226*, 410-417.
Stocum, D.L., and Cameron, J.A. (2011). Looking proximally and distally: 100 years of limb regeneration and beyond. Dev Dyn *240*, 943-968.
Suzuki, M., Satoh, A., Ide, H., and Tamura, K. (2007). Transgenic Xenopus with prx1 limb enhancer reveals crucial contribution of MEK/ERK and PI3K/AKT pathways in blastema formation during limb regeneration. Dev Biol *304*, 675-686.
Suzuki, Y., Chou, J., Garvey, S.L., Wang, V.R., and Yanes, K.O. (2019). Evolution and Regulation of Limb Regeneration in Arthropods. Results Probl Cell Differ *68*, 419-454.
Tasaki, J., Shibata, N., Nishimura, O., Itomi, K., Tabata, Y., Son, F., Suzuki, N., Araki, R., Abe, M., Agata, K.*, et al.* (2011). ERK signaling controls blastema cell differentiation during planarian regeneration. Development *138*, 2417-2427.
Tseng, A.S., Adams, D.S., Qiu, D., Koustubhan, P., and Levin, M. (2007). Apoptosis is required during early stages of tail regeneration in Xenopus laevis. Dev Biol *301*, 62-69.
Wehner, D., Cizelsky, W., Vasudevaro, M.D., Ozhan, G., Haase, C., Kagermeier-Schenk, B., Röder, A., Dorsky, R.I., Moro, E., Argenton, F.*, et al.* (2014). Wnt/β-catenin signaling

defines organizing centers that orchestrate growth and differentiation of the regenerating zebrafish caudal fin. Cell Rep *6*, 467-481.
Wilkie, I.C. (2001). Autotomy as a prelude to regeneration in echinoderms. Microsc Res Tech *55*, 369-396.
Yeganeh, P.N., and Mostafavi, M.T. (2020). Causal Disturbance Analysis: A Novel Graph Centrality Based Method for Pathway Enrichment Analysis. IEEE/ACM Trans Comput Biol Bioinform *17*, 1613-1624.
Yokoyama, H., Ogino, H., Stoick-Cooper, C.L., Grainger, R.M., and Moon, R.T. (2007). Wnt/beta-catenin signaling has an essential role in the initiation of limb regeneration. Dev Biol *306*, 170-178.
Yun, M.H., Gates, P.B., and Brockes, J.P. (2013). Regulation of p53 is critical for vertebrate limb regeneration. Proc Natl Acad Sci U S A *110*, 17392-17397.
Yun, M.H., Gates, P.B., and Brockes, J.P. (2014). Sustained ERK activation underlies reprogramming in regeneration-competent salamander cells and distinguishes them from their mammalian counterparts. Stem Cell Reports *3*, 15-23.
Zaffaroni, G., Okawa, S., Morales-Ruiz, M., and Del Sol, A. (2019). An integrative method to predict signalling perturbations for cellular transitions. Nucleic Acids Res *47*, e72.
Zhang, D.L., Gu, L.J., Liu, L., Wang, C.Y., Sun, B.S., Li, Z., and Sung, C.K. (2009). Effect of Wnt signaling pathway on wound healing. Biochem Biophys Res Commun *378*, 149-151.
Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res *43*, D76-81.

5 Discussion

Cellular transition is a complex, dynamic process that involves several layers of regulations. Not only the intrinsic genetic programme of the cell, but also external influences such as chemical compounds, infections and tissue injury, can induce the fluctuation of microenvironment (niche) and in turn trigger a transition from one cellular state to another. The influence of a niche is transmitted through intracellular signal transduction pathways, which in turn brings about changes in the cellular GRN. Niche effects on cellular transitions including the differentiation tendency of various stem cells, cell reprogramming, generation of cellular disease phenotypes, cell migration and regeneration. This is particularly valuable for clinical applications, as it allows to derive required cells and tissues for regenerative medicine, or to revert cellular disease phenotypes to their healthy counterparts. In the recent decade, chemical compounds that target siganlling pathways have been widely used for cell engineering, including reprogramming and transdifferentiation (Qin et al., 2017). In addition, disease treatment often uses drugs, which most likely target proteins, to improve cellular function or interevent disease. However, identification of optimal chemical compounds for chemical-induced cellular conversion or disease treatment usually rely on exhaustive trial-and-error testing of a large scale of compounds, which is both inefficient and resource intensive.

In this regards, computational methods taking advantage of high-throughput data is particularly appealing since they are capable of screening candidates rapidly and reducing the number of possible candidates in an unbiased way. To date, multiple computational methods have been developed to predict chemical compounds reverting disease phenotypes by mapping to a perturbation compendium which mainly consists of cancer cells (Lamb et al., 2006; Subramanian et al., 2017). However, this is not suitable to infer signalling perturbation for cellular conversions or non-cancer diseases, since cancer cells often exhibit significant differences in the signalling mechanism compared to the non-cancer counterparts. In addition, these methods often do not consider the cellular initial state, which is important for the faithful generation of target cells. Another category of existing computational methods uses gene expression data as the proxy of signalling activity to identify signalling pathways directly that are activated or inhibited in disease conditions or other perturbations (Bao et al., 2016; Fakhry et al., 2016). It is ambiguous whether the signalling activity can be reflected without considering the post-translational modification.

To address the limitation of current computational methods, the primary contribution of this thesis is to introduce comprehensive computational methods to identify signalling perturbations, including chemical compounds, signalling pathways and signalling molecules, that can induce the cellular conversion, revert disease phenotypes and trigger regeneration.

5.1 Integration of prior knowledge and network model

The methods, SiPer and ChemPert, proposed in this thesis integrate the prior knowledge about perturbation with a network model considering the initial cellular state to identify signalling perturbations for cellular transition. First, the manually curated perturbation databases of transcriptional signatures are constructed in this thesis. The methods first infer signalling proteins in the databases whose perturbations result in similar transcriptional signatures as desired. Furthermore, since the transcriptional responses are also determined by initial cellular states besides perturbagens, the network model is integrated into the methods to ensure the predictions that are specific to the query cellular state.

The inference of exact signalling paths acting on downstream GRN requires not only the gene expression (protein abundance) information, but also protein activity measures such as phosphorylation levels. However, phosphorylation data before and after cell perturbation is not widely available. Thus, instead of attempting to infer the exact signalling paths that could reach the desired transcriptional signatures, the proposed methods construct databases that establish the relation between signalling perturbations and the corresponding cell responses in terms of transcriptional signatures. Then, a pattern-matching algorithm is developed to detect the similarities between the query signatures and the reference signatures in the database and identify the potential candidates inducing the desired transcriptional signatures. Indeed, taking advantage of prior knowledge not only circumvent the complex inference of signalling paths, but also significantly improves the predictive power of proposed methods as shown in the Results section.

The initial cell state/type is important for the faithful generation of the target cells, as well as the development of precise therapeutic strategy. Hence, the proposed methods of this thesis further integrate gene expression of the initial cellular state into a network model. Although gene expression data are doubtful to represent protein activity directly, it still can capture cell state/type-specific information on top of signalling perturbations inferred from prior knowledge. In this thesis, different network models are designed for SiPer and ChemPert by leveraging the characterizations of different kinds of gene expression data, including

scRNA-seq and bulk RNA-seq data. SiPer is designed to predict cell state-specific chemical compounds specifically targeting desired sets of TFs required for different kinds of cell conversions. This ranges from cell type conversions, including cell subtypes, to cell phenotypic changes, such as different functional or morphological states. Therefore, the network-based model of SiPer uses scRNA-seq data by leveraging the heterogeneity of single cells to measure the interactions between signalling proteins across cells. SiPer simulates signalling transduction and draws on the concept of signalling entropy by taking the advantage of scRNA-seq. On the other hand, ChemPert is sought to identify drugs that can revert the cellular disease phenotype to their healthy counterparts, which can be broadly captured by bulk RNA-seq. Compared to scRNA-seq, bulk RNA-seq blurs the heterogeneity across cells and therefore we cannot use it to measure the interaction strength between two connective genes in the network as scRNA-seq. Instead of simulating the signalling transduction in this case, we simply select the paths that are enriched for highly expressed genes.

After validating the capacity of our integrative methods that combines the prior knowledge with a network model in predicting chemical compounds using benchmarking datasets with known perturbations, SiPer and ChemPert are applied to cellular conversion and disease treatment respectively. SiPer recapitulates a considerable number of experimentally validated chemical compounds in diverse cellular conversion experiments, including the generation of functional cell subtypes or distinct phenotypic states of a same cell type. Moreover, by applying SiPer, we successfully develop a novel and efficient protocol to drive the conversion of hepatic progenitors into functional human induced hepatocytes, resembling primary hepatocytes in cellular identity and functionality. ChemPert is applied to different types of diseases, including age-related and infectious diseases, and predicted current state-of-the-art therapeutics alongside potentially novel candidates. These results demonstrate that integration of the prior knowledge and a network model shows superior performance in identifying chemical compounds for different purposes. The particular advantages and limitations of this integrative strategy will be discussed in the following.

5.1.1 Advantages

The presented methods offer several advantages in identifying chemical compounds by integrating the prior knowledge with a network model. First, unlike the existing perturbation databases that mainly focus on cancer cells, we manually curate large-scale perturbation databases solely consisting of non-cancer cells. These databases provide as important resources

to study signalling perturbations for non-cancer cells, as opposed to cancer cells, which often exhibit significant differences in the signalling mechanisms. Indeed, using these non-cancer perturbation databases significantly improves the predictive performance in comparison to using the catalog of cancer data as shown in this thesis. Therefore, these non-cancer perturbation databases will be benefit for a wide range of scientific communities. Second, the transcriptional signatures are booleanized (inhibition or activation) in our databases. While Boolean values are a rather coarse approximation of transcriptional responses, their utility in describing cellular phenotypes and transitions upon perturbation has been highlighted previously (Albert and Othmer, 2003). Moreover, since the perturbation datasets in our databases are collected from different resources and profiled using different sequencing platforms, the Boolean transcriptional signatures can decrease the technical biases so that the predictions are more robust. Third, the existing compendium-based methods to infer chemical compounds only consider the transcriptional signatures without taking the initial cellular state into account. In contrast, our methods consider both which increase the specificity of the predictions. Indeed, generating one cell type from different starting cell types/states requires different combinations of chemical compounds. For example, the differentiation of beta cell-like cells from MSCs requires two main stages: adding nicotinamide in high glucose culture to generate pancreatic progenitors and then maturation to beta-like cells by using nicotinamide together with exendin-4 (Xin et al., 2016). However, ESCs or iPSCs induced beta cells need three to five differentiation steps with different chemical cocktails targeting specific signalling pathways at each stage (Pavathuparambil Abdul Manaph et al., 2019).

5.1.2 Limitations

While the proposed methods have shown high accuracy by taking advantage of transcriptome signatures of normal cell types, they still have some limitations. First of all, the number of unique perturbagens and total datasets is much smaller than the database LINCS L1000 (mainly including cancer cells). This is due to the fact that current available perturbation experiments mainly focus on cancer cells. We believe that the performance of our methods can be further improved by generating and adding more perturbation datasets, especially perturbagens that are not included in our current databases. In addition, due to the scarcity of perturbation datasets with different concentrations, the proposed methods are not able to predict the optimal concentrations of chemical compounds. Therefore, these methods solely enable the prioritization of the promising chemical compounds for further experimental screening.

Nevertheless, our comprehensive analysis suggest that the developed methods present high value for significantly narrowing down the number of screened chemical compounds for cell engineering and disease treatment. The generation of more datasets and subsequent modelling of dose-dependent transcriptional responses constitutes another future challenge for these methods. Finally, the established experimental protocols often involve multiple chemical compounds and the combination of these compounds with synergistic effect normally increase the efficiency of cellular conversion (Cao et al., 2016; Kunisada et al., 2012). However, the synergistic or redundant effects among the signalling cues is not considered in our methods. Even though SiPer allows user to design a combination of perturbagens covering distinct signalling pathways, further development for these methods to model those effects will enable the identification of more optimal combinations of chemical compounds.

5.2 Integration of signalling and gene regulation networks

Another de novo prediction method is devised in this thesis to predict the signalling pathways to induce salamander limb regeneration. The method models the interactions between the upstream signalling network and the downstream GRN to predict signalling perturbations inducing cellular transitions by altering the GRN states, which conceptually differs from existing methods of de nove predictions. The two different regulatory layers are represented with different models which inherently reflect the difference of mechanism and time scale in two layers.

The gene regulation layer is delineated using a Boolean network model. In order to recapitulate the cellular changes required for the desired cellular transition, the GRN is composed of the TFs that shift their expression Boolean states from the initial to the target cellular states. A raw GRN is inferred by integration of different statistical measures in this study, due to the scarcity of the prior knowledge for salamander. For the well-studied organisms, the GRN can also be obtained from database containing manually curated transcriptional interactions from literature, for example MetaCore. To obtain a network that does reflect the corresponding phenotype, the inferred network needs to be contextualized to the booleanized gene expression data of the initial and desired cellular states (Crespo et al., 2013). We employ a Boolean model, as it represent a simple yet powerful approach to modelling GRNs (Albert and Othmer, 2003). Although continuous models can model gene expression dynamics precisely, it requires a large amount of transcriptome data for the parameter inference, which often takes a huge amount of time even for small networks. The

cellular transition is modeled by GRN state transition from the initial to the target state by perturbing interface TFs in silico. Then the interface TFs that are most likely to induce GRN state transition are identified.

The signalling regulatory layer models the signal transduction using a probabilistic model. While gene expression data cannot represent protein activity directly, transcriptomics data still can gain insights into signalling pathways to some extent. In this thesis, we aim at inducing the cellular transitions between two cellular states. Therefore, we can postulate that the proteins that transmit the signal should present in the initial stable state and therefore their gene expression should be detectable. Therefore, in the initial cellular state, the probability of a gene being expressed is used to measure the probability of the signalling protein being exploited for signal transduction. We simulate the signal transduction along the most probably expressed signalling paths (MPPs) and identify the signalling proteins that are most specific and likely to transmit the signal to the inferred interface TFs in gene regulatory layer.

5.2.1 Advantages

The main contribution of this method is the introduction of a general method that integrates the signalling and gene regulatory networks to systematically describe and predict key signalling pathways to induce cellular transitions. This method explicitly models the regulatory activity of signalling on transcription by considering the transition between GRN states corresponding to the initial and target cellular phenotypes. To our knowledge, no existing method considers the change of GRN caused by the signalling perturbation, but solely rely on downstream dysregulated genes. Furthermore, the main application of this framework is to identify signalling pathways inducing salamander limb regeneration. All previous bioinformatic studies on salamander molecular interactions relied on orthologues to other model organisms, such as human or mouse, as this information is hardly available in axolotls. However, this can result in the missing of interactions that are specific to the regeneration, since none of the model organisms is known to be able to regenerate their limbs in the same way as salamander does. Therefore, we reconstruct a salamander regeneration-specific GRN using the blastema specific time-series microarray dataset. Notably, this dataset is the first time-series gene expression profile specific to salamander regeneration in the connective tissue cells. The predicted signaling pathways and molecules for salamander limb regeneration were substantially consistent with the literature. In conclusion, the integration of the signalling network and the

GRN allows recapitulating signalling pathways and molecules for cellular transitions by explicitly modelling the regulatory activity of signalling on transcription.

5.2.2 Limitations

For the method that integrates the signalling network into the GRN, the Boolean state and expression probability are required to be estimated as accurately as possible. These values play important roles in the whole pipeline since they are used to contextualize the signalling network and GRN for the cellular transition. While a data-driven and platform-free approach is used in the proposed method, it requires replicates for each cellular state and a certain number of DEGs. Alternatively, for well-studied organisms, the databases including a large number of RNA-seq experiments can be used to define gene-specific expression distributions more precisely and several such databases indeed already exist for human, mouse and rat (Lachmann et al., 2018; Söllner et al., 2017). However, they are not suitable for the study of salamander. In addition, another limitation of the present method is that the simulation of the combined perturbation of interface TFs in-silico is very computationally expensive. Here, the effects of combinations on the GRN state are calculated exhaustively, including the low-efficiency TFs, since these TFs might have synergistic effects together with other TFs that induce significant change of the GRN state. As a consequence, only combinations of up to 4 interface TFs will be considered to calculate the frequency of the interface TFs. The size of combination can be increased if the computing is efficient. In addition, this method cannot be applied to the transitions between very similar cellular states, where the TFs with gene expression state change in cellular transition are not able to form a connected GRN with a reasonable size (>=10 TFs).

5.3 Specificity of signalling perturbation

The signalling pathways are often highly interconnected, which results in the potential for the undesirable crosstalk between pathways. A major drawback of perturbing signalling cues to induce the desired cellular transition is the triggering of unintended downstream response. For example, many drugs show side effects or cause preclinical and clinical toxic events due to acting on undesired biological targets (Whitebread et al., 2005). Hence, it is challenge to identify signalling perturbations acting on specific sets of targets, while minimizing the off-target effects. Indeed, this is a difficult task, as downstream target genes of each signalling perturbation in each cell type are largely unknown and experimental screening for every cell type would be too resource intensive and impractical. In this thesis, the proposed computational

methods measure how specifically the predicted signalling molecules target desired genes in addition to considering reachability between the signalling molecules and the desired targets. The specificity of each signalling molecule is measured by computing the similarity between its downstream targets and the desired targets by using Jensen-Shannon divergence. By doing so, we prioritize the signalling molecules specifically act on the intended targets, while have a minimized effect on undesirable ones.

5.4 Outlook

In the future, the subsequent research of this thesis could focus on two aspects, 1) the further development of computational frameworks due to the aforementioned limitations and 2) experimental validation for the predictions to proof the general applicability of the presented methods.

5.4.1 Improvement of computational frameworks

5.4.1.1 Prediction of the combination of signalling proteins

One limitation of the proposed methods in this thesis is that they did not consider the potential combined effect of multiple signalling protein targets of chemical compound on downstream targets. However, as mentioned above, the chemical cocktails of cellular conversions normally involve multiple chemical compounds and the optimal combination of these compounds with synergistic effect can reach higher efficiency for cell engineering (Cao et al., 2016; Kunisada et al., 2012). Therefore, the ability to predict combination effects of signalling perturbations can increase the usability of our methods. An intuitive strategy is to predict combinations of signalling proteins based on their addictive effect on the downstream genes, which was not successful in our initial trial. This suggests that a deeper insight into the interparty of different signal transduction paths to model their combination is required. Especially, it is necessary to consider the extensive cross-talk between signal transduction paths. To extend the current methods that take the signalling cross-talk into account, it would be great helpful to learn signal transduction from phospho-proteomics data for combination of perturbations. For example, different molecules might use the same path or different paths to have a synergistic or redundant effect on interface TFs. Using the signal transduction pattern inferred from phospho-proteomics data, we could mimic the signal transduction in silico using the mathematical model. This strategy could be hindered by the scarcity of phospho-proteomics data of combined

perturbations. Alternatively, instead of simulating the exact signal transduction, we might predict the combination of molecules relying on the transcriptional responses, which is the most widely available data to date. Machine learning/ deep learning algorithms could be applied since they might capture response patterns for the combination of perturbations more accurately when a large scale of transcriptomics data is available.

5.4.1.2 Integration of other regulatory layers

Cellular transition can be induced by multiple regulatory layers, including cells metabolic or epigenetic modifications in addition to the signalling and transcriptional regulations that have been considered in this thesis. Therefore, the metabolic and epigenetic regulations can be integrated in the current models of this thesis, allowing the discovery of more different types of interventions.

Metabolism has orchestrated interplay with singalling and transcriptional regulations. While metabolism is regulated by signalling and transcriptional regulators, metabolism has also been shown to regulate gene expression through metabolic enzymes and metabolites, which can affect the chromatin directly or indirectly (Li et al., 2018a). Moreover, metabolism also exhibits feedback regulatory effects on signalling pathways, such as mTOR, AMPK and p53 (Lorendeau et al., 2015). Therefore, integrating metabolic network with signalling network could potentially identify molecules have effects on downstream GRN more comprehensively.

To date, chemical compounds that modulate the cellular epigenetic landscape have also been used for cellular reprogramming (Shi et al., 2008; Tran et al., 2015) and disease treatment (Wouters and Delwel, 2016). These compounds regulate the activity of DNA methyltransferases or enzymes for histone-modifications, such as acetyltransferases, methyltransferases and deacetylases. Therefore, epigenetic data, such as Chip-seq and ATAC-seq, could be integrated into the proposed methods. In particular, we can compare the epigenetic profiles of two cellular state and then combine the change of epigenetic level with the change of transcriptomic level to construct a more comprehensive GRN to represent the cellular phenotype. In additional, we could also try to predict the chemical compounds that are able to induce the change of the accessibility or activity state of the regulatory regions of the specific TFs that can trigger the cellular state transitions.

5.4.2 Experimental validation for the computational predictions

5.4.2.1 Experimental validation of chemical compounds for disease treatment

In this thesis, we have experimentally validated that the chemical compounds predicted by SiPer can induce the maturation of hepatocytes. While extensive literatures confirm the predictive power of the ChemPert for the disease treatment, a solid experimental validation will be carried out in collaboration with Prof. Dr. Jan Rehwinkel in Oxford University. ChemPert will be used to predict chemical compounds as replacement of interferons (IFNs) in anti-viral immunotherapy. IFNs are cytokines that have strong ability to promote the immune response to confer consistent to viral infections. However, the limited production of IFNs hinders their use in clinical broadly. In addition, IFNs affect a wide range of cell types, resulting in side effects for patients (McNab et al., 2015). Therefore, it would be of clinical importance to identify chemical compounds addressing these limitations, while existing the same or similar therapeutic effect as IFNs. We have identified a set of chemical compounds which potentially have similar effects as IFNs using transcriptomics data before and after treatment of interferon-β (IFNb) in monocytes. The top 20 predicted chemical compounds will be used for further experimental validation. Notably, the top three predictions, poly I:C, emetine and azacytidine, have been shown to induce IFNs and interferon stimulated genes (Jeong et al., 2015; Raj and Pitha, 1981; Schellekens et al., 1975). This experimental validation will further the predictive power of ChemPert.

5.4.2.2 Experimental validation of predictions for salamander limb regeneration

We have shown that our computational method correctly recapitulates the key signalling pathways along the limb regeneration process from literatures. To further validate the predictions and identify novel cues for regeneration, further in vivo experiments are required during salamander limb regeneration. We will collaborate with Prof. Dr. Elly Tanaka and apply three strategies that have been developed in her lab to modulate gene expression. The first is CRISPR-mediate deletions of the gene after injection of Cas9 and gRNA into the egg. This was previously shown to be successful in yielding regeneration-specific knockout phenotypes (Fei et al., 2014). Second, morpholinos will be electroporated into the regenerating limb (Mercader et al., 2005). Third, baculovirus mediated gene delivery will be used to express Cas9 and gRNA specifically in connective tissue cells to obtain cell-type specific knockout of the candidate gene. For overexpression of genes, baculovirus mediated gene over-expression will be used. Notably, Tanaka's lab has already made a series of transgenic lines for canonical Wnt,

BMP, Notch or TGF-B that, based on a tamoxifen-inducible cre activity, could overexpress a dominant negative or constitutively active member of the pathways in cell types of interest. In addition, we will also identify chemical compounds which can block or activate signalling pathways to validate the role of canonical signalling pathways.

The effect of the gene perturbation or chemical perturbation will be evaluated using functional assays. First, the morphology of the regenerating tissue will be imaged for live limb to determine if the tempo and overall morphology are normal or perturbed (Kragl and Tanaka, 2009). In our experiments, we will aim not only to inhibit regeneration but also to activate signalling pathways and genes that should accelerate the timing of regeneration. Those samples that show an overall suppression/delay or acceleration of regeneration will be examined in more detail for cellular parameters. The samples will be examined for the number and timing of the accumulation of connective tissue cells at the amputation plane as an assay for cell migration to the wound. BrdU labelling will be used to assay effects on proliferation. To assay whether connective tissue cells turn on a limb blastema expression profile, immunohistochemistry will be employed by using a series of antibodies generated against a number of axolotls blastema markers, including prrx1, msx1, hoxA9, hoxA11, hoxA13, and meis. Perturbations of genes or modulation of predicted key signalling pathways that are important for cell decision to become a blastema cell should alter the induction of at least one of the genes. Through such an analysis we should be able to validate the novel predictions and link specific signalling pathways and GRNs with specific cellular behaviors required for blastema formation.

5.5 Conclusion

The induction of cellular transition, including the change of cellular identity or phenotype, is of clinical interest, as it allows to derive required cells or tissues for regenerative medicine or to revert cells from disease states to normally functional states. It can be achieved by perturbing multiple regulatory layers, such as the direct ectopic expression of transcriptional regulators, or the fluctuation of signalling transduction which in turn regulates the transcriptional layer. Forced ectopic expression of genes involves the transfer of genetic material, which has raised safety concerns to translate them into clinical applications. In contrast, chemical compound targeting signaling pathways is a safer and more easily controlled and cost-effective strategy for cellular transition. However, to identify optimal chemical compounds, cell reprogramming protocols as well as drug discovery currently rely on exhaustive trial-and-error testing of a large

scale of compounds, which is both inefficient and resource intensive. A systematic guidance with computational methods is beneficial to this process. To date, some of the existing methods predict signalling perturbation relying on databases containing cancer cells, which is not suitable for the inference of non-cancer cells signalling perturbation. Another category of methods focuses on GRN or signalling networks separately, without analysing the effect of signalling perturbations on the GRN. The purpose of this thesis is to develop more systematically computational methods that address the limitations of existing methods. In conclusion, the contributions of this thesis are listed as follows:

- **Perturbation databases solely consisting of non-cancer cells are constructed.** The existing perturbation databases mainly focus on cancer cells. As we known, cancer cells exist widespread signalling rewiring and therefore exhibit significant differences in the signalling mechanisms compared to non-cancer cells. Therefore, it is necessary to construct such databases to understand the signalling perturbation of non-cancer cell. In this thesis, we construct the perturbation databases only invloving non-cancer cells and have shown that using our databases indeed significantly outperforms others for the prediction of signalling perturbation for normal cells. The databases constructed in this thesis will be highly useful resources for a wide range of scientific communities.
- **Integration of the non-cancer cell perturbation database and a network model identifies cell type specific chemical compounds.** Perturbation databases collect transcriptional signatures across different kinds of cell types. While this offers as valuable prior knowledge, some false positive signalling molecules for the corresponding initial cellular state could be inferred from the databases. The integration of network model by considering the gene expression profile of initial cellular state allows the prioritization of cell type/state-specific signalling molecules.
- **Integration of two regulatory layers predicts signalling pathways whose perturbations could induce the cellular transition.** The method applied to salamander limb regeneration integrates two distinct models for signalling and transcriptional regulatory layers. These two regulatory layers are connected by interface TFs. The signal transduction and its effect on the interface TFs are mimicked by using a probabilistic model. The key interface TFs that determine the state of the GRN is identified by simulating the perturbations in silico on a Boolean network model.
- **The predicted signalling molecules are optimized to specifically target downstream GRN or desired TFs.** One key issue needs to be addressed to predict

signalling cues is to circumvent targeting undesired downstream genes. In this thesis, the proposed methods consider the specificity of the predicted signalling cues that act on desired genes, while minimize the effects on non-desired ones. This is achieved calculating the Jensen-Shannon's divergence between the predicted targets of each signalling molecule and the desired targets.

- **Chemical compounds inducing cellular conversions can be consistently recapitulated**. The proposed method in this thesis, SiPer, is applied to chemical-based cellular conversion, including conversion between different cell (sub)types or distinct phenotypic states of a same cell type. A considerable number of experimentally used chemical compounds in corresponding protocols are predicted. In addition, by applying SiPer, we successfully induce the conversion from hepatic progenitors to mature hepatocytes, which not only resemble primary hepatocytes in cellular identity and functionality, but also circumvent the abnormal lipid metabolism in the previous protocol. SiPer can be easily used for experimentalists through a web interface at https://siper.uni.lu.
- **Chemical compounds applied in the state-of-the-art therapeutics for various diseases can be predicted.** The present method of this thesis, ChemPert, is applied to various diseases, including aging-associated diseases and infectious diseases, to identify chemical compounds reverting disease phenotypes to their normal counterparts. The predicted chemical compounds have been shown to have therapeutic effect on the corresponding disease with extensive literature evidences Therefore, by applying the proposed method, the number of screening compounds for the identification of new therapeutic interventions could be reduced substantially. The method is also freely available as a web application at https://siper.uni.lu/chempert.
- **Previous knowledge related to limb regeneration is consistently captured in the predictions.** Previous studies identify the key factors during regeneration largely by trial and error, which requires a large amount of time and resources. In this thesis, we propose another computational method to predict signalling pathways and molecules whose perturbations can induce salamander limb regeneration. The first connective tissue-specific time serious data after amputation of salamander limb is generated in this study. Using this dataset, our method identifies intracellular signalling cues during salamander limb regeneration, such as the initiation of proliferation, migration and patterning, which are consistent with literatures.

In conclusion, the computational methods present in this thesis are of great value to identify signalling perturbations as a replacement for genetic manipulations. They are generally applicable methods that can guide the experimentalists or pharmaceutical researchers to identify chemical compounds, signalling molecules and pathways for the induction of the desired cellular transitions. This ranges from cellular conversions, including differentiation, reprogramming and conversions between phenotypic states in vivo or in vitro, to the reversal of pathological phenotypes, with high potential applications in regenerative medicine and disease treatment.

6 Reference

Abeyrathna, P., and Su, Y. (2015). The critical role of Akt in cardiovascular function. Vascul Pharmacol *74*, 38-48.

Addis, R.C., Ifkovits, J.L., Pinto, F., Kellam, L.D., Esteso, P., Rentschler, S., Christoforou, N., Epstein, J.A., and Gearhart, J.D. (2013). Optimization of direct fibroblast reprogramming to cardiomyocytes using calcium activity as a functional measure of success. J Mol Cell Cardiol *60*, 97-106.

Ahmadzadeh, A. (2014). Papaverine increases human serum albumin glycation. J Biol Phys *40*, 97-107.

Al Haj Baddar, N.W., Chithrala, A., and Voss, S.R. (2019). Amputation-induced reactive oxygen species signaling is required for axolotl tail regeneration. Dev Dyn *248*, 189-196.

Al-Beltagi, S., Preda, C.A., Goulding, L.V., James, J., Pu, J., Skinner, P., Jiang, Z., Wang, B.L., Yang, J., Banyard, A.C*., et al.* (2021). Thapsigargin Is a Broad-Spectrum Inhibitor of Major Human Respiratory Viruses: Coronavirus, Respiratory Syncytial Virus and Influenza A Virus. Viruses *13*.

Albert, R., and Othmer, H.G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol *223*, 1-18.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600-1607.

Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., and Califano, A. (2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet *48*, 838-847.

Angelopoulou, A., Alexandris, N., Konstantinou, E., Mesiakaris, K., Zanidis, C., Farsalinos, K., and Poulas, K. (2020). Imiquimod - A toll like receptor 7 agonist - Is an ideal option for management of COVID 19. Environ Res *188*, 109858.

Anthony, M.S., Clarkson, T.B., and Williams, J.K. (1998). Effects of soy isoflavones on atherosclerosis: potential mechanisms. Am J Clin Nutr *68*, 1390s-1393s.

Arel-Dubeau, A.M., Longpré, F., Bournival, J., Tremblay, C., Demers-Lamarche, J., Haskova, P., Attard, E., Germain, M., and Martinoli, M.G. (2014). Cucurbitacin E has neuroprotective properties and autophagic modulating activities on dopaminergic neurons. Oxid Med Cell Longev *2014*, 425496.

Avior, Y., Sagi, I., and Benvenisty, N. (2016). Pluripotent stem cells in disease modelling and drug discovery. Nat Rev Mol Cell Biol *17*, 170-182.

Baghaki, S., Yalcin, C.E., Baghaki, H.S., Aydin, S.Y., Daghan, B., and Yavuz, E. (2020). COX2 inhibition in the treatment of COVID-19: Review of literature to propose repositioning of celecoxib for randomized controlled studies. Int J Infect Dis *101*, 29-32.

Bao, Z., Li, X., Zan, X., Shen, L., Ma, R., and Liu, W. (2016). Signalling pathway impact analysis based on the strength of interaction between genes. IET Syst Biol *10*, 147-152.

Barh, D., Tiwari, S., Weener, M.E., Azevedo, V., Góes-Neto, A., Gromiha, M.M., and Ghosh, P. (2020). Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19. Comput Biol Med *126*, 104051.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M*., et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res *41*, D991-995.

Basson, M.A. (2012). Signaling in cell differentiation and morphogenesis. Cold Spring Harb Perspect Biol *4*.

Baumgartner, C., Toifl, S., Farlik, M., Halbritter, F., Scheicher, R., Fischer, I., Sexl, V., Bock, C., and Baccarini, M. (2018). An ERK-Dependent Feedback Mechanism Prevents Hematopoietic Stem Cell Exhaustion. Cell Stem Cell *22*, 879-892.e876.
Beck, C.W., Christen, B., and Slack, J.M. (2003). Molecular pathways needed for regeneration of spinal cord and muscle in a vertebrate. Dev Cell *5*, 429-439.
Ben-David, U., and Benvenisty, N. (2011). The tumorigenicity of human embryonic and induced pluripotent stem cells. Nat Rev Cancer *11*, 268-277.
Bernitz, J.M., Kim, H.S., MacArthur, B., Sieburg, H., and Moore, K. (2016). Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. Cell *167*, 1296-1309.e1210.
Berretta, A.A., Silveira, M.A.D., Cóndor Capcha, J.M., and De Jong, D. (2020). Propolis and its potential against SARS-CoV-2 infection mechanisms and COVID-19 disease: Running title: Propolis against SARS-CoV-2 infection and COVID-19. Biomed Pharmacother *131*, 110622.
Biteau, B., Hochmuth, C.E., and Jasper, H. (2011). Maintaining tissue homeostasis: dynamic control of somatic stem cell activity. Cell Stem Cell *9*, 402-411.
Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K.*, et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.
Brockes, J.P., Kumar, A., and Velloso, C.P. (2001). Regeneration as an evolutionary variable. J Anat *199*, 3-11.
Brooks, M.B. (2013). Erlotinib and gefitinib, epidermal growth factor receptor kinase inhibitors, may treat non-cancer-related tumor necrosis factor-α mediated inflammatory diseases. Oncologist *18*, e3-5.
Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods *17*, 159-162.
Bucan, V., Peck, C.T., Nasser, I., Liebsch, C., Vogt, P.M., and Strauß, S. (2018). Identification of axolotl BH3-only proteins and expression in axolotl organs and apoptotic limb regeneration tissue. Biol Open *7*.
Burleigh, M.E., Babaev, V.R., Patel, M.B., Crews, B.C., Remmel, R.P., Morrow, J.D., Oates, J.A., Marnett, L.J., Fazio, S., and Linton, M.F. (2005). Inhibition of cyclooxygenase with indomethacin phenethylamide reduces atherosclerosis in apoE-null mice. Biochem Pharmacol *70*, 334-342.
Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F.B., Pastor-Flores, D., Roma, L.P., Renders, S., Zeisberger, P.*, et al.* (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. Cell *169*, 807-823.e819.
Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., and Collins, J.J. (2014). CellNet: network biology applied to stem cell engineering. Cell *158*, 903-915.
Cao, N., Huang, Y., Zheng, J., Spencer, C.I., Zhang, Y., Fu, J.D., Nie, B., Xie, M., Zhang, M., Wang, H.*, et al.* (2016). Conversion of human fibroblasts into functional cardiomyocytes by small molecules. Science *352*, 1216-1220.
Cao, Y., Xie, L., Shi, F., Tang, M., Li, Y., Hu, J., Zhao, L., Zhao, L., Yu, X., Luo, X.*, et al.* (2021). Targeting the signaling in Epstein-Barr virus-associated diseases: mechanism, regulation, and clinical study. Signal Transduct Target Ther *6*, 15.
Caruso, F., Singh, M., Belli, S., Berinato, M., and Rossi, M. (2020). Interrelated Mechanism by Which the Methide Quinone Celastrol, Obtained from the Roots of Tripterygium wilfordii, Inhibits Main Protease 3CL(pro) of COVID-19 and Acts as Superoxide Radical Scavenger. Int J Mol Sci *21*.

Castelo-Branco, G., Rawal, N., and Arenas, E. (2004). GSK-3beta inhibition/beta-catenin stabilization in ventral midbrain precursors increases differentiation into dopamine neurons. J Cell Sci *117*, 5731-5737.
Caubit, X., Nicolas, S., and Le Parco, Y. (1997). Possible roles for Wnt genes in growth and axial patterning during regeneration of the tail in urodele amphibians. Dev Dyn *210*, 1-10.
Chakraborty, S., Ji, H., Kabadi, A.M., Gersbach, C.A., Christoforou, N., and Leong, K.W. (2014). A CRISPR/Cas9-based system for reprogramming cell lineage specification. Stem Cell Reports *3*, 940-947.
Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol *27*, 275-280.
Chang, K.H., Lee-Chen, G.J., Huang, C.C., Lin, J.L., Chen, Y.J., Wei, P.C., Lo, Y.S., Yao, C.F., Kuo, M.W., and Chen, C.M. (2019). Modeling Alzheimer's Disease by Induced Pluripotent Stem Cells Carrying APP D678H Mutation. Mol Neurobiol *56*, 3972-3983.
Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Chen, J., Wu, Y.*, et al.* (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nat Genet *45*, 34-42.
Chen, J., Sathiyamoorthy, K., Zhang, X., Schaller, S., Perez White, B.E., Jardetzky, T.S., and Longnecker, R. (2018). Ephrin receptor A2 is a functional entry receptor for Epstein-Barr virus. Nat Microbiol *3*, 172-180.
Chen, M., Lu, P., Ma, Q., Cao, Y., Chen, N., Li, W., Zhao, S., Chen, B., Shi, J., Sun, Y.*, et al.* (2020). CTNNB1/β-catenin dysfunction contributes to adiposity by regulating the cross-talk of mature adipocytes and preadipocytes. Sci Adv *6*, eaax9605.
Chen, T., Shen, L., Yu, J., Wan, H., Guo, A., Chen, J., Long, Y., Zhao, J., and Pei, G. (2011). Rapamycin and other longevity-promoting compounds enhance the generation of mouse induced pluripotent stem cells. Aging Cell *10*, 908-911.
Cheung, T.H., Quach, N.L., Charville, G.W., Liu, L., Park, L., Edalati, A., Yoo, B., Hoang, P., and Rando, T.A. (2012). Maintenance of muscle stem-cell quiescence by microRNA-489. Nature *482*, 524-528.
Chiang, C., Litingtung, Y., Lee, E., Young, K.E., Corden, J.L., Westphal, H., and Beachy, P.A. (1996). Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. Nature *383*, 407-413.
Chiba, T., Kondo, Y., Shinozaki, S., Kaneko, E., Ishigami, A., Maruyama, N., Umezawa, K., and Shimokado, K. (2006). A selective NFkappaB inhibitor, DHMEQ, reduced atherosclerosis in ApoE-deficient mice. J Atheroscler Thromb *13*, 308-313.
Cho, Y.L., Min, J.K., Roh, K.M., Kim, W.K., Han, B.S., Bae, K.H., Lee, S.C., Chung, S.J., and Kang, H.J. (2015). Phosphoprotein phosphatase 1CB (PPP1CB), a novel adipogenic activator, promotes 3T3-L1 adipogenesis. Biochem Biophys Res Commun *467*, 211-217.
Chong, R., Wakade, C., Seamon, M., Giri, B., Morgan, J., and Purohit, S. (2021). Niacin Enhancement for Parkinson's Disease: An Effectiveness Trial. Front Aging Neurosci *13*, 667032.
Christman, J.K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene *21*, 5483-5495.
Chua, X.Y., Chai, Y.L., Chew, W.S., Chong, J.R., Ang, H.L., Xiang, P., Camara, K., Howell, A.R., Torta, F., Wenk, M.R.*, et al.* (2020). Immunomodulatory sphingosine-1-phosphates as plasma biomarkers of Alzheimer's disease and vascular cognitive impairment. Alzheimers Res Ther *12*, 122.
Chung, S., Leung, A., Han, B.S., Chang, M.Y., Moon, J.I., Kim, C.H., Hong, S., Pruszak, J., Isacson, O., and Kim, K.S. (2009). Wnt1-lmx1a forms a novel autoregulatory loop and

controls midbrain dopaminergic differentiation synergistically with the SHH-FoxA2 pathway. Cell Stem Cell *5*, 646-658.
Chung, S.T., Huang, Y.T., Hsiung, H.Y., Huang, W.H., Yao, C.W., and Lee, A.R. (2015). Novel daidzein analogs and their in vitro anti-influenza activities. Chem Biodivers *12*, 685-696.
Cieślar-Pobuda, A., Knoflach, V., Ringh, M.V., Stark, J., Likus, W., Siemianowicz, K., Ghavami, S., Hudecki, A., Green, J.L., and Łos, M.J. (2017). Transdifferentiation and reprogramming: Overview of the processes, their similarities and differences. Biochim Biophys Acta Mol Cell Res *1864*, 1359-1369.
Codega, P., Silva-Vargas, V., Paul, A., Maldonado-Soto, A.R., Deleo, A.M., Pastrana, E., and Doetsch, F. (2014). Prospective identification and purification of quiescent adult neural stem cells from their in vivo niche. Neuron *82*, 545-559.
Colasante, G., Lignani, G., Rubio, A., Medrihan, L., Yekhlef, L., Sessa, A., Massimino, L., Giannelli, S.G., Sacchetti, S., Caiazzo, M.*, et al.* (2015). Rapid Conversion of Fibroblasts into Functional Forebrain GABAergic Interneurons by Direct Genetic Reprogramming. Cell Stem Cell *17*, 719-734.
Corciulo, C., Lendhey, M., Wilder, T., Schoen, H., Cornelissen, A.S., Chang, G., Kennedy, O.D., and Cronstein, B.N. (2017). Endogenous adenosine maintains cartilage homeostasis and exogenous adenosine inhibits osteoarthritis progression. Nat Commun *8*, 15019.
Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M.*, et al.* (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med *23*, 405-408.
Crespo, I., and Del Sol, A. (2013). A general strategy for cellular reprogramming: the importance of transcription factor cross-repression. Stem Cells *31*, 2127-2135.
Crespo, I., Perumal, T.M., Jurkowski, W., and del Sol, A. (2013). Detecting cellular reprogramming determinants by differential stability analysis of gene regulatory networks. BMC Syst Biol *7*, 140.
Czubowicz, K., Jęśko, H., Wencel, P., Lukiw, W.J., and Strosznajder, R.P. (2019). The Role of Ceramide and Sphingosine-1-Phosphate in Alzheimer's Disease and Other Neurodegenerative Disorders. Mol Neurobiol *56*, 5436-5455.
D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M.*, et al.* (2015). A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity. Stem Cell Reports *5*, 763-775.
Dai, P., Harada, Y., and Takamatsu, T. (2015). Highly efficient direct conversion of human fibroblasts to neuronal cells by chemical compounds. J Clin Biochem Nutr *56*, 166-170.
Dai, S., Wang, B., Li, W., Wang, L., Song, X., Guo, C., Li, Y., Liu, F., Zhu, F., Wang, Q.*, et al.* (2016). Systemic application of 3-methyladenine markedly inhibited atherosclerotic lesion in ApoE(-/-) mice by modulating autophagy, foam cell formation and immune-negative molecules. Cell Death Dis *7*, e2498.
Dao, T.T., Nguyen, P.H., Lee, H.S., Kim, E., Park, J., Lim, S.I., and Oh, W.K. (2011). Chalcones as novel influenza A (H1N1) neuraminidase inhibitors from Glycyrrhiza inflata. Bioorg Med Chem Lett *21*, 294-298.
Darnet, S., Dragalzew, A.C., Amaral, D.B., Sousa, J.F., Thompson, A.W., Cass, A.N., Lorena, J., Pires, E.S., Costa, C.M., Sousa, M.P.*, et al.* (2019). Deep evolutionary origin of limb and fin regeneration. Proc Natl Acad Sci U S A *116*, 15106-15115.
Dastan, Z., Pouramir, M., Ghasemi-Kasman, M., Ghasemzadeh, Z., Dadgar, M., Gol, M., Ashrafpour, M., Pourghasem, M., Moghadamnia, A.A., and Khafri, S. (2019). Arbutin reduces cognitive deficit and oxidative stress in animal model of Alzheimer's disease. Int J Neurosci *129*, 1145-1153.

De, D., Halder, D., Shin, I., and Kim, K.K. (2017). Small molecule-induced cellular conversion. Chem Soc Rev *46*, 6241-6254.
Dhanya, R., Arya, A.D., Nisha, P., and Jayamurthy, P. (2017). Quercetin, a Lead Compound against Type 2 Diabetes Ameliorates Glucose Uptake via AMPK Pathway in Skeletal Muscle Cell Line. Front Pharmacol *8*, 336.
Dimos, J.T., Rodolfa, K.T., Niakan, K.K., Weisenthal, L.M., Mitsumoto, H., Chung, W., Croft, G.F., Saphier, G., Leibel, R., Goland, R.*, et al.* (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science *321*, 1218-1221.
Ding, Y., Kong, D., Zhou, T., Yang, N.D., Xin, C., Xu, J., Wang, Q., Zhang, H., Wu, Q., Lu, X.*, et al.* (2020). α-Arbutin Protects Against Parkinson's Disease-Associated Mitochondrial Dysfunction In Vitro and In Vivo. Neuromolecular Med *22*, 56-67.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
Dolan, C.P., Dawson, L.A., and Muneoka, K. (2018). Digit Tip Regeneration: Merging Regeneration Biology with Regenerative Medicine. Stem Cells Transl Med *7*, 262-270.
Dunn, S.J., Martello, G., Yordanov, B., Emmott, S., and Smith, A.G. (2014). Defining an essential transcription factor program for naïve pluripotency. Science *344*, 1156-1160.
Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184-1191.
Dutta, B., Wallqvist, A., and Reifman, J. (2012). PathNet: a tool for pathway analysis using topological information. Source Code Biol Med *7*, 10.
Eder, J., Sedrani, R., and Wiesmann, C. (2014). The discovery of first-in-class drugs: origins and evolution. Nat Rev Drug Discov *13*, 577-587.
Espay, A.J., LeWitt, P.A., and Kaufmann, H. (2014). Norepinephrine deficiency in Parkinson's disease: the case for noradrenergic enhancement. Mov Disord *29*, 1710-1719.
Esteban, M.A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S.*, et al.* (2010). Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. Cell Stem Cell *6*, 71-79.
Eyal, S. (2018). The Fever Tree: from Malaria to Neurological Diseases. Toxins (Basel) *10*.
Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol *5*, e8.
Fakhry, C.T., Choudhary, P., Gutteridge, A., Sidders, B., Chen, P., Ziemek, D., and Zarringhalam, K. (2016). Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. BMC Bioinformatics *17*, 318.
Farkas, J.E., Freitas, P.D., Bryant, D.M., Whited, J.L., and Monaghan, J.R. (2016). Neuregulin-1 signaling is essential for nerve-dependent axolotl limb regeneration. Development *143*, 2724-2731.
Federation, A.J., Bradner, J.E., and Meissner, A. (2014). The use of small molecules in somatic-cell reprogramming. Trends Cell Biol *24*, 179-187.
Fei, J.F., Schuez, M., Tazaki, A., Taniguchi, Y., Roensch, K., and Tanaka, E.M. (2014). CRISPR-mediated genomic deletion of Sox2 in the axolotl shows a requirement in spinal cord neural stem cell amplification during tail regeneration. Stem Cell Reports *3*, 444-459.
Fisher, M.C., Clinton, G.M., Maihle, N.J., and Dealy, C.N. (2007). Requirement for ErbB2/ErbB signaling in developing cartilage and bone. Dev Growth Differ *49*, 503-513.

Fogarty, M.J., Marin Mathieu, N., Mantilla, C.B., and Sieck, G.C. (2020). Aging reduces succinate dehydrogenase activity in rat type IIx/IIb diaphragm muscle fibers. J Appl Physiol (1985) *128*, 70-77.
Foitzik, K., Paus, R., Doetschman, T., and Dotto, G.P. (1999). The TGF-beta2 isoform is both a required and sufficient inducer of murine hair follicle morphogenesis. Dev Biol *212*, 278-289.
Francesconi, M., Di Stefano, B., Berenguer, C., de Andrés-Aguayo, L., Plana-Carmona, M., Mendez-Lago, M., Guillaumet-Adkins, A., Rodriguez-Esteban, G., Gut, M., Gut, I.G.*, et al.* (2019). Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. Elife *8*.
Fu, Y., Huang, C., Xu, X., Gu, H., Ye, Y., Jiang, C., Qiu, Z., and Xie, X. (2015). Direct reprogramming of mouse fibroblasts into cardiomyocytes with chemical cocktails. Cell Res *25*, 1013-1024.
Fukushima, T. (2005). Niacin metabolism and Parkinson's disease. Environ Health Prev Med *10*, 3-8.
Fukushima, T., Tanaka, Y., Hamey, F.K., Chang, C.H., Oki, T., Asada, S., Hayashi, Y., Fujino, T., Yonezawa, T., Takeda, R.*, et al.* (2019). Discrimination of Dormant and Active Hematopoietic Stem Cells by G(0) Marker Reveals Dormancy Regulation by Cytoplasmic Calcium. Cell Rep *29*, 4144-4158.e4147.
Gao, F., Wu, D.Q., Hu, Y.H., Jin, G.X., Li, G.D., Sun, T.W., and Li, F.J. (2008). In vitro cultivation of islet-like cell clusters from human umbilical cord blood-derived mesenchymal stem cells. Transl Res *151*, 293-302.
Gao, J., Ding, Y., Wang, Y., Liang, P., Zhang, L., and Liu, R. (2021). Oroxylin A is a severe acute respiratory syndrome coronavirus 2-spiked pseudotyped virus blocker obtained from Radix Scutellariae using angiotensin-converting enzyme II/cell membrane chromatography. Phytother Res *35*, 3194-3204.
Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res *29*, 1363-1375.
Geng, J., Liu, W., Gao, J., Jiang, C., Fan, T., Sun, Y., Qin, Z.H., Xu, Q., Guo, W., and Gao, J. (2019). Andrographolide alleviates Parkinsonism in MPTP-PD mice via targeting mitochondrial fission mediated by dynamin-related protein 1. Br J Pharmacol *176*, 4574-4591.
Gerber, T., Murawala, P., Knapp, D., Masselink, W., Schuez, M., Hermann, S., Gac-Santel, M., Nowoshilow, S., Kageyama, J., Khattak, S.*, et al.* (2018). Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science *362*.
Gharagozloo, P., Lazareno, S., Popham, A., and Birdsall, N.J. (1999). Allosteric interactions of quaternary strychnine and brucine derivatives with muscarinic acetylcholine receptors. J Med Chem *42*, 438-445.
Giri, B., Belanger, K., Seamon, M., Bradley, E., Purohit, S., Chong, R., Morgan, J.C., Baban, B., and Wakade, C. (2019). Niacin Ameliorates Neuro-Inflammation in Parkinson's Disease via GPR109A. Int J Mol Sci *20*.
Gnad, T., Navarro, G., Lahesmaa, M., Reverte-Salisa, L., Copperi, F., Cordomi, A., Naumann, J., Hochhäuser, A., Haufs-Brusberg, S., Wenzel, D.*, et al.* (2020). Adenosine/A2B Receptor Signaling Ameliorates the Effects of Aging and Counteracts Obesity. Cell Metab *32*, 56-70.e57.
Godwin, J.W., Pinto, A.R., and Rosenthal, N.A. (2013). Macrophages are required for adult salamander limb regeneration. Proc Natl Acad Sci U S A *110*, 9415-9420.
Godwin, J.W., and Rosenthal, N. (2014). Scar-free wound healing and regeneration in amphibians: immunological influences on regenerative success. Differentiation *87*, 66-75.

Goswami, R., Gershburg, S., Satorius, A., and Gershburg, E. (2012). Protein kinase inhibitors that inhibit induction of lytic program and replication of Epstein-Barr virus. Antiviral Res *96*, 296-304.
Goulding, L.V., Yang, J., Jiang, Z., Zhang, H., Lea, D., Emes, R.D., Dottorini, T., Pu, J., Liu, J., and Chang, K.C. (2020). Thapsigargin at Non-Cytotoxic Levels Induces a Potent Host Antiviral Response that Blocks Influenza A Virus Replication. Viruses *12*.
Granados, K., Poelchen, J., Novak, D., and Utikal, J. (2020). Cellular Reprogramming-A Model for Melanoma Cellular Plasticity. Int J Mol Sci *21*.
Guo, S., Jiang, Q., Chen, L., and Guo, D. (2016). Gene regulatory network inference using PLS-based methods. BMC Bioinformatics *17*, 545.
Gustavsen, J.A., Pai, S., Isserlin, R., Demchak, B., and Pico, A.R. (2019). RCy3: Network biology using Cytoscape from within R. F1000Res *8*, 1774.
Ha, Y.J., Choi, Y.S., Oh, Y.R., Kang, E.H., Khang, G., Park, Y.B., and Lee, Y.J. (2021). Fucoxanthin Suppresses Osteoclastogenesis via Modulation of MAP Kinase and Nrf2 Signaling. Mar Drugs *19*.
Hamoda, A.M., Fayed, B., Ashmawy, N.S., El-Shorbagi, A.A., Hamdy, R., and Soliman, S.S.M. (2021). Marine Sponge is a Promising Natural Source of Anti-SARS-CoV-2 Scaffold. Front Pharmacol *12*, 666664.
Han, D.W., Tapia, N., Hermann, A., Hemmer, K., Höing, S., Araúzo-Bravo, M.J., Zaehres, H., Wu, G., Frank, S., Moritz, S.*, et al.* (2012). Direct reprogramming of fibroblasts into neural stem cells by defined factors. Cell Stem Cell *10*, 465-472.
Han, J., Kim, S.J., Ryu, M.J., Jang, Y., Lee, M.J., Ju, X., Lee, Y.L., Cui, J., Shong, M., Heo, J.Y.*, et al.* (2019). Chloramphenicol Mitigates Oxidative Stress by Inhibiting Translation of Mitochondrial Complex I in Dopaminergic Neurons of Toxin-Induced Parkinson's Disease Model. Oxid Med Cell Longev *2019*, 4174803.
Han, L.P., Li, C.J., Sun, B., Xie, Y., Guan, Y., Ma, Z.J., and Chen, L.M. (2016). Protective Effects of Celastrol on Diabetic Liver Injury via TLR4/MyD88/NF-κB Signaling Pathway in Type 2 Diabetic Rats. J Diabetes Res *2016*, 2641248.
Hankenson, K.D., Dishowitz, M., Gray, C., and Schenker, M. (2011). Angiogenesis in bone regeneration. Injury *42*, 556-561.
Haridhasapavalan, K.K., Borgohain, M.P., Dey, C., Saha, B., Narayan, G., Kumar, S., and Thummer, R.P. (2019). An insight into non-integrative gene delivery approaches to generate transgene-free induced pluripotent stem cells. Gene *686*, 146-159.
Haudenschild, D.R., Carlson, A.K., Zignego, D.L., Yik, J.H.N., Hilmer, J.K., and June, R.K. (2019). Inhibition of early response genes prevents changes in global joint metabolomic profiles in mouse post-traumatic osteoarthritis. Osteoarthritis Cartilage *27*, 504-512.
Haury, A.C., Mordelet, F., Vera-Licona, P., and Vert, J.P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. BMC Syst Biol *6*, 145.
He, Q., Ding, G., Zhang, M., Nie, P., Yang, J., Liang, D., Bo, J., Zhang, Y., and Liu, Y. (2021). Trends in the Use of Sphingosine 1 Phosphate in Age-Related Diseases: A Scientometric Research Study (1992-2020). J Diabetes Res *2021*, 4932974.
Hirsch, T., Rothoeft, T., Teig, N., Bauer, J.W., Pellegrini, G., De Rosa, L., Scaglione, D., Reichelt, J., Klausegger, A., Kneisz, D.*, et al.* (2017). Regeneration of the entire human epidermis using transgenic stem cells. Nature *551*, 327-332.
Ho, D.M., and Whitman, M. (2008). TGF-beta signaling is required for multiple processes during Xenopus tail regeneration. Dev Biol *315*, 203-216.
Hockemeyer, D., and Jaenisch, R. (2016). Induced Pluripotent Stem Cells Meet Genome Editing. Cell Stem Cell *18*, 573-586.

Hodos, R., Zhang, P., Lee, H.C., Duan, Q., Wang, Z., Clark, N.R., Ma'ayan, A., Wang, F., Kidd, B., Hu, J.*, et al.* (2018). Cell-specific prediction and application of drug-induced gene expression profiles. Pac Symp Biocomput *23*, 32-43.
Hong, C.T., Chan, L., and Bai, C.H. (2020). The Effect of Caffeine on the Risk and Progression of Parkinson's Disease: A Meta-Analysis. Nutrients *12*.
Hori, K., Cholewa-Waclaw, J., Nakada, Y., Glasgow, S.M., Masui, T., Henke, R.M., Wildner, H., Martarelli, B., Beres, T.M., Epstein, J.A.*, et al.* (2008). A nonclassical bHLH Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling. Genes Dev *22*, 166-178.
Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K.*, et al.* (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science *341*, 651-654.
Hou, Y., Ge, S., Li, X., Wang, C., He, H., and He, L. (2021). Testing of the inhibitory effects of loratadine and desloratadine on SARS-CoV-2 spike pseudotyped virus viropexis. Chem Biol Interact *338*, 109420.
Hu, W., Qiu, B., Guan, W., Wang, Q., Wang, M., Li, W., Gao, L., Shen, L., Huang, Y., Xie, G.*, et al.* (2015). Direct Conversion of Normal and Alzheimer's Disease Human Fibroblasts into Neuronal Cells by Small Molecules. Cell Stem Cell *17*, 204-212.
Huang, D.D., Yan, X.L., Fan, S.D., Chen, X.Y., Yan, J.Y., Dong, Q.T., Chen, W.Z., Liu, N.X., Chen, X.L., and Yu, Z. (2020). Nrf2 deficiency promotes the increasing trend of autophagy during aging in skeletal muscle: a potential mechanism for the development of sarcopenia. Aging (Albany NY) *12*, 5977-5991.
Huang, P., He, Z., Ji, S., Sun, H., Xiang, D., Liu, C., Hu, Y., Wang, X., and Hui, L. (2011). Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. Nature *475*, 386-389.
Huang, S., Jiang, L., Cheon, I.S., and Sun, J. (2019). Targeting Peroxisome Proliferator-Activated Receptor-Gamma Decreases Host Mortality After Influenza Infection in Obese Mice. Viral Immunol *32*, 161-169.
Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*.
Ieda, M., Fu, J.D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G., and Srivastava, D. (2010). Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. Cell *142*, 375-386.
Imokawa, Y., and Brockes, J.P. (2003). Selective activation of thrombin is a critical determinant for vertebrate lens regeneration. Curr Biol *13*, 877-881.
Indari, O., Jakhmola, S., Manivannan, E., and Jha, H.C. (2021). An Update on Antiviral Therapy Against SARS-CoV-2: How Far Have We Come? Front Pharmacol *12*, 632677.
Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533-538.
Jeong, H.Y., Kang, W.S., Hong, M.H., Jeong, H.C., Shin, M.G., Jeong, M.H., Kim, Y.S., and Ahn, Y. (2015). 5-Azacytidine modulates interferon regulatory factor 1 in macrophages to exert a cardioprotective effect. Sci Rep *5*, 15768.
Jimenez, M.A., Akerblad, P., Sigvardsson, M., and Rosen, E.D. (2007). Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. Mol Cell Biol *27*, 743-757.
Jin, J., Chen, S., Wang, D., Chen, Y., Wang, Y., Guo, M., Zhou, C., and Dou, J. (2018). Oroxylin A suppresses influenza A virus replication correlating with neuraminidase inhibition and induction of IFNs. Biomed Pharmacother *97*, 385-394.
Jung, S., Potapov, I., Chillara, S., and Del Sol, A. (2021). Leveraging systems biology for predicting modulators of inflammation in patients with COVID-19. Sci Adv *7*.

Kandwal, S., and Fayne, D. (2020). Repurposing drugs for treatment of SARS-CoV-2 infection: computational design insights into mechanisms of action. J Biomol Struct Dyn, 1-15.
Kapoor, N., Ghorai, S.M., Kushwaha, P.K., Shukla, R., Aggarwal, C., and Bandichhor, R. (2020). Plausible mechanisms explaining the role of cucurbitacins as potential therapeutic drugs against coronavirus 2019. Inform Med Unlocked *21*, 100484.
Kawakami, Y., Rodriguez Esteban, C., Raya, M., Kawakami, H., Martí, M., Dubova, I., and Izpisúa Belmonte, J.C. (2006). Wnt/beta-catenin signaling regulates vertebrate limb regeneration. Genes Dev *20*, 3232-3237.
Khalili, N., Karimi, A., Moradi, M.T., and Shirzad, H. (2018). In vitro immunomodulatory activity of celastrol against influenza A virus infection. Immunopharmacol Immunotoxicol *40*, 250-255.
Kim, J.B., and Spiegelman, B.M. (1996). ADD1/SREBP1 promotes adipocyte differentiation and gene expression linked to fatty acid metabolism. Genes Dev *10*, 1096-1107.
Kim, N.Y., Kohn, J.C., Huynh, J., Carey, S.P., Mason, B.N., Vouyouka, A.G., and Reinhart-King, C.A. (2015). Biophysical induction of vascular smooth muscle cell podosomes. PLoS One *10*, e0119008.
Kim, Y., Jeong, J., and Choi, D. (2020). Small-molecule-mediated reprogramming: a silver lining for regenerative medicine. Exp Mol Med *52*, 213-226.
Kishida, T., Ejima, A., Yamamoto, K., Tanaka, S., Yamamoto, T., and Mazda, O. (2015). Reprogrammed Functional Brown Adipocytes Ameliorate Insulin Resistance and Dyslipidemia in Diet-Induced Obesity and Type 2 Diabetes. Stem Cell Reports *5*, 569-581.
Klein, T.W., and Newton, C.A. (2007). Therapeutic potential of cannabinoid-based drugs. Adv Exp Med Biol *601*, 395-413.
Knapp, D., Schulz, H., Rascon, C.A., Volkmer, M., Scholz, J., Nacu, E., Le, M., Novozhilov, S., Tazaki, A., Protze, S.*, et al.* (2013). Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. PLoS One *8*, e61352.
Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T.*, et al.* (2015). ArrayExpress update--simplifying data submissions. Nucleic Acids Res *43*, D1113-1116.
Komine, O., Nagaoka, M., Hiraoka, Y., Hoshino, M., Kawaguchi, Y., Pear, W.S., and Tanaka, K. (2011). RBP-J promotes the maturation of neuronal progenitors. Dev Biol *354*, 44-54.
Kotzsch, A., Nickel, J., Seher, A., Sebald, W., and Müller, T.D. (2009). Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5-type I receptor specificity. Embo j *28*, 937-947.
Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res *25*, 1860-1872.
Kragl, M., and Tanaka, E.M. (2009). Axolotl (Ambystoma mexicanum) limb and tail amputation. Cold Spring Harb Protoc *2009*, pdb.prot5267.
Krämer, A., Green, J., Pollard, J., Jr., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics *30*, 523-530.
Kumar, M., Kulshrestha, R., Singh, N., and Jaggi, A.S. (2019). Expanding spectrum of anticancer drug, imatinib, in the disorders affecting brain and spinal cord. Pharmacol Res *143*, 86-96.
Kunisada, Y., Tsubooka-Yamazoe, N., Shoji, M., and Hosoya, M. (2012). Small molecules induce efficient differentiation into insulin-producing cells from human induced pluripotent stem cells. Stem Cell Res *8*, 274-284.

Kunitomi, H., Oki, Y., Onishi, N., Kano, K., Banno, K., Aoki, D., Saya, H., and Nobusue, H. (2020). The insulin-PI3K-Rac1 axis contributes to terminal adipocyte differentiation through regulation of actin cytoskeleton dynamics. Genes Cells *25*, 165-174.
Kuo, C.Y., and Kohn, D.B. (2016). Gene Therapy for the Treatment of Primary Immune Deficiencies. Curr Allergy Asthma Rep *16*, 39.
Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun *9*, 1366.
Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N.*, et al.* (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929-1935.
Leem, Y.H., Park, J.S., Park, J.E., Kim, D.Y., Kang, J.L., and Kim, H.S. (2020). Papaverine inhibits α-synuclein aggregation by modulating neuroinflammation and matrix metalloproteinase-3 expression in the subacute MPTP/P mouse model of Parkinson's disease. Biomed Pharmacother *130*, 110576.
Lévesque, M., Gatien, S., Finnson, K., Desmeules, S., Villiard, E., Pilote, M., Philip, A., and Roy, S. (2007). Transforming growth factor: beta signaling is essential for limb regeneration in axolotls. PLoS One *2*, e1227.
Li, H., Jiang, H., Zhang, B., and Feng, J. (2018a). Modeling Parkinson's Disease Using Patient-specific Induced Pluripotent Stem Cells. J Parkinsons Dis *8*, 479-493.
Li, W., and Ding, S. (2010). Small molecules that modulate embryonic stem cell fate and somatic cell reprogramming. Trends Pharmacol Sci *31*, 36-45.
Li, Y., Qin, R., Yan, H., Wang, F., Huang, S., Zhang, Y., Zhong, M., Zhang, W., and Wang, Z. (2018b). Inhibition of vascular smooth muscle cells premature senescence with rutin attenuates and stabilizes diabetic atherosclerosis. J Nutr Biochem *51*, 91-98.
Li, Y., Zhang, Q., Yin, X., Yang, W., Du, Y., Hou, P., Ge, J., Liu, C., Zhang, W., Zhang, X.*, et al.* (2011). Generation of iPSCs from mouse fibroblasts with a single gene, Oct4, and small molecules. Cell Res *21*, 196-204.
Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res *47*, W199-w205.
Lin, Y.F., Jing, W., Wu, L., Li, X.Y., Wu, Y., Liu, L., Tang, W., Long, J., Tian, W.D., and Mo, X.M. (2008). Identification of osteo-adipo progenitor cells in fat tissue. Cell Prolif *41*, 803-812.
Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A., Jørgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K.*, et al.* (2007). Systematic discovery of in vivo phosphorylation networks. Cell *129*, 1415-1426.
Linhart, H.G., Ishimura-Oka, K., DeMayo, F., Kibe, T., Repka, D., Poindexter, B., Bick, R.J., and Darlington, G.J. (2001). C/EBPalpha is required for differentiation of white, but not brown, adipose tissue. Proc Natl Acad Sci U S A *98*, 12532-12537.
Liu, D.D., Zhang, B.L., Yang, J.B., and Zhou, K. (2020). Celastrol ameliorates endoplasmic stress-mediated apoptosis of osteoarthritis via regulating ATF-6/CHOP signalling pathway. J Pharm Pharmacol *72*, 826-835.
Liu, Y., Yu, C., Daley, T.P., Wang, F., Cao, W.S., Bhate, S., Lin, X., Still, C., 2nd, Liu, H., Zhao, D.*, et al.* (2018). CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. Cell Stem Cell *23*, 758-771.e758.
Liu, Z., Wang, L., Welch, J.D., Ma, H., Zhou, Y., Vaseghi, H.R., Yu, S., Wall, J.B., Alimohamadi, S., Zheng, M.*, et al.* (2017). Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. Nature *551*, 100-104.

Lokhande, K.B., Doiphode, S., Vyas, R., and Swamy, K.V. (2020). Molecular docking and simulation studies on SARS-CoV-2 M(pro) reveals Mitoxantrone, Leucovorin, Birinapant, and Dynasore as potent drugs against COVID-19. J Biomol Struct Dyn, 1-12.
Lorendeau, D., Christen, S., Rinaldi, G., and Fendt, S.M. (2015). Metabolic control of signalling pathways and metabolic auto-regulation. Biol Cell *107*, 251-272.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.
Love, N.R., Chen, Y., Ishibashi, S., Kritsiligkou, P., Lea, R., Koh, Y., Gallop, J.L., Dorey, K., and Amaya, E. (2013). Amputation-induced reactive oxygen species are required for successful Xenopus tadpole tail regeneration. Nat Cell Biol *15*, 222-228.
Lyssiotis, C.A., Lairson, L.L., Boitano, A.E., Wurdak, H., Zhu, S., and Schultz, P.G. (2011). Chemical control of stem cell fate and developmental potential. Angew Chem Int Ed Engl *50*, 200-242.
Mach, F., Montecucco, F., and Steffens, S. (2008). Cannabinoid receptors in acute and chronic complications of atherosclerosis. Br J Pharmacol *153*, 290-298.
Madrigal-Matute, J., López-Franco, O., Blanco-Colio, L.M., Muñoz-García, B., Ramos-Mozo, P., Ortega, L., Egido, J., and Martín-Ventura, J.L. (2010). Heat shock protein 90 inhibitors attenuate inflammatory responses in atherosclerosis. Cardiovasc Res *86*, 330-337.
Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics *21*, 3448-3449.
Makanae, A., Hirata, A., Honjo, Y., Mitogawa, K., and Satoh, A. (2013). Nerve independent limb induction in axolotls. Dev Biol *381*, 213-226.
Makanae, A., Mitogawa, K., and Satoh, A. (2014). Co-operative Bmp- and Fgf-signaling inputs convert skin wound healing to limb formation in urodele amphibians. Dev Biol *396*, 57-66.
Makanae, A., and Satoh, A. (2012). Early regulation of axolotl limb regeneration. Anat Rec (Hoboken) *295*, 1566-1574.
Makki, M.S., and Haqqi, T.M. (2016). Histone Deacetylase Inhibitor Vorinostat (SAHA) Suppresses IL-1β-Induced Matrix Metallopeptidase-13 Expression by Inhibiting IL-6 in Osteoarthritis Chondrocyte. Am J Pathol *186*, 2701-2708.
Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat Methods *9*, 796-804.
Marciniec, K., Beberok, A., Pęcak, P., Boryczka, S., and Wrześniok, D. (2020). Ciprofloxacin and moxifloxacin could interact with SARS-CoV-2 protease: preliminary in silico analysis. Pharmacol Rep *72*, 1553-1561.
Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Mendanha Falcão, A., Xiao, L., Li, H., Häring, M., Hochgerner, H., Romanov, R.A.*, et al.* (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science *352*, 1326-1329.
Martello, G., Sugimoto, T., Diamanti, E., Joshi, A., Hannah, R., Ohtsuka, S., Göttgens, B., Niwa, H., and Smith, A. (2012). Esrrb is a pivotal target of the Gsk3/Tcf3 axis regulating embryonic stem cell self-renewal. Cell Stem Cell *11*, 491-504.
Massa, M.S., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. BMC Syst Biol *4*, 121.
Matsumoto, A., Takeishi, S., Kanie, T., Susaki, E., Onoyama, I., Tateishi, Y., Nakayama, K., and Nakayama, K.I. (2011). p57 is required for quiescence and maintenance of adult hematopoietic stem cells. Cell Stem Cell *9*, 262-271.

Mayo, J.C., Sainz, R.M., Tan, D.X., Antolín, I., Rodríguez, C., and Reiter, R.J. (2005). Melatonin and Parkinson's disease. Endocrine *27*, 169-178.
McNab, F., Mayer-Barber, K., Sher, A., Wack, A., and O'Garra, A. (2015). Type I interferons in infectious disease. Nat Rev Immunol *15*, 87-103.
Mercader, N., Leonardo, E., Piedra, M.E., Martínez, A.C., Ros, M.A., and Torres, M. (2000). Opposing RA and FGF signals control proximodistal vertebrate limb development through regulation of Meis genes. Development *127*, 3961-3970.
Mercader, N., Tanaka, E.M., and Torres, M. (2005). Proximodistal identity during vertebrate limb regeneration is regulated by Meis homeodomain proteins. Development *132*, 4131-4142.
Michelini, E., Cevenini, L., Mezzanotte, L., Coppa, A., and Roda, A. (2010). Cell-based assays: fuelling drug discovery. Anal Bioanal Chem *398*, 227-238.
Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49-55.
Miyamoto, K., Akiyama, M., Tamura, F., Isomi, M., Yamakawa, H., Sadahiro, T., Muraoka, N., Kojima, H., Haginiwa, S., Kurotsu, S.*, et al.* (2018). Direct In Vivo Reprogramming with Sendai Virus Vectors Improves Cardiac Function after Myocardial Infarction. Cell Stem Cell *22*, 91-103.e105.
Monaghan, J.R., Athippozhy, A., Seifert, A.W., Putta, S., Stromberg, A.J., Maden, M., Gardiner, D.M., and Voss, S.R. (2012). Gene expression patterns specific to the regenerating limb of the Mexican axolotl. Biol Open *1*, 937-948.
Morris, M.C., Evans, D.A., Bienias, J.L., Scherr, P.A., Tangney, C.C., Hebert, L.E., Bennett, D.A., Wilson, R.S., and Aggarwal, N. (2004). Dietary niacin and the risk of incident Alzheimer's disease and of cognitive decline. J Neurol Neurosurg Psychiatry *75*, 1093-1099.
Morris, S.A., and Daley, G.Q. (2013). A blueprint for engineering cell fate: current technologies to reprogram cell identity. Cell Res *23*, 33-48.
Motyl, J., and Strosznajder, J.B. (2018). Sphingosine kinase 1/sphingosine-1-phosphate receptors dependent signalling in neurodegenerative diseases. The promising target for neuroprotection in Parkinson's disease. Pharmacol Rep *70*, 1010-1014.
Mounsey, R.B., Mustafa, S., Robinson, L., Ross, R.A., Riedel, G., Pertwee, R.G., and Teismann, P. (2015). Increasing levels of the endocannabinoid 2-AG is neuroprotective in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine mouse model of Parkinson's disease. Exp Neurol *273*, 36-44.
Muneoka, K., and Bryant, S.V. (1982). Evidence that patterning mechanisms in developing and regenerating limbs are the same. Nature *298*, 369-371.
Muneoka, K., Fox, W.F., and Bryant, S.V. (1986). Cellular contribution from dermis and cartilage to the regenerating limb blastema in axolotls. Dev Biol *116*, 256-260.
Musa, A., Ghoraie, L.S., Zhang, S.D., Glazko, G., Yli-Harja, O., Dehmer, M., Haibe-Kains, B., and Emmert-Streib, F. (2017). A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinform *18*, 903.
Nacu, E., Gromberg, E., Oliveira, C.R., Drechsel, D., and Tanaka, E.M. (2016). FGF8 and SHH substitute for anterior-posterior tissue interactions to induce limb regeneration. Nature *533*, 407-410.
Nakae, J., Kitamura, T., Kitamura, Y., Biggs, W.H., 3rd, Arden, K.C., and Accili, D. (2003). The forkhead transcription factor Foxo1 regulates adipocyte differentiation. Dev Cell *4*, 119-129.
Nayak, D.P., and Rasmussen, A.F., Jr. (1966). Influence of mitomycin C on the replication of influenza viruses. Virology *30*, 673-683.

Nguyen, M., Singhal, P., Piet, J.W., Shefelbine, S.J., Maden, M., Voss, S.R., and Monaghan, J.R. (2017). Retinoic acid receptor regulation of epimorphic and homeostatic regeneration in the axolotl. Development *144*, 601-611.
Nguyen, T.T.H., Jung, J.H., Kim, M.K., Lim, S., Choi, J.M., Chung, B., Kim, D.W., and Kim, D. (2021). The Inhibitory Effects of Plant Derivate Polyphenols on the Main Protease of SARS Coronavirus 2 and Their Structure-Activity Relationship. Molecules *26*.
Nguyen-Chi, M., Laplace-Builhé, B., Travnickova, J., Luz-Crawford, P., Tejedor, G., Lutfalla, G., Kissa, K., Jorgensen, C., and Djouad, F. (2017). TNF signaling and macrophages govern fin regeneration in zebrafish larvae. Cell Death Dis *8*, e2979.
Nielsen, R., Pedersen, T.A., Hagenbeek, D., Moulos, P., Siersbaek, R., Megens, E., Denissov, S., Børgesen, M., Francoijs, K.J., Mandrup, S.*, et al.* (2008). Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. Genes Dev *22*, 2953-2967.
Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes Dev *12*, 2048-2060.
Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118-122.
Okawa, S., Saltó, C., Ravichandran, S., Yang, S., Toledo, E.M., Arenas, E., and Del Sol, A. (2018). Transcriptional synergy as an emergent property defining cell subpopulation identity enables population shift. Nat Commun *9*, 2595.
Oshimori, N., and Fuchs, E. (2012). Paracrine TGF-β signaling counterbalances BMP-mediated repression in hair follicle stem cell activation. Cell Stem Cell *10*, 63-75.
Owlarn, S., Klenner, F., Schmidt, D., Rabert, F., Tomasso, A., Reuter, H., Mulaw, M.A., Moritz, S., Gentile, L., Weidinger, G.*, et al.* (2017). Generic wound signals initiate regeneration in missing-tissue contexts. Nat Commun *8*, 2282.
Pal, B., Endisha, H., Zhang, Y., and Kapoor, M. (2015). mTOR: a potential therapeutic target in osteoarthritis? Drugs R D *15*, 27-36.
Pan, X., Kaminga, A.C., Wen, S.W., Wu, X., Acheampong, K., and Liu, A. (2019). Dopamine and Dopamine Receptors in Alzheimer's Disease: A Systematic Review and Network Meta-Analysis. Front Aging Neurosci *11*, 175.
Pang, Y., Gan, L., Wang, X., Su, Q., Liang, C., and He, P. (2019). Celecoxib aggravates atherogenesis and upregulates leukotrienes in ApoE(-/-) mice and lipopolysaccharide-stimulated RAW264.7 macrophages. Atherosclerosis *284*, 50-58.
Papageorgiou, N., Zacharia, E., Briasoulis, A., Charakida, M., and Tousoulis, D. (2016). Celecoxib for the treatment of atherosclerosis. Expert Opin Investig Drugs *25*, 619-633.
Parikh, J.R., Klinger, B., Xia, Y., Marto, J.A., and Blüthgen, N. (2010). Discovering causal signaling pathways through gene-expression patterns. Nucleic Acids Res *38*, W109-117.
Park, G., Yoon, B.S., Kim, Y.S., Choi, S.C., Moon, J.H., Kwon, S., Hwang, J., Yun, W., Kim, J.H., Park, C.Y.*, et al.* (2015). Conversion of mouse fibroblasts into cardiomyocyte-like cells using small molecule treatments. Biomaterials *54*, 201-212.
Park, T.S., Zimmerlin, L., Evans-Moses, R., and Zambidis, E.T. (2018). Chemical Reversion of Conventional Human Pluripotent Stem Cells to a Naïve-like State with Improved Multilineage Differentiation Potency. J Vis Exp.
Pavathuparambil Abdul Manaph, N., Sivanathan, K.N., Nitschke, J., Zhou, X.-F., Coates, P.T., and Drogemuller, C.J. (2019). An overview on small molecule-induced differentiation of mesenchymal stem cells into beta cells for diabetic therapy. Stem Cell Research & Therapy *10*, 293.

Pedraza, C.E., Taylor, C., Pereira, A., Seng, M., Tham, C.S., Izrael, M., and Webb, M. (2014). Induction of oligodendrocyte differentiation and in vitro myelination by inhibition of rho-associated kinase. ASN Neuro *6*.
Pirinen, E., Auranen, M., Khan, N.A., Brilhante, V., Urho, N., Pessia, A., Hakkarainen, A., Kuula, J., Heinonen, U., Schmidt, M.S.*, et al.* (2020). Niacin Cures Systemic NAD(+) Deficiency and Improves Muscle Performance in Adult-Onset Mitochondrial Myopathy. Cell Metab *31*, 1078-1090.e1075.
Pittenger, M.F., Mackay, A.M., Beck, S.C., Jaiswal, R.K., Douglas, R., Mosca, J.D., Moorman, M.A., Simonetti, D.W., Craig, S., and Marshak, D.R. (1999). Multilineage potential of adult human mesenchymal stem cells. Science *284*, 143-147.
Postuma, R.B., Lang, A.E., Munhoz, R.P., Charland, K., Pelletier, A., Moscovich, M., Filla, L., Zanatta, D., Rios Romenets, S., Altman, R.*, et al.* (2012). Caffeine for treatment of Parkinson disease: a randomized controlled trial. Neurology *79*, 651-658.
Qin, H., Zhao, A., and Fu, X. (2017). Small molecules for reprogramming and transdifferentiation. Cell Mol Life Sci *74*, 3553-3575.
Rackham, O.J., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., Shin, J.W.*, et al.* (2016). A predictive computational framework for direct reprogramming between human cell types. Nat Genet *48*, 331-335.
Raj, N.B., and Pitha, P.M. (1981). Analysis of interferon mRNA in human fibroblast cells induced to produce interferon. Proc Natl Acad Sci U S A *78*, 7426-7430.
Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B.*, et al.* (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun *6*, 7866.
Reddien, P.W., and Sánchez Alvarado, A. (2004). Fundamentals of planarian regeneration. Annu Rev Cell Dev Biol *20*, 725-757.
Rhim, J.H., Luo, X., Xu, X., Gao, D., Zhou, T., Li, F., Qin, L., Wang, P., Xia, X., and Wong, S.T. (2015). A High-content screen identifies compounds promoting the neuronal differentiation and the midbrain dopamine neuron specification of human neural progenitor cells. Sci Rep *5*, 16237.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.
Rivera, D.S., Lindsay, C., Codocedo, J.F., Morel, I., Pinto, C., Cisternas, P., Bozinovic, F., and Inestrosa, N.C. (2016). Andrographolide recovers cognitive impairment in a natural model of Alzheimer's disease (Octodon degus). Neurobiol Aging *46*, 204-220.
Rivetti di Val Cervo, P., Besusso, D., Conforti, P., and Cattaneo, E. (2021). hiPSCs for predictive modelling of neurodegenerative diseases: dreaming the possible. Nat Rev Neurol, 1-12.
Rocha, S., Ribeiro, D., Fernandes, E., and Freitas, M. (2020). A Systematic Review on Anti-diabetic Properties of Chalcones. Curr Med Chem *27*, 2257-2321.
Rojas-Muñoz, A., Rajadhyksha, S., Gilmour, D., van Bebber, F., Antos, C., Rodríguez Esteban, C., Nüsslein-Volhard, C., and Izpisúa Belmonte, J.C. (2009). ErbB2 and ErbB3 regulate amputation-induced proliferation and migration during vertebrate regeneration. Dev Biol *327*, 177-190.
Rostam, M.A., Shajimoon, A., Kamato, D., Mitra, P., Piva, T.J., Getachew, R., Cao, Y., Zheng, W., Osman, N., and Little, P.J. (2018). Flavopiridol Inhibits TGF-β-Stimulated Biglycan Synthesis by Blocking Linker Region Phosphorylation and Nuclear Translocation of Smad2. J Pharmacol Exp Ther *365*, 156-164.

Ruffell, D., Mourkioti, F., Gambardella, A., Kirstetter, P., Lopez, R.G., Rosenthal, N., and Nerlov, C. (2009). A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. Proc Natl Acad Sci U S A *106*, 17475-17480.
Sader, F., Denis, J.F., Laref, H., and Roy, S. (2019). Epithelial to mesenchymal transition is mediated by both TGF-β canonical and non-canonical signaling during axolotl limb regeneration. Sci Rep *9*, 1144.
Saisho, Y., Hirose, H., Horimai, C., Miyashita, K., Takei, I., Umezawa, K., and Itoh, H. (2008). Effects of DHMEQ, a novel nuclear factor-kappab inhibitor, on beta cell dysfunction in INS-1 cells. Endocr J *55*, 433-438.
Sakaki-Yumoto, M., Liu, J., Ramalho-Santos, M., Yoshida, N., and Derynck, R. (2013). Smad2 is essential for maintenance of the human and mouse primed pluripotent stem cell state. J Biol Chem *288*, 18546-18560.
Salama, M., and Arias-Carrión, O. (2011). Colchicine as a promising drug for Parkinson's disease. Med Hypotheses *76*, 150.
Salama, M., Ellaithy, A., Helmy, B., El-Gamal, M., Tantawy, D., Mohamed, M., Sheashaa, H., Sobh, M., and Arias-Carrión, O. (2012). Colchicine protects dopaminergic neurons in a rat model of Parkinson's disease. CNS Neurol Disord Drug Targets *11*, 836-843.
Sampogna, G., Guraya, S.Y., and Forgione, A. (2015). Regenerative medicine: Historical roots and potential strategies in modern medicine. J Microsc Ultrastruct *3*, 101-107.
Sartor, M.A., Leikauf, G.D., and Medvedovic, M. (2009). LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics *25*, 211-217.
Satoh, A., makanae, A., Hirata, A., and Satou, Y. (2011). Blastema induction in aneurogenic state and Prrx-1 regulation by MMPs and FGFs in Ambystoma mexicanum limb regeneration. Dev Biol *355*, 263-274.
Schellekens, H., Huffmeyer, J.H., and Van Griensven, L.J. (1975). The influence of emetine on the induction of interferon by poly-I: poly-C in Swiss mice. J Gen Virol *26*, 197-200.
Schiffman, S.S., Clark, C.M., and Warwick, Z.S. (1990). Gustatory and olfactory dysfunction in dementia: not specific to Alzheimer's disease. Neurobiol Aging *11*, 597-600.
Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun *9*, 20.
Sergushichev, A.A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv, 060012.
Shahzad, S., and Willcox, M. (2020). Immuno-pathogenesis of nCOVID-19 and a possible host-directed therapy including anti-inflammatory and anti-viral prostaglandin (PG J(2)) for effective treatment and reduction in the death toll. Med Hypotheses *143*, 110080.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Sharma, S., and Petsalaki, E. (2019). Large-scale datasets uncovering cell signalling networks in cancer: context matters. Curr Opin Genet Dev *54*, 118-124.
Shi, Y., Desponts, C., Do, J.T., Hahm, H.S., Schöler, H.R., and Ding, S. (2008). Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. Cell Stem Cell *3*, 568-574.
Shi, Y., Inoue, H., Wu, J.C., and Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. Nat Rev Drug Discov *16*, 115-130.
Shtro, A.A., Zarubaev, V.V., Luzina, O.A., Sokolov, D.N., and Salakhutdinov, N.F. (2015). Derivatives of usnic acid inhibit broad range of influenza viruses and protect mice from lethal influenza infection. Antivir Chem Chemother *24*, 92-98.

Si-Tayeb, K., Noto, F.K., Nagaoka, M., Li, J., Battle, M.A., Duris, C., North, P.E., Dalton, S., and Duncan, S.A. (2010). Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. Hepatology *51*, 297-305.
Singh, B.N., Doyle, M.J., Weaver, C.V., Koyano-Nakagawa, N., and Garry, D.J. (2012). Hedgehog and Wnt coordinate signaling in myogenic progenitors and regulate limb regeneration. Dev Biol *371*, 23-34.
Singh, B.N., Koyano-Nakagawa, N., Donaldson, A., Weaver, C.V., Garry, M.G., and Garry, D.J. (2015). Hedgehog Signaling during Appendage Development and Regeneration. Genes (Basel) *6*, 417-435.
Söllner, J.F., Leparc, G., Hildebrandt, T., Klein, H., Thomas, L., Stupka, E., and Simon, E. (2017). An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Sci Data *4*, 170185.
Song, G., Pacher, M., Balakrishnan, A., Yuan, Q., Tsay, H.C., Yang, D., Reetz, J., Brandes, S., Dai, Z., Pützer, B.M.*, et al.* (2016). Direct Reprogramming of Hepatic Myofibroblasts into Hepatocytes In Vivo Attenuates Liver Fibrosis. Cell Stem Cell *18*, 797-808.
Spigset, O., and Mjörndal, T. (1999). Increased glucose intolerance related to digoxin treatment in patients with type 2 diabetes mellitus. J Intern Med *246*, 419-422.
Stefanachi, A., Leonetti, F., Pisani, L., Catto, M., and Carotti, A. (2018). Coumarin: A Natural, Privileged and Versatile Scaffold for Bioactive Compounds. Molecules *23*.
Stephens, J.M., Morrison, R.F., Wu, Z., and Farmer, S.R. (1999). PPARgamma ligand-dependent induction of STAT1, STAT5A, and STAT5B during adipogenesis. Biochem Biophys Res Commun *262*, 216-222.
Stock, S.R., Blackburn, D., Gradassi, M., and Simon, H.G. (2003). Bone formation during forelimb regeneration: a microtomography (microCT) analysis. Dev Dyn *226*, 410-417.
Stocum, D.L., and Cameron, J.A. (2011). Looking proximally and distally: 100 years of limb regeneration and beyond. Dev Dyn *240*, 943-968.
Strasen, J., Sarma, U., Jentsch, M., Bohn, S., Sheng, C., Horbelt, D., Knaus, P., Legewie, S., and Loewer, A. (2018). Cell-specific responses to the cytokine TGFβ are determined by variability in protein levels. Mol Syst Biol *14*, e7733.
Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e1821.
Su, Z., Niu, W., Liu, M.L., Zou, Y., and Zhang, C.L. (2014). In vivo conversion of astrocytes to neurons in the injured adult spinal cord. Nat Commun *5*, 3338.
Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K.*, et al.* (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell *171*, 1437-1452.e1417.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.
Sun, B., Roh, K.H., Lee, S.R., Lee, Y.S., and Kang, K.S. (2007). Induction of human umbilical cord blood-derived stem cells with embryonic stem cell phenotypes into insulin producing islet-like structure. Biochem Biophys Res Commun *354*, 919-923.
Sun, H., Wu, Y., Pan, Z., Yu, D., Chen, P., Zhang, X., Wu, H., Zhang, X., An, C., Chen, Y.*, et al.* (2018). Gefitinib for Epidermal Growth Factor Receptor Activated Osteoarthritis Subpopulation Treatment. EBioMedicine *32*, 223-233.
Sun, K., Luo, J., Guo, J., Yao, X., Jing, X., and Guo, F. (2020a). The PI3K/AKT/mTOR signaling pathway in osteoarthritis: a narrative review. Osteoarthritis Cartilage *28*, 400-409.

Sun, Y., He, Z., Li, J., Gong, S., Yuan, S., Li, T., Ning, N., Xing, L., Zhang, L., Chen, F.*, et al.* (2020b). Gentamicin Induced Microbiome Adaptations Associate With Increased BCAA Levels and Enhance Severity of Influenza Infection. Front Immunol *11*, 608895.
Suzuki, M., Satoh, A., Ide, H., and Tamura, K. (2007). Transgenic Xenopus with prx1 limb enhancer reveals crucial contribution of MEK/ERK and PI3K/AKT pathways in blastema formation during limb regeneration. Dev Biol *304*, 675-686.
Suzuki, Y., Chou, J., Garvey, S.L., Wang, V.R., and Yanes, K.O. (2019). Evolution and Regulation of Limb Regeneration in Arthropods. Results Probl Cell Differ *68*, 419-454.
Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res *44*, D380-384.
Taechowisan, T., Samsawat, T., Puckdee, W., and Phutdhawong, W.S. (2020). Antiviral activity of geldanamycin and its derivatives against influenza virus. Journal of Applied Pharmaceutical Science *10*, 113-120.
Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.
Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W.*, et al.* (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. Cell *158*, 1254-1269.
Takato, T., Iwata, K., Murakami, C., Wada, Y., and Sakane, F. (2017). Chronic administration of myristic acid improves hyperglycaemia in the Nagoya-Shibata-Yasuda mouse model of congenital type 2 diabetes. Diabetologia *60*, 2076-2083.
Takeda, Y., Harada, Y., Yoshikawa, T., and Dai, P. (2017). Direct conversion of human fibroblasts to brown adipocytes by small chemical compounds. Sci Rep *7*, 4304.
Takizawa, H., Regoes, R.R., Boddupalli, C.S., Bonhoeffer, S., and Manz, M.G. (2011). Dynamic variation in cycling of hematopoietic stem cells in steady state and inflammation. J Exp Med *208*, 273-284.
Talasaz, A.H., Kakavand, H., Van Tassell, B., Aghakouchakzadeh, M., Sadeghipour, P., Dunn, S., and Geraiely, B. (2021). Cardiovascular Complications of COVID-19: Pharmacotherapy Perspective. Cardiovasc Drugs Ther *35*, 249-259.
Tamada, K., Nakajima, S., Ogawa, N., Inada, M., Shibasaki, H., Sato, A., Takasawa, R., Yoshimori, A., Suzuki, Y., Watanabe, N.*, et al.* (2019). Papaverine identified as an inhibitor of high mobility group box 1/receptor for advanced glycation end-products interaction suppresses high mobility group box 1-mediated inflammatory responses. Biochem Biophys Res Commun *511*, 665-670.
Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P., and Romero, R. (2009). A novel signaling pathway impact analysis. Bioinformatics *25*, 75-82.
Tasaki, J., Shibata, N., Nishimura, O., Itomi, K., Tabata, Y., Son, F., Suzuki, N., Araki, R., Abe, M., Agata, K.*, et al.* (2011). ERK signaling controls blastema cell differentiation during planarian regeneration. Development *138*, 2417-2427.
Taura, D., Noguchi, M., Sone, M., Hosoda, K., Mori, E., Okada, Y., Takahashi, K., Homma, K., Oyamada, N., Inuzuka, M.*, et al.* (2009a). Adipogenic differentiation of human induced pluripotent stem cells: comparison with that of human embryonic stem cells. FEBS Lett *583*, 1029-1033.
Taura, D., Sone, M., Homma, K., Oyamada, N., Takahashi, K., Tamura, N., Yamanaka, S., and Nakao, K. (2009b). Induction and isolation of vascular cells from human induced pluripotent stem cells--brief report. Arterioscler Thromb Vasc Biol *29*, 1100-1103.
Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L.*, et al.* (2014). Systematic identification of culture

conditions for induction and maintenance of naive human pluripotency. Cell Stem Cell *15*, 471-487.
Tran, K.A., Jackson, S.A., Olufs, Z.P., Zaidan, N.Z., Leng, N., Kendziorski, C., Roy, S., and Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. Nat Commun *6*, 6188.
Trukhan, D.I., Mazurov, A.L., and Rechapova, L.A. (2016). [Acute respiratory viral infections: Topical issues of diagnosis, prevention and treatment in therapeutic practice]. Ter Arkh *88*, 76-82.
Tseng, A.S., Adams, D.S., Qiu, D., Koustubhan, P., and Levin, M. (2007). Apoptosis is required during early stages of tail regeneration in Xenopus laevis. Dev Biol *301*, 62-69.
Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods *13*, 966-967.
Valenti, M.T., Serena, M., Carbonare, L.D., and Zipeto, D. (2019). CRISPR/Cas system: An emerging technology in stem cell research. World J Stem Cells *11*, 937-956.
Vasanthakumar, A., Moro, K., Xin, A., Liao, Y., Gloury, R., Kawamoto, S., Fagarasan, S., Mielke, L.A., Afshar-Sterle, S., Masters, S.L.*, et al.* (2015). The transcriptional regulators IRF4, BATF and IL-33 orchestrate development and maintenance of adipose tissue-resident regulatory T cells. Nat Immunol *16*, 276-285.
Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Südhof, T.C., and Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. Nature *463*, 1035-1041.
Vijayakumar, B.G., Ramesh, D., Joji, A., Jayachandra Prakasan, J., and Kannan, T. (2020). In silico pharmacokinetic and molecular docking studies of natural flavonoids and synthetic indole chalcones against essential proteins of SARS-CoV-2. Eur J Pharmacol *886*, 173448.
Vivar, J.C., Pemu, P., McPherson, R., and Ghosh, S. (2013). Redundancy control in pathway databases (ReCiPa): an application for improving gene-set enrichment analysis in Omics studies and "Big data" biology. Omics *17*, 414-422.
Wang, C., Liu, P., Luo, J., Ding, H., Gao, Y., Sun, L., Luo, F., Liu, X., and He, H. (2017). Geldanamycin Reduces Acute Respiratory Distress Syndrome and Promotes the Survival of Mice Infected with the Highly Virulent H5N1 Influenza Virus. Front Cell Infect Microbiol *7*, 267.
Wang, H., Yang, Y., Liu, J., and Qian, L. (2021). Direct cell reprogramming: approaches, mechanisms and progress. Nat Rev Mol Cell Biol *22*, 410-424.
Wang, H.W., Lin, L.M., He, H.Y., You, F., Li, W.Z., Huang, T.H., Ma, G.X., and Ma, L. (2011). Human umbilical cord mesenchymal stem cells derived from Wharton's jelly differentiate into insulin-producing cells in vitro. Chin Med J (Engl) *124*, 1534-1539.
Wang, S., Zhang, J., and Ye, X. (2012). [Protein kinase inhibitor flavopiridol inhibits the replication of influenza virus in vitro]. Wei Sheng Wu Xue Bao *52*, 1137-1142.
Wang, X., Cao, R., Zhang, H., Liu, J., Xu, M., Hu, H., Li, Y., Zhao, L., Li, W., Sun, X.*, et al.* (2020). The anti-influenza virus drug, arbidol is an efficient inhibitor of SARS-CoV-2 in vitro. Cell Discov *6*, 28.
Watanabe, T., Sato, Y., Masud, H., Takayama, M., Matsuda, H., Hara, Y., Yanagi, Y., Yoshida, M., Goshima, F., Murata, T.*, et al.* (2020). Antitumor activity of cyclin-dependent kinase inhibitor alsterpaullone in Epstein-Barr virus-associated lymphoproliferative disorders. Cancer Sci *111*, 279-287.
Wehner, D., Cizelsky, W., Vasudevaro, M.D., Ozhan, G., Haase, C., Kagermeier-Schenk, B., Röder, A., Dorsky, R.I., Moro, E., Argenton, F.*, et al.* (2014). Wnt/β-catenin signaling defines organizing centers that orchestrate growth and differentiation of the regenerating zebrafish caudal fin. Cell Rep *6*, 467-481.

Whitebread, S., Hamon, J., Bojanic, D., and Urban, L. (2005). Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. Drug Discov Today *10*, 1421-1433.
Wilkie, I.C. (2001). Autotomy as a prelude to regeneration in echinoderms. Microsc Res Tech *55*, 369-396.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z.*, et al.* (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-d1082.
Woo, J.H., Shimoni, Y., Yang, W.S., Subramaniam, P., Iyer, A., Nicoletti, P., Rodríguez Martínez, M., López, G., Mattioli, M., Realubit, R.*, et al.* (2015). Elucidating Compound Mechanism of Action by Network Perturbation Analysis. Cell *162*, 441-451.
Wouters, B.J., and Delwel, R. (2016). Epigenetics and approaches to targeted epigenetic therapy in acute myeloid leukemia. Blood *127*, 42-52.
Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol *11*, R53.
Xiao, Y., Gong, Y., Lv, Y., Lan, Y., Hu, J., Li, F., Xu, J., Bai, J., Deng, Y., Liu, L.*, et al.* (2015). Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. Sci Rep *5*, 10889.
Xie, B., Sun, D., Du, Y., Jia, J., Sun, S., Xu, J., Liu, Y., Xiang, C., Chen, S., Xie, H.*, et al.* (2019). A two-step lineage reprogramming strategy to generate functionally competent human hepatocytes from fibroblasts. Cell Res *29*, 696-710.
Xin, Y., Jiang, X., Wang, Y., Su, X., Sun, M., Zhang, L., Tan, Y., Wintergerst, K.A., Li, Y., and Li, Y. (2016). Insulin-Producing Cells Differentiated from Human Bone Marrow Mesenchymal Stem Cells In Vitro Ameliorate Streptozotocin-Induced Diabetic Hyperglycemia. PLoS One *11*, e0145838.
Xing, J., Shankar, R., Drelich, A., Paithankar, S., Chekalin, E., Dexheimer, T., Chua, M.S., Rajasekaran, S., Tseng, C.K., and Chen, B. (2020). Analysis of Infected Host Gene Expression Reveals Repurposed Drug Candidates and Time-Dependent Host Response Dynamics for COVID-19. bioRxiv.
Xu, J., Du, Y., and Deng, H. (2015). Direct lineage reprogramming: strategies, mechanisms, and applications. Cell Stem Cell *16*, 119-134.
Xu, Y., Shi, Y., and Ding, S. (2008). A chemical approach to stem-cell biology and regenerative medicine. Nature *453*, 338-344.
Yamanaka, S. (2020). Pluripotent Stem Cell-Based Cell Therapy-Promise and Challenges. Cell Stem Cell *27*, 523-531.
Yang, H., Adam, R.C., Ge, Y., Hua, Z.L., and Fuchs, E. (2017). Epithelial-Mesenchymal Micro-niches Govern Stem Cell Lineage Choices. Cell *169*, 483-496.e413.
Yang, L., Zhang, J., and Wang, G. (2015). The effect of sodium hyaluronate treating knee osteoarthritis on synovial fluid interleukin -1β and clinical treatment mechanism. Pak J Pharm Sci *28*, 407-410.
Ye, S., Li, P., Tong, C., and Ying, Q.L. (2013). Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. Embo j *32*, 2548-2560.
Yee, J., White, R.E., Anderton, E., and Allday, M.J. (2011). Latent Epstein-Barr virus can inhibit apoptosis in B cells by blocking the induction of NOXA expression. PLoS One *6*, e28506.
Yeganeh, P.N., and Mostafavi, M.T. (2020). Causal Disturbance Analysis: A Novel Graph Centrality Based Method for Pathway Enrichment Analysis. IEEE/ACM Trans Comput Biol Bioinform *17*, 1613-1624.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.
Yokoyama, H., Ogino, H., Stoick-Cooper, C.L., Grainger, R.M., and Moon, R.T. (2007). Wnt/beta-catenin signaling has an essential role in the initiation of limb regeneration. Dev Biol *306*, 170-178.
Youngblood, B., Hale, J.S., Kissick, H.T., Ahn, E., Xu, X., Wieland, A., Araki, K., West, E.E., Ghoneim, H.E., Fan, Y.*, et al.* (2017). Effector CD8 T cells dedifferentiate into long-lived memory cells. Nature *552*, 404-409.
Yu, M., Zhang, H., Wang, B., Zhang, Y., Zheng, X., Shao, B., Zhuge, Q., and Jin, K. (2021). Key Signaling Pathways in Aging and Potential Interventions for Healthy Aging. Cells *10*.
Yun, M.H., Gates, P.B., and Brockes, J.P. (2013). Regulation of p53 is critical for vertebrate limb regeneration. Proc Natl Acad Sci U S A *110*, 17392-17397.
Yun, M.H., Gates, P.B., and Brockes, J.P. (2014). Sustained ERK activation underlies reprogramming in regeneration-competent salamander cells and distinguishes them from their mammalian counterparts. Stem Cell Reports *3*, 15-23.
Zaffaroni, G., Okawa, S., Morales-Ruiz, M., and Del Sol, A. (2019). An integrative method to predict signalling perturbations for cellular transitions. Nucleic Acids Res *47*, e72.
Zhang, D.L., Gu, L.J., Liu, L., Wang, C.Y., Sun, B.S., Li, Z., and Sung, C.K. (2009). Effect of Wnt signaling pathway on wound healing. Biochem Biophys Res Commun *378*, 149-151.
Zhang, H., Li, Y., Wang, H.B., Zhang, A., Chen, M.L., Fang, Z.X., Dong, X.D., Li, S.B., Du, Y., Xiong, D.*, et al.* (2018). Ephrin receptor A2 is an epithelial cell receptor for Epstein-Barr virus entry. Nat Microbiol *3*, 1-8.
Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res *43*, D76-81.
Zhang, J.W., Klemm, D.J., Vinson, C., and Lane, M.D. (2004). Role of CREB in transcriptional regulation of CCAAT/enhancer-binding protein beta gene during adipogenesis. J Biol Chem *279*, 4471-4478.
Zhang, M., Lin, Y.H., Sun, Y.J., Zhu, S., Zheng, J., Liu, K., Cao, N., Li, K., Huang, Y., and Ding, S. (2016a). Pharmacological Reprogramming of Fibroblasts into Neural Stem Cells by Signaling-Directed Transcriptional Activation. Cell Stem Cell *18*, 653-667.
Zhang, T., Hu, Q., Shi, L., Qin, L., Zhang, Q., and Mi, M. (2016b). Equol Attenuates Atherosclerosis in Apolipoprotein E-Deficient Mice by Inhibiting Endoplasmic Reticulum Stress via Activation of Nrf2 in Endothelial Cells. PLoS One *11*, e0167020.
Zhang, X., Yalcin, S., Lee, D.F., Yeh, T.Y., Lee, S.M., Su, J., Mungamuri, S.K., Rimmelé, P., Kennedy, M., Sellers, R.*, et al.* (2011). FOXO1 is an essential regulator of pluripotency in human embryonic stem cells. Nat Cell Biol *13*, 1092-1099.
Zhu, S., Li, W., Zhou, H., Wei, W., Ambasudhan, R., Lin, T., Kim, J., Zhang, K., and Ding, S. (2010). Reprogramming of human primary somatic cells by OCT4 and chemical compounds. Cell Stem Cell *7*, 651-655.
Zimmerlin, L., Park, T.S., Huo, J.S., Verma, K., Pather, S.R., Talbot, C.C., Jr., Agarwal, J., Steppan, D., Zhang, Y.W., Considine, M.*, et al.* (2016). Tankyrase inhibition promotes a stable human naïve pluripotent state with improved functionality. Development *143*, 4368-4380.
Zimmerlin, L., Park, T.S., and Zambidis, E.T. (2017). Capturing Human Naïve Pluripotency in the Embryo and in the Dish. Stem Cells Dev *26*, 1141-1161.

7 Appendices

7.1 Supplementary Tables

Table S1. Information on collected transcriptome profiles of siganlling perturbations in NCPC of SiPer

Table S2. List of interactions in prior knowledge networks of SiPer

Table S3. Information on benchmarking datasets in SiPer

Table S4. List of perturbations used for both normal and cancer cells in SiPer

Table S5. SiPer perdictions for cellular conversion cases

Table S6. SiPer predictions and experimental validationfor hepatocyte maturation

Table S7. Information on collected transcriptome profiles of siganlling perturbations in database of ChemPert

Table S8. Some examples of datasets that their perturbation targets can only be predicted by perturbation database of non-cancer cells

Table S9. ChemPert predictions for aged-related diseases with evidence and the corresponding dataset information

Table S10. ChemPert predictions for infectious diseases with evidence and the corresponding dataset information

Table S11. Full list of enriched GO terms for each time interval

Table S12. Full list of predicted signalling pathways and key molecules for each time interval

Table S13. Canonical pathways present in MetaCore from Clarivate Analytics included in the signalling network