Stem Cell Reports
Article

ISSCR

OPEN ACCESS

A single cell-based computational platform to identify chemical compounds targeting desired sets of transcription factors for cellular conversion

Menglin Zheng,[1,7] Bingqing Xie,[2,6,7] Satoshi Okawa,[1,3] Soon Yi Liew,[2] Hongkui Deng,[2,*] and Antonio del Sol[1,4,5,*]

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Esch-sur-Alzette, 4367 Belvaux, Luxembourg
[2]MOE Engineering Research Center of Regenerative Medicine, 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, 3555 Dudelange, Luxembourg
[4]CIC bioGUNE-BRTA (Basque Research and Technology Alliance), Bizkaia Technology Park, 801 Building, 48160 Derio, Spain
[5]IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain
[6]Institute of Epigenetics and Brain Science, Southwest Medical University, Luzhou 646099, China
[7]These authors contributed equally
*Correspondence: hongkui_deng@pku.edu.cn (H.D.), antonio.delsol@uni.lu (A.d.S.)
https://doi.org/10.1016/j.stemcr.2022.10.013

SUMMARY

Cellular conversion can be induced by perturbing a handful of key transcription factors (TFs). Replacement of direct manipulation of key TFs with chemical compounds offers a less laborious and safer strategy to drive cellular conversion for regenerative medicine. Nevertheless, identifying optimal chemical compounds currently requires large-scale screening of chemical libraries, which is resource intensive. Existing computational methods aim at predicting cell conversion TFs, but there are no methods for identifying chemical compounds targeting these TFs. Here, we develop a single cell-based platform (SiPer) to systematically prioritize chemical compounds targeting desired TFs to guide cellular conversions. SiPer integrates a large compendium of chemical perturbations on non-cancer cells with a network model and predicted known and novel chemical compounds in diverse cell conversion examples. Importantly, we applied SiPer to develop a highly efficient protocol for human hepatic maturation. Overall, SiPer provides a valuable resource to efficiently identify chemical compounds for cell conversion.

INTRODUCTION

The generation of desired functional 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. It has been observed that a handful of specific transcription factors (TFs) is usually sufficient to trigger cellular conversion (Morris and Daley, 2013). Indeed, researchers have developed experimental cell conversion protocols with forced ectopic expression of specific sets of TFs in target cells (Colasante et al., 2015; Takahashi et al., 2007). In contrast to conventional cellular conversion by genetic methods, many studies have revealed that chemical compounds alone can induce cell fate conversion without manipulation of genetic materials (Ye et al., 2016; Hou et al., 2013; Cheng et al., 2015). Compared with TF-based protocols, chemical compounds are more easily controlled by adjusting the concentrations, durations, and combinations (Ye et al., 2016) and are usually more cost effective and cell permeable (Hou et al., 2013; Cao et al., 2016). In addition, they can avoid risks related to genetic manipulations such as unwanted long-term expression of the delivered TFs, inflammatory responses, and the possibility for insertional mutagenesis (Cieślar-Pobuda et al., 2017). However, a major drawback of chemical compounds is that they usually act on cellular signaling networks, which could lead to undesired, potentially detrimental activation/inactivation of downstream TFs, while researchers have revealed that the regulation of specific TFs is essential for the generation of desired cell types (Treutlein et al., 2016; Morris et al., 2014). Thus, ideally one wishes to perturb a specific set of target TFs using chemical compounds, while minimizing the effect on non-target TFs. Nevertheless, this is a challenging task, as downstream target TFs of each small molecule in each cell type are largely unknown. Currently, chemical-induced cellular conversion protocols rely on large-scale screening of small molecules, which is inefficient and resource intensive. In this regard, computational guidance can be of great help prioritize chemical compounds for cellular conversion. Nevertheless, existing computational methods, which are designed to identify cellular conversion TFs (Cahan et al., 2014; D'Alessio et al., 2015; Ribeiro et al., 2021), do not provide information on chemical compounds targeting these TFs.

To date, some computational methods, such as Connectivity Map (CMap) (Lamb et al., 2006; Subramanian et al., 2017) and DECCODE (Napolitano et al., 2021), have been designed to predict chemical compounds on the basis of the similarity between pre-compiled transcriptional changes before and after perturbations and query gene signatures. Although these methods have an advantage in

making accurate predictions for chemical compounds present in the compendium, they are unable to predict compounds that do not exist in the compendium and consequently their top predictions may be suboptimal. In addition, these methods rely on differentially expressed genes (DEGs), which are not optimal for recognizing small differences in cell conversion TF cocktails that result in distinct cell types or phenotypes. Furthermore, the current existing compendia of gene expression profiles analyzed after cell perturbations consist mostly of cancer cells (Lamb et al., 2006; Subramanian et al., 2017), which are known to possess signal transduction pathways and transcriptional logics that are significantly different from those of non-cancer cells (Sharma and Petsalaki, 2019; Pawson and Warner, 2007). Therefore, the compendia are not suitable for identifying chemical compounds for cellular conversions on normal cells.

Here, we present SiPer (signaling proteins and chemical perturbagens), a single-cell RNA sequencing (scRNA-seq)-based computational platform that identifies signaling proteins and chemical perturbagens (including small molecules, drugs, and cytokines) for targeting desired sets of TFs to induce conversion of cell populations. SiPer integrates an initial cell-specific network model with a manually collected compendium of 5,591 experimentally generated transcriptional signatures of non-cancer cells within 6 h before and after chemical perturbations. The application of SiPer correctly identifies perturbagens and their corresponding protein targets of known chemical-induced cell conversion examples, including cellular phenotypic state conversion and cell type conversion. Furthermore, we apply SiPer to identify chemical compounds for driving the differentiation of hepatic progenitors into functional human induced hepatocytes (hiHeps), which resemble primary hepatocytes (PHs) in cell identify and functionality. Finally, we demonstrate that the approach followed by SiPer, which relies on targeting cellular conversion TFs, generates more accurate predictions for chemical compounds than the existing methods (Subramanian et al., 2017; Napolitano et al., 2021).

In summary, SiPer is a valuable resource for funneling efforts toward the establishment of high-quality chemical cellular conversion protocols, enabling the design of new experimental strategies in regenerative medicine. SiPer is freely available as a web application at https://siper.uni.lu.

RESULTS

Overview of SiPer algorithm

SiPer is designed to prioritize chemical perturbagens and their corresponding protein targets in the intracellular signaling network (signaling proteins) that have high possibilities to activate/inhibit desired sets of query TFs for engineering cell populations (Figure 1A). The algorithm of SiPer is composed of three major stages (Figure 1B; experimental procedures). First, SiPer pre-selects candidate signaling proteins from the built-in perturbation database focusing on non-cancer cells, which is called non-cancer cell perturbation compendium (NCPC). NCPC is composed of two parts, Perturb-reTFs and Perturb-targets (Figure 1C). Perturb-reTFs is a compendium of response transcriptional signatures (differentially expressed TFs [DETFs]) to each signaling perturbagen (supplemental experimental procedures), and Perturb-targets is a compendium of the protein targets of each signaling perturbagen (supplemental experimental procedures). NCPC contained 5,591 transcriptional profiles in response to chemical perturbations across 134 different types of non-cancer cells from three species, human, mouse, and rat (Table S1). These profiles were derived from 2,386 unique signaling perturbagens, which covered 2,072 TFs in both activation (up) and inhibition (down) directions. In particular, we counted the frequencies of TFs that are known to be important for developmental processes (developmental TFs) and pioneer transcription factors (PFs), as these TFs carry the ability to alternate cell fates and should be contained in NCPC in sufficient amounts. The definitions of developmental TFs and PFs were obtained from Huilgol et al. (2019) and from Ribeiro et al. (2021) and Sunkel and Stanton (2021), respectively. Except for the three TFs (*HMGN1*, *DEC1*, and *DEC2*) that are not defined as TFs in our TF list, all the developmental TFs were observed multiple times in both activation and inhibition directions and all the PFs were observed multiple times (Table S2). In fact, the developmental TFs and PFs are significantly more frequently observed in NCPC than the other TFs (Figure S1). Thus, the presented evidence suggests that NCPC contains a sufficient number of each of the TFs that are important for cell conversion. Second, the candidate signaling proteins identified by stage 1 do not consider the initial state of the query cell type. In stage 2, SiPer further filters out the signaling proteins predicted from stage 1 that are not specific to the query cell type. SiPer combines scRNA-seq data of the initial cell population and the prior knowledge network (PKN) (supplemental experimental procedures) to prioritize signaling proteins that are specific to the initial cellular state (Figure 1B). The PKN consists of two layers, the upstream signaling network and downstream TF-TF interactions. These two layers of networks are combined by interface TFs, which mediate the signal transduction from cytoplasm to the nucleus. This resulted in complete PKNs including 6,633 genes and 96,137 interactions for human and 6,825 genes and 109,789 interactions for mouse and rat (Table S3). Finally, SiPer divides the final set of predicted signaling proteins from stage 2 into different functional

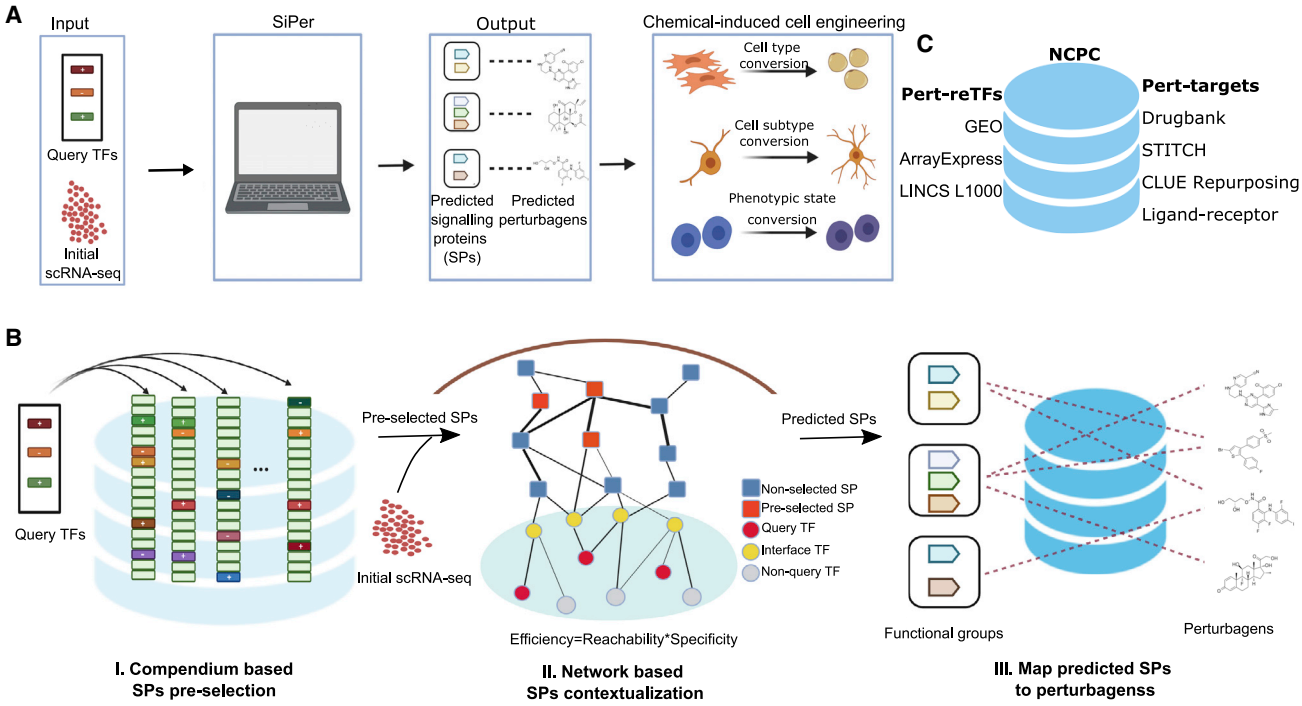


Figure 1. Schematic outline of SiPer

(A) The workflow of SiPer. The input of SiPer includes the initial scRNA-seq and desired TFs provided by the user. Given input, SiPer identifies chemical perturbagens and corresponding signaling protein targets (SPs) to different kinds of cellular conversions, including conversions between cell types, cell subtypes and phenotypic states.

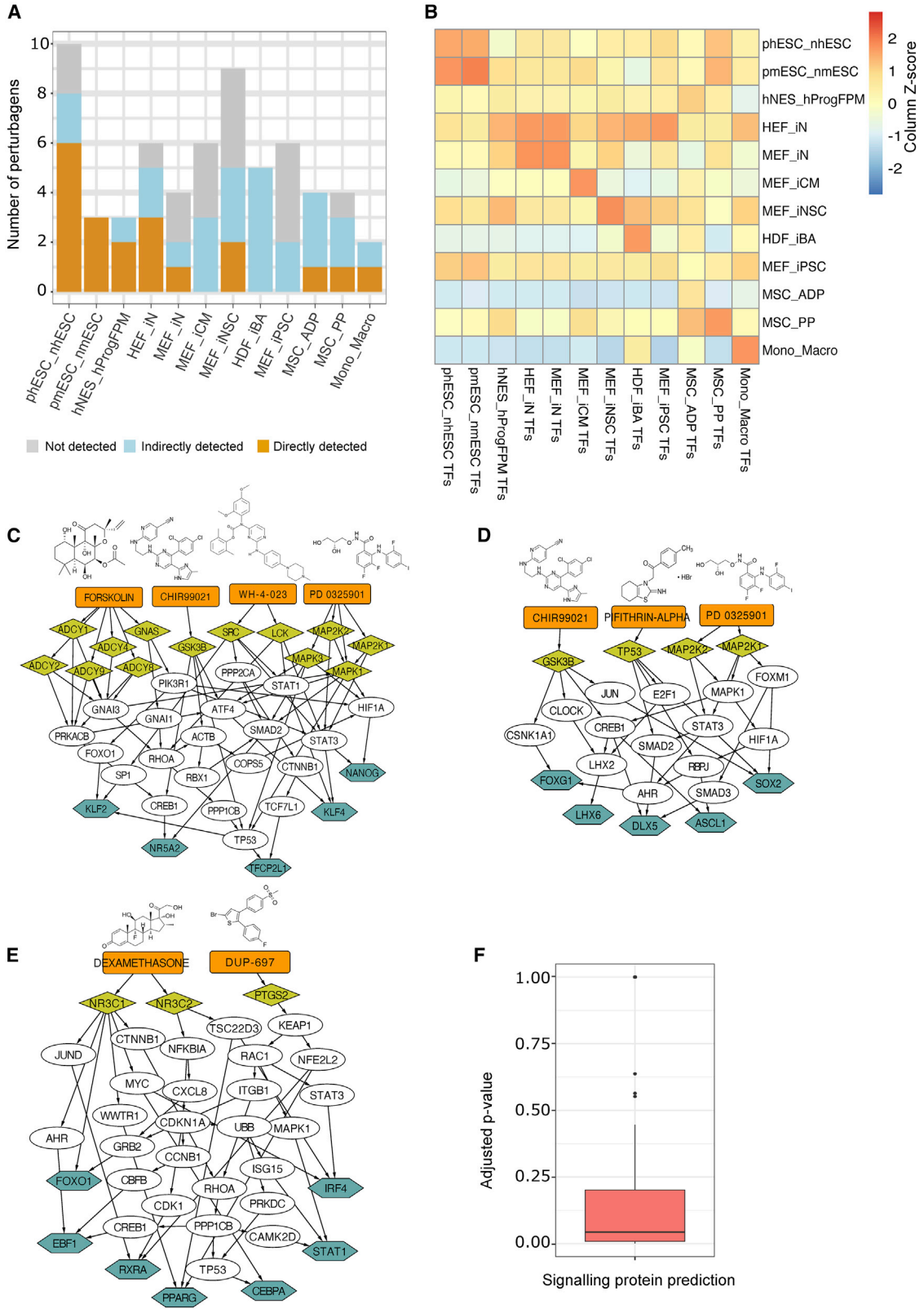
(B) SiPer contains three major stages: (1) pre-selection of candidate signaling proteins from the built-in NCPC by using query TFs, (2) network-based modeling to predict initial cellular state-specific signaling proteins targeting the query set of TFs using scRNA-seq of initial cell type/state, and (3) identification of perturbagens targeting the predicted signaling proteins.

(C) 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, ArrayExpress and LINCS L1000. Perturb-targets contains perturbagens and their signaling protein targets from Drugbank, STITCH, CLUE Repurposing, and manually curated ligand-receptor interactions.

groups on the basis of Reactome signaling pathways. For each group of signaling proteins, SiPer identifies optimal perturbagens on the basis of the similarity between their target proteins and the predicted signaling proteins (Figure 1B). In addition, SiPer annotates the functional mechanism of perturbagens and automatically separates perturbagens into different groups on the basis of their similarities in target proteins, which allows users to design an optimal combination of perturbagens targeting distinct functional groups.

SiPer accurately predicts perturbagens for various cellular conversions

SiPer was first applied to the conversion between different phenotypic states of the same cell type, where the conversion chemical cocktails are experimentally validated (Figure 2A; Table S4). Predicted signaling proteins exhibited high specificity to the respective query TFs (Figure 2B) and allowed the identification of perturbagens that target these proteins. For example, for the conversion of lineage primed embryonic stem cells (ESCs) to a more plastic naive state in both human and mouse, SiPer was able to predict both perturbagens and signaling proteins. In particular, the two classical inhibitors (2i) CHIR99021 and PD0325901 (Ying et al., 2008; Zimmerlin et al., 2016) were correctly identified. In addition to these two inhibitors, SiPer predicted LIF in both mouse ESCs (mESCs) and human ESCs (hESCs), which is an important cytokine required for sustaining ESC self-renewal (Niwa et al., 1998). Furthermore, as the 2i is not sufficient to maintain a stable naive state of hESCs (Zimmerlin et al., 2017), additional chemical compounds for the maintenance of the ground state of naive hESCs, such as forskolin (Park et al., 2018) and WH-4-023 (Theunissen et al., 2014), were identified by SiPer as well. Moreover, Theunissen et al. (2014) showed that activin A enhanced the efficiency of naive hESCs conversion from the primed state, which was also predicted by SiPer (Theunissen et al., 2014). SiPer network



(legend on next page)

visualization of the underlying putative signaling cascades (supplemental experimental procedures) also identified CTNNB1 (Figure 2C), the central effector of WNT signaling 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 2C), consistent with their key role in maintaining pluripotency (Sakaki-Yumoto et al., 2013; Zhang et al., 2011). We further investigated the predictive power of SiPer in broad cell type conversions with reported chemical conversion cocktails, including cell reprogramming and differentiation (Figure 2A; Table S4). Specifically, given the conversion TFs for GABAergic neurons from human fibroblasts reported by (Colasante et al., 2015), multiple essential chemical compounds and corresponding protein targets reported in previous study were also 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 2D) (Hori et al., 2008; Komine et al., 2011). Next, we applied SiPer to identify perturbagens and key signaling proteins in the context of cellular differentiation (Table S4). 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 signaling 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, 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 2E), indicating that they might regulate the differentiation TFs in a coordinated fashion.

We performed Fisher's exact test to examine whether SiPer's predictions were significantly enriched for correct perturbagens in comparison with the total number of predicted perturbagens and to the total number of correct perturbagens, which resulted in a significant p value ($<2.7 \times 10^{-11}$). Note that to compensate for the small numbers of correct perturbagens and example cases, the test was carried out by aggregating all the examples. It is also important to emphasize that this p value was computed under the assumption that all unknown ones were false positives. However, many perturbagens had similar functions or protein targets as the correct ones and are potentially novel hits that have not been reported previously (e.g., light blue bars in Figure 2A). Therefore, the enrichment for true hits is likely higher.

Taken together, these results demonstrate that the current challenge of replacing TFs with chemical compounds in cellular conversion protocols, especially conversions between phenotypic states and cell subtypes, could be potentially addressed by SiPer.

Statistical evaluation of SiPer performance

Next, we performed a large-scale statistical assessment of SiPer's predictive accuracy to assess SiPer's general applicability beyond the example cases above. To this end, we collected 200 benchmarking datasets in which the perturbagen and signaling protein targets of the perturbagen are known and the scRNA-seq data of initial cellular state is available (supplemental experimental procedures; Table S1). Note, these datasets are not necessarily related to cell identity conversion (e.g., differentiation, reprogramming) but can also be related to phenotypic changes, such as cell activation or disease treatment. The performance of signaling protein prediction by SiPer was examined using Fisher's exact test for both individual datasets and the

Figure 2. Application of SiPer to different kinds of cellular conversion

(A) Number of experimentally validated perturbagens predicted by SiPer in different cellular conversion cases. Orange, experimentally validated perturbagens that are directly predicted. Blue, perturbagens that are not predicted directly, but their protein targets or the other perturbagens with similar function are predicted. Gray, perturbagens are not predicted in either way.

(B) SiPer's efficiency score matrix denoting the regulatory potential of signaling proteins to downstream TFs. Each cell in the heatmap represents the average efficiency score of top 20 predicted signaling proteins to corresponding conversion TF sets. For example, "phESC_nhESC" means primed hESC converts to naive hESC. Abbreviations for (A) and (B): phESC/pmESC, primed hESC/mESC; nhESC/nmESC, naive hESC/mESC; hNES, human neuroepithelial; hProgFPM, human floor plate midbrain progenitor; HEF/MEF, human/mouse embryonic fibroblast; iN, induced GABAergic neuron; iCM, induced cardiomyocytes; iNSC, induced neural stem cell; HDF, human dermal fibroblast; iBA, induced brown adipocyte; iPSC, induced pluripotent stem cell; MSC, mesenchymal stem cell; ADP, adipocyte; PP, pancreatic progenitor; Mono, monocyte; Macro, macrophage.

(C–E) SiPer network visualization of putative signaling cascades between predicted perturbagens (orange rectangle), predicted signaling protein targets (yellow diamond), intermediate signaling proteins (white ellipse), and query TFs (blue hexagon). (C) Phenotypic state conversion of primed ESCs to naive ESCs. (D) Reprogramming into GABAergic neurons from HEFs. (E) differentiation of MSCs to adipocytes.

(F) Adjusted p value of Fisher's exact test for 200 benchmarking datasets in the prediction of signaling proteins.

aggregation of the 200 datasets. The p values for the individual examples were corrected for multiple testing using the Benjamini-Hochberg method. The result revealed that 100 of the 200 datasets had adjusted p values ≤ 0.05, with a median adjusted p value of 0.043 (Figure 2F). As each of the 200 datasets has only one correct perturbagen, we could not calculate the enrichment p value for individual datasets and only the aggregation of all the datasets was considered for the enrichment of perturbagen prediction. The resultant p value was significant ($<2.2 \times 10^{-16}$). Thus, our large-scale benchmarking corroborates that SiPer preferentially predicts both correct signaling proteins and perturbagens.

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. We previously established a two-step lineage reprogramming strategy to generate functional human hepatocytes (Xie et al., 2019): human fibroblasts were first reprogrammed into hepatic progenitors and then, in a second step, induced into functional hiHeps by a two-chemical combination of forskolin and SB431542 (2C). However, in the established protocol, hiHeps cultured long term on the basis of the hepatocyte culture medium gradually show excessive accumulation of lipid droplets (Figure S2A), while the widely used Williams' E medium does not lead to abnormal lipid metabolism but cannot support full maturation of hiHeps (Figure S2B). Thus, we applied SiPer to predict additional perturbagens that may activate key hepatic TFs 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 DETFs in PHs compared with hepatic progenitors ($\log_2$ fold change > 4) were selected (supplemental experimental procedures) as query TFs for SiPer analysis (Figure 3A). scRNA-seq of hepatic progenitor cells together with eight hepatic key TFs were used as input 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. Moreover, additional thirteen chemical compounds of ten different functional mechanism groups were identified (Figure S2C; Table S5).

To promote hiHeps maturation on the basis of Williams' E medium with 2C (W2C), 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 with W2C (Figure 3B). 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 S2D). To find the essential combination of the seven compounds, omitting each factor determined TSA to be dispensable for hepatic functional gene expression (Figure S2E). Thus, the identity and function of hiHeps cultured in W2C together with hydrocortisone, progesterone, Bio and U0126 (W6C) were further evaluated.

First of all, morphological examination revealed that W6C-cultured hiHeps showed typical hepatocyte morphology, without presenting signs of abnormal lipid metabolism (Figure 3C). Global hierarchical clustering revealed that W6C-cultured hiHeps were clustered closely with PHs, while separately from hepatic progenitors and W2C-cultured hiHeps (Figure 3D). Moreover, the expression levels of hepatic TFs and key hepatic functional gene sets involved in glucose, lipid cholesterol, fatty acid, and drug metabolism (Table S5) were similar between W6C-cultured hiHeps and PHs (Figure 3E), which were further validated by qPCR analysis (Figure 3F). Besides, we confirmed the upregulation of protein expression of ALB, drug- metabolizing enzymes CYP3A4 and CYP1A2, and glutamine synthetase in W6C-cultured hiHeps by immunostaining (Figure S2F).

We further characterized the functionality of W6C-cultured hiHeps. Results showed that W6C supported better glycogen synthesis and lipoprotein uptake in hiHeps (Figure 4A) and increased bile acid synthesis activity (Figure 4B). Importantly, W6C enabled long-term culture of hiHeps up to at least 28 days with stably maintained albumin secretion at levels similar to PHs (Figure 4C), good hepatocyte morphology, and cell survival, whereas W2C-cultured hiHeps showed low albumin secretion with gradual cell death (Figure S2G). In addition, W6C-cultured hiHeps presented similar levels of drug-metabolizing activity of CYP3A4 and CYP1A2, key hepatocyte functions compared with those of PHs (Figure 4D).

To investigate the engraftment ability of W6C-cultured hiHeps *in vivo*, we transplanted them into the Tet-uPA (urokinase-type plasminogen activator)/Rag2$^{-/-}$/γc$^{-/-}$ liver injury mouse model. The secreted human albumin levels in mouse serum 6 weeks post-transplantation were examined. We found that the human albumin levels were comparable between mice transplanted with W6C-cultured hiHeps and PHs (Figure 4E). Importantly, the mature hepatocyte markers ALB, CYP3A4, and CYP2C9 were detected by immunostaining in mouse livers transplanted with W6C-cultured hiHeps (Figure 4F).

At last, expression levels of query TFs were also verified to be upregulated in W6C-cultured hiHeps, along with other

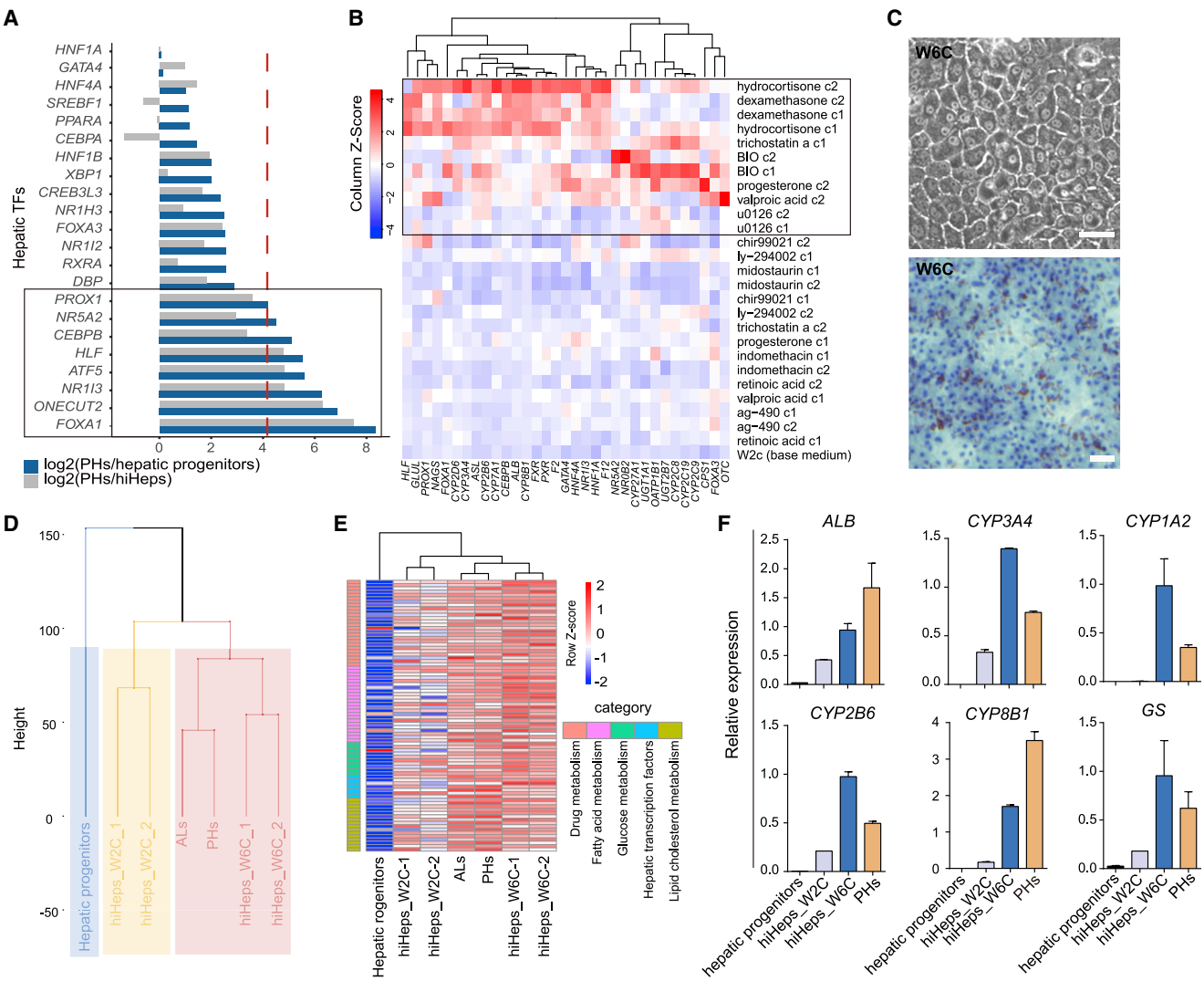


Figure 3. Generation of hiHeps by applying SiPer predicted chemical perturbagens

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

(B) 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.

(C) A representative bright field image and oil-red staining image of lipid synthesis and accumulation in hiHeps cultured in W6C. Scale bar, 50 μm.

(D) Hierarchical clustering of global gene expression of hepatic progenitors, hiHeps cultured in W2C and W6C, PHs, and adult liver tissues (ALs) by RNA-seq analysis.

(E) 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. Panel of genes analyzed listed in Table S5.

(F) qRT-PCR analysis of gene expression of key hepatocyte markers. Relative expression was normalized to hiHeps_W2C. n = 2 technical replicates. Data are mean ± MSE.

key hepatic TFs (Figure S3A), which corroborated that SiPer predicted chemical compounds targeting desired TFs. Through SiPer's network visualization, the intermediate signaling proteins connecting the predicted perturbagens and query TFs were illustrated, including known nuclear receptor proteins such as PPARA, RXRA, and AHR, key regulators of hepatocyte metabolic network, suggesting their importance as effectors of hepatic maturation (Figure S3B).

Collectively, these results showed that SiPer effectively identified perturbagens to induce hiHeps maturation in a

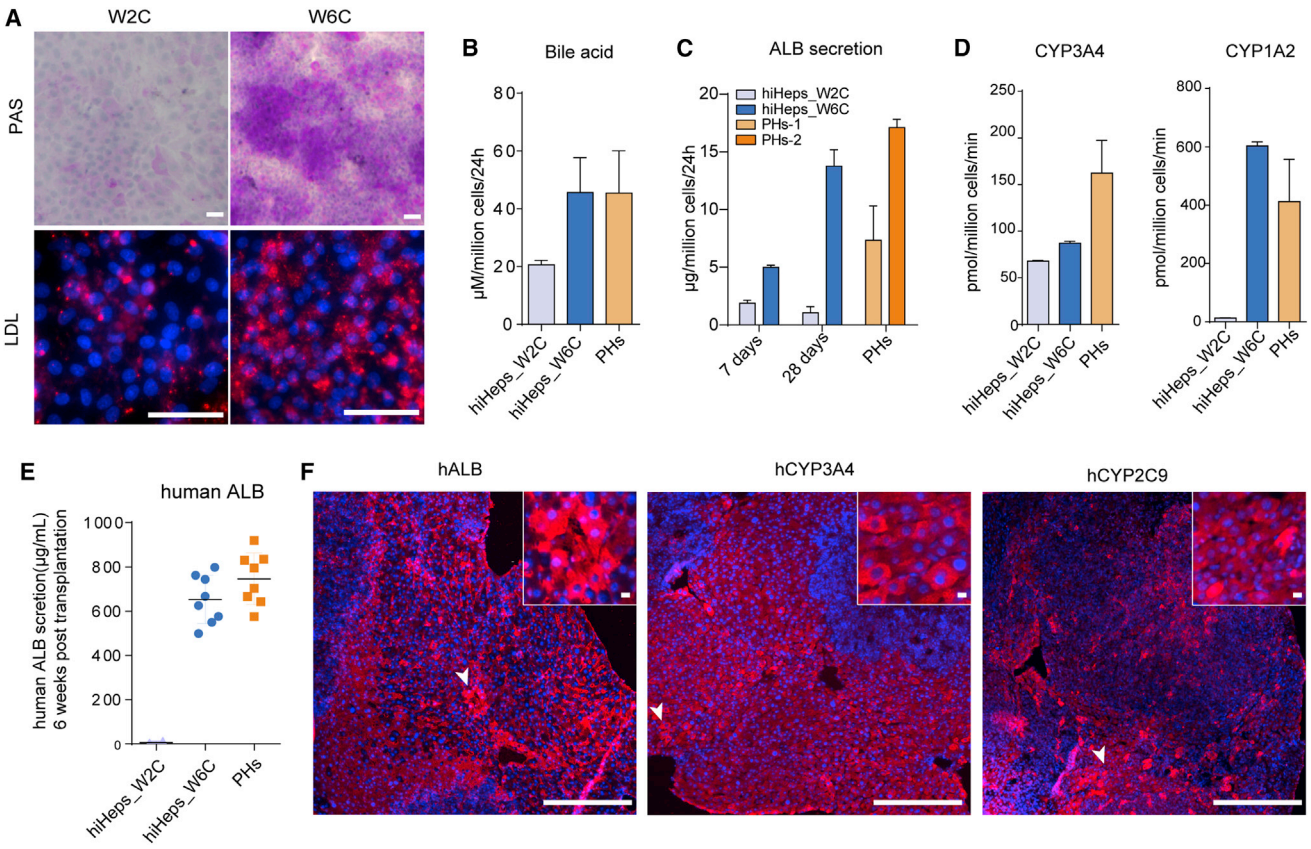


Figure 4. Functional characterization of the W6C-cultured hiHeps

(A) Analysis of key hepatic functions of hiHeps cultured in W2C and W6C: PAS staining of glycogen synthesis (upper panel) and low-density lipoprotein (LDL) uptake staining (lower panel). Scale bar, 50 μm.

(B) Bile acid secretion of PHs (n = 6 independent experiments) and hiHeps (n = 3 independent experiments) cultured in W2C and W6C. Data are mean ± MSE.

(C) ALB secretion of PHs (n = 6 independent experiments) and hiHeps (n = 3 independent experiments) cultured in W2C and W6C at days 7 and 28 of culture. Data are mean ± MSE.

(D) UPLC/MS/MS (ultra performance liquid chromatography-tandem mass spectrometer analysis) of drug-metabolizing activity of CYP3A4 and CYP1A2. n = 3 independent experiments. Data are mean ± MSE.

(E) Human ALB secretion level in mouse serum after 6 weeks of transplantation of hiHeps cultured in W2C (n = 2 independent experiments) and W6C (n = 8 independent experiments), and PHs (n = 8 independent experiments) in URG mice.

(F) Immunofluorescent staining of human specific hALB, hCYP3A4, and hCYP2C9 in mouse liver transplanted with hiHeps cultured in W6C. Arrows indicte the enlarged area. Scale bar, 100 μm.

Williams' E medium. Importantly, the new formulation was identified through a very simplified, straightforward experimental process because of the integration of SiPer in the workflow.

Comparison with existing methods for the prediction of chemical compounds

To date, no existing methods are designed to predict chemical compounds for cellular conversion on the basis of a handful of desired TFs. The existing predict chemical compounds solely on the basis of DEGs between initial and target cell types by using CMap database methods (Subramanian et al., 2017; Napolitano et al., 2021). However, DEGs usually contain many non-conversion TFs that are more commonly expressed in other cell types and targeting all DEGs could lead to the identification of chemical compounds that do not act on the desired cell conversion TFs. Moreover, these methods have the requirement of minimum number of input genes. Therefore, we cannot modify the input by using the small number of conversion TFs. Nevertheless, a comparison was performed between SiPer and other two computational methods, CMap query (Subramanian et al., 2017) and DECCODE (Napolitano et al., 2021) using DEGs (supplemental experimental

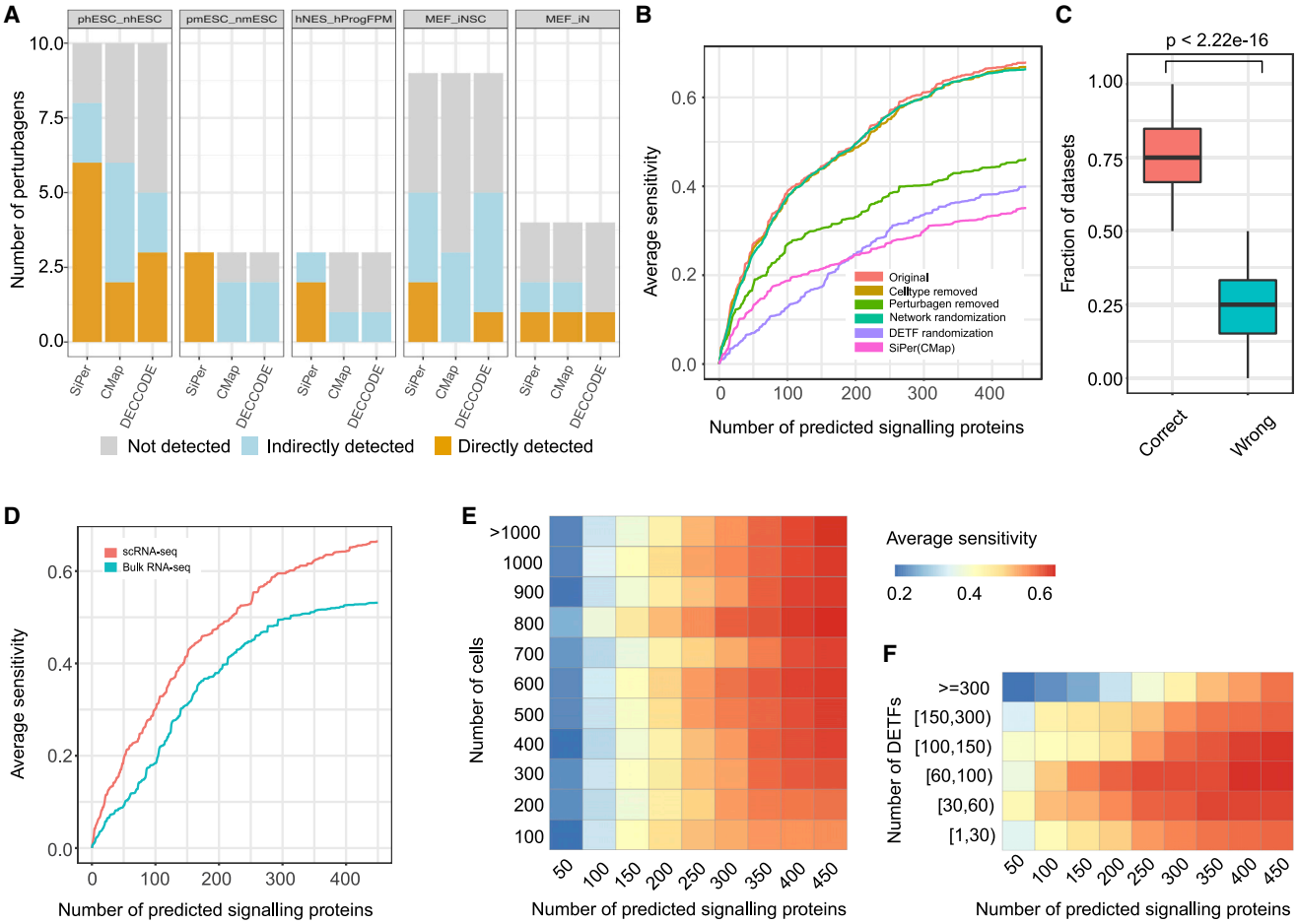


Figure 5. Comparison with existing methods and robustness evaluation of SiPer

(A) Comparison between SiPer and two existing computational methods for the prediction of chemical compounds in five cellular conversion cases. Orange, experimentally validated perturbagens that are directly predicted. Blue, perturbagens that are not predicted directly, but their protein targets or the other perturbagens with similar function are predicted. Gray, perturbagens are not predicted.

(B) Robustness analysis of SiPer in terms of average sensitivity tested by removing NCPC datasets with same cell type as test dataset (cell type removed), removing NCPC datasets with same perturbagen as test dataset (perturbagen removed), random removal of 10% interactions in PKN (network randomization), randomly selecting same number of TFs as the number of test DETFs (DETF randomization), and replacing the NCPC compendium with the CMap database (SiPer [CMap]).

(C) The fraction of datasets correctly clustered to their corresponding class (i.e., normal or cancer) (one-sided Wilcoxon test, $p < 2.22 \times 10^{-16}$).

(D) Evaluation of SiPer by replacing the scRNA-seq data of initial cellular state with bulk RNA-seq data.

(E) Robustness evaluation of SiPer with respect to input cell number by randomly selecting cells from the input scRNA-seq data.

(F) Robustness evaluation of SiPer with respect to DETF number by splitting the benchmarking datasets into different subsets on the basis of the number of DETFs.

procedures). Among all the cellular conversion cases presented in Table S4, five had the gene expression profiles of both initial and target cell types. We applied both CMap and DECCODE to these five examples by using their DEGs as input to predict the chemical compounds for the respective conversions. We examined the number of experimentally used perturbagens that were predicted by each method and SiPer showed overall better performance than the other two methods (Figure 5A). Specifically, SiPer predicted not only the perturbagens identified by CMap and DECCODE but also the ones that were not predicted by them, including LCK/SRC inhibitor, HMG-CoA reductase inhibitor and Wnt antagonist (Table S6). These results indicate that SiPer generated more accurate predictions for chemical compounds than those inferred by the existing computational methods that rely on DEGs.

As the cell conversion examples mentioned above are relatively small in size, we further performed the comparison analysis using the 200 benchmarking datasets described above. The result revealed that both CMap and DECCODE had only two datasets with correct perturbagen prediction, whereas SiPer had 123. This huge performance difference was partly because the former two methods are not designed to predict protein ligand perturbagens and their prediction is limited to only chemical perturbagens. However, even when only considering the 92 datasets in which the correct perturbagen is a chemical compound, SiPer correctly predicted 30 datasets (Fisher's exact $p < 2.2 \times 10^{-12}$) compared with 1 and 2 datasets for DECCODE and CMap, respectively (Fisher's exact $p > 0.99$). These results suggest that SiPer significantly outperforms both methods in chemical perturbagen prediction.

Robustness evaluation of SiPer

SiPer first predicts signaling proteins from stage 1 and stage 2 and then predicts the chemical compounds targeting these predicted signaling proteins in stage 3. The parameters including features in initial gene expression profile, NCPC and background network only have effects on the prediction of signaling proteins. Therefore, the robustness of SiPer to these different parameters for the prediction of signaling proteins were assessed by using the 200 benchmarking datasets (supplemental experimental procedures). The average sensitivity (supplemental experimental procedures) in identifying the direct protein targets of the correct perturbagens across all the datasets was evaluated to assess the robustness of SiPer. We showed that SiPer was robust to the change in the PKN, the cell types in NCPC and the number of cells (Figures 5B and 5E). However, when NCPC was replaced with the cancer datasets (Subramanian et al., 2017), the performance of SiPer significantly decreased (Figure 5B). To investigate the reason for this decreased performance, we performed a hierarchical clustering analysis on the basis of DETFs of 1,956 perturbagens (Table S1), by which both non-cancer and cancer cell types were stimulated (supplemental experimental procedures). The fraction of cells correctly clustered to their respective class (i.e., non-cancer or cancer) was significantly higher than mis-clustered ones (one-sided Wilcoxon test, $p < 2.22 \times 10^{-16}$) (Figure 5C). This result indicates the responses between non-cancer cells are similar than those between cancer and non-cancers under same perturbation. Therefore, it is essential to construct a perturbation compendium focusing on non-cancer cells, such as NCPC, for accurate prediction of signaling perturbagens when targeting non-cancer cells. We also showed that the performance of SiPer consistently decreased when using bulk RNA-seq data of initial cellular state instead of using scRNA-seq data (Figure 5D). In addition, to validate the efficiency of using the gene profiles of the initial cell type for the prediction of perturbagens, we compared the performance of SiPer between the true scRNA-seq data and randomly selecting the scRNA-seq data from benchmarking datasets. Only 72 of 200 datasets whose true positive perturbagens were predicted in the randomly selected scRNA-seq data, while the perturbagen of 123 datasets were predicted using their corresponding true scRNA-seq data. These results show that it is essential to consider the initial cellular state for the prediction. Moreover, we also showed that the predictions of SiPer are specific to query TFs (Figure 5B) and the overall performance was maintained across datasets with numbers of query TFs less than 100 (Figure 5F), which is presumably because of the lack of specificity of target signaling proteins affecting a large number of downstream TFs. Altogether, these systematic assessments demonstrate that SiPer is a robust platform for the changes in its parameters and able to obtain the predictions that are specific to query TFs.

DISCUSSION

Studies have shown that the cellular conversion can be induced by the perturbation of a handful of key TFs. Instead of direct manipulation of these cellular conversion TFs, using chemical compounds targeting signaling pathways and in turn inducing the expression change of desired TFs is a safer and less laborious strategy. Therefore, in this study, we have developed a computational platform to identify chemical compounds targeting desired TFs for cellular conversions. However, the inference of the exact signaling pathways acting on specific TFs remains a challenge because of the scarcity of protein activity data, such as protein phosphorylation measurements. Thus, instead of predicting exact signaling pathways, SiPer infers signaling proteins from the pre-compiled perturbation compendium, whose perturbations were shown to result in the dysregulation of similar sets of TFs to the query TFs. Although a similar strategy has been taken in previous studies (Lamb et al., 2006; Schubert et al., 2018; Subramanian et al., 2017), their compendia were derived mainly from cancer cells, whose signaling pathways has exhibited significant rewiring from the normal counterparts (Sharma and Petsalaki, 2019). Therefore, NCPC provides a valuable resource for research focusing on non-cancer cells. Moreover, SiPer integrates NCPC with a network model by taking advantage of scRNA-seq data that reveals the heterogeneity and asynchrony of individual cells. This allows SiPer to be applied to any novel cell populations identified by

scRNA-seq. Indeed, many known signaling proteins and chemical compounds were correctly identified by SiPer for conversion between different cell types and closely related cell populations. Because of the scarcity of available cellular conversion cases for method validation, we further collected 200 benchmarking datasets which are not necessarily related to cell identity conversion but related to phenotypic changes. We showed that the predictions by SiPer targeting specific TFs were more accurate than those generated by other computational methods on the basis of DEGs. Indeed, DEGs may not include cell conversion TFs, and usually contain many non-conversion TFs that are more commonly expressed in other cell types. For this reason, targeting all DEGs could lead to the identification of chemical compounds that do not act on the desired cell conversion TFs. Moreover, although we demonstrated the importance of initial scRNA-seq data on the basis of benchmarking datasets, the in-silico validation to demonstrate the efficiency of identifying different chemical compounds for cellular conversion cases depending on the initial cellular state-specific signaling proteins is still lacking. This is because no clearly true negative chemical compounds (the chemical compounds used in one case are not suitable in another case) in the cellular conversion are available to evaluate the performance currently. Overall, further validation on the basis of more benchmarking datasets and cellular conversion cases can be carried out as the development of data in the future.

Moreover, we applied SiPer for the development of a novel protocol for human hepatic maturation. A recently reported protocol has enabled the successful generation of hiHeps (Xie et al., 2019). However, this protocol uses 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 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 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 in terms of molecular identity and functionality without abnormal lipid accumulation (Figures 4D–4K). Notably, integrating SiPer predictions into the workflow dramatically simplified the experimental design, ultimately resulting in a straightforward, streamlined experimental process.

SiPer is designed to predict chemical compounds in one step from initial to target cell type given the desired target TFs. However, many long-term cellular conversions are induced by more than one transition waves, which normally requires precise manipulation of the cell fate transition in a stepwise manner (Zhao et al., 2015; Touboul et al., 2016; Guan et al., 2022). In this scenario, SiPer can be also generalized to apply in each conversion step only if users have prior knowledge about intermediate stages and the target TFs for each stage.

In summary, SiPer constitutes a valuable computational platform to facilitate the design of cell conversion protocols using chemical compounds, which holds great promise for both basic cell research and regenerative medicine. Users can easily apply SiPer to any cellular systems of their interest by accessing the web interface.

EXPERIMENTAL PROCEDURES

Resource availability

Corresponding author

Further information and requests for resources and reagents should be directed to and will be fulfilled by the corresponding author, Antonio del Sol (antonio.delsol@uni.lu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

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

SiPer was implemented in R, and the 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.

Algorithm of SiPer

Stage 1: Identification of candidates of signaling proteins from NCPC

SiPer first selects candidates of signaling proteins from NCPC on the basis of a set of query TFs with the expression direction information (i.e., up- or down-regulation). SiPer calculates the similarity between query TFs and each reference in Perturb-reTFs of NCPC by a modified Jaccard similarity coefficient (supplemental experimental procedures), which ensures SiPer to consider the number as well as the effects of common TFs between the query and reference. The perturbagens with the modified Jaccard index larger than 2 Z score were selected (supplemental experimental procedures). The signaling proteins of these selected perturbagens are then ranked by their frequency and the ones ranked in top 40% are retrieved as candidate signaling proteins whose perturbations could affect the query set of TFs. The details of the parameter optimization and selection are described in supplemental experimental procedures. The "activation" sign is assigned to candidate signaling 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.

Stage 2: Prediction of signaling proteins with network model

Candidate signaling proteins identified from NCPC are further filtered to predict the final set of signaling proteins by taking into

account the gene expression state of starting cells and PKN among signaling proteins and TFs. The algorithm consists of two major steps as following. First, SiPer simulates the signal transduction from candidate signaling proteins from stage 1 to initial response TFs and calculates an efficiency score $EF_{s\to t}$ defined as

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

where the $R_{s\to t}$ is the reachability score (supplemental experimental procedures) measured by the weighted shortest path from signaling protein to a response TF t and *reverted* $S_{s\to t}$ is the specificity score (supplemental experimental procedures) quantified by the inverse of signaling entropy to measure the certainty of the signal transmitting along with the shortest path. The efficiency score quantifies both the strength and specificity with which a signaling protein acts on a downstream TF. Second, SiPer also ensures that the predicted signaling proteins specifically act on the query set of TFs and have a minimized effect on non-query ones by computing Jensen-Shannon divergence (JSD) for each signaling protein

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

where P, Q are the observed and ideal efficiency score vectors, respectively, and $M = \frac{1}{2}(\mathrm{P}+Q)$ and D is Kullback-Leibler divergence (supplemental experimental procedures). The lower JSD value indicates that the signaling protein can more specifically target the query TFs, which implies that the phenotypic response would also be closer to our expectation. Once JSD value is calculated for all candidate signaling proteins inferred from NCPC, the signaling proteins with JSD value not equal to 1 are selected as final candidates.

Stage 3: Identification of perturbagens targeting predicted signaling proteins

SiPer further predicts the perturbagens targeting predicted signaling proteins. SiPer first performs over-representation analysis on predicted signaling proteins to first divide the proteins into different functional groups on the basis of Reactome signaling pathways. Then for each group of signaling proteins, SiPer identifies perturbagens by calculating the similarity between their target proteins and predicted signaling proteins with another modified Jaccard index (supplemental experimental procedures), which ensures SiPer to identify the perturbagens targeting predicted signaling proteins with consistent modes as predicted. Finally, SiPer further merges the top predicted perturbagens on the basis of their target similarity. This allows users to design optimal combination of chemical compounds from different functional groups. The details of stage 3 are described in supplemental experimental procedures.

Ethical statement

The present study was approved by the Clinical Research Ethics Committee of China-Japan Friendship Hospital (approval number 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.

Cell culture

The methods for culturing primary human hepatocytes and hiHeps have been described previously (Xie et al., 2019). The details are described in supplemental experimental procedures.

Measurements of drug-metabolizing activity of CYP450s

Cells were incubated with indicated substrates (testosterone for CYP3A4, phenacetin for CYP1A2), and the level of products was then analyzed by ultraperformance liquid chromatography-tandem mass spectrometry (UPLC/MS/MS) (supplemental experimental procedures).

Transplantation

Tet-uPA/Rag2$^{-/-}$/γc−/− (URG) mice on a BALB/c background were purchased from Beijing Vitalstar Biotechnology. Cells for transplantation were suspended into single cells in HCM medium; 2×10^6 cells in 200 μL suspension were injected into the spleen of the mice. The detection of human albumin secretion and CYP450 expression in mouse liver are described in supplemental experimental procedures.

RNA sequencing and bioinformatics analysis

The details of generation and pre-processing for bulk RNA-seq and scRNA-seq data for the hepatic maturation are described in supplemental experimental procedures.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stemcr.2022.10.013.

AUTHOR CONTRIBUTIONS

A.d.S. conceived the overall study and supervised its computational part. M.Z. developed the computational method and compiled datasets. M.Z. and S.O. performed the analysis. H.D. designed and supervised the cell culture experiments. B.X. performed the experiments. M.Z., B.X., S.O., S.L., H.D., and A.d.S. wrote the manuscript.

ACKNOWLEDGMENTS

M.Z. is supported by CORE grant from Fonds National de la Recherche Luxembourg (C15/BM/10397420). S.O. is supported by CORE grant from Fonds National de la Recherche Luxembourg (C19/BM/13624979). H.D. is supported by the National Natural Science Foundation of China (32288102) and the National Key R&D Program of China (2018YFA0108102). We thank Jacek Lebioda for the development of web application.

CONFLICTS OF INTEREST

The authors declare no competing interests.

Received: April 10, 2022
Revised: October 18, 2022
Accepted: October 19, 2022
Published: November 17, 2022

REFERENCES

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., Nie, B., Xie, M., Zhang, M., Wang, H., et al. (2016). Conversion of human fibroblasts into functional cardiomyocytes by small molecules. Science *6290*, 1216–1220.

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.

Cheng, L., Gao, L., Guan, W., Mao, J., Hu, W., Qiu, B., Zhao, J., Yu, Y., and Pei, G. (2015). Direct conversion of astrocytes into neuronal cells by drug cocktail. Cell Res. *25*, 1269–1272.

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.

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.

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 Rep. *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.

Guan, J., Wang, G., Wang, J., Zhang, Z., Fu, Y., Cheng, L., Meng, G., Lyu, Y., Zhu, J., Li, Y., et al. (2022). Chemical reprogramming of human somatic cells to pluripotent stem cells. Nature *605*, 325–331.

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.

Huilgol, D., Venkataramani, P., Nandi, S., and Bhattacharjee, S. (2019). Transcription factors that govern development and disease: an achilles heel in cancer. Genes *10*, 794.

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.

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. Gene Cell. *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.

Morris, S.A., Cahan, P., Li, H., Zhao, A.M., San Roman, A.K., Shivdasani, R.A., Collins, J.J., and Daley, G.Q. (2014). Dissecting engineered cell types and enhancing cell fate conversion via CellNet. Cell *158*, 889–902.

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.

Napolitano, F., Rapakoulia, T., Annunziata, P., Hasegawa, A., Cardon, M., Napolitano, S., Vaccaro, L., Iuliano, A., Wanderlingh, L.G., Kasukawa, T., et al. (2021). Automatic identification of small molecules that promote cell conversion and reprogramming. Stem Cell Rep. *16*, 1381–1390.

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.

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. *136*, 57921.

Pawson, T., and Warner, N. (2007). Oncogenic re-wiring of cellular signaling pathways. Oncogene *26*, 1268–1275.

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.

Ribeiro, M.M., Okawa, S., and Del Sol, A. (2021). TransSynW: a single-cell RNA-sequencing based web application to guide cell conversion experiments. Stem Cells Transl. Med. *10*, 230–238.

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.

Sharma, S., and Petsalaki, E. (2019). Large-scale datasets uncovering cell signalling networks in cancer: context matters. Curr. Opin. Genet. Dev. *54*, 118–124.

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.e17.

Sunkel, B.D., and Stanton, B.Z. (2021). Pioneer factors in development and cancer. iScience *24*, 103132.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

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*, 524–526.

Treutlein, B., Lee, Q.Y., Camp, J.G., Mall, M., Koh, W., Shariati, S.A.M., Sim, S., Neff, N.F., Skotheim, J.M., Wernig, M., and Quake, S.R. (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature *534*, 391–395.

Touboul, T., Chen, S., To, C.C., Mora-Castilla, S., Sabatini, K., Tukey, R.H., and Laurent, L.C. (2016). Stage-specific regulation of the WNT/β-catenin pathway enhances differentiation of hESCs into hepatocytes. J. Hepatol. *64*, 1315–1326.

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.

Ye, J., Ge, J., Zhang, X., Cheng, L., Zhang, Z., He, S., Wang, Y., Lin, H., Yang, W., Liu, J., et al. (2016). Pluripotent stem cells induced from mouse neural stem cells and small intestinal epithelial cells by small molecule compounds. Cell Res. *26*, 34–45.

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, 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, X., Yalcin, S., Lee, D.F., Yeh, T.Y.J., 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.

Zhao, Y., Zhao, T., Guan, J., Zhang, X., Fu, Y., Ye, J., Zhu, J., Meng, G., Ge, J., Yang, S., et al. (2015). A XEN-like state bridges somatic cells to pluripotency during chemical reprogramming. Cell *163*, 1678–1691.

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.