



Cell fate conversion prediction by group sparse optimization method utilizing single-cell and bulk OMICs data

Jing Qin , Yaohua Hu, Jen-Chih Yao, Ricky Wai Tak Leung, Yongqiang Zhou, Yiming Qin and Junwen Wang 

Corresponding author: Jing Qin, School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University, Shenzhen, 518107, China.
Email: qinj29@mail.sysu.edu.cn

Abstract

Cell fate conversion by overexpressing defined factors is a powerful tool in regenerative medicine. However, identifying key factors for cell fate conversion requires laborious experimental efforts; thus, many of such conversions have not been achieved yet. Nevertheless, cell fate conversions found in many published studies were incomplete as the expression of important gene sets could not be manipulated thoroughly. Therefore, the identification of master transcription factors for complete and efficient conversion is crucial to render this technology more applicable clinically. In the past decade, systematic analyses on various single-cell and bulk OMICs data have uncovered numerous gene regulatory mechanisms, and made it possible to predict master gene regulators during cell fate conversion. By virtue of the sparse structure of master transcription factors and the group structure of their simultaneous regulatory effects on the cell fate conversion process, this study introduces a novel computational method predicting master transcription factors based on group sparse optimization technique integrating data from multi-OMICs levels, which can be applicable to both single-cell and bulk OMICs data with a high tolerance of data sparsity. When it is compared with current prediction methods by cross-referencing published and validated master transcription factors, it possesses superior performance. In short, this method facilitates fast identification of key regulators, give raise to the possibility of higher successful conversion rate and in the hope of reducing experimental cost.

Key words: cell fate conversion; master transcription factor; group sparse optimization; integrative OMICs; gene regulatory network; single-cell genomics

Jing Qin is an associate professor at the School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University. Her research focuses mainly on integrative OMICs approaches for gene regulation studies.

Yaohua Hu is an associate professor at the College of Mathematics and Statistics, Shenzhen University. His research focuses mainly on theory and algorithms for large-scale optimization, and their various applications in statistics, machine learning and bioinformatics.

Jen-Chih Yao is a professor at the Research Center for Interneural Computing, China Medical University. His research focuses mainly on theory and applications of mathematical programming and operations research.

Ricky Wai Tak Leung is a post-doctoral fellow at the School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University. His research focuses on gene regulation studies.

Yongqiang Zhou is a postgraduate student at the School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University. His research focuses on gene regulatory network reconstruction.

Yiming Qin is a post-doctoral fellow at the Center for Genomic Sciences & School of Biomedical Sciences, the University of Hong Kong. Her research focuses on computational analysis of genomic big data.

Junwen Wang is a professor at the Department of Quantitative Health Sciences and Center for Individualized Medicine, Mayo Clinic. His research focuses mainly on genomics, development of bioinformatics tools and precision medicine.

Submitted: 19 May 2021; Received (in revised form): 6 July 2021

Introduction

Multicellular organisms are consisted of several cell types, where different types of cell express specific sets of proteins and RNAs that control respective cell morphology. The specification of a particular cell type typically involves the programmed expression of several thousand genes controlled by regulators, such as transcription factors (TFs). Recent technologies that enable cells converting from one type into another, called cell fate conversion, provide a promising strategy of regenerative medicine to produce functional cells that are lost in diseases such as Parkinson disease [1] and corneal epithelial stem cell deficiency [2]. There are a large number of cell fate conversion examples achieved just by the overexpression of a few key TFs. This suggests that cell identity defining expression profiles can be orchestrated by the combination of a small number of TFs, which are called master TFs [3]. For example, POU5F1 (OCT-4), SOX2, NANOG, KLF4 in embryonic stem cells [4, 5]; MYOD1, MYOG and MYF5 in muscle cells [6, 7]; GATA1, CEBPA and SFPI1 in blood cells [8], TBX21 and FOXP3 in T cells [9] and CDX2 in intestine cells [10]. These respective sets of master TFs function at the apex of transcriptional hierarchies, regulate downstream tissue-specific genes, and eventually determine the respective cell fates and identities. The identification of master TFs in different cell types is vital in both developmental biology and regenerative medicine, yet the traditional approach requires tedious experiments and laborious efforts [4, 11–13]. Despite great efforts made by researchers, many of the fate converted cells published were indeed not completely converted [14], for example, as these converted cells were unable to fully replicate the cell function or gene expression pattern of their target cells [15].

Computational methods to predict master TFs for cell fate conversion

Recent advancement of high-throughput sequencing technologies generates an enormous amount of OMICs (genomics, transcriptomics, epigenomics, proteomics and metabolomics) data in various cell types. Systematic analysis on these genome-wide OMICs data grants the possibility to predict master TFs during cell fate conversion. Shmulevich and colleagues ranked the expression difference between the two cell types to identify expression-reversed TF-TF pairs, based on the observation that a pair of mutually repressive TFs are often a toggle switch to control the cell fate of ‘sister’ lineages [16]. However, this method purely relies on the expression of TFs and does not take into account the regulatory relationship between them and cell-type specific genes, thus these factors may not achieve satisfactory cell fate conversion results. In fact, gene expression profile comparisons between engineered cells and their *in vivo* counterparts showed that most published cell fate conversions failed to properly silence or activate many of the key genes, despite the evident phenotype and marker expression in these engineered cells [14]. Therefore, not only the expression of TFs but also their downstream regulatory networks should be considered during the prediction of master TFs.

With the network-based strategy, two steps are usually required for the identification of master TFs: (i) construct high-quality transcriptional regulatory network and (ii) search potential master TFs from the constructed network. To construct transcriptional regulatory network, reverse engineering was commonly performed on both bulk transcriptome data [14, 17] and single-cell transcriptome data [18, 19], although it suffers from the drawback that direct and indirect TF-to-target

interactions are often indistinguishable. Networks are also constructed from tissue-specific binding of TFs on target genes’ regulatory regions, which reveal the regulatory direction but not the subsequent transcription effects on target genes [20]. Our previous studies have shown that only less than 20% of TF bound targets to have their expression level significantly altered under TF perturbation in mammalian cells [21, 22]. To conclude, integrating both transcriptome data and TF binding information is necessary to provide more accurate networks than those derived from either of them alone [23]. To improve the accuracy of master TF prediction during cell fate conversion, one should consider cell-type specific gene regulatory network incorporating both TF binding and expression dependency among TFs and targets.

Within a cell-type specific network, highly-connected hub TFs are usually considered as master TFs, since they regulate more genes [24]. Manipulating a few hub TFs may then be able to control most cell-type specific genes needed in cell fate conversion. In addition, TFs can also be classified into three levels by comparing their in- and out-degrees in these networks. TFs at the top level, which have higher out-degree than in-degree, are found to be important for cell identity [20, 25]. Thus, previous studies have applied these two methods on TF binding networks for master TF screening [20, 26]. However, it is difficult to distinguish cell-type specific master TFs from non-specific TFs with ubiquitous binding or regulations in many cell types. For instance, CTCF, a chromosomal architectural protein expressed in many tissues/cell types, is always in the pool of either hub TFs [26] or top TFs [20].

Therefore, even with a high-quality cell-type specific network considering both expression dependency and TF binding, detecting master TFs from hub TFs or top TFs may still be difficult. Firstly, it is because many hub/top TFs may be redundant as they share a similar downstream target group. To alleviate this problem, Mogrify, a network-based method, added an extra step to remove redundant TFs [27]. Secondly, this target group controlled by several hub/top TFs may not cover sufficient lineage-specific genes that are required for complete cell fate conversion. Consequently, most of the current cell fate conversions were found to be incomplete, in which the acquired engineered cells still possessed hundreds of differentially expressed genes (DEGs) when they were compared to their target *in vivo* counterparts [28]. In order to mitigate this shortcoming, CellNet provides a utility to find master TF so as to improve these incomplete cell fate conversions. But prediction performance of both methods was not satisfactory [14, 27, 28]. This is probably due to the low accuracy of the predefined networks used by these two methods. Mogrify implemented TF-target interactions from database STRING and Motif Activity Response Analysis (MARA), in which most of TF-target interactions are predicted and not cell specific. CellNet constructs specific networks for 20 cell types through correlation-based analyses on transcriptomes, whose quality was shown to be only slightly higher than random guessing [17]. Moreover, for those cell types lacking predefined cell-specific regulatory network information, these two methods may be more difficult to get satisfactory results.

Multi-OMICs integration improve prediction performance

In addition to transcriptome and TF binding information, recent discovery of super-enhancers gives rise to an opportunity to integrate epigenetic information and improve master TF

identification. Super-enhancers are large clusters of enhancers, whose activities can be identified by epigenetic markers or TF binding. It has been reported that super-enhancers are occupied by master TFs, which drives the expression of cell identity defining genes [29, 30]. Based on histone modification or mediator binding signals, super-enhancers could be detected in various tissues or cell types [31].

In the past decade, single-cell RNA sequencing (scRNA-seq) technology allows profiling transcriptomes of hundreds to thousands of single cells in one single experiment. This means that one scRNA-seq experiment can produce as many or even more transcriptome profiles than hundreds of the previous bulk sequencing experiments. scRNA-seq is capable of capturing cell heterogeneity and stochastic gene expression, which records more expression variations in a cell population. Given that a sufficient number of transcriptome profiles and significant expression variations are critical to network construction in all reverse engineering methods, single-cell transcriptomes would be a suitable data source for such purpose, as well as inferring master TF in cell fate conversion.

As shown in above, different types of OMICs data could be utilized to provide useful information in master TFs identification. However, efficient computational methods are lacking to integrate these data for accurate predictions. A desirable method should consider the quality of cell-specific transcriptional regulatory network, while providing an efficient way to find master TFs from the network, and make sure that all genes needed to be converted are controlled by the predicted master TFs directly or indirectly.

Mathematical modeling for integrating Multi-OMICs data

Sparse optimization, also called LASSO (least absolute shrinkage and selection operator) [32] in statistics or basis pursuit [33] in compressed sensing, is one of the most popular and practical computational methods for data analysis and machine learning. Particularly, by virtue of the sparse structure ubiquitously arising from practical applications, the principle of sparse optimization technique aims to find a few major factors in data fitting mechanism. In mathematics, the sparsity of variables is usually measured by the original L_0 penalty (i.e. the number of nonzeros of the variable) or the relaxed L_1 penalty (see section Methods and Discussion for the theoretical comparison); the resulting sparse optimization techniques are called L_0 regularized and L_1 regularized sparse optimization methods, respectively. The sparse optimization technique has also been applied in bioinformatics, such as gene regulatory networks inference [23].

Besides the sparse structure, group structure is another common structure in various applications, in which the solution has a natural grouping of its components, and the components within each group are likely to be either zero or nonzero simultaneously. In mathematics, the group sparsity of variables can be measured by the $L_{2,0}$ penalty (i.e. the number of nonzero groups) [34] or the $L_{2,1}$ penalty [35]. By employing the group sparse penalty, group sparse optimization technique can promote the group structure and reduce the degrees of freedom of the solution, thereby leading to better recovery or prediction performance in various disciplines [35–37]. Group sparse optimization technique has also been applied to characterize the structure of TF complex and improve the accuracy of GRN inference [34].

In a cell fate conversion, all genes needed to be converted are simultaneously changed by only a few master TFs from one

cell type (donor cell) into another cell type (target cell), which meets the essentials of group and sparse structures. Motivated by the sparse structure of master TFs and the group structure of the transcriptional regulatory network controlled by master TFs, this study developed a novel bioinformatics method based on the group sparse optimization technique to address the problem of inefficient and incomplete cell fate conversion. The methods using the $L_{2,0}$ and $L_{2,1}$ penalties are called group sparse optimization (GSO) and group LASSO (gLASSO), respectively. In details, it treats the regulatory coefficients of each TF to all genes as a group, formulates the gene regulatory network during cell fate conversion into GSO model and predicts master TFs by integrating bulk/single-cell transcriptome, TF binding and super-enhancer information. To evaluate the performance of different master TF prediction methods, a scoring system based on published, wet-lab validated master TFs, which we call standard TFs was utilized. When this method was compared with other methods utilizing different combinations of OMICs datasets by cross-referencing their results with existing standard TFs, it demonstrated superior performance. This superiority was achieved in several aspects: (i) high quality of cell-specific regulatory information is extracted by integrating multiple cell-specific OMICs data; (ii) network construction and master TF prediction are processed simultaneously; (iii) all genes, whose expression are altered in cell fate conversion, are considered as a whole target gene group; this makes sure all essential genes were covered as targets of the inferred master TFs; (iv) regulatory effect of master TFs on their targets are quantified and (v) genomic regions of super-enhancers can be incorporated to narrow down the candidates of master TFs. Furthermore, this method was found to be applicable to both single-cell and bulk OMICs data and achieves similar performance. It does not require preliminary gene regulatory information of the target cells. This method can help researchers to find master TFs when new cell fate conversions are needed.

Results

To assess the efficiency of different approaches in master TF inference, we took the most well-known cell fate conversion case - induced pluripotent stem cell (iPSC) as an example. Takahasi and Yamanaka firstly reported a successful cell reprogramming from mouse fibroblasts into embryonic stem cell (mESC)-like iPSCs through forced overexpression of four TFs, Pou4f1 (Oct-4), Sox2, Klf4 and Myc (OSKM) [4]. Other TF combinations, such as Sall4, Nanog, Esrrb and Lin28 (SNEL), can also reprogram mouse fibroblasts into iPSCs [5, 38, 39]. Besides, overexpression of some chromatin modifiers, for instance, Tet1 [40], Prdm14 and Jarid2 [41] were found to play important roles in cell reprogramming. We searched all TFs (including chromatin modifiers) that have been used to induce iPSC, and listed the number of publications that used such TF for iPSC induction in Figure 2. We will call them standard iPSC factors hereafter. Even though other untested TFs might also have the potential for cell programming, this iPSC factor list is able to act as a referencing standard to evaluate different master TF prediction methods.

We have collected predicted TF binding sites (TFBSs), TFBSs identified by chromatin immunoprecipitation coupled with sequencing or microarray (ChIP-seq/chip), single-cell and bulk transcriptome data, super-enhancer regions and high-quality mESC networks for iPSC factor prediction (data were all derived from mESCs and can be downloaded from <https://qinlab.sysu.edu.cn/GSO>). To assess their contribution to iPSC factor prediction, we used different data combinations and

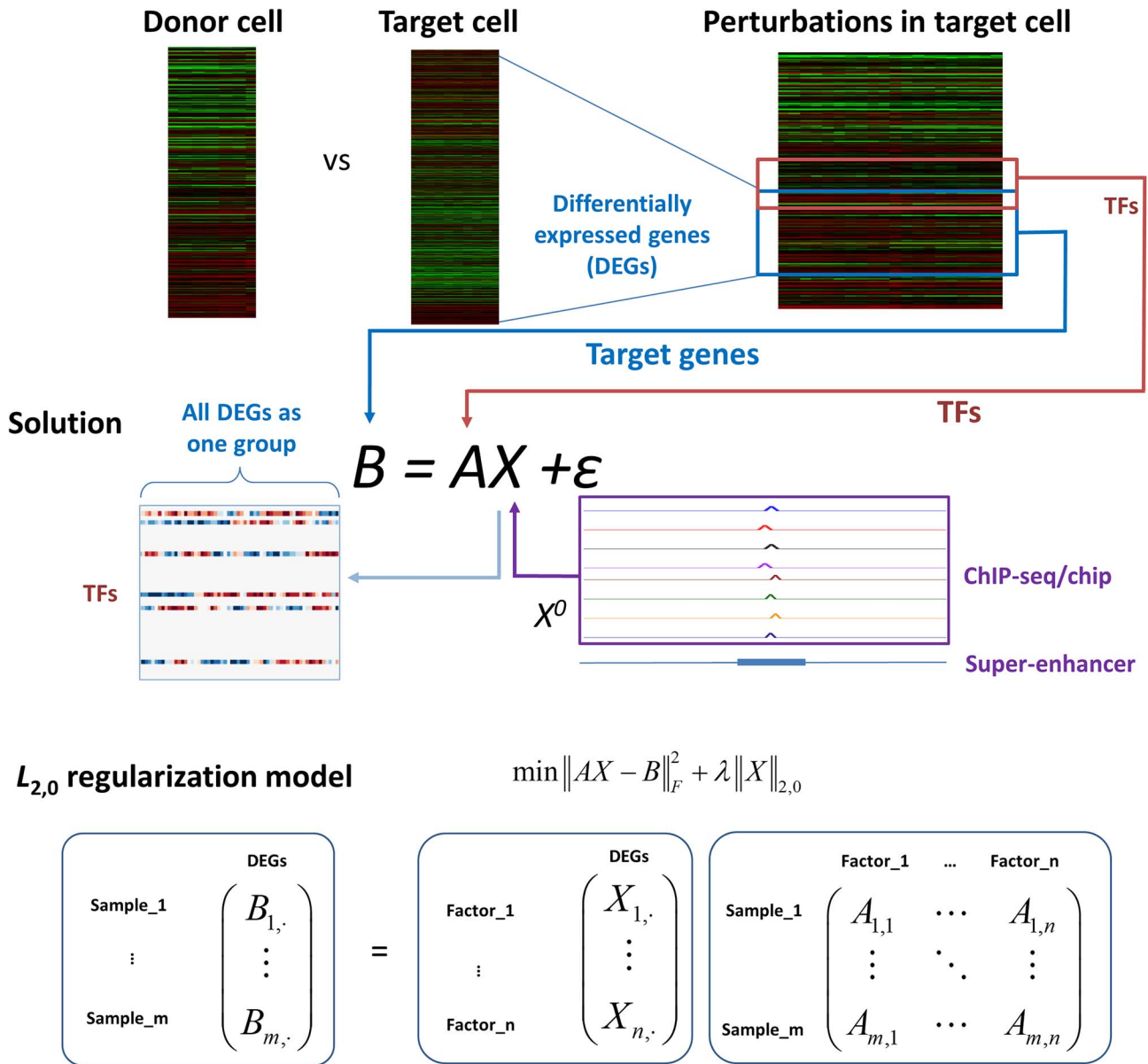


Figure 1. Workflow of the master TF inference with GSO. DEGs between donor cells and target cells are identified by comparing expression profiles of two cell types. Cell-type-specific gene regulatory networks are hidden in the transcriptomes derived from perturbations on the target/donor cells. From these transcriptomes (usually >50 samples/cells), expression profiles of TFs form matrix A , while expression profiles of DEGs form matrix B . The linear model $AX = B + \epsilon$ approximately describes the expression dependency between TFs and DEGs, in which X represents the regulatory strength while ϵ is the matrix of noise. TF binding and super-enhancer information are transformed into X^0 as an initial guess for the solution searching. Master TF inference is an optimization problem to find an X to minimize the difference between AX and B with only a small number of selected TFs, which its regulatory strength on DEGs ($X_{i,\cdot}$) are non-zeros. Red color in the solution matrix X means positive regulation, blue color means negative regulation and white means no regulation. TFs with colors (non-zeros in the solution X) are the predicted master TFs that show regulatory effects on DEGs that need to be changed from donor cells into target cells. The middle panel shows the $L_{2,0}$ regularization model (a GSO model), while the lower panel lists the structures of matrix A , B and X .

computational methods to predict and rank the TFs. Figure 2 compares the results of different approaches utilizing diverse data and methods, including Mogrify [27], CellNet [14] and our GSO method (Figure 1).

To quantify the performance of each method and data combination, a total score of each method summarizes the rankings of all iPSC factors; on the other hand, a total weighted score further considers the usage of an iPSC factor in cell reprogramming (see section Methods). Consistent with the previous comparison between Mogrify and CellNet, Mogrify obtained a

higher total score and total weighted score than CellNet [27] (Figure 2, Supplementary Table S2A and S2B). Top 20 Mogrify TFs contain two Yamanak factors (OSKM), two SNEL factors and two other iPSC factors, but only three iPSC factors were reported by CellNet (Figure 2). Our method, GSO with the integration of multiple OMICs data, ranked top 20 TFs out of 939 candidates and achieved the highest total scores and total weighted scores. Its performance on single-cell data was as good as that on bulk-cell data. In the following sections, we analyzed and compared the contributions of various OMICs data to prediction accuracy, as

TF	Pubmed	Mogrify	CellNet	pTFBS_SuperEnh	ChIP_SuperEnh	HubTF	TopTF	gLASSO	GSO	GSO_SuperEnh	GSO_sc	GSO_sc_SuperEnh	GSO_sc_pCRISPR
Pou5f1	306	3	2	13	5	6		3	2	7	8	1	
Sox2	285	2		14	14	15		7	3	10	7	9	
Klf4	229			15	15	14		10	8	9	11	6	
Myc	226			16	17	16		9	15	16	17	13	
Nanog	129	1	3	19	8	8		8	6	2	1	3	
Esrrb	2				2	2		2	5	1	2	5	
Sall4	2	16	14					17	10	11	4	17	
Lin28	2												
Tet1	9												
Prdm14	4	17											
Jarid2	2												
Zic3	1	8											
Glis	1												
Total Score		79	44	0	28	65	65	0	91	98	91	97	93
Total Weighted Score		449	275	0	210	369	365	0	534	571	483	495	559

Figure 2. Predictions of master TFs for mouse fibroblast to mESC conversion by various methods assessed by standard iPSC factors. Standard iPSC factors are listed in the first column, which includes Pou4f1 (Oct-4), Sox2, Klf4 and Myc (OSKM, green) [4], Sall4, Nanog, Esrrb and Lin28 (SNEL, red) [5, 38, 39], and other TFs and chromatin modifiers (yellow), Tet1 [40], Prdm14 [41], Jarid2 [41], Zic3 [39] and Glis [38]. The numbers of publications that used the iPSC factors for iPSC induction are listed in the second column. To quantify the performance of each method, all rankings of iPSC factors were summarized as the total score, while the total weighted score further takes the prevalence of usage of an iPSC factor for cell reprogramming into consideration. The 3rd to 14th columns compares the prediction results of 12 methods. The numbers in red grids show the ranking of iPSC factors derived from each method. White grids mean the iPSC factors are not predicted as the top 20 TFs. The abbreviations of each method listed in the first row are described as follows. pTFBS_SuperEnh: enrichment analysis of predicted TFBSs in super-enhancers; ChIP_SuperEnh: enrichment analysis of TFBSs identified by ChIP-seq/chip in super-enhancers; HubTF: hub TFs in L_0 network of mESC; TopTF: top TFs in L_0 network of mESC; gLASSO: group LASSO integrating transcriptomes and TF binding information; GSO: GSO via $L_{2,0}$ regularization model integrating bulk transcriptomes and TF binding information; GSO_SuperEnh: GSO model integrating bulk transcriptomes, TF binding and super-enhancer information; GSO_sc: GSO model integrating single-cell transcriptomes of cell reprogramming process and TF binding information; GSO_sc_SuperEnh: GSO model integrating single-cell transcriptomes of cell reprogramming process, TF binding and super-enhancer information; GSO_sc_pCRISPR: GSO single-cell model integrating single-cell transcriptomes after pooled CRISPR and TF binding information.

well as the advantages and disadvantages of different prediction methods.

Enriched TFs in super-enhancers

First, we tested whether TFs with enriched binding sites in super-enhancers were more likely to be master TFs, since super-enhancers are frequently occupied by master TFs regulating lineage-specific genes [30]. Within the regions of mESC super-enhancers, putative TFBSs were predicted. However, this binding site enrichment analysis had poor predictive powers on the bounded master TFs (Figure 2 column pTFBS_SuperEnh, Supplementary Table S2C). It may be because of the large genomic regions of super-enhancers, as they are with average lengths of dozens of thousands of base pairs and the predicted TFBSs are usually of high false positives.

Then, we performed enrichment analysis on ChIP-seq/chip binding sites within super-enhancers (see section Methods). It gives a better prediction than the predicted TFBSs (Figure 2 column ChIP_SuperEnh). All of four Yamanaka factors ranked in the top 20. Besides Yamanaka factors, one of SNEL factors was also found to be enriched in super-enhancers. It indicates that super-enhancers were indeed enriched with iPSC factors binding when compared to random sequences. To summarize, super-enhancer data could be useful for iPSC factor prediction, provided that the quality of TFBS information is high.

When only super-enhancers and TFBS information were used, all reported iPSC factors were ranked beyond the top 10 list (Figure 2 column ChIP_SuperEnh, Supplementary Table

S2D). Both the total score and total weighted score were lower when compared to those of network-based methods, e.g. Mogrify and CellNet. This is because TFBS information alone without considering the target expression changes usually leads to an overestimation of the regulatory effect of TFs to the genes they bound to. Besides, limited to the availability of ChIP-seq/chip data, some iPSC factors were not being able to be analyzed (Supplementary Table S1).

Hub TFs and top TFs in mESC network

Hub/top TFs of a transcriptional regulatory network are commonly regarded as master TFs [24, 25]. Thus, identification of hub/top TFs from constructed mESC network could be an alternative approach for iPSC factor prediction. We tested these two traditional methods for master TF prediction from a high-quality transcription regulatory network of mESC inferred through L_0 regularized sparse optimization method in our previous study [23]. We call it L_0 network hereafter. Hub TFs and top TFs of this network were ranked (Figure 2 column HubTF and TopTF). These two methods showed similar predictive power over iPSC factors. All four Yamanaka factors and two of the SNEL factors were predicted as top 20 master TFs by both methods. The total scores and total weighted scores were improved when compared to enrichment analyses of TFBSs (Figure 2 column ChIP_SuperEnh and pTFBS_SuperEnh), because L_0 network has high quality under the consideration on the regulatory consequence of TF binding by integrating transcriptomic data [23].

To evaluate the influence of network quality to iPSC factor prediction, we compared the hub/top TFs of L_0 network and those of two other mESC networks inferred using $L_{1/2}$ and L_1 regularized sparse optimization methods in our previous study [23], named $L_{1/2}$ network and L_1 network, respectively. Previous validation using two independent sets of known regulatory connections has shown that the quality of L_0 network is higher than that of $L_{1/2}$ and L_1 network [23]. Comparing the hub/top TFs of these three networks, a similar trend was observed on both total scores and total weighted scores, both of which summarizes the rankings of all iPSC factors, and the later further considers the popularity of an iPSC factor using in cell reprogramming (Supplementary Table S2E and S3F). Hub/top TFs from L_0 network achieved a higher score than those from $L_{1/2}$ network and L_1 network. It indicates that the quality of cell-specific networks is crucial for master TF inference.

However, hub/top TFs may not be good enough for master TF inference. Even though the prediction utilized a high-quality network considering both target expression dependency and TF binding, the results of hub/top TFs were not satisfactory. As some irrelevant TFs, such as CTCF, may also be identified as hub TFs or top TFs [20, 26]. Because cell fate conversion is a process changing the expression status of a set of genes (DEGs) from one cell type (donor cell) into that of another cell type (target cell), the gene expressions of the donor cell should be taken into consideration too. Besides, many hub/top TFs may only control a similar target group that even when they are considered together, it would still be possible not covering all the genes needed in cell fate conversion. Therefore, many genes in current cell fate conversions were not converted properly into the right expression states as those in the target cell type [14, 28].

Group sparse optimization for master TF inference from bulk and single-cell data

To improve master TF inference, we designed a new method directly predicting master TFs and quantifying their regulatory effects on downstream targets by using transcriptomes and TF binding information (Figure 1). This method takes both the donor cell and target cell expression profiles into consideration. Firstly, it identifies DEGs between fibroblast (donor cell) and mESC (target cell). In the cell fate conversion, all DEGs are regarded as a group because most of them are regulated by master TFs simultaneously. GSO was applied to find a small number of master TFs out of 939 candidates that can control the expression of all DEGs. It calculates the regulatory effect of these TFs by considering the expression changes of the whole group of DEGs between the donor cell and the target cell based on the expression profiles of 939 candidate TFs (matrix A in Figure 1) and DEGs (matrix B in Figure 1) in a number of transcriptome samples. To obtain gene regulatory network specifically activated in target cells, these transcriptome data of TFs and DEGs would be better if they were derived from perturbations in the target cells. Moreover, group sparse penalty in the GSO model promotes the number of selected TFs below a defined sparsity level, for example, four out of all TFs. TF binding information was also integrated to provide an initial guess (matrix X^0) for solution searching (see section Methods).

First, we tested our method using bulk transcriptomes. Bulk transcriptome data from 245 perturbation experiments on mESCs were curated, which were conducted by the same research group [42–44]. Matrices A , B and X^0 were generated with these bulk transcriptome data (see section Methods). Our

method successfully predicted all Yamanaka factors and three of SNEL factors (Figure 2 column GSO) as part of the top 20 master TFs when the mentioned bulk transcriptome data and TF binding information from ChIP-seq/chip were integrated (see section Methods), while most of the Yamanaka factors were ranked in the top 10. The Yamanaka and SNEL factors predicted in our method also achieved a better ranking when compared to these factors predicted in hub/top TFs (Figure 2 column HubTF and TopTF). On top of that, the total score and total weighted score were higher than those from the other two network-based methods, Mogrify and CellNet. gLASSO failed to predict any of the 13 standard iPSC factors (Figure 2 column gLASSO), although it used the same data combination as that of GSO.

We also tested the performance of this method on single-cell data. From Guo et al. [45], 912 single-cell transcriptomes of fibroblasts (74 cells), mESCs (82 cells), iPSCs (65 cells) and reprogramming processes (691 cells) were collected [45]. Then, matrices A , B and X^0 were generated with normalized read counts from these 912 cells (see section Methods). GSO achieved good results similar to that of using bulk transcriptome data (Figure 2 column GSO_sc). However, results from bulk transcriptomes preferred Yamanaka factor OSKM, while those from single-cell data preferred iPSC factor SNEL. This is probably due to the fact that these single-cell sequencing data were generated from overexpressing of Yamanaka factors, and these overexpressed RNAs lacks poly-A tails, they would not be quantified by scRNA-seq. The underestimation of Yamanaka factors expression in the infected cells by scRNA-seq might mislead the relationship between these TF and their target genes, resulting the underestimated ranking of Yamanaka factors. Thus, the total weighted score considering the prevalence of usage of SNEL TFs was much lower, since OSKM was still the most popular TF set for cell reprogramming from fibroblast to iPSC.

Single-cell transcriptomes of reprogramming processes documented the necessary gene regulatory network in cell fate conversion. Yet, for most cell types, such data is not available. Moreover, due to the low reprogramming efficiency, only a small fraction of the cell population obtained the target-cell-like gene expression profile, while a majority of cells went through a deviated path and differentiate into non-desirable cell types [45]. The rest cells directing into other cell fate might interfere with the network construction. Fortunately, perturbation experiments on target cells, especially those which knock-down/out or overexpress TFs, would reveal the importance of such TFs in the target cell's gene regulatory networks. Therefore, data with pooled CRISPR (clustered regularly interspaced short palindromic repeat) targeting TFs coupled with scRNA-seq (such as CROP-seq [46] and Perturb-seq [47]) in target cells allows us to obtain enough transcriptome data for master TF prediction, while this pipeline can also be applied to any cell type. By using such CROP-seq data in mESC collected from Yang et al. [48], master TFs were predicted and compared with those predicted from previously mentioned single-cell and bulk transcriptome data (Figure 2 column GSO_sc_pCRISPR). The master TFs predicted from CROP-seq data were found to be different from those predicted from scRNA-seq data of reprogramming process, while they are more similar to those predicted from bulk transcriptomes, as these bulk transcriptomes were also derived from gene perturbation experiments on mESCs (Figure 2).

To test the reproducibility of our method and observe whether the same result can be obtained by inputting data with similar origins, predictions were run on different single-cell transcriptome matrices of mESC generated from the same treatments. Prediction results from two single-cell

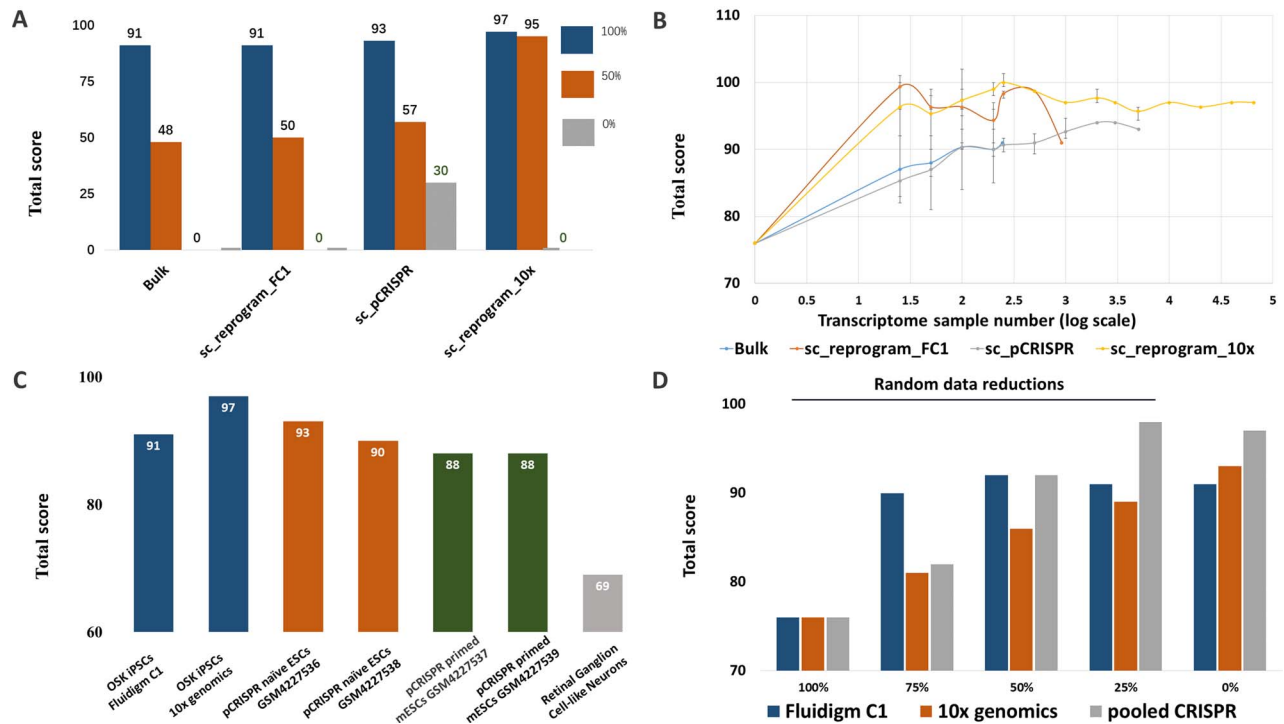


Figure 3. Performance of master TF prediction by GSO under different data input conditions assessed by total score. (A) Different quantity of ChIP-seq/chip data (0, 50 and 100%) was randomly selected as data reduction and incorporated as the initial guess X^0 combined with bulk transcriptomes or single-cell transcriptome for GSO method. (B) Different amount of transcriptome sample reduction was performed for each dataset. The abbreviations of each method listed in the first row are described as follows. Bulk: bulk RNA sequencing; sc_reprogram_FC1: single-cell sequencing data from Fluidigm C1 platform generated from reprogramming processes; sc_pCRISPR: GSO single-cell pooled CRISPR; sc_reprogram_10x: single-cell sequencing data from 10x genomics platform generated from reprogramming processes. (C) Performance differences between different datasets for both single-cell and bulk cell transcriptomes as well as those from different cell types to assess the reproducibility and the cell type specificity of GSO. (D) Performance comparison between different data sparsity level generated by random data reduction at 0, 25, 50, 75 and 100%.

transcriptome matrices were compared, which contain 5073 and 3855 cells respectively derived from repeated pooled CRISPR experiments on naïve mESC (GSM4227536 and GSM4227538 from Yang et al. [48]). Another two datasets, which were both derived from OSK-mediated reprogramming process (GSE103221 from Guo et al. [45]) but detected by two different scRNA-seq platforms- Fluidigm C1 and 10x Genomics respectively, were also compared. Their total scores and TF rankings were similar, but a higher single-cell sample number has been observed to achieve a better performance (Figure 3C).

Contributions of various OMICs data types to the GSO prediction

ChIP-seq/chip and Initial X

To assess the contributions of different OMICs data integrated by the GSO model, we reran the predictions by reducing the information or replacing with unrelated data for each data type. First, initial guess matrix X^0 derived from TF binding information was tested by replacing it with each of the following matrices: (1) $X_{zero} := 0$, in which each element is zero, representing that no TF binding information was utilized; (2) $X_{50\%}$, representing that randomly selected 50% of TF binding information (non-zero elements) was utilized (see section Methods). When reduced TF binding information from ChIP-seq/chip was integrated in GSO, the total scores were found to drop notably. When TF binding data was fully unutilized, randomly reduced to 50% and fully reduced, the total scores dropped from 91 to 48 and 0, respectively, for the case of bulk transcription data; from 91 to 50 to 0,

respectively, for single-cell transcriptome of cell reprogramming process derived from Fluidigm C1 platform; from 97 to 95 to 0, respectively, for single-cell transcriptome of cell reprogramming process derived from 10x Genomics platform; from 93 to 57 to 30 for single-cell transcriptome obtained from pooled CRISPR (Figure 3A). Second, we also tested other initial X s without TF binding information, but their prediction results were all poor (Supplementary Table S10).

Transcriptome sparsity level

It is well-known that single-cell transcriptome matrix is highly sparse due to missing data. The sparsity of single-cell transcriptome matrices used in this study is all above 0.99, where more than 99% of elements in the matrices are zeroes. To test how sparsity level affects the prediction result, sparsity levels of different single-cell datasets were artificially increased by randomly reducing non-zero data quantity to 75, 50 and 25% of the original single-cell transcriptome matrices. Surprisingly, the artificial increasing of sparsity level only reduced the total score slightly even when the data was reduced to 25% (Figure 3D). Furthermore, to test whether decreasing sparsity level by imputation methods inferring the missing gene expression values can improve the prediction performance, we compared the results from single-cell transcriptome matrix of normalized raw reads and those from imputed matrices. Three imputation methods, Knn-smooth2, DrImpute [49] and SAVER [50], were compared (Figure 4). Yet none of the missing data imputations improved the performance of master TF prediction. Instead, Knn-smooth and DrImpute had even slightly reduced the prediction total

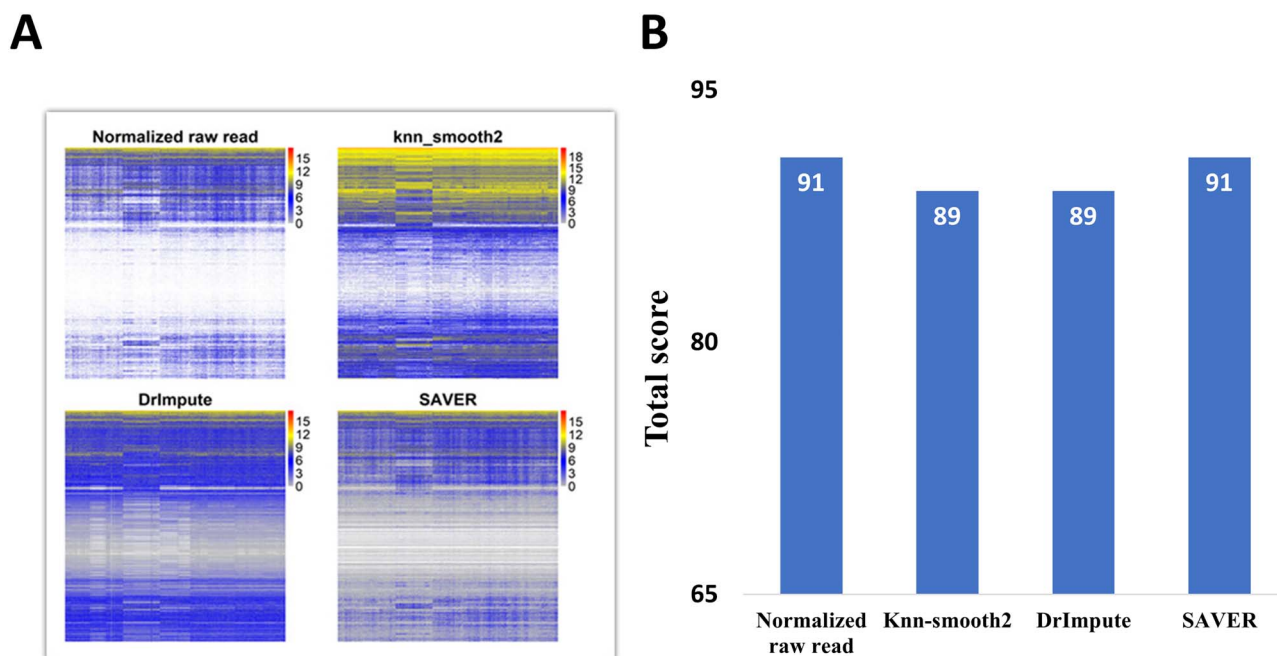


Figure 4. Effects of data imputation on master TF inference. (E) Single-cell expression profiles from platform Fluidigm C1 (912 cells) were imputed with three imputation methods, Knn-smooth2, DrImpute and SAVER. The single-cell expression matrix heatmaps and total scores of master TF inference from different imputation methods were compared.

score and total weighted score, while SAVER did not affect the master TF prediction results (Supplementary Table S4A-F). It may be because these methods relied on existing information to infer missing data, and they did not provide additional information but introduced false signals, which led to spurious gene network interactions [51–53].

Cell type specificity of transcriptomes

As mentioned above, the cell-type-specific network information constructed from transcriptomes is important for master TF inference. Therefore, when we replaced the single-cell transcriptome matrix with those from other cell types while remain using mESC derived TFBS information, the resulting total scores decreased as expected. When single-cell transcriptomes (3425 and 2795 cells in two repeated experiments, GSM4227537 and GSM4227539) from primed mESC (a cell type very similar to naïve mESC but in a primed state) were used, the total score slightly reduced to 88 for both primed mESC datasets. When the data of reprogramming processes from fibroblast to retinal ganglion cell-like neuron (a cell type very different from naïve mESC) was used, the total score has been decreased to 69, which is even lower when we ranked TFs based on TF binding information alone (total score 76) (Figure 3C). This demonstrated that gene regulatory network is highly cell-type specific. The combination of cell-type-specific OMICs data is important to the construction of cell-type-specific network, as well as the prediction of master TFs for a specific cell type. When data from different cell types were mixed like the above example, connections in the network would be disrupted, resulting in poor performance.

Transcriptome sample number

To assess how the change of transcriptome sample number might affect the result of GSO, we artificially reduced the sample number of different datasets (numbers of samples were

randomly selected from each of the four sets of bulk/single-cell transcriptomes), and then compared the prediction results between different number of samples. Surprisingly, all types of dataset achieved satisfactory result (total score around 90) when the sample number reached 100 only, further increase of sample number up to 65,068 did not result in a notable raise in total score (Figure 3B).

Super-enhancer region enriches the master TFs

Since super-enhancer regions were enriched with master TF binding, we further tested whether this information could improve the prediction. GSO method allows the incorporation of super-enhancer information (see section Methods). Results demonstrated that the incorporation of super-enhancer information further raised the ranking of iPSC factors, thus both total score and total weighted score were increased (Figure 2 column GSO_SuperEnh, and Supplementary Table S2I and S2K).

Running time

The running time for one prediction by GSO is quite fast, which is about 20 minutes by using Matlab R2020a on a personal computer with i7–9700 CPU and 24 GB RAM. The number of cells hardly affected the running time required. With the increase of cell number from 25 to ~60 000 cells, the running time only increased slightly from 20 to 22 minutes (Figure 5). This fast running time can be explained by the state-of-the-art iterative thresholding algorithm (ITA) applied to solve the GSO problem [34], that consists of a gradient descent step of linear least-squares and a group hard-thresholding step of the $L_{2,0}$ penalty alternatively. It was shown in [34] that the ITA has a fast linear convergence rate and very low computational complexity (as both steps in ITA have analytical formulas). Moreover, the major part of computational cost of each gradient descent step is $(A^T A)X$, and the one of the group hard-thresholding step is $H(\cdot)$

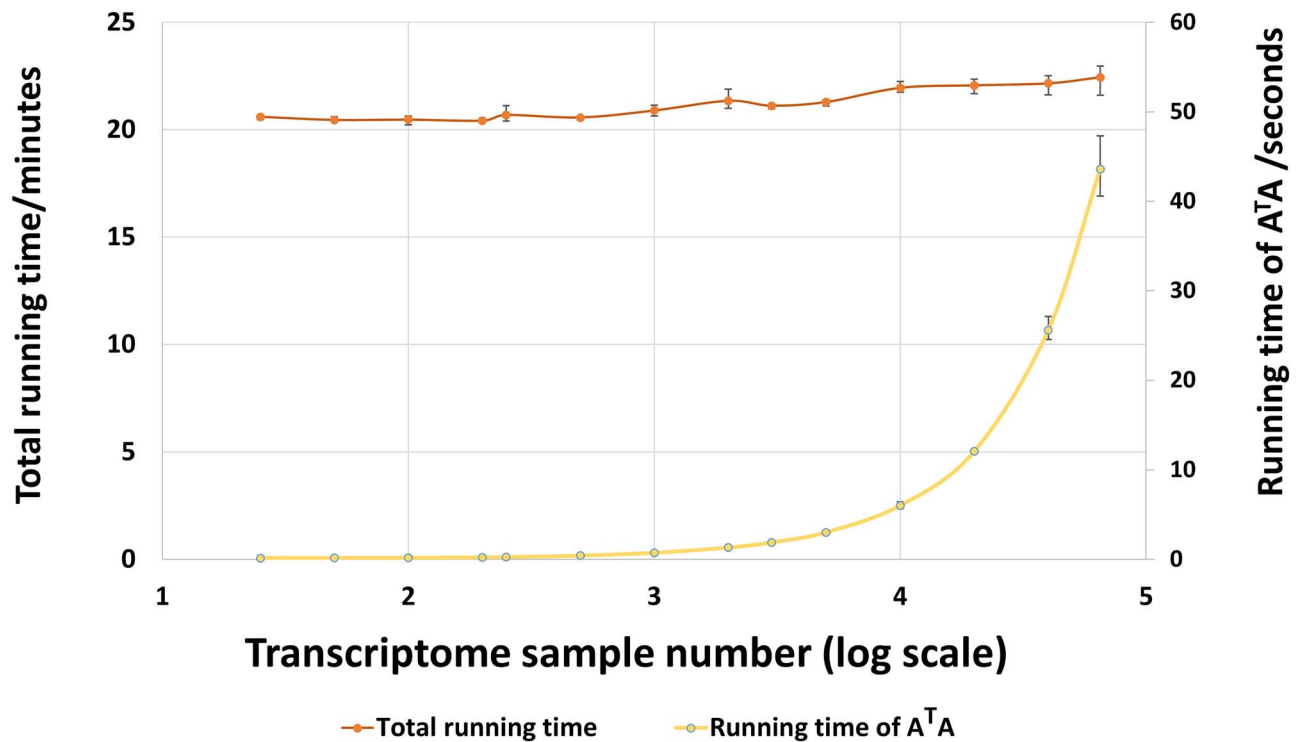


Figure 5. Running time of GSO method with various sample/cell number. The experiment was repeated three times. With the increase of cell number from 25 to ~60,000 cells, the total running time only increased slightly. $A^T A$ is the major part relevant to sample number. The running time it costed increased significantly as the number of samples increased. But it costed just several seconds, so the total running time did not increase much as the number of samples increased.

(see section IHTA in Methods), both of which depends on the dimension of variables (TFs) but rarely on the number of samples (cells). The major part relevant to sample number is $A^T A$. The running time that it costed increased significantly as the number of samples increased (Figure 5). However, it was calculated only once in the whole algorithm and costed just several seconds. Hence, the total running time did not increase much as the number of samples increased.

Discussion

This study evaluated the efficiency and proficiency of different data combinations and methods in predicting master TFs during cell fate conversion. We introduced a new bioinformatics method based on GSO technique that is able to integrate multiple OMICs data and has superior performances when compared to several state-of-the-art prediction methods, including CellNet and Mogrify, by a semi-quantitative scoring system in the case of conversion from mouse fibroblast to mESC. As this conversion is the most well studied case so far and possess abundant TFBS, expression and epigenetic information that can be used as input data, while there are also numerous studies discovering wet-lab proven master TFs for this conversion. Therefore, the performance scoring system takes advantage of these published and wet-lab validated master TFs and consider them as the standard master TFs in mouse fibroblast to mESC conversion. To assess and compare the efficiencies of our GSO method and other methods, we developed a semi-quantitative scoring system naming total score and total weighted score (see section Methods). Prediction methods achieved a higher score when more of these standard TFs were predicted in their top 20 master TF list, those TFs ranked higher, and a higher

number of publications used the respected TFs for the cell fate conversion.

In brief, methods utilizing TFBS and super-enhancer information alone made poor results: the one using TFBS information based on predictions (Figure 2 column pTFBS_SuperEnh) even got 0 total score and total weighted score while the one using TFBS information derived from ChIP/ChIP-seq (Figure 2 column ChIP_SuperEnh) had the third least total score and total weighted score. In comparison, methods utilizing gene network information and traditional master TF identification methods made better results (Figure 2 column HubTF and TopTF). Method integrating gene network information and better master TF identification method achieve even more satisfactory results (Figure 2 column Mogrify). Surprisingly, the method CellNet, despite of its integration of gene network information and better master TF identification method, it resulted with poor performance comparative to using ChIP/ChIP-seq and super-enhancer data alone (Figure 2 column CellNet). Most importantly, our GSO methods, regardless of whether super-enhancer information was integrated or scRNA-seq data was used instead, managed to result notably high total and total weighted scores when compared to any of the above methods. Furthermore, after testing and running our method with different data replacement, there are several key points: (i) the integration of ChIP-seq/chip data in form of initial X value is important; (ii) different data to be integrated and utilized in our method are suggested to be obtained from the same cell type; (iii) this method has a high tolerance on single-cell transcriptome data sparsity; (iv) our result is reproducible when the data were substituted by different datasets from the same cell type.

From these results, it was also found that methods integrating data with more OMICs levels properly will generally result in better performance. Therefore, methods using one

level of data type alone resulted relatively poor performance. While methods integrating multiple OMICs levels logically improves the conversion performance. In the process of master TF inference in cell fate conversion, both network construction and master TF prediction are important steps to achieve satisfactory results. For network construction, CellNet implements correlation-based method to infer cell-specific network from transcriptomes of perturbation experiments in a certain cell type. It incorporates TFBS information identified by ChIP-seq to select a proper cutoff of TF-gene expression correlation. However, correlation-based methods have been shown to have low efficiency for gene regulatory network prediction regardless of which cutoff is selected [17, 23]. That might be the reason that CellNet has surprisingly unsatisfactory performance during the prediction of TFs required for reprogramming fibroblasts to iPSCs. Mogrify adopted an improved method for network construction. It integrates protein-protein interactions (PPIs) from STRING and transcriptional activity of predicted TFBSs from MARA by FANTOM consortium. However, PPI networks and predicted TFBSs are not cell-specific and have high false-positive rates. Both CellNet and Mogrify rely on pre-existing networks, new cell types whose gene regulatory networks have not been analyzed might lack necessary information for further master TF identification. Therefore, building a high-quality cell-specific network for target cell of the cell reprogramming is crucial for more accurate master TF prediction. This was also supported by our test, in which hub/top TFs from mESC network of high quality achieved a higher score than those from those of relatively lower quality (see section Results). While during the second step - master TF inference, when master TFs were searched from the built network, TFs are usually weighted by their targets in the network. Hub/top TFs, CellNet and Mogrify all use similar strategies in this step. However, these strategies did not quantitatively assess the expression dependency between TFs and their targets. To summarize, a refined master TF prediction method should improve the above drawbacks as follows: (i) a better integration of data from multi-OMICs levels; (ii) improving the inference of TF-gene regulatory relationships; (iii) data utilized for gene regulatory network constructions should be cell-specific; (iv) expression dependency between TFs and their targets should be assessed. A new method addressing the above areas was hence developed in this study.

Our method utilizes a group sparse optimization model: the $L_{2,0}$ regularization model, in which an $L_{2,0}$ penalty is used to characterize the simultaneous regulations of master TFs to all DEGs (group structure) and to select a few master TFs (sparse structure). $L_{2,0}$ regularization model is a combination of the regression-based method and the group sparsity penalty method, which minimizes the regression residual and the group sparsity of variables simultaneously. Regression-based methods have been shown to enjoy better performance than correlation-based methods [17]; the L_0 -type penalty is an exact measure of the sparse structure and thus has a significantly stronger sparsity promoting capability when compared to the relaxed L_1 -type penalty [34, 54]. Consequently, the L_0 regularization model showed better performance when constructing cell-specific network from transcriptomes of perturbation experiments together with TF binding information in our previous study [23]. Motivated by the benefits of regression-based method and L_0 -type penalty, this study utilized the GSO method (L_0 -type penalty) to infer master TFs for cell fate conversion, which merges network construction and master TF prediction into one step and showed better performance than gLASSO (L_1 -type penalty).

The comparisons of mathematical theory between GSO ($L_{2,0}$ regularization model) and gLASSO ($L_{2,1}$ regularization model) were summarized in two perspectives of consistency theory of models and numerical theory of algorithms:

(i) Consistency theory of models. Both GSO and gLASSO are mathematical models for real-life problems, and the global solutions of both models are not the exact solution of real-life problems. By definition, the $L_{2,0}$ norm is an exact measure of the group sparsity, while the $L_{2,1}$ norm is a relaxation of the $L_{2,0}$ norm; hence GSO (Eq. (5) in section Methods) provides a more accurate representation of group sparse structure than gLASSO. Moreover, it was reported in [34, 54] that, contrasted with gLASSO, GSO allows a weaker condition (RIP or REC) on matrix A (Figure 1) to guarantee the perfect recovery and consistency theory. In a word, GSO is a more accurate and stable model for group sparse optimization problems than gLASSO.

(ii) Numerical theory of algorithms. Suffering from the non-convexity of $L_{2,0}$ norm, it is intractable to design algorithms to find a global solution of GSO [55]; fortunately, several fast iterative algorithms were designed to approach a local solution of GSO [34, 56]. Alternatively, gLASSO inherits the benefit of convexity of $L_{2,1}$ norm in designing fast algorithms to find a global solution of the relaxed problem. In a word, fast algorithms are applicable for approaching the local and global solutions of GSO and gLASSO so as to approximate the solution of real-life problems, respectively. Moreover, initial point is sensitive for GSO algorithms but insensitive for gLASSO algorithms.

From the above comparisons, there is still no mathematical theory to identify the superiority of GSO or gLASSO; one is a more accurate model, while another enjoys the benefit of numerical algorithms for finding global solutions. Extensive empirical studies revealed the significant advantages of GSO (or L_0 regularization model) in terms of model reliability, strong sparsity promoting capability and achieving solution with biological sense; see, e.g. [23, 34, 54, 57, 58]. Particularly, when a good initial point is provided, the GSO will achieve a high-quality local solution of (Eq. 5), which could be even better than the global solution of gLASSO. Therefore, by virtue of TF binding information derived from ChIP-seq/chip as a good initial guess, we adopted GSO in this study because of its advantages in model reliability and strong group sparsity promoting capability over gLASSO. Comparison of numerical results between these two models in the case of conversion from mouse fibroblast to mESC was consistent with previous observation that GSO performed better than gLASSO when the solution has high sparsity, since cell fate conversion needs only a small number of TFs out of hundreds to thousands of candidates.

In addition, our method does not require predefined gene regulatory network. It utilizes the inherent cell-specific regulatory information hidden in the OMICs data, and directly infers master TFs. Thus it is applicable to more cell types of clinical potentials. It predicts master TFs that could control all DEGs by taking all DEGs between donor cells and target cells as one target group. It is also able to quantify the dependency between master TFs and all DEGs that are needed to be changed from the donor cells into the target cells. This method considered multiple aspects thoroughly, thus shows better performance.

Furthermore, our method provides a more flexible platform for master TFs prediction. It is able to integrate not only transcriptome data and TF binding information but also other epigenetic information, which can provide cell-specific activities of DNA regulatory elements. Multiple tests described above have shown that the addition of each OMICs data type makes the prediction result better. TF binding information could be derived

from either ChIP-seq/chip of TFs or a combination of predicted TFBS and epigenetic data, such as DNase-seq (DNase I hypersensitive sites sequencing). Besides, super-enhancer regions identified from epigenetic data were found to be significantly enriched with the binding of our predicted master TFs, when super-enhancer information were further integrated, the total score and total weighted score were increased, improving the performance of master TF prediction when compared to utilizing TFBSs alone. To test the importance of using TFBS information as the initial guess X^0 , several test runs with completely and partially abandoning TFBS information were performed. Generally, the reduction of TFBS information led to a notable decrease in total score, suggesting that the integration of TFBS information is crucial for accurate master TF inference using our GSO method, and the absent of TFBS information resulted in terrible performance (Figure 3A). This is consistent with results of gene regulatory network construction in previous studies [17, 23], where the accuracy of networks constructed from transcriptome data alone was similar to those from random guessing. However, integration of transcriptome data increased the total score of master TF prediction from 76 to 91~97, which indicates the contribution of transcriptome data by providing the cell-type-specific gene regulatory information. The cell type specificity of transcriptomic data utilized is also important, it is because when the transcriptomic data was originated from a completely different cell type, the total score dramatically reduced to 69, which is even lower than that when the use of transcriptomic data was absent.

Our method is also highly reproducible, since using data from different sources always produce comparable results regardless of sequencing technologies or platforms (total score ranging from 91 to 97), as long as they are from the same cell type (Figure 3C). More than that, our method has a relatively high tolerance on transcriptome sparsity level. Artificial data reductions on different single-cell datasets were made, total scores were found to suffer a slight reduction (less than 7.5% reduction) when 50% of transcriptome data points was randomly selected, and a mild reduction (1–15% reduction) when 25% of transcriptome data points was randomly selected (Figure 3D).

Thanks to the recent advancements in single-cell sequencing technology, scRNA-seq can generate transcriptomes from thousands of cells in one single experiment. DrivAER [59], a machine learning method for gene set analysis on scRNA-seq data, is able to detect driving TFs during cell fate conversion when it runs on data derived from cell reprogramming process. However, as mentioned above, for most cell types, such data is not available since corresponding cell fate conversion has not yet succeed. Other than utilizing single-cell transcriptome data detected from cell reprogramming process, single-cell transcriptome data from different perturbation experiments targeting individual TFs (pooled CRISPR) can also provide vital TF-to-gene regulatory relationship information in the target cell, where they can be implemented in our method and achieve good results. Our results also indicate that using scRNA-seq in one single experiment alone can generate sufficient transcriptome profiles for master TF inference, thus reducing cost when compared with using bulk sequencing technology. Although scRNA-seq transcriptomes suffers from high data sparsity, our method has a high tolerance on data sparsity level. In addition, single-cell epigenetic technology, such as single-cell assay of transposase accessible chromatin with high-throughput sequencing (scATAC-seq), allows detection of open chromatin regions of the same population of cells or even the same single cells with scRNA-seq. There are existing methods making an afford to integrate these single-cell multi-OMICs data,

yet they mainly focus on cell population analysis and suffer greatly from high sparsity level [18, 60]. Our method has the potential to make use of these data for master TF identification with high tolerance on data sparsity level. When combined with pooled CRISPR and single-cell multi-OMICs, our method could also be applied to cell fate conversion between other cell types, even if they are not well studied or do not have known gene regulation information.

Methods

Enrichment analysis on TF binding in super-enhancers

Super-enhancers of mESCs were downloaded from dbSUPER [31]. Putative TFBSs within super-enhancers were predicted by MISP (<https://bitbucket.org/hanfeisun/misp>) with TFBS motifs collected from JASPAR [61], UniPROBE [62] and CIS-BP [63]. Each TFBS has a prediction score reported by MISP. To assess the enrichment of TFBSs in super-enhancers, an accumulative score (AS) of TF binding in all super-enhancers was calculated by:

$$AS^{\text{enh}} = \sum_{i=1}^n \max(S_i^{\text{enh}}), \quad (1)$$

where n is the total number of super-enhancers in a tissue/cell type, and $\max(S_i^{\text{enh}})$ is the maximum TFBS prediction score of the TF within the i^{th} super-enhancer. To normalize the ASs of different TFs, n intergenic sequences were randomly selected. TFBSs in all random sequences were predicted. And then AS of each TF in all random sequences were calculated. By sampling n random sequences for m times, the average ASs of m samplings were used to normalize the ASs of different TFs in super-enhancers. This normalized accumulative score (NAS) quantifies the enrichment of predicted TFBS in super-enhancers:

$$NAS = \ln \left(\frac{\sum_{i=1}^n \max(S_i^{\text{enh}})}{\sum_{j=1}^m \left(\sum_{i=1}^n \max(S_i^{\text{rand}_j}) \right) / m} \right), \quad (2)$$

where $\max(S_i^{\text{rand}_j})$ is the maximum TFBS prediction score of the TF within the i^{th} random sequence in the j^{th} sampling.

NASs were also calculated for 66 TFs' TFBS (including chromatin modifiers) identified by ChIP-seq/chip in mESC (Supplementary Table S1). S_i^{enh} and $S_i^{\text{rand}_j}$ were ChIP-seq/chip fold enrichment of TFBSs within super-enhancers and random sequences, respectively. TFs were then ranked by NASs.

Hub TFs and top TFs

Hub TFs were ranked by the number of targets [24] in mESC-specific networks, which was a directed network constructed by combining transcriptomes of perturbation experiments and TF binding data in our previous study [23]. Top TFs were ranked according to the difference between their out-degree and in-degree in the same networks, which is calculated by:

$$S_{\text{top}} = \frac{O - I}{O + I}, \quad (3)$$

where O is the out-degree and I is the in-degree [25].

Bulk and single-cell transcriptome data processing

DEGs between mouse fibroblasts and mESCs were determined by comparing their expression profiles, which were downloaded from Gene Expression Omnibus (GEO) [64]. Mouse embryonic fibroblast (MEF) samples included GSM1100519, GSM1100520, GSM1100521, GSM1310500, GSM853462, GSM853463, GSM853464, GSM1297612, GSM1297613 and GSM1297612; while mESC samples included GSM1053554, GSM1197055, GSM1197056, GSM1297603, GSM1297604, GSM1297605, GSM1168397, GSM1168398 and GSM1168399. DEGs between the two cell types were identified by Limma [65] with the cutoff of log₂ fold change >1 or < -1 and adjust P-value <0.05.

Bulk transcriptome data of 245 perturbation experiments in mESC were also collected from GEO (GSE16375 [42], GSE31381 [43], GSE26520 [44]). For each experiment, gene expression fold changes between mESC (control) and treatment samples were log₂ transformed. Combining log₂-transformed expression fold change data of all experiments, profiles of TFs (939 TFs) formed matrix A, and those of DEGs between fibroblasts and mESCs formed matrix B (Figure 1). Both matrices contain data from the 245 experiments (245 rows).

Single-cell transcriptome data of mouse fibroblasts, mESCs, iPSC and different time points of reprogramming processes detected by Fluidigm C1 (912 cells) or 10× Genomics (65 068 cells) were collected from GSE103221 [45]. Single-cell transcriptome data of naïve and primed mESC from pooled CRISPR of 25 TFs followed by scRNA-seq were downloaded from GSE142451 [48]. Single-cell transcriptome data of direct induction of retinal ganglion cell-like neurons were downloaded from GSE140128 [66]. For each dataset, normalized read counts of all genes were downloaded and log₂ transformed (log₂(count+1)). Similar to how we handled the bulk transcriptome data, log₂ transformed profiles of TFs in all cells formed matrix A, and those of DEGs between fibroblasts and mESCs formed matrix B (Figure 1). Both matrices contain data from all cells. How master TF predicted will be described in the following sections.

Since missing data is a major problem with single-cell transcriptome, to test whether the data imputation improves prediction performance, imputation was conducted for normalized read count matrix using three algorithms, Knn-smooth2 (<https://github.com/yanailab/knn-smoothing>), DrImpute [49] and SAVER [50], with default settings. Then matrices A and B were also extracted from the imputed matrices of three algorithms according to TF and DEG names needed for further prediction, respectively. The influence of sparsity level on the master TF prediction was tested by artificially increasing sparsity levels of different single-cell transcriptome matrices, which is to reduce data quantity to 75, 50 and 25%. In details, for each cell in a single-cell transcriptome matrix, 75, 50 or 25% of genes whose expression is non-zero were randomly selected and retained respectively, and the rest of genes were forced set as 0. Then matrices A and B were again extracted from the new matrices of increased sparsity for further prediction.

To assess the effect of transcriptome sample number on the result of master TF prediction, we calculated the total scores for a series of sample numbers for bulk transcriptome and each of single-cell transcriptome datasets (Figure 3B). Numbers of samples/cells were randomly selected from each of the four sets of bulk/single-cell transcriptomes, and formed new transcriptome matrices. Matrices A and B were again extracted from the new matrices of reduced sample/cell numbers. Then master TFs were predicted using each new matrix, and the prediction results between different numbers of samples/cells were

compared. Each experiment was repeat three times independently. When the sample number was reduced to 0, only TF binding information was used. TFs were ranked according to the number of their targets as DEGs between mouse fibroblasts and mESCs.

Group sparse optimization

As shown in Figure 1, regulatory relationship between TFs and targets were formulated approximately by a linear system

$$B = AX + \varepsilon, \quad (4)$$

where $A \in \mathbb{R}^{m \times r}$ denotes the expression data matrix of r TFs in m bulk transcriptome studies or single cells, $B \in \mathbb{R}^{m \times n}$ denotes the expression data matrix of n DEGs between mouse fibroblasts and mESCs in m bulk transcriptome studies or single cells, ε denotes the matrix of noise and $X \in \mathbb{R}^{r \times n}$ denotes the regulation matrix that describes the regulatory relationship between these r TFs and n DEGs. Matrices A and B were derived from bulk transcriptome data in perturbation experiments or single-cell transcriptomes as described in the previous section.

With these matrices, we then proceed to the inference of master TFs via the GSO method. The master TF inference is aimed to find a small number of TFs targeting most of the DEGs simultaneously. It can be described as an optimization problem to find an X such that the difference between AX and B is minimized with only a small number of selected TFs, whose regulatory strength on DEGs (i.e. $X_{i \cdot}$) are nonzeros. Note that the number of master TFs is measured by the $L_{2,0}$ norm of X , which is defined by the number of non-zeros rows of X (namely, the group sparsity of X , see Figure 1). Hence master TF inference can be formulated by a (nonconvex) group sparse optimization (GSO) problem, that minimizes the regression residual and the group sparse penalty simultaneously:

$$\min_X \|AX - B\|_F^2 + \lambda \|X\|_{2,0}, \quad (5)$$

where the Frobenius norm of X is defined by $\|X\|_F := \sqrt{\sum_{i=1}^r \sum_{j=1}^n X_{ij}^2}$, $\|X\|_{2,0}$ is the number of non-zeros rows, and $\lambda > 0$ is the regularization parameter providing a trade-off between accuracy of regression and group sparsity of variables. Note that the $L_{2,0}$ norm handles each row of X (regulatory relationships between one TF and all DEGs) as one group. Hence, by introducing the group sparse penalty, GSO selects the groups that quantify the regulatory strength of master TFs to all targeted DEGs. It is worth mentioning that gLASSO [35] is a well-known convex relaxation of GSO, which adopts an $L_{2,1}$ norm (i.e. $\|X\|_{2,1} := \sum_{i=1}^r \|X_{i \cdot}\|$) as the group sparsity promoting penalty instead of the $L_{2,0}$ norm in the problem (Eq. 5). The comparisons of mathematical theory between GSO and gLASSO were discussed in section Discussion.

Algorithms for solving GSO and gLASSO

In this study, we applied the iterative thresholding algorithms to solve both GSO (Eq. 5) and gLASSO. All experiments were performed with Matlab R2020a on personal desktop (Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz). An R package with the same functions is also available (in R version 4.0.2), which requires packages 'data.table', 'stringr', 'doParallel', 'foreach' and 'parallel'. Both codes are deposited in the GSO homepage and GitHub (see section Data Availability).

Iterative thresholding algorithms are one type of the most popular optimization algorithms for sparse optimization [34, 56, 67]. It has a fast convergence rate (at a linear convergence rate) and is of very low computational complexity (having analytical formulae at each iteration) [34]. Benefitting from its fast computing, simple formulation and low storage requirement, it is applicable and effective even for large-scale problems. Utilizing iterative thresholding algorithms to solve GSO has been investigated in a uniform framework of the proximal gradient method in our previous work [34]. In particular, the iterative hard thresholding algorithm (IHTA) [57] was proposed to solve GSO (Eq. 5), while the iterative soft thresholding algorithm (ISTA) was introduced [67] to solve gLASSO, which were formally described as follows.

IHTA

Set the step size $\nu = 1/(2\|A\|^2)$, start with an initial matrix $X^0 \in \mathbb{R}^{n \times r}$ and generate a sequence $\{X^k\} \subseteq \mathbb{R}^{n \times r}$ via the iteration:

$$X^{k+1} = H\left(X^k - \nu A^T (AX^k - b)\right), \quad (6)$$

where $H(\cdot)$ is the hard group thresholding operator, defined by

$$(H(X))_{i,\cdot} := \begin{cases} X_{i,\cdot}, & \text{if } \|X_{i,\cdot}\| > \sqrt{2\nu\lambda}, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

ISTA

Set the step size $\nu = 1/(2\|A\|^2)$, start with an initial matrix $X^0 \in \mathbb{R}^{n \times r}$ and generate a sequence $\{X^k\} \subseteq \mathbb{R}^{n \times r}$ via the iteration:

$$X^{k+1} = S\left(X^k - \nu A^T (AX^k - b)\right), \quad (8)$$

where $S(\cdot)$ is the soft group thresholding operator, defined by

$$(S(X))_{i,\cdot} := \begin{cases} \left(1 - \frac{\nu\lambda}{\|X_{i,\cdot}\|}\right) X_{i,\cdot}, & \text{if } \|X_{i,\cdot}\| > \nu\lambda, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Parameter update scheme

The regularization parameter λ plays an important role in adjusting a tradeoff between the fidelity of linear system and the group sparsity of the solution. After pre-setting K (i.e. the number of master TFs to be searched), we designed a dynamic scheme to update parameter λ such that an approximate solution within a given group sparsity level K of problem (Eq. 5) is obtained. The principle of the dynamic parameter updating scheme is as follows. Because both IHTA and ISTA include a thresholding operator, in which each group of variables is updated to zeros if its norm is smaller than a given threshold. Then, by virtue of the thresholding operator, parameter λ can be dynamically determined via the consideration of the pre-defined group sparsity level. This process dynamically sets the value of threshold as the norm of the $(K+1)^{\text{th}}$ most dominant row of the iterate such that only K master TFs of variables are selected while the other groups are vanishing. In particular, we denote $Z^k := X^k - \nu A^T (AX^k - b)$ and $R_k := (K+1)^{\text{th}}$ largest value of $\{\|Z_i^k\|\}$, and set the

dynamic update scheme for IHTA and ISTA respectively by

$$\lambda(H) = \frac{R_k^2}{2\nu} \text{ and } \lambda(S) = \frac{R_k}{\nu}. \quad (10)$$

TF scoring for GSO, gLASSO

To score and rank the master TFs reported by GSO and gLASSO, we selected a series of group sparsity level K from 1 to 20 and ran one GSO (or gLASSO) for each K . The group sparsity level K is the number of rows in matrix X that are non-zero, which is the number of selected master TFs, since each row is the regulatory coefficients of one TF to all DEGs (Figure 1). For each TF, we used K_{\min} to denote the smallest value of K when this TF is selected as master TF. Due to the principle that GSO (or gLASSO) selects the K most important TFs when the group sparsity is set as K , we assumed that TFs selected by GSO (or gLASSO) are more important when K is smaller, so a TF got a higher score if its K_{\min} was smaller. Therefore, the score of each TF selected by GSO (or gLASSO) is defined by

$$S_G = \begin{cases} \frac{1}{K_{\min}}, & \text{if this TF is selected in any GSO trial,} \\ 0, & \text{if it is not selected in all GSO trials.} \end{cases} \quad (11)$$

Please refer to Supplementary Table S3B for more details.

Integration of TF binding information and super-enhancer regions

TF binding information was transformed into a matrix X^0 that served as an initial guess to guide the approximation to the solution. TFBSs within each gene promoter (from -10 to $+10$ kbp of gene transcription start sites) were derived from ChIP-seq/chip data of the target cell mESCs as described in [23]. For TF i and gene j , if binding site of TF i is present at the promoter of gene j , X_{ij}^0 is assigned as the Pearson correlation coefficient (PCC) of the expression profiles of TF i ($A_{i,\cdot}$) and gene j ($B_{j,\cdot}$) across all samples or cells; otherwise, X_{ij}^0 is assigned as 0. When epigenetic information was incorporated, super-enhancer regions were used to filter the TFBSs. When a TFBS is within super-enhancer regions, X_{ij}^0 defined by this TFBS was assigned as PCC of the corresponding TF and gene; otherwise, it was reset as 0.

Other initial guess matrices were also generated to test the influence of different initial points to the final prediction results:

- (1) $X_{\text{zero}} := 0$, where each element in the matrix is zero, representing that no TF binding information was utilized;
- (2) $X_{50\%}$, representing randomly selected 50% of TF binding information from the original X^0 ;
- (3) $X_{\text{one}} := 1$, where each element in the matrix is one, assuming that all TFs are regulating all genes;
- (4) X_{Gaussian} , where each element is randomly generated from Gaussian distribution;
- (5) X_{uniform} , where each element is generated from uniform distribution.
- (6) $X_{\text{regression}} := (A^T A)^{-1} A^T B$, which is the solution of linear regression;
- (7) X_{PCC} : PCCs between TFs and genes, that is

$$(X_{\text{PCC}})_{ij} := \frac{\text{cov}(A_{i,\cdot}, B_{j,\cdot})}{\sigma_{A_i} \sigma_{B_j}}, \quad (12)$$

where σ denotes the standard deviation.

Mogrify and CellNet

Prediction of Mogrify was performed for the conversion between 'fibroblast' and 'embryonic stem cell lines'. Results were directly downloaded from <http://www.mogrify.net/>. Similar to our assumption, TFs selected by Mogrify are assumed to be more important when the total number of selected TFs is smaller. Thus, we chose the number of TFs from 5 to 20 (below 5 is not permitted by Mogrify) and ran Mogrify. Predicted TFs were ranked by the same scoring mechanism as S_C (Supplementary Table S3A). CellNet prediction was done with the transcriptomes of MEFs and mESCs mentioned in the preceding section using the CellNet online tool (<http://cellnet.hms.harvard.edu/>). Top 20 TFs were recorded.

Comparison of different approaches

To quantify the efficiencies of different methods, a total score was calculated for each method to summarize the ranking of all iPSC factors:

$$S_T = \sum_{i=1}^{20} (21 - R_i), \quad (13)$$

where i from 1 to 20 represents the top 20 TFs predicted as master TFs by each method, R_i is equal to the rank of TF i if TF i has been used to induce iPSC (standard TFs) and R_i is equal to 21 otherwise. This scoring system allows the ranking of standard iPSC TFs to be quantified and summed up, so that the more standard factors, the higher their ranking, the higher the total score. It considers only the existence and ranking of standard TFs in the predicted top 20 TF list, so the number 21 was used to make standard iPSC TFs ranking after 20 score 0.

To take the popularity of a TF for cell fate conversion into consideration, TFs with the ability to induce iPSCs were ranked according to the number of original PubMed publications that have used them for iPSC induction. A total weighted score was calculated for each method:

$$S_{TW} = \sum_{i=1}^{20} (10 - r_i) (21 - R_i), \quad (14)$$

where r_i is the rank of the standard TF i in a descending order list ranking the number of publications each standard TF has been used in wet lab iPSC inductions. Compared to the above scoring method, this method in additionally considered the popularity the standard TFs emerged in the prediction results. Similarly, the number 10 was used, since only 9 out of these 13 validated TFs were being predicted by all tested methods. It makes sure that the more publications using the standard iPSC TFs, the higher their scores.

Authors' contributions

J.Q. conceived of the study, designed the bioinformatics workflow, carried out the OMICs data collection, processing and integration, participated in the computational experiments of prediction and comparison of different methods, and drafted the manuscript. Y.H. designed and coded the mathematical model and algorithm, and participated in the computational experiments of prediction and comparison of different methods. J.C.Y. proposed the mathematical idea and designed the mathematical

model. R.W.T.L. participated in the comparison and analyze on the different substitution of initial matrix X^0 and drafted the manuscript. Y.Z. tested the contributions of various OMICs on the master TF prediction. Y.Q. participated in transcriptome data processing and analyses. J.W. conceived of the study, and participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Key Points

- This study introduces a novel computational method predicting master transcription factors in cell type conversion based on group sparse optimization technique and integration of multi-OMICs, which can be applicable to both single-cell and bulk OMICs data with high tolerance on data sparsity level.
- When it is compared with the other state-of-the-art prediction methods by a scoring system cross-referencing published and validated master TFs, it demonstrated superior performance.
- This method facilitates fast identification of key regulators, in hope of increasing successful cell identity conversion rates and reducing costs from experimental trials.

Data availability

The data and codes underlying this article are available at <https://qinlab.sysu.edu.cn/GSO> and <https://github.com/jingQinlab/GSO>.

Funding

The National Natural Science Foundation of China (41606143 to J.Q.); research grants from Research Grants Council, Hong Kong (17121414 M to J.W.); startup funds from Mayo Clinic, USA (Mayo Clinic Arizona and Center for Individualized Medicine to J.W.); National Natural Science Foundation of China (12071306, 11871347 to Y.H.); Natural Science Foundation of Guangdong Province of China (2019A1515011917, 2020B1515310008 to Y.H.); Natural Science Foundation of Shenzhen (JCYJ20190808173603590 to Y.H.); National Science Council of Taiwan (MOST 102-2115-M-039-003-MY3 to J.C.Y.); China Postdoctoral Science Foundation (grant 2019TQ0397 to R.W.T.L.).

References

1. Barker RA, Parmar M, Studer L, et al. Human trials of stem cell-derived dopamine neurons for Parkinson's disease: dawn of a new era. *Cell Stem Cell* 2017;21:569–73.
2. Sareen D, Saghizadeh M, Ornelas L, et al. Differentiation of human limb-derived induced pluripotent stem cells into limb-like epithelium. *Stem Cells Transl Med* 2014;3:1002–12.
3. Graf T, Enver T. Forcing cells to change lineages. *Nature* 2009;462:587–94.
4. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;126:663–76.
5. Buganim Y, Markoulaki S, van Wietmarschen N, et al. The developmental potential of iPSCs is greatly influenced by reprogramming factor selection. *Cell Stem Cell* 2014;15:295–309.

6. Davis RL, Weintraub H, Lassar AB. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* 1987;51:987–1000.
7. Olson EN. MyoD family: a paradigm for development? *Genes Dev* 1990;4:1454–61.
8. Kulesa H, Frampton J, Graf T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboplasts, and erythroblasts. *Genes Dev* 1995;9:1250–62.
9. Oestreich KJ, Weinmann AS. Transcriptional mechanisms that regulate T helper 1 cell differentiation. *Curr Opin Immunol* 2012;24:191–5.
10. Verzi MP, Shin H, San Roman AK, et al. Intestinal master transcription factor CDX2 controls chromatin access for partner transcription factor binding. *Mol Cell Biol* 2013;33:281–92.
11. Ieda M, Fu JD, Delgado-Olguin P, et al. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* 2010;142:375–86.
12. Vierbuchen T, Ostermeier A, Pang ZP, et al. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* 2010;463:1035–41.
13. Sekiya S, Suzuki A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* 2011;475:390–3.
14. Cahan P, Li H, Morris SA, et al. CellNet: network biology applied to stem cell engineering. *Cell* 2014;158:903–15.
15. Narsinh KH, Plews J, Wu JC. Comparison of human induced pluripotent and embryonic stem cells: fraternal or identical twins? *Mol Ther* 2011;19:635–8.
16. Heinaniemi M, Nykter M, Kramer R, et al. Gene-pair expression signatures reveal lineage control. *Nat Methods* 2013;10:577–83.
17. Marbach D, Costello JC, Kuffner R, et al. Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;9:796–804.
18. Hu X, Hu Y, Wu F, et al. Integration of single-cell multi-omics for gene regulatory network inference. *Comput Struct Biotechnol J* 2020;18:1925–38.
19. Pratapa A, Jaliyal AP, Law JN, et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;17:147–54.
20. Zhang S, Tian D, Tran NH, et al. Profiling the transcription factor regulatory networks of human cell types. *Nucleic Acids Res* 2014;42:12380–7.
21. Qin J, Li MJ, Wang P, et al. ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res* 2011;39:W430–6.
22. Wang P, Qin J, Qin Y, et al. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Res* 2015;43:W264–9.
23. Qin J, Hu Y, Xu F, et al. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* 2014;67:294–303.
24. Borneman AR, Leigh-Bell JA, Yu H, et al. Target hub proteins serve as master regulators of development in yeast. *Genes Dev* 2006;20:435–48.
25. Gerstein MB, Kundaje A, Hariharan M, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012;489:91–100.
26. Neph S, Stergachis AB, Reynolds A, et al. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 2012;150:1274–86.
27. Rackham OJ, Firas J, Fang H, et al. A predictive computational framework for direct reprogramming between human cell types. *Nat Genet* 2016;48:331–5.
28. Morris SA, Cahan P, Li H, et al. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* 2014;158:889–902.
29. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell* 2013;155:934–47.
30. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 2013;153:307–19.
31. Khan A, Zhang X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* 2015;44:D164–71.
32. Przednowek K, Iskra J, Maszczyk A, et al. Regression shrinkage and neural models in predicting the results of 400-metres hurdles races. *Biol Sport* 2016;33:415–21.
33. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 1998;20:33–61.
34. Hu Y, Li C, Meng K, et al. Group sparse optimization via lp, q regularization. *J Mach Learn Res* 2017;18:960–1011.
35. Ming Y, Yi L. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B* 2006;68:49–67.
36. Bach FR. Consistency of the group Lasso and multiple kernel learning. *J Mach Learn Res* 2008;9:1179–225.
37. Scardapane S, Comminiello D, Hussain A, et al. Group sparse regularization for deep neural networks. *Neurocomputing* 2017;241:81–9.
38. Maekawa M, Yamaguchi K, Nakamura T, et al. Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. *Nature* 2011;474:225–9.
39. Declercq J, Sheshadri P, Verfaillie CM, et al. Zic3 enhances the generation of mouse induced pluripotent stem cells. *Stem Cells Dev* 2013;22:2017–25.
40. Chen J, Gao Y, Huang H, et al. The combination of Tet1 with Oct4 generates high-quality mouse-induced pluripotent stem cells. *Stem Cells* 2015;33:686–98.
41. Iseki H, Nakachi Y, Hishida T, et al. Combined overexpression of JARID2, PRDM14, ESRRB, and SALL4A dramatically improves efficiency and kinetics of reprogramming to induced pluripotent stem cells. *Stem Cells* 2016;34:322–33.
42. Nishiyama A, Xin L, Sharov AA, et al. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* 2009;5:420–33.
43. Correa-Cerro LS, Piao Y, Sharov AA, et al. Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Sci Rep* 2011;1:167.
44. Nishiyama A, Sharov AA, Piao Y, et al. Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Sci Rep* 2013;3:1390.
45. Guo L, Lin L, Wang X, et al. Resolving cell fate decisions during somatic cell reprogramming by single-cell RNA-seq. *Mol Cell* 2019;73:815–29 e817.
46. Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 2017;14:297–301.
47. Dixit A, Parnas O, Li B, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 2016;167:1853–66 e1817.
48. Yang L, Zhu Y, Yu H, et al. scMAGECK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biol* 2020;21:19.

49. Gong W, Kwak IY, Pota P, et al. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 2018;**19**:220.
50. Huang M, Wang J, Torre E, et al. SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**:539–42.
51. Jackson CA, Castro DM, Saldi GA, et al. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *elife* 2020;**9**:e51254.
52. Andrews TS, Hemberg M. False signals induced by single-cell imputation. *F1000Res* 2018;**7**:1740.
53. Zhang R, Atwal GS, Lim WK. Noise regularization removes correlation artifacts in single-cell RNA-seq data preprocessing. *Patterns (N Y)* 2021;**2**:100211.
54. Chartrand R, Staneva V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Probl* 2008;**24**:035020.
55. Natarajan BK. Sparse approximate solutions to linear systems. *SIAM J Comput* 1995;**24**:227–34.
56. Blumensath T, Davies ME. Iterative thresholding for sparse approximations. *J Fourier Anal Appl* 2008;**14**: 629–54.
57. Blumensath T, Davies ME. Iterative hard thresholding for compressed sensing. *Appl Comput Harmon Anal* 2009;**27**:265–74.
58. Xu L, Zheng SC, Jia JY. Unnatural L0 sparse representation for natural image deblurring. *IEEE Proc CVPR* 2013; 1107–14.
59. Simon LM, Yan F, Zhao Z. DrivAER: Identification of driving transcriptional programs in single-cell RNA sequencing data. *Gigascience* 2020;**9**:1–10.
60. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform* 2020;**22**:1–13.
61. Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2016;**44**:D110–5.
62. Hume MA, Barrera LA, Gisselbrecht SS, et al. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 2015;**43**:D117–22.
63. Weirauch MT, Yang A, Albu M, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 2014;**158**:1431–43.
64. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
65. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
66. Wang J, He Q, Zhang K, et al. Quick commitment and efficient reprogramming route of direct induction of retinal ganglion cell-like neurons. *Stem Cell Rep* 2020;**15**:1095–110.
67. Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun Pure Appl Math* 2004;**57**:1413–57.