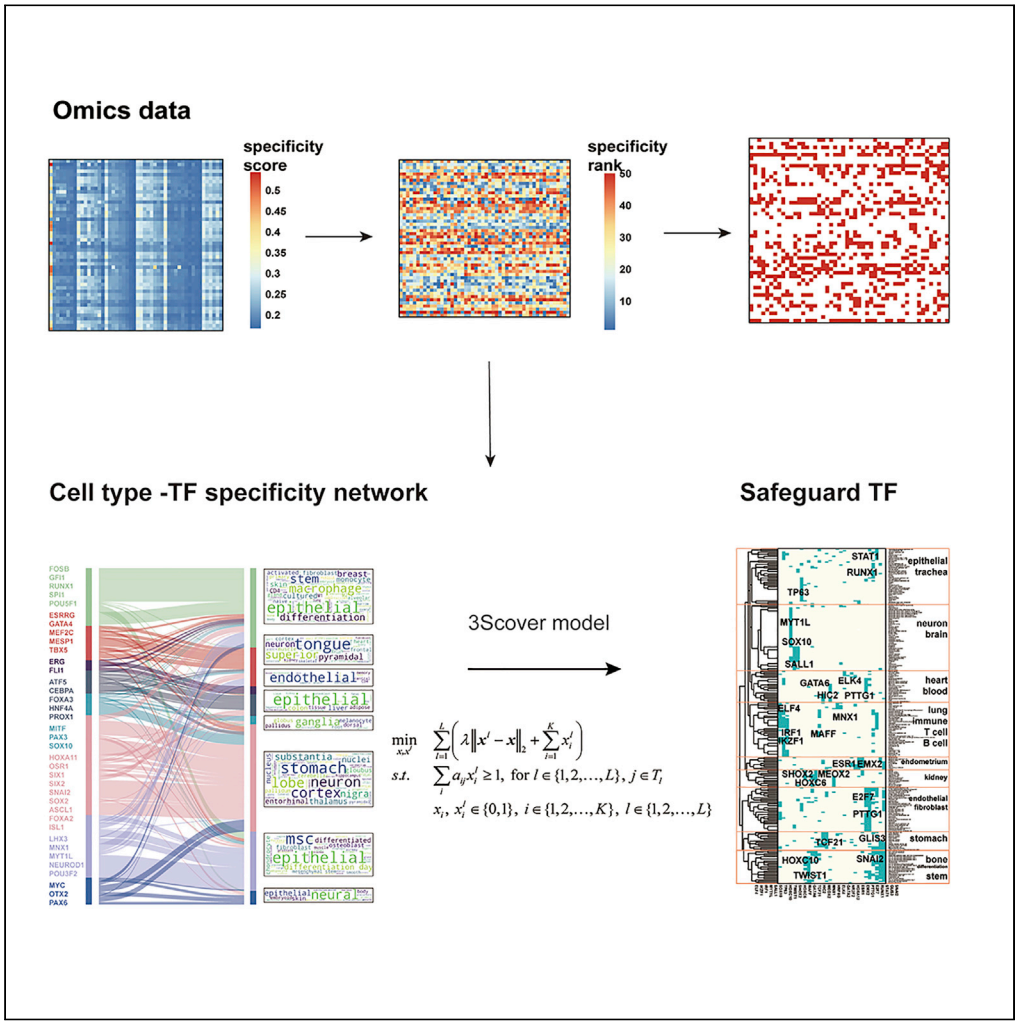
iScience

CellPress
OPEN ACCESS

Article

3Scover: Identifying Safeguard TF from Cell Type-TF Specificity Network by an Extended Minimum Set Cover Model



$$\min_{\lambda,x'} \sum_{l=1}^{L}\left(\lambda\left\|x^{l}-x\right\|_{2}+\sum_{i=1}^{K}x_{i}^{l}\right)$$
$$s.t. \quad \sum_{i} a_{ij}x_{i}^{l} \geq 1, \text{ for } l \in \{1,2,\ldots,L\}, j \in T_{l}$$
$$x_{i}, x_{i}^{l} \in \{0,1\}, \; i \in \{1,2,\ldots,K\}, \; l \in \{1,2,\ldots,L\}$$

Qiuyue Yuan,
Yong Wang

ywang@amss.ac.cn

HIGHLIGHTS

Cell type-TF specificity networks reveal the relationships among TF and cell identity

3SCover extracts safeguard TFs by "many but one specificity" and parsimony principle

Safeguard TFs are enriched in reprogramming panel and interact closely with CR

Safeguard TFs are conserved in mouse and human

Yuan & Wang, iScience *23*, 101227
June 26, 2020
https://doi.org/10.1016/j.isci.2020.101227

iScience

CellPress
OPEN ACCESS

Article

3Scover: Identifying Safeguard TF from Cell Type-TF Specificity Network by an Extended Minimum Set Cover Model

Qiuyue Yuan[1,2,3,4] and Yong Wang[1,2,3,4,5,*]

SUMMARY

Transcription factors (TFs) define cellular identity either by activating target cell program or by silencing donor program as demonstrated by intensive cell reprogramming studies. Here, we propose an extended minimum set cover model with stable selection (3Scover) to systematically identify silencing TFs, named safeguard TFs, from omics data. First, a cell type-TF specificity network is constructed to systematically link cell types with their specifically expressed TFs. Then we search the minimum TF set to cover this network with "many but one specificity" characteristic and integrate many subsampling models for a stable solution. 3Scover identified 30 safeguard TFs in human and mouse. These safeguard TFs are significantly enriched in the experimentally discovered reprogramming panel with their protein-protein interactors. In addition, they tend to interact closely with chromatin regulators, negatively regulate transcription, and function earlier in development. Collectively, 3Scover allows us to probe master TFs and combinatorial regulation in controlling cell identity.

INTRODUCTION

Transcription factors (TFs) are the master regulators for many important biological processes. Specifically, cell identity is controlled to a large extent by the TFs, which bind specific sequence, recruit chromatin regulators (CRs), turn on and off the target genes, and finally change the cell fate (Corces et al., 2018; D'Alessio et al., 2015; Duren et al., 2017). This fact is revealed by the seminal induced pluripotent stem cell experiments: that a small number of TFs are sufficient to establish gene expression profiles, which define pluripotent cell identity (Yamanaka, 2012). Further investigations confirmed that ectopic expression of TF converts cells from one type to another by many cellular reprogramming experiments. For example, the combination of the three TFs (ASCL1, BRN2, and MYT1L) has been shown to reprogram fibroblasts and other somatic cells into induced neuronal (iN) cells (Masserdotti et al., 2016). A pool of six genes (TFs: SIX1, SIX2, HOXA11, and SNAI2; transcriptional co-activator: OSR1, EYA1) were found to activate nephron progenitor phenotype in the adult proximal tubule cell line (Hendry et al., 2013).

Those cell reprogramming experiments imply that TFs work in different combinations with other TFs or co-factors to enact a vast repertoire of cellular fates, and combinatorial regulations among several TFs are critical to convert one cell type to another (Li et al., 2019; Wang et al., 2009; Zhang et al., 2018). However, the detailed underlying mechanism remains elusive, for example, which TFs are important in the 1,500–2,000 TFs encoded in the genome, how to form the right combination panel, what is each TF's role in the panel, etc.

Recently, Mall et al. proposed that reprogramming requires the activation of target cell programs and silencing of donor cell programs (Mall et al., 2017) by taking the reprogramming experiment from fibroblasts and other somatic cells to iN cells as an illustration. ASCL1 acts as a pioneer TF to activate the neuronal program, whereas MYT1L acts as a safeguard TF to directly repress other non-neuronal somatic lineages to maintain neuronal identity (Mall et al., 2017). Systematic approaches are proposed to identify pioneer TFs for most cell types in humans according to two characteristics: typically expressed at relatively high levels and in a quite strict cell-type-specific fashion (D'Alessio et al., 2015). These pioneer TF characteristics were widely accepted and used to identify regulon, a group of genes that are regulated by TF as a unit, in all cell types for mouse (Suo et al., 2018). Compared with pioneer TFs, safeguard TFs are difficult to

[1]CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[2]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[3]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, 650223, China

[4]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, 310024, China

[5]Lead Contact

*Correspondence: ywang@amss.ac.cn

https://doi.org/10.1016/j.isci.2020.101227

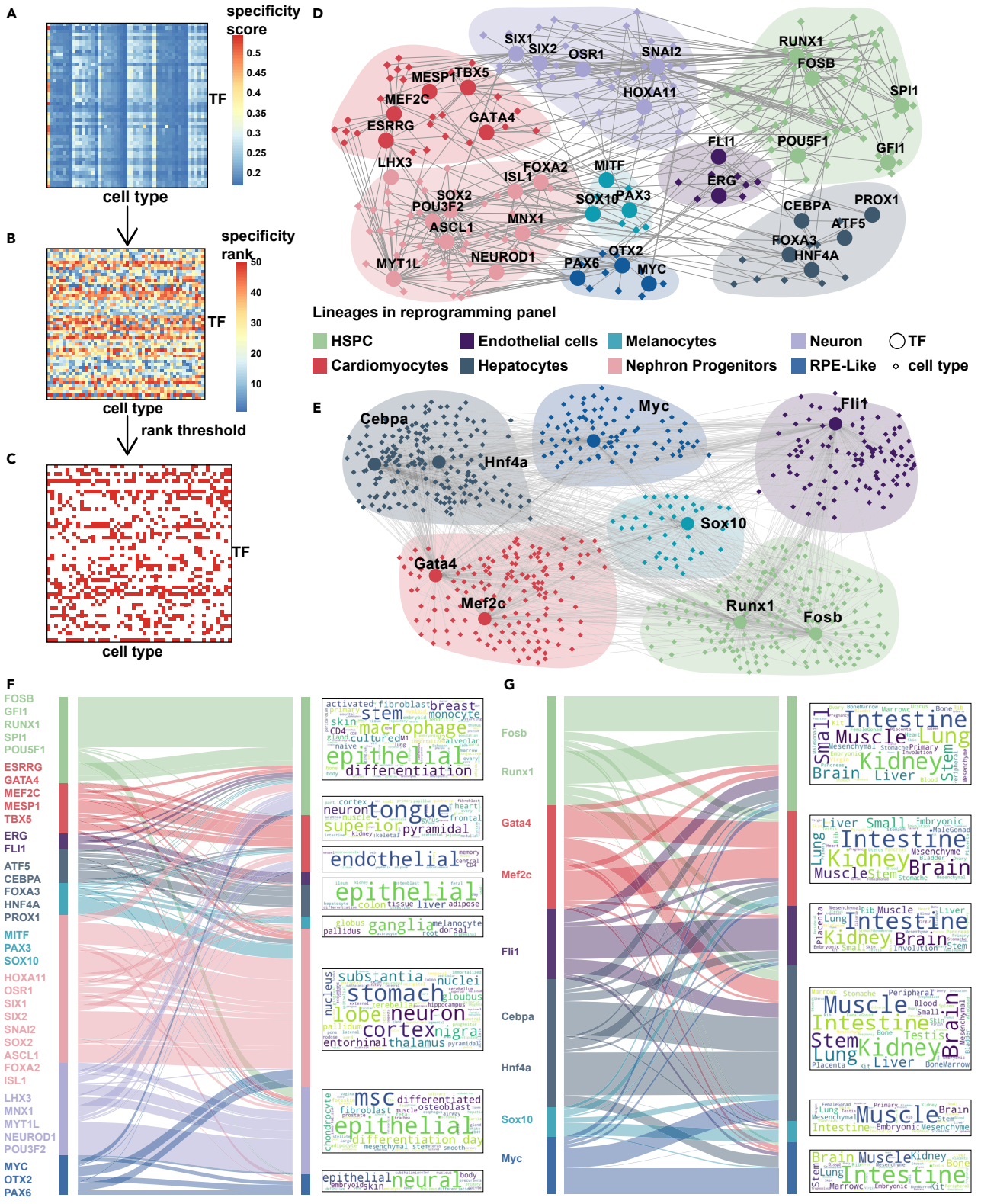


Figure 1. Cell Type-TF Specificity Network Construction and the Reprogramming TF Panel Induced Subnetwork

(A) TF is associated with cell types by the defined specificity scores, which are organized into a matrix with rows and columns indicating TFs and cell types.
(B) Specificity score is ranked for all TFs within each cell type (column).

Figure 1. *Continued*

(C) Adjacency matrix of the cell type-TF network by thresholding the specificity rank data. It is derived from specificity rank matrix by setting a threshold and binarized into 1 and 0 indicating if TF is present and absent in a given cell type. This adjacency matrix can be easily represented as a bipartite graph wherein TF and cell types are nodes and their connections are edges. The cell type-TF specificity network is a bipartite network and can induce the TF-TF network if two TFs share the same cell type.

(D) Reprogramming panel inducted TF-TF subnetwork. We select the TFs used in reprogramming experiment from literature, called *reprogramming panel TF*, and extract the induced subnetwork. TFs are grouped by cell type to which they are used to reprogram, called *preprogramming panel lineage*. Color of TFs and cell types correspond to their reprogramming panel lineage.

(E) Reprogramming panel TF induced subnetwork from mouse cell type-TF specificity network.

(F) Sankey plot and word cloud of (D). Sankey plot shows the relationship of TF groups and cell type groups in (D); each group of cell type is indicated by the corresponding word cloud.

(G) Sankey plot and word cloud of (E).

See also Figure S1.

be defined in biology and characterized from high-throughput data. Thus, computational method to identify safeguard TFs is in pressing need and expected to reduce barriers in understanding the mechanisms underlying TF combinatorial regulation.

In this study, we propose a Set cover model with Stable Selection (3Scover) to identify safeguard TFs across a large collection of different cell types or tissues in a robust and parsimony way. 3Scover takes the TF ranking or scoring data in each cell type as input, reconstructs a cell type-TF network, extracts patterns from the network, and then identifies safeguard TFs as output. To test the validity, we apply 3Scover to two large-scale transcriptomic datasets and identify 30 safeguard TFs in human and mouse (available at https://github.com/AMSSwanglab/3SCover/blob/master/Human_safeguard_TF.txt; https://github.com/AMSSwanglab/3SCover/blob/master/Mouse_safeguard_TF.txt). Those safeguard TFs serve as distinctive signatures of lineages and group similar cell types together. The experimentally verified TFs in cell reprogramming panels are enriched in safeguard TF-TF protein-protein interaction (PPI) network. We further explore the biological properties for safeguard TFs by public omics data. The regulatory pattern of safeguard TFs is different from that of other TFs, with a higher percentage of distal enhancers. Safeguard TFs closely interact with the CRs with negative regulation of gene expression in epigenomics. Those are consistent with the concept of safeguard TF functioning in early development and low differentiation context and playing a negative regulatory role.

RESULTS

Constructing Cell Type-TF Specificity Network

We construct a cell type-TF specificity network, i.e., a cell type-TF bipartite graph, to connect TFs with cell types (nodes in the network) by quantifying how "specific" a TF belongs to a cell type (edges in the network). This network provides a global landscape and useful resource to study TF and cell type relationships and allows us to explore the hidden patterns and systematically dissect the TF "specificity" across various cell types. Figure 1 shows the procedure of constructing the cell type-TF specificity network and modeling structures in the network.

As the first step, we quantify TF's "specificity" in a given cell type by two properties derived from its expression pattern: "high expression" and "cell type-specific fashion." "High expression" is a prerequisite for specificity, and TF with "high expression" is more likely to play major roles in gene regulation. To exclude the housekeeping TFs, which are highly expressed in all cell types, we introduce "cell type-specific fashion" to remove TFs with basic cellular function. An entropy-based measure of Jensen-Shannon divergence is calculated as specificity score, and TFs are ranked in each cell type (see Methods for details). These specificity score data are shown in Figure 1A as a matrix where the row denotes TFs and column denotes cell types. This score matrix is converted into TF specificity rank matrix in each cell type as shown in Figure 1B. We then choose a rank threshold (top 30 in our study) and output the cell type-TF adjacency matrix in Figure 1C by setting items below the threshold to 1, whereas other items to 0. This sparse matrix can be naturally reorganized into a bipartite graph linking cell types with the TFs, named *cell type-TF specificity network*.

Following the aforementioned procedure, we calculate the specificity rank matrix and construct the human and mouse cell type-TF specificity networks from two datasets, i.e., 1,055 TFs across 233 cell types ranked by gene expression in human (D'Alessio et al., 2015) and 202 TFs (regulons) across 818 cell types ranked by

the regulatory strength in mouse (Suo et al., 2018). To visualize the two large networks, we focus on the TFs previously used in lineage reprogramming experiment and extract the reprogramming panel TF-induced subnetworks as shown in Figures 1D and 1E. TFs are classified into eight groups based on their reprogramming panel lineages, including hematopoietic multipotent progenitor cell, cardiomyocytes, endothelial cells, hepatocytes, melanocytes, nephron progenitors, neuron, and retinal pigment epithelium-like cells (RPE-Like). We next classify cell types into eight groups according to the number of their linked TFs in the cell type-TF specificity network (see Methods for details). We observe that the cell type-TF specificity networks are well organized in modular structure in both human and mouse. For example, ASCL1, POU3F2, SOX2, MYT1L, NEUROD1, ISL1, MNX1, and FOXA2 are highly associated with neuron cell types in human. Those cell type-TF module structures are highly conserved in human and mouse. GATA4 and MEF2C are associated with cardiomyocytes, and CEBPA and HNF4A are associated with hepatocytes. The Sankey plot in Figures 1F and 1G summarizes the relationship between cell types and TFs. We observe that the numbers of cell types linked with TFs vary widely. In the human subnetwork, SNAI2 and RUNX1 are linked with maximal number of 51 cell types, whereas LHX3 is linked with only one cell type. For mouse, Hnf4a is linked with a maximal number of 230 cell types and the least number is 97 by Sox10. We generate word cloud for each group of cell type to display the cell type annotations by their frequency in the group. Figures 1F and 1G indicate that the constructed cell type-TF specificity networks are consistent with the reprogramming experiments. For example, in human subnetwork, "endothelial" is of high appearance in the cell types in ERG and FLI1 linking group and they are known TFs used to reprogram donor cell types to endothelial cells. In mouse subnetwork, Cebpa and Hnf4a connect with liver and are known as factors to reprogram to hepatocytes. Taken together, the reconstructed cell type-TF specificity network is in high quality and the encoded high-level relationships among TFs and cell types need to be explored further. We next develop systematic method to mine the knowledge from the network.

Characterizing Safeguard TFs in Cell Type-TF Specificity Network

We observe in the reconstructed cell type-TF specificity network that some TFs have high degree by specifically expressing in many cell types, but some TFs do not. We first compare the known safeguard TF MYT1L (Mall et al., 2017) and pioneer TF NEUROD1 (Guo et al., 2014) for their degrees in the network. Both TFs are used to reprogram fibroblast to neuron but show different specificity pattern. We extract the subnetwork induced by MYT1L and NEUROD1, respectively, from the human cell type-TF specificity network (Figure S1A), and most of the cell types included are neuron related. Figure 2A shows that the degree of MYT1L in human cell type-TF specificity network is 35, whereas the degree is 4 for NEUROD1. We further observe their distinct expression patterns across tissues. Figure 2B shows that NEUROD1 is specifically expressed in "brain-cerebellar hemisphere" and "brain-cerebellum," whereas safeguard TF MYT1L is turned on in almost all neuronal cell types (brain amygdala, brain anterior cingulate cortex, brain caudate, brain cerebellar hemisphere, brain cerebellum, brain cortex, brain frontal cortex, brain hippocampus, brain hypothalamus, brain nucleus accumbens, brain putamen, and pituitary) and is turned off in other tissues. The contrast of MYT1L and NEUROD1 in Figures 2A and 2C suggests an important characteristic for safeguard TF: "many but one specificity" at gene expression level, which requires TF to be expressed with high level in a cell type-specific way in many cell types, but not just in one cell type. This is different from pioneer TF by requiring narrow cell type specificity. This "many but one specificity" property is well explained by the fact that the maintenance of the neuronal lineage context not only needs the activation of neuron functional properties but also needs repression properties of other lineages (Mall et al., 2017).

In addition to "many but one specificity" for a single TF, we observe in Figures 1D and 1E that a small set of TFs can cover almost all the cell types. This motivates us to deduce the coverage property from the whole cell type-TF specificity network. We make a reasonable assumption that safeguard TFs are indispensable in each lineage to switch off the context of other lineages. Putting together, safeguard TFs from all lineages constitute the safeguard TF set. Motivated by the law of parsimony, we hypothesize that the safeguard TF set is organized under the overall goal of parsimony principle as the evolution outcome: the lineage program maintaining relies on a minimal set of safeguard TFs repressing conversion to other lineages.

Collectively, we propose two quantitative characteristics that safeguard TF set should meet. (1) "Many but one specificity": Specificity can be quantified by entropy-based measure to assess whether the TF is specifically expressed in a broad lineage, but not just in a single cell type. (2) Parsimony: All cell type identities

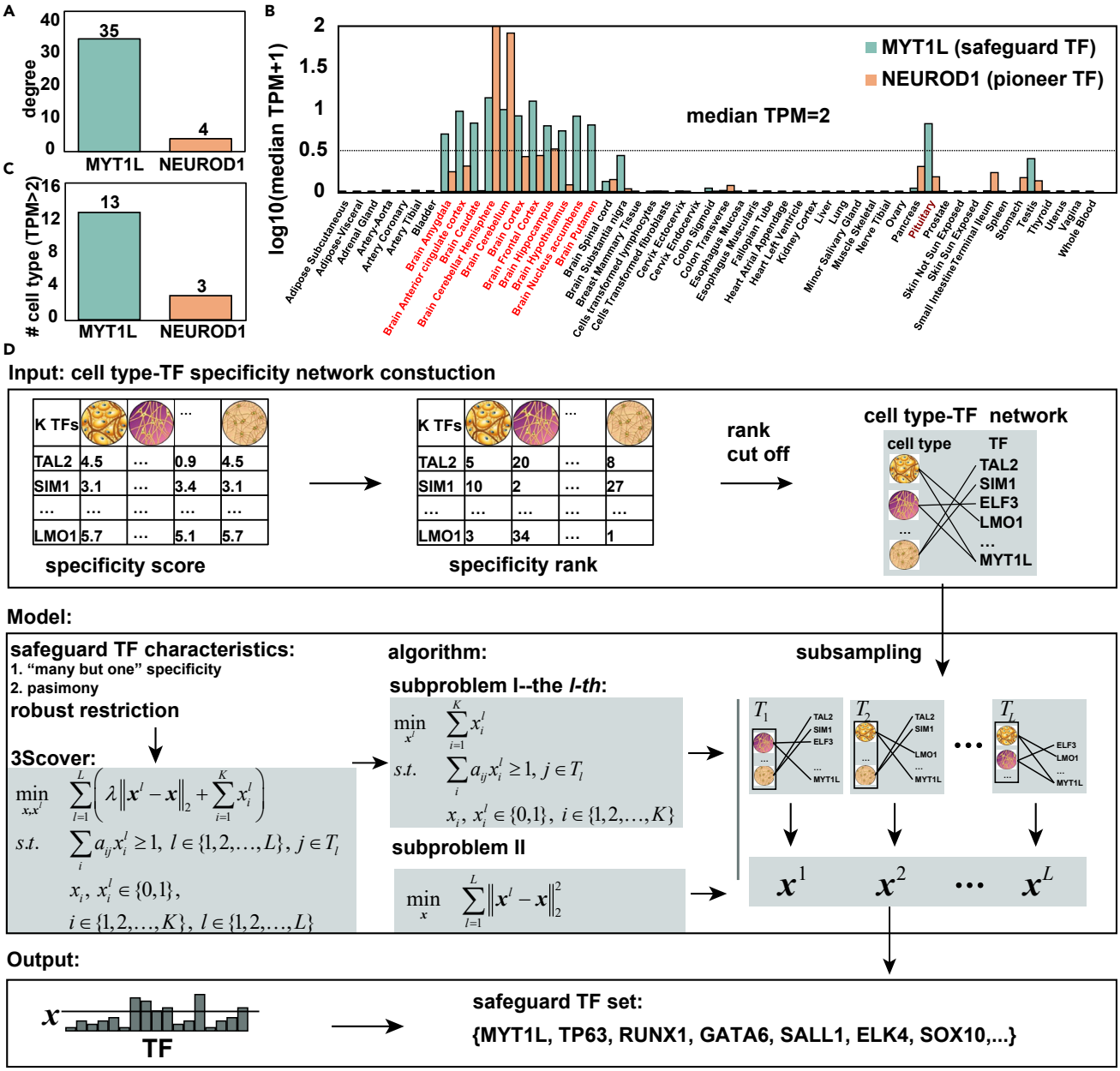


Figure 2. Safeguard TF Characterization and 3Scover Framework

(A) The degree of MYT1L and NEUROD1 in human cell type-TF specificity network.

(B) Comparison of expression pattern for safeguard TF and pioneer TF across tissues. Bar plot shows the median TPM of MYT1L (safeguard TF, green) and Neurod1 (pioneer TF, orange) in 53 tissues from GTEx dataset. The line indicates that TPM value is 2. In 12 tissues, MYT1L is above the line (red), in which 11 tissues are brain related.

(C) Number of cell types whose median expression is above the line. There are 13 tissues for MYT1L (green) and only 3 for NEUROD1 (orange).

(D) Overview of 3Scover framework. 3Scover takes cell type-TF specificity network as input, models two major characteristics of safeguard TF, combines the minimum set cover problem model with ensemble strategy, and prioritizes safeguard TFs as output. Specifically, the solution stability is achieved by subsampling the input network, finding the minimum set cover solution for each subnetwork (subproblem I), and then integrating them into a stable safeguard TF list (subproblem II).

in an organism should be safeguarded by a minimal set of TFs, in which a combination of TFs maintains the context of corresponding lineage. We note that the parsimony characteristic is a global property for the safeguard TF set to cover all lineages, whereas the "many but one specificity" characteristic is a local property for a certain TF to cover some broadly specific lineage.

Identifying Safeguard TF by Set Cover Model with Stability Selection (3Scover)

Based on the deduced properties of safeguard TF, we propose a model, 3Scover (Set Cover Problem with Stability Selection), to systematically identify the safeguard TFs, taking cell type-TF specificity network as input. The main idea behind 3Scover is illustrated in Figure 2D. Classical set cover model is introduced to find the minimum safeguard TF set to cover the cell type-TF specificity network. To get a robust safeguard TF set, we extend the classical model by introducing the stability selection. This is an ensemble strategy by subsampling the cell type-TF specificity network and aggregating the minimum safeguard TF sets based on each subnetwork (Figure 2D and the motivations are in Methods).

3Scover is formally described as follows:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{x}^l} \quad & \sum_{l=1}^{L}\left(\lambda\|\mathbf{x}^l-\mathbf{x}\|_2+\sum_{i=1}^{K}x_i^l\right) \\ \text{s.t.} \quad & \sum_i a_{ij}x_i^l \geq 1, \ \text{for } l\in\{1,2,\ldots,L\}, \ j\in T_l \\ & x_i, \ x_i^l\in\{0,1\}, \ i\in\{1,2,\ldots,K\}, \ l\in\{1,2,\ldots,L\} \end{aligned}$$

where L is the number of subsampling times; K and N are, respectively, the number of TFs and cell types in the network; $\mathbf{x}^l$ represents the solution for l-th set cover subproblem by subsampling; and $\mathbf{x}$ is the consistent or stable solution for all the subproblems. The term in objective function $\sum_{i=1}^{K}x_i^l$ finds the minimum set cover for the l-th subproblem; the term in objective function $\sum_{l=1}^{L}\|\mathbf{x}^l-\mathbf{x}\|$ minimizes the distance between the consistent solution with subsampling solutions; the constraints $\sum_i a_{ij}x_i^l\geq 1, \ \text{for } l\in\{1,2,\ldots,L\}, \ \text{for } j\in T_l$ are introduced to restrict the l-th solution to cover all cell types in T_l; and T_l is the set of cell types in the l-th random subsampling.

We tackle the large scale 0-1 integer linear programming by iteratively solving two sub-optimization problems (Figure 2C and Methods). The set cover sub-optimization problem, which is known as one of Karp's 21 NP complete problem, is solved by branch and bound algorithm in CPLEX for MATLAB.

3Scover Uncovers Distinctive Signatures of Lineages

We first apply 3Scover to identify safeguard TFs in human. We subsample from the human cell type-TF specificity network for 1,000 times as input, and 3Scover ranks the TFs by the probability of the TF being a safeguard TF (Table 1). We select the top 30 TFs as our final safeguard TF set. This includes 5 TFs in existing reprogramming panels—MYT1L, RUNX1, MNX1, SNAI2, and SOX10 (Table 2). And we can reasonably speculate that those safeguard TFs work with pioneer TFs when reprogramming to "melanocytes," "nephron progenitors," "neuron," and "haematopoietic multipotent progenitor cell."

Those 30 safeguard TFs serve as distinctive signatures of lineages. We group cell types into nine lineage groups (Figure S2), and cell types in the same group perform similar functions. The expression patterns in Figure 3A clearly show that 30 safeguard TFs are representative in the cell type-TF specificity network. We further annotate each group by summarizing the included cell types and identify the signature safeguard TF for each group (see Methods for details). For example, there are 46 cell types in the neuron group (Figure 3A), in which 31 are brain regions, such as "prefrontal cortex," "midbrain," and "cerebellar hemisphere." Of the remaining 15 cell types 7 are neuronal cell types. Thus, this group is annotated as neuron/brain and associates to safeguard TF MYT1L, SOX10, and SALL1. When we focus on MYT1L as positive control, we find that 33 of 35 MYT1L linking cell types are included in this group and all 33 cell types are related to brain or neuron.

Figure 3B shows the heatmap of safeguard TFs' expression pattern using the independent GTEx expression data across 53 tissues. We find that all tissues from brain are in the first group. Also, MYT1L, SOX10, and SALL1 are the safeguard TFs in this group. This is in accordance with the result in Figure 3A. From the clustering result of the cell types, we can conclude that similar safeguard TF profiles lead to similar cell types.

TF	x_i in the Optimal Solution	Cell Type Group	TF Family	Canonical Sequence Length	Length	# Linked Cell Type	Reprogramming Panel	Degree in PPI Network	Interacting CR	# of CR	Phylostratum	Conserved in Mouse?
MYT1L	1	Neuron	C2H2 ZF	1,186	542,162	35	Neuron (Marro et al., 2011)	3		0	6	No
TP63	0.933	Epithelial	p53	680	265,854	28		188	BRD8, SMARCD2, MDM2, MDM4, CARM1, KAT2B	6	2	No
IKZF1	0.926	immune	C2H2 ZF	519	129121	39		173	HDAC1, CHD4, HDAC2, HDAC4, HDAC5, HDAC7, HDAC3, RBBP4, CHD3, SMARCA4, HDAC11, CBX8	12	2	No
RUNX1	0.914	Epithelial	Runt	453	261499	51	HMPC (Sandler et al., 2014)	151	SUV39H1, KAT6B, HDAC1, HDAC3, KAT6A, HDAC2, TRIM33, SMARCA4, SMARCB1, KMT2A, DNMT1, HDAC11	12	5	Yes
PTTG1	0.891	Endothelial	Unknown	202	6,939	43		136		0	12	No
GATA6	0.828	Heart	GATA	595	33,095	22		10		0	2	Yes
SALL1	0.817	Neuron	C2H2 ZF	1,324	15,299	25		25		0	2	No
MNX1	0.569	Immune	Homeodomain	401	5,802	9	neuron (Lee et al., 2009)	1		0	3	No
SHOX2	0.555	Kidney	Homeodomain	331	10,154	26		5		0	2	No
STAT1	0.424	Epithelial	STAT	750	45,216	53		295	SMARCA4, JAK2, SMARCA2, DOT1L, HDAC4, HDAC1, HDAC3	7	2	No
GLIS3	0.419	Stomach	C2H2 ZF	775	475,909	41		24		0	2	No
HOXC6	0.343	Kidney	Homeodomain	235	13,973	26		9		0	2	No
HOXC10	0.272	Bone	Homeodomain	342	5,118	22		10		0	2	No

Table 1. Safeguard TFs in Human Identified by 3Scover and Their Genomic Features

(Continued on next page)

TF	x_i in the Optimal Solution	Cell Type Group	TF Family	Canonical Sequence Length	Length	# Linked Cell Type	Reprogramming Panel	Degree in PPI Network	Interacting CR	# of CR	Phylostratum	Conserved in Mouse?
GATA2	0.271	Endometrium	GATA	480	13,767	3		61	HDAC3, HDAC5, KAT2A	3	2	Yes
ELF4	0.257	Immune	Ets	663	45,795	37		14	MDM2	1	5	No
ELK4	0.244	Heart	Ets	431	24,931	11		11	SIRT7	1	5	Yes
HOXA13	0.244	Endometrium	Homeodomain	388	3,228	13		0		0	3	No
TCF21	0.234	Bone	bHLH	179	6,418	24		15		0	6	No
MAFF	0.21	Immune	bZIP	164	14,580	28		70	HDAC5	1	6	No
HNF4G	0.188	Immune	Nuclear receptor	408	26,860	13		15		0	5	No
SNAI2	0.161	Bone	C2H2 ZF	268	3,762	51	nephron progenitors (Hendry et al., 2013)	32	HDAC2, HDAC1, KDM1A, CHD4, MDM2	5	2	No
ESR1	0.156	Endometrium	Nuclear receptor	595	412,779	15		1,330	SMARCA4, SMARCD1, SMARCA2, MDM2, TADA3, KAT5, HDAC7, HDAC4, HDAC5, HDAC9, RBBP4, HDAC1, HDAC3, KAT6A, HDAC2, KDM1A, KMT2D, RBBP5, ASH2L, WDR5, TRRAP, EHMT2, CHD4, EP400, RUVBL1, RUVBL2, EZH2, SUV39H1, SMARCD3, SUPT6H, WHSC1L1, SUZ12, CHD6	33	5	No
IRF1	0.151	Immune	IRF	325	9,166	27		73	KAT2B, KAT2A, SMARCA4, MDM2	4	5	No
EMX2	0.149	Endometrium	Homeodomain	252	7,103	16		3		0	2	No

Table 1. *Continued*

(Continued on next page)

TF	x_i in the Optimal Solution	Cell Type Group	TF Family	Canonical Sequence Length	Length	# Linked Cell Type	Reprogramming Panel	Degree in PPI Network	Interacting CR	# of CR	Phylostratum	Conserved in Mouse?
TWIST1	0.139	Bone	bHLH	202	2,206	27		72	KAT2B, HDAC2, CHD4, HDAC3, WDR5, SETD8, CHD3, HDAC6	8	6	No
NR2F2	0.138	Stomach	Nuclear receptor	414	14,337	9		73	HDAC1, SMARCAD1	2	5	No
SOX10	0.135	Neuron	HMG/Sox	466	12,222	20	Melanocytes (Yang et al., 2014)	16		0	2	No
HIC2	0.118	Heart	C2H2 ZF	615	34,059	15		22		0	2	No
E2F7	0.106	Endothelial	E2F	911	44,336	49		17	RUVBL1	1	2	No
MEOX2	0.103	Kidney	Homeodomain	304	75,473	12		220	KAT5, CXXC1	2	2	No

Table 1. Continued

x_i value in the optimal solution denotes the probability of the TF being a safeguard TF, and 30 safeguard TFs are ranked by x_i; cell type group corresponds to cell type-TF heatmap in Figure 3A; canonical sequence is the sequence of the TF to describe all the protein products encoded by one gene in a given species in a single entry; number of linked cell type is the number of linked cell type of TF in the cell type-TF specificity network; reprogramming panel is the lineage to which TF is used to reprogram; degree in PPI network is the degree of TF in human PPI network; interacting CR is the linked CRs of TF in the PPI network; phylostratum is the evolutionary emergence level of TF (Domazet-Loso and Tautz, 2008), and a higher level corresponds to a later emergence of gene.

Reprogramming Program Lineage	Combination of TFs	Safeguard TF
Melanocytes	MITF, PAX3, SOX10	SOX10
Nephron progenitors	HOXA11, OSR1, SIX1, SIX2, SNAI2	SNAI2
Neuron	ASCL1, FOXA2, ISL1, LHX3, MNX1, MYT1L, NEUROD1, POU3F2	MYT1L
Hematopoietic	FOSB, GFI1, RUNX1, SPI1, POU5F1	RUNX1

Table 2. Safeguard TFs Identified by 3Scover that Have Been Used for Lineage Reprogramming in Human
TF combination includes all TFs used to reprogram to the reprogramming penal lineage. Safeguard TF identified in our work included in the TF combination is listed in the third column.

The Genomic Features of Safeguard TFs

TFs in Reprogramming Panel Are Enriched in PPI Network of Safeguard TFs

TFs tend to co-regulate transcription by interacting with other TFs, CRs, and co-factors. We reconstruct the PPI network of safeguard TFs with other TFs based on the PPI repository BioGRID. We further label safeguard TFs and TFs in the list of 35 experimentally verified TFs used in cell reprogramming in Figure 4A. We ask if the TFs in reprogramming panel are enriched in this network. Thirteen TFs are included in the network for the total 205 TFs, and eight are expected by chance (Figure 4B). This gives a fold change 1.87 and p value 0.0068 (hypergeometric test). It supports that safeguard TFs are strongly implicated in reprogramming. We can define two types of proteins based on the network structure. The type I protein interacts with only one safeguard TF and type II protein interacts with more than two safeguard TFs. Most safeguard TFs interact with both type I and type II proteins, suggesting that safeguard TFs can function independently and cooperatively with other TFs.

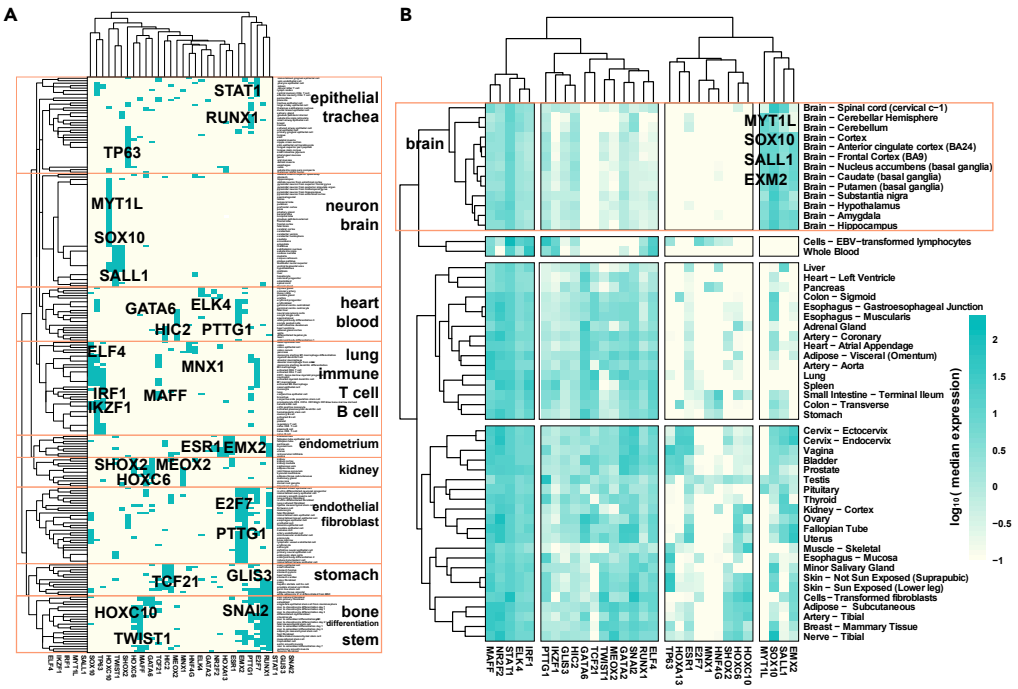


Figure 3. Human Safeguard TFs Serve Distinctive Signatures of Lineages
(A) Heatmap based on adjacency matrix of mouse safeguard TF-induced specificity network. Green represents that the cell type (row) is linked by the TF (column). We identify nine groups of cell types and assign each TF in the group according to the heatmap. We annotate each group by summarizing its included cell types.
(B) Heatmap of independent expression data from GTEx database. According to the median expression of safeguard TFs, we can identify brain group and the three signature TFs consistently show up— MYT1L, SOX10, and SALL1.
See also Figure S2.

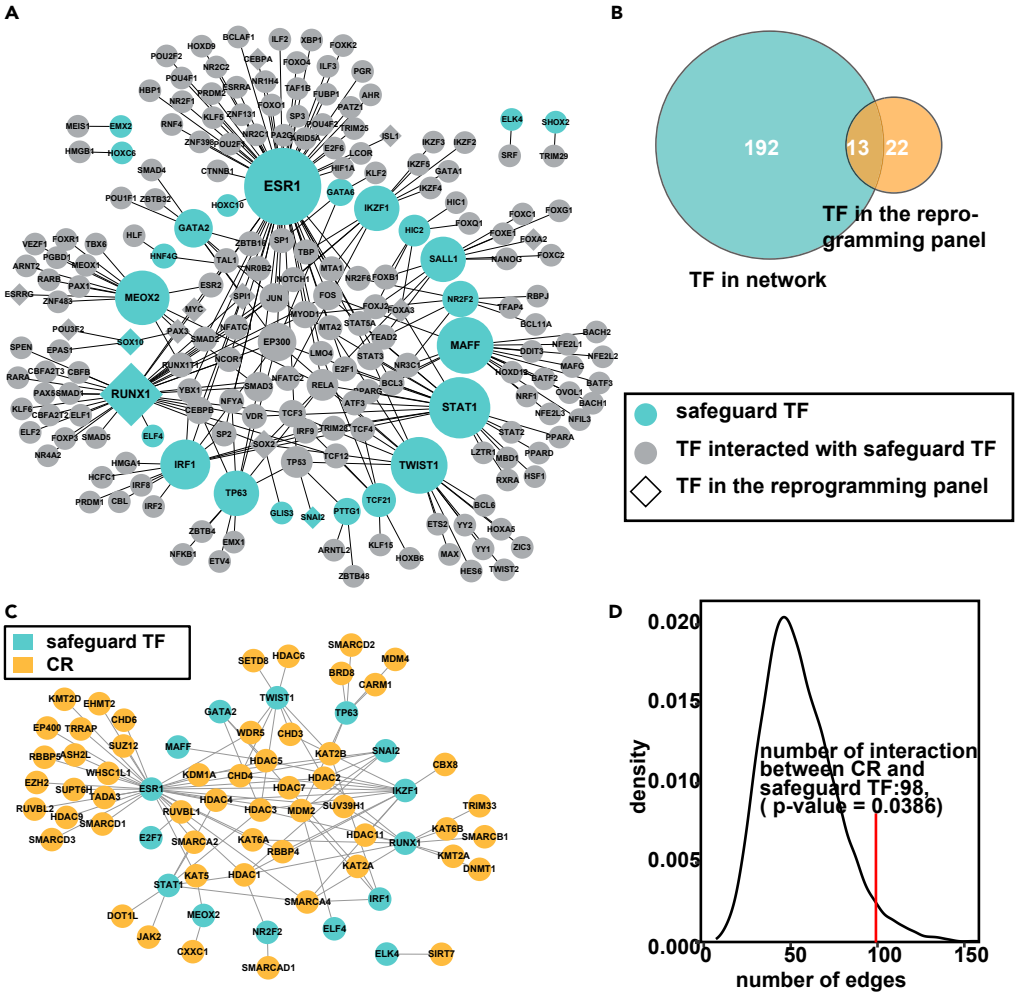


Figure 4. Safeguard TFs Interact with Other TFs and CRs in the Protein-Protein Interaction Network

(A) PPI network among safeguard TFs and other TFs. Green node represents safeguard TF, and rhombus represents TF in reprogramming panel.

(B) Overlap of TFs in the network and TFs in reprogramming panel. Reprogramming panel TFs are significantly enriched in safeguard TFs and its protein-protein interactors.

(C) PPI network among safeguard TFs and CRs. Safeguard TFs are denoted by green nodes, and CRs are denoted by orange nodes.

(D) Distribution of the number of edges in the PPI network. The distribution is constructed by randomly sampling 30 TFs and CRs 10,000 times. Red line marks number of edges for network in (C) with a number of 98 and p value 0.0386.

In the PPI network, MYC has interaction with safeguard TF RNUX1, protecting the first group annotated with epithelial and trachea. It was previously used to reprogram to RPE-Like cells by cooperating with NCoR/SMRT co-repressors to create an epigenetic barrier to somatic cell reprogramming (Zhuang et al., 2018). It suggests that RNUX1 may safeguard epithelial lineage and repress other lineages by recruiting repressors such as MYC.

Safeguard TFs Interact Closely with Chromatin Regulator

We then check one subclass of proteins in the PPI network. We hypothesize that safeguard TF should recruit CRs to regulate the chromatin accessibility and to repress cell type differentiation to other lineages. We construct a safeguard TF-CR PPI network by extracting the interactions among TFs and CRs (Figure 4C). Figure 4D shows that the number of edges in the network of safeguard TF is significantly larger than networks constructed by randomly picking 30 TFs (p value, 0.0386, 10,000 random networks). It suggests that safeguard TF and CR tend to interact closely. Furthermore, functional enrichment analysis shows that CRs in

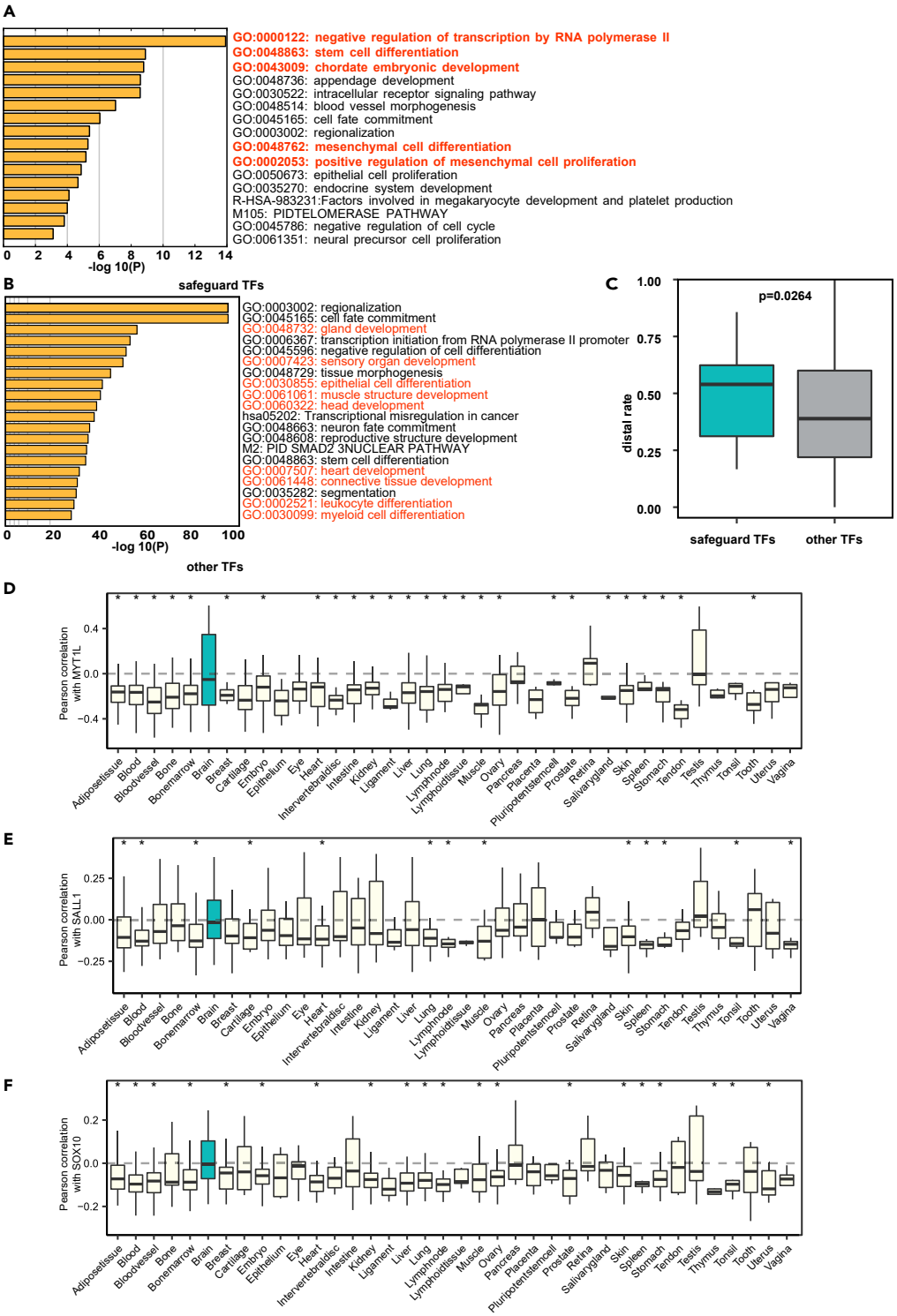


Figure 5. Safeguard TFs Tend to Take Effects in Early Stage of Development and Differentiation and Repress Other Lineages

(A) Function enrichment for human safeguard TFs. Function annotations specifically for safeguard TF are marked in red.

(B) Function enrichment for other TFs and function annotations specifically for other TFs are marked by red.

(C) Comparing the percentage of distal enhancers (intergenic enhancers) between safeguard TFs and other TFs by t test. A p value 0.0264 shows that the percentage of distal enhancer of safeguard TFs is significantly higher than other TFs.

Figure 5. *Continued*

(D) Mean correlation between safeguard TF MYT1L and cell type makers. For each tissue type, we calculate the Pearson correlation coefficients on median TPM of MYT1L and tissue type markers across all cell types in GTEx datasets. The cell type markers are from CellMarker database.

(E) Mean correlation between safeguard TF SALL1 and cell type markers.

(F) Mean correlation between safeguard TF SOX10 and cell type makers.

See also Figure S3.

this network are enriched in negative regulation of expression (Figure S1C), whereas all CRs aggregate into epigenetic regulation of gene expression (Figure S1D).

Safeguard TFs Tend to Function in Early Stage in Development and Differentiation

To explore the function of safeguard TF in detail, we perform enrichment on both 30 safeguard TFs and other TFs in the human cell type-TF specificity network (D'Alessio et al., 2015) (Figures 5A and 5B). The most enriched function is negative regulation of transcription by RNA polymerase II, which offers the potential for the safeguard TFs to repress expression of core genes in other cell types. And 19 of 30 safeguard TFs have this function. Compared with the enriched function of other TFs, chordate embryonic development is only enriched in safeguard TFs, whereas the exclusively enriched functions of other TFs are gland development, sensory organ development, muscle structure development, head development, heart development, and myeloid cell differentiation. All of those are late specific tissue or lineage development. This suggests that safeguard TFs may take effect earlier than other TFs. Besides, mesenchymal cell (which can differentiate into a variety of cell types) differentiation and positive regulation of mesenchymal cell proliferation are only enriched in safeguard TFs, whereas epithelial cell differentiation and myeloid cell differentiation, as well as leukocyte differentiation, are only enriched in other TFs. These also provide evidence that safeguard TFs take effect in earlier differentiation context.

Safeguard TFs Tend to Be Regulated by Distal Regulatory Elements

We further compare safeguard TFs and other TFs for nine genomic features by unpaired t test. Those features include regulatory complexity (defined as number of regulatory enhancers associated with this TF), number of proximal enhancers, number of distal enhancers, number of exons, exon length, PPI degree, gene length, ratio of proximal enhancer, and ratio of distal enhancer. Figure 5C shows that safeguard TF is significantly higher than other TFs in the ratio of distal enhancers, which are key contributors to gene expression specificity among cell types (Bulger and Groudine, 2011). It suggests that safeguard TFs might be elaborately regulated by distal enhancers and are significantly different from other TFs.

Safeguard TFs Tend to Repress Other Lineages

An essential feature of safeguard TF is repressing differentiation to other lineages. We use the correlation of safeguard TF and the tissue type marker genes (CellMarker database, Zhang et al., 2019) across cell types in GTEx data to check if safeguard TF's expression level is negatively correlated to those marker genes' expression (see Methods). Figure 5D shows that safeguard TF in neuron group, MYT1L, has significantly negative mean correlation with cell type marker genes in 26 of 37 tissues other than brain (t test, p value $<$ 0.05, 23 for q-value $<$ 0.05). Figures 5E and 5F show the similar results for SALL1 (13 of 37) and SOX10 (20 of 37). This indicates that SOX10, MYT1L, and SALL1 tend to repress neuron lineages differentiation to other lineages. We can get similar results in some other safeguard TFs (Figure S3), supporting that safeguard TFs may repress many different somatic lineages using an independent large-scale gene expression dataset.

Safeguard TFs Tend to Be Conserved in Mouse and Human

In addition to human dataset, we apply 3Scover to reveal 30 mouse safeguard TFs by randomly subsampling for 1,000 times in the mouse cell type-TF specificity network (Table 3). Figure 6A shows the heatmap of adjacency matrix for mouse safeguard TF-induced specificity network. Clearly those 30 TFs are very representative. Figures 6B and 6C are the enriched function for mouse safeguard TFs and other TFs. We can find that embryonic morphogenesis is only enriched in mouse safeguard TFs and the corresponding function for other TFs is morphogenesis of epithelium. This suggests that mouse safeguard TFs take effect in an earlier development stage. Besides, mouse safeguard TFs work in less differentiated stage, such as hematopoietic progenitor cell differentiation, compared with other TFs with the function of myeloid cell differentiation. These two observations are consistent with the human safeguard TFs.

TF	x_i	Family	Canonical Sequence Length	Length	# Linked Cell Type	Reprogramming Panel Lineage	PPI Degree	CR	#CR	Phylostratum	Conserved in Human?
Rel	0.781	Rel	587	34,120	206		5		0	3	No
Egr1	0.704	C2H2 ZF	533	3,750	178		9		0	2	No
Elf2	0.685	Ets	593	87,925	141		7		0	5	No
Sox4	0.634	HMG/Sox	390	4,780	202		0		0	2	No
Elk4	0.548	Ets	430	18,080	135		0		0	5	Yes
Pole3	0.447	Unknown	145	1,214	195		3	Chrac1	1	2	No
Hnf4a	0.416	Nuclear receptor	474	66,051	230	Hepatocytes (Yang et al., 2014)	7	Sirt1	1	5	No
Pparg	0.416	Nuclear receptor	505	129,258	207		87	Men1, dac1, Sirt1	3	5	No
Gata6	0.394	GATA	589	33,126	225		15	Bmi1	1	2	Yes
Runx3	0.307	Runt	409	57,346	241		1		0	5	No
Klf10	0.295	C2H2 ZF	479	6,306	159		4		0	2	No
Cebpd	0.253	bZIP	269	2,260	162		7	Rb1	1	3	No
Hoxa10	0.211	Homeodomain	416	3,743	142		6		0	2	No
Gtf2i	0.209	GTF2I-like	998	76,929	157		3		0	2	No
Tal1	0.193	bHLH	329	13,945	156		18	Hdac1, Kat2b	2	6	No
Tead2	0.172	TEA	445	17,867	211		4		0	3	No
Zfp580	0.153	C2H2 ZF	172	2,192	175		0		0	2	No
Gata4	0.142	GATA	441	46,358	196	Cardiomyocytes (Chen et al., 2012)	27	Smarca4, Eed	2	2	No
Foxo3	0.136	Forkhead	672	90,957	132		8		0	3	No
Gata2	0.126	GATA	480	8,369	172		3		0	2	Yes
Xbp1	0.124	bZIP	267	5,353	169		2		0	6	No
E2f4	0.115	E2F	410	7,708	212		4		0	2	No
Tfdp1	0.114	E2F	410	39,698	221		10		0	2	No
Trps1	0.114	GATA	1281	234,291	123		6		0	3	No

Table 3. Safeguard TFs in Mouse Identified by 3Scover and Their Genomic Features

(Continued on next page)

TF	x_i	Family	Canonical Sequence Length	Length	# Linked Cell Type	Reprogramming Panel Lineage	PPI Degree	CR	#CR	Phylostratum	Conserved in Human?
Pou2f1	0.105	Homeodomain; POU	770	137,481	94		21		0	3	No
Nfyb	0.103	Unknown	207	15,442	192		0		0	2	No
Nr2f6	0.098	Nuclear receptor	390	7,834	141		4		0	5	No
Creb3l1	0.087	bZIP	519	41,843	191		1		0	2	No
Myb	0.085	Myb/SANT	636	36,055	192		21		0	2	No
Runx1	0.085	Runt	451	95,864	177	Hematopoietic stem cell (Sandler et al., 2014)	44	Hdac1, Hdac2, Rbbp4, Smarca4, Smarcb1, Ash2l, Phc1, Smarcd1, Bmi1	10	5	Yes

Table 3. Continued

The genomic feature descriptions are the same as in Table 1.

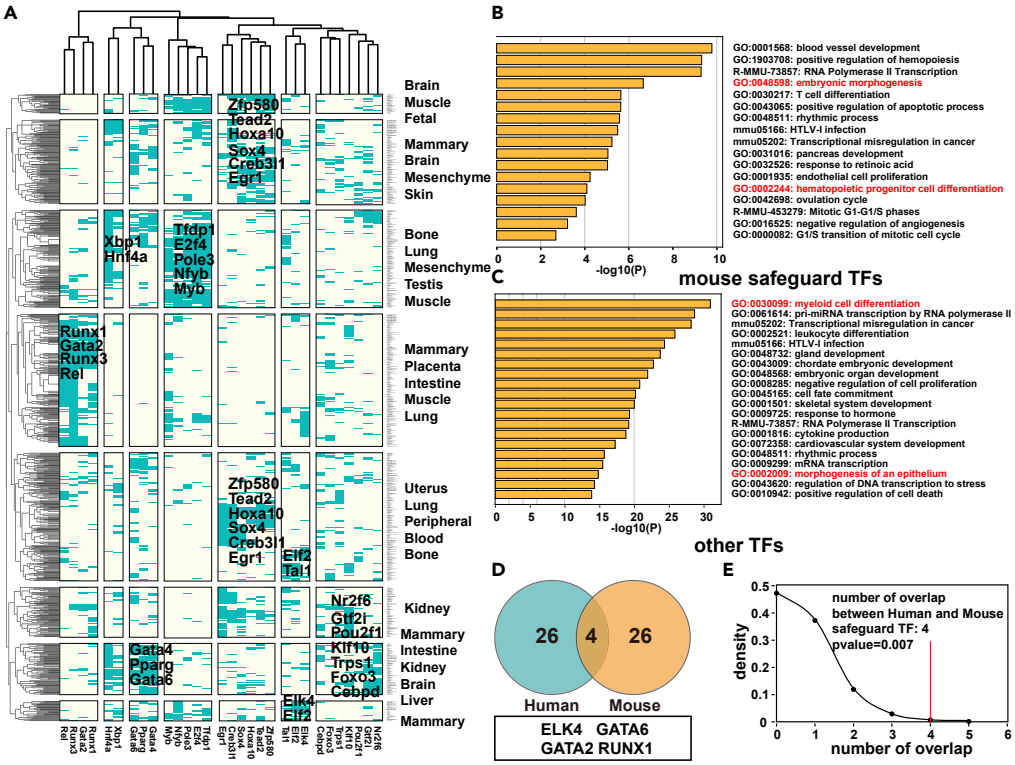


Figure 6. Safeguard TFs Are Conserved in Mouse and Human

(A) Heatmap of adjacency matrix for mouse safeguard TF-induced specificity network. Green represents that cell type (row) is associated to the TF (column). We identify eight groups of cell types and TFs according to the heatmap. We annotate each group by summarizing its included cell types.

(B) Function enrichment of mouse safeguard TFs. Red represents mouse safeguard TF-specific functions.

(C) Function enrichment of other non-safeguard mouse TFs. Red represents non-safeguard TF-specific functions.

(D) Conserved safeguard TFs in mouse and human (ELK4, GATA6, GATA2, and RNUX1).

(E) The number of overlapped safeguard TFs between mouse and human are significantly larger than random. We count the overlap of randomly picked 30 mouse TFs and random picked 30 human TFs for 1,000 times and then generate the null distribution. Red line marks the observed number of overlaps of safeguard TF with a p value 0.007.

Figure 6D shows that four safeguard TFs are conserved in mouse and human, ELK4, GATA6, GATA2, and RUNX1; three of them are experimentally verified reprogramming TFs, HNF4A, GATA4, and RUNX1. The conservation of safeguard TFs in human and mouse are statistically enriched with a p value 0.007 (Figure 6E) (see Methods). However, all human safeguard TFs in the brain and neuron—MYT1L, SOX10, and SALL1—are human specific, suggesting the regulatory difference of nervous system between human and mouse.

3Scover Predicts Pou2f1 as Muscle Candidate Safeguard TFs

An important application of 3Scover is to predict the safeguard TF for a certain tissue. We take muscle as an example. We first find the safeguard TF specifically linked with muscle lineage by a fold change (see Methods). Figure 7A shows that the fold changes of Rel and Pou21 between muscle lineage and all cell types are ectopic high. Besides, we take advantage of the fact that there are some muscle linages that appear during reprogramming from mouse embryonic fibroblasts (MEFs) to iN cells (Treutlein et al., 2016). We utilize the single-cell RNA sequencing data at multiple time points for iN reprogramming to explore the expression pattern of different TFs in the iN reprogramming progress. Figures 7B–7D show the expression pattern of Myod1, Rel, and Pou2f1. It suggests that the expression pattern of Pou2f1 is similar to that of Myod1, which is a muscle pioneer reprogramming factor (Davis et al., 1987). In addition, it has been suggested that Oct-1 (Pou2f1) is involved in the specification of myogenic phenotypes (Lakich et al., 1998) and causes disorganization when under-expressed (Columbaro et al., 2013). Bringing the aforementioned observations together, we suggest Pou2f1 as the candidate safeguard TF for muscle.

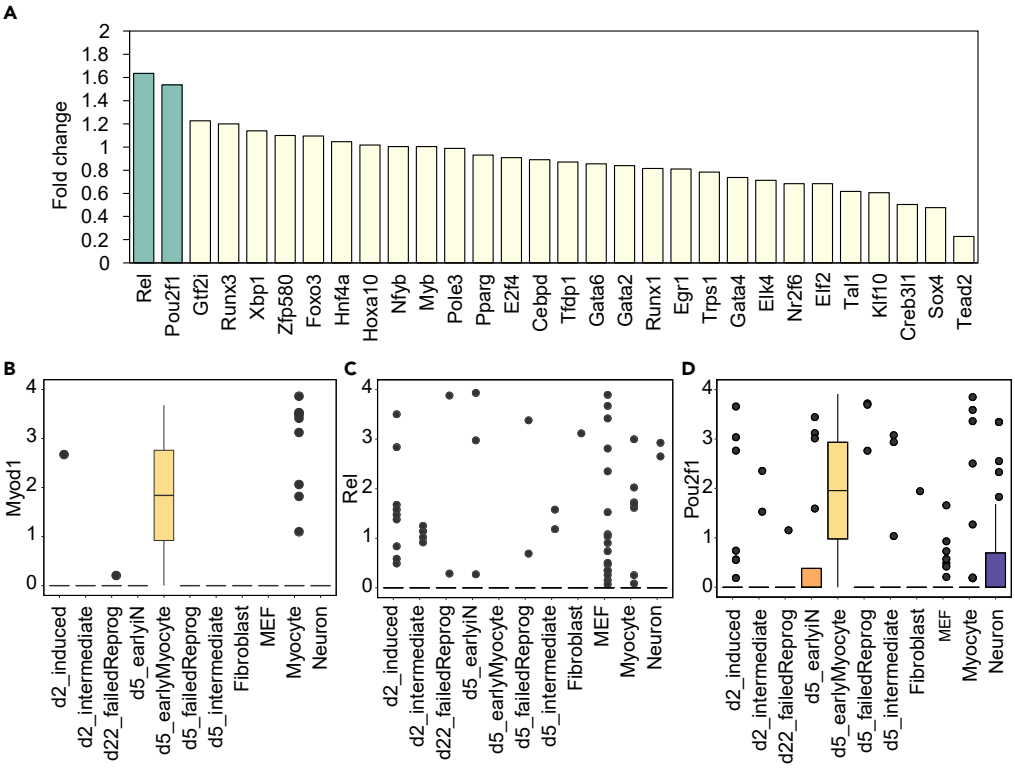


Figure 7. Pou2f1 Is a Candidate Muscle Safeguard TF

(A) Fold change of mouse safeguard TFs in muscle based on cell type-TF specificity network. Fold change is computed by dividing the percentage of cell types linked with the TF in muscle by the percentage in all cell types. Green represents Rel and Pou2f1.

(B) Boxplot of Myod1's expression in the single cell RNA sequencing data for reprogramming form MEF to iN. Each point represents a single cell, and all cells are classified to nine groups.

(C) Boxplot of Rel's expression.

(D) Boxplot of Pou2f1's expression.

DISCUSSION

We propose 3Scover, an ensemble method based on set cover model, to find a robust set of TFs in the reconstructed cell type-TF specificity network taking advantage of two quantitative characteristics of safeguard TF. 3Scover takes cell type-cell type interaction, cell type-TF interaction, and TF-TF interaction into account and is different from the naive strategy to identify safeguard TFs only in one cell type. 3Scover relies on the cell type-TF specificity network as the only input. In our test, this network is constructed by gene expression data (human) or regulatory strength data (mouse). It can be generalized to motif enrichment score given chromatin accessibility data or even experimentally measured cell type-TF specificity network in future.

At present, MYT1L is the only known and experimentally validated safeguard TF. 3SCover, as the first computational method for identifying safeguard TFs, extracts two quantitative characteristics of safeguard TF based on the observation of MYT1L. There are some further evidences for "many but one" specificity we observed in the literature. In reprogramming, safeguard TFs are often used to reprogram many different donor cell types (Xu et al., 2015). Besides, MYT1L is a pan neuron-specific TF (Mall et al., 2017). These facts suggest that safeguard TFs are lineage specific but not specific in only one cell type. More experimentally verified safeguard TFs will enhance this "many but one" specificity observation and help us to better define and capture the main safeguard TF's character.

For the human and mouse data, we choose 30 as a rank cutoff to construct the cell type-TF specificity network. We tried cutoff larger than 30, but the number of TFs in solution of set cover problem remains

about the same (Figure S1C). Here, 30 is chosen as a demonstration because dozens of candidate TFs are reported as signatures of the cell type in literature. The choice can be guided by the knowledge of the number of reprogramming factors for a cell type in practice.

Several genomic evidences suggest that safeguard TFs act as a repressor, including the negative regulation of safeguard TFs and CR having interaction, the negative correlation of safeguard TF with cell type marker genes, as well as safeguard TF RNUX1 recruiting MYC by PPI, forming a barrier to other lineages. However, how to define "repression" from data and how the safeguard TFs repress other lineages remains a question waiting for more wet-laboratory experiments and data analysis to explore.

3Scover ranks MYT1L as the first in our human safeguard TF list, and MYT1L is the only known and experimentally validated safeguard TF for neuron. It is well known that the combination of ASCL1 as a pioneer TF, MYT1L as a safeguard TF, and BRN2 as a context-dependent TF reprograms fibroblasts into iN cells. In addition, MYT1L mutations can cause intellectual disability (Blanchet et al., 2017) suggesting that genetic mutations on a safeguard TF lead to disease related to cell types protected by it. Based on our pilot study of the safeguard TFs in the whole cell type-TF network, we propose a working model to help to understand the role of safeguard TFs in the development and differentiation. When pluripotent stem cells differentiate to a certain type of mature cell, safeguard TF should be turned on first to maintain the lineage including the target mature cell type and repress other lineages that stem cells may convert to. After that, pioneer TFs and other context-dependent TFs are turned on to direct the cell to differentiation target cell type. This model is conceptional and needs more independent data and evidence to support and verify. We believe that many safeguard TFs will be revealed in future, and this allows better model and mechanism understanding.

Limitations of the Study

Our work is limited as a pure computational prediction for safeguard TF by omics data. 3Scover provides promising safeguard TF candidates, which should be further supported and validated by well-designed wet-laboratory functional experiments. We believed the perturbation experiments (knock-down or overexpression) and genomic binding experiments (chromatin immunoprecipitation sequencing) in the right cellular context will be extremely useful. Second, "Safeguard TF" is a limited term in this study because we define the safeguard TF only through the comparison of MYT1L as safeguard TF and NEUROND1 as pioneer TF based on the expression across certain cell types. However, we highlight that the "many but one specificity" and parsimony principle in the cell type-TF specificity network are quite general concepts and may reveal other important TFs in a more complete dataset with more cell types. Third, safeguard TF has the implication to "repress" other lineages in addition to "many but one specificity." How to define "repression" from other omics data and how the safeguard TFs repress other lineages remains a question waiting for more wet-laboratory experiments and data analysis to explore.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yong Wang (ywang@amss.ac.cn).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The datasets and code generated during this study are available at https://github.com/AMSSwanglab/3SCover. The identified 30 safeguard TFs in human and mouse are available at https://github.com/AMSSwanglab/3SCover/blob/master/Human_safeguard_TF.txt; https://github.com/AMSSwanglab/3SCover/blob/master/Mouse_safeguard_TF.txt. All code related to the Data visualization techniques is available at https://github.com/AMSSwanglab/3SCover/tree/master/figure.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101227.

ACKNOWLEDGMENTS

This work was supported by Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01), CAS "Light of West China" Program (No.xbzg-zdsys-201913), National Key R&D Program of China (No. 2017YFC0908400), and the National Natural Science Foundation of China (NSFC) under Grant Nos. 11871463, 61671444, and 61621003. The computations were (partly) done on the high-performance computers of State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences. We would like to thank Dr. Qianyi Lee and lab member for discussion and the reviewers for their helpful comments.

AUTHOR CONTRIBUTIONS

Conceptualization, Q.Y. and Y.W.; Methodology, Q.Y. and Y.W.; Data analysis, Q.Y.; Writing, Q.Y. and Y.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 22, 2020
Revised: April 27, 2020
Accepted: May 28, 2020
Published: June 26, 2020

REFERENCES

Blanchet, P., Bebin, M., Bruet, S., Cooper, G.M., Thompson, M.L., Duban-Bedu, B., Gerard, B., Piton, A., Suckno, S., Deshpande, C., et al. (2017). MYT1L mutations cause intellectual disability and variable obesity by dysregulating gene expression and development of the neuroendocrine hypothalamus. PLoS Genet. *13*, e1006957.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers (vol 144, pg 327, 2011). Cell *144*, 825.

Chen, J.X., Krane, M., Deutsch, M.A., Wang, L., Rav-Acha, M., Gregoire, S., Engels, M.C., Rajarajan, K., Karra, R., Abel, E.D., et al. (2012). Inefficient reprogramming of fibroblasts into cardiomyocytes using Gata4, Mef2c. and Tbx5. Circ Res *111*, 50–55.

Columbaro, M., Mattioli, E., Maraldi, N.M., Ortolani, M., Gasparini, L., D'Apice, M.R., Postorivo, D., Nardone, A.M., Avnet, S., Cortelli, P., et al. (2013). Oct-1 recruitment to the nuclear envelope in adult-onset autosomal dominant leukodystrophy. Biochim. Biophys. Acta *1832*, 411–420.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W.D., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, 420.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M., et al. (2015). A systematic approach to identify candidate transcription factors that control cell identity. Stem Cell Reports *5*, 763–775.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987–1000.

Domazet-Loso, T., and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. Mol. Biol. Evol. *25*, 2699–2707.

Duren, Z.N., Chen, X., Jiang, R., Wang, Y., and Wong, W.H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. Proc. Natl. Acad. Sci. U S A *114*, E4914–E4923.

Guo, Z.Y., Zhang, L., Wu, Z., Chen, Y.C., Wang, F., and Chen, G. (2014). In vivo direct reprogramming of reactive Glial cells into functional neurons after brain injury and in an Alzheimer's disease model. Cell Stem Cell *14*, 188–202.

Hendry, C.E., Vanslambrouck, J.M., Ineson, J., Suhaimi, N., Takasato, M., Rae, F., and Little, M.H. (2013). Direct transcriptional reprogramming of adult cells to embryonic nephron progenitors. J. Am. Soc. Nephrol. *24*, 1424–1434.

Lakich, M.M., Diagana, T.T., North, D.L., and Whalen, R.G. (1998). MEF-2 and Oct-1 bind to two homologous promoter sequence elements and participate in the expression of a skeletal muscle-specific gene. J. Biol. Chem. *273*, 15217–15226.

Lee, S., Lee, B., Lee, J.W., and Lee, S.K. (2009). Retinoid signaling and neurogenin2 function are coupled for the specification of spinal motor neurons through a chromatin modifier CBP. Neuron *62*, 641–654.

Li, L.J., Wang, Y., Torkelson, J.L., Shankar, G., Pattison, J.M., Zhen, H.H., Fang, F.Q., Duren, Z., Xin, J.X., Gaddam, S., et al. (2019). TFAP2C-and p63-dependent networks sequentially rearrange chromatin landscapes to drive human epidermal lineage commitment. Cell Stem Cell *24*, 271.

Mall, M., Kareta, M.S., Chanda, S., Ahlenius, H., Perotti, N., Zhou, B., Grieder, S.D., Ge, X.C., Drake, S., Ang, C.E., et al. (2017). Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. Nature *544*, 245.

Marro, S., Pang, Z.P., Yang, N., Tsai, M.C., Qu, K., Chang, H.Y., Sudhof, T.C., and Wernig, M. (2011). Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. Cell Stem Cell *9*, 374–382.

Masserdotti, G., Gascon, S., and Gotz, M. (2016). Direct neuronal reprogramming: learning from and for development. Development *143*, 2494–2510.

Sandler, V.M., Lis, R., Liu, Y., Kedem, A., James, D., Elemento, O., Butler, J.M., Scandura, J.M., and Rafii, S. (2014). Reprogramming human endothelial cells to haematopoietic cells requires vascular induction. Nature *511*, 312–318.

Suo, S.B., Zhu, Q., Saadatpour, A., Fei, L.J., Guo, G.J., and Yuan, G.C. (2018). Revealing the critical regulators of cell identity in the mouse cell Atlas. Cell Rep. *25*, 1436.

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

Wang, Y., Zhang, X.S., and Xia, Y. (2009). Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. Nucleic Acids Res. *37*, 5943–5958.

Xu, J., Du, Y., and Deng, H. (2015). Direct lineage reprogramming: strategies, mechanisms, and applications. Cell Stem Cell *16*, 119–134.

Yamanaka, S. (2012). Induced pluripotent stem cells: past, present, and future. Cell Stem Cell *10*, 678–684.

Zhang, L.R., Xue, G.G., Liu, J.J., Li, Q.Z., and Wang, Y. (2018). Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data. BMC Genomics *19*, 914.

Yang, R., Zheng, Y., Li, L., Liu, S., Burrows, M., Wei, Z., Nace, A., Herlyn, M., Cui, R., Guo, W., et al. (2014). Direct conversion of mouse and human fibroblasts to functional melanocytes by defined factors. Nat Commun *5*, 5807.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. *47*, D721–D728.

Zhuang, Q., Li, W., Benda, C., Huang, Z., Ahmed, T., Liu, P., Guo, X., Ibañez, D.P., Luo, Z., Zhang, M., et al. (2018). NCoR/SMRT co-repressors cooperate with c-MYC to create an epigenetic barrier to somatic cell reprogramming. Nat. Cell Biol. *20*, 400–412.

iScience, Volume *23*

Supplemental Information

3Scover: Identifying Safeguard TF from Cell Type-TF Specificity Network by an Extended Minimum Set Cover Model

Qiuyue Yuan and Yong Wang

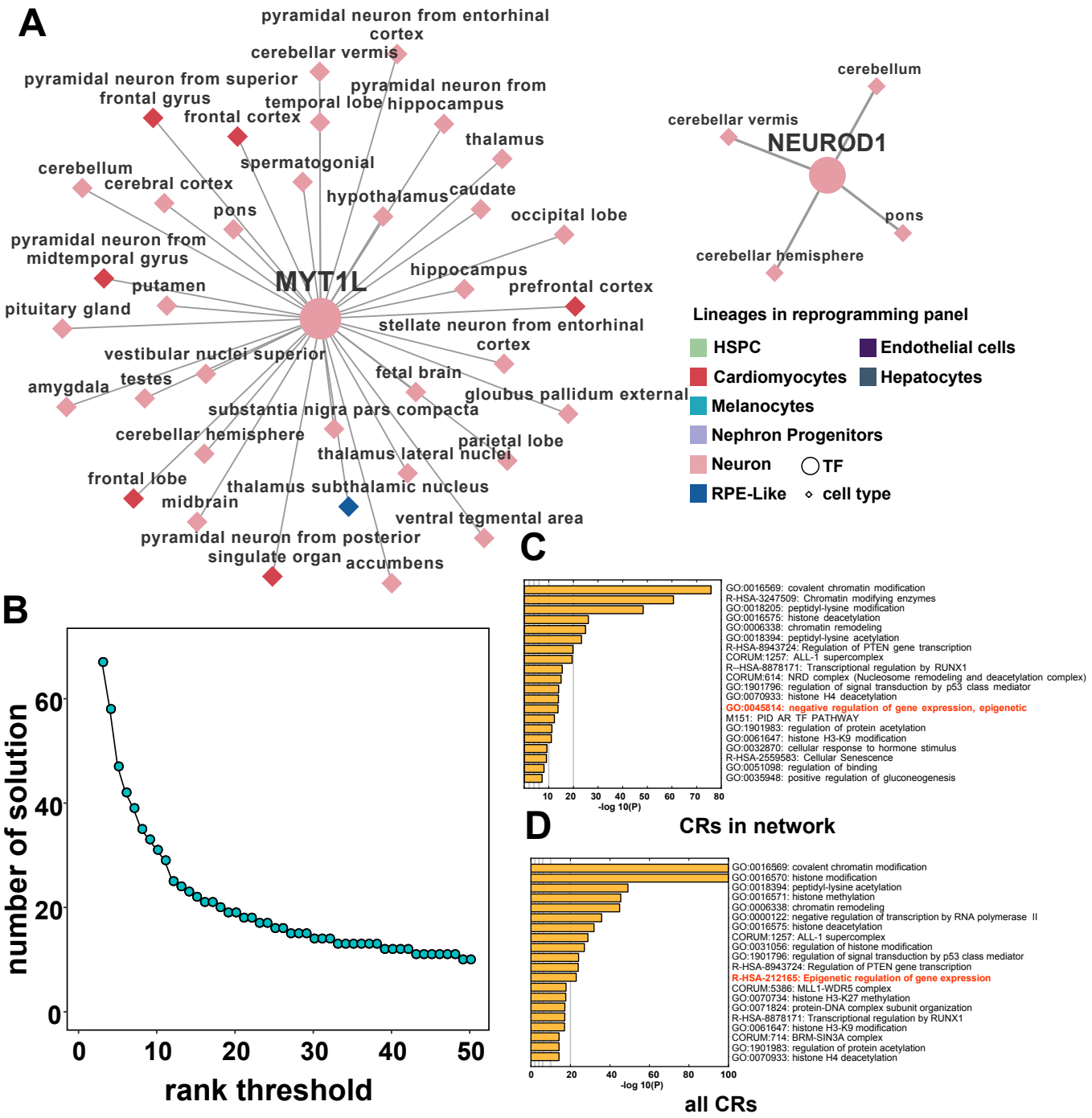


Figure S1. MYT1L and NEUROD1 induced cell type-TF subnetworks. Related to Figure 1. (A) MYT1L and NEUROD1 induced subnetwork of human cell type-TF specificity network. (B) The number of TFs in minimum set cover based on different rank thresholds. The number remains unchanged when the rank threshold m >30. (C) Function enrichment result of CRs in network Figure 4C. (D) Function enrichment of all CRs.

0

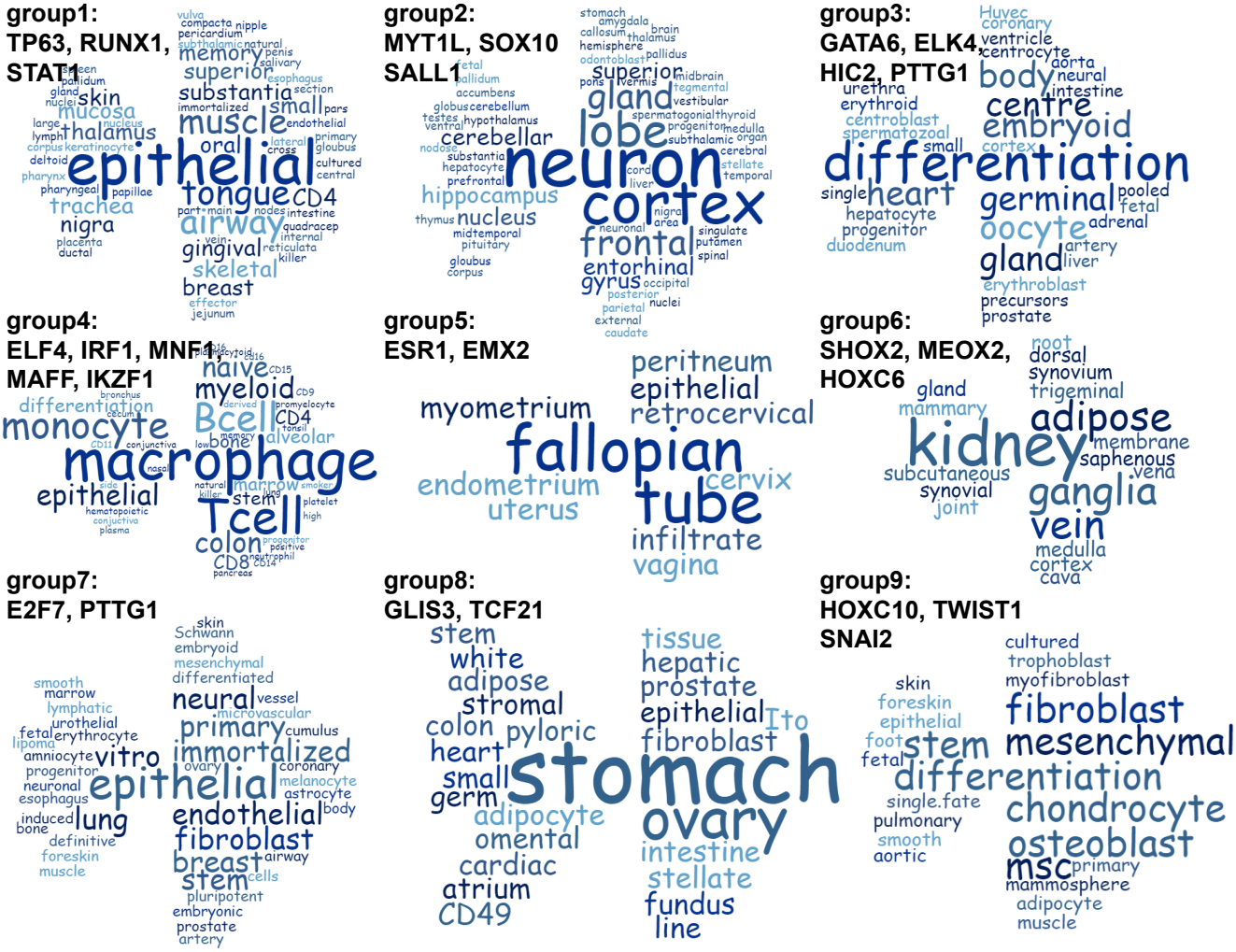


Figure S2. Detailed information for the heatmap in Figure 3A. Related to Figure 3. Word cloud is used (https://www.wordclouds.com/) to show the major content of each group of cell types in the heatmap, where a larger word represents for a higher frequency.

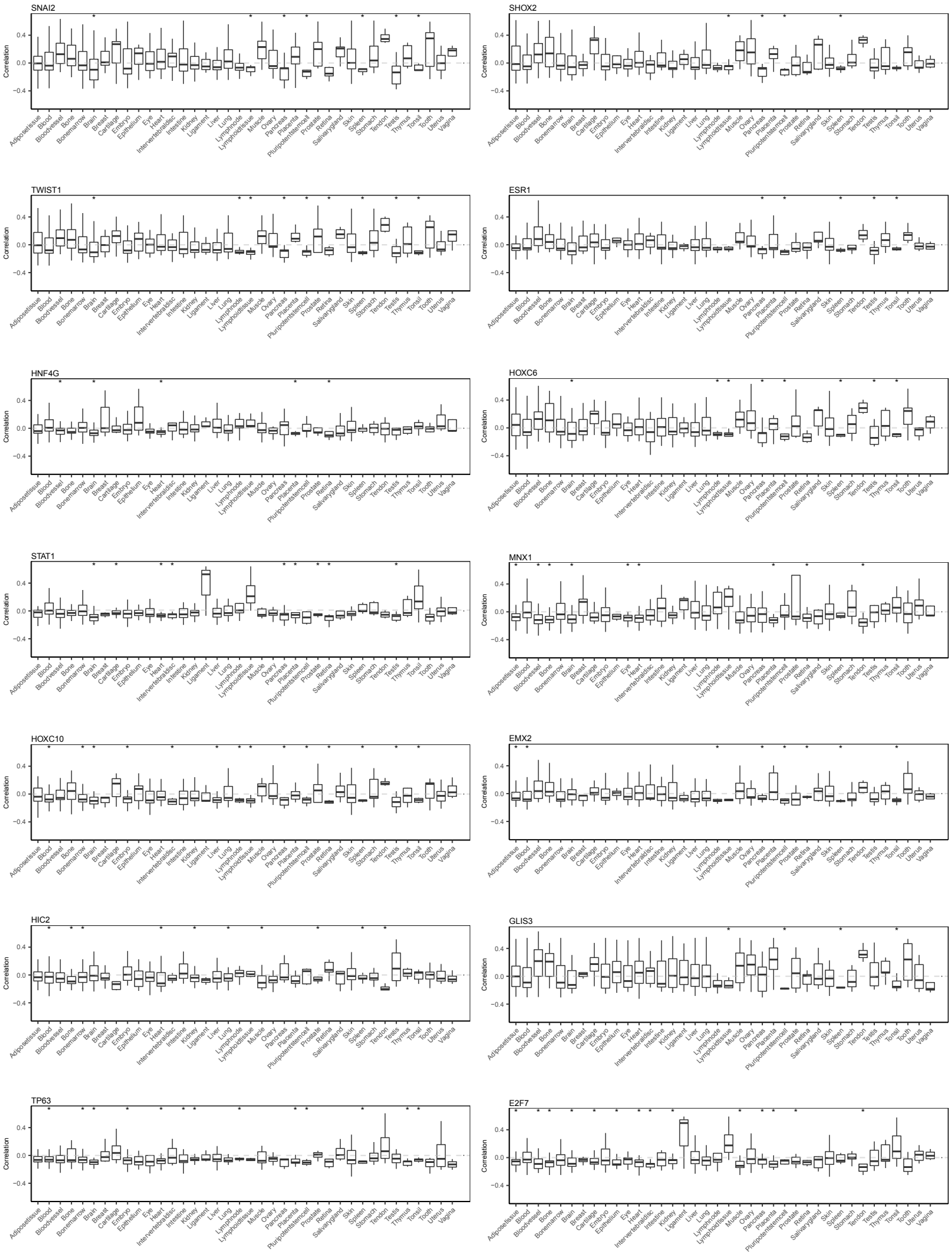


Figure S3. Correlation between safeguard TF and cell type marker genes. Related to Figure 5. Mean correlation between safeguard TFs and cell type makers. For each cell type, we get the Pearson correlation coefficients based on median TPM of safeguard TF and cell type marker genes across all cell types in GTEx datasets. The cell type marker genes are from CellMarker database. More details are in Figure 5D.

Transparent Methods

Public data collection

To observe the distinct expression patterns across all cell types of MYT1L and NEUROD1, we collected gene medium TPMs across 53 tissues from Genotype-Tissue Expression (GTEx) portal (Carithers et al., 2014). GTEx data resource was also used to provide the independent evidence for the ability of safeguard TF determining distinctive signatures of lineages. Protein-protein interaction repository were downloaded from BIOGRID (https://thebiogrid.org/). Functional enrichment analysis is given by Metascape (http://metascape.org). To predict the safeguard TF in muscle, we used single-cell RNA-seq at multiple time points in direct reprogramming from fibroblast to neuron, because some fibroblast cells transform to muscle in this process (Treutlein et al., 2016).

Specificity score

The method of calculating specificity scores is presented by D'Alessio, A.C., et al. (D'Alessio et al., 2015). The specificity score for a certain TF in the ith cell type is a $-log_{10}(JSD)$ and JSD is defined as follows:

$$JSD = \frac{1}{2}\left(\sum_{k=1}^{N} x_k log \frac{2x_k}{x_k + y_k} + \sum_{k=1}^{N} y_k log \frac{2y_k}{x_k + y_k}\right)$$

x_k represents the normalized expression of the TF in the kth cell type and $x_k = \frac{e_k}{\sum_{k=1}^{N} e_k}$, where e_k represents the expression of the TF in the k-th cell type.

y is the background vector and $y_k = \begin{cases} 0, k \neq i \\ 1, k = i \end{cases}$.

The score can be simplified as:

$$JSD = \frac{1}{2}\left(\sum_{k \neq i} x_k log\ 2 + log \frac{2x_i}{x_i + 1} + log \frac{2}{x_i + 1}\right)$$

Classifying cell types to reprogramming panel lineage

After extracting the reprogramming panel TF induced subnetwork from cell type-TF specificity network, we classify TFs into 8 groups based on the lineage to which reprograming panel TFs are used to reprogram (reprogramming panel lineages). We also classify cell types to 8 groups. For a certain cell type, we count the number of TFs in each group linked to this cell type. Then, we classify the cell type to the group with the largest counting number.

Motivation to build the 3Scover model

We explain the main idea by set covering and ensemble strategy respectively to extract safeguard TFs from the constructed cell type-TF network. Let's first consider the set cover problem without subsampling:

$$\min_{\mathbf{x}} \quad \sum_{i=1}^{K} x_i$$

$$s.t. \quad \sum_{i=1}^{K} a_{ij} x_i \geq 1,\ j \in \{1,2,\ldots,N\}$$

$$x_i \in \{0,1\},\ i \in \{1,2,\ldots,K\}$$

Where, *K* and *N* are respectively the number of TFs and cell types in the network; $x_i = 1$, if the *i-th* TF is a safeguard TF and $x_i = 0$, if not; $A = (a_{ij})_{K\times N}$ is the adjacency matrix of cell type-TF network and $a_{ij} = 1$, if the *i-th* TF is linked with the *j-th* cell type; $\min_x \sum_{i=1}^{K} x_i$ aims to minimize the number of identified safeguard TFs; $\sum_{i=1}^{K} a_{ij} x_i \geq 1, j \in \{1,2,\ldots,N\}$ requires at least one safeguard TF linked with each cell type.

This is a 0-1 integer linear programming, or more precisely, a set cover problem. This is a classical model in combinatorics, computer science, operations research, and complexity theory and known as one of Karp's 21 NP-problem. In this set cover problem, the set of cell types is the universe; each TF represent for a set including cell types linked with the TF; the safeguard TF list the minimum set cover, i.e., we aim to minimize the number of safeguard TFs to cover every cell type in the universe. We note that the first characteristic, "many but one specificity", is modeled explicitly in the objective function. However, in a general sense, the probability of a TF to be safeguard TF increases along with the increased numbers cell types covered by this TF, suggesting that the model tends to identify a TF of "many but one specificity" as a safeguard TF.

The set cover model is deterministic and has some limitations: 1) There may be more than one minimum set cover. 2) Minimum set cover is easily influenced by the noise in the data. To overcome these difficulties, we introduce the ensemble strategy. In statistics and machine learning, ensemble strategies use multiple learning algorithms (set cover problem in our case) to obtain better predictive performance than could be obtained from any of the constituent (individual strategies) learning algorithms. For example, in reconstructing gene regulatory network, the ensemble method obtained the most consistent network structure with respect to all of the datasets, thereby not only significantly alleviating the problem of dimensionality but also remarkably improving the prediction reliability (Wang et al., 2006). This motivates us to introduce the ensemble strategy to solve a multitude of set cover subsampling models for a more stable solution.

3Scover model

We propose Set cover problem with Stability Selection model (3Scover) as follows:

$$\min_{\mathbf{x},\mathbf{x}^l} \quad \sum_{l=1}^{L}\left(\lambda\left\|\mathbf{x}^l - \mathbf{x}\right\|_2 + \sum_{i=1}^{K} x_i^l\right)$$

$$s.t. \quad \sum_i a_{ij} x_i^l \geq 1,\ \text{for } l \in \{1,2,\ldots,L\},\ j \in T_l \qquad (1)$$

$$x_i,\ x_i^l \in \{0,1\},\ i \in \{1,2,\ldots,K\},\ l \in \{1,2,\ldots,L\}$$

Where *L* is the number of subsampling times; *K* and *N* are respectively the number of TFs and

cell types in the network; $\boldsymbol{x}^l$ represents the solution for *l-th* set cover sub-problem by subsampling; $\boldsymbol{x}$ is the consistent or stable solution for all the subproblems; There are two terms in the objective function. $\sum_{i=1}^{K} x_i^l$ finds the minimum set cover for the *l-th* sub-problem; and $\sum_{l=1}^{L} \|\boldsymbol{x}^l - \boldsymbol{x}\|$ minimizes distance between the consistent solution with subsampling solutions; We introduce the constraint $\sum_i a_{ij} x_i^l \geq 1, \text{for } l \in \{1,2,\dots,L\}, \text{for } j \in T_l$ to restrict the *l-th* solution to cover all cell types in T_l. T_l is the set of remaining cell types at the *l-th* randomly subsampling;

3Scover quantifies the two characteristics of safeguard TF based on the cell type-TF specificity network. "Many but one specificity" can be measured by the degree of TF and the safeguard TF set satisfying parsimony must be a set of TFs with minimal size whose induced subnetwork including all cell types. In addition, 3Scover uses "stability selection" method based on subsampling cell types to get TFs which most possibly be safeguard TF. It suggests that if a TF is of essential importance, it should be included in the minimum set cover even though we remove some cell types' information from our data (Guimaraes et al., 2006). By randomly removing, or subsampling in other words, a fixed number of cell types from the network, solving the problem, gaining the minimum TF sets, and integrating them, we can find the stable safeguard TFs which are seldom influenced by the absence of several cell types (or data noise) and non-uniqueness of the solution.

3Scover overcomes the impact of noise in data by subsampling, removes non-uniqueness of set cover problem by synthesizing multiple solutions, finally outputs a robust safeguard TF set. Importantly, 3Scover takes into account the cell type-TF, TF-TF, and cell type-cell type interaction. As the input of the model, cell type-TF interaction is spontaneously involved in the model. Additionally, the model could avoid closely related TFs concomitantly selected as safeguard TF (see proof in **Methods**). Compared with previous work to identify important TFs separately for every cell type, we conceive all cell types as a whole and thus could identify TFs at the systems level.

Decomposition algorithm

The objective function of 3Scover is convex but the variables are restricted to be binary. It's an extension from the classical set cover problem known as a NP-hard problem. Accurately solving (1) needs a lot of computer capacity due to a large number of variables. To overcome this difficulty, we propose a decomposition algorithm as follows.

Specifically, we decompose (1) into 2 relatively easy sub-problems:

Sub-optimization problem I:

$$\begin{aligned} \min_{x^l} \quad & \sum_{l=1}^{L} \left(\lambda \left\| \boldsymbol{x}^l - \boldsymbol{x} \right\|_2 + \sum_{i=1}^{K} x_i^l \right) \\ s.t. \quad & \sum_i a_{ij} x_i^l \geq 1,\ l \in \{1,2,\dots,L\},\ j \in T_l \\ & x_i,\ x_i^l \in \{0,1\},\ i \in \{1,2,\dots,K\},\ l \in \{1,2,\dots,L\} \end{aligned} \qquad (2)$$

It can be divided into L independent problems and the *l-th* is as follows:

$$\begin{aligned} \min_{x^l} \quad & \sum_{i=1}^{K} x_i^l \\ s.t. \quad & \sum_i a_{ij} x_i^l \geq 1,\ j \in T_l \\ & x_i,\ x_i^l \in \{0,1\},\ i \in \{1,2,\ldots,K\} \end{aligned} \tag{3}$$

This is the classical set cover problem. We solve the problem by branch-and-cut algorithm with CPLEX.

Sub-optimization problem II:

$$\min_{\boldsymbol{x}} \quad \sum_{l=1}^{L} \left\| \boldsymbol{x}^l - \boldsymbol{x} \right\|_2^2 \tag{4}$$

This subproblem has a closed-form solution as the mean of the L sub-solutions
We solve the two sub-problems iteratively.

- STEP-0: Initialization. Solve *L* set cover problems, with the *l-th* problems as

  $$\begin{aligned} \min_{x^l} \quad & \sum_{i=1}^{K} x_i^l \\ s.t. \quad & \sum_i a_{ij} x_i^l \geq 1,\ j \in T_l \\ & x_i,\ x_i^l \in \{0,1\},\ \text{for } i \in \{1,2,\ldots,K\} \end{aligned} ,$$

  taking the T_l-induced subnetwork as input for the *l-th* set cover problem and get solutions $\boldsymbol{x}^{l(0)}, i \in \{1, \ldots, L\}$.

- STEP-1: Fix $\boldsymbol{x}^{l(k)}, i \in \{1, \ldots, L\}$, and solve $\min_{\boldsymbol{x}} \quad \sum_{l=1}^{L} \left\| \boldsymbol{x}^l - \boldsymbol{x} \right\|_2^2$, in which the solution is $\boldsymbol{x}^{(k+1)} = \sum_{l=1}^{L} \boldsymbol{x}^{l(k)} / L$.
- STEP-2: Fix $\boldsymbol{x}^{(k)}$, solving $x^{l(k+1)}$ by

  $$\begin{aligned} \min_{\boldsymbol{x}^{l(k+1)}} \quad & \sum_{l=1}^{L} \left(\lambda \left\| \boldsymbol{x}^{l(k+1)} - \boldsymbol{x}^{(k)} \right\|_2 + \sum_{i=1}^{K} x_i^{l(k+1)} \right) \\ s.t. \quad & \sum_i a_{ij} x_i^{l(k+1)} \geq 1,\ \text{for } l \in \{1,2,\ldots,L\},\ j \in T_l \\ & x_i,\ x_i^{l(k+1)} \in \{0,1\},\ \text{for } i \in \{1,2,\ldots,K\},\ l \in \{1,2,\ldots,L\} \end{aligned}$$

By choosing the parameters λ, subsampling times *L*, and subsampling size $|T_l|$ as well as the number of iterations, we can gain the stable solution $\boldsymbol{x}$, where x_i is an estimator of the probability of the *i-th* TF to be a safeguard TF.

Algorithm complexity analysis

The core subroutine in our algorithm is to solve a set cover problem with K elements and N sets. Set cover problem is known as one of Karp's 21 NP-complete problems. We used the CPLEX solver in MATLAB for the set cover problem by branch and bound algorithm. There are 2^K leaf nodes in a binary tree and in theory the computational complexity $O(2^K)$. In practice, the solver is very efficient due to the sparsity of cell type-TF bipartite graph. To achieve a stable solution, we solved the set cover problem for L times in the algorithm of 3SCover. For each

sub-solution, the algorithm was used to identify the safeguard TF in human and mouse where K is the number of TF and N is the number of resampling cell type. For the human case, we have K=1055, N=210, and L=1000. The total running time is 2 days and 56 minutes (0.51s for single set cover subproblem on average). For the mouse case, we have K=202, N=736, and L=1000. The total running time is 2 days and 56 minutes (227s for single set cover subproblem on average).

Proof of avoiding redundancy of 3Scover

3Scover could avoid closely related TFs concomitantly selected as safeguard TF and we explain this by reduction to absurdity. We consider two closely related TFs, TF1 and TF2, and the linked cell types of TF1 are included in the linked cell types of TF2. If both TFs are in safeguard TF set, we can construct a smaller set by removing TF1. This will generate a smaller safeguard TF set and reduce the objective function.

Conservation of safeguard TF in human and mouse

We observed that 3 of the 4 safeguard TFs conserved in human and mouse are experimentally verified reprogramming TFs. We ask the question if this 75% ratio (3/4) is different with the situation by chance in statistics. We compute the conservation p-value by a permutation based test. Specifically, we transformed 202 mouse TFs and 1,055 human TFs to the homologous gene id. Then we randomly chose 30 IDs in human and mouse for 1,000 times and calculated the number of overlapped TFs. There are 7 times in which the number of overlapped TF is bigger than 4. This suggests that the conservation of safeguard TF in human and mouse are statistically enriched with a p-value 0.007.

Correlation of safeguard TF and the tissue type markers

According to the CellMarker database (Zhang et al., 2019), we extract the markers for 70 cell types by filtering the samples by the following steps.

1. Annotated by “Cancer cell” as its cell type
2. With “Undefined” tissue type
3. With none gene symbols for cell marker
4. With “single-cell sequencing” as “markerResource”
5. With a cell name as “cancer cell” or “cancer stem cell”

The resulting sample comprises 1,596 cells from 70 cell types. To find markers for a certain cell type, we collect all markers of cells included in the cell type. Then we remove markers with <1 median TPM values in all tissues from The Genotype-Tissue Expression (GTEx) project. Cell marker are intuitively specific for cell type. Therefore, we remove cell markers identified as marker in more than 40 cell types. Finally, we filter the tissues with less than 3 markers. To the end, we get 37 cell types and the number of their cell markers ranges from 4 (Vagina) to 254 (Blood).

For a certain safeguard TF, we calculate the Pearson correlation coefficient between it and each marker of a cell type based on the GTEx expression dataset. A negative mean correlation of tissue type markers (p-value < 0.05, t-test) suggests that the safeguard TF tends to repress the cell type.

Measure safeguard TF specifically linked with lineage

We use a fold change to measure the specificity of a safeguard TF linked with a lineage. The foreground percentage of a lineage is the percentage of cell types linked with the TF among the cell types in the lineage. And the background percentage is the cell types linked with TF among all cell types. This fold change is defined by dividing the foreground by the background percentage.

Data visualization techniques

All heatmaps in Figure 1, Figure 3, and Figure 6 are performed by pheatmap R package. All networks included in Figure 1, Figure 4, and Figure S1 are performed by Cytoscape (Shannon et al., 2003). Sankey diagrams of Figure 1F and Figure 1G are performed by ggalluvial R package. Functional enrichment is implemented by Metascape (http://metascape.org). Histogram, boxplot, and line graph in Figure 4, Figure 5, Figure 6, Figure 7, Figure S1 and Figure S3 are performed by ggplot2 R package. Venn diagrams in Figure 4 and Figure 6 are implemented by VennDiagram R package. Word cloud diagrams are performed by a online tool (https://www.wordclouds.com/). All related code is available at https://github.com/AMSSwanglab/3SCover/tree/master/figure.

Supplemental References

Carithers, L., Rao, A., and Moore, H. (2014). Acquisition of biospecimens to support the Genotype-Tissue Expression (GTEx) project. Cancer Res *74*.

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

Guimaraes, K.S., Jothi, R., Zotenko, E., and Przytycka, T.M. (2006). Predicting domain-domain interactions using a parsimony approach. Genome Biology *7*.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

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

Wang, Y., Joshi, T., Zhang, X.S., Xu, D., and Chen, L.N. (2006). Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics *22*, 2413-2420.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M.*, et al.* (2019). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res *47*, D721-D728.