Molecular Therapy
Nucleic Acids
Original Article

Deciphering the decisive factors driving fate bifurcations in somatic cell reprogramming

Chunshen Long,[1,3] Hanshuang Li,[1,3] Pengfei Liang,[1] Lemuge Chao,[1] Yan Hong,[1] Junping Zhang,[1,2] Qilemuge Xi,[1] and Yongchun Zuo[1]

[1]State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, School of Life Sciences, Inner Mongolia University, Hohhot 010070, China; [2]School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

Single-cell studies have demonstrated that somatic cell reprogramming is a continuous process of cell fates transition. Only partial reprogramming intermediates can overcome the molecular bottlenecks to acquire pluripotency. To decipher the underlying decisive factors driving cell fate, we identified induced pluripotent stem cells or stromal-like cells (iPSCs/SLCs) and iPSCs or trophoblast-like cells (iPSCs/TLCs) fate bifurcations by reconstructing cellular trajectory. The mesenchymal-epithelial transition and the activation of pluripotency networks are the main molecular series in successful reprogramming. Correspondingly, intermediates diverge into SLCs accompanied by the inhibition of cell cycle genes and the activation of extracellular matrix genes, whereas the TLCs fate is characterized by the up-regulation of placenta development genes. Combining putative gene regulatory networks, seven (*Taf7*, *Ezh2*, *Klf2*, etc.) and three key factors (*Cdc5l*, *Klf4*, and *Nanog*) were individually identified as drivers of the successful reprogramming by triggering downstream pluripotent networks during iPSCs/SLCs and iPSCs/TLCs fate bifurcation. Conversely, 11 factors (*Cebpb*, *Sox4*, *Junb*, etc.) and four factors (*Gata2*, *Jund*, *Ctnnb1*, etc.) drive SLCs fate and TLCs fate, respectively. Our study sheds new light on the understanding of decisive factors driving cell fate, which is helpful for improving reprogramming efficiency through manipulating cell fates to avoid alternative fates.

INTRODUCTION

Yamanaka factors *Oct4*, *Sox2*, *Klf4,* and *cMyc* (OSKM) which reprogrammed somatic cells into induced pluripotent stem cells (iPSCs) was a milestone breakthrough in the field of cell biology. It holds great promise for applications in cell transplantation therapy, disease modeling, and drug screening.[1,2] Over the past decade, advancements in high-throughput sequencing technologies have deepened our understanding of this process.[3]

In general, somatic cell reprogramming involves the down-regulation of somatic programs and activation of pluripotency networks. This process is accompanied by a series of necessary molecular events, such as mesenchyme-epithelial transition (MET),[4,5] and resetting of global histone and DNA methylation patterns.[6,7] The discovery of induced pluripotency has further proved that transcription factors (TFs) are key determinants of cell fate, which can induce the expression of genes necessary for obtaining a pluripotent state.[8–12] Some researchers have tried to replace or improve the classical OSKM reprogramming system, including chemical cocktails[13,14] and the seven-factor system (*Jdp2*, *Jhdm1b*, *Mkk6*, *Glis1*, *Nanog*, *Essrb*, and *Sall4*).[8] The success of somatic cell reprogramming alternatives and optimization cocktails have proven that the identification and application of key factors can further overcome the reprogramming obstacles and improve reprogramming efficiency.[15,16] Analysis based on single-cell data indicated that only a small fraction of intermediate-state cells can successfully overcome molecular barriers to be reprogrammed into iPSCs. Many cells stagnate on this path or deviate from it to other cell states. These alternative fates include stromal-like cells in the early stage of reprogramming, trophectoderm-, extraembryonic endoderm-, and neural-like cells in the late stage of reprogramming.[17–19] Some studies have reconstructed the continuous trajectory of cell fate transformation during somatic cell programming through densely sampled single-cell sequencing data and identified some key bifurcations of cell fate.[20,21] However, the key cell fate decision events and their detailed molecular mechanisms in somatic reprogramming remain largely unknown.

In this study, we combined relevant algorithms to construct the cell trajectory in the OSKM-induced secondary reprogramming system, and identified two key fate decision events in this process, which led to the successful reprogramming of intermediate-state cells or the alternative fates of stromal-like cells (SLCs) and trophoblast-like cells (TLCs). By dissecting the molecular processes, we deciphered the decisive factors that can drive pluripotency and alternative fate. This study has improved our understanding for the molecular mechanism of cell fate decisions in the reprogramming process.

Received 6 July 2023; accepted 28 September 2023;
https://doi.org/10.1016/j.omtn.2023.102044.

[3]These authors contributed equally

Correspondence: Qilemuge Xi, State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, School of Life Sciences, Inner Mongolia University, Hohhot 010070, China.
E-mail: qlmgxi@imu.edu.cn
Correspondence: Yongchun Zuo, State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Institutes of Biomedical Sciences, School of Life Sciences, Inner Mongolia University, Hohhot 010070, China.
E-mail: yczuo@imu.edu.cn

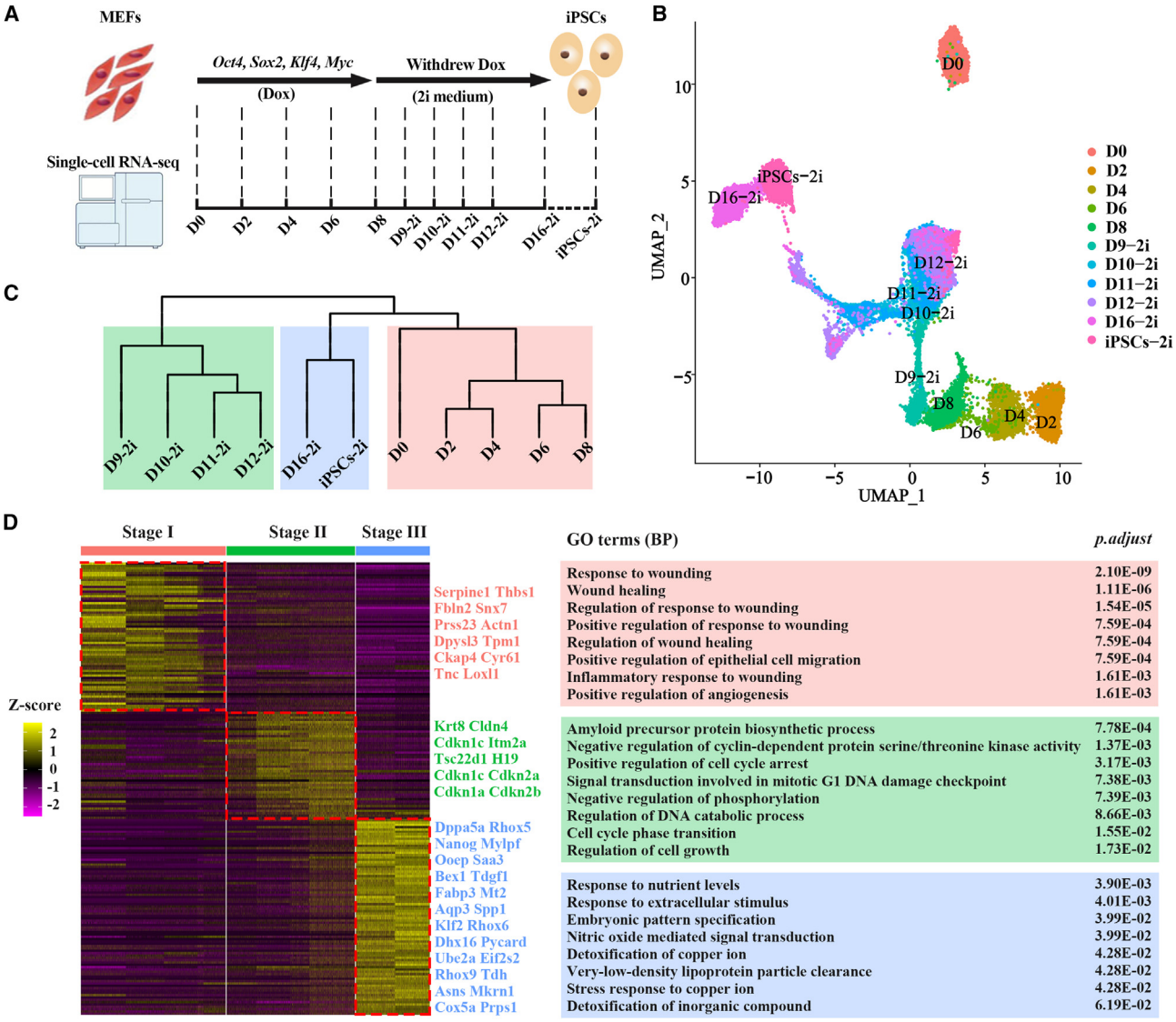

Figure 1. The global transcriptome profiles of somatic cell reprogramming

(A) Overview of single-cell mRNA-seq experimental workflow in raw data. (B) The UMAP analysis of all single cells. Cells sampled at different time points are represented by different colors. (C and D) Hierarchical clustering showing that the reprogramming process can be divided into three stages. The representative markers and enriched Gene ontology (GO) terms are listed to the right.

RESULTS

The somatic cell reprogramming can be divided into three stages

To explore the global molecular dynamics during Yamanaka factors-induced somatic cell reprogramming, we collected the single-cell RNA sequencing (scRNA-seq) dataset of mouse embryo fibroblasts (MEFs) induced into iPSCs by ectopic expression of OSKM via a secondary reprogramming system from the Gene Expression Omnibus (GEO) database.[20] The dataset covers cells intensively sampled from key time points during reprogramming, as well as individually picked Oct4-EGFP^{+} cells (iPSCs-2i, Figure 1A). After preprocessing, 23,760 high-quality single cells were left, and the cumulative number of detected genes was 16,662. Uniform Manifold Approximation and Projection (UMAP) analysis showed that D0 to D8 cells separated well from each other, and D9-2i to D12-2i cells quite mixed together. Part of the D10-2i, D11-2i, and D12-2i cells were separated from the scattered D9-2i to D12-2i mixture and moved closer to the D16-2i/iPSCs (Figure 1B). The hierarchical clustering analysis revealed that cells sampled at these time points can be divided into three stages: stage I (D0 to D8 cells), stage II (D9-2i to D12-2i cells), and stage III (D16-2i cells and iPSCs) (Figure 1C).

Next, 194 differentially expressed genes (DEGs, log2FC > 0.5 and adjusted p value <0.05) were obtained among the three stages using Seurat[22] analysis (Figure 1D). These DEGs were specifically expressed in corresponding stages. The stage I markers mainly exhibited MEF related characteristics, such as *Serpine1*, *Thbs1,* and *Fbln2*,[23,24] which were largely involved in the regulation of "wound healing." The stage II cells exhibited epithelial (*Krt8* and *Cldn4*),[4,25] trophoblast (*Krt18*, *Cdkn1c*, *Hspb1*, *Phlda2*, *Tmem37*, and *Car2*)[26] and embryonic mesenchyme (*Cdkn1c*, *Itm2a*, *Tsc22d1*, and *H19*)[26] signatures, and the enriched biological processes were mainly associated with cell cycle. The stage III cells were characterized by pluripotency markers (e.g., *Dppa5a*, *Nanog*, and *Klf2*), and the embryonic pattern specification was the representative enriched gene ontology (GO) term.

Two fate bifurcations were detected during reprogramming

Our understanding of the potential molecular mechanism that controls cell fate decisions during reprogramming is limited because the intermediates toward different fates are difficult to distinguish. Hence, we reconstructed a cellular trajectory in which cells were arranged in pseudotime according to the similarity of gene expression patterns[27] (Figure 2A). Notably, two critical cell fate decision points were detected, which began to appear at D8 and D10-2i, respectively (Figure S1). The cellular trajectory can be divided into five branches, namely pre-branch 1 (PB1), failed branch 1 (FB1), pre-branch 2 (PB2), failed branch 2 (FB2), and successful branch (SB) (Figure 2B). PB1 was mainly composed of D0 to D6 cells, and the majority of cells in SB were derived from D16-2i cells and iPSCs-2i. FB1, PB2, and FB2 were part of D8 to D12-2i cells, illustrating stage II was a critical stage of cell fate decision (Figures 1C and 2C). Given that Dppa5a is an indicator of successful reprogramming,[21] we checked the proportion of Dppa5a$^+$ cells in each branch. As expected, Dppa5a$^+$ cells sequentially increased in PB1, FB1, PB2, FB2, and SB branches (Figure 2D). Additionally, only the cell cycle activity of SB cells increased significantly compared with PB1 cells. The proportion of FB1, PB2, and FB2 cells in the G1 phase increased to varying degrees, and the proportion of cells in the S/G2/M phase decreased (Figure 2E).

To quantitatively assess the differentiation potential changes of the cells along the FB1, FB2, and SB branch, we introduced the SLICE algorithm.[28] The SLICE utilizes scRNA-seq to quantitatively measure cell differentiation potential based on single-cell entropy (scEntropy), and a high scEntropy corresponds to a high differentiation potential. The analyses revealed that FB2 and SB branch cells obtained gradually increased differentiation potential along the cellular trajectory, consistent with the progressive activation of the pluripotency program (Figure 2F), while the differentiation potential of FB1 branch cells does not fluctuate much.

The reported markers also reflected the distinct molecular mechanisms among different branches.[29] Specifically, fibroblast markers (*Prrx1*, *Thy1*, *Col1a2*, etc.) were expressed in PB1 and FB1 cells (Figure 2G). Epithelial markers had higher expression levels in PB2 and FB2 cells, and some of them (*Cldn3*, *Cldn4*, and *Epcam*) also had higher expression levels in SB cells. Embryonic development markers (*Msx2* and *Gata2*) were specifically expressed in FB2 cells. Early pluripotent and late pluripotent genes were specifically expressed in SB cells as expected.

These findings have connected the cell fate branches with differential gene expression patterns during somatic cell reprogramming, providing new clues for further exploration of the key factors driving cell fate.

Pluripotency and alternative cell fates in somatic cell reprogramming

The reprogramming intermediates and fate terminal cells often do not belong to a certain specific cell type.[3,20] The accurate description of the cell state transition is crucial to the analysis of the molecular mechanism of the reprogramming process. We mentioned above that Oct4-EGFP$^+$ cells (iPSCs-2i) and most of D16-2i cells were concentrated at the end of the SB branch, implying that the SB branch was a successfully reprogramming trajectory (Figure 2B). However, the other two alternative fates still need to be further identified.

To this end, we identified 33, 34, and 98 marker genes (Table S1) in FB1 terminal cells (FB1TCs), FB2 terminal cells (FB2TCs), and SB terminal cells (SBTCs), respectively. The expression levels of FB1TC and SBTC markers gradually increased along their respective branches, and FB2TC markers also activated in FB2TCs as expected (Figures 3A–3C). Stromal, trophoblast, and pluripotency signatures accounted for the highest proportion in markers of FB1TCs, FB2TCs, and SBTCs, respectively. Thus, FB1TCs, FB2TCs, and SBTCs were defined as SLCs, TLCs, and iPSCs, respectively.

The UMAP showed that D0 MEFs and terminal cells of the three fates were well separated from each other (Figure 3D). As expected, fibroblast genes (e.g., *Thy1*, *Sdpr*, and *Tagln2*), stromal cell genes (e.g., *1500015O10Rik*, *Serpine2*, and *Ephx1*), trophoblast genes (e.g., *Krt18* and *Krt8*) and pluripotency genes (e.g., *Sox2*, *Nanog* and *Dppa5a*) were specifically expressed in MEFs, SLCs, TLCs, and iPSCs, respectively (Figure S2). Correspondingly, D0 markers were involved in typical fibroblast characteristics, such as response to wounding, extracellular matrix organization, and collagen fibril organization (Figure 3E). SLC markers were enriched in several processes related to Schwann cell migration, regulation of hemopoiesis, and epithelial cell proliferation. TLC markers were mainly enriched in some processes related to placenta development, response to wounding, and epithelial cell migration. The iPSC markers were enriched in some processes closely related to pluripotency, such as DNA methylation, gastrulation, regulation of cell cycle, and embryonic pattern specification.

Further, we found SLCs and iPSCs contained finer subpopulations (Figure S3). SLCs can be divided into two subclusters (Figure S3A). Subcluster 1 mainly contains partial D8 to D12-2i cells. Subcluster 2 contains partial D8 to D12-2i cells, D16-2i cells, and iPSCs-2i cells (Figure S3B). Eighteen DEGs could clearly specify the two SLC subclusters. For example, *Ly6a* and *Fth1* have higher expression levels

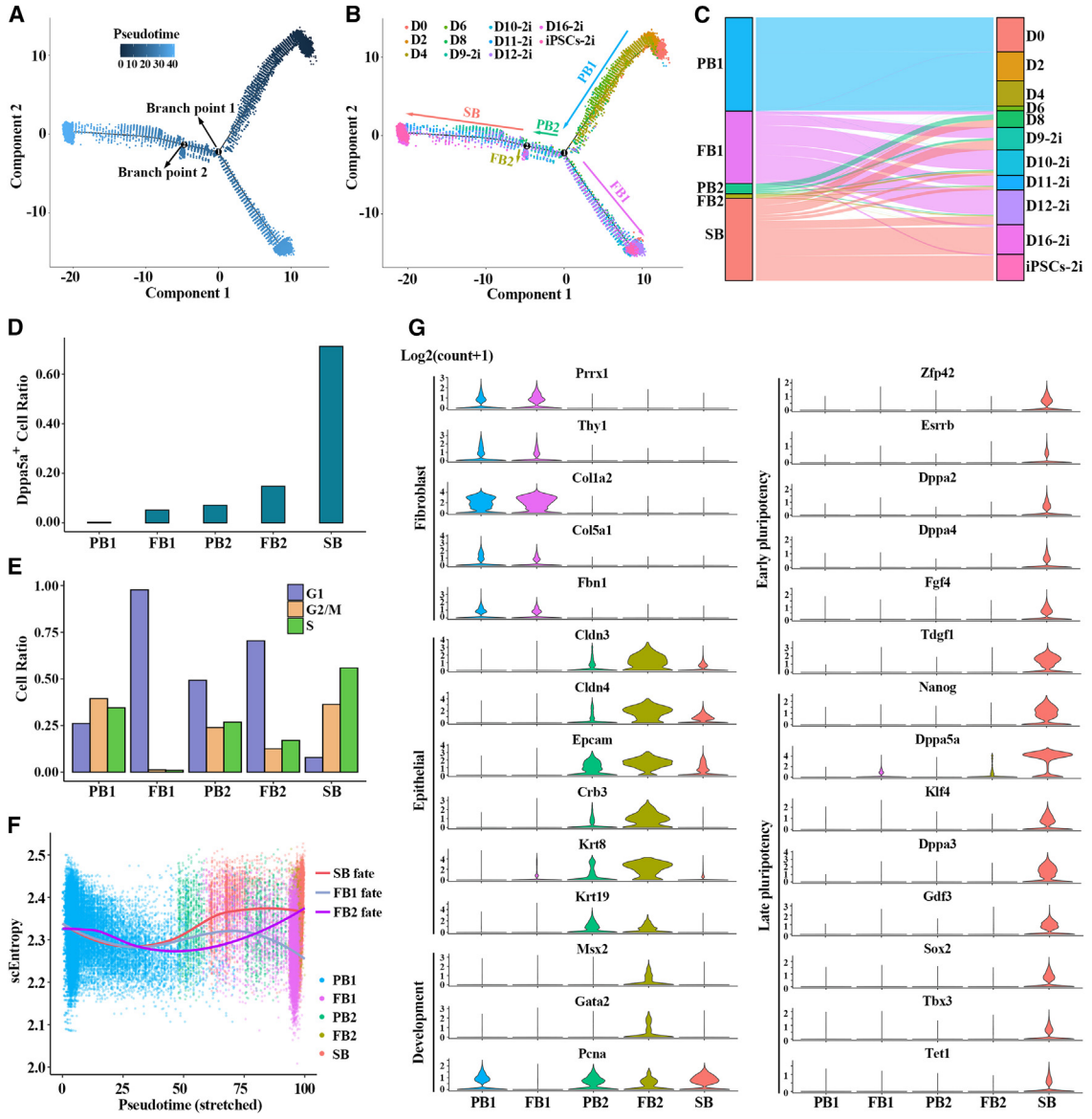

Figure 2. Identification of critical cell fate bifurcations during somatic cell reprogramming

(A) Pseudotime analysis was performed in the reprogramming progress based on monocle2. (B) The different branches in the cellular trajectory. PB1, pre-branch 1; FB1, failed branch 1; PB2, pre-branch 2; FB2, failed branch 2; SB, successful branch. (C) The Sankey diagram showing the distribution of the cells in different branches at each sampling time point. (D) Histogram showing the ratio of Dppa5a^{+} cells in different branches. (E) Histogram showing the cell ratio in different cell cycle stages. (F) The fitting curve represents the dynamic trend of single-cell entropy during MEFs to three different cell fates. (G) The violin diagram showing the expression pattern of representative genes in different branches.

in subcluster 1, while genes such as *Col3a1* and *Bgn* have higher expression levels in subcluster 2 (Figure S3C). The iPSCs can be divided into three subclusters, and subclusters 1 to 3 gradually approach the end of iPSC fate in the cellular trajectory (Figure S3D). Subcluster 1 mainly included D10-2i to D12-2i cells. Subclusters 2 and 3 mainly included D16-2i and iPSCs-2i cells (Figure S3E), respectively. Twenty-four DEGs were identified among the three subclusters (Figure S3F). It is reported that sustained co-expression of *Epcam*, *Nanog*, and *Sox2* is thought to be required to progress toward iPSCs.[30] Of note, *Epcam* was significantly down-regulated in subclusters 2 and 3, which indicates that *Epcam* may play a role in pluripotency acquisition rather than pluripotency maintenance.

Compared with iPSCs, the DEGs up-regulated in SLCs and TLCs also displayed their unique stromal-like and trophoblast-like characteristics (Figures 3F and 3G), while the DEGs up-regulated in iPSCs

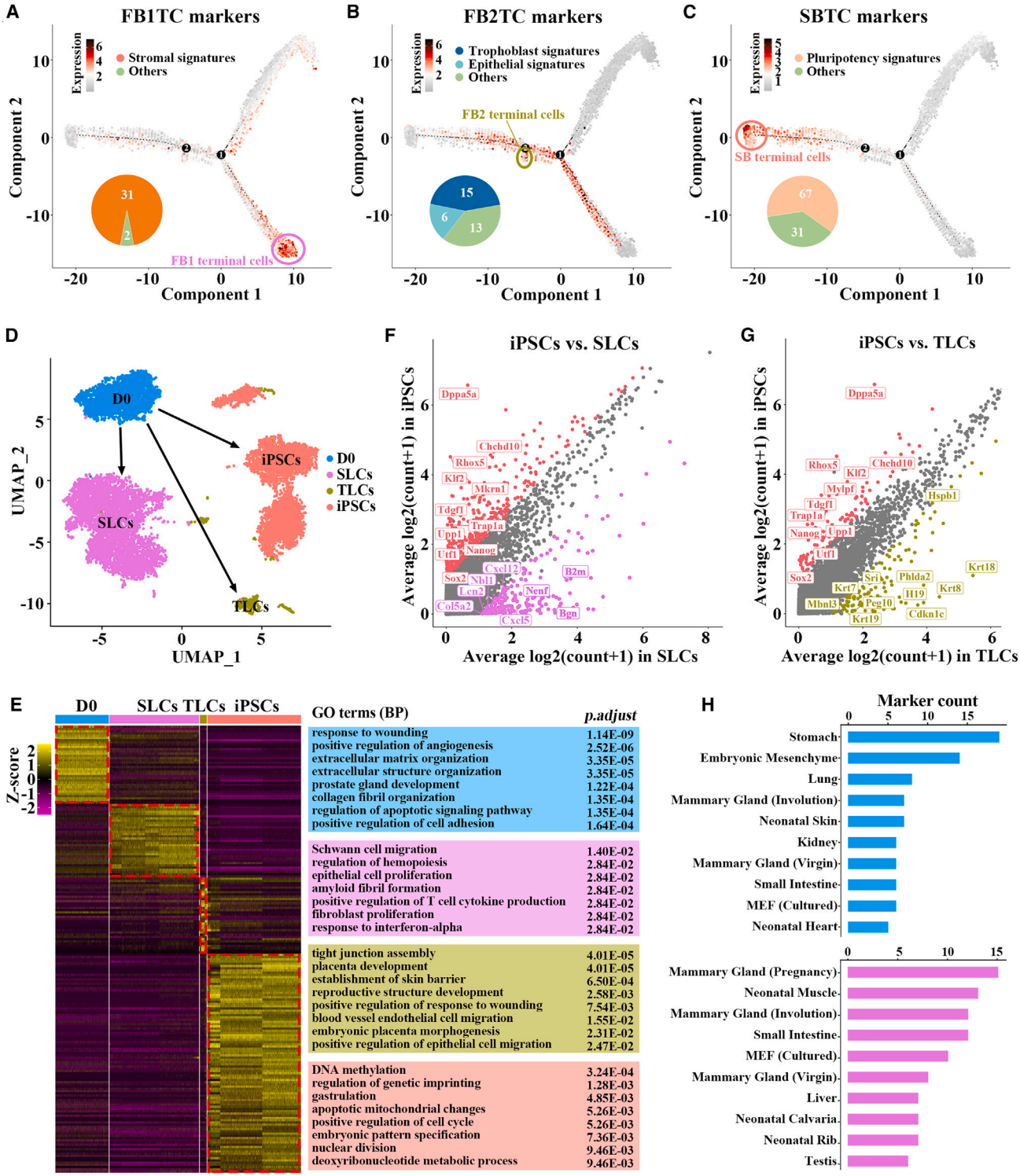

Figure 3. Characterization of branch terminal cells at different reprogramming fates

(A–C) The average expression levels of terminal cell markers are displayed on the cellular trajectory. Among them, the stromal, trophoblast, and pluripotency signatures accounted for the highest proportion, respectively. (D) UMAP analysis showing the distribution of MEFs (D0), SLCs, TLCs, and iPSCs. (E) Heatmap showing the differentially expression genes among MEFs (D0), SLCs, TLCs, and iPSCs. The representative genes and gene ontology (GO) terms were listed to the right. (F) Scatterplot showing the

(legend continued on next page)

contained representative pluripotent genes (e.g., *Dppa5a*, *Nanog,* and *Sox2*). Given that MEFs belong to stromal cells, and FB1TCs are also identified as stromal-like cells, hence, the potential molecular difference under the significant separation of MEFs and SLCs in UMAP results needs to be further clarified. We sought to annotate their molecular characteristics by comparing their shared markers with those of all tissue-specific stromal cells.[26] The results indicated that MEFs exhibited the highest similarity to stomach and embryonic mesenchyme cells, sharing 35 common markers. Meanwhile, SLCs demonstrated their closest resemblance to mammary gland and neonatal muscle cells, sharing 33 overlapping markers (Figure 3H).

Molecular roadmap among different cell fates

To investigate the temporal dynamics of DEGs between successful and failed reprogramming fates, we identified the significant branch-dependent DEGs and divided them into different clusters according to kinetic trends. A total of 2,809 DEGs were filtered in iPSCs/SLCs bifurcate and divided into three clusters (Figure 4A). The global kinetic patterns of genes in cluster 1 (n = 1483) were rapidly down-regulated in iPSCs fate and up-regulated in SLCs fate (Figures 4A and 4B). The DEGs in cluster 1 were largely involved in the regulation of stromal cells related to biological processes such as extracellular matrix organization, response to wounding, and collagen fibril organization (Figure 4C). The global dynamics of DEGs in cluster 2 (n = 205) and cluster 3 (n = 1121) were progressively up-regulated in iPSCs fate and down-regulated in SLCs fate (Figures 4A and 4B). Cluster 2 genes contained some placenta-specific TFs (*Rhox6*, *Tfap2c*, *Philda2*, and *Hic2*)[31] and were enriched in "embryonic placenta development" process (Figure 4C). Cluster 3 contains many pluripotency TFs, and these DEGs were mainly involved in biological processes such as chromosome segregation, DNA replication, and mitotic cell cycle phase transition (Figure 4C).

Notably, we found that MEF genes (n = 117) only appear in cluster 1, epithelial genes (n = 8) only appear in cluster 2, and pluripotency genes (n = 146) only appear in cluster 3. As expected, the expression patterns of these genes were consistent with the overall kinetic patterns of their respective clusters (Figure 4D). However, we found that most genes have different expression patterns only when they are close to the branch terminals of the two fate bifurcations, which cannot reflect the molecular basis near the fate bifurcations. So, 25 DEGs were identified near the iPSCs/SLCs bifurcation, of which seven genes (subset of cluster 3) were continuously expressed in iPSCs fate and immediately down-regulated in SLCs fate after iPSCs/SLCs bifurcation, respectively (Figure 4D). Functional annotation showed that these seven genes were mainly involved in cell cycle processes such as "positive regulation of mitotic cell cycle phase transition" (*Cdk1*, *Rrm2*, *Ube2c*, *Cenpa*, and *Pclaf*) and "nucleobase-containing compound biosynthetic process" (*Rrm2*, *Pclaf*, *Dut*, and *Hmgb2*), corresponding to the decrease of cell cycle activity in FB1 cells. The remaining 18 genes (*Acta2*, *Car4*, *Ccl2*, etc.) were subsets of cluster 1, which were immediately down-regulated in iPSCs fate and up-regulated in SLCs fate after iPSCs/SLCs bifurcation (Figure S4A). They were mainly involved in the biological process of regulation of leukocyte cell-cell adhesion, regulation of protein localization to membrane, and neutrophil degranulation (Figure S4B). These results indicated that the inhibition of cell cycle program and the activation of stromal program represent the main molecular series in SLCs fate. The rapid down-regulation of stromal program and up-regulation of the epithelial program, as well as sustained expression of cell cycle genes represent the early molecular series in iPSCs fate.

For the second fate bifurcation, 178 DEGs were identified toward iPSCs or TLCs fates and divided into two clusters (Figure S5). Cluster 1 (n = 86) was gradually up-regulated in iPSCs fate and down-regulated in TLCs fate, which contains 40 pluripotency genes (e.g., *Klf4*, *Nanog*, *Nr0b1*, *Trim28*, *Klf2*, *Rpp25*, and *Cdc5l*), 16 trophoblast genes (*Cacybp*, *Ube2c*, *Cenpa*, etc.) and 15 common genes of pluripotency and trophoblast (*Dtymk*, *AA467197*, *Dut*, etc.) (Table S1). The GO analysis showed cluster 1 was mainly involved in pluripotency-related biological processes, such as "positive regulation of mitotic cell cycle" and "embryonic pattern specification" (Figure S5). Cluster 2 (n = 92) was progressively down-regulated in iPSCs fate and up-regulated in TLCs fate, which included seven epithelial genes (e.g., *Cldn4*, *Cldn3*, and *Epcam*) and 28 trophoblast genes (e.g., *Mxd4*, *Gata2*, and *Phlda2*) (Table S1). Cluster 2 was mainly involved in biological processes related to trophoblast, such as "embryonic placenta development" and "cell differentiation involved in embryonic" (Figure S5).

Deciphering decisive factors driving cell fate bifurcations

Studies for the induced pluripotency and transdifferentiation have revealed that TFs are master regulators of cell fate decisions.[7,32] To explore how the key TFs determine gene expression programs and further establish cell fates during reprogramming, we used SCENIC[33] to infer gene regulatory networks (GRNs) from the scRNA-seq data. Based on default filter parameters, 439 regulons (i.e., TFs and their target genes) were identified. The hierarchical clustering for 14,048 selected cells (MEFs, SLCs, TLCs, and iPSCs) based on the binarized regulon activity showed most of the same cell types were clustered together except a small part of MEFs separated from others (Figure 5A). Meanwhile, we found that some regulons display cell type-specific activation patterns, and the motifs of representative TFs were listed next to them. Consistent with iPSCs identity, nearly all iPSCs showed high pluripotency regulons activity, such as Nanog(+), Klf4(+), Sox2(+). The other three cell types also have highly active regulons corresponding to cell identity signatures.

We then focused on cells that followed the successful reprogramming trajectory, which can be further classified into five clusters based on pseudotime order and their activity score (Figures S6A and S6B).

differentially expressed genes between iPSCs and SLCs, with some previously reported genes highlighted. (G) Scatterplot showing the differentially expressed genes between iPSCs and TLCs, with some previously reported genes highlighted. (H) Top: Histogram exhibiting the count of overlapping markers between MEFs and tissue-specific stromal cells. Bottom: Histogram exhibiting the count of overlapping markers between SLCs and tissue-specific stromal cells.

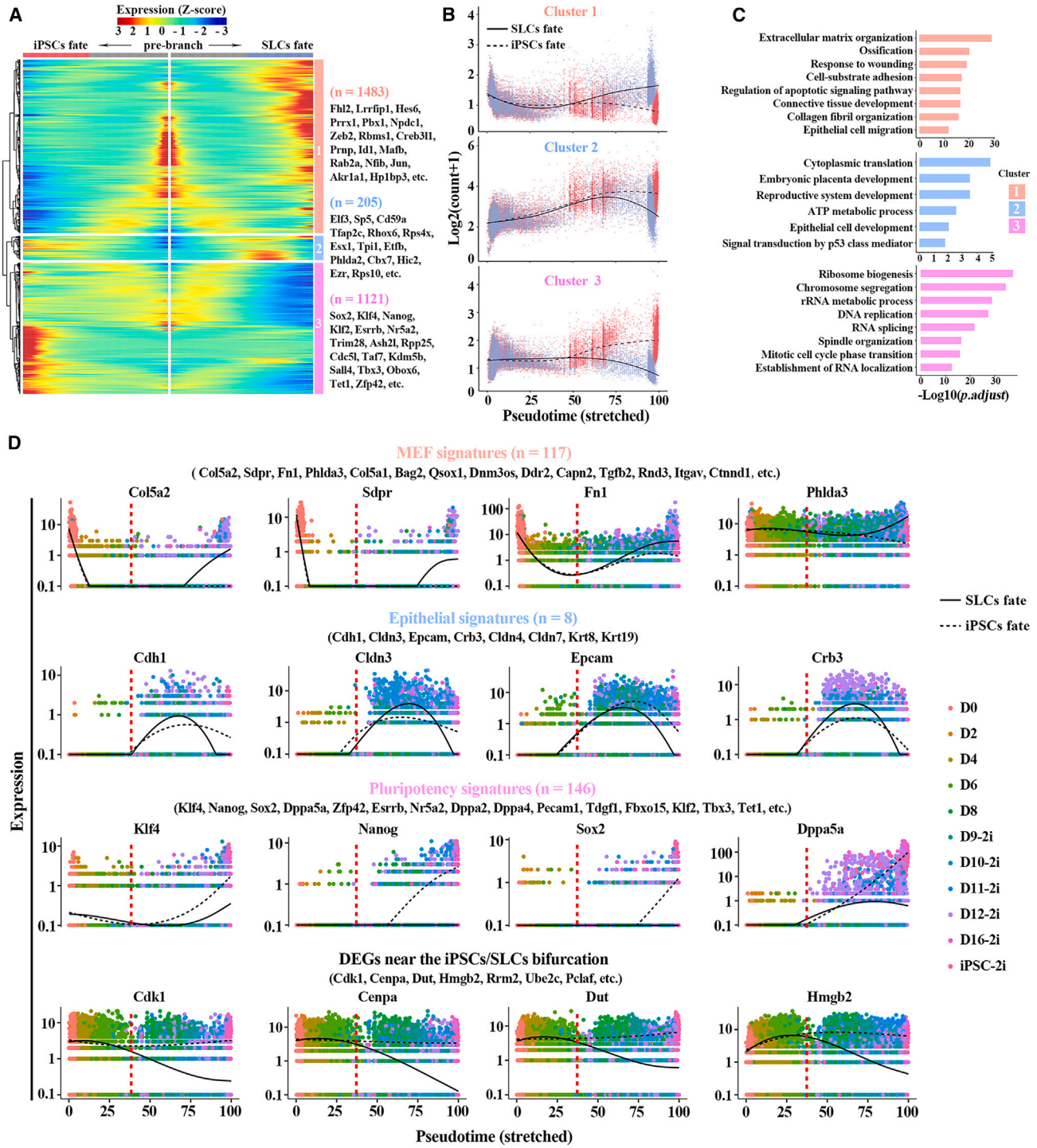

Figure 4. Dynamic trends of gene expression among different cell fates

(A) Gene expression heatmap of 2809 DEGs between the iPSCs and SLCs fates. The cellular trajectory of iPSCs and SLCs fates are displayed on the left and right, respectively. These DEGs were divided into three clusters (1, 2, 3) according to their expression patterns, and the representative DEGs are listed. (B) The kinetic trends (fitted using the Loess algorithm) of DEGs in the three clusters depicted in (A). The cell types are shown in different colors. (C) Gene ontology analyses of each gene cluster depicted in (A). (D) The expression dynamics of representative genes in different fate branches. The red dashed line represents the time point of the iPSCs/SLCs bifurcation occurring on the stretched pseudotime axis of iPSCs fate.

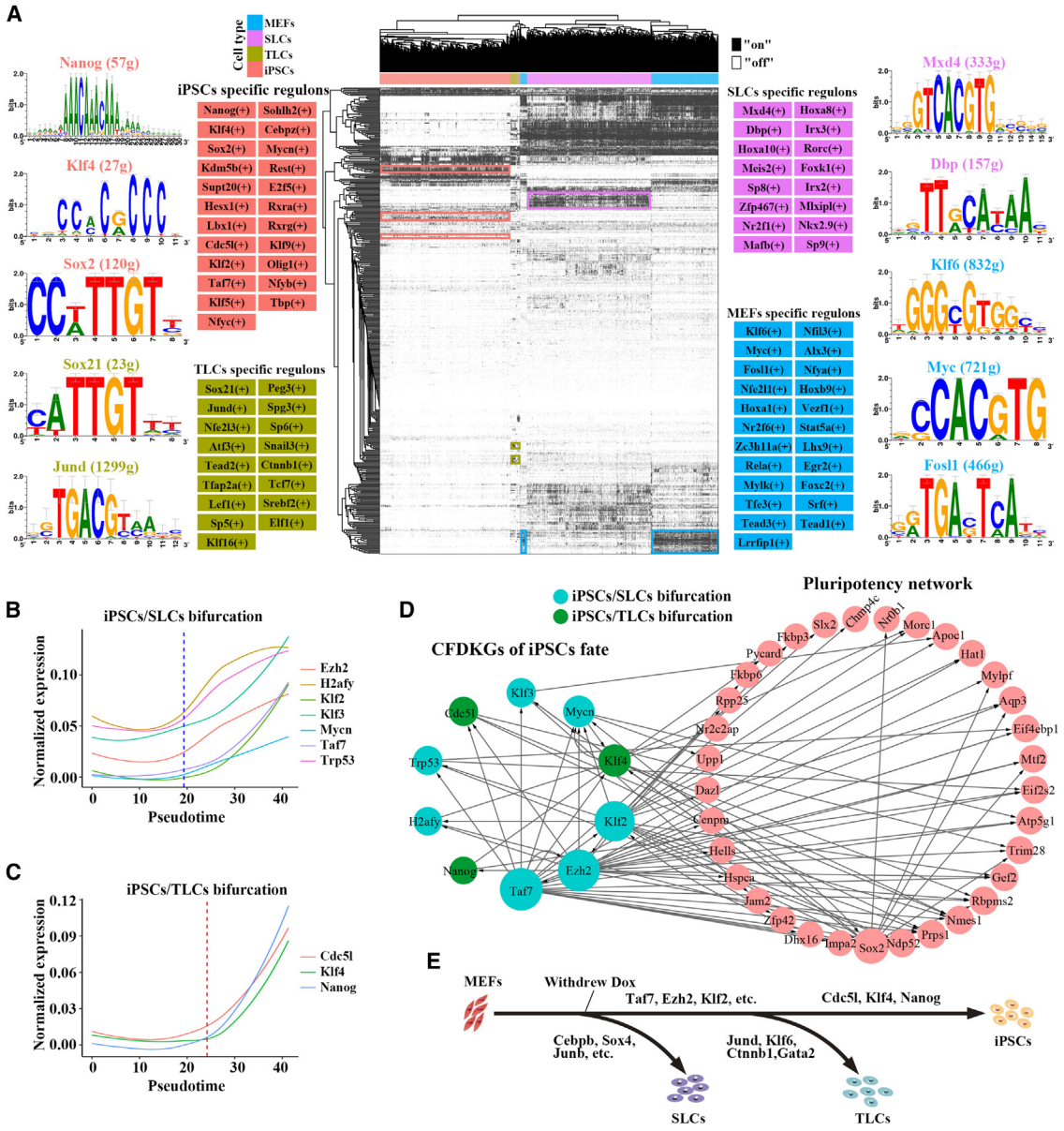

Figure 5. The decisive factors driving fate bifurcations

(A) The heatmap displayed the binarized activity scores of regulons among MEFs, SLCs, TLCs, and iPSCs. The cell-specific regulons were listed adjacent to the heatmap, with colors matching their respective cell types. The representative motifs of key TFs are also listed beside. (B) The expression patterns of cell fate decision key genes (CFDKGs) for iPSCs fate at the iPSCs/SLCs bifurcation. (C) The expression patterns of cell fate decision key genes (CFDKGs) for iPSCs fate at the iPSCs/TLCs bifurcation. (D) The putative CFDKGs of iPSCs fate can target and activate iPSCs-specific network. The larger the node, the more genes interact with in the GRN. (E) Summary model of cell fate bifurcation during reprogramming.

A total of 269 regulons showed temporal activation patterns in these five cell clusters (Figure S6C). Overall, somatic-specific regulons were gradually down-regulated (e.g., Runx1(+), Foxc2(+))[34] and pluripotency regulons were up-regulated in the five clusters sorted by pseudotime. Some regulons associated with the embryo development process were transiently activated in this process, such as regionalization, formation of primary germ layer, and embryonic morphogenesis (Figure S6D). Our findings are consistent with the reported the activation of developmental programs in parallel to iPSC formation.[3]

Finally, combining DEGs between different fates and putative GRNs, we focused on identifying cell fate decision key genes (CFDKGs). CFDKGs were defined to fulfill two conditions, i.e., exhibiting differential expression patterns between fate bifurcations, and acting as

upstream TFs to activate specific GRNs of fate terminal cells. Successfully reprogrammed cell fates undergo two critical cell fate decisions (Figure 2B). Seven (*Ezh2*, *H2afy*, *Klf2*, *Klf3*, *Mycn*, *Taf7*, and *Trp53*) and three (*Cdc5l*, *Klf4*, and *Nanog*) CFDKGs of iPSCs fate were identified at the iPSCs/SLCs and iPSCs/TLCs bifurcation, respectively (Figures 5B, 5C, and S7). Among them, Taf7 had the highest number of iPSCs-specific downstream target genes and *Nanog* had the lowest number of iPSCs-specific downstream target genes (Figure 5D). Most of the genes in the 10 CFDKGs belong to the pluripotency-related TFs, including *Mycn*, *Taf7*, *Klf2*, *Trp53*, *Cdc5l*, *Klf4*, and *Nanog*,[20] while *Ezh2*, *H2afy*, and *Klf3* were reported to be highly expressed in trophoblast progenitor cells, neutrophils, and gastric epithelial cells, respectively.[26] Here, we showed for the first time that *Ezh2*, *H2afy*, and *Klf3* have important effects on promoting somatic cell reprogramming. Similarly, we also identified their respective CFDKGs in two alternative cell fates. Eleven TFs (*Cebpb*, *Mxd4*, *Creb3*, *Fos*, *Hoxa7*, *Sox4*, *Jun*, *Hoxa9*, *Dbp*, *Junb*, and *Hoxa10*) were identified as CFDKGs of SLCs, which were up-regulated in SLCs fate branch, and as upstream TFs activating specific GRNs of SLCs (Figure S8). Among them, *Mxd4*, *Fos*, *Sox4*, *Jun*, *Dbp*, *Junb*, and *Hoxa10* have been identified as stromal cell-specific TFs in previous studies.[26] Four TFs (*Gata2*, *Jund*, *Ctnnb1*, and *Klf6*) were identified as CFDKGs of TLCs, which were up-regulated in TLCs fate branch and function as upstream TFs to activate specific GRNs of TLCs (Figure S9). Among them, *Gata2* has been identified as a trophoblast-specific TF in previous reports.[26] Collectively, we identified two key cell fate decision events in the secondary reprogramming system in this study. The first cell fate decision occurs after dox withdrawal at day 8, and the reprogramming intermediates expressing 11 key genes (*Cebpb*, *Sox4*, *Junb*, etc.) tend to SLCs fate, whereas the reprogramming intermediates expressing 7 key genes (*Taf7*, *Ezh2*, *Klf2*, etc.) tend to become precursor cells for successful reprogramming. Subsequently, reprogramming intermediates underwent a second cell fate decision, with reprogramming intermediates expressing *Cdc5l*, *Klf4*, and *Nanog* tending to successful reprogramming, and those expressing *Gata2*, *Jund*, *Ctnnb1*, and *Klf6* tending to TLCs fate (Figure 5E).

DISCUSSION

In recent years, single-cell analysis has revealed some rare but important mechanisms that have evaded us so far during somatic cell reprogramming, but it has also demonstrated the unexpected complexity of their molecular mechanisms. This study was built upon a published dataset and revealed previously unnoticed patterns and mechanisms through novel perspectives and diverse methodologies.

First, we identified the three stages during reprogramming: stage I (D0 to D8 cells), stage II (D9-2i to D12-2i cells), and stage III (D16-2i/iPSCs cells). These three stages can be characterized by MEFs, epithelial/trophoblastic/mesenchymes, and pluripotency signatures, respectively. By reconstructing the cellular trajectory, iPSCs/SLCs and iPSCs/TLCs cell fate bifurcation were detected at D8 and D10-2i, respectively. The fate that tends to be successfully reprogrammed at the iPSCs/SLCs bifurcation showed typical MET characteristics, which was characterized by down-regulated expression of MEF genes and up-regulated expression of epithelial genes.[5] Reprogramming intermediate cells diverge into SLCs at the iPSCs/SLCs bifurcation with the up-regulated expression of extracellular matrix gene and down-regulated expression of cell cycle genes. Considering that MEFs also belong to stromal cells, the differences between MEFs and SLCs are worth discussing. By comparing their shared markers with those of tissue-specific stromal cells, we found that MEFs are most similar to stomach and embryo mesenchyme-specific stromal cells, while SLCs are most similar to the gland and neural muscle-specific stromal cells. The second fate bifurcation represents the diversion of iPSCs and TLCs. The iPSCs fate was indicated by the up-regulated expression of pluripotency signatures, while the TLCs fate was characterized by the up-regulated expression of trophoblast signatures.

Combining trajectory reconstruction and GRNs inference algorithms, we putatively identified key TFs that drive different cell fates. The activation of seven key TFs (*Taf7*, *Ezh2*, *Klf2*, etc.) at the iPSCs/SLCs bifurcation and three key TFs (*Cdc5l*, *Klf4*, and *Nanog*) at the iPSCs/TLCs bifurcation greatly facilitates the successful reprogramming of intermediate-state cells. In addition to some widely reported pluripotency factors,[20] we have also identified some previously overlooked crucial regulators, such as *Ezh2*, *H2afy*, and *Cdc5l*. Notably, 269 regulons exhibited temporal activation patterns in the successful reprogramming trajectory, with the majority enriched in embryo development-related pathways, demonstrating shared regulatory mechanisms between somatic cell reprogramming and embryonic development. However, the up-regulation of 11 key TFs (*Cebpb*, *Mxd4*, *Creb3*, etc.) and four TFs (*Gata2*, *Jund*, *Ctnnb1*, and *Klf6*) drive cells toward alternative fates at the iPSCs/SLCs and iPSCs/TLCs fate bifurcations, respectively. Among them, four TFs (*Cebpb*, *Creb3*, *Hoxa7*, and *Hoxa9*) and three TFs (*Jund*, *Ctnnb1*, and *Klf6*) were the first reported regulators of stromal and trophoblast fates in this study, respectively.

Our study has illuminated critical cell fate bifurcations, depicted the molecular route along pluripotency or alternative fates, and further identified key TFs driving different fates. These findings provided a reference for experimental biologists to manipulate cell fates and improve reprogramming efficiency.

MATERIALS AND METHODS

Dataset collection

The single-cell RNA sequencing (scRNA-seq) data of mouse somatic cell reprogramming were collected from the GEO database under accession number GEO: GSE106340.[20] The sampling process for the somatic cell reprogramming included two stages. The first stage included day 0 (D0), day 2 (D2), day 4 (D4), day 6 (D6), and day 8 (D8) cells that were cultured in the serum environment (containing dox). The second stage is day 9 (D9-2i), day 10 (D10-2i), day 11 (D11-2i), day 12 (D12-2i), and day 16 (D16-2i) when dox was removed and cells were transferred to the 2i environment. The iPSCs-2i were selected with endogenous Oct4 locus (Oct4-IRES-EGFP) as a reporter.

The cells sampled at the above time points were sequenced together with iPSCs-2i.

scRNA-seq data preprocessing

For scRNA-seq data preprocessing, all raw data were controlled by FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and were trimmed based on Trimmomatic (version 0.38)[35] to remove low-quality reads.[36,37] Next, filtered reads were mapped to the UCSC mm10 genome using genome-aligner STAR (version 2.5.1b) with default parameters. Then, UMI counts were performed with cellranger from the 10X Genomics pipeline (version 1.1.0, with default parameters). All cells with fewer than 1,000 UMIs and 1,000 genes were detected per cell in total and all genes that were expressed in fewer than 50 cells were discarded. Excluding the D10-serum, D12-serum, D16-serum, and iPSCs-serum cells, 23,760 cells and 16,662 genes were retained for further analysis.

scRNA-seq dimensionality reduction, differentiation potential assessment, and trajectory inference

For the dimension reduction, principal-component analysis (PCA) was performed on the scaled expression of 2,000 highly variable genes using the RunPCA function in Seurat (version 4.3.0).[22] Following that, UMAP was implemented on the top 10 PCs via the RunUMAP functions. The SLICE[28] was used to quantitatively measure cell differentiation potential based on single-cell entropy with default parameters. Monocle2[27] was utilized for inference single cell trajectory with default parameters.

Identification of DEGs

The DEGs analysis was performed using Seurat.[22] The Seurat function "find_all_markers" (thresh.test = 1, test.use = "roc") was used to identify unique cluster-specific marker genes. For two given clusters, DEGs were identified by the find.markers function with the following parameters: thresh.use = 1, test.use = "roc." For a certain gene, the roc test generated a value ranging from 0 (for "random") to 1 (for "perfect"), representing the "classification power." Genes with a fold change ≥ 2 or ≤ 0.5 and a power ≥ 0.4 were regarded as DEGs.

Single-cell gene regulatory network inference

The workflow of pySCENIC[33] (https://pypi.org/project/pyscenic/0.6.6/#tutorial) was used to inference the GRNs. In pySCENIC workflow, RcisTarget[38] package determined TFs and their predicted target genes (i.e., targetomes) based on the correlation of gene expression across cells, and GRNBoost[39] identified whether the predicted target genes have the corresponding TF motifs to refine targetomes. The RaacFold predicted functional domain and active sites.[40–42] Finally, active targetomes were recognized in each cell. The regulatory networks driving cell fates were screened out and visualized by Cytoscape (version 3.7.0).[43]

GO enrichment and statistical analysis

GO enrichment analysis was performed based on the R package clusterProfiler (version 3.14.3).[44] Representative GO terms with p value <0.05 were summarized. Statistical analyses were implemented with R (version 4.0.3, http://www.r-project.org).

DATA AND CODE AVAILABILITY

All sequencing data associated with this study were downloaded from the Gene Expression Omnibus (GEO) database under accession number GSE106340.[20]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2023.102044.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Eric S. Lander (Broad Institute of MIT and Harvard) for sharing their scRNA-seq data of somatic cell reprogramming in the GEO database under accession number GEO: GSE106340. This work was financially supported by the National Natural Science Foundation of China (62061034, 62171241); the Natural Science Foundation Project of Inner Mongolia Autonomous Region (2022ZD13); the key technology research program of Inner Mongolia Autonomous Region (2021GG0398); the Science and Technology Leading Talent Team in Inner Mongolia Autonomous Region (2022LJRC0009); and the innovation team development plan of colleges and universities in Inner Mongolia Autonomous Region (NMGIRT2204).

AUTHOR CONTRIBUTIONS

Y.Z. and C.L. conceived and designed the study. C.L. and H.L. performed the bioinformatics and statistical analysis of the data. C.L. and H.L. wrote the manuscript. P.L., L.C., Y.H., J.Z., and Q.X. contributed to the revision of the manuscript. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Ohnuki, M., and Takahashi, K. (2015). Present and future challenges of induced pluripotent stem cells. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140367. https://doi.org/10.1098/rstb.2014.0367.

2. Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676. https://doi.org/10.1016/j.cell.2006.07.024.

3. Apostolou, E., and Stadtfeld, M. (2018). Cellular trajectories and molecular mechanisms of iPSC reprogramming. Curr. Opin. Genet. Dev. *52*, 77–85. https://doi.org/10.1016/j.gde.2018.06.002.

4. Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., et al. (2010). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell *7*, 51–63. https://doi.org/10.1016/j.stem.2010.04.014.

5. Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell *7*, 64–77.

6. Papp, B., and Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. Cell *152*, 1324–1343. https://doi.org/10.1016/j.cell.2013.02.043.

7. Hochedlinger, K., and Jaenisch, R. (2015). Induced Pluripotency and Epigenetic Reprogramming. Cold Spring Harbor Perspect. Biol. *7*, a019448. https://doi.org/10.1101/cshperspect.a019448.

8. Wang, B., Wu, L., Li, D., Liu, Y., Guo, J., Li, C., Yao, Y., Wang, Y., Zhao, G., Wang, X., et al. (2019). Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Jdp2-Jhdm1b-Mkk6-Glis1-Nanog-Essrb-Sall4. Cell Rep. *27*, 3473–3485.e5.

9. Li, H., Long, C., Hong, Y., Chao, L., Peng, Y., and Zuo, Y. (2022). The Cumulative Formation of R-loop Interacts with Histone Modifications to Shape Cell Reprogramming. Int. J. Mol. Sci. *23*, 1567. https://doi.org/10.3390/ijms23031567.

10. Chao, L., Yang, S., Li, H., Long, C., Xi, Q., and Zuo, Y. (2022). Competitive binding of TET1 and DNMT3A/B cooperates the DNA methylation pattern in human embryonic stem cells. Biochim. Biophys. Acta. Gene Regul. Mech. *1865*, 194861. https://doi.org/10.1016/j.bbagrm.2022.194861.

11. Zhao, L., Long, C., Zhao, G., Su, J., Ren, J., Sun, W., Wang, Z., Zhang, J., Liu, M., Hao, C., et al. (2022). Reprogramming barriers in bovine cells nuclear transfer revealed by single-cell RNA-seq analysis. J. Cell Mol. Med. *26*, 4792–4804. https://doi.org/10.1111/jcmm.17505.

12. Xu, B., Liu, D., Wang, Z., Tian, R., and Zuo, Y. (2021). Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. Cell. Mol. Life Sci. *78*, 129–141.

13. Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., et al. (2013). Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. Science *341*, 651–654. https://doi.org/10.1126/science.1239278.

14. Guan, J., Wang, G., Wang, J., Zhang, Z., Fu, Y., Cheng, L., Meng, G., Lyu, Y., Zhu, J., Li, Y., et al. (2022). Chemical reprogramming of human somatic cells to pluripotent stem cells. Nature *605*, 325–331. https://doi.org/10.1038/s41586-022-04593-5.

15. Li, H., Song, M., Yang, W., Cao, P., Zheng, L., and Zuo, Y. (2020). A Comparative Analysis of Single-Cell Transcriptome Identifies Reprogramming Driver Factors for Efficiency Improvement. Mol. Ther. Nucleic Acids *19*, 1053–1064.

16. Li, H., Ta, N., Long, C., Zhang, Q., Li, S., Liu, S., Yang, L., and Zuo, Y. (2019). The spatial binding model of the pioneer factor Oct4 with its target genes during cell reprogramming. Comput. Struct. Biotechnol. J. *17*, 1226–1233.

17. Liu, X., Ouyang, J.F., Rossello, F.J., Tan, J.P., Davidson, K.C., Valdes, D.S., Schröder, J., Sun, Y.B.Y., Chen, J., Knaupp, A.S., et al. (2020). Reprogramming roadmap reveals route to human induced trophoblast stem cells. Nature *586*, 101–107. https://doi.org/10.1038/s41586-020-2734-6.

18. Xing, Q.R., El Farran, C.A., Gautam, P., Chuah, Y.S., Warrier, T., Toh, C.X.D., Kang, N.Y., Sugii, S., Chang, Y.T., Xu, J., et al. (2020). Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. Sci. Adv. *6*, eaba1190. https://doi.org/10.1126/sciadv.aba1190.

19. Parenti, A., Halbisen, M.A., Wang, K., Latham, K., and Ralston, A. (2016). OSKM Induce Extraembryonic Endoderm Stem Cells in Parallel to Induced Pluripotent Stem Cells. Stem Cell Rep. *6*, 447–455. https://doi.org/10.1016/j.stemcr.2016.02.003.

20. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell *176*, 928–943.e22. https://doi.org/10.1016/j.cell.2019.01.006.

21. Guo, L., Lin, L., Wang, X., Gao, M., Cao, S., Mai, Y., Wu, F., Kuang, J., Liu, H., Yang, J., et al. (2019). Resolving cell fate decisions during somatic cell reprogramming by single-cell RNA-Seq. Mol. Cell *73*, 815–829.e7.

22. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

23. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinf. *14*, 128. https://doi.org/10.1186/1471-2105-14-128.

24. Lattin, J.E., Schroder, K., Su, A.I., Walker, J.R., Zhang, J., Wiltshire, T., Saijo, K., Glass, C.K., Hume, D.A., Kellie, S., and Sweet, M.J. (2008). Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res. *4*, 5. https://doi.org/10.1186/1745-7580-4-5.

25. Takaishi, M., Tarutani, M., Takeda, J., and Sano, S. (2016). Mesenchymal to Epithelial Transition Induced by Reprogramming Factors Attenuates the Malignancy of Cancer Cells. PLoS One *11*, e0156904.

26. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091–1107.e17.

27. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982. https://doi.org/10.1038/nmeth.4402.

28. Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A., and Xu, Y. (2017). SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res. *45*, e54. https://doi.org/10.1093/nar/gkw1278.

29. Zhao, T., Fu, Y., Zhu, J., Liu, Y., Zhang, Q., Yi, Z., Chen, S., Jiao, Z., Xu, X., Xu, J., et al. (2018). Single-cell RNA-seq reveals dynamic early embryonic-like programs during chemical reprogramming. Cell Stem Cell *23*, 31–45.e7.

30. Tran, K.A., Pietrzak, S.J., Zaidan, N.Z., Siahpirani, A.F., McCalla, S.G., Zhou, A.S., Iyer, G., Roy, S., and Sridharan, R. (2019). Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. Cell Rep. *27*, 1726–1741.e5. https://doi.org/10.1016/j.celrep.2019.04.056.

31. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

32. Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

33. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. https://doi.org/10.1038/nmeth.4463.

34. Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell *168*, 442–459.e20. https://doi.org/10.1016/j.cell.2016.12.016.

35. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

36. Zheng, L., Liang, P., Long, C., Li, H., Li, H., Liang, Y., He, X., Xi, Q., Xing, Y., and Zuo, Y. (2023). EmAtlas: a comprehensive atlas for exploring spatiotemporal activation in mammalian embryogenesis. Nucleic Acids Res. *51*, D924–D932. https://doi.org/10.1093/nar/gkac848.

37. Li, H., Long, C., Hong, Y., Luo, L., and Zuo, Y. (2023). Characterizing Cellular Differentiation Potency and Waddington Landscape via Energy Indicator. Research *6*, 0118.

38. Verfaillie, A., Imrichova, H., Janky, R., and Aerts, S. (2015). iRegulon and i-cisTarget: Reconstructing Regulatory Networks Using Motif and Track Enrichment. Curr. Protoc. Bioinformatics *52*, 2.16.1–2.16.39.

39. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, e12776.

40. Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2021). RaacLogo: a new sequence logo generator by using reduced amino acid clusters. Briefings Bioinf. *22*, bbaa096.

41. Zheng, L., Liu, D., Li, Y.A., Yang, S., Liang, Y., Xing, Y., and Zuo, Y. (2022). RaacFold: a webserver for 3D visualization and analysis of protein structure by using reduced amino acid alphabets. Nucleic Acids Res. *50*, W633–W638.

42. Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics *33*, 122–124. https://doi.org/10.1093/bioinformatics/btw564.

43. 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. https://doi.org/10.1101/gr.1239303.

44. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.