

High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells

Gang Chang^{1,2,*}, Shuai Gao^{1,2,*}, Xinfeng Hou², Zijian Xu², Yanfeng Liu³, Lan Kang², Yu Tao², Wenqiang Liu², Bo Huang², Xiaochen Kou², Jiayu Chen², Lei An¹, Kai Miao¹, Keqian Di¹, Zhilong Wang¹, Kun Tan¹, Tao Cheng³, Tao Cai², Shaorong Gao², Jianhui Tian¹

¹Ministry of Agriculture Key Laboratory of Animal Genetics, Breeding and Reproduction, National Engineering Laboratory for Animal Breeding, College of Animal Sciences and Technology, China Agricultural University, 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China; ²National Institute of Biological Sciences, 7 Science Park Road, Zhongguancun Life Science Park, Beijing

It remains controversial whether the abnormal epigenetic modifications accumulated in the induced pluripotent stem cells (iPSCs) can ultimately affect iPSC pluripotency. To probe this question, iPSC lines with the same genetic background and proviral integration sites were established, and the pluripotency state of each iPSC line was characterized using tetraploid (4N) complementation assay. Subsequently, gene expression and global epigenetic modifications of “4N-ON” and the corresponding “4N-OFF” iPSC lines were compared through deep sequencing analyses of mRNA expression, small RNA profile, histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation. We found that methylation of an imprinted gene, *Zrsr1*, was consistently disrupted in the iPSC lines with reduced pluripotency. Furthermore, the disrupted methylation could not be rescued by improving culture conditions or subcloning of iPSCs. Moreover, the relationship between hypomethylation of *Zrsr1* and pluripotency state of iPSCs was further validated in independent iPSC lines derived from other reprogramming systems.

Keywords: imprinted gene; somatic reprogramming; iPSCs; tetraploid complementation

Cell Research (2014) 24:293-306. doi:10.1038/cr.2013.173; published online 31 December 2013

Introduction

Induced pluripotent stem cells (iPSCs) can be generated from differentiated somatic cells through the ectopic expression of a set of transcription factors that are enriched in embryonic stem cells (ESCs) [1]. Recent studies have raised serious concerns regarding the further use of iPSCs and the major concern involves the inferior quality of iPSCs. For example, differences in the transcriptomes, chromatin structures, methylomes, and even

variations in differentiation potential have been observed in iPSCs [2-6]. Therefore, screening high-quality iPSCs and identifying key factors associated with pluripotency are prerequisites for further application of iPSCs.

The *Dlk1-Dio3* gene cluster has been identified to correlate with the pluripotency of iPSCs [7]. However, it was reported that the stoichiometry of transcription factors influenced the state of the *Dlk1-Dio3* locus and some *Gtl2*-LOW (not OFF) iPSCs can support the full-term development of all-iPSC mice [8]. These studies indicate that many other factors contribute to the pluripotency state of iPSCs. Transcription factor-mediated reprogramming is a complex process, and the pluripotency of iPSCs can be influenced by the reprogramming strategies and culture conditions employed [9, 10]. Therefore, systematic deciphering of genetic or epigenetic alternations in the iPSCs with consistent genetic background would help to identify some hotspots of iPSCs.

To this end, we established genetically identical iPSCs

*These two authors contributed equally to this work.

Correspondence: Jianhui Tian^a, Shaorong Gao^b

^aTel: 86-10-62896057; Fax: 86-10-62733856

E-mail: tianjh@cau.edu.cn

^bTel: 86-10-80728967; Fax: 86-10-80727535

E-mail: gaoshaorong@nibs.ac.cn

Received 13 August 2013; revised 23 October 2013; accepted 25 November 2013; published online 31 December 2013

with well-defined pluripotency to enable the identification of hitherto unconsidered pluripotency-related factors at the genome-wide levels of mRNA expression, small RNA profile, core histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation using high-throughput (HT) sequencing strategies.

Results

High similarity of the transcriptomes, chromatin structures, and methylomes in iPSCs

Genome-wide analyses of human iPSCs have been well performed [2-4, 11]. However, systematic dissection of the relationship between the pluripotency of iPSCs and genome-wide deviations at different levels is still lacking. It becomes feasible for analyzing mouse iPSCs with combined tetraploid complementation validation and genome-wide deep sequencing. To this aim, we performed present study using a well-defined system in which the genetic background and proviral integrations in the genome of all iPSC lines were the same. Moreover, the pluripotency state of all iPSC lines used was characterized by tetraploid complementation assay [12, 13].

Then high-throughput sequencing of mRNA, small RNA, core histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation was applied to each cell line using RNA sequencing (RNA-Seq), small RNA sequencing (sRNA-Seq), chromatin immunoprecipitation sequencing (ChIP-Seq), and methylated DNA immunoprecipitation sequencing (MeDIP-Seq) strategies, respectively (Figure 1A). Overall, we sequenced eight iPSC lines, including 1°-iPSC-37 (4N-ON), 2°-iPSC-32 (4N-ON), 2°-iPSC-16 (4N-ON), 3°-iPSC-3 (4N-ON), 1°-iPSC-42 (4N-OFF), 2°-iPSC-4 (4N-OFF), 2°-iPSC-18 (4N-OFF), 3°-iPSC-9 (4N-OFF), with four somatic cells and one ESC line (R1) as controls. All iPSC lines were cultured for several passages (approximately p12) under standardized conditions, followed by the validation of pluripotency through 4N assay, and then cultures were collected for genome-wide sequencing. According to data published by us and others [8, 10, 12], transplanting at least two hundred reconstructed tetraploid embryos is used as a criterion to determine whether an iPSC line is “4N-ON” or “4N-OFF”. The “4N-ON” iPSCs could support the full-term development of all-iPSC mice (Supplementary information, Figure S1A). The simple sequence length polymorphism (SSLP) analysis results further confirmed the genetic background of all-iPSC mice (Supplementary information, Figure S1B). Characteristics of the iPSC lines used in HT analyses are summarized in Supplementary information, Table S1.

The comprehensive sequencing generated high-quality data with approximately 2.3 billion mapped reads. A detailed evaluation of the HT sequencing data is shown in Supplementary information, Table S2.

The transcriptomes, genome-wide maps of nucleosomes with H3K27me3, H3K4me3, and H3K4me2 modifications, and DNA methylation were compared among the iPSC lines. Our results showed that all iPSC lines clustered together with R1 but separated from somatic cells; especially, no well-separated clustering emerged to distinguish the “4N-ON” from “4N-OFF” iPSC lines (Figure 1B). The genome-wide similarities were further demonstrated by the examination of the representative pluripotency gene *Pou5f1* and the lineage differentiation-specific gene *Irx2* (Figure 1C and 1D). Transcriptional activation is controlled by the “open” chromatin state, which is remodeled from the folded state to ensure the access of transcription factors to the transcription start site (TSS) [14]. Thus, we next examined the distributions of core epigenetic modifications proximal to the TSS. Specifically, we analyzed the read coverage located from upstream 2 kb to downstream 2 kb region of TSS. Similar distributions of H3K27me3, H3K4me3, H3K4me2, and DNA methylation (5mC) were observed in these iPSC lines (Supplementary information, Figure S1C). These findings indicate that iPSC lines are globally similar by analyzing the above various molecular characteristics.

The dissection of cell line-specific deviations of iPSCs

Then we try to explore the cell line-specific deviations in iPSCs, which may account for the characteristic of individual iPSC line. The pairwise comparison at the aforementioned six levels showed the presence of some internal differences in each sub-group (Figure 2A-2F). These results indicate that cell line-specific deviations still exist despite the global similarity among iPSC lines.

The standard for screening cell line-specific deviations was designed based on the previous report with small modifications [11]. In detail, a reference is first extracted from the four “4N-ON” iPSC lines, and the resulting reference threshold quantifies the range of each gene or locus. All transcripts or loci for which the mRNA, microRNA (miRNA), histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation values fall outside of the center quartiles by more than 1.5 times the interquartile range, simultaneously, read count is 2 times more or 0.5 time less than the average of eight cell lines, are regarded as the candidate deviations. By this standard, we observed that each iPSC line had certain specific modification and expression pattern (Supplementary information, Table S3).

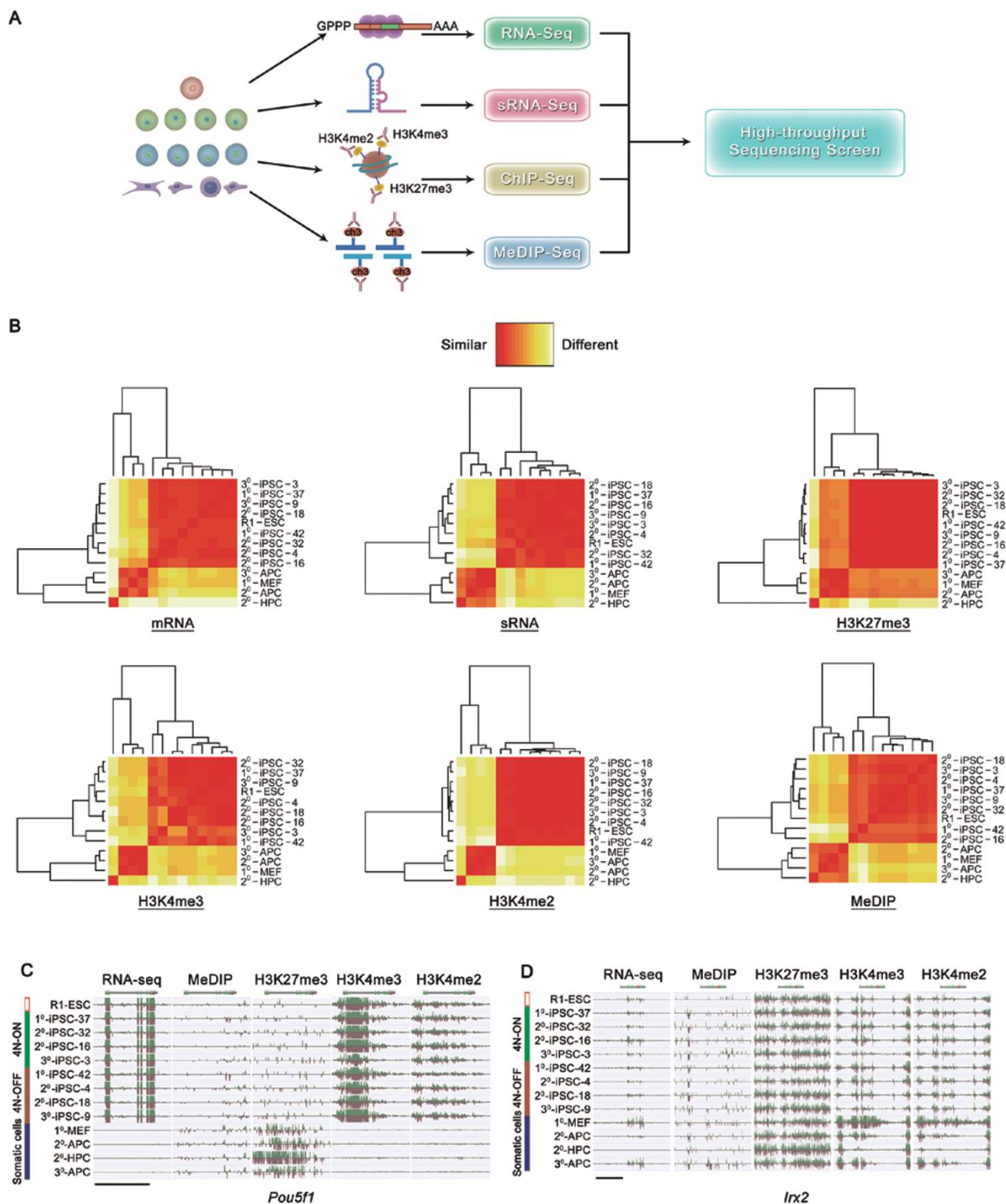


Figure 1 High similarity of the transcriptomes, chromatin structures, and methylomes in iPSCs. **(A)** Schematic depicting of the genome-wide analyses of well-defined iPSCs. In total, four “4N-ON” iPSC lines, four corresponding “4N-OFF” iPSC lines, four somatic cells, and one ESC line were used for sequencing. **(B)** The hierarchical cluster at different levels based on our datasets. The gradient color from red to the orange on each spot indicates the change of relation from similarity to difference between two samples. The x and y axis are symmetrical. **(C, D)** The presentation of RNA-Seq, MeDIP-Seq, H3K27me3, H3K4me3, and H3K4me2 ChIP-Seq profiles in the region of the *Pou5f1* and *Irx2* genes. Scale bars in **C** and **D**, 5 kb.

Notably, we identified many cell line-specific transcripts in each cell line (Supplementary information, Table S3); and six genes, such as *Coll1a1*, *Thy1*, *Ctgf*, *Cav1*, *Lox* and *Col5a1*, could be validated using quantitative PCR (Q-PCR) (Supplementary information, Figure S2A-S2F). The quantitative PCR results corresponded well to the general distribution patterns of read counts (Supplementary information, Figure S2G). Collectively,

our results showed that the cell line-specific deviations were present in our HT iPSC lines, which made it possible to further investigate the relationship between these deviations and cell line characteristics.

High-throughput screening reveals the loss of methylation of imprinted gene in “4N-OFF” iPSC lines

Generalized linear models, which were designed to

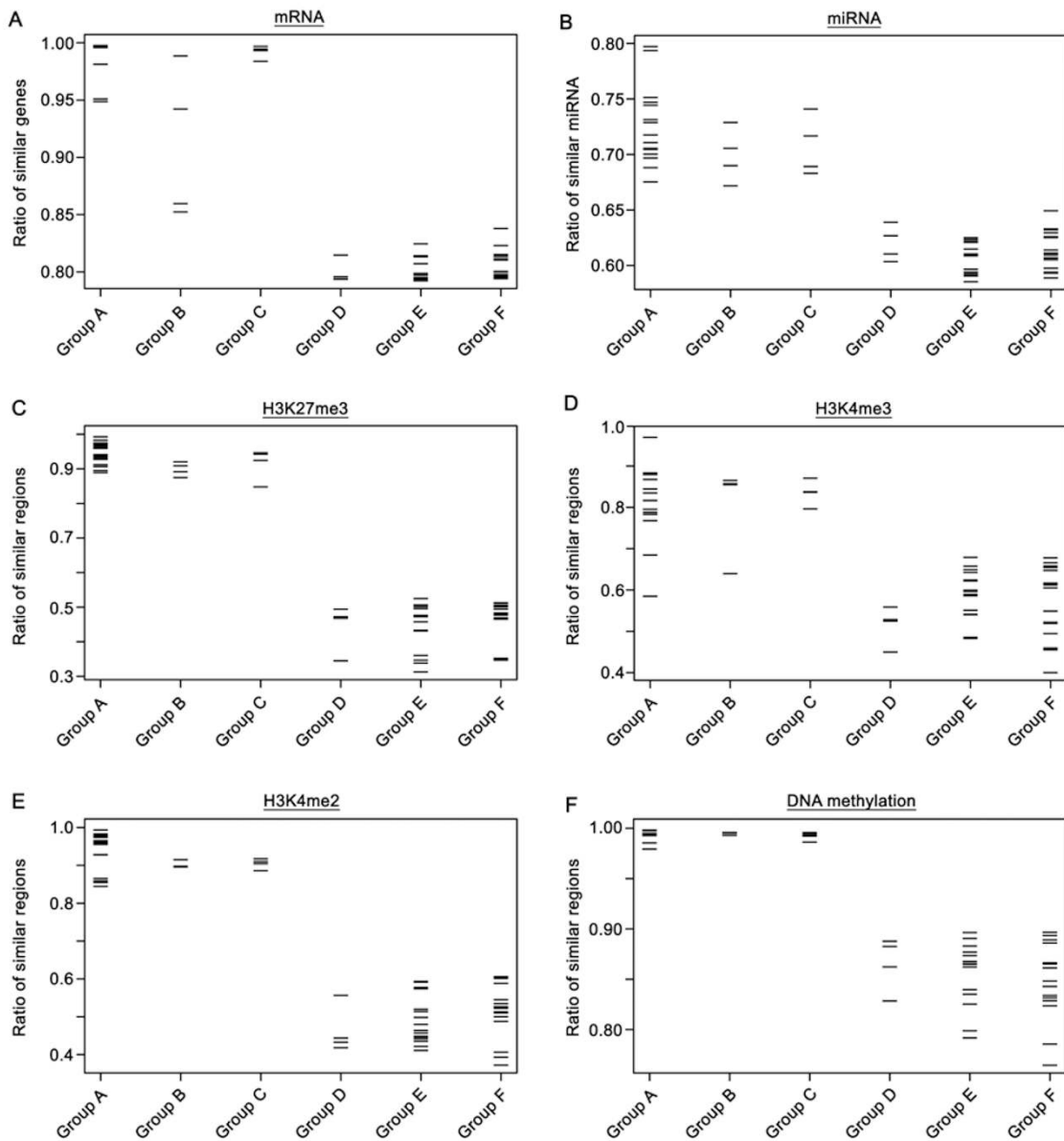


Figure 2 The pairwise comparison reveals the presence of cell line-specific deviations. **(A)** mRNA expression dataset. **(B)** miRNA expression dataset. **(C)** H3K27me3 ChIP dataset. **(D)** H3K4me3 ChIP dataset. **(E)** H3K4me2 ChIP dataset. **(F)** MeDIP dataset. Each short bar in the figure shows one minus ratio of significantly different genes/loci by the total considered genes/loci comparing two samples in Group A, between “4N-ON” iPSC lines and “4N-OFF” iPSC lines; Group B, between “4N-ON” iPSC lines and R1 ESC line; Group C, between “4N-OFF” iPSC lines and R1 ESC line; Group D, between somatic cells and R1 ESC line; Group E, between somatic cells and “4N-ON” iPSC lines; Group F, between somatic cells and “4N-OFF” iPSC lines.

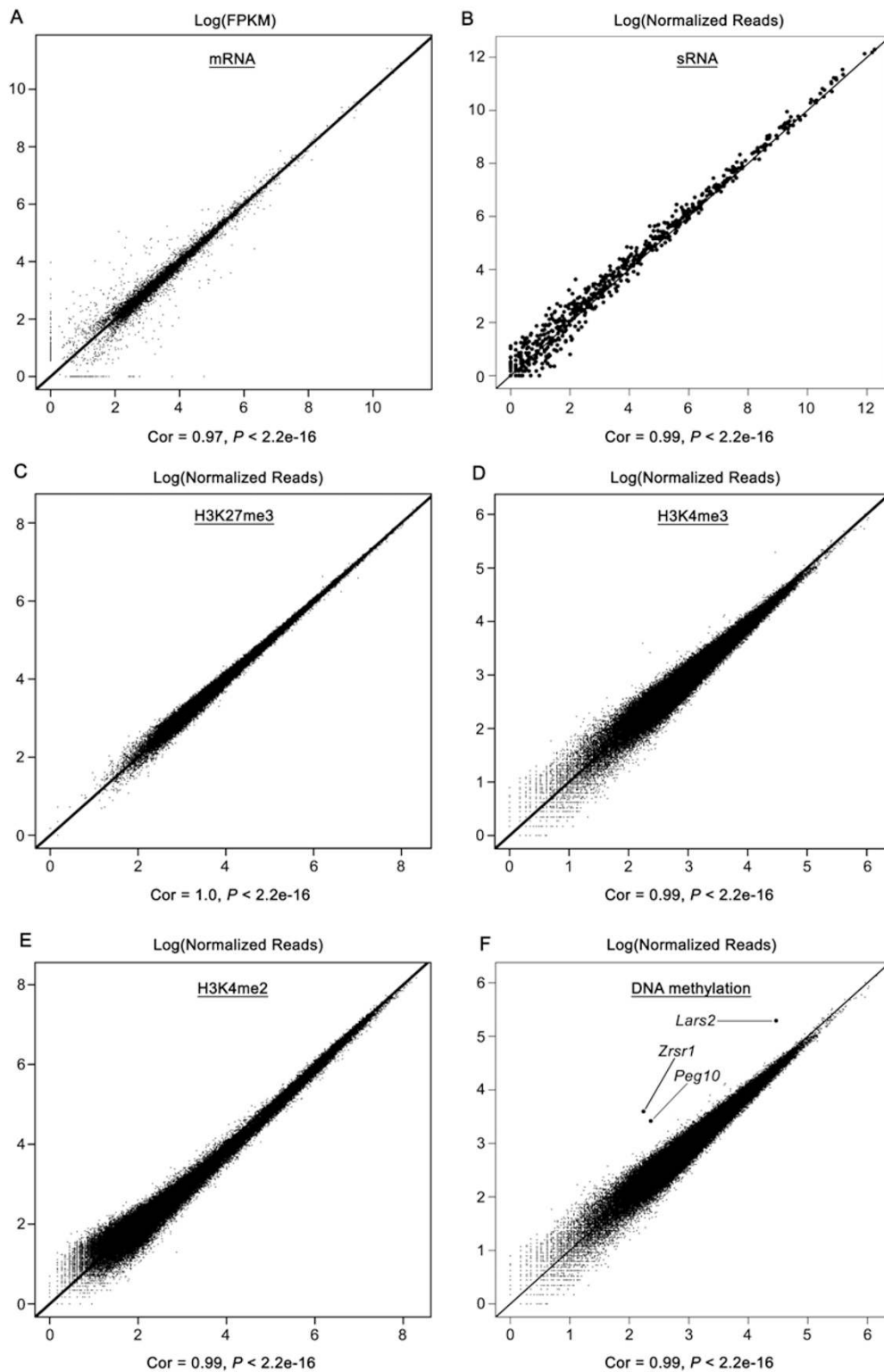


Figure 3 High-throughput sequencing reveals three genes with aberrant DNA methylation. **(A-F)** The scatter plot between the “4N-ON” and “4N-OFF” iPSC lines at the average levels of mRNA, small RNA (sRNA), core histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation based on our datasets, with significant loci highlighted with their names. Cor, Pearson correlation coefficient; P , P -value (X, Y: log transformed average reads or expression value).

estimate the effects of “4N-OFF”-specific characteristics, were further applied to the six aforementioned deep-sequencing datasets. The standards used to determine significance included multiple test-corrected *P*-values (< 0.05) and the similarity to R1 ESCs.

Among the top ten candidates at different levels summarized in Supplementary information, Table S4, one locus was finally identified to be of interest, which corresponded to an imprinted gene, *Zrsr1* (Figure 3A-3F). There are consistently high MeDIP signals enriched in the locus in all the four “4N-ON” iPSC lines and also R1

ESC line, which can be “good status” reference, and low signals in all the four corresponding “4N-OFF” iPSC lines (Figure 4A). Notably, the hypomethylation of *Zrsr1* has been reported to correlate with the decreased pluripotency in nuclear transfer-mediated reprogramming [15]. Other modifications (H3K27me3, H3K4me3, and H3K4me2) and transcripts of this locus showed no difference between the “4N-ON” and “4N-OFF” iPSC lines (Figure 4A). The locus mapping to *Zrsr1* was divided into three regions for the following validation: Z1R2A, Z1R2B, and Z1R3. Z1R3 was located in an exon of *Zrsr1*, while

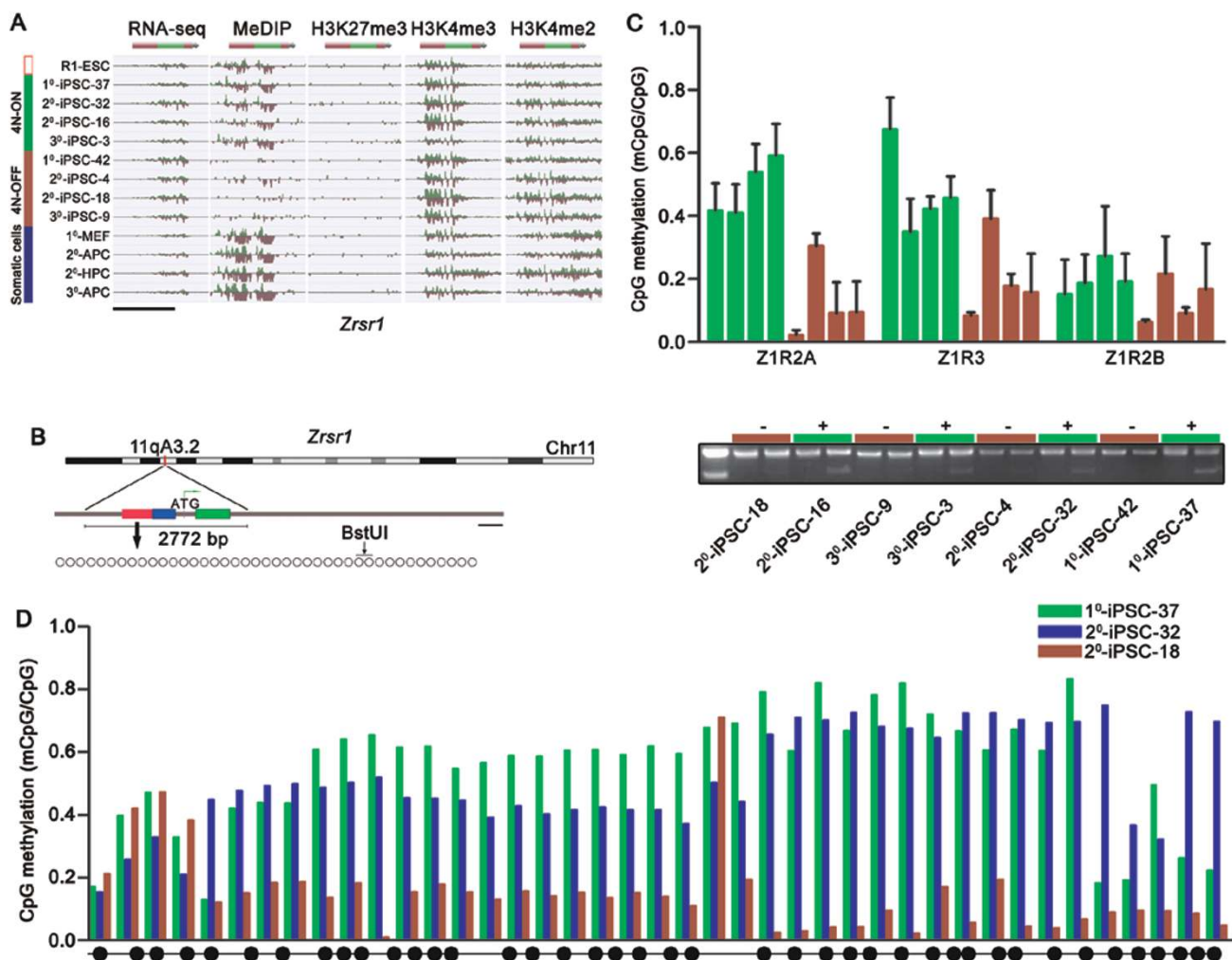


Figure 4 The confirmation of hypomethylation of *Zrsr1* in the HT “4N-OFF” iPSCs. **(A)** The read distribution of RNA-Seq, H3K27me3, H3K4me3, H3K4me2 ChIP-Seq, and MeDIP-Seq at the locus corresponding to the upstream region of *Zrsr1*. Scale bar, 3 kb. **(B)** Schematic representation of the bisulfite sequencing validation of locus located at *Zrsr1* (red box, Z1R2A; blue box, Z1R2B; green box, Z1R3). The arrowheads indicate the CpG dinucleotides analyzed in COBRA. BstUI was used to cut Z1R2A. Scale bar, 0.5 kb. **(C)** Bisulfite genomic sequencing of Z1R2A, Z1R2B, Z1R3, and COBRA of Z1R2A. Green columns indicate the four “4N-ON” iPSC lines; brown columns indicate the four “4N-OFF” iPSC lines. (+) enzyme added, (–) enzyme not added. Error bars, SD (*n* = 3 measurements). **(D)** Ion torrent sequencing analyses of the DNA methylation of Z1R2A. The black circles at the bottom stand for the CpG islands located in the Z1R2A region. The column groups in the histogram from the left to right sequentially correspond to the CpG islands.

Z1R2A and Z1R2B were located within the promoter region (Figure 4B). Z1R2A is located in the upstream regulatory regions with annotated CpG islands (Figure 4B). Bisulfite genomic sequencing was performed to confirm

the HT data, and Z1R2A was finally confirmed to be the actual differentially methylated regions (DMRs) of *Zrsr1* (Figure 4C, top). Combined bisulfite restriction analysis (COBRA) of Z1R2A further confirmed these results

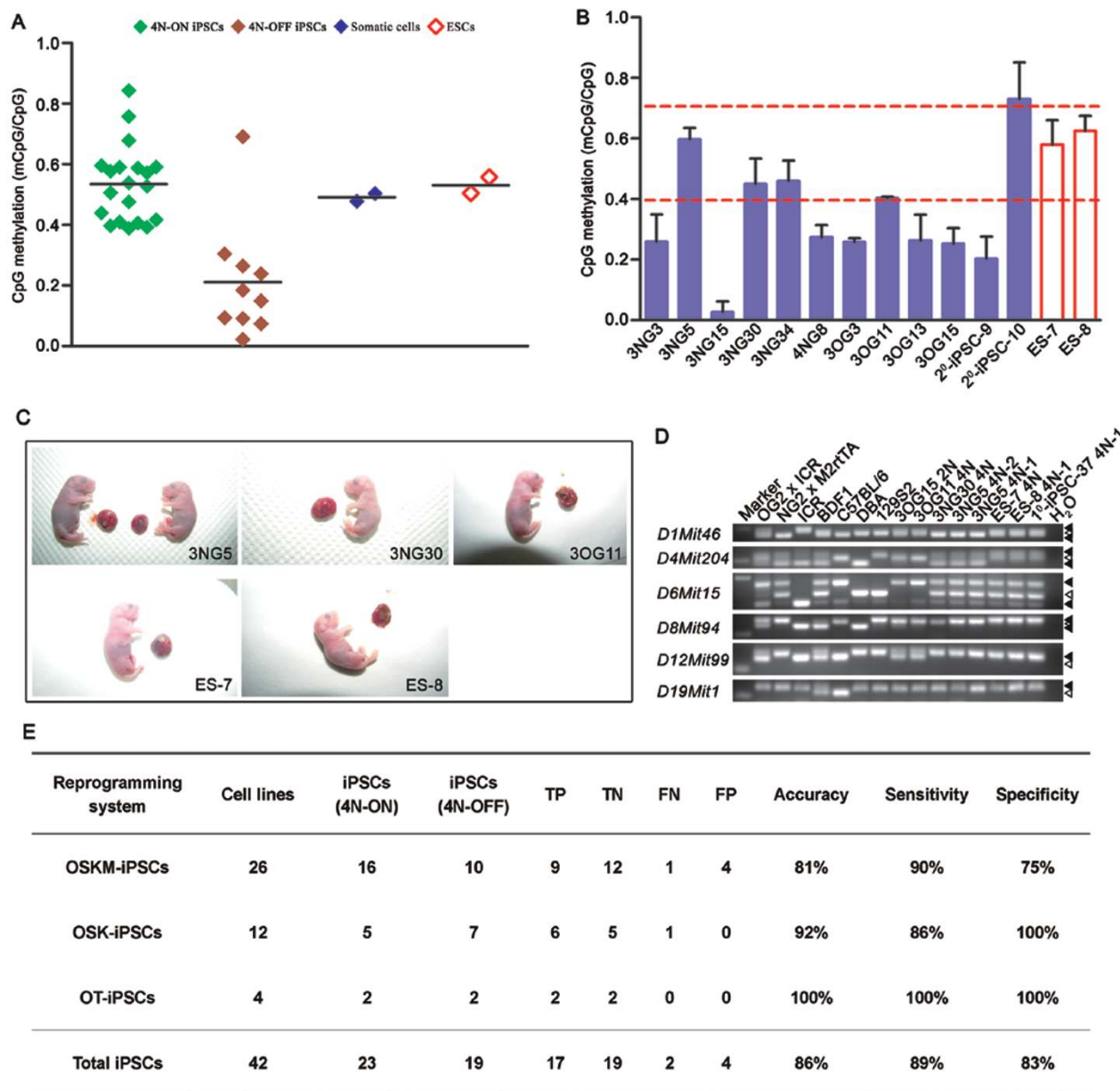


Figure 5 The confirmation of relationship between hypomethylation of *Zrsr1* and the reduced pluripotency of iPSCs. **(A)** Bisulfite genomic sequencing of *Zrsr1* in the bisulfite validation set ($n = 30$). **(B)** Bisulfite genomic sequencing of *Zrsr1* in the test set ($n = 12$) with two ESC lines as controls. The blue columns indicate the iPSC lines; the red-frame columns indicate the ESC lines. The red dashed lines represent the top and bottom boundaries of DNA methylation of *Zrsr1*. Error bars, SD ($n = 3$ measurements). **(C)** The viable neonatal all-iPSC mice derived from the iPSC lines in the test set and control ESC lines through 4N assay. **(D)** SSLP analyses for the identification of the all-iPSC mice as indicated. **(E)** Summary of *Zrsr1* hypomethylation in distinguishing “4N-OFF” iPSC lines (true positives, TP) from “4N-ON” iPSC lines (true negatives, TN), cell lines that are incorrectly regarded as “4N-OFF” iPSC lines (false positives, FP) or incorrectly regarded as “4N-ON” iPSC lines (false negatives, FN).

(Figure 4C, bottom). Notably, the ion torrent sequencing platform was also used to verify the observed results, and DNA methylation patterns of Z1R2A in three typical iPSC lines were consistent with the above bisulfite genomic sequencing results (Figure 4D).

Thus far, these results indicate that loss of methylation of *Zrsr1*, which was screened by HT sequencing, might correlate with the reduced pluripotency of “4N-OFF” iPSC lines.

Confirmation of the relationship between hypomethylation of Zrsr1 and the reduced pluripotency of iPSCs

To further confirm the relationship between the observed hypomethylation of *Zrsr1* and impaired pluripotency of “4N-OFF” iPSC lines, another twenty-two iPSC lines with known pluripotency status were included. These iPSC lines included iPSCs generated using four factors (OSKM), three factors (OSK), and two factors (OT, *Oct4* and *Tet1*; Chen *et al.*, unpublished). Taken together, we extensively analyzed a comprehensive validation set containing thirty iPSC lines (20 “4N-ON” and 10 “4N-OFF”), together with two ESC lines and two somatic cell lines as controls. The bisulfite genomic sequencing results showed that Z1R2A could efficiently distinguish “4N-ON” from “4N-OFF” iPSCs (Figure 5A), but Z1R3 could not (Supplementary information, Figure S3A). Thus, hypomethylation of *Zrsr1* (stands for Z1R2A when mentioning DNA methylation) was associated with the reduced pluripotency of iPSCs.

Next, a test set containing twelve independent iPSC lines with unknown pluripotency and another two ESC lines as controls, ES-7 and ES-8 (with genetic background of 129S2×M2rtTA), was used to further verify the observed association. These iPSC lines included two cell lines from APCs (2°-iPSC-9 and -10), one cell line carrying the genetic background of Nanog-GFP derived with a polycistronic cassette expressing *Oct4*, *Sox2*, *Klf4*, and *c-Myc* (OSKM), five cell lines carrying the genetic background of Nanog-GFP derived with a polycistronic cassette expressing *Oct4*, *Sox2*, and *Klf4* (OSK), and four cell lines carrying the genetic background of Pou5f1-EGFP derived with a polycistronic cassette expressing *Oct4*, *Sox2*, and *Klf4* (OSK). The methylation level of *Zrsr1* from 40 to 70%, which corresponded to the mean ± SD of our “4N-ON” validation set, is considered the normal range. Any iPSC line falling outside of the normal range was considered “4N-OFF”. Eight iPSC lines among the twelve cell lines were found to be *Zrsr1* hypomethylated (Figure 5B). Indeed, these eight iPSC lines failed the stringent test of 4N assay (Supplementary information, Table S1). Three of the four remaining iPSC lines were verified to be “4N-ON” (Figure 5C and 5D).

Notably, one outlier in test set, 3NG34, is “4N-OFF” but its methylation level is within the normal range, suggesting the involvement of other unknown factors in the failure of 4N assay. The performance of *Zrsr1* hypomethylation was summarized, and an average accuracy of 86%, sensitivity of 89%, and specificity of 83% were achieved among the total forty-two iPSC lines (Figure 5E). This result strongly suggests the inherent association between hypomethylation of *Zrsr1* and impaired pluripotency of iPSCs. Interestingly, the expression level of *Zrsr1* in different iPSC lines showed no significant difference (Supplementary information, Figure S3B). Similarly, the expression of *Murr1*, also known as *Commd1*, in which *Zrsr1* located [16], also showed no consistent difference among “4N-ON” and “4N-OFF” iPSCs (Supplementary information, Figure S3C). No detectable 5mC signal was observed on the DMR2 of *Murr1* (Supplementary information, Table S5).

Moreover, we also investigated the state of *Dlk1-Dio3* locus in our system. We analyzed DNA methylation at the *Gtl2* DMR, which was reported to correlate with the expression of *Gtl2* and *Rian* [17]. In agreement with recently published results [8], we detected hypermethylation of the *Gtl2* DMR in some “4N-ON” iPSC lines (Supplementary information, Figure S3D and S3E). The gene expression levels of *Gtl2* and *Rian* showed no significant differences between the “4N-ON” and “4N-OFF” iPSC lines (Supplementary information, Figure S3F-S3I). Thus, aberrant methylation of *Zrsr1*, not *Dlk1-Dio3* locus, was found in “4N-OFF” iPSCs in our system.

Aberrant DNA methylation status of Zrsr1 could not be rescued by improving the culture condition or subcloning of iPSCs

To determine whether *in vitro* passaging might impact DNA methylation status of *Zrsr1*, early-passage (p5) and later-passage (p12) iPSCs were used, and the results showed that DNA methylation of *Zrsr1* was stable (Figure 6A). Therefore, transcription factor-mediated reprogramming was presumed to be the major reason for hypomethylation. Considering the heterogeneity of the cell populations at later stages of reprogramming, early-stage intermediates were selected and the results showed that the DNA methylation of *Zrsr1* dropped sharply on induction day 7 (D7) (Figure 6B).

To further study whether hypomethylation of *Zrsr1* was the characteristic of whole cell population or some given cells, single cell-derived subclones of “4N-ON” and “4N-OFF” iPSCs were analyzed. We found that DNA methylation of *Zrsr1* in the subclones of “4N-OFF” iPSC lines still maintained the hypomethylation state resembling original iPSCs (Figure 6C and 6D). This result

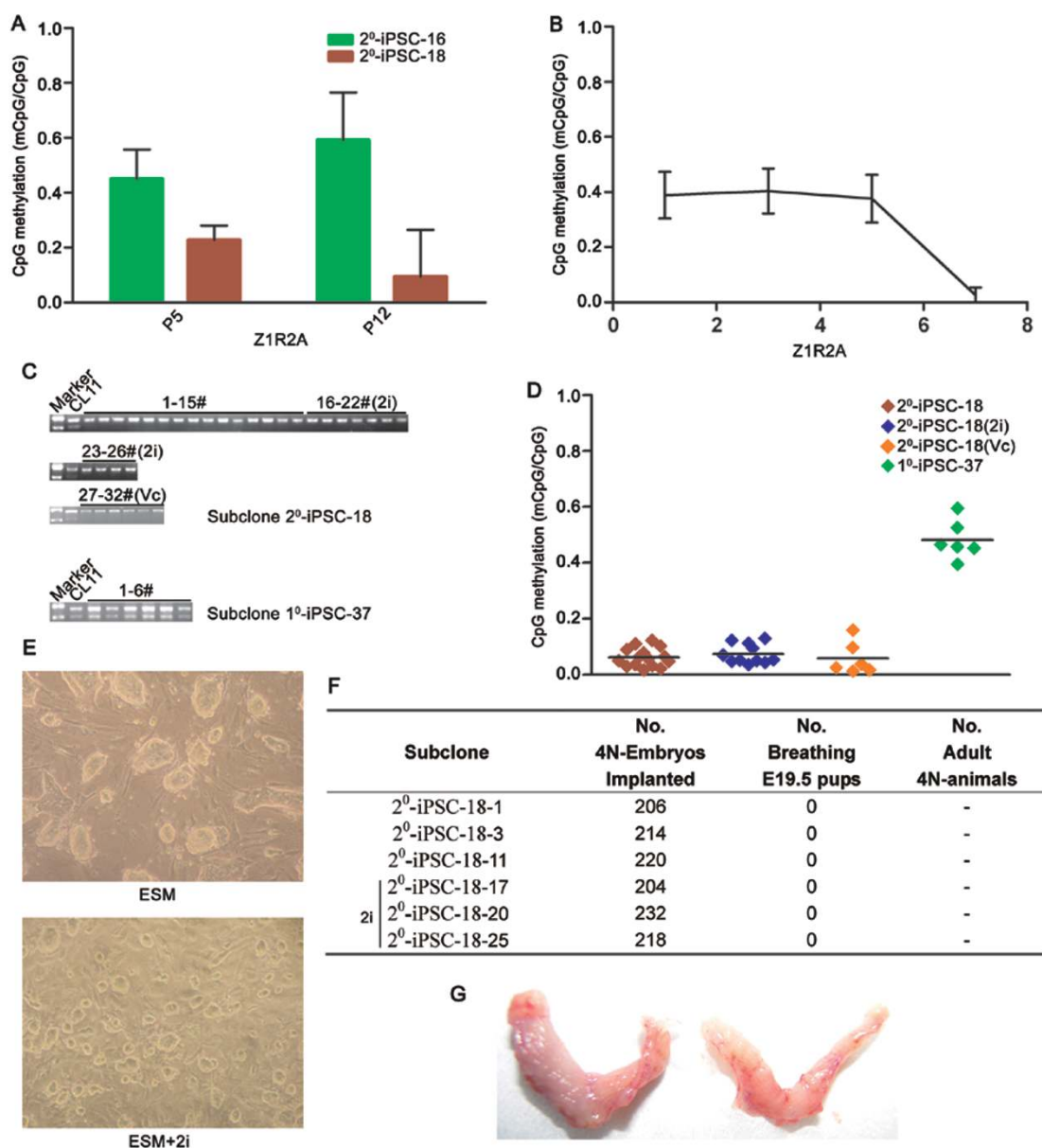


Figure 6 DNA methylation status of *Zrsr1* during the process of *in vitro* culture, transcription factor-mediated reprogramming, and subcloning of iPSCs with or without small molecules; and the failure of generating all-iPSC mice with subclones of “4N-OFF” iPSCs. **(A)** DNA methylation of Z1R2A in early- and later-passage HPC-iPSCs. **(B)** DNA methylation changes of Z1R2A in the process of cellular reprogramming. **(C)** Combined bisulfite restriction analysis of Z1R2A in the subclones of 2⁰-iPSC-18, 2⁰-iPSC-18 with 2i, 2⁰-iPSC-18 with Vc, and 1⁰-iPSC-37. CL11 (ESCs) was used as a control. Error bars, SD ($n = 3$ measurements). **(D)** Bisulfite genomic sequencing of Z1R2A in the subclones of 2⁰-iPSC-18, 2⁰-iPSC-18 with 2i, 2⁰-iPSC-18 with Vc, and 1⁰-iPSC-37. **(E)** The morphology changes in the subclones of 2⁰-iPSC-18 cultured in ESC medium without or with 2i. **(F)** Summary of tetraploid complementation in subclones of 2⁰-iPSC-18. **(G)** The uterus of pseudo-pregnant mice implanted with tetraploid-aggregated embryos injected with the iPSCs of 2⁰-iPSC-18 subclones.

indicates the homogeneous hypomethylation pattern of *Zrsr1* in “4N-OFF” iPSCs. Then we examined whether the quality of iPSCs could be improved by optimizing the reprogramming conditions. It was previously reported that ascorbic acid (Vc) could prevent the aberrant DNA

methylation of imprinted genes during the transcription factor-mediated reprogramming process [18]. Therefore, we tried to test whether Vc could also prevent the hypomethylation of *Zrsr1* observed in this study. Vc was used in the process of establishment of subclones and *de novo*

iPSC generation. However, we found that Vc could not rescue the hypomethylation of *Zrsr1* in the establishment of subclones (Figure 6C and 6D). Similarly, Vc could not prevent the appearance of high proportion of iPSCs with hypomethylated *Zrsr1* locus in the process of *de novo* iPSC generation (Supplementary information, Figure S4). Then, we tried to use small molecule inhibitors (2i) (PD0325901 and CHIR99021) to improve the pluripotency of “4N-OFF” iPSCs. Our results showed that 2i could change the morphology of iPSCs (Figure 6E); however, the rescued pluripotency was not observed in the subclones cultured with 2i (Figure 6F and 6G). These results demonstrate that hypomethylation of *Zrsr1* is very stable once it was studied during the reprogramming process. Taken all together, genome-scale sequencing of genetically well-defined iPSCs reveals that loss of methylation of *Zrsr1* correlated with the reduced pluripotency of iPSCs.

Discussion

In this study, to further explore the candidate factors correlating with the pluripotency of iPSCs, we utilized genetically well-defined iPSCs combined with high-throughput sequencing to characterize the genome-wide differences between “4N-ON” and “4N-OFF” iPSC lines. The total 105 sequencing samples deposited in GEO will be valuable data source for the community. To our knowledge, the present genome-wide analyses provide the most detailed information in understanding the pluripotency of iPSCs thus far. Our study demonstrates that iPSCs are globally similar at the levels of mRNA, small RNA, core histone modifications (H3K27me3, H3K4me3, and H3K4me2), and DNA methylation, which is similar to a recent report in which undistinguishable gene expression and chromatin modification differences were observed between human ESCs and iPSCs [4]. However, each iPSC line has cell line-specific deviations that may be associated with the specific cellular identity. Most importantly, we have discovered that the consistent loss of DNA methylation of *Zrsr1* correlated with the reduced pluripotency of iPSCs derived by different combinations of transcription factors.

As a pivotal epigenetic modification, DNA methylation plays a wide range of roles in regulation of gene expression, genomic imprinting, X chromosome inactivation and tumorigenesis [19, 20]. Abnormal imprinting has been found in the process of transcription factor-mediated reprogramming [7, 21]. Here, our result showed that hypomethylation of *Zrsr1* was associated with the impaired pluripotency of iPSCs. In consistence with our finding, the hypomethylation of *Zrsr1* was previously

found to be correlated with the reduced pluripotency of NT-ESCs [15]. In addition, hypermethylation of *Zrsr1* was detected in parthenogenetic ESCs with decreased pluripotency [22, 23]. Taken together, accumulated evidence suggests that the DNA methylation status of *Zrsr1* appears to be pivotal for successful reprogramming. However, the reasons accounting for hypomethylation of *Zrsr1* were still unknown. We propose that some *de novo* DNA methyltransferases, such as *Dnmt3a*, which have been proposed to be correlated with imprinting in iPSCs [18], might play a role in this process.

Zrsr1, also known as *U2af1-rs1*, was found to be a neomorphic maternally imprinted gene, and the CpG island (CGI) of *U2af1-rs1* was the imprinting control region (ICR) of *Murr1/U2af1-rs1* locus [24–26]. Notably, the maternal-predominant expression of *Murr1* was found to result from the transcriptional interference of *Zrsr1*, demonstrating the dominant role of *Zrsr1* in controlling the *Murr1/U2af1-rs1* locus [16]. Our observation that the expressions of *Zrsr1* and *Murr1* showed no difference between “4N-ON” and “4N-OFF” iPSCs was consistent with this notion [16]. Although the mechanism of how hypomethylation of *Zrsr1* impact the pluripotency of iPSCs remains unclear, it is worth noting that the connection between hypomethylation and upregulation of gene expression is not completely strict [27]. It was proposed that hypomethylated DNA may exert its roles through long-range chromosomal interactions.

Ascorbic acid (Vc) and valproic acid (VPA) exert important roles in optimizing the transcription factor-mediated reprogramming [28, 29], and were reported to successively prevent the hypermethylation of the *Dkl1-Dio3* locus [7, 18]. However, the direct evidence of linking the normal *Dkl1-Dio3* imprinting in the drug-treated iPSCs with the improved pluripotency verified through 4N assay is still lacking. Our result showed that the supplementation of Vc could not prevent the hypomethylation of *Zrsr1*, and the dominating role of Vc in decreasing DNA methylation of imprinted genes may account for this inconsistency. Then, we tried to improve the quality of iPSCs by using 2i. The employment of 2i, which selectively targeted mitogen-activated protein kinase kinase and glycogen synthase kinase-3, was reported to revert primed ESCs to naïve pluripotency state [30]; however, 2i showed no effect on the recovery of *Zrsr1* hypomethylation nor the improved pluripotency of iPSCs as manifested through 4N assay. The biological relevance of this discrepancy of 2i in ESCs and iPSCs remains elusive.

Up to now, the discovered pluripotency-related factors, *Dkl1-Dio3* and *Zrsr1*, are both imprinted genes; although the mechanism is not completely understood,

these findings indicate the underlying relationship between imprinted genes and pluripotency of iPSCs. Thus, other recently reported DMRs [31] were also scrutinized once more. However, except for *Zrsr1*, no other consistent DNA methylation loci were found to be associated with the pluripotency of iPSCs (Supplementary information, Table S5). Of note, the 5mC signal was not detected at twenty-six loci in the present study (Supplementary information, Table S5). Therefore, it will be intriguing to evaluate whether other high-resolution sequencing methods, such as bisulfite or single-cell sequencing [32], could provide more information about these loci.

The absolute association of any gene or loci with pluripotency of iPSCs is still not well understood. It is evident that the circuitry of pluripotency in iPSCs is more complex than we imagined. Epigenetic modifications, such as DNA methylation and core histone modifications, have been proposed to play important roles in maintaining the state of pluripotent stem cells [3, 33]. Therefore, in addition to the absolute pluripotency-related factors, the underlying relationship between cell line-specific epigenetic abnormalities and the identity of iPSCs is also worthy of study. Our present study analyzed the cell line-specific deviations at the levels of transcriptomes, genome-wide maps of nucleosomes with H3K27me3, H3K4me3, and H3K4me2 modifications, and DNA methylation. Further investigation of these cell line-specific deviations might be informative for elaborating the functional differences using more sophisticated and integrated analysis approaches. These studies would shed light on the mechanisms of reprogramming and improve the quality of iPSCs.

Materials and Methods

Cell culture

Mouse embryonic fibroblasts (MEFs) were derived from 13.5 dpc embryos collected from female 129S2/Sv, OG2 (*Tg(Pou5f1-EGFP)2Mnn*), and NG2 (*Tg(Nanog-GFP,Puro)*) mice that were mated with *Rosa26-M2rtTA* transgenic mice. ESCs and iPSCs were cultured on mitomycin C-treated MEFs in ES medium, which contained DMEM (Life technologies) supplemented with 15% FBS, 1 mM L-glutamine, 0.1 mM mercaptoethanol, 1% non-essential amino acid, and 1 000 U/ml LIF (all from Chemicon).

The generation of iPSCs

The viral supernatants containing the TetO-FUW-*Oct4*, *Sox2*, *Klf4*, and *c-Myc* plasmids and the packaging plasmids ps-PAX-2 and pMD2G were harvested, and the MEFs were infected with supernatants containing viruses at a density of $\sim 5 \times 10^5$ cells per 6-cm dish. The infection medium was replaced with ES medium supplemented with 1 μ g/ml doxycycline (Dox) 12 h after infection. The ES-like colonies appeared at approximately 12 days, and 4 days after the withdrawal of Dox, smooth colonies were isolated

and passaged 3 days later for the derivation of 1°-iPSC lines. Then somatic cells of 1°-all-iPSC mice were separately isolated, and the addition of Dox to the medium for the generation of 2°-iPSCs. 3°, 4°, 5°, and 6°-iPSCs were all derived from the preceding all-iPSC mice using the same induction method.

For the generation of Vc-iPSCs, Vc was added to the ESM with Dox to a final concentration of 10 ng/ml. And ES-like colonies were picked up on day 14 after two days withdrawal of Dox and Vc, and eight Vc-iPSC lines (iPSC-V1-V8) were established for subsequent study.

SSLP and q-PCR analyses

The primer sequences for SSLP analysis were obtained from the Mouse Genome Informatics website (<http://www.informatics.jax.org>). q-PCR was performed using a SYBR Green-based PCR Master Mix (Takara) and the ABI7500 Real-Time PCR System (Applied Biosystems).

Diploid blastocyst injection and tetraploid complementation

All animal procedures were performed according to the National Institute of Biological Sciences Guide for the care and use of laboratory animals. Teratoma formation, diploid blastocyst injection, and tetraploid complementation were performed as previously described [12].

Bisulfite genomic sequencing and combined bisulfite restriction analysis

Bisulfite genomic sequencing and combined bisulfite restriction analysis were conducted as previously described [34]. Three biological repeats were carried out for bisulfite genomic sequencing in each cell line. The PCR products cloned to the vector by the pEASYTM-T5 Zero cloning kit (TransGen Biotech) were sequenced in Invitrogen (Life Technologies) and SangonBiotech (Shanghai, China).

Ion torrent bisulfite genomic sequencing

The PCR products of Z1R2A from nested PCR were sonicated to about 300 bp using BioRuptor UCD-200 TS Sonication System (Diagenode) as verified through gelose gel electrophoresis. The fragmented DNA was end-repaired and purified to prepare for ligation to Ion adapters. Then the adapter-ligated and nick-translated DNA was recovered using Agencourt AMPure XP Reagent (Beckman Coulter). The 200 base-read unamplified libraries (330 bp target peak) were established using E-Gel iBase Size-select library system (Life Technologies) and quantified using a 2100 Bioanalyzer (Agilent Technologies). Finally, the libraries fixed on the magnetic beads were further amplified, recovered using the Ion OneTouch and Ion OneTouch ES (Life Technologies), respectively. The sequencing and data analyses were performed at CapitalBio Corporation (Beijing, China) using the Ion Torrent system developed by Life Technologies.

RNA-Seq library generation

Total RNA was isolated from cell pellets using TRIzol reagent (Life Technologies) according to the manufacturer's instructions. The RNA integrity was confirmed using a 2100 Bioanalyzer (Agilent Technologies) with a minimum RNA integrity number (RIN) of 8. The mRNA was enriched using oligo(dT) magnetic beads and sheared to create short fragments of approximately 200 bp. cDNA

was synthesized using random hexamer primers and purified using PCR product extraction kit (Qiagen). Finally, the sequencing primers linked to the cDNA fragments (approximately 200 bp) were isolated by gel electrophoresis and enriched by PCR amplification to construct the library.

MeDIP-Seq library generation

Genomic DNA was extracted from the cell pellets using the DNeasy Mini Kit (Qiagen). The DNA quality was analyzed using a Qubit 2.0 Fluorometer (Life Technologies). The gDNA was sonicated to 100-500 bp and repaired to contain a 3'-dA overhang; adapters were then ligated to the end of the DNA fragments according to the Paired-End DNA Sample Prep Kit (Illumina). For immunoprecipitation, the DNA was first denatured, then immunoprecipitated with the 5mC antibody using the Magnetic Methylated DNA Immunoprecipitation kit (Diagenod). q-PCR was performed to validate the enrichment efficiency. Next, the immunoprecipitated DNA was amplified for approximately 12-15 cycles, and fragments of the proper size (usually 200-300 bp) were gel-purified using the Gel Extraction Kit (Qiagen) and quantified using a 2100 Bioanalyzer (Agilent Technologies).

sRNA library generation

sRNAs of approximately 18-30 nt were first separated from the 5-10 µg of total RNA by size fractionation with a 15% TBE urea polyacrylamide gel (TBU). Next, the 5' RNA adapter (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3') was ligated to the RNA pool with T4 RNA ligase. The ligated RNA was size-fractionated on a 15% agarose gel, and the 40-65 nt fraction was excised. The 3' RNA adapter (5'-pUCGUAUGCCGU-CUUCUGCUUGidT-3'; p, phosphate; idT, inverted deoxythymidine) was subsequently ligated to the precipitated RNA using T4 RNA ligase. The ligated RNA was size-fractionated on a 10% agarose gel, and the 70-90 nt fraction was excised. The samples were reverse transcribed, and amplified for about 15 cycles to generate the sRNA libraries.

Chromatin immunoprecipitation (ChIP) and ChIP-Seq library generation

ChIP experiments were performed as previously described (<http://www.abcam.com>). Briefly, $\sim 1.5 \times 10^8$ cells were resuspended in lysis buffer and digested with micrococcal nuclease (Takara) for approximately 5 min at 37 °C. Then, the lysate was immunoprecipitated with the following antibodies: anti-H3K4me2 (Millipore 07-030), anti-H3K4me3 (Abcam ab8580), and anti-H3K27me3 (Millipore 07-449). A fraction of input "whole-cell extract" was retained as a sequencing control. The DNA isolated from the ChIP was quantified using a Qubit 2.0 Fluorometer (Life Technologies), and q-PCR was performed to validate the enrichment efficiency. Then, the enriched DNA was sonicated to 100-500 bp fragments. The DNA ends were repaired to create 3'-dA overhangs, and the adapters were then ligated to the ends of the DNA fragments. DNA fragments of approximately 100-300 bp were recovered and amplified to construct the sequencing library. Thirty-nine ChIP libraries and thirteen input controls were used for sequencing.

Illumina HiSeq 2000 sequencing

The sequencing was performed at the Beijing Genomics Insti-

tute (BGI) (Shenzhen, Guangdong, China) using the HiSeq 2000 system developed by Illumina. Paired-end was used for MeDIP-Seq; Single-end sequencing was applied to RNA-Seq, sRNA-Seq, and ChIP-Seq.

Sequence analysis

The MeDIP/ChIP sequencing reads were mapped to the mouse reference genome (mm9/NCBI37) using Bowtie (v0.12.7) software allowing the max mismatch of 3 nt [35]. For the paired-end sequences, the maximum insert length constraint was 1 000 nt. Only unique mapped reads were extracted for the subsequent analysis. The dinucleotide frequency in the 200 nt range that was centered in the mapped reads was estimated, and the dinucleotide enrichment was calculated as the ratio of dinucleotide frequency in each sample to the frequency of dinucleotides in the whole genome background. The reads that overlapped in different annotation categories (such as exons, introns, and intergenic regions, etc.) were counted, and the relative enrichment was calculated as the ratio of the proportion of reads falling into a certain annotation category under the real sample and under the assumption that the reads were uniformly distributed along the whole genome. The genome annotation datasets were generated using the Ensembl BioMart service (<http://www.ensembl.org>), except that the repeats were downloaded from the UCSC genome browser (<http://genome.ucsc.edu>). The overlapping annotation regions were reduced, and the R (version 2.13.0) and Bioconductor (release 2.8) packages (<http://r-project.org>, <http://www.bioconductor.org>) were used for the statistical calculation.

The aligned reads were further fed into MACS (v1.4.1) software to identify the regions enriched for DNA methylation (peaks) [36]. The MACS software used the default parameters, and the cutoff *P*-value was $1e-5$. For the histone modification, which has a different peak shape and profile, the software CCAT (v3.0) was chosen to identify the modification-enriched regions (peaks) [37]. The parameters from the original paper were used [38]. The sliding window was 2 000 nt for the H3K27me3 modification and 500 nt for the H3K4me3 and H3K4me2 modifications. Then, the FDR cutoff was 0.05. The peaks that were identified in each individual sample were finally merged together; the read count in each peak was taken for the subsequent statistical comparison.

The RNA-Seq reads were mapped to the mouse genome using the Tophat (v1.3.3) software and the Ensembl genome annotation (Mus_musculus.NCBIM37.64.gtf) [39]. The default parameters were used. The read coverage in the exons, introns, and intergenic regions was calculated. Then, the expression level (FPKM, Fragments Per Kilobase of exon per Million fragments mapped) was estimated by Cufflinks (v1.2.0) software [40].

The raw reads from the sRNA library were processed to remove the adapter sequences, and classified according to the barcodes. sRNAs with lengths of ≥ 18 nt were mapped to the mouse genome, and only perfectly matched reads were extracted for subsequent analysis. The sRNA sequences were then categorized step-by-step using a different annotation dataset, including the miRBase miRNA database (miRBase release 18, <http://www.mirbase.org/>), ncRNA (Ensembl release 64), repeats (<http://genome.ucsc.edu>) and protein coding genes (Ensembl release 64). Reads matching the miRBase mature miRNA were accumulated as the expression abundance of miRNA.

Statistical analysis

Hierarchical clustering was performed to describe the relationship among samples. The Bioconductor “edgeR” or “limma” package was applied for the pairwise sample comparison [41]. Generalized linear models were designed to identify the differences between “4N-OFF” and “4N-ON” iPSC lines. The multiple tests were adjusted using FDR methods, and the cutoff of $FDR \leq 0.05$ was chosen for the significance.

Data access

All the high-throughput sequencing data sets can be accessed as the GEO reference GSE36294. (URL: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=nlapfkwsgeaiho&acc=GSE36294>)

Acknowledgments

We thank our colleague Dr Bing Zhu for his critical comments on the manuscript. We thank Prof Rudolf Jaenisch (the Whitehead Institute of MIT) for generously supplying the lentivirus vectors. We are also grateful to our laboratory colleagues for their assistance with experiments and preparation of the manuscript. This work was supported by the Ministry of Science and Technology (2011CB812700, 2010CB944900, and 2011CB964800 to SG, 2011AA100303 and 2011BAD19B01 to JT, 2011ZX09102-010-04 to TC) and the National Natural Science Foundation of China (31000656 to GC).

References

- 1 Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006; **126**:663-676.
- 2 Chin MH, Mason MJ, Xie W, *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* 2009; **5**:111-123.
- 3 Doi A, Park IH, Wen B, *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* 2009; **41**:1350-1353.
- 4 Guenther MG, Frampton GM, Soldner F, *et al.* Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* 2010; **7**:249-257.
- 5 Kim K, Doi A, Wen B, *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* 2010; **467**:285-290.
- 6 Polo JM, Liu S, Figueroa ME, *et al.* Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat Biotechnol* 2010; **28**:848-855.
- 7 Stadtfeld M, Apostolou E, Akutsu H, *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* 2010; **465**:175-181.
- 8 Carey BW, Markoulaki S, Hanna JH, *et al.* Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell Stem Cell* 2011; **9**:588-598.
- 9 Jiang J, Lv W, Ye X, *et al.* *Zscan4* promotes genomic stability during reprogramming and dramatically improves the quality of iPSC cells as demonstrated by tetraploid complementation. *Cell Res* 2013; **23**:92-106.
- 10 Zhao XY, Li W, Lv Z, *et al.* iPSC cells produce viable mice through tetraploid complementation. *Nature* 2009; **461**:86-90.
- 11 Bock C, Kiskinis E, Verstaepen G, *et al.* Reference maps of human ES and iPSC cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* 2011; **144**:439-452.
- 12 Kang L, Wang J, Zhang Y, Kou Z, Gao S. iPSC cells can support full-term development of tetraploid blastocyst-complemented embryos. *Cell Stem Cell* 2009; **5**:135-138.
- 13 Kou Z, Kang L, Yuan Y, *et al.* Mice cloned from induced pluripotent stem cells (iPSCs). *Biol Reprod* 2010; **83**:238-243.
- 14 Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 2006; **7**:29-59.
- 15 Hikichi T, Kohda T, Wakayama S, Ishino F, Wakayama T. Nuclear transfer alters the DNA methylation status of specific genes in fertilized and parthenogenetically activated mouse embryonic stem cells. *Stem Cells* 2008; **26**:783-788.
- 16 Wang Y, Joh K, Masuko S, *et al.* The mouse *Murr1* gene is imprinted in the adult brain, presumably due to transcriptional interference by the antisense-oriented *U2af1-rs1* gene. *Mol Cell Biol* 2004; **24**:270-279.
- 17 Lin SP, Youngson N, Takada S, *et al.* Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the *Dkl1-Gtl2* imprinted cluster on mouse chromosome 12. *Nat Genet* 2003; **35**:97-102.
- 18 Stadtfeld M, Apostolou E, Ferrari F, *et al.* Ascorbic acid prevents loss of *Dkl1-Dio3* imprinting and facilitates generation of all-iPSC cell mice from terminally differentiated B cells. *Nat Genet* 2012; **44**:398-405, S391-S392.
- 19 Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev* 2002; **16**:6-21.
- 20 Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. *Science* 2001; **293**:1089-1093.
- 21 Hiura H, Toyoda M, Okae H, *et al.* Stability of genomic imprinting in human induced pluripotent stem cells. *BMC Genet* 2013; **14**:32.
- 22 Jiang H, Sun B, Wang W, *et al.* Activation of paternally expressed imprinted genes in newly derived germline-competent mouse parthenogenetic embryonic stem cell lines. *Cell Res* 2007; **17**:792-803.
- 23 Li C, Chen Z, Liu Z, *et al.* Correlation of expression and methylation of imprinted genes with pluripotency of parthenogenetic embryonic stem cells. *Hum Mol Genet* 2009; **18**:2177-2187.
- 24 Hatada I, Sugama T, Mukai T. A new imprinted gene cloned by a methylation-sensitive genome scanning method. *Nucleic Acids Res* 1993; **21**:5577-5582.
- 25 Nabetani A, Hatada I, Morisaki H, Oshimura M, Mukai T. Mouse *U2af1-rs1* is a neomorphic imprinted gene. *Mol Cell Biol* 1997; **17**:789-798.
- 26 Zhang Z, Joh K, Yatsuki H, *et al.* Comparative analyses of genomic imprinting and CpG island-methylation in mouse *Murr1* and human *MURR1* loci revealed a putative imprinting control region in mice. *Gene* 2006; **366**:77-86.
- 27 Ficiz G, Hore TA, Santos F, *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* 2013; **13**:351-359.

- 28 Esteban MA, Wang T, Qin B, *et al.* Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. *Cell Stem Cell* 2010; **6**:71-79.
- 29 Huangfu D, Osafune K, Maehr R, *et al.* Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat Biotechnol* 2008; **26**:1269-1275.
- 30 Marks H, Kalkan T, Menafra R, *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* 2012; **149**:590-604.
- 31 Xie W, Barr CL, Kim A, *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* 2012; **148**:816-831.
- 32 Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* 2012; **150**:402-412.
- 33 Bernstein BE, Mikkelsen TS, Xie X, *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006; **125**:315-326.
- 34 Chang G, Miao YL, Zhang Y, *et al.* Linking incomplete reprogramming to the improved pluripotency of murine embryonal carcinoma cell-derived pluripotent stem cells. *PLoS One* 2010; **5**:e10320.
- 35 Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**:R25.
- 36 Zhang Y, Liu T, Meyer CA, *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**:R137.
- 37 Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. *Nat Immunol* 2011; **12**:918-922.
- 38 Xu H, Handoko L, Wei X, *et al.* A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* 2010; **26**:1199-1204.
- 39 Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009; **25**:1105-1111.
- 40 Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 2011; **27**:2325-2329.
- 41 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**:139-140.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)