Stem Cell Reports, Volume *7*

Supplemental Information

Regulation of the DNA Methylation Landscape in Human Somatic Cell Reprogramming by the miR-29 Family

Eriona Hysolli, Yoshiaki Tanaka, Juan Su, Kun-Yong Kim, Tianyu Zhong, Ralf Janknecht, Xiao-Ling Zhou, Lin Geng, Caihong Qiu, Xinghua Pan, Yong-Wook Jung, Jijun Cheng, Jun Lu, Mei Zhong, Sherman M. Weissman, and In-Hyun Park

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. miR-29 targets *in silico* TET1/2/3, DNMTA/B, and TDG. miR-29 – target alignment through microRNA.org algorithm.

Figure S2. Change in DNA methylation proteins regulated by miR-29a family. (A) *TET1/3* and *DNMT3A/3B* are highly expressed in in pluripotent ESCs (H1, H9), iPSCs (BJ1-iPSC, PGP9-iPSC) and embryonic carcinoma cell line (NCCIT), but not in fibroblasts (D551 = Detroit 551 fibroblasts) (n=3, independent experiments). (B) Luciferase reporter with 3'UTR of *TET1* shows reduced activity after miR-29a overexpression (retrovirus construct) in 293T cells (n=3, independent experiments). (C) *TET1/3*, *DNMT3A/3B* and *TDG* are induced during reprogramming (n=3, independent experiments). (D) Dot Blot with serial dilution of gDNA extracted from D551 fibroblasts 3 days post antagomir (250nM) or mimic (50nM), and subsequent incubation with DNA modification antibodies shows increase and decrease of 5hmC, 5fC, 5caC, respectively. (E) Relative expression level of miR-29 family to U6 small RNA in fibroblast. (n=3, technical replicates). (F) qPCR confirmation of miR-29a downregulation by miR-29a inhibitor. (n=3, technical replicates). (G-H) Expression differences of pluripotent genes in miR-29a depleted fibroblasts by (G) RNA-Seq data and (H) qPCR. (n=3, independent experiments). Error bars (S2A-C, E, F, and H) represent standard deviation. (I) DNA methylation status in miR-29a depleted fibroblast in *KLF4* locus. (J) The cumulative distribution of RNA and protein fold change upon miR-29a inhibition. Genes are classified by the number of miR-29a target sites. The classification was also performed to the according to three types of target sites (8mer, 7mer-m8 and 7mer-1a). * $p<0.05$ by one-side T test.

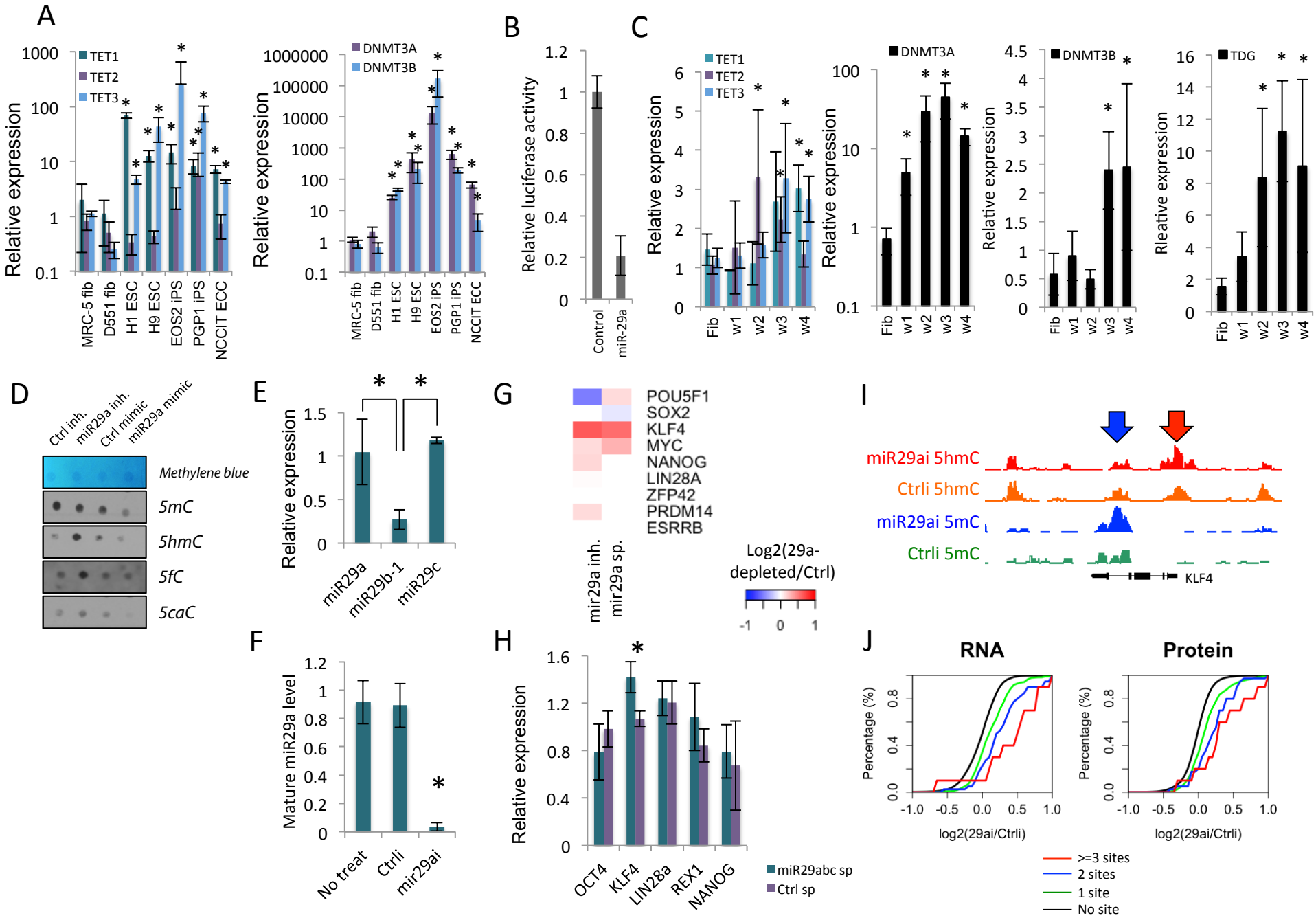
Figure S3. Change in DNA methylation by mir-29a depletion in fibroblast.
(A) Venn diagrams show overlap of hyper- and hypo-DMRs of 5hmC between miR-29a depleted cells and human ESCs. (B) DNA methylation patterns around reprogramming factor binding sites. Log2(5hmC/5mC) in Ctrli (black), 29ai (red) and ESC (green) over OCT4, SOX2, KLF4 and MYC binding sites at 2 days after reprogramming induction (Soufi et al., 2012) and ESCs (Lister et al., 2009) were plotted from -5kbp to 5kbp relative to the peak centers. (C) ChIP-qPCR of SOX2 in miR-29a-depleted and control fibroblast cells with SOX2 transgene. * $p<0.05$ by

one-side T test. n=3, technical replicates. Error bars represent standard deviation. (D) GO analysis for genes regulated by common 5mC hypo-DMRs in miR-29a-depleted firoblast and ESC. Dashed line represents 0.05 FDR. (E) Bisulfite sequencing of *NANOG* promoter in miR-29ai and Ctrli fibroblast cells. (F) Comparison of DMRs in fibroblasts after infection with DNA methylases, demethylases, pluripotency factors, and miR-29abc sponge assessed by Illumina's 450K array. Red: methylation. Blue: demethylation. Heat color represents difference of beta value between miR-29a/b/c and control sponge or between protein-overexpressed cells and empty vector-integrated cells. (G) Overrepresented GO terms in differentially-methylated or -demethylated regions by overexpression of DNMT and TET proteins. * $p<0.05$ by one-side T test.

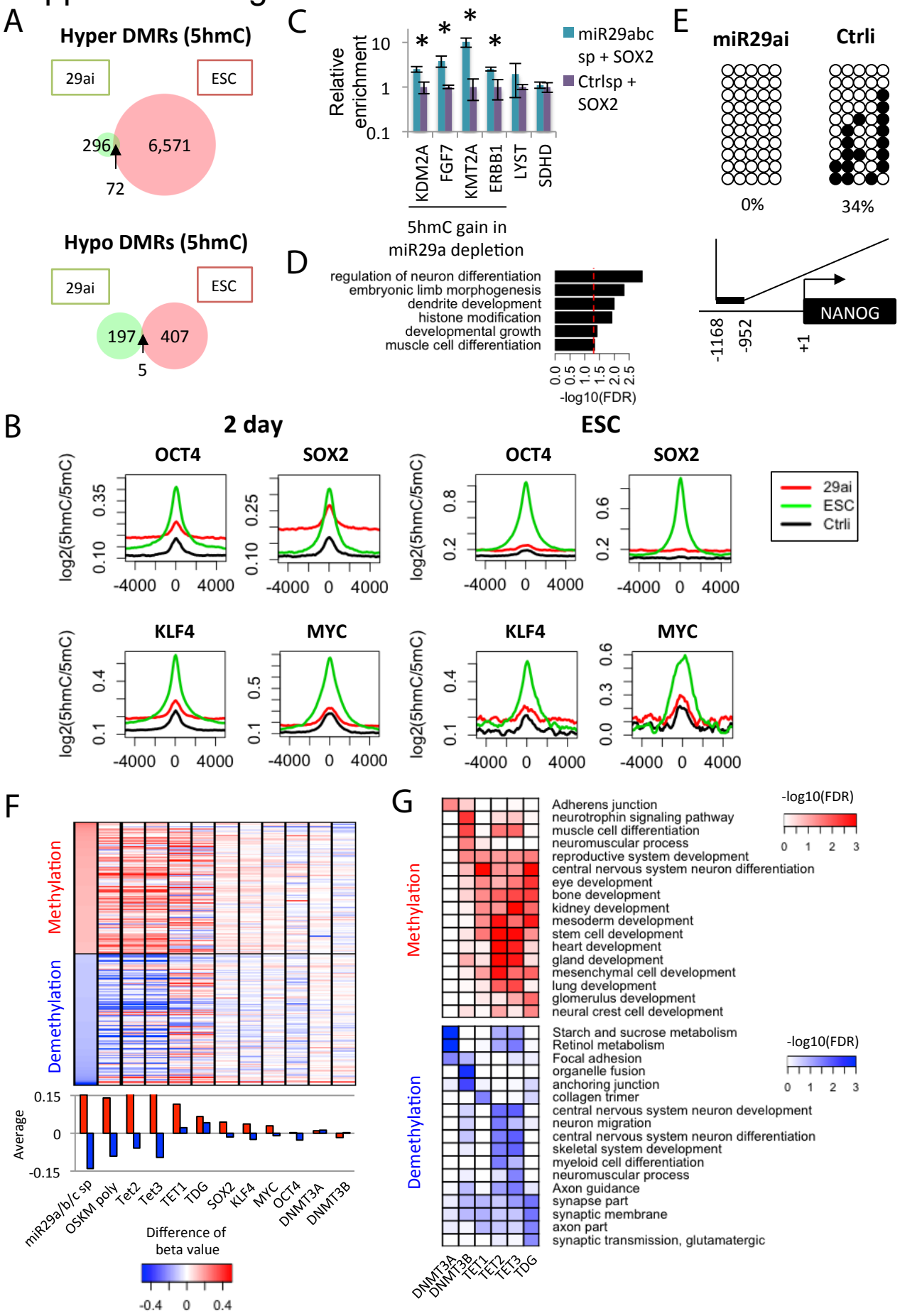
Figure S4. miR-29a targets modulate reprogramming efficiency and aberrantly expressed ESC-specific genes. (A) Validation of AP staining method with true pluripotency TRA-160 surface marker staining. (B) qPCR confirmation of ectopic expression of *DNMTs*, *TETs* and *TDG*, and downregulation of *p53* by short hairpin RNA (n=3, technical replicates). (C) Overexpression of TET1 increases reprogramming efficiency in human fibroblasts (n=3, independent experiments). (D) DNMT3B increases reprogramming efficiency (2 different fibroblast lines and FM-1, n=7, independent experiments). (E) TP53 knockdown counteracts effect of miR-29a overexpression in reprogramming (FM-1, n=3, independent experiments). (F) qPCR of ESC-specific genes *TCERG1L*, *FAM19A5*, and *TMEM132D* in ESCs (H1, H7, H9), iPSCs (EOS2, PGP1, MRC-5, D551), D551 and FM-1 control derived iPSCs, and D551 and FM-1 miR-29a depleted iPSCs. (n=3, technical replicates). (G) qPCR of iPSC-specific genes *ZNF454* and *ZNF 572* reveals their high levels across all iPSCs. (n=3, technical replicates). Error bars (S4B-G) represent standard deviation. (H) Model of miR-29a-mediated regulation of epigenetic landscape during reprogramming. miR-29a family maintains the cellular epigenetic states by suppressing proteins for DNA methylation. Suppression of miR-29a family members by reprogramming factors results in the global DNA methylation change in somatic cells, leading to ES-like DNA methylation state in iPSCs. Blue circle represents methylated DNA.

Supplemental Figure 1

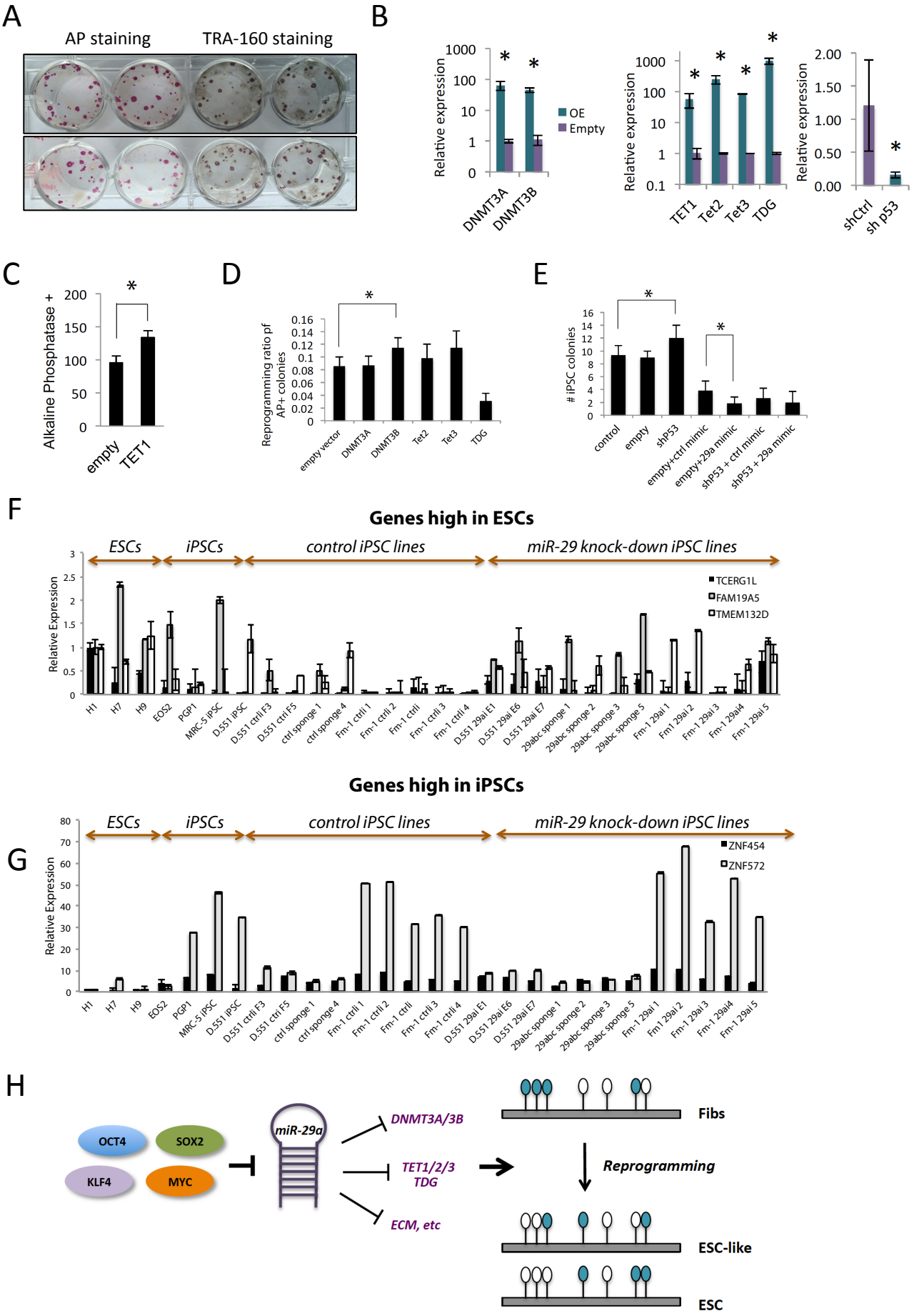
Supplemental Figure 2



Supplemental Figure 3



Supplemental Figure 4



SUPPLEMENTAL TABLES

Supplemental Table 1. miRNAs that suppress the luciferase report for 3'UTR of TET1

Supplemental Table 2. Genes up- or downregulated by miR-29a depletion by miR-29a antagomir.

Supplemental Table 3. Proteins up- or downregulated by miR-29a depletion by miR-29a antagomir.

Supplemental Table 4. S4A. List of primers used for RT-qPCR, ChIP-qPCR, and bi-sulfite sequencing.
S4B. Total and mapped reads in RNA-seq, (h)MeDIP-seq, and ChIP-seq.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Reprogramming

Reprogramming experiments were carried out using the four human transcription factors OCT4, SOX2, KLF4, MYC in pMSCV retrovirus backbone as previously described (Park et al., 2008). Briefly VSV-G peudotyped retrovirus was generated in 293T cells using GAG-POL, VSV-G and pMIG vector expressing each for four factors. The virus was collected, filtered in 0.45 um filtering system and centrifuged for 1.5 hours at 23,000 rpm in a Beckman L80 ultracentrifuge. Detroit 551 human fibroblasts cells (ATCC CCL-110) were transduced at multiplicity of infection (MOI) 5. Cells were passaged into 0.2% gelatin-coated plates containing irradiated or mitomycin-C treated mouse embryonic feeder cells (Millipore, EmbryoMax CF1 strain cat# PMEF-CF). The medium was switched from 10% FBS in DMEM into human embryonic stem cell medium containing KSR and bFGF in DMEM/F12 on day 5, and cells were maintained in this medium for 3-4 weeks until iPSC colonies formed. Lentivirus was generated using the packaging vectors MDL, REV and VSV-G.

Plasmids

Full length TET1 wild-type and mutant (Anjana Rao) were subcloned from pEF-HA-FLAG TET1 into the pLenti 7.3 EGFP/V5 backbone using the gateway system (Tahiliani et al., 2009). Full length DNMT3A and Dnmt3b were cloned into pMIG retroviral vector after PCR amplification. TET1 shRNA and scramble control sequences in PLKO.1-puro plasmids were purchased from Sigma Aldrich Mission shRNA library. pMSCV-control and –miR-29abc sponge constructs were generated as described before (Cheng et al., 2013). 3'UTR of TET1 was cloned into the luciferase containing plasmid psi-CHECK2.

Transfection with miRNA mimic, antagomir, and sponge

Following miR-29a mimic and antagomir were purchased from Dharmacon and transfected into fibroblasts using Invitrogen's RNAi MAX or Lipofectamine 2000 (cat # 13778-150, 11668-019) at 100 – 250nM concentration: miRIDIAN mimic negative control #1 cat# CN-001000-01-05
Dharmacon miRIDIAN mimic human hsa-miR-29a-3p cat# C-300504-07
Dharmacon miRIDIAN hairpin inhibitor negative control cat# IN-001005-01-20
Dharmacon miRIDIAN hairpin inhibitor hsa-miR-29a cat# IH-300504-08

To test the effect of miR-29a mimic and antagomir on reprogramming, cells were infected with retroviral vectors expressing reprogramming factors two days post transfection.

Chromatin Immunoprecipitation

Detroit 551 fibroblasts were infected with pMIG-SOX2 and pMSCV-miR-29abc sponge by FuGENE HD transfection reagent with manufacturer's protocol (Promega, cat # E2311). Three days later cells were fixed with 1% formaldehyde final concentration and quenched with glycine. ChIP experiments were performed using manufacturer's guidelines (Cell Signaling, SimpleChIP cat# 8980S). In brief, three days after infection, cells were cross-linked with 1% formaldehyde for 10 min, and cross-linking was terminated with glycine. After harvesting cells, nuclei were isolated and treated with micrococcal nuclease. After sonication to break the nuclear membrane ChIP was performed using SOX2 (Cell Signaling, cat #5024). %input values in miR-29a-depleted cells were normalized to those in control cells.

qPCR

Total RNA was extracted using TRIzol and miRNeasy Mini Kit (Qiagen cat# 217004). For mRNA reverse transcription (RT) reactions Bio-Rad iScript cDNA synthesis kit (cat# 170-8890) with random primers was used according to manufacturer's instructions. For miRNA synthesis, miScript II RT kit (Qiagen cat# 218161) was used. Immunoprecipitated DNA was used to quantify the enrichments using qPCR. qPCR reactions were performed with iQ SYBR Green Super Mix (Bio-Rad cat# 170-8880) in a Real-Time CFX96 thermal cycler machine. The list of primers for qPCR is available in Table S4A.

Dot Blot

Genomic DNA (gDNA) was extracted from cells using Qiagen's DNeasy Blood and Tissue Kit (cat# 69506), eluted in AE buffer, and concentration determined using Thermo Scientific NanoDrop 2000 Spectrophotometer (Model#: ND-2000). gDNA was denatured at 100 °C for 10 min, then spotted in 2ul aliquots in Amersham Hybond N+ nylon membrane cutouts (GE Life Sciences Cat# RPN119B), or vacuumed through a Bio-Dot apparatus (Bio-Rad 170-6545). The membrane was washed with 6X SSC buffer prior to blotting, and 2X SSC buffer after sample loading. DNA was crosslinked to membrane using a UV crosslinker (Fisher Scientific), then

blocked for 30 min in 5% milk in 1X TBS buffer. Rabbit 5-hmC antibody (Active Motif cat# 39769) was used at 1:5000 – 1:10000 concentration overnight at 4°C. Rabbit 5-caC and 5-fC were also from Active Motif (cat # 61225 and 61227).

Proteomics analysis and Data Processing

Detroit 551 cells in10 cm dish were transfected with 250 nM control or miR-29a inhibitor, and collected three days after transfection. Differentially expressed proteins were determined by performing peptide identification at the Taplin Biological Mass Spectrometry Facility at Harvard Medical School. Three biological replicates were performed. Cells were lysed and protein extracted and precipitated followed by digestion. For multiplexing, samples were labeled with TMT 126-131, then combined, fractionated, and submitted to LC-MS3. Labeling efficiency was between 80-99%. Total amount of peptide extracted was between 180-200ug. Benjamini-Hochberg (BH) corrected p-values were used to determine proteins with significant changes; k-means clustering was used to separate the proteins with significant changes (BH p-value < 0.05). Signal-to-noise ratio (the ratio of mean to standard deviation from three replicates) was then calculated in each protein. Proteins with less than 20 signal-to-noise ratio in both 29ai and Ctrli sample were retained for subsequent analysis. Target genes and target site categories of miR-29a were obtained from TargetScan (Release 6.2). Genes were classified by the count of conserved target sites. Average protein expression changes of three replicates were then compared to RNA expression changes, which were calculated from RNA-seq.

RNA-Sequencing and Data Processing

1ug of Total RNA from 1ug ESCs (H1 and H9), iPSCs (D551-iPSCs and FM-1-iPSC), D551 fibroblasts transfected with control and miR-29a antagomir, and infected with control or miR-29abc sponge were submitted for sequencing using Illumina's Hi-Seq 2000 Sequencer. Library preparation and sequencing were carried out at Yale Sequencing Facility using Illumina's instructions. Genome sequence and genomic coordinate of RefSeq genes and CGIs in human (hg19) were obtained from UCSC database. RNA-seq reads were aligned to the human genome by Tophat (v2.0.12) with default option (Trapnell et al., 2009), resulting in more than 90% of mapping ratio. Gene expression values (RPKM value) were calculated by Cufflinks (v1.2.1) by using RefSeq genes as reference annotation (Trapnell et al., 2010). Gene ontology analysis was

performed by GOstats in Bioconductor packages. Multiple test correction was performed by Benjamini-Hochberg (BH) method. Significant GO terms were obtained by FDR threshold 0.05.

(h)MeDIP-sequencing and Data Processing

gDNA was isolated from ESCs (H1 and H9), D551-iPSCs and D551 fibroblasts transfected with control and miR-29a antagomir and processed for MeDIP and hMeDIP (Taiwo et al., 2012). 2 ug of gDNA was sonicated,and IP was performed using antibodies for 5hmC (EF-118-0032, Diagenode), or 5mC (AIP-206-025, Diagenode). For sequencing, library was generated by Illumina Truseq DNA LT Sample Prep kit and sequenced by HiSeq2000. (h)MeDIP-seq reads were mapped to the human genome by Bowtie2 (v2.1.0) with the parameter "--local -D 15 -R 3 -N 1 -L 20 -i S,1,0.50 -k 1" (Langmead and Salzberg, 2012). Differentially-methylated regions of 5hmC and 5mC were identified by MACS2 (v2.1) peak caller with the option "--bw=200 -g hs –q 0.05 --broad" (Feng et al., 2012). Differentially-hypermethylated regions (hyper-DMRs) in 29ai fibroblasts were detected by running MACS2 using 29ai and Ctrli as treatment and control, respectively. In contrast, differentially-hypomethylated regions (hypo-DMRs) were obtained using 29ai and Ctrli as control and treatment, respectively. We also identified hyper- and hypo-DMRs in ESCs by comparing public H9 ESC (h)MeDIP-seq data (SRX189181 and SRX189182) to our Ctrli datasets (Gao et al., 2013). In addition, 5hmC and 5mC-enriched regions in 29ai and Ctrli fibroblast were also identified by MACS2 using input DNA as control. Genes within 3k bp from 5(h)mC-enriched regions were defined as target genes of DNA methylation.

Binding sites of OCT4, SOX2, KLF4 and MYC at 2 days after reprogramming induction and ESC were defined from public ChIP-seq data (SRP011557, SRP001251 and SRP001264) (Lister et al., 2009; Soufi et al., 2012). Peak calling was performed by MACS2 with the option "--bw=200 -g hs –q 0.05". Numbers for total reads, mapped reads, and mapping percentages from RNA-seq, (h)MeDIP-seq and ChIP-seq are provided in Table S4B.

Illumina 450K sequencing and Data Processing

gDNA from D551 fibroblasts transduced for 3 days with retrovirus pMIG-DNMT3A/3B, lentivirus pLenti/V5/DEST TET1, retrovirus pMSCV mTet2/mTet3, lentivirus pRVYtet-TDG, retrovirus pMIG polycistronic OSKM or individually pMIG OCT4, SOX2, KLF4, AND MYC, and pMSCV-miR-29abc sponge, was harvested with PureLink Kit from Life Technologies (cat#

K1820-02) and prepared for sequencing by the Yale Center for Genome Analysis according to manufacturer's protocol using the Illumina's Infinium HumanMethylation450 BeadChip kit (cat# WG-314-1003). IDAT files were preprocessed and converted into beta value by minfi Bioconductor package. Beta value in 29abc sponge was then subtracted from that in Ctrl sponge to measure DNA methylation change. Beta values in DNMTs, TETs, TDG and OSKM-transduced cells were subtracted from that in empty vector-transduced cells. Top and bottom 500 DNA methylation changes in 29abc sponge were used as differentially-methylated regions.

Clustering of DNA methylation profiles of ESCs, iPSCs and sponge-treated iPSCs.

In each library, the number of sequence reads within TSS±3Kbp, gene body and CpGI shore and CpGI was counted and normalized to the total number of mapped reads and the region size (RPKM value). Here, the CpGI shore was defined by 3Kbp flanking region of CpGI. Hierarchical clustering was then performed by heatmap.2 function in R. Differentially methylated regions (DMRs) between ESC/29ai and iPSC/Ctrli were identified by more than 1.25 fold change and one-side T test p-value cutoff 0.05. DMRs with less than 0.1 average RPKM are removed from subsequent analysis. Genes, which are nearest to differentially methylated CpGI shores, were used for GO analysis.

Bisulfite-sequencing

0.5ug of genomic DNA was used for bisulfite treatment by EZ DNA Methylation-Lighting^{TM} kit with manufacturer's protocol (ZYMO RESEARCH cat # D5030). Converted DNA was then amplified by PCR with Quick-Load *Taq* 2X Master Mix (NEB cat # M0271L) and bisulfite-conversion-based primers (Table S4A). Amplified DNA was purified by 2% agarose gel and cloned into pCR2.1-TOPO vector. Subsequently, cytosine conversion and retention were detected by Sanger sequencing using M13 forward primers.

SUPPLEMENTAL REFERENCES

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol *11*, R90.

Cheng, J., Guo, S., Chen, S., Mastriano, S.J., Liu, C., D'Alessio, A.C., Hysolli, E., Guo, Y., Yao, H., Megyola, C.M.*, et al.* (2013). An Extensive Network of TET2-Targeting MicroRNAs Regulates Malignant Hematopoiesis. Cell Rep.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728-1740.

Gao, F., Xia, Y., Wang, J., Luo, H., Gao, Z., Han, X., Zhang, J., Huang, X., Yao, Y., Lu, H.*, et al.* (2013). Integrated detection of both 5-mC and 5-hmC by high-throughput tag sequencing technology highlights methylation reprogramming of bivalent genes during cellular differentiation. Epigenetics *8*.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Park, I.H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., and Daley, G.Q. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*, 141-146.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994-1004.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L.*, et al.* (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930-935.

Taiwo, O., Wilson, G.A., Morris, T., Seisenberger, S., Reik, W., Pearce, D., Beck, S., and Butcher, L.M. (2012). Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc *7*, 617-636.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.